Bernoulli **11**(4), 2005, 747–758

Bootstrap prediction and Bayesian prediction under misspecified models

TADAYOSHI FUSHIKI

Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan. E-mail: fushiki@ism.ac.jp

We consider a statistical prediction problem under misspecified models. In a sense, Bayesian prediction is an optimal prediction method when an assumed model is true. Bootstrap prediction is obtained by applying Breiman's 'bagging' method to a plug-in prediction. Bootstrap prediction can be considered to be an approximation to the Bayesian prediction under the assumption that the model is true. However, in applications, there are frequently deviations from the assumed model. In this paper, both prediction methods are compared by using the Kullback–Leibler loss under the assumption that the model does not contain the true distribution. We show that bootstrap prediction is asymptotically more effective than Bayesian prediction under misspecified models.

Keywords: bagging; Bayesian prediction; bootstrap; Kullback–Leibler divergence; misspecification; prediction

1. Introduction

In this paper, we consider a statistical prediction problem under misspecified models. Observations $x^N = \{x_1, \ldots, x_N\}$ are independent and identically distributed according to $q(x)$. The problem is the probabilistic prediction of a future observation x_{N+1} based on x^N. We assume a statistical model $\{p(x; \theta)|\theta = (\theta^a) \in \Theta, a = 1, \ldots, m\}$, where Θ is an open set in m-dimensional Euclidean space $\mathbb{R}^m$. The true distribution $q(x)$ is not necessarily in the assumed model $\{p(x; \theta)\}$. The performance of a predictive distribution $\hat{p}(x_{N+1}; x^N)$ is measured by using the Kullback–Leibler divergence, that is, we use the loss function

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

The risk function is

$$\mathrm{E}_{x^N}\{D(q(\cdot), \hat{p}(\cdot; x^N))\} = \int q(x^N) \left\{ \int q(x_{N+1}) \log \frac{q(x_{N+1})}{\hat{p}(x_{N+1}; x^N)} \, \mathrm{d}x_{N+1} \right\} \mathrm{d}x^N,$$

where $q(x^N) = \prod_{i=1}^N q(x_i)$.

For the statistical prediction problem, a predictive distribution often used naively is a plug-in distribution with the maximum likelihood estimator (MLE)

$$p(x_{N+1}; \hat{\theta}_{\mathrm{MLE}}(x^N)),$$

1350–7265

where

$$\hat{\theta}_{\mathrm{MLE}}(x^N) = \underset{\theta}{\mathrm{argmax}}\{\log p(x^N; \theta)\}.$$

Akaike's information criterion is derived from the viewpoint of minimizing the risk of the plug-in distribution with the MLE (Akaike 1973). Takeuchi's information criterion is a criterion for the plug-in distribution when the assumed model does not contain the true distribution (Takeuchi 1976).

Bayesian methods for prediction also have been considered (Geisser 1993). When the true distribution belongs to the statistical model $\{p(x; \theta)\}$, the Bayes risk with a proper prior distribution $\pi(\theta)$,

$$\mathrm{E}_\theta[\mathrm{E}_{x^N}\{D(p(\cdot; \theta), \hat{p}(\cdot; x^N)\}] = \int \pi(\theta)\left[\int p(x^N; \theta)\left\{\int p(x_{N+1}; \theta)\log\frac{p(x_{N+1}; \theta)}{\hat{p}(x_{N+1}; x^N)}\,\mathrm{d}x_{N+1}\right\}\mathrm{d}x^N\right]\mathrm{d}\theta,$$

is minimized by the Bayesian predictive distribution (Aitchison 1975)

$$p_\pi(x_{N+1}|x^N) = \int p(x_{N+1}; \theta)\pi(\theta|x^N)\mathrm{d}\theta,$$

where

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

is the posterior distribution. Komaki (1996) proved that the Bayesian predictive distribution includes 'the vector orthogonal to the model' and this vector is effective for prediction. Shimodaira (2000) evaluated the risk of the Bayesian predictive distribution when the true distribution does not belong to the assumed model.

In the field of machine learning, 'bagging' was proposed by Breiman (1996). Bagging provides a stable prediction by averaging many predictions based on bootstrap data. Bootstrap predictive distributions are derived by applying the bagging method to the plug-in distribution with the MLE (Harris 1989; Fushiki *et al.* 2004; 2005). Fushiki *et al.* (2004, 2005) investigated the relationship between the bootstrap predictive distribution and the Bayesian predictive distribution and evaluated the predictive performance.

