Ann. Inst. Statist. Math.
Vol. 43, No. 4, 735–746 (1991)

# ESTIMATING A MODEL THROUGH THE CONDITIONAL MLE

TAKEMI YANAGIMOTO

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

(Received September 26, 1989; revised July 30, 1990)

**Abstract.** The estimation problem of a model through the conditional maximum likelihood estimator (MLE) is explored. The estimated model is compared using the two dual Kullback-Leibler losses with that through the unconditional MLE. The former is found to be superior to the latter under familiar models. This result is applicable to the model selection problem. These suggest a novel extensive use of the conditional likelihood, since the traditional use of the conditional likelihood was restricted only on inference for the structural parameter.

*Key words and phrases*: AIC, conditional inference, exponential family, Kullback-Leibler loss, model selection.

## 1. Introduction

Conditional inference has been focused for the structural parameter $\theta$ under the existence of the remaining parameter $\mu$, which is usually referred to as a nuisance. In practical situations, however, the parameter $\mu$ is not always nuisance but more important than $\theta$ in familiar examples such as a normal model. In familiar models, the so called structural parameter is usually the dispersion one, and the so called nuisance parameter is the mean (or location) one. Therefore, unless conditional inference for $\theta$ is applicable to inference for $\mu$ or $(\theta, \mu)$, conditional inference is less attractive in practice. The aim of this paper is to claim superiority of the estimated model through the conditional maximum likelihood estimator (MLE) over that through the unconditional MLE under selected models.

Let $x_1, \dots, x_n$ be a sample of size $n$ from a population having the density (or probability) function $p(x; \theta, \mu)$. Write $\boldsymbol{x} = (x_1, \dots, x_n)'$ for $\theta \in \Theta$ and $\mu \in M$. Suppose there exists a statistic $t$ such that

$$p(\boldsymbol{x}; \theta, \mu) \left(= \prod p(x_i; \theta, \mu)\right) = pc(\boldsymbol{x}; \theta \mid t) pr(t; \theta, \mu). \tag{1.1}$$

Inference of $\theta$ is recommended to be based only on the conditional likelihood $pc(\boldsymbol{x}; \theta \mid t)$. A lot of works have been devoted to conditional inference, which includes Fisher (1935), Kalbfleish and Sprott (1970), Godambe (1980), Lindsay

(1982) and Cox and Reid (1987). Their attention, however, was restricted on inference only for $\theta$.

When we obtain the conditional MLE of $\theta$, $\hat{\theta}_c$, a naive way to estimate $\mu$ is to maximize the residual likelihood, $pr(t; \hat{\theta}_c, \mu)$. Such a treatment is actually employed in the survival analysis, for example Lawless ((1982), p. 360). The claim addressed here is that $(\hat{\theta}_c, \hat{\mu}_c)$ provides us with a better estimate of the model than $(\hat{\theta}_u, \hat{\mu}_u)$ under some familiar models, that is, $p(\boldsymbol{x}; \hat{\theta}_c, \hat{\mu}_c)$ can be a better estimate of $p(\boldsymbol{x}; \theta, \mu)$ than $p(\boldsymbol{x}; \hat{\theta}_u, \hat{\mu}_u)$. This claim looks reasonable, but has not been addressed explicitly. This naive extension makes conditional inference much more useful in practice.

Models in the study and preliminaries are given in Section 2. In Section 3 the two dual Kullback-Leibler losses of the two MLE's, the KL and the KS losses, are explored, and superiority of the conditional MLE is shown. We observe that the estimated likelihood in terms of the conditional MLE reduces rightly that in terms of the unconditional MLE. Section 4 is devoted to a criterion for comparing estimated models. The selection problem of the order of the normal polynomial regression model is discussed in Section 5. In the final section we refer to the other two models briefly.

## 2. Models in the study and preliminaries

Our interest will be focused on the models where the conditional MLE of $\theta$ is recommended. Yanagimoto and Anraku (1989) gave seven models where the conditional MLE is properly superior to the unconditional MLE for a finite sample size and a finite number of strata. Bar-Lev (1984) and Jorgensen (1987) discussed the restricted exponential families of distributions, and studied the conditional MLE. In what will follow, for simplicity we will not distinguish a model, its distribution or its density function, unless any confusion is anticipated. Our primary interest will be placed on the normal, the inverse Gaussian and the gamma models. These three models are practically important, and have many favorable properties. Blaesild and Jensen (1985) gave a characterization of them as members of the exponential family satisfying a favorable property. In these three models the MLE of $\mu$ is $\hat{\mu} = \bar{x}$ for an arbitrarily fixed $\theta$, and consequently the conditional and the unconditional MLE's are identical. Setting $t = \bar{x}$, the density function is factored as in (1.1).

Other models in the study are the $2 \times 2$ table and the two-parameter exponential distribution models. The multiple $2 \times 2$ tables model is the example where conditional inference is most widely employed. A reason is that there often exist many strata compared with the total sample size.

