Ann Inst Stat Math (2018) 70:421–438
https://doi.org/10.1007/s10463-016-0592-7

An information criterion for model selection with missing data via complete-data divergence

Hidetoshi Shimodaira[1,2] · **Haruyoshi Maeda**[1,3]

Received: 26 September 2015 / Revised: 4 November 2016 / Published online: 21 January 2017

Abstract We derive an information criterion to select a parametric model of complete-data distribution when only incomplete or partially observed data are available. Compared with AIC, our new criterion has an additional penalty term for missing data, which is expressed by the Fisher information matrices of complete data and incomplete data. We prove that our criterion is an asymptotically unbiased estimator of complete-data divergence, namely the expected Kullback–Leibler divergence between the true distribution and the estimated distribution for complete data, whereas AIC is that for the incomplete data. The additional penalty term of our criterion for missing data turns out to be only half the value of that in previously proposed information criteria PDIO and AICcd. The difference in the penalty term is attributed to the fact that our criterion is derived under a weaker assumption. A simulation study with the weaker assumption shows that our criterion is unbiased while the other two criteria are biased. In addition, we review the geometrical view of alternating minimizations of the EM algorithm. This geometrical view plays an important role in deriving our new criterion.

The research was supported in part by JSPS KAKENHI Grant (24300106, 16H02789).

✉ Hidetoshi Shimodaira
shimo@sigmath.es.osaka-u.ac.jp

[1] Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, Japan

[2] RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

[3] Present Address: Kawasaki Heavy Industries, Ltd., 1-1 Kawasaki-cho, Akashi, Hyogo 673-8666, Japan

Keywords Akaike information criterion · Alternating projections · Data manifold · EM algorithm · Fisher information matrix · Incomplete data · Kullback–Leibler divergence · Misspecification · Takeuchi information criterion

1 Introduction

Modeling complete data $X = (Y, Z)$ is often preferable to modeling incomplete or partially observed data Y when missing data Z are not observed. The expectation-maximization (EM) algorithm (Dempster et al. 1977) computes the maximum likelihood estimate of parameter vector $\boldsymbol{\theta}$ for a parametric model of the probability distribution of X. In this research, we consider the problem of model selection in such situations. For mathematical simplicity, we assume that X consists of independent and identically distributed random vectors. More specifically, $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, and the complete-data distribution is modeled as $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \sim p_x(\boldsymbol{x}; \boldsymbol{\theta})$. Each vector is decomposed as $\boldsymbol{x}^T = (\boldsymbol{y}^T, \boldsymbol{z}^T)$, and the marginal distribution is expressed as $p_y(\boldsymbol{y}; \boldsymbol{\theta}) = \int p_x(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})\, d\boldsymbol{z}$, where T denotes the matrix transpose and the integration is over all possible values of $\boldsymbol{z}$. We formally treat $\boldsymbol{y}, \boldsymbol{z}$ as continuous random variables with the joint density function p_x. However, when they are discrete random variables, the integration should be replaced with a summation of the probability functions. We use symbols such as p_x and p_y for both the continuous and discrete cases, and simply refer to them as distributions.

The log-likelihood function is $\ell_y(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log p_y(\boldsymbol{y}_t; \boldsymbol{\theta})$ with the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T \in \mathbb{R}^d$. We assume that the model is identifiable, and the parameter is restricted to $\Theta \subset \mathbb{R}^d$. Then, the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is defined by $\hat{\boldsymbol{\theta}}_y = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell_y(\boldsymbol{\theta})$. The dependence of $\ell_y(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_y$ on $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ is suppressed in the notation. Akaike (1974) proposed the information criterion

$$\mathrm{AIC} = -2\ell_y(\hat{\boldsymbol{\theta}}_y) + 2d$$

for model selection. The first term measures the goodness of fit, whereas the second term is interpreted as a penalty for model complexity. The AIC values for candidate models are computed, and then the model that minimizes AIC is selected. This information criterion estimates the expected discrepancy between the unknown true distribution of $\boldsymbol{y}$, which is denoted as q_y, and the estimated distribution $p_y(\hat{\boldsymbol{\theta}}_y)$. This discrepancy is measured by the incomplete-data Kullback–Leibler divergence.

In this study, we work on the complete-data Kullback–Leibler divergence instead of the incomplete-data counterpart. An information criterion to estimate the expected discrepancy between the unknown true distribution of $\boldsymbol{x}$, which is denoted as q_x, and the estimated distribution $p_x(\hat{\boldsymbol{\theta}}_y)$ is derived. This approach makes sense when modeling complete data more precisely describes the part being examined. Similar attempts are found in the literature. Shimodaira (1994) proposed the information criterion PDIO (predictive divergence for incomplete observation models)

$$\mathrm{PDIO} = -2\ell_y(\hat{\boldsymbol{\theta}}_y) + 2\mathrm{tr}(I_x(\hat{\boldsymbol{\theta}}_y)I_y(\hat{\boldsymbol{\theta}}_y)^{-1}).$$

The two matrices in the penalty term are the Fisher information matrices for complete data and incomplete data. They are defined by

$$I_x(\boldsymbol{\theta}) = -\int p_x(\boldsymbol{x};\boldsymbol{\theta})\frac{\partial^2 \log p_x(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\,\mathrm{d}\boldsymbol{x},$$
$$I_y(\boldsymbol{\theta}) = -\int p_y(\boldsymbol{y};\boldsymbol{\theta})\frac{\partial^2 \log p_y(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\,\mathrm{d}\boldsymbol{y}.$$

Let $p_{z|y}(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}) = p_x(\boldsymbol{y},\boldsymbol{z};\boldsymbol{\theta})/p_y(\boldsymbol{y};\boldsymbol{\theta})$ be the conditional distribution of $\boldsymbol{z}$ given $\boldsymbol{y}$, and $I_{z|y}(\boldsymbol{\theta}) = I_x(\boldsymbol{\theta}) - I_y(\boldsymbol{\theta})$ be the Fisher information matrix for $p_{z|y}$. Since $I_{z|y}(\boldsymbol{\theta})$ is nonnegative definite, we have $\mathrm{tr}(I_x(\boldsymbol{\theta})I_y(\boldsymbol{\theta})^{-1}) = \mathrm{tr}((I_y(\boldsymbol{\theta}) + I_{z|y}(\boldsymbol{\theta}))I_y(\boldsymbol{\theta})^{-1}) = d + \mathrm{tr}(I_{z|y}(\boldsymbol{\theta})I_y(\boldsymbol{\theta})^{-1}) \geq d$. Thus, the nonnegative difference

$$\mathrm{PDIO} - \mathrm{AIC} = 2\mathrm{tr}(I_{z|y}(\hat{\boldsymbol{\theta}}_y)I_y(\hat{\boldsymbol{\theta}}_y)^{-1})$$

is interpreted as the additional penalty for missing data. There are similar attempts in the literature (Cavanaugh and Shumway 1998; Seghouane et al. 2005; Claeskens and Consentino 2008; Yamazaki 2014). In particular, Cavanaugh and Shumway (1998) proposed another information criterion

$$\mathrm{AIC}_{cd} = -2Q(\hat{\boldsymbol{\theta}}_y;\hat{\boldsymbol{\theta}}_y) + 2\mathrm{tr}(I_x(\hat{\boldsymbol{\theta}}_y)I_y(\hat{\boldsymbol{\theta}}_y)^{-1})$$

by replacing $\ell_y(\hat{\boldsymbol{\theta}}_y)$ in PDIO with $Q(\hat{\boldsymbol{\theta}}_y;\hat{\boldsymbol{\theta}}_y)$ to measure the goodness of fit. It should be noted that cd stands for complete data. This is the function introduced in Dempster et al. (1977) for the EM algorithm, and is defined by

$$Q(\boldsymbol{\theta}_2;\boldsymbol{\theta}_1) = \sum_{t=1}^{n}\int p_{z|y}(\boldsymbol{z}|\boldsymbol{y}_t;\boldsymbol{\theta}_1)\log p_x(\boldsymbol{y}_t,\boldsymbol{z};\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{z}.$$

We recently found that the assumption in Shimodaira (1994) to derive PDIO is unnecessarily strong. Additionally, the same assumption explains the derivation of AIC_{cd}. In this paper, we derive a new information criterion under a weaker assumption. The updated version of PDIO is

$$\mathrm{AIC}_{x;y} = -2\ell_y(\hat{\boldsymbol{\theta}}_y) + d + \mathrm{tr}(I_x(\hat{\boldsymbol{\theta}}_y)I_y(\hat{\boldsymbol{\theta}}_y)^{-1}).$$

The first suffix $\boldsymbol{x}$ indicates that a random variable is used to measure the discrepancy, while the second suffix $\boldsymbol{y}$ indicates a random variable is used for the observation. Then, the additional penalty for missing data becomes

$$\mathrm{AIC}_{x;y} - \mathrm{AIC} = \mathrm{tr}(I_{z|y}(\hat{\boldsymbol{\theta}}_y)I_y(\hat{\boldsymbol{\theta}}_y)^{-1}). \tag{1}$$

The additional penalty is only half the value of that in PDIO. In practice, the computation of $\mathrm{AIC}_{x;y}$ as well as the related criteria PDIO and AIC_{cd} is not very difficult. The SEM algorithm of Meng and Rubin (1991) provides a shortcut to compute the penalty term $\mathrm{tr}(I_x(\hat{\boldsymbol{\theta}}_y)I_y(\hat{\boldsymbol{\theta}}_y)^{-1})$ without computing the two Fisher information matrices as described in Shimodaira (1994) and Cavanaugh and Shumway (1998).

