

A family of the adjusted estimators maximizing the asymptotic predictive expected log-likelihood

Haruhiko Ogasawara¹

Received: 5 May 2016 / Accepted: 29 September 2016 / Published online: 9 December 2016
© The Behaviormetric Society 2016

Abstract A family of the estimators adjusting the maximum likelihood estimator by a higher-order term maximizing the asymptotic predictive expected log-likelihood is introduced under possible model misspecification. The negative predictive expected log-likelihood is seen as the Kullback–Leibler distance plus a constant between the adjusted estimator and the population counterpart. The vector of coefficients in the correction term for the adjusted estimator is given explicitly by maximizing a quadratic form. Examples using typical distributions in statistics are shown.

Keywords Expected log-likelihood · Kullback–Leibler distance · Mean square error · Asymptotic expectation · Shrinkage · Asymptotic bias

1 Introduction

It is known that the maximum likelihood estimator (MLE) does not necessarily give the smallest values of typical indexes of errors, where an error is defined as the deviation of a parameter estimate from its true value, which is typically squared and averaged over a range of the associated observable variables. For instance, consider $X_i \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ ($i = 1, \dots, n$) and $\bar{X} = n^{-1} \sum_{j=1}^n X_j$ with n being the sample size. Then, it is known that among the family of the estimators of σ^2 given by $c_n \sum_{j=1}^n (X_j - \bar{X})^2$, where c_n is a constant depending on n , $c_n = 1/(n + 1)$ gives the

Communicated by Shuichi Kawano.

✉ Haruhiko Ogasawara
hogasa@res.otaru-uc.ac.jp

¹ Department of Information and Management Science, Otaru University of Commerce, 3-5-21, Midori, Otaru 047-8501, Japan

smallest mean square error (see, e.g., DeGroot and Schervish 2002, p. 431; for the associated problems about variance estimation, see Ogasawara 2015). Note that the above estimator is $n/(n+1)$ times the normal-theory MLE.

Discrepancy functions are defined in many ways using different definitions of distances. In the above case, the squared Euclid distance averaged over the distribution is used. This example is a fortunate one in that the constant c_n is obtained as an exact one. In many other cases, it is difficult to obtain exact solutions. Using an additive term of order $O_p(n^{-1})$ to the MLE, Ogasawara (2015) obtained a solution minimizing the asymptotic mean square error up to order $O(n^{-2})$ denoted by $\text{MSE}_{\rightarrow O(n^{-2})}$ in this paper.

The Euclid distance is an intuitively appealing natural one, especially for a single parameter. On the other hand, in the multi-parameter case, the overall error index for the MLEs of the parameters is defined in many ways since the MSEs for several estimators are given by the matrix MSE, which is the sum of the corresponding covariance matrix and the outer product of the same vectors of biases (see, e.g., Giles and Rayner 1979). One of the natural definitions of the overall index is the sum of the MSEs of the estimators, which is called the total MSE (TMSE) by Gruber (1998, p. 117). Ogasawara (2014b) obtained an optimal value of the added term of order $O_p(n^{-1})$ similar to the single parameter case, which minimizes the $\text{TMSE}_{\rightarrow O(n^{-2})}$ and a similar solution for the linear predictor given by the multiple MLEs with known weights for the predictor. Another overall index appropriate especially in the multi-parameter case is the log-likelihood, which will be used in this paper. Note that using this index, the problem of different ways of summarizing the matrix MSE is avoided. Further, since likelihood is used, the log-likelihood index is unchanged by reparametrization.

In the following section, the predictive expected log-likelihood will be defined and used. Then, a family of the estimators adjusting the MLE by adding a higher-order term is defined, whose asymptotically optimal value will be derived. The family has two typical sub-families giving shrinkage (or expanded) estimators and bias-adjusted estimators. Examples will be given using frequently used distributions (gamma, exponential, Poisson, Bernoulli and normal), where the corresponding optimal adjustment using the $\text{MSE}_{\rightarrow O(n^{-2})}$ will also be shown for comparison. The conditions for the relative sizes of the optimal coefficients by maximizing the asymptotic predictive expected log-likelihood and by minimizing the $\text{MSE}_{\rightarrow O(n^{-2})}$ will be given for a simple case.

2 The asymptotic predictive expected log-likelihood and its maximization

Let $\boldsymbol{\theta}$ be a $q \times 1$ vector of parameters with $\hat{\boldsymbol{\theta}}_{\text{ML}}$ and $\boldsymbol{\theta}_0$ being its MLE and the population counterpart, respectively. Let \mathbf{X} be a $n \times p$ matrix of n realized values of the random vector $\mathbf{x}^* = (x_1^*, \dots, x_p^*)'$. Then, the log-likelihood of $\boldsymbol{\theta}$ given \mathbf{X} averaged over n observations is denoted by $\bar{l}(\boldsymbol{\theta}, \mathbf{X})$. When \mathbf{X} is replaced by the corresponding

random matrix \mathbf{X}^* with its rows being independent copies of \mathbf{x}^* , we have $\bar{l}(\boldsymbol{\theta}, \mathbf{X}^*)$ whose expectation over its range is

$$\bar{l}^*(\boldsymbol{\theta}) \equiv E_{\mathbf{X}^*}\{\bar{l}(\boldsymbol{\theta}, \mathbf{X}^*)\} = \int_{R(\mathbf{Z})} \bar{l}(\boldsymbol{\theta}, \mathbf{Z})f(\mathbf{Z}|\boldsymbol{\theta}_0)d\mathbf{Z}, \tag{1}$$

where $f(\mathbf{Z}|\boldsymbol{\theta})$ is the density of $\mathbf{X}^* = \mathbf{Z}$. In the case of discrete distributions, (1) should be replaced by the corresponding summation with $f(\mathbf{Z}|\boldsymbol{\theta}_0)$ seen as a probability mass. The expectation (1) is maximized when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ by Jensen’s inequality. That is, $\hat{\boldsymbol{\theta}}_{ML}$ does not maximize $\bar{l}^*(\boldsymbol{\theta})$, which is the expected log-likelihood of $\boldsymbol{\theta}$.

It is known that under regularity conditions

$$E_{\mathbf{X}^*}\{\bar{l}^*(\hat{\boldsymbol{\theta}}_{ML})\} = \bar{l}^*(\boldsymbol{\theta}_0) - n^{-1}(q/2) + O(n^{-2}) \tag{2}$$

(see, e.g., Sakamoto et al. 1986, Eq. (4.21)).

The asymptotic difference $n^{-1}(q/2)$, when multiplied by 2, is a half of n^{-1} times the correction term in the Akaike information criterion (AIC; Akaike 1973). The amount of the difference is seen as an undesirable bias of $\hat{\boldsymbol{\theta}}_{ML}$. In this paper by a modification of $\hat{\boldsymbol{\theta}}_{ML}$, say $\tilde{\boldsymbol{\theta}}$, we will maximize $\bar{l}^*(\tilde{\boldsymbol{\theta}})$ among a family of the estimators using optimal coefficients yielding $\tilde{\boldsymbol{\theta}}$ from $\hat{\boldsymbol{\theta}}_{ML}$. So, we will maximize the following quantity:

$$E_{\mathbf{X}^*}\{\bar{l}^*\{\tilde{\boldsymbol{\theta}}(\mathbf{X}^*)\}\} = \int_{R(\mathbf{X})} \left[\int_{R(\mathbf{Z})} \bar{l}\{\tilde{\boldsymbol{\theta}}(\mathbf{X}), \mathbf{Z}\}f(\mathbf{Z}|\boldsymbol{\theta}_0)d\mathbf{Z} \right] f(\mathbf{X}|\boldsymbol{\theta}_0)d\mathbf{X}, \tag{3}$$

which is called the mean expected log-likelihood (Sakamoto et al. 1986, p. 60) when $\tilde{\boldsymbol{\theta}}$ is $\hat{\boldsymbol{\theta}}_{ML}$. Since \mathbf{Z} in (3) can be seen as a set of quantities associated with prediction to be given in the future, (3) is called the predictive expected log-likelihood in this paper.

When \mathbf{Z} is a set of quantities in the future, (1) to (3) are associated with predictive likelihood or the predictive posterior distribution though the direction of prediction is reversed, since \mathbf{Z} in the future is integrated out (see Fisher 1956, Chapter 5, Section 7; Hinkley 1979; Lejeune and Faulkenberry 1982; Leonard 1982; Bjornstad 1990; Konishi and Kitagawa 1996, 2008, Section 9.3).

Recall that when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\bar{l}^*(\boldsymbol{\theta})$ in (1) is maximized. Consequently, (3) is also maximized when $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Since $\boldsymbol{\theta}_0$ does not depend on \mathbf{X} , the maximum is $E_{\mathbf{X}^*}\{\bar{l}^*(\boldsymbol{\theta}_0)\} = \bar{l}^*(\boldsymbol{\theta}_0)$. Take the difference

$$\begin{aligned} \bar{l}^*(\boldsymbol{\theta}_0) - E_{\mathbf{X}^*}\{\bar{l}^*(\tilde{\boldsymbol{\theta}})\} &= E_{\mathbf{X}^*}\{\bar{l}^*(\boldsymbol{\theta}_0) - \bar{l}^*(\tilde{\boldsymbol{\theta}})\} \\ &\equiv \int_{R(\mathbf{X})} f(\mathbf{X}|\boldsymbol{\theta}_0) \log \frac{L^*(\boldsymbol{\theta}_0)}{L^*(\tilde{\boldsymbol{\theta}})} d\mathbf{X}, \end{aligned} \tag{4}$$

where $L^*(\boldsymbol{\theta}_0)$ and $L^*(\tilde{\boldsymbol{\theta}})$ are $\exp\{\bar{l}^*(\boldsymbol{\theta}_0)\}$ and $\exp\{\bar{l}^*(\tilde{\boldsymbol{\theta}})\}$, respectively. The quantity $\exp\{\bar{l}^*(\boldsymbol{\theta})\}$ is interpreted as a pseudo-likelihood of $\boldsymbol{\theta}$ corresponding to the predictive expected log-likelihood. The value of (4) is non-negative, and is seen as a discrepancy of $\boldsymbol{\theta}_0$ and $\tilde{\boldsymbol{\theta}}$ based on the Kullback and Leibler (1951) distance. So, the maximization of

the predictive expected log-likelihood in terms of $\tilde{\boldsymbol{\theta}}$ is given by finding $\tilde{\boldsymbol{\theta}}$ such that the distance of (5) is minimized. Incidentally, Lawless and Fredette (2005) used the Kullback–Leibler distance for evaluation of the goodness of an estimator.

In practice, $f(\mathbf{X}|\boldsymbol{\theta}_0)$ or $f(\mathbf{Z}|\boldsymbol{\theta}_0)$ may not be true, where another alternative true density of \mathbf{X}^* is denoted by $f_T(\mathbf{X}|\boldsymbol{\zeta}_0)$ with $\boldsymbol{\zeta}_0$ being a $q^* \times 1$ vector of parameters for the alternative model. The vector $\boldsymbol{\theta}_0$ under model misspecification is defined as $\hat{\boldsymbol{\theta}}_{ML}$ when infinitely many observations are available. Under model misspecification, we deal with the cases satisfying

$$E_T \left\{ \frac{\partial \bar{l}(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right\} = \mathbf{0} \quad (5)$$

where the subscript “T” indicates that the subscripted operator is defined under the alternative true distribution. That is, in this case the expectation is taken using the density $f_T(\mathbf{X}|\boldsymbol{\zeta}_0)$ rather than $f(\mathbf{X}|\boldsymbol{\theta}_0)$. Examples satisfying (5) will be illustrated in Sect. 4.5. In the following, the possible model misspecification is considered unless otherwise stated.

3 A family of the estimators maximizing the asymptotic predictive expected log-likelihood

3.1 The asymptotic predictive expected log-likelihood for the family of estimators

Define a family of estimators

$$\hat{\boldsymbol{\theta}}_{Ak} = \hat{\boldsymbol{\theta}}_{ML} - n^{-1} \text{diag}(\mathbf{k}) \hat{\boldsymbol{\alpha}}, \quad (6)$$

where $\text{diag}(\mathbf{k}) = \begin{bmatrix} k_1 & & & 0 \\ & k_2 & & \\ & & \ddots & \\ 0 & & & k_q \end{bmatrix}$, $\mathbf{k} = (k_1, \dots, k_q)'$ is the vector of

unknown coefficients to be derived, and $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}(\hat{\boldsymbol{\theta}}_{ML})$ is a differentiable arbitrary function of $\hat{\boldsymbol{\theta}}_{ML}$. Two typical cases are $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\theta}}_{ML}$ and $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\beta}}_1$, where $n^{-1} \hat{\boldsymbol{\beta}}_1$ is the vector of the asymptotic biases of the elements of $\hat{\boldsymbol{\theta}}_{ML}$ with the assumption that $E_T(\hat{\boldsymbol{\theta}}_{ML}) = \boldsymbol{\theta}_0 + n^{-1} \boldsymbol{\beta}_1 + O(n^{-2})$, and $\hat{\boldsymbol{\beta}}_1$ is a sample version of $\boldsymbol{\beta}_1$. As mentioned earlier, these cases give two sub-families of estimators. When $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\theta}}_{ML}$ and $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\beta}}_1$, the following notations are used:

$$\hat{\boldsymbol{\theta}}_{Sk} = \{\mathbf{I}_{(q)} - n^{-1} \text{diag}(\mathbf{k})\} \hat{\boldsymbol{\theta}}_{ML} \quad \text{and} \quad \hat{\boldsymbol{\theta}}_{Bk} = \hat{\boldsymbol{\theta}}_{ML} - n^{-1} \text{diag}(\mathbf{k}) \hat{\boldsymbol{\beta}}_1, \quad (7)$$

where S in $\hat{\boldsymbol{\theta}}_{Sk}$ indicates the shrinkage when k_i 's are hopefully positive, B in $\hat{\boldsymbol{\theta}}_{Bk}$ indicates bias adjustment which does not necessarily mean bias reduction, and $\mathbf{I}_{(q)}$ is the $q \times q$ identity matrix.

Recalling the definition of $\bar{l}^*(\boldsymbol{\theta})$ in (1), let

$$\frac{\partial^j \bar{l}^*}{(\partial \boldsymbol{\theta}_0)^{(j)}} \equiv \frac{\partial^j \bar{l}^*(\boldsymbol{\theta})}{(\partial \boldsymbol{\theta})^{(j)}} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \quad (j = 1, 2, 3) \tag{8}$$

for simplicity of notation, where $\mathbf{x}^{(j)} = \mathbf{x} \otimes \cdots \otimes \mathbf{x}$ (j times of \mathbf{x}) is the j -fold Kronecker product of \mathbf{x} . Then,

$$\frac{\partial \bar{l}^*}{\partial \boldsymbol{\theta}_0} = \mathbf{0} \quad \text{and} \quad \frac{\partial^j \bar{l}^*}{(\partial \boldsymbol{\theta}_0)^{(j)}} = O(1) \quad (j = 2, 3), \tag{9}$$

where $\bar{l}^*(\boldsymbol{\theta})$ does not include stochastic quantities \mathbf{X}^* or \mathbf{Z}^* until $\boldsymbol{\theta}$ is evaluated at, e.g., stochastic $\hat{\boldsymbol{\theta}}_{ML}$.

The predictive expected log-likelihood of $\hat{\boldsymbol{\theta}}_{Ak}$ under possible model misspecification is written as:

$$E_T \{ \bar{l}^*(\hat{\boldsymbol{\theta}}_{Ak}) \} = \int_{R(\mathbf{X})} \bar{l}^* \{ \boldsymbol{\theta}_{Ak}(\mathbf{X}) \} f_T(\mathbf{X} | \boldsymbol{\zeta}_0) d\mathbf{X}. \tag{10}$$

Note that

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{Ak} &= \hat{\boldsymbol{\theta}}_{ML} - n^{-1} \text{diag}(\mathbf{k}) \hat{\boldsymbol{\alpha}} \\ &= \hat{\boldsymbol{\theta}}_{ML} - n^{-1} \{ \text{diag}(\mathbf{k}) \boldsymbol{\alpha}_0 + \text{diag}(\mathbf{k}) (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \} \\ &= \hat{\boldsymbol{\theta}}_{ML} - n^{-1} \{ \text{diag}(\boldsymbol{\alpha}_0) + \text{diag}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \} \mathbf{k}, \end{aligned} \tag{11}$$

where $\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}(\boldsymbol{\theta}_0)$. Then, using the Taylor series expansion about $\boldsymbol{\theta}_0$ under regularity conditions with the assumptions $\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0 = O_p(n^{-1/2})$ and $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = O_p(n^{-1/2})$, we have

$$\begin{aligned} \bar{l}^* \{ \boldsymbol{\theta}_{Ak}(\mathbf{X}^*) \} &= \bar{l}^*(\hat{\boldsymbol{\theta}}_{Ak}) \equiv \hat{l}_k^* \\ &= \bar{l}^*(\boldsymbol{\theta}_0) + \sum_{j=2}^4 \left\{ \frac{\partial^j \bar{l}^*}{j! (\partial \boldsymbol{\theta}'_0)^{(j)}} (\hat{\boldsymbol{\theta}}_{Ak} - \boldsymbol{\theta}_0)^{(j)} \right\}_{O_p(n^{-j/2})} + O_p(n^{-5/2}) \\ &\equiv (\bar{l}^*)_{O(1)} + \sum_{j=2}^4 \left\{ \frac{\partial^j \bar{l}^*}{j! (\partial \boldsymbol{\theta}'_0)^{(j)}} (\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)^{(j)} \right\}_{O_p(n^{-j/2})} \\ &\quad - \left\{ n^{-1} \mathbf{k}' \text{diag}(\boldsymbol{\alpha}_0) \frac{\partial^2 \bar{l}^*}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}'_0} (\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) \right\}_{O_p(n^{-3/2})} \end{aligned} \tag{12}$$

$$\begin{aligned}
 & - \left\{ n^{-1} \mathbf{k}' \text{diag}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \frac{\partial^2 \bar{l}^*}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} (\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}_0) \right\}_{O_p(n^{-2})} \\
 & + \left\{ n^{-2} \frac{1}{2} \mathbf{k}' \text{diag}(\boldsymbol{\alpha}_0) \frac{\partial^2 \bar{l}^*}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} \text{diag}(\boldsymbol{\alpha}_0) \mathbf{k} \right\}_{O(n^{-2})} \\
 & - \left\{ n^{-1} \frac{1}{2} \mathbf{k}' \text{diag}(\boldsymbol{\alpha}_0) \frac{\partial^3 \bar{l}^*}{\partial \boldsymbol{\theta}_0 (\partial \boldsymbol{\theta}_0')^2} (\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}_0)^{(2)} \right\}_{O_p(n^{-2})} + O_p(n^{-5/2}),
 \end{aligned}$$

where $\{\cdot\}_{O_p(\cdot)}$ and $\{\cdot\}_{O(\cdot)}$ indicate the orders of the quantities in braces for clarity.

