Statistics and Probability Letters 173 (2021) 109064

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Correcting the corrected AIC

Timothy DelSole [a,*], Michael K. Tippett [b]

[a] Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, VA, USA
[b] Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA

ARTICLE INFO

Article history:
Received 9 October 2020
Received in revised form 1 February 2021
Accepted 5 February 2021
Available online 17 February 2021

Keywords:
Model selection
Akaike's Information Criterion
AICc
ANCOVA

ABSTRACT

The standard correction to Akaike's Information Criterion, AICc, assumes the same predictors for training and verification and therefore underestimates prediction error for random predictors. A corrected AIC for regression models containing a mix of random and fixed predictors is derived.

1. Introduction

Akaike's Information Criterion (AIC) has a known tendency to select overfitted models. Hurvich and Tsai (1989) showed that the cause of this overfitting tendency lies in the asymptotic approximations used to derive AIC. To derive a bias-corrected version of AIC, Hurvich and Tsai (1989) evaluated the Kullback–Leibler (KL) divergence *exactly* for normal distributions, assuming the candidate family of models includes the true model. The resulting criterion, AICc, often outperforms its competitors (McQuarrie and Tsai, 1998) and has become a standard criterion recommended by many investigators (e.g., Burnham and Anderson, 2002, p66). However, an assumption that is not always emphasized in the derivation of AICc is that predictor values are the *same* in the training and validation samples. Rosset and Tibshirani (2020) call this the "Same-X" assumption, and note that many model selection criteria implicitly assume Same-X. In contrast, many applications of model selection fall under the "Random-X" assumption, in which predictor values differ from training to validation. Although the Same-X and Random-X distinction has been known for some time (see Rosset and Tibshirani, 2020, for a review of this literature), the generalization of standard model selection criteria to Random-X is more recent. For instance, the extension of Mallows' Cp to Random-X has appeared only recently (Rosset and Tibshirani, 2020). In this paper, we derive a new criterion, AICm, which is an exactly unbiased estimate of the Kullback–Leibler-based criterion for regression models containing an arbitrary mix of Same-X and Random-X predictors. Such models include the Analysis of Covariance (ANCOVA) model. The multivariate generalization of AICm also is derived.

Under Same-X, AICm equals AICc. Under Random-X, AICm leads to a new criterion that we call AICr. We use the same numerical model as (Hurvich and Tsai, 1989) to show that AICc is indeed biased for Random-X and that it is more likely to select overfitted models than AICr. This paper complements Tian et al. (2020), who derive several model selection criterion under Random-X. Their RAIC differs from our AICr only by the fact that AICr accounts for the intercept. A notable fact is that Fujikoshi (1985) derived a criterion for selecting X-variables in Canonical Correlation Analysis. That criterion is equivalent to selection based on *differences* of AICr derived in this paper.

* Corresponding author.
E-mail address: tdelsole@gmu.edu (T. DelSole).

https://doi.org/10.1016/j.spl.2021.109064

2. The corrected AIC for Random-X

We consider the problem of predicting y based on x. Let the conditional PDF of y given x be $p(y|x)$, where x and y denote explanatory and response variables, respectively. We call $p(y|x)$ the *true* PDF. The *candidate* PDF is denoted $q(y|x;\phi)$, where ϕ denotes model parameters. In this notation, y, x, and ϕ could be multivariate. By familiar arguments, the Kullback–Leibler divergence leads to a model selection criterion based on the expected value of $-2\log q(y|x;\phi)$ (Akaike, 1974; Hurvich and Tsai, 1989; McQuarrie and Tsai, 1998).

To estimate the KL divergence, let ϕ_* denote an estimate of ϕ derived from the training sample (x_*, y_*), and let (x_0, y_0) denote the validation sample. Both samples are drawn from the true PDF, but y_* and y_0 are conditionally independent given (x_*, x_0); i.e., $p(y_*, y_0|x_*, x_0) = p(y_*|x_*)p(y_0|x_0)$. Then, we consider

$$\Delta(X_0, X_*, y_*) = -2\mathbb{E}_{Y_0|X_0,X_*,y_*}[\log q(y_0|x_0;\phi_*)], \tag{1}$$

where $\mathbb{E}_{Y_0|X_0,X_*,y_*}[\cdot]$ denotes the expectation over $p(y_0|x_0, x_*, y_*)$.

