An Akaike Information Criterion for Model Selection in the Presence of Incomplete Data

by

Joseph E. Cavanaugh

Department of Statistics, University of Missouri, Columbia, MO 65211

and

Robert H. Shumway

Division of Statistics, University of California, Davis, CA 95616

Abstract

We derive and investigate a variant of AIC, the Akaike information criterion, for model selection in settings where the observed data is incomplete. Our variant is based on the motivation provided for the PDIO ("predictive divergence for incomplete observation models") criterion of Shimodaira (1994, in *Selecting Models from Data: Artificial Intelligence and Statistics IV, Lecture Notes in Statistics* **89**, Springer-Verlag, New York, 21–29). However, our variant differs from PDIO in its "goodness-of-fit" term. Unlike AIC and PDIO, which require the computation of the observed-data empirical log-likelihood, our criterion can be evaluated using only complete-data tools, readily available through the EM algorithm and the SEM ("supplemented" EM) algorithm of Meng and Rubin (1991, *Journal of the American Statistical Association* **86**, 899–909). We compare the performance of our AIC variant to that of both AIC and PDIO in simulations where the data being modeled contains missing values. The results indicate that our criterion is less prone to overfitting than AIC and less prone to underfitting than PDIO.

AMS Subject Classification: 62B10, 94A17, 62E25.

Keywords: AIC, EM algorithm, information theory, Kullback-Leibler information, model selection criteria, PDIO criterion, SEM algorithm.

Abbreviated Title: An AIC for Model Selection with Incomplete Data.

1. Introduction

Modeling in the presence of incomplete or partially observed data arises in a large variety of practical problems, including applications involving ANOVA and regression models (Rubin, 1976; Little, 1979), state-space models (Shumway and Stoffer, 1982), latent class models (Goodman, 1974), and mixture models (Titterington, Smith, and Makov, 1985). In such settings, we generally view the observed, incomplete data $\mathbf{Y}_{obs}$ together with unobserved, missing data $\mathbf{Y}_{mis}$ as comprising the complete data $\mathbf{Y}$. A parametric family of models $f(\mathbf{Y}|\,\boldsymbol{\theta})$ is postulated for the complete data $\mathbf{Y}$, where the size of the parameter vector $\boldsymbol{\theta}$ dictates the complexity of the corresponding model. The complete-data model $f(\mathbf{Y}|\,\boldsymbol{\theta})$ implies a model for the incomplete data, $f(\mathbf{Y}_{obs}|\,\boldsymbol{\theta})$, although the latter is often more difficult to represent or to work with than the former. In many frameworks, a fitted model for the complete data $f(\mathbf{Y}|\,\hat{\boldsymbol{\theta}})$ can be conveniently found through utilizing the well-known EM (expectation/maximization) algorithm (Dempster, Laird, and Rubin, 1977).

In most applications, finding a suitable dimension for the parameter vector $\boldsymbol{\theta}$ is an important component of the modeling problem. A common approach is to choose several different dimensions for $\boldsymbol{\theta}$, find the fitted models corresponding to these choices, compute a model selection criterion for each of the fitted candidate models, and determine the dimension of $\boldsymbol{\theta}$ for the final model based on the values of the criterion. The Akaike information criterion (Akaike, 1973, 1974), or AIC, is the most widely known and used of the criteria which have been proposed for this purpose.

In the present context, AIC can be interpreted as a measure of separation between the fitted model for the incomplete data, $f(\mathbf{Y}_{obs}|\,\hat{\boldsymbol{\theta}})$, and the "true" or generating model which presumably gave rise to the incomplete data, say $f(\mathbf{Y}_{obs}|\,\boldsymbol{\theta}_o)$. Yet as indicated by Shimodaira (1994), in many applications it may be more natural or desirable to use a criterion based on the complete data, which assesses the separation between the fitted model $f(\mathbf{Y}|\,\hat{\boldsymbol{\theta}})$ and the generating model $f(\mathbf{Y}|\,\boldsymbol{\theta}_o)$. There are several arguments to be made in defense of this idea. First of all, the implementation of the EM algorithm is based on the premise that a convenient class of models can be specified for the complete data $\mathbf{Y}$, whereas the corresponding class of models for the incomplete data $\mathbf{Y}_{obs}$ may be difficult to exhibit or to work with. Since it is the complete data for which the investigator postulates the family of models, it seems reasonable to base model selection on measures which assess the propriety of fitted candidate models within this family. Secondly, as pointed out by Meng and Rubin (1991, page 899), the EM algorithm essentially involves "capitalizing on computing power and complete-data tools to handle missing-data problems." Because the EM algorithm utilizes complete-data tools, it may be more computationally convenient to calculate a selection criterion based on these quantities rather than analogous incomplete-data quantities. And finally, the complete-data density $f(\mathbf{Y}|\,\boldsymbol{\theta})$ is composed of the product of the incomplete-data density $f(\mathbf{Y}_{obs}|\,\boldsymbol{\theta})$ and the conditional density of the missing data $\mathbf{Y}_{mis}$ given the incomplete data $\mathbf{Y}_{obs}$; i.e.,

$$f(\mathbf{Y}|\,\boldsymbol{\theta}) = f(\mathbf{Y}_{obs}|\,\boldsymbol{\theta})\ f(\mathbf{Y}_{mis}|\,\mathbf{Y}_{obs}, \boldsymbol{\theta}).$$

Suppose that the density $f(\mathbf{Y}_{mis}|\,\mathbf{Y}_{obs}, \boldsymbol{\theta})$ is substantially affected by deviations of $\boldsymbol{\theta}$ from the "true" parameter vector $\boldsymbol{\theta}_o$. Model selection based on the discrepancy between $f(\mathbf{Y}|\,\hat{\boldsymbol{\theta}})$ and $f(\mathbf{Y}|\,\boldsymbol{\theta}_o)$ would incorporate this information; it is not clear that model selection based on the discrepancy between $f(\mathbf{Y}_{obs}|\,\hat{\boldsymbol{\theta}})$ and $f(\mathbf{Y}_{obs}|\,\boldsymbol{\theta}_o)$ would do the same.

The last of the aforementioned arguments is further explored in the next section. In Section 3, we present an informal derivation of a model selection criterion which is analogous to AIC, yet is based on complete-data rather than incomplete-data concepts and tools. We call this criterion AICcd, where the notation "cd" stands for "complete data". Our criterion is motivated by the same principle as the PDIO ("predictive divergence for incomplete observation models") criterion introduced by Shimodaira (1994), yet differs from PDIO in its goodness-of-fit term.

In Section 4, we describe the evaluation of AICcd, and indicate the computational advantage the criterion holds over PDIO and AIC. We contrast the forms of AICcd, PDIO, and AIC in Section 5, and discuss several key principles related to the behavior of these criteria. These principles are illustrated by the simulation sets presented in Sections 6 and 7. In these simulations, we compare the effectiveness of AICcd, PDIO, and AIC at selecting a model of correct dimension within a candidate class, where the data being modeled contains various degrees of missing values. Our results demonstrate that AICcd is generally less prone to underfitting than PDIO and less prone to overfitting than AIC. The simulations in Section 6 involve modeling bivariate normal data whereas the simulations in Section 7 are based on multivariate regression models. Section 8 concludes.

2. Complete-Data versus Incomplete-Data Kullback-Leibler Discrepancy

Let $f(\mathbf{Y}|\,\boldsymbol{\theta})$ and $f(\mathbf{Y}_{obs}|\,\boldsymbol{\theta})$ respectively denote parametric densities for the complete data $\mathbf{Y}$ and the incomplete data $\mathbf{Y}_{obs}$. Assume that the parameter vector $\boldsymbol{\theta}$ is d-dimensional. Let $\boldsymbol{\theta}_o$ denote the "true" parameter vector, so that $f(\mathbf{Y}|\,\boldsymbol{\theta}_o)$ and $f(\mathbf{Y}_{obs}|\,\boldsymbol{\theta}_o)$ respectively represent the generating densities for the complete and the incomplete data.

A well-known measure of separation between two models is given by the non-normalized Kullback-Leibler information (Kullback, 1968), also known as the cross entropy or discrepancy. The complete-data Kullback-Leibler discrepancy between a candidate model $f(\mathbf{Y}|\, \boldsymbol{\theta})$ and the generating model $f(\mathbf{Y}|\, \boldsymbol{\theta}_o)$ is defined by

$$D_{\mathbf{Y}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) = \mathrm{E}_{\mathbf{Y}}\{-2 \ln f(\mathbf{Y}|\, \boldsymbol{\theta})\}, \tag{2.1}$$

where $\mathrm{E}_{\mathbf{Y}}$ denotes the expected value with respect to the density $f(\mathbf{Y}|\, \boldsymbol{\theta}_o)$. Similarly, the incomplete-data Kullback-Leibler discrepancy between a candidate model $f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta})$ and the generating model $f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta}_o)$ is defined by

$$D_{\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) = \mathrm{E}_{\mathbf{Y}_{obs}}\{-2 \ln f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta})\}, \tag{2.2}$$

where $\mathrm{E}_{\mathbf{Y}_{obs}}$ denotes the expected value with respect to the density $f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta}_o)$. Since the evaluation of (2.1) and (2.2) requires knowledge of $\boldsymbol{\theta}_o$, these quantities are not directly accessible. Thus, it is not possible to assess the exact discrepancy between a fitted candidate model, parameterized by $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and the corresponding generating model.

An important contribution of Akaike (1973, 1974) was in showing that in certain large-sample settings, the expected value of

$$D_{\mathbf{Y}_{obs}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_o) = \mathrm{E}_{\mathbf{Y}_{obs}}\{-2 \ln f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{2.3}$$

(with respect to $f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta}_o)$) is approximately the same as the expected value of

$$\mathrm{AIC} = -2 \ln L(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}) + 2d, \tag{2.4}$$

where $L(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$ denotes the incomplete-data empirical likelihood. In other words, AIC serves as an approximately unbiased estimator of the expected incomplete-data discrepancy

$$\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o) = \mathrm{E}_{\mathbf{Y}_{obs}}\left\{\mathrm{E}_{\mathbf{Y}_{obs}}\{-2 \ln f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right\}. \tag{2.5}$$

The terms $-2 \ln L(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$ and $2d$ in AIC are commonly referred to as the "goodness-of-fit" and "penalty" terms, respectively.