Define

$$\boldsymbol{\Lambda} \equiv \frac{\partial^2 \bar{l}^*}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} = O(1) \quad \text{and} \quad \boldsymbol{\Gamma} \equiv n E_T \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0'} \right) = O(1) \tag{13}$$

where $\partial \bar{l} / \partial \boldsymbol{\theta}_0 = \partial \bar{l}(\boldsymbol{\theta}, \mathbf{X}^*) / \partial \boldsymbol{\theta} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = O_p(n^{-1/2})$ (compare with $\partial \bar{l}^* / \partial \boldsymbol{\theta}_0 = \mathbf{0}$). Let $\hat{l}_{\text{ML}}^* \equiv \bar{l}^*(\hat{\boldsymbol{\theta}}_{\text{ML}})$. Then, the expectation of (12) is

$$\begin{aligned}
 E_T(\hat{l}_k^*) &= E_T(\hat{l}_{\text{ML}}^*) \\
 &+ n^{-2} \mathbf{k}' \left[-\text{diag}(\boldsymbol{\alpha}_0) \boldsymbol{\Lambda} \boldsymbol{\beta}_1 - n E_T \{ (\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}_0)' \otimes \text{diag}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \} \text{vec}(\boldsymbol{\Lambda}) \right. \\
 &\quad \left. - \frac{1}{2} \{ \text{vec}'(\boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1}) \otimes \text{diag}(\boldsymbol{\alpha}_0) \} \frac{\partial^3 \bar{l}^*}{(\partial \boldsymbol{\theta}_0)^{(3)}} \right] \\
 &- n^{-2} \frac{1}{2} \mathbf{k}' \text{diag}(\boldsymbol{\alpha}_0) (-\boldsymbol{\Lambda}) \text{diag}(\boldsymbol{\alpha}_0) \mathbf{k} + O(n^{-3})
 \end{aligned} \tag{14}$$

with

$$\begin{aligned}
 & n E_T \{ (\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}_0)' \otimes \text{diag}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \} \\
 &= \left[\text{diag}[n \text{acov}\{(\hat{\boldsymbol{\theta}}_{\text{ML}})_1, \hat{\boldsymbol{\alpha}}\}], \dots, \text{diag}[n \text{acov}\{(\hat{\boldsymbol{\theta}}_{\text{ML}})_q, \hat{\boldsymbol{\alpha}}\}] \right] + O(n^{-1}) \\
 &= \left[\text{diag} \left\{ \left(\boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right)_1 \right\}, \dots, \text{diag} \left\{ \left(\boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right)_q \right\} \right] + O(n^{-1}) \tag{15} \\
 &\equiv \text{diag} \left\{ \text{row} \left(\boldsymbol{\Lambda}^{-1} \boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \right\} + O(n^{-1}),
 \end{aligned}$$

where $\text{vec}(\cdot)$ is the vectorizing operator stacking the columns of a matrix sequentially with $\text{vec}'(\cdot) = \{\text{vec}(\cdot)\}'$, $\text{acov}(\cdot, \cdot)$ is the asymptotic covariance of order $O(n^{-1})$ for two variables, $\text{diag}(\mathbf{x}) = \text{diag}(\mathbf{x}')$, $(\cdot)_j$ indicates the j -th element of a vector, and $(\cdot)_j$ is the j -th row of a matrix. Since the a -th element of

$$\text{diag}\left\{\text{row}\left(\Lambda^{-1}\Gamma\Lambda^{-1}\frac{\partial\alpha'_0}{\partial\theta_0}\right)\right\}\text{vec}(\Lambda)$$

is

$$\begin{aligned} \sum_{b=1}^q \left\{ \Lambda^{-1}\Gamma\Lambda^{-1}\frac{\partial(\alpha_0)_a}{\partial\theta_0} \right\}_b (\Lambda)_{ab} &= \left\{ \Lambda\Lambda^{-1}\Gamma\Lambda^{-1}\frac{\partial(\alpha_0)_a}{\partial\theta_0} \right\}_a \\ &= \left\{ \Gamma\Lambda^{-1}\frac{\partial(\alpha_0)_a}{\partial\theta_0} \right\}_a, \end{aligned} \tag{16}$$

where $(\cdot)_{ab}$ indicates the (a, b) th element of a matrix, we have

$$\begin{aligned} &\text{diag}\left\{\text{row}\left(\Lambda^{-1}\Gamma\Lambda^{-1}\frac{\partial\alpha'_0}{\partial\theta_0}\right)\right\}\text{vec}(\Lambda) \\ &= \left\{ \left(\Gamma\Lambda^{-1}\frac{\partial\alpha'_0}{\partial\theta_0}\right)_{11}, \dots, \left(\Gamma\Lambda^{-1}\frac{\partial\alpha'_0}{\partial\theta_0}\right)_{qq} \right\}' = \text{Diag}\left(\Gamma\Lambda^{-1}\frac{\partial\alpha'_0}{\partial\theta_0}\right)\mathbf{1}_{(q)}, \end{aligned} \tag{17}$$

where $\text{Diag}(\cdot)$ is the diagonal matrix whose diagonal elements are those of a matrix in parentheses and $\mathbf{1}_{(q)}$ is the $q \times 1$ vector of 1's.

It is known that β_1 under possible model misspecification is given as:

$$\begin{aligned} \beta_1 &= \Lambda^{-1}nE_T\left(\frac{\partial\bar{l}}{\partial\theta'_0} \otimes \mathbf{M}\right)\text{vec}(\Lambda^{-1}) \\ &\quad - \frac{1}{2}\Lambda^{-1}\frac{\partial^3\bar{l}^*}{\partial\theta_0(\partial\theta'_0)^{(2)}}\text{vec}(\Lambda^{-1}\Gamma\Lambda^{-1}), \end{aligned} \tag{18}$$

where $\partial^2\bar{l}/\partial\theta_0\partial\theta'_0 - E_T(\partial^2\bar{l}/\partial\theta_0\partial\theta'_0) \equiv (\mathbf{M})_{O_p(n^{-1/2})}$. The result of (18) is typically derived by the inverse expansion of $\partial\bar{l}/\partial\theta|_{\theta=\hat{\theta}_{ML}} = \mathbf{0}$ at $\theta = \theta_0$ in terms of $\hat{\theta}_{ML} - \theta_0$ (see, e.g., Ogasawara, 2010, Eq. (2.4)). Using (18), the first term in brackets on the right-hand side of (14) becomes

$$\begin{aligned} -\text{diag}(\alpha_0)\Lambda\beta_1 &= -\text{diag}(\alpha_0)\Lambda\left\{ \Lambda^{-1}nE_T\left(\frac{\partial\bar{l}}{\partial\theta'_0} \otimes \mathbf{M}\right)\text{vec}(\Lambda^{-1}) \right. \\ &\quad \left. - \frac{1}{2}\Lambda^{-1}\frac{\partial^3\bar{l}^*}{\partial\theta_0(\partial\theta'_0)^{(2)}}\text{vec}(\Lambda^{-1}\Gamma\Lambda^{-1}) \right\} \\ &= -\text{diag}(\alpha_0)nE_T\left(\frac{\partial\bar{l}}{\partial\theta'_0} \otimes \mathbf{M}\right)\text{vec}(\Lambda^{-1}) \\ &\quad + \frac{1}{2}\left\{ \text{vec}'(\Lambda^{-1}\Gamma\Lambda^{-1}) \otimes \text{diag}(\alpha_0) \right\} \frac{\partial^3\bar{l}^*}{(\partial\theta_0)^{(3)}}. \end{aligned} \tag{19}$$

From (17) and (19), we have

$$\begin{aligned}
 E_T(\hat{l}_k^*) &= E_T(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} \\
 &+ n^{-2} \left[\mathbf{k}' \left\{ -\text{diag}(\boldsymbol{\alpha}_0) n E_T \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\boldsymbol{\Lambda}^{-1}) \right. \right. \\
 &\quad \left. \left. - \text{Diag} \left(\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \mathbf{1}_{(q)} \right\} - \frac{1}{2} \mathbf{k}' \text{diag}(\boldsymbol{\alpha}_0) (-\boldsymbol{\Lambda}) \text{diag}(\boldsymbol{\alpha}_0) \mathbf{k} \right] \\
 &+ O(n^{-3}) \\
 &\equiv E_T(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} + n^{-2} (\mathbf{k}' \mathbf{b}_A - \mathbf{k}' \mathbf{A}_A \mathbf{k}) + O(n^{-3}),
 \end{aligned} \tag{20}$$

where $E_T(\cdot)_{\rightarrow O(n^{-2})}$ indicates that the expectation is taken up to order $O(n^{-2})$, and

$$\begin{aligned}
 \mathbf{b}_A &= -\text{diag}(\boldsymbol{\alpha}_0) n E_T \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\boldsymbol{\Lambda}^{-1}) - \text{Diag} \left(\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \mathbf{1}_{(q)}, \\
 \mathbf{A}_A &= \frac{1}{2} \text{diag}(\boldsymbol{\alpha}_0) (-\boldsymbol{\Lambda}) \text{diag}(\boldsymbol{\alpha}_0).
 \end{aligned} \tag{21}$$

3.2 The vector of maximizing coefficients

The \mathbf{k} maximizing (20) without the remainder term is given by

$$\begin{aligned}
 \mathbf{k}_{Amax} &= \frac{1}{2} \mathbf{A}_A^{-1} \mathbf{b}_A \\
 &= \text{diag}^{-1}(\boldsymbol{\alpha}_0) \boldsymbol{\Lambda}^{-1} \text{diag}^{-1}(\boldsymbol{\alpha}_0) \\
 &\quad \times \left\{ \text{diag}(\boldsymbol{\alpha}_0) n E_T \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\boldsymbol{\Lambda}^{-1}) + \text{Diag} \left(\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \mathbf{1}_{(q)} \right\} \\
 &= \text{diag}^{-1}(\boldsymbol{\alpha}_0) \boldsymbol{\Lambda}^{-1} \\
 &\quad \times \left\{ n E_T \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\boldsymbol{\Lambda}^{-1}) + \text{diag}^{-1}(\boldsymbol{\alpha}_0) \text{Diag} \left(\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \mathbf{1}_{(q)} \right\},
 \end{aligned} \tag{22}$$

where $\text{diag}^{-1}(\boldsymbol{\alpha}_0) = \{\text{diag}(\boldsymbol{\alpha}_0)\}^{-1}$ with the assumption of its existence.

The maximizing value without the remainder term is

$$\begin{aligned}
 E_T(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} &+ \frac{n^{-2}}{4} \mathbf{b}'_A \mathbf{A}_A^{-1} \mathbf{b}_A \\
 &= E_T(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} \\
 &+ \frac{n^{-2}}{2} \left\{ \text{vec}'(\boldsymbol{\Lambda}^{-1}) n E_T \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) + \mathbf{1}'_{(q)} \text{Diag} \left(\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \text{diag}^{-1}(\boldsymbol{\alpha}_0) \right\} \\
 &\times (-\boldsymbol{\Lambda}^{-1}) \\
 &\times \left\{ n E_T \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\boldsymbol{\Lambda}^{-1}) + \text{diag}^{-1}(\boldsymbol{\alpha}_0) \text{Diag} \left(\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \mathbf{1}_{(q)} \right\}.
 \end{aligned} \tag{23}$$

As will be addressed later, (22) and (23) become considerably simplified in the case of the canonical parameters of the exponential family since $\partial^j \bar{l} / (\partial \boldsymbol{\theta}_0)^{(j)} = \partial^j \bar{l}^* / (\partial \boldsymbol{\theta}_0)^{(j)}$ ($j = 2, 3, \dots$) and $\mathbf{M} = \mathbf{O}$. The result is summarized.

Result 1 *The vector \mathbf{k} of coefficients maximizing the asymptotic predictive expected log-likelihood averaged over observations up to order $O(n^{-2})$ among the family of the estimators $\hat{\boldsymbol{\theta}}_{Ak} = \hat{\boldsymbol{\theta}}_{ML} - n^{-1} \text{diag}(\mathbf{k}) \hat{\boldsymbol{\alpha}}$ using a differentiable vector function $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}(\hat{\boldsymbol{\theta}}_{ML})$ under possible model misspecification is given by \mathbf{k}_{Amax} in (22) with the maximized value being (23).*

When a common coefficient k in \mathbf{k} is used, \mathbf{k} becomes $\mathbf{k} = k\mathbf{1}_{(q)}$. The result in this case is summarized as:

Result 2 *When $\mathbf{k} = k\mathbf{1}_{(q)}$ in Result 1, b_A and a_A become scalars as:*

$$b_A = -\boldsymbol{\alpha}'_0 n E_T \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\boldsymbol{\Lambda}^{-1}) - \text{tr} \left(\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \tag{24}$$

$$\text{and } a_A = \frac{1}{2} \boldsymbol{\alpha}'_0 (-\boldsymbol{\Lambda}) \boldsymbol{\alpha}_0, \tag{25}$$

respectively. Then, the maximizing k in $\mathbf{k} = k\mathbf{1}_{(q)}$ is

$$\begin{aligned} k_{Amax} &= \frac{b_A}{2a_A} \\ &= \frac{1}{\boldsymbol{\alpha}'_0 \boldsymbol{\Lambda} \boldsymbol{\alpha}_0} \left\{ \boldsymbol{\alpha}'_0 n E_T \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\boldsymbol{\Lambda}^{-1}) + \text{tr} \left(\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \right\} \end{aligned} \tag{26}$$

with the maximized value without the remainder term being

$$\begin{aligned} &E_T(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} + n^{-2} \frac{b_A^2}{4a_A} \\ &= E_T(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} \\ &\quad - \frac{n^{-2}}{2\boldsymbol{\alpha}'_0 \boldsymbol{\Lambda} \boldsymbol{\alpha}_0} \left\{ \boldsymbol{\alpha}'_0 n E_T \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\boldsymbol{\Lambda}^{-1}) + \text{tr} \left(\boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \right\}^2. \end{aligned} \tag{27}$$

In the case of the canonical parameters in the exponential family, \mathbf{M} vanishes in Result 2 as in Result 1. When $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\theta}}_{ML}$, $\partial \boldsymbol{\alpha}'_0 / \partial \boldsymbol{\theta}_0$ becomes $\mathbf{I}_{(q)}$ and can be omitted in Results 1 and 2. In this case, the contribution of the trace term in (26) to k_{Amax} is

$$\frac{\text{tr}(\Gamma\Lambda^{-1})}{\alpha_0'\Lambda\alpha_0} = \frac{\text{tr}\{-\Lambda(\Lambda^{-1}\Gamma\Lambda^{-1})\}}{\alpha_0'(-\Lambda)\alpha_0} = \frac{\text{tr}\{n^{-1} \text{acov}_{\theta_0}^{-1}(\hat{\theta}_{ML})n \text{acov}_T(\hat{\theta}_{ML})\}}{\alpha_0'n^{-1} \text{acov}_{\theta_0}^{-1}(\hat{\theta}_{ML})\alpha_0}, \tag{28}$$

where $\text{acov}_{\theta_0}(\hat{\theta}_{ML})$ is the asymptotic covariance matrix of order $O(n^{-1})$ for $\hat{\theta}_{ML}$ under correct model specification, $\text{acov}_T(\hat{\theta}_{ML})$ is the corresponding matrix under possible model misspecification, and $\text{acov}_{\theta_0}^{-1}(\hat{\theta}_{ML}) = \{\text{acov}_{\theta_0}(\hat{\theta}_{ML})\}^{-1}$. Since the trace in (28) becomes q under correct model specification, the scaled trace value $\text{tr}\{n^{-1} \text{acov}_{\theta_0}^{-1}(\hat{\theta}_{ML})n \text{acov}_T(\hat{\theta}_{ML})\}/q$ is called the dispersion ratio in this paper. When the dispersion ratio is relatively large, $k_{A_{\max}}$ in (26) tends to be large.