To derive $\mathrm{AIC}_{x;y}$, we first review the basic properties of Kullback–Leibler divergence for incomplete data in Sect. 2. Section 3 considers those for complete data. Although these results are not new, they are crucial for the argument in later sections. In particular, the geometrical view of alternating minimizations (Csiszár and Tusnády, 1984; Amari, 1995) in Sect. 3.3 is important to understand why the goodness of fit term of $\mathrm{AIC}_{x;y}$ is expressed by the incomplete-data likelihood function instead of the complete-data counterpart.

Section 4, which begins the argument of model selection, discusses what the information criteria should estimate. In general, parametric models are misspecified, and we do *not* assume that the true distribution is expressed as $q_x = p_x(\boldsymbol{\theta}_0)$ using the "true" parameter value $\boldsymbol{\theta}_0$. However, the unbiasedness of $\mathrm{AIC}_{x;y}$ is based on the assumption that $p_{z|y}(\boldsymbol{\theta})$ is correctly specified for $q_{z|y}$. In Sect. 5, we derive our new information criterion. The argument is very straightforward; it simply follows the argument for the robust version of AIC, which is also known as the Takeuchi information criterion (TIC) that is described in Burnham and Anderson (2002) and Konishi and Kitagawa (2008). Section 6 compares the assumptions used to derive PDIO and AIC_{cd} to those of $\mathrm{AIC}_{x;y}$. Section 7 presents a simulation study to verify the theory. Finally, Sect. 8 contains some concluding remarks. Proofs are deferred to the Appendix.

2 Incomplete-data divergence

Here, we review Kullback–Leibler divergence and the asymptotic distribution of MLE under model misspecification (White 1982). Let g_y and f_y be the arbitrary probability distributions of incomplete data. The incomplete-data Kullback–Leibler divergence from g_y to f_y is

$$D_y(g_y;\, f_y) = -\int g_y(\boldsymbol{y})(\log f_y(\boldsymbol{y}) - \log g_y(\boldsymbol{y}))\,\mathrm{d}\boldsymbol{y},$$

where $D_y(g_y;\, f_y) \geq 0$ and the equality holds for $g_y = f_y$ (Csiszár 1975; Amari and Nagaoka 2007). The cross-entropy is

$$L_y(g_y;\, f_y) = -\int g_y(\boldsymbol{y}) \log f_y(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$$

and the entropy is $L_y(g_y) = L_y(g_y;\, g_y)$. Instead of minimizing $D_y(g_y;\, f_y) = L_y(g_y;\, f_y) - L_y(g_y)$ with respect to f_y, we minimize $L_y(g_y;\, f_y)$, because $L_y(g_y)$ is independent of f_y.

For the true distribution q_y and the parametric model $p_y(\boldsymbol{\theta})$, we consider the minimization of $D_y(q_y;\, p_y(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$. The optimal parameter value is defined by

$$\bar{\boldsymbol{\theta}}_y = \arg\min_{\boldsymbol{\theta}\in\Theta} L_y(q_y;\, p_y(\boldsymbol{\theta})).$$

This minimization is interpreted geometrically as a "projection" of q_y to the model manifold $M_y(p_y)$ as illustrated in Fig. 1a. Let $M_y(p_y) = \{p_y(\boldsymbol{\theta}) : \forall\boldsymbol{\theta} \in \Theta\}$ be the set of $p_y(\boldsymbol{\theta})$ with all possible parameter values. Then, the projection is defined as:

$$\min_{f_y\in M_y(p_y)} D_y(q_y;\, f_y) = D_y(q_y;\, p_y(\bar{\boldsymbol{\theta}}_y)). \tag{2}$$

The projection $p_y(\bar{\boldsymbol{\theta}}_y)$ is the best approximation of q_y in $M_y(p_y)$ when the discrepancy is measured by the Kullback–Leibler divergence. We assume that the parametric model is generally misspecified and $q_y \notin M_y(p_y)$. Later, we also consider the situation where the parametric model is correctly specified and $q_y \in M_y(p_y)$. In the correctly specified case, $\bar{\boldsymbol{\theta}}_y$ is the true parameter value in the sense that $q_y = p_y(\bar{\boldsymbol{\theta}}_y)$.

Similar to the optimal parameter value, the maximum likelihood estimator is interpreted as a projection of $\hat{q}_y$ to $M_y(p_y)$. Let $\hat{q}_y(\boldsymbol{y}) = \frac{1}{n}\sum_{t=1}^{n}\delta(\boldsymbol{y}-\boldsymbol{y}_t)$ be the empirical distribution of $\boldsymbol{y}$ for the observed incomplete data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. Here, $\delta(\cdot)$ denotes the Dirac delta function for continuous random variables, or is simply the indicator function for discrete random variables such that $\delta(\boldsymbol{y}-\boldsymbol{y}_t) = 1$ for $\boldsymbol{y} = \boldsymbol{y}_t$ and $\delta(\boldsymbol{y}-\boldsymbol{y}_t) = 0$ otherwise. Then, we can write $\ell_y(\boldsymbol{\theta}) = -nL_y(\hat{q}_y;\, p_y(\boldsymbol{\theta}))$. Thus,

$$\hat{\boldsymbol{\theta}}_y = \arg\min_{\boldsymbol{\theta}\in\Theta} L_y(\hat{q}_y;\, p_y(\boldsymbol{\theta})). \tag{3}$$

We assume the regularity conditions of White (1982) for consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_y$. More specifically, we assume all the regularity conditions (A1) to (A6) for the true distribution q_y and the model distribution $p_y(\boldsymbol{\theta})$. In particular, $\bar{\boldsymbol{\theta}}_y$ is determined uniquely (i.e., identifiable) and is interior to the parameter space Θ. We assume that $I_y(\boldsymbol{\theta})$, $G_y(q_y;\boldsymbol{\theta})$ and $H_y(q_y;\boldsymbol{\theta})$ defined below are nonsingular in the neighborhood of $\bar{\boldsymbol{\theta}}_y$. Then, White (1982) showed that, as $n \to \infty$ asymptotically, $\hat{\boldsymbol{\theta}}_y \overset{a.s.}{\to} \bar{\boldsymbol{\theta}}_y$ and

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_y - \bar{\boldsymbol{\theta}}_y) \overset{d}{\to} N(\mathbf{0},\, H_y^{-1}G_yH_y^{-1}). \tag{4}$$

The matrices are defined as $G_y = G_y(q_y;\bar{\boldsymbol{\theta}}_y)$ and $H_y = H_y(q_y;\bar{\boldsymbol{\theta}}_y)$, where

$$G_y(g_y;\boldsymbol{\theta}) = \int g_y(\boldsymbol{y})\frac{\partial \log p_y(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \log p_y(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\,\mathrm{d}\boldsymbol{y},$$
$$H_y(g_y;\boldsymbol{\theta}) = -\int g_y(\boldsymbol{y})\frac{\partial^2 \log p_y(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\,\mathrm{d}\boldsymbol{y}.$$

In the case of the correct specification $q_y = p_y(\bar{\boldsymbol{\theta}}_y)$, the matrices become $G_y = H_y = I_y(\bar{\boldsymbol{\theta}}_y)$.

