MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Asymptotically Constant-Risk Predictive Densities When the Distributions of Data and Target Variables are Different

Keisuke YANO and Fumiyasu KOMAKI

METR 2014–09 March 2014

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

The METR technical reports are published as a means to ensure timely dissemination of scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the authors or by other copyright holders, notwithstanding that they have offered their works here electronically. It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may not be reposted without the explicit permission of the copyright holder.

Asymptotically Constant-Risk Predictive Densities When the Distributions of Data and Target Variables are Different

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 investigate asymptotic construction of constant-risk Bayesian predictive densities under the Kullback–Leibler risk when the distributions of data and target variables are different and have a common unknown parameter. It is known that the Kullback–Leibler risk is asymptotically equal to a trace of the product of two matrices: the inverse of the Fisher information matrix for the data, and the Fisher information matrix for the target variables. We assume that the trace has an unique maximum point with respect to the parameter. We construct asymptotically constant-risk Bayesian predictive densities using a prior depending on the sample size. Further, we apply the theory to the subminimax estimator problem and the prediction based on the binary regression model.

1 Introduction

Let $x^{(N)} = (x_1, \cdots, x_N)$ be independent N data distributed according to a probability density $p(x|\theta)$ that belongs to a d-dimensional parametric model $\{p(x|\theta) : \theta \in \Theta\}$, where $\theta = (\theta^1, \cdots, \theta^d)$ is an unknown d-dimensional parameter, and Θ is the parameter space. Let y be a target variable distributed according to a probability density $q(y|\theta)$ that belongs to a d-dimensional parametric model $\{q(y|\theta) : \theta \in \Theta\}$ with the same parameter θ. Here, we assume that the distributions of the data and the target variables $p(x|\theta)$ and $q(y|\theta)$ are different. For simplicity, we assume that the data and the target variables are independent given by θ.

We construct predictive densities for target variables based on the data. We measure the performance of the predictive density $\hat{q}(y|x^{(N)})$ by the Kullback–Leibler divergence $D(q(\cdot|\theta), \hat{q}(\cdot|x^{(N)}))$ from the true density $q(y|\theta)$ to the predictive density $\hat{q}(y|x^{(N)})$:

$$D(q(\cdot|\theta), \hat{q}(\cdot|x^{(N)})) = \int q(y|\theta) \log \frac{q(y|\theta)}{\hat{q}(y|x^{(N)})} \mathrm{d}y.$$

Then, the risk function $R(\theta, \hat{q}(y|x^{(N)}))$ of the predictive density $\hat{q}(y|x^{(N)})$ is given by

$$\begin{aligned} R(\theta, \hat{q}(y|x^{(N)})) &= \int p(x^{(N)}|\theta) D(q(\cdot|\theta), \hat{q}(\cdot|x^{(N)})) \mathrm{d}x^{(N)} \\ &= \int p(x^{(N)}|\theta) \int q(y|\theta) \log \frac{q(y|\theta)}{\hat{q}(y|x^{(N)})} \mathrm{d}y \mathrm{d}x^{(N)}. \end{aligned}$$

For the construction of predictive densities, we consider the Bayesian predictive density defined by

$$q_\pi(y|x^{(N)}) = \frac{\int q(y|\theta) p(x^{(N)}|\theta) \pi(\theta; N) \mathrm{d}\theta}{\int p(x^{(N)}|\theta) \pi(\theta; N) \mathrm{d}\theta},$$

where $\pi(\theta; N)$ is a prior density for θ possibly depending on the sample size N. Aitchison (1975) showed that, for a given prior density $\pi(\theta; N)$, the Bayesian predictive density $q_\pi(y|x^{(N)})$ is a Bayes solution under

the Kullback–Leibler risk. Based on the asymptotics as the sample size goes to infinity, Komaki (1996) and Hartigan (1998) showed its superiority over any plug-in predictive density $q(y|\hat{\theta})$ with any estimator $\hat{\theta}$. However, there remains a problem of prior selection for constructing better Bayesian predictive densities. Thus, a prior $\pi(\theta; N)$ must be chosen based on an optimality criterion for actual applications.

Among various criteria, we focus on a criterion of constructing minimax predictive densities under the Kullback–Leibler risk. For simplicity, we refer to the priors generating minimax predictive densities as minimax priors. Minimax priors have been previously studied in various predictive settings; see Bernardo (1979), Clarke and Barron (1994), Aslan (2006), and Komaki (2011, 2012). Except for Komaki (2011), these studies are based on the assumption that the distributions $p(x|\theta)$ and $q(y|\theta)$ are identical. Let us consider the prediction based on the logistic regression model where the covariates of the data and the target variables are not identical. In this predictive setting, the assumption that the distributions $p(x|\theta)$ and $q(y|\theta)$ are identical is no longer valid.

We focus on the minimax priors in predictions where the distributions $p(x|\theta)$ and $q(y|\theta)$ are different and have a common unknown parameter. Such a predictive setting has traditionally been considered in the statistical prediction and the experiment design. It has recently been studied in the statistical learning theory (e.g., Kanamori and Shimodaira, 2003). Predictive densities where the distributions $p(x|\theta)$ and $q(y|\theta)$ are different and have a common unknown parameter are studied by Shimodaira (2000), Fushiki, Komaki, and Aihara (2004), Suzuki and Komaki (2010), and Komaki (2013).

Let $g^X_{ij}(\theta)$ be the (i,j)-component of the Fisher information matrix of the distribution $p(x|\theta)$, and let $g^Y_{ij}(\theta)$ be the (i,j)-component of the Fisher information matrix of the distribution $q(y|\theta)$. Let $g^{X,ij}(\theta)$ and $g^{Y,ij}(\theta)$ denote the (i,j)-components of their inverse matrices. We adopt the Einstein's summation convention: if the same indices appear twice in any one term, it implies summation over that index from 1 to d. For the asymptotics below, we assume that the prior densities $\pi(\theta; N)$ are smooth.

On the asymptotics as the sample size N goes to infinity, we construct asymptotically constant-risk priors $\pi(\theta; N)$ in the sense that the asymptotic risk

$$R(\theta, q_\pi(y|x^{(N)})) = \frac{1}{N} R_1(\theta, q_\pi(y|x^{(N)})) + \frac{1}{N\sqrt{N}} R_2(\theta, q_\pi(y|x^{(N)})) + O(N^{-2})$$

is constant up to $\mathrm{O}(N^{-3/2})$. Since the proper prior with the constant risk is a minimax prior for any finite sample size, the asymptotically constant-risk prior relates to the minimax prior; in Section 4, we verify that the asymptotically constant-risk prior agrees with the exact minimax prior in binomial examples.

It is known that the N^{-1}-order term $R_1(\theta, q_\pi(y|x^{(N)}))$ of the Kullback–Leibler risk is equal to the trace $g^{X,ij}(\theta) g^Y_{ij}(\theta)$. If the trace does not depend on the parameter θ, the construction of asymptotically constant-risk priors is parallel to Aslan (2006); see also Komaki (2013).

However, we consider the setting where there exists a unique maximum point of the trace $g^{X,ij}(\theta) g^Y_{ij}(\theta)$; for example, this setting appears in predictions based on the binary regression model where the covariates of the data and the target variables are not identical. In this setting, there does not exist asymptotically constant-risk priors among the priors independent of the sample size N. The reason is as follows: we consider the proper priors $\pi(\theta)$. Although the asymptotic Bayes risk $\int (1/N) g^{X,ij}(\theta) g^Y_{ij}(\theta) \pi(\theta) \mathrm{d}\theta$ is maximized by the discrete priors, the discrete priors violate the smoothness condition of the priors for the asymptotics.

When there exists a unique maximum point of the trace $g^{X,ij}(\theta) g^Y_{ij}(\theta)$, we construct the asymptotically constant-risk prior $\pi(\theta; N)$ up to $\mathrm{O}(N^{-2})$ by making the prior dependent on the sample size N as

$$\frac{\pi(\theta; N)}{|g^X(\theta)|^{1/2}} \quad \propto \quad \{f(\theta)\}^{\sqrt{N}} h(\theta),$$

where $f(\theta)$ and $h(\theta)$ are the scalar functions of θ independent of N, and $|g^X(\theta)|$ denotes the determinant of the Fisher information matrix $g^X(\theta)$.

The key idea is that, if the specified parameter point has the more undue risk than the other parameter points, then the more prior weights should be concentrated on that point.

Further, we clarify the subminimax estimator problem based on the mean squared error from the viewpoint of the prediction where the distributions of data and target variables are different and have

a common unknown parameter. We obtain the improvement achieved by the minimax estimator over the subminimax estimators up to $\mathrm{O}(N^{-2})$. The subminimax estimator problem (Hodges and Lehmann, 1950; Ghosh, 1964) is the problem that, at first glance, there seems to exist asymptotically dominating estimators of the minimax estimator. However, any relationship between such subminimax estimator problems and predictions have not been investigated, and further, in general, the improvement by the minimax estimator over the subminimax estimators have not been investigated.

2 Information Geometrical Notations

In this section, we prepare the information geometrical notations; see Amari (1985) for details. We abbreviate $\partial/\partial\theta^i$ to ∂_i where the indices $i, j, k, \ldots$ run from 1 to d. Similarly, we abbreviate $\partial^2/\partial\theta^i\partial\theta^j$, $\partial^3/\partial\theta^i\partial\theta^j\partial\theta^k$, and $\partial^4/\partial\theta^i\partial\theta^j\partial\theta^k\partial\theta^l$ to ∂_{ij}, ∂_{ijk}, and ∂_{ijkl}, respectively. We denote the expectations of the random variables X, Y, and $X^{(N)}$ by $\mathrm{E}_X[\cdot]$, $\mathrm{E}_Y[\cdot]$, and $\mathrm{E}_{X^{(N)}}[\cdot]$, respectively. We denote their probability densities by $p(x|\theta)$, $q(y|\theta)$, and $p(x^{(N)}|\theta)$, respectively.

We define the predictive metric proposed by Komaki (2013) as

$$\mathring{g}_{ij}(\theta) \quad = \quad g^X_{ik}(\theta) g^{Y,kl}(\theta) g^X_{lj}(\theta).$$

When the parameter is one-dimensional, $g_{\theta\theta}(\theta)$ denotes Fisher information, and $g^{\theta\theta}(\theta)$ denotes its inverse. Let $\overset{\mathrm{e}}{\Gamma}{}^X_{ij,k}(\theta)$ and $\overset{\mathrm{m}}{\Gamma}{}^X_{ij,k}(\theta)$ be the quantities given by

$$\overset{\mathrm{e}}{\Gamma}{}^X_{ij,k}(\theta) \quad := \quad \mathrm{E}_X[\partial_{ij} \log p(x|\theta) \partial_k \log p(x|\theta)]$$

and

$$\overset{\mathrm{m}}{\Gamma}{}^X_{ij,k}(\theta) \quad := \quad \int \frac{1}{p(x|\theta)} [\partial_{ij} p(x|\theta) \partial_k p(x|\theta)] \mathrm{d}x.$$

Using these quantities, the e-connection and m-connection coefficients with respect to the parameter θ for the model $\{p(x|\theta) : \theta \in \Theta\}$ are given by

$$\overset{\mathrm{e}}{\Gamma}{}^{X,k}_{ij}(\theta) \quad := \quad g^{X,lk}(\theta) \overset{\mathrm{e}}{\Gamma}{}^X_{ij,l}(\theta)$$

and

$$\overset{\mathrm{m}}{\Gamma}{}^{X,k}_{ij}(\theta) \quad := \quad g^{X,kl}(\theta) \overset{\mathrm{m}}{\Gamma}{}^X_{ij,l}(\theta),$$

respectively.

The (0, 3)-tensor $T^X_{ijk}(\theta)$ is defined by

$$T^X_{ijk}(\theta) \quad := \quad \mathrm{E}_X[\partial_i \log p(x|\theta) \partial_j \log p(x|\theta) \partial_k \log p(x|\theta)].$$

The tensor $T^X_{ijk}(\theta)$ also produces a (0, 1)-tensor

$$T^X_i(\theta) \quad := \quad T^X_{ijk}(\theta) g^{X,jk}(\theta).$$

In the same manner, the information geometrical quantities $\overset{\mathrm{e}}{\Gamma}{}^Y_{ij,k}(\theta)$, $\overset{\mathrm{m}}{\Gamma}{}^Y_{ij,k}(\theta)$, and $T^Y_{ijk}(\theta)$ are defined for the model $\{q(y|\theta) : \theta \in \Theta\}$.

