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

Hidetoshi Shimodaira
The Institute of Statistical Mathematics, Tokyo, JAPAN
Email: shimo@ism.ac.jp

RM–712 December 1998

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 are shown for the polynomial regression, and geometrical interpretations are given for a better understanding.

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 covariates, and y be the response variable. In predictive inference, such as the regression analysis, we are interested in estimating the conditional density $q(y|x)$ of y given x, using a parametric model; a family of the conditional densities $p(y|x,\theta)$ parameterized by $\theta = (\theta^1, \ldots, \theta^m)' \in \Theta \subset \mathcal{R}^m$. Having observed i.i.d. samples of size n, denoted by $(x^{(n)}, y^{(n)}) = ((x_t, y_t) : t = 1, \ldots, n)$, 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 of (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

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

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 as 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 of 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 $\hat{\theta}_1$. MWLE turns out to be down-weighting the observed samples which are not important in fitting the model with respect to the population. 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.

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 x. The same idea is also seen in Rao (1991), where weighted maximum likelihood estimation is considered for unequally spaced time series data.

Note that 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.

The construction of the rest of the paper is as follows. In spite of the asymptotic optimality of $w(x) = q_1(x)/q_0(x)$, another choice of the weight function can improve the expected loss for moderate sample size. The optimal weight is obtained by compromising the bias and the variance of $\hat{\theta}_w$ as explained in the asymptotic expansion of the expected loss given in Section 4. Then, in Section 5, an information criterion is derived to find a good $w(x)$ as well as a good form of $p(y|x, \theta)$ from data, and the numerical example is revisited in Section 6. In Section 7, we show the Bayesian predictive density is also improved by considering the weight function. In Section 8, we try to give geometrical interpretations of our approach. 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 $y \in \mathcal{R}$ using a polynomial function of $x \in \mathcal{R}$. Let the model $p(y|x, \theta)$ be the polynomial regression

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

where $\theta = (\beta_0, \ldots, \beta_d, \sigma)$ and $N(a, b)$ 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 $d = 3$:

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

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

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

where $\mu_0 = 0.5$, $\tau_0^2 = 0.5^2$. This corresponds to the sampling scheme of x or the design of experiments, and 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.

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

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

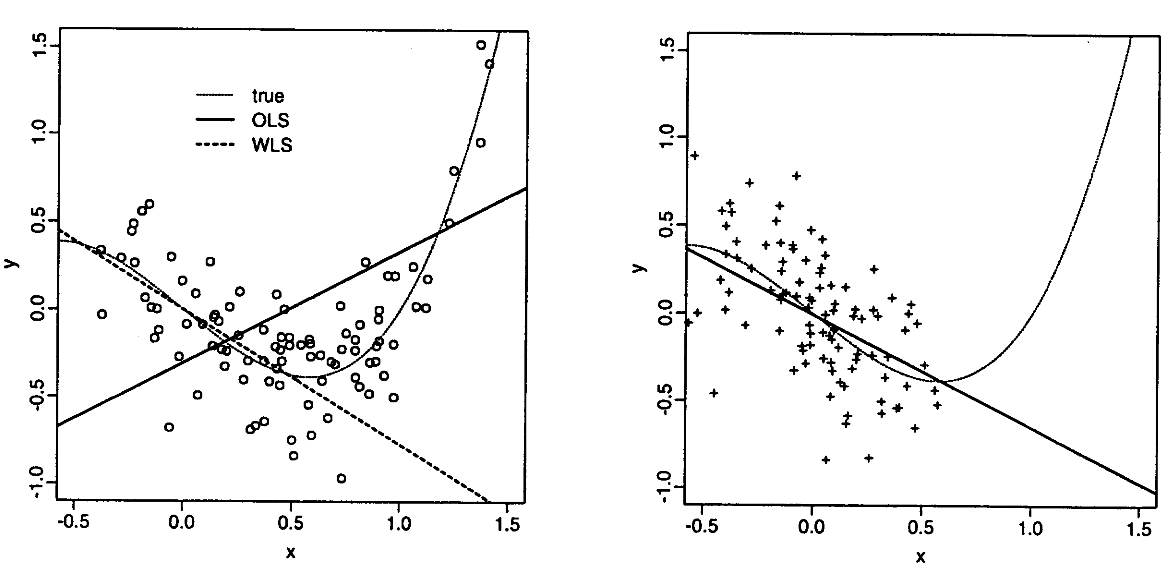
where $\mu_1 = 0.0$, $\tau_1^2 = 0.3^2$. The ratio of $q_1(x)$ to $q_0(x)$ is

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

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. 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. We obtained 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. Note that "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.



