The AIC Criterion and Symmetrizing the Kullback–Leibler Divergence

Abd-Krim Seghouane, *Member, IEEE*, and Shun-Ichi Amari, *Life Fellow, IEEE*

***Abstract*—The Akaike information criterion (AIC) is a widely used tool for model selection. AIC is derived as an asymptotically unbiased estimator of a function used for ranking candidate models which is a variant of the Kullback–Leibler divergence between the true model and the approximating candidate model. Despite the Kullback–Leibler's computational and theoretical advantages, what can become inconvenient in model selection applications is their lack of symmetry. Simple examples can show that reversing the role of the arguments in the Kullback–Leibler divergence can yield substantially different results. In this paper, three new functions for ranking candidate models are proposed. These functions are constructed by symmetrizing the Kullback–Leibler divergence between the true model and the approximating candidate model. The operations used for symmetrizing are the average, geometric, and harmonic means. It is found that the original AIC criterion is an asymptotically unbiased estimator of these three different functions. Using one of these proposed ranking functions, an example of new bias correction to AIC is derived for univariate linear regression models. A simulation study based on polynomial regression is provided to compare the different proposed ranking functions with AIC and the new derived correction with AIC_c.**

***Index Terms*—Akaike information criterion (AIC), geometric and harmonic means, Kullback–Leibler divergence, model selection.**

I. Introduction

A COMMON and widespread problem in science and engineering is determining a suitable model to describe or characterize an experimental data set. This determination consists of two tasks, the choice of an appropriate model structure, and the estimation of its parameters. The task of parameter estimation is generally done by maximum likelihood or least squares procedures given the structure or dimension of the model. The choice of the dimension of a model is often facilitated by the use of a model selection criterion where one only has to evaluate two simple terms [1]–[3]. The underlying idea of model selection criteria is the parsimonious principle which says that there should be a tradeoff between data fitting and complexity. Thus, all criteria have one term defining a measure of fit, typically a deviance statistic and one term characterizing the complexity, a multiple of the number of free parameters in the model.

Different philosophies have been used to derive different model selection criteria [4]. The minimum description length (MDL) suggested in [5] implements the parsimony principle of economy to code a data set using the idea of universal coding introduced by Kolmogorov. Therefore, the model chosen with the MDL can be considered as providing the best explanation of the data in terms of the code length. In [6] and [7], based on Bayesian arguments and maximum *a posteriori* probability, the Bayesian information criterion (BIC) was introduced. A number of other criteria of this type have been proposed in the literature; from them one can cite, the final prediction error (FPE) [8], the combination among p variables (Cp or Mallows statistic) [9], and the Hannan and Quinn (HQ) [10].

Another way used for deriving model selection criteria is based on the quantification of "how close are" the probability density of the generating model (the true density) and the probability density of the fitted approximating model. Several coefficients have been introduced in the literature for this quantification. Such coefficients have been variously named divergence, distance, and measure, depending on whether they satisfy all the properties of a metric or not. Due to its computational and theoretical advantages, the Kullback–Leibler divergence [11] is perhaps the most frequently used information theoretic coefficient for measuring divergence, separation, or disparity between two probability densities. The Akaike information criterion (AIC) [12], which is the first theoretic criterion to have gained widespread acceptance, is based on the concept of Kullback–Leibler directed divergence between two probability density functions. It is derived from consideration of the asymptotic behavior of the Kullback–Leibler directed divergence between the true density and the probability density of the fitted approximating model under the assumption that the true density is correctly specified or overfitted. Despite its computational and theoretical advantages, the Kullback–Leibler divergence, which is also known as Kullback–Leibler information, relative entropy, or the I-divergence, is a nonsymmetric divergence or measure. Its asymmetry means that reversing the roles of the two arguments in the Kullback–Leibler divergence can yield substantially different results, as can be shown using simple examples. This raises questions about which orientation should be considered more appropriate for model selection and the principles by which this choice should be made.

Manuscript received July 14, 2005; revised June 21, 2006. This work was supported by the National ICT Australia, which is funded by the Australian Department of Communications, Information Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Center of Excellence Program.

A.-K. Seghouane is with the Systems Engineering and Complex Systems Research Program, National ICT Australia, Canberra Research Laboratory, Canberra, ACT 2601, Australia, and also with the Research School of Information Sciences and Engineering, The Australian National University, Canberra, ACT 0200, Australia (e-mail: Abd-krim.seghouane@nicta.com.au).

S.-I. Amari is with the Mathematical Neuroscience Laboratory, RIKEN Brain Science Institute, Saitama 351-0198, Japan (e-mail: amari@brain.riken.go.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2006.882813

1045-9227/$20.00 © 2006 IEEE

In evaluating the adequacy of a fitted model, results from [13] indicate that the directed Kullback–Leibler divergence may better reflect the error due to overfitting, whereas the alternative directed divergence may better reflect the error due to underfitting. This suggests that there might be an advantage to combining the directed and the alternative directed divergences by symmetrizing them. The composite measure may provide a more balanced gauge of model disparity than either of its individual components. Thus, in settings where the collection of candidate models consists of both underfitted and overfitted models a symmetrized version of the Kullback–Leibler divergence may serve to better indicate which models are improperly specified than the Kullback–Leibler divergence. As a consequence, an estimator of a symmetrized version of the Kullback–Leibler divergence may be preferable to an estimator of the Kullback–Leibler divergence as model selection criterion. A simple example of symmetric measure of model separation can be obtained by the sum of the two directed divergences. This divergence is known as Kullback's symmetric divergence or J-divergence [14]. The Kullback information criterion (KIC) proposed in [15] is an asymptotically unbiased estimator of a variant (within a constant) of the J-divergence between the generating model and the fitted approximating model. However, the J-divergence is just one particular case of mixture of the two directed divergences. We can consider alternative ways of mixing the two directed divergences to create a symmetric divergence. What is needed, is a symmetric divergence that can be easily evaluated analytically and estimated to derive a model selection criterion.

In this paper, we propose three alternate ways for symmetrizing the Kullback–Leibler divergence to create three new ranking functions (also called discrepancies). The operations used for symmetrizing are the average geometric and harmonic means. Their asymptotic estimates are derived by using similar arguments, used by Akaike in the derivation of AIC. It is found that AIC is also an asymptotically unbiased estimator of these three ranking functions. This explains why AIC provides good performance also when the collection of candidate models consists of both underfitted and overfitted models. Based on this, other forms of correction to the AIC criterion different from those that already exist can be derived. An example of a new correction, named AIC_c^*, which is different from AIC_c [16], is derived for the univariate linear regression model using the ranking function obtained by averaging the two directed divergences.

The remainder of this paper is organized as follows. In Section II, we discuss and extend the derivation of AIC and KIC criteria. In Section III, we introduce two other ways for symmetrizing the Kullback–Leibler divergence, namely the geometric and harmonic means and derive asymptotic estimates of functions which are variants of these symmetrized quantities. An example of a new bias corrected AIC adapted to univariate linear regression model is derived in Section IV. This new bias corrected AIC is compared to AIC_c in a simulation experiment described in Section V. This simulation also investigates the effectiveness of AIC in estimating the different proposed ranking functions. Concluding remarks are given in Section VI. Theoretical justifications of the proposed results are also presented.