Let $M^k_{ij}(\theta)$ be a (1, 2)-tensor defined by

$$M^k_{ij}(\theta) := \overset{\mathrm{m}}{\Gamma}{}^{Y,k}_{ij}(\theta) - \overset{\mathrm{m}}{\Gamma}{}^{X,k}_{ij}(\theta).$$

For a derivative $(\partial_1 v(\theta), \cdots, \partial_d v(\theta))$ of the scalar function $v(\theta)$, the e-covariant derivative is given by

$$\overset{\mathrm{e}}{\nabla}_i v_j(\theta) \quad := \quad \partial_{ij} v(\theta) - \overset{\mathrm{e}}{\Gamma}{}^{X,k}_{ij}(\theta) v_k(\theta).$$

3 Asymptotically constant-risk priors when the distributions of data and target variables are different

In this section, we consider the setting where the trace $g^{X,ij}(\theta)g^Y_{ij}(\theta)$ has an unique maximum point. We construct asymptotically constant-risk priors under the Kullback–Leibler risk in the sense that the asymptotic risk up to $\mathrm{O}(N^{-2})$ is constant. We find asymptotically constant-risk priors up to $\mathrm{O}(N^{-2})$ in two steps: first, expand the Kullback–Leibler risk of Bayesian predictive densities. Second, find the prior having an asymptotically constant risk using this expansion.

From now on, we assume following two conditions for the prior $\pi(\theta; N)$:

(C1) The prior $\pi(\theta; N)$ has the form

$$\frac{\pi(\theta; N)}{|g^X(\theta)|^{1/2}} \quad \propto \quad \exp\{\sqrt{N}\log f(\theta) + \log h(\theta)\},$$

where $f(\theta)$ and $h(\theta)$ are smooth scalar functions of θ independent of N.

(C2) The unique maximum point of the scalar function $f(\theta)$ is equal to the unique maximum point of the trace $g^{X,ij}(\theta)g^Y_{ij}(\theta)$.

Based on conditions (C1) and (C2), we expand the Kullback–Leibler risk of a Bayesian predictive density up to $\mathrm{O}(N^{-2})$.

Theorem 3.1. *The Kullback–Leibler risk of a Bayesian predictive density based on the prior $\pi(\theta; N)$ satisfying condition (C1) is expanded as*

$$\begin{aligned}
&R(\theta, q_\pi(y|x^{(N)})) \\
&= \frac{1}{2N} g^Y_{ij}(\theta) g^{X,ij}(\theta) + \frac{1}{2N}\mathring{g}^{ij}(\theta)\partial_i \log f(\theta)\partial_j \log f(\theta) - \frac{1}{N\sqrt{N}} T^Y_{ijk}(\theta) g^{X,ij}(\theta) g^{X,kl}(\theta)\partial_l \log f(\theta) \\
&\quad + \frac{1}{N\sqrt{N}}\mathring{g}^{ij}(\theta)\overset{\mathrm{e}}{\nabla}_i \partial_j \log f(\theta) + \frac{1}{N\sqrt{N}}\mathring{g}^{ij}(\theta) g^{X,kl}(\theta)\left\{\overset{\mathrm{e}}{\nabla}_i \partial_k \log f(\theta)\right\}\partial_j \log f(\theta)\partial_l \log f(\theta) \\
&\quad - \frac{1}{3N\sqrt{N}} T^Y_{ijk}(\theta) g^{X,is}(\theta) g^{X,jt}(\theta) g^{X,ku}(\theta)\partial_s \log f(\theta)\partial_t \log f(\theta)\partial_u \log f(\theta) \\
&\quad + \frac{1}{2N\sqrt{N}} g^Y_{kl}(\theta) M^l_{ij}(\theta) g^{X,is}(\theta) g^{X,jt}(\theta) g^{X,ku}(\theta)\partial_s \log f(\theta)\partial_t \log f(\theta)\partial_u \log f(\theta) \\
&\quad + \frac{1}{2N\sqrt{N}} g^{X,ij}(\theta) g^Y_{kl}(\theta) g^{X,kl}(\theta) M^m_{ij}(\theta)\partial_m \log f(\theta) + \frac{1}{N\sqrt{N}}\mathring{g}^{ij}(\theta) M^k_{ij}(\theta)\partial_k \log f(\theta) \\
&\quad + \frac{1}{2N\sqrt{N}}\mathring{g}^{ij}(\theta) T^X_i(\theta)\partial_j \log f(\theta) + \frac{1}{2N\sqrt{N}} g^{X,im}(\theta) g^Y_{ij}(\theta) g^{X,kl}(\theta) M^j_{kl}(\theta)\partial_m \log f(\theta) \\
&\quad + \frac{1}{N\sqrt{N}}\mathring{g}^{ij}(\theta)\partial_i \log f(\theta)\partial_j \log h(\theta) + \mathrm{O}(N^{-2}).
\end{aligned} \tag{1}$$

The proof is given in the Appendix.

Remark 3.1. For the subsequent theorem, it is important that at the point θ_f maximizing the scalar function $\log f(\theta)$, $R(\theta_f, q_\pi(y|x^N))$ is given by

$$\begin{aligned}
&R(\theta_f, q_\pi(y|x^N)) \\
&= \frac{1}{2N}\sup_{\theta\in\Theta}\{g^{X,ij}(\theta) g^Y_{ij}(\theta)\} + \frac{1}{N\sqrt{N}}\mathring{g}^{ij}(\theta_f)\partial_{ij}\log f(\theta_f) + \mathrm{O}(N^{-2}).
\end{aligned} \tag{2}$$

The $N^{-3/2}$-order term of this risk is common whenever we use the same scalar function $\log f(\theta)$. This term is negative because of the definition of the point θ_f. Under condition (C2), θ_f is equal to the unique maximum point $\theta_{\max}$ of the trace $g^{X,ij}(\theta)g^Y_{ij}(\theta)$.

Based on (1) and (2), we construct asymptotically constant-risk priors using the solutions of the partial differential equations.

Theorem 3.2. *Suppose that the scalar functions* $\log \tilde{f}(\theta)$ *and* $\log \tilde{h}(\theta)$ *satisfy the following conditions:*

(A1) $\log \tilde{f}(\theta)$ *is the solution of the Eikonal equation given by*

$$\mathring{g}^{ij}(\theta)\partial_i \log \tilde{f}(\theta)\partial_j \log \tilde{f}(\theta) \quad = \quad g^{X,ij}(\theta_{\max})g^Y_{ij}(\theta_{\max}) - g^{X,ij}(\theta)g^Y_{ij}(\theta), \tag{3}$$

where $\theta_{\max}$ *is the unique maximum point of the scalar function* $g^{X,ij}(\theta)g^Y_{ij}(\theta)$.

(A2) $\log \tilde{h}(\theta)$ *is the solution of the first-order linear partial equation given by*

$$\begin{aligned}
\mathring{g}^{ij}\partial_i \log \tilde{f}(\theta)\partial_j \log \tilde{h}(\theta) = &-\mathring{g}^{ij}(\theta)\overset{\mathrm{e}}{\nabla}_i \partial_j \log \tilde{f}(\theta) \\
&- \mathring{g}^{ij}(\theta)g^{X,kl}(\theta)\left\{\overset{\mathrm{e}}{\nabla}_i \partial_k \log \tilde{f}(\theta)\right\}\partial_j \log \tilde{f}(\theta)\partial_l \log \tilde{f}(\theta) \\
&+ T^Y_{ijk}(\theta)g^{X,ij}(\theta)g^{X,kl}(\theta)\partial_l \log \tilde{f}(\theta) \\
&+ \frac{1}{3}T^Y_{ijk}(\theta)g^{X,is}(\theta)g^{X,jt}(\theta)g^{X,ku}(\theta)\partial_s \log \tilde{f}(\theta)\partial_t \log \tilde{f}(\theta)\partial_u \log \tilde{f}(\theta) \\
&- \frac{1}{2}g^Y_{kl}(\theta)M^l_{ij}(\theta)g^{X,is}(\theta)g^{X,jt}(\theta)g^{X,ku}(\theta)\partial_s \log \tilde{f}(\theta)\partial_t \log \tilde{f}(\theta)\partial_u \log \tilde{f}(\theta) \\
&- \frac{1}{2}g^{X,ij}(\theta)g^Y_{kl}(\theta)g^{X,kl}(\theta)M^m_{ij}(\theta)\partial_m \log \tilde{f}(\theta) - \mathring{g}^{ij}(\theta)M^k_{ij}(\theta)\partial_k \log \tilde{f}(\theta) \\
&- \frac{1}{2}\mathring{g}^{ij}(\theta)T^X_i(\theta)\partial_j \log \tilde{f}(\theta) - \frac{1}{2}g^{X,im}(\theta)g^Y_{ij}(\theta)g^{X,kl}(\theta)M^j_{kl}(\theta)\partial_m \log \tilde{f}(\theta) \\
&+ \mathring{g}^{ij}(\theta_{\max})\partial_{ij} \log \tilde{f}(\theta_{\max}).
\end{aligned} \tag{4}$$

Let $\pi(\theta; N)$ *be the prior that is constructed as*

$$\frac{\pi(\theta; N)}{|g^X(\theta)|^{1/2}} \quad \propto \quad \exp\{\sqrt{N}\log \tilde{f}(\theta) + \log \tilde{h}(\theta)\}.$$

Further, suppose that $\log \tilde{f}(\theta)$ *satisfies condition (C2).*

Then, the Bayesian predictive density based on the prior $\pi(\theta; N)$ *has the asymptotically smallest constant risk up to* $\mathrm{O}(N^{-2})$.

Proof. First, we consider the prior $\phi(\theta; N)$ constructed as

$$\frac{\phi(\theta; N)}{|g^X(\theta)|^{1/2}} \quad \propto \quad \exp\{\sqrt{N}\log \tilde{f}(\theta)\}.$$

From Theorem 3.1, the Kullback–Leibler risk $R(\theta, q_\phi(y|x^{(N)}))$ based on the prior $\phi(\theta; N)$ is given by

$$R(\theta, q_\phi(y|x^{(N)})) \quad = \quad \frac{1}{2N}g^{X,ij}(\theta_{\max})g^Y_{ij}(\theta_{\max}) + \mathrm{o}(N^{-1}). \tag{5}$$

This is constant up to $\mathrm{o}(N^{-1})$.

Suppose that there exists another prior $\varphi(\theta; N)$ constructed as

$$\frac{\varphi(\theta; N)}{|g^X(\theta)|^{1/2}} \quad \propto \quad \exp\{\sqrt{N}\log f(\theta)\},$$

and the Bayesian predictive density based on the prior $\varphi(\theta; N)$ has the asymptotically constant risk

$$R(\theta, q_\varphi(y|x^{(N)})) = \frac{k}{2N} + \mathrm{o}(N^{-1}).$$

From Theorem 3.1, the prior $\varphi(\theta; N)$ must satisfy the equation

$$\mathring{g}^{ij}(\theta)\partial_i \log f(\theta)\partial_j \log f(\theta) \quad = \quad k - g^{X,ij}(\theta)g^Y_{ij}(\theta).$$

The left-hand side of the above equation is non-negative, because the matrix $\mathring{g}^{ij}(\theta)$ is positive-definite. Hence, the infimum of the constant k is equal to $g^{X,ij}(\theta_{\max})g^Y_{ij}(\theta_{\max})$. From (5), the N^{-1}-order term of the risk based on the prior $\phi(\theta; N)$ achieves the infimum $g^{X,ij}(\theta_{\max})g^Y_{ij}(\theta_{\max})$. Thus, the Bayesian predictive density based on the prior $\phi(\theta; N)$ has the asymptotically smallest constant risk up to $\mathrm{o}(N^{-1})$.