Our objective is to propose a version of AIC that will have an expected value (with respect to $f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta}_o)$) which, in large-sample settings, is approximately the same as the expected value of

$$D_{\mathbf{Y}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_o) = \mathrm{E}_{\mathbf{Y}}\{-2 \ln f(\mathbf{Y}|\, \boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{2.6}$$

(Note that (2.6) is a function of $\mathbf{Y}_{obs}$ through its dependence on $\hat{\boldsymbol{\theta}}$, yet does not involve the missing data $\mathbf{Y}_{mis}$.) Equivalently, we wish to propose an approximately unbiased estimator of the expected complete-data discrepancy

$$\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o) = \mathrm{E}_{\mathbf{Y}_{obs}} \left\{ \mathrm{E}_{\mathbf{Y}}\{-2 \ln f(\mathbf{Y}|\, \boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\}. \qquad (2.7)$$

The relationship between (2.1) and (2.2) provides an important insight into why it may be preferable to base model selection on the former as opposed to the latter. To establish this relationship, recall that

$$f(\mathbf{Y}|\, \boldsymbol{\theta}) = f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta})\ f(\mathbf{Y}_{mis}|\, \mathbf{Y}_{obs}, \boldsymbol{\theta}), \qquad (2.8)$$

meaning

$$\mathrm{E}_{\mathbf{Y}}\{-2 \ln f(\mathbf{Y}|\, \boldsymbol{\theta})\} = \mathrm{E}_{\mathbf{Y}}\{-2 \ln f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta})\} + \mathrm{E}_{\mathbf{Y}}\{-2 \ln f(\mathbf{Y}_{mis}|\, \mathbf{Y}_{obs}, \boldsymbol{\theta})\}. \qquad (2.9)$$

Using (2.8), it is easily shown that

$$\mathrm{E}_{\mathbf{Y}}\{-2 \ln f(\mathbf{Y}_{obs}|\, \boldsymbol{\theta})\} = D_{\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o),$$

meaning that (2.9) can be written as

$$D_{\mathbf{Y}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) = D_{\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) + \mathrm{E}_{\mathbf{Y}}\{-2 \ln f(\mathbf{Y}_{mis}|\, \mathbf{Y}_{obs}, \boldsymbol{\theta})\}. \qquad (2.10)$$

Now consider the second of the two terms on the right-hand side of (2.10). Define

$$D_{\mathbf{Y}_{mis}|\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o|\, \mathbf{Y}_{obs}) = \mathrm{E}_{\mathbf{Y}_{mis}|\mathbf{Y}_{obs}}\{-2 \ln f(\mathbf{Y}_{mis}|\, \mathbf{Y}_{obs}, \boldsymbol{\theta})\}, \qquad (2.11)$$

where $\mathrm{E}_{\mathbf{Y}_{mis}|\mathbf{Y}_{obs}}$ denotes the expected value with respect to the density $f(\mathbf{Y}_{mis}|\, \mathbf{Y}_{obs}, \boldsymbol{\theta}_o)$. We will refer to (2.11) as the conditional missing-data discrepancy. Using (2.8), it is easily shown that

$$\mathrm{E}_{\mathbf{Y}}\{-2 \ln f(\mathbf{Y}_{mis}|\, \mathbf{Y}_{obs}, \boldsymbol{\theta})\} = \mathrm{E}_{\mathbf{Y}_{obs}}\{D_{\mathbf{Y}_{mis}|\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o|\, \mathbf{Y}_{obs})\},$$

meaning that (2.10) can be written as

$$D_{\mathbf{Y}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) = D_{\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) + \mathrm{E}_{\mathbf{Y}_{obs}}\{D_{\mathbf{Y}_{mis}|\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o|\, \mathbf{Y}_{obs})\}. \qquad (2.12)$$

Now one can easily establish using Jensen's inequality that for any $\mathbf{Y}_{obs}$ and any $\boldsymbol{\theta}$,

$$D_{\mathbf{Y}_{mis}|\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o|\, \mathbf{Y}_{obs}) \geq D_{\mathbf{Y}_{mis}|\mathbf{Y}_{obs}}(\boldsymbol{\theta}_o, \boldsymbol{\theta}_o|\, \mathbf{Y}_{obs}). \qquad (2.13)$$

If we then define $k(\boldsymbol{\theta}_o) \equiv \mathrm{E}_{\mathbf{Y}_{obs}}\{D_{\mathbf{Y}_{mis}|\mathbf{Y}_{obs}}(\boldsymbol{\theta}_o, \boldsymbol{\theta}_o|\ \mathbf{Y}_{obs})\}$, by (2.12) and (2.13), we have for any $\boldsymbol{\theta}$

$$D_{\mathbf{Y}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) \geq D_{\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) + k(\boldsymbol{\theta}_o). \tag{2.14}$$

Thus as a function of $\boldsymbol{\theta}$, the complete-data discrepancy $D_{\mathbf{Y}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o)$ is always *at least as great* as the incomplete-data discrepancy $D_{\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o)$, adjusted by the constant $k(\boldsymbol{\theta}_o)$.

From (2.12) and (2.14), we can infer that the complete-data discrepancy is potentially more sensitive than the incomplete-data discrepancy to deviations of $\boldsymbol{\theta}$ from $\boldsymbol{\theta}_o$ which affect the conditional missing-data discrepancy (2.11). This implies that in the presence of missing data, $D_{\mathbf{Y}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o)$ may be preferable to $D_{\mathbf{Y}_{obs}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o)$ for assessing the separation between a model parameterized by $\boldsymbol{\theta}$ and one parameterized by $\boldsymbol{\theta}_o$. As a consequence, an estimator of $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ may be preferable to an estimator of $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ as a model selection criterion, provided of course that the former is accurate enough to sufficiently reflect the sensitivity of $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$.

In the next section, we introduce and derive the AICcd statistic, which in large-sample settings, serves as an approximately unbiased estimator of $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$. This criterion has different goodness-of-fit and penalty terms than AIC, yet both terms reduce to their AIC counterparts when $\mathbf{Y} = \mathbf{Y}_{obs}$. The criterion shares the penalty term of Shimodaira's (1994) PDIO, yet differs in the goodness-of-fit term, where PDIO and AIC agree.

3. Derivation of AICcd

We seek an approximately unbiased estimator of $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$. We will require that the parameter space for the candidate model under consideration includes $\boldsymbol{\theta}_o$ as an interior point. (This strong assumption is also used in the derivation of AIC. See Linhart and Zucchini, 1986, page 245.) We will assume that the fitted parameter vector $\hat{\boldsymbol{\theta}}$ is obtained using the EM algorithm, making $\hat{\boldsymbol{\theta}}$ a maximum likelihood estimator of $\boldsymbol{\theta}_o$. We will require the usual regularity conditions needed to ensure the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$.

Following conventions similar to those of Meng and Rubin (1991), let

$$\begin{aligned}
Q(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2) &= \int_{\mathbf{Y}_{mis}} \{\ln f(\mathbf{Y}|\ \boldsymbol{\theta}_1)\}\ f(\mathbf{Y}_{mis}|\ \mathbf{Y}_{obs}, \boldsymbol{\theta}_2)\ d\mathbf{Y}_{mis}, && (3.1)\\
\mathbf{I}_o(\boldsymbol{\theta} \mid \mathbf{Y}) &= -\frac{\partial^2 \ln f(\mathbf{Y}|\ \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}, && (3.2)\\
\mathbf{I}_o(\boldsymbol{\theta} \mid \mathbf{Y}_{obs}) &= -\frac{\partial^2 \ln f(\mathbf{Y}_{obs}|\ \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}, && (3.3)\\
\mathbf{I}_{oc}(\boldsymbol{\theta} \mid \mathbf{Y}_{obs}) &= \int_{\mathbf{Y}_{mis}} \left\{-\frac{\partial^2 \ln f(\mathbf{Y}|\ \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\}\ f(\mathbf{Y}_{mis}|\ \mathbf{Y}_{obs}, \boldsymbol{\theta})\ d\mathbf{Y}_{mis}. && (3.4)
\end{aligned}$$

To begin, expand $\mathrm{E}_{\mathbf{Y}}\{-2\ln f(\mathbf{Y}|\ \boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ about $\boldsymbol{\theta}_o$ to obtain

$$\mathrm{E}_{\mathbf{Y}}\{-2\ln f(\mathbf{Y}|\ \boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \approx \mathrm{E}_{\mathbf{Y}}\{-2\ln f(\mathbf{Y}|\ \boldsymbol{\theta}_o)\} + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)'\mathrm{E}_{\mathbf{Y}}\{\mathbf{I}_o(\boldsymbol{\theta}_o \mid \mathbf{Y})\}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o). \tag{3.5}$$

Now using (2.8), one can show that

$$\mathrm{E}_{\mathbf{Y}}\{-2\ln f(\mathbf{Y}|\ \boldsymbol{\theta}_o)\} = \mathrm{E}_{\mathbf{Y}_{obs}}\{-2Q(\boldsymbol{\theta}_o \mid \boldsymbol{\theta}_o)\}, \tag{3.6}$$

and that

$$\mathrm{E}_{\mathbf{Y}}\{\mathbf{I}_o(\boldsymbol{\theta}_o \mid \mathbf{Y})\} = \mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\boldsymbol{\theta}_o \mid \mathbf{Y}_{obs})\}. \tag{3.7}$$

