Journal of Statistical Planning and Inference 90 (2000) 227–244

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Improving predictive inference under covariate shift by weighting the log-likelihood function

Hidetoshi Shimodaira *

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan

Received 17 December 1998; received in revised form 21 January; accepted 25 February 2000

Abstract

A class of predictive densities is derived by weighting the observed samples in maximizing the log-likelihood function. This approach is effective in cases such as sample surveys or design of experiments, where the observed covariate follows a different distribution than that in the whole population. Under misspecification of the parametric model, the optimal choice of the weight function is asymptotically shown to be the ratio of the density function of the covariate in the population to that in the observations. This is the pseudo-maximum likelihood estimation of sample surveys. The optimality is defined by the expected Kullback–Leibler loss, and the optimal weight is obtained by considering the importance sampling identity. Under correct specification of the model, however, the ordinary maximum likelihood estimate (i.e. the uniform weight) is shown to be optimal asymptotically. For moderate sample size, the situation is in between the two extreme cases, and the weight function is selected by minimizing a variant of the information criterion derived as an estimate of the expected loss. The method is also applied to a weighted version of the Bayesian predictive density. Numerical examples as well as Monte-Carlo simulations are shown for polynomial regression. A connection with the robust parametric estimation is discussed. © 2000 Elsevier Science B.V. All rights reserved.

MSC: 62B10; 62D05

Keywords: Akaike information criterion; Design of experiments; Importance sampling; Kullback–Leibler divergence; Misspecification; Sample surveys; Weighted least squares

1. Introduction

Let x be the explanatory variable or the covariate, and y be the response variable. In predictive inference with the regression analysis, we are interested in estimating the conditional density $q(y|x)$ of y given x, using a parametric model. Let $p(y|x,\theta)$ be the model of the conditional density which is parameterized by $\theta=(\theta^1,\dots,\theta^m)' \in \Theta \subset \mathscr{R}^m$.

* Correspondence address: Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford University, Stanford, CA 94305-4065, USA.

E-mail address: shimo@ism.ac.jp (H. Shimodaira).

0378-3758/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0378-3758(00)00115-4

Having observed i.i.d. samples of size n, denoted by $(x^{(n)}, y^{(n)}) = ((x_t, y_t)\colon t = 1, \ldots, n)$, we obtain a predictive density $p(y|x, \hat{\theta})$ by giving an estimate $\hat{\theta} = \hat{\theta}(x^{(n)}, y^{(n)})$. In this paper, we discuss improvement of the maximum likelihood estimate (MLE) under both (i) *covariate shift* in distribution and (ii) *misspecification* of the model as explained below.

Let $q_1(x)$ be the density of x for evaluation of the predictive performance, while $q_0(x)$ be the density of x in the observed data. We consider the Kullback–Leibler loss function

$$\mathrm{loss}_i(\theta) := -\int q_i(x) \int q(y|x) \log p(y|x, \theta)\, \mathrm{d}y\, \mathrm{d}x$$

for $i = 0, 1$, and then employ $\mathrm{loss}_1(\hat{\theta})$ for evaluation of $\hat{\theta}$, rather than the usual $\mathrm{loss}_0(\hat{\theta})$. The situation $q_0(x) \neq q_1(x)$ will be called covariate shift in distribution, which is one of the premises of this paper.

This situation is not so odd as it might look at first. In fact, it is seen in various fields as follows. In sample surveys, $q_0(x)$ is determined by the sampling scheme, while $q_1(x)$ is determined by the population. In regression analysis, covariate shift often happens because of the limitation of resources, or the design of experiments. In artificial neural networks literature, "active learning" is the typical situation where we control $q_0(x)$ for better prediction. We could say that the distribution of x in future observations is different from that of the past observations; x is not necessarily distributed as $q_1(x)$ in future, but we can give imaginary $q_1(x)$ to specify the region of x where the prediction accuracy should be controlled. Note that $q_0(x)$ and/or $q_1(x)$ are often estimated from data, but we assume they are known or estimated reasonably in advance.

The second premise of this paper is misspecification of the model. Let $\hat{\theta}_0$ be the MLE of θ, and θ_0^* be the asymptotic limit of $\hat{\theta}_0$ as $n \to \infty$. Under certain regularity conditions, MLE is consistent and $p(y|x, \theta_0^*) = q(y|x)$ provided that the model is correctly specified. In practice, however, $p(y|x, \theta_0^*)$ deviates more or less from $q(y|x)$.

Under both the covariate shift and the misspecification, MLE does not necessarily provide a good inference. We will show that MLE is improved by giving a weight function $w(x)$ of the covariate in the log-likelihood function

$$L_w(\theta|x^{(n)}, y^{(n)}) := -\sum_{t=1}^{n} l_w(x_t, y_t|\theta), \tag{1.1}$$

where $l_w(x, y|\theta) = -w(x)\log p(y|x, \theta)$. Then the maximum weighted log-likelihood estimate (MWLE), denoted by $\hat{\theta}_w$, is obtained by maximizing (1.1) over Θ. It will be seen that the weight function $w(x) = q_1(x)/q_0(x)$ is the optimal choice for sufficiently large n in terms of the expected loss with respect to $q_1(x)$. We denote MWLE with this weight function by $\hat{\theta}_1$. A comparison between $\hat{\theta}_0$ and $\hat{\theta}_1$ is made in the numerical example of polynomial regression of Section 2, and the asymptotic optimality of $\hat{\theta}_1$ is shown in Section 3. Note that MWLE turns out to be downweighting the observed samples which are not important in fitting the model with respect to the population. An interpretation of MWLE as one of the robust estimation techniques is given in Section 9.

This type of estimation is not new in statistics. Actually, $\hat{\theta}_1$ is regarded as a generalization of the pseudo-maximum likelihood estimation in sample surveys (Skinner et al., 1989, p. 80; Pfeffermann et al., 1998); the log likelihood is weighted inversely proportional to $q_0(x)$, the probability of selecting unit x, while $q_1(x)$ is equal probability for all possible values of x. The same idea is also seen in Rao (1991), where weighted maximum likelihood estimation is considered for unequally spaced time-series data.

The local likelihoods or the weighted likelihoods formally similar to (1.1) are found in the literature for semi-parametric inference. However, $\hat{\theta}_w$ is estimated using a weight function concentrated locally around each x or (x, y) in the semi-parametric approach; thus $\hat{\theta}_w$ in $p(y|x, \hat{\theta}_w)$ will depend on (x, y) as well as the data $(x^{(n)}, y^{(n)})$. On the other hand, we restrict our attention to a rather conventional parametric modeling approach here, and $\hat{\theta}_w$ depends only on the data.

In spite of the asymptotic optimality of $w(x) = q_1(x)/q_0(x)$ mentioned above, another choice of the weight function can improve the expected loss for moderate sample size by compromising the bias and the variance of $\hat{\theta}_w$. We develop a practical method for this improvement in Sections 4–7. The asymptotic expansion of the expected loss is given in Section 4, and a variant of the information criterion is derived as an estimate of the expected loss in Section 5. This new criterion is used to find a good $w(x)$ as well as a good form of $p(y|x, \theta)$. The numerical example is revisited in Section 6, and a simulation study is given in Section 7.

In Section 8, we show the Bayesian predictive density is also improved by considering the weight function. Finally, concluding remarks are given in Section 9. All the proofs are deferred to the appendix.