II. AIC Criterion and the Weighted Average Mixture

Suppose a collection of data $\mathbf{y}_n = (y_1, \ldots, y_n)^\top$ has been generated according to an unknown parametric model or density $p(\mathbf{y}|\theta_0)$. We try to find a parametric model which provides a suitable approximation for $p(\mathbf{y}|\theta_0)$ where

$$p(\mathbf{y}|\theta_0) = \prod_{i=1}^{n} p(y_i|\theta_0).$$

Let $\mathcal{M}_k = \{p(\mathbf{y}|\theta_k)|\theta_k \in \Theta_k\}$ denote a k-dimensional parametric family, where θ_k consists of k-independent elements that correspond to the model's parameters. Let $\hat{\theta}_k$ denote the vector of parameter estimates obtained by maximizing the likelihood function $p(\mathbf{y}_n|\theta_k)$ over Θ_k, and let $p(\mathbf{y}|\hat{\theta}_k)$ denote the corresponding fitted model. For simplicity, we will assume $k = 1, 2, \ldots, k_{\max}$, so the collection $\mathcal{M}_k$'s consists of nested families, i.e, $\Theta_1 \subset \Theta_2 \subset \ldots \subset \Theta_{k_{\max}}$ of dimension one through $k_{\max}$ [12]. It is also assumed as in [12] that the search is carried out in a parametric family of distribution including the true model.

To determine which candidate model best approximates the unknown generating model $p(\mathbf{y}|\theta_0)$, we need a real coefficient which provides a suitable reflection of the disparity between $p(\mathbf{y}|\theta_0)$ and an approximating model $p(\mathbf{y}|\hat{\theta}_k)$. The Kullback–Leibler directed divergence is one such coefficient.

The Kullback–Leibler divergence between two parametric densities $p(\mathbf{y}|\theta_k)$ and $p(\mathbf{y}|\theta_0)$, with respect to $p(\mathbf{y}|\theta_0)$ is defined as

$$\begin{aligned} 2I_n(\theta_0, \theta_k) &= E_{\theta_0}\left\{2\ln\frac{p(\mathbf{y}|\theta_0)}{p(\mathbf{y}|\theta_k)}\right\} \\ &= E_{\theta_0}\{-2\ln p(\mathbf{y}|\theta_k)\} - E_{\theta_0}\{-2\ln p(\mathbf{y}|\theta_0)\} \\ &= d_n(\theta_0, \theta_k) - d_n(\theta_0, \theta_0) \end{aligned} \tag{1}$$

where $d_n(\theta_i, \theta_j) = E_{\theta_i}\{-2\ln p(\mathbf{y}|\theta_j)\}$ and the expectation $E_{\theta_i}\{.\}$ is taken with respect to $p(\mathbf{y}|\theta_i)$.

The Kullback–Leibler divergence is one example of the Ali–Silvey class of measures of divergence [17], which are defined to be of the form $D(p(\mathbf{y}|\theta_0), p(\mathbf{y}|\theta_k)) = f[E_{\theta_0}\{C(\phi[p(\mathbf{y}|\theta_0), p(\mathbf{y}|\theta_k)])\}]$, where $\phi(.,.)$ represents the ratio $p(\mathbf{y}|\theta_k)/p(\mathbf{y}|\theta_0)$, $C(.)$ is a continuous convex function and f is an increasing function. The Kullback–Leibler divergence corresponds to $C(x) = -\log x$ and $f(x) = x$.

Since $d_n(\theta_0, \theta_0)$ does not depend on θ_k, any ranking of the candidate models according to $2I_n(\theta_0, \theta_k)$ would be identical to ranking them according to $d_n(\theta_0, \theta_k)$ or

$$D_{1n}(\theta_0, \theta_k) = 2I_n(\theta_0, \theta_k) + d_n(\theta_0, \theta_0). \tag{2}$$

Therefore, $D_{1n}(\theta_0, \hat{\theta}_k)$ would provide a suitable measure of a variant of Kullback–Leibler directed divergence between the generating model $p(\mathbf{y}|\theta_0)$ and the fitted model $p(\mathbf{y}|\hat{\theta}_k)$. Yet, evaluating $D_{1n}(\theta_0, \hat{\theta}_k)$ is not possible, since doing so requires the knowledge of θ_0.

In [12], it was argued that $-2\log p(\mathbf{y}_n|\hat{\theta}_k)$ is a biased estimator of $D_{1n}(\theta_0,\hat{\theta}_k)$ and an *asymptotic* bias correction was proposed, leading to [12]

$$\text{AIC} = -2\ln p(\mathbf{y}_n|\hat{\theta}_k) + 2k. \tag{3}$$

If we write

$$\Delta_n(k,\theta_0) = E_{\theta_0}\left\{D_{1n}(\theta_0,\hat{\theta}_k)\right\}$$

then, one can establish that

$$\Delta_n(k,\theta_0) = E_{\theta_0}\{\text{AIC}\} + o(1).$$

The alternate divergence $I_n(\hat{\theta}_k,\theta_0)$ provides extra information that may be useful to take into account in model selection problems. Therefore, what is needed, is a symmetric divergence that can be asymptotically estimated to provide an information theoretic criterion for model selection.

To address the symmetry problem, various solutions can be considered. The weighted average, defined as

$$M_\eta(\theta_0,\theta_k) = \eta I_n(\theta_0,\theta_k) + (1-\eta) I_n(\theta_k,\theta_0), \qquad \eta\in[0,1]. \tag{4}$$

is the simplest solution to generate a symmetric divergence for a particular value of η. Indeed, the average of the Kullback–Leibler divergences which is symmetric is defined as

$$\begin{aligned} M_n(\theta_0,\theta_k) &= \frac{I_n(\theta_0,\theta_k) + I_n(\theta_k,\theta_0)}{2} \\ &= \frac{1}{4}\left(d_n(\theta_0,\theta_k) - d_n(\theta_0,\theta_0)\right. \\ &\qquad \left. + d_n(\theta_k,\theta_0) - d_n(\theta_k,\theta_k)\right). \end{aligned} \tag{5}$$

It corresponds to the particular case $\eta = 1/2$ of (4) and equals $J_n(\theta_0,\theta_k)/2$ [14].

The coefficient $M_n(\theta_0,\theta_k)$ is also an example of the Ali–Silvey class of measures of divergence; it corresponds to $C(x) = ((x-1)/2)\log x$ and $f(x) = x$.

Now, since $d_n(\theta_0,\theta_0)$ does not depend on θ_k and by analogy to (2), the basis for deriving AIC, any ranking of the candidate models according to $2M_n(\theta_0,\theta_k)$ would be identical to ranking them according to

$$D_{2n}(\theta_0,\theta_k) = 2M_n(\theta_0,\theta_k) + d_n(\theta_0,\theta_0). \tag{6}$$

Evaluating $D_{2n}(\theta_0,\hat{\theta}_k)$ is not possible, since doing so requires the knowledge of θ_0. However, an asymptotically unbiased estimator can be derived.

Under the assumptions mentioned in the beginning of this section, and the usual regularity conditions required to ensure the consistency and asymptotic normality of the maximum-likelihood estimator $\hat{\theta}_k$, we have the following statistic with an expectation which is within $o(1)$ of $E_{\theta_0}\{D_{2n}(\theta_0,\hat{\theta}_k)\}$.

Proposition 1: The AIC criterion

$$\text{AIC} = -2\ln p(\mathbf{y}_n|\hat{\theta}_k) + 2k \tag{7}$$

also satisfies

$$E_{\theta_0}\left\{D_{2n}(\theta_0,\hat{\theta}_k)\right\} = E_{\theta_0}\{\text{AIC}\} + o(1).$$

Proof: See Appendix I.

The AIC is also an asymptotically unbiased estimator of a variant within a constant of the average of the Kullback–Leibler divergences. Therefore, it provides information on the behavior of $I_n(\theta_k,\theta)$, a term that is more sensitive to underfitting [13].