For normal distributions, the PDF $q(y|x;\phi)$ can be derived from the model

$$\underset{N\times 1}{\mathbf{y}} = \underset{N\times M}{\mathbf{X}}\ \underset{M\times 1}{\boldsymbol{\beta}} + \underset{N\times 1}{\boldsymbol{\epsilon}}, \tag{2}$$

where N is the sample size, M is the number of explanatory variables, $\boldsymbol{\beta}$ contains the regression coefficients, $\boldsymbol{\epsilon}$ is a random vector, and the dimension of each term is indicated below it. The elements of $\boldsymbol{\epsilon}$ are independent and identically distributed normal random variables with zero mean and variance σ^2. Let $\boldsymbol{\beta}_*$ and σ_*^2 denote the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2, respectively, derived from the sample $(\mathbf{X}_*, \mathbf{y}_*)$. Then,

$$\Delta(X_0, X_*, y_*) = N\log 2\pi + N\log\sigma_*^2 + \mathbb{E}_{Y_0|X_0,X_*,y_*}\left[\|\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}_*\|^2\right]/\sigma_*^2, \tag{3}$$

where conditional independence of $(\boldsymbol{\beta}_*, \sigma_*^2, \mathbf{y}_0)$ given $(\mathbf{X}_0, \mathbf{X}_*, \mathbf{y}_*)$ has been used. It is understood that $\mathbf{X}_0$ and $\mathbf{X}_*$ are each of dimension $N \times M$, and $\mathbf{y}_0$ and $\mathbf{y}_*$ are each of dimension $N \times 1$.

We assume that the candidate family of models includes the true model. Therefore, $\mathbf{y}_0 = \mathbf{X}_0\boldsymbol{\beta} + \boldsymbol{\epsilon}_0$, where $\boldsymbol{\epsilon}_0$ has the same distribution as $\boldsymbol{\epsilon}$ and is independent of $\boldsymbol{\epsilon}$, and

$$\begin{aligned}\mathbb{E}_{Y_0|X_0,X_*,y_*}\left[\|\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}_*\|^2\right] &= \mathbb{E}_{Y_0|X_0,X_*,y_*}\left[\|\mathbf{X}_0\boldsymbol{\beta} + \boldsymbol{\epsilon}_0 - \mathbf{X}_0\boldsymbol{\beta}_*\|^2\right] = \mathbb{E}_{Y_0|X_0,X_*,y_*}\left[\|\boldsymbol{\epsilon}_0\|^2 + \|\mathbf{X}_0\left(\boldsymbol{\beta}_* - \boldsymbol{\beta}\right)\|^2\right]\\ &= N\sigma^2 + \|\mathbf{X}_0\left(\boldsymbol{\beta}_* - \boldsymbol{\beta}\right)\|^2.\end{aligned}$$

Following Akaike, we take the expectation of (3) with respect to $p(y_*|x_*)$. From standard regression theory (Seber and Lee, 2003, theorem 3.5),

$$\boldsymbol{\beta}_* - \boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\left(\mathbf{X}_*^T\mathbf{X}_*\right)^{-1}\right), \tag{4}$$

and hence

$$\mathbb{E}_{Y_*|X_0,X_*}\left[\|\mathbf{X}_0\left(\boldsymbol{\beta}_* - \boldsymbol{\beta}\right)\|^2\right] = \sigma^2\,\mathrm{tr}\left[\mathbf{X}_0\left(\mathbf{X}_*^T\mathbf{X}_*\right)^{-1}\mathbf{X}_0^T\right]. \tag{5}$$