2. Illustrative example in regression

Here we consider the normal regression to predict the response $y \in \mathscr{R}$ using a polynomial function of $x \in \mathscr{R}$. Let the model $p(y|x, \theta)$ be the polynomial regression

$$y = \beta_0 + \beta_1 x + \cdots + \beta_d x^d + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2), \tag{2.1}$$

where $\theta = (\beta_0, \ldots, \beta_d, \sigma)$ and $\mathrm{N}(a, b)$ denotes the normal distribution with mean a and variance b. In the numerical example below, we assume the true $q(y|x)$ is also given by (2.1) with $d = 3$:

$$y = -x + x^3 + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, 0.3^2). \tag{2.2}$$

The density $q_0(x)$ of the covariate x is

$$x \sim \mathrm{N}(\mu_0, \tau_0^2), \tag{2.3}$$

where $\mu_0 = 0.5$, $\tau_0^2 = 0.5^2$. This corresponds to the sampling scheme of x or the design of experiments. A dataset $(x^{(n)}, y^{(n)})$ of size $n = 100$ is generated from (2.2) and (2.3), and plotted by circles in Fig. 1a. MLE $\hat{\theta}_0$ is obtained by the ordinary least squares (OLS) for the normal regression; we consider a model of the form (2.1) with $d = 1$, and the regression line fitted by OLS is drawn in solid line in Fig. 1a.

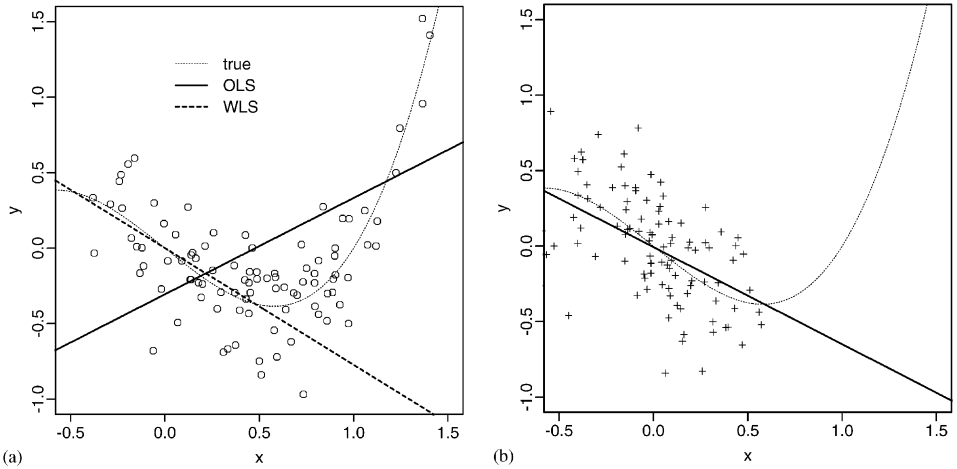

Fig. 1. Fitting of polynomial regression with degree $d=1$. (a) Samples (x_t, y_t) of size $n=100$ are generated from $q(y|x)q_0(x)$ and plotted as circles, where the underlying true curve is indicated by the thin dotted line. The solid line is obtained by OLS, and the dotted line is WLS with weight $q_1(x)/q_0(x)$. (b) Samples of $n=100$ are generated from $q(y|x)q_1(x)$, and the regression line is obtained by OLS.

On the other hand, MWLE $\hat{\theta}_w$ is obtained by weighted least squares (WLS) with weights $w(x_t)$ for the normal regression. We again consider the model with $d=1$, and the regression line fitted by WLS with $w(x)=q_1(x)/q_0(x)$ is drawn in dotted line in Fig. 1a. Here, the density $q_1(x)$ for imaginary "future" observations or that for the whole population in sample surveys is specified in advance by

$$x \sim \mathrm{N}(\mu_1, \tau_1^2), \tag{2.4}$$

where $\mu_1 = 0.0$, $\tau_1^2 = 0.3^2$. The ratio of $q_1(x)$ to $q_0(x)$ is

$$\frac{q_1(x)}{q_0(x)} = \frac{\exp(-(x-\mu_1)^2/2\tau_1^2)/\tau_1}{\exp(-(x-\mu_0)^2/2\tau_0^2)/\tau_0} \propto \exp\left(-\frac{(x-\bar{\mu})^2}{2\bar{\tau}^2}\right), \tag{2.5}$$

where $\bar{\tau}^2 = (\tau_1^{-2} - \tau_0^{-2})^{-1} = 0.38^2$, and $\bar{\mu} = \bar{\tau}^2(\tau_1^{-2}\mu_1 - \tau_0^{-2}\mu_0) = -0.28$.

The obtained lines in Fig. 1a are very different for OLS and WLS. The question is: which is better than the other? It is known that OLS is the best linear unbiased estimate and makes small mean squared error of prediction in terms of $q(y|x)q_0(x)$ which generated the data. On the other hand, WLS with weight (2.5) makes small prediction error in terms of $q(y|x)q_1(x)$ which will generate future observations, and thus WLS is better than OLS here. To confirm this, a dataset of size $n=100$ is generated from $q(y|x)q_1(x)$ specified by (2.2) and (2.4). The regression line of $d=1$ fitted by OLS is shown in Fig. 1b, which is considered to have small prediction error for the "future" data. The regression line of WLS fitted to the past data in Fig. 1a is quite similar to the line of OLS fitted to the future data in Fig. 1b. In practice, only the past data is available. The WLS gave almost the equivalent result to the future OLS by using only the past data.

The underlying true curve is the polynomial with $d = 3$, and thus the regression line of $d = 1$ cannot be fitted to it nicely over all the region of x. However, the true curve is almost linear in the region of $\mu_1 \pm 2\tau_1$, and the nice fit of the WLS in this region is obtained by throwing away the observed samples which are outside of this region. The "effective sample size" may be defined in terms of the entropy by $n_e = \exp(-\sum_{t=1}^{n} p_t \log p_t)$, where $p_t = w(x_t)/\sum_{t'=1}^{n} w(x_{t'})$. In the WLS above, $n_e = 49.3$, which is about the half of the original sample size $n = 100$, and then increases the variance of the WLS. This is discussed later in detail.

3. Asymptotic properties of MWLE

Let $E_i(\cdot)$ denote the expectation with respect to $q(y|x)q_i(x)$ for $i = 0, 1$. Considering $-L_w(\theta)$ as the summation of i.i.d. random variables $l_w(x_t, y_t|\theta)$, it follows from the law of large numbers that $-L_w(\theta)/n \to E_0(l_w(x, y|\theta))$ as n grows to infinity. Then we have $\hat{\theta}_w \to \theta_w^*$ in probability as $n \to \infty$, where θ_w^* is the minimizer of $E_0(l_w(x, y|\theta))$ over $\theta \in \Theta$. Hereafter, we restrict our attention to proper $w(x)$ such that $E_0(l_w(x, y|\theta))$ exists for all $\theta \in \Theta$ and that the Hessian of $E_0(l_w(x, y|\theta))$ is non-singular at θ_w^*, which is uniquely determined and interior to Θ.

If the above result is applied to $w(x) = q_1(x)/q_0(x)$, we find that $\hat{\theta}_1$ converges in probability to the minimizer of $\mathrm{loss}_1(\theta)$ over $\theta \in \Theta$, which we denote θ_1^*. Here the key idea is the importance sampling identity:

$$E_0\left\{\frac{q_1(x)}{q_0(x)} \log p(y|x, \theta)\right\} = \int q(y|x) q_0(x) \frac{q_1(x)}{q_0(x)} \log p(y|x, \theta)\, \mathrm{d}x\, \mathrm{d}y$$

$$= E_1(\log p(y|x, \theta)), \tag{3.1}$$

which implies $E_0(l_w(x, y|\theta)) \equiv \mathrm{loss}_1(\theta)$ and $\theta_w^* = \theta_1^*$ when $w(x) = q_1(x)/q_0(x)$.

Except for the equivalent weight $w(x) \propto q_1(x)/q_0(x)$, we have $\theta_w^* \neq \theta_1^*$ under misspecification in general. From the definition of θ_1^*, therefore, $\mathrm{loss}_1(\theta_w^*) > \mathrm{loss}_1(\theta_1^*)$. This immediately implies the asymptotic optimality of the weight $w(x) = q_1(x)/q_0(x)$, because $\hat{\theta}_w \to \theta_w^*$ and $\hat{\theta}_1 \to \theta_1^*$ and thus $\mathrm{loss}_1(\hat{\theta}_w) > \mathrm{loss}_1(\hat{\theta}_1)$ for sufficiently large n.