3 Complete-data divergence

Here, we review Kullback–Leibler divergence for complete data when only incomplete data can be observed (Csiszár and Tusnády 1984; Amari 1995).

3.1 Projection to the model manifold

Let g_x and f_x be the arbitrary probability distributions of complete data. The complete-data Kullback–Leibler divergence from g_x to f_x is

$$D_x(g_x;\, f_x) = -\int g_x(\boldsymbol{x})(\log f_x(\boldsymbol{x}) - \log g_x(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}.$$

All the arguments of incomplete data in Sect. 2 apply to complete data by replacing $\boldsymbol{y}$ with $\boldsymbol{x}$ in the notation. For example, we write $D_x(g_x;\, f_x) = L_x(g_x;\, f_x) - L_x(g_x)$ with $L_x(g_x;\, f_x) = -\int g_x(\boldsymbol{x}) \log f_x(\boldsymbol{x})\, d\boldsymbol{x}$ and $L_x(g_x) = L_x(g_x;\, g_x)$. The projection of q_x to the model manifold $M_x(p_x) = \{p_x(\boldsymbol{\theta}) : \forall \boldsymbol{\theta} \in \Theta\}$ is defined as:

$$\min_{f_x \in M_x(p_x)} D_x(q_x;\, f_x) = D_x(q_x;\, p_x(\bar{\boldsymbol{\theta}}_x)) \tag{5}$$

with $\bar{\boldsymbol{\theta}}_x = \arg\min_{\boldsymbol{\theta}} L_x(q_x;\, p_x(\boldsymbol{\theta}))$. Figure 1b shows a geometric illustration. Note that $\bar{\boldsymbol{\theta}}_x \neq \bar{\boldsymbol{\theta}}_y$ and $p_x(\bar{\boldsymbol{\theta}}_x) \neq p_x(\bar{\boldsymbol{\theta}}_y)$ in general.

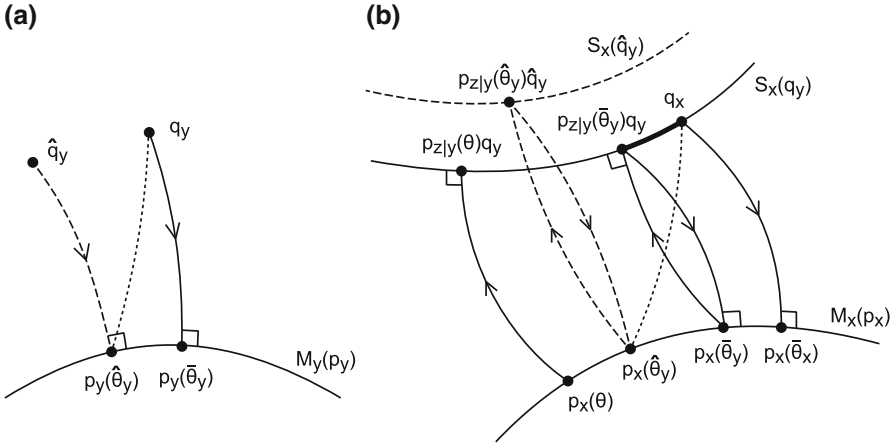

Fig. 1 **a** Space of incomplete-data probability distributions. Projection from q_y to the model manifold $M_y(p_y)$ (*arrow* with a *solid line*), and that from $\hat{q}_y$ (*arrow* with a *broken line*) using Eqs. (2) and (3) in Sect. 2, respectively. The *dotted line* indicates $D_y(q_y;\, p_y(\hat{\boldsymbol{\theta}}_y))$, which is the loss function for $\mathrm{risk}_{y;y}$. **b** Space of complete-data probability distributions. Projection from q_x to the model manifold $M_x(p_x)$ using Eq. (5) in Sect. 3.1. Projection from $p_x(\boldsymbol{\theta})$ to the data manifold $S_x(q_y)$ using Eq. (9) in Sect. 3.2. Alternating projections between the two manifolds using Eq. (10) in Sect. 3.3. The *dotted line* indicates $D_x(q_x;\, p_x(\hat{\boldsymbol{\theta}}_y))$, which is the loss function for $\mathrm{risk}_{x;y}$. The *bold* segment indicates $D_x(q_x;\, p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y)$, which is assumed to be zero in (15).

3.2 Projection to the data manifold

The following simple lemma helps understand how the incomplete-data divergence and the complete-data divergence are related.

Lemma 1 *For two distributions $g_x(\boldsymbol{x})$ and $f_x(\boldsymbol{x})$, we have*

$$D_x(g_x;\, f_x) = D_x(g_x;\, f_{z|y}g_y) + D_y(g_y;\, f_y), \tag{6}$$

where $f_{z|y}g_y$ represents the distribution $f_{z|y}(\boldsymbol{z}|\boldsymbol{y})g_y(\boldsymbol{y})$. Therefore, the difference of the two divergences is $D_x(g_x;\, f_x) - D_y(g_y;\, f_y) = D_x(g_x;\, f_{z|y}g_y)$, which is zero if $g_{z|y} = f_{z|y}$. For an arbitrary distribution $h_x(\boldsymbol{x})$, the last term in (6) is expressed as:

$$D_y(g_y;\, f_y) = D_x(h_{z|y}g_y;\, h_{z|y}f_y). \tag{7}$$

In particular, choosing $h_x = f_x$ gives $D_y(g_y;\, f_y) = D_x(f_{z|y}g_y;\, f_x)$, and

$$D_x(g_x;\, f_x) = D_x(g_x;\, f_{z|y}g_y) + D_x(f_{z|y}g_y;\, f_x). \tag{8}$$

We consider the set of all probability distributions g_x with the same marginal distribution $g_y = q_y$ for a specified q_y. This set is denoted as $S_x(q_y) = \{g_{z|y}q_y : \forall g_{z|y}\}$. Note that the elements of $S_x(q_y)$ are written as $g_{z|y}q_y$ with arbitrary $g_{z|y}$ because $\int g_{z|y}(\boldsymbol{z}|\boldsymbol{y})q_y(\boldsymbol{y})\,d\boldsymbol{z} = q_y(\boldsymbol{y})$. Equations (88) and (57) in Amari (1995) are $S_x(\hat{q}_y)$ and its restriction to a finite dimensional model, respectively, and are called the observed data (sub)manifold there. Here, we call $S_x(q_y)$ the expected data manifold and $S_x(\hat{q}_y)$ the observed data manifold, although it may be abuse of the word "manifold" for subsets with infinite dimensions.

The projection of $p_x(\boldsymbol{\theta})$ to $S_x(q_y)$ should be defined to minimize the complete-data divergence over $S_x(q_y)$, but the roles of g_x and f_x in $D_x(g_x;\, f_x)$ are exchanged from those of (5). We minimize $D_x(g_x;\, p_x(\boldsymbol{\theta}))$ over $g_x \in S_x(q_y)$. By letting $g_x \in S_x(q_y)$ and $f_x = p_x(\boldsymbol{\theta})$ in (6),

$$D_x(g_x;\, p_x(\boldsymbol{\theta})) = D_x(g_{z|y}q_y;\, p_{z|y}(\boldsymbol{\theta})q_y) + D_y(q_y;\, p_y(\boldsymbol{\theta})),$$

which is minimized when $g_{z|y} = p_{z|y}(\boldsymbol{\theta})$. Therefore, the projection gives the minimum value as:

$$\min_{g_x \in S_x(q_y)} D_x(g_x;\, p_x(\boldsymbol{\theta})) = D_y(q_y;\, p_y(\boldsymbol{\theta})). \tag{9}$$

Using (8), the minimum value can also be written as $D_y(q_y;\, p_y(\boldsymbol{\theta})) = D_x(p_{z|y}(\boldsymbol{\theta})q_y;\, p_x(\boldsymbol{\theta}))$.

3.3 Alternating projections between the two manifolds

The optimal parameter $\bar{\boldsymbol{\theta}}_y$ of the incomplete data is interpreted as a dual or alternate minimization problem of complete-data divergence. By minimizing (9) over $\boldsymbol{\theta} \in \Theta$, we define the alternating projections between $S_x(q_y)$ and $M_x(p_x)$ as:

$$\min_{f_x \in M_x(p_x)} \min_{g_x \in S_x(q_y)} D_x(g_x; f_x) = D_y(q_y; p_y(\bar{\boldsymbol{\theta}}_y)), \tag{10}$$

where the minimum is attained by $g_x = p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y$ and $f_x = p_x(\bar{\boldsymbol{\theta}}_y)$. See Eq. (65) in Amari (1995). This implies that $p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y$ is the best approximation of q_x when the two manifolds $S_x(q_y)$ and $M_x(p_x)$ are known, while $p_x(\bar{\boldsymbol{\theta}}_y)$ is the best approximation of q_x in $M_x(p_x)$. This interpretation is the key to understanding our problem.

