Comput Stat (2016) 31:643–669
DOI 10.1007/s00180-015-0584-8

ORIGINAL PAPER

A multistage algorithm for best-subset model selection based on the Kullback–Leibler discrepancy

Tao Zhang[1] · **Joseph E. Cavanaugh**[2]

Received: 11 October 2013 / Accepted: 10 April 2015 / Published online: 24 April 2015

Abstract The selection of a best-subset regression model from a candidate family is a common problem that arises in many analyses. The Akaike information criterion (AIC) and the corrected AIC (AIC$_c$) are frequently used for this purpose. AIC and AIC$_c$ are designed to estimate the expected Kullback–Leibler discrepancy. For best-subset selection, both AIC and AIC$_c$ are negatively biased, and the use of either criterion will lead to the selection of overfitted models. To correct for this bias, we introduce an "improved" AIC variant, AIC$_i$, which has a penalty term evaluated using Monte Carlo simulation. A multistage model selection procedure AIC$_{\text{aps}}$, which utilizes AIC$_i$, is proposed for best-subset selection. Simulation studies are compiled to compare the performances of the different model selection methods.

Keywords Akaike information criterion · Linear models · Monte Carlo simulation · Variable selection

1 Introduction

The selection of a best-subset regression model from a candidate family is a common analytical problem. In best-subset model selection, we consider all possible subsets (APS) of regressor variables; thus, numerous candidate models may need to be fit and compared. One of the main challenges of best-subset selection arises from the size of the candidate model family: specifically, the probability of selecting an inappropriate

✉ Tao Zhang
tao.zhang@boehringer-ingelheim.com

[1] Biostatistics, Boehringer Ingelheim, 29/F, Park Place, 1601 Nanjing Road (West), Shanghai 200040, China

[2] Department of Biostatistics, The University of Iowa, Iowa City, IA, USA

model generally increases as the size of the family grows. For this reason, it is usually difficult to select an optimal model when best-subset selection is attempted based on a moderate to large number of regressor variables.

If one is inclined to assume that a true model exists, then the goal of model selection is to search among a candidate family to find a model that is "closest" to the true model. The notion of closeness is quantified by a measure that reflects the disparity between each fitted candidate model and the true model. Such a measure is called a *discrepancy*. One of the most popular discrepancies is the Kullback–Liebler (K–L) discrepancy, also known as the K–L (1951) information. The K–L discrepancy is applicable in nearly all parametric frameworks, and because of its close connection to likelihood-based principles, is considered one of the most important discrepancies in model selection.

In practice, it is impossible to evaluate the exact value of a discrepancy, since such a measure depends on the generating model. However, under appropriate conditions, one can often formulate a statistic to estimate a discrepancy. Such a statistic may be used as a model selection criterion. A model selection criterion is designed to reflect the propriety of a fitted candidate model. It reflects both the conformity of the fitted model to the data, and the competing objective of model parsimony. If the value of the criterion is small, then the objectives of conformity and parsimony are well balanced. Such a candidate model will often satisfy the attribute of generalizability, and may therefore be viewed as providing an adequate approximation to the generating model.

The most widely known and used model selection criterion is the Akaike (1973, 1974) information criterion (AIC). AIC is formulated as an asymptotically unbiased estimator of the K–L discrepancy. The broad acceptance of AIC can be attributed to its computational simplicity and its connection to likelihood theory. AIC can be applied in any framework where the candidate models are fit using maximum likelihood (ML), and the sample size is large enough to ensure the conventional large-sample properties of maximum likelihood estimators.

Sugiura (1978), and later Hurvich and Tsai (1989), investigated the small-sample properties of AIC in the framework of Gaussian linear regression models. They proposed a corrected version of AIC, AIC$_c$, to provide a more accurate estimator of the K–L discrepancy in small-sample settings. Extensive simulation studies have demonstrated the small-sample superiority of AIC$_c$ over AIC, and the criterion has been extended to many modeling frameworks beyond that of Gaussian linear regression (Hurvich et al. 1990; Hurvich and Tsai 1993; Bedrick and Tsai 1994).

The Schwarz (1978) information criterion (SIC), more commonly known as the Bayesian information criterion (BIC), is designed to provide an approximation to a transformation of the posterior probability of a candidate model. When the sample size is large, BIC tends to select the candidate model that is a posteriori most probable. BIC is a popular competitor to AIC, partly because of its Bayesian justification, and partly for its tendency to favor parsimonious models.

AIC and AIC$_c$ are both developed as estimators of the K–L discrepancy. In essence, these criteria are comprised of two parts, the goodness-of-fit term and the penalty term. The goodness-of-fit term is designed to measure the conformity of the fitted model to the data at hand. However, the goodness-of-fit term serves as a negatively biased estimator of the targeted discrepancy. Efron (1983, 1986) refers to the bias as the expected optimism. The penalty term is designed to correct for this bias.

The penalty term reflects model complexity. In the linear regression setting, the complexity of a candidate model is dictated by the rank of the design matrix. In such a framework, the penalty term of AIC is $2(p + 1)$, where p denotes the rank and $(p + 1)$ corresponds to the number of parameters in the fitted model. AIC$_c$ employs the penalty term $[2(p+1)n]/(n-p-2)$, where n is the sample size. In large-sample settings, the penalizations of AIC and AIC$_c$ are essentially the same. BIC utilizes a penalty that increases in accordance with the sample size, $(p+1)\log(n)$. In the setting of best-subset model selection, the number of candidate models for each subset size plays a large part in determining the expected optimism. Unfortunately, none of these traditional model selection criteria take this factor into account. Thus, these criteria might be inappropriate criteria in the best-subset setting.

Improved AIC, or AIC$_i$, is another variant of AIC that is based on a flexible and accurate estimator of the expected optimism. The criterion was first proposed by Hurvich et al. (1990) as a refinement to AIC in selecting univariate Gaussian autoregressive models. With AIC$_i$, the penalty term is based on simulation, where the generating model is assumed to yield a Gaussian white noise process. Using simulation to approximate the expected optimism avoids the need for an analytic derivation and any accompanying large-sample assumptions. Simulation results featured in Hurvich et al. (1990) demonstrate that when the sample size is small and the candidate models are estimated by ML, AIC$_i$ outperforms AIC as well as AIC$_c$. However, the authors only examine simulation settings where the true autoregressive order is small. The selection behavior of AIC$_i$ has not yet been assessed for larger generating orders.

The literature that explicitly addresses the problem of best-subset model selection is somewhat scant. A notable contribution is the criterion proposed by Tibshirani and Knight (1999), the covariance inflation criterion (CIC). The computation of CIC is intensive, requiring permuted versions of the data set. The estimate of the expected optimism is based on the covariance between the responses and their corresponding predicted values. Unfortunately, CIC only tends to work well when the generating model is null. When the generating model is not null, CIC often fails to protect against the inclusion of spurious regressors.

The major objective of our study is to develop improved model selection methods for best-subset regression. We focus on the important case where the size of the sample is moderate relative to the number of parameters in the largest candidate models. The method we propose is designed to outperform the procedure that is conventionally used in best-subset applications: i.e., choosing the fitted model among the entire candidate collection that minimizes a traditional criterion such as AIC or BIC. Our method aims to increase the probability of selecting an appropriate model structure, while striking a balance between overfitting and underfitting.

As standard model selection criteria, AIC and AIC$_c$ are designed to estimate the expected K–L discrepancy. For best-subset selection, these standard criteria exhibit estimation bias that depends on the number of candidate models of a particular subset size. The development of our method is motivated by the need to adjust for this bias. We accomplish this objective by selectively employing a penalty term based on Monte Carlo simulation. When some of the models of a particular subset size are overfit, our penalty term is devised to provide a more accurate approximation to the expected optimism than that provided by AIC or AIC$_c$. Accordingly, we anticipate that our method will lead to improved model selections.

Our work is organized as follows. Section 1 serves as an introduction to this paper. In Sect. 2, we provide necessary background and preliminary concepts. An overview of the K–L discrepancy is presented in the framework of linear models. Model selection criteria based on estimation of the K–L discrepancy are then introduced; specifically AIC, AIC$_c$, and AIC$_i$. Section 3 discusses the best-subset model selection problem along with its challenges. In particular, for overspecified models, we illustrate the bias inherent when standard criteria are used to estimate their corresponding discrepancies. Section 4 is devoted to the development of a best-subset model selection criterion AIC$_i$. To define the target K–L discrepancy, we propose the concept of a representative model as the target model. If a particular representative model is overspecified, the criterion AIC$_i$ is designed to approximate the corresponding K–L discrepancy. In Sect. 5, we devise a multistage model selection procedure, AIC$_{\text{aps}}$, that adaptively compares the criterion values of AIC$_c$ and AIC$_i$ across model collections of progressively larger subset sizes. Simulation studies are reported in Sect. 6 to investigate the selection behavior of AIC$_{\text{aps}}$, and to compare its performance to that of standard criteria. Finally, Sect. 7 presents conclusions and future research directions.

2 The Kullback–Leibler discrepancy, AIC, AIC$_c$, and AIC$_i$

In this section, we focus on the setting of normal linear models. Suppose a collection of data $\boldsymbol{y}$ is generated from a linear model

$$\boldsymbol{y} = \boldsymbol{X}_o\boldsymbol{\beta}_o + \boldsymbol{\epsilon}_o, \tag{1}$$

where $\boldsymbol{y}$ is an $n \times 1$ outcome vector; $\boldsymbol{X}_o$ is an $n \times p_o$ design matrix of full column rank, with the first column consisting of 1s; $\boldsymbol{\beta}_o$ is a $p_o \times 1$ vector; and $\boldsymbol{\epsilon}_o$ is an $n \times 1$ error vector distributed as $N(\boldsymbol{0}, \sigma_o^2\boldsymbol{I})$. Let $f(\boldsymbol{y} \mid \boldsymbol{\theta}_o, \boldsymbol{X}_o)$ denote the joint density of $\boldsymbol{y}$ corresponding to this model; i.e., $N(\boldsymbol{X}_o\boldsymbol{\beta}_o, \sigma_o^2\boldsymbol{I})$. Here, $\boldsymbol{\theta}_o = (\boldsymbol{\beta}_o^T, \sigma_o^2)^T$.