$\hat{\theta}_1$ has consistency in a sense that it converges to the optimal parameter value. However, $\hat{\theta}_0$ is more efficient than $\hat{\theta}_1$ in terms of the asymptotic variance. This will be important for moderate sample size, where n is large enough for the asymptotic expansions to be allowed, but not enough for the optimality of $\hat{\theta}_1$ to hold. The following lemma, which is used in the subsequent sections, gives the asymptotic distribution of $\hat{\theta}_w$. The derivation is parallel to that of MLE under misspecification given in White (1982); we replace $\log p(y|x, \theta) q_0(x)$ of MLE with $w(x) \log p(y|x, \theta) q_0(x)$ of MWLE.

Lemma 1. *Assume the regularity conditions similar to those of White* (1982), *i.e., the model is sufficiently smooth and the support of $p(y|x, \theta)$ is the same as that of $q(y|x)$ for all $\theta \in \Theta$. Also assume θ_w^* is an interior point of Θ. Then,* $n^{1/2}(\hat{\theta}_w - \theta_w^*)$

is asymptotically normally distributed as $\mathrm{N}(0, H_w^{-1} G_w H_w^{-1})$, *where* G_w *and* H_w *are* $m \times m$ *matrices defined by*

$$G_w = E_0 \left\{ \left. \frac{\partial l_w(x,y|\theta)}{\partial \theta} \right|_{\theta_w^*} \left. \frac{\partial l_w(x,y|\theta)}{\partial \theta'} \right|_{\theta_w^*} \right\}, \quad H_w = E_0 \left\{ \left. \frac{\partial^2 l_w(x,y|\theta)}{\partial \theta \partial \theta'} \right|_{\theta_w^*} \right\}, \tag{3.2}$$

which are assumed to be non-singular.

4. Expected loss

In the previous section, optimal choice of $w(x)$ was discussed in terms of the asymptotic bias $\theta_w^* - \theta_1^*$. For moderate sample size, however, the variance of $\hat{\theta}_w$ due to the sampling error should be considered. In order to take account of both the bias and the variance, we employ the expected loss $E_0^{(n)}(\mathrm{loss}_1(\hat{\theta}_w))$ to determine the optimal weight; $E_0^{(n)}(\cdot)$ denotes the expectation with respect to $(x^{(n)}, y^{(n)})$ which follows $\prod_{t=1}^n q(y_t|x_t) q_0(x_t)$.

Lemma 2. *The expected loss is asymptotically expanded as*

$$E_0^{(n)}(\mathrm{loss}_1(\hat{\theta}_w)) = \mathrm{loss}_1(\theta_w^*) + \frac{1}{n} \left\{ K_w^{[1]\prime} b_w + \frac{1}{2} \mathrm{tr}(K_w^{[2]} H_w^{-1} G_w H_w^{-1}) \right\} + \mathrm{o}(n^{-1}), \tag{4.1}$$

where the elements of $K_w^{[1]}$ *and* $K_w^{[2]}$ *are defined by*

$$(K_w^{[k]})_{i_1 \cdots i_k} = -E_0 \left\{ \frac{q_1(x)}{q_0(x)} \left. \frac{\partial^k \log p(y|x,\theta)}{\partial \theta^{i_1} \cdots \partial \theta^{i_k}} \right|_{\theta_w^*} \right\}$$

and b_w *is the asymptotic limit of* $nE_0^{(n)}(\hat{\theta}_w - \theta_w^*)$, *which is of order* $\mathrm{O}(1)$.

The expression for b_w is given in the following lemma. We use the summation convention $A_i B^i = \sum_{i=1}^m A_i B^i$ in the formula.

Lemma 3. *The elements of* $b_w = \lim_{n\to\infty} nE_0^{(n)}(\hat{\theta}_w - \theta_w^*)$ *are given by*

$$b_w^{i_1} := H_w^{i_1 i_2} H_w^{j_1 j_2} \{ (H_w^{[2\cdot 1]})_{i_2 j_1 \cdot j_2} - \tfrac{1}{2} (H_w^{[3]})_{i_2 j_1 k_1} H_w^{k_1 k_2} (H_w^{[1\cdot 1]})_{k_2 \cdot j_2} \}, \tag{4.2}$$

where H_w^{ij} *denotes the* (i,j) *element of* H_w^{-1}, *and*

$$(H_w^{[k]})_{i_1 \cdots i_k} = E_0 \left\{ \left. \frac{\partial^k l_w(x,y|\theta)}{\partial \theta^{i_1} \cdots \partial \theta^{i_k}} \right|_{\theta_w^*} \right\},$$

$$(H_w^{[k\cdot l]})_{i_1 \cdots i_k \cdot j_1 \cdots j_l} = E_0 \left\{ \left. \frac{\partial^k l_w(x,y|\theta)}{\partial \theta^{i_1} \cdots \partial \theta^{i_k}} \right|_{\theta_w^*} \left. \frac{\partial^l l_w(x,y|\theta)}{\partial \theta^{j_1} \cdots \partial \theta^{j_l}} \right|_{\theta_w^*} \right\}.$$

Note that the matrices defined in (3.2) *are written as* $G_w = H_w^{[1\cdot 1]}$ *and* $H_w = H_w^{[2]}$.

For sufficiently large n, $\mathrm{loss}_1(\theta_w^*)$ is the dominant term on the right-hand side of (4.1), and the optimal weight is $w(x) = q_1(x)/q_0(x)$ as seen in Section 3. If n is not large enough compared with the extent of the misspecification, the $\mathrm{O}(n^{-1})$ terms related to the first and second moments of $\hat\theta_w - \theta_w^*$ cannot be ignored in (4.1), and the optimal weight changes. In an extreme case where the model is correctly specified, we only have to look at the $\mathrm{O}(n^{-1})$ terms as shown in the following lemma.

Lemma 4. *Assume there exists $\theta^* \in \Theta$ such that $q(y|x) = p(y|x,\theta^*)$. Then, $\theta_w^* = \theta^*$ and $q(y|x) = p(y|x,\theta_w^*)$ for all proper $w(x)$. The expected loss $E_0^{(n)}(\mathrm{loss}_1(\hat\theta_w))$ is minimized when $w(x) \equiv 1$ for sufficiently large n.*

5. Information criterion

The performance of MWLE for a specified $w(x)$ is given by (4.1). However, we cannot calculate the value of the expected loss from it in practice, because $q(y|x)$ is unknown. We provide a variant of the information criterion as an estimate of (4.1).

Theorem 1. *Let the information criterion for MWLE be*

$$\mathrm{IC}_w := -2L_1(\hat\theta_w) + 2\,\mathrm{tr}(J_w H_w^{-1}), \tag{5.1}$$

where

$$L_1(\theta) = \sum_{t=1}^{n} \frac{q_1(x)}{q_0(x)} \log p(y_t|x_t,\theta),$$

$$J_w = -E_0\left\{ \frac{q_1(x)}{q_0(x)} \frac{\partial \log p(y|x,\theta)}{\partial\theta}\bigg|_{\theta_w^*} \frac{\partial l_w(x,y|\theta)}{\partial\theta'}\bigg|_{\theta_w^*} \right\}.$$

The matrices J_w and H_w may be replaced by their consistent estimates

$$\hat J_w = -\frac{1}{n}\sum_{t=1}^{n} \frac{q_1(x_t)}{q_0(x_t)} \frac{\partial \log p(y_t|x_t,\theta)}{\partial\theta}\bigg|_{\hat\theta_w} \frac{\partial l_w(x_t,y_t|\theta)}{\partial\theta'}\bigg|_{\hat\theta_w},$$

$$\hat H_w = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial^2 l_w(x_t,y_t|\theta)}{\partial\theta\partial\theta'}\bigg|_{\hat\theta_w}.$$

Then, $\mathrm{IC}_w/2n$ is an estimate of the expected loss unbiased up to $\mathrm{O}(n^{-1})$ term:

$$E_0^{(n)}(\mathrm{IC}_w/2n) = E_0^{(n)}(\mathrm{loss}_1(\hat\theta_w)) + \mathrm{o}(n^{-1}). \tag{5.2}$$

Fortunately, expression (5.1) turns out to be rather simpler than that of (4.1), and we do not have to worry about the calculation of the third derivatives appeared in (4.2). The explicit form of (5.1) for the normal regression is given in (6.1) and (6.2).