Also by standard regression theory,

$$N\frac{\sigma_*^2}{\sigma^2} \sim \chi^2_{N-M}, \tag{6}$$

and hence

$$\mathbb{E}_{Y_*|X_*}\left[\frac{1}{\sigma_*^2}\right] = \left(\frac{N}{N-M-2}\right)\frac{1}{\sigma^2}. \tag{7}$$

Consolidating these results gives

$$\mathbb{E}_{Y_*|X_*,X_0}\left[\Delta(X_0, X_*, y_*)\right] = N\log 2\pi + N\mathbb{E}_{Y_*|X_*}\left[\log\sigma_*^2\right] + \frac{N}{N-M-2}\left[N + \mathrm{tr}\left[\mathbf{X}_0\left(\mathbf{X}_*^T\mathbf{X}_*\right)^{-1}\mathbf{X}_0^T\right]\right]. \tag{8}$$

In general, explanatory variables may consist of a mix of random and fixed variables. Accordingly, partition $\mathbf{X}_0$ as

$$\mathbf{X}_0 = \begin{bmatrix}\mathbf{F} & \mathbf{R}_0\end{bmatrix}, \tag{9}$$

where $\mathbf{F}$ is a fixed $N \times M_F$ matrix, and $\mathbf{R}_0$ is a random $N \times M_R$ matrix, with $M = M_F + M_R$. The rows of $\mathbf{R}_0$ are independent realizations from a multivariate normal distribution. Exactly one column of $\mathbf{F}$ is a vector of ones corresponding to the intercept, hence $1 \le M_F \le M$. The resulting design matrix (9) includes the Analysis of Covariance (ANCOVA) model. Similarly, define

$$\mathbf{X}_* = \begin{bmatrix}\mathbf{F} & \mathbf{R}_*\end{bmatrix}, \tag{10}$$

where $\mathbf{R}_*$ is drawn from the same distribution as $\mathbf{R}_0$ but is independent of $\mathbf{R}_0$. $\mathbf{F}$ is the same in (9) and (10). The following lemma is proven in the appendix.

Lemma 1. *Let* $\mathbf{X}_0$ *and* $\mathbf{X}_*$ *be defined as in the previous paragraph. Then*

$$\mathbb{E}_{R_0,R_*}\left[\mathrm{tr}\left[\mathbf{X}_0\left(\mathbf{X}_*^T\mathbf{X}_*\right)^{-1}\mathbf{X}_0^T\right]\right] = M_F + \frac{M_R(N+M_F)}{N-M-1}, \tag{11}$$

where $1 \le M_F \le M$ *and* $M_R = M - M_F$.

To estimate (8), we use $\log\sigma_*^2$ as an unbiased estimate of $\mathbb{E}_{Y_*|X_*}\left[\log\sigma_*^2\right]$, and invoke Lemma 1, which yields the following proposition.

Proposition 1. *An unbiased estimate of* (8) *under* (9) and (10) *is*

$$AICm(Y|X) = N\log 2\pi + N\log\sigma_*^2 + \frac{N(N+M_F)}{N-M-2}\left(1+\frac{M_R}{N-M-1}\right), \tag{12}$$

where "m" emphasizes a mix of Random-X and Same-X explanatory variables.

Two special cases are of interest. The first is $\mathbf{X}_0 = \mathbf{X}_* = \mathbf{F}$, which is the case for both Fixed-X and Same-X as defined by Rosset and Tibshirani (2020). Fixed-X means $\mathbf{F}$ is fixed and Same-X means $\mathbf{F}$ is random. The appendix shows that Lemma 1 holds true for both Same-X and Fixed-X. Therefore, both Fixed-X and Same-X correspond to $M_F = M$ and $M_R = 0$ in Proposition 1.

Proposition 2. *An unbiased estimate of* (8) *under Same-X or Fixed-X is*

$$AICc(Y|X) = N\log 2\pi + N\log\sigma_*^2 + \frac{N(N+M)}{N-M-2}. \tag{13}$$

This expression is precisely the AICc derived in Hurvich and Tsai (1989). Therefore, the standard correction for AIC corresponds to Fixed-X and Same-X.