The explicit forms of the density functions in the study are presented in Table 1. It gives also the forms of the MLE's or the estimating equations. Note that the parametrization may be a little different from a usual one. We first employ a parameter representing the mean. Then the parameter orthogonal to the mean parameter is chosen, which represents magnitude of the dispersion. Our parametrization is designed also for being associated neatly with population moments. In this concern recall that Gauss originally employed the notation $h = 1/\sqrt{2}\sigma$ in the

Table 1. Models in the study and the conditional and the unconditional MLE's. The statistics $\bar{x}$, $\tilde{x}$ and $x_{(1)}$ denote the sample mean, the sample geometric mean and the smallest order statistic.

| Model | Density function | $\hat{\theta}_c$ | $\hat{\theta}_u$ | $\hat{\mu}$ |
|---|---|---|---|---|
| Normal | $\frac{1}{\sqrt{2\pi\theta}} e^{-(x-\mu)^2/2\theta}$ | $\sum(x_i-\bar{x})^2/(n-1)$ | $\sum(x_i-\bar{x})^2/n$ | $\bar{x}$ |
| Inverse Gaussian | $\sqrt{1/2\pi\theta x^3}\, e^{-(x-\mu)^2/2\theta\mu^2 x}$ | $\sum\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)/(n-1)$ | $\sum\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)/n$ | $\bar{x}$ |
| Gamma | $\frac{x^{1/\theta-1}}{\Gamma(1/\theta)(\theta\mu)^{1/\theta}} e^{-x/\theta\mu}$ | *1 | *2 | $\bar{x}$ |
| Two-parameter exponential | $\frac{1}{\theta}\exp -(x-\mu)/\theta$ | $n(\bar{x}-x_{(1)})/(n-1)$ | $\bar{x}-x_{(1)}$ | $x_{(1)}$ |
| 2 × 2 table | $\binom{n}{x}\frac{e^{x(\alpha+\theta)}}{(1+e^{\alpha+\theta})^n}\binom{m}{y}\frac{e^{y\alpha}}{(1+e^{\alpha})^m}$ | *3 | $\log\frac{x(m-y)}{y(n-x)}$ | *4 |

*1 $\log\bar{x}/\tilde{x} = \xi(\hat{\theta}_c) - \xi(\hat{\theta}_c/n)$ with $\xi(\theta) = \log 1/\theta - \psi(1/\theta)$.

*2 $\log\bar{x}/\tilde{x} = \xi(\hat{\theta}_u)$.

*3 $x = \sum\binom{n}{z}\binom{m}{x+y-z} z e^{z\hat{\theta}_c} / \sum\binom{n}{z}\binom{m}{x+y-z} e^{z\hat{\theta}_c}$.

*4 $\hat{\mu} = (x+y)/(n+m)$ with $\mu = n\frac{e^{\alpha+\theta}}{1+e^{\alpha+\theta}} + m\frac{e^{\alpha}}{1+e^{\alpha}}$.

normal distribution as the precision parameter (Davis (1857), p. 259), while the parameter $\sigma^2$ is employed now.

The polynomial regression model with the normal error is also explored to discuss the model selection problem, where the parameter $\mu$ is a vector. In such a case we will write $\mu$ as $\boldsymbol{\mu}$ to emphasize a vector.

## 3. Kullback-Leibler loss

When all the parameters in a model are estimated, we need to use a suitable loss of the estimated model $p(\boldsymbol{z}; \hat{\theta}(\boldsymbol{x}), \hat{\mu}(\boldsymbol{x}))$ to the true model $p(\boldsymbol{z}; \theta, \mu)$. Here $\boldsymbol{z}$ denotes a latent (or unobserved) sample vector to express a model. We will suppress the sample $\boldsymbol{x}$ in the estimators, unless any confusion is anticipated. The loss introduced by Kullback and Leibler (1951) is most convenient for our comparison study. There are the two dual types of the loss; one is

$$KL(\hat{\theta}(\boldsymbol{x}), \hat{\mu}(\boldsymbol{x}); \theta, \mu) = \int \log \frac{p(\boldsymbol{z}; \hat{\theta}(\boldsymbol{x}), \hat{\mu}(\boldsymbol{x}))}{p(\boldsymbol{z}; \theta, \mu)} p(\boldsymbol{z}; \hat{\theta}(\boldsymbol{x}), \hat{\mu}(\boldsymbol{x})) d\boldsymbol{z}, \tag{3.1}$$

and the other is

$$KS(\hat{\theta}(\boldsymbol{x}), \hat{\mu}(\boldsymbol{x}); \theta, \mu) = \int \log \frac{p(\boldsymbol{z}; \theta, \mu)}{p(\boldsymbol{z}; \hat{\theta}(\boldsymbol{x}), \hat{\mu}(\boldsymbol{x}))} p(\boldsymbol{z}; \theta, \mu) d\boldsymbol{z}. \tag{3.2}$$

