Accepted Manuscript

Generalized ridge estimator and model selection criteria in multivariate linear regression

Yuichi Mori, Taiji Suzuki

PII: S0047-259X(17)30781-9
DOI: https://doi.org/10.1016/j.jmva.2017.12.006
Reference: YJMVA 4315

To appear in: *Journal of Multivariate Analysis*

Received date : 17 May 2016

Please cite this article as: Y. Mori, T. Suzuki, Generalized ridge estimator and model selection criteria in multivariate linear regression, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2017.12.006

Generalized ridge estimator and model selection criteria in multivariate linear regression

Yuichi Mori[a], Taiji Suzuki[b,c,d,*]

[a] Department of Mathematical and Computing Science, School of Computing, Tokyo Institute of Technology
[b] Department of Mathematical Informatics, Graduate School of Information Science and Technology, University of Tokyo
[c] PRESTO, Japan Science and Technology Agency, Japan
[d] Center for Advanced Integrated Intelligence Research, RIKEN, Tokyo, Japan

Abstract

We propose new model selection criteria based on generalized ridge estimators dominating the maximum likelihood estimator under the squared risk and the Kullback–Leibler risk in multivariate linear regression. Our model selection criteria have the following desirable properties: consistency, unbiasedness, and uniformly minimum variance. Consistency is proven under an asymptotic structure $p/n \to c$, where n is the sample size and p is the parameter dimension of the response variables. In particular, our proposed class of estimators dominates the maximum likelihood estimator under the squared risk, even when the model does not include the true model. Experimental results show that the risks of our model selection criteria are smaller than those based on the maximum likelihood estimator, and that our proposed criteria specify the true model under some conditions.

Keywords: Consistency, generalized ridge estimators, high-dimensional statistics, modified information criteria, modified model selection criteria, multivariate linear regression,
MSC 2000 subject classifications: Primary 62C10, secondary 62J05.

1. Introduction

Model selection criteria such as Akaike's information criterion (AIC) [1] and Mallows's C_p (Cp) [11] are frequently used in applications; their theoretical properties have also been studied extensively. In this paper, we consider model selection issues in the context of multivariate linear regression based on a type of generalized ridge estimator. Applications of multivariate linear regression include genetic data analysis and multiple brain scans; see, e.g., [2, 8].

We consider a multivariate linear regression model with p response variables, k explanatory variables, and multivariate Gaussian error terms. Specifically, we assume that $Y \sim \mathcal{N}_{n\times p}(AB, \Sigma \otimes \mathrm{I}_n)$, where Y is an $n \times p$ observation matrix of p response variables, A is an $n\times k$ observation matrix of k explanatory variables, B is a $k\times p$ unknown matrix of regression coefficients that does not include intercepts, Σ is a $p\times p$ unknown covariance matrix, k is a non-stochastic number, and n is the sample size. We assume that $k = \mathrm{rank}(A)$ is fixed, and that both $k \le n$ and $n - p - k - 1 > 0$.

Our interest focuses on the problem of selecting an appropriate set of explanatory variables for Y. To fix ideas, let $F = \{1, \ldots, k\}$ denote the index set of coefficients and $\mathcal{J}$ stand for its power set. For any $J \in \mathcal{J}$, let $k_J = |J|$ denote the number of elements in J. Then, the candidate model corresponding to the subset J can be expressed as $Y \sim \mathcal{N}_{n\times p}(A_J B_J, \Sigma \otimes \mathrm{I}_n)$, where A_J is an $n \times k_J$ matrix consisting of the columns of A indexed by the elements of J, and B_J is a $k_J \times p$ unknown matrix of regression coefficients. We assume that the candidate model corresponding to $J_* \in \mathcal{J}$ is the "true model", i.e., the model that generates the observations.

One way to perform model selection in multivariate linear regression is to apply well-known model selection criteria such as AIC [1], AICc [3], Cp [11], and MCp [6]. These criteria are unbiased or asymptotically unbiased

*Corresponding author
Email address: taiji@mist.i.u-tokyo.ac.jp (Taiji Suzuki)