Second, we consider the prior $\pi(\theta; N)$ constructed as

$$\frac{\pi(\theta; N)}{|g^X(\theta)|^{1/2}} \quad \propto \quad \exp\{\sqrt{N}\log \tilde{f}(\theta) + \log \tilde{h}(\theta)\}.$$

The above argument ensures that the prior $\pi(\theta; N)$ has the asymptotically smallest constant risk up to $\mathrm{o}(N^{-1})$. Thus, we only have to check if the $N^{-3/2}$-order term of the risk is the smallest constant. From (2), the $N^{-3/2}$-order term of the risk at the point $\theta_{\max}$ is unchanged by the choice of the scalar function $\log h(\theta)$. In other words, the constant $N^{-3/2}$-order term must agree with the quantity $\mathring{g}^{ij}(\theta_{\max})\partial_{ij}\log \tilde{f}(\theta_{\max})$. From Theorem 3.1, if we choose the prior $\pi(\theta; N)$, the $N^{-3/2}$-order term of the risk is the smallest constant, and it agrees with the quantity $\mathring{g}^{ij}(\theta_{\max})\partial_{ij}\log \tilde{f}(\theta_{\max})$. Thus, the prior $\pi(\theta; N)$ has the asymptotically smallest constant risk up to $\mathrm{O}(N^{-2})$. □

Remark 3.2. In particular, we consider the model with a one-dimensional parameter θ. From condition (C2), $\partial_\theta \log \tilde{f}(\theta)$ is specified as

$$\begin{aligned}
\sqrt{\mathring{g}^{\theta\theta}(\theta)}\partial_\theta \log \tilde{f}(\theta) &= \sqrt{g^{X,\theta\theta}(\theta_{\max})g^Y_{\theta\theta}(\theta_{\max}) - g^{X,\theta\theta}(\theta)g^Y_{\theta\theta}(\theta)} \quad \text{if} \quad \theta \le \theta_{\max}, \\
\sqrt{\mathring{g}^{\theta\theta}(\theta)}\partial_\theta \log \tilde{f}(\theta) &= -\sqrt{g^{X,\theta\theta}(\theta_{\max})g^Y_{\theta\theta}(\theta_{\max}) - g^{X,\theta\theta}(\theta)g^Y_{\theta\theta}(\theta)} \quad \text{if} \quad \theta \ge \theta_{\max}.
\end{aligned} \tag{6}$$

Integrating both sides of equation (6), the unique function $\log \tilde{f}(\theta)$ is obtained. By substituting $\log \tilde{f}(\theta)$ in (4), the unique function $\log \tilde{h}(\theta)$ is obtained.

Remark 3.3. Compare the Kullback–Leibler risk based on the asymptotically constant-risk priors $\pi(\theta; N)$ with that based on the priors $\lambda(\theta)$ independent of the sample size N. From Theorem 3.1 and Theorem 3.2, the Kullback–Leibler risk based on the asymptotically constant-risk priors $\pi(\theta; N)$ is given as

$$\begin{aligned}
R(\theta, q_\pi(y|x^{(N)})) &= \frac{1}{2N}g^{X,ij}(\theta_{\max})g^Y_{ij}(\theta_{\max}) \\
&\quad + \frac{1}{N\sqrt{N}}\mathring{g}^{ij}(\theta_{\max})\partial_{ij}\log \tilde{f}(\theta_{\max}) + \mathrm{O}(N^{-2}).
\end{aligned} \tag{7}$$

In contrast, the Kullback–Leibler risk based on the priors $\lambda(\theta)$ is given as

$$R(\theta, q_\lambda(y|x^{(N)})) \quad = \quad \frac{1}{2N}g^{X,ij}(\theta)g^Y_{ij}(\theta) + \mathrm{O}(N^{-2}). \tag{8}$$

The N^{-1}-order term in (8) is above the N^{-1}-order term in (7); although the $N^{-3/2}$-order term in (8) does not exist, the $N^{-3/2}$-order term in (7) is negative. Thus, the maximum of the risk based on the asymptotically constant-risk priors $\pi(\theta; N)$ is smaller than that of the risk based on the priors $\lambda(\theta)$.

4 Subminimax estimator problem based on the mean squared error

In this section, we refer to the subminimax estimator problem based on the mean squared error, from the viewpoint of the prediction where the distributions of data and target variables are different and have a common unknown parameter.

Let us consider the binomial estimation based on the mean squared error $R_{\mathrm{MSE}}(\theta, \hat{\theta})$. For any finite sample size N, the Bayes estimator $\hat{\theta}_\pi$ based on the Beta prior $\pi(\theta; N) \propto \theta^{\sqrt{N}/2-1}(1-\theta)^{\sqrt{N}/2-1}$ is minimax under the mean squared error. The mean squared error of the minimax Bayes estimator $\hat{\theta}_\pi$ is given by

$$R_{\mathrm{MSE}}(\theta, \hat{\theta}_\pi) = \frac{N}{4(\sqrt{N}+N)^2} = \frac{1}{4N} - \frac{1}{2N\sqrt{N}} + \mathrm{O}(N^{-2}). \quad (9)$$

In contrast, the mean squared error of the maximum likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$ is given by

$$R_{\mathrm{MSE}}(\theta, \hat{\theta}_{\mathrm{MLE}}) = \frac{\theta(1-\theta)}{N}.$$

We compare the two estimators $\hat{\theta}_\pi$ and $\hat{\theta}_{\mathrm{MLE}}$. In comparison of the N^{-1}-order terms of the mean squared errors, it seems that the maximum likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$ dominates the minimax Bayes estimator $\hat{\theta}_\pi$. In other words, the N^{-1}-order term of $R_{\mathrm{MSE}}(\theta, \hat{\theta}_{\mathrm{MLE}})$ is not greater than that of $R_{\mathrm{MSE}}(\theta, \hat{\theta}_\pi)$ for every $\theta \in \Theta$, and the equality holds when $\theta = 1/2$. This seeming paradox is known as the subminimax estimator problem; see Robbins (1950), Hodges and Lehmann (1950), and Frank and Kiefer (1951) for details. See also Ghosh (1964) for the conditions that such problems do not occur in estimation.

However, this paradox does not mean the inferiority of the minimax Bayes estimator, because, although the mean squared error of the minimax Bayes estimator $\hat{\theta}_\pi$ has the negative $N^{-3/2}$-order term, the mean squared error of the maximum likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$ does not have the $N^{-3/2}$-order term. Hence, in comparison of the mean squared errors up to $\mathrm{O}(N^{-2})$, the maximum of the mean squared error $R_{\mathrm{MSE}}(\theta, \hat{\theta}_\pi)$ is below the maximum of the mean squared error $R_{\mathrm{MSE}}(\theta, \hat{\theta}_{\mathrm{MLE}})$.

We consider the prior $\lambda(\theta)$ independent of the sample size. In the content of the estimation based on the mean squared error, the mean squared error $R_{\mathrm{MSE}}(\theta, \hat{\theta}_\lambda)$ of the Bayes estimator $\hat{\theta}_\lambda$ based on the prior $\lambda(\theta)$ is expanded as

$$R_{\mathrm{MSE}}(\theta, \hat{\theta}_\lambda) = \frac{1}{N} \sum_{i=1}^{d} g^{X,ii}(\theta) + \mathrm{O}(N^{-2}).$$

In the content of the prediction when the distributions of data and target variables are different and have a common unknown parameter, the Kullback–Leibler risk $R(\theta, q_\lambda(y|x^{(N)}))$ of the Bayesian predictive density $q_\lambda(y|x^{(N)})$ based on the prior $\lambda(\theta)$ is expanded as

$$R(\theta, q_\lambda(y|x^{(N)})) = \frac{1}{2N} g^{X,ij}(\theta) g^{Y}_{ij}(\theta) + \mathrm{O}(N^{-2}).$$

If the target variable y is a d-dimensional Gaussian random variable with the mean vector θ and unit variance, then the Kullback–Leibler risk of the predictive density $q_\lambda(y|x^{(N)})$ is equivalent to the mean squared error of the Bayesian estimator based on the prior $\lambda(\theta)$ up to $\mathrm{O}(N^{-2})$:

$$R(\theta, q_\lambda(y|x^{(N)})) = \frac{1}{2} R_{\mathrm{MSE}}(\theta, \theta_\lambda) + \mathrm{O}(N^{-2}).$$

Based on this equivalence, we consider the estimation based on the mean squared error when $\Sigma_{i=1}^{d} g^{X,ii}(\theta)$ has a unique maximum point $\theta_{\max}$. By substituting the identity matrix δ_{ij} in the expansion of the quantity $g^{Y}_{ij}(\theta)\mathrm{E}_{X^{(N)}}[(\theta^i_\pi - \theta^i)(\theta^j_\pi - \theta^j)]$, we obtain the asymptotically constant-risk priors $\pi(\theta; N)$ up to $\mathrm{O}(N^{-2})$; see Lemma A2.

We compare the mean squared error of the asymptotically constant-risk Bayes estimator $\hat{\theta}_\pi$ with that of the maximum likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$. The mean squared error of the maximum likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$ is given as

$$R_{\mathrm{MSE}}(\theta, \hat{\theta}_{\mathrm{MLE}}) = \frac{1}{N} \Sigma_{i=1}^{d} g^{X,ii}(\theta) + \mathrm{O}(N^{-2}).$$

See Efron (1975) and Amari (1985). The mean squared error of the asymptotically constant-risk Bayes estimator $\hat{\theta}_\pi$ is given as

$$R_{\mathrm{MSE}}(\theta, \hat{\theta}_\pi) = \frac{1}{N}\sum_{i=1}^{d} g^{X,ii}(\theta_{\max}) + \frac{2}{N\sqrt{N}} g^{X,ik}(\theta_{\max}) g^{Y}_{kl}(\theta_{\max}) g^{X,lj}(\theta_{\max}) \partial_{ij} \log \tilde{f}(\theta_{\max}) + \mathrm{O}(N^{-2}).$$

Thus, the maximum of the mean squared error of the asymptotically constant-risk Bayes estimator is smaller than that of estimators by the improvement of order $N^{-3/2}$ in proportion to the Hessian of the scalar function $\log \tilde{f}(\theta)$ at $\theta_{\max}$. In the prediction where the trace $g^{X,ij}(\theta) g^{Y}_{ij}(\theta)$ has a unique maximum point, the same improvement holds (Remark 3.3).

For example, we consider the binomial estimation based on the mean squared error. The geometrical quantities to be used are given by

$$\begin{aligned} g^{X}_{\theta\theta}(\theta) &= \frac{1}{\theta(1-\theta)}, & g^{Y}_{\theta\theta}(\theta) &= 1, \\ \overset{\mathrm{m}}{\Gamma}{}^{X}_{\theta\theta,\theta}(\theta) &= 0, & \overset{\mathrm{m}}{\Gamma}{}^{Y}_{\theta\theta,\theta}(\theta) &= 0, \\ \overset{\mathrm{e}}{\Gamma}{}^{X}_{\theta\theta,\theta}(\theta) &= -\frac{1-2\theta}{\theta^2(1-\theta)^2}, & \overset{\mathrm{e}}{\Gamma}{}^{Y}_{\theta\theta,\theta}(\theta) &= 0, \\ T^{X}_{\theta\theta\theta}(\theta) &= \frac{1-2\theta}{\theta^2(1-\theta)^2}, & \text{and} \quad T^{Y}_{\theta\theta\theta}(\theta) &= 0, \end{aligned}$$

respectively. Since $\overset{\mathrm{m}}{\Gamma}{}^{X,\theta}_{\theta\theta}$, $\overset{\mathrm{m}}{\Gamma}{}^{Y,\theta}_{\theta\theta}$, and $T^{Y}_{\theta\theta\theta}$ vanish, the asymptotically constant-risk prior in the estimation is identical to the asymptotically constant-risk prior in the prediction; compare Theorem 3.1 with the expansion of $g^{Y,ij}(\theta)\mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)]$ in Lemma A2.