In this paper, we investigate the bootstrap prediction under misspecified models. In Section 2 the predictive performance of the Bayesian predictive distribution and that of the bootstrap predictive distribution are compared. We show that the bootstrap predictive distribution asymptotically provides better prediction than the Bayesian predictive distribution. In Section 3 some examples are given; in particular, a regression problem is considered. Section 4 is a concluding discussion.

2. Bootstrap prediction under misspecified models

The bootstrap predictive distribution (Fushiki *et al.* 2005) is defined by

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

where $\hat{p}$ is the empirical distribution $\hat{p}(x) = (1/N)\sum_{i=1}^{N}\delta(x - x_i)$ and the bootstrap estimator $\hat{\theta}^*_{\mathrm{MLE}} = \hat{\theta}_{\mathrm{MLE}}(x^{*N})$ is the MLE based on a bootstrap sample x^{*N}. This predictive distribution is obtained by applying the bagging method to the plug-in distribution with the MLE. We investigate the bootstrap prediction under misspecified models.

This paper adopts the tensor notation. A partial derivative with respect to parameter θ^a is written as ∂_a. 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. For example,

$$f^a \partial_a p(x;\, \theta) = \sum_{a=1}^{m} f^a \frac{\partial p(x;\, \theta)}{\partial \theta^a}$$

(see also Amari and Nagaoka 2000; McCullagh 1987).

The following theorem provides an asymptotic expansion of the bootstrap predictive distribution.

Theorem 1. *The bootstrap predictive distribution is asymptotically expanded as*

$$p^*(x_{N+1};\, x^N) = p(x_{N+1};\, \hat{\theta}_{\mathrm{MLE}}) + \frac{1}{2N} h^{ac}(\hat{\theta}_{\mathrm{MLE}}) g_{cd}(\hat{\theta}_{\mathrm{MLE}}) h^{bd}(\hat{\theta}_{\mathrm{MLE}}) \partial_a \partial_b p(x_{N+1};\, \hat{\theta}_{\mathrm{MLE}})$$

$$+ \frac{1}{N} k_2^a(\hat{\theta}_{\mathrm{MLE}}) \partial_a p(x_{N+1};\, \hat{\theta}_{\mathrm{MLE}}) + o_p(N^{-1}),$$

where

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

$$h_{ab}(\theta) = \mathrm{E}_q\{-\partial_a \partial_b \log p(x;\, \theta)\},$$

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

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

$$K_{abc}(\theta) = \mathrm{E}_q\{-\partial_a \partial_b \partial_c \log p(x;\, \theta)\}$$

and $(g^{ab}(\theta))$ *and* $(h^{ab}(\theta))$ *are the inverse matrices of* $(g_{ab}(\theta))$ *and* $(h_{ab}(\theta))$*, respectively.*

Proof. By Theorem 1 in Fushiki *et al.* (2005), which remains valid when the true distribution does not belong to the assumed statistical model, the bootstrap predictive distribution is asymptotically expanded as

$$p^*(x_{N+1};\, x^N) = p(x_{N+1};\, \hat{\theta}_{\mathrm{MLE}}) + \frac{1}{N}\,\overline{k}_2^{N,a}(\hat{\theta}_{\mathrm{MLE}})\partial_a p(x_{N+1};\, \hat{\theta}_{\mathrm{MLE}}) + \frac{1}{2N}\overline{s}^{N,ab}(\hat{\theta}_{\mathrm{MLE}})\partial_a\partial_b p(x_{N+1};\, \hat{\theta}_{\mathrm{MLE}}) + O_p(N^{-2}),$$

with $\overline{s}^{N,ab}$ and $\overline{k}_2^{N,a}$ as defined in Fushiki *et al.* (2005). Since

$$\overline{s}^{N,ab}(\hat{\theta}_{\mathrm{MLE}}) = h^{ac}(\hat{\theta}_{\mathrm{MLE}})g_{cd}(\hat{\theta}_{\mathrm{MLE}})h^{bd}(\hat{\theta}_{\mathrm{MLE}}) + O_p(N^{-1/2})$$

and

$$\overline{k}_2^{N,a}(\hat{\theta}_{\mathrm{MLE}}) = k_2^a(\hat{\theta}_{\mathrm{MLE}}) + O_p(N^{-1/2}),$$

the theorem is obtained. □

The risk of the bootstrap predictive distribution is evaluated as follows. Let θ_0 be the closest parameter to the true distribution in the following sense:

$$\theta_0 = \operatorname*{argmin}_{\theta}\{D(q(\cdot),\, p(\cdot;\, \theta))\}.$$