Other forms of weighted average can be imagined to create symmetric divergences; for example, $S_\eta(\theta_0,\theta_k)$ defined as

$$S_\eta(\theta_0,\theta_k) = \eta I_n(\theta_0,\theta_k) + (2-\eta) I_n(\theta_k,\theta_0), \qquad \eta\in[0,2]$$

can also be used to generate a symmetric divergence. Indeed, this mixture of $I_n(\theta_0,\theta_k)$ and $I_n(\theta_k,\theta_0)$ is a generalization of the J-divergence which is symmetric [14]

$$\begin{aligned} S_n(\theta_0,\theta_k) &= 2S_{\eta=1}(\theta_0,\theta_k) \\ &= 2J_n(\theta_0,\theta_k) \\ &= 2I_n(\theta_0,\theta_k) + 2I_n(\theta_k,\theta_0) \\ &= d_n(\theta_0,\theta_k) - d_n(\theta_0,\theta_0) \\ &\quad + d_n(\theta_k,\theta_0) - d_n(\theta_k,\theta_k). \end{aligned}$$

The coefficient $S_n(\theta_0,\hat{\theta}_k)$ is another example of the Ali–Silvey class of measures of divergence, it corresponds to $C(x) = (x-1)\log x$ and $f(x) = x$.

By analogy to (2), the basis of AIC, one can imagine ranking the candidate models according to

$$D^*_{2n}(\theta_0,\theta_k) = 2S_n(\theta_0,\theta_k) + d_n(\theta_0,\theta_0). \tag{8}$$

Then, $D^*_{2n}(\theta_0,\theta_k)$ constitutes a basis for deriving a model selection criterion.

Under the previous assumptions, an asymptotically unbiased estimator within $o(1)$ of $E_{\theta_0}\{D^*_{2n}(\theta_0,\hat{\theta}_k)\}$ can be derived.

Proposition 2: The KIC criterion introduced in [15]

$$\text{KIC} = -2\ln p(\mathbf{y}_n|\hat{\theta}_k) + 3k \tag{9}$$

satisfy

$$E_{\theta_0}\left\{D^*_{2n}(\theta_0,\hat{\theta}_k)\right\} = E_{\theta_0}\{\text{KIC}\} + o(1).$$

Proof: Can easily be derived by following the same lines of derivation of proposition 1 in Appendix I.

We should not expect that every reasonable mixture of the directed and alternate Kullback–Leibler divergences to be of the form $f[E\{C(\phi)\}]$. Other mixtures of the directed and alternate Kullback–Leibler divergences that are not member of the Ali–Silvey class can be constructed for a model selection purpose. Two different mixtures are proposed in Section III.

III. AIC Criterion and the Geometric and Harmonic Means

To address the symmetry problem of the Kullback–Leibler divergence, we can consider different forms of mixture of the directed and alternate Kullback–Leibler divergences or different ways of averaging them. Naturally, we immediately think of the geometric and harmonic means. Based on this, in what follows, we define two new symmetric measures of divergence between probabilities and derive their asymptotic estimates. These estimators can be used as model selection criteria.

The geometric mean of $I_n(\theta_0, \theta_k)$ and $I_n(\theta_k, \theta_0)$ is defined by

$$G_n(\theta_0, \theta_k) = \sqrt{I_n(\theta_0, \theta_k) I_n(\theta_k, \theta_0)}$$

and it corresponds to the particular case of $\eta = 1/2$ of the more general mixture

$$G_\eta(\theta_0, \theta_k) = I_n(\theta_0, \theta_k)^\eta I_n(\theta_k, \theta_0)^{1-\eta}, \qquad \eta \in [0, 1].$$

Its weighted version is

$$\ln G_\eta(\theta_0, \theta_k) = \eta \ln I_n(\theta_0, \theta_k) + (1-\eta) \ln I_n(\theta_k, \theta_0), \qquad \eta \in [0, 1].$$

As in Section II, since $d_n(\theta_0, \theta_0)$ does not depend on θ_k and by analogy to (2), any ranking of the candidate models according to $2G_n(\theta_0, \theta_k)$ would be identical to ranking them according to

$$D_{3n}(\theta_0, \theta_k) = 2G_n(\theta_0, \theta_k) + d_n(\theta_0, \theta_0). \tag{10}$$

Then, this coefficient measure can be used for model selection within the limit of the derivation of a model selection criterion. However, and as earlier, evaluating $D_{3n}(\theta_0, \hat{\theta}_k)$ is not possible, since doing so requires the knowledge of θ_0, but an asymptotically unbiased estimator can be derived.

Under previous assumptions, we have the following statistic with an expectation which is within $O(n^{-1/2})$ of $E_{\theta_0}\{D_{3n}(\theta_0, \hat{\theta}_k)\}$.

Proposition 3: The AIC criterion

$$\text{AIC} = -2 \ln p(\mathbf{y}_n | \hat{\theta}_k) + 2k \tag{11}$$

also satisfies

$$E_{\theta_0}\left\{D_{3n}(\theta_0, \hat{\theta}_k)\right\} = E_{\theta_0}\{\text{AIC}\} + O(n^{-1/2}).$$

Proof: See Appendix II.

The harmonic mean of $I_n(\theta_0, \theta_k)$ and $I_n(\theta_k, \theta_0)$ is defined by

$$H_n(\theta_0, \theta_k) = \frac{2}{\dfrac{1}{I_n(\theta_0, \theta_k)} + \dfrac{1}{I_n(\theta_k, \theta_0)}}.$$

By analogy to (2), any ranking of the candidate models according to $2H_n(\theta_0, \theta_k)$ would be identical to ranking them according to

$$D_{4n}(\theta_0, \theta_k) = 2H_n(\theta_0, \theta_k) + d_n(\theta_0, \theta_0). \tag{12}$$

This measure can also be used for model selection within the limit of the derivation of a model selection criterion. Under previous assumptions, we have the following statistic with an expectation which is within $O(n^{-1/2})$ of $E_{\theta_0}\{D_{4n}(\theta_0, \hat{\theta}_k)\}$.

Proposition 4: The AIC criterion

$$\text{AIC} = -2 \ln p(\mathbf{y}_n | \hat{\theta}_k) + 2k \tag{13}$$

also satisfies

$$E_{\theta_0}\left\{D_{4n}(\theta_0, \hat{\theta}_k)\right\} = E_{\theta_0}\{\text{AIC}\} + O(n^{-1/2}).$$

Proof: See Appendix III.

Therefore, in the dominant order sense the Kullback–Leibler divergence, the average, geometric, and harmonic means of the two Kullback–Leibler divergences are equivalent.

The geometric $G_n(\theta_0, \theta_k)$ and harmonic $H_n(\theta_0, \theta_k)$ means are not members of the Ali–Silvey class of measures of divergence, but because of their direct relationship to the Kullback–Leibler divergence, they do have its two distance properties plus the symmetric property as does the J-divergence.

All of the aforementioned mixtures means can be considered as particular cases of the one-parameter family of means, named the α-mean. This concept originates from information geometry [18], which has a family of invariance structures (α-affine connections together with the Fisher Riemannian metric) in the manifold of probability distributions. The α-mean of two positives numbers x and y is defined by

$$m_\alpha(x, y) = t_\alpha \left[x^{(1-\alpha)/2} + y^{(1-\alpha)/2}\right]^{2/(1-\alpha)}$$

where $t_\alpha = 2^{-2/(1-\alpha)}$ is a constant that makes sure $m_\alpha(x, x) = x$ is satisfied.