Substituting (3.6) and (3.7) into (3.5), we obtain

$$\mathrm{E}_{\mathbf{Y}}\{-2\ln f(\mathbf{Y}|\ \boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \approx \mathrm{E}_{\mathbf{Y}_{obs}}\{-2Q(\boldsymbol{\theta}_o \mid \boldsymbol{\theta}_o)\} + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)'\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\boldsymbol{\theta}_o \mid \mathbf{Y}_{obs})\}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o). \tag{3.8}$$

Yet for large n, it is justifiable to replace $\mathbf{I}_{oc}(\boldsymbol{\theta}_o \mid \mathbf{Y}_{obs})$ in (3.8) with $\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$. This leads to the large-sample approximation

$$\mathrm{E}_{\mathbf{Y}}\{-2\ln f(\mathbf{Y}|\ \boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \approx \mathrm{E}_{\mathbf{Y}_{obs}}\{-2Q(\boldsymbol{\theta}_o \mid \boldsymbol{\theta}_o)\} + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)'\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o). \tag{3.9}$$

Next, expand the first argument of $-2Q(\boldsymbol{\theta}_o \mid \hat{\boldsymbol{\theta}})$ about $\hat{\boldsymbol{\theta}}$ to obtain

$$\begin{aligned} -2Q(\boldsymbol{\theta}_o \mid \hat{\boldsymbol{\theta}}) &\approx -2Q(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}}) - 2(\boldsymbol{\theta}_o - \hat{\boldsymbol{\theta}})' \left\{ \left\{ \frac{\partial Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right\} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\} \\ &\quad + (\boldsymbol{\theta}_o - \hat{\boldsymbol{\theta}})' \left\{ \left\{ -\frac{\partial^2 Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right\} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\} (\boldsymbol{\theta}_o - \hat{\boldsymbol{\theta}}). \end{aligned} \tag{3.10}$$

Now on the right-hand side of (3.10), the second of the three terms is zero, since

$$\frac{\partial Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}. \tag{3.11}$$

(In the EM algorithm, the point of convergence $\hat{\boldsymbol{\theta}}$ provides a solution to the equation $(\partial Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}))/(\partial \boldsymbol{\theta})$ $= \mathbf{0}$.) We can rewrite the third of these three terms by noting that

$$-\frac{\partial^2 Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}). \tag{3.12}$$

Using (3.11) and (3.12), we can express (3.10) as

$$-2Q(\boldsymbol{\theta}_o \mid \hat{\boldsymbol{\theta}}) \approx -2Q(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)'\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o). \tag{3.13}$$

Yet for large n, it is justifiable to replace $\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$ in (3.13) by $\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}$. This leads to the large-sample approximation

$$-2Q(\boldsymbol{\theta}_o \mid \hat{\boldsymbol{\theta}}) \approx -2Q(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)'\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o). \tag{3.14}$$

Now consider using $Q(\boldsymbol{\theta}_o \mid \hat{\boldsymbol{\theta}})$ as an approximation to $Q(\boldsymbol{\theta}_o \mid \boldsymbol{\theta}_o)$ in (3.9). We obtain

$$\mathrm{E}_{\mathbf{Y}}\{-2\ln f(\mathbf{Y}|\,\boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \approx \mathrm{E}_{\mathbf{Y}_{obs}}\{-2Q(\boldsymbol{\theta}_o \mid \hat{\boldsymbol{\theta}})\} + (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)'\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o). \quad (3.15)$$

If we then substitute the right-hand side of (3.14) for $-2Q(\boldsymbol{\theta}_o \mid \hat{\boldsymbol{\theta}})$ in (3.15), we have

$$\begin{aligned}\mathrm{E}_{\mathbf{Y}}\{-2\ln f(\mathbf{Y}|\,\boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \approx \mathrm{E}_{\mathbf{Y}_{obs}}\{-2Q(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})\} + \mathrm{E}_{\mathbf{Y}_{obs}}\left\{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)'\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)\right\}\\ + (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)'\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o). \quad (3.16)\end{aligned}$$

Taking expectations of both sides of (3.16) with respect to $f(\mathbf{Y}_{obs}|\,\boldsymbol{\theta}_o)$ yields the following useful large-sample approximation for $\Delta_{\mathbf{Y}}(d,\boldsymbol{\theta}_o)$:

$$\begin{aligned}\Delta_{\mathbf{Y}}(d,\boldsymbol{\theta}_o) &= \mathrm{E}_{\mathbf{Y}_{obs}}\left\{\mathrm{E}_{\mathbf{Y}}\{-2\ln f(\mathbf{Y}|\,\boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right\}\\ &\approx \mathrm{E}_{\mathbf{Y}_{obs}}\{-2Q(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})\} + 2\mathrm{E}_{\mathbf{Y}_{obs}}\left\{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)'\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)\right\}. \quad (3.17)\end{aligned}$$

Consider the estimation of the two terms on the right-hand side of (3.17). The first of these terms can be estimated by $-2Q(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$, which is easily evaluated after the last iteration of the EM algorithm. For the second of these terms, we will use the well-known fact that the large-sample variance/covariance matrix of $(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)$ is approximated by $\mathbf{I}_o^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$. Thus, we can write

$$\begin{aligned}&2\mathrm{E}_{\mathbf{Y}_{obs}}\{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)'\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)\}\\ &= 2\ \mathrm{trace}\left\{\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}\mathrm{E}_{\mathbf{Y}_{obs}}\{(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)'\}\right\}\\ &\approx 2\ \mathrm{trace}\left\{\mathrm{E}_{\mathbf{Y}_{obs}}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}\ \mathbf{I}_o^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\right\}. \quad (3.18)\end{aligned}$$

A natural estimator for (3.18) is given by

$$2\ \mathrm{trace}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\ \mathbf{I}_o^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}. \quad (3.19)$$

Using $-2Q(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$ and (3.19) to approximate the terms on the right-hand side of (3.17) suggests the following large-sample estimator for $\Delta_{\mathbf{Y}}(d,\boldsymbol{\theta}_o)$:

$$\mathrm{AICcd} = -2Q(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}}) + 2\ \mathrm{trace}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\ \mathbf{I}_o^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}. \quad (3.20)$$

Shimodaira's (1994) PDIO criterion, written in the present notation, has the form

$$\mathrm{PDIO} = -2\ln L(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}) + 2\ \mathrm{trace}\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\ \mathbf{I}_o^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\}. \quad (3.21)$$

Our derivation of AICcd is similar to Shimodaira's derivation of PDIO, yet differs is several key aspects: most noticeably in the development of the goodness-of-fit term. The difference between these terms in AICcd and PDIO causes the criteria to behave quite differently, as the discussion in Section 5 and the simulations in Sections 6 and 7 will indicate.

4. Evaluating AICcd

The penalty term (3.19) of AICcd and PDIO involves the information matrix $\mathbf{I}_o(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$. An explicit expression for $\mathbf{I}_o(\boldsymbol{\theta} \mid \mathbf{Y}_{obs})$ via (3.3) can be difficult to obtain directly, since $f(\mathbf{Y}_{obs} | \, \boldsymbol{\theta})$ is often inaccessible or cumbersome to work with. Fortunately, the "supplemented" EM or SEM algorithm of Meng and Rubin (1991) provides a convenient mechanism for evaluating both $\mathbf{I}_o^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$ and the penalty term (3.19) without the need for such an expression. (Evaluating $\mathbf{I}_o^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$, the approximate large-sample variance/covariance matrix of $\hat{\boldsymbol{\theta}}$, is the motivation behind the SEM algorithm. The EM algorithm alone does not provide this matrix.)

Let $\hat{\boldsymbol{\theta}}^{(t)}$ denote the estimate of $\boldsymbol{\theta}_o$ obtained on the t^{th} iteration of the EM algorithm. As indicated by Meng and Rubin (1991), the EM algorithm defines a mapping $\mathbf{M}(\boldsymbol{\theta}) = (M_1(\boldsymbol{\theta}), \ldots, M_d(\boldsymbol{\theta}))'$ such that $\hat{\boldsymbol{\theta}}^{(t+1)} = \mathbf{M}(\hat{\boldsymbol{\theta}}^{(t)})$ for $t = 0, 1, \ldots$. If $\hat{\boldsymbol{\theta}}^{(t)}$ converges to $\hat{\boldsymbol{\theta}}$ (and $\mathbf{M}(\boldsymbol{\theta})$ is continuous), we must have $\hat{\boldsymbol{\theta}} = \mathbf{M}(\hat{\boldsymbol{\theta}})$. A first-order expansion of $\mathbf{M}(\hat{\boldsymbol{\theta}}^{(t)})$ about $\hat{\boldsymbol{\theta}}$ leads to the approximation

$$(\hat{\boldsymbol{\theta}}^{(t+1)} - \hat{\boldsymbol{\theta}})' \approx (\hat{\boldsymbol{\theta}}^{(t)} - \hat{\boldsymbol{\theta}})' \; \mathbf{DM},$$

where $\mathbf{DM}$ is a $d \times d$ matrix having $\{(\partial M_j(\boldsymbol{\theta}))/(\partial \theta_i)\} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ in row i and column j; i.e.,

$$\mathbf{DM} = \left[\frac{\partial M_j(\boldsymbol{\theta})}{\partial \theta_i} \right] |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad 1 \leq i, j \leq d.$$

Thus, as Meng and Rubin state (1991, page 901), in a neighborhood of $\hat{\boldsymbol{\theta}}$, "the EM algorithm is essentially a linear iteration with rate matrix $\mathbf{DM}$, since $\mathbf{DM}$ is typically nonzero."