An advantage of the restricted vector $\mathbf{k} = k\mathbf{1}_{(q)}$ over the unrestricted one is that the solution $k_{A_{\max}}$ is available even when some but not all of the elements of α_0 are zero. In this case, $\text{diag}^{-1}(\alpha_0)$ was unavailable unless \mathbf{k} is shortened by omitting its elements corresponding to the zero elements of α_0 .

When the model is correctly specified, we have $-\Lambda = \Gamma = \mathbf{I}_0$, where \mathbf{I}_0 is the Fisher information matrix per observation and the dispersion ratio becomes 1, which gives

Result 3 *The vector \mathbf{k} of coefficients maximizing the asymptotic predictive expected log-likelihood under the conditions of Result 1 and, in addition, under correct model specification is given by*

$$\mathbf{b}_A = \text{diag}(\alpha_0)nE_{\theta_0}\left(\frac{\partial \bar{l}}{\partial \theta'} \otimes \mathbf{M}\right)\text{vec}(\mathbf{I}_0^{-1}) + \text{Diag}\left(\frac{\partial \alpha'_0}{\partial \theta_0}\right)\mathbf{1}_{(q)} \tag{29}$$

$$\text{and } \mathbf{A}_A = \frac{1}{2}\text{diag}(\alpha_0)\mathbf{I}_0\text{diag}(\alpha_0) \tag{30}$$

yielding

$$\begin{aligned} \mathbf{k}_{A_{\max}} &= \frac{1}{2}\mathbf{A}_A^{-1}\mathbf{b}_A \\ &= \text{diag}^{-1}(\alpha_0)\mathbf{I}_0^{-1} \\ &\times \left\{ nE_{\theta_0}\left(\frac{\partial \bar{l}}{\partial \theta'} \otimes \mathbf{M}\right)\text{vec}(\mathbf{I}_0^{-1}) + \text{diag}^{-1}(\alpha_0)\text{Diag}\left(\frac{\partial \alpha'_0}{\partial \theta_0}\right)\mathbf{1}_{(q)} \right\}, \end{aligned} \tag{31}$$

where $E_{\theta_0}(\cdot)$ indicates that the expectation is taken under correct model specification. The maximized value without the remainder term is

$$\begin{aligned}
 & E_{\theta_0}(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} + \frac{n^{-2}}{4} \mathbf{b}'_A \mathbf{A}_A^{-1} \mathbf{b}_A \\
 &= E_{\theta_0}(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} \\
 &+ \frac{n^{-2}}{2} \left\{ \text{vec}'(\mathbf{I}_0^{-1}) n E_{\theta_0} \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0} \otimes \mathbf{M} \right) + \mathbf{1}'_{(q)} \text{Diag} \left(\frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \text{diag}^{-1}(\boldsymbol{\alpha}_0) \right\} \\
 &\times \mathbf{I}_0^{-1} \left\{ n E_{\theta_0} \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\mathbf{I}_0^{-1}) + \text{diag}^{-1}(\boldsymbol{\alpha}_0) \text{Diag} \left(\frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \mathbf{1}_{(q)} \right\}.
 \end{aligned} \tag{32}$$

As in Result 2, we have

Result 4 When $\mathbf{k} = k\mathbf{1}_{(q)}$ in Result 3 under correct model specification,

$$\begin{aligned}
 b_A &= \boldsymbol{\alpha}'_0 n E_{\theta_0} \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\mathbf{I}_0^{-1}) + \text{tr} \left(\frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right), \\
 a_A &= \boldsymbol{\alpha}'_0 \mathbf{I}_0 \boldsymbol{\alpha}_0 / 2
 \end{aligned} \tag{33}$$

and

$$k_{Amax} = \frac{b_A}{2a_A} = \frac{1}{\boldsymbol{\alpha}'_0 \mathbf{I}_0 \boldsymbol{\alpha}_0} \left\{ \boldsymbol{\alpha}'_0 n E_{\theta_0} \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\mathbf{I}_0^{-1}) + \text{tr} \left(\frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \right\}. \tag{34}$$

The maximized value without the remainder term is

$$\begin{aligned}
 & E_{\theta_0}(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} + n^{-2} \frac{b_A^2}{4a_A} \\
 &= E_{\theta_0}(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} + \frac{n^{-2}}{2\boldsymbol{\alpha}'_0 \mathbf{I}_0 \boldsymbol{\alpha}_0} \left\{ \boldsymbol{\alpha}'_0 n E_{\theta_0} \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \otimes \mathbf{M} \right) \text{vec}(\mathbf{I}_0^{-1}) + \text{tr} \left(\frac{\partial \boldsymbol{\alpha}'_0}{\partial \boldsymbol{\theta}_0} \right) \right\}^2.
 \end{aligned} \tag{35}$$

As before, canonical parametrization in the exponential family gives simplification. Under non-canonical parametrization, an example of the vanishing term with \mathbf{M} after taking the expectation is the normal distribution as will be illustrated later.

When $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\theta}}_{ML}$ with the vanishing \mathbf{M} , k_{Amax} of (34) becomes as simple as:

$$k_{Amax} = \frac{q}{\boldsymbol{\alpha}'_0 \mathbf{I}_0 \boldsymbol{\alpha}_0} \tag{36}$$

with the maximized value

$$E_{\theta_0}(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})} + n^{-2} \frac{q^2}{2\boldsymbol{\alpha}'_0 \mathbf{I}_0 \boldsymbol{\alpha}_0}. \tag{37}$$

Note that the term $E_{\mathbf{T}}(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})}$ or $E_{\theta_0}(\hat{l}_{ML}^*)_{\rightarrow O(n^{-2})}$ appears throughout the associated results for clarity and completeness though the term is irrelevant for

deriving \mathbf{k}_{Amax} . So, the term is seen as a constant. However, when the evaluation of the term is required, the result can be obtained, which is given in the appendix. The term of order $O(n^{-1})$ in $E_{\theta_0}(\hat{\mathbf{I}}_{\text{ML}}^*)_{\rightarrow O(n^{-2})}$ was given in (2) as $-n^{-1}q/2$, where q is the q times the unit dispersion ratio under correct model specification. Under possible model misspecification, the term is $n^{-1}\text{tr}(\mathbf{\Gamma}\mathbf{\Lambda}^{-1})/2$, where $-\text{tr}(\mathbf{\Gamma}\mathbf{\Lambda}^{-1})$ is also the q times the dispersion ratio.

The terms $E_{\text{T}}(\hat{\mathbf{I}}_{\text{ML}}^*)_{\rightarrow O(n^{-2})}$ and $E_{\theta_0}(\hat{\mathbf{I}}_{\text{ML}}^*)_{\rightarrow O(n^{-2})}$ are associated with the asymptotic bias correction higher than those in the AIC and Takeuchi information criterion (TIC; Takeuchi 1976; Stone 1977; Konishi and Kitagawa 2008), which was derived by Konishi and Kitagawa (2003), and Ogasawara (2017) in different expressions. In this bias correction, many of the terms in $E_{\theta_0}(\hat{\mathbf{I}}_{\text{ML}}^*)_{\rightarrow O(n^{-2})}$ and $E_{\text{T}}(\hat{\mathbf{I}}_{\text{ML}}^*)_{\rightarrow O(n^{-2})}$ are cancelled since the amount of bias correction is $E_{\theta_0}(\hat{\mathbf{I}}_{\text{ML}}^*)_{\rightarrow O(n^{-2})} - E_{\theta_0}(\hat{\mathbf{I}}_{\text{ML}}^*)_{\rightarrow O(n^{-2})}$ or $E_{\text{T}}(\hat{\mathbf{I}}_{\text{ML}}^*)_{\rightarrow O(n^{-2})} - E_{\text{T}}(\hat{\mathbf{I}}_{\text{ML}}^*)_{\rightarrow O(n^{-2})}$.

The maximizing coefficients in Results 1–4 depend on unknown population values, e.g., $\boldsymbol{\alpha}_0$. Although we can have their sample counterparts, it can be shown that when the population values are replaced by their estimates, the associated asymptotic results do not hold. This difficulty can be solved by several ways, which will be addressed after (46) in the next section.

4 Examples

In this section, examples are given using typical distributions in statistics. The results in Sects. 4.1 to 4.4 are obtained under correct model specification while those in Sect. 4.5 are shown under model misspecification. In Sect. 4.4, the multi-parameter cases are illustrated using the normal distribution. In these examples, the corresponding results derived by minimizing $\text{MSE}_{\rightarrow O(n^{-2})}$ are also given for comparison. Ogasawara (2015) obtained the coefficient of the bias adjustment minimizing the $\text{MSE}_{\rightarrow O(n^{-2})}$ for a single parameter and a parameter in the parameter vector. Ogasawara (2014b) extended the result to the multi-parameter case, where all the parameters in a vector parameter are simultaneously considered. Ogasawara (2014b) used $\text{TMSE}_{\rightarrow O(n^{-2})}$ (see Sect. 1) and the linear predictor, where only a single coefficient k was used. For comparison to the results in Sect. 4.4, Ogasawara's (2014b) result is extended using unconstrained vector \mathbf{k} of coefficients and an arbitrary vector $\hat{\boldsymbol{\alpha}}$ as follows.

Let $\hat{\boldsymbol{\theta}}_{Ak} = \hat{\boldsymbol{\theta}}_{\text{ML}} - n^{-1}\text{diag}(\mathbf{k})\hat{\boldsymbol{\alpha}}$ as before. Then, under possible model misspecification

$$\begin{aligned}
 \text{TMSE} &= E_T\{(\hat{\boldsymbol{\theta}}_{Ak} - \boldsymbol{\theta}_0)'(\hat{\boldsymbol{\theta}}_{Ak} - \boldsymbol{\theta}_0)\} \\
 &= n^{-1}\mathbf{1}'_{(q)}\boldsymbol{\beta}_2 + n^{-2}\left[\mathbf{1}'_{(q)}\boldsymbol{\beta}_{\Delta 2} - 2\text{tr}\{\text{diag}(\mathbf{k})n\text{acov}_T(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}'_{ML})\} \right. \\
 &\quad \left. + \{\boldsymbol{\beta}_1 - \text{diag}(\mathbf{k})\boldsymbol{\alpha}_0\}'\{\boldsymbol{\beta}_1 - \text{diag}(\mathbf{k})\boldsymbol{\alpha}_0\}\right] + O(n^{-3}) \\
 &= n^{-1}\mathbf{1}'_{(q)}\boldsymbol{\beta}_2 \\
 &\quad + n^{-2}\left[\mathbf{1}'_{(q)}\boldsymbol{\beta}_{\Delta 2} + \boldsymbol{\beta}'_1\boldsymbol{\beta}_1 - 2\mathbf{k}'\text{Diag}\{n\text{acov}_T(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}'_{ML}) + \boldsymbol{\alpha}_0\boldsymbol{\beta}'_1\}\mathbf{1}_{(q)} \right. \\
 &\quad \left. + \mathbf{k}'\text{Diag}(\boldsymbol{\alpha}_0\boldsymbol{\alpha}'_0)\mathbf{k}\right] + O(n^{-3}) \\
 &= \text{TMSE}_{-O(n^{-2})} + O(n^{-3}),
 \end{aligned}
 \tag{38}$$

where $n^{-1}\boldsymbol{\beta}_2$ is the $q \times 1$ vector of the asymptotic variances of order $O(n^{-1})$ for the elements of $\hat{\boldsymbol{\theta}}_{ML}$, and $n^{-2}\boldsymbol{\beta}_{\Delta 2}$ is the corresponding vector of the added higher-order asymptotic variances of order $O(n^{-2})$. Then,

Result 5 *The vector of the coefficients minimizing $\text{TMSE}_{-O(n^{-2})}$ defined in (38) under possible model misspecification is*

$$\mathbf{k}_{Amin} = \text{Diag}^{-1}(\boldsymbol{\alpha}_0\boldsymbol{\alpha}'_0)\text{Diag}\{n\text{acov}_T(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}'_{ML}) + \boldsymbol{\alpha}_0\boldsymbol{\beta}'_1\}\mathbf{1}_{(q)}
 \tag{39}$$

with the minimized $\text{TMSE}_{-O(n^{-2})}$ being

$$\begin{aligned}
 n^{-1}\mathbf{1}'_{(q)}\boldsymbol{\beta}_2 + n^{-2}\left[\mathbf{1}'_{(q)}\boldsymbol{\beta}_{\Delta 2} + \boldsymbol{\beta}'_1\boldsymbol{\beta}_1 - \mathbf{1}'_{(q)}\text{Diag}\{n\text{acov}_T(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}'_{ML}) + \boldsymbol{\alpha}_0\boldsymbol{\beta}'_1\} \right. \\
 \left. \times \text{Diag}^{-1}(\boldsymbol{\alpha}_0\boldsymbol{\alpha}'_0)\text{Diag}\{n\text{acov}_T(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}'_{ML}) + \boldsymbol{\alpha}_0\boldsymbol{\beta}'_1\}\mathbf{1}_{(q)}\right]
 \end{aligned}
 \tag{40}$$

Result 6 *When $\mathbf{k} = k\mathbf{1}_{(q)}$ in Result 5,*

$$k_{Amin} = \frac{1}{\boldsymbol{\alpha}'_0\boldsymbol{\alpha}_0} [\text{tr}\{n\text{acov}_T(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}'_{ML})\} + \boldsymbol{\alpha}'_0\boldsymbol{\beta}_1]
 \tag{41}$$

with the minimized $\text{TMSE}_{-O(n^{-2})}$ being

$$n^{-1}\mathbf{1}'_{(q)}\boldsymbol{\beta}_2 + n^{-2}\left[\mathbf{1}'_{(q)}\boldsymbol{\beta}_{\Delta 2} + \boldsymbol{\beta}'_1\boldsymbol{\beta}_1 - \frac{1}{\boldsymbol{\alpha}'_0\boldsymbol{\alpha}_0} [\text{tr}\{n\text{acov}_T(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}'_{ML})\} + \boldsymbol{\alpha}'_0\boldsymbol{\beta}_1]^2\right]
 \tag{42}$$

Result 5 is new. Result 6 is slightly generalized than that in Ogasawara (2014b) in that an arbitrary vector $\hat{\boldsymbol{\alpha}}$ is used. When $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\theta}}_{ML}$ and $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\beta}}_1$, \mathbf{k}_{Amin} is written as \mathbf{k}_{Smin} and \mathbf{k}_{Bmin} , respectively.

As addressed earlier, canonical parametrization in the exponential family gives simplified results. Two additional results by this parametrization under correct model specification are given. For this case, it is known that

$$\boldsymbol{\beta}_1 = -\frac{1}{2}\mathbf{I}_0^{-1}\mathbf{I}_0^{(D1)}\text{vec}(\mathbf{I}_0^{-1})
 \tag{43}$$

(see, e.g., Ogasawara, 2013), where $\mathbf{I}_0^{(Dj)} = -\partial^{j+2}\bar{l}/\partial\boldsymbol{\theta}_0(\partial\boldsymbol{\theta}'_0)^{(j+1)}$ ($j = 1, 2, \dots$) with $j = 2$ for later use, which is the j -th derivative of the information matrix evaluated at the population value. From (43), after some algebra, we have

$$\begin{aligned} \frac{\partial\boldsymbol{\beta}'_1}{\partial\boldsymbol{\theta}_0} &= \frac{1}{2}\mathbf{I}_0^{(D1)}[\{\mathbf{I}_0^{-1}\mathbf{I}_0^{(D1)}\text{vec}(\mathbf{I}_0^{-1})\} \otimes \mathbf{I}_0^{-1}] - \frac{1}{2}\mathbf{I}_0^{(D2)}\{\text{vec}(\mathbf{I}_0^{-1}) \otimes \mathbf{I}_0^{-1}\} \\ &\quad + \frac{1}{2}\mathbf{I}_0^{(D1)}\mathbf{I}_0^{-1(2)}\mathbf{I}_0^{(D1)'}\mathbf{I}_0^{-1}. \end{aligned} \tag{44}$$

Recalling Result 4 and defining $k_{B\max}$, b_B and a_B as $k_{A\max}$, b_A and a_A , respectively, when $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\beta}}_1$ ($k_{S\max}$, b_S and a_S are similarly defined for later use when $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\theta}}_{ML}$), (44) gives

$$\begin{aligned} k_{B\max} &= \frac{b_B}{2a_B} = \frac{\text{tr}(\partial\boldsymbol{\beta}'_1/\partial\boldsymbol{\theta}_0)}{\boldsymbol{\beta}'_1\mathbf{I}_0\boldsymbol{\beta}_1} \\ &= \frac{2}{\text{vec}'(\mathbf{I}_0^{-1})\mathbf{I}_0^{(D1)'}\mathbf{I}_0^{-1}\mathbf{I}_0^{(D1)}\text{vec}(\mathbf{I}_0^{-1})} \text{tr}\left[\mathbf{I}_0^{(D1)}[\{\mathbf{I}_0^{-1}\mathbf{I}_0^{(D1)}\text{vec}(\mathbf{I}_0^{-1})\} \otimes \mathbf{I}_0^{-1}] \right. \\ &\quad \left. - \mathbf{I}_0^{(D2)}\{\text{vec}(\mathbf{I}_0^{-1}) \otimes \mathbf{I}_0^{-1}\} + \mathbf{I}_0^{(D1)}\mathbf{I}_0^{-1(2)}\mathbf{I}_0^{(D1)'}\mathbf{I}_0^{-1}\right]. \end{aligned} \tag{45}$$

In the case of a single parameter, it is known that for $\hat{\theta}_{Bk} = \hat{\theta}_{ML} - n^{-1}k\hat{\beta}_1$,

$$k_{B\min} = 5 - 2[k/ss] = 5 - 2\frac{i_0^{(D2)}/i_0^2}{(i_0^{(D1)}/i_0^{3/2})^2} = 5 - 2\frac{i_0 i_0^{(D2)}}{(i_0^{(D1)})^2} \tag{46}$$

(Ogasawara, 2013, Result 2; Ogasawara, 2014a, Erratum), where $[k/ss]$ stands for the ratio of the excess kurtosis to the squared skewness of the single sufficient statistic, and i_0 , $i_0^{(D1)}$ and $i_0^{(D2)}$ are scalar counterparts of \mathbf{I}_0 , $\mathbf{I}_0^{(D1)}$ and $\mathbf{I}_0^{(D2)}$, respectively.