The above-mentioned geometrical interpretation corresponds to the well-known fact that the EM algorithm of Dempster et al. (1977) is alternating projections between $S_x(\hat{q}_y)$ and $M_x(p_x)$. See Csiszár and Tusnády (1984), Byrne (1992), Amari (1995), and Ip and Lalwani (2000). Starting from the initial value $\boldsymbol{\theta}^{(1)}$, the EM algorithm computes a sequence of the parameter values $\{\boldsymbol{\theta}^{(s)}; s = 1, 2, \ldots\}$ by the updating formula $\boldsymbol{\theta}^{(s+1)} = \arg\max_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(s)})$. It follows from $L_x(p_{z|y}(\boldsymbol{\theta}_1)\hat{q}_y; p_x(\boldsymbol{\theta}_2)) = -Q(\boldsymbol{\theta}_2; \boldsymbol{\theta}_1)/n$ that

$$\boldsymbol{\theta}^{(s+1)} = \arg\min_{\boldsymbol{\theta} \in \Theta} L_x(p_{z|y}(\boldsymbol{\theta}^{(s)})\hat{q}_y; p_x(\boldsymbol{\theta})),$$

meaning $p_x(\boldsymbol{\theta}^{(s+1)})$ is the projection from $p_{z|y}(\boldsymbol{\theta}^{(s)})\hat{q}_y$ to $M_x(p_x)$. Alternatively, $p_{z|y}(\boldsymbol{\theta}^{(s)})\hat{q}_y$ is the projection from $p_x(\boldsymbol{\theta}^{(s)})$ to $S_x(\hat{q}_y)$. Thus, the converging point of the alternating projections satisfies

$$\hat{\boldsymbol{\theta}}_y = \arg\min_{\boldsymbol{\theta} \in \Theta} L_x(p_{z|y}(\hat{\boldsymbol{\theta}}_y)\hat{q}_y; p_x(\boldsymbol{\theta})). \tag{11}$$

4 Risk functions for model selection

By looking at the incomplete-data distributions, the discrepancy between the true distribution q_y and our estimation $p_y(\hat{\boldsymbol{\theta}}_y)$ is measured by the incomplete-data divergence $D_y(q_y; p_y(\hat{\boldsymbol{\theta}}_y))$. If we take it as the loss function, the expected loss-function, or the risk function, will measure the discrepancy in the long run. Then, AIC and its variants are derived as estimators of

$$\text{risk}_{y;y} = E\{D_y(q_y; p_y(\hat{\boldsymbol{\theta}}_y))\}. \tag{12}$$

The expectation is evaluated with respect to q_x, although it involves only q_y here. This is the standard approach in the literature (Akaike 1974; Bozdogan 1987; Burnham and Anderson 2002; Konishi and Kitagawa 2008).

Shimodaira (1994) and Cavanaugh and Shumway (1998) proposed another approach, which employs the complete-data divergence $D_x(q_x; p_x(\hat{\boldsymbol{\theta}}_y))$ to measure the discrepancy between the complete-data distributions q_x and $p_x(\hat{\boldsymbol{\theta}}_y)$. Using the complete-data divergence as the loss function, the risk function becomes

$$\text{risk}_{x;y} = E\{D_x(q_x; p_x(\hat{\boldsymbol{\theta}}_y))\}. \tag{13}$$

The first suffix $\boldsymbol{x}$ indicates the random variable for the loss function, while the second suffix $\boldsymbol{y}$ indicates the random variable for the observation.

However, estimating (13) is difficult. The complete-data empirical distribution $\hat{q}_x(\boldsymbol{x}) = \frac{1}{n}\sum_{t=1}^{n}\delta(\boldsymbol{x}-\boldsymbol{x}_t)$ is unknown; we only know that $\hat{q}_x$ is somewhere in the observed data manifold $S_x(\hat{q}_y)$. Considering the limiting situation of $n \to \infty$, we may only know that the true distribution is somewhere in the expected data manifold: $q_x \in S_x(q_y)$. Then, the best substitute for q_x is

$$q_x = p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y \tag{14}$$

as suggested by (10) from the viewpoint of the alternating projections in Sect. 3.3. To estimate (13), we assume that (14) holds in this paper. This assumption is rephrased as:

$$D_x(q_x;\, p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y) = 0$$

or equivalently

$$q_{z|y} = p_{z|y}(\bar{\boldsymbol{\theta}}_y), \tag{15}$$

implying that $p_{z|y}(\boldsymbol{\theta})$ is correctly specified for $q_{z|y}$ and that $\bar{\boldsymbol{\theta}}_x = \bar{\boldsymbol{\theta}}_y$, because the two projections from q_x and $p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y$ to $M_x(p_x)$ become identical as illustrated in Fig. 1b. Because it is impossible to know how much $q_{z|y}$ actually deviates from $p_{z|y}(\bar{\boldsymbol{\theta}}_y)$ when $Z = (z_1, \ldots, z_n)$ is missing *completely*, we assume (15) in the following argument to derive $\mathrm{AIC}_{x;y}$. Note that assumption (15) holds with $\bar{\boldsymbol{\theta}}_x = \bar{\boldsymbol{\theta}}_y = \boldsymbol{\theta}_0$ in the case of the correct specification where $q_x = p_x(\boldsymbol{\theta}_0)$.

We are now ready to derive $\mathrm{AIC}_{x;y}$ as an estimator of $2n\,\mathrm{risk}_{x;y}$. The arguments in Lemma 2 and Theorem 1 almost duplicate that used to derive TIC mentioned in Burnham and Anderson (2002) and Konishi and Kitagawa (2008). However, it should be noted that in Lemma 2 the first term of $\mathrm{risk}_{x;y}$ is expressed by the incomplete-data divergence instead of the complete-data divergence. A point for proving the lemma is that

$$D_x(q_x;\, p_x(\bar{\boldsymbol{\theta}}_y)) = D_x(q_x;\, p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y) + D_y(q_y;\, p_y(\bar{\boldsymbol{\theta}}_y)) = D_y(q_y;\, p_y(\bar{\boldsymbol{\theta}}_y)), \tag{16}$$

which follows from Lemma 1 and the assumption (15). $D_x(q_x;\, p_x(\bar{\boldsymbol{\theta}}_y))$ on the left-hand side is the amount of misspecification of $p_x(\boldsymbol{\theta})$, and can be decomposed into two parts: $D_x(q_x;\, p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y)$ and $D_y(q_y;\, p_y(\bar{\boldsymbol{\theta}}_y))$, which are the contribution of $p_{z|y}(\boldsymbol{\theta})$ and $p_y(\boldsymbol{\theta})$, respectively. To estimate (13), instead of estimating $D_x(q_x;\, p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y)$, we ignore it.

Lemma 2 *Assume the regularity conditions of* White (1982) *mentioned in Sect. 2, and also assume that* (15) *holds. Then, the expected loss is asymptotically expanded as:*

$$\mathrm{risk}_{x;y} = D_y(q_y;\, p_y(\bar{\boldsymbol{\theta}}_y)) + \frac{1}{2n}\mathrm{tr}(H_x H_y^{-1} G_y H_y^{-1}) + O(n^{-3/2}). \tag{17}$$

The matrices G_y and H_y are those defined in Sect. 2, and $H_x = H_x(p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y; \bar{\boldsymbol{\theta}}_y)$ with

$$H_x(g_x; \boldsymbol{\theta}) = -\int g_x(\boldsymbol{x}) \frac{\partial^2 \log p_x(\boldsymbol{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} d\boldsymbol{x}.$$

The dominant term in (17) is also expressed as $D_y(q_y; p_y(\bar{\boldsymbol{\theta}}_y)) = L_y(q_y; p_y(\bar{\boldsymbol{\theta}}_y)) - L_y(q_y)$ using the cross-entropy.

