Bernoulli **11**(2), 2005, 293–307

Nonparametric bootstrap prediction

TADAYOSHI FUSHIKI[1], FUMIYASU KOMAKI[2] and KAZUYUKI AIHARA[3,4]

[1]*Institute of Statistical Mathematics, 4-6-7, Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan. E-mail: fushiki@ism.ac.jp*
[2]*Department of Mathematical Informatics, Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan. E-mail: komaki@mist.i.u-tokyo.ac.jp*
[3]*Department of Complex Science and Engineering, Graduate School of Frontier Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan. E-mail: aihara@sat.t.u-tokyo.ac.jp*
[4]*ERATO Aihara Complexity Modelling Project, JST, 45-18 Oyama, Shibuya-ku, Tokyo 151-0065, Japan*

Ensemble learning has recently been intensively studied in the field of machine learning. 'Bagging' is a method of ensemble learning and uses bootstrap data to construct various predictors. The required prediction is then obtained by averaging the predictors. Harris proposed using this technique with the parametric bootstrap predictive distribution to construct predictive distributions, and showed that the parametric bootstrap predictive distribution gives asymptotically better prediction than a plug-in distribution with the maximum likelihood estimator. In this paper, we investigate nonparametric bootstrap predictive distributions. The nonparametric bootstrap predictive distribution is precisely that obtained by applying bagging to the statistical prediction problem. We show that the nonparametric bootstrap predictive distribution gives predictions asymptotically as good as the parametric bootstrap predictive distribution.

Keywords: asymptotic theory; bagging; bootstrap predictive distribution; information geometry; Kullback–Leibler divergence

1. Introduction

Let us suppose that observations $x^N = \{x_1, \ldots, x_N\}$ are independently distributed according to a distribution $p(x; \omega)$ that belongs to a statistical model

$$\{p(x; \omega) | \omega = (\omega^a) \in U, a = 1, \ldots, m\},$$

where U is a subset of m-dimensional Euclidean space. A future observation x_{N+1} is independent of x^N and has the same distribution $p(x_{N+1}; \omega)$. We predict x_{N+1} by a distribution $\hat{p}(x_{N+1}, x^N)$. The loss of a predictive distribution $\hat{p}(x_{N+1}, x^N)$ is measured by using the Kullback–Leibler divergence (Aitchison 1975)

$$D\{p(x_{N+1}; \omega) \| \hat{p}(x_{N+1}, x^N)\} = \int p(x_{N+1}; \omega) \log \frac{p(x_{N+1}; \omega)}{\hat{p}(x_{N+1}, x^N)} \mathrm{d}x_{N+1}.$$

1350–7265

The risk function is given by

$$\mathrm{E}_{x^N}[D\{p(x_{N+1};\,\omega)\|\hat{p}(x_{N+1},\,x^N)\}] = \int p(x^N;\,\omega)\int p(x_{N+1};\,\omega)\log\frac{p(x_{N+1};\,\omega)}{\hat{p}(x_{N+1},\,x^N)}\mathrm{d}x_{N+1}\mathrm{d}x^{\mathrm{N}}.$$

Methods of ensemble learning, such as bagging (Breiman 1996) and boosting (Freund and Schapire 1997), have recently been intensively studied in the field of machine learning. Breiman's bagging uses bootstrap data to construct various predictors. Prediction is obtained by averaging the predictors.

Harris (1989) proposed using the bagging technique (at that time unnamed) with the parametric bootstrap to construct bootstrap predictive distributions; in this paper, we call Harris's bootstrap predictive distributions parametric bootstrap predictive distributions to distinguish them from nonparametric bootstrap predictive distributions. He showed that the parametric bootstrap predictive distribution asymptotically dominates the estimative distribution that is a plug-in distribution with the maximum likelihood estimator (MLE) when the model is a one-parameter exponential family. Vidoni (1995) proposed an approximation of parametric bootstrap predictive distributions by using the p^*-formula. Fushiki *et al*. (2004) clarified the relationship between the parametric bootstrap distribution and Bayesian predictive distribution and showed that the parametric bootstrap predictive distribution asymptotically dominates the estimative distribution in general distributions

In this paper, we investigate nonparametric bootstrap predictive distributions constructed by using nonparametric bootstrapping. The nonparametric bootstrap predictive distribution is precisely that obtained by applying bagging to the statistical prediction problem.

The information-geometric framework (Amari 1985; Amari and Nagaoka 2000) briefly explained below is used to show the results in the present paper, as is done by Fushiki *et al*. (2004). In information geometry, a statistical model is considered as a manifold, which is called a statistical manifold. A Riemannian metric of a statistical manifold is given by the Fisher information matrix whose (a, b)th element is

$$g_{ab}(\omega) = \mathrm{E}\{\partial_a \log p(x;\,\omega)\partial_b \log p(x;\,\omega)\},$$

where ∂_a is an abbreviation for the derivative operator $\partial/\partial\omega^a$. The inverse matrix of $(g_{ab}(\omega))$ is written as $(g^{ab}(\omega))$. The e-connection coefficients and the m-connection coefficients are defined by

$$\overset{\mathrm{e}}{\Gamma}_{ab,c}(\omega) = \int \partial_a\partial_b \log p(x;\,\omega)\partial_c p(x;\,\omega)\mathrm{d}x$$

and

$$\begin{aligned}\overset{\mathrm{m}}{\Gamma}_{ab,c}(\omega) &= \int \partial_a\partial_b p(x;\,\omega)\partial_c \log p(x;\,\omega)\mathrm{d}x \\ &= \overset{\mathrm{e}}{\Gamma}_{ab,c}(\omega) + T_{abc}(\omega),\end{aligned}$$

respectively. Here,

$$T_{abc}(\omega) = \mathrm{E}\{\partial_a \log p(x;\, \omega)\partial_b \log p(x;\, \omega)\partial_c \log p(x;\, \omega)\}$$