(a)

(b)

Figure 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 in 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.

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 $-L_w(\theta)$ is 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 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 $\text{loss}_1(\theta)$ over $\theta \in \Theta$, which we denote θ^*_1. Here the key idea is the importance sampling identity:

$$E_0\left\{\frac{q_1(x)}{q_0(x)} \log p(y|x, \theta)\right\} = \int q(y|x)q_0(x)\frac{q_1(x)}{q_0(x)} \log p(y|x, \theta)\, dx\, dy = E_1(\log p(y|x, \theta)), \tag{3.1}$$

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

From the definition of θ^*_1, $\text{loss}_1(\theta^*_w) \geq \text{loss}_1(\theta^*_1)$ for other choice of $w(x)$. Except for the equivalent weight $w(x) \propto q_1(x)/q_0(x)$, the equality $\text{loss}_1(\theta^*_w) = \text{loss}_1(\theta^*_1)$ holds not for all $q(y|x)$, and so we have $\text{loss}_1(\theta^*_w) > \text{loss}_1(\theta^*_1)$ in general. Thus $\text{loss}_1(\hat{\theta}_w) > \text{loss}_1(\hat{\theta}_1)$ for sufficiently large n, and the asymptotically optimal weight of MWLE is

$w(x) = q_1(x)/q_0(x)$.

$\hat{\theta}_1$ 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. Hence we give the asymptotic distribution of $\hat{\theta}_w$ for the subsequent sections. The derivation is parallel to that of MLE under misspecification given in White (1982); we replace $\log p(y|x,\theta)q_0(x)$ for 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), for example, 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, $n^{\frac{1}{2}}(\hat{\theta}_w - \theta^*_w)$ is asymptotically normally distributed as $N(0, H_w^{-1}G_wH_w^{-1})$, where G_w and H_w are $m \times m$ matrices defined by*

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

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 of 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*

$$E_0^{(n)}(\mathrm{loss}_1(\hat{\theta}_w)) = \mathrm{loss}_1(\theta^*_w) + \frac{1}{n}\left\{ K_w^{[1]\prime} b_w + \frac{1}{2}\,\mathrm{tr}(K_w^{[2]} H_w^{-1} G_w H_w^{-1}) \right\} + O(n^{-\frac{3}{2}}), \tag{4.1}$$

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

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

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

Although it is not essential to the rest of the paper, the expression for b_w is given in the following lemma. We use the summation convention $A_iB^i = \sum_{i=1}^m A_iB^i$ 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*

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

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

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

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 in 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 $O(n^{-1})$ 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 $O(n^{-1})$ terms.

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

$$L_1(\theta) = \sum_{t=1}^n \frac{q_1(x)}{q_0(x)}\log p(y_t|x_t,\theta), \quad J_w = E_0\Big\{\frac{q_1(x)}{q_0(x)}\frac{\partial \log p(y|x,\theta)}{\partial\theta}\Big|_{\theta_w^*}\frac{\partial l_w(x,y|\theta)}{\partial\theta_w'}\Big|_{\theta_w^*}\Big\}.$$

The term $\mathrm{tr}(J_w H_w^{-1})$ *may be called as "effective dimension," and the matrices* J_w *and* H_w *may be replaced by their consistent estimates*

$$\hat{J}_w = \frac{1}{n}\sum_{t=1}^n \frac{q_1(x_t)}{q_0(x_t)}\frac{\partial\log p(y_t|x_t,\theta)}{\partial\theta}\Big|_{\hat{\theta}_w}\frac{\partial l_w(x_t,y_t|\theta)}{\partial\theta'}\Big|_{\hat{\theta}_w}, \quad \hat{H}_w = \frac{1}{n}\sum_{t=1}^n \frac{\partial^2 l_w(x_t,y_t|\theta)}{\partial\theta\partial\theta'}\Big|_{\hat{\theta}_w}.$$

