Journal of Statistical Planning and Inference 142 (2012) 3134–3145

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

A model selection criterion for discriminant analysis of high-dimensional data with fewer observations

Masashi Hyodo [a], Takayuki Yamada [b,*], Muni S. Srivastava [c]

[a] JSPS Research Fellow, Graduate School of Economics, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[b] Risk Analysis Research Center, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[c] Department of Statistics, University of Toronto, 100 St. George Street, Toronto, Ontario, Canada M5S 3G3

ARTICLE INFO

Article history:
Received 9 February 2010
Received in revised form
12 March 2012
Accepted 3 June 2012
Available online 13 June 2012

Keywords:
Akaike information criterion
Discriminant analysis
Ridge-type estimator
High dimensional data

ABSTRACT

This paper is concerned with the problem of selecting variables in two-group discriminant analysis for high-dimensional data with fewer observations than the dimension. We consider a selection criterion based on approximately unbiased for AIC type of risk. When the dimension is large compared to the sample size, AIC type of risk cannot be defined. We propose AIC by replacing maximum likelihood estimator with ridge-type estimator. This idea follows Srivastava and Kubokawa (2008). It has been further extended by Yamamura et al. (2010). Simulation revealed that the proposed AIC performs well.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

One of the problems in the discriminant analysis is to find the variables which characterize the difference between groups. A well-known procedure is to use the canonical variate analysis. Our interest is to remove a set of redundant variables, and find the best subset of variables in the procedure of discriminant analysis. The approach for selection of variables dealt in this paper is to consider a family of variable selection models based on the theory of additional information due to Rao (1948, 1970). The selection criterion is based on the idea of Akaike (1973). Fujikoshi (2002) gave an approximately unbiased estimator for the AIC type of risk defined by expected log-predictive likelihood or equivalently the expected Kullbac–Leibler information for a candidate model for the case in which dimension p and total sample size N are both large with $p \leq N$. When $p > N$, the AIC type of risk cannot be defined, because the maximum likelihood estimator of $\boldsymbol{\Sigma}$ cannot be defined. In order to evaluate the risk, we use the ridge-type estimator. This idea follows Srivastava and Kubokawa (2008). It has been further extended by Yamamura et al. (2010). The AIC type criterion is defined by the sum of -2 log likelihood and the estimated bias. In Section 2, we evaluate the bias asymptotically as p and N go to infinity together with $p > N$, and obtain a criterion. In Section 3, we do a simulation study to demonstrate the performance of the proposed criterion. In Section 4, we give an example to apply the AIC. Concluding remarks are given in Section 5.

2. AIC type criterion for discriminant analysis

2.1. Redundancy model for discriminant analysis

We first describe the redundancy model for discriminant analysis written in Fujikoshi (2002). Let $\boldsymbol{Y} = (\boldsymbol{y}_1^{(1)}, \ldots, \boldsymbol{y}_{N_1}^{(1)}, \boldsymbol{y}_1^{(2)}, \ldots, \boldsymbol{y}_{N_2}^{(2)})$ be the $p \times N$ observation matrix, where $\boldsymbol{y}_j^{(i)}, j = 1, \ldots, N_i, i = 1, 2$, are independently and identically

[*] Corresponding author.
E-mail addresses: caicmhy@gmail.com (M. Hyodo), yma801228@gmail.com (T. Yamada), srivasta@utstat.toronto.edu (M.S. Srivastava).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.06.002

distributed (hereafter, i.i.d.) according to p-dimensional normal distribution with the mean vector $\boldsymbol{\mu}^{(i)}$ and the covariance matrix $\boldsymbol{\Sigma}$ (hereafter, $N_p(\boldsymbol{\mu}^{(i)},\boldsymbol{\Sigma})$), which is assumed to be positive definite, and $N=N_1+N_2$. Suppose that $\boldsymbol{Y}$ has the probability density function $g(\boldsymbol{Y})$ under the true model M^*, and

$$M^* : E[\boldsymbol{y}|\Pi_i]=\boldsymbol{\mu}_*^{(i)}, \quad \mathrm{Var}(\boldsymbol{y}|\Pi_i)=\boldsymbol{\Sigma}_*,$$

where E and Var denote the expectation and covariance matrix under the true model M^*.

Let K be any subset of $P=\{1,2,\ldots,p\}$ with $\sharp K=k$ members. For any subset K, we define a candidate model M_k. Suppose that we are interested in selecting the best subset from a family $\mathcal{K}$ of all subsets of P. In the following we fix a subset K, and assume that $K=\{1,2,\ldots,k\}$. This assumption comes only from notational simplicity, and results are applicable for any subset. Further, we shall identify K as k simply.