We assume that such θ_0 exists uniquely in Θ. In the following, we also assume that $\mathrm{E}\{o_p(N^{-1})\} = o(N^{-1})$. The difference between the risk of the bootstrap predictive distribution and that of the plug-in distribution with the MLE is

$$\begin{aligned}
&\mathrm{E}_{x^N}\{D(q(\cdot),\, p(\cdot;\, \hat{\theta}_{\mathrm{MLE}}(x^N)))\} - \mathrm{E}_{x^N}\{D(q(\cdot),\, p^*(\cdot;\, x^N))\} \\
&\quad = \int q(x^N)\int q(x_{N+1})\log\frac{p^*(x_{N+1};\, x^N)}{p(x_{N+1};\, \hat{\theta}_{\mathrm{MLE}}(x^N))}\,\mathrm{d}x_{N+1}\,\mathrm{d}x^N \\
&\quad = \int q(x^N)\int q(x_{N+1})\frac{p^*(x_{N+1};\, x^N) - p(x_{N+1};\, \hat{\theta}_{\mathrm{MLE}}(x^N))}{p(x_{N+1};\, \hat{\theta}_{\mathrm{MLE}}(x^N))}\,\mathrm{d}x_{N+1}\,\mathrm{d}x^N + o(N^{-1}) \\
&\quad = \frac{1}{2N}\,h^{ac}(\theta_0)g_{cd}(\theta_0)h^{bd}(\theta_0)\int q(x_{N+1})\frac{\partial_a\partial_b p(x_{N+1};\, \theta_0)}{p(x_{N+1};\, \theta_0)}\,\mathrm{d}x_{N+1} \\
&\qquad + \frac{1}{2N}\,k_2^a(\theta_0)\int q(x_{N+1})\frac{\partial_a p(x_{N+1};\, \theta_0)}{p(x_{N+1};\, \theta_0)}\,\mathrm{d}x_{N+1} + o(N^{-1}),
\end{aligned} \tag{2}$$

where we use the well-known fact that $\hat{\theta}_{\mathrm{MLE}}$ converges to θ_0 in probability (White 1982). From the definition of θ_0,

$$\int q(x)\partial_a \log p(x;\, \theta_0)\mathrm{d}x = 0.$$

Then, the second term of (2) is 0. Using the relation

$$\int q(x)\frac{\partial_a\partial_b p(x;\,\theta)}{p(x;\,\theta)}\,\mathrm{d}x = \int q(x)\partial_a \log p(x;\,\theta)\partial_b \log p(x;\,\theta)\mathrm{d}x + \int q(x)\partial_a\partial_b \log p(x;\,\theta)\mathrm{d}x$$

$$= g_{ab}(\theta) - h_{ab}(\theta),$$

we can obtain the following theorem.

Theorem 2. *The risk of the bootstrap predictive distribution is asymptotically expanded as*

$$\mathrm{E}_{x^N}\{D(q(\cdot),\, p^*(\cdot;\, x^N))\} = \mathrm{E}_{x^N}\{D(q(\cdot),\, p(\cdot;\, \hat{\theta}_{\mathrm{MLE}}(x^N)))\}$$

$$-\frac{1}{2N}\{h^{ac}(\theta_0)g_{cd}(\theta_0)h^{bd}(\theta_0)g_{ab}(\theta_0) - h^{ab}(\theta_0)g_{ab}(\theta_0)\} + o(N^{-1}).$$

According to Shimodaira (2000), the risk of the Bayesian predictive distribution is given by

$$\mathrm{E}_{x^N}\{D(q(\cdot),\, p_\pi(\cdot|x^N))\} = \mathrm{E}_{x^N}\{D(q(\cdot),\, p(\cdot;\, \hat{\theta}_{\mathrm{MLE}}(x^N)))\} - \frac{1}{2N}\{h^{ab}(\theta_0)g_{ab}(\theta_0) - m\}$$

$$+ o(N^{-1}),$$

where $m = \dim(\theta)$.