Note that the minimizing coefficients in Results 5 and 6 given above generally depend on unknown population values, which are not available in practice and cannot automatically be replaced by their sample counterparts. The situation is similar to that of the maximizing coefficients in Sect. 3. This difficulty can be solved in several ways (see Ogasawara 2014b, pp. 203–204; Ogasawara 2016, Section 7). That is, (i) in some fortunate cases (e.g., Examples 1.1 to 2.3, 4.2, 4.4 and 4.5 shown later), the minimizing coefficients do not depend on unknown values. (ii) Some cases have lower/upper bounds for the minimizing coefficients (e.g., Example 3.1 shown later). (iii) In many cases, we have prior information on the range of unknown values. When the minimizing coefficients are monotonic functions of the population values, the range gives bounds similar to those of (ii). (iv) When another independent sample of size $O(n)$ is available as in cross validation, the estimated minimizing coefficients using the second sample can be used without changing the asymptotic results given in Results 5 and 6.

4.1 Gamma distribution

Example 1.1 The canonical parameter $(-1/\beta)$, when the shape parameter (α) is given, with its biased MLE in the gamma distribution.

The density of the gamma distribution using the scale parameter β , given the shape parameter α , is $f(x^* = x|\alpha, \beta) = x^{\alpha-1} \exp(-x/\beta)/\{\beta^\alpha \Gamma(\alpha)\}$ ($x > 0$), where the notations similar to $\hat{\alpha}$ and β_1 used earlier are used for their familiarity as long as confusion does not occur. The case of $\alpha = 1$ gives the exponential distribution. When α is a positive integer, we have the Erlangian distribution.

The canonical parameter is $\theta = -1/\beta$ (negative rate). Then, $f(x^* = x|\alpha, \theta) = x^{\alpha-1}(-\theta)^\alpha \exp(\theta x)/\Gamma(\alpha)$ ($x > 0$). From, e.g., Ogasawara (2014a, Example 3), $\hat{\theta}_{ML} = -\alpha/\bar{x}$, $\beta_1 = \theta_0/\alpha$, $i_0 = \text{var}(x^*) = \alpha/\theta_0^2$, $\beta_2 = i_0^{-1} = \theta_0^2/\alpha$ and $[k/ss] = 3/2$, where \bar{x} is the sample mean. Then,

1. $\hat{\theta}_{Smax}$ (see (36)): The notation $\hat{\theta}_{Smax}$ synonymous with $\hat{\theta}_{Sk_{Smax}}$ is used for simplicity with other simple ones defined similarly:

$$k_{Smax} = 1/(\theta_0^2 i_0) = 1/\alpha. \tag{47}$$

2. $\hat{\theta}_{Bmax}$ (see (34)): Use $\partial\beta_1/\partial\theta_0 = 1/\alpha$, then $k_{Bmax} = (\partial\beta_1/\partial\theta_0)/(\beta_1^2 i_0) = (1/\alpha)/(1/\alpha) = 1$. Since $\hat{\beta}_1 = \hat{\theta}_{ML}/\alpha$, we find that

$$\hat{\theta}_{Smax} = \hat{\theta}_{Bmax} = \left(1 - \frac{n^{-1}}{\alpha}\right) \hat{\theta}_{ML}. \tag{48}$$

3. $\hat{\theta}_{Smin}$ (see (41)): $k_{Smin} = \theta_0^{-2}(\beta_2 + \theta_0\beta_1) = \theta_0^{-2}(\theta_0^2\alpha^{-1} + \theta_0\theta_0\alpha^{-1}) = 2/\alpha$.
4. $\hat{\theta}_{Bmin}$ (see (41)): Since $\hat{\beta}_1 = \hat{\theta}_{ML}/\alpha$, we have $k_{Bmin} = \alpha k_{Smin} = 2$, which is also given by $k_{Bmin} = 5 - 2[k/ss] = 2$.

From 3 and 4,

$$\hat{\theta}_{Smin} = \hat{\theta}_{Bmin} = \left(1 - n^{-1} \frac{2}{\alpha}\right) \hat{\theta}_{ML}. \tag{49}$$

It is found that the amount of the correction in (49) is two times that of (48). The relatively larger correction in $\hat{\theta}_{Smin}$ and $\hat{\theta}_{Bmin}$ than in $\hat{\theta}_{Smax}$ and $\hat{\theta}_{Bmax}$ is a typical tendency. It is to be noted that the formulas of (48) and (49) do not depend on the unknown canonical parameter θ_0 .

Example 1.2 The scale parameter (β) , when the shape parameter (α) is given in the gamma distribution (a non-canonical parameter with its unbiased MLE).

Since $E_{\theta_0}(x^*) = \alpha\beta_0$ using the population scale parameter β_0 , the scale parameter is a scaled expectation parameter. So, only k_{Smax} and k_{Smin} are obtained. Preliminary results are

$$\begin{aligned} \frac{\partial \bar{l}}{\partial \beta_0} &= \frac{\bar{x}}{\beta_0^2} - \frac{\alpha}{\beta_0}, \quad \hat{\beta}_{\text{ML}} = \frac{\bar{x}}{\alpha}, \quad \frac{\partial^2 \bar{l}}{\partial \beta_0^2} = -2\frac{\bar{x}}{\beta_0^3} + \frac{\alpha}{\beta_0^2}, \quad i_0 = -E_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \beta_0^2} \right) = \frac{\alpha}{\beta_0^2}, \\ E_{\theta_0}(\hat{\beta}_{\text{ML}}) &= \beta_0, \quad \beta_2 = n\text{var}(\hat{\beta}_{\text{ML}}) = \beta_0^2/\alpha, \quad \beta_1 = 0, \\ nE_{\theta_0} \left[\frac{\partial \bar{l}}{\partial \beta_0} \left\{ \frac{\partial^2 \bar{l}}{\partial \beta_0^2} - E_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \beta_0^2} \right) \right\} \right] &= nE_{\theta_0} \left[\frac{\bar{x} - \alpha\beta_0}{\beta_0^2} \left\{ -\frac{2(\bar{x} - \alpha\beta_0)}{\beta_0^3} \right\} \right] \\ &= -\frac{2}{\beta_0^5} n\text{var}(\bar{x}) = -\frac{2}{\beta_0^5} \alpha\beta_0^2 = -\frac{2\alpha}{\beta_0^3}. \end{aligned} \quad (50)$$

1. $\hat{\theta}_{\text{Smax}}$ (see (34)):

$$k_{\text{Smax}} = \frac{1}{\beta_0^2(\alpha\beta_0^{-2})} \left\{ \beta_0 \left(-\frac{2\alpha}{\beta_0^3} \right) \frac{\beta_0^2}{\alpha} + 1 \right\} = -\frac{1}{\alpha} \quad (51)$$

and

$$\hat{\theta}_{\text{Smax}} = (1 - n^{-1}k_{\text{Smax}})\hat{\theta}_{\text{ML}} = \left(1 + \frac{n^{-1}}{\alpha} \right) \hat{\theta}_{\text{ML}} = \left(1 + \frac{n^{-1}}{\alpha} \right) \frac{\bar{x}}{\alpha} \quad (52)$$

2. $\hat{\theta}_{\text{Smin}}$ (see (41)):

$$k_{\text{Smin}} = \beta_2\theta_0^{-2} + \beta_1\theta_0^{-1} = \beta_2\theta_0^{-2} = 1/\alpha \quad (53)$$

$$\text{and } \hat{\theta}_{\text{Smin}} = \left(1 - \frac{n^{-1}}{\alpha} \right) \hat{\theta}_{\text{ML}} = \left(1 - \frac{n^{-1}}{\alpha} \right) \frac{\bar{x}}{\alpha}. \quad (54)$$

Note that (52) is an expected result since (52) should be asymptotically equal to the negative of the reciprocal of (48). The results are puzzling in that $\hat{\theta}_{\text{Smax}}$ in (52) is an inflated estimator while $\hat{\theta}_{\text{Smin}}$ in (54) is a shrinkage estimator. Since the amount of the added bias in (52) and (54) is the same with the reversed direction, (54) looks better than (52) since the variance of (54) is smaller than that of (52). Note, however, that this is based on the viewpoint of the Euclid distance. When the viewpoint of the Kullback–Leibler distance (see (4)) is employed, (52) is better than (54).

Example 1.3 The exponential distribution under arbitrary parametrization.

This case partially reduces to those in Examples 1.1 and 1.2 when $\alpha = 1$. That is, under canonical parametrization ($\theta = 1/\beta$) using the rate parameter θ ,

1.

$$\hat{\theta}_{S_{\max}} = \hat{\theta}_{B_{\max}} = (1 - n^{-1})\hat{\theta}_{ML} = (1 - n^{-1})/\bar{x}. \tag{55}$$

2.

$$\hat{\theta}_{S_{\min}} = \hat{\theta}_{B_{\min}} = (1 - 2n^{-1})\hat{\theta}_{ML} = (1 - 2n^{-1})/\bar{x}. \tag{56}$$

The pleasantly simple result of (55) was obtained by Takezawa (2014; see also 2015) as an exact one in that (55) maximizes the exact predictive expected log-likelihood among the family of $k_n \hat{\theta}_{ML}$ rather than the asymptotic expectation. Recall that the parametrization $\theta = -1/\beta$ was used in Example 1.1. When $\theta = 1/\beta$ is used, the reflected observable variable $-x^*$ is to be used, if necessary, without changing the essential results.

When the scale or expectation parameter β is used, from (52) and (54) with $\alpha = 1$, we have

3.

$$\hat{\theta}_{S_{\max}} = (1 + n^{-1})\hat{\theta}_{ML} = (1 + n^{-1})\bar{x}, \tag{57}$$

4.

$$\hat{\theta}_{S_{\min}} = (1 - n^{-1})\hat{\theta}_{ML} = (1 - n^{-1})\bar{x}. \tag{58}$$

Again simple results are obtained. Note that Takezawa (2014, 2015) gave only (55) among (55) to (58). However, (57) is not an exact one. For illustration, we derive the exact solution $k_{n_{\max}}$ maximizing the exact predictive expected log-likelihood among the family of $k_n \hat{\theta}_{ML} = k_n \hat{\beta}_{ML}$.

First, we have

$$\begin{aligned} & E_{\theta_0} \{ \bar{l}^* (k_n \hat{\theta}_{ML}) \} \\ &= \int_0^{+\infty} \left\{ \log \left(\frac{1}{k_n \hat{\theta}_{ML}} \right) - \frac{1}{k_n \hat{\theta}_{ML}} \beta \right\} \frac{(1/\beta)^n x^{n-1}}{(n-1)!} \exp \left(-\frac{x}{\beta} \right) dx. \tag{59} \\ &= \int_0^{+\infty} \left\{ \log \left(\frac{1}{k_n x / n} \right) - \frac{n}{k_n x} \beta \right\} \frac{\beta^{-n} x^{n-1}}{(n-1)!} \exp \left(-\frac{x}{\beta} \right) dx, \end{aligned}$$

where x is the variable corresponding to the sum of n observed variables, which follows the Erlangian distribution with the parameters n and β .

Differentiating (59) with respect to k_n and setting the result to zero, we obtain

$$\begin{aligned} & \int_0^{+\infty} \left(-\frac{1}{k_n} + \frac{n}{k_n^2 x} \beta \right) \frac{\beta^{-n} x^{n-1}}{(n-1)!} \exp\left(-\frac{x}{\beta}\right) dx. \\ &= -\frac{1}{k_n} + \frac{1}{k_n^2} \frac{n}{n-1} \int_0^{+\infty} \frac{\beta^{-(n-1)} x^{n-2}}{(n-2)!} \exp\left(-\frac{x}{\beta}\right) dx. = -\frac{1}{k_n} + \frac{1}{k_n^2} \frac{n}{n-1} = 0, \end{aligned} \quad (60)$$

giving

$$k_{n\max} = \frac{n}{n-1} = 1 + \frac{1}{n} + O(n^{-2}). \quad (61)$$

The above result obtained directly was given for illustration, which is also derived by using Takezawa's (2014) exact solution as:

$$\{(1 - n^{-1})\hat{\theta}_{\text{ML}}^{-1}\}^{-1} = \frac{n}{n-1} \hat{\theta}_{\text{ML}} = k_{n\max} \hat{\theta}_{\text{ML}}, \quad (62)$$

where $\hat{\theta}_{\text{ML}}^{-1}$ is the MLE of the canonical parameter. Equation (62) is based on the invariant property of the likelihood irrespective of parametrization.

When $\beta^* \equiv \log \beta$ (the log scale or the log inter-event time parameter) is used,

$$\begin{aligned} \hat{\theta}_{\text{Smax}} &= \log\{(1 - n^{-1})^{-1} \exp(\hat{\theta}_{\text{ML}})\} = \hat{\theta}_{\text{ML}} - \log(1 - n^{-1}) \\ &= \{1 - \hat{\theta}_{\text{ML}}^{-1} \log(1 - n^{-1})\} \hat{\theta}_{\text{ML}} = \hat{\theta}_{\text{ML}} + n^{-1} + O(n^{-2}). \end{aligned} \quad (63)$$

From (63), $k_{n\max} = 1 - \hat{\theta}_{\text{ML}}^{-1} \log(1 - n^{-1})$ is obtained, though the additive exact solution $\hat{\theta}_{\text{ML}} - \log(1 - n^{-1})$ and its asymptotic approximation $\hat{\theta}_{\text{ML}} + n^{-1}$ are simpler.

When we again compare the results in (55) and (57), (55) gives the unbiased estimator up to order $O(n^{-1})$ and simultaneously gives variance reduction. On the other hand, bias and variance are both inflated in (57). It is of interest to see that the correction factor $(1 - n^{-1})$ in (55) and (58) is the same though the parametrization is different.

4.2 Poisson distribution

Example 2.1 The canonical parameter ($\log \lambda$) in the Poisson distribution with its biased MLE.

Let λ be the source or the expectation parameter in the Poisson distribution. Then, $\theta = \log \lambda$ is the canonical parameter with its probability function $\Pr(x^* = x|\theta) = \exp(-e^\theta) e^{\theta x} / x! = \exp(\theta x - e^\theta) / x!$ ($x = 0, 1, 2, \dots$). Preliminary results are

$$\hat{\theta}_{\text{ML}} = \log \bar{x} \quad (\bar{x} \neq 0), \quad i_0 = \lambda_0 = e^{\theta_0}, \quad \beta_2 = i_0^{-1}, \quad \beta_1 = -e^{-\theta_0} / 2 \quad \text{and} \quad [k/\text{ss}] = 1 \quad (64)$$

(Ogasawara 2014a, Example 2).

1.

$$\hat{\theta}_{Smax} \text{ (see (36))}: k_{Smax} = \frac{1}{\theta_0^2 i_0} = \frac{1}{\theta_0^2 e^{\theta_0}}. \tag{65}$$

2. $\hat{\theta}_{Bmax}$ (see (34)): Use $\partial\beta_1/\partial\theta_0 = e^{-\theta_0}/2$, then,

$$k_{Bmax} = \frac{\partial\beta_1/\partial\theta_0}{\beta_1^2 i_0} = \frac{e^{-\theta_0}/2}{(e^{-\theta_0}/2)^2 e^{\theta_0}} = 2. \tag{66}$$

3. $\hat{\theta}_{Smin}$ (see (41)):

$$\begin{aligned} k_{Smin} &= \beta_2\theta_0^{-2} + \beta_1\theta_0^{-1} = e^{-\theta_0}\theta_0^{-2} - (e^{-\theta_0}/2)\theta_0^{-1} \\ &= e^{-\theta_0}\theta_0^{-1}\{\theta_0^{-1} - (1/2)\} = k_{Smax} - (e^{-\theta_0}/2)\theta_0^{-1} \quad (\theta_0 \neq 0). \end{aligned} \tag{67}$$

4.

$$\hat{\theta}_{Bmin} \text{ (see (46))}: k_{Bmin} = 5 - 2[k/ss] = 3. \tag{68}$$

Note that as in the gamma distribution, the amount of correction in (68) is larger than that of (66). However, (67) is smaller (larger) than (65) when $\theta_0 > 0$ ($\theta_0 < 0$). The value of (67) can be negative, zero and positive.