is the skewness tensor. We can calculate curvatures of the manifold from the connection coefficients. In differential geometry, the manifold is said to be flat when connection coefficients vanish in some coordinate systems. If the model is an exponential family (or a mixture family) and the natural parameter (or the mixture parameter) is adopted as a coordinate system, the e-connection coefficients (or the m-connection coefficients) become 0. In information geometry, we say that the model is e-flat (or m-flat). Indices of connection coefficients are raised or lowered by the metric or the inverse. For example,

$$\overset{\mathrm{m}}{\Gamma}{}_{ab}^{c}(\omega) = \overset{\mathrm{m}}{\Gamma}_{ab,d}(\omega)g^{dc}(\omega), \qquad \overset{\mathrm{m}}{\Gamma}{}^{a}(\omega) = \overset{\mathrm{m}}{\Gamma}{}_{bc}^{a}(\omega)g^{bc}(\omega),$$

where Einstein's summation convention is used: if an index appears twice in any one term, once as an upper and once as a lower index, summation over the index is implied (see also McCullagh 1987). By using the information-geometric framework, Komaki (1996) showed that the predictive distribution obtained by adding 'a vector orthogonal to the model' to an estimative distribution dominates the estimative distribution.

This paper is organized as follows. We introduce nonparametric bootstrap predictive distributions in Section 2. In Section 3 an asymptotic expansion of nonparametric bootstrap predictive distributions is calculated. As a result, it is shown that the nonparametric bootstrap predictive distribution is equivalent to the parametric bootstrap predictive distribution up to second order. In Section 4 we show that the nonparametric bootstrap predictive distribution asymptotically dominates the estimative distribution. Some examples are given in Section 5, the paper concludes with a discussion.

2. Nonparametric bootstrap predictive distributions

Let $\hat{\omega}(x^N)$ be the MLE based on the observations x^N. We abbreviate $\hat{\omega}(x^N)$ to $\hat{\omega}$ when there is no ambiguity. The parametric bootstrap predictive distribution (Harris 1989) is defined by

$$p^{\dagger}(x_{N+1};\, \hat{\omega}(x^N)) = \mathrm{E}_{y^N}\{p(x_{N+1};\, \hat{\omega}(y^N))\} = \int p(x_{N+1};\, \hat{\omega}(y^N))p(y^N;\, \hat{\omega}(x^N))\mathrm{d}y^N.$$

Here, we introduce nonparametric bootstrap predictive distributions. Let $\hat{\omega}^*$ be the MLE calculated from a nonparametric bootstrap sample $x^{*N} = \{x_1^*, \ldots, x_N^*\}$ independently obtained from the empirical distribution

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - x_i).$$

Definition. *The nonparametric bootstrap predictive distribution is defined by*

$$p^*(x_{N+1};\, x^N) = \mathrm{E}_{\hat{p}}\{p(x_{N+1};\, \hat{\omega}^*)\} = \int p(x_{N+1};\, \hat{\omega}^*(x^{*N}))\hat{p}(x^{*N})\, \mathrm{d}x^{*N}. \tag{1}$$

A Monte Carlo estimate of the nonparametric bootstrap predictive distribution is obtained by the following procedure:

1. For $t = 1$ to T:
   (a) generate bootstrap data $x^{*N}_{(t)} = \{x^*_{(t),1}, \ldots, x^*_{(t),N}\}$ from the empirical distribution $\hat{p}(x)$;
   (b) calculate the MLE $\hat{\omega}^*_{(t)}$ from $x^{*N}_{(t)}$.
2. Output the predictive distribution

$$p^*_T(x_{N+1};\, x^N) = \frac{1}{T}\sum_{t=1}^{T} p(x_{N+1};\, \hat{\omega}^*_{(t)}). \tag{2}$$

This procedure is precisely the one obtained by applying bagging to the statistical prediction problem. When T tends to infinity, (2) converges to (1).

The nonparametric bootstrap predictive distribution does not necessarily belong to the model $\{p(x;\, \omega)\}$ like the Bayesian predictive distribution (Komaki 1996). We will show that the part that deviates from the model can be effectively used in prediction.

3. An asymptotic expansion of nonparametric bootstrap predictive distributions

3.1. Moments of the maximum likelihood estimator when the true distribution does not belong to the statistical model

In this subsection only, we assume that observations $x^N = \{x_1, \ldots, x_N\}$ are independently obtained from a distribution p_0 that does not necessarily belong to the statistical model $\{p(x;\, \omega) | \omega \in U\}$. Let ω_0 be the parameter of the distribution closest to p_0 in the model, that is,

$$\omega_0 = \operatorname*{argmin}_{\omega \in U} D\{p_0(x) \| p(x;\, \omega)\}.$$

The MLE $\hat{\omega}$ is given by

$$\hat{\omega} = \operatorname*{argmax}_{\omega \in U} \{\log p(x^N;\, \omega)\}.$$

We calculate the asymptotic moments of the MLE. Let us consider a normalized MLE $\tilde{\omega} = \sqrt{N}(\hat{\omega} - \omega_0)$. The first three moments of the normalized MLE are asymptotically given as follows:

$$\mathrm{E}_{p_0}(\tilde{\omega}^a) = \frac{1}{\sqrt{N}} k_2^a(\omega_0) + O(N^{-3/2}),$$

$$\mathrm{E}_{p_0}(\tilde{\omega}^a \tilde{\omega}^b) = s^{ab}(\omega_0) + O(N^{-1}),$$

$$\mathrm{E}_{p_0}(\tilde{\omega}^a \tilde{\omega}^b \tilde{\omega}^c) = O(N^{-1/2}).$$

Here,

$$s_{ab}(\omega) = J_{ac}(\omega) J_{bd}(\omega) I^{cd}(\omega),$$

$$k_2^a(\omega) = J^{ab}(\omega) J^{cd}(\omega) \Gamma_{bc,d}(\omega) + \frac{1}{2} J^{ab}(\omega) J^{ce}(\omega) J^{df}(\omega) I_{ef}(\omega) K_{bcd}(\omega),$$

$$I_{ab}(\omega) = \mathrm{E}_{p_0}\{\partial_a \log p(x; \omega) \partial_b \log p(x; \omega)\},$$

$$J_{ab}(\omega) = \mathrm{E}_{p_0}\{-\partial_a \partial_b \log p(x; \omega)\},$$

$$K_{abc}(\omega) = \mathrm{E}_{p_0}\{\partial_a \partial_b \partial_c \log p(x; \omega)\},$$

$$\Gamma_{ab,c}(\omega) = \mathrm{E}_{p_0}\{\partial_a \partial_b \log p(x; \omega) \partial_c \log p(x; \omega)\},$$

and $(J^{ab}(\omega))$, $(I^{ab}(\omega))$ and $(s^{ab}(\omega))$ are the inverse matrices of $(J_{ab}(\omega))$, $(I_{ab}(\omega))$ and $(s_{ab}(\omega))$, respectively.