Let us introduce the α-representation of a number u by

$$l_\alpha(u) = \begin{cases} \dfrac{2}{1-\alpha} u^{(1-\alpha)/2}, & \alpha \neq 1 \\ \log u, & \alpha = 1 \end{cases}$$

when $\alpha = 1$; the limit of $(2/(1-\alpha))u^{(1-\alpha)/2}$ is $\log u$. Then, the α-representation of the α-mean of two numbers is the mean of the α-representation of respective numbers

$$l_\alpha\left(m_\alpha(x, y)\right) = \frac{1}{2}\left[l_\alpha(x) + l_\alpha(y)\right].$$

From this, we derive the special cases

$$\begin{aligned}
&\alpha = 3 \quad m_3(x,y) = \frac{2}{\frac{1}{x}+\frac{1}{y}} \\
&\alpha = 1 \quad m_1(x,y) = \sqrt{xy} \\
&\alpha = -1 \quad m_{-1}(x,y) = \frac{1}{2}(x+y) \\
&\alpha = -\infty \quad m_{-\infty}(x,y) = \max(x,y) \\
&\alpha = +\infty \quad m_{+\infty}(x,y) = \min(x,y).
\end{aligned}$$

The α-mean is monotone with respect α, $m_\alpha(x,y) \geq m_{\alpha'}(x,y)$, for $\alpha \leq \alpha'$. Therefore, the relation among the various symmetric mixtures of the Kullback–Leibler divergences is

$$\max\left[I_n(\theta_0,\theta_k), I_n(\theta_k,\theta_0)\right] \geq M_n(\theta_0,\theta_k) \geq G_n(\theta_0,\theta_k)$$

and

$$G_n(\theta_0,\theta_k) \geq H_n(\theta_0,\theta_k) \geq \min\left[I_n(\theta_0,\theta_k), I_n(\theta_k,\theta_0)\right].$$

IV. Example of a New Bias Corrected AIC

Based on one of the previous mentioned ranking functions, a new bias correction to AIC is derived for univariate linear regression model. The correction is of particular use when the sample size is small or when the number of fitted parameters is a moderate to large fraction of the sample size. The proposed correction is different from

$$\mathrm{AIC}_c = -2\ln p(\mathbf{y}|\hat{\theta}_k) + \frac{2n(k+1)}{n-k-2}$$

proposed in [16] and derived from $D_{1n}(\theta_0,\hat{\theta_k})$. For linear univariate regression models, the corrected criterion, called AIC_c^*, is an asymptotically exact unbiased estimator of $D_{2n}(\theta_0,\hat{\theta_k})$, assuming the true model is correctly specified or overfitted.

As in [13], suppose that the generating model for the data and the kth candidate model are, respectively, given by

$$\begin{aligned}
\mathbf{y}_n &= X\beta_0 + \boldsymbol{\varepsilon}_0, \qquad \boldsymbol{\varepsilon}_0 \sim N\left(0, \sigma_0^2 I_n\right) \\
\mathbf{y}_n &= X\beta_k + \boldsymbol{\varepsilon}_k, \qquad \boldsymbol{\varepsilon}_k \sim N\left(0, \sigma_k^2 I_n\right).
\end{aligned} \tag{14}$$

Here, $\mathbf{y}_n$ is an $n \times 1$ observation vector, $\boldsymbol{\varepsilon}_0$ and $\boldsymbol{\varepsilon}$ are $n \times 1$ noise vectors, β_0 and β_k are $k \times 1$ parameter vectors, and X is an $n \times k$ design matrix of full rank. The vector of parameters is $\theta_k = [\beta_k\ \sigma_k^2]$, where $k = \dim(\beta_k)$, and the maximum-likelihood estimate $\hat{\theta}_k$ of θ_0 is $\hat{\beta}_k = (X'X)^{-1}X'\mathbf{y}_n$ and $\hat{\sigma}_k = (\mathbf{y}_n - X\hat{\beta}_k)'(\mathbf{y}_n - X\hat{\beta}_k)/n$. In this case, the total number of parameter is $k+1 = \dim(\theta_k)$.

It is also assumed (as in [12] and [16]) that the true model is correctly specified or orverfitted by all the candidate models, i.e., $\theta_0 \in \Theta_k, \forall k$.

Proposition 5: Let

$$\mathrm{AIC}_c^* = -2\ln p(\mathbf{y}|\hat{\theta}_k) + \frac{n(k+1)}{n-k-2} - n\psi\left(\frac{n-k}{2}\right) + n\ln\frac{n}{2} \tag{15}$$

where ψ is the *digamma* function. Then

$$E_{\theta_0}\left\{D_{2n}(\theta_0,\hat{\theta}_k)\right\} = E_{\theta_0}\left\{\mathrm{AIC}_c^*\right\}.$$

Proof: See Appendix IV.

Using this approximation [13]

$$\psi\left(\frac{n-k}{2}\right) \simeq \ln\left(\frac{n}{2}\right) - \frac{k}{n} - \frac{1}{n-k}$$

in (14), we obtain

$$\mathrm{AIC}_c^* \simeq -2\ln p(\mathbf{y}|\hat{\theta}_k) + \frac{(k+1)(2n-k-2)}{n-k-2} + \frac{k}{n-k}. \tag{16}$$

It is worth mentioning that asymptotically AIC_c^* will converge to AIC.

V. Simulation Results

In this example, we propose to compare the two univariate corrected versions of AIC, i.e., AIC_c and AIC_c^*, and investigate the effectiveness of AIC in estimating the different aforementioned ranking functions. For this, we consider the classical problem of fitting polynomial functions to a set of points over a compact support. The motivation for studying this example is that polynomials create a difficult model selection problem with a marked tendency to overfitting, especially when the data set is small. In addition, polynomials constitute an interesting class of univariate linear models, for which there exist efficient techniques for computing the best fit.

Two polynomials of order three and six with parameters $\beta = [1, -0.5, -5, -1.5]^T$ and $\beta = [1, -0.5, -5, -1.5, 1, 2, -4]^T$ are used to generate eight simulation data sets. The sample sizes, $n = 15$ and $n = 20$ are used with each polynomial. The noise levels $\mathrm{snr} = 0$ dB and $\mathrm{snr} = 10$ dB are used with the polynomial of order three and $\mathrm{snr} = 10$ dB and $\mathrm{snr} = 15$ dB are used with the polynomial of order six. The design points are equally spaced over the interval $[-x_{\max}, x_{\max}]$ with $x_{\max} = 3$. Data are made noisy by adding independent identically distributed (i.i.d.) samples sampled from a zero mean Gaussian random variable with variance σ_0^2 adjusted to have the required value of $\mathrm{snr} = 10\log(\sigma_s^2/\sigma_0^2)$. Here

$$\sigma_s^2 = \frac{1}{2x_{\max}} \int_{-x_{\max}}^{x_{\max}} f(x)^2 dx.$$

A Monte-Carlo simulation consisting of 1000 runs was conducted. For each realization of the data set, a least squares algorithm was used to fit candidate polynomials of degree 0 to 10. Table I gives a comparison of AIC, AIC_c, AIC_c^*, KIC, FPE

TABLE I
FREQUENCY OF THE MODEL ORDER SELECTED BY EACH CRITERION FOR 1000 REALIZATIONS

Polynomial order	snr	N	Order	AIC	AIC_c	AIC_c^*	KIC	FPE	BIC
3	0	15	$< k_0$	12	109	70	45	15	38
3	0	15	$= k_0$	246	**800**	**749**	453	331	391
3	0	15	$> k_0$	742	91	181	502	654	571
3	10	15	$< k_0$	0	0	0	0	0	0
3	10	15	$= k_0$	264	**888**	**804**	512	359	452
3	10	15	$> k_0$	736	112	196	488	641	548
3	0	20	$< k_0$	7	35	23	33	7	33
3	0	20	$= k_0$	422	**791**	**725**	638	471	636
3	0	20	$> k_0$	571	174	252	329	522	331
3	10	20	$< k_0$	0	0	0	0	0	0
3	10	20	$= k_0$	429	**851**	**782**	692	473	690
3	10	20	$> k_0$	571	149	218	308	527	310
3	10	30	$< k_0$	0	0	0	0	0	0
3	10	30	$= k_0$	549	**819**	**766**	799	566	860
3	10	30	$> k_0$	451	181	234	201	434	140
3	10	100	$< k_0$	0	0	0	0	0	0
3	10	100	$= k_0$	688	**768**	**747**	871	690	958
3	10	100	$> k_0$	312	232	253	129	310	42
6	10	15	$< k_0$	54	683	421	152	87	120
6	10	15	$= k_0$	213	**284**	**461**	306	300	280
6	10	15	$> k_0$	733	33	118	542	613	600
6	15	15	$< k_0$	0	177	42	6	1	1
6	15	15	$= k_0$	253	**781**	**834**	415	377	378
6	15	15	$> k_0$	747	42	124	579	622	621
6	10	20	$< k_0$	70	371	211	156	88	156
6	10	20	$= k_0$	385	**559**	**606**	512	455	512
6	10	20	$> k_0$	545	70	183	332	457	332
6	15	20	$< k_0$	36	266	137	96	41	96
6	15	20	$= k_0$	399	**637**	**647**	535	460	535
6	15	20	$> k_0$	565	97	216	369	499	369

[8], and BIC [7] in terms of relative frequencies of the selected model order.