Example 2.2 The source or the expectation parameter (λ) in the Poisson distribution (a non-canonical parameter).

Basic results are

$$\begin{aligned} \Pr(x^* = x|\lambda) &= \exp(-\lambda)\lambda^x/x! \quad (x = 0, 1, 2, \dots), \\ \hat{\lambda}_{ML} &= \bar{x}, \quad \beta_1 = 0, \quad \beta_2 = \lambda_0, \quad i_0 = \lambda_0^{-1}, \\ nE_{\theta_0} \left[\frac{\partial \bar{l}}{\partial \lambda_0} \left\{ \frac{\partial^2 \bar{l}}{\partial \lambda_0^2} - E_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \lambda_0^2} \right) \right\} \right] &= nE_{\theta_0} \left\{ \left(\frac{\bar{x}}{\lambda_0} - 1 \right) \left(-\frac{\bar{x}}{\lambda_0^2} + \frac{1}{\lambda_0} \right) \right\} \\ &= -\frac{1}{\lambda_0^3} n\text{var}(\bar{x}) = -\frac{1}{\lambda_0^2}. \end{aligned} \tag{69}$$

1.

$$\hat{\theta}_{Smax} \text{ (see (34))}: k_{Smax} = \frac{1}{\lambda_0^2 \lambda_0^{-1}} \left\{ \lambda_0 \left(-\frac{1}{\lambda_0^2} \right) \lambda_0 + 1 \right\} = 0. \tag{70}$$

2.

$$\hat{\theta}_{Smin} \text{ (see (41))}: k_{Smin} = \beta_2\theta_0^{-2} + \beta_1\theta_0^{-1} = \beta_2\theta_0^{-2} = \lambda_0\lambda_0^{-2} = \lambda_0^{-1}. \tag{71}$$

Equation (70) indicates that the MLE is optimal. The amount of correction in $\hat{\theta}_{Smin}$ is larger than that of $\hat{\theta}_{Smax}$.

Example 2.3 The parameter of the mean inter-event time ($\beta = 1/\lambda$) (a non-canonical parameter with its biased MLE).

Preliminary results are

$$\begin{aligned} \Pr(x^* = x|\beta) &= \exp(-1/\beta)\beta^{-x}/x! \quad (x = 0, 1, 2, \dots), \\ \hat{\beta}_{ML} &= 1/\bar{x}, \quad \beta_1 = \beta_0^2, \quad \beta_2 = \beta_0^3, \quad i_0 = \beta_0^{-3}, \quad n \text{ acov}(\hat{\theta}_{ML}, \hat{\beta}_1) = 2\beta_0^4, \\ nE_{\theta_0} \left[\frac{\partial \bar{l}}{\partial \beta_0} \left\{ \frac{\partial^2 \bar{l}}{\partial \beta_0^2} - E_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \beta_0^2} \right) \right\} \right] &= nE_{\theta_0} \left\{ \left(\frac{1}{\beta_0^2} - \frac{\bar{x}}{\beta_0} \right) \left(\frac{\bar{x}}{\beta_0^2} - \frac{1}{\beta_0^3} \right) \right\} \\ &= -\frac{1}{\beta_0^3} n\text{var}(\bar{x}) = -\frac{1}{\beta_0^4} \end{aligned} \tag{72}$$

(see Ogasawara 2015).

1.

$$\hat{\theta}_{Smax} \text{ (see (34))}: k_{Smax} = \frac{1}{\beta_0^2 \beta_0^{-3}} = \left\{ \beta_0 \left(-\frac{1}{\beta_0^4} \right) \beta_0^3 + 1 \right\} = 0. \tag{73}$$

2. $\hat{\theta}_{Bmax}$ (see (34)): Since $\partial \beta_1 / \partial \theta_0 = 2\beta_0$, we have

$$k_{Bmax} = \frac{1}{\beta_0^4 \beta_0^{-3}} = \left\{ \beta_0^2 \left(-\frac{1}{\beta_0^4} \right) \beta_0^3 + 2\beta_0 \right\} = 1. \tag{74}$$

3.

$$\hat{\theta}_{Smin} \text{ (see (41))}: k_{Smin} = \beta_2 \theta_0^{-2} + \beta_1 \theta_0^{-1} = \beta_0^3 \beta_0^{-2} + \beta_0^2 \beta_0^{-1} = 2\beta_0. \tag{75}$$

4. $\hat{\theta}_{Bmin}$ (see (41)): Using $\alpha_0 = \beta_1 = \beta_0^2$ under correct model specification,

$$k_{Bmin} = \frac{1}{\beta_1^2} \{ n \text{ acov}_{\theta_0}(\hat{\theta}_{ML}, \hat{\beta}_1) + \beta_1^2 \} = 2 + 1 = 3. \tag{76}$$

Note that again $k_{Smax} = 0$ as is expected.

4.3 Bernoulli distribution

Example 3.1 The canonical parameter (logit) with its biased MLE.

Basic results are

$$\begin{aligned}
 \Pr(x^* = x|\eta) &= \left\{ \frac{1}{1 + \exp(-\eta)} \right\}^x \left\{ \frac{\exp(-\eta)}{1 + \exp(-\eta)} \right\}^{1-x} \quad (x = 0, 1), \\
 \pi_0 &\equiv \frac{1}{1 + \exp(-\eta_0)}, \quad \eta_0 = \log \frac{\pi_0}{1 - \pi_0}, \quad \hat{\pi}_{ML} = \bar{x}, \\
 \hat{\eta}_{ML} &= \log \{ \bar{x} / (1 - \bar{x}) \}, \quad i_0 = \pi_0(1 - \pi_0), \quad \beta_2 = \{ \pi_0(1 - \pi_0) \}^{-1}, \\
 \beta_1 &= -\frac{i_0^{-2} \kappa_3(x^*)}{2} = -\frac{1 - 2\pi_0}{2\pi_0(1 - \pi_0)}, \quad \text{sk}(x^*) = \frac{1 - 2\pi_0}{\{ \pi_0(1 - \pi_0) \}^{1/2}}, \\
 \text{kt}(x^*) &= \frac{1 - 6\pi_0 + 6\pi_0^2}{\pi_0(1 - \pi_0)}, \quad [\text{k/SS}] = \frac{1 - 6\pi_0 + 6\pi_0^2}{(1 - 2\pi_0)^2} = 1 - \frac{2\pi_0(1 - \pi_0)}{(1 - 2\pi_0)^2} < 1 \\
 &(\bar{x} \neq 0, 1; \pi_0 \neq 0, 0.5, 1)
 \end{aligned}
 \tag{77}$$

(Ogasawara 2014a, Example 1), where $\kappa_j(\cdot)$ indicates the j -th cumulant of a variable; $\text{sk}(\cdot)$ is the skewness of a variable; and $\text{kt}(\cdot)$ is the excess kurtosis of a variable.

1.

$$\hat{\theta}_{Smax} \text{ (see (36)): } k_{Smax} = \frac{1}{\eta_0^2 \pi_0 (1 - \pi_0)}. \tag{78}$$

2. $\hat{\theta}_{Bmax}$ (see (34)): Since

$$\begin{aligned}
 \frac{\partial \beta_1}{\partial \theta_0} &= -\frac{1}{2} \frac{\partial}{\partial \eta_0} \frac{1 - 2\pi_0}{\pi_0(1 - \pi_0)} = \frac{\pi_0(1 - \pi_0)}{\pi_0(1 - \pi_0)} + \frac{(1 - 2\pi_0)^2 \pi_0(1 - \pi_0)}{2\{ \pi_0(1 - \pi_0) \}^2} \\
 &= 1 + \frac{(1 - 2\pi_0)^2}{2\pi_0(1 - \pi_0)} > 1, \\
 k_{Bmax} &= \frac{\partial \beta_1 / \partial \theta_0}{\beta_1^2 i_0} = \left\{ 1 + \frac{(1 - 2\pi_0)^2}{2\pi_0(1 - \pi_0)} \right\} / \left[\left\{ \frac{1 - 2\pi_0}{2\pi_0(1 - \pi_0)} \right\}^2 \pi_0(1 - \pi_0) \right] \\
 &= \frac{2\{ 2\pi_0(1 - \pi_0) + (1 - 2\pi_0)^2 \}}{(1 - 2\pi_0)^2} = 2 + \frac{4\pi_0(1 - \pi_0)}{(1 - 2\pi_0)^2} > 2 \quad (\pi_0 \neq 0, 0.5, 1).
 \end{aligned}
 \tag{79}$$

3. $\hat{\theta}_{Smin}$ (see (41)):

$$\begin{aligned}
 k_{Smin} &= \beta_2 \theta_0^{-2} + \beta_1 \theta_0^{-1} = \{ \pi_0(1 - \pi_0) \}^{-1} \eta_0^{-2} - \frac{1 - 2\pi_0}{2\pi_0(1 - \pi_0)} \eta_0^{-1} \\
 &= \frac{1}{\pi_0(1 - \pi_0) \eta_0^2} \left\{ 1 - \frac{1}{2} (1 - 2\pi_0) \eta_0 \right\} > \frac{1}{\pi_0(1 - \pi_0) \eta_0^2} = k_{Smax} \\
 &(\pi_0 \neq 0, 0.5, 1; (1 - 2\pi_0) \eta_0 < 0).
 \end{aligned}
 \tag{80}$$

4. $\hat{\theta}_{\text{Bmin}}$ (see (46)):

$$k_{\text{Bmin}} = 5 - 2[k/ss] = 5 - 2 \left\{ 1 - \frac{2\pi_0(1 - \pi_0)}{(1 - 2\pi_0)^2} \right\} = 3 + \frac{4\pi_0(1 - \pi_0)}{(1 - 2\pi_0)^2} > 3 \quad (81)$$

(Ogasawara 2014a, 2015, Example 2).

In the above results, the relative sizes of corrections by optimal coefficients are found to be similar to those in the Poisson distribution. All the optimal coefficients depend on the unknown parameter. The lower bound 3 for k_{Bmin} is a known one while the lower bound 2 for k_{Bmax} is a new one.

Example 3.2 The source or the expectation parameter (π) (a non-canonical parameter).

Basic results are

$$\begin{aligned} \Pr(x^* = x|\pi) &= \pi^x(1 - \pi)^{1-x} \quad (x = 0, 1), \quad \hat{\pi}_{\text{ML}} = \bar{x}, \quad \beta_1 = 0, \\ i_0 &= \{\pi_0(1 - \pi_0)\}^{-1}, \quad \beta_2 = \pi_0(1 - \pi_0), \\ n\mathbf{E}_{\theta_0} \left[\frac{\partial \bar{l}}{\partial \pi_0} \left\{ \frac{\partial^2 \bar{l}}{\partial \pi_0^2} - \mathbf{E}_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \pi_0^2} \right) \right\} \right] \\ &= n\mathbf{E}_{\theta_0} \left\{ \left(\frac{\bar{x}}{\pi_0} - \frac{1 - \bar{x}}{1 - \pi_0} \right) \left(-\frac{\bar{x}}{\pi_0^2} - \frac{1 - \bar{x}}{(1 - \pi_0)^2} + \frac{1}{\pi_0(1 - \pi_0)} \right) \right\} \quad (82) \\ &= \frac{1}{\pi_0(1 - \pi_0)\pi_0^2(1 - \pi_0)^2} n\mathbf{E}_{\theta_0}[(\bar{x} - \pi_0)\{\pi_0^2 - (1 - \pi_0)^2\}(\bar{x} - \pi_0)] \\ &= -\frac{1 - 2\pi_0}{\pi_0^2(1 - \pi_0)^2}. \end{aligned}$$

1. $\hat{\theta}_{\text{Smax}}$ (see (34)):

$$\begin{aligned} k_{\text{Smax}} &= \frac{1}{\pi_0^2\{\pi_0(1 - \pi_0)\}^{-1}} \left[\pi_0 \left\{ -\frac{1 - 2\pi_0}{\pi_0^2(1 - \pi_0)^2} \right\} \pi_0(1 - \pi_0) + 1 \right] \\ &= \frac{1 - \pi_0}{\pi_0} \left(-\frac{1 - 2\pi_0}{1 - \pi_0} + 1 \right) = 1. \quad (83) \end{aligned}$$

2. $\hat{\theta}_{\text{Smin}}$ (see (41)):

$$k_{Smin} = \beta_2\theta_0^{-2} + \beta_1\theta_0^{-1} = \beta_2\theta_0^{-2} = \pi_0(1 - \pi_0)\pi_0^{-2} = \frac{1 - \pi_0}{\pi_0} > 0 \quad (\pi_0 \neq 0, 1). \tag{84}$$

We find that k_{Smax} does not depend on the unknown parameter. The relative size of k_{Smax} and k_{Smin} depends on the unknown parameter, which is given as follows:

$$\begin{array}{l} \pi_0: \quad (0, 1/2) \quad \quad \quad 1/2 \quad \quad \quad (1/2, 1) \\ \text{Property: } \quad k_{Smin} > k_{Smax} \quad k_{Smin} = k_{Smax} \quad k_{Smin} < k_{Smax} \end{array} \tag{85}$$

4.4 The univariate normal distribution

For convenience, the variance parameter σ^2 is also denoted by ϕ . Preliminary results are

$$f(x^* = x|\mu, \sigma^2(= \phi)) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\phi^{1/2}}} \exp\left\{-\frac{(x - \mu)^2}{2\phi}\right\},$$

$$\theta_0 = (\mu_0, \phi_0)', \quad \frac{\partial \bar{l}}{\partial \theta_0} = \left\{ \frac{\bar{x} - \mu_0}{\phi_0}, \quad -\frac{1}{2\phi_0} + \frac{n^{-1} \sum_{j=1}^n (x_j - \mu_0)^2}{2\phi_0^2} \right\}',$$

$$\hat{\theta}_{ML} = \left\{ \bar{x}, \quad n^{-1} \sum_{j=1}^n (x_j - \bar{x})^2 \right\}',$$

$$\frac{\partial^2 \bar{l}}{\partial \theta_0 \partial \theta_0'} = \begin{bmatrix} -\frac{1}{\phi_0} & -\frac{\bar{x} - \mu_0}{\phi_0^2} \\ -\frac{\bar{x} - \mu_0}{\phi_0^2} & \frac{1}{2\phi_0^2} - \frac{n^{-1} \sum_{j=1}^n (x_j - \mu_0)^2}{\phi_0^3} \end{bmatrix},$$

$$-E_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \theta_0 \partial \theta_0'} \right) = \begin{pmatrix} \frac{1}{\phi_0} & 0 \\ 0 & \frac{1}{2\phi_0^2} \end{pmatrix} = \mathbf{I}_0, \quad n \text{ acov}(\hat{\theta}_{ML}) = \mathbf{I}_0^{-1}, \quad \beta_1 = (0, -\phi_0)',$$

$$nE_{\theta_0} \left(\frac{\partial \bar{l}}{\partial \mu_0} \mathbf{M} \right) = \begin{pmatrix} 0 & -\frac{1}{\phi_0^2} \\ -\frac{1}{\phi_0^2} & 0 \end{pmatrix}, \quad nE_{\theta_0} \left(\frac{\partial \bar{l}}{\partial \phi_0} \mathbf{M} \right) = \begin{pmatrix} 0 & 0 \\ 0 & -\frac{1}{\phi_0^3} \end{pmatrix},$$

(86)

where

$$\begin{aligned}
 & nE_{\theta_0} \left[\frac{\partial \bar{l}}{\partial \phi_0} \left\{ \frac{\partial^2 \bar{l}}{\partial \phi_0^2} - E_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \phi_0^2} \right) \right\} \right] \\
 &= -\frac{n}{2\phi_0^5} E_{\theta_0} \left[\left\{ n^{-1} \sum_{j=1}^n (x_j - \mu_0)^2 - \phi_0 \right\}^2 \right] = -\frac{1}{2\phi_0^5} (3\phi_0^2 - \phi_0^2) = -\frac{1}{\phi_0^3}
 \end{aligned} \tag{87}$$

is used.

Example 4.1 The parameters μ and $\phi (= \sigma^2)$ with unconstrained \mathbf{k} (a non-canonical parameter vector with its partially biased MLE).

Since $\beta_1 = (0, -\phi_0)'$, $\hat{\theta}_{Bmax}$ and $\hat{\theta}_{Bmin}$ are not available.