Meng and Rubin (1991) go on to show

$$\mathbf{I}_o^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}) = \mathbf{I}_{oc}^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}) \; (\mathbf{I} - \mathbf{DM})^{-1}, \tag{4.1}$$

and

$$\mathbf{I}_o^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}) = \mathbf{I}_{oc}^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}) + \mathbf{I}_{oc}^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\mathbf{DM}(\mathbf{I} - \mathbf{DM})^{-1}. \tag{4.2}$$

(See page 901, (2.3.1), (2.3.4), (2.4.6), and (2.4.7).) This means that the penalty term (3.19) can be written using (4.1) as

$$2 \text{ trace}\{(\mathbf{I} - \mathbf{DM})^{-1}\}, \tag{4.3}$$

or written using (4.2) as

$$2d + 2 \text{ trace}\Big\{\mathbf{I}_{oc}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\{\mathbf{I}_{oc}^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}) \; \mathbf{DM}(\mathbf{I} - \mathbf{DM})^{-1}\}\Big\} \tag{4.4}$$

$$= 2d + 2 \text{ trace}\{\mathbf{DM}(\mathbf{I} - \mathbf{DM})^{-1}\}. \tag{4.5}$$

The computation of $\mathbf{DM}$ is discussed in subsection 3.3 of Meng and Rubin (1991). Once $\mathbf{DM}$ is obtained, the penalty term of AICcd and PDIO can be easily evaluated using (4.3). ($\mathbf{I}_o^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$ is typically computed using (4.2).)

Expression (4.4) is useful for the purpose of comparing the penalty term of AIC ($2d$) to the penalty term of AICcd and PDIO. The matrix $\mathbf{I}_{oc}^{-1}(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})\ \mathbf{DM}(\mathbf{I} - \mathbf{DM})^{-1}$ is described by Meng and Rubin (1991, page 901) as representing "the increase in variance [of $\hat{\boldsymbol{\theta}}$] due to missing information". Thus, the trace term in (4.4) can be conveniently viewed as a measure of the amount of data which is missing in $\mathbf{Y}$; or more precisely, as a measure of the extent to which the missing data $\mathbf{Y}_{mis}$ affects the fitted model. If $\mathbf{Y} = \mathbf{Y}_{obs}$, this trace term will be zero (since $\mathbf{DM} = \mathbf{0}$); otherwise, it will be positive. Moreover, this term will be substantial in settings where the amount of missing data is large relative to the complexity of the fitted model. Thus, (4.4) implies that the penalty term of AICcd and PDIO is composed of the penalty term of AIC together with a term which assesses an additional penalty in accordance to the impact of the missing data on the fitted model.

In discussing the evaluation of AICcd, it is important to note that its goodness-of-fit term is based on the complete-data function $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta})$, which is the principal tool used by the EM algorithm. The evaluation of this term should always be straightforward. The same cannot be said of the goodness-of-fit term of AIC and PDIO, which is based on the incomplete-data log-likelihood $\ln L(\boldsymbol{\theta} \mid \mathbf{Y}_{obs})$. We feel that one of the most compelling features of AICcd is that its computation involves only complete-data quantities which arise naturally in the execution of the EM and SEM algorithms. It is therefore readily accessible in any of the wide variety of incomplete-data problems for which the EM algorithm has been proposed.

5. Contrasting AICcd, PDIO, and AIC

An evaluation of comparable expressions for AICcd, PDIO, and AIC can serve as a starting point to an investigation of the behavior of these criteria. A convenient representation for the goodness-of-fit term of AICcd is obtained by defining

$$H(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2) = \int_{\mathbf{Y}_{mis}} \{\ln f(\mathbf{Y}_{mis} | \ \mathbf{Y}_{obs}, \boldsymbol{\theta}_1)\}\ f(\mathbf{Y}_{mis} | \ \mathbf{Y}_{obs}, \boldsymbol{\theta}_2)\ d\mathbf{Y}_{mis},$$

and by utilizing (3.1) and (2.8) to show that

$$-2Q(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2) = -2H(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2) + \{-2 \ln f(\mathbf{Y}_{obs} | \ \boldsymbol{\theta}_1)\}. \tag{5.1}$$

By (3.20), (3.21), and (2.4), along with (5.1) and representation (4.5) for (3.19), we have

$$\begin{aligned} \text{AICcd} &= \Big\{-2\ln L(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}) + \{-2H(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})\}\Big\} + \Big\{2d + 2\ \text{trace}\{\mathbf{DM}(\mathbf{I}-\mathbf{DM})^{-1}\}\Big\}, \quad (5.2) \\ \text{PDIO} &= -2\ln L(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}) + \Big\{2d + 2\ \text{trace}\{\mathbf{DM}(\mathbf{I}-\mathbf{DM})^{-1}\}\Big\}, \quad (5.3) \\ \text{AIC} &= -2\ln L(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs}) + 2d. \quad (5.4) \end{aligned}$$

Each of the preceding criteria is comprised of the sum of a goodness-of-fit term and a penalty term. As the fitted model becomes more complex, the penalty term increases, whereas the goodness-of-fit term tends to decrease. (The latter behavior is a reflection of the improvement in fit which results from using larger, more flexible models.) Ideally, the fitted model which provides the optimal balance between fidelity to the data and parsimony is identified by the minimum criterion value.

The goodness-of-fit term of AIC and PDIO, $-2\ln L(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$, measures the conformity of the observed data $\mathbf{Y}_{obs}$ to the fitted model $f(\mathbf{Y}_{obs}|\ \hat{\boldsymbol{\theta}})$. This term as well as the additional component $-2H(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$ comprise the goodness-of-fit term of AICcd. The component $-2H(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$ measures the conformity of the missing data $\mathbf{Y}_{mis}$ to the fitted model $f(\mathbf{Y}_{mis}|\ \mathbf{Y}_{obs}, \hat{\boldsymbol{\theta}})$ in the following sense: if many realizations of $\mathbf{Y}_{mis}$ were generated according to the density $f(\mathbf{Y}_{mis}|\ \mathbf{Y}_{obs}, \hat{\boldsymbol{\theta}})$, and the goodness-of-fit measure $-2\ln f(\mathbf{Y}_{mis}|\ \mathbf{Y}_{obs}, \hat{\boldsymbol{\theta}})$ was averaged over these realizations, this average would approximate $-2H(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$.

First, we consider the relative behavior of PDIO and AIC. These criteria share the same goodness-of-fit term, yet a comparison of (5.3) and (5.4) indicates that the penalty term of PDIO is always at least as large as the penalty term of AIC. (Recall from the discussion in Section 4 that $\text{trace}\{\mathbf{DM}(\mathbf{I}-\mathbf{DM})^{-1}\}$ is always nonnegative, and is positive when $\mathbf{Y} \neq \mathbf{Y}_{obs}$.) Thus, PDIO will always choose a fitted model in a candidate class which is no larger than the model chosen by AIC. This implies that PDIO is more prone than AIC to underfitting (i.e., to choosing a model of lower dimension than the generating model), whereas AIC is more prone than PDIO to overfitting (i.e., to choosing a model of higher dimension than the generating model). These tendencies become more extreme as the amount of missing data increases, since the trace component in the penalty term of PDIO grows in relation to the amount of missing information.

We next consider the relative behavior of AICcd and PDIO. These criteria share the same penalty term, but an inspection of (5.2) and (5.3) reveals that the goodness-of-fit terms differ by the component $-2H(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$. In applications where the degree of missing information is large relative to the degree of complete information, the penalty term of PDIO often dominates its goodness-of-fit term; as a result, the criterion may tend to underfit excessively. (This behavior is exhibited in the simulations reported by Shimodaira, 1994.) The additional component $-2H(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$ in the goodness-of-fit term of AICcd counteracts this underfitting tendency, thus providing AICcd with a certain level of protection against choosing models which are too small. This tendency will be illustrated by the simulations which follow in Sections 6 and 7.

Finally, we consider the relative behavior of AICcd and AIC. By comparing (5.2) and (5.4), we note that AICcd contains extra components in both the goodness-of-fit term and the penalty term, each of which involve the missing information. Note that the component $-2H(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$ provides a missing-data supplement to the goodness-of-fit term of AIC in the same way that the component $2\ \mathrm{trace}\{\mathbf{DM}(\mathbf{I} - \mathbf{DM})^{-1}\}$ provides a missing-data supplement to the penalty term of AIC. It would be difficult to give a general characterization of the contribution made by the sum of these components to AICcd. However, extensive simulation results (including those which follow) indicate that in settings where the criteria do not exhibit similar selection behavior, AICcd generally tends to overfit to a lesser degree than AIC. In such instances, AICcd may tend to underfit more frequently than AIC, yet rarely to the same extent as PDIO.

To summarize, the additional penalty term which AICcd and PDIO share over AIC, $2\ \mathrm{trace}\{\mathbf{DM}(\mathbf{I} - \mathbf{DM})^{-1}\}$, penalizes a fitted model in accordance to the impact the missing data has on the model. Thus, in applications where the observed data is incomplete, AICcd and PDIO often favor lower-dimensional models than AIC. Yet unlike AICcd, PDIO does not contain the additional goodness-of-fit term $-2H(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$. The inclusion of this term attenuates the effect of the extra penalty term, and protects AICcd from the type of excessive underfitting which PDIO may exhibit when there exists a significant amount of missing information.

6. Simulations: Modeling Bivariate Normal Data

Let μ_1, μ_2, σ_1^2, σ_2^2, and σ_{12} denote, respectively, the two means, two variances, and covariance for a general bivariate normal model.

Suppose we collect a data set consisting of observations on a pair of random variables (y_1, y_2). To model this data, we consider a candidate class consisting of four types of bivariate normal models corresponding to certain parameter constraints. The constraints, along with the implied dimensions and estimation requirements of the associated models, are as follows:

Dimension	Parameter Constraints	Parameters to be Estimated
5	None	$\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12}$
4	$\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$	$\mu_1, \mu_2, \sigma^2, \sigma_{12}$
3	$\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$, $\mu_1 = \mu_2 \equiv \mu$	$\mu, \sigma^2, \sigma_{12}$
2	$\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$, $\mu_1 = \mu_2 \equiv \mu$, $\sigma_{12} = 0$	μ, σ^2

In each of our simulation sets, 1000 samples of size 50 are generated using a known bivariate normal model in the preceding candidate class. In some sets, certain data pairs within each sample are made incomplete by eliminating either the first or the second observation. Whether a data pair is made incomplete is determined at random according to specified discard probabilities. We will use $\Pr(y_1 \text{ mis})$ to denote the probability that the first observation is discarded and the second is retained, and $\Pr(y_2 \text{ mis})$ to denote the probability that the second observation is discarded and the first is retained. (Thus, in a simulation set where $\Pr(y_1 \text{ mis})$ and $\Pr(y_2 \text{ mis})$ are both set at 0.30, for each sample of size 50, one would expect roughly 15 pairs where y_1 is missing but y_2 is observed, 15 pairs where y_1 is observed but y_2 is missing, and 20 pairs where y_1 and y_2 are both observed.)