Given the model $p(y|x,\theta)$ and the data $(x^{(n)}, y^{(n)})$, we choose a weight function $w(x)$ which attains the minimum of IC_w over a certain class of weights. This is selection of the weight rather than model selection. Searching the optimal weight over all the possible forms of $w(x)$ is equivalent to n-dimensional optimization problem with respect to $(w(x_t)\colon t=1,\dots,n)$. But we do not take this line here, because of the computational cost as well as a conceptual difficulty which will be mentioned in Section 9. Rather than the global search, we shall pick a better one from the two extreme cases of $w(x)\equiv 1$ and $w(x)=q_1(x)/q_0(x)$, or consider a class of weights by connecting the two extremes continuously:

$$w(x)=\left(\frac{q_1(x)}{q_0(x)}\right)^{\lambda}, \quad \lambda\in[0,1], \tag{5.3}$$

where $\lambda=0$ corresponds to $\hat{\theta}_0$ and $\lambda=1$ corresponds to $\hat{\theta}_1$. In the next section, we numerically find $\hat{\lambda}$ which minimizes IC_w by searching over $\lambda\in[0,1]$. Note that (5.3) is proportional to $\mathrm{N}(\bar{\mu},\bar{\tau}^2/\lambda)$ in the case of (2.5), and $\lambda^{-1/2}$ is the window scale parameter.

When we have several candidate forms of $p(y|x,\theta)$, the model and the weight are selected simultaneously by minimizing IC_w. A similar idea of the simultaneous selection is found in Shibata (1989), where an information criterion RIC is derived for the penalized likelihood. A crucial distinction, however, is that the weight for θ is selected in RIC, whereas the weight for x is selected in IC_w. Another distinction is that the weight is additive to the log likelihood in RIC, while it is multiplicative in IC_w.

Akaike (1974) gave an information criterion

$$\mathrm{AIC}=-2L_0(\hat{\theta}_0)+2\dim\theta, \tag{5.4}$$

where $L_0(\theta)$ is the log-likelihood function. AIC is intended for MLE, and it is obtained as a special case of IC_w. When $q_1(x)=q_0(x)$ and $w(x)\equiv 1$, IC_w reduces to

$$\mathrm{TIC}=-2L_0(\hat{\theta}_0)+2\,\mathrm{tr}(G_0H_0^{-1}),$$

where $G_0=G_w$ and $H_0=H_w$ when $w(x)\equiv 1$. TIC is derived by Takeuchi (1976) as a precise version of AIC, and it is equivalent to the criterion of Linhart and Zucchini (1986). If $p(y|x,\theta_0^*)$ is sufficiently close to $q(y|x)$, $\mathrm{tr}(G_0H_0^{-1})\approx\dim\theta$ and TIC reduces to AIC.

6. Numerical example revisited

For the normal linear regression, such as the polynomial regression given in (2.1), β-components of $\hat{\theta}_w$ are obtained by WLS with weights $w(x_t)$. σ-component of $\hat{\theta}_w$ is then given by $\hat{\sigma}^2=\sum_{t=1}^n w(x_t)\hat{\varepsilon}_t^2/\hat{c}_w$, where $\hat{c}_w=\sum_{t=1}^n w(x_t)$ and $\hat{\varepsilon}_t$ is the residual. Letting $\hat{h}_t$, $t=1,\dots,n$ be the diagonal elements of the hat matrix used in the WLS, the information criterion (5.1) is calculated from

$$-L_1(\hat{\theta}_w)=\frac{1}{2}\sum_{t=1}^{n}\frac{q_1(x_t)}{q_0(x_t)}\left\{\frac{\hat{\varepsilon}_t^2}{\hat{\sigma}^2}+\log(2\pi\hat{\sigma}^2)\right\}, \tag{6.1}$$

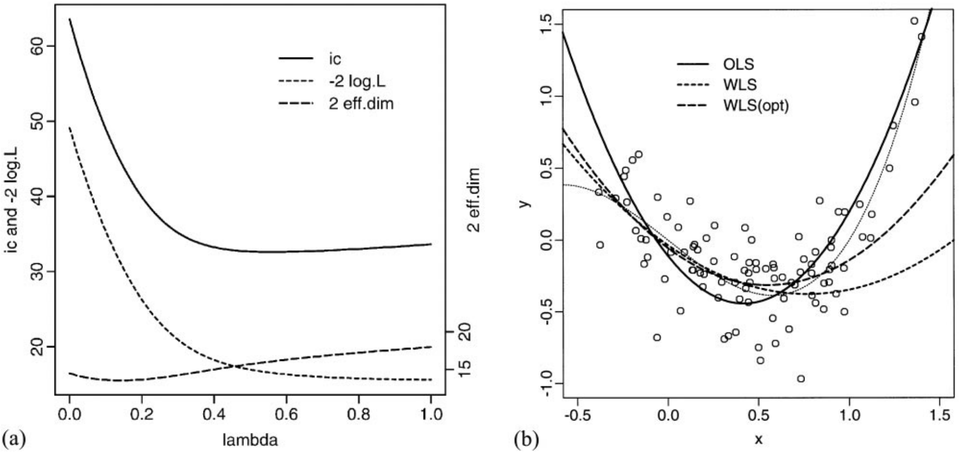

Fig. 2. (a) Curve of IC_w versus $\lambda \in [0,1]$ for the model of Section 2 with $d=2$. The weight function (5.3) connecting from $w(x) \equiv 1$ (i.e. $\lambda=0$) to $w(x)=q_1(x)/q_0(x)$ (i.e. $\lambda=1$) was used. Also shown are $-2L_1(\hat{\theta}_w)$ in dotted lines, and $2\,\mathrm{tr}(J_w H_w^{-1})$ in broken lines. (b) The regression curves for $d=2$. The WLS curve with the optimal $\hat{\lambda}$ as well as those for OLS ($\lambda=0$) and WLS ($\lambda=1$) are drawn.

Table 1
IC_w values with weight (5.3) for $\lambda=0$, $\lambda=1$, and $\lambda=\hat{\lambda}$. Also shown is $\hat{\lambda}$ value. Calculated for the polynomial regression example of Section 2 with $d=0,\ldots,4$

	$d=0$	$d=1$	$d=2$	$d=3$	$d=4$
$\lambda=0$	138.72	174.02	63.59	28.97	31.75
$\lambda=1$	73.96	33.23	33.64	34.80	34.98
$\lambda=\hat{\lambda}$	73.92	32.68	32.62	28.96	31.75
$\hat{\lambda}$	0.95	0.77	0.56	0.01	0.00

$$\mathrm{tr}(\hat{J}_w \hat{H}_w^{-1}) = \sum_{t=1}^{n} \frac{q_1(x_t)}{q_0(x_t)} \left\{ \frac{\hat{\varepsilon}_t^2}{\hat{\sigma}^2} \hat{h}_t + \frac{w(x_t)}{2\hat{c}_w} \left(\frac{\hat{\varepsilon}_t^2}{\hat{\sigma}^2} - 1 \right)^2 \right\}. \tag{6.2}$$

We apply the above formulas to the data generated from (2.2) and (2.3) in Section 2. Fig. 2a shows the plot of the information criterion and its two components for $d=2$. By increasing λ from 0 to 1, the first term of (5.1) decreases while the second term increases in general. We numerically find $\hat{\lambda}$ so that the two terms balance. For $d=2$, IC_w takes the minimum 32.62 at $\hat{\lambda}=0.56$. The regression curves obtained by this method are shown in Fig. 2b.