The second special case is Random-X, defined as follows.

Definition 1 (*Random-X*)**.** Random-X means that $\mathbf{X}_0$ and $\mathbf{X}_*$ are defined as in (9) and (10) with $M_F = 1$, where $\mathbf{F}$ is an N-dimensional vector of ones to account for the intercept.

Our definition of Random-X differs from that of Rosset and Tibshirani (2020) by including an intercept term. We include the intercept in Random-X so that the expectation of (8) does not depend on the mean of $\mathbf{X}_0$ and $\mathbf{X}_*$. Random-X corresponds to $M_F = 1$ and $M_R = M - 1$ in Proposition 1.

Proposition 3. *An unbiased estimate of* (8) *under Random-X is*

$$AICr(Y|X) = N\log 2\pi + N\log\sigma_*^2 + \frac{N(N+1)}{N-M-2}\left(1+\frac{M-1}{N-M-1}\right), \tag{14}$$

where the "r" is to emphasize that the explanatory variables are random.

The difference between AICr and AICc is

$$\mathrm{AICr}(Y|X) - \mathrm{AICc}(Y|X) = \frac{(M-1)(M+2)}{(N-M-1)(N-M-2)}, \tag{15}$$

which is positive for all $M > 1$. It follows that AICc underestimates the out-of-sample prediction uncertainty under Random-X. This is to be expected: if X is random, then its difference between training and validation samples contributes a source of prediction uncertainty that is missing when X is assumed fixed. For given N, this bias grows faster-than-quadratically with M. The fact that the bias grows with M means that AICc is more likely than AICr to select overfitted models under Random-X.

Hurvich and Tsai (1989) also consider model selection for autoregressive (AR) models, for which Random-X is clearly more appropriate than Same-X. However, past justifications of AICc for AR processes have been based on asymptotic arguments (Hurvich and Tsai, 1989; Brockwell and Davis, 2002). For serially correlated processes, Lemma 1 does not hold, for reasons discussed in Appendix 1. The exact information criterion for AR processes is not known for small N, even for Gaussian processes.

3. Numerical simulations

Following Hurvich and Tsai (1989), we compare selection criteria using realizations from model (2) with $\boldsymbol{\beta} = (1,2,3,0,0,0,0)^T$, $\sigma^2 = 1$, and $N = 10$. The candidate models contain at most seven explanatory variables, hence $\mathbf{X}$ is $N \times 7$. The candidate models are evaluated sequentially such that the m'th model uses the first m columns of $\mathbf{X}$.

In contrast to Hurvich and Tsai (1989), we use 10 000 realizations instead of 100 (for more accuracy), and the first column of $\mathbf{X}$ is a vector of ones corresponding to the intercept term. For Random-X, we generate 10 000 realizations of $\mathbf{X}_0$,

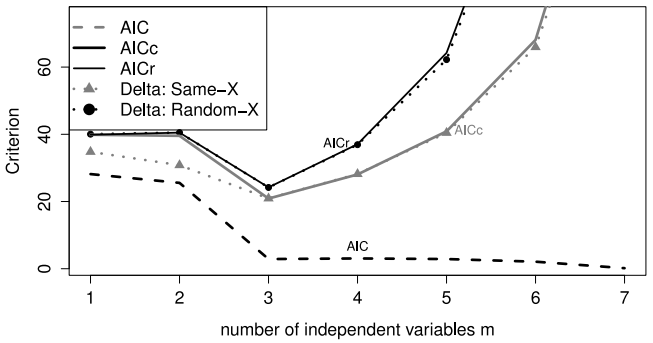

Fig. 1. Selection criteria and the Kullback–Leibler divergence averaged over 10 000 realizations from the regression model (2) with $M = 7$, $N = 10$, and other details described in the text. The example is similar to that in Fig. 1 of Hurvich and Tsai (1989)..