The former was discussed in Kullback ((1959), Chapter 3), and the latter is more widely employed in the recent literature. To distinguish them, we call the former the Kullback-Leibler loss (KL loss), and the latter the Kullback-Leibler separator (KS loss). By definition it holds that $KL(\hat{\theta}, \hat{\mu}; \theta, \mu) = KS(\theta, \mu; \hat{\theta}, \hat{\mu})$, if both the losses exist. They are invariant with strictly monotone transformations of $\theta$, $\mu$ and $x$. This, together with the invariance property of likelihood inference, permits us to extend the results obtained here to other transformed models. Define the risk induced from the KL loss as $RKL(\hat{\theta}, \hat{\mu}; \theta, \mu) = E(KL(\hat{\theta}(\boldsymbol{x}), \hat{\mu}(\boldsymbol{x}); \theta, \mu) \mid p(\boldsymbol{x}; \theta, \mu))$. We will write the risk as $C\cdot RKL$ for $\hat{\theta} = \hat{\theta}_c$ and so forth.

*Example* 1. Consider the losses of the estimator $(\hat{\theta}, \hat{\mu}) = (a\bar{x}, bs^2)$ with $s^2 = \sum(x_i - \bar{x})^2/(n-1)$ in the normal model for $a$ and $b > 0$. Then the minimum of $RKL(\hat{\theta}, \hat{\mu}; \theta, \mu)$ is attained at $a = 1$ and $b = 1$, and that of $RKS(\hat{\theta}, \hat{\mu}; \theta, \mu)$ is attained at $a = 1$ and $b = (n+1)(n-1)/n(n-2)$ $(> 1)$. Recall that $\hat{\theta}_u$ is obtained by setting $b = (n-1)/n < 1$. It appears that the selection of $a = 1$ and $b = 1$ is appealing in this example.

When $p(x; \theta, \mu)$ is a member of the exponential family, the KL loss is associated with the likelihood ratio statistic. In fact it is known (Kullback (1959), p. 95) that

$$KL(\hat{\theta}_u(\boldsymbol{x}), \hat{\mu}(\boldsymbol{x}); \theta, \mu) = \log \frac{p(\boldsymbol{x}; \hat{\theta}_u(\boldsymbol{x}), \hat{\mu}(\boldsymbol{x}))}{p(\boldsymbol{x}; \theta, \mu)}, \tag{3.3}$$

for every $\theta$, $\mu$ and $\boldsymbol{x}$. The equality (3.3) means that the likelihood ratio statistic has two different characteristics. While a larger likelihood ratio statistic looks appealing, it also indicates a larger KL loss of the estimated model. The following proposition shows that a result corresponding to the equality (3.3) holds for the conditional MLE in a weaker manner.

PROPOSITION 1. *Suppose that* $p(x;\theta,\mu)$ *is either the normal distribution or the inverse Gaussian. Then it holds that*

$$RKL(\hat{\theta}_c,\hat{\mu};\theta,\mu) = E\left(\log\frac{p(\boldsymbol{x};\hat{\theta}_c(\boldsymbol{x}),\hat{\mu}(\boldsymbol{x}))}{p(\boldsymbol{x};\theta,\mu)} \mid p(\boldsymbol{x};\theta,\mu)\right), \tag{3.4}$$

*for every* $\theta$ *and* $\mu$.

The proof follows straightforwardly. Proposition 1, together with the equality (3.3), yields an elegant proof of Proposition 2.

PROPOSITION 2. (Yanagimoto (1987)) *Under the same assumption in Proposition* 1 *it holds that* $RKL(\hat{\theta}_c,\hat{\mu};\theta,\mu) < RKL(\hat{\theta}_u,\hat{\mu};\theta,\mu)$ *for every* $\theta$ *and* $\mu$.

For ease of our understandings we give explicit forms of the quantities for the normal model.

*Example* 2. (normal model) The above quantities for the normal model are expressed explicitly as follows:

$$\begin{aligned}
2\times KL(\hat{\theta}_c,\hat{\mu};\theta,\mu) &= n\log\frac{\theta}{s^2} - n + \frac{n\{s^2+(\bar{x}-\mu)^2\}}{\theta},\\
2\times CLR &= n\log\frac{\theta}{s^2} - (n-1) + \frac{(n-1)s^2+n(\bar{x}-\mu)^2}{\theta},\\
2\times KL(\hat{\theta}_u,\hat{\mu};\theta,\mu) &= 2\times ULR\\
&= n\log\frac{n\theta}{(n-1)s^2} - n + \frac{(n-1)s^2+n(\bar{x}-\mu)^2}{\theta},
\end{aligned}$$

where $CLR$ and $ULR$ denote the log-likelihood ratio statistics based on the conditional and the unconditional MLE's, respectively. Then it follows that $2C\cdot RKL = n\eta((n-1)/2)+1$, and $2U\cdot RKL = n\eta((n-1)/2)+n\log n/(n-1)$, where $\eta(u)=\log u-\psi(u)$ with $\psi(\cdot)$ being the digamma function. The difference of the latter risk to the former is $n\log n/(n-1)-1$ ($>0$).