Using the matrices $G(\theta) = (g_{ab}(\theta))$ and $H(\theta) = (h_{ab}(\theta))$, the above results can be written as

$$\mathrm{E}_{x^N}\{D(q(\cdot),\, p_\pi(\cdot|x^N))\} = \mathrm{E}_{x^N}\{D(q(\cdot),\, p(\cdot;\, \hat{\theta}_{\mathrm{MLE}}(x^N)))\}$$

$$-\frac{1}{2N}\{\mathrm{tr}[G(\theta_0)H^{-1}(\theta_0)] - m\} + o(N^{-1}),$$

$$\mathrm{E}_{x^N}\{D(q(\cdot),\, p^*(\cdot;\, x^N))\} = \mathrm{E}_{x^N}\{D(q(\cdot),\, p(\cdot;\, \hat{\theta}_{\mathrm{MLE}}(x^N)))\}$$

$$-\frac{1}{2N}\{\mathrm{tr}[G(\theta_0)H^{-1}(\theta_0)G(\theta_0)H^{-1}(\theta_0)] - \mathrm{tr}[G(\theta_0)H^{-1}(\theta_0)]\}$$

$$+ o(N^{-1}).$$

Since $H(\theta_0)$ is the Hessian matrix of $D(q(\cdot),\, p(\cdot;\, \theta))$ at θ_0, it is a symmetric positive definite matrix. We assume that $G(\theta_0)$ is also positive definite. Since $H(\theta_0)$ is positive definite, it can be written as UDU^{T}, where U is an orthogonal matrix, $D = \mathrm{diag}\{\delta_1, \ldots, \delta_m\}$, and $\delta_1, \ldots, \delta_m$ are the eigenvalues of $H(\theta_0)$. If we define $H^{1/2}(\theta_0) = UD^{1/2}U^{\mathrm{T}}$, then $H(\theta_0) = H^{1/2}(\theta_0)H^{1/2}(\theta_0)$ and $\mathrm{tr}[G(\theta_0)H^{-1}(\theta_0)] = \mathrm{tr}[H^{-1/2}(\theta_0)G(\theta_0)H^{-1/2}(\theta_0)]$. Let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $H^{-1/2}(\theta_0)G(\theta_0)H^{-1/2}(\theta_0)$; then $\lambda_1, \ldots, \lambda_m$ are real and positive because $H^{-1/2}(\theta_0)G(\theta_0)H^{-1/2}(\theta_0)$ is a symmetric positive definite matrix. Since

$$
\begin{aligned}
&\mathrm{E}_{x^N}\{D(q(\cdot),\ p_\pi(\cdot|x^N))\} - \mathrm{E}_{x^N}\{D(q(\cdot),\ p^*(\cdot;\ x^N))\} \\
&= \frac{1}{2N}\left\{\mathrm{tr}[G(\theta_0)H^{-1}(\theta_0)G(\theta_0)H^{-1}(\theta_0)] - \mathrm{tr}[G(\theta_0)H^{-1}(\theta_0)]\right\} \\
&\quad - \frac{1}{2N}\left\{\mathrm{tr}[G(\theta_0)H^{-1}(\theta_0)] - m\right\} + o(N^{-1}) \\
&= \frac{1}{2N}\left\{\mathrm{tr}[(H^{-1/2}(\theta_0)G(\theta_0)H^{-1/2}(\theta_0))^2] - 2\mathrm{tr}[H^{-1/2}(\theta_0)G(\theta_0)H^{-1/2}(\theta_0)] + m\right\} \\
&\quad + \ o(N^{-1}) \\
&= \frac{1}{2N}\sum_{i=1}^{m}\left(\lambda_i^2 - 2\lambda_i + 1\right) + o(N^{-1}) \\
&= \frac{1}{2N}\sum_{i=1}^{m}(\lambda_i - 1)^2 + o(N^{-1}),
\end{aligned}
$$

we can obtain the following theorem.

Theorem 3. *The bootstrap predictive distribution asymptotically provides better prediction than the Bayesian predictive distribution when* $G(\theta_0) \neq H(\theta_0)$.

3. Examples

Example 1 *Normal distribution with mean parameter.* On the true distribution $q(x)$, let

$$\mu_\mathrm{t} = \mathrm{E}_q(x), \qquad \sigma_\mathrm{t}^2 = \mathrm{var}_q(x).$$

We consider a statistical model $N(\mu, \sigma_\mathrm{m}^2)$, where σ_m is given. Then,

$$
\begin{aligned}
\mu_0 &= \underset{\mu}{\mathrm{argmin}}\{D(q(\cdot),\ p(\cdot;\ \mu))\} \\
&= \mu_\mathrm{t}
\end{aligned}
$$

and

$$G(\mu_0) = \frac{\sigma_\mathrm{t}^2}{\sigma_\mathrm{m}^4}, \qquad H(\mu_0) = \frac{1}{\sigma_\mathrm{m}^2}.$$

The difference between the risk of the bootstrap predictive distribution and that of the Bayesian predictive distribution is

$$\frac{1}{2N}\left\{G(\mu_0)H^{-1}(\mu_0) - 1\right\}^2 + o(N^{-1}) = \frac{1}{2N}\left(\frac{\sigma_\mathrm{t}^2}{\sigma_\mathrm{m}^2} - 1\right)^2 + o(N^{-1}).$$

Figure 1 shows the results of numerical experiments when the true distribution is $N(\mu_\mathrm{t}, \sigma_\mathrm{t}^2)$.