Table 1
Probability of selecting a candidate model of order m when the true order is $m = 3$. Probabilities are estimated from 10 000 simulations.

m	AIC	AICc	AICr
1	0.00	0.00	0.01
2	0.00	0.00	0.01
3	0.32	0.97	0.98
4	0.10	0.03	0.00
5	0.09	0.00	0.00
6	0.14	0.00	0.00
7	0.34	0.00	0.00

and *another* 10 000 independent realizations of $\mathbf{X}_*$, such that each row of the last $M-1$ columns are independent identically distributed normal random variables. The corresponding estimate of Δ is

$$\hat{\Delta}_{\text{Random-X}}(X_0, X_*, y_*) = \frac{1}{10000}\sum_{k=1}^{10000}\left(N\log\left(\sigma_*^{(k)}\right)^2 + \frac{\|\mathbf{y}_0^{(k)} - \mathbf{X}_0^{(k)}\boldsymbol{\beta}_*^{(k)}\|^2}{\left(\sigma_*^{(k)}\right)^2}\right),$$

where superscript (k) indicates the estimate derived from the k'th realization. For Same-X, 10 000 realizations of $\mathbf{X}_*$ are generated in the same way, and $\mathbf{X}_0$ is set equal to $\mathbf{X}_*$. The corresponding sample estimate of Δ is

$$\hat{\Delta}_{\text{Same-X}}(X_*, X_*, y_*) = \frac{1}{10000}\sum_{k=1}^{10000}\left(N\log\left(\sigma_*^{(k)}\right)^2 + \frac{N\sigma^2}{\left(\sigma_*^{(k)}\right)^2} + \frac{\|\mathbf{X}_*^{(k)}\left(\boldsymbol{\beta}_*^{(k)} - \boldsymbol{\beta}\right)\|^2}{\left(\sigma_*^{(k)}\right)^2}\right).$$

It is understood that terms like $\mathbf{X}_0\boldsymbol{\beta}_*$ and $\mathbf{X}_*\boldsymbol{\beta}_*$ are evaluated using only the first m columns of the candidate model.

The average values of AIC, AICc, AICr, $\hat{\Delta}_{\text{Random-X}}$ and $\hat{\Delta}_{\text{Same-X}}$ are shown in Fig. 1. The results for AIC, AICc, and $\hat{\Delta}_{\text{Random-X}}$ essentially reproduce those of Fig. 1 in Hurvich and Tsai (1989), with our $\hat{\Delta}_{\text{Random-X}}$ corresponding to Δ in Hurvich and Tsai (1989). As can be seen, AIC is a strongly negatively biased estimator of $\Delta_{\text{Random-X}}$, whereas AICc is less biased. Nevertheless, AICc is *still negatively biased* relative to $\Delta_{\text{Random-X}}$. This bias also is evident in Fig. 1 in Hurvich and Tsai (1989). For $m \geq 3$, AICc is an unbiased estimate of $\Delta_{\text{Same-X}}$. These results confirm that AICc is an unbiased estimate of the information criterion for Same-X, but not for Random-X. In contrast, AICr is an unbiased estimate of the information criterion for Random-X.

Table 1 shows the probability of choosing the candidate model of order m over the true model of order 3 for the three criteria. As can be seen, AIC has large probability of selecting overfitted models (i.e., 67% for $m > 3$). In contrast, AICc selects overfitted models 3% of the time, while AICr selects overfitted models 0% of the time, confirming that AICr is less likely to select overfitted models than AICc. In this particular example from Hurvich and Tsai (1989), the probability of selecting overfitted models is small, but other examples that highlight the overfitting tendencies of AICc could be contrived.