Assume the proposed candidate model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{X}$ is an $n \times p$ design matrix of full column rank, with the first column consisting of 1s; $\boldsymbol{\beta}$ is a $p \times 1$ vector; and $\boldsymbol{\epsilon}$ is an $n \times 1$ error vector distributed as $N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Let $f\ (\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X})$ denote the joint density of $\boldsymbol{y}$ corresponding to this model; i.e., $N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Here, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$. Also, let $l\ (\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})$ denote the log likelihood corresponding to $f\ (\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X})$.

The propriety of the candidate model can be assessed through the use of a *discrepancy*, a measure that reflects the disparity between the candidate model and the generating model. A well-known discrepancy is the K–L discrepancy, derived from the K–L information (1968), which is defined as

$$d_{\text{KL}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) = E_*\{-2l\ (\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})\}.$$

Here, E_* denotes the expectation under $f(\boldsymbol{y} \mid \boldsymbol{\theta}_o, \boldsymbol{X}_o)$. For linear models, if a constant involving 2π is neglected, we have $-2l\ (\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) = n \log \sigma^2 + \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2/\sigma^2$. Thus, $d_{\mathrm{KL}}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) = n \log \sigma^2 + n\sigma_o^2/\sigma^2 + \|\boldsymbol{X}_o\boldsymbol{\beta}_o - \boldsymbol{X}\boldsymbol{\beta}\|^2/\sigma^2$.

Let $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}^T, \hat{\sigma}^2\right)^T$ denote the ML estimator of $\boldsymbol{\theta}$. Then for a fitted candidate model $f\left(\boldsymbol{y} \,\middle|\, \hat{\boldsymbol{\theta}},\ \boldsymbol{X}\right)$, the K–L discrepancy is given by

$$d_{\mathrm{KL}}\left(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_o\right) = E_*\{-2l\ (\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

The measure $d_{\mathrm{KL}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_o)$ is a random variable that depends on the true model, and therefore cannot be evaluated in practical applications. Thus, model selection criteria are often developed by constructing estimators of the expected value of $d_{\mathrm{KL}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_o)$, say Δ_{KL}.

Let $\boldsymbol{y}^+$ denote a hypothetical future set of data, which is generated from $f(\boldsymbol{y} \mid \boldsymbol{\theta}_o, \boldsymbol{X}_o)$ but is independent of $\boldsymbol{y}$. The expected K–L discrepancy is given by

$$\begin{aligned} \Delta_{\mathrm{KL}} &= E_{*+}\left\{-2l\left(\hat{\boldsymbol{\theta}} \,\middle|\, \boldsymbol{y}^+, \boldsymbol{X}\right)\right\} && (3) \\ &= E_*\left\{E_+\left[-2l\left(\boldsymbol{\theta} \mid \boldsymbol{y}^+, \boldsymbol{X}\right)\right]|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right\}. && (4) \end{aligned}$$

In (3), E_{*+} denotes the expectation under the joint distribution of $(\boldsymbol{y}, \boldsymbol{y}^+)$. In (4), the inner expectation E_+ is taken under the distribution of $\boldsymbol{y}^+$, and the outer expectation E_* is taken under the distribution of $\boldsymbol{y}$.

Now consider writing Δ_{KL} as

$$\Delta_{\mathrm{KL}} = E_*\left\{-2l\left(\hat{\boldsymbol{\theta}} \,\middle|\, \boldsymbol{y}, \boldsymbol{X}\right)\right\} + E_*\left\{E_+\left[-2l\left(\boldsymbol{\theta} \mid \boldsymbol{y}^+, \boldsymbol{X}\right)\right]|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \left[-2l\left(\hat{\boldsymbol{\theta}} \,\middle|\, \boldsymbol{y}, \boldsymbol{X}\right)\right]\right\}. \quad (5)$$

The statistic $-2l\left(\hat{\boldsymbol{\theta}} \,\middle|\, \boldsymbol{y}, \boldsymbol{X}\right)$ is an unbiased estimator of $E_*\left\{-2l\left(\hat{\boldsymbol{\theta}} \,\middle|\, \boldsymbol{y}, \boldsymbol{X}\right)\right\}$. In practice, $-2l\left(\hat{\boldsymbol{\theta}} \,\middle|\, \boldsymbol{y}, \boldsymbol{X}\right)$ reflects the conformity of the fitted model to the data $\boldsymbol{y}$, and therefore measures goodness-of-fit. For linear models, we have

$$-2l\left(\boldsymbol{\theta} \mid \boldsymbol{y}^+, \boldsymbol{X}\right) = n \log \hat{\sigma}^2 + \|\boldsymbol{y}^+ - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/\hat{\sigma}^2 \quad \text{and} \quad -2l\left(\hat{\boldsymbol{\theta}} \,\middle|\, \boldsymbol{y}, \boldsymbol{X}\right) = n \log \hat{\sigma}^2 + n.$$

Therefore, for linear models, the K–L discrepancy can be expressed as

$$\Delta_{\mathrm{KL}} = E_*\left\{n \log \hat{\sigma}^2 + n\right\} + E_{*+}\left\{\frac{\left\|\boldsymbol{y}^+ - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right\|^2}{\hat{\sigma}^2} - n\right\}. \quad (6)$$

By writing Δ_{KL} in the form of (5), we see that we can use $-2l\left(\hat{\boldsymbol{\theta}}\,\middle|\,\boldsymbol{y},\boldsymbol{X}\right)$ as a platform for approximating Δ_{KL}. The corresponding bias is given by

$$B = E_{*}\left\{E_{+}\left[-2l\left(\boldsymbol{\theta}\mid\boldsymbol{y}^{+},\boldsymbol{X}\right)\right]\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \left[-2l\left(\hat{\boldsymbol{\theta}}\,\middle|\,\boldsymbol{y},\boldsymbol{X}\right)\right]\right\}.$$

Efron (1983, 1986) refers to B as the *expected optimism*. The difference B is positive, since it compensates for the fact that the goodness-of-fit term has a lower value for the data $\boldsymbol{y}$ from which $\hat{\boldsymbol{\theta}}$ is obtained, than for the future data $\boldsymbol{y}^{+}$.

Our goal is to find information criteria that precisely estimate the K–L discrepancy. If we have a reasonable estimator $\hat{B}$ for the expected optimism B, an estimator of Δ_{KL} could be constructed as

$$-2l\left(\hat{\boldsymbol{\theta}}\,\middle|\,\boldsymbol{y},\boldsymbol{X}\right)+\hat{B}. \tag{7}$$

In the case of linear models, (7) reduces to

$$n\log\hat{\sigma}^{2}+n+\hat{B}.$$

Akaike (1973, 1974) demonstrates that B can often be asymptotically approximated by $2(p+1)$, where $(p+1)$ is the dimension of $f\,(\boldsymbol{y}\mid\boldsymbol{\theta},\boldsymbol{X})$. More specifically, if the following two assumptions hold, $2(p+1)$ serves as an asymptotically unbiased estimator of the expected optimism.

(a) The generating model $f(\boldsymbol{y}\mid\boldsymbol{\theta}_{o},\boldsymbol{X}_{o})$ is a member of the candidate class $\{f\,(\boldsymbol{y}\mid\boldsymbol{\theta},\boldsymbol{X})\mid\boldsymbol{\theta}\in\Theta\}$, where Θ denotes the $(p+1)$-dimensional parameter space. Thus, the candidate model $f\,(\boldsymbol{y}\mid\boldsymbol{\theta},\boldsymbol{X})$ is correctly specified or overspecified.

(b) An appropriate set of regularity conditions hold so that the traditional asymptotic properties of the ML estimator $\hat{\boldsymbol{\theta}}$ are ensured.

Under assumptions (a) and (b), an asymptotically unbiased estimator of Δ_{KL} is provided by AIC:

$$\mathrm{AIC}=-2l\left(\hat{\boldsymbol{\theta}}\,\middle|\,\boldsymbol{y},\boldsymbol{X}\right)+2(p+1).$$

As a consequence of (a) and (b), AIC works very well for many model selection applications provided that the sample size is large. It tends to select a model $f\left(\boldsymbol{y}\,\middle|\,\hat{\boldsymbol{\theta}},\ \boldsymbol{X}\right)$ that minimizes the mean squared error of prediction (Shibata 1981).

One drawback of AIC is noted by Sugiura (1978), and later by Hurvich and Tsai (1989). In settings where the sample size n is relatively small compared to the dimension $(p+1)$, AIC has a potentially high degree of negative bias. This bias usually leads to severe overfitting. For correctly specified or overspecified normal linear models, the aforementioned authors derive an exactly unbiased estimator of B, which is

$$B = E_{*+}\left\{\frac{\left\|\boldsymbol{y}^{+}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right\|^{2}}{\hat{\sigma}^{2}}-n\right\}=\frac{2(p+1)n}{n-p-2}. \tag{8}$$

They refer to their criterion as "corrected" AIC, AIC$_c$:

$$\mathrm{AIC}_c = -2l\left(\hat{\boldsymbol{\theta}}\,\middle|\,\boldsymbol{y}, \boldsymbol{X}\right) + \frac{2(p+1)n}{n-p-2}.$$

The penalty term of AIC$_c$ provides a more accurate estimator of the expected optimism than AIC. Through small-sample simulation studies, Hurvich and Tsai (1989) convincingly demonstrate that AIC$_c$ outperforms AIC at selecting the correct model.

The penalty terms of AIC and AIC$_c$ are needed to correct for the bias represented by the expected optimism. In Sect. 3, we will demonstrate that for best-subset model selection, these penalizations do not adequately approximate the expected optimism. However, we will first introduce another criterion AIC$_i$, which serves as one of the fundamental tools in devising our best-subset selection procedure.

Improved AIC, or AIC$_i$, first proposed by Hurvich et al. (1990), is based on the following motivation. As an estimator of the K–L discrepancy, AIC$_c$ is exactly unbiased only in the Gaussian linear modeling framework. In other modeling frameworks, the exact unbiasedness property of AIC$_c$ does not hold. However, under the assumption that the candidate model is either correctly specified or overspecified, it may be argued that the expected optimism only loosely depends on the generating model. The derivation of AIC implies that in large-sample settings, the expected optimism is approximated by $2(p+1)$, a term that does not involve the characteristics of the generating model. For relatively small samples, as evidenced by simulation studies, the expected optimism only weakly depends on the generating model.