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

$$E_0^{(n)}(\mathrm{IC}_w/2n) = E_0^{(n)}(\mathrm{loss}_1(\hat{\theta}_w)) + O(n^{-\frac{3}{2}}). \tag{5.2}$$

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 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 $(w(x_t) : t = 1, \ldots, n)$. But we do not take this line here, because of the computational cost as well as a conceptual difficulty in it. 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 $N(\bar{\mu}, \bar{\tau}^2/\lambda)$ in the case of (2.5), and $\lambda^{-\frac{1}{2}}$ 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

$$\mathrm{TIC} = -2L_0(\hat{\theta}_0) + 2\,\mathrm{tr}(G_0 H_0^{-1}),$$

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 & Zucchini (1986). If $p(y|x,\theta_0^*)$ is sufficiently close to $q(y|x)$, $\mathrm{tr}(G_0 H_0^{-1}) \approx \dim\theta$ 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 $\hat{\theta}_w$ is obtained by WLS with weights $w(x_t)$. σ-component of $\hat{\theta}_w$ is then given by $\hat{\sigma}^2 = \sum_{t=1}^n w(x_t)\hat{\epsilon}_t^2/\hat{c}_w$, where $\hat{c}_w = \sum_{t=1}^n w(x_t)$ and $\hat{\epsilon}_t$ is the residual. Letting $\hat{h}_t$, $t = 1, \ldots, n$ be the diagonal elements of the hat matrix used in the WLS, the information criterion (5.1) is calculated from

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

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

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 $d \geq 3$. Even in such a situation, the appropriate $\hat{\lambda}$ is selected by minimizing IC_w.

In practical data analysis, it would be rare to have correctly specified models at hand. Therefore, we exclude $d \geq 3$ 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 that (4.1) is minimized when $\lambda = 1$ and $d = 1$ over $d \in \{0, 1, 2\}$, for sufficiently large n.

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

7 Bayesian inference

We have been working on the predictive density

$$p(y|x, \hat{\theta}_w), \tag{7.1}$$

which is based on MWLE $\hat{\theta}_w$. This type of predictive density is occasionally called as estimative density in 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

$$p_w(\theta|x^{(n)}, y^{(n)}) \propto p(\theta) \exp L_w(\theta|x^{(n)}, y^{(n)}). \tag{7.2}$$