Hurvich and Tsai (1989) also consider the autoregressive model

$$y_t = 0.99y_{t-1} - 0.8y_{t-2} + \epsilon_t \quad (t = 0, \ldots, N-1), \tag{16}$$

where ϵ_t is a Gaussian white noise process with zero mean and unit variance. We fit realizations from this model to an order-m autoregressive model, where order-0 corresponds to the intercept-only model. The fit is based on least squares

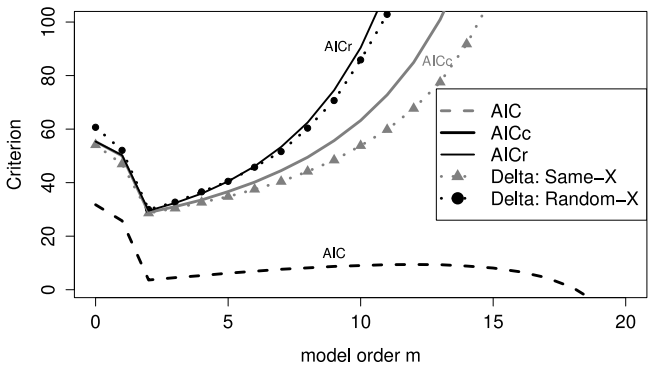

Fig. 2. Average criteria and Kullback–Leibler divergence averaged over 10 000 realizations from the autoregressive model (16) based on $N = 23$. The example is similar to that in Fig. 2 of Hurvich and Tsai (1989)..

Table 2
Probability of selecting a particular candidate AR model of order p when the true order is $p = 2$. Probabilities are estimated from 10 000 simulations from model (16).

p	AIC	AICc	AICr
0	0.00	0.00	0.00
1	0.00	0.01	0.01
2	0.24	0.88	0.92
3	0.04	0.07	0.05
4	0.02	0.03	0.01
5	0.01	0.01	0.00
⋮	⋮	⋮	⋮
17	0.04	0.00	0.00
18	0.11	0.00	0.00
19	0.43	0.00	0.00

estimation of a model of the form (2), where **y** is a 23-dimensional vector from the process, the first column of **X** is a vector of ones corresponding to the intercept, and the next two columns of **X** are 1-step and 2-step lagged versions of the process. Therefore, $M = 3$ and $N = 23$. The corresponding estimates of the criteria are shown in Fig. 2. The results are similar to those of Fig. 2 in Hurvich and Tsai (1989), particularly in showing that AIC fails to reach a minimum and that AICc follows the shape of $\hat{\Delta}_{\text{random-X}}$. The probability of selecting particular models are shown in Table 2. The results are similar to those in Table 1, particularly in showing that AIC tends to select the maximum order and that AICr tends to select more parsimonious models than AICc. However, AICc is negatively biased relative to $\Delta_{\text{Same-X}}$. This discrepancy is presumably due to serial correlation in the data, which violates the assumptions under which AICc was derived. AICr appears to be a nearly unbiased estimate of $\hat{\Delta}_{\text{random-X}}$, but this is fortuitous to this example. For instance, if an AR(1) model is used, then the bias of AICr is evident, and AICc is even more biased (not shown). These results show that neither AICc nor AICr are exactly unbiased for AR processes. These biases reflect the fact that neither criterion is derived by exact integration of an information criterion for AR processes.

4. Multivariate AICm

We now derive AICm for the multivariate regression model

$$\underset{N\times P}{\mathbf{Y}} = \underset{N\times M}{\mathbf{X}}\;\underset{M\times P}{\mathbf{B}} + \underset{N\times P}{\mathbf{E}}\,, \tag{17}$$

where N is sample size, P is the number of response variables in **Y**, M is the number of explanatory variables in **X**, **B** contains regression coefficients, **E** is a random matrix. Each row of **E** is independently distributed as a multivariate normal with zero mean and covariance matrix $\boldsymbol{\Sigma}$.