3.2. An asymptotic expansion of nonparametric bootstrap predictive distributions

The moments of $\tilde{\omega}^* = \sqrt{N}(\hat{\omega}^* - \hat{\omega})$ can be obtained by replacing ω_0 by the MLE $\hat{\omega}$ and p_0 by the empirical distribution $\hat{p}$ in the moments of $\tilde{\omega}$ shown by the previous subsection. Then the moments of the normalized MLE calculated from a nonparametric bootstrap sample are given by

$$\mathrm{E}_{\hat{p}}(\tilde{\omega}^{*a}) = \frac{1}{\sqrt{N}} \overline{k}_2^{N,a}(\hat{\omega}) + O_p(N^{-3/2}),$$

$$\mathrm{E}_{\hat{p}}(\tilde{\omega}^{*a} \tilde{\omega}^{*b}) = \bar{s}^{N,ab}(\hat{\omega}) + O_p(N^{-1}),$$

$$\mathrm{E}_{\hat{p}}(\tilde{\omega}^{*a} \tilde{\omega}^{*b} \tilde{\omega}^{*c}) = O_p(N^{-1/2}).$$

Here,

$$\bar{s}^N_{ab}(\omega) = \bar{J}^N_{ac}(\omega)\bar{J}^N_{bd}(\omega)\bar{I}^{N,bd}(\omega),$$

$$\overline{k}^{N,a}_2(\omega) = \bar{J}^{N,ab}(\omega)\bar{J}^{N,cd}(\omega)\overline{\Gamma}^N_{bc,d}(\omega) + \frac{1}{2}\bar{J}^{N,ab}(\omega)\bar{J}^{N,ce}(\omega)\bar{J}^{N,df}(\omega)\bar{I}^N_{ef}(\omega)\overline{K}^N_{bcd}(\omega),$$

$$\bar{I}^N_{ab}(\omega) = \mathrm{E}_{\hat{p}}\{\partial_a \log p(x;\omega)\partial_b \log p(x;\omega)\} = \frac{1}{N}\sum_{\alpha=1}^{N}\partial_a \log p(x_\alpha;\omega)\partial_b \log p(x_\alpha;\omega),$$

and $\bar{J}^N_{ab}(\omega)$, $\overline{K}^N_{abc}(\omega)$ and $\overline{\Gamma}^N_{ab,c}(\omega)$ are defined in the same way. $(\bar{J}^{N,ab}(\omega))$, $(\bar{I}^{N,ab}(\omega))$ and $(\bar{s}^{N,ab}(\omega))$ are the inverse matrices of $(\bar{J}^N_{ab}(\omega))$, $(\bar{I}^N_{ab}(\omega))$ and $(\bar{s}^N_{ab}(\omega))$, respectively.

From the above, we can prove the following theorem.

Theorem 1. *The nonparametric bootstrap predictive distribution* $p^*(x_{N+1};x^N)$ *has the following third-order asymptotic expansion:*

$$p^*(x_{N+1};x^N) = p(x_{N+1};\hat{\omega}) + \frac{1}{N}\overline{k}^{N,a}_2(\hat{\omega})\partial_a p(x_{N+1};\hat{\omega}) + \frac{1}{2N}\bar{s}^{N,ab}(\hat{\omega})\partial_a\partial_b p(x_{N+1};\hat{\omega}) + O_p(N^{-2}). \tag{3}$$

Proof. Using the Taylor expansion, the theorem is easily obtained. □

According to Fushiki *et al*. (2004), an asymptotic expansion of the parametric bootstrap predictive distribution is given by

$$p^\dagger(x_{N+1};\hat{\omega}) = p(x_{N+1};\hat{\omega}) + \frac{1}{2N}g^{ab}(\hat{\omega})\left\{\partial_a\partial_b p(x_{N+1};\hat{\omega}) - \overset{\mathrm{m}}{\Gamma}{}^c_{ab}(\hat{\omega})\partial_c p(x_{N+1};\hat{\omega})\right\} + O_p(N^{-2}),$$

where the second term is 'the vector orthogonal to the model' (Komaki 1996). Since

$$\overline{k}^{N,a}_2(\hat{\omega}) = -\frac{\overset{\mathrm{m}}{\Gamma}{}^a(\hat{\omega})}{2} + O_{\mathrm{p}}(\mathrm{N}^{-1/2})$$

and

$$\bar{s}^{N,ab}(\hat{\omega}) = g^{ab}(\hat{\omega}) + O_p(N^{-1/2}),$$

the nonparametric bootstrap predictive distribution coincides with the parametric bootstrap predictive distribution up to second order.

4. Risk evaluation

In this section, we evaluate the prediction accuracy of the nonparametric bootstrap predictive distribution.

We use two lemmas to evaluate the risk. The following lemma is proved in the same way as the proof of the expectation lemma in Hartigan (1998).