For each of the 1000 samples in a set, all four models in the candidate class are fit to the data using the SEM algorithm; the criteria AIC, PDIO, and AICcd are evaluated using (2.4), (3.21), and (3.20); and the candidate model selected by each criterion is determined. The distribution of selections by each criterion is recorded for the 1000 samples and presented in Table 1. In this table, the dimension of the generating model is listed in the second column, and the discard probabilities are listed in the third column.

We include four simulation sets for each of the following generating models:

Set	True	True		
Numbers	Dimension	Parameter Values		
1–4	3	$\mu_1 = \mu_2 \equiv \mu = 0$,	$\sigma_1^2 = \sigma_2^2 \equiv \sigma^2 = 10$,	$\sigma_{12} = 6$
5–8	3	$\mu_1 = \mu_2 \equiv \mu = 0$,	$\sigma_1^2 = \sigma_2^2 \equiv \sigma^2 = 10$,	$\sigma_{12} = 8$
9–12	4	$\mu_1 = 0$, $\mu_2 = 2$,	$\sigma_1^2 = \sigma_2^2 \equiv \sigma^2 = 10$,	$\sigma_{12} = 6$
13–16	4	$\mu_1 = 0$, $\mu_2 = 2$,	$\sigma_1^2 = \sigma_2^2 \equiv \sigma^2 = 10$,	$\sigma_{12} = 8$

For the first of the four sets corresponding to a generating model, none of the data is discarded. For the second, third, and fourth sets, the discard probabilities $\Pr(y_1 \text{ mis})$ and $\Pr(y_2 \text{ mis})$ are both set at 0.15, 0.30, and 0.40, respectively.

When none of the data is missing, AIC, PDIO and AICcd are all equivalent, and therefore all yield the same selection results. As the discard probabilities are increased, differences in the behavior of the criteria become more apparent. The simulation results support the following conclusions.

(i) The selection performance of the criteria improves as the correlation between y_1 and y_2 is increased. Each criterion performs more effectively in sets 5 through 8 than in sets 1 through 4, and more effectively in sets 13 through 16 than in sets 9 through 12. In the sets where data is discarded, this behavior can be easily explained. When the correlation is high, incomplete data pairs are less costly since it is possible to accurately impute the missing elements. All criteria should benefit in such a setting.

(ii) In every simulation set where data is discarded, AICcd underfits to a lesser degree than PDIO, and overfits to a comparable or to a slightly lesser degree than AIC.

(iii) As the discard probabilities are increased, PDIO becomes more prone towards selecting lower dimensional models, which results in excessive underfitting in sets 4, 8, 12, and 16. AICcd exhibits this propensity to a much lesser extent.

As mentioned in Section 5, as the discard probabilities are increased, the trace component $2\,\mathrm{trace}\{\mathbf{DM}(\mathbf{I}-\mathbf{DM})^{-1}\}$ in the penalty term (4.5) tends to increase. The goodness-of-fit term of PDIO, $-2\ln L(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$, does not compensate for this behavior and as a result, becomes increasingly less competitive with the penalty term. This causes PDIO to become more prone towards selecting lower dimensional models, which results in excessive underfitting in sets 4, 8, 12, and 16. The additional goodness-of-fit component in AICcd, $-2H(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$, provides protection against this tendency.

Figures 1 through 5 provide some insight into the nature of the expected complete-data discrepancy $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$, the expected incomplete-data discrepancy $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$, and the estimators of these quantities provided by AICcd, PDIO, and AIC.

Figures 1, 2, and 3 illustrate the changes which occur in $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ and $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ as the discard probabilities are increased. The samples from sets 1 through 4 are used in simulating the expected discrepancies. To serve as a reference in each figure, the simulated $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ curve from set 1 (based on no missing data) is plotted against the candidate model dimensions $d = 2, 3, 4, 5$. The simulated $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ and $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ curves from sets 2, 3, and 4 are overlaid in Figures 1, 2, and 3, respectively. (For comparison purposes, each curve is translated so that its minimum at d

$= 3$ is set at zero. The curves are then scaled by dividing each value by the difference between the maximum and the minimum of the reference curve, $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ from set 1.)

As the discard probabilities are increased, the values of $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ and $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ decrease for $d = 2$ and increase for $d = 4, 5$. However, $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ decreases to a lesser extent than $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ for $d = 2$, and increases to a greater extent for $d = 4, 5$. Note that as a result, the minimum of the $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ curve becomes sharper and better defined, whereas the minimum of the $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ curve becomes less pronounced. This phenomenon suggests that in the presence of incomplete data, an estimator of $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ may be preferable to an estimator of $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ for the purpose of model selection, provided that the former adequately reflects the discriminatory behavior of $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$.

Figure 4 illustrates how effectively AICcd and PDIO serve as approximately unbiased estimators of $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$, and Figure 5 illustrates how effectively AIC serves as an approximately unbiased estimator of $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$. Figure 4 features the simulated $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ curve for simulation set 3 plotted for $d = 2, 3, 4, 5$ along with the curves which represent the average values of AICcd and PDIO. In Figure 5, the simulated $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ curve for set 3 is plotted for $d = 2, 3, 4, 5$ along with the curve which represents the average values of AIC. (Each curve has been translated so that its minimum at $d = 3$ is set at zero. The curves in Figure 4 are then scaled by dividing each value by the difference between the maximum and the minimum of the reference $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ curve. The curves in Figure 5 are similarly scaled using the reference $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ curve.)

Note that the average AICcd and AIC curves respectively follow the simulated $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ and $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ curves to a comparable degree. Both the AICcd and AIC curves exhibit a more gradual slope than the corresponding expected discrepancy curves over $d = 3, 4, 5$. The average PDIO curve tracks $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ more effectively than the AICcd curve over these dimensions, yet assumes a value at $d = 2$ which is much lower than $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$. This type of behavior is more pronounced in simulation sets with higher discard probabilities, and helps to illustrate why the PDIO criterion is often prone to underfitting.

The preceding section of simulations considers the performance of the criteria in a simplistic, illustrative setting, albeit one where the use of model selection criteria would not be typically employed. In the next section, we consider the behavior of the criteria in a more practical, traditional framework.

Table 1. Dimension Selections for Bivariate Normal Simulations

			Dimension Selections											
Set	True	Pr(y_1 mis),	AIC				PDIO				AICcd			
	Dim.	Pr(y_2 mis)	2	3	4	5	2	3	4	5	2	3	4	5
1	3	0.00, 0.00	1	799	118	82	1	799	118	82	1	799	118	82
2	3	0.15, 0.15	10	776	123	91	18	849	84	49	11	783	119	87
3	3	0.30, 0.30	51	718	129	102	213	730	35	22	54	714	127	105
4	3	0.40, 0.40	207	573	119	101	739	252	5	4	193	605	109	93
5	3	0.00, 0.00	0	813	122	65	0	813	122	65	0	813	122	65
6	3	0.15, 0.15	0	800	130	70	0	891	77	32	0	797	129	74
7	3	0.30, 0.30	0	791	131	78	11	942	38	9	0	783	139	78
8	3	0.40, 0.40	16	735	143	106	389	600	10	1	15	738	149	98
9	4	0.00, 0.00	0	0	850	150	0	0	850	150	0	0	850	150
10	4	0.15, 0.15	1	0	844	155	1	3	882	114	1	2	846	151
11	4	0.30, 0.30	8	11	830	151	108	39	794	59	17	13	812	158
12	4	0.40, 0.40	56	32	738	174	672	38	277	13	105	85	660	150
13	4	0.00, 0.00	0	0	860	140	0	0	860	140	0	0	860	140
14	4	0.15, 0.15	0	0	852	148	0	0	905	95	0	0	863	137
15	4	0.30, 0.30	0	0	829	171	10	9	934	47	0	0	835	165
16	4	0.40, 0.40	6	7	807	180	461	65	465	9	11	30	789	170

