Journal of Multivariate Analysis 123 (2014) 364–379

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A variable selection criterion for linear discriminant rule and its optimality in high dimensional and large sample data

Masashi Hyodo [a,*], Tatsuya Kubokawa [b]

[a] Department of Mathematical Information Science, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan

[b] Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

ARTICLE INFO

Article history:
Received 11 December 2012
Available online 28 October 2013

AMS subject classifications:
62H30
62H12
62E20

Keywords:
Asymptotic optimality
High dimension and large sample
Linear discriminant analysis
Misclassification error
Multivariate normal
Second-order approximation
Variable selection

ABSTRACT

In this paper, we suggest the new variable selection procedure, called MEC, for linear discriminant rule in the high dimensional and large sample setup. MEC is derived as a second-order unbiased estimator of the misclassification error probability of the linear discriminant rule (LDR). It is shown that MEC not only asymptotically decomposes into 'fitting' and 'penalty' terms like AIC and Mallows C_p, but also possesses an asymptotic optimality in the sense that MEC achieves the smallest possible conditional probability of misclassification in candidate variable sets. Through simulation studies, it is shown that MEC has good performances in the sense of selecting the true variable sets.

1. Introduction

In this paper, we consider the problem of classifying a future observation vector into one of the two population groups Π_1 and Π_2. For each $i = 1, 2$, Π_i denotes a population from a multivariate normal distribution $\mathcal{N}_p(\boldsymbol{\mu}_i, \Sigma)$, and it is supposed that $\boldsymbol{x}_{ij}$, $j = 1, \ldots, N_i$, are observed from the population Π_i. Here, $\boldsymbol{\mu}_i$, $i = 1, 2$, and Σ are unknown parameters, and they are estimated by the sample mean $\bar{\boldsymbol{x}}_i = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}$ $i = 1, 2$, and the pooled sample covariance matrix $S = n^{-1} \sum_{i=1}^{2} \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)'$ for $n = N - 2 N = N_1 + N_2$. Shao et al. [14] show that the misclassification error rate of the linear discriminant rule (LDR) is asymptotically close to ones of the optimal rule when p diverges to infinity at a rate slower than $n^{1/2}$. From this perspective, LDR works well for the situation where the number of variables used for classification is much smaller than the training sample size. Thus, it is desired to find an optimal subset of variables in the sense that misclassification error rate can be small. Variable selection methods for discriminant analysis have been studied by Fujikoshi [1,3], Sakurai et al. [13], Wilbur et al. [16] and others. Related to this issue, multiple testing problems for no additional information have been discussed by Rao [10,11] and Kshirsagar [5]. In this paper, we suggest a new variable selection procedure based on misclassification error rate and establish the optimality in high dimensional and large sample setting.

* Corresponding author.
E-mail addresses: hyodoh_h@yahoo.co.jp (M. Hyodo), tatsuya@e.u-tokyo.ac.jp (T. Kubokawa).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.10.005

To explain the new variable selection procedure, consider the following linear discriminant rule. Let $\boldsymbol{x} = (x_1, \ldots, x_p)'$ be a future observation in the full model. Let $\boldsymbol{j} = (j_1, \ldots, j_{k(j)})'$ be a subset of the set $\{1, 2, \ldots, p\}$, and let $\boldsymbol{x}(\boldsymbol{j}) = (x_{j_1}, \ldots, x_{j_{k(j)}})'$ be the corresponding sub vector of $\boldsymbol{x}$. The model based on the variable $\boldsymbol{x}(\boldsymbol{j})$ is denoted by $\boldsymbol{j}$. Let $\mathcal{J}$ be a suitable family of subsets of $\{1, \ldots, p\}$. The LDR for classifying $\boldsymbol{x}$ based on the model $\boldsymbol{j}$ is that $\boldsymbol{x}$ is classified as coming from Π_1, if $W(\boldsymbol{j}) > \alpha$, and from Π_2, if $W(\boldsymbol{j}) < \alpha$, where α is a cut off point for classification rule, and

$$W(\boldsymbol{j}) = (\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j}))' S(\boldsymbol{j})^{-1} \left\{ \boldsymbol{x}(\boldsymbol{j}) - \frac{1}{2}(\bar{\boldsymbol{x}}_1(\boldsymbol{j}) + \bar{\boldsymbol{x}}_2(\boldsymbol{j})) \right\}.$$

Here, $\bar{\boldsymbol{x}}_i(\boldsymbol{j})$ $i = 1, 2$, and $S(\boldsymbol{j})$ are the sample mean and the pooled sample covariance matrix in the model $\boldsymbol{j}$. Then the problem of variable selection in LDR is regarded as how to select the best subset $\boldsymbol{j}$ from $\mathcal{J}$. To this end, we consider the conditional error probabilities of misallocation $L_1(\boldsymbol{j}) = P[W(\boldsymbol{j}) < \alpha | \boldsymbol{x}(\boldsymbol{j}) \in \Pi_1, \bar{\boldsymbol{x}}_1(\boldsymbol{j}), \bar{\boldsymbol{x}}_2(\boldsymbol{j}), S(\boldsymbol{j})]$ and $L_2(\boldsymbol{j}) = P[W(\boldsymbol{j}) \geq \alpha | \boldsymbol{x}(\boldsymbol{j}) \in \Pi_2, \bar{\boldsymbol{x}}_1(\boldsymbol{j}), \bar{\boldsymbol{x}}_2(\boldsymbol{j}), S(\boldsymbol{j})]$, which can be expressed as