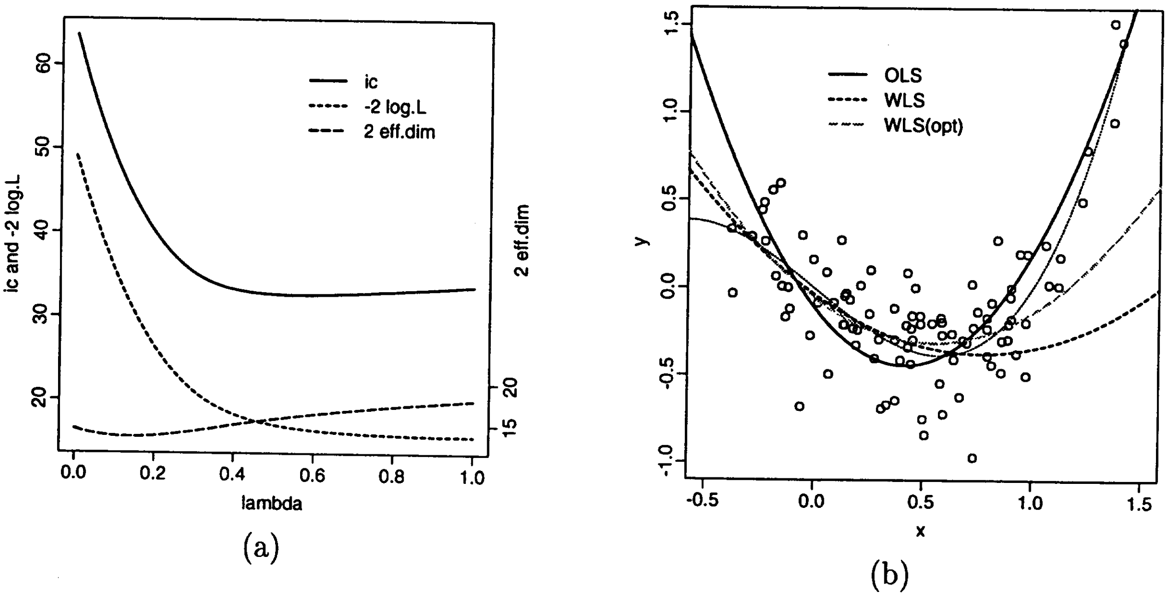


Figure 2: (a) Curve of IC_w versus $\lambda \in [0,1]$ for the model of Section 2 with $d = 2$. The weight function (5.3) connecting from $w(x) \equiv 1$ (i.e. $\lambda = 0$) to $w(x) = q_1(x)/q_0(x)$ (i.e. $\lambda = 1$) was used. Also shown are $-2L_1(\hat{\theta}_w)$ in dotted lines, and $2\,\mathrm{tr}(J_w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

Then the predictive density will be

$$p_w(y|x, x^{(n)}, y^{(n)}) = \int p(y|x,\theta) p_w(\theta|x^{(n)}, y^{(n)})\, d\theta. \tag{7.3}$$

In the case of $w(x) \equiv 1$, (7.2) reduces to the ordinary posterior density, and (7.3) reduces to the ordinary Bayesian predictive density.

The Kullback-Leibler loss of (7.3) with respect to $q(y|x)q_1(x)$ is

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

and thus the expected loss is given by

$$E_0^{(n)}\Big(E_1(-\log p_w(y|x, x^{(n)}, y^{(n)}))\Big). \tag{7.4}$$

Lemma 5 *For sufficiently large n, (7.4) is asymptotically expanded as*

$$E_0^{(n)}(\mathrm{loss}_1(\hat{\theta}_w)) + \frac{1}{n}\Big\{K_w^{[1]\prime} a_w - \frac{1}{2}\,\mathrm{tr}\Big((K_w^{[1\cdot 1]} - K_w^{[2]})H_w^{-1}\Big)\Big\} + o(n^{-1}), \tag{7.5}$$

where

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

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

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

Furthermore, (7.4) is estimated by an information criterion

$$-2\sum_{t=1}^{n}\frac{q_1(x_t)}{q_0(x_t)}\log p_w(y_t|x_t,x^{(n)},y^{(n)})+2\,\mathrm{tr}(J_wH_w^{-1}). \tag{7.6}$$

In fact, the expectation of (7.6), if divided by $2n$, is equal to (7.5) up to $O(n^{-1})$ terms.

When $q_1(x)=q_0(x)$ and $w(x)\equiv 1$, (7.6) reduces to the information criterion for the Bayesian predictive density given in Konishi & Kitagawa (1996). Selection of $w(x)$ as well as selection of $p(\theta)$ and $p(y|x,\theta)$ becomes possible by minimizing (7.6). Comparing the values of (5.1) and (7.6), we can also choose which to use from (7.1) and (7.3). In this selection of predictive method, we may not have to calculate (7.3) explicitly, because the difference between (5.1) and (7.6) would be obtained as a consistent estimate of the second term of (7.5); this is computationally advantageous when $w(x)=q_1(x)/q_0(x)$ and thus $K_w^{[1]}=0$.

8 Geometrical interpretations

Geometrical interpretations of statistical methods are often helpful for a better understanding and further development. Here we attempt to give intuitive interpretations of MWLE using the terminology of Efron (1978) and Amari (1985).