Lemma 2. *The following relation holds:*

$$\int (\hat{\omega}^a - \omega^a)\Big\{ g_{ij}(\hat{\omega}) - \bar{I}^N_{ij}(\hat{\omega}) \Big\} p(x^N; \omega) \mathrm{d}x^N = O(N^{-3/2}).$$

Although $(\hat{\omega}^a - \omega^a)\{ g_{ij}(\hat{\omega}) - \bar{I}^N_{ij}(\hat{\omega})\} = O_p(N^{-1})$, the expectation of the $1/N$-order term of $(\hat{\omega}^a - \omega^a)\{ g_{ij}(\hat{\omega}) - \bar{I}^N_{ij}(\hat{\omega})\}$ vanishes. We can also prove that

$$\int (\hat{\omega}^a - \omega^a)\Big\{ g_{ij}(\hat{\omega}) - \bar{J}^N_{ij}(\hat{\omega}) \Big\} p(x^N; \omega) \mathrm{d}x^N = O(N^{-3/2}),$$

$$\int (\hat{\omega}^a - \omega^a)\Big\{ 3 \overset{\mathrm{m}}{\Gamma}_{(ij,k)}(\hat{\omega}) - 2T_{ijk}(\hat{\omega}) - \overline{K}^N_{ijk}(\hat{\omega}) \Big\} p(x^N; \omega) \mathrm{d}x^N = O(N^{-3/2}),$$

$$\int (\hat{\omega}^a - \omega^a)\Big\{ \overset{\mathrm{e}}{\Gamma}_{ij,k}(\hat{\omega}) - \overline{\Gamma}^N_{ij,k}(\hat{\omega}) \Big\} p(x^N; \omega) \mathrm{d}x^N = O(N^{-3/2}).$$

Here, the use of parentheses implies symmetrization with respect to the indices inside them, for example,

$$A_{(ijk)} = \frac{1}{3!}(A_{ijk} + A_{ikj} + A_{jik} + A_{jki} + A_{kij} + A_{kji}).$$

It is easy to prove that the above relation holds with respect to the inverse matrix $(\bar{I}^{N,ij}(\hat{\omega}))$.

Lemma 3. *The following relation holds:*

$$\int (\hat{\omega}^a - \omega^a)\big\{ g^{ij}(\hat{\omega}) - \bar{I}^{N,ij}(\hat{\omega}) \big\} p(x^N; \omega) \mathrm{d}x^N = O(N^{-3/2}).$$

We can also prove that

$$\int (\hat{\omega}^a - \omega^a)\big\{ g^{ij}(\hat{\omega}) - \bar{J}^{N,ij}(\hat{\omega}) \big\} p(x^N; \omega) \mathrm{d}x^N = O(N^{-3/2}).$$

Theorem 4. *The difference between the risk function of the estimative distribution* $p(x_{N+1}; \hat{\omega})$ *and that of the nonparametric bootstrap predictive distribution* $p^*(x_{N+1}; x^N)$ *is given by*

$$\mathrm{E}_{x^N}[D\{p(x_{N+1}; \omega) \| p(x_{N+1}; \hat{\omega})\} - D\{p(x_{N+1}; \omega) \| p^*(x_{N+1}; x^N)\}]$$
$$= \frac{1}{8N^2} \int \frac{1}{p(x_{N+1}; \omega)} \left[g^{ab}(\omega) \Big\{ \partial_a \partial_b p(x_{N+1}; \omega) - \overset{\mathrm{m}}{\Gamma}{}^c_{ab}(\omega) \partial_c p(x_{N+1}; \omega) \Big\} \right]^2 \mathrm{d}x_{N+1} + o(N^{-2}). \tag{4}$$

Therefore the nonparametric bootstrap predictive distribution $p^*(x_{N+1}; x^N)$ asymptotically dominates the estimative distribution $p(x_{N+1}; \hat{\omega})$.

Proof. The left-hand side of (4) can be rewritten as

$$\begin{aligned}
&\int p(x^N; \omega) \int p(x_{N+1}; \omega) \log \frac{p^*(x_{N+1}; x^N)}{p(x_{N+1}; \hat{\omega})} \mathrm{d}x_{N+1} \mathrm{d}x^N \\
&\quad = \int p(x^N; \omega) \int \{p(x_{N+1}; \omega) - p(x_{N+1}; \hat{\omega})\} \log \frac{p^*(x_{N+1}; x^N)}{p(x_{N+1}; \hat{\omega})} \mathrm{d}x_{N+1} \mathrm{d}x^N \\
&\qquad + \int p(x^N; \omega) \int p(x_{N+1}; \hat{\omega}) \log \frac{p^*(x_{N+1}; x^N)}{p(x_{N+1}; \hat{\omega})} \, \mathrm{d}x_{N+1} \mathrm{d}x^N. \qquad (5)
\end{aligned}$$

The second term of the right-hand side of (5) is

$$\begin{aligned}
&\int p(x^N; \omega) \int p(x_{N+1}; \hat{\omega}) \log \left\{ 1 + \frac{p^*(x_{N+1}; x^N) - p(x_{N+1}; \hat{\omega})}{p(x_{N+1}; \hat{\omega})} \right\} \mathrm{d}x_{N+1} \mathrm{d}x^N \\
&= \int p(x^N; \omega) \int p(x_{N+1}; \hat{\omega}) \left\{ \frac{p^*(x_{N+1}; x^N) - p(x_{N+1}; \hat{\omega})}{p(x_{N+1}; \hat{\omega})} \right\} \mathrm{d}x_{N+1} \mathrm{d}x^N \\
&\quad - \frac{1}{2} \int p(x^N; \omega) \int p(x_{N+1}; \hat{\omega}) \left\{ \frac{p^*(x_{N+1}; x^N) - p(x_{N+1}; \hat{\omega})}{p(x_{N+1}; \hat{\omega})} \right\}^2 \mathrm{d}x_{N+1} \mathrm{d}x^N + o(N^{-2}) \\
&= -\frac{1}{8N^2} \int \frac{1}{p(x_{N+1}; \omega)} \left[g^{ab}(\omega) \left\{ \partial_a \partial_b p(x_{N+1}; \omega) - \overset{\mathrm{m}}{\Gamma}{}^{c}_{ab}(\omega) \partial_c p(x_{N+1}; \omega) \right\} \right]^2 \mathrm{d}x_{N+1} + o(N^{-2}).
\end{aligned}$$