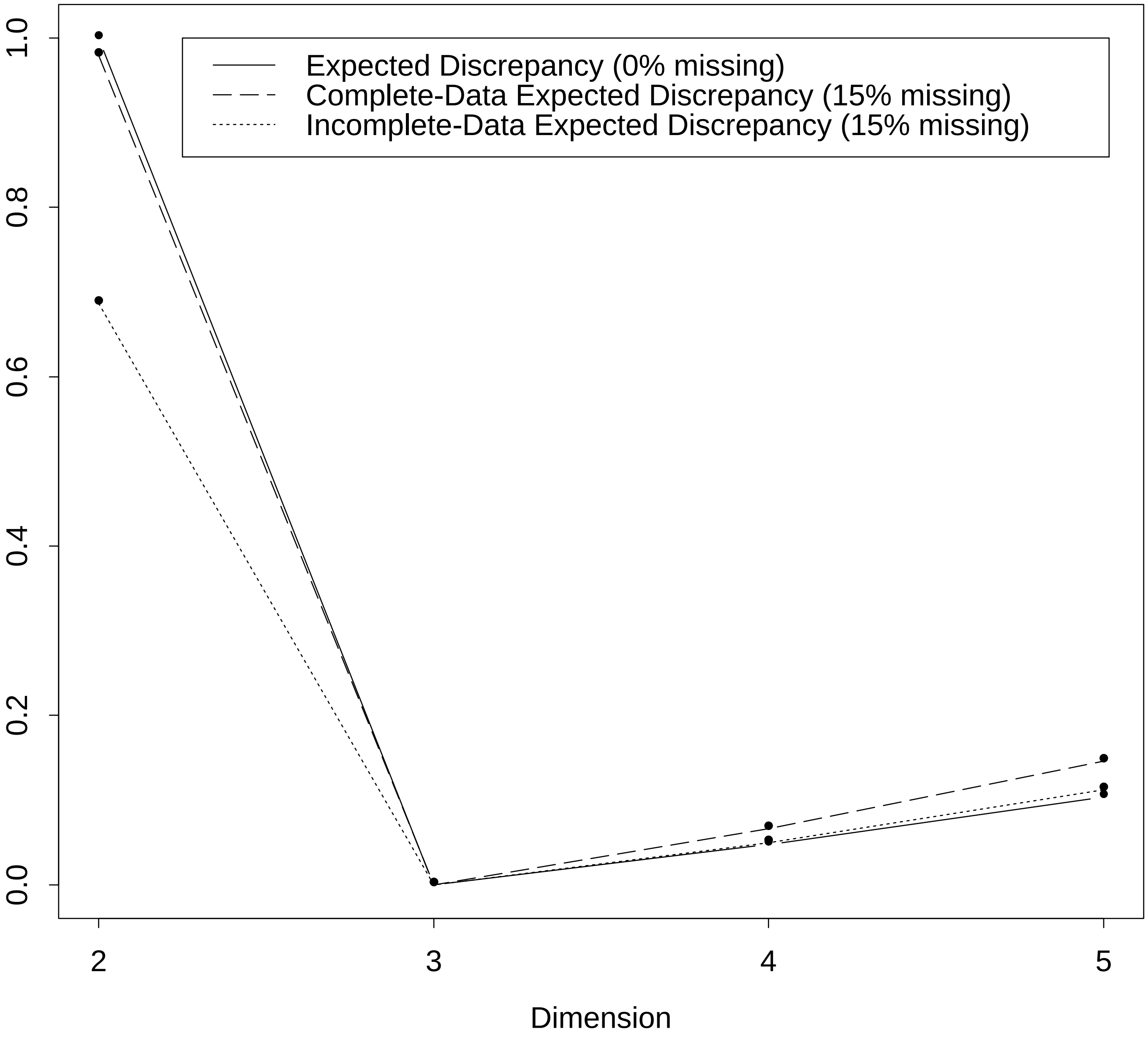

Figure 1. Simulated $\triangle_{\mathbf{Y}}(d,\boldsymbol{\theta}_o)$ and $\triangle_{\mathbf{Y}_{obs}}(d,\boldsymbol{\theta}_o)$ Curves
(bivariate normal simulation sets 1, 2)

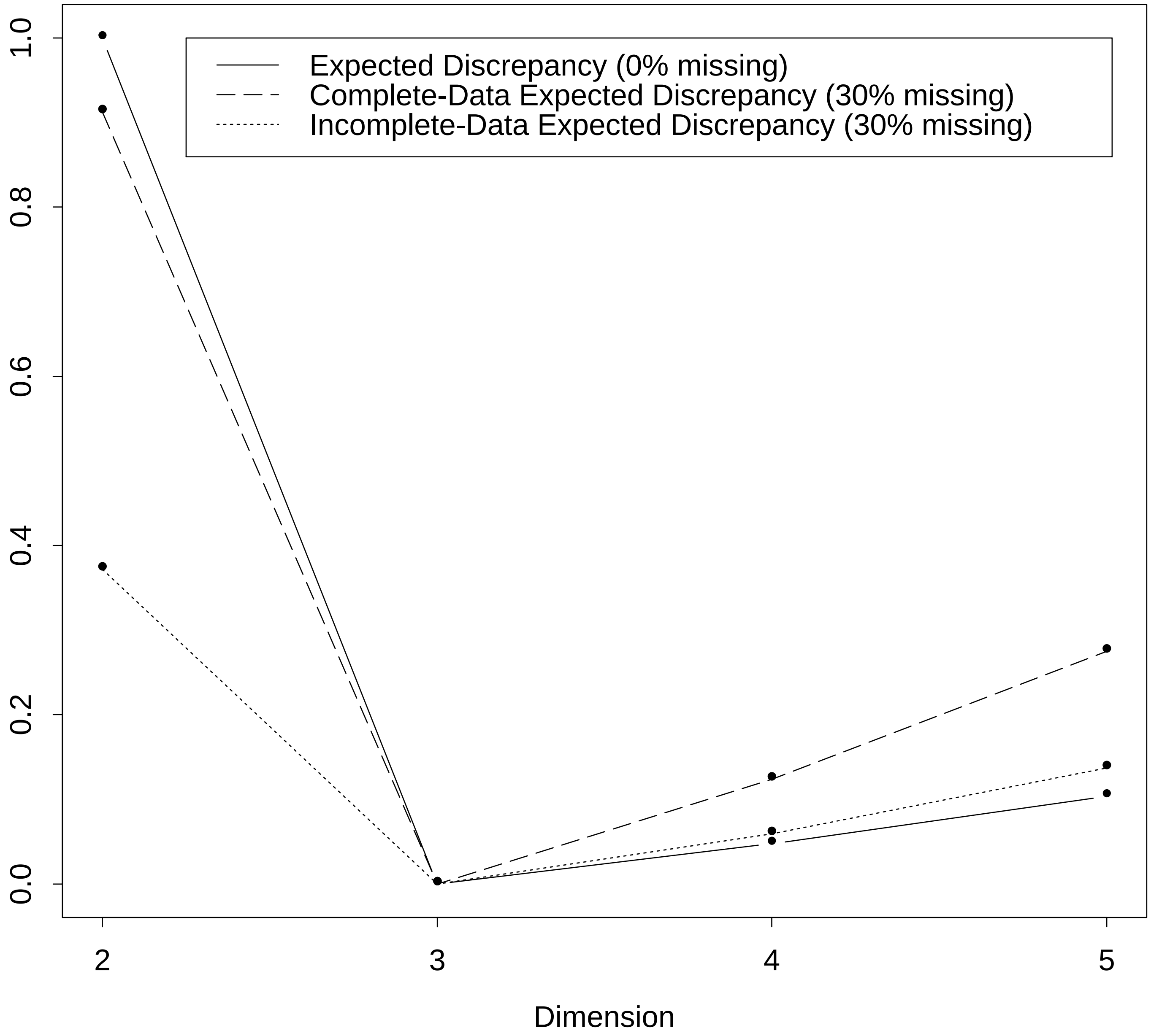

Figure 2. Simulated $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ and $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ Curves
(bivariate normal simulation sets 1, 3)

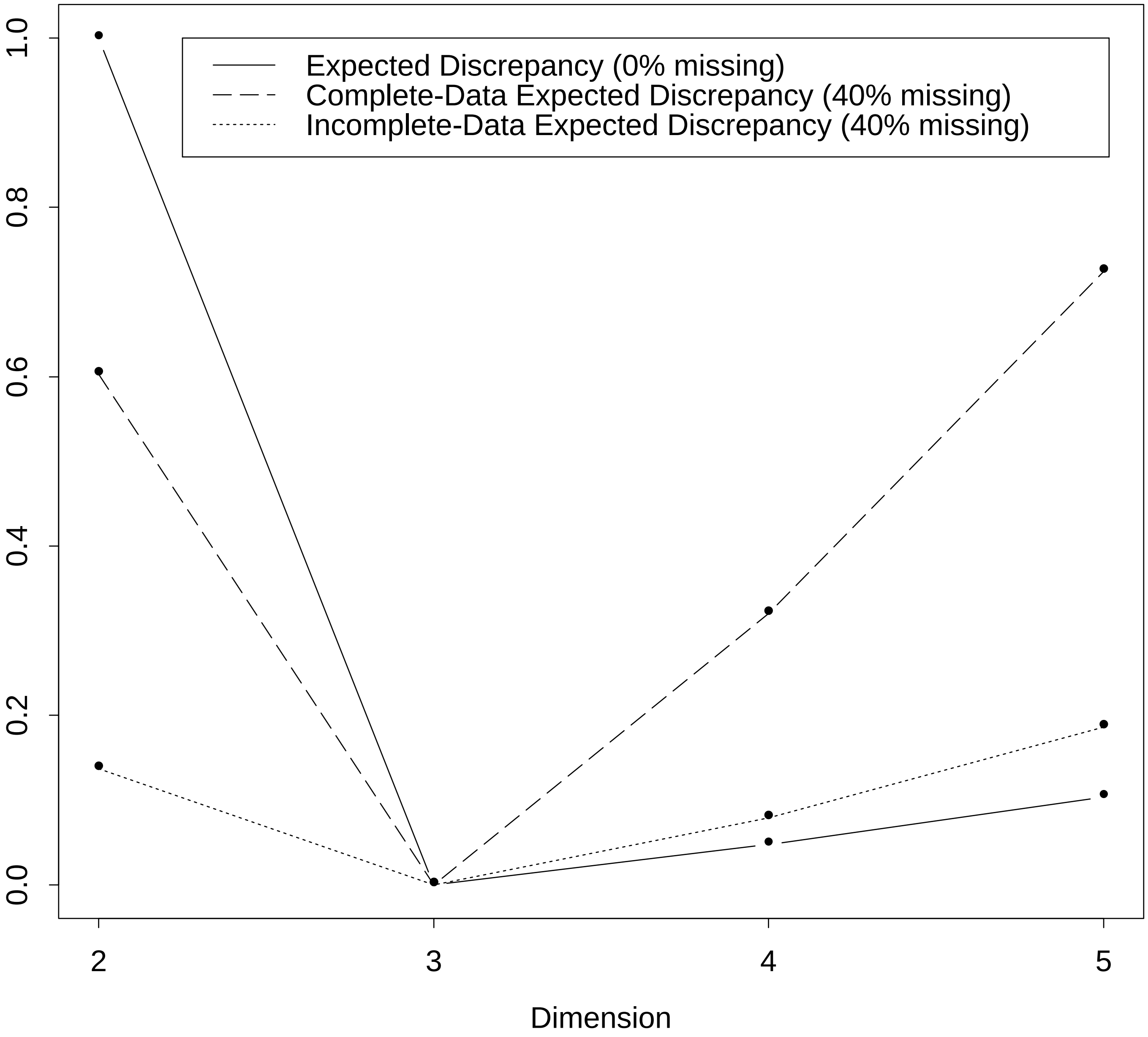

Figure 3. Simulated $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ and $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ Curves (bivariate normal simulation sets 1, 4)