Let $\mathcal{D}$ be the space of all joint densities of (x,y). Each element $r(x,y)\in\mathcal{D}$ is a density function, and it is represented as a point in Fig. 3. For example, $q(y|x)q_0(x)\in\mathcal{D}$ is indicated as a point, which is the intersection of the two sets labeled $q(y|x)$ and $q_0(x)$. The sets of densities in Fig. 3 have the obvious meanings as the labels indicate: $q(y|x)$ denotes $\left\{r(x,y)\in\mathcal{D}\mid r(x,y)/(\int r(x,y)\,dy)=q(y|x)\right\}$, and $q_i(x)$ denotes $\{r(x,y)\in\mathcal{D}\mid \int r(x,y)\,dy=q_i(x)\}$. The manifolds (i.e. smooth sets) of model

$$\{p(y|x,\theta)q_i(x)\in\mathcal{D}\mid\theta\in\Theta\} \tag{8.1}$$

are shown as curves on $q_i(x)$ for $i=0,1$.

The minimization of $\mathrm{loss}_i(\theta)$ is equivalent to the minimization of the Kullback-Leibler divergence of $q(y|x)q_i(x)$ from (8.1). This is represented as "projection" of the point $q(y|x)q_i(x)$ to the manifold $p(y|x,\theta)q_i(x)$ in Fig. 3. The minimum is attained at θ_i^*, and so the projected point is $p(y|x,\theta_i^*)q_i(x)$. In general, $\theta_0^*\neq\theta_1^*$, or equivalently $p(y|x,\theta_0^*)q_1(x)\neq p(y|x,\theta_1^*)q_1(x)$ in Fig. 3. This is because the metric structure is different in each foliation of $q_i(x)$.

MLE $\hat{\theta}_0$ is known to be interpreted as the projection of the empirical distribution

$$\hat{q}_0(x,y)=\frac{1}{n}\sum_{t=1}^{n}\delta(x-x_t)\delta(y-y_t)$$

to the manifold $p(y|x,\theta)q_0(x)$. This is because $-L_0(\theta)/n$ is regarded as the expectation of $-\log p(y|x,\theta)$ with respect to $\hat{q}_0(x,y)$. Quite similarly, MWLE $\hat{\theta}_w$ is interpreted as the projection of

$$w(x)\hat{q}_0(x,y)/\hat{c}_w \tag{8.2}$$

to the model manifold. By using the weight $w(x)$, the point $\hat{q}_0(x,y)$ is "shifted" to (8.2). Let $\hat{q}_1(x,y)$ denote (8.2) with $w(x) = q_1(x)/q_0(x)$ as shown in Fig. 3. $\hat{q}_1(x,y)$ imitates the empirical distribution that would be obtained from i.i.d. samples following $q(y|x)q_1(x)$. Note that $\hat{q}_1(x,y) \to q(y|x)q_1(x)$ in distribution as $n \to \infty$.

We next consider an interpretation of the Bayesian predictive density. The decrease of the expected loss of (7.3) from that of (7.1) is of order $O(n^{-1})$ as shown in (7.5), 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 $\Delta/2n + o(n^{-1})$, where $\Delta = (\mathrm{tr}(G_0 H_0^{-1}) - \dim\theta)$. As shown in Fig. 4, 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 $\hat{p}_B$ 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; $\hat{p}_B$ deviates from $\hat{p}$ of order $O(n^{-1})$ as shown in Davison (1986). Therefore, $\hat{p}_B$ has larger expected loss than $\hat{p}$ if $q(y|x)$ is located in the outside of the model (i.e. $\Delta < 0$), because $\hat{p}_B$ is located in the opposite side of q. This does not contradict with the classical result that the expected loss of $\hat{p}_B$ 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 $O(n^{-2})$ as seen in Komaki (1996).