The first term on the right-hand side of (5) is expanded as

$$\begin{aligned}
&\int p(x^N; \omega) \int (\omega^a - \hat{\omega}^a) \partial_a p(x_{N+1}; \hat{\omega}) \log \frac{p^*(x_{N+1}; x^N)}{p(x_{N+1}; \hat{\omega})} \mathrm{d}x_{N+1} \mathrm{d}x^N \\
&\quad + \frac{1}{2} \int p(x^N; \omega) \int (\omega^a - \hat{\omega}^a)(\omega^b - \hat{\omega}^b) \partial_a \partial_b p(x_{N+1}; \hat{\omega}) \log \frac{p^*(x_{N+1}; x^N)}{p(x_{N+1}; \hat{\omega})} \mathrm{d}x_{N+1} \mathrm{d}x^N \\
&\quad + o(N^{-2}). \qquad (6)
\end{aligned}$$

The second term of (6) is

$$\frac{1}{4N^2} \int \frac{1}{p(x_{N+1}; \omega)} \left[g^{ab}(\omega) \left\{ \partial_a \partial_b p(x_{N+1}; \omega) - \overset{\mathrm{m}}{\Gamma}{}^{c}_{ab}(\omega) \partial_c p(x_{N+1}; \omega) \right\} \right]^2 \mathrm{d}x_{N+1} + o(N^{-2}).$$

It remains to evaluate the first term of (6).

The asymptotic expansion (3) of the nonparametric bootstrap predictive distribution can be rewritten as

$$\begin{aligned}\hat{p}^*(x_{N+1};\, x^N) = {} & p(x_{N+1};\, \hat{\omega}) + \frac{1}{2N} g^{ab}(\hat{\omega})\left\{\partial_a\partial_b p(x_{N+1};\, \hat{\omega}) - \overset{\mathrm{m}}{\Gamma}{}^{c}_{ab}(\hat{\omega})\partial_c p(x_{N+1};\, \hat{\omega})\right\} \\ & + \frac{1}{N}\left\{\overline{k}_2^{N,a}(\hat{\omega}) + \frac{\overset{\mathrm{m}}{\Gamma}{}^{a}(\hat{\omega})}{2}\right\}\partial_a p(x_{\mathrm{N}+1};\, \hat{\omega}) \\ & + \frac{1}{2N}\left\{\bar{s}^{N,ab}(\hat{\omega}) - g^{ab}(\hat{\omega})\right\}\partial_a\partial_b p(x_{N+1};\, \hat{\omega}) \\ & + O_p(N^{-2}), \end{aligned} \tag{7}$$

where the third and fourth terms of (7) are of third order.

Since the second-order term of the first term of (6) is 0 due to orthogonality, we only evaluate the third-order term:

$$\begin{aligned} & \int p(x^N;\, \omega)\int(\omega^a - \hat{\omega}^a)\partial_a p(x_{N+1};\, \hat{\omega})\log\frac{p^*(x_{N+1};\, x^N)}{p(x_{N+1};\, \hat{\omega})}\,\mathrm{d}x_{N+1}\mathrm{d}x^N \\ & = \int p(x^N;\, \omega)\int(\omega^c - \hat{\omega}^c)\frac{\partial_c p(x_{N+1};\, \hat{\omega})}{p(x_{N+1};\, \hat{\omega})}\left[\frac{1}{N}\left\{\overline{k}_2^{N,a}(\hat{\omega}) + \frac{\overset{\mathrm{m}}{\Gamma}{}^{a}(\hat{\omega})}{2}\right\}\partial_a p(x_{N+1};\, \hat{\omega})\right. \\ & \quad \left. + \frac{1}{2N}\left\{\bar{s}^{N,ab}(\hat{\omega}) - g^{ab}(\hat{\omega})\right\}\partial_a\partial_b p(x_{N+1};\, \hat{\omega})\right]\mathrm{d}x_{N+1}\mathrm{d}x^N + o(N^{-2}) \\ & = \frac{1}{N} g_{ab}(\omega)\int(\omega^b - \hat{\omega}^b)\left\{\overline{k}_2^{N,a}(\hat{\omega}) + \frac{\overset{\mathrm{m}}{\Gamma}{}^{a}(\hat{\omega})}{2}\right\}p(x^N;\, \omega)\mathrm{d}x^N \\ & \quad + \frac{1}{2N}\overset{\mathrm{m}}{\Gamma}{}_{ab,c}(\omega)\int(\omega^c - \hat{\omega}^c)\left\{\bar{s}^{N,ab}(\hat{\omega}) - g^{ab}(\hat{\omega})\right\}p(x^N;\, \omega)\mathrm{d}x^N + o(N^{-2}). \end{aligned} \tag{8}$$