Table 1 shows IC_w values for $d=0,\ldots,4$. For each d, IC_w is minimized at $\lambda=\hat{\lambda}$. Then, IC_w of $\hat{\lambda}$ is minimized at the model $d=3$. By minimizing IC_w, λ and d are simultaneously selected. For $d=3$, it turns out that $\hat{\lambda}=0.01 \approx 0$. In fact, the model of $d=3$ is correctly specified in this dataset, and it follows from Lemma 4 that $\hat{\theta}_0$ is optimal for $d \geqslant 3$. Even in such a situation, the appropriate $\hat{\lambda}$ is selected by minimizing IC_w.

Table 2
Asymptotic convergence of the second term of (4.1). $2n$ times of $\text{loss}_1(\hat{\theta}_w) - \text{loss}_1(\theta_w^*)$ is calculated for the Monte-Carlo replicates, and its average is tabulated. For every simulation ($d = 0, 1, 2$ and $\lambda = 0, 1$), the values are showing a good convergence as $n \to \infty$

n	$d = 0$		$d = 1$		$d = 2$	
	$\lambda = 0$	$\lambda = 1$	$\lambda = 0$	$\lambda = 1$	$\lambda = 0$	$\lambda = 1$
50	−3.9	5.5	−2.0	8.9	19.0	15.9
100	−5.1	4.8	−3.1	7.6	17.7	11.3
300	−7.0	4.5	−4.5	6.8	17.7	9.0
1000	−8.0	4.4	−4.9	6.6	19.3	8.5

Table 3
Asymptotic convergence of (5.2). $2n\,\text{loss}_1(\hat{\theta}_w) + 2L_1(\hat{\theta}_w)$ is calculated for the Monte-Carlo replicates, and its average is tabulated in the left columns. For $n \geqslant 300$, this agrees very well with the average of $2\,\text{tr}(\hat{J}_w \hat{H}_w^{-1})$ tabulated in the right columns. $2\,\text{tr}(J_w H_w^{-1})$ is shown in the $n = \infty$ row

n	$d = 0$				$d = 1$				$d = 2$			
	$\lambda = 0$		$\lambda = 1$		$\lambda = 0$		$\lambda = 1$		$\lambda = 0$		$\lambda = 1$	
50	1.8	1.7	9.9	7.7	6.8	6.1	15.7	11.1	26.4	13.1	24.8	13.4
100	1.5	1.4	9.2	8.1	6.0	5.7	14.2	12.0	22.6	14.9	19.8	14.9
300	1.3	1.3	8.8	8.4	5.4	5.3	13.3	12.6	18.5	15.7	17.3	16.0
1000	1.2	1.2	8.7	8.6	5.1	5.1	13.1	12.9	16.2	15.4	16.7	16.3
∞		1.2		8.6		5.0		13.0		14.9		16.4

In practical data analysis, it would be rare to have correctly specified models at hand. Therefore, we exclude $d \geqslant 3$ from the above example, and restrict the candidates to $d < 3$. Then, $d = 2$ is selected, and $d = 1$ has almost the same IC_w value, while $d = 0$ has significantly larger IC_w value. This agrees with the asymptotic result verified by extensive Monte-Carlo simulations in Shimodaira (1997) that (4.1) is minimized when $\lambda = 1$ and $d = 1$ over $d \in \{0, 1, 2\}$, for sufficiently large n.

7. Simulation study

First we show simulation results in Tables 1–3 which confirm the theory of Sections 4 and 5. A large number N of replicates of the dataset of size n are generated from (2.2) and (2.3). We used (2.4) as $q_1(x)$. Four simulations of $n = 50$, 100, 300, and 1000 are done with $N = 10^5$ for $n = 50$–300 and $N = 10^6$ for $n = 1000$. For each replicate of the dataset, $\hat{\theta}_w$ is calculated for $\lambda = 0, 1$ and $d = 0, 1, 2$. Then, $\text{loss}_1(\hat{\theta}_w)$, $L_1(\hat{\theta}_w)$, and $\text{tr}(\hat{J}_w \hat{H}_w^{-1})$ are calculated, and their averages over the N replicates are obtained for each simulation. The tables show nice agreement between the simulations and the theory.

Next, we show the results of another set of simulations in Table 4 which confirms the practical performance of the weight selection procedure. We used (2.3) and (2.4) as before, but (2.2) is replaced by $y = -x + \beta_3 x^3 + \varepsilon$ for generating replicates of the

Table 4
The expected loss for the selected weight and the selected model. $2n$ times of $\text{loss}_1(\hat{\theta}_w) + E_1(\log q(y|x))$ is calculated for the replicates, and its average is tabulated in the columns of $\lambda = 0, 1$ for $d = 0, 1, 2$. The average of the loss of the selected weight is shown in the columns of $\hat{\lambda}$. The right most columns show the average of the loss of the selected model

β_3	$d=0$			$d=1$			$d=2$			$\hat{d}$		
	0	1	$\hat{\lambda}$	0	1	$\hat{\lambda}$	0	1	$\hat{\lambda}$	0	1	$\hat{\lambda}$
1.0	98.0	50.7	50.9	123.8	12.3	11.9	71.7	16.0	16.7	73.2	15.0	15.3
0.5	96.0	61.1	61.1	49.9	9.0	8.2	25.6	11.8	11.5	26.9	11.0	10.8
0.2	142.2	68.4	68.4	12.6	7.8	6.4	8.5	10.4	8.1	10.1	9.6	7.9
0.1	152.8	71.1	71.0	6.0	7.8	5.7	5.9	10.4	7.2	6.3	9.6	6.9
0.0	162.2	73.8	73.7	3.6	7.7	4.8	5.0	10.2	6.6	4.4	9.5	6.0

dataset of size $n = 100$. Five simulations of $\beta_3 = 1.0$, 0.5, 0.2, 0.1, and 0.0 are done with $N = 10^4$. In the table, $E_1(-\log q(y|x))$ is subtracted from the loss to make comparisons easier.

The expected loss of MLE ($\lambda = 0$) is 123.8 for $\beta_3 = 1.0$, $d = 1$, while that of MWLE ($\lambda = 1$) is 12.3, showing a great improvement as we have observed in the numerical example. The expected loss of MWLE with the selected weight ($\hat{\lambda}$) is 11.9, which is not significantly different from that of $\lambda = 1$. This is a consequence of the large sample size $n = 100$, where the first term of (4.1) or (5.1) is dominant. The same observation holds for $\beta_3 = 1.0$–0.0, $d = 0$, and $\beta_3 = 1.0, 0.5$, $d = 1, 2$.

On the other hand, $\lambda = 0$ is significantly better than $\lambda = 1$ for $\beta_3 = 0.0$, $d = 1, 2$, where the model is correctly specified. In this case the second term of (4.1) is dominant and $\lambda = 0$ is the optimal choice as mentioned in Lemma 4. The selected weight $\hat{\lambda}$ performs close to $\lambda = 0$, but with slightly larger expected loss. This difference is the price we pay for the weight selection using (5.1) which is an estimate of (4.1), not the true value of (4.1).

For all the cases of $\beta_3 = 1.0$–0.0, $d = 0, 1, 2$, the weight selection procedure gives the expected loss close to the optimal choice. Fixing λ to either of 0 or 1 can lead to very poor performance. The same observation holds for the model selection as shown in the columns of $\hat{d}$.

8. Bayesian inference

We have been working on the predictive density

$$p(y|x, \hat{\theta}_w), \tag{8.1}$$

which is based on MWLE $\hat{\theta}_w$. This type of predictive density is occasionally called as an estimative density in the literature. Another possibility is the Bayesian predictive density. Here we consider a weighted version of it, and examine its performance in prediction.