The equality (3.4) gives us a striking fact supporting possible superiority of the conditional MLE over the unconditional MLE under selected models. In fact we observe that the excess of the KL loss of $(\hat{\theta}_u,\hat{\mu})$ to that of $(\hat{\theta}_c,\hat{\mu})$ is equal to the difference between the averages of the log-likelihood ratio. Recall that the likelihood of the estimated model through the unconditional MLE is greater than that through the conditional MLE for any $\boldsymbol{x}$. Therefore, this means that the excess of the loss is caused by maximizing the unconditional likelihood without disregarding $pr(\boldsymbol{x};\theta,\mu)$.

The results concerning the KS loss corresponding to Proposition 2 are given as follow, which are shown by easy calculations.

PROPOSITION 3. *Suppose that $p(x;\theta,\mu)$ is either the normal or the inverse Gaussian density function. Then it holds that*

$$RKS(\hat{\theta}_c,\hat{\mu};\theta,\mu) < RKS(\hat{\theta}_u,\hat{\mu};\theta,\mu),$$

*for every $\theta$ and $\mu$.*

*Example* 2 (continued). The KS losses are expressed as

$$2\times KS(\hat{\theta}_c(\boldsymbol{x}),\hat{\mu}(\boldsymbol{x});\theta,\mu) = n\log\frac{s^2}{\theta} - n + \frac{n(\theta+(\bar{x}-\mu)^2)}{s^2},$$
$$2\times KS(\hat{\theta}_u(\boldsymbol{x}),\hat{\mu}(\boldsymbol{x});\theta,\mu) = n\log\frac{(n-1)s^2}{n\theta} - n + \frac{n^2(\theta+(\bar{x}-\mu)^2)}{(n-1)s^2}.$$

The risks induced from the two losses are

$$2\times RKS(\hat{\theta}_c,\hat{\mu};\theta,\mu) = -n\eta((n-1)/2) + 1 + \frac{2(n+1)}{n-3},$$
$$2\times RKS(\hat{\theta}_u,\hat{\mu};\theta,\mu) = -n\eta((n-1)/2) + n\log\frac{n-1}{n} + \frac{4n}{(n-3)},$$

for $n > 4$. Note that the difference of the twice risk induced from the unconditional MLE to that induced from the conditional MLE is $n\log(n-1)/n-1+2(n-1)/(n-3)$. This difference is greater than that between the twice risks induced from the two KL losses.

Since the explicit forms of $\hat{\theta}_u$ and $\hat{\theta}_c$ are unavailable for the gamma model, analytical approach looks severely tough. Simulation study in Yanagimoto (1988) suggests that the results in Propositions 2 and 3 hold for various values of $\theta$. (The KL and the KS losses in this paper were called the K-L type and the K-L losses, respectively.) Our additional extensive simulation study supports this conjecture. Simulation results of the average of the KL losses through the conditional and the unconditional MLE's and $\log\{p(\boldsymbol{x};\hat{\theta}_c,\hat{\mu})/p(\boldsymbol{x};\theta,\mu)\}$ are presented in Table 2 for various values of $\theta$. The computer program was prepared using the special functions and random number generators in IMSL Library (IMSL (1980)) and formula in Yanagimoto (1988). The figures in Table 3 suggest that the equality (3.4) holds at least approximately.

Table 2. Empirical estimates of the three quantities selected values of the sample size and the dispersion parameter in the gamma distribution with 10,000 iterations.

| $n$ | $\theta$ | $C\cdot RKL$ | $E(CLR)$ | $U\cdot RKL$ |
|---|---|---|---|---|
| 6 | 0.05 | 1.145 | 1.144 | 1.191 |
| | 0.1 | 1.154 | 1.152 | 1.198 |
| | 0.2 | 1.156 | 1.160 | 1.204 |
| | 0.5 | 1.148 | 1.156 | 1.198 |
| | 1 | 1.138 | 1.147 | 1.186 |
| | 2 | 1.143 | 1.153 | 1.189 |
| | 4 | 1.177 | 1.175 | 1.210 |
| 20 | 0.05 | 1.032 | 1.034 | 1.047 |
| | 0.1 | 1.034 | 1.034 | 1.046 |
| | 0.2 | 1.129 | 1.140 | 1.184 |
| | 0.5 | 1.123 | 1.134 | 1.176 |
| | 1 | 1.154 | 1.162 | 1.201 |
| | 2 | 1.150 | 1.152 | 1.187 |
| | 4 | 1.061 | 1.063 | 1.071 |

