Journal of Multivariate Analysis 140 (2015) 234–244

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Asymptotic properties of the misclassification rates for Euclidean Distance Discriminant rule in high-dimensional data

Hiroki Watanabe [a], Masashi Hyodo [b,*], Takashi Seo [a], Tatjana Pavlenko [c]

[a] *Department of Mathematical Information Science, Tokyo University of Science, Japan*

[b] *Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University, Japan*

[c] *Department of Mathematics, KTH Royal Institute of Technology, KTH Royal Institute of Technology, Sweden*

ARTICLE INFO

Article history:
Received 23 February 2014
Available online 27 May 2015

AMS subject classifications:
62H30
62E20
62H10
62H12

Keywords:
High-dimensional framework
Conditional error rate
Expected error rate

ABSTRACT

Performance accuracy of the Euclidean Distance Discriminant rule (EDDR) is studied in the high-dimensional asymptotic framework which allows the dimensionality to exceed sample size. Under mild assumptions on the traces of the covariance matrix, our new results provide the asymptotic distribution of the conditional misclassification rate and the explicit expression for the consistent and asymptotically unbiased estimator of the expected misclassification rate. To get these properties, new results on the asymptotic normality of the quadratic forms and traces of the higher power of Wishart matrix, are established. Using our asymptotic results, we further develop two generic methods of determining a cut-off point for EDDR to adjust the misclassification rates. Finally, we numerically justify the high accuracy of our asymptotic findings along with the cut-off determination methods in finite sample applications, inclusive of the large sample and high-dimensional scenarios.

1. Introduction

In this paper, we focus on the discrimination problem which is concerned with the allocation of a given object, $\boldsymbol{x}$, a random vector represented by a set of features $(x_1, \ldots, x_p)$, to one or two populations, Π_1 and Π_2 given by $\mathcal{N}_p(\boldsymbol{\mu}_1, \Sigma)$ and $\mathcal{N}_p(\boldsymbol{\mu}_2, \Sigma)$, respectively, where $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and common covariance matrix Σ is non-singular. Let $\{\boldsymbol{x}_{gj}\}_{j=1}^{N_g}$ be a random sample of independent observations drawn from gth population $\mathcal{N}_p(\boldsymbol{\mu}_g, \Sigma),\ g = 1, 2$. Let also $N = N_1 + N_2$ denote the total sample size and set $n = N - 2$. We are interested to explore the discrimination procedure that can accommodate $p > n$ cases, with the main focus on the performance accuracy in the asymptotic framework that allows p to grow together with n.

Clearly, the classical discriminant procedures, like Fisher linear discriminant rule, cannot be used when $p > n$ since the sample covariance matrix is singular and hence cannot be inverted. An intuitively appealing alternative considered in this study focuses on geometrical properties of the sample space and re-formulates the classification problem in terms of the *Euclidean distance discriminant rule* (EDDR): assign a new observation $\boldsymbol{x}$ to the "nearest" population Π_g, i.e. assign to Π_g if it is on average closer to the data from Π_g than to the data from the other population. Matusita's papers (see [3,4]) are perhaps the oldest references dealing with the discriminant rule based on distance measures, including the case when the multivariate distributions underlying the data are not specified.

[*] Corresponding author.
E-mail address: hyodoh_h@yahoo.co.jp (M. Hyodo).

http://dx.doi.org/10.1016/j.jmva.2015.05.008
0047-259X/© 2015 Elsevier Inc. All rights reserved.

Recently, Aoshima and Yata [2] have been considered the EDDR for the high-dimensional multi-class problem with different class covariance matrices. In particular, they derived asymptotic conditions which ensure that the expected misclassification rate converges to zero. Recent paper by Srivastava [8] used the Moore–Penrose inverse of the estimated covariance matrix and suggested a second-order approximation of the expected error rate in high-dimensional data.

We, in this study, focus on the asymptotic behavior of the misclassification rates of EDDR. Continuing with the normality assumption, with $\boldsymbol{\mu}_g$ acting as the center of the Π_g's distribution we define

$$T_0(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{\mu}_2\|^2 - \|\boldsymbol{x} - \boldsymbol{\mu}_1\|^2, \tag{1.1}$$

and its sample based version as

$$\widetilde{T}(\boldsymbol{x}) = \|\boldsymbol{x} - \bar{\boldsymbol{x}}_2\|^2 - \|\boldsymbol{x} - \bar{\boldsymbol{x}}_1\|^2 \tag{1.2}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\bar{\boldsymbol{x}}_g$'s denote the sample mean vectors, $g = 1, 2$. Hence, each term in (1.1) and (1.2) represents the distance between the observed vector $\boldsymbol{x}$ and the centroid of Π_g's or its sample based counterpart.