Let $p(\theta)$ be the prior density of θ. Given the data $(x^{(n)}, y^{(n)})$, we shall define the weighted posterior density by

$$p_w(\theta|x^{(n)}, y^{(n)}) \propto p(\theta)\exp L_w(\theta|x^{(n)}, y^{(n)}). \tag{8.2}$$

Then the predictive density will be

$$p_w(y|x, x^{(n)}, y^{(n)}) = \int p(y|x,\theta)p_w(\theta|x^{(n)}, y^{(n)})\,\mathrm{d}\theta. \tag{8.3}$$

In the case of $w(x) \equiv 1$, (8.2) reduces to the ordinary posterior density, and (8.3) reduces to the ordinary Bayesian predictive density.

The Kullback–Leibler loss of (8.3) with respect to $q(y|x)q_1(x)$ is

$$-\int q_1(x)\int q(y|x)\log p_w(y|x, x^{(n)}, y^{(n)})\,\mathrm{d}y\,\mathrm{d}x$$

and thus the expected loss is given by

$$E_0^{(n)}(E_1(-\log p_w(y|x, x^{(n)}, y^{(n)}))). \tag{8.4}$$

Lemma 5. *For sufficiently large n,* (8.4) *is asymptotically expanded as*

$$E_0^{(n)}(\mathrm{loss}_1(\hat\theta_w)) + \frac{1}{n}\left\{K_w^{[1]\prime}a_w - \frac{1}{2}\mathrm{tr}((K_w^{[1\cdot 1]} - K_w^{[2]})H_w^{-1})\right\} + \mathrm{o}(n^{-1}), \tag{8.5}$$

where

$$K_w^{[1\cdot 1]} = E_0\left\{\frac{q_1(x)}{q_0(x)}\left.\frac{\partial \log p(y|x,\theta)}{\partial\theta}\right|_{\theta_w^*}\left.\frac{\partial \log p(y|x,\theta)}{\partial\theta'}\right|_{\theta_w^*}\right\}$$

and $a_w = \mathrm{plim}_{n\to\infty}\,\hat a_w$ *is the probability limit of*

$$\hat a_w = n\int(\theta - \hat\theta_w)p_w(\theta|x^{(n)}, y^{(n)})\,\mathrm{d}\theta.$$

Furthermore, (8.4) *is estimated by an information criterion*

$$-2\sum_{t=1}^{n}\frac{q_1(x_t)}{q_0(x_t)}\log p_w(y_t|x_t, x^{(n)}, y^{(n)}) + 2\,\mathrm{tr}(J_wH_w^{-1}). \tag{8.6}$$

In fact, the expectation of (8.6), *if divided by* $2n$, *is equal to* (8.5) *up to* $\mathrm{O}(n^{-1})$ *terms.*

When $q_1(x) = q_0(x)$ and $w(x) \equiv 1$, (8.6) reduces to the information criterion for the Bayesian predictive density given in Konishi and Kitagawa (1996). Selection of $w(x)$ as well as selection of $p(\theta)$ and $p(y|x,\theta)$ becomes possible by minimizing (8.6). Comparing the values of (5.1) and (8.6), we can also choose which to use from (8.1) and (8.3).

The decrease of the expected loss of (8.3) from that of (8.1) is of order $\mathrm{O}(n^{-1})$ as shown in (8.5), which can be positive or negative depending on $q(y|x)$. For brevity sake, we assume $q_1(x) = q_0(x)$ and $w(x) \equiv 1$ below. Then the decrease in the expected loss is $\Delta/2n + \mathrm{o}(n^{-1})$, where $\Delta = (\mathrm{tr}(G_0H_0^{-1}) - \dim\theta)$. $G_0 - H_0 \neq 0$ and $\Delta \neq 0$ under

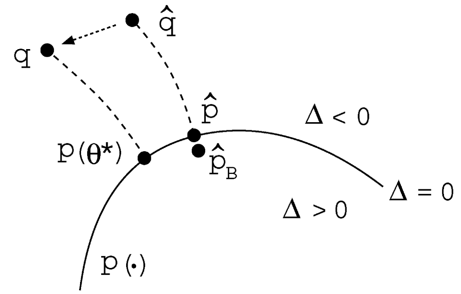

Fig. 3. The curvature of the model in relation to the location of the true density q. On the parametric model denoted by $p(\cdot)$, we have $\Delta = 0$, and $|\Delta|$ increases as q deviates from $p(\cdot)$. The region of $\Delta > 0$ is in the inside direction of the model, and $\Delta < 0$ is in the outside direction of the model. $\hat{q}$ denotes the empirical distribution, and the projection of $\hat{q}$ to the model is the estimative density $\hat{p} = p(y|x, \hat{\theta}_0)$. $\hat{p}_{\mathrm{B}}$ denotes the Bayesian predictive density.

misspecification in general, which is utilized in the information matrix tests (White, 1982) for detecting the misspecification.

An enlightening interpretation of Δ may be possible by the following geometric argument using the terminology of Efron (1978) and Amari (1985). Fig. 3 shows the space of all conditional densities. Δ is determined by the extent of the misspecification multiplied by the "embedding mixture curvature" of the model (S. Amari, personal communication). Bayesian predictive density $\hat{p}_{\mathrm{B}}$ is a mixture of $p(y|x, \theta)$ around $\hat{\theta}_0$, and thus it is located in the inside of the model because of the curvature; $\hat{p}_{\mathrm{B}}$ deviates from $\hat{p}$ of order $\mathrm{O}(n^{-1})$ as shown in Davison (1986). Therefore, $\hat{p}_{\mathrm{B}}$ has larger expected loss than $\hat{p}$ if $q(y|x)$ is located in the outside of the model (i.e. $\Delta < 0$), because $\hat{p}_{\mathrm{B}}$ is located in the opposite side of q. This does not contradict the classical result that the expected loss of $\hat{p}_{\mathrm{B}}$ is asymptotically smaller than that of $\hat{p}$ for some prior. In Bayesian literature, the case of correct specification (i.e. $\Delta = 0$) is discussed and the difference of the expected loss is of order $\mathrm{O}(n^{-2})$ as seen in Komaki (1996). Note that the quadratic form of Δ is relevant when the curvature vanishes, and $\Delta \geqslant 0$ if all the eigenvalues are non-negative; see Shimodaira (1997) for details.

9. Concluding remarks

Although the ratio $q_1(x)/q_0(x)$ has been assumed to be known, it is often estimated from data in practice. Assuming $q_1(x)$ is known, we tried three possibilities in the numerical example of Section 2: (i) $q_0(x)$ is specified correctly without unknown parameters. (ii) Assuming the normality of $q_0(x)$, the unknown μ_0 and τ_0 are estimated. (iii) Non-parametric kernel density estimation is applied to $q_0(x)$. Then, it turns out that MWLE is robust against the estimation of $q_1(x)/q_0(x)$ and the results are almost identical in the three cases. This may be because the form of $q_0(x)$ is quite simple and the sample size $n = 100$ is rather large.

A parametric approach to take account of estimation of $q_1(x)/q_0(x)$ is considered as follows. Let the observed data z_t, $t=1,\dots,n$, follow $p_0(z|\theta)$, while future observations will follow $p_1(z|\theta)$. Then a possible estimating equation will be

$$\sum_{t=1}^{n} w(z_t|\theta)\frac{\partial \log p_1(z_t|\theta)}{\partial\theta} = 0, \tag{9.1}$$

where $w(z|\theta) = (p_1(z|\theta)/p_0(z|\theta))^{\lambda}$. The solution of (9.1) reduces to the MWLE discussed in this paper by letting $z = (x, y)$, $p_0(z|\theta) = p(y|x,\theta)q_0(x)$, and $p_1(z|\theta) = p(y|x,\theta)q_1(x)$.