Table 3. The Kullback-Leibler risks of the simultaneous estimation in the multiple normal populations under the assumption that the means of $n - k$ out of $n$ populations are known.

| $n$ | $k$ | $2 \times RKL$ | | $2 \times RKS$ | |
|---|---|---|---|---|---|
| | | $CMLE$ | $UMLE$ | $CMLE$ | $UMLE$ |
| 10 | 1 | 2.152 | 2.206 | 2.991 | 3.509 |
| | 2 | 3.318 | 3.533 | 4.698 | 6.467 |
| | 3 | 4.496 | 5.063 | 6.704 | 10.937 |
| | 5 | 7.131 | 9.063 | 12.869 | 30.937 |
| 20 | 1 | 2.071 | 2.097 | 2.399 | 2.609 |
| | 2 | 3.132 | 3.239 | 3.618 | 4.261 |
| | 3 | 4.200 | 4.450 | 4.867 | 6.217 |
| | 5 | 6.363 | 7.117 | 7.483 | 11.345 |
| | 10 | 12.066 | 15.929 | 15.434 | 39.070 |

## 4. Estimation of the likelihood of a model

In practical situations an assumed model is not always the unique candidate model. It is more realistic that we have various candidate models potentially fitted well to data in the study. Then we need to discuss the goodness of fit of an assumed model or to compare it with other ones. To prepare comparisons of candidate models, we discuss an estimator of the likelihood of a fitted model.

Assume that the true model is $p(\boldsymbol{x}; c, m)$ for some $c \in \Theta$ and $m \in M$. Consider the measure of the goodness of a fitted model,

$$E\left\{\int \log p(\boldsymbol{z}; \hat{\theta}(\boldsymbol{x}), \hat{\mu}(\boldsymbol{x})) p(\boldsymbol{z}; c, m) d\boldsymbol{z} \mid p(\boldsymbol{x}; c, m)\right\}, \tag{4.1}$$

which will be written as $H(\hat{\theta}, \hat{\mu}; c, m)$. The measure is the average of the log-likelihood of the estimated model. Therefore, it is regarded as that of reproducibility of the true model by the estimated model.

It follows that $H(c, m; c, m) - H(\hat{\theta}, \hat{\mu}; c, m) = RKS(\hat{\theta}, \hat{\mu}; c, m) > 0$, if the left-hand side exists. Using Proposition 3, we obtain

$$H(\hat{\theta}_c, \hat{\mu}; c, m) > H(\hat{\theta}_u, \hat{\mu}; c, m) \tag{4.2}$$

for every $c$ and $m$, if $p(x; \theta, \mu)$ is normal or inverse Gaussian. The inequality is expected to hold for the gamma model.

Next we discuss the two estimates of $H(\hat{\theta}, \hat{\mu}; c, m)$ using the conditional and the unconditional MLE's. The following propositions are useful for constructing them. The proof follows from the equality (3.3) and Proposition 1.

PROPOSITION 4. (i) *Suppose $p(x; \theta, \mu)$ is a member in the exponential family. Then it holds that $H(\hat{\theta}_u, \hat{\mu}; c, m) = E\{\log p(\boldsymbol{x}; \hat{\theta}_u, \hat{\mu}) \mid p(\boldsymbol{x}; c, m)\} - RKL(\hat{\theta}_u, \hat{\mu}; c, m) - RKS(\hat{\theta}_u, \hat{\mu}; c, m)$ for every $c$ and $m$.*

(ii) *Suppose that $p(x; \theta, \mu)$ is normal or inverse Gaussian. Then the above equality for $\hat{\theta}_c$ in place of $\hat{\theta}_u$ also holds for every $c$ and $m$.*

Recall that Proposition 1 holds approximately for the gamma model. Therefore, we can reasonably expect that Proposition 4(ii) approximately holds for the gamma density function.

Since the unknown parameters still remain in the two risks, it is necessary to estimate them. Note that the large sample approximation yields $RKL(\hat{\theta}, \hat{\mu}; c, m) \fallingdotseq RKS(\hat{\theta}, \hat{\mu}; c, m) \fallingdotseq 1$ for $\hat{\theta}$ being either the conditional MLE or the unconditional. Set $T_c = \log p(\boldsymbol{x}; \hat{\theta}_c, \hat{\mu}) - 2$ and $T_u = \log p(\boldsymbol{x}; \hat{\theta}_u, \hat{\mu}) - 2$, the latter of which is $-1/2$ times as large as the AIC in Akaike (1973). Then $T_c$ and $T_u$ are approximately unbiased estimates of $H(\hat{\theta}_c, \hat{\mu}; c, m)$ and $H(\hat{\theta}_u, \hat{\mu}; c, m)$, respectively.