Using Lemmas 2 and 3,

$$
\begin{aligned}
&\int (\omega^c - \hat{\omega}^c)\{\bar{s}^{N,ab}(\hat{\omega}) - g^{ab}(\hat{\omega})\}\, p(x^N;\, \omega)\mathrm{d}x^N \\
&\quad = \int (\omega^c - \hat{\omega}^c)\Big\{\bar{J}^{N,ai}(\hat{\omega})\bar{J}^{N,bj}(\hat{\omega})\bar{I}^N_{ij}(\hat{\omega}) - g^{ai}(\hat{\omega})g^{bj}(\hat{\omega})g_{ij}(\hat{\omega})\Big\}\, p(x^N;\, \omega)\mathrm{d}x^N \\
&\quad = \int (\omega^c - \hat{\omega}^c)\Big\{\bar{J}^{N,ai}(\hat{\omega})\bar{J}^{N,bj}(\hat{\omega})\bar{I}^N_{ij}(\hat{\omega}) - \bar{J}^{N,ai}(\hat{\omega})\bar{J}^{N,bj}(\hat{\omega})g_{ij}(\hat{\omega})\Big\}\, p(x^N;\, \omega)\mathrm{d}x^N \\
&\qquad + \int (\omega^c - \hat{\omega}^c)\{\bar{J}^{N,ai}(\hat{\omega})\bar{J}^{N,bj}(\hat{\omega})g_{ij}(\hat{\omega}) - \bar{J}^{N,ai}(\hat{\omega})g^{bj}(\hat{\omega})g_{ij}(\hat{\omega})\}\, p(x^N;\, \omega)\mathrm{d}x^N \\
&\qquad + \int (\omega^c - \hat{\omega}^c)\{\bar{J}^{N,ai}(\hat{\omega})g^{bj}(\hat{\omega})g_{ij}(\hat{\omega}) - g^{ai}(\hat{\omega})g^{bj}(\hat{\omega})g_{ij}(\hat{\omega})\}\, p(x^N;\, \omega)\mathrm{d}x^N \\
&\quad = g^{ai}(\omega)g^{bj}(\omega)\int (\hat{\omega}^c - \omega^c)\Big\{g_{ij}(\hat{\omega}) - \bar{I}^N_{ij}(\hat{\omega})\Big\}\, p(x^N;\, \omega)\mathrm{d}x^N \\
&\qquad + g^{ai}(\omega)g_{ij}(\omega)\int (\hat{\omega}^c - \omega^c)\{g^{bj}(\hat{\omega}) - \bar{J}^{N,bj}(\hat{\omega})\}\, p(x^N;\, \omega)\mathrm{d}x^N \\
&\qquad + g^{bj}(\omega)g_{ij}(\omega)\int (\hat{\omega}^c - \omega^c)\{g^{ai}(\hat{\omega}) - \bar{J}^{N,ai}(\hat{\omega})\}\, p(x^N;\, \omega)\mathrm{d}x^N + o(N^{-1}) \\
&\quad = o(N^{-1}).
\end{aligned}
$$

Therefore the second term of (8) is $o(N^{-2})$. We can prove the same thing for the first term of (8). Finally, the right-hand side of (8) is $o(N^{-2})$. □

Fushiki *et al*. (2004) evaluated the risk of the parametric bootstrap predictive distribution. From the result, the risks of both bootstrap predictive distributions are equal up to order $1/N^2$.

5. Numerical experiments

Example 1 *Normal distribution* $N(\mu, \sigma^2)$. The Fisher information matrix and the connection coefficients are given by

$$g_{\mu\mu} = \frac{1}{\sigma^2}, \qquad g_{\mu\sigma} = 0, \qquad g_{\sigma\sigma} = \frac{2}{\sigma^2},$$

$$\overset{\mathrm{m}}{\Gamma}_{\mu\mu,\mu} = \overset{\mathrm{m}}{\Gamma}_{\mu\sigma,\mu} = \overset{\mathrm{m}}{\Gamma}_{\mu\sigma,\sigma} = \overset{\mathrm{m}}{\Gamma}_{\sigma\mu,\mu} = \overset{\mathrm{m}}{\Gamma}_{\sigma\mu,\sigma} = \overset{\mathrm{m}}{\Gamma}_{\sigma\sigma,\mu} = 0, \qquad \overset{\mathrm{m}}{\Gamma}_{\mu\mu,\sigma} = \overset{\mathrm{m}}{\Gamma}_{\sigma\sigma,\sigma} = \frac{2}{\sigma^3},$$

$$\overset{\mathrm{e}}{\Gamma}_{\mu\mu,\mu} = \overset{\mathrm{e}}{\Gamma}_{\mu\sigma,\sigma} = \overset{\mathrm{e}}{\Gamma}_{\sigma\mu,\sigma} = \overset{\mathrm{e}}{\Gamma}_{\sigma\sigma,\mu} = \overset{\mathrm{e}}{\Gamma}_{\mu\mu,\sigma} 0, \qquad \overset{\mathrm{e}}{\Gamma}_{\mu\sigma,\mu} = \overset{\mathrm{e}}{\Gamma}_{\sigma\mu,\mu} = -\frac{2}{\sigma^3}, \qquad \overset{\mathrm{e}}{\Gamma}_{\sigma\sigma,\sigma} = -\frac{6}{\sigma^3}.$$

Here, $g_{ab}(\omega)$, $\overset{\mathrm{e}}{\Gamma}_{ab,c}(\omega)$ and $\overset{\mathrm{m}}{\Gamma}_{ab,c}(\omega)$ are abbreviated to g_{ab}, $\overset{\mathrm{e}}{\Gamma}_{ab,c}$ and $\overset{\mathrm{m}}{\Gamma}_{ab,c}$ respectively, and we use these abbreviations in the following. The improvement of risk is

$$\frac{3}{4N^2} + o(N^{-2}).$$

The comparison between the predictive performance of the parametric bootstrap predictive distribution and that of the nonparametric bootstrap predictive distribution is shown in Figure 1, where the true distribution is $N(0, 1)$ and 5000 bootstrap samples are used to calculate the nonparametric bootstrap predictive distribution. The loss function is calculated by numerical integration and the expectation of the loss is calculated by 10 000 Monte Carlo iterations. In