Now we consider a model M_k, which means that the first k variate $\boldsymbol{y}_1=(y_1,\ldots,y_k)'$, $k\le n=N-2$, is sufficient, and the remainder variate $\boldsymbol{y}_2=(y_{k+1},\ldots,y_p)'$ is redundant, i.e., the $p-k$ variate $\boldsymbol{y}_2$ has no additional information in canonical variate analysis, in the presence of $\boldsymbol{y}_1$. Here, we assume that the random vector $\boldsymbol{y}$ under Π_i is distributed as $N_p(\boldsymbol{\mu}^{(i)},\boldsymbol{\Sigma})$. In order to write the model M_k in a parametric expression, let us consider the partitions

$$\boldsymbol{\mu}^{(i)}=\begin{pmatrix}\boldsymbol{\mu}_1^{(i)}\\ \boldsymbol{\mu}_2^{(i)}\end{pmatrix}, \quad (i=1,2), \quad \boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix},$$

where $\boldsymbol{\mu}_1^{(i)}$ is $k\times 1$ vector and $\boldsymbol{\Sigma}_{11}$ is $k\times k$ matrix. Further, let

$$\boldsymbol{\mu}_{2\cdot 1}^{(i)}=\boldsymbol{\mu}_2^{(i)}-\boldsymbol{\Gamma}\boldsymbol{\mu}_1^{(i)}, \quad (i=1,2),\ \boldsymbol{\Gamma}=\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}.$$

Then the candidate model M_k can be written as

$$M_k : \boldsymbol{\mu}_{2\cdot 1}^{(1)}=\boldsymbol{\mu}_{2\cdot 1}^{(2)}=\boldsymbol{\mu}_{2\cdot 1}.$$

Note that $M_1\subset M_2\subset\cdots\subset M_{p-1}$.

In variable the selection problem for high-dimensional case, it is frequently reported that the size of the set of variables set is smaller than sample size N. For example, Fan and Fan (2008) also have selected 11 variables for classification problem in leukemia data (p=3571, N =72). In this paper, we restrict the candidate models to

$$M_1,M_2,\ldots,M_{N-2}.$$

2.2. Risk for candidate model with fewer observations than the dimension

Let $f(\boldsymbol{Y};\Theta_k)$ be the density function of $\boldsymbol{Y}$ under the model M_k, where

$$\Theta_k=\{\boldsymbol{\Sigma}_{11},\boldsymbol{\Sigma}_{22\cdot 1},\boldsymbol{\Gamma},\boldsymbol{\mu}_1^{(1)},\boldsymbol{\mu}_1^{(2)},\boldsymbol{\mu}_{2\cdot 1}\},$$

with $\boldsymbol{\Sigma}_{22\cdot 1}=\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. Ordinary, the risk for the model M_k is defined by expected log-predictive likelihood, which is as follows:

$$R_k=E_{\boldsymbol{Y}}E_{\boldsymbol{X}}[-2\log f(\boldsymbol{X};\hat{\Theta}_k)],$$

where $\boldsymbol{X}$ is a $p\times N$ random matrix that has the same distribution as $\boldsymbol{Y}$, and is independent of $\boldsymbol{Y}$; $\hat{\Theta}_k$ is the maximum likelihood estimator of Θ_k. When $p>n$, however, $\hat{\Theta}_k$ cannot be defined, because within-group matrix

$$\boldsymbol{W}=\sum_{i=1}^{2}\sum_{j=1}^{N_i}(\boldsymbol{y}_j^{(i)}-\overline{\boldsymbol{y}}^{(i)})(\boldsymbol{y}_j^{(i)}-\overline{\boldsymbol{y}}^{(i)})'$$

becomes singular. So we use the following ridge-type estimator

$$\boldsymbol{W}_\lambda=\boldsymbol{W}+\lambda\boldsymbol{I}_p$$

for regularization of with-in group matrix. Here, λ is called ridge parameter. From Srivastava and Kubokawa (2007) and Kubokawa and Srivastava (2008), the following estimation of the ridge parameter is chosen by the empirical Bayes method:

$$\hat{\lambda}=\frac{\mathrm{tr}(\boldsymbol{W})}{n\sqrt{p}}.$$

For notational simplicity, we write $\hat{\lambda}$ as λ when confusion is not caused. Using the above ridge-type estimator, we suggest the estimation of Θ as

$$\hat{\Theta}_{k,\lambda}=\left\{\hat{\boldsymbol{\Sigma}}_{11},\hat{\boldsymbol{\Sigma}}_{22\cdot 1,\lambda},\hat{\boldsymbol{\Gamma}}_\lambda,\hat{\boldsymbol{\mu}}_1^{(1)},\hat{\boldsymbol{\mu}}_1^{(2)},\hat{\boldsymbol{\mu}}_{2\cdot 1,\lambda}\right\},$$

Download English Version:

https://daneshyari.com/en/article/10524889

Download Persian Version:

https://daneshyari.com/article/10524889

Daneshyari.com