The natural advantage of using $\widetilde{T}(\boldsymbol{x})$ for classifying high-dimensional data is its ability to mitigate the effect of dimensionality on the performance accuracy. Indeed, as it is seen from (1.2), $\widetilde{T}(\boldsymbol{x})$ utilizes only the marginal distribution of the p variables, thereby naturally reducing the effect of large p in implementations. But the dimensionality has impact on the classification accuracy. To show this, we first point out that classifier $\widetilde{T}(\boldsymbol{x})$ has a bias. In fact,

$$\mathrm{E}[\widetilde{T}(\boldsymbol{x})|\boldsymbol{x} \in \Pi_g] = (-1)^{g-1}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 + \frac{N_1 - N_2}{N_1 N_2}\mathrm{tr}\,\Sigma, \quad g = 1, 2,$$

and thus the impact of dimensionality is implied by the quantity $(N_1 - N_2)\mathrm{tr}\,\Sigma/(N_1N_2)$. In this study, we introduce the bias-corrected version $\widetilde{T}(\boldsymbol{x})$ defined as

$$T(\boldsymbol{x}) = \|\boldsymbol{x} - \bar{\boldsymbol{x}}_2\|^2 - \|\boldsymbol{x} - \bar{\boldsymbol{x}}_1\|^2 - \frac{N_1 - N_2}{N_1 N_2}\mathrm{tr}\,S, \tag{1.3}$$

where the subtraction of $(N_1 - N_2)/(N_1N_2)\mathrm{tr}\,S$ in (1.3) is to guarantee that $\mathrm{E}[T(\boldsymbol{x})|\boldsymbol{x} \in \Pi_g] = (-1)^{g-1}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, $g = 1, 2$. Here, $S = (1/n)\sum_{g=1}^{2}\sum_{j=1}^{N_g}(\boldsymbol{x}_{gj} - \bar{\boldsymbol{x}}_g)(\boldsymbol{x}_{gj} - \bar{\boldsymbol{x}}_g)'$.

Now, the EDDR given by $T(\boldsymbol{x})$ places a new observation $\boldsymbol{x}$ to Π_1 if $T(\boldsymbol{x}) > \tilde{c}$, and to Π_2 otherwise, where $\tilde{c}$ is an appropriate cut-off point. Then, for a specific $\tilde{c}$, the performance accuracy of EDDR will be represented by the pair of misclassification rates that result. Precisely, we define the conditional misclassification rate of EDDR by

$$ce(2|1) = \Pr(T(\boldsymbol{x}) \le \tilde{c}|\boldsymbol{x} \in \Pi_1,\ \bar{\boldsymbol{x}}_1,\ \bar{\boldsymbol{x}}_2,\ S)$$

and its expected version by $e(2|1) = \mathrm{E}[ce(2|1)]$, where the expectation is taken with respect to $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2$ and S. Our main objective is to derive characteristic properties of both conditional and expected misclassification rate in high-dimensional data.

In many practical problems one type of misclassification rate is generally regarded as more serious than the other, examples include e.g. medical applications associated with the diagnosis of diseases. In such a case, it might be desired to determine the cut-off $\tilde{c}$ to obtain a specified probability of the error, or at least to approximate a specified probability. Then, one might base the choice of $\tilde{c}$ on the expected misclassification rate. This method, denoted in what follows by **M1**, suggests to set a cut-off point $\tilde{c}$ such that

$$\textbf{M1}: \ e(2|1) = \mathrm{E}[ce(2|1)] = \alpha,$$

where α is a value given by experimenters.

On the other hand, one may exploit the confidence of the conditional error rate when determining $\tilde{c}$; we denote this method by

$$\textbf{M2}: \ \Pr(ce(2|1) < eu) = 1 - \beta,$$

where $1 - \beta$ is the desired level of confidence and eu is an upper bound.

Both determination methods **M1** and **M2** have been established by using large sample approximation, see [1,5,6]. In this study, we extend the consideration to the high-dimensional case. Our main theoretical results provide the asymptotically unbiased and consistent estimator of $e(2|1)$ and the limit distribution of $ce(2|1)$ under general assumptions covering the case when $p > n$. In fact, **M1** and **M2** procedures can be considered as specific examples of using our generic results in the theory of EDDR in high-dimensions.

The remaining part of the paper is organized as follows. In Section 2, we derived the asymptotically unbiased and consistent estimator of $e(2|1)$. Further, the limiting approximations of the cut-off point defined by **M1** are established by using this estimator. In Section 3, two estimators of the confidence-based cut-off point defined by **M2** are proposed, for which the asymptotic normality of the conditional error rate is shown. Section 4 summaries the results of numerical experiments justifying the validity of the suggested cut-off estimators for various strength of dependence underlying the data along with a number of high-dimensional scenarios where p far exceeds the sample size. We conclude in Section 5.