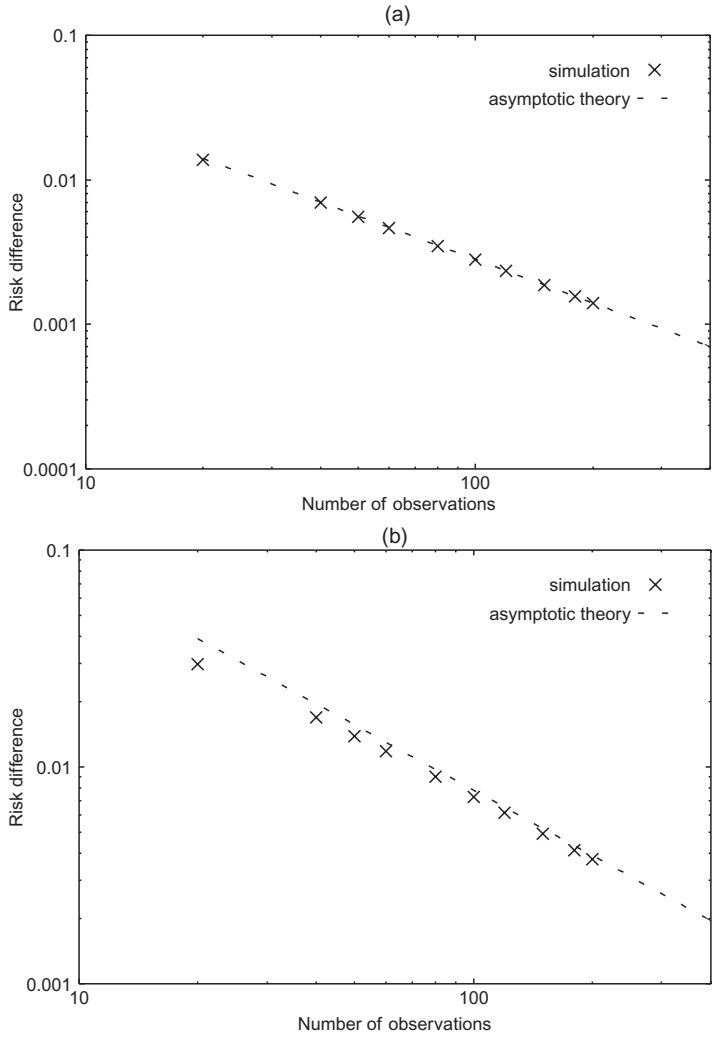

Figure 1. Difference between the risk of the bootstrap predictive distribution and that of the Bayesian predictive distribution. The true distribution is $N(0, \sigma_t^2)$ and the assumed model is $N(\mu, 1)$ One thousand bootstrap samples are used to calculate the bootstrap predictive distribution. The loss function is calculated by numerical integration and the expectation of the loss is calculated by 10 000 Monte Carlo iterations. (a) $\sigma_t = 0.5$, (b) $\sigma_t = 1.5$.

The uniform prior on μ is used to calculate the Bayesian predictive distribution. It is known that the Bayesian predictive distribution with the uniform prior dominates the plug-in distribution with the MLE when this assumed normal model is true. Although the uniform prior is improper, the Bayesian predictive distribution with the uniform prior is admissible and natural from the viewpoint of invariance. However, as shown in Figure 1, the bootstrap

predictive distribution is more effective than the Bayesian predictive distribution under the misspecified model.

Example 2 *Gamma distribution*. Let

$$p(x;\lambda)=\frac{\lambda^r}{\Gamma(r)}x^{r-1}\exp(-\lambda x),$$

where r is fixed. Then,

$$\lambda_0=\operatorname*{argmin}_{\lambda}\{D(q(\cdot),\,p(\cdot;\lambda))\}=\frac{r}{\mathrm{E}_q(x)}$$

and

$$G(\lambda_0)=\mathrm{var}_q(x),\qquad H(\lambda_0)=\frac{r}{\lambda_0^2}.$$

The difference between the risk of the bootstrap predictive distribution and that of the Bayesian predictive distribution is

$$\frac{1}{2N}\left\{G(\lambda_0)H^{-1}(\lambda_0)-1\right\}^2+o(N^{-1})=\frac{1}{2N}\left\{\frac{r\,\mathrm{var}_q(x)}{\mathrm{E}_q(x)^2}-1\right\}^2+o(N^{-1}).$$

Figure 2 shows the results of numerical experiments when the true distribution is a lognormal distribution. In Figure 2, the Jeffreys prior $\pi(\lambda)\propto\lambda^{-1}$ is used to calculate the Bayesian predictive distribution.