In this example, equation (3) is given by

$$\theta^2(1-\theta)^2\{\partial_\theta \log \tilde{f}(\theta)\}^2 = \sqrt{\frac{1}{4} - \theta(1-\theta)},$$

and the solution $\log \tilde{f}(\theta)$ is $(1/2)\log\{\theta(1-\theta)\}$. Here, the second-order derivative of the function $\log \tilde{f}(\theta)$ is given by

$$\partial_{\theta\theta} \log \tilde{f}(\theta) = -\frac{1-2\theta+2\theta^2}{2\theta^2(1-\theta)^2}.$$

From this, equation (4) is given by

$$\frac{1}{2}\theta(1-\theta)(1-2\theta)\partial_\theta \log \tilde{h}(\theta) + \theta^2 - \theta = -\frac{1}{4},$$

and the solution $\log \tilde{h}(\theta)$ is $(1/2)\log\{\theta(1-\theta)\}$. Hence, the asymptotically constant-risk prior $\pi(\theta; N)$ is a Beta prior with the parameters $\alpha = \sqrt{N}/2$ and $\beta = \sqrt{N}/2$. Note that the asymptotically constant-risk prior coincides with the exact minimax prior. Since $g^{X,\theta\theta}(\theta_{\max}) = 1/2$ and $g^{X,\theta\theta}(\theta_{\max})\partial_{\theta\theta} \log \tilde{f}(\theta_{\max}) = -1$, the mean squared error of the asymptotically constant-risk Bayes estimator $\hat{\theta}_\pi$ agrees with (9) up to $\mathrm{O}(N^{-2})$.

5 Application to the prediction of the binary regression model under the covariate shift

In this section, we construct asymptotically constant-risk priors in the prediction based on the binary regression model under the covariate shift; see Shimodaira (2000).

We consider that we predict a binary response variable y based on the binary response variables $x^{(N)}$. We assume that the target variable y and the data $x^{(N)}$ follow the logistic regression models with the same parameter β given by

$$\log \frac{\Pi_x}{1-\Pi_x} = \alpha + z\beta$$

and

$$\log \frac{\Pi_y}{1-\Pi_y} = \tilde{\alpha} + \tilde{z}\beta,$$

where Π_x is the success probability of the data, and Π_y is the success probability of the target variable. Let α and $\tilde{\alpha}$ denote known constant terms, and let β denotes the common unknown parameter. Further, we assume that the covariates z and $\tilde{z}$ are different.

Using the parameter $\theta = \Pi_x$, we convert this predictive setting to binomial prediction where the data x and the target variable y are distributed according to

$$p(x|\theta) := \begin{cases} \theta & \text{if } x = 1, \\ 1-\theta & \text{if } x = 0, \end{cases}$$

and

$$q(y|\theta) := \begin{cases} \mathrm{e}^{\tilde{\alpha}-\tilde{z}z^{-1}\alpha}\theta^{\tilde{z}z^{-1}} / \left\{(1-\theta)^{\tilde{z}z^{-1}} + \mathrm{e}^{\tilde{\alpha}-\tilde{z}z^{-1}\alpha}\theta^{\tilde{z}z^{-1}}\right\} & \text{if } y = 1, \\ (1-\theta)^{\tilde{z}z^{-1}} / \left\{(1-\theta)^{\tilde{z}z^{-1}} + \mathrm{e}^{\tilde{\alpha}-\tilde{z}z^{-1}\alpha}\theta^{\tilde{z}z^{-1}}\right\} & \text{if } y = 0, \end{cases}$$

respectively. We obtain two Fisher informations for x and y as

$$g^X_{\theta\theta}(\theta) = \frac{1}{\theta(1-\theta)}$$

and

$$g^Y_{\theta\theta}(\theta) = \left(\frac{\tilde{z}}{z}\right)^2 \mathrm{e}^{-\tilde{\alpha}+\tilde{z}z^{-1}\alpha} \frac{(1-\theta)^{\tilde{z}z^{-1}-2}\theta^{\tilde{z}z^{-1}-2}}{\left\{\theta^{\tilde{z}z^{-1}} + \mathrm{e}^{-\tilde{\alpha}+\tilde{z}z^{-1}\alpha}(1-\theta)^{\tilde{z}z^{-1}}\right\}^2},$$

respectively.

For simplicity, we consider the setting where $z = 1$, $\tilde{z} = 2$, and $\alpha = \tilde{\alpha} = 0$. The geometrical quantities to be used are given by

$$\begin{aligned} g^X_{\theta\theta}(\theta) &= \frac{1}{\theta(1-\theta)}, & g^Y_{\theta\theta}(\theta) &= \frac{4}{\{(1-\theta)^2+\theta^2\}^2}, \\ \overset{\mathrm{m}}{\Gamma}{}^X_{\theta\theta,\theta}(\theta) &= 0, & \overset{\mathrm{m}}{\Gamma}{}^Y_{\theta\theta,\theta}(\theta) &= 4\frac{(1-2\theta)(1+2\theta-2\theta^2)}{\theta(1-\theta)\{(1-\theta)^2+\theta^2\}^3}, \\ \overset{\mathrm{e}}{\Gamma}{}^X_{\theta\theta,\theta}(\theta) &= -\frac{1-2\theta}{\theta^2(1-\theta)^2}, & \overset{\mathrm{e}}{\Gamma}{}^Y_{\theta\theta,\theta}(\theta) &= -4\frac{1-2\theta}{\theta(1-\theta)\{(1-\theta)^2+\theta^2\}^2}, \\ T^X_{\theta\theta\theta}(\theta) &= \frac{1-2\theta}{\theta^2(1-\theta)^2}, & \text{and} \quad T^Y_{\theta\theta\theta}(\theta) &= 8\frac{1-2\theta}{\theta(1-\theta)\{(1-\theta)^2+\theta^2\}^3}, \end{aligned}$$

respectively. Using these quantities, equation (3) is given by

$$4\frac{\theta^2(1-\theta)^2}{\{\theta^2+(1-\theta)^2\}^2}(\partial_\theta \log \tilde{f}(\theta))^2 = 4 - 4\frac{\theta(1-\theta)}{\{\theta^2+(1-\theta)^2\}^2}.$$

By noting that the maximum point of $g^{X,\theta\theta}(\theta)g^Y_{\theta\theta}(\theta)$ is $1/2$, the solution $\log \tilde{f}(\theta)$ of this equation is given by

$$\begin{aligned} \log \tilde{f}(\theta) &= 2\sqrt{1-\theta+\theta^2} + \log\{\theta(1-\theta)\} \\ &\quad - \log(2-\theta+2\sqrt{1-\theta+\theta^2}) - \log(1+\theta+2\sqrt{1-\theta+\theta^2}). \end{aligned}$$

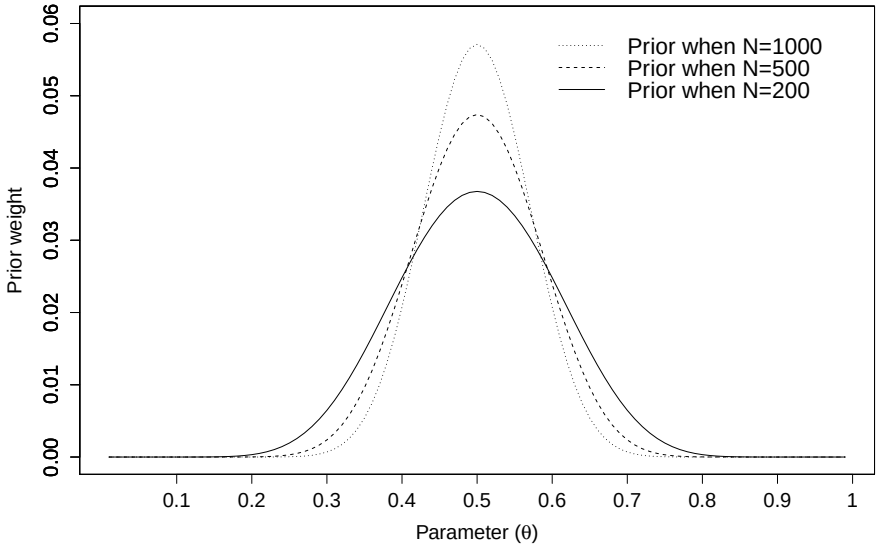

Figure 1: Asymptotically constant-risk prior in the prediction where the data are distributed according to the binomial distribution $\mathrm{Bin}(N, \theta)$ and the target variable is distributed according to the binomial distribution $\mathrm{Bin}(1, \theta^2/(\theta^2 + (1-\theta)^2))$

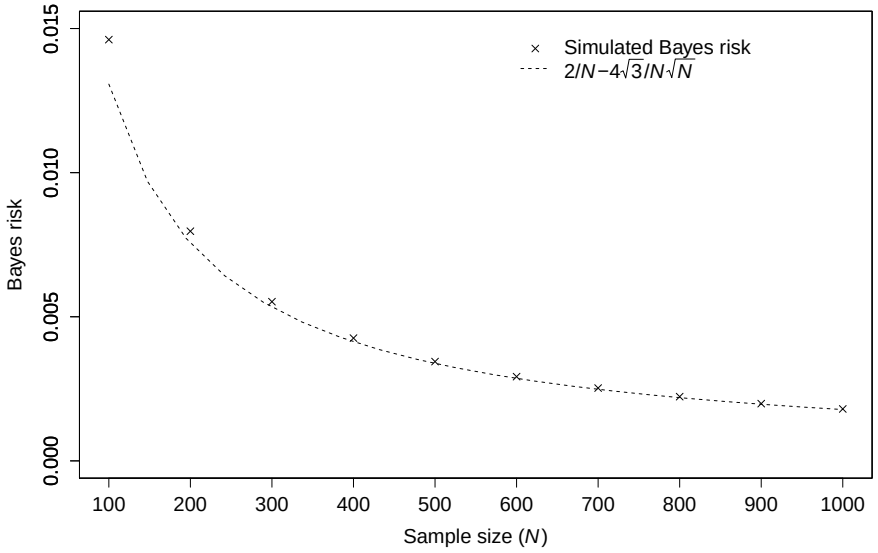

Figure 2: Bayes risk based on the asymptotically constant-risk prior in the prediction where the data are distributed according to the binomial distribution $\mathrm{Bin}(N, \theta)$ and the target variable is distributed according to the binomial distribution $\mathrm{Bin}(1, \theta^2/(\theta^2 + (1-\theta)^2))$

Using this solution, we obtain the solution of equation (4) given by

$$\begin{aligned}\log \tilde{h}(\theta) \;=\; & \frac{1}{6}\Bigg[-\frac{1}{1-\theta} - \frac{1}{\theta} - 12\theta(1-\theta) - 12\sqrt{3}\sqrt{1-\theta+\theta^2} \\ & +(3-6\sqrt{3})\{\log\theta + \log(1-\theta)\} - 3\log(1-\theta+\theta^2) + 10\log\{(1-\theta)^2+\theta^2\} \\ & -6\log(\sqrt{3}+2\sqrt{1-\theta+\theta^2}) + 6\sqrt{3}\log\{1+(1-\theta)+2\sqrt{1-\theta+\theta^2}\} \\ & +6\sqrt{3}\log\{1+\theta+2\sqrt{1-\theta+\theta^2}\}\Bigg].\end{aligned}$$

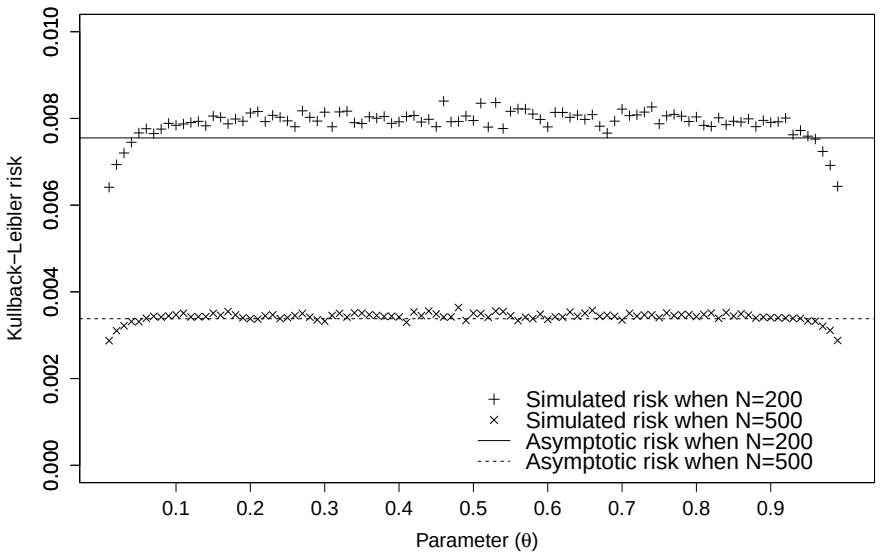

Figure 3: Comparison of the Kullback–Leibler risk calculated using the Monte Carlo simulations and the asymptotic risk $2/N - (4\sqrt{3})/(N\sqrt{N})$ in the prediction where the data are distributed according to the binomial distribution $\mathrm{Bin}(N, \theta)$ and the target variable is distributed according to the binomial distribution $\mathrm{Bin}(1, \theta^2/(\theta^2 + (1-\theta)^2))$

The asymptotically constant-risk priors for the different sample sizes are shown in Fig.1. The prior weight is found to be more concentrated to 1/2 as the sample size N grows.