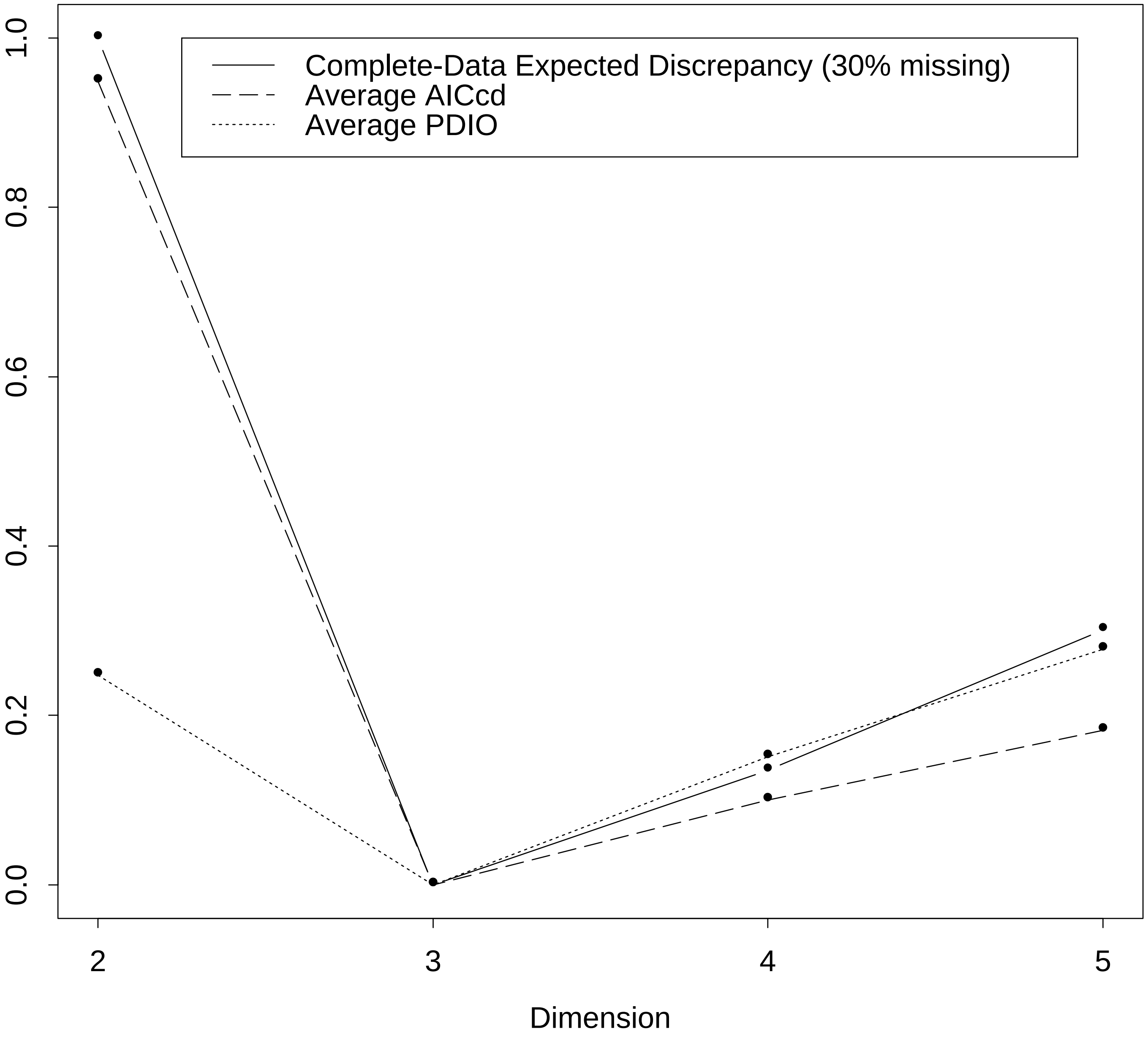

Figure 4. Simulated $\Delta_{\mathbf{Y}}(d, \boldsymbol{\theta}_o)$ Curve and Average AICcd, PDIO Curves (bivariate normal simulation set 3)

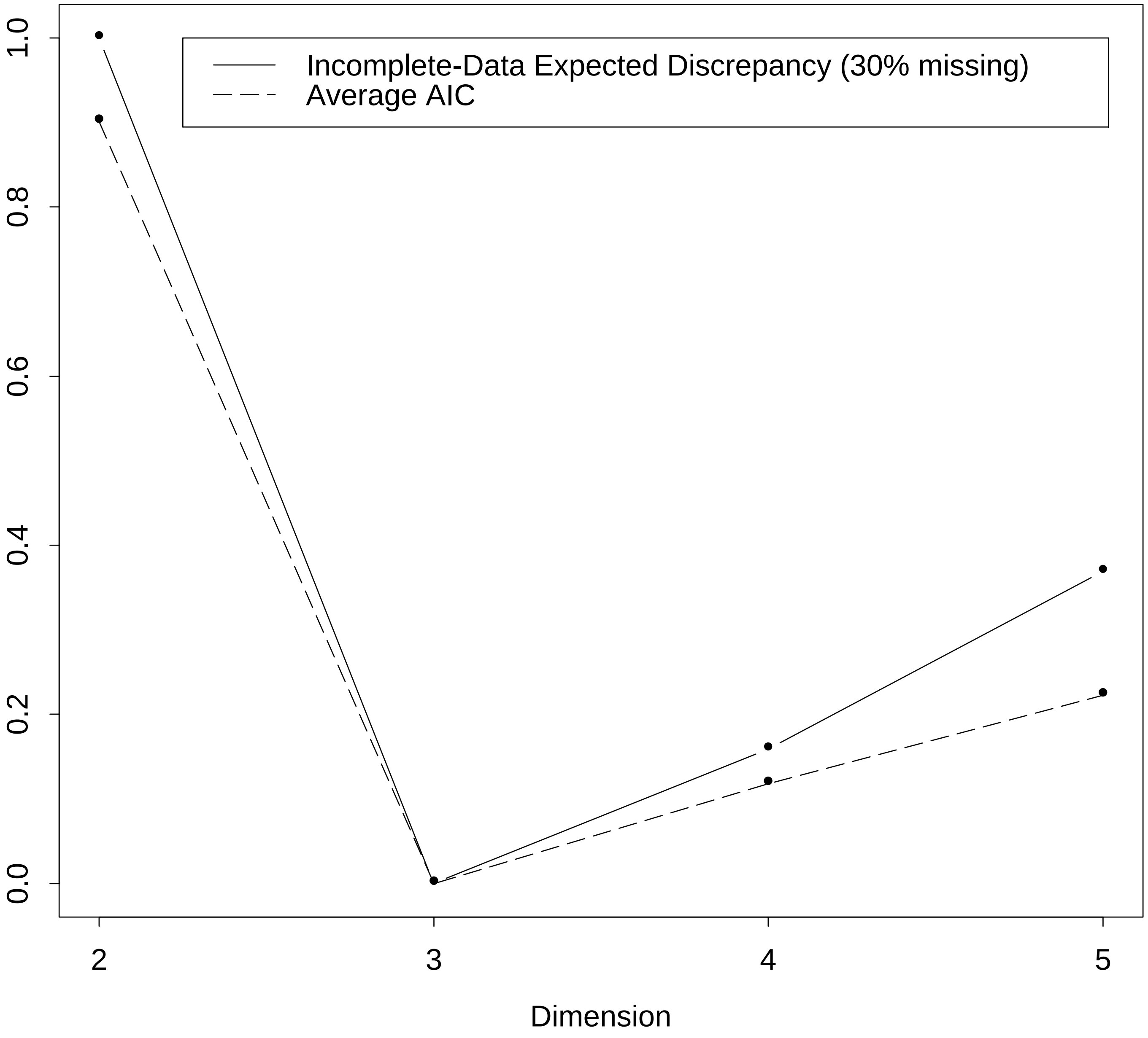

Figure 5. Simulated $\Delta_{\mathbf{Y}_{obs}}(d, \boldsymbol{\theta}_o)$ Curve
and Average AIC Curve
(bivariate normal simulation set 3)

7. Simulations: Multivariate Regression Modeling

Consider the multivariate regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U},$$

where the $n \times p$ matrix $\mathbf{Y}$ consists of rows which are independent, p-dimensional random vectors, the $n \times m$ matrix $\mathbf{X}$ is a known design matrix of covariate values, and the $m \times p$ matrix $\boldsymbol{\beta}$ is an unknown matrix of regression parameters. The $n \times p$ matrix $\mathbf{U}$ is comprised of rows which are independent p-variate normal random vectors, each with a mean vector of zero and a variance/covariance matrix $\boldsymbol{\Sigma}$.

One of the most important problems in regression modeling is that of choosing the number of predictors to include in the model; i.e., of determining the size of the design matrix $\mathbf{X}$. If m regressors are retained for a candidate regression model, the overall dimension of the model is given by $d = mp + p(p+1)/2$.

We consider a setting where $p = 2$, so that the rows of $\mathbf{Y}$ represent bivariate data pairs. The design matrices $\mathbf{X}$ for our class of candidate models range in size from $m = 1$ column to $m = 8$ columns, representing models of dimension $d = 5$ through $d = 19$. We assume that the candidate models are nested; i.e., if $\mathbf{X}_1$ has m_1 columns and $\mathbf{X}_2$ has m_2 columns where $m_1 < m_2$, the columns of $\mathbf{X}_1$ comprise the first m_1 columns of $\mathbf{X}_2$.