The estimating equation (9.1) is often seen in the literature of the robust parametric estimation; e.g., Green (1984), Hampel et al. (1986), Lindsay (1994), Basu and Lindsay (1994), Field and Smith (1994) and Windham (1995). In this context, the samples which are not concordant to the model $p_1(z|\theta)$ will be regarded as "outliers" and downweighted to reduce the impact on the parameter estimation. The specification of the weight function is thus the focal point of the argument. Although the covariate shift is a mechanism different from the outliers, an interpretation of MWLE in terms of the robust estimation can be given as follows. For simplicity, let z be a discrete random variable and $\hat{p}_0(z)$ be the observed relative frequency which estimates the contaminated distribution $p_0(z)$ consistently. The weight $(p_1(z|\theta)/\hat{p}_0(z))^{\lambda}$ in (9.1) then leads to the minimum disparity estimator obtained by minimizing the power divergence

$$2nI^{-\lambda} = \frac{2}{\lambda(\lambda-1)}\sum_{z}\hat{p}_0(z)\left\{\left(\frac{\hat{p}_0(z)}{p_1(z|\theta)}\right)^{-\lambda} - 1\right\}$$

(Cressie and Read, 1984; Lindsay, 1994). The uniform weight $\lambda = 0$ is sensitive to outliers, and a positive value $\lambda = 0.5$, say, improves the robustness. In the regression analysis, the deviation of $p_0(z)$ from $p_1(z)$ is decomposed into two parts: the misspecification of $p(y|x,\theta)$ and the covariate shift. MWLE is obtained by applying the power weighting scheme to the second part where $\lambda = 1$ is asymptotically optimal. It may be interesting to consider a robust version of MWLE by applying the weight, say $(p_1(y|x,\theta)/\hat{p}_0(y|x))^{v}$ with $v = 0.5$, to the first part as well.

A numerical example of simultaneous selection of the weight and the model by the information criterion is shown in Section 6. The information criterion takes account of the selection bias caused by estimation of the parameter, but it does not take account the bias caused by the selection of the weight and the model. It is important to evaluate the expected loss of the predictive density obtained after these selection. The simulation study of Section 7 indicates that the method presented in this paper is effective, and the final expected loss of the selected weight and/or model is reasonably small.

Although we have employed a specific type of one-parameter connection in (5.3), other types of connection may work similarly. However, increasing the number of connection parameters will increase the final expected loss because of the sampling error of (5.1) as an estimator of (4.1). We observed the slight increase of the expected loss of $\hat{\lambda}$ in Table 4, and this increase can be much larger for multi-parameter connections.

We derived a variant of AIC for MWLE under covariate shift. On the other hand, Shimodaira (1994) and Cavanaugh and Shumway (1998) discussed variants of AIC for MLE in the presence of incomplete data. Information criteria have to be tailored for different styles of sampling scheme, and the unified approach for them is left as a future work.

A software for calculating IC_w and $\hat{\lambda}$ of normal regression will be available in S language at `http://www.ism.ac.jp/~shimo/`.

Acknowledgements

I would like to thank John Copas, Tony Hayter, Motoaki Kawanabe, Shinto Eguchi, and the reviewers for helpful comments and suggestions.

Appendix A.

Proof of Lemma 1. Since θ^*_w is interior to Θ, so is $\hat{\theta}_w$ for sufficiently large n. Then, $\hat{\theta}_w$ is obtained as a solution of the estimating equation

$$\sum_{t=1}^{n} \left.\frac{\partial l_w(x_t, y_t|\theta)}{\partial \theta}\right|_{\hat{\theta}_w} = 0. \tag{A.1}$$

The Taylor expansion of (A.1) leads to

$$n^{-1}\sum_{t=1}^{n} \left.\frac{\partial^2 l_w(x_t, y_t|\theta)}{\partial \theta \partial \theta'}\right|_{\theta^*_w} n^{1/2}(\hat{\theta}_w - \theta^*_w) = -n^{-1/2}\sum_{t=1}^{n} \left.\frac{\partial l_w(x_t, y_t|\theta)}{\partial \theta}\right|_{\theta^*_w} + \mathrm{O_p}(n^{-1/2}). \tag{A.2}$$

It follows from the central limit theorem that the right-hand side is asymptotically distributed as $\mathrm{N}(0, G_w)$, while the left-hand side converges to $H_w n^{1/2}(\hat{\theta}_w - \theta^*_w)$. Thus we obtained the desired result. □

Proof of Lemma 2. The Taylor expansion of $\mathrm{loss}_1(\hat{\theta}_w)$ around θ^*_w is

$$\mathrm{loss}_1(\theta^*_w) + \frac{1}{n}\left\{(K^{[1]}_w)_i n^{1/2}\dot{\theta}^i_w + \frac{1}{2}(K^{[2]}_w)_{ij}\dot{\theta}^i_w\dot{\theta}^j_w\right\} + \mathrm{O_p}(n^{-3/2}), \tag{A.3}$$

where $\dot{\theta}_w = n^{1/2}(\hat{\theta}_w - \theta^*_w)$, and the summation convention $A_iB^i = \sum_{i=1}^m A_iB^i$ is used. Considering Lemma 1, the expectation of (A.3) gives (4.1). By taking expectation of (A.2), we observe $b_w = \mathrm{O}(1)$. □

Proof of Lemma 3. Considering the $\mathrm{O_p}(n^{-1/2})$ term in (A.2) explicitly, the Taylor expansion of (A.1) gives

$$\hat{H}^*_w\dot{\theta}_w = -n^{-1/2}\sum_{t=1}^{n} \left.\frac{\partial l_w(x_t, y_t|\theta)}{\partial \theta}\right|_{\theta^*_w} + n^{-1/2}e_w,$$

where

$$\hat{H}^*_w = \frac{1}{n}\sum_{t=1}^{n} \left.\frac{\partial^2 l_w(x_t, y_t|\theta)}{\partial\theta\partial\theta'}\right|_{\theta^*_w}, \quad (e_w)_i = -\frac{1}{2}(H^{[3]}_w)_{ijk}\dot{\theta}^j_w\dot{\theta}^k_w + \mathrm{O_p}(n^{-1/2}).$$

Noting $\hat{H}^{*-1}_w = H_w^{-1} - H_w^{-1}(\hat{H}^*_w - H_w)H_w^{-1} + \mathrm{O_p}(n^{-1})$, $n^{1/2}E^{(n)}_0(\dot{\theta}_w)$ is written as

$$E_0\left\{H_w^{-1}\left.\frac{\partial^2 l_w(x, y|\theta)}{\partial\theta\partial\theta'}\right|_{\theta^*} H_w^{-1}\left.\frac{\partial l_w(x, y|\theta)}{\partial\theta}\right|_{\theta^*}\right\} + H_w^{-1}E^{(n)}_0(e_w) + \mathrm{O}(n^{-1/2}),$$

which immediately implies (4.2). □

Proof of Lemma 4. For any x, the conditional Kullback–Leibler loss

$$\mathrm{loss}(\theta|x) = -\int q(y|x)\log p(y|x,\theta)\,\mathrm{d}y$$

is minimized at θ^* if $q(y|x) = p(y|x,\theta^*)$. Then $\theta^*_w = \theta^*$ for any $w(x)$, because $E_0(l_w(x, y|\theta)) = \int q(x)w(x)\mathrm{loss}(\theta|x)\,\mathrm{d}x$. Thus $\mathrm{loss}_1(\theta^*_w)$ in (4.1) is equal for any $w(x)$.