Based on these findings, Hurvich et al. (1990) and Bengtsson and Cavanaugh (2006) argue that the expected optimism can be accurately approximated by Monte Carlo simulation, using an arbitrary but convenient choice of a surrogate model in place of the true model, such as the null model. This Monte Carlo approximation is valid as long as the candidate model is either correctly specified or overspecified.

The structure of AIC$_i$ can be outlined as follows. Let $\boldsymbol{y}(1), \boldsymbol{y}(2), \ldots, \boldsymbol{y}(M)$ denote M fitting samples generated as i.i.d. under the null model, and let $\hat{\boldsymbol{\theta}}(1), \hat{\boldsymbol{\theta}}(2), \ldots, \hat{\boldsymbol{\theta}}(M)$ represent the corresponding parameter estimators for each sample. Also, let $\boldsymbol{y}^{+}(1), \boldsymbol{y}^{+}(2), \ldots, \boldsymbol{y}^{+}(M)$ denote M validation or future samples, additionally generated as i.i.d. under the null model. AIC$_i$ is defined as

$$\mathrm{AIC}_i = -2l\left(\hat{\boldsymbol{\theta}}\,\middle|\,\boldsymbol{y}, \boldsymbol{X}\right) + \frac{1}{M}\sum_{j=1}^{M}\left\{\left[-2l\left(\boldsymbol{y}^{+}(j)\mid\hat{\boldsymbol{\theta}}(j)\right)\right] - \left[-2l\left(\boldsymbol{y}(j)\mid\hat{\boldsymbol{\theta}}(j)\right)\right]\right\}.$$

Note that $-2l\left(\hat{\boldsymbol{\theta}}\,\middle|\,\boldsymbol{y}, \boldsymbol{X}\right)$ again serves as the goodness-of-fit term. As indicated by the structure of AIC$_i$, the criterion shares the same goodness-of-fit term as AIC and AIC$_c$, yet involves a penalty term derived from Monte Carlo simulation. Using simulation to characterize model complexity is a flexible approach that can be adapted to accommodate various modeling settings, including the setting for best-subset regression.

3 The best-subset model selection framework

A common problem that arises in many analyses is the selection of a best-subset model from the entire collection of candidate models based on APS of regressor variables.

Let $\boldsymbol{X}_P$ denote an $n \times P$ design matrix of full column rank that includes all $(P-1)$ possible regressor variables. Consider a subset candidate model with design matrix $\boldsymbol{X}$, $f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X})$, where $\boldsymbol{X}$ is an $n \times p$ design matrix of full column rank. Here, $(p-1)$ denotes the number of regressors in $\boldsymbol{X}_P$ that are included in $\boldsymbol{X}$. In total, there are $2^{(P-1)}$ possible subsets of the regressors represented in $\boldsymbol{X}_P$, and each of these subsets could be used to construct a candidate model. Best-subset model selection proceeds by searching among these $2^{(P-1)}$ candidate models and finding the best fit according to some criterion.

As previously mentioned, in the normal linear modeling framework when the candidate model of interest is correctly specified or overspecified, AIC$_c$ is an unbiased estimator of the K–L discrepancy. However, for best-subset selection, AIC$_c$ is biased. Thus, AIC$_c$ alone is an inappropriate criterion in this setting.

To explain this bias, we will first provide an illustrative example. In this example, we consider the use of order statistics to estimate the mean of a series of random variables. Suppose we have a collection of random variables $y_1, y_2, \ldots, y_n$ with a common mean: $E(y_i) = \mu$, for $i = 1, 2, \ldots, n$. Each of these random variables serves as an unbiased estimator of μ. However, the minimum of these random variables is a biased estimator of μ, because the expectation of the minimum order statistic is less than μ. Moreover, the size of the bias depends on n. Thus, the bias becomes more prominent as n increases.

In best-subset model selection, the problem of estimating the K–L discrepancy could be viewed as analogous to the preceding problem. Suppose for a particular subset size, several candidate models are overspecified. We will demonstrate in Proposition 1 that the K–L discrepancies of these models are all equal. The AIC$_c$ for each of these overfitted models provides an unbiased estimator of the common K–L discrepancy. However, if we focus on the fitted model corresponding to the minimum AIC$_c$, then this minimum AIC$_c$ is biased for the common K–L discrepancy. The bias is again caused by the ordering of the AIC$_c$ values (or equivalently, the goodness-of-fit terms). As the number of overspecified models grows, the difference increases between the expectation of the minimum AIC$_c$ and the common K–L discrepancy.

In relating the problem of estimating a common K–L discrepancy using the minimum AIC$_c$ to the problem of estimating a common mean using the minimum order statistic, we note that the latter development is typically based on a simplistic setting where the variates are assumed independent. For dependent variates, the bias of the minimum order statistic is not only governed by the size of the sample, but also by the nature of the dependence among the variates (see, for instance, Maurer and Margolin 1976). For overspecified models of a particular subset size, the AIC$_c$ values will obviously be dependent, since they are all based on empirical log-likelihoods arising from the same response data. Nonetheless, the analogy to conventional order statistics is helpful in conceptualizing how bias arises in best-subset model selection.

Now suppose that a large number of candidate models exist for a subfamily of a certain subset size, and that in turn, these models comprise a substantial proportion of all the candidate models. Potentially, there is a high likelihood of choosing a model from this subfamily, regardless of whether the generating model is a member of this subfamily. To avoid this problem, model selection criteria should penalize more heavily for a subset size that represents a large number of candidate models. In other words, for a particular subset size, the magnitude of the penalization should be positively related to the number of models. With AIC$_c$, note that the penalty term is merely based on n and p, and does not depend on the number of models of size s. Thus, AIC$_c$ is an inappropriate criterion for best-subset selection.

In this work, we used R (R Core Team, version 2.15.3) to perform simulations. The R package leaps (Lumley 2009) was applied to find the best model for each subset size. Implemented by the branch-and-bound algorithm, the leaps package was first utilized in subset selection by Beale et al. (1967), and later by Hocking and Leslie (1967) and LaMotte and Hocking (1970). This preceding algorithm is feasible on modern PCs when the largest model size S is not too large: e.g., $S < 20$. We emphasize that best-subset selection is computationally expensive. For every additional regressor variable included for analyses, the computational cost will roughly double.

4 AIC$_i$ in best-subset model selection

For best-subset model selection, our goal is to find a criterion that can accurately estimate the K–L discrepancy. Once we have such a criterion, a selection procedure can be proposed that effectively utilizes this statistic. Yet before developing the criterion, we will need to answer an important question: i.e., what type of model should serve as the target for the estimation of the K–L discrepancy?

Consider a candidate subset of shared dimension that contains multiple overspecified models. The following proposition demonstrates that these models share the same K–L discrepancy. Thus, the estimation of the common K–L discrepancy could be based on any of the overspecified models. The proof of the proposition appears in the appendix.

Proposition 1 *Assume data* $\boldsymbol{y}$ *is generated from a linear model* $\boldsymbol{y} = \boldsymbol{X}_o\boldsymbol{\beta}_o + \boldsymbol{\epsilon}_o$, *where* $\boldsymbol{X}_o$ *is an* $n \times p_o$ *design matrix with full column rank, and* $\boldsymbol{\epsilon}_o \sim N(\boldsymbol{0}, \sigma_o^2\boldsymbol{I})$. *Let* $C(\boldsymbol{X}_o)$ *denote the column space of* $\boldsymbol{X}_o$. *Suppose two overspecified candidate models are given by* $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$, *with* $\boldsymbol{\epsilon}_1 \sim N(\boldsymbol{0}, \sigma_1^2\boldsymbol{I})$, *and* $\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2$, *with* $\boldsymbol{\epsilon}_2 \sim N(\boldsymbol{0}, \sigma_2^2\boldsymbol{I})$. *Here,* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *are both* $n \times p$ *matrices with full column rank, and* $p_o < p$. *Since both candidate models are overspecified,* $C(\boldsymbol{X}_o) \subseteq C(\boldsymbol{X}_1)$ *and* $C(\boldsymbol{X}_o) \subseteq C(\boldsymbol{X}_2)$. *Let* $\Delta_{KL}(1)$ *and* $\Delta_{KL}(2)$ *denote the K–L discrepancies of the two candidate models. Then we have* $\Delta_{KL}(1) = \Delta_{KL}(2)$.

In practice, although the K–L discrepancies are identical for overspecified models of a common dimension, the realized values of a model selection criterion such as AIC$_c$ will vary. We tend to favor the model with the minimum sum of squared errors (SSE), because for a certain subset size, the minimum SSE model provides the best goodness-of-fit. Unfortunately, as we have previously discussed, AIC$_c$ for the minimum SSE model is negatively biased for the K–L discrepancy.

We are now in a position to answer the question raised earlier in this section: i.e., what type of model should serve as the target for the estimation of the K–L discrepancy? We define a *representative sufficient* model, or simply a *representative* model, as any model within a collection of a certain size that contains all the variables in the true model. A comparison of K–L discrepancies among the representative models from different subset sizes would help us to decide the best subset size. For a certain subset size, Proposition 1 demonstrates that all the representative models have the same K–L discrepancy. For this reason, if we use AIC$_c$ for the minimum SSE model to estimate this common K–L discrepancy, the estimate will be biased. Moreover, the bias will increase in accordance with the number of representative models.

