Linear Algebra and its Applications 455 (2014) 82–106

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Jackknife bias correction of the AIC for selecting variables in canonical correlation analysis under model misspecification ☆

Yusuke Hashiyama [1], Hirokazu Yanagihara [*,2], Yasunori Fujikoshi [3]

Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

ARTICLE INFO

Article history:
Received 23 May 2012
Accepted 28 April 2014
Available online 20 May 2014
Submitted by L.-H. Lim

MSC:
62F40
62H20

Keywords:
Bias correction
Canonical correlation analysis
Jackknife method
Model misspecification
Nonnormality
Variable selection

ABSTRACT

In this paper, we deal with a bias correction of Akaike's information criterion (AIC) for selecting variables in the canonical correlation analysis when a goodness of fit of the model is assessed by the risk function consisting of the expected Kullback–Leibler loss function with a normal assumption. Although the bias of the AIC to the risk function is $O(n^{-1})$ when the model is correctly specified, its order turns into $O(1)$ when the model is misspecified, where n is the sample size. By using the leave-two-out jackknife method with a constant adjustment, we propose a new criterion that reduces the AIC's bias to $O(n^{-2})$ even when the model is misspecified, and is an exact unbiased estimator of the risk function when data is generated from the normal distribution. Additionally, by applying basic theorems of linear algebra, e.g., the formula of an inverse of the sum of matrices and a simple property of an inverse matrix, to our problem, we

☆ The authors thank Prof. Hirofumi Wakaki, Hiroshima University, for helpful comments on the moment condition of the expansion. Also the authors would like to thank the two referees for helpful suggestions.

* Corresponding author.
E-mail address: yanagi@math.sci.hiroshima-u.ac.jp (H. Yanagihara).

[1] Current address: Osaka Prefectural Ibaraki High School, 12-1 Shinjo-cho, Ibaraki, Osaka 567-0884, Japan.

[2] Supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Challenging Exploratory Research, #25540012, 2013–2015.

[3] Supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Scientific Research (C), #25330038, 2013–2015.

http://dx.doi.org/10.1016/j.laa.2014.04.028

obtain strict conditions to guarantee the validity of the bias correction, and another expression of the proposed criterion to reduce computational time tremendously, which does not contain any jackknife estimators.

1. Introduction

Canonical correlation analysis (CCA), which analyzes the correlation of two linearly combined variables, is an important method in multivariate analysis, and a solution of its optimization problem can be solved by an eigenvalue problem. CCA has been introduced in many textbooks for applied statistical analysis (see e.g., [1, Chap. 14.7], [2, Chap. 8.7]), and even now it is widely used in many applied fields (e.g., [3–5]). Determining the variables to be used is an important problem in CCA as well as selecting the variables in the multivariate linear regression model. Hence, variable selection in CCA has been investigated in many papers, e.g., [6–12].

The choice of variables based on the minimization of an information criterion is one of the major variable selection methods. Under such a variable selection method, a model consisting of a combination of interested variables is called a candidate model, and a candidate model having the smallest information criterion among all candidate models is regarded as the empirical best (or simply best) model. Then the set of variables in the best model is regarded as the best subset of variables. The most famous and widely-used information criterion for variable selection is Akaike's information criterion (AIC), which was proposed by [13] and [14]. In the AIC, a goodness of fit of a candidate model is measured by the Kullback–Leibler (KL) discrepancy [15] function. The AIC is defined by adding an estimator of a bias to the risk function consisting of the expected KL loss function, which is called a bias-correction term, to a negative twofold maximum log-likelihood. It is well known fact in the field of statistics that the AIC is an asymptotic unbiased estimator of the risk function when the model being considered is completely specified, i.e., the following two assumptions are satisfied simultaneously:

A1. A structure of the model being considered (e.g., a mean structure or covariance structure) is specified.
A2. A distribution of the model being considered is specified.

Since the AIC is an asymptotic unbiased estimator of the risk function, the bias of the AIC to the risk function may become large when the sample size becomes small and the number of parameters used in the candidate model becomes large. This will cause a fault that the AIC often selects the model with the many parameters as the best model. The fault of the AIC is avoided by using the bias-corrected AIC, which is derived by correcting the bias to the risk function, instead of the original AIC. Hence, the bias correction has been studied under various models, conditions and correction methods.