The corrected versions AIC_c and AIC_c^* show clearly better performances, followed by KIC or BIC depending on the data set. The performances of AIC_c in comparison to AIC_c^* depend also on the data set. This behavior is directly linked to the behavior of the ranking functions $D_{1n}(\theta_0, \hat{\theta}_k)$ and $D_{2n}(\theta_0, \hat{\theta}_k)$ used in their derivations. We can also remark from these simulations that AIC_c^* underfit less than AIC_c when the polynomial order equal six, due to the presence in $D_{2n}(\theta_0, \hat{\theta}_k)$ of the additional term which measures model dissimilarity in a way which is more sensitive to underfitting. This explains why the perfor-

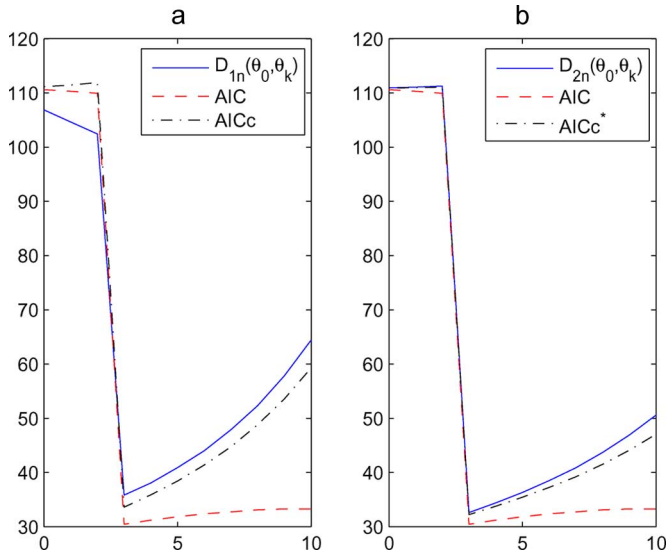


Fig. 1. For the third-order polynomial with $n = 20$ and $\mathrm{snr} = 10$ dB: (a) averages of AIC, AIC_c, and $D_{1n}(\theta_0, \hat{\theta}_k)$; and (b) averages of AIC, AIC_c^*, and $D_{2n}(\theta_0, \hat{\theta}_k)$. The x-axis represents the model order "k."

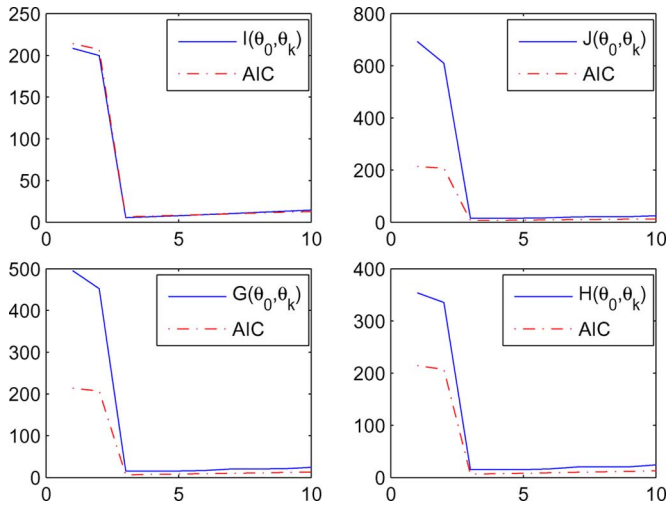


Fig. 2. $I_n(\theta_0, \hat{\theta}_k)$, $J_n(\theta_0, \hat{\theta}_k)$, $G_n(\theta_0, \hat{\theta}_k)$, $H_n(\theta_0, \hat{\theta}_k)$, and AIC for the third-order polynomial with $n = 100$ and $\mathrm{snr} = 10$ dB. The x-axis represents the model order "k."

mance of AIC_c^* is better in comparison to AIC_c in this particular case. Therefore, depending on the situation there will be ranking functions that are more suited than others for the derivation of appropriate model selection criteria. Note also that no single model selection criterion will always be better than another. Certain criteria perform best for specific models types.

Fig. 1 provides some insight as to why AIC_c and AIC_c^* tend to outperform AIC in selecting the correct order of the true model. Consider the fourth simulation data set of Table I. Simulated values of $E_{\theta_0}\{\mathrm{AIC}_c\}$, $E_{\theta_0}\{D_{1n}(\theta_0, \hat{\theta}_k)\}$, $E_{\theta_0}\{\mathrm{AIC}_c^*\}$, $E_{\theta_0}\{D_{2n}(\theta_0, \hat{\theta}_k)\}$, and $E_{\theta_0}\{\mathrm{AIC}\}$ are obtained by averaging AIC_c, $D_{1n}(\theta_0, \hat{\theta}_k)$, AIC_c^*, $D_{2n}(\theta_0, \hat{\theta}_k)$, and AIC, respectively, over the 1000 realizations. These average values are plotted versus k.

Figs. 2 and 3 are obtained with the two other simulation data sets (as described in the figures). The different curves

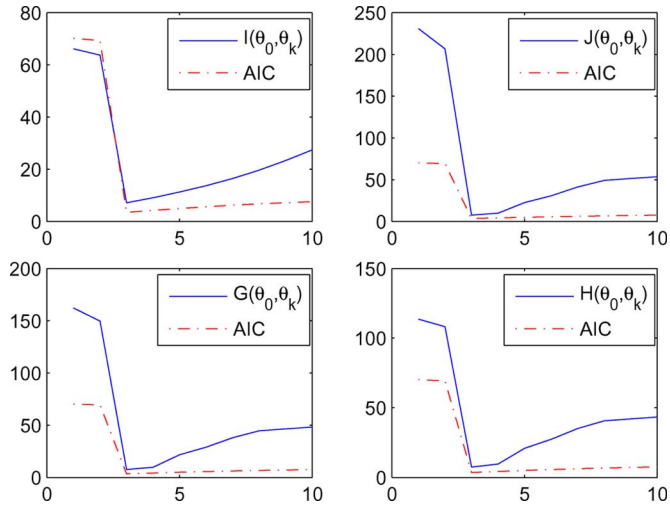


Fig. 3. $I_n(\theta_0, \hat{\theta}_k)$, $J_n(\theta_0, \hat{\theta}_k)$, $G_n(\theta_0, \hat{\theta}_k)$, $H_n(\theta_0, \hat{\theta}_k)$, and AIC for the third-order polynomial with $n = 30$ and $\mathrm{snr} = 10$ dB. The x-axis represents the model order "k."