To address this problem, we will use the idea behind AIC$_i$ to derive an appropriate estimator of the K–L discrepancy for a representative model. Let $\hat{\boldsymbol{\theta}}_r = (\hat{\boldsymbol{\beta}}_r^T, \hat{\sigma}_r^2)^T$ denote the ML estimator for a representative model, and let $\Delta_{\mathrm{KL}}(r)$ denote the corresponding K–L discrepancy. In reference to (6) and (8), we have

$$\begin{aligned}\Delta_{\mathrm{KL}}(r) &= E_{*+}\left[-2l\left(\hat{\boldsymbol{\theta}}_r \mid \boldsymbol{y}^+, \boldsymbol{X}_r\right)\right]\\ &= E_*\left\{n\log\hat{\sigma}_r^2 + n\right\} + E_{*+}\left\{\frac{\left\|\boldsymbol{y}^+ - \boldsymbol{X}_r\hat{\boldsymbol{\beta}}_r\right\|^2}{\hat{\sigma}_r^2} - n\right\}\\ &= E_*\left\{n\log\hat{\sigma}_r^2 + n\right\} + \frac{2(p+1)n}{n-p-2}.\end{aligned}$$

Also, let $\hat{\boldsymbol{\theta}}_{min} = \left(\hat{\boldsymbol{\beta}}_{min}^T, \hat{\sigma}_{min}^2\right)^T$ denote the ML estimator for the minimum SSE model, and let $-2l\left(\hat{\boldsymbol{\theta}}_{min} \mid \boldsymbol{y}, \boldsymbol{X}_{min}\right)$ denote the corresponding goodness-of-fit term. Suppose we use $-2l\left(\hat{\boldsymbol{\theta}}_{min} \mid \boldsymbol{y}, \boldsymbol{X}_{min}\right)$ in the estimation of $\Delta_{\mathrm{KL}}(r)$. The bias B is then given by

$$\begin{aligned}B &= \Delta_{\mathrm{KL}}(r) - E_*\left\{-2l\left(\hat{\boldsymbol{\theta}}_{min} \mid \boldsymbol{y}, \boldsymbol{X}_{min}\right)\right\}\\ &= E_*\left\{n\log\hat{\sigma}_r^2 + n\right\} + \frac{2(p+1)n}{n-p-2} - E_*\left\{n\log\hat{\sigma}_{min}^2 + n\right\}\\ &= E_*\left\{n\log\frac{\hat{\sigma}_r^2}{\hat{\sigma}_{min}^2}\right\} + \frac{2(p+1)n}{n-p-2}.\end{aligned}$$

For candidate models that are overspecified, the preceding expectation in B could be approximated through Monte Carlo simulation, based on the idea behind AIC$_i$. By substituting E_* with an estimator $\hat{E}_*$, obtained through Monte Carlo simulation based on the null model, we have

$$\begin{aligned}\hat{B} &= \hat{E}_*\left\{n\log\frac{\hat{\sigma}_r^2}{\hat{\sigma}_{min}^2}\right\} + \frac{2(p+1)n}{n-p-2}\\ &= \frac{1}{M}\sum_{j=1}^{M} n\log\frac{\hat{\sigma}_r^2(j)}{\hat{\sigma}_{min}^2(j)} + \frac{2(p+1)n}{n-p-2}.\end{aligned} \tag{9}$$

Specifically, let $\boldsymbol{y}(1), \boldsymbol{y}(2), \ldots, \boldsymbol{y}(M)$ be M vectors of data generated i.i.d. from $N(\boldsymbol{0}, \boldsymbol{I})$. Thus, we employ a null model as a surrogate for the true model (1), where $\boldsymbol{\beta}_o$ is $\boldsymbol{0}$ and σ_o^2 is one. The justification of AIC$_i$ ensures that $\hat{B}$ accurately estimates the bias B regardless of the parameter specification for the surrogate model, provided that the fitted candidate model is either correctly specified or overspecified. We choose the $N(\mathbf{0}, \boldsymbol{I})$ surrogate model merely for convenience.

For each subset size, $\hat{\sigma}_r^2(j)$ and $\hat{\sigma}_{min}^2(j)$ are the ML estimators of error variance corresponding to the representative model and the minimum SSE model, respectively.

In the best-subset setting, AIC$_i$ can therefore be defined as

$$\begin{aligned}\text{AIC}_i &= \left\{n \log \hat{\sigma}^2 + n\right\} + \hat{B} \\ &= \left\{n \log \hat{\sigma}^2 + n\right\} + \left\{\frac{1}{M}\sum_{j=1}^{M} n \log \frac{\hat{\sigma}_r^2(j)}{\hat{\sigma}_{min}^2(j)} + \frac{2(p+1)n}{n-p-2}\right\} \qquad (10)\\ &= n \log \hat{\sigma}^2 + \frac{1}{M}\sum_{j=1}^{M} n \log \frac{\hat{\sigma}_r^2(j)}{\hat{\sigma}_{min}^2(j)} + \frac{n(n+p)}{n-p-2}, \qquad (11)\end{aligned}$$

where $\hat{\sigma}^2$ is the ML estimator of error variance corresponding to the candidate model of interest.

For a certain subset size, AIC$_i$ measures the proximity of a representative model to the underlying true model, in terms of the K–L discrepancy. As evident from (10), in addition to the penalty term in AIC$_c$, AIC$_i$ has an extra penalty,

$$\frac{1}{M}\sum_{j=1}^{M} n \log \frac{\hat{\sigma}_r^2(j)}{\hat{\sigma}_{min}^2(j)},$$

that accounts for the bias caused by best-subset selection. One appealing aspect of this penalty term can be attributed to its generation based on the null model. Consequently, this term is independent of the true model $f(\boldsymbol{y} \mid \boldsymbol{\theta}_o, \boldsymbol{X}_o)$, and is therefore independent of the goodness-of-fit term $n \log \hat{\sigma}^2$. For convenience, we can construct a table that lists the simulated penalties corresponding to various combinations of n and p. By looking up the penalty terms from such a table, AIC$_i$ could be computed as readily as other standard model selection criteria, such as AIC and AIC$_c$.

In order to provide a quantitative illustration of the differences among the criteria, Table 1 presents the penalty terms employed by AIC, AIC$_c$, and AIC$_i$. In this example, we consider orders from 0 to 10, and a sample size of $n = 100$. Note that order refers to the number of candidate regressor variables (i.e., $(p-1)$). In Table 1, the penalty term is $2(p+1)$ in AIC and $[2(p+1)n]/(n-p-2)$ in AIC$_c$. Through all the orders, AIC$_c$ always provides larger penalties than AIC, which reflects AIC$_c$'s correction for AIC's bias in small samples. In AIC$_i$, the penalty term is evaluated from (9) over 1000 simulated samples. As shown in Table 1, when the order is 0 or 10, the penalty term of AIC$_i$ is exactly equal to that of AIC$_c$. Since for each of these subset sizes, the intercept-only model or the full model is the only candidate model, the bias issue that arises in best-subset selection becomes moot. When the order is neither 0 nor 10, the penalty term of AIC$_i$ is always larger than that of AIC$_c$. The magnitudes of the difference are most significant for orders 4, 5, and 6. Among these orders, the numbers of candidate models are considerable, which leads to pronounced bias corrections for AIC$_i$.

Table 1 Penalty terms of AIC, AIC$_c$, and AIC$_i$: $n = 100$

Order	AIC	AIC$_c$	AIC$_i$
0	4	4.12	4.12
1	6	6.25	9.12
2	8	8.42	12.58
3	10	10.64	15.32
4	12	12.90	17.62
5	14	15.22	19.62
6	16	17.58	21.41
7	18	20.00	23.05
8	20	22.47	24.60
9	22	25.00	26.10
10	24	27.59	27.59

5 The multistage model selection procedure AIC$_{\text{aps}}$

In best-subset selection, for a candidate subfamily of a certain size, we tend to focus on the fitted model corresponding to the minimum SSE. The following proposition indicates that, for a given size, a representative model's K–L discrepancy is smaller than that of an underspecified model. Thus, a representative model is prone to produce the minimum SSE. The proof of the proposition appears in the Appendix.

Proposition 2 *Assume data* $\mathbf{y}$ *is generated from a linear model* $\mathbf{y} = \boldsymbol{X}_o\boldsymbol{\beta}_o + \boldsymbol{\epsilon}_o$, *where* $\boldsymbol{X}_o$ *is an* $n \times p_o$ *design matrix with full column rank, and* $\boldsymbol{\epsilon}_o \sim N(\mathbf{0}, \sigma_o^2\boldsymbol{I})$. *Suppose a correctly specified or overspecified candidate model* (1) *is given by* $\mathbf{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$, *with* $\boldsymbol{\epsilon}_1 \sim N(\mathbf{0}, \sigma_1^2\boldsymbol{I})$, *and an underspecified candidate model* (2) *is given by* $\mathbf{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2$, *with* $\boldsymbol{\epsilon}_2 \sim N(\mathbf{0}, \sigma_2^2\boldsymbol{I})$. *Here,* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *are both* $n \times p$ *matrices with full column rank, and* $p_o \leq p$. *Since model* (1) *is correctly specified or overspecified,* $C(\boldsymbol{X}_o) \subseteq C(\boldsymbol{X}_1)$; *and since model* (2) *is underspecified,* $C(\boldsymbol{X}_o) \nsubseteq C(\boldsymbol{X}_2)$. *Let* $\Delta_{KL}(1)$ *and* $\Delta_{KL}(2)$ *denote the K–L discrepancies of the two candidate models. Also, let* $\bar{\Delta}_{KL}(1) = \Delta_{KL}(1)/n$ *and* $\bar{\Delta}_{KL}(2) = \Delta_{KL}(2)/n$ *denote the mean K–L discrepancies of the two models. Then as* $n \to \infty$, $\bar{\Delta}_{KL}(1) - \bar{\Delta}_{KL}(2) \to C$, *where* $C < 0$.

Proposition 2 demonstrates that for a certain subset size, a representative model tends to be the minimum SSE model. The remaining problem is to use an appropriate criterion to estimate a representative model's K–L discrepancy. In the preceding section, we propose a criterion AIC$_i$ to estimate this measure. AIC$_i$ is most appropriate for the setting where several representative models exist within a certain subset size. However, if the generating model is the only representative model for a subset size, AIC$_c$ actually outperforms AIC$_i$ in terms of approximating the K–L discrepancy. In this case, the minimum SSE model should be the correctly specified model, and all the other models should be underspecified. Since these underspecified models have different K–L discrepancies that should exceed the K–L discrepancy for the correctly specified model, no bias is induced by using the minimum order statistic to estimate a common target based on multiple representative models.