$$L_g(\boldsymbol{j}) = \Phi\left((-1)^g \frac{(\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j}))' S(\boldsymbol{j})^{-1} \{\boldsymbol{\mu}_g(\boldsymbol{j}) - (\bar{\boldsymbol{x}}_1(\boldsymbol{j}) + \bar{\boldsymbol{x}}_2(\boldsymbol{j}))/2\} - \alpha}{\sqrt{(\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j}))' S(\boldsymbol{j})^{-1} \Sigma(\boldsymbol{j}) S(\boldsymbol{j})^{-1} (\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j}))}} \right) \tag{1.1}$$

for $g = 1, 2$, where $\Phi(\cdot)$ is the standard normal distribution function, and $\boldsymbol{\mu}_g(\boldsymbol{j})$ and $\Sigma(\boldsymbol{j})$ denote the population mean and covariance matrix in the model $\boldsymbol{j}$. When π_i, $i = 1, 2$, is a prior probability of the group membership, the expected error rate is given by

$$R(\boldsymbol{j}) = \pi_1 R_1(\boldsymbol{j}) + \pi_2 R_2(\boldsymbol{j})$$

where $R_g(\boldsymbol{j})$ is the unconditional error of misallocation given by $R_g(\boldsymbol{j}) = \mathrm{E}[L_g(\boldsymbol{j})]$ for $g = 1, 2$. The variable selection procedure proposed in this paper is an asymptotically unbiased estimator of the misclassification error $R(\boldsymbol{j})$ in the high dimensional and large sample setting.

A naive procedure for selection of variables is the method of minimizing

$$\Phi\left(-D(\boldsymbol{j})/2\right), \tag{1.2}$$

with respect to $\boldsymbol{j} \in \mathcal{J}$, where $D(\boldsymbol{j})$ is the sample Mahalanobis distance based on $\boldsymbol{x}(\boldsymbol{j})$, namely,

$$D(\boldsymbol{j})^2 = (\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j}))' S(\boldsymbol{j})^{-1} (\bar{\boldsymbol{x}}_1(\boldsymbol{j}) - \bar{\boldsymbol{x}}_2(\boldsymbol{j})). \tag{1.3}$$

However, $\Phi\left(-D(\boldsymbol{j})/2\right)$ has the bias $R(\boldsymbol{j}) - \mathrm{E}[\Phi(-D(\boldsymbol{j})/2)]$ which is not negligible. McLachlan [8,9] derived a second order asymptotic unbiased estimator of $R(\boldsymbol{j})$ under the large sample framework, namely,

(A0): $n \to \infty$, but p is bounded.

Fujikoshi [1] applied the estimator given by McLachlan [8,9] to the variable selection problem, and investigated the asymptotic properties and the relationship with AIC. On the other hand, Raudys [12] and Wyman et al. [17] derived the asymptotic approximations of the error probability under the high dimensional and large sample setting given by

(A1): $(n, p) \to \infty$ with $p/n \to c_0 \in [0, 1)$.

This setup not only includes the large sample setting (A0) as $c_0 = 0$, but also covers the case of large dimension p subject to $p < n$. This is practically important, because, as pointed out by Siotani [15], it is known that large sample approximations under (A0) are not good when p is large. In fact, as seen from Fujikoshi et al. [4] and Kubokawa et al. [6], the approximations under the setting (A1) give good approximations for large p less than n as well as small p.

In this paper, we derive a second-order unbiased estimator of $R(\boldsymbol{j})$ in the high dimensional and large sample setting (A1). The unbiased estimator is here called the *Misclassification Error Criterion* (MEC), which is useful for selecting variables in the linear discriminant rule. We show that MEC can be asymptotically decomposed into the 'fitting' and 'penalty' terms, namely,

$$\mathrm{MEC} = \Phi\left(-D(\boldsymbol{j})/2\right) + (\text{penalty}) + o_p(1),$$

where the penalty term increases in the dimension of model $\boldsymbol{j}$. This is a desirable property that variable selection procedures like AIC and C_p should possess. We also show that MEC has an asymptotical optimality as a variable selection procedure in (A1). Such optimality in the high dimensional and large sample setting is not known as long as we know.

Recently, Kubokawa et al. [6] derived a second-order approximation of the error probability of misclassification (EPMC) for the ridge-type linear discriminant rule in the high dimensional and large sample setting, and derived a second-order unbiased estimator of EPMC. Since the ridge-type linear discriminant rule is not invariant under scale transformations, their approach needs to calculate various kinds of fourth moments of the inverted Wishart matrix. It was hard to obtain such fourth moments, so that the approach used by Kubokawa et al. [6] cannot be used for developing an asymptotic optimality of MEC. Instead, the method used in this paper is to express $L_g(\boldsymbol{j})$ based on nine primitive random variables, namely four random variables having the standard normal distribution and five random variables having chi-square distributions. These stochastic expressions are essentially derived by Fujikoshi [3]. This approach not only makes it easier to derive the second-order approximation and the second-order unbiased estimator of $R(\boldsymbol{j})$, but also enables us to establish the asymptotic optimality of MEC as a variable selection procedure in both high dimensional and large sample settings.