1. $\hat{\theta}_{Smax}$ (see (31)):

$$\begin{aligned}
 \mathbf{k}_{Smax} &= \text{diag}^{-1}(\mu_0, \phi_0) \begin{pmatrix} \phi_0 & 0 \\ 0 & 2\phi_0^2 \end{pmatrix} \\
 &\times \left\{ \begin{pmatrix} 0 & -1/\phi_0^2 & 0 & 0 \\ -1/\phi_0^2 & 0 & 0 & -1/\phi_0^3 \end{pmatrix} (\phi_0, 0, 0, 2\phi_0^2)' + \text{diag}^{-1}(\mu_0, \phi_0) \mathbf{1}_{(2)} \right\} \\
 &= \begin{pmatrix} \phi_0/\mu_0 & 0 \\ 0 & 2\phi_0 \end{pmatrix} \begin{pmatrix} 1/\mu_0 \\ -2/\phi_0 \end{pmatrix} = \begin{pmatrix} \phi_0/\mu_0^2 \\ -4 \end{pmatrix} = \begin{pmatrix} c_v^2 \\ -4 \end{pmatrix} \quad (\mu_0 \neq 0),
 \end{aligned} \tag{88}$$

where c_v is the coefficient of variation assuming $\mu_0 \neq 0$. It is found that the second element -4 of \mathbf{k}_{Smax} gives an estimator asymptotically equal to the estimator called “the third variance” by Takezawa (2012, Eq. (21)) using a different method though with a common predictive viewpoint. Note that the adjusted estimator for a variance in (88) and Takezawa’ third variance are inflated estimators over the usual unbiased one and the normal-theory MLE.

2. $\hat{\theta}_{Smin}$ (see (39)):

$$\begin{aligned}
 \mathbf{k}_{Smin} &= \text{Diag}^{-1} \left\{ \begin{pmatrix} \mu_0 \\ \phi_0 \end{pmatrix} (\mu_0, \phi_0) \right\} \left\{ \begin{pmatrix} n\text{var}(\hat{\mu}_{ML}) & 0 \\ 0 & n\text{var}(\hat{\phi}_{ML}) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & -\phi_0^2 \end{pmatrix} \right\} \mathbf{1}_{(2)} \\
 &= \begin{pmatrix} 1/\mu_0^2 & 0 \\ 0 & 1/\phi_0^2 \end{pmatrix} \begin{pmatrix} \phi_0 & 0 \\ 0 & 2\phi_0^2 - \phi_0^2 \end{pmatrix} \mathbf{1}_{(2)} = \begin{pmatrix} \phi_0/\mu_0^2 \\ 1 \end{pmatrix} = \begin{pmatrix} c_v^2 \\ 1 \end{pmatrix}.
 \end{aligned} \tag{89}$$

It is of interest to see that both the elements of \mathbf{k}_{Smax} and \mathbf{k}_{Smin} for μ_0 are the same and depend on only c_v . Since c_v^2 in (88) and (89) is positive under the

conditions $\mu_0 \neq 0$ and $\phi_0 = \sigma_0^2 > 0$, the adjusted estimator $(1 - n^{-1}c_V^2)\bar{x}$ is a shrinkage estimator when $1 - n^{-1}c_V^2 \geq 0$. If μ_0 is close to zero, c_V^2 becomes very large. While this may seem odd, the result is reasonable, which can be explained as follows. When μ_0 is close to zero, the large c_V^2 gives small $1 - n^{-1}c_V^2$ as long as $1 - n^{-1}c_V^2 \geq 0$. Consequently, this gives a smaller adjusted estimator $(1 - n^{-1}c_V^2)\bar{x}$, which is closer to the small population value than the usual estimator \bar{x} . Since $1 - n^{-1}c_V^2$ is an asymptotic result, $1 - n^{-1}c_V^2$ can become negative with finite n . In this case, $1 - n^{-1}c_V^2$ may be replaced by zero.

When \mathbf{k} for $\hat{\theta}_{\text{Bmax}}$ and $\hat{\theta}_{\text{Bmin}}$ is shortened as a scalar only for ϕ_0 , the solutions become similar to those in Example 4.2 shown next.

Example 4.2 The parameters μ and $\phi (= \sigma^2)$ with $\mathbf{k} = k\mathbf{1}_{(2)}$ (a non-canonical parameter vector with its partially biased MLE).

This is given for illustration using $\mathbf{k} = k\mathbf{1}_{(2)}$. Note that generally the equal value k for μ and $\phi = \sigma^2$ is meaningless. However, when c_V is known to be proportional to $1/\sigma$ with $\mu > 0$, the equal k can be employed.

1. $\hat{\theta}_{\text{Smax}}$ (see (34)):

$$\begin{aligned}
 k_{\text{Smax}} &= \frac{1}{(\mu_0, \phi_0) \begin{pmatrix} 1/\phi_0 & 0 \\ 0 & 1/(2\phi_0^2) \end{pmatrix} \begin{pmatrix} \mu_0 \\ \phi_0 \end{pmatrix}} \\
 &\times \left\{ (\mu_0, \phi_0) \begin{pmatrix} 0 & -1/\phi_0^2 & 0 & 0 \\ -1/\phi_0^2 & 0 & 0 & -1/\phi_0^3 \end{pmatrix} (\phi_0, 0, 0, 2\phi_0^2)' + 2 \right\} \\
 &= \left(\frac{\mu_0^2}{\phi_0} + \frac{1}{2} \right)^{-1} \left\{ \phi_0 \left(-\frac{1}{\phi_0} - \frac{2}{\phi_0} \right) + 2 \right\} = -\frac{2\phi_0}{2\mu_0^2 + \phi_0} = -\frac{2c_V^2}{2 + c_V^2} < 0.
 \end{aligned}
 \tag{90}$$

2. $\hat{\theta}_{\text{Bmax}}$ (see (34)): Using $\frac{\partial \beta_1'}{\partial \theta_0} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$,

$$\begin{aligned}
 k_{\text{Bmax}} &= \frac{1}{(0, -\phi_0) \begin{pmatrix} 1/\phi_0 & 0 \\ 0 & 1/(2\phi_0^2) \end{pmatrix} \begin{pmatrix} 0 \\ -\phi_0 \end{pmatrix}} \\
 &\times \left\{ (0, -\phi_0) \begin{pmatrix} 0 & -1/\phi_0^2 & 0 & 0 \\ -1/\phi_0^2 & 0 & 0 & -1/\phi_0^3 \end{pmatrix} (\phi_0, 0, 0, 2\phi_0^2)' - 1 \right\} \quad (91) \\
 &= 2 \left\{ \phi_0 \left(\frac{1}{\phi_0} + \frac{2}{\phi_0} \right) - 1 \right\} = 4.
 \end{aligned}$$

3. $\hat{\boldsymbol{\theta}}_{\text{Smin}}$ (see (41)):

$$\begin{aligned}
 k_{\text{Smin}} &= \frac{1}{\boldsymbol{\theta}'_0 \boldsymbol{\theta}_0} (\mathbf{1}'_{(2)} \boldsymbol{\beta}_2 + \boldsymbol{\theta}'_0 \boldsymbol{\beta}_1) \\
 &= \frac{1}{(\mu_0, \phi_0)(\mu_0, \phi_0)'} \{ \mathbf{1}'_{(2)} (\phi_0, 2\phi_0^2)' + (\mu_0, \phi_0)(0, -\phi_0)' \} = \frac{\phi_0(1 + \phi_0)}{\mu_0^2 + \phi_0^2} > 0. \quad (92)
 \end{aligned}$$

4. $\hat{\boldsymbol{\theta}}_{\text{Bmin}}$ (see (41)): Using $\boldsymbol{\beta}_1 = (0, -\phi_0)'$ and

$$\begin{aligned}
 n \operatorname{acov}_{\theta_0}(\hat{\boldsymbol{\theta}}_{\text{ML}}, \hat{\boldsymbol{\beta}}'_1) &= \begin{pmatrix} 0 & 0 \\ 0 & -n \operatorname{avar}(\hat{\phi}_{\text{ML}}) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & -2\phi_0^2 \end{pmatrix}, \quad (93) \\
 k_{\text{Bmin}} &= \frac{1}{\boldsymbol{\beta}'_1 \boldsymbol{\beta}_1} (-2\phi_0^2 + \boldsymbol{\beta}'_1 \boldsymbol{\beta}_1) = \frac{1}{\phi_0^2} (-2\phi_0^2) + 1 = -1.
 \end{aligned}$$

Note that $-\{n/(n+1)\}k_{\text{Bmin}} = n/(n+1)$ is the exact solution as explained in Sect. 1.

Example 4.3 The mean parameter μ when the variance $\phi (= \sigma^2)$ is known (the canonical and expectation parameter).

We have

$$\begin{aligned}
 \theta_0 &= \mu_0, \quad \frac{\partial \bar{l}}{\partial \theta_0} = \frac{\bar{x} - \mu_0}{\phi_0}, \quad \hat{\theta}_{\text{ML}} = \bar{x}, \quad \frac{\partial^2 \bar{l}}{\partial \theta_0^2} = -\frac{1}{\phi_0}, \quad i_0 = \frac{1}{\phi_0}, \quad (94) \\
 n \operatorname{var}(\hat{\theta}_{\text{ML}}) &= i_0^{-1} = \phi_0, \quad \boldsymbol{\beta}_1 = 0.
 \end{aligned}$$

Then,

1.

$$\hat{\boldsymbol{\theta}}_{\text{Smax}} \text{ (see (36)): } k_{\text{Smax}} = \frac{1}{\mu_0^2 \phi_0^{-1}} = \frac{\phi_0}{\mu_0^2} = c_v^2. \quad (95)$$

2.

$$\hat{\theta}_{Smin} \text{ (see (41))}: k_{Smin} = \frac{1}{\mu_0^2} \{n \text{ var}(\hat{\theta}_{ML}) + \mu_0 \times 0\} = \frac{\phi_0}{\mu_0^2} = c_v^2. \tag{96}$$

We see that the optimal coefficients are the same. Note that the results hold even under model misspecification, since $n \text{ var}(\bar{x}) = \phi_0$ holds under arbitrary distributions when it exists with $\Gamma = -\Lambda = \phi_0^{-1}$.

Example 4.4 The variance parameter $\phi (= \sigma^2)$ when the mean μ is known (a non-canonical parameter with its unbiased MLE).

Basic results are

$$\begin{aligned} \frac{\partial \bar{l}}{\partial \theta_0} &= -\frac{1}{2\phi_0} + \frac{n^{-1} \sum_{j=1}^n (x_j - \mu_0)^2}{2\phi_0^2}, \quad \hat{\phi}_{ML} = n^{-1} \sum_{j=1}^n (x_j - \mu_0)^2, \\ \beta_1 &= 0, \quad n \text{ var}(\hat{\theta}_{ML}) = 2\phi_0^2, \quad \frac{\partial^2 \bar{l}}{\partial \theta_0^2} = \frac{1}{2\phi_0^2} - \frac{n^{-1} \sum_{j=1}^n (x_j - \mu_0)^2}{\phi_0^3}, \\ E_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \theta_0^2} \right) &= -\frac{1}{2\phi_0^2}, \quad i_0 = \frac{1}{2\phi_0^2}, \quad E_{\theta_0} \left[\frac{\partial \bar{l}}{\partial \theta_0} \left\{ \frac{\partial^2 \bar{l}}{\partial \theta_0^2} - E_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \theta_0^2} \right) \right\} \right] = -\frac{1}{\phi_0^3}, \end{aligned} \tag{97}$$

where the last equation is the same as (87). Then,

1.

$$\hat{\theta}_{Smax} \text{ (see (34))}: k_{Smax} = \frac{1}{\phi_0^2(2\phi_0^2)^{-1}} \{ \phi_0(-\phi_0^3)^{-1} 2\phi_0^2 + 1 \} = -2. \tag{98}$$

2.

$$\hat{\theta}_{Smin} \text{ (see (41))}: k_{Smin} = \frac{1}{\phi_0^2} \{n \text{ avar}(\hat{\theta}_{ML}) + \phi_0 \times 0\} = 2. \tag{99}$$

Example 4.5 The precision parameter $\theta = 1/\sigma^2 = 1/\phi$, when the mean μ is known (the canonical parameter with its biased MLE; Ogasawara 2013, Example 5).

The vector of the canonical parameters, when μ and ϕ are unknown, is $\{\mu/\phi, -1/(2\phi)\}'$ with the vector of sufficient statistics $(x^*, x^{*2})'$. Note, however, that the use of the parameter μ/ϕ may be limited. Let $\theta = 1/\phi$. Then, θ is the precision parameter (the reciprocal of variance). When μ_0 is known, define $\theta(\mu_0, -0.5)' \equiv \theta \mathbf{d}$, where θ is a single parameter with the sufficient statistic

$$s_d \equiv (x^*, x^{*2}) \mathbf{d} = (x^*, x^{*2})(\mu_0, -0.5)', \tag{100}$$

whose mean over observations is $-\frac{1}{2} n^{-1} \sum_{j=1}^n (x_j^* - \mu_0)^2 + \frac{\mu_0^2}{2}$.

Basic results are

$$\begin{aligned} E_{\theta_0}(s_d) &= \frac{1}{2}(\mu_0^2 - \phi_0), \quad i_0 = \text{var}(s_d) = \phi_0^2/2, \quad \kappa_3(s_d) = -\phi_0^3, \quad \kappa_4(s_d) = 3\phi_0^4, \\ \text{sk}(s_d) &= -\sqrt{8}, \quad \text{kt}(s_d) = 12, \quad \beta_1 = 2/\phi_0 = 2\theta_0, \quad [\text{k/ss}] = 3/2, \\ \beta_2 &= i_0^{-1} = 2\phi_0^{-2} = 2\theta_0^2, \quad \hat{\theta}_{\text{ML}} = 1/\hat{\phi}_{\text{ML}} \end{aligned} \quad (101)$$

(Ogasawara 2014a, Example 5). Then,

1.

$$\hat{\theta}_{\text{Smax}} \text{ (see (36))}: k_{\text{Smax}} = \frac{1}{\theta_0^2(\phi_0^2/2)} = 2. \quad (102)$$

2. $\hat{\theta}_{\text{Bmax}}$ (see (34)): Using $\beta_1 = 2\theta_0$,

$$k_{\text{Bmax}} = \frac{2}{(2\theta_0)^2(\phi_0^2/2)} = 1. \quad (103)$$

3.

$$\hat{\theta}_{\text{Smin}} \text{ (see (41))}: k_{\text{Smin}} = \beta_2\theta_0^{-2} + \beta_1\theta_0^{-1} = 2\theta_0^2\theta_0^{-2} + 2\theta_0\theta_0^{-1} = 4. \quad (104)$$

4.

$$\hat{\theta}_{\text{Bmin}} \text{ (see (46))}: k_{\text{Bmin}} = 5 - 2[\text{k/ss}] = 2. \quad (105)$$

The result of (102) is an expected one from (98) in Example 4.4. Note that the distribution of x^* is the same in Examples 4.4 and 4.5 with different parametrization, i.e., $\theta = \phi$ or $\theta = 1/\phi$. Then, from (98), $\hat{\theta}_{\text{Smax}}$ in Example 4.5 is asymptotically equal to the reciprocal of $\hat{\theta}_{\text{Smax}}$ in Example 4.4, which gives, for Example 4.5, $\{(1 + 2n^{-1})\hat{\phi}_{\text{ML}}\}^{-1} = (1 - 2n^{-1})\hat{\theta}_{\text{ML}} + O_p(n^{-2})$ yielding (102).

4.5 Misspecified models

Example 5.1 The misspecified exponential distribution when the gamma distribution with $\alpha \neq 1$ is true.

This example is the case when one of the parameters in a distribution is fixed at an incorrect value. In this example, the density

$$f(x^* = x|\lambda) = \lambda \exp(-\lambda x) \quad (x > 0) \quad (106)$$

is used irrespective of the situation where the true density is

$$f_T(x^* = x|\lambda_1, \alpha) = x^{\alpha-1} \lambda_1^\alpha \exp(-\lambda_1 x) / \Gamma(\alpha) \quad (x > 0, \alpha \neq 1). \quad (107)$$

Then,

$$\begin{aligned}
 \theta_0 &= \lambda_0 = \lambda_1/\alpha, \quad \hat{\theta}_{ML} = 1/\bar{x}, \\
 E_{\theta_0}(\bar{x}) &= 1/\lambda_0, \quad n \text{ var}_{\theta_0}(\bar{x}) = 1/\lambda_0^2, \\
 E_{\Gamma}(\bar{x}) &= \alpha/\lambda_1 = 1/\lambda_0, \quad n \text{ var}_{\Gamma}(\bar{x}) = \alpha/\lambda_1^2, \\
 n \text{ avar}_{\theta_0}(\hat{\theta}_{ML}) &= \left(\frac{\partial \hat{\theta}_{ML}}{\partial \bar{x}} \Big|_{\bar{x}=1/\lambda_0} \right)^2 n \text{ var}_{\theta_0}(\bar{x}) = \lambda_0^4 \lambda_0^{-2} = \lambda_0^2, \\
 n \text{ avar}_{\Gamma}(\hat{\theta}_{ML}) &= \left(\frac{\partial \hat{\theta}_{ML}}{\partial \bar{x}} \Big|_{\bar{x}=1/\lambda_0} \right)^2 n \text{ var}_{\Gamma}(\bar{x}) = \lambda_0^4 (\alpha/\lambda_1^2) = \frac{\lambda_1^2}{\alpha^3},
 \end{aligned} \tag{108}$$

$$\begin{aligned}
 \frac{\partial \bar{l}}{\partial \theta_0} &= \frac{1}{\lambda_0} - \bar{x}, \quad \frac{\partial^2 \bar{l}}{\partial \theta_0^2} = -\frac{1}{\lambda_0^2} = -i_0 = O(1), \\
 \Gamma = \gamma &= n E_{\Gamma} \left\{ \left(\frac{\partial \bar{l}}{\partial \theta_0} \right)^2 \right\} = E_{\Gamma} \left\{ \left(\frac{1}{\lambda_0} - \bar{x} \right)^2 \right\} = \frac{\alpha}{\lambda_1^2}, \\
 \beta_1 &= \frac{1}{2} \frac{\partial^2 \hat{\theta}_{ML}}{\partial \bar{x}^2} \Big|_{\bar{x}=1/\lambda_0} n \text{ var}_{\Gamma}(\bar{x}) = \lambda_0^3 (\alpha/\lambda_1^2) = \lambda_1/\alpha^2.
 \end{aligned}$$

The dispersion ratio defined earlier is

$$\frac{n \text{ avar}_{\Gamma}(\hat{\theta}_{ML})}{n \text{ avar}_{\theta_0}(\hat{\theta}_{ML})} = \frac{1}{\alpha}. \tag{109}$$

Since $\beta_1 = \lambda_1/\alpha^2$ depends on α and λ_1 , only k_{Smax} and k_{Smin} are considered.