5 Information criteria

Let us define an information criterion as an estimator of $\mathrm{risk}_{x;y}$.

$$\widehat{\mathrm{risk}}_{x;y} = L_y(\hat{q}_y; p_y(\hat{\boldsymbol{\theta}}_y)) - L_y(q_y) + \frac{1}{2n}\mathrm{tr}(G_y H_y^{-1}) + \frac{1}{2n}\mathrm{tr}(H_x H_y^{-1} G_y H_y^{-1}), \tag{18}$$

where the matrices G_y, H_y and H_x may be replaced by their consistent estimators with error $O_p(n^{-1/2})$. When $\boldsymbol{x} \equiv \boldsymbol{y}$, (18) reduces to

$$\widehat{\mathrm{risk}}_{y;y} = L_y(\hat{q}_y; p_y(\hat{\boldsymbol{\theta}}_y)) - L_y(q_y) + \frac{1}{n}\mathrm{tr}(G_y H_y^{-1}), \tag{19}$$

which corresponds to the Takeuchi information criterion (TIC) for estimating $\mathrm{risk}_{y;y}$ mentioned in Burnham and Anderson (2002) and Konishi and Kitagawa (2008). In model selection, we ignore $L_y(q_y)$, because all candidate models have the same value. The first term $L_y(\hat{q}_y; p_y(\hat{\boldsymbol{\theta}}_y)) = -\ell_y(\hat{\boldsymbol{\theta}}_y)/n$ of order $O_p(1)$ measures the goodness of fit, while the last two terms of order $O(n^{-1})$ are interpreted as the penalty of model complexity. Our estimator is justified by the following theorem.

Theorem 1 *Assume the regularity conditions of* White (1982) *mentioned in Sect. 2, and also assume that (15) holds. Then, we have*

$$L_y(q_y; p_y(\bar{\boldsymbol{\theta}}_y)) = E\{L_y(\hat{q}_y; p_y(\hat{\boldsymbol{\theta}}_y))\} + \frac{1}{2n}\mathrm{tr}(G_y H_y^{-1}) + O(n^{-3/2}), \tag{20}$$

and therefore

$$E\{\widehat{\mathrm{risk}}_{x;y}\} = \mathrm{risk}_{x;y} + O(n^{-3/2}). \tag{21}$$

Thus, the estimator is unbiased asymptotically up to terms of order $O(n^{-1})$.

In the case of the correct specification where $q_y = p_y(\bar{\boldsymbol{\theta}}_y)$ for the incomplete-data distribution, we have $G_y = H_y = I_y(\bar{\boldsymbol{\theta}}_y)$, and the information matrix is consistently estimated by $I_y(\hat{\boldsymbol{\theta}}_y)$. Assuming (15), this implies that $q_x = p_x(\bar{\boldsymbol{\theta}}_x)$ is correctly specified for the complete-data distribution. Hence, $H_x = I_x(\bar{\boldsymbol{\theta}}_y)$ is consistently estimated by $I_x(\hat{\boldsymbol{\theta}}_y)$. For model selection, we assume that $p_y(\boldsymbol{\theta})$ is misspecified for q_y in general. However, these equations may approximately hold if $p_y(\bar{\boldsymbol{\theta}}_y)$ is a good approximation

of q_y. By substituting $G_y \approx H_y \approx I_y(\bar{\boldsymbol{\theta}}_y)$ and $H_x \approx I_x(\bar{\boldsymbol{\theta}}_y)$ into (18) and (19), we have

$$\widehat{\text{risk}}_{x;y} \approx L_y(\hat{q}_y; p_y(\hat{\boldsymbol{\theta}}_y)) - L_y(q_y) + \frac{d}{2n} + \frac{1}{2n}\text{tr}(I_x(\hat{\boldsymbol{\theta}}_y)I_y(\hat{\boldsymbol{\theta}}_y)^{-1}),$$

and

$$\widehat{\text{risk}}_{y;y} \approx L_y(\hat{q}_y; p_y(\hat{\boldsymbol{\theta}}_y)) - L_y(q_y) + \frac{d}{n},$$

where $L_y(q_y)$ is ignored for model selection. Multiplying by $2n$ converts these approximations to $\text{AIC}_{x;y}$ and AIC, respectively.

6 PDIO and AICcd

The idea behind the derivation of PDIO and AIC_{cd} is to replace $\hat{q}_x$ by

$$\hat{q}_x = p_{z|y}(\hat{\boldsymbol{\theta}}_y)\hat{q}_y. \tag{22}$$

This implies (14) by considering the limiting situation of $n \to \infty$. Thus, the assumption for PDIO and AIC_{cd} is stronger than the assumption for $\text{AIC}_{x;y}$. Substituting (22) into the complete-data MLE gives

$$\hat{\boldsymbol{\theta}}_x = \arg\min_{\boldsymbol{\theta}\in\Theta} L_x(\hat{q}_x; p_x(\boldsymbol{\theta})). \tag{23}$$

Comparing (23) with (11) gives $\hat{\boldsymbol{\theta}}_x = \hat{\boldsymbol{\theta}}_y$. Therefore, there should not be any missing data, or at least $p_{z|y}(\boldsymbol{\theta})$ should not involve the parameter $\boldsymbol{\theta}$. Consequently, AIC, PDIO, AIC_{cd}, and $\text{AIC}_{x;y}$ are equivalent when PDIO and AIC_{cd} are justified under (22).

Although assumption (22) is too strong to work with, it is interesting to see how PDIO and AIC_{cd} would be derived if (22) is formally accepted. The argument below to derive PDIO and AIC_{cd} is rather confusing because $\hat{q}_x$ is interpreted interchangeably as the complete-data empirical distribution or the right-hand side of (22).

By a similar argument to the proof of Theorem 1, the Taylor expansion of $L_x(\hat{q}_x; p_x(\boldsymbol{\theta}))$ around $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_y$ is

$$L_x(\hat{q}_x; p_x(\boldsymbol{\theta})) = L_x(\hat{q}_x; p_x(\hat{\boldsymbol{\theta}}_y)) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_y)^T \hat{H}_x(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_y) + O_p(n^{-3/2}) \tag{24}$$

with $\hat{H}_x = H_x(\hat{q}_x; \hat{\boldsymbol{\theta}}_y)$. Its expectation with $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}_y$ gives

$$L_x(q_x; p_x(\bar{\boldsymbol{\theta}}_y)) = E\{L_x(\hat{q}_x; p_x(\hat{\boldsymbol{\theta}}_y))\} + \frac{1}{2n}\text{tr}(H_x H_y^{-1} G_y H_y^{-1}) + O(n^{-3/2}). \tag{25}$$

This corresponds to (20) of Theorem 1. Noticing (16) and thus, $D_y(q_y; p_y(\bar{\boldsymbol{\theta}}_y)) = L_x(q_x; p_x(\bar{\boldsymbol{\theta}}_y)) - L_x(q_x)$, and then substituting (25) into (17) gives the estimator of $\text{risk}_{x;y}$ unbiased up to $O(n^{-1})$ under (22) as:

$$\widehat{\mathrm{risk}}_{x;y} = L_x(\hat{q}_x;\, p_x(\hat{\boldsymbol{\theta}}_y)) - L_x(q_x) + \frac{1}{n}\mathrm{tr}(H_x H_y^{-1} G_y H_y^{-1}). \tag{26}$$

The goodness of fit term is $L_x(p_{z|y}(\hat{\boldsymbol{\theta}}_y)\hat{q}_y;\, p_x(\hat{\boldsymbol{\theta}}_y)) = -Q(\hat{\boldsymbol{\theta}}_y; \hat{\boldsymbol{\theta}}_y)/n$ under (22). Therefore, (26) gives AIC_{cd} by the same approximation used to derive $\mathrm{AIC}_{x;y}$.

In Cavanaugh and Shumway (1998), for evaluating (3.15) there, they assumed that $E\{Q(\boldsymbol{\theta}_0; \hat{\boldsymbol{\theta}}_y)\} \approx E\{Q(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0)\}$ or $E(L_x(p_{z|y}(\hat{\boldsymbol{\theta}}_y)\hat{q}_y;\, p_x(\boldsymbol{\theta}_0))) \approx L_x(p_{z|y}(\boldsymbol{\theta}_0)q_y;\, p_x(\boldsymbol{\theta}_0))$ under the correct specification $q_x = p_x(\boldsymbol{\theta}_0)$. The equality holds exactly under (22) because $E(L_x(\hat{q}_x;\, p_x(\boldsymbol{\theta}_0))) = L_x(q_x;\, p_x(\boldsymbol{\theta}_0))$ if $\hat{q}_x$ is interpreted as the empirical distribution. Unfortunately, the difference is $E\{Q(\boldsymbol{\theta}_0; \hat{\boldsymbol{\theta}}_y)\} - E\{Q(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0)\} = O(1)$ in general without assuming (22), leading to the bias of AIC_{cd} even when (15) holds.