In this example, we obtain the Kullback–Leibler risk of the Bayesian predictive density based on the asymptotically constant-risk prior $\pi(\theta; N)$ as

$$R(\theta, q_\pi(y|x^{(N)})) \quad = \quad \frac{2}{N} - \frac{4\sqrt{3}}{N\sqrt{N}} + \mathrm{O}(N^{-2}).$$

We compare this value with the Bayes risk calculated using the Monte Carlo simulation; see Fig.2. As the sample size N grows, the difference appears negligible. Further, we compare this value with the risk itself calculated by the Monte Carlo simulation; see Fig.3. As the sample size N grows, the risk becomes more constant.

6 Discussion

We have considered the setting where the quantity $g^{X,ij}(\theta) g^Y_{ij}(\theta)$ – the trace of the product of the inverse Fisher information matrix $g^{X,ij}(\theta)$ and the Fisher information matrix $g^Y_{ij}(\theta)$– has a unique maximum point, and we have investigated asymptotically constant-risk priors in the sense that the asymptotic risk is constant up to $\mathrm{O}(N^{-2})$.

In Section 3, we have considered the prior depending on the sample size N and constructed the asymptotically constant-risk prior using the partial differential equations (3) and (4). In Section 4, we have clarified the relationship between the subminimax estimator problem based on the mean squared error and the prediction where the distributions of data and target variables are different. In Section 5, we have constructed asymptotically constant-risk priors in the prediction based on the logistic regression model under the covariate shift.

We have assumed that the trace $g^{X,ij}(\theta) g^Y_{ij}(\theta)$ is finite. However, the trace may diverge in the non-compact parameter space; for example, it diverges under the predictive setting where the distribution $q(y|\theta)$ of the target variable is the Poisson distribution and the data distribution $p(x|\theta)$ is the exponential distribution with Θ equivalent to $\mathbb{R}$. Therefore, for our future work, in such a setting, we should adopt the criteria other than minimaxity.

Appendix

We prove Theorem 3.1. First, we introduce some lemmas for the proof. For the expansion, we follow the following six steps (the first five steps are arranged in the form of lemmas): the first is to expand the MAP estimator, the second is to calculate their bias and mean squared error, the third is to expand the Kullback–Leibler risk using $\hat{\theta}_\pi$-plugin predictive density $q(y|\hat{\theta}_\pi)$, the fourth is to expand the Bayesian predictive density based on the prior $\pi(\theta; N)$, the fifth is to expand the Bayesian estimator minimizing the Bayes risk, and the last is to prove Theorem 3.1 using these lemmas.

We use some additional notations for the expansion. Let $\hat{\theta}_\pi$ be the maximum point of the scalar function $\log p(x^{(N)}|\theta) + \log\{\pi(\theta; N)/|g^X(\theta)|^{1/2}\}$. Let $l(\theta|x^{(N)})$ denote the log likelihood of the data $x^{(N)}$. Let $l_{ij}(\theta|x^{(N)})$, $l_{ijk}(\theta|x^{(N)})$, and $l_{ijkl}(\theta|x^{(N)})$ be the derivatives of order 2, 3, and 4 of the log likelihood $l(\theta|x^{(N)})$. Let $H_{ij}(\theta|x^{(N)})$ denote the quantity $l_{ij}(\theta|x^{(N)}) + N g^X_{ij}(\theta)$. Let $\tilde{l}_i(\theta|x^{(N)})$ and $\tilde{H}_{ij}(\theta|x^{(N)})$ denote $(1/\sqrt{N})l_i(\theta|x^{(N)})$ and $(1/\sqrt{N})H_{ij}(\theta|x^{(N)})$, respectively. In addition, the brackets ( ) denotes the symmetrization: for any two tensors a_{ij} and b_{ij}, $a_{i(j}b_{k)l}$ denotes $a_{i(j}b_{k)l} = (a_{ij}b_{kl} + a_{ik}b_{jl})/2$.

Lemma A1. *Let $\hat{\theta}_\pi$ be the maximum point of $\log p(x^{(N)}|\theta) + \log\{\pi(\theta; N)/|g^X(\theta)|^{1/2}\}$. Then, the i-th component of this estimator $\hat{\theta}_\pi$ is expanded as follows:*

$$
\begin{aligned}
\hat{\theta}^i_\pi &= \theta^i + \frac{1}{\sqrt{N}} g^{X,ik}(\theta)\tilde{l}_k(\theta|x^{(N)}) + \frac{1}{\sqrt{N}} g^{X,ik}(\theta)\partial_k \log f(\theta) \\
&+ \frac{1}{N} g^{X,ik}(\theta)\tilde{H}_{km}(\theta|x^{(N)}) g^{X,mr}(\theta)\tilde{l}_r(\theta|x^N) \\
&+ \frac{1}{2N} g^{X,ik}(\theta) L^X_{kmr}(\theta) g^{X,mq}(\theta) g^{X,rs}(\theta)\tilde{l}_q(\theta|x^N)\tilde{l}_s(\theta|x^{(N)}) \\
&+ \frac{1}{N} g^{X,ik}(\theta)\tilde{H}_{km}(\theta|x^N) g^{X,mr}(\theta)\partial_r \log f(\theta) \\
&+ \frac{1}{N} g^{X,ik}(\theta) L^X_{kmr}(\theta) g^{X,mq}(\theta) g^{X,rs}(\theta)\tilde{l}_q(\theta|x^N)\partial_s \log f(\theta) \\
&+ \frac{1}{2N} g^{X,ik}(\theta) L^X_{kmr}(\theta) g^{X,mq}(\theta) g^{X,rs}(\theta)\partial_q \log f(\theta)\partial_s \log f(\theta) \\
&+ \frac{1}{N} g^{X,ik}(\theta) g^{X,mq}(\theta)\partial_{km} \log f(\theta)\tilde{l}_q(\theta|x^{(N)}) \\
&+ \frac{1}{N} g^{X,ik}(\theta) g^{X,mq}(\theta)\partial_{km} \log f(\theta)\partial_q \log f(\theta) \\
&+ \frac{1}{N} g^{X,ik}(\theta)\partial_k \log h(\theta) + \mathrm{O_P}(N^{-3/2}).
\end{aligned} \tag{10}
$$

Proof. By the definition of $\hat{\theta}_\pi$, we get the equation given by

$$
\partial_i \log p(x^{(N)}|\hat{\theta}_\pi) + \partial_i \log \frac{\pi(\hat{\theta}_\pi; N)}{|g^X(\hat{\theta}_\pi)|^{1/2}} = 0.
$$

From our assumption that prior $\pi(\theta; N)$ has the form given by

$$
\frac{\pi(\theta; N)}{|g^X(\theta)|^{1/2}} \propto \exp\{\sqrt{N}\log f(\theta) + \log h(\theta)\},
$$

we rewrite this equation as

$$
\partial_i \log p(x^{(N)}|\hat{\theta}_\pi) + \sqrt{N}\partial_i \log f(\hat{\theta}_\pi) + \partial_i \log h(\hat{\theta}_\pi) = 0.
$$

By applying Taylor expansion around θ to this new equation, we derive the following expansion:

$$
\begin{aligned}
&\partial_i \log p(x^{(N)}|\theta) + \{\partial_{ij} \log p(x^{(N)}|\theta)\}(\hat{\theta}^j_\pi - \theta^j) \\
&+ \frac{1}{2}\{\partial_{ijk} \log p(x^{(N)}|\theta)\}(\hat{\theta}^j_\pi - \theta^j)(\hat{\theta}^k_\pi - \theta^k) + \sqrt{N}\partial_i \log f(\theta) \\
&+ \sqrt{N}\{\partial_{ij} \log f(\theta)\}(\hat{\theta}^j_\pi - \theta^j) + \partial_i \log h(\theta) + \mathrm{o_P}(1) = 0.
\end{aligned}
$$

From the law of large numbers and the central limit theorem, we rewrite the above expansion as

$$\begin{aligned} Ng^X_{ij}(\theta)(\hat{\theta}^j_\pi - \theta^j) &= \partial_i \log p(x^{(N)}|\theta) + \sqrt{N}\partial_i \log f(\theta) + H_{ij}(\theta|x^{(N)})(\hat{\theta}^j_\pi - \theta^j) \\ &\quad + \frac{N}{2} L_{ijk}(\theta)(\hat{\theta}^j_\pi - \theta^j)(\hat{\theta}^k_\pi - \theta^k) + \sqrt{N}\partial_{ij} \log f(\theta)(\hat{\theta}^j_\pi - \theta^j) \\ &\quad + \partial_i \log h(\theta) + \mathrm{o_P}(1). \end{aligned} \tag{11}$$

By substituting the deviation $\hat{\theta}_\pi - \theta$ recursively into expansion (11), we obtain expansion (10).

□

Lemma A2. *Let $\hat{\theta}_\pi$ be the maximum point of $\log p(x^{(N)}|\theta) + \log\{\pi(\theta; N)/|g^X(\theta)|^{1/2}\}$. Then, the i-th component of the bias of the estimator $\hat{\theta}_\pi$ is given by*

$$\begin{aligned} \mathrm{E}_{X^{(N)}}[\hat{\theta}^i_\pi] &= \theta^i + \frac{1}{\sqrt{N}} g^{X,ik}\partial_k \log f(\theta) \\ &\quad - \frac{1}{2N}\overset{\mathrm{m}}{\Gamma}{}^{X,i}(\theta) + \frac{1}{2N} g^{X,ik}(\theta) g^{X,mq}(\theta) g^{X,rs}(\theta) L^X_{kmr}(\theta)\partial_q \log f(\theta)\partial_s \log f(\theta) \\ &\quad + \frac{1}{N} g^{X,ik}(\theta) g^{X,mq}(\theta)\partial_{km} \log f(\theta)\partial_q \log f(\theta) \\ &\quad + \frac{1}{N} g^{X,ik}(\theta)\partial_k \log h(\theta) + \mathrm{O}(N^{-3/2}). \end{aligned} \tag{12}$$