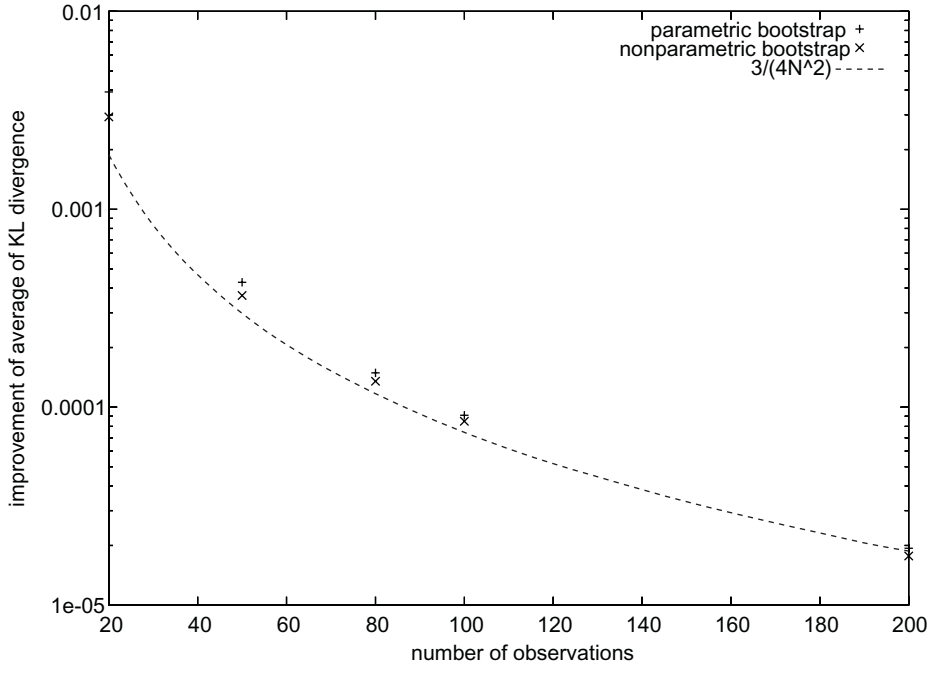

Figure 1. Comparison between the predictive performance of the parametric bootstrap predictive distribution and that of the nonparametric bootstrap predictive distribution when the model is $N(\mu, \sigma^2)$ and the true distribution is $N(0, 1)$. We set $T = 5000$. The loss function is calculated by numerical integration and the expectation of the loss is calculated by 10 000 Monte Carlo iterations.

this model, the parametric bootstrap predictive distribution showed slightly better performance than the nonparametric bootstrap predictive distribution.

Example 2 *Normal distribution* $N(\omega, \omega^2)$. The Fisher information and the connection coefficients are given by

$$g_{\omega\omega} = \frac{3}{\omega^2}, \qquad \overset{\mathrm{m}}{\Gamma}_{\omega\omega,\omega} = \frac{4}{\omega^3}, \qquad \overset{\mathrm{e}}{\Gamma}_{\omega\omega,\omega} = -\frac{10}{\omega^3}.$$

The improvement of risk is

$$\frac{19}{27N^2} + o(N^{-2}).$$

Figure 2 shows the result of the numerical experiment in this model. In the simulation, the true distribution is $N(1, 1)$ and 5000 bootstrap samples are used to calculate the nonparametric bootstrap predictive distribution. The loss function is calculated by numerical integration and the expectation of the loss is calculated by 10 000 Monte Carlo iterations.

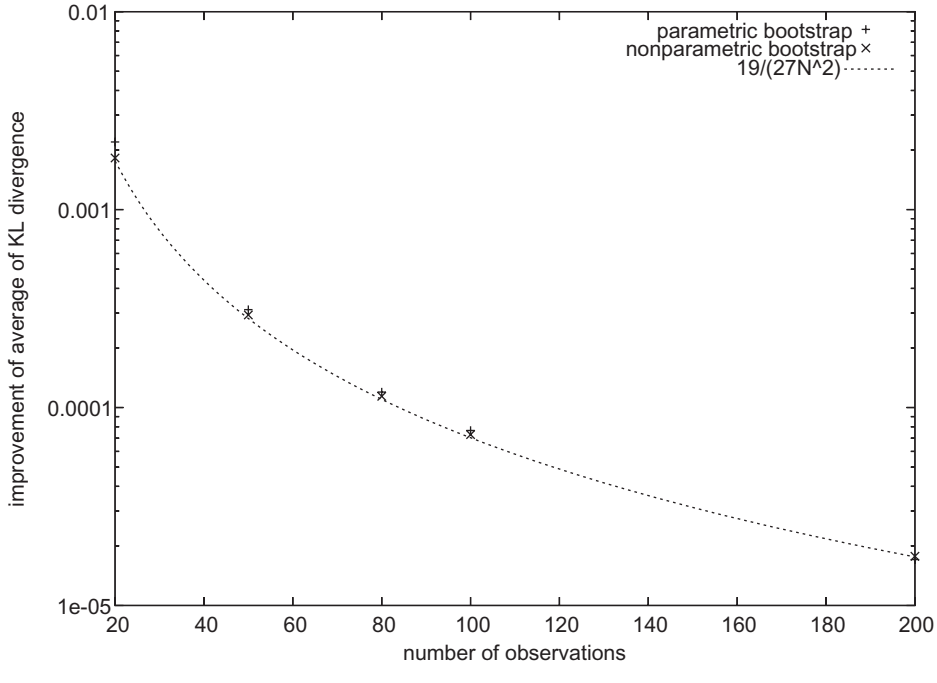

Figure 2. Comparison between the predictive performance of the parametric bootstrap predictive distribution and that of the nonparametric bootstrap predictive distribution when the model is $N(\omega, \omega^2)$ and the true distribution is $N(1, 1)$. We set $T = 5000$. The loss function is calculated by numerical integration and the expectation of the loss is calculated by 10 000 Monte Carlo iterations.

Example 3 *Poisson distribution* Po(λ). The Fisher information matrix and the connection coefficients are given by

$$g_{\lambda\lambda} = \frac{1}{\lambda}, \qquad \overset{\mathrm{e}}{\Gamma}{}^{\lambda}_{\lambda\lambda} = -\frac{1}{\lambda}, \qquad \overset{\mathrm{m}}{\Gamma}{}^{\lambda}_{\lambda\lambda} = 0.$$

The improvement of risk is

$$\frac{1}{4N^2} + o(N^{-2}).$$

Figure 3 shows the result of the numerical experiment in this model. In the simulation, the true distribution is Po(1) and 5000 bootstrap samples are used to calculate the nonparametric bootstrap predictive distribution. The loss function is calculated by numerical summation and the expectation of the loss is calculated by 100 000 Monte Carlo iterations.