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 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, $t = 1, \ldots, n$ follow $p_0(z|\theta)$, while future observations will follow $p_1(z|\theta)$. Then a possible estimating equation will be

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

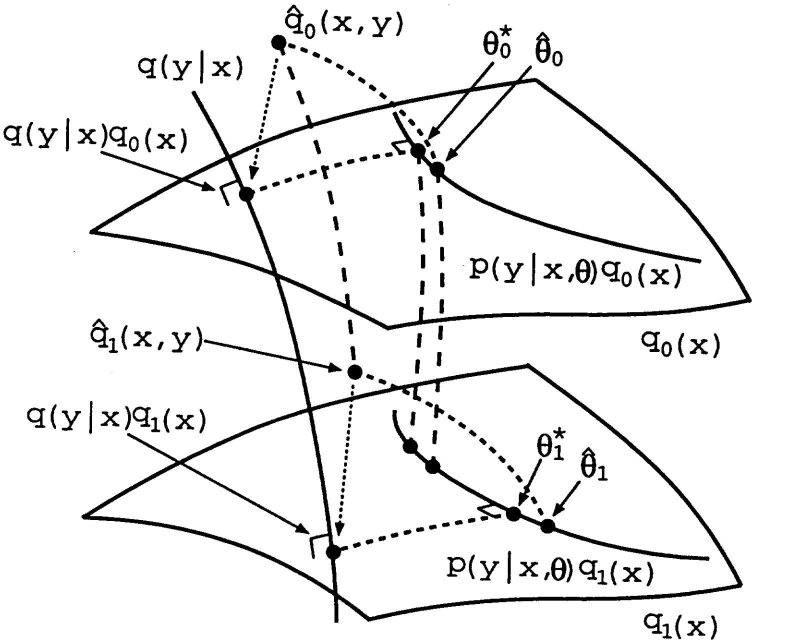


Figure 3: Schematic diagram of the space of joint densities of (x, y) in the sense of Amari (1985); each point represents a joint density of (x, y).

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 study along this line needs further consideration.

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 of those caused by the selection of weight and model. It is important to evaluate the expected loss of the predictive density obtained after these selection. An extensive Monte-Carlo simulation in Shimodaira (1997) indicates that the method presented in this paper is effective, and the final expected loss of MWLE is smaller than that of MLE in most situations.

We derived a variant of AIC for MWLE under covariate shift. On the other hand, Shimodaira (1994) and Cavanaugh & 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.

Acknowledgment

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

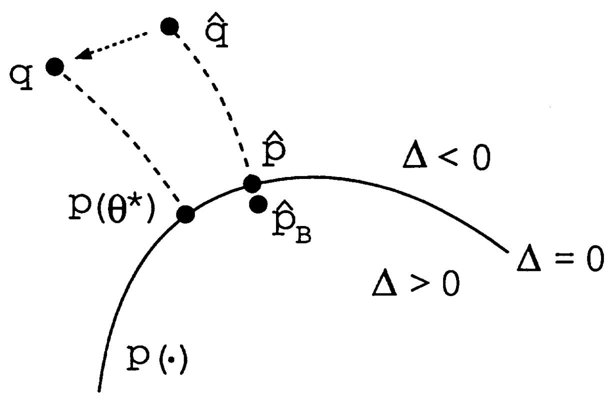


Figure 4: The curvature of the model in relation to the location of the true density q. On the parametric model denoted by $p(\cdot)$, we have $\Delta = 0$, and $|\Delta|$ increases as q deviates from $p(\cdot)$. The region of $\Delta > 0$ is in the inside direction of the model, and $\Delta < 0$ is 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 $\hat{p} = p(y|x, \hat{\theta}_0)$. $\hat{p}_B$ denotes the Bayesian predictive density.

A Appendix

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

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

The Taylor expansion of (A.1) leads to

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

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

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

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

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

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

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

where

$$\hat{H}_w^* = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial^2 l_w(x_t, y_t|\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta_w^*}, \qquad (e_w)_i = -\frac{1}{2}(H_w^{[3]})_{ijk}\dot{\theta}_w^j\dot{\theta}_w^k + O_p(n^{-\frac{1}{2}}).$$

Noting $\hat{H}_w^{*-1} = H_w^{-1} - H_w^{-1}(\hat{H}_w^* - H_w)H_w^{-1} + O_p(n^{-1})$, $n^{\frac{1}{2}}E_0^{(n)}(\dot{\theta}_w)$ is written as