In Shimodaira (1994), (3.5) corresponds to our (24), where $\hat{\boldsymbol{\theta}}_x = \hat{\boldsymbol{\theta}}_y$ is assumed implicitly to ignore the first derivative. Although $L_x(\hat{q}_x)$ diverges for continuous random variable $\boldsymbol{x}$, $D_x(\hat{q}_x;\, p_x(\hat{\boldsymbol{\theta}}_y)) = L_x(\hat{q}_x;\, p_x(\hat{\boldsymbol{\theta}}_y)) - L_x(\hat{q}_x)$ is formally considered. Similar to (16), we then have $D_x(\hat{q}_x;\, p_x(\hat{\boldsymbol{\theta}}_y)) = D_y(\hat{q}_y;\, p_y(\hat{\boldsymbol{\theta}}_y))$ in (3.6) there. From this argument, the goodness of fit term of (26) is $L_x(\hat{q}_x;\, p_x(\hat{\boldsymbol{\theta}}_y)) = L_y(\hat{q}_y;\, p_y(\hat{\boldsymbol{\theta}}_y)) + L_x(\hat{q}_x) - L_y(\hat{q}_y)$, where $L_x(\hat{q}_x) - L_y(\hat{q}_y)$ is independent of the model specification if $\hat{q}_x$ is interpreted as the empirical distribution. Therefore, (26) gives PDIO because $L_x(\hat{q}_x;\, p_x(\hat{\boldsymbol{\theta}}_y))$ can be replaced with $L_y(\hat{q}_y;\, p_y(\hat{\boldsymbol{\theta}}_y))$ for model selection.

7 Simulation study

7.1 Simulation 1

To verify Theorem 1, we performed a simulation study of the two-component normal mixture model defined as follows. Let $z \in \{1, 2\}$ be a discrete random variable for the component label, and $y \in \mathbb{R}$ be a continuous random variable for the observation. The distribution of z is $P(z = i) = \pi_i$ and the conditional distribution of y given $z = i$ is the normal distribution with mean μ_i and variance σ_i^2. The true parameter for data generation is specified as $\boldsymbol{\theta}_0^T = (\pi_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = (0.6, -1, 1, 0.7^2, 0.7^2)$. We consider two candidate models for selection. Model 1 is a two-component normal mixture model with a constraint $\sigma_1^2 = \sigma_2^2$ ($d = 4$), whereas Model 2 is the same model without the constraint ($d = 5$). Because these two models are correctly specified, (15) holds. However, (22) obviously does not.

We generated $B = 4000$ datasets with sample size $n = 100, 200, 500, 1000, 2000, 5000, 10000$. They are denoted as $X^{(b)} = (\boldsymbol{x}_1^{(b)}, \ldots, \boldsymbol{x}_n^{(b)})$, $b = 1, \ldots, B$. We also generated datasets of sample size $\tilde{n} = 15000$, which are denoted as $\tilde{X}^{(b)} = (\tilde{\boldsymbol{x}}_1^{(b)}, \ldots, \tilde{\boldsymbol{x}}_{\tilde{n}}^{(b)})$ for computing the loss functions. For each $X^{(b)} = (Y^{(b)}, Z^{(b)})$ and Model k, $k = 1, 2$, we computed the information criteria $\mathrm{AIC}(Y^{(b)}, k)$, $\mathrm{PDIO}(Y^{(b)}, k)$, $\mathrm{AIC}_{cd}(Y^{(b)}, k)$, $\mathrm{AIC}_{x;y}(Y^{(b)}, k)$, and the loss functions $\mathrm{loss}_{y;y}(Y^{(b)}, k) = L_y(q_y;\, p_y(\hat{\boldsymbol{\theta}}_y^{(b)}))$, $\mathrm{loss}_{x;y}(X^{(b)}, k) = L_x(q_x;\, p_x(\hat{\boldsymbol{\theta}}_y^{(b)}))$, where $\hat{\boldsymbol{\theta}}_y^{(b)}$ is computed from $Y^{(b)}$. In the formulas below, $:\approx$ denotes that the expectation on the

left-hand side is computed numerically by the simulation on the right-hand side. The loss functions are computed numerically by

$$\text{loss}_{y;y}(Y^{(b)}, k) :\approx -\frac{1}{\tilde{n}}\sum_{t=1}^{\tilde{n}} \log p_y(\tilde{\boldsymbol{y}}_t^{(b)}; \hat{\boldsymbol{\theta}}_y^{(b)}),$$

$$\text{loss}_{x;y}(X^{(b)}, k) :\approx -\frac{1}{\tilde{n}}\sum_{t=1}^{\tilde{n}} \log p_x(\tilde{\boldsymbol{x}}_t^{(b)}; \hat{\boldsymbol{\theta}}_y^{(b)}),$$

where p_x, p_y, and $\hat{\boldsymbol{\theta}}_y^{(b)}$ are for Model k. Then, the expectation with respect to $q_x = p_x(\boldsymbol{\theta}_0)$ is computed by the simulation average. For example,

$$E(\Delta\text{AIC}) :\approx \frac{1}{B}\sum_{b=1}^{B}(\text{AIC}(Y^{(b)}, 1) - \text{AIC}(Y^{(b)}, 2)),$$

$$\Delta\text{risk}_{x;y} :\approx \frac{1}{B}\sum_{b=1}^{B}(\text{loss}_{x;y}(X^{(b)}, 1) - \text{loss}_{x;y}(X^{(b)}, 2)).$$

This Monte Carlo method calculates the expectation accurately for sufficiently large $\tilde{n}$ and B.

The result shown in Table 1 verifies Theorem 1. For sufficiently large n, $E(\Delta\text{AIC}) = 2n\Delta\text{risk}_{y;y}$ and $E(\Delta\text{AIC}_{x;y}) = 2n\Delta\text{risk}_{x;y}$ hold very well. On the other hand, $E(\Delta\text{PDIO})$ differs significantly from $2n\Delta\text{risk}_{y;y}$ and $2n\Delta\text{risk}_{x;y}$. Thus, PDIO is not a good estimator of either of these risk functions. In addition, the expected value of AIC_{cd} is similar to that of PDIO, but its variation is larger than PDIO, as seen in the standard errors.

Let us consider the difference PDIO $-$ AIC_{cd}

$$\text{diff}(Y, \hat{\boldsymbol{\theta}}_y) = 2Q(\hat{\boldsymbol{\theta}}_y; \hat{\boldsymbol{\theta}}_y) - 2\ell_y(\hat{\boldsymbol{\theta}}_y) = 2\sum_{t=1}^{n}\int p_{z|y}(\boldsymbol{z}|\boldsymbol{y}_t; \hat{\boldsymbol{\theta}}_y)\log p_{z|y}(\boldsymbol{z}|\boldsymbol{y}_t; \hat{\boldsymbol{\theta}}_y)\,d\boldsymbol{z},$$

and its difference between the two models, which is denoted as $\Delta\text{diff}(Y, \hat{\boldsymbol{\theta}}_y) = \Delta\text{PDIO} - \Delta\text{AIC}_{cd}$. $\Delta\text{diff}(Y, \hat{\boldsymbol{\theta}}_y)$ and $E(\Delta\text{diff}(Y, \hat{\boldsymbol{\theta}}_y))$ can be very large, and they are $O(n)$ under model misspecification. If (15) holds, as is the case of Table 1, $E(\text{diff}(Y, \bar{\boldsymbol{\theta}}_y)) = 2n\int q_x(x)\log q_{z|y}(z|y)\,dx$ is independent of the model. Therefore, the difference becomes smaller; $\Delta\text{diff}(Y, \hat{\boldsymbol{\theta}}_y) = O_p(\sqrt{n})$ and $E(\Delta\text{diff}(Y, \hat{\boldsymbol{\theta}}_y)) = O(1)$.