Considering $K^{[1]}_w = 0$, the second term in (4.1) is written as

$$n^{-1}\,\mathrm{tr}(K^{[2]}_w Q(w)^{-1}Q(w^2)Q(w)^{-1}), \tag{A.4}$$

where $K^{[2]}_w = Q(q_1/q_0)$ and $Q(a)$ is defined for any $a(x)$ by

$$Q(a) = E_0\left\{a(x)\left.\frac{\partial \log p(y|x,\theta)}{\partial\theta}\right|_{\theta^*}\left.\frac{\partial \log p(y|x,\theta)}{\partial\theta'}\right|_{\theta^*}\right\}.$$

It it easy to verify that $Q(w)^{-1}Q(w^2)Q(w)^{-1} - Q(1)^{-1}$ is non-negative definite for any $w(x)$, and so (A.4) is minimized when $w(x) \equiv 1$. □

Proof of Theorem 1. The Taylor expansion of $\log p(y|x,\hat{\theta}_w)$ around θ^*_w gives

$$L_1(\hat{\theta}_w) = L_1(\theta^*_w) + \frac{1}{n}\left\{\left.\frac{\partial L_1(\theta)}{\partial\theta'}\right|_{\theta^*_w} n^{1/2}\dot{\theta}_w + \frac{1}{2}\left.\frac{\partial^2 L_1(\theta)}{\partial\theta^i\partial\theta^j}\right|_{\theta^*_w}\dot{\theta}^i_w\dot{\theta}^j_w\right\} + \mathrm{O_p}(n^{-1/2})$$

and thus $E^{(n)}_0(-L_1(\hat{\theta}_w)/n)$ is expanded as

$$\mathrm{loss}_1(\theta^*_w) - \frac{1}{n}E^{(n)}_0\left\{\frac{1}{n}\left.\frac{\partial L_1(\theta)}{\partial\theta'}\right|_{\theta^*_w} n^{1/2}\dot{\theta}_w\right\} + \frac{1}{2n}\mathrm{tr}(K^{[2]}_w H_w^{-1}G_w H_w^{-1}) + \mathrm{O}(n^{-3/2}). \tag{A.5}$$

Considering $-n^{-1}\partial L_1(\theta)/\partial\theta|_{\theta^*_w} = K^{[1]}_w + \mathrm{O_p}(n^{-1/2})$, the second term of (A.5) becomes

$$\begin{aligned}&\frac{1}{n}K^{[1]\prime}_w b_w + \frac{1}{n}E^{(n)}_0\left\{n^{1/2}\left(-\frac{1}{n}\left.\frac{\partial L_1(\theta)}{\partial\theta^i}\right|_{\theta^*_w} - (K^{[1]}_w)_i\right)\right.\\ &\qquad \left.\times H^{ij}_w\left(n^{-1/2}\left.\frac{\partial L_w(\theta)}{\partial\theta^j}\right|_{\theta^*_w} + \mathrm{O_p}(n^{-1/2})\right)\right\}\\ &= \frac{1}{n}K^{[1]\prime}_w b_w - \frac{1}{n}H^{ij}_w(J_w)_{ij} + \mathrm{O}(n^{-3/2}).\end{aligned}$$

Combining this with (A.5) and (4.1) completes the proof. □

Proof of Lemma 5. Assuming certain regularity conditions similar to those of Johnson (1970), we have the asymptotic limit of (8.2) is normal with mean $\hat{\theta}_w$ and covariance matrix $\hat{H}_w^{-1}/n$, since $\log p_w(\theta|x^{(n)}, y^{(n)})$ is expanded as

$$-\tfrac{1}{2}n^{1/2}(\theta-\hat{\theta}_w)'\hat{H}_w n^{1/2}(\theta-\hat{\theta}_w)+\mathrm{O_p}(n^{-1/2}),$$

where the terms independent of θ are omitted. Then, (8.3) is asymptotically expanded as

$$p(y|x,\hat{\theta}_w)+\frac{1}{n}\left.\frac{\partial p(y|x,\theta)}{\partial\theta'}\right|_{\hat{\theta}_w}\hat{a}_w+\frac{1}{2n}\mathrm{tr}\left(\left.\frac{\partial^2 p(y|x,\theta)}{\partial\theta\partial\theta'}\right|_{\hat{\theta}_w}\hat{H}_w^{-1}\right)+\mathrm{o_p}(n^{-1}). \quad (A.6)$$

Note that Dunsmore (1976) gave the unweighted version of (A.6) when the model specification is correct, but the term of $\hat{a}_w$ was missing as indicated by Komaki (1996). Applying the identity

$$\frac{1}{p}\frac{\partial^2 p}{\partial\theta\partial\theta'}=\frac{\partial\log p}{\partial\theta}\frac{\partial\log p}{\partial\theta'}+\frac{\partial^2\log p}{\partial\theta\partial\theta'}$$

to the third term of (A.6), we obtain

$$\begin{aligned}&\log p_w(y|x,x^{(n)},y^{(n)})\\&=\log p(y|x,\hat{\theta}_w)+\frac{1}{n}\left.\frac{\partial\log p(y|x,\theta)}{\partial\theta'}\right|_{\theta_w^*}a_w+\frac{1}{2n}\mathrm{tr}\left\{\left(\left.\frac{\partial\log p(y|x,\theta)}{\partial\theta}\right|_{\theta_w^*}\right.\right.\\&\quad\left.\left.\times\left.\frac{\partial\log p(y|x,\theta)}{\partial\theta'}\right|_{\theta_w^*}+\left.\frac{\partial^2\log p(y|x,\theta)}{\partial\theta\partial\theta'}\right|_{\theta_w^*}\right)H_w^{-1}\right\}+\mathrm{o_p}(n^{-1}). \quad (A.7)\end{aligned}$$

Thus (8.5) immediately follows from (A.7). The last statement of the lemma is verified by combining (A.7) with Theorem 1. □

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19, 716–723.

Amari, S., 1985. Differential-Geometrical Methods in Statistics. Lecture Notes in Statistics, Vol. 28. Springer, Berlin.

Basu, A., Lindsay, B.G., 1994. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. Ann. Inst. Statist. Math. 46, 683–705.

Cavanaugh, J.E., Shumway, R.H., 1998. An Akaike information criterion for model selection in the presence of incomplete data. J. Statist. Plann. Infererence 67, 45–65.

Cressie, N., Read, T.R.C., 1984. Multinomial goodness-of-fit tests. J. Roy. Statist. Soc. Ser. B 46, 440–464.

Davison, A.C., 1986. Approximate predictive likelihood. Biometrika 73, 323–332.

Dunsmore, I.R., 1976. Asymptotic prediction analysis. Biometrika 63, 627–630.

Efron, B., 1978. The geometry of exponential families. Ann. Statist. 6, 362–376.

Field, C., Smith, B., 1994. Robust estimation – a weighted maximum likelihood approach. Internat. Statist. Rev. 62, 405–424.

Green, P.J., 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). J. Roy. Statist. Soc. Ser. B 46, 149–192.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

Johnson, R.A., 1970. Asymptotic expansions associated with posterior distributions. Ann. Math. Statist. 41, 851–864.

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

Konishi, S., Kitagawa, G., 1996. Generalised information criteria in model selection. Biometrika 83, 875–890.

Lindsay, B.G., 1994. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. Ann. Statist. 22, 1081–1114.

Linhart, H., Zucchini, W., 1986. Model Selection. Wiley, New York.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., Rasbash, J., 1998. Weighting for unequal selection probabilities in multilevel models. J. Roy. Statist. Soc. Ser. B 60, 23–56.

Shibata, R., 1989. Statistical aspects of model selection. In: Willems, J.C. (Ed.), From Data to Model. Springer, Berlin, pp. 215–240.

Shimodaira, H., 1994. A new criterion for selecting models from partially observed data. In: Cheeseman, P., Oldford, R.W. (Eds.), Selecting Models from Data: AI and Statistics IV. Springer, Berlin, pp. 21–30 (Chapter 3).

Shimodaira, H., 1997. Predictive inference under misspecification and its model selection. Research Memorandum 642, The Institute of Statistical Mathematics, Tokyo, Japan.

Skinner, C.J., Holt, D., Smith, T.M.F., 1989. Analysis of Complex Surveys. Wiley, New York.

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

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

Windham, M.P., 1995. Robustifying model fitting. J. Roy. Statist. Soc. Ser. B 57, 599–609.