By definition it holds that $T_c < T_u$ for any $\boldsymbol{x}$. It may look that $T_u$ is superior to $T_c$. Recall, however, that the reverse inequality (4.2) also holds. This superficial confusion comes from inaccuracy of the above large sample approximations. In fact when $p(x; \theta, \mu)$ is normal, inverse Gaussian or gamma, we know that $RKS(\hat{\theta}_u, \hat{\mu}; c, m) > RKS(\hat{\theta}_c, \hat{\mu}; c, m) > 1$ and $RKL(\hat{\theta}_u, \hat{\mu}; c, m) > RKL(\hat{\theta}_c, \hat{\mu}; c, m) > 1$ for every $c$ and $m$. These inequalities show that $T_c$ is a less biased estimator of $H(\hat{\theta}_c, \hat{\mu}; c, m)$ than $T_u$ is that of $H(\hat{\theta}_u, \mu; c, m)$.

The above results support superiority of the use of the conditional MLE for estimating $H(\hat{\theta}, \hat{\mu}; c, m)$ over that of the unconditional MLE. It is surprising that any estimator other than the unconditional MLE has been out of interest, though much work has been devoted to the AIC and related criteria.

We assumed that a true model is involved in an assumed model in the previous section. Next we suppose that a true model having the density function $g(\boldsymbol{x})$ is incorrect, that is, $g(\boldsymbol{x})$ is not included in the family $p(\boldsymbol{x};\theta,\mu)$ for $\theta\in\Theta$ and $\mu\in M$. Let $c_g$ and $m_g$ be the values attaining the maximum of $E(\log p(\boldsymbol{x};\theta,\mu)\mid g(\boldsymbol{x}))$. Then $p(\boldsymbol{x};c_g,m_g)$ is regarded as the nearest density function to $g(\boldsymbol{x})$. An extension of Proposition 4 is possible as follows.

PROPOSITION 5. *Suppose that $p(\boldsymbol{x};\theta,\mu)$ is either normal or inverse Gaussian. Then it holds that*

$$E\{\log p(\boldsymbol{x};\hat{\theta},\hat{\mu})-KL(\hat{\theta},\hat{\mu};c_g,m_g)-KS(\hat{\theta},\hat{\mu};c_g,m_g)\mid g(\boldsymbol{x})\}$$
$$=E\left\{\int\log p(z;\hat{\theta}(\boldsymbol{x}),\hat{\mu}(\boldsymbol{x}))p(z;c_g,m_g)dz\mid g(\boldsymbol{x})\right\}$$

*for either $\hat{\theta}=\hat{\theta}_c$ or $\hat{\theta}_u$, if all the expectations exist.*

## 5. Model selection

In this section we will discuss the model selection problem. The inequalities in the previous section suggest superiority of the conditional MLE also in this problem. However, it is necessary to pursue whether this conjecture is actually true or not. Since the model selection problem is much complicated, we will restrict our attention on a family of normal models.

Consider $K$ models such that $\boldsymbol{x}\sim N(\boldsymbol{\mu},\theta I)$, $\boldsymbol{\mu}\in M_k$, $M_k=\{(\mu_1,\ldots,\mu_n)'\mid\mu_{k+1}=\cdots=\mu_n=0\}$ for $k=1,\ldots,K$. We write the true model as $\boldsymbol{\mu}=\boldsymbol{m}$ and $\theta=c$, and assume that $|m_k|$ is decreasing in some sense. Let $k_0$ be the maximum $k$ such that $m_k\neq 0$. The models, $M_k$ for $k<k_0$, are then incorrect. Our problem is to select the optimum model $M_k$, and then to estimate parameters in the model. Note that the estimation problem of the normal polynomial regression model is analytically equivalent with the sequential simultaneous estimation of means of many normal models. The model $M_k$ corresponds with the $(k-1)$-order polynomial regression model. The problem of determining an optimum fitting order of the normal polynomial regression model is a familiar, important one in practice.

Consider a model $M_k$ for a fixed $k$, and the partition of vectors of $\boldsymbol{x}$ and $\boldsymbol{\mu}$ into the first $k$-dimensional vectors and the remaining $(n-k)$-dimensional ones, for example $\boldsymbol{x}'=(\boldsymbol{x}_1',\boldsymbol{x}_2')$. Then we obtain that $\hat{\theta}_c=\|\boldsymbol{x}_2\|^2/(n-k)$, $\hat{\theta}_u=\|\boldsymbol{x}_2\|^2/n$, $\hat{\boldsymbol{\mu}}_1=\boldsymbol{x}_1$ and $\hat{\boldsymbol{\mu}}_2=\boldsymbol{0}_{(n-k)}$. It follows that

$$KL(\hat{\theta},\hat{\boldsymbol{\mu}};\theta,\boldsymbol{\mu})=-n\log\frac{\hat{\theta}}{\theta}+\frac{n\hat{\theta}+\|\boldsymbol{x}_1-\boldsymbol{\mu}_1\|^2+\|\boldsymbol{\mu}_2\|^2}{\theta}-n,$$
$$KS(\hat{\theta},\hat{\boldsymbol{\mu}};\theta,\boldsymbol{\mu})=n\log\frac{\hat{\theta}}{\theta}+\frac{n\theta+\|\boldsymbol{x}_1-\boldsymbol{\mu}_1\|^2+\|\boldsymbol{\mu}_2\|^2}{\hat{\theta}}-n.$$