7.2 Simulation 2

We next performed a simulation study on the three-component normal mixture model to examine how well the information criteria work for model selection

Table 1 Expected values of the information criteria and the risk functions in Simulation 1

n	100	200	500	1000	2000	5000	10000
$E(\Delta\mathrm{AIC})$	0.810	0.898	0.982	0.978	0.986	0.982	1.04
	(0.027)	(0.025)	(0.023)	(0.023)	(0.023)	(0.023)	(0.022)
$E(\Delta\mathrm{PDIO})$	43.5	41.1	37.0	36.0	34.9	34.4	34.2
	(1.64)	(0.716)	(0.344)	(0.220)	(0.141)	(0.088)	(0.064)
$E(\Delta\mathrm{AIC}_{cd})$	42.3	41.0	37.2	36.6	35.2	35.5	33.5
	(1.67)	(0.793)	(0.518)	(0.494)	(0.573)	(0.812)	(1.08)
$E(\Delta\mathrm{AIC}_{x;y})$	22.1	21.0	19.0	18.5	18.0	17.7	17.6
	(0.821)	(0.361)	(0.174)	(0.113)	(0.074)	(0.049)	(0.037)
$2n\Delta\mathrm{risk}_{y;y}$	1.83	1.47	1.15	1.08	1.03	1.02	0.967
	(0.052)	(0.040)	(0.030)	(0.027)	(0.026)	(0.030)	(0.033)
$2n\Delta\mathrm{risk}_{x;y}$	100.9	28.9	20.3	18.6	18.2	17.5	17.0
	(40.3)	(1.39)	(0.620)	(0.487)	(0.456)	(0.464)	(0.430)

These values are differences between the two models with standard errors in parentheses

Table 2 Frequency of model selection in Simulation 2

	Model 1 ($d = 6$)	Model 2 ($d = 7$)	Model 3 ($d = 7$)	Model 4[a] ($d = 7$)	Model 5[a] ($d = 8$)
AIC	881	2419	262	5600	838
PDIO	5442	16	4	4534	4
AIC_{cd}	2063	2	974	6551	410
$\mathrm{AIC}_{x;y}$	3704	65	15	6190	26

[a] Correctly specified model

in a practical situation where some candidate models do not satisfy assumption (15). The true parameter value is $\boldsymbol{\theta}_0^T = (\pi_1, \pi_2, \mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2) = (0.5, 0.3, -2, 0, 3, 0.7^2, 0.7^2, 1^2)$. We consider five candidates with the following constraints. Model 1 is $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ ($d = 6$). Model 2 is $\sigma_2^2 = \sigma_3^2$ ($d = 7$). Model 3 is $\sigma_1^2 = \sigma_3^2$ ($d = 7$). Model 4 is $\sigma_1^2 = \sigma_2^2$ ($d = 7$), and Model 5 has no constraint ($d = 8$). Models 1, 2 and 3 are misspecified and do not satisfy (15). Models 4 and 5 are correctly specified and satisfy (15). None of the models satisfy (22). We have generated $B = 10000$ datasets of $n = 500$ and $\tilde{n} = 2000$.

Table 2 shows the model selection results. Model 4 is the best model in the sense that it minimizes both $\mathrm{risk}_{y;y}$ and $\mathrm{risk}_{x;y}$ (Table 3). All the information criteria tend to select Model 4. AIC tends to choose a more complex model (i.e., Model 2 or Model 5) than the other criteria, indicating a smaller penalty for model complexity. PDIO tends to choose a simpler model (i.e., Model 1), implying a larger penalty for model complexity.

To compare candidate models in the long run, the expected loss of each Model k relative to that of Model 4 is computed by

Table 3 Risk functions for models and those for information criteria in Simulation 2

	$2n\Delta\text{risk}_{y;y}$		$2n\Delta\text{risk}_{x;y}$	
Model 1	6.60	(0.04)	33.2	(0.21)
Model 2	1.40	(0.02)	59.2	(0.71)
Model 3	7.86	(0.04)	80.7	(0.80)
Model 4[a]	0	(0.00)	0	(0.00)
Model 5[a]	1.32	(0.02)	45.6	(0.87)
AIC	**1.44**	(0.03)	39.6	(0.91)
PDIO	3.57	(0.04)	19.6	(0.30)
AIC_{cd}	2.33	(0.04)	28.2	(0.72)
$\text{AIC}_{x;y}$	2.36	(0.04)	**14.8**	(0.43)

These values are relative to Model 4 with standard errors in parentheses
[a] Correctly specified model

$$\Delta\text{risk}_{x;y}(k) :\approx \frac{1}{B}\sum_{b=1}^{B}(\text{loss}_{x;y}(X^{(b)}, k) - \text{loss}_{x;y}(X^{(b)}, 4)).$$

Table 3 (upper) shows the results. The most complex model (Model 5) is the second best in terms of $\text{risk}_{y;y}$, but the simplest model (Model 1) is the second best in terms of $\text{risk}_{x;y}$, indicating a large contribution of $p_{z|y}(\boldsymbol{\theta})$ to the second term of (17).

The information criterion performance is measured by the expected loss of the selected model. For example, the performance of AIC in terms of complete data is measured by

$$\Delta\text{risk}_{x;y}(\text{AIC}) :\approx \frac{1}{B}\sum_{b=1}^{B}(\text{loss}_{x;y}(X^{(b)}, \hat{k}^{(b)}) - \text{loss}_{x;y}(X^{(b)}, 4)),$$

where $\hat{k}^{(b)}$ is the minimum AIC model computed from $Y^{(b)}$. Table 3 (lower) shows the results, where the value in bold denotes the minimum value of each column. AIC outperforms the other criteria in terms of $\text{risk}_{y;y}$, and $\text{AIC}_{x;y}$ outperforms the other criteria in terms of $\text{risk}_{x;y}$. In this example, some models do not satisfy assumption (15), but AIC and $\text{AIC}_{x;y}$ work very well as expected.

8 Concluding remarks

We derived $\text{AIC}_{x;y}$ as an unbiased estimator of the expected Kullback–Leibler divergence between the true distribution and the estimated distribution of complete data when only incomplete data are available. In Simulation 1, $\text{AIC}_{x;y}$ and AIC are unbiased up to the penalty terms, whereas PDIO and AIC_{cd} are not.

To derive $\text{AIC}_{x;y}$, we assumed (15), meaning that the conditional distribution $p_{z|y}(\boldsymbol{\theta})$ of the missing data given the incomplete data is correctly specified, while the marginal distribution $p_y(\boldsymbol{\theta})$ of the incomplete data is misspecified in general. However, the conditional distribution is misspecified in practice. In Simulation 2, we observed that $\text{AIC}_{x;y}$ and AIC perform better than the other criteria even if some

models are misspecified. Without assumption (15), the dominant term in (17) is $D_x(q_x; p_x(\bar{\boldsymbol{\theta}}_y)) = D_x(q_x; p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y) + D_y(q_y; p_y(\bar{\boldsymbol{\theta}}_y)) \geq D_y(q_y; p_y(\bar{\boldsymbol{\theta}}_y))$. Thus, $\mathrm{AIC}_{x;y}$ estimates the lower bound of $2n\,\mathrm{risk}_{x;y}$. It is impossible to reasonably estimate the ignored term $D_x(q_x; p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y)$ in our setting where $z_1, \ldots, z_n$ are missing completely.

Although we assume that $p_{z|y}(\boldsymbol{\theta})$ is correctly specified, it is beneficial to include $p_{z|y}(\boldsymbol{\theta})$ as a part of $p_x(\boldsymbol{\theta}) = p_{z|y}(\boldsymbol{\theta})p_y(\boldsymbol{\theta})$ for model selection. The variance of $\hat{\boldsymbol{\theta}}_y$ causes $p_{z|y}(\hat{\boldsymbol{\theta}}_y)$ to fluctuate even if $p_{z|y}(\bar{\boldsymbol{\theta}}_y) = q_{z|y}$. The amount of this random variation is measured by the additional penalty term (1) in $\mathrm{AIC}_{x;y}$.

In the future, we plan to work on more complicated missing mechanisms or combine a missing mechanism with other sampling mechanisms, such as the covariate-shift (Shimodaira 2000) problem. One important extension is semi-supervised learning (Chapelle et al. 2006; Kawakita and Takeuchi 2014), where the log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log p_y(\boldsymbol{y}_t; \boldsymbol{\theta}) + \sum_{t=n+1}^{n+n'} \log p_x(\boldsymbol{x}_t; \boldsymbol{\theta}).$$

In this case, the additional complete data $\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+n'}$ help estimate conditional distribution $q_{z|y}$. We may reasonably estimate $D_x(q_x; p_{z|y}(\bar{\boldsymbol{\theta}}_y)q_y)$ without assuming (15), leading to a new information criterion, which will be the subject in future research.