The (i,j)-component of the mean squared error of $\hat{\theta}_\pi$ is given by

$$\begin{aligned} &\mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)] \\ &= \frac{1}{N} g^{X,ij}(\theta) + \frac{1}{N} g^{X,ik}(\theta) g^{X,jl}(\theta)\partial_k \log f(\theta)\partial_l \log f(\theta) \\ &\quad - \frac{1}{N\sqrt{N}} g^{X,k(i}(\theta)\overset{\mathrm{m}}{\Gamma}{}^{X,j)}(\theta)\partial_k \log f(\theta) + \frac{2}{N\sqrt{N}} g^{X,k(i}(\theta) g^{X,j)l}(\theta)\partial_{kl} \log f(\theta) \\ &\quad + \frac{2}{N\sqrt{N}} g^{X,k(i}(\theta)\partial_k g^{X,j)l}(\theta)\partial_l \log f(\theta) \\ &\quad + \frac{1}{N\sqrt{N}} g^{X,k(i}(\theta) g^{X,j)l}(\theta) g^{X,nr}(\theta) g^{X,pt}(\theta) L^X_{lrt}(\theta)\partial_k \log f(\theta)\partial_n \log f(\theta)\partial_p \log f(\theta) \\ &\quad + \frac{2}{N\sqrt{N}} g^{X,k(i}(\theta) g^{X,j)l}(\theta) g^{X,nr}(\theta)\partial_{ln} \log f(\theta)\partial_r \log f(\theta)\partial_k \log f(\theta) \\ &\quad + \frac{2}{N\sqrt{N}} g^{X,k(i}(\theta) g^{X,j)l}(\theta)\partial_k \log f(\theta)\partial_l \log h(\theta) \\ &\quad + \mathrm{O}(N^{-2}), \end{aligned} \tag{13}$$

where $g^{X,k(i}(\theta)\overset{\mathrm{m}}{\Gamma}{}^{X,j)}(\theta)$ denotes $(1/2)\{g^{X,ki}(\theta)\overset{\mathrm{m}}{\Gamma}{}^{X,j}(\theta) + g^{X,ki}(\theta)\overset{\mathrm{m}}{\Gamma}{}^{X,j}(\theta)\}$ and $g^{X,k(i}(\theta)\partial_k g^{X,j)l}(\theta)$ denotes $(1/2)\{g^{X,ki}(\theta)\partial_k g^{X,jl}(\theta) + g^{X,kj}(\theta)\partial_k g^{X,il}(\theta)\}$. The (i,j,k)-component of the mean of the third power of the deviation $\hat{\theta}_\pi - \theta$ is given by

$$\begin{aligned} &\mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)(\hat{\theta}^k_\pi - \theta^k)] \\ &= \frac{1}{N\sqrt{N}} g^{X,is}(\theta) g^{X,jt}(\theta) g^{X,ku}(\theta)\partial_s \log f(\theta)\partial_t \log f(\theta)\partial_u \log f(\theta) \\ &\quad + \frac{3}{N\sqrt{N}} g^{X,(ij}(\theta) g^{X,k)l}(\theta)\partial_l \log f(\theta) + \mathrm{O}(N^{-2}). \end{aligned} \tag{14}$$

Proof. First, using Lemma A1, we determine the i-th component of the bias of $\hat{\theta}_\pi$ given by

$$
\begin{aligned}
&\mathrm{E}_{X^{(N)}}[\hat{\theta}^i_\pi - \theta^i] \\
&= \frac{1}{\sqrt{N}} g^{X,ik} \partial_k \log f(\theta) \\
&\quad - \frac{1}{2N} \overset{\mathrm{m}}{\Gamma}{}^{X,i}(\theta) + \frac{1}{2N} g^{X,ik}(\theta) g^{X,mq}(\theta) g^{X,rs}(\theta) L^X_{kmr}(\theta) \partial_q \log f(\theta) \partial_s \log f(\theta) \\
&\quad + \frac{1}{N} g^{X,ik}(\theta) g^{X,mq}(\theta) \partial_{km} \log f(\theta) \partial_q \log f(\theta) \\
&\quad + \frac{1}{N} g^{X,ik}(\theta) \partial_k \log h(\theta) + \mathrm{O}(N^{-3/2}).
\end{aligned}
$$

Second, consider the following relationship:

$$
\begin{aligned}
&\mathrm{E}_{X^{(N)}} \left[\left\{ \hat{\theta}^i_\pi - \theta^i - \frac{1}{\sqrt{N}} g^{X,ik}(\theta) \tilde{l}_k(\theta | x^{(N)}) - \frac{1}{\sqrt{N}} g^{X,ik}(\theta) \partial_k \log f(\theta) \right\} \right. \\
&\qquad \left. \times \left\{ \hat{\theta}^j_\pi - \theta^j - \frac{1}{\sqrt{N}} g^{X,jl}(\theta) \tilde{l}_l(\theta | x^N) - \frac{1}{\sqrt{N}} g^{X,jl}(\theta) \partial_l \log f(\theta) \right\} \right] \\
&= \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)] + \frac{1}{N} g^{X,ij}(\theta) + \frac{1}{N} g^{X,ik}(\theta) g^{X,jl}(\theta) \partial_k \log f(\theta) \partial_l \log f(\theta) \\
&\quad - \frac{1}{\sqrt{N}} g^{X,ki}(\theta) \mathrm{E}_{X^{(N)}}[(\hat{\theta}^j_\pi - \theta^j) \tilde{l}_k(\theta | x^{(N)})] \\
&\quad - \frac{1}{\sqrt{N}} g^{X,kj}(\theta) \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i) \tilde{l}_k(\theta | x^{(N)})] \\
&\quad - \frac{1}{\sqrt{N}} g^{X,ki}(\theta) \mathrm{E}_{X^{(N)}}[(\hat{\theta}^j_\pi - \theta^j) \partial_k \log f(\theta)] \\
&\quad - \frac{1}{\sqrt{N}} g^{X,kj}(\theta) \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i) \partial_k \log f(\theta)]. \qquad (15)
\end{aligned}
$$

By differentiating the j-th component of the bias $\mathrm{E}_{X^{(N)}}[\hat{\theta}^j_\pi - \theta^j]$, we obtain the equation given by

$$
\frac{1}{N} \partial_k \mathrm{E}_{X^{(N)}}[\hat{\theta}^j_\pi - \theta^j] \quad = \quad - \frac{1}{N} \delta^j_k + \frac{1}{\sqrt{N}} \mathrm{E}_{X^{(N)}}[(\hat{\theta}^j_\pi - \theta^j) \tilde{l}_k(\theta | x^N)], \qquad (16)
$$

where δ^i_j denotes the delta function: if the upper and the lower indices agree then the value of this function is 1 and otherwise 0. Equation (16) has been used by Efron (1975), Amari (1985), and Komaki (1996). By substituting equations (16) and (12) into relationship (15), we obtain the (i, j)-component of the mean squared error of $\hat{\theta}_\pi$ given by

$$
\begin{aligned}
&\mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)] \\
&= \frac{1}{N} g^{X,ij}(\theta) + \frac{1}{N} g^{X,ik}(\theta) g^{X,jl}(\theta) \partial_k \log f(\theta) \partial_l \log f(\theta) \\
&\quad - \frac{1}{N\sqrt{N}} g^{X,k(i}(\theta) \overset{\mathrm{m}}{\Gamma}{}^{X,j)}(\theta) \partial_k \log f(\theta) + \frac{2}{N\sqrt{N}} g^{X,k(i}(\theta) g^{X,j)l}(\theta) \partial_{kl} \log f(\theta) \\
&\quad + \frac{2}{N\sqrt{N}} g^{X,k(i}(\theta) \partial_k g^{X,j)l}(\theta) \partial_l \log f(\theta) \\
&\quad + \frac{1}{N\sqrt{N}} g^{X,k(i}(\theta) g^{X,j)l}(\theta) g^{X,nr}(\theta) g^{X,pt}(\theta) L^X_{lrt}(\theta) \partial_k \log f(\theta) \partial_n \log f(\theta) \partial_p \log f(\theta) \\
&\quad + \frac{2}{N\sqrt{N}} g^{X,k(i}(\theta) g^{X,j)l}(\theta) g^{X,nr}(\theta) \partial_{ln} \log f(\theta) \partial_r \log f(\theta) \partial_k \log f(\theta) \\
&\quad + \frac{2}{N\sqrt{N}} g^{X,k(i}(\theta) g^{X,j)l}(\theta) \partial_k \log f(\theta) \partial_l \log h(\theta) \\
&\quad + \mathrm{O}(N^{-2}).
\end{aligned}
$$

Finally, by taking the expectation of the third power of the deviation $\hat{\theta}^i_\pi - \theta^i$, we obtain the following expansion:

$$
\begin{aligned}
&\mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)(\hat{\theta}^k_\pi - \theta^k)] \\
&= \frac{1}{N\sqrt{N}} g^{X,is}(\theta) g^{X,jt}(\theta) g^{X,ku}(\theta) \partial_s \log f(\theta) \partial_t \log f(\theta) \partial_u \log f(\theta) \\
&\quad + \frac{3}{N\sqrt{N}} g^{X,(ij}(\theta) g^{X,k)l}(\theta) \partial_l \log f(\theta) + \mathrm{O}(N^{-2}).
\end{aligned}
$$

□

Lemma A3. *Let $\hat{\theta}_\pi$ be the maximum point of $\log p(x^{(N)}|\theta) + \log\{\pi(\theta; N)/|g^X(\theta)|^{1/2}\}$. The Kullback–Leibler risk of the plug-in predictive density $q(y^{(N)}|\hat{\theta}_\pi)$ with the estimator $\hat{\theta}_\pi$ is expanded as follows:*

$$
\begin{aligned}
&R(\theta, q(y|\hat{\theta}_\pi)) \\
&= \frac{1}{2N} g^Y_{ij}(\theta) g^{X,ij}(\theta) + \frac{1}{2N} \mathring{g}^{ij}(\theta) \partial_i \log f(\theta) \partial_j \log f(\theta) \\
&\quad + \frac{1}{N\sqrt{N}} \mathring{g}^{ij}(\theta) \left\{ \partial_{ij} \log f(\theta) - \overset{\mathrm{e}}{\Gamma}{}^{X,k}_{ij}(\theta) \partial_k \log f(\theta) \right\} \\
&\quad + \frac{1}{N\sqrt{N}} \mathring{g}^{ij}(\theta) g^{X,kl}(\theta) \left\{ \partial_{ik} \log f(\theta) - \overset{\mathrm{e}}{\Gamma}{}^{X,m}_{ik} \partial_m \log f(\theta) \right\} \partial_j \log f(\theta) \partial_l \log f(\theta) \\
&\quad - \frac{1}{N\sqrt{N}} T^Y_{ijk}(\theta) g^{X,ij}(\theta) g^{X,kl}(\theta) \partial_l \log f(\theta) \\
&\quad - \frac{1}{3N\sqrt{N}} T^Y_{ijk}(\theta) g^{X,is}(\theta) g^{X,jt}(\theta) g^{X,ku}(\theta) \partial_s \log f(\theta) \partial_t \log f(\theta) \partial_u \log f(\theta) \\
&\quad + \frac{1}{2N\sqrt{N}} g^Y_{kl}(\theta) \left\{ \overset{\mathrm{m}}{\Gamma}{}^{Y,l}_{ij}(\theta) - \overset{\mathrm{m}}{\Gamma}{}^{X,l}_{ij}(\theta) \right\} g^{X,is}(\theta) g^{X,jt}(\theta) g^{X,ku}(\theta) \partial_s \log f(\theta) \partial_t \log f(\theta) \partial_u \log f(\theta) \\
&\quad + \frac{1}{2N\sqrt{N}} g^{X,ij}(\theta) g^Y_{kl}(\theta) g^{X,kl}(\theta) \left\{ \overset{\mathrm{m}}{\Gamma}{}^{Y,m}_{ij}(\theta) - \overset{\mathrm{m}}{\Gamma}{}^{X,m}_{ij}(\theta) \right\} \partial_m \log f(\theta) \\
&\quad + \frac{1}{N\sqrt{N}} \mathring{g}^{ij}(\theta) \left\{ \overset{\mathrm{m}}{\Gamma}{}^{Y,k}_{ij}(\theta) - \overset{\mathrm{m}}{\Gamma}{}^{X,k}_{ij}(\theta) \right\} \partial_k \log f(\theta) \\
&\quad + \frac{1}{N\sqrt{N}} \mathring{g}^{ij}(\theta) \partial_i \log f(\theta) \partial_j \log h(\theta) + \mathrm{O}(N^{-2}).
\end{aligned} \tag{17}
$$