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

which immediately implies (4.2). ▮

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

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

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

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

$$n^{-1}\,\text{tr}\Big(K_w^{[2]}Q(w)^{-1}Q(w^2)Q(w)^{-1}\Big), \tag{A.4}$$

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

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

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

$$L_1(\hat{\theta}_w) = L_1(\theta_w^*) + \frac{1}{n}\left\{\frac{\partial L_1(\theta)}{\partial\theta'}\bigg|_{\theta_w^*} n^{\frac{1}{2}}\dot{\theta}_w + \frac{1}{2}\frac{\partial^2 L_1(\theta)}{\partial\theta^i\partial\theta^j}\bigg|_{\theta_w^*}\dot{\theta}_w^i\dot{\theta}_w^j\right\} + O_p(n^{-\frac{1}{2}}),$$

and thus $E_0^{(n)}(-L_1(\hat{\theta}_w)/n)$ is expanded as

$$\text{loss}_1(\theta_w^*) - \frac{1}{n}E_0^{(n)}\left\{\frac{1}{n}\frac{\partial L_1(\theta)}{\partial\theta'}\bigg|_{\theta_w^*} n^{\frac{1}{2}}\dot{\theta}_w\right\} + \frac{1}{2n}\text{tr}(K_w^{[2]}H_w^{-1}G_wH_w^{-1}) + O(n^{-\frac{3}{2}}) \tag{A.5}$$

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

$$\begin{aligned} &\frac{1}{n}K_w^{[1]\prime}b_w + \frac{1}{n}E_0^{(n)}\left\{n^{\frac{1}{2}}\left(-\frac{1}{n}\frac{\partial L_1(\theta)}{\partial\theta^i}\bigg|_{\theta_w^*} - (K_w^{[1]})_i\right)\times\right. \\ &\qquad\left. H_w^{ij}\left(n^{-\frac{1}{2}}\frac{\partial L_w(\theta)}{\partial\theta^j}\bigg|_{\theta_w^*} + O_p(n^{-\frac{1}{2}})\right)\right\} \\ =\;&\frac{1}{n}K_w^{[1]\prime}b_w - \frac{1}{n}H_w^{ij}(J_w)_{ij} + O(n^{-\frac{3}{2}}). \end{aligned}$$

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 (7.2) 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

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

where terms independent of θ are omitted. Then, (7.3) is asymptotically expanded as

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

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

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

Thus (7.5) 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*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin/New York.

CAVANAUGH, J. E. & SHUMWAY, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *J. Statist. Plann. Infer.* **67**, 45–65.

DAVISON, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73**, 323–332. (Correction: V77 p667).

DUNSMORE, I. R. (1976). Asymptotic prediction analysis. *Biometrika* **63**, 627–630.

EFRON, B. (1978). The geometry of exponential families. *Ann. Statist.* **6**, 362–376.

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.

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., editor, *From Data to Model*, pages 215–240. Springer-Verlag, Berlin/New York.

SHIMODAIRA, H. (1994). A new criterion for selecting models from partially observed data. In Cheeseman, P. & Oldford, R. W., editors, *Selecting Models from Data: AI and Statistics IV*, chapter 3, pages 21–30. Springer-Verlag.

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. *Mathematical Sciences* **No. 153**, 12–18. (in Japanese).

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

REGISTRATION
of
Research Memorandum

(month) (day) (year)
Research Memorandum NO. 712 Received on 12 , 7 , 1998 by

Section of Research Information, the Institute of Statistical Mathematics

(Do not write above this line)

(Fill in the blanks below and return to Section of Research Information)

Title

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

Author(s) and Affiliation(s)

Hidetoshi Shimodaira
The Institute of Statistical Mathematics
4–6–7 Minami-Azabu, Minato-ku, Tokyo 106-8569, JAPAN.
Email: `shimo@ism.ac.jp`

Key Words

Akaike information criterion; Design of experiments; Importance sampling; Kullback-Leibler divergence; Misspecification; Sample surveys; Weighted least squares.

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 are shown for the polynomial regression, and geometrical interpretations are given for a better understanding.