The first column of each $\mathbf{X}$ is taken to be a vector of ones. The covariate values are generated by taking independent measurements on a random variable having a uniform distribution on the interval (0,5). Setting up the design matrices in this simplistic fashion ensures that the simulation results are not unduly influenced by such factors as multicollinearity and high-leverage cases.

For each simulation set, 1000 response matrices $\mathbf{Y}$ of dimension 50×2 are generated from a model having the form

$$\mathbf{Y} = \mathbf{X}_o\boldsymbol{\beta}_o + \mathbf{U}, \quad \text{where} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 7 \\ 7 & 16 \end{bmatrix}.$$

Thus, the response variable represented in the first column of $\mathbf{Y}$, y_1, is much less variable than the response variable represented in the second column of $\mathbf{Y}$, y_2.

Three different parameter matrices $\boldsymbol{\beta}_o$ are used. For each $\boldsymbol{\beta}_o$, a collection of five simulation sets are run with the pair of discard probabilities $(\Pr(y_1 \text{ mis}), \Pr(y_2 \text{ mis}))$ set at $(0.00, 0.00)$, $(0.00, 0.60)$, $(0.20, 0.40)$, $(0.40, 0.20)$, and $(0.60, 0.00)$. The $m_o \times 2$ parameter matrices $\boldsymbol{\beta}_o$ have all elements set equal to 1, and have sizes determined by $m_o = 3$ ($d_o = 9$), $m_o = 5$ ($d_o = 13$), and $m_o = 7$ ($d_o = 17$). The fifteen sets corresponding to these three $\boldsymbol{\beta}_o$ are labeled 1 - 5, 6 - 10, and 11 - 15, respectively.

For each of the 1000 samples in a set, all 8 models in the candidate class are fit to the data using the SEM algorithm; the criteria AIC, PDIO, and AICcd are evaluated using (2.4), (3.21), and (3.20); and the candidate model selected by each criterion is determined. The distribution of selections by each criterion is recorded for the 1000 samples and presented in Table 2. In this table, the dimension of the generating model d_o is listed in the second column, and the discard probabilities for the samples are listed in the third column. For brevity, we group the dimension selections into three categories: "$< d_o$" (underfitting), "d_o" (correct dimension), and "$> d_o$" (overfitting).

The results of the simulations support the following conclusions.

(i) As the dimension of the generating model d_o is increased, each criterion tends to become more prone towards underfitting and less prone towards overfitting.

(ii) In every simulation set where data is discarded, AICcd overfits to a lesser degree than AIC and underfits to a lesser degree than PDIO. Moreover, although PDIO often obtains more correct dimension selections than AICcd, AICcd maintains the greatest level of consistency as a selection criterion. In sets where AIC demonstrates a strong propensity towards overfitting (e.g., 7), and in sets where PDIO demonstrates a strong propensity towards underfitting (e.g., 15), AICcd exhibits these tendencies to a much lesser extent.

(iii) In the simulation sets where data is discarded, $\Pr(y_1 \text{ mis}) + \Pr(y_2 \text{ mis})$ is held constant at 0.60. Yet as $\Pr(y_1 \text{ mis})$ is increased and $\Pr(y_2 \text{ mis})$ is decreased, PDIO becomes more prone towards selecting lower dimensional models; this results in excessive underfitting in sets such as 5, 10, and 15. AICcd exhibits this tendency to a much lesser extent.

Since $\text{Var}(y_1) = 4$ and $\text{Var}(y_2) = 16$, increasing $\Pr(y_1 \text{ mis})$ and decreasing $\Pr(y_2 \text{ mis})$ results in discarding a larger percentage of the less variable data and retaining a higher percentage of the noisier data. This causes the goodness-of-fit term of PDIO, $-2 \ln L(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}_{obs})$, to become less effective against its penalty term, which will be large since $\Pr(y_1 \text{ mis}) + \Pr(y_2 \text{ mis})$ is substantial. As a result, PDIO increasingly favors lower dimensional models. In sets such as 5, 10, and 15, this tendency results in PDIO underfitting to an excessive degree. Here again, the additional goodness-of-fit component in AICcd, $-2H(\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}})$, provides protection against this behavior.

Table 2. Dimension Selections for Multivariate Regression Simulations

			Dimension Selections								
Set	True	Pr(y_1 mis),		AIC			PDIO			AICcd	
	Dim.: d_o	Pr(y_2 mis)	$< d_o$	d_o	$> d_o$	$< d_o$	d_o	$> d_o$	$< d_o$	d_o	$> d_o$
1	9	0.00, 0.00	0	713	287	0	713	287	0	713	287
2	9	0.00, 0.60	1	606	393	31	912	57	13	722	265
3	9	0.20, 0.40	3	618	379	41	904	55	13	724	263
4	9	0.40, 0.20	2	636	362	76	874	50	21	739	240
5	9	0.60, 0.00	9	611	380	188	768	44	45	697	258
6	13	0.00, 0.00	0	710	290	0	710	290	0	710	290
7	13	0.00, 0.60	0	598	402	72	885	43	31	759	210
8	13	0.20, 0.40	0	628	372	85	862	53	27	758	215
9	13	0.40, 0.20	3	636	361	112	837	51	23	771	206
10	13	0.60, 0.00	8	598	394	305	660	35	84	709	207
11	17	0.00, 0.00	0	792	208	0	792	208	0	792	208
12	17	0.00, 0.60	0	697	303	219	760	21	127	725	148
13	17	0.20, 0.40	0	731	269	177	782	41	73	778	149
14	17	0.40, 0.20	9	726	265	289	680	31	89	778	133
15	17	0.60, 0.00	17	685	298	462	509	29	157	704	139

8. Conclusion

In Sections 2 through 5, we derived and discussed an analogue of AIC for model selection in applications where the observed data is incomplete. Our criterion estimates the expected complete-data Kullback-Leibler discrepancy in the same manner that Akaike's (1973, 1974) AIC estimates the expected incomplete-data discrepancy.

AIC lacks the property of consistency, but is *asymptotically efficient* in the sense of Shibata (1980), which is arguably a property of greater practical value. (See, for example, Hurvich and Tsai, 1989; Bhansali, 1993.) AICcd should possess the same asymptotic properties as AIC under the assumption that the proportion of missing information to complete information tends to zero as the degree of complete information (i.e., the overall sample size) tends to infinity. Establishing the properties of AICcd when this assumption is not met is a topic for future investigation.

As a model selection criterion, AIC performs effectively in a large variety of applications. However, recent work has shown that in settings where the sample size is small relative to the dimension of the largest model in the candidate class, AIC provides an estimator of the expected discrepancy which is significantly negatively biased. "Corrected" variants of AIC which compensate for this small-sample bias have been developed for such applications: see Hurvich and Tsai (1989); Hurvich, Shumway, and Tsai (1990); and Bedrick and Tsai (1994). In future work, we hope to develop analogous "corrected" variants of AICcd, since AICcd itself will exhibit substantial negative bias in the type of settings previously mentioned.

Our simulations in Sections 6 and 7 indicate that in the presence of incomplete data, AICcd tends to underfit to a lesser degree than PDIO, and tends to overfit to a lesser degree than AIC. AICcd achieves the latter by incorporating a penalization for missing information which is lacking in AIC; it achieves the former by incorporating a goodness-of-fit term for missing information which is lacking in PDIO.

AICcd is based entirely on complete-data tools. Unlike AIC and PDIO, it does not require the evaluation of the observed-data empirical log-likelihood, which may be problematic or burdensome to compute. Thus, AICcd can be easily evaluated in the framework of the SEM algorithm without any additional programming. This important property of AICcd, along with its promising performance in our simulation sets, will hopefully encourage the usage and further investigation of this criterion as well as others based on complete-data tools and principles.

Acknowledgements

We wish to extend our sincere appreciation to an associate editor and a referee for carefully reading the original version of this manuscript, and for preparing helpful and constructive critiques which served to greatly improve the exposition and content. We also extend our thanks to the editor.

The work of the first author was supported by grants from the Research Board of the University of Missouri and the Research Council of the University of Missouri-Columbia.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov and F. Csaki, Eds., *Second International Symposium on Information Theory*, Akademia Kiado, Budapest, 267–281.

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

Bedrick, E. J. and C. L. Tsai (1994). Model selection for multivariate regression in small samples. *Biometrics* **50**, 226–231.

Bhansali, R. J. (1993). Order selection for linear time series models: A review. In: T. S. Rao, Ed., *Developments in Time Series Analysis*, Chapman and Hall, London, 50–66.

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

Goodman, L. A. (1974). Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika* **61**, 315–331.

Hurvich, C. M., R. H. Shumway and C. L. Tsai (1990). Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. *Biometrika* **77**, 709–719.

Hurvich, C. M. and C. L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Kullback, S. (1968). *Information Theory and Statistics.* Dover, New York.

Linhart, H. and W. Zucchini (1986). *Model Selection.* John Wiley, New York.

Little, R. J. A. (1979). Maximum likelihood inference for multiple regression with missing values: A simulation study. *Journal of the Royal Statistical Society, Series B* **41**, 76–87.

Little, R. J. A. and D. B. Rubin (1987). *Statistical Analysis with Missing Data.* John Wiley, New York.

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

Rubin, D. B. (1976). Noniterative least squares estimates, standard errors and F-tests for any analysis of variance with missing data. *Journal of the Royal Statistical Society, Series B* **38**, 270–274.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* **80**, 147–164.

Shimodaira, H. (1994). A new criterion for selecting models from partially observed data. In: P. Cheeseman and R. W. Oldford, Eds., *Selecting Models from Data: Artificial Intelligence and Statistics IV, Lecture Notes in Statistics* **89**, Springer-Verlag, New York, 21–29.

Shumway, R. H. and D. S. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* **3**, 253–264.

Titterington, D. M., A. F. M. Smith and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley, New York.