Suppose $\|\boldsymbol{m}_2\|^2=0$, that is, the assumed model is true. Proposition 4 still holds by replacing $\hat{\mu}$ and $m$ with $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{m}$, respectively. The risks are then written as $2C{\cdot}RKL = n\eta((n-k)/2)+k$, $2U{\cdot}RKL = n\eta((n-k)/2)+\log n/(n-k)$, $2C{\cdot}RKS = -n\eta((n-k)/2)+k+2(n+k)/(n-k-2)$ and $2U{\cdot}RKS = -n\eta((n-k)/2)+n\log(n-k)/n+2n(k+1)/(n-k-2)$. Note that $\eta(u)$ is approximated by $2/u$ for a large $u$. Table 3 presents numerical values of the above four quantities. Sugiura (1978) gave an explicit form of $2U{\cdot}RKL+2U{\cdot}RKS$, and pointed out that it could take a value differing greatly from $2(k+1)$, which is employed in defining the AIC. We observe that $U{\cdot}RKS$ is much larger than $C{\cdot}RKS$, when $k$ is large. This is a strong reason why the use of the conditional MLE is recommended.

Next we consider the selection problem of the order $k$. The goodness of fit of the model $M_k$ is evaluated by $T_k = \log p(\boldsymbol{x};\hat{\theta},\hat{\mu}) - (k+1)$. We select the $k$ maximizing $T_k$ over all the integers less than or equal to $K$. Then we estimate the parameter by specifying the model $M_k$. The use of the unconditional MLE of $\theta$ in $T_k$ yields the minimum AIC procedure. Our assertion is to use $\hat{\theta}_c$ in place of $\hat{\theta}_u$. The analytical comparison study of the two procedures looks severely complicated, and we conduct the simulation study. As usual in the regression model, we maximize $T_k$ sequentially, that is, the minimum $k$ satisfying $T_k > T_{k+1}$ is chosen. In addition, we set $K = n/2$. Note that the use of the unconditional MLE makes the procedure prodigal, which can be proved analytically. Table 4 presents the results. As expected, the results on risk comparisons are parallel with those in the case of a fixed $k$. It should be noted that $k$ often takes a fairly large number in the regression model. Consequently, the difference of a risk of the procedure through the conditional MLE to that through the unconditional MLE can be large.

Table 4. Estimated average Kullback-Leibler losses in the simultaneous estimation of multiple normal populations with the variance 1 with 10,000 iterations.

| Means configuration† | $2 \times KL$ | | $2 \times KS$ | |
|---|---|---|---|---|
| | *CMLE* | *UMLE* | *CMLE* | *UMLE* |
| $20 * 0$ | 1.800 | 1.820 | 2.239 | 2.423 |
| $2 * 10$, $18 * 0$ | 4.000 | 4.219 | 4.804 | 6.119 |
| $2 * 10$, $48 * 0$ | 3.808 | 3.873 | 4.065 | 4.437 |
| 10, 5, 1.5, 1.0,<br>0.5(0.1)0.1, $11 * 0$ | 6.616 | 6.869 | 7.317 | 9.283 |

†$20 * 0$ means 20 0's, and 0.5(0.1)0.1 means numbers from 0.5 to 0.1 at a step size 0.1.

## 6. Other models

In this section we discuss the two-parameter exponential and the logit models.

### 6.1 *Two-parameter exponential model*

As in Table 1 the density function of the two-parameter exponential distribution is of a simple form, and both the conditional and the unconditional MLE's of $\theta$ can be expressed in an explicit form. Since the support of the distribution depends on the unknown parameter $\mu$, it is not a member of the regular exponential family.

Similarly to the normal and the inverse Gaussian distributions, it holds that $\hat{\theta}_u = (n-1)\hat{\theta}_c/n$. It is easily shown that the equality (3.3) and Proposition 1 hold, which yield $KL(\hat{\theta}_c, \hat{\mu}; \theta, \mu) < KL(\hat{\theta}_u, \hat{\mu}; \theta, \mu)$ for every $\theta$ and $\mu$. The KS loss, however, does not exist for both the MLE's. This is because the common support of the estimated models through the two MLE's is properly included in that of the true model. It seems philosophically reasonable that the estimated model shrinks the support. Therefore, it is undesirable that the KS loss takes infinity, when the support is shrunk. Consequently, the estimation of the measure (4.1) is impossible. The statistic $T_k$, however, can be calculated, and looks applicable to the model selection.