Figure 1 displays the patterns of AIC, AIC$_c$, and AIC$_i$ as estimators of the K–L discrepancy. In this example, we first consider models with nested design matrices comprised of the first m regressors of $\boldsymbol{X}$, for $m = 1, \ldots, 10$. Among these models, the generating model has order 3. Thus, the model of order 3 is correctly specified, and the models of order >3 are overspecified. Since the true parameters are known, the K–L discrepancies for these 10 models can be computed. We then consider APS models in the candidate family, and compute criterion values for the minimum SSE model of each subset size. Figure 1 plots the average K–L discrepancies, and the average values of AIC, AIC$_c$ and AIC$_i$ over 1000 simulations.

One should expect that the K–L discrepancy achieves its minimum at order 3. At order 3, AIC$_c$ is closer to the K–L discrepancy than AIC$_i$. For this order, AIC$_c$ is an approximately unbiased estimator of the K–L discrepancy, whereas AIC$_i$ is positively biased. When the penalty term of AIC$_i$ is simulated from the null model, we assume all the candidate models are overspecified. However, at order 3, no candidate model is overspecified. Thus, the penalty of AIC$_i$ induces a positive bias.

When the order is larger than 3, AIC$_i$ is generally closer to the K–L discrepancy than AIC$_c$. As evident from the figure, AIC$_i$ tracks the K–L discrepancy curve very closely. Thus, AIC$_i$ can estimate the K–L discrepancy with the least amount of bias. In contrast, AIC$_c$ tends to underestimate the K–L discrepancy for subfamilies featuring overspecified models, even though the underestimation is less prominent than that exhibited by AIC. Therefore, it is more appropriate to use AIC$_i$ in estimating the K–L discrepancy for subfamilies with overspecified models.

An important conclusion derived from Fig. 1 can be summarized as follows. When there are several representative models in the subfamily, AIC$_i$ is an approximately unbiased estimator of the K–L discrepancy, and AIC$_c$ is negatively biased. On the other hand, when the generating model is the only representative model in the subfamily, AIC$_c$ is an approximately unbiased estimator of the K–L discrepancy, and AIC$_i$ is positively biased. In this setting, the generating model should correspond to the minimum SSE model, and all other models should be underspecified.

Clearly, for both AIC$_c$ and AIC$_i$, specific settings exist where one of the two criteria provides a more accurate approximation to the K–L discrepancy and exhibits less bias. In order to combine the strengths of the two criteria, we propose a multistage selection procedure, AIC$_{\text{aps}}$. Unlike conventional stepwise procedures, our method considers APS of regressors, and therefore constitutes an exhaustive search.

Under the assumption that one of the candidate models (of orders 1 through S) is correctly specified, a description of the procedure AIC$_{\text{aps}}$ is as follows.

Step 1: The true model is either the minimum SSE model of order 1, or some other model of order >1. Compare AIC$_c$ for the minimum SSE model of order 1, to the AIC$_i$ for all the models of order >1. If there is any model of order >1 that has AIC$_i$ smaller than AIC$_c$ for the order 1 model, go to step 2. Otherwise select the minimum SSE model of order 1.

Step 2: At this stage, we assume the true model is either the minimum SSE model of order 2, or some other model of order >2. Compare AIC$_c$ for the minimum SSE model of order 2, to the AIC$_i$ for all the models of order >2. If there is any model of order >2 that has AIC$_i$ smaller than

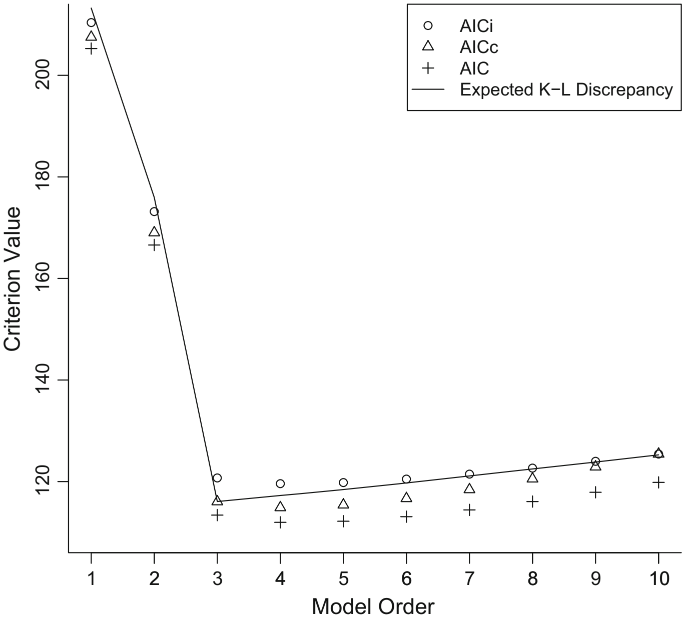

Fig. 1 Average K–L discrepancy for representative models ($p \geq 3$), and criterion values for minimum SSE models: true order $= 3$ and $n = 100$

AIC$_c$ for the order 2 model, go to step 3. Otherwise select the minimum SSE model of order 2.

⋮

Step $S - 1$: At this stage, we assume the true model is either the minimum SSE model of order $S - 1$, or the full model of order S. Among these two models, select the model with the smallest AIC$_c$.

This multistage process is illustrated by Fig. 2.

Based on the preceding discussion, if the generating model is a member of subset size s, our multistage procedure, AIC$_{\text{aps}}$, is expected to stop at step s. Before step s, we are essentially comparing AIC$_c$ for underspecified models to AIC$_i$ for representative models. With these two types of models, the representative models usually produce much lower criterion values. Thus, the procedure should proceed to the next step. When the procedure arrives at step s, we are comparing AIC$_c$ for the correctly specified model to AIC$_i$ for overspecified models. At this stage, based on the properties of AIC$_c$ and AIC$_i$, both of these criteria should accurately estimate their corresponding K–L discrepancies. Thus, subtle differences among the discrepancy measures can be identified, which should, in principle, lead to the identification of the correct model.

Our proposed algorithm is designed as a refinement of the conventional best-subsets selection procedure based on a traditional model selection criterion such as AIC or AIC$_c$. In the usual procedure, for each candidate subfamily of a certain order, ranging from 1 to S, the smallest SSE model is identified. These S models are then compared

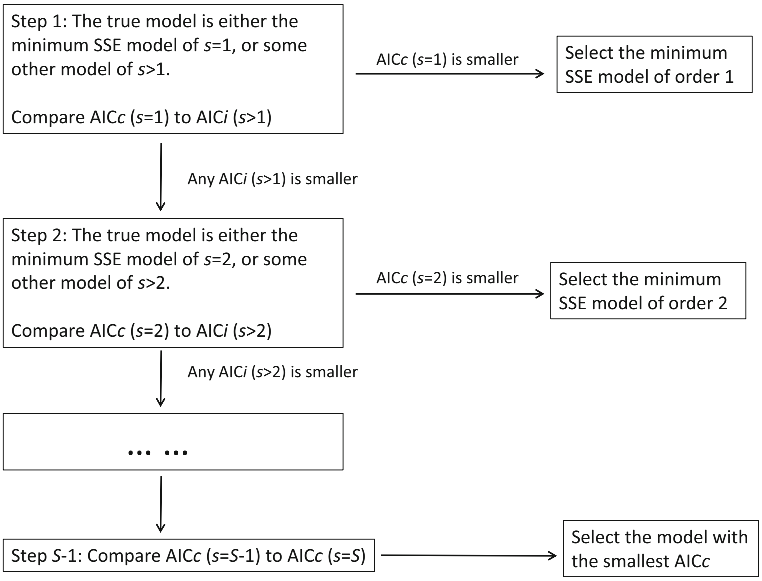
Step 1: The true model is either the minimum SSE model of *s*=1, or some other model of *s*>1.
Compare AIC*c* (*s*=1) to AIC*i* (*s*>1)
AIC*c* (*s*=1) is smaller
Select the minimum SSE model of order 1
Any AIC*i* (*s*>1) is smaller
Step 2: The true model is either the minimum SSE model of *s*=2, or some other model of *s*>2.
Compare AIC*c* (*s*=2) to AIC*i* (*s*>2)
AIC*c* (*s*=2) is smaller
Select the minimum SSE model of order 2
Any AIC*i* (*s*>2) is smaller
… …
Step *S*-1: Compare AIC*c* (*s*=*S*-1) to AIC*c* (*s*=*S*)
Select the model with the smallest AIC*c*

Fig. 2 Multistage process of $\mathrm{AIC}_{\mathrm{aps}}$

using the values of the criterion, which penalize for complexity, and the model corresponding to the smallest value is selected. Our algorithm also focuses on the minimum SSE models, yet adaptively compares AIC_c and AIC_i values, so as to minimize the bias that would result from using either criterion exclusively.

6 Simulation study

In our simulation study, we compare the performances of different model selection methods. The selection methods under consideration include (1) AIC; (2) AIC_c; (3) AIC_i; (4) BIC; (5) $\mathrm{AIC}_{\mathrm{aps}}$. Our study is comprised of three collections of simulation sets. For each collection, we compile 10 simulation sets based on 1000 replications (i.e., samples), where each set is characterized by the order of the generating model. Here, order refers to the number of regressor variables in the model. For each generated sample, the first step is to construct an $n \times 10$ full design matrix $\boldsymbol{X}$, where n denotes the sample size. Every row of $\boldsymbol{X}$ represents the collection of covariates for a particular subject, generated as i.i.d. replicates from a multivariate normal distribution with mean vector $\mathbf{0}$ and identity covariance matrix. The columns of $\boldsymbol{X}$ can be written as $[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{10}]$, with the corresponding 10×1 coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{10})^T$. In simulation set m ($m \in \{1, \ldots, 10\}$), the first m elements of $\boldsymbol{\beta}$ are set to 1, and the remaining $(10 - m)$ elements are set to 0. Assuming the intercept is 0, the generating model can then be written as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$. Based on this configuration, in simulation set m, only the first m columns of $\boldsymbol{X}$ are involved in the generation of $\boldsymbol{y}$.

The generating models can be presented as follows:

$$\begin{aligned} y_i &= x_{i1} + \epsilon_i \\ y_i &= x_{i1} + x_{i2} + \epsilon_i \\ &\vdots \\ y_i &= x_{i1} + x_{i2} + \ldots + x_{i10} + \epsilon_i, \end{aligned} \tag{12}$$

where the random errors ϵ_i are i.i.d. distributed as $N(0, \sigma^2)$.