3.1. Conditional prediction

We now turn to a prediction problem in the conditional setting. This setting includes regression and classification where bagging is mainly used. Let $z=(x,\,y)$, where y is a response variable and x is a covariate. We consider the problem of predicting y_{N+1} given x_{N+1} based on data $z^N=\{(x_1,\,y_1),\ldots,(x_N,\,y_N)\}$. The true distribution is $q(x,\,y)=q(y|x)q(x)$, and a conditional model $\{p(y|x;\theta)\}$ is assumed. The loss function of a predictive distribution $\hat{p}(y_{N+1}|x_{N+1};\,z^N)$ is given by

$$\int q(x_{N+1})\int q(y_{N+1}|x_{N+1})\log\frac{q(y_{N+1}|x_{N+1})}{\hat{p}(y_{N+1}|x_{N+1};\,z^N)}\,\mathrm{d}y_{N+1}\,\mathrm{d}x_{N+1}.$$

The risk function is

$$\int q(z^N)\int q(x_{N+1})\int q(y_{N+1}|x_{N+1})\log\frac{q(y_{N+1}|x_{N+1})}{\hat{p}(y_{N+1}|x_{N+1};\,z^N)}\,\mathrm{d}y_{N+1}\,\mathrm{d}x_{N+1}\,\mathrm{d}z^N. \tag{3}$$

The conditional bootstrap predictive distribution is defined by

$$p^*(y_{N+1}|x_{N+1};\,z^N)=\mathrm{E}_{\hat{p}}\{p(y_{N+1}|x_{N+1};\,\hat{\theta}^*_{\mathrm{MLE}})\}. \tag{4}$$

Here, $\hat{\theta}^*_{\mathrm{MLE}}$ does not depend on $p(x)$, which is a statistical model of x. Then, the conditional

