ARTICLE IN PRESS

Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

# Accuracy of regularized D-rule for binary classification

Won Son [a], Johan Lim [a,*], Xinlei Wang [b]

[a] Department of Statistics, Seoul National University, Seoul, Republic of Korea
[b] Department of Statistical Science, Southern Methodist University, Dallas, TX, USA

ARTICLE INFO

*Article history:*
Received 21 December 2016
Accepted 7 November 2017
Available online xxxx

*AMS 2000 subject classifications:*
62H30
62H99

*Keywords:*
Classification
High dimensional data
Linear shrinkage covariance matrix estimator
Random matrix theory
Regularized D-rule

ABSTRACT

We consider a regularized D-classification rule for high dimensional binary classification, which adapts the linear shrinkage estimator of a covariance matrix as an alternative to the sample covariance matrix in the D-classification rule (D-rule in short). We find an asymptotic expression for misclassification rate of the regularized D-rule, when the sample size $n$ and the dimension $p$ both increase and their ratio $p/n$ approaches a positive constant $\gamma$. In addition, we compare its misclassification rate to the standard D-rule under various settings via simulation.



## 1. Introduction

Classification for high dimensional data is one of main research topics in last several decades, where the high dimensional data are now prevalent in every discipline. The binary classification is a multivariate procedure, in which we first build a classifier from training samples with known memberships, and then assign a new observation $\mathbf{x}$ into one of two disjoint populations $\pi_i$, $i = 1, 2$ using its features. A large number of methods for binary classification have been proposed in the literature, many of which based on a Bayes classifier (under 0-1 loss). The Bayes classifier is the optimal classifier in terms of minimizing the misclassification rate and is often set as the benchmark when solving classification problems. However, it depends on unknown model parameters including those in a covariance matrix. The estimation of the covariance matrix from high dimensional data and its application to the classification is the main theme of this paper. Throughout the paper, we assume that sample size $n$ (more precisely, both $n_1$ and $n_2$ increase with the same rate and $n = n_1 + n_2$) and the dimension $p$ increase with the rate that $p/n \to \gamma$, where $\gamma \in (0, 1)$; and we do not make any structural assumptions including the sparsity on the model or classifier.

Suppose $\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}$ from class 1 are distributed as $(\boldsymbol{\mu}_1, \Sigma)$, and $\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2}$ from class 2 are distributed as $(\boldsymbol{\mu}_2, \Sigma)$. The D-classification rule (or simply D-rule) is the classifier that assigns a new observation $\mathbf{x}$ into class $i$, if

$$\det(\mathbf{A}_i) = \min_j \det(\mathbf{A}_j),$$

* Corresponding author.
*E-mail address:* johanlim@snu.ac.kr (J. Lim).

https://doi.org/10.1016/j.jkss.2017.11.002
1226-3192/

Please cite this article in press as: Son, W., et al., Accuracy of regularized D-rule for binary classification. Journal of the Korean Statistical Society (2017), https://doi.org/10.1016/j.jkss.2017.11.002.

where $\mathbf{A}_j = \mathbf{A} + \alpha_j(\mathbf{x} - \bar{\mathbf{x}}_j)(\mathbf{x} - \bar{\mathbf{x}}_j)^\intercal$, $\mathbf{A} = \sum_{l=1}^{2}\sum_{k=1}^{n_l}(\mathbf{x}_{lk} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lk} - \bar{\mathbf{x}}_l)^\intercal$ and $\alpha_j = n_j/(n_j + 1)$ with $j = 1, 2$. It also can be expressed as

$$\alpha_i(\mathbf{x} - \bar{\mathbf{x}}_i)^\intercal\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i) = \min_j \alpha_j(\mathbf{x} - \bar{\mathbf{x}}_j)^\intercal\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_j), \tag{1}$$

where $\mathbf{S} = \mathbf{A}/n$ with $n = n_1 + n_2$. The D-rule is not a new classifier, but appears in the literature with several other names. We can show that it is asymptotically equivalent to the well-known Fisher's linear discriminant with simple algebra. Also, when the training data are from normal distributions, it is the Bayes classifier when assuming an equal cost for false positives and negatives and a uniform prior for two populations, where the mean and covariance matrices are estimated with their sample counterparts.