Following the univariate derivation, let $\mathbf{B}_*$ and $\boldsymbol{\Sigma}_*$ denote the maximum likelihood estimators of **B** and $\boldsymbol{\Sigma}$, respectively, derived from the training sample $(\mathbf{X}_*, \mathbf{Y}_*)$, and let $(\mathbf{X}_0, \mathbf{Y}_0)$ be the validation sample used to estimate KL divergence. Then,

$$-2\log q(y_0|x_0;\phi^*) = NP\log(2\pi) + N\log|\boldsymbol{\Sigma}_*| + Q\,. \tag{18}$$

where $|\cdot|$ denotes the determinant and

$$Q = \mathrm{tr}\left[(\mathbf{Y}_0 - \mathbf{X}_0\mathbf{B}_*)\,\boldsymbol{\Sigma}_*^{-1}\,(\mathbf{Y}_0 - \mathbf{X}_0\mathbf{B}_*)^T\right]. \tag{19}$$

Assuming the candidate family of models includes the true model, $\mathbf{Y}_0 = \mathbf{X}_0\mathbf{B} + \mathbf{E}_0$, where the rows of $\mathbf{E}_0$ and $\mathbf{E}$ are independent, then

$$\begin{aligned}\mathbb{E}_{Y_0|X_0,X_*,Y_*}[Q] &= \mathbb{E}_{Y_0|X_0,X_*,Y_*}\left[\mathrm{tr}\left[\boldsymbol{\Sigma}_*^{-1}\mathbf{E}_0^T\mathbf{E}_0 + \boldsymbol{\Sigma}_*^{-1}\,(\mathbf{B}_* - \mathbf{B})^T\,\mathbf{X}_0^T\mathbf{X}_0\,(\mathbf{B}_* - \mathbf{B})\right]\right]\\ &= N\,\mathrm{tr}\left[\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Sigma}\right] + \mathrm{tr}\left[\boldsymbol{\Sigma}_*^{-1}\,(\mathbf{B}_* - \mathbf{B})^T\,\mathbf{X}_0^T\mathbf{X}_0\,(\mathbf{B}_* - \mathbf{B})\right].\end{aligned}$$

From standard regression theory, $(\mathbf{B}_*, \mathbf{X}_0, \boldsymbol{\Sigma}_*)$ are conditionally independent given $(\mathbf{X}_*, \mathbf{X}_0)$, and $N\boldsymbol{\Sigma}_*$ has Wishart distribution $\mathcal{W}_P[N - M, \boldsymbol{\Sigma}]$ (theorems 6.2.2 and 6.2.3 in Mardia et al., 1979). Standard properties of Wishart-distributed matrices (Muirhead, 2009, section 3.2.3) give

$$\mathbb{E}_{Y^*|X^*}\left[\boldsymbol{\Sigma}_*^{-1}\right] = \frac{N}{N - M - P - 1}\boldsymbol{\Sigma}^{-1}. \tag{20}$$

Using the fact that the covariance between $(\mathbf{B}_*)_{ij}$ and $(\mathbf{B}_*)_{kl}$ is $(\boldsymbol{\Sigma})_{ik}((\mathbf{X}_*^T\mathbf{X}_*)^{-1})_{jl}$ (theorem 6.2.3 in Mardia et al., 1979), it follows that

$$\mathbb{E}_{Y_0Y_*|X_0X_*}[Q] = \frac{NP}{N - M - P - 1}\left(N + \mathrm{tr}\left[\mathbf{X}_0\left(\mathbf{X}_*^T\mathbf{X}_*\right)^{-1}\mathbf{X}_0^T\right]\right). \tag{21}$$

Therefore, the information criterion for multivariate regression model (17) is

$$\mathbb{E}_{Y_0Y_*|X_0X_*}\left[-2\log q(y_0|x_0;\phi^*)\right] = NP\log(2\pi) + N\mathbb{E}_{Y_*|X_*}\log|\boldsymbol{\Sigma}_*| + \frac{NP}{N - M - P - 1}\left(N + \mathrm{tr}\left[\mathbf{X}_0\left(\mathbf{X}_*^T\mathbf{X}_*\right)^{-1}\mathbf{X}_0^T\right]\right), \tag{22}$$

which is the multivariate generalization of Δ in (8) (to which it reduces when $P = 1$). Invoking Lemma 1 gives the following selection criterion for multivariate regression with an arbitrary mix of Random-X and Same-X predictors.