Proof. By applying the Taylor expansion, the Kullback–Leibler risk $R(\theta, q(y|\hat{\theta}_\pi))$ is expanded as

$$
\begin{aligned}
&E_{x^{(N)}}[D(q(\cdot|\theta), q(\cdot|\hat{\theta}_\pi))] \\
&= \mathrm{E}_{X^{(N)}} \left[\int q(y|\theta) \left\{ -l_i(\theta|y) \tilde{\theta}^i_\pi - \frac{1}{2} l_{ij}(\theta|y) (\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j) \right.\right. \\
&\qquad \left.\left. - \frac{1}{6} l_{ijk}(\theta|y) (\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)(\hat{\theta}^k_\pi - \theta^k) + \mathrm{O_P}(N^{-2}) \right\} \mathrm{d}y \right] \\
&= \frac{1}{2} g^Y_{ij}(\theta) \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)] - \frac{1}{6} L^Y_{ijk}(\theta) \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)(\hat{\theta}^k_\pi - \theta^k)] + \mathrm{O}(N^{-2}) \\
&= \frac{1}{2} g^Y_{ij}(\theta) \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)] \\
&\quad + \left\{ \frac{3}{2} \overset{\mathrm{m}}{\Gamma}{}^Y_{(ij,k)}(\theta) - \frac{1}{3} T^Y_{ijk}(\theta) \right\} \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)(\hat{\theta}^k_\pi - \theta^k)] + \mathrm{O}(N^{-2}) \\
&= \frac{1}{2} g^Y_{ij}(\theta) \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)] - \frac{1}{3} T^Y_{ijk}(\theta) \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)(\hat{\theta}^k_\pi - \theta^k)] \\
&\quad + \frac{1}{2} \left\{ g^Y_{kl}(\theta) \overset{\mathrm{m}}{\Gamma}{}^{Y,l}_{ij}(\theta) - g^Y_{kl}(\theta) \overset{\mathrm{m}}{\Gamma}{}^{X,l}_{ij}(\theta) \right\} \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)(\hat{\theta}^k_\pi - \theta^k)] \\
&\quad + \frac{1}{2} g^Y_{kl}(\theta) \overset{\mathrm{m}}{\Gamma}{}^{X,l}_{ij}(\theta) \mathrm{E}_{X^{(N)}}[(\hat{\theta}^i_\pi - \theta^i)(\hat{\theta}^j_\pi - \theta^j)(\hat{\theta}^k_\pi - \theta^k] + \mathrm{O}(N^{-2}),
\end{aligned} \tag{18}
$$

where $\overset{\mathrm{e}}{\Gamma}{}^{Y}_{(ij,k)}$ denotes $(1/3)\{\overset{\mathrm{e}}{\Gamma}{}^{Y}_{ij,k}+\overset{\mathrm{e}}{\Gamma}{}^{Y}_{jk,i}+\overset{\mathrm{e}}{\Gamma}{}^{Y}_{ki,j}\}$.

By the definition of the predictive metric $\mathring{g}_{ij}(\theta)=g^{X}_{ik}(\theta)g^{Y,kl}(\theta)g^{X}_{lj}(\theta)$, by expansions (13) and (14), and by the relationship $L^{X}_{ijk}(\theta)=-\overset{\mathrm{e}}{\Gamma}{}^{X}_{ij,k}(\theta)-\overset{\mathrm{e}}{\Gamma}{}^{X}_{jk,i}(\theta)-\overset{\mathrm{e}}{\Gamma}{}^{X}_{ki,j}(\theta)-T^{X}_{ijk}(\theta)$, the last two terms of the above expansion (18) are expanded as

$$
\begin{aligned}
&\frac{1}{2}g^{Y}_{ij}(\theta)\mathrm{E}_{X^{(N)}}[(\hat{\theta}^{i}_{\pi}-\theta^{i})(\hat{\theta}^{j}_{\pi}-\theta^{j})]+\frac{1}{2}g^{Y}_{kl}(\theta)\overset{\mathrm{m}}{\Gamma}{}^{X,l}_{ij}(\theta)\mathrm{E}_{X^{(N)}}[(\hat{\theta}^{i}_{\pi}-\theta^{i})(\hat{\theta}^{j}_{\pi}-\theta^{j})(\hat{\theta}^{k}_{\pi}-\theta^{k})]\\
&=\frac{1}{2N}g^{Y}_{ij}(\theta)g^{X,ij}(\theta)+\frac{1}{2N}\mathring{g}^{ij}(\theta)\partial_i\log f(\theta)\partial_j\log f(\theta)\\
&\quad+\frac{1}{N\sqrt{N}}\mathring{g}^{ij}(\theta)\left\{\partial_{ij}\log f(\theta)-\overset{\mathrm{e}}{\Gamma}{}^{X,k}_{ij}(\theta)\partial_k\log f(\theta)\right\}\\
&\quad+\frac{1}{N\sqrt{N}}\mathring{g}^{ij}(\theta)g^{X,kl}(\theta)\left\{\partial_{ik}\log f(\theta)-\overset{\mathrm{e}}{\Gamma}{}^{X,m}_{ik}\partial_m\log f(\theta)\right\}\partial_j\log f(\theta)\partial_l\log f(\theta)\\
&\quad+\frac{1}{N\sqrt{N}}\mathring{g}^{ij}(\theta)\partial_i\log f(\theta)\partial_j\log h(\theta)+\mathrm{O}(N^{-2}).
\end{aligned}
\tag{19}
$$

By substituting expansion (19) into expansion (18), expansion (17) is obtained.

□

Note that expansion (17) is invariant up to $\mathrm{O}(N^{-2})$ under the reparametrization so that each term of this expansion is a scalar function of θ.

Lemma A4. *Let $\hat{\theta}_{\pi}$ be the maximum point of $\log p(x^{(N)}|\theta)+\log\{\pi(\theta;N)/|g^{X}(\theta)|^{1/2}\}$. The Bayesian predictive density based on the prior $\pi(\theta;N)$ is expanded as*

$$
\begin{aligned}
q_{\pi}(y|x^{(N)}) &= q(y|\hat{\theta}_{\pi})+\frac{1}{N}g^{X,ij}(\hat{\theta}_{\pi})\left\{\partial_i\log|g^{X}(\hat{\theta}_{\pi})|^{\frac{1}{2}}-\overset{\mathrm{e}}{\Gamma}{}^{X,k}_{ik}(\hat{\theta}_{\pi})\right\}\partial_j q(y|\hat{\theta}_{\pi})\\
&\quad+\frac{1}{2N}g^{X,ij}(\hat{\theta}_{\pi})\left\{\partial_{ij}q(y|\hat{\theta}_{\pi})-\overset{\mathrm{m}}{\Gamma}{}^{X,k}_{ij}(\hat{\theta}_{\pi})\partial_k q(y|\hat{\theta}_{\pi})\right\}\\
&\quad+\mathrm{O}_{\mathrm{P}}(N^{-3/2}).
\end{aligned}
\tag{20}
$$

Proof. Let $\tilde{\theta}_{\pi}$ denote $\hat{\theta}_{\pi}-\theta$. First, using a Taylor expansion twice, we expand the posterior density $\pi(\theta|x^{(N)})$ as

$$
\begin{aligned}
\pi(\theta|x^{(N)}) &= |g^{X}(\hat{\theta}_{\pi})|^{\frac{1}{2}}\frac{\pi(\hat{\theta}_{\pi})}{|g^{X}(\hat{\theta}_{\pi})|^{\frac{1}{2}}}p(x^{(N)}|\hat{\theta}_{\pi})\exp\left[-\frac{1}{2}\{-l_{ij}(\hat{\theta}_{\pi}|x^{(N)})\}\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}\right]\\
&\quad\times\left[1-\{\partial_i\log|g^{X}(\hat{\theta}_{\pi})|^{\frac{1}{2}}\}\tilde{\theta}^{i}_{\pi}+\frac{1}{2}\left\{\frac{\partial_{ij}|g^{X}(\hat{\theta}_{\pi})|^{\frac{1}{2}}}{|g^{X}(\hat{\theta}_{\pi})|^{\frac{1}{2}}}\right\}\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}+\mathrm{O}_{\mathrm{P}}(N^{-3/2})\right]\\
&\quad\times\left(1+\frac{1}{2}\{\sqrt{N}\partial_{ij}\log f(\hat{\theta}_{\pi})\}\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}-\frac{1}{6}\{l_{ijk}(\hat{\theta}_{\pi}|x^{(N)})\}\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}\tilde{\theta}^{k}_{\pi}\right.\\
&\qquad+\frac{1}{2}\{\log h(\hat{\theta}_{\pi})\}\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}\\
&\qquad-\frac{1}{6}\{\sqrt{N}\partial_{ijk}\log f(\hat{\theta}_{\pi})\}\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}\tilde{\theta}^{k}_{\pi}+\frac{1}{24}l_{ijkl}(\hat{\theta}_{\pi}|x^{(N)})\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}\tilde{\theta}^{k}_{\pi}\tilde{\theta}^{l}_{\pi}\\
&\qquad+\frac{1}{2}\left[\frac{1}{2}\{\sqrt{N}\partial_{ij}\log f(\hat{\theta}_{\pi})\}\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}-\frac{1}{6}l_{ijk}(\hat{\theta}_{\pi}|x^{N})\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}\tilde{\theta}^{k}_{\pi}\right]\\
&\qquad\quad\left.\times\left[\frac{1}{2}\{\sqrt{N}\partial_{ij}\log f(\hat{\theta}_{\pi})\}\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}-\frac{1}{6}l_{ijk}(\hat{\theta}_{\pi}|x^{(N)})\tilde{\theta}^{i}_{\pi}\tilde{\theta}^{j}_{\pi}\tilde{\theta}^{k}_{\pi}\right]+\mathrm{O}_{\mathrm{P}}(N^{-3/2})\right)\\
&\quad\times\left\{\int p(x^{(N)}|\theta)\frac{\pi(\theta;N)}{|g^{X}(\theta)|^{\frac{1}{2}}}|g^{X}(\theta)|^{\frac{1}{2}}\mathrm{d}\theta\right\}^{-1}.
\end{aligned}
$$

We denote the $N^{-\frac{1}{2}}$-order, N^{-1}-order, and $N^{-3/2}$-order terms by $(N^{-1/2})a_0(\tilde{\theta}_\pi;\hat{\theta}_\pi)$, $(N^{-1})a_1(\tilde{\theta}_\pi;\hat{\theta}_\pi)$, and $(N^{-3/2})a_2(\tilde{\theta}_\pi;\hat{\theta}_\pi)$, respectively. Then, this expansion is rewritten as

$$\begin{aligned}
\pi(\theta|x^{(N)}) &= |g^X(\hat{\theta}_\pi)|^{\frac{1}{2}}\frac{\pi(\hat{\theta}_\pi)}{|g^X(\hat{\theta}_\pi)|^{\frac{1}{2}}}p(x^{(N)}|\hat{\theta}_\pi)\exp\left[-\frac{1}{2}\{-l_{ij}(\hat{\theta}_\pi|x^{(N)})\}\tilde{\theta}^i_\pi\tilde{\theta}^j_\pi\right] \\
&\quad\times\left[1+\frac{1}{\sqrt{N}}a_0(\tilde{\theta}_\pi;\hat{\theta}_\pi)\right. \\
&\quad\left.+\frac{1}{N}a_1(\tilde{\theta}_\pi;\hat{\theta}_\pi)+\frac{1}{N\sqrt{N}}a_2(\tilde{\theta}_\pi;\hat{\theta}_\pi)+\mathrm{O_P}(N^{-2})\right] \\
&\quad\times\left\{\int p(x^{(N)}|\theta)\frac{\pi(\theta;N)}{|g^X(\theta)|^{\frac{1}{2}}}|g^X(\theta)|^{\frac{1}{2}}\mathrm{d}\theta\right\}^{-1}.
\end{aligned}$$

To make the expansion easier to see, the following notations are used. Let $\phi(\eta;-l_{ij}(\hat{\theta}_\pi|x^{(N)}))$ be the probability density function of d-dimensional normal distribution with the precision matrix whose (i,j)-component is $-l_{ij}(\hat{\theta}_\pi|x^{(N)})$. Let $\eta=(\eta^1,\cdots,\eta^d)$ be a d-dimensional random vector distributed according to the normal density $\phi(\eta;-l_{ij}(\hat{\theta}_\pi|x^{(N)}))$ The notations $\bar{a}_0(\hat{\theta}_\pi)$, $\bar{a}_1(\hat{\theta}_\pi)$, $\bar{a}_2(\hat{\theta}_\pi)$, and $\hat{\omega}^{ij}(\hat{\theta}_\pi)$ denote the expectations of $a_0(\eta;\hat{\theta}_\pi)$, $a_1(\eta;\hat{\theta}_\pi)$, $a_2(\eta;\hat{\theta}_\pi)$, and $\eta^i\eta^j$, respectively.