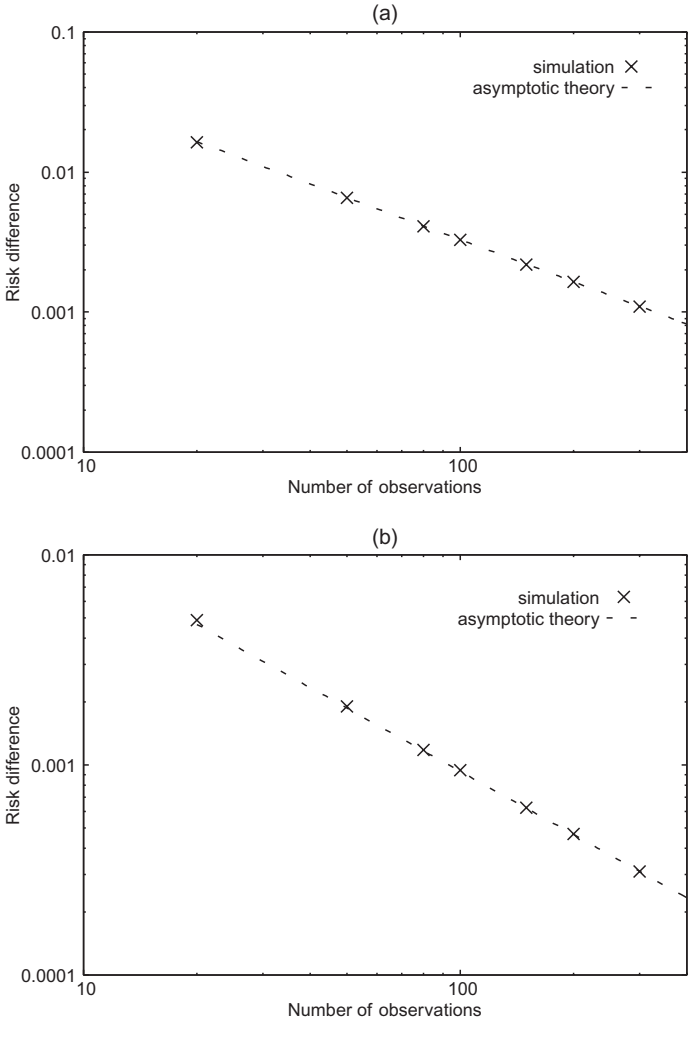

Figure 2. Difference between the risk of the bootstrap predictive distribution and that of the Bayesian predictive distribution. The true distribution is $q(x) = 1/(\sqrt{2\pi\sigma_t^2}x)\exp[-\{\log(x) - \mu_t\}^2/(2\sigma_t^2)]$ and the assumed model is $p(x; \lambda) = \lambda^2 x \exp(-\lambda x)$. One thousand bootstrap samples are used to calculate the 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. (a) $(\mu_t, \sigma_t) = (0.5, 0.3)$, (b) $(\mu_t, \sigma_t) = (0.3, 0.5)$.

bootstrap predictive distribution does not depend on $p(x)$. The Bayesian predictive distribution

$$p_\pi(y_{N+1}|x_{N+1}, z^N) = \int p(y_{N+1}|x_{N+1}; \theta)\pi(\theta|z^N)\mathrm{d}\theta$$

also does not depend on $p(x)$ because the posterior

$$\pi(\theta|z^N) \propto p(y^N|x^N;\, \theta)\pi(\theta).$$

does not depend on $p(x)$. Then the results in the previous section remain valid in the conditional setting.

Example 3 *Linear regression*. On the true distribution $q(x, y)$, let

$$\mu(x) = \mathrm{E}_q(Y|X = x), \qquad \sigma^2(x) = \mathrm{E}_q\{(Y - \mathrm{E}_q[Y|X = x])^2|X = x\}.$$

We consider a statistical model

$$y = ax + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_0^2),$$

where σ_0^2 is given and a is the only parameter. Then

$$a_0 = \operatorname*{argmin}_a \left(\iint q(y, x) \log \frac{q(y|x)}{p(y|x;\, a)} \, \mathrm{d}y \, \mathrm{d}x \right) = \frac{\mathrm{E}_q\{x\mu(x)\}}{\mathrm{E}_q(x^2)}$$

and

$$G(a_0) = \mathrm{E}_q\{x^2\mu(x)^2 + x^2\sigma^2(x) - 2a_0x^3\mu(x) + a_0^2x^4\}, \qquad H(a_0) = \frac{\mathrm{E}_q(x^2)}{\sigma_0^2}.$$

In particular, when $x \sim U(0, 1)$ and $y|x \sim N(x^\alpha, \sigma_0^2)$, the difference between the risk of the bootstrap predictive distribution and that of the Bayesian predictive distribution is

$$\frac{9(\alpha - 1)^4(5\alpha + 8)^2}{50(\alpha + 2)^4(\alpha + 4)^2(2\alpha + 3)^2\sigma_0^4 N} + o(N^{-1}).$$

The results of numerical experiments are shown in Figure 3.

4. Discussion

We have considered the predictive performance of the bootstrap prediction under misspecified models.

When the true distribution belongs to an assumed model, the Bayesian predictive distribution with an appropriate prior dominates the plug-in distribution and is considered as an optimal predictive distribution in some sense (Aitchison 1975). The bootstrap predictive distribution can be considered to be an approximation to the Bayesian predictive distribution and asymptotically dominates the plug-in distribution with the MLE (Fushiki *et al.* 2004; 2005).

However, when the assumed model does not include the true distribution, it was shown that the bootstrap predictive distribution is asymptotically more effective than the Bayesian predictive distribution. This result can be understood in the following way. Each bootstrap estimator is obtained based on a random sample from the empirical distribution which is close to the true distribution, and then the variance of the bootstrap estimator is