Acknowledgements We would like to thank the reviewers for their comments to improve the manuscript. We appreciate Kei Hirose and Shinpei Imori for their suggestions and comments. While preparing an earlier version of the manuscript, which was published as Shimodaira (1994), Hidetoshi Shimodaira is indebted to Shun-ichi Amari for the geometrical view of the EM algorithm and to Noboru Murata for the derivation of the Takeuchi information criterion.

Appendix A: Technical details

A.1 Proof of Lemma 1

For brevity, we omit $(\boldsymbol{y}, \boldsymbol{z})$ of $f_x(\boldsymbol{y}, \boldsymbol{z})$ in the integrals below. $D_x(g_x; f_x) = \int\int g_{z|y}g_y(\log g_{z|y} + \log g_y - \log f_{z|y} - \log f_y)d\boldsymbol{z}d\boldsymbol{y} = \int g_y \int g_{z|y}(\log g_{z|y} - \log f_{z|y})d\boldsymbol{z}d\boldsymbol{y} + \int g_y(\int g_{z|y}d\boldsymbol{z})(\log g_y - \log f_y)d\boldsymbol{y} = \int g_y \int g_{z|y}(\log g_{z|y}g_y - \log f_{z|y}g_y)d\boldsymbol{z}d\boldsymbol{y} + \int g_y(\log g_y - \log f_y)d\boldsymbol{y} = D_x(g_{z|y}g_y; f_{z|y}g_y) + D_y(g_y; f_y)$, thus showing (6). $D_y(g_y; f_y) = \int\int h_{z|y}g_y(\log g_y - \log f_y + \log h_{z|y} - \log h_{z|y})d\boldsymbol{z}d\boldsymbol{y} = D_x(h_{z|y}g_y; h_{z|y}f_y)$, which shows (7). □

A.2 Proof of Lemma 2

We assume $q_{z|y} = p_{z|y}(\bar{\boldsymbol{\theta}}_y)$ and $\bar{\boldsymbol{\theta}}_x = \bar{\boldsymbol{\theta}}_y$. From the definitions of $\bar{\boldsymbol{\theta}}_x$ and H_x, we have

$$\frac{\partial D_x(q_x;\, p_x(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\Big|_{\bar{\boldsymbol{\theta}}_y} = 0, \quad \frac{\partial^2 D_x(q_x;\, p_x(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\Big|_{\bar{\boldsymbol{\theta}}_y} = H_x.$$

Hence, the Taylor expansion of $D_x(q_x;\, p_x(\boldsymbol{\theta}))$ around $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}_y$ is

$$D_x(q_x;\, p_x(\boldsymbol{\theta})) = D_x(q_x;\, p_x(\bar{\boldsymbol{\theta}}_y)) + \frac{1}{2}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y)^T H_x(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y) + O(n^{-3/2})$$

for $\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_y = O(n^{-1/2})$. The first term on the right-hand side is $D_y(q_y;\, p_y(\bar{\boldsymbol{\theta}}_y))$ as shown in (16). Substituting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_y$ in $D_x(q_x;\, p_x(\boldsymbol{\theta}))$ and taking its expectation gives (17) by noting

$$E\left\{(\hat{\boldsymbol{\theta}}_y - \bar{\boldsymbol{\theta}}_y)^T H_x(\hat{\boldsymbol{\theta}}_y - \bar{\boldsymbol{\theta}}_y)\right\} = \mathrm{tr}\left(H_x\, E\left\{(\hat{\boldsymbol{\theta}}_y - \bar{\boldsymbol{\theta}}_y)(\hat{\boldsymbol{\theta}}_y - \bar{\boldsymbol{\theta}}_y)^T\right\}\right),$$

which becomes $\mathrm{tr}\left(H_x H_y^{-1} G_y H_y^{-1}\right)/n + O(n^{-2})$ from (4). □

A.3 Proof of Theorem 1

From the definitions of $\hat{\boldsymbol{\theta}}_y$ and $\hat{H}_y = H_y(\hat{q}_y; \hat{\boldsymbol{\theta}}_y)$, we have

$$\frac{\partial L_y(\hat{q}_y;\, p_y(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\Big|_{\hat{\boldsymbol{\theta}}_y} = 0, \quad \frac{\partial^2 L_y(\hat{q}_y;\, p_y(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\Big|_{\hat{\boldsymbol{\theta}}_y} = \hat{H}_y.$$

Hence, the Taylor expansion of $L_y(\hat{q}_y;\, p_y(\boldsymbol{\theta}))$ around $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_y$ is

$$L_y(\hat{q}_y;\, p_y(\boldsymbol{\theta})) = L_y(\hat{q}_y;\, p_y(\hat{\boldsymbol{\theta}}_y)) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_y)^T \hat{H}_y(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_y) + O_p(n^{-3/2})$$

for $\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_y = O_p(n^{-1/2})$. Substituting $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}_y$ in $L_y(\hat{q}_y;\, p_y(\boldsymbol{\theta}))$, we take its expectation below. By noting $\hat{H}_y = H_y + O_p(n^{-1/2})$, we have

$$E\left\{(\bar{\boldsymbol{\theta}}_y - \hat{\boldsymbol{\theta}}_y)^T \hat{H}_y(\bar{\boldsymbol{\theta}}_y - \hat{\boldsymbol{\theta}}_y)\right\} = \mathrm{tr}\left(H_y\, E\left\{(\hat{\boldsymbol{\theta}}_y - \bar{\boldsymbol{\theta}}_y)(\hat{\boldsymbol{\theta}}_y - \bar{\boldsymbol{\theta}}_y)^T\right\}\right) + O(n^{-3/2}),$$

which becomes $\mathrm{tr}(H_y H_y^{-1} G_y H_y^{-1})/n + O(n^{-3/2})$ from (4). This proves (20) because

$$E\{L_y(\hat{q}_y;\, p_y(\bar{\boldsymbol{\theta}}_y))\} = E\{L_y(\hat{q}_y;\, p_y(\hat{\boldsymbol{\theta}}_y))\} + \frac{1}{2n}\mathrm{tr}(G_y H_y^{-1}) + O(n^{-3/2}),$$

and $E\{L_y(\hat{q}_y;\, p_y(\bar{\boldsymbol{\theta}}_y))\} = L_y(q_y;\, p_y(\bar{\boldsymbol{\theta}}_y))$. Substituting (20) into (17) and comparing it with (18) yields (21). □

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Amari, S. (1995). Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, *8*, 1379–1408.

Amari, S., Nagaoka, H. (2007). *Methods of information geometry 191*. Providence, RI: American Mathematical Society.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

Burnham, K. P., Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

Byrne, W. (1992). Alternating minimization and Boltzmann machine learning. *IEEE Transactions on Neural Networks*, *3*, 612–620.

Cavanaugh, J. E., Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference*, *67*, 45–65.

Chapelle, O., Schölkopf, B., Zien, A. (2006). *Semi-supervised learning*. Cambridge: The MIT Press.

Claeskens, G., Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, *64*, 1062–1069.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, *3*, 146–158.

Csiszár, I., Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statistics and decisions, Supplement Issue*, *1*, 205–237.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, *39*, 1–38.

Ip, E. H., Lalwani, N. (2000). A note on the geometric interpretation of the EM algorithm in estimating item characteristics and student abilities. *Psychometrika*, *65*, 533–537.

Kawakita, M., Takeuchi, J. (2014). Safe semi-supervised learning based on weighted likelihood. *Neural Networks*, *53*, 146–164.

Konishi, S., Kitagawa, G. (2008). *Information criteria and statistical modeling*. New York: Springer.

Meng, X.-L., Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*, 899–909.

Seghouane, A. K., Bekara, M., Fleury, G. (2005). A criterion for model selection in the presence of incomplete data based on Kullback's symmetric divergence. *Signal Processing*, *85*, 1405–1417.

Shimodaira, H. (1994). A new criterion for selecting models from partially observed data. *Selecting Models from Data* (pp. 21–29). New York: Springer.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*, 227–244.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Yamazaki, K. (2014). Asymptotic accuracy of distribution-based estimation of latent variables. *The Journal of Machine Learning Research*, *15*, 3541–3562.