Note that the magnitude of ϵ_i might obscure the underlying regression surface for the sample, and thus affect model selection. In order to control the impact of the generating model error, we will determine σ^2 from the signal-to-noise ratio (SNR). As defined by Cavanaugh (2004), SNR is "a ratio of two variances: the variance of the linear form in the regressor variables relative to the variance of the error component." SNR can be written as

$$\begin{aligned} \text{SNR} &= \frac{\text{var}(\boldsymbol{X}\boldsymbol{\beta})}{\sigma^2} \\ &= \frac{\boldsymbol{\beta}^T \text{var}(\boldsymbol{X})\boldsymbol{\beta}}{\sigma^2}. \end{aligned}$$

For linear models, the coefficient of determination, or R^2, is a useful measure of goodness-of-fit. It can be shown that if the correctly specified model is fit to the data, then R^2 is approximately SNR/(1 + SNR). Thus, we can determine σ^2 from R^2 by noting that

$$\begin{aligned} \sigma^2 &= \frac{\boldsymbol{\beta}^T \text{var}(\boldsymbol{X})\boldsymbol{\beta}}{\text{SNR}} \\ &\approx \frac{\boldsymbol{\beta}^T \text{var}(\boldsymbol{X})\boldsymbol{\beta}}{R^2/(1 - R^2)}. \end{aligned}$$

Based on the preceding relationship between σ^2, R^2, and SNR, we will derive σ^2 for SNR = 9. Therefore, R^2 for the correctly specified fitted model is approximately 0.9.

For each collection in our study, recall that we compile 10 simulation sets based on 1000 replications. We employ a sample size of $n = 100$ for the first collection, $n = 75$ for the second, and $n = 50$ for the third. Therefore, we can assess the impact of sample size on model selection.

For each replication, APS of regressors are considered to define the family of candidate models. After the models are fit to the data, we use model selection methods to search for the fitted candidate model that provides the best approximation to the generating model. Note that the consideration of all possible subset models ensures that one of the candidate models is correctly specified. In other words, the generating model is a member of the candidate family.

For every candidate model, we assume the existence of an intercept. Since there are 10 regressor variables, say $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{10})$, the total number of possible subset models is $2^{10} = 1024$. Let $s = 0, 1, \ldots, 10$ denote the subset size. The number of candidate models for subset size s is then $\binom{10}{s}$. Note that when the subset size is $s = 0$, the $\binom{10}{0} = 1$ candidate model is the intercept-only model. Also, when the subset size is $s = 10$, the $\binom{10}{10} = 1$ candidate model is the full model. According to our design, in simulation set m, the generating regressors involve $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$. Thus, candidate models that do not contain *all* the variables $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ are underspecified. On the other hand, candidate models that contain *all* the variables $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ plus some additional variables are overspecified.

In order to examine the selection behaviors of the different methods, the criterion values of AIC, AIC$_c$, BIC, CIC, and AIC$_i$ are calculated for every fitted candidate model, and the model favored by each method is recorded. Our multistage model selection procedure AIC$_{\text{aps}}$ selects the best model through adaptively comparing AIC$_c$ and AIC$_i$.

Over 1000 simulations, the model selection results are displayed in Table 2. Each row of the table indicates the order of the generating model, and each cell entry shows the number of times (out of 1000) the correct model is selected by a certain method.

As evident from Table 2, over nearly the entire range of the generating order m, AIC$_{\text{aps}}$ obtains more correct selections than either AIC or AIC$_c$. When m is small, say from 1 to 5, the performances of AIC and AIC$_c$ are very poor, with the criteria selecting the correct model structure <50 % of the time. In contrast, AIC$_{\text{aps}}$ performs quite well for all values of m, selecting the correct model structure more than 80 % of the time. We notice that the performances of AIC and AIC$_c$ are worse when m is small as compared to when m is relatively large. This phenomenon indicates that both AIC and AIC$_c$ tend to select overspecified models. When m is small, most of the candidate models are overspecified, which increases the risk of overfitting. Whereas for larger m, the number of overspecified models is reduced, which protects against overfitting. In fact, when m is between 9 and 10, there is little chance for any of the methods to select overspecified models; thus, in these sets, the performances of AIC and AIC$_c$ are comparable to that of AIC$_{\text{aps}}$.

The criterion AIC$_i$ performs better than AIC and AIC$_c$ only for small values of m. As m increases, AIC$_i$ begins to select overspecified models. As illustrated by Table 1, for large orders m, the separation between consecutive penalty terms for AIC$_i$ becomes less pronounced. Consequently, AIC$_i$ tends to favor more complex models, since the increase in penalization for the more complex models is insufficient to compensate for the improvement in goodness-of-fit.

Compared to AIC, AIC$_c$, and AIC$_i$, BIC achieves better selection results due to its larger penalty term. However, when m is small, BIC also exhibits a tendency to overfit, and does not yield results as favorable as AIC$_{\text{aps}}$. As m becomes larger, the selection results for BIC improve, and in some sets, the criterion marginally outperforms AIC$_{\text{aps}}$.

CIC provides acceptable results for small values of m, yet the performance of this criterion quickly deteriorates as m increases. For m between 4 and 9, CIC obtains relatively few correct selections.

Table 2 Frequencies of the correct model selected by AIC, AIC$_c$, AIC$_i$, BIC, CIC, and AIC$_{\mathrm{aps}}$: 1000 replications

Order	AIC	AIC$_c$	AIC$_i$	BIC	CIC	AIC$_{\mathrm{aps}}$
$n = 100$						
1	195	222	496	717	624	820
2	233	276	368	722	392	845
3	279	342	344	766	250	862
4	330	420	313	789	146	873
5	378	461	332	842	98	885
6	461	558	371	849	77	864
7	550	646	448	886	73	848
8	662	746	532	911	92	846
9	819	874	756	965	169	874
10	1000	1000	1000	1000	1000	1000
$n = 75$						
1	186	243	487	669	612	826
2	237	314	429	703	393	855
3	244	330	368	693	226	872
4	302	416	345	730	136	849
5	381	495	357	787	95	887
6	474	602	390	834	84	888
7	511	644	477	852	62	868
8	651	764	613	884	80	853
9	812	880	759	938	181	880
10	1000	1000	1000	1000	1000	1000
$n = 50$						
1	177	244	522	575	612	837
2	207	314	424	591	402	863
3	241	377	391	626	249	878
4	284	474	368	682	161	897
5	338	530	409	705	107	888
6	395	572	477	713	74	873
7	508	732	529	813	64	891
8	619	783	664	833	74	871
9	789	905	824	920	186	903
10	1000	1000	1000	1000	1000	1000

The aforementioned findings are illuminated by Fig. 3, where the average model orders selected by each of the methods are plotted against the generating orders, for a sample size of $n = 100$. The solid line represents the correct orders. As indicated from Fig. 3, AIC, AIC$_c$, and AIC$_i$ have propensities to select orders higher than the correct order, and this propensity is most pronounced when the generating order is small. On the other hand, AIC$_{\mathrm{aps}}$ outperforms AIC, AIC$_c$, and AIC$_i$ by systematically tending to select an order which is close to that of the generating model.

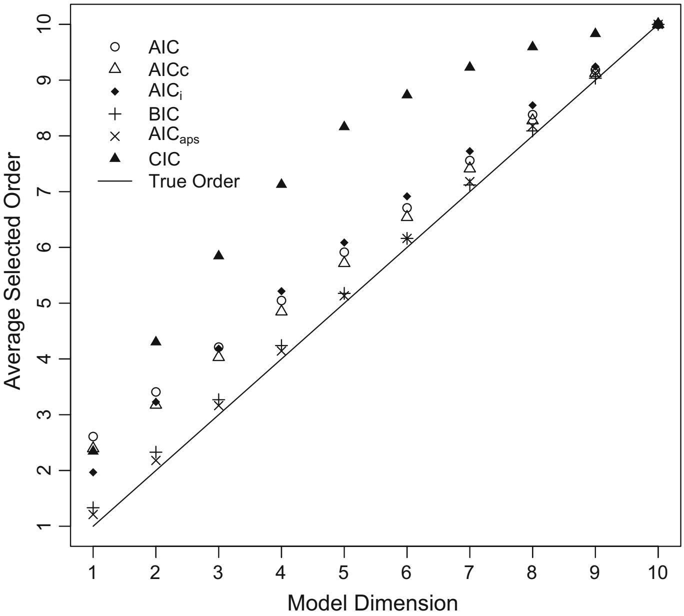

Fig. 3 Average orders selected by each of the criteria: $n = 100$, 1000 replications

In order to illustrate more detailed selection results, for the special case when the generating order is 4, Table 3 presents the frequencies of all the orders selected by each of the methods over the 1000 replications. AIC, AIC$_c$, and AIC$_i$ exhibit strong tendencies to favor larger models, selecting models of order 5 more often than models of the correct order 4. The selection pattern for BIC is much more favorable; BIC shows less of a propensity to include extraneous variables, choosing models of the correct order 4 more often than the higher orders. AIC$_{\text{aps}}$ also favors models of the correct order quite frequently, outperforming BIC and exhibiting more parsimonious selections.