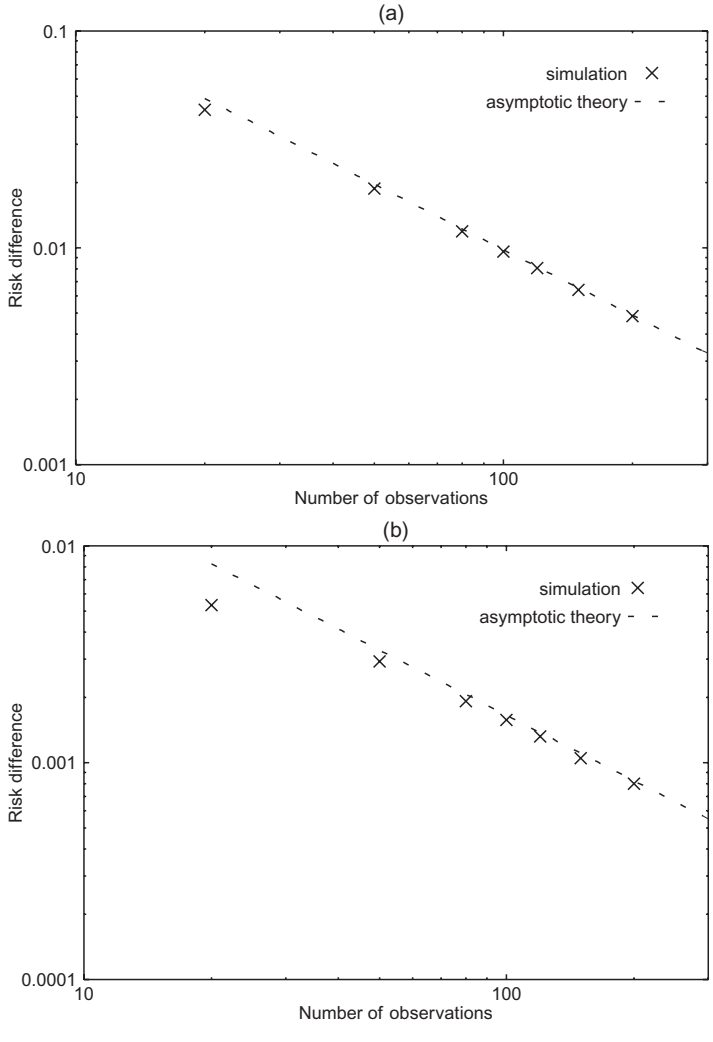

Figure 3. Difference between the risk of the bootstrap predictive distribution and that of the Bayesian predictive distribution. The true structure is $y = x^{\alpha} + \varepsilon$, $\varepsilon \sim N(0, 0.01)$, $x \sim U(0, 1)$, and the assumed model is $y = ax + \varepsilon$, $\varepsilon \sim N(0, 0.01)$. The uniform prior on a is used to calculate the Bayesian predictive distribution. One thousand bootstrap samples are used to calculate the bootstrap predictive distribution. To calculate the risk function, 10 0000 Monte Carlo samples are used. (a) $\alpha = 0.5$, (b) $\alpha = 1.5$.

asymptotically the variance of the MLE in misspecified models and contains information on misspecification. On the other hand, the asymptotic variance of the posterior distribution does not have enough information on misspecification because the posterior distribution $\pi(\theta|x^N)(\propto p(x^N; \theta)\pi(\theta))$ strongly depends on the assumed statistical model.

The asymptotic risk of the Bayesian predictive distribution in Shimodaira (2000) is valid when $\log \pi(\theta)$ is not too large. When there is strong prior information, $\log \pi(\theta)$ becomes large. In such a case, the Bayesian prediction is quite different from the plug-in prediction and sometimes works very well. We think that the bootstrap predictive distribution is more effective than the Bayesian predictive distribution if there is no strong prior information and the correctness of the assumed model is suspected. When model misspecification is large, the difference between the risk of the bootstrap predictive distribution and that of the Bayesian predictive distribution is not important. However, in real problems, such a model will not be used, and a more appropriate model will be explored. We think that the result in this paper is meaningful in realistic situations where the assumed model is misspecified 'moderately', but we have not confirmed the effectiveness of the results in real problems. This will be the subject of future work. In Section 3, we analysed the risk difference by means of numerical experiments. Low-dimensional models were used to confirm the theory. In high-dimensional models such as are needed for real problems, the difference is considered to be much larger.

Acknowledgements

I would like to thank Fumiyasu Komaki, Shinto Eguchi, Hironori Fujisawa, two anonymous referees and the editor for helpful comments.

References

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

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csáki (eds), *Proceedings of the Second International Symposium on Information Theory*, pp. 267–281, Budapest: Akadémiai Kiadó.

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.

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

Fushiki, T., Komaki, F. and Aihara, K. (2005) Nonparametric bootstrap prediction. *Bernoulli*, **11**, 293–307.

Geisser, S. (1993) *Predictive Inference*: *An Introduction*. New York: Chapman & Hall.

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

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

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

Shimodaira, H. (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function, *J. Statist. Plann. Inference*, **90**, 227–240.

Takeuchi, K. (1976) Distributions of information statistics and criteria for adequacy of models (in Japanese). *Math. Sci.*, **153**, 12–18.

White, H. (1982) Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 1–26.

Received September 2004 revised March 2005