1. $\hat{\theta}_{Smax}$ (see (26)):

$$k_{Smax} = \frac{\text{tr}(\Gamma \Lambda^{-1})}{\theta_0' \Lambda \theta_0} = \frac{(\alpha/\lambda_1^2)(-i_0)^{-1}}{\theta_0^2(-i_0)} = \frac{1}{\alpha} \neq 1. \tag{110}$$

2. $\hat{\theta}_{Smin}$ (see (41)):

$$k_{Smin} = \frac{1}{\theta_0^2} \{ n \text{ avar}_{\Gamma}(\hat{\theta}_{ML}) + \theta_0^2 \alpha^{-1} \} = \frac{1}{\lambda_0^2} \times \frac{\alpha \lambda_0^4}{\lambda_1^2} + \alpha^{-1} = \frac{2}{\alpha} \neq 2. \tag{111}$$

It is found that that k_{Smax} is equal to the dispersion ratio, which is not equal to the unit value given by the true model. Similarly, k_{Smin} in (111) is two times the dispersion ratio, which is not equal to 2 given by the true model (see (56)).

Example 5.2 The misspecified Poisson distribution when the Bernoulli distribution is true.

This example is the case when a distribution is used under the situation where the true distribution is a different type of distribution. In this example, the source or expectation parameter in the Poisson distribution is used:

$$\Pr(x^* = x|\lambda) = \exp(-\lambda)\lambda^x/x! \quad (x = 0, 1, 2, \dots), \quad (112)$$

when the true distribution is

$$\Pr_{\Gamma}(x^* = x|\pi) = \pi^x(1 - \pi)^{1-x} \quad (x = 0, 1). \quad (113)$$

Basic results are

$$\begin{aligned} \theta_0 &= \lambda_0 = \pi_0, \quad \hat{\theta}_{\text{ML}} = \bar{x}, \quad \beta_1 = 0, \quad n \text{var}_{\theta_0}(\hat{\theta}_{\text{ML}}) = n \text{var}_{\theta_0}(\bar{x}) = \pi_0, \\ \beta_2 &= n \text{var}_{\Gamma}(\hat{\theta}_{\text{ML}}) = n \text{var}_{\Gamma}(\bar{x}) = \pi_0(1 - \pi_0) \neq \pi_0 \quad (\pi_0 \neq 0), \\ \frac{\partial \bar{l}}{\partial \theta_0} &= -1 + \frac{\bar{x}}{\pi_0}, \quad \frac{\partial^2 \bar{l}}{\partial \theta_0^2} = -\frac{\bar{x}}{\pi_0^2}, \end{aligned} \quad (114)$$

$$E_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \theta_0^2} \right) = -\frac{1}{\pi_0} = -i_o, \quad E_{\Gamma} \left(\frac{\partial^2 \bar{l}}{\partial \theta_0^2} \right) = -\frac{1}{\pi_0} = E_{\theta_0} \left(\frac{\partial^2 \bar{l}}{\partial \theta_0^2} \right),$$

$$\Gamma = \gamma = nE_{\Gamma} \left\{ \left(\frac{\partial \bar{l}}{\partial \theta_0} \right)^2 \right\} = \frac{1}{\pi_0^2} nE_{\Gamma} \{ (\bar{x} - \pi_0)^2 \} = \frac{1 - \pi_0}{\pi_0} < \pi_0^{-1} = i_o,$$

$$nE_{\Gamma} \left[\frac{\partial \bar{l}}{\partial \theta_0} \left\{ \frac{\partial^2 \bar{l}}{\partial \theta_0^2} - E_{\Gamma} \left(\frac{\partial^2 \bar{l}}{\partial \theta_0^2} \right) \right\} \right] = -\frac{1}{\pi_0 \pi_0^2} nE_{\Gamma} \{ (\bar{x} - \pi_0)^2 \} = -\frac{1 - \pi_0}{\pi_0^2}.$$

Note that the dispersion ratio $\pi_0(1 - \pi_0)/\pi_0 = 1 - \pi_0$ ($\pi_0 \neq 0$) is smaller than 1. Then,

1. $\hat{\theta}_{\text{Smax}}$ (see (26)):

$$k_{\text{Smax}} = \frac{1}{\pi_0^2(-\pi_0^{-1})} \left\{ \pi_0 \left(-\frac{1 - \pi_0}{\pi_0^2} \right) (-\pi_0) - (1 - \pi_0) \right\} = 0. \quad (115)$$

2. $\hat{\theta}_{\text{Smin}}$ (see (41)):

$$k_{\text{Smin}} = \frac{1}{\theta_0^2} \{ n \text{var}_{\Gamma}(\hat{\theta}_{\text{ML}}) + \theta_0 \times 0 \} = \frac{1}{\pi_0^2} \pi_0(1 - \pi_0) = \frac{1 - \pi_0}{\pi_0}. \quad (116)$$

Recall that the zero value of k_{Smax} holds when the Poisson distribution is true (see (70)).

5 Different criteria

5.1 Comparison of optimal coefficients using different criteria

In Sect. 4, the tendency of the relatively large amount of correction by k_{Amin} over that by k_{Amax} was observed. The difference can be partially explained using the case of a single parameter.

Theorem 1 *For a single parameter under correct model specification*

$$k_{Amin} - k_{Amax} = -\frac{i_0^{(D1)}}{2i_0^2\alpha_0} \quad (\alpha_0 \neq 0). \tag{117}$$

Proof From (34) in Result 4 and (41) in Result 6, we have

$$k_{Amin} - k_{Amax} = \frac{1}{\alpha_0^2} \{n \operatorname{acov}_{\theta_0}(\hat{\theta}_{ML}, \hat{\alpha}) + \alpha_0 \beta_1\} - \frac{1}{\alpha_0^2 i_0} \left\{ \alpha_0 n E_{\theta_0} \left(\frac{\partial \bar{l}}{\partial \theta_0} \otimes m \right) \frac{1}{i_0} + \frac{\partial \alpha_0}{\partial \theta_0} \right\}, \tag{118}$$

where $m = \mathbf{M}$ in the case of a single parameter. Since $\beta_1 = i_0^{-2} n E_{\theta_0} (\partial \bar{l} / \partial \theta_0 \otimes m) - \{1 / (2i_0^2)\} (-\partial^3 \bar{l}^* / \partial \theta_0^3)$ (see (18)) and $n \operatorname{acov}_{\theta_0}(\hat{\theta}_{ML}, \hat{\alpha}) = i_0^{-1} \partial \alpha_0 / \partial \theta_0$, (118) becomes

$$k_{Amin} - k_{Amax} = -\frac{1}{2i_0^2\alpha_0} \left(-\frac{\partial^3 \bar{l}^*}{\partial \theta_0^3} \right) = -\frac{i_0^{(D1)}}{2i_0^2\alpha_0}. \text{ Q.E.D.} \tag{119}$$

Let $\operatorname{sign}(\cdot) = -1, 0, 1$ depending on the sign of the real quantity in parentheses, then the following result is obtained.

$$\begin{array}{l} \operatorname{sign}(i_0^{(D1)} \alpha_0): \quad \quad \quad -1 \quad \quad \quad 0 \quad \quad \quad 1 \\ \text{Property:} \quad \quad k_{Amin} > k_{Amax} \quad k_{Amin} = k_{Amax} \quad k_{Amin} < k_{Amax} \end{array} \tag{120}$$

Corollary 1 *For a single canonical parameter in the exponential family under correct model specification*

$$k_{Amin} - k_{Amax} = \beta_1 / \alpha_0. \tag{121}$$

Proof From (34) with the vanishing m and (41) as in Theorem 1, we have

$$\begin{aligned}
 k_{Amin} - k_{Amax} &= \frac{1}{\alpha_0^2} \{n \operatorname{acov}_{\theta_0}(\hat{\theta}_{ML}, \hat{\alpha}) + \alpha_0 \beta_1\} - \frac{\partial \alpha_0 / \partial \theta_0}{\alpha_0^2 i_0} \\
 &= \frac{1}{\alpha_0^2} n \operatorname{avar}_{\theta_0}(\hat{\theta}_{ML}) \frac{\partial \alpha_0}{\partial \theta_0} + \frac{\beta_1}{\alpha_0} - \frac{\partial \alpha_0 / \partial \theta_0}{\alpha_0^2 i_0} \tag{122} \\
 &= \beta_1 / \alpha_0 \quad (\alpha_0 \neq 0). \quad \text{Q.E.D.}
 \end{aligned}$$

Corollary 1 is a special case of Theorem 1 and is alternatively obtained by $\beta_1 = -i_0^{(D1)} / (2i_0^2)$ with the vanishing m in (18). Using Corollary 1, the following results are obtained.

Corollary 2 Under the same conditions as in Corollary 1,

$$k_{Smin} - k_{Smax} = \beta_1 / \theta_0 \quad (\theta_0 \neq 0) \quad \text{and} \quad k_{Bmin} - k_{Bmax} = 1. \tag{123}$$

The second equation of (123) shows that k_{Bmin} , when available, is always greater than k_{Bmax} by 1. The first equation shows the following result.

Corollary 3 Under the same conditions as in Corollary 1, the relative size of k_{Smin} and k_{Smax} is given as follows:

$$\begin{array}{l}
 \operatorname{sign}\{\operatorname{sk}(x^*)\theta_0\}: \quad \quad \quad -1 \quad \quad \quad 0 \quad \quad \quad 1 \\
 \text{Property:} \quad \quad \quad k_{Smin} > k_{Smax} \quad k_{Smin} = k_{Smax} \quad k_{Smin} < k_{Smax}
 \end{array} \tag{124}$$

Proof It is known that

$$\beta_1 = -\frac{1}{2} i_0^{-2} i_0^{(D1)} = -\frac{1}{2} i_0^{-1/2} \operatorname{sk}(x^*) \tag{125}$$

(Ogasawara 2013, Eq. 3.1). Using (125) and Corollary 2, (124) follows. Q.E.D.

In Example 1.1 for the gamma distribution, $\beta_1 = \theta_0 / \alpha$ gives $\beta_1 / \theta_0 = 1 / \alpha > 0$ yielding $k_{Smin} > k_{Smax}$. In Example 2.1 for the Poisson distribution, $\beta_1 = -e^{-\theta_0} / 2 < 0$ (see (64)). However, $\theta_0 (= \log \lambda_0)$ can be negative, 0 or positive (see the comment after (68)), giving an undetermined property of (120) or relative size. In Example 3.1 for the Bernoulli distribution, $\beta_1 / \theta_0 = -(1 - 2\pi_0) / \{2\pi_0(1 - \pi_0)\} \eta_0^{-1} > 0$ ($\pi_0 \neq 0, 0.5, 1$), which gives $k_{Smin} > k_{Smax}$. In Example 4.3 for the univariate normal distribution, $\beta_1 = 0$ and $\operatorname{sk}(x^*) = 0$, giving $k_{Smin} = k_{Smax}$. On the other hand, in Example 4.5 for the same normal distribution under a condition different from that in Example 4.3, $\operatorname{sk}(s_d) = -\sqrt{8} < 0$, where s_d corresponds to x^* in Example 4.3. Since in Example 4.5 $\theta_0 = \phi_0^{-1} > 0$, (124) gives $k_{Smin} > k_{Smax}$ (see (102) and (104)).

Recall that the term with $i_0^{(D1)}$ or $-\partial^3 \bar{l}^* / \partial \theta_0^3$ comes from the last term except the remainder term on the right-hand side of the last equation of (12). Since the other terms including \mathbf{k} have the common factor $-\partial^2 \bar{l}^* / \partial \theta_0 \partial \theta_0' = O(1)$, we find that in the case of a single parameter, the minimization of $MSE \rightarrow_{O(n^{-2})}$ is equal to the maximization of the asymptotic predictive expected log-likelihood up to order $O(n^{-2})$ when we neglect the last term with the third derivative.

In the case of multiple parameters, these two criteria do not give the same optimal sets of coefficients \mathbf{k} . This suggests the following new distances to be minimized.

Definition 1 The generalized mean square error and the scale-free mean square error up to order $O(n^{-2})$ are defined as:

$$GMSE \rightarrow_{O(n^{-2})} \equiv E_T \{ (\hat{\theta}_{Ak} - \theta_0)' (-\Lambda) (\hat{\theta}_{Ak} - \theta_0) \} \rightarrow_{O(n^{-2})} \tag{126}$$

and

$$SMSE \rightarrow_{O(n^{-2})} \equiv E_T \{ (\hat{\theta}_{Ak} - \theta_0)' \text{Diag}(-\Lambda) (\hat{\theta}_{Ak} - \theta_0) \} \rightarrow_{O(n^{-2})}, \tag{127}$$

respectively.

It is obvious that (126) is a Mahalanobis distance under correct model specification, which is also scale free as for (127). In (126) and (127), Λ is an unknown non-stochastic quantity like the target vector θ_0 . Minimization of (126) and (127) can be done in similar manners as before with the explicit solutions minimizing the corresponding quadratic forms. Note that minimizing $GMSE \rightarrow_{O(n^{-2})}$ is equivalent to maximizing the asymptotic predictive expected log-likelihood up to order $O(n^{-2})$ neglecting the term of the third derivative but considering the off-diagonal elements of the second-derivative matrix with the different diagonal elements in the case of multiple parameters. So, the optimal coefficients obtained by this criterion may be situated between those by minimizing the $TMSE \rightarrow_{O(n^{-2})}$ and those by maximizing the asymptotic predictive expected log-likelihood.

5.2 Composite correction vector

So far, the functional form of $\hat{\alpha} = \alpha(\hat{\theta}_{ML})$ is assumed to be given, whose typical cases are $\hat{\theta}_{ML}$ and $\hat{\beta}_1 (= \beta_1(\hat{\theta}_{ML}))$. Let $\hat{\alpha}^{(1)}$ and $\hat{\alpha}^{(2)}$ are two arbitrary vectors similar to $\hat{\alpha}$. Then, define the composite vector using a fixed weight w as follows:

$$\hat{\alpha}_w = w \hat{\alpha}^{(1)} + (1 - w) \hat{\alpha}^{(2)} \tag{128}$$

The vector $\hat{\alpha}_w$ can be used as a special case of $\hat{\alpha}$. An example of $\hat{\alpha}_w$ is $\hat{\alpha}_w = w \hat{\theta}_{ML} + (1 - w) \hat{\beta}_1$, which is expected to yield an intermediate effect between those when $\hat{\alpha} = \hat{\theta}_{ML}$ for shrinkage and when $\hat{\alpha} = \hat{\beta}_1$ for bias adjustment.

Acknowledgements This work was partially supported by a Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology (JSPS KAKENHI, Grant No. 26330031).