Our simulation sets are based on small to moderate sample sizes. As the sample size n is reduced from 100 to 50, the performances of AIC and BIC deteriorate. This phenomenon could be explained by dependence of these criteria on the large-sample assumption: AIC is an asymptotically unbiased estimator of the K–L discrepancy, while BIC is a large-sample approximation to a transformed Bayesian posterior probability. When the sample size is not sufficiently large, optimality properties based on asymptotic justifications may fail to hold. In particular, AIC is unable to approximate the K–L discrepancy accurately, and thus cannot provide meaningful comparisons between the fitted candidate models. Also, although BIC is a consistent criterion, BIC does not tend to choose the correctly specified model, despite the inclusion of the true model in the candidate family. On the other hand, AIC$_c$ is designed as an exactly unbiased estimator of the K–L discrepancy; thus, its performance is not substantially affected by the reduction of the sample size. The penalty term of AIC$_c$ is employed as part of the penalization for AIC$_i$, which protects the method from the deleterious

Table 3 Frequencies of the model orders selected by AIC, AIC$_c$, AIC$_i$, BIC, CIC, and AIC$_{\text{aps}}$: true order = 4, 1000 replications

Order	AIC	AIC$_c$	AIC$_i$	BIC	CIC	AIC$_{\text{aps}}$
$n = 100$						
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	324	411	318	803	150	869
5	382	378	328	169	128	112
6	219	173	230	27	130	19
7	63	32	94	1	114	0
8	12	6	26	0	108	0
9	0	0	4	0	157	0
10	0	0	0	0	213	0
$n = 75$						
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	279	397	304	748	118	873
5	380	390	333	211	146	112
6	244	175	245	35	102	8
7	75	30	84	3	103	4
8	17	7	24	3	137	3
9	5	1	10	0	134	0
10	0	0	0	0	260	0
$n = 50$						
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	274	452	389	678	149	885
5	373	371	340	252	159	100
6	244	145	177	63	105	10
7	81	29	78	6	111	3
8	20	3	14	1	111	0
9	8	0	2	0	132	0
10	0	0	0	0	233	0

impact of a small sample. As a result, in nearly all of the simulation sets, the multistage procedure AIC$_{\text{aps}}$ outperforms the remaining other criteria by choosing a model of correct structure most frequently.

7 Discussion

We have proposed a multistage model selection procedure for best-subset selection in the linear modeling framework. Our algorithm is developed using estimates of the K–L discrepancy for representative models, which include both the correctly specified and overspecified models. For K–L discrepancy estimation, we argue it is more appropriate to use AIC_c for a correctly specified model, and AIC_i for overspecified models. In order to combine the strengths of these criteria, we devise a multistage selection procedure $\mathrm{AIC}_{\mathrm{aps}}$ through adaptively comparing the criterion values of AIC_c and AIC_i. Our simulation results show that our procedure performs well in terms of selecting the correct model structure.

The conventional approach to best-subset selection amounts to choosing the fitted model among the entire candidate collection that minimizes a traditional criterion such as AIC or BIC. The problem of overfitting is endemic to this approach. The success of our algorithm results from reducing this propensity, by appropriately utilizing the properties of AIC_c and AIC_i in a manner suggested by the theoretical results established in Propositions 1 and 2. Note that Proposition 2 indicates that the K–L discrepancy can effectively delineate underspecified models from correctly specified and overspecified models, at least asymptotically. This notion is crucial to the conceptual development of our procedure. In small-sample settings where the SNR is weak, underfitting might be as problematic as overfitting. The behavior of $\mathrm{AIC}_{\mathrm{aps}}$ in such settings merits further investigation.

Model selection procedures are often characterized by their asymptotic behaviors. BIC is *consistent* whereas the AIC family of criteria is *asymptotically efficient* in the sense of Shibata (1981). Assuming that the generating model is represented in the collection of candidate models, a consistent criterion will asymptotically select the fitted model having the correct structure with probability one. On the other hand, assuming that the generating model lies outside the collection of candidate models, an asymptotically efficient criterion will asymptotically select the fitted model that minimizes the mean squared error of prediction. Asymptotically efficient criteria have a tendency to choose overspecified models, even in large-sample applications. This problem is exacerbated in best-subset selection, especially in to small to moderate sample size settings. Although our multistage procedure is based on an adaptive utilization of AIC_c and AIC_i, it substantially reduces the overfitting propensity of using AIC, AIC_c, or AIC_i individually for best-subset selection. In this sense, it exhibits selection behaviors typically associated with a consistent criterion, such as BIC, in large-sample applications.

Although our procedure effectively addresses the overfitting problem that arises in best-subset selection, further improvements and refinements of the algorithm may warrant investigation. For instance, the penalty term of AIC_i is simulated assuming that all candidate models of a particular size are overspecified; i.e., that all candidate models are representative. As suggested by a referee, at any particular stage, the minimum AIC_c model could be used to identify a necessary (yet perhaps not sufficient) subset of regressors. The subset could then be used to determine which larger models might be representative and which might be underspecified. Such a delineation would allow for a refinement of the AIC_i penalty term based on a smaller number of representative models.

In future work, we hope to explore such refinements. We also hope to theoretically investigate the large-sample optimality of our selection procedure. Finally, we plan to extend our methodology to the framework of generalized linear models (GLMs), in order to account for a broader array of modeling problems.

Acknowledgments The authors wish to extend their appreciation to the associate editor and to an anonymous reviewer for carefully reading the original version of this manuscript, and for providing constructive suggestions that served to improve the presentation and content.

Appendix

Proof of Proposition 1

We will follow the derivation of AIC$_c$ in Davison (2003, pp. 402–403). Consider writing the K–L discrepancy of the first candidate model as

$$\begin{aligned}\Delta_{\mathrm{KL}}(1) &= E_{*+}\left\{-2l\left(\hat{\boldsymbol{\beta}}_1, \hat{\sigma}_1^2 \left| \boldsymbol{y}^+, \boldsymbol{X}_1\right.\right)\right\} \\ &= E_*\left\{n \log \hat{\sigma}_1^2 + n\right\} + E_{*+}\left\{\frac{\left\|\boldsymbol{y}^+ - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right\|^2}{\hat{\sigma}_1^2} - n\right\}. \end{aligned} \tag{13}$$

We will demonstrate the two terms that appear in (13) are equal for both candidate models.

For the first term $E_*\left\{n \log \hat{\sigma}_1^2 + n\right\}$, E_* corresponds to the expectation under the distribution of $\boldsymbol{y}$. We have

$$\begin{aligned}E_*\left\{n \log \hat{\sigma}_1^2 + n\right\} &= nE_*\left\{\log \hat{\sigma}_1^2\right\} + n \\ &= nE_*\left\{\log\left(\frac{\mathrm{SSE}_1}{n}\right)\right\} + n \\ &= nE_*\left\{\log \mathrm{SSE}_1\right\} - n \log n + n.\end{aligned}$$

From McQuarrie and Tsai (1998, p. 67),

$$E_*\left\{\log \mathrm{SSE}_1\right\} = \log \sigma_o^2 + \log 2 + \psi\left(\frac{n-p}{2}\right), \tag{14}$$

where ψ denotes Euler's *psi* function, which has no closed-form solution. Equation (14) indicates $E_*\left\{\log \mathrm{SSE}_1\right\}$ only depends on σ_o^2, n, and p. Since the two candidate models have the same dimension,

$$E_*\left\{\log \mathrm{SSE}_1\right\} = E_*\left\{\log \mathrm{SSE}_2\right\}.$$

Therefore,

$$E_*\left\{n \log \hat{\sigma}_1^2 + n\right\} = E_*\left\{n \log \hat{\sigma}_2^2 + n\right\}. \tag{15}$$

Next, let us consider the second term in (13), $E_{*+}\{\|\boldsymbol{y}^+ - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\|^2/\hat{\sigma}_1^2 - n\}$, where E_{*+} corresponds to the expectation under the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{y}^+$. It can be argued that $\left\|\boldsymbol{y}^+ - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right\|^2$ and $\hat{\sigma}_1^2$ are independent. Thus,

$$E_{*+}\left\{\frac{\left\|\boldsymbol{y}^+ - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right\|^2}{\hat{\sigma}_1^2} - n\right\} = \frac{E_{*+}\left\{\left\|\boldsymbol{y}^+ - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right\|^2\right\}}{E_*\left\{\hat{\sigma}_1^2\right\}} - n. \tag{16}$$

Further, it can be shown that $E_{*+}\left\{\left\|\boldsymbol{y}^+ - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right\|^2\right\} = \sigma_o^2(n+p)$. Since $n\hat{\sigma}_1^2 \sim \sigma_o^2\chi^2_{n-p}$, we have

$$E_*\left\{\frac{1}{\hat{\sigma}_1^2}\right\} = \frac{n}{\sigma_o^2(n-p-2)}.$$

It follows that (16) can be derived as

$$E_{*+}\left\{\frac{\left\|\boldsymbol{y}^+ - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right\|^2}{\hat{\sigma}_1^2} - n\right\} = \frac{2(p+1)n}{n-p-2}, \tag{17}$$

which is the penalty term of AIC$_c$. Since the column ranks of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are both p,

$$E_{*+}\left\{\frac{\left\|\boldsymbol{y}^+ - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right\|^2}{\hat{\sigma}_1^2} - n\right\} = E_{*+}\left\{\frac{\left\|\boldsymbol{y}^+ - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\right\|^2}{\hat{\sigma}_2^2} - n\right\}. \tag{18}$$

Combining (15) and (18), we see that $\varDelta_{\mathrm{KL}}(1) = \varDelta_{\mathrm{KL}}(2)$. □

Proof of Proposition 2

The proof of Proposition 2 requires the following lemma, which is presented without proof.

Lemma 1 *Suppose u_n and v_n are sequences of random variables such that as $n \to \infty$, $u_n - E(u_n) = o_p(1)$ and $v_n - E(v_n) = o_p(1)$. Then, based on the continuous mapping theorem, as $n \to \infty$,*

$$E\left(\frac{u_n}{v_n}\right) = \frac{E(u_n)}{E(v_n)} + o(1).$$

To prove Proposition 2, We will follow the notation from McQuarrie and Tsai (1998, Chapter 2). Based on Hurvich and Tsai (1989), the mean K–L discrepancies of the two candidate models can be expressed as

$$\bar{\Delta}_{KL}(1) = E_* \left\{ \log \hat{\sigma}_1^2 + \frac{\sigma_o^2}{\hat{\sigma}_1^2} + \frac{\|\boldsymbol{X}_o\boldsymbol{\beta}_o - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\|^2/n}{\hat{\sigma}_1^2} \right\},$$

$$\bar{\Delta}_{KL}(2) = E_* \left\{ \log \hat{\sigma}_2^2 + \frac{\sigma_o^2}{\hat{\sigma}_2^2} + \frac{\|\boldsymbol{X}_o\boldsymbol{\beta}_o - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\|^2/n}{\hat{\sigma}_2^2} \right\}.$$

Thus,

$$\begin{aligned} &\bar{\Delta}_{KL}(1) - \bar{\Delta}_{KL}(2) \\ &\quad = E_* \left\{ \log \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} + \frac{\sigma_o^2 + \|\boldsymbol{X}_o\boldsymbol{\beta}_o - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\|^2/n}{\hat{\sigma}_1^2} - \frac{\sigma_o^2 + \|\boldsymbol{X}_o\boldsymbol{\beta}_o - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\|^2/n}{\hat{\sigma}_2^2} \right\}. \end{aligned} \tag{19}$$

Consider the three terms that appear in the expectation in (19). We will show that the first term is negative, the second term converges to 1, and the third term converges to -1.