$E_{\theta_0}\{D_{1n}(\theta_0, \theta_k)\}$, $E_{\theta_0}\{D_{2n}(\theta_0, \theta_k)\}$ (a particular case of the weighted average), $E_{\theta_0}\{D_{3n}(\theta_0, \theta_k)\}$, and $E_{\theta_0}\{D_{4n}(\theta_0, \theta_k)\}$, which are plotted versus k are obtained by averaging the exact expression of these quantities over the 1000 replications. These expressions can easily be derived in the linear case. We can observe on the figure that all these divergences are minimized at $k = 3$. Note, as illustrated on Fig. 2, that AIC provides a good estimate of a variant of these divergences for large n. However, as illustrated on Fig. 3, AIC is a negatively biased estimator of the variant of these divergences for small n, especially when k increases.

VI. Conclusion

This paper had an objective of searching for easily estimated symmetric divergences based on the Kullback–Leibler divergence that can be used to derive model criteria. The obtained results suggest that symmetrizing the Kullback–Leibler divergence may provide a foundation for the development of other model selection criteria which can be preferable to that provided by the Kullback–Leibler divergence. This motivates the need to further explore the properties of $M_n(\theta, \theta_k)$, $G_n(\theta, \theta_k)$, and $H_n(\theta, \theta_k)$ as a measure of model disparity, as well as the need to develop small sample estimator of their variant functions.

By symmetrizing the Kullback–Leibler divergence in different ways, different functions for ranking candidate models can be constructed and thus different model selection criteria derived. Each of these criteria will have properties that are linked to the function from which it has been derived. Obviously, other forms of symmetrizing the Kullback–Leibler divergence different from those proposed in this paper can be imagined; however, it is not evident that all possible symmetrizing operations will be useful for model selection.

The results derived in this paper provide support to the idea affirming that in the dominant order sense, the Kullback–Leibler divergence, the average, and the geometric and harmonic means of Kullback–Leibler's divergences are equivalent since all of them can be estimated using the AIC criterion.

As it can be seen in Section V, in settings where the sample size is small and the candidate models consists of families which are excessively overparameterized, the AIC criterion may exhibit a negative bias in estimating the different variants of the proposed divergences. In such cases, this criterion provides an insufficient degree of bias correction for the larger fitted models; as a result, it tends to underestimate the ranking function from which it is derived. Recent work has led to successful small sample refinements of the penalty terms of AIC [16], [19], [20] and KIC [13], [21], [22]. Two types of refinements are used in the literature. The first one is based on assuming a particular form for candidate model families and using the characteristics of this class of model families to derive more precise approximation for the penalty term. The second refinement is based on using bootstrap to approximation the bias adjustment. In small applications where excessively overparameterized families are candidates, AIC tends to underestimate $D_{3n}(\theta_0, \hat{\theta}_k)$ and $D_{4n}(\theta_0, \hat{\theta}_k)$. In future work, we expect to use such approaches to develop new small sample refinements for the penalty term of AIC to estimate more accurately $D_{3n}(\theta_0, \hat{\theta}_k)$ and $D_{4n}(\theta_0, \hat{\theta}_k)$.

APPENDIX I

The following Taylor approximation results are used to approximate the expectation of $D_{2n}(\theta_0, \hat{\theta}_k)$ within $o(1)$. Define

$$I(\theta_k) = E_{\theta_k}\left\{-\frac{\partial^2 \ln p(\mathbf{y}|\theta_k)}{\partial\theta_k \partial\theta_k'}\right\}$$

and

$$F(\theta_k, \mathbf{y}) = \left\{-\frac{\partial^2 \ln p(\mathbf{y}|\theta_k)}{\partial\theta_k \partial\theta_k'}\right\}.$$

Thus, $I(\theta_0)$, $I(\hat{\theta}_k)$, and $F(\hat{\theta}_k, \mathbf{y})$ denote, respectively, the true, the expected, and the observed Fisher information matrix.

By expanding $-2\ln p(\mathbf{y}|\theta_0)$ around $\hat{\theta}_k$, we have

$$-2\ln p(\mathbf{y}|\theta_0) = -2\ln p(\mathbf{y}|\hat{\theta}_k) + (\hat{\theta}_k - \theta_0)' F(\theta_k^*, \mathbf{y})(\hat{\theta}_k - \theta_0)$$

where θ_k^* is between $\hat{\theta}_k$ and θ_0 and ε_1 is $o_p(1)$. Usual regularity conditions justify the expansion

$$\begin{aligned} d_n(\theta_0, \theta_0) = {} & E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\} \\ & + E_{\theta_0}\left\{(\hat{\theta}_k - \theta_0)' F(\hat{\theta}_k, \mathbf{y})(\hat{\theta}_k - \theta_0)\right\} + o(1). \quad (17) \end{aligned}$$

From the second-order expansion of $d_n(\theta_0, \hat{\theta}_k)$ around θ_0 and the usual regularity conditions, we have

$$d(\theta_0, \hat{\theta}_k) = d(\theta_0, \theta_0) + (\hat{\theta}_k - \theta_0)' I(\theta_0)(\hat{\theta}_k - \theta_0) + \varepsilon_2$$

where ε_2 is $o_p(1)$.

Then, from (1)

$$2I_n(\theta_0, \hat{\theta}_k) = \left\{(\hat{\theta}_k - \theta_0)' I(\theta_0)(\hat{\theta}_k - \theta_0)\right\} + \varepsilon_2. \quad (18)$$

From the second-order expansion of $d_n(\hat{\theta}_k, \theta_0)$ around $\hat{\theta}_k$ and the usual regularity conditions, we have

$$d(\hat{\theta}_k, \theta_0) = d(\hat{\theta}_k, \hat{\theta}_k) + (\hat{\theta}_k - \theta_0)' I(\hat{\theta}_k)(\hat{\theta}_k - \theta_0) + \varepsilon_3$$

where ε_3 is $o_p(1)$.

Then

$$\begin{aligned} 2I_n(\hat{\theta}_k, \theta_0) &= d(\hat{\theta}_k, \theta_0) - d(\hat{\theta}_k, \hat{\theta}_k) \\ &= \left\{(\hat{\theta}_k - \theta_0)' I(\hat{\theta}_k)(\hat{\theta}_k - \theta_0)\right\} + \varepsilon_3. \quad (19) \end{aligned}$$

The result follows from the fact that under the usual regularity conditions, the terms

$$(\hat{\theta}_k - \theta_0)' F(\hat{\theta}_k, \mathbf{y})(\hat{\theta}_k - \theta_0), \quad (\hat{\theta}_k - \theta_0)' I(\theta_0)(\hat{\theta}_k - \theta_0)$$

and

$$(\hat{\theta}_k - \theta_0)' I(\hat{\theta}_k)(\hat{\theta}_k - \theta_0)$$

converge to centrally χ^2 random variables with k degrees of freedom. Thus, the expectation of each of these terms is within $o(1)$ of k. Then

$$\begin{aligned} & E_{\theta_0}\left\{D_{2n}(\theta_0, \hat{\theta}_k)\right\} \\ &= E_{\theta_0}\left\{2M_n(\theta_0, \hat{\theta}_k)\right\} + d_n(\theta_0, \theta_0) \\ &= E_{\theta_0}\left\{I_n(\theta_0, \hat{\theta}_k) + I_n(\hat{\theta}_k, \theta_0)\right\} + d_n(\theta_0, \theta_0), \quad \eta \in [0,1] \\ &= \frac{1}{2}E_{\theta_0}\left\{(\hat{\theta}_k - \theta_0)' I(\theta_0)(\hat{\theta}_k - \theta_0)\right\} \\ &\quad + \frac{1}{2}E_{\theta_0}\left\{(\hat{\theta}_k - \theta_0)' I(\hat{\theta}_k)(\hat{\theta}_k - \theta_0)\right\} \\ &\quad + E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\} \\ &\quad + E_{\theta_0}\left\{(\hat{\theta}_k - \theta_0)' F(\hat{\theta}_k, \mathbf{y})(\hat{\theta}_k - \theta_0)\right\} + o(1) \\ &= E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\} + \frac{k}{2} + \frac{k}{2} + k + o(1) \\ &= E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\} + 2k + o(1). \end{aligned}$$

This concludes the proof.