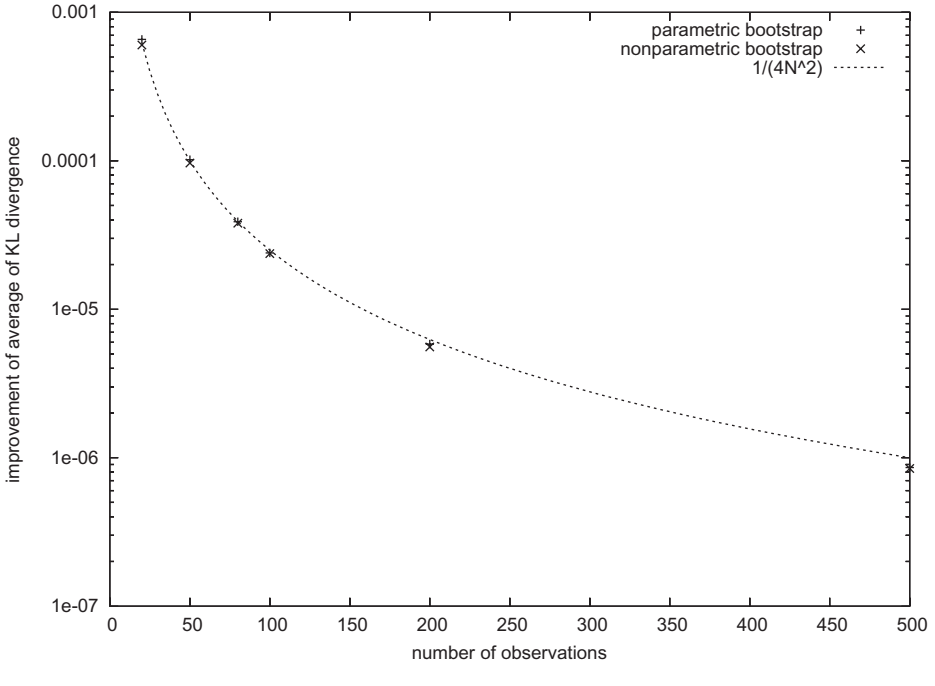

Figure 3. Comparison between the predictive performance of the parametric bootstrap predictive distribution and that of the nonparametric bootstrap predictive distribution when the model is Po(λ) and the true distribution is Po(1). We set $T = 5000$. The loss function is calculated by numerical summation and the expectation of the loss is calculated by 100 000 Monte Carlo iterations.

6. Discussion

We have investigated nonparametric bootstrap prediction by using asymptotic theory. Up to second order, the nonparametric bootstrap predictive distribution coincides with the parametric bootstrap predictive distribution. Up to order $1/N^2$, the risks of both bootstrap predictive distributions are same. In our numerical experiments, parametric bootstrap predictive distributions showed slightly better performance than nonparametric bootstrap predictive distributions. However, we have not calculated the higher-order risk analytically. On the other hand, from a computational viewpoint, the nonparametric bootstrap predictive distribution is preferable because it is difficult to generate random numbers according to $p(x; \hat{\omega})$ if $p(x; \hat{\omega})$ is not a simple distribution. In applications, it is important to evaluate the appropriate number of bootstraps. This is a problem for future investigation.

The prediction problem in the conditional setting is also important. This setting includes regression and classification where bagging is mainly used. Let $x = (y, z)$, where y is a response variable and z is a covariate. We assume that y has a condistional distribution $p(y|z; \omega)$ and z has a distribution $p(z)$. We consider the problem of predicting y_{N+1} based on data $x^N = \{(y_1, z_1), \ldots, (y_N, z_N)\}$. The risk function is defined by

$$\int \left[\int \left\{ \int p(y_{N+1}|z_{N+1}; \omega) \log \frac{p(y_{N+1}|z_{N+1}; \omega)}{\hat{p}(y_{N+1}|z_{N+1}, x^N)} \mathrm{d}y_{N+1} \right\} p(z_{N+1}) \mathrm{d}z_{N+1} \right] p(x^N; \omega) \mathrm{d}x^N,$$

which can be rewritten as

$$\int \left\{ \int \int p(y_{N+1}, z_{N+1}; \omega) \log \frac{p(y_{N+1}, z_{N+1}; \omega)}{\hat{p}(y_{N+1}|z_{N+1}, x^N) p(z_{N+1})} \mathrm{d}y_{N+1} \mathrm{d}z_{N+1} \right\} p(x^N; \omega) \mathrm{d}x^N.$$

If x^N and z_{N+1} are given, the conditional estimative distribution $p(y_{N+1}|z_{N+1}; \hat{\omega})$ and the conditional nonparametric bootstrap predictive distribution

$$p^*(y_{N+1}|z_{N+1}; x^N) = \mathrm{E}_{\hat{p}}\{p(y_{N+1}|z_{N+1}; \hat{\omega}^*)\}$$

do not depend on $p(z)$. Then

$$\begin{aligned} p^*(y_{N+1}, z_{N+1}; x^N) &= \mathrm{E}_{\hat{p}}\{p(y_{N+1}, z_{N+1}; \hat{\omega}^*)\} = \mathrm{E}_{\hat{p}}\{p(y_{N+1}|z_{N+1}; \hat{\omega}^*)\} p(z_{N+1}) \\ &= p^*(y_{N+1}|z_{N+1}; x^N) p(z_{N+1}). \end{aligned}$$

Therefore, the conditional nonparametric bootstrap predictive distribution asymptotically dominates the conditional estimative distribution.

Acknowledgements

The authors would like to thank Akimichi Takemura, two anonymous refrees and the asociate editor for helpful comments.

References

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

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

Amari, S. and Nagaoka, H. (2000) *Methods of Information Geometry*. New York: AMS and Oxford University Press.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123–140.

Freund, Y. and Schapire, R. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*, **55**, 119–139.

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

Harris, I.R. (1989) Predictive fit for natural exponential families. *Biometrika*, **76**, 675–684.

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

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

McCullagh, P. (1987) *Tensor Methods in Statistics*. London: Chapman & Hall.

Vidoni, P. (1995) A simple predictive density based on the p^*-formula. *Biometrika*, **82**, 855–863.

Received January 2003 and revised September 2004