### 6.2 *Logit model*

Consider a simple 2 × 2 table model. Let $x$ and $y$ be outcomes from the two binomial distributions, $Bi(n, e^{\alpha+\theta}/(1+e^{\alpha+\theta}))$ and $Bi(m, e^{\alpha}/(1+e^{\alpha}))$. As in the introduction the MLE of $\mu = ne^{\alpha+\theta}/(1+e^{\alpha+\theta}) + me^{\alpha}/(1+e^{\alpha})$ is $x+y$ for an arbitrarily fixed $\theta$. The conditional likelihood given $t = x+y$ depends only on $\theta$. Though this model is a two sample model, the factorization property (1.1) holds. The extension of our treatments is straightforward. Note that the equality (3.3) holds, since the joint distribution of $x$ and $y$ is in the exponential family.

Table 5. Averages of the three quantities to selected values of the odds ratio $\theta$ in the logit model. Both the sample sizes are 10, and the parameter $\alpha$ is 0.

| $\theta$ | $C{\cdot}RKL$ | $E(CLR)$ | $U{\cdot}RKL$ |
|---|---|---|---|
| 0 | 1.009 | 1.059 | 1.061 |
| 0.5 | 1.016 | 1.064 | 1.066 |
| 1.0 | 1.040 | 1.081 | 1.083 |
| 1.5 | 1.081 | 1.104 | 1.178 |
| 2.0 | 1.114 | 1.116 | 1.119 |
| 2.5 | 1.109 | 1.094 | 1.097 |
| 3.0 | 1.060 | 1.036 | 1.039 |
| 3.5 | .980 | .954 | .956 |
| 4.0 | .890 | .867 | .868 |

As Yanagimoto and Anraku (1989) noted, there is no sufficient evidence showing superiority of the conditional MLE over the unconditional MLE in the logit model. We compare the $RKL(\hat{\theta}, \hat{\mu}; \theta, \mu)$ of the two MLE's. The results are given in Table 5, which shows that neither the equality in Proposition 1 nor the inequality in Proposition 2 holds for this model. When $\theta$ is small, it holds that $RKL(\hat{\theta}_c, \hat{\mu}; \theta, \mu) < RKL(\hat{\theta}_u, \hat{\mu}; \theta, \mu)$. Table 5 and simulation results, not included here, show that such a region of $\theta$ is $|\theta| < 2$ as a rule of thumb. Since our interest is unlikely to be paid to a large absolute value of $\theta$, the result supports the actual preference of the conditional MLE. Note also that the increase of the number of strata makes the region wider.

## Acknowledgements

The author thanks the referees for their generous comments, which are helpful in improving the presentation.

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd Inter. Symp. on Information Theory* (eds. B. N. Petroc and F. Csak), 267–281, Akademia Kiado, Budapest.

Bar-Lev, S. K. (1984). Asymptotic behaviour of conditional maximum likelihood estimators in a certain exponential model, *J. Roy. Statist. Soc. Ser. B*, **46**, 425–430.

Blaesild, P. and Jensen, J. L. (1985). Saddlepoint formulas for reproductive exponential models, *Scand. J. Statist.*, **12**, 193–202.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximated conditional inference (with discussion), *J. Roy. Statist. Soc. Ser. B*, **49**, 1–39.

Davis, C. H. (1857). *Theory of the Motions of the Heavenly Bodies Moving about the Sun in Conic Sections*, Little Brown, Boston.

Fisher, R. A. (1935). The logic of inductive inference, *J. Roy. Statist. Soc.*, **98**, 39–54.

Godambe, V. P. (1980). On sufficiency and ancillary in the presence of nuisance parameter, *Biometrika*, **67**, 155–162.

IMSL (1980). *User's Manual Stat/Library*, IMSL Inc., Houston.

Jorgensen, B. (1987). Exponential dispersion models (with discussion), *J. Roy. Statist. Soc. Ser. B*, **49**, 127–162.

Kalbfleish, J. D. and Sprott, D. A. (1970). Application of likelihood method to models involving large number of parameters (with discussion), *J. Roy. Statist. Soc. Ser. B*, **32**, 175–208.

Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New York.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Ann. Math. Statist.*, **22**, 79–86.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

Lindsay, B. (1982). Conditional score functions: Some optimality results, *Biometrika*, **69**, 503–512.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criticism and the finite corrections, *Comm. Statist. Theory Methods*, **7**, 13–26.

Yanagimoto, T. (1987). A notion of an obstructive residual likelihood, *Ann. Inst. Statist. Math.*, **39**, 247–261.

Yanagimoto, T. (1988). The conditional maximum likelihood estimator of the shape parameter in the gamma distribution, *Metrika*, **35**, 161–175.

Yanagimoto, T. and Anraku, K. (1989). Possible superiority of the conditional MLE over the unconditional MLE, *Ann. Inst. Statist. Meth.*, **41**, 269–278.