APPENDIX II

Using (17)–(19), the expectation of Geometric mean and $D_{3n}(\theta_0,\hat{\theta}_k)$ can be approximated as follows:

$$\begin{aligned}
&E_{\theta_0}\left\{D_{3n}(\theta_0,\hat{\theta}_k)\right\}\\
&\quad= E_{\theta_0}\left\{2G_n(\theta_0,\hat{\theta}_k)\right\}+d_n(\theta_0,\theta_0)\\
&\quad= E_{\theta_0}\left\{\sqrt{2I_n(\theta_0,\hat{\theta}_k)2I_n(\hat{\theta}_k,\theta_0)}\right\}+d_n(\theta_0,\theta_0)\\
&\quad= E_{\theta_0}\left\{\sqrt{\left[(\hat{\theta}_k-\theta_0)'I(\theta_0)(\hat{\theta}_k-\theta_0)+\varepsilon_2\right]}\right.\\
&\qquad\left.\cdot\sqrt{\left[(\hat{\theta}_k-\theta_0)'I(\hat{\theta}_k)(\hat{\theta}_k-\theta_0)+\varepsilon_3\right]}\right\}\\
&\qquad+E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\}\\
&\qquad+E_{\theta_0}\left\{(\hat{\theta}_k-\theta_0)'F(\hat{\theta}_k,\mathbf{y})(\hat{\theta}_k-\theta_0)\right\}+o(1)
\end{aligned}$$

A first-order Taylor expansion of $I(\hat{\theta}_k)$ around θ_0 (and assuming that $I(.)$ is sufficiently smooth around θ_0) gives

$$\begin{aligned}
I(\hat{\theta}_k)&\simeq I(\theta_0)+\frac{\partial I(\theta_0)}{\partial\theta}(\hat{\theta}_k-\theta_0)'\\
&= I(\theta_0)+I'(\theta_0)(\hat{\theta}_k-\theta_0)'. \qquad (20)
\end{aligned}$$

Using this and the approximation $(1+x)^{1/2}\simeq 1+x/2$ for small x, we have

$$\begin{aligned}
&\sqrt{(\hat{\theta}_k-\theta_0)'I(\theta_0)(\hat{\theta}_k-\theta_0).(\hat{\theta}_k-\theta_0)'I(\hat{\theta}_k)(\hat{\theta}_k-\theta_0)}\\
&\quad\simeq\left\{(\hat{\theta}_k-\theta_0)'I(\theta_0)(\hat{\theta}_k-\theta_0).(\hat{\theta}_k-\theta_0)'\right.\\
&\qquad\left.\times\left\{I(\theta_0)+I'(\theta_0)(\hat{\theta}_k-\theta_0)'\right\}(\hat{\theta}_k-\theta_0)\right\}^{1/2}\\
&\quad\simeq(\hat{\theta}_k-\theta_0)'I(\theta_0)(\hat{\theta}_k-\theta_0).\sqrt{1+\varepsilon_4}\\
&\quad\simeq(\hat{\theta}_k-\theta_0)'I(\theta_0)(\hat{\theta}_k-\theta_0)+\varepsilon_5,
\end{aligned}$$

where ε_4 and ε_5 are $O_p(n^{-1/2})$. This allows us to write

$$\begin{aligned}
E_{\theta_0}\left\{D_{3n}(\theta_0,\hat{\theta}_k)\right\}&=E_{\theta_0}\left\{2G_n(\theta_0,\hat{\theta}_k)\right\}+d_n(\theta_0,\theta_0)\\
&\simeq E_{\theta_0}\left\{(\hat{\theta}_k-\theta_0)'I(\theta_0)(\hat{\theta}_k-\theta_0)\right\}\\
&\quad+O(n^{-1/2})+E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\}\\
&\quad+E_{\theta_0}\left\{(\hat{\theta}_k-\theta_0)'F(\hat{\theta}_k,\mathbf{y})(\hat{\theta}_k-\theta_0)\right\}\\
&=E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\}+2k+O(n^{-1/2}).
\end{aligned}$$

This concludes the proof.

APPENDIX III

Using (17)–(19), the expectation of harmonic mean and $D_{4n}(\theta_0,\hat{\theta}_k)$ can be approximated as follows:

$$\begin{aligned}
&E_{\theta_0}\left\{D_{4n}(\theta_0,\hat{\theta}_k)\right\}\\
&\quad= E_{\theta_0}\left\{2H_n(\theta_0,\hat{\theta}_k)\right\}+d_n(\theta_0,\theta_0)\\
&\quad= E_{\theta_0}\left\{\frac{2}{\frac{1}{2I_n(\theta_0,\hat{\theta}_k)}+\frac{1}{2I_n(\hat{\theta}_k,\theta_0)}}\right\}+d_n(\theta_0,\theta_0)\\
&\quad\simeq E_{\theta_0}\left\{\frac{2}{\frac{1}{a}+\frac{1}{b}}\right\}+E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\}\\
&\qquad+E_{\theta_0}\left\{(\hat{\theta}_k-\theta_0)'F(\hat{\theta}_k,\mathbf{y})(\hat{\theta}_k-\theta_0)\right\}+o(1).
\end{aligned}$$

Using (20) and the approximation $(1+x)^{-1}\simeq 1-x$ for small x, we have

$$\begin{aligned}
E_{\theta_0}\left\{\frac{2}{\frac{1}{a}+\frac{1}{b}}\right\}&\simeq E_{\theta_0}\left\{\frac{2}{\frac{1}{a}+\frac{1}{c}}\right\}\\
&\simeq k+O(n^{-1/2})
\end{aligned}$$

where $a=(\hat{\theta}_k-\theta_0)'I(\theta_0)(\hat{\theta}_k-\theta_0)+\varepsilon_2$, $b=(\hat{\theta}_k-\theta_0)'I(\hat{\theta}_k)(\hat{\theta}_k-\theta_0)+\varepsilon_3$ and $c=(\hat{\theta}_k-\theta_0)'\{I(\theta_0)+I'(\theta_0)(\hat{\theta}_k-\theta_0)'\}(\hat{\theta}_k-\theta_0)+\varepsilon_3$.

Then

$$\begin{aligned}
E_{\theta_0}\left\{D_{4n}(\theta_0,\hat{\theta}_k)\right\}&=E_{\theta_0}\left\{2H_n(\theta_0,\hat{\theta}_k)\right\}+d_n(\theta_0,\theta_0)\\
&\simeq k+E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\}\\
&\quad+k+O(n^{-1/2})\\
&=E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\}+2k+O(n^{-1/2}).
\end{aligned}$$

This concludes the proof.