The first term $\log\left(\hat{\sigma}_1^2/\hat{\sigma}_2^2\right)$ can be written as $\log(\text{SSE}_1/n) - \log(\text{SSE}_2/n)$. To evaluate this difference, we will need the expectation of SSE for both model (1) and model (2). Since model (1) is correctly specified or overspecified, it can be shown that,

$$E_* \{\log(\text{SSE}_1/n)\} = \log \sigma_o^2 + \log 2 - \log n + \psi\left(\frac{n-p}{2}\right), \tag{20}$$

where ψ is Euler's *psi* function (McQuarrie and Tsai 1998, p. 67). The *psi* function has a useful recursive property, such that $\psi(v+1) = \psi(v) + 1/v$, for $v > 0$. Since model (2) is underspecified, SSE_2/σ_o^2 follows a noncentral $\chi^2(n-p, \lambda)$ distribution, where the noncentrality parameter $\lambda = E(\boldsymbol{y}')(I - \boldsymbol{X}(\boldsymbol{X}_2'\boldsymbol{X}_2)\boldsymbol{X}_2')E(\boldsymbol{y})/\sigma_o^2$. It can be shown that

$$E_* \{\log(\text{SSE}_2/n)\} = \log \sigma_o^2 + \log 2 - \log n + \sum_{r=0}^{\infty} e^{-\lambda/2} \frac{(\lambda/2)^r}{r!} \psi\left(\frac{n-p}{2} + r\right) \tag{21}$$

(McQuarrie and Tsai 1998, p. 47).

Since both model (1) and model (2) have the same dimension p, we can let $k = (n-p)/2$ in both (20) and (21). Substitution of $(n-p)/2$ with k yields

$$\begin{aligned} &E_* \{\log(\text{SSE}_1/n) - \log(\text{SSE}_2/n)\} \\ &\quad = \psi(k) - \sum_{r=0}^{\infty} e^{-\lambda/2} \frac{(\lambda/2)^r}{r!} \psi\,(k+r) \\ &\quad = \sum_{r=0}^{\infty} e^{-\lambda/2} \frac{(\lambda/2)^r}{r!} \psi(k) - \sum_{r=0}^{\infty} e^{-\lambda/2} \frac{(\lambda/2)^r}{r!} \psi\,(k+r) \\ &\quad = \sum_{r=0}^{\infty} e^{-\lambda/2} \frac{(\lambda/2)^r}{r!} \left(\psi(k) - \psi\,(k+r)\right). \end{aligned}$$

The preceding derivation uses the property of the Poisson distribution,

$$\sum_{r=0}^{\infty} e^{-\lambda/2} \frac{(\lambda/2)^r}{r!} = 1.$$

From the recursive property of the ψ function, we have

$$\psi(k+r) = \psi(k) + \frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{k+r-1}.$$

Thus,

$$E_* \{\log(\mathrm{SSE}_1/n) - \log(\mathrm{SSE}_2/n)\} < 0. \tag{22}$$

Now consider the second term in (19). Since model (1) is correctly specified or overspecified, as $n \to \infty$,

$$\begin{aligned} E_* \left\{ \frac{\sigma_o^2 + \left\| \boldsymbol{X}\boldsymbol{\beta}_o - \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1 \right\|^2 /n}{\hat{\sigma}_1^2} \right\} &= \frac{n+p}{n-p-2} \\ &= \frac{1+p/n}{1-p/n-2/n} \\ &\to 1. \end{aligned} \tag{23}$$

For the last term in (19),

$$E_* \left\{ - \frac{\sigma_o^2 + \|\boldsymbol{X}_o\boldsymbol{\beta}_o - \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2\|^2/n}{\hat{\sigma}_2^2} \right\},$$

we will first consider the expectations of the numerator and the denominator, respectively.

Let H_2 denote the projection matrix onto the column space of $\boldsymbol{X}_2$; i.e., $H_2 = \boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T$. For the numerator, we have

$$\begin{aligned} &E_*\{\sigma_o^2 + \|\boldsymbol{X}\boldsymbol{\beta}_o - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\|^2/n\} \\ &\quad = \sigma_o^2 + E_*\{\|\boldsymbol{X}\boldsymbol{\beta}_o - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\|^2\}/n \\ &\quad = \sigma_o^2 + \sigma_o^2 \mathrm{tr}(H_2)/n + (\boldsymbol{X}_o\boldsymbol{\beta}_o - H_2\boldsymbol{X}_o\boldsymbol{\beta}_o)^T(\boldsymbol{X}_o\boldsymbol{\beta}_o - H_2\boldsymbol{X}_o\boldsymbol{\beta}_o)/n \\ &\quad = \sigma_o^2(1 + p/n) + (\boldsymbol{X}_o\boldsymbol{\beta}_o)^T(I - H_2)(\boldsymbol{X}_o\boldsymbol{\beta}_o)/n. \end{aligned}$$

For the denominator, we have

$$\begin{aligned} E_*\{\hat{\sigma}_2^2\} &= \frac{1}{n} E_*\{\boldsymbol{y}^T(I - H_2)\boldsymbol{y}\} \\ &= \frac{1}{n}\{\sigma_o^2 \mathrm{tr}(I - H_2) + (\boldsymbol{X}_o\boldsymbol{\beta}_o)^T(I - H_2)(\boldsymbol{X}_o\boldsymbol{\beta}_o)\} \end{aligned}$$

$$= \sigma_o^2(1 - p/n) + (\boldsymbol{X}_o\boldsymbol{\beta}_o)^T(I - H_2)(\boldsymbol{X}_o\boldsymbol{\beta}_o)/n.$$

To combine the expectations of the numerator and the denominator as previously derived, we apply Lemma 1:

$$\begin{aligned} &E_*\left\{-\frac{\sigma_o^2 + \|\boldsymbol{X}_o\boldsymbol{\beta}_o - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\|^2/n}{\hat{\sigma}_2^2}\right\} \\ &= -\frac{E_*\{\sigma_o^2 + \|\boldsymbol{X}_o\boldsymbol{\beta}_o - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\|^2/n\}}{E_*\{\hat{\sigma}_2^2\}} + o(1) \\ &= -\frac{\sigma_o^2(1 + p/n) + (\boldsymbol{X}_o\boldsymbol{\beta}_o)^T(I - H_2)(\boldsymbol{X}_o\boldsymbol{\beta}_o)/n}{\sigma_o^2(1 - p/n) + (\boldsymbol{X}_o\boldsymbol{\beta}_o)^T(I - H_2)(\boldsymbol{X}_o\boldsymbol{\beta}_o)/n} + o(1). \end{aligned}$$

It is generally assumed that $(\boldsymbol{X}_o\boldsymbol{\beta}_o)^T(I - H_2)(\boldsymbol{X}_o\boldsymbol{\beta}_o)$ is $O(n)$ (Fujikoshi and Satoh 1997). Thus, as $n \to \infty$,

$$E_*\left\{-\frac{\sigma_o^2 + \|\boldsymbol{X}_o\boldsymbol{\beta}_o - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\|^2/n}{\hat{\sigma}_2^2}\right\} \to -1 \tag{24}$$

Combining (22), (23), and (24), we see that as $n \to \infty$, $\bar{\Delta}_{\mathrm{KL}}(1) - \bar{\Delta}_{\mathrm{KL}}(2) \to C$, where $C < 0$. □

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F (eds) 2nd international symposium on information theory. Akadémia Kiadó, Budapest, pp 267–281

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control AC–19:716–723

Beale EML, Kendall MG, Mann DW (1967) The discarding of variables in multivariate analysis. Biometrika 54:357–366

Bedrick EJ, Tsai CL (1994) Model selection for multivariate regression in small samples. Biometrics 50:226–231

Bengtsson T, Cavanaugh JE (2006) An improved Akaike information criterion for state-space model selection. Comput Stat Data Anal 50:2635–2654

Cavanaugh JE (2004) Criteria for linear model selection based on Kullback's symmetric divergence. Aust N Z J Stat 46:257–274

Davison AC (2003) Statistical models. Cambridge University Press, Cambridge

Efron B (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 78:316–331

Efron B (1986) How biased is the apparent error rate of a prediction rule? J Am Stat Assoc 81:461–470

Fujikoshi Y, Satoh K (1997) Modified AIC and C_p in multivariate linear regression. Biometrika 84:707–716

Hocking RR, Leslie RN (1967) Selection of the best subset in regression analysis. Technometrics 9:531–540

Hurvich CM, Shumway RH, Tsai CL (1990) Improved estimators of Kullback–Leibler information for autoregressive model selection in small samples. Biometrika 77:709–719

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76:297–307

Hurvich CM, Tsai CL (1993) A corrected Akaike information criterion for vector autoregressive model selection. J Time Ser Anal 14:271–279

Kullback S (1968) Information theory and statistics. Dover, New York

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

LaMotte LR, Hocking RR (1970) Computational efficiency in the selection of regression variables. Technometrics 12:83–93

Lumley T, using Fortran code by A. Miller (2009) Leaps: regression subset selection. R package version 2.9. http://CRAN.R-project.org/web/packages/leaps

Maurer W, Margolin BH (1976) The multivariate inclusion–exclusion formula and order statistics from dependent variates. Ann Stat 4:1190–1199

McQuarrie ADR, Tsai C-L (1998) Regression and time series model selection. World Scientific, River Edge

R Core Team (2013) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Shibata R (1981) An optimal selection of regression variables. Biometrika 68:45–54

Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. Commun Stat A7:13–26

Tibshirani R, Knight K (1999) The covariance inflation criterion for adaptive model selection. J R Stat Soc Ser B 61:529–546