Proposition 4. *An unbiased estimate of the information criterion (22) for the multivariate regression model (17) under (9) and (10) is*

$$AICm(Y|X) = N\log|\boldsymbol{\Sigma}_*| + NP\log 2\pi + \frac{N(N + M_F)P}{N - M - P - 1}\left(1 + \frac{M_R}{N - M - 1}\right).$$

Substituting $M_F = M$ and $M_R = 0$ gives the Same-X AICc,

$$\mathrm{AICc}(Y|X) = N\log|\boldsymbol{\Sigma}_*| + NP\log 2\pi + \frac{(N + M)NP}{N - M - P - 1},$$

which agrees with the AICc in Bedrick and Tsai (1994). Substituting $M_F = 1$ and $M_R = M - 1$ gives the multivariate generation of AICr for Random-X,

$$\mathrm{AICr}(Y|X) = N\log|\boldsymbol{\Sigma}_*| + NP\log 2\pi + \frac{N(N + 1)P}{N - M - P - 1}\left(1 + \frac{M - 1}{N - M - 1}\right),$$

which agrees with (14) for $P = 1$. An equivalent expression for AICr is

$$\mathrm{AICr}(Y|X) = N\log 2\pi + N\log\sigma_*^2 + (N + 1)\left(\frac{N(M_R + P)}{N - M_R - P - 2} - \frac{N(M_R)}{N - M_R - 2}\right).$$

Using this expression to compute the difference in AICr between nested models yields (5.17) of Fujikoshi (1985), where the latter is the AIC criterion for selecting the "best subset" of X-variables in Canonical Correlation Analysis. This reveals that Fujikoshi (1985) derived the corrected AIC for Random-X well before (Rosset and Tibshirani, 2020) and related papers. DelSole and Tippett (2020) give yet another derivation of AICr from a different perspective.

Acknowledgments

We thank Anthony Davison and Richard Davis for helpful comments on an early version of this paper. This research was supported primarily by the National Science Foundation, United States (AGS-1822221). Additional support was provided from National Science Foundation, United States (AGS-1338427), National Aeronautics and Space Administration, United States (NNX14AM19G), the National Oceanic and Atmospheric Administration, United States (NA14OAR4310160).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2021.109064.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723.
Bedrick, E.J., Tsai, C.-L., 1994. Model selection for multivariate regression in small samples. Biometrics 50 (1), 226–231, URL http://www.jstor.org/stable/2533213.
Brockwell, P.J., Davis, R.A., 2002. Introduction to Time Series and Forecasting. Springer.
Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, second ed. Springer.
DelSole, T., Tippett, M.K., 2020. A selection criterion for canonical correlation analysis, covariance structure, and graphical models. Stat (submitted for publication).
Fujikoshi, Y., 1985. Selection of variables in discriminant analysis and canonical correlation analysis. In: Krishnaiah, P.R. (Ed.), Multivariate Analysis VI: Proceedings of the Sixth International Symposium on Multivariate Analysis. Elsevier, pp. 219–236.
Hurvich, C.M., Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.
Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press.
McQuarrie, A.D.R., Tsai, C.-L., 1998. Regression and Time Series Model Selection. World Scientific Publishing.
Muirhead, R.J., 2009. Aspects of Multivariate Statistical Theory. John Wiley & Sons.
Rosset, S., Tibshirani, R.J., 2020. From fixed-x to random-x regression: Bias–variance decompositions, covariance penalties, and prediction error estimation. J. Amer. Statist. Assoc. 115 (529), 138–151, URL https://doi.org/10.1080/01621459.2018.1424632.
Seber, G.A.F., Lee, A.J., 2003. Linear Regression Analysis. Wiley-Interscience.
Tian, S., Hurvich, C., Simonoff, J., 2020. Selection of regression models under linear restrictions for fixed and random designs. ArXiv e-prints:2009.10029.