APPENDIX IV

From the model candidate (14), we have

$$\begin{aligned}
d_n(\theta_i,\theta_j)&=E_{\theta_i}\left\{-2\ln p(\mathbf{y}|\theta_j)\right\}\\
&=n\ln(2\pi)+n\ln\sigma_j^2+n\frac{\sigma_i^2}{\sigma_j^2}\\
&\quad+\frac{1}{\sigma_j^2}(\beta_i-\beta_j)^TX^TX(\beta_i-\beta_j).
\end{aligned}$$

Using this expression in (6) leads to

$$\begin{aligned}
D_{2n}(\theta_0,\hat{\theta}_k)&=\frac{1}{2}\left[d(\theta_0,\hat{\theta}_k)+d(\theta_0,\theta_0)\right.\\
&\qquad\left.+d(\hat{\theta}_k,\theta_0)-d(\hat{\theta}_k,\hat{\theta}_k)\right]\\
&=\frac{1}{2}\left[2n\ln\sigma_k^2+2n\ln(2\pi)+n\frac{\hat{\sigma}_k^2}{\sigma_0^2}+n\frac{\sigma_0^2}{\hat{\sigma}_k^2}\right.\\
&\qquad+\frac{1}{\sigma_0^2}(\hat{\beta}_k-\beta_0)^TX^TX(\hat{\beta}_k-\beta_0)\\
&\qquad+\frac{1}{\hat{\sigma}_k^2}(\hat{\beta}_k-\beta_0)^TX^TX(\hat{\beta}_k-\beta_0)\\
&\qquad\left.-2n\ln\left(\frac{\hat{\sigma}_k^2}{\sigma_0^2}\right)\right]. \qquad (21)
\end{aligned}$$

We have the following results (see [13] for details):

$$E_{\theta_0}\left\{n\frac{\hat{\sigma}_k^2}{\sigma_0^2}\right\} = n-k, \quad E_{\theta_0}\left\{n\frac{\sigma_0^2}{\hat{\sigma}_k^2}\right\} = \frac{n^2}{n-k-2}$$
$$E_{\theta_0}\left\{-2n\ln\left(\frac{\hat{\sigma}_k^2}{\sigma_0^2}\right)\right\} = -2n\psi\left(\frac{n-k}{2}\right) + 2n\ln\frac{n}{2}$$
$$E_{\theta_0}\left\{\frac{1}{\sigma_0^2}(\hat{\beta}_k-\beta_0)^T X^T X(\hat{\beta}_k-\beta_0)\right\} = k$$

and

$$E_{\theta_0}\left\{\frac{1}{\hat{\sigma}_k^2}(\hat{\beta}_k-\beta_0)^T X^T X(\hat{\beta}_k-\beta_0)\right\} = \frac{nk}{n-k-2}.$$

Substituting these results in (21), we obtain

$$\begin{aligned} E_{\theta_0}\left\{D_{2n}(\theta_0,\hat{\theta}_k)\right\} &= E_{\theta_0}\left\{n\ln\sigma_k^2 + n\ln(2\pi) + n\right\} \\ &\quad + \frac{n(k+1)}{n-k-2} - n\psi\left(\frac{n-k}{2}\right) \\ &\quad + n\ln\left(\frac{n}{2}\right) \\ &= E_{\theta_0}\left\{-2\ln p(\mathbf{y}|\hat{\theta}_k)\right\} + \frac{n(k+1)}{n-k-2} \\ &\quad - n\psi\left(\frac{n-k}{2}\right) + n\ln\left(\frac{n}{2}\right) \end{aligned}$$

This concludes the proof.

REFERENCES

[1] N. Murata, S. Yoshizawa, and S. I. Amari, "Network information criterion-determining the number of hidden units for an artificial neural network model," *IEEE Trans. Neural Netw.*, vol. 5, no. 6, pp. 865–872, Nov. 1994.

[2] K. Tsuda, M. Sugiyama, and K. R. Muller, "Subspace information criterion for nonquadratic regularizers-model selection regressors," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 70–80, Jan. 2002.

[3] J. I. Arribas and J. Cid-Sueiro, "A model selection algorithm for a posteriori probability estimation with neural networks," *IEEE Trans. Neural Netw.*, vol. 16, no. 4, pp. 799–809, Jul. 2005.

[4] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[5] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[6] H. Akaike, "Time series analysis and control through parametric models," in *Applied Time Series Analysis*, D. F. Findley, Ed. New York: Academic, 1978, pp. 1–23.

[7] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.

[8] H. Akaike, "Statistical predictor identification," *Ann. Inst. Stat. Math.*, vol. 22, pp. 203–217, 1970.

[9] C. L. Mallows, "Some comments on Cp," *Technometrics*, vol. 15, pp. 661–675, 1975.

[10] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Stat. Soc.*, ser. B, vol. 41, pp. 190–195, 1979.

[11] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 76–86, 1951.

[12] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[13] A. K. Seghouane and M. Bekara, "A small sample model selection criterion based on the Kullback symmetric divergence," *IEEE Trans. Signal Process.*, vol. 52, no. 12, pp. 3314–3323, Dec. 2004.

[14] H. Jeffreys, "An invariant form of the prior probability in estimation problems," *J. Roy. Stat. Soc.*, vol. A, pp. 453–469, 1946.

[15] J. E. Cavanaugh, "A large-sample model selection criterion based on Kullback's symmetric divergence," *Stat. Probab. Lett.*, vol. 42, pp. 333–343, 1999.

[16] C. M. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, pp. 297–307, 1989.

[17] S. M. Ali and S. D. Silvey, "A class of coefficients of divergence of one distribution from another," *J. Roy. Stat. Soc.*, ser. B, vol. 28, pp. 131–142, 1966.

[18] S. Amari and H. Nagoaka, *Methods of Information Geometry*. London, U.K.: Oxford Univ. Press, 1993.

[19] C. M. Hurvich and C. L. Tsai, "A corrected Akaike information criterion for vector autoregressive model selection," *J. Time Ser. Anal.*, vol. 14, pp. 271–279, 1993.

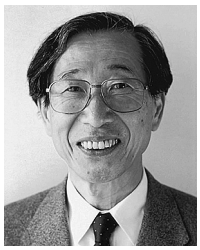
[20] E. J. Bedrik and C. L. Tsai, "Model selection for multivariate regression in small smaples," *Biometrics*, vol. 50, pp. 226–231, 1994.

[21] A. K. Seghouane, "Autoregressive model order selection from small samples using Kullback's symmetric divergence," *IEEE Trans. Circuits Syst. I, Reg. Papers*, 2006, to be published.

[22] ——, "Multivariate regression model selection from small samples using Kullback's symmetric divergence," *Signal Process.*, vol. 86, pp. 2074–2084, 2006.

Abd-Krim Seghouane (M'05) was born in Azazga, Algeria, in 1973. He received the engineer degree in electronics from Institut d'Electronique de l'Université Mouloud Mammeri, Tizi-ouzou, Algeria, in 1996, the Magistère degree in signal processing from Ecole Militaire Polytechnique, Bordj El Bahri, Algeria, in 2000, and the Ph.D degree in control and signal processing from Université de Paris Sud, Orsay, France, in 2002.

From 2003 to 2004, he was a Posdoctoral Researcher at INRIA Rocquencourt, Le Chesnay, France. In 2004, he joined National ICT Australia (NICTA), Canberra, Australia, where he is currently a Senior Researcher within the Systems Engineering and Complex Systems Research Program, Canberra Research Laboratory. His research interests are within statistical signal and image processing.



Shun-Ichi Amari (M'71–SM'92–F'94–LF'06) was born in Tokyo, Japan, on January 3, 1936. He received the Dr. Eng. degree in mathematical engineering from the Graduate School, University of Tokyo, Tokyo, Japan, in 1963.

He worked as an Associate Professor at Kyushu University and the University of Tokyo, and then a Full Professor at the University of Tokyo. He is now Professor-Emeritus. He serves now as Director of RIKEN Brain Science Institute, Saitama, Japan. He has been engaged in research in wide areas of mathematical engineering, such as topological network theory, differential geometry of continuum mechanics, pattern recognition, and information sciences. In particular, he has devoted himself to mathematical foundations of neural networks, including statistical neurodynamics, dynamical theory of neural fields, associative memory, self-organization, and general learning theory. Another main subject of his research is information geometry initiated by himself, which applies modern differential geometry to statistical inference, information theory, control theory, stochastic reasoning, and neural networks, providing a new powerful method of information sciences and probability theory.

Dr. Amari is the Past President of International Neural Networks Society. He received the Emanuel A. Piore Award and the Neural Networks Pioneer Award from the IEEE, the Japan Academy Award, and the C&C Award, among many others. He has served as a founding Co-Editor-in-Chief of *Neural Networks*, among many other journals.