Using the above notations, we get the following posterior expansion:

$$\begin{aligned}
\pi(\theta|x^{(N)}) &= \phi(\hat{\theta}_\pi;-l_{ij}(\hat{\theta}_\pi|x^{(N)})) \\
&\quad\times\left[1+\frac{1}{\sqrt{N}}\{a_0(\tilde{\theta}_\pi;\hat{\theta}_\pi)-\bar{a}_0(\hat{\theta}_\pi)\}\right. \\
&\quad+\frac{1}{N}\{a_1(\tilde{\theta}_\pi;\hat{\theta}_\pi)-\bar{a}_1(\hat{\theta}_\pi)\}-\frac{1}{N}\bar{a}_0(\hat{\theta}_\pi)\{a_0(\tilde{\theta}_\pi;\hat{\theta}_\pi)-\bar{a}_0(\hat{\theta}_\pi)\} \\
&\quad+\frac{1}{N\sqrt{N}}\{a_2(\tilde{\theta}_\pi;\hat{\theta}_\pi)-\bar{a}_2(\hat{\theta}_\pi)\}-\frac{1}{N\sqrt{N}}\bar{a}_0(\hat{\theta}_\pi)\{a_1(\tilde{\theta}_\pi;\hat{\theta}_\pi)-\bar{a}_1(\hat{\theta}_\pi)\} \\
&\quad-\frac{1}{N\sqrt{N}}\bar{a}_1(\hat{\theta}_\pi)\{a_0(\tilde{\theta}_\pi;\hat{\theta}_\pi)-\bar{a}_0(\hat{\theta}_\pi)\} \\
&\quad\left.+\frac{1}{N\sqrt{N}}\bar{a}_0^2(\hat{\theta}_\pi)\{a_1(\tilde{\theta}_\pi;\hat{\theta}_\pi)-\bar{a}_1(\hat{\theta}_\pi)\}+\mathrm{O_P}(N^{-2})\right]. \qquad (21)
\end{aligned}$$

Second, using (21), the Bayesian predictive density $q_\pi(y|x^{(N)})$ based on the prior $\pi(\theta;N)$ is expanded as

$$\begin{aligned}
&q_\pi(y|x^{(N)}) \\
&=\int q(y|\hat{\theta}_\pi)\left[1-\{\partial_i\log q(y|\hat{\theta}_\pi)\}\tilde{\theta}^i_\pi+\frac{1}{2}\frac{\partial_{ij}q(y|\hat{\theta}_\pi)}{q(y|\hat{\theta}_\pi)}\tilde{\theta}^i_\pi\tilde{\theta}^j_\pi+o_\mathrm{P}(N^{-1})\right]\pi(\theta|x^N)\mathrm{d}\theta \\
&=\int q(y|\hat{\theta}_\pi)\left[1+\{\partial_i\log|g^X(\hat{\theta}_\pi)|^{\frac{1}{2}}\}\{\partial_j\log q(y|\hat{\theta}_\pi)\}\tilde{\theta}^i_\pi\tilde{\theta}^j_\pi\right. \\
&\quad+\frac{1}{6}\{\partial_{ijk}\log p(x^{(N)}|\hat{\theta}_\pi)+\sqrt{N}\partial_{ijk}\log f(\hat{\theta}_\pi)\}\{\partial_l\log q(y|\hat{\theta}_\pi)\}\tilde{\theta}^i_\pi\tilde{\theta}^j_\pi\tilde{\theta}^k_\pi\tilde{\theta}^l_\pi \\
&\quad\left.+\frac{1}{2}\frac{\partial_{ij}q(y|\hat{\theta}_\pi)}{q(y|\hat{\theta}_\pi)}\tilde{\theta}^i_\pi\tilde{\theta}^j_\pi+\mathrm{o_P}(N^{-1})\right]\phi(\tilde{\theta}_\pi;-l_{ij}(\hat{\theta}_\pi|x^N))\mathrm{d}\tilde{\theta}_\pi \\
&=q(y|\hat{\theta}_\pi)+\hat{\omega}^{ij}(\hat{\theta}_\pi)\{\partial_i\log|g^X(\hat{\theta}_\pi)|^{\frac{1}{2}}\}\partial_j q(y|\hat{\theta}_\pi)+\frac{1}{2}\hat{\omega}^{ik}(\hat{\theta}_\pi)\hat{\omega}^{jl}(\hat{\theta}_\pi)l_{ijk}(\hat{\theta}_\pi|x^N)\partial_l q(y|\hat{\theta}_\pi) \\
&\quad+\frac{1}{2}\hat{\omega}^{ij}(\hat{\theta}_\pi)\partial_{ij}q(y|\hat{\theta}_\pi)+\mathrm{O_P}(N^{-3/2}). \qquad (22)
\end{aligned}$$

Here, the following two equations hold:

$$-l_{ij}(\hat{\theta}_\pi|x^{(N)}) = Ng^X_{ij}(\hat{\theta}_\pi)-\sqrt{N}\tilde{H}_{ij}(\hat{\theta}_\pi|x^N)+\mathrm{O_P}(1), \qquad (23)$$

$$l_{ijk}(\hat{\theta}_\pi|x^{(N)}) = -2N\overset{\mathrm{e}}{\Gamma}{}^{X}_{ij,k}(\hat{\theta}_\pi) - N\overset{\mathrm{m}}{\Gamma}{}^{X}_{ik,j}(\hat{\theta}_\pi) + \sqrt{N}\tilde{H}_{ijk}(\hat{\theta}|x^N). \tag{24}$$

By combining equation (23) with the Sherman–Morrison–Woodbury formula, the following expansion is obtained:

$$\hat{\omega}^{ij}(\hat{\theta}_\pi) = \frac{1}{N}g^{X,ij}(\hat{\theta}_\pi) + \frac{1}{N\sqrt{N}}g^{X,ik}(\hat{\theta}_\pi)g^{X,jl}(\hat{\theta}_\pi)H_{kl}(\hat{\theta}_\pi|x^{(N)}) + \mathrm{O_P}(N^{-2}). \tag{25}$$

By substituting equations (23), (24), and (25) into expansion (22), expansion (20) is obtained. □

Note that the integration of expansion (20) is 1 up to $\mathrm{O_P}(N^{-2})$. Further, expansion (20) is similar to the expansion in Komaki (1996). However, the estimator that is a center of the expansion is different because of the dependence of the prior on the sample size.

Lemma A5. *The Bayesian estimator $\hat{\theta}_{\mathrm{opt}}$ minimizing the Bayes risk $\int R(\theta, q(y|\hat{\theta}))\mathrm{d}\pi(\theta; N)$ among plug-in predictive densities is given by*

$$\begin{aligned}\hat{\theta}^i_{\mathrm{opt}} &= \hat{\theta}^i_\pi + \frac{1}{2N}g^{X,ij}(\hat{\theta}_\pi)T^X_j(\hat{\theta}_\pi) \\ &\quad + \frac{1}{2N}g^{X,jk}(\hat{\theta}_\pi)\left\{\overset{\mathrm{m}}{\Gamma}{}^{Y,i}_{jk}(\hat{\theta}_\pi) - \overset{\mathrm{m}}{\Gamma}{}^{X,i}_{jk}(\hat{\theta}_\pi)\right\} + \mathrm{O_P}(N^{-3/2}).\end{aligned} \tag{26}$$

Proof. The Bayes risk $\int R(\theta, q(y|\hat{\theta}))\mathrm{d}\pi(\theta; N)$ is decomposed as

$$\begin{aligned}\int R(\theta, q(y|\hat{\theta}))\mathrm{d}\pi(\theta; N) &= \int \pi(\theta; N)\int p(x^{(N)}|\theta)\int q(y|\theta)\log\frac{q(y|\theta)}{q_\pi(y|x^{(N)})}\mathrm{d}y\mathrm{d}x^{(N)}\mathrm{d}\theta \\ &\quad + \int \pi(\theta; N)\int p(x^{(N)}|\theta)\int q(y|\theta)\log\frac{q_\pi(y|x^{(N)})}{q(y|\hat{\theta})}\mathrm{d}y\mathrm{d}x^{(N)}\mathrm{d}\theta.\end{aligned}$$

The first term of this decomposition is not dependent on $\hat{\theta}$. From Fubini's theorem and Lemma A4, the proof is completed.

□

Using these lemmas, we prove Theorem 3.1. First, we find that the Kullback–Leibler risk of the plug-in predictive density with the estimator $\hat{\theta}_{\mathrm{opt}}$ defined in Lemma A5 is given by

$$\begin{aligned}R(\theta, q(y|\hat{\theta}_{\mathrm{opt}})) &= R(\theta, q(y|\hat{\theta}_\pi)) + \frac{1}{2N\sqrt{N}}\mathring{g}^{ij}(\theta)T^X_i(\theta)\partial_j\log f(\theta) \\ &\quad + \frac{1}{2N\sqrt{N}}g^{X,im}(\theta)g^Y_{ij}(\theta)g^{X,kl}(\theta) \\ &\quad \times\left\{\overset{\mathrm{m}}{\Gamma}{}^{Y,j}_{kl}(\theta) - \overset{\mathrm{m}}{\Gamma}{}^{X,j}_{kl}(\theta)\right\}\partial_m\log f(\theta).\end{aligned} \tag{27}$$

Using the expansion (27) and Lemma A3, we expand the Kullback–Leibler risk $R(\theta, q_\pi(y|x^{(N)}))$. Here, the risk $R(\theta, q_\pi(y|x^{(N)}))$ is equal to the risk $R(\theta, q(y|\hat{\theta}_{\mathrm{opt}}))$ up to $\mathrm{O}(N^{-2})$, because we expand the risk $R(\theta, q_\pi(y|x^{(N)}))$ as

$$R(\theta, q_\pi(y|x^{(N)})) = R(\theta, q(y|\hat{\theta}_{\mathrm{opt}})) + \mathrm{O}(N^{-2}). \tag{28}$$

Thus, we obtain expansion (1).

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, pp. 547–554.

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Springer–Verlag, New York.

Aslan, M. (2006). Asymptotically minimax Bayes predictive densities. *Ann. Statist.* **34**, pp. 2921–2938.

Bernardo, J. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, pp. 113–147.

Clarke, B. and Barron, A. (1994). Jeffreys prior is asymptotically least favorable under entropy risk. *J. Statist. Plann. Infer.* **41**, pp. 37–60.

Efron, B. (1975). Defining curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3**, pp. 1189–1372.

Frank, P. and Kiefer, J. (1951). Almost submminimax and biased minimax procedures. *Ann. Math. Statist.* **22**, pp. 465–468.

Fushiki, T., Komaki, F., and Aihara, K. (2004). On parametric bootstrapping and Bayesian prediction. *Scand. J. Statist.* **31**, pp. 403–416.

Ghosh, M. N. (1964). Uniform approximation of minimax point estimates. *Ann. Math. Statist.* **35**, pp. 1031–1047.

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

Hodges, J. L. and Lehmann, E. L. (1950). Some problems in minimax point estimation. *Ann. Math. Statist.* **21**, pp. 182–197.

Kanamori, T. and Shimodaira, H. (2003). Active learning algorithm using the maximum weighted log-likelihood estimator. *J. Statist. Plann. Infer.* **116**, pp. 149–162.

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

Komaki, F. (2011). Bayesian predictive densities based on latent information priors. *J. Statist. Plann. Infer.* **141**, pp. 3705–3715.

Komaki, F. (2012). Asymptotically minimax Bayesian predictive densities for multinomial models. *Electron. J. Statist.* **6**, pp. 934–957.

Komaki, F. (2013). Asymptotic properties of Bayesian predictive densities when the distributions of data and target variables are different. *submitted* .

Robbins, H. (1950). Asymptotically subminimax solutions of compound statistical decision problems. In *Proc. Second Berkley Symp. Math. Statist. Prob.* Univ. of California Press., pp. 131–148.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Infer.* **90**, pp. 227–244.

Suzuki, T. and Komaki, F. (2010). On prior selection and covariate shift of β-Bayesian prediction under α-divergence risk. *Commun. Statist. Theory* **39**, pp. 1655–1673.