Appendix

The asymptotic predictive expected log-likelihood for the maximum likelihood estimator

In this appendix, $E_T(\hat{\tilde{I}}_{ML}^*)_{-O(n^{-2})}$ in (14) is obtained, where

$E_T(\hat{\tilde{I}}_{ML}^*) = E_T\{\bar{I}^*(\hat{\theta}_{ML})\} = \int_{R(\mathbf{X})} \bar{I}^*\{\theta_{ML}(\mathbf{X})\}f_T(\mathbf{X}|\zeta_0)d\mathbf{X}$ (see (10)). For this expectation, we use the expansion of $\hat{\theta}_{ML}$ by Ogasawara (2010, p. 2151) as follows:

$$\hat{\theta}_{ML} - \theta_0 = \sum_{j=1}^3 \Lambda^{(j)} \mathbf{I}_0^{(j)} + O_p(n^{-2}) \quad (\Lambda^{(j)} = O(1), \mathbf{I}_0^{(j)} = O_p(n^{-j/2}), j = 1, 2, 3),$$

$$\Lambda^{(1)} \mathbf{I}_0^{(1)} = -\Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0},$$

$$\Lambda^{(2)} \mathbf{I}_0^{(2)} = \Lambda^{-1} \mathbf{M} \Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0} - \frac{1}{2} \Lambda^{-1} E_T(\mathbf{J}_0^{(3)}) \left(\Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0} \right)^{(2)}$$

$$\begin{aligned} \Lambda^{(3)} \mathbf{I}_0^{(3)} &= -\Lambda^{-1} \mathbf{M} \Lambda^{-1} \mathbf{M} \Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0} + \frac{1}{2} \Lambda^{-1} \mathbf{M} \Lambda^{-1} E_T(\mathbf{J}_0^{(3)}) \left(\Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0} \right)^{(2)} \\ &\quad + \Lambda^{-1} E_T(\mathbf{J}_0^{(3)}) \left\{ \left(\Lambda^{-1} \mathbf{M} \Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0} \right) \otimes \left(\Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0} \right) \right\} - \frac{1}{2} \Lambda^{-1} \{ \mathbf{J}_0^{(3)} \\ &\quad - E_T(\mathbf{J}_0^{(3)}) \} \left(\Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0} \right)^{(2)} \\ &\quad - \frac{1}{2} \Lambda^{-1} E_T(\mathbf{J}_0^{(3)}) \left[\left(\Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0} \right) \otimes \left\{ \Lambda^{-1} E_T(\mathbf{J}_0^{(3)}) \left(\Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0} \right)^{(2)} \right\} \right] \\ &\quad + \frac{1}{6} \Lambda^{-1} E_T(\mathbf{J}_0^{(4)}) \left(\Lambda^{-1} \frac{\partial \bar{I}}{\partial \theta_0} \right)^{(3)}, \end{aligned}$$

$$\mathbf{J}_0^{(j)} \equiv \frac{\partial^j \bar{I}}{\partial \theta_0 (\partial \theta_0')^{(j-1)}} = O_p(1) \quad (j = 3, 4),$$

(129)

$$\begin{aligned} \mathbf{I}_0^{(1)} &= \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0}, \quad \mathbf{I}_0^{(2)} = \left\{ \mathbf{v}'(\mathbf{M}) \otimes \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0}, \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \right)^{(2)} \right\}' \equiv (\mathbf{I}_0^{(2-1)'}, \mathbf{I}_0^{(2-2)'})', \\ \mathbf{I}_0^{(3)} &= \left[\mathbf{v}'(\mathbf{M})^{(2)} \otimes \frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0}, \mathbf{v}'(\mathbf{M}) \otimes \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \right)^{(2)}, \right. \\ &\quad \left. \text{vec}'\{\mathbf{J}_0^{(3)} - \mathbf{E}_T(\mathbf{J}_0^{(3)})\} \otimes \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \right)^{(2)}, \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}'_0} \right)^{(3)} \right]' \\ &\equiv (\mathbf{I}_0^{(3-1)'}, \mathbf{I}_0^{(3-2)'}, \mathbf{I}_0^{(3-3)'}, \mathbf{I}_0^{(3-4)'})', \end{aligned}$$

where $\mathbf{v}(\cdot)$ is the vectorizing operator taking the non-duplicated elements of a symmetric matrix and $\mathbf{v}'(\cdot) = \{\mathbf{v}(\cdot)\}'$.

Using (129), the matrices $\boldsymbol{\Lambda}^{(2-j)}$ ($j = 1, 2$) and $\boldsymbol{\Lambda}^{(3-j)}$ ($j = 1, \dots, 4$) are implicitly defined by

$$\boldsymbol{\Lambda}^{(2)} \mathbf{I}_0^{(2)} = \sum_{j=1}^2 \boldsymbol{\Lambda}^{(2-j)} \mathbf{I}_0^{(2-j)}, \quad \boldsymbol{\Lambda}^{(3)} \mathbf{I}_0^{(3)} = \sum_{j=1}^4 \boldsymbol{\Lambda}^{(3-j)} \mathbf{I}_0^{(3-j)}. \tag{130}$$

The expectation to be derived is

$$\begin{aligned} \mathbf{E}_T(\hat{\mathbf{l}}_{\text{ML}}^*) &= \{\bar{l}^*(\boldsymbol{\theta}_0)\}_{O(1)} \\ &+ \mathbf{E}_T \left[\sum_{j=2}^4 \left\{ \frac{\partial^j \bar{l}^*}{j! (\partial \boldsymbol{\theta}'_0)^{(j)}} (\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}_0)^{(j)} \right\}_{O_p(n^{-j/2})} \right]_{\rightarrow O(n^{-2})} + O(n^{-3}). \end{aligned} \tag{131}$$

In (131), the asymptotic expectation is derived term by term. In the following, the notation, e.g., $(\boldsymbol{\Lambda}^{(2-1)})_{(e:ab,c,d)}$ indicates an element of $\boldsymbol{\Lambda}^{(2-1)}$ corresponding to the e -th row and the column denoted by “ ab, c, d ” which corresponds to $(\mathbf{M})_{ab}$ ($a \geq b$), $\partial \bar{l} / \partial (\boldsymbol{\theta}_0)_c$ and $\partial \bar{l} / \partial (\boldsymbol{\theta}_0)_d$ in $\mathbf{I}_0^{(2-1)}$. The notation, e.g., $\sum_{(g,h)}^{(2)}$

indicates the summation of two terms exchanging g and h ; $\sum_{a \geq b} (\cdot) \equiv \sum_{b=1}^q \sum_{a=b}^q (\cdot)$; and

$$\lambda^{ab} = (\boldsymbol{\Lambda}^{-1})_{ab}.$$

1.

$$\begin{aligned}
 & E_T \left\{ \frac{1}{2} \frac{\partial^2 \bar{l}^*}{(\partial \theta'_0)^{(2)}} (\hat{\theta}_{ML} - \theta_0)^{(2)} \right\} \\
 &= \frac{1}{2} n^{-1} \text{vec}'(\Lambda) n E_T \{ (\Lambda^{(1)} \mathbf{I}_0^{(1)})^{(2)} \} \\
 & \quad + \frac{1}{2} n^{-2} \text{vec}'(\Lambda) \left[2n^2 E_T \{ (\Lambda^{(2)} \mathbf{I}_0^{(2)}) \otimes (\Lambda^{(1)} \mathbf{I}_0^{(1)}) \} \right. \\
 & \quad \left. + 2n^2 E_T \{ (\Lambda^{(3)} \mathbf{I}_0^{(3)}) \otimes (\Lambda^{(1)} \mathbf{I}_0^{(1)}) \} + n^2 E_T \{ (\Lambda^{(2)} \mathbf{I}_0^{(2)})^{(2)} \} \right] \\
 & \quad + O(n^{-3})
 \end{aligned} \tag{132}$$

$$\begin{aligned}
 &= \frac{n^{-1}}{2} \text{tr}(\Gamma \Lambda^{-1}) \\
 & \quad + n^{-2} \sum_{a,b=1}^q \lambda_{ab} \left[\sum_{\substack{c \geq d \\ e,f=1}}^q n^2 E_T \{ m_{cd}(\mathbf{I}_0^{(1)})_e (\mathbf{I}_0^{(1)})_f \} (\Lambda^{(2-1)})_{(a:cd,e)} (-\lambda^{bf}) \right. \\
 & \quad \left. + \sum_{c,d,e=1}^q n^2 E_T \{ (\mathbf{I}_0^{(1)})_c (\mathbf{I}_0^{(1)})_d (\mathbf{I}_0^{(1)})_e \} (\Lambda^{(2-2)})_{(a:c,d)} (-\lambda^{be}) \right. \\
 & \quad \left. + \sum_{c \geq d} \sum_{e \geq f} \sum_{g,h=1}^q (\Lambda^{(3-1)})_{(a:cd,ef,g)} (-\lambda^{bh}) \left[n \text{acov}\{m_{cd}, m_{ef}\} \gamma_{gh} \right. \right. \\
 & \quad \left. \left. + \sum_{(g,h)}^2 n \text{acov}\{m_{cd}, (\mathbf{I}_0^{(1)})_g\} n \text{acov}\{m_{ef}, (\mathbf{I}_0^{(1)})_h\} \right] \right. \\
 & \quad \left. + \sum_{c \geq d} \sum_{e,f,g=1}^q (\Lambda^{(3-2)})_{(a:cd,ef,g)} (-\lambda^{bg}) \sum_{(ef,g)}^3 n \text{acov}\{m_{cd}, (\mathbf{I}_0^{(1)})_e\} \gamma_{fg} \right. \\
 & \quad \left. + \sum_{c,d,ef,g,h=1}^q (\Lambda^{(3-3)})_{(a:cdef,g,h)} (-\lambda^{bh}) \sum_{(f,g,h)}^3 n \text{acov}\{(\mathbf{J}_0^{(3)})_{(c,d,e)}, (\mathbf{I}_0^{(1)})_f\} \gamma_{gh} \right. \\
 & \quad \left. + \sum_{c,d,ef=1}^q (\Lambda^{(3-4)})_{(a:c,d,e)} (-\lambda^{bf}) (\gamma_{cd} \gamma_{ef} + \gamma_{ce} \gamma_{df} + \gamma_{cf} \gamma_{de}) \right]
 \end{aligned}$$

$$\begin{aligned}
 & + \frac{1}{2} \left[\sum_{\substack{c \geq d \\ (B)}} \sum_{e \geq f} \sum_{g,h=1}^q (\mathbf{\Lambda}^{(2-1)})_{(a: cd,g)} (\mathbf{\Lambda}^{(2-1)})_{(b: ef,h)} \left[n \operatorname{acov}(m_{cd}, m_{ef}) \gamma_{gh} \right. \right. \\
 & \qquad \qquad \qquad \left. \left. + \sum_{(g,h)}^2 n \operatorname{acov}\{m_{cd}, (\mathbf{I}_0^{(1)})_g\} n \operatorname{acov}\{m_{ef}, (\mathbf{I}_0^{(1)})_h\} \right] \right. \\
 & + 2 \sum_{c \geq d} \sum_{e,f,g=1}^q (\mathbf{\Lambda}^{(2-1)})_{(a: cd,e)} (\mathbf{\Lambda}^{(2-2)})_{(b: f,g)} \sum_{(e,f,g)}^3 n \operatorname{acov}\{m_{cd}, (\mathbf{I}_0^{(1)})_e\} \gamma_{fg} \\
 & \left. \left. + \sum_{c,d,e,f=1}^q (\mathbf{\Lambda}^{(2-2)})_{(a: c,d)} (\mathbf{\Lambda}^{(2-2)})_{(b: e,f)} (\gamma_{cd}\gamma_{ef} + \gamma_{ce}\gamma_{df} + \gamma_{cf}\gamma_{de}) \right] \right] + O(n^{-3}), \tag{B} \tag{A}
 \end{aligned}$$

where, e.g., $\left[\cdot \right]_{(A)}$ is for ease of finding correspondence.

2.

$$\begin{aligned}
 & E_T \left\{ \frac{1}{6} \frac{\partial^3 \bar{l}^*}{(\partial \theta_0^{(3)})} (\hat{\theta}_{ML} - \theta_0)^{(3)} \right\} \\
 & = n^{-2} \frac{\partial^3 \bar{l}^*}{(\partial \theta_0^{(3)})} \left[\frac{n^2}{6} E_T \{ (\mathbf{\Lambda}^{(1)} \mathbf{I}_0^{(1)})^{(3)} \} + \frac{n^2}{2} E_T \{ (\mathbf{\Lambda}^{(2)} \mathbf{I}_0^{(2)}) \otimes (\mathbf{\Lambda}^{(1)} \mathbf{I}_0^{(1)})^{(2)} \} \right] \\
 & \quad + O(n^{-3}) \tag{133}
 \end{aligned}$$

$$\begin{aligned}
 & = n^{-2} \sum_{a,b,c=1}^q \frac{\partial^3 \bar{l}^*}{\partial \theta_{0a} \partial \theta_{0b} \partial \theta_{0c}} \left[\frac{1}{6} \sum_{d,e,f=1}^q (-\lambda^{ad})(-\lambda^{be})(-\lambda^{cf}) \right. \\
 & \qquad \qquad \qquad \left. \times n^2 E_T \{ (\mathbf{I}_0^{(1)})_d (\mathbf{I}_0^{(1)})_e (\mathbf{I}_0^{(1)})_f \} \right. \\
 & + \frac{1}{2} \left[\sum_{\substack{d \geq e \\ (B)}} \sum_{f,g,h=1}^q (\mathbf{\Lambda}^{(2-1)})_{(a: def)} (-\lambda^{bg})(-\lambda^{ch}) \sum_{(f,g,h)}^3 n \operatorname{acov}\{m_{de}, (\mathbf{I}_0^{(1)})_f\} \gamma_{gh} \right. \\
 & \left. \left. + \sum_{d,e,f,g=1}^q (\mathbf{\Lambda}^{(2-2)})_{(a: d,e)} (-\lambda^{bf})(-\lambda^{cg}) (\gamma_{de}\gamma_{fg} + \gamma_{df}\gamma_{eg} + \gamma_{dg}\gamma_{ef}) \right] \right] + O(n^{-3}). \tag{B} \tag{A}
 \end{aligned}$$

3.

$$\begin{aligned}
& E_T \left\{ \frac{1}{24} \frac{\partial^4 \bar{l}^*}{(\partial \boldsymbol{\theta}'_0)^{(4)}} (\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)^{(4)} \right\} \\
& = \frac{n^{-2}}{8} \text{vec}'(\Lambda^{-1} \Gamma \Lambda^{-1}) \frac{\partial^4 \bar{l}^*}{(\partial \boldsymbol{\theta}_0)^{(2)} (\partial \boldsymbol{\theta}'_0)^{(2)}} \text{vec}(\Lambda^{-1} \Gamma \Lambda^{-1}) + O(n^{-3}).
\end{aligned} \tag{134}$$

In 1, 2 and 3, when the model is true $E_T(\cdot) = E_{\theta_0}(\cdot)$ and $-\Lambda = \Gamma = \mathbf{I}_0$. Especially, the term of order $O(n^{-1})$ becomes

$$\frac{n^{-1}}{2} \text{tr}(\Gamma \Lambda^{-1}) = -n^{-1} \frac{q}{2}. \tag{135}$$

That is, the expectation is asymptotically smaller than $\bar{l}^*(\boldsymbol{\theta}_0) = \bar{l}_0^*$ by $n^{-1}q/2$ up to this order.

References

- Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F (eds) Proceedings of the 2nd international symposium on information theory. Akadémiai Kiado, Budapest, pp 267–281
- Bjornstad JF (1990) Predictive likelihood: a review. *Stat Sci* 5:242–265
- DeGroot MH, Schervish MJ (2002) Probability and statistics, 3rd edn. Addison-Wesley, Boston
- Fisher RA (1956) Statistical methods and scientific inference. Oliver and Boyd, Edinburgh
- Giles DEA, Rayner AC (1979) The mean squared errors of the maximum likelihood and natural-conjugate Bayes regression estimators. *J Econometr* 11:319–334
- Gruber MHJ (1998) Improving efficiency by shrinkage: The James-Stein and ridge regression estimators. Marcel Dekker, New York
- Hinkley D (1979) Predictive likelihood. *Ann Stat* 7:718–728
- Konishi S, Kitagawa G (1996) Generalized information criteria in model selection. *Biometrika* 83:875–890
- Konishi S, Kitagawa G (2003) Asymptotic theory for information criteria in model selection—functional approach. *J Stat Plan Inference* 114:45–61
- Konishi S, Kitagawa G (2008) Information criteria and statistical modeling. Springer, New-York
- Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22:79–86
- Lawless JF, Fredette M (2005) Frequentist prediction intervals and predictive distributions. *Biometrika* 92:529–542
- Lejeune M, Faulkenberry GD (1982) A simple predictive density function. *J Am Stat Assoc* 77:654–657
- Leonard T (1982) Comment (on Lejeune & Faulkenberry, 1982). *J Am Stat Assoc* 77:657–658
- Ogasawara H (2010) Asymptotic expansions for the pivots using log-likelihood derivatives with an application in item response theory. *J Multivar Anal* 101:2149–2167
- Ogasawara H (2013) Asymptotic cumulants of the estimator of the canonical parameter in the exponential family. *J Stat Plan Inference* 143:2142–2150
- Ogasawara H (2014a) Supplement to the paper “Asymptotic cumulants of the estimator of the canonical parameter in the exponential family”. *Econ Rev (Otaru University of Commerce)*, 65 (2 & 3), 3–16. Permalink: <http://hdl.handle.net/10252/5399>. Accessed 20 Nov 2016
- Ogasawara H (2014b) Optimization of the Gaussian and Jeffreys power priors with emphasis on the canonical parameters in the exponential family. *Behaviormetrika* 41:195–223
- Ogasawara H (2015) Bias adjustment minimizing the asymptotic mean square error. *Commun Stat Theory Methods* 44:3503–3522

- Ogasawara H (2016) Optimal information criteria minimizing their asymptotic mean square errors. *Sankhyā B* 78:152–182
- Ogasawara H (2017) Asymptotic cumulants of some information criteria. *J Jpn Soc Comput Stat* (to appear)
- Sakamoto Y, Ishiguro M, Kitagawa G (1986) Akaike information criterion statistics. Reidel, Dordrecht
- Stone M (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J R Stat Soc B* 39:44–47
- Takeuchi K (1976) Distributions of information statistics and criteria of the goodness of models. *Math Sci* 153:12–18 (in Japanese)
- Takezawa K (2012) A revision of AIC for normal error models. *Open J Stat* 2:309–312
- Takezawa K (2014) Estimation of the exponential distribution from a viewpoint of prediction. In: *Proceedings of 2014 Japanese Joint Statistical Meeting*, 305. University of Tokyo, Tokyo, Japan (in Japanese)
- Takezawa K (2015) Estimation of the exponential distribution in the light of future data. *Br J Math Comput Sci* 5:128–132