The D-rule above contains the sample covariance matrix, which is known to be inconsistent for high dimensional data. Many alternative estimators are proposed for either structural (*e.g.*, sparsity or bandedness) or non-structural covariance matrices (Bickel & Levina, 2008a, 2008b; El Karoui, 2008; Won, Lim, Kim, & Rajaratnam, 2013). Here, we make no structural assumption on the model or covariance matrix. The linear shrinkage covariance matrix estimator by Ledoit and Wolf (2004) is the most popular non-structural estimator,

$$\mathbf{S}_\delta = \mathbf{S} + \delta\mathbf{I}_p, \tag{2}$$

where $\mathbf{I}_p$ is a $p \times p$ identity matrix and $\delta$ is a positive constant. It linearly shrinks the sample eigenvalues to those of the target matrix and reduces the expected estimation loss of the sample covariance matrix (Ledoit & Wolf, 2004). It is also successfully applied to many high-dimensional procedures to resolve the difficulties from the high dimensionality. For example, Lanckriet, El Ghaoui, Bhattacharyya, and Jordan (2002) intuitively propose to use $\mathbf{S}_\delta$ in building a robust classifier when $\mathbf{S}$ is not well defined. Schafer and Strimmer (2005) reconstruct a gene regulatory network from microarray gene expression data using the inverse of a regularized covariance matrix. Guo, Hastie, and Tibshirani (2007) suggest the regularized discriminant analysis and apply it to classifying the data from a microarray experiment. Pyun, Lim, and Gray (2009) applies the linear shrinkage estimator $\mathbf{S}_\delta$ to finding a vector quantizer that is robust to various noisy sources. Chen, Paul, Prentice, and Wang (2011) and Lee, Lim, Son, Jung, and Park (2015) propose a modified Hotelling's $T^2$-statistic for testing high dimensional mean vectors and finding differentially expressed gene sets. Recently, Choi, Ng, and Lim (2017) propose to modify the likelihood ratio test for testing the covariance structure and show that the modified LRT significantly improves the power when both $n$ and $p$ are large.

Not surprisingly, the regularized D-rule with the linear shrinkage covariance matrix estimator is also studied by several authors including Friedman (1989) and Kubokawa, Hyodo, and Srivastava (2013). In particular, Kubokawa et al. (2013) studies the expected misclassification rate of the regularized D-rule, and compare it to the standard D-rule as we do in this paper. However, there are two major differences between our work here and the results by Kubokawa et al. (2013). First, in Kubokawa et al. (2013), the regularization parameter $\delta$ is an order of $O(n^{-1})$, whereas it is an order of $\delta = O(1)$ in this paper. Thus, the regularization by Kubokawa et al. (2013) becomes infinitesimal as $n$ increases. Second, due to the magnitude of the regularization for large $n$, the approximate expected misclassification rate by Kubokawa et al. (2013) depends on $\tau$, the limit of $n_1/(n_1 + n_2)$, and there are cases when the regularized D-rule performs worse than the original D-rule. Unlike Kubokawa et al. (2013), by using $\delta = O(1)$, the asymptotic results in Section 3 do not depend on $\tau$ and the regularized D-rule always has smaller expected misclassification rate than the original D-rule in asymptotic regardless of $\tau$. We add the detailed comparisons in Section 5.

The remainder of the paper is organized as follows. The regularized D-rule of this paper is formally introduced, and its misclassification rate is expressed in Section 2. In Section 3, we study the asymptotic misclassification rate for the case $\Sigma = \mathbf{I}_p$ and compare it to the original D-rule. In Section 4, we numerically compare the misclassification rates of the regularized D-rule and original D-rule for various choices of general $\Sigma$. In Section 5, we compare our results in this paper to those by Kubokawa et al. (2013). An example of improvements of accuracy by the regularized D-rule is illustrated with real data in Section 6. We conclude the paper in Section 7 with a brief summary and discussion on the choice of $\delta$.

## 2. Regularized D-rule: General

Suppose $\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}$ are drawn from $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2}$ are drawn from $N_p(\boldsymbol{\mu}_2, \Sigma)$ with $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\Sigma$ unspecified.

The regularized D-rule classifies an observation $\mathbf{x}$ to class 1 if

$$\alpha_1(\mathbf{x} - \bar{\mathbf{x}}_1)^\intercal\mathbf{S}_\delta^{-1}(\mathbf{x} - \bar{\mathbf{x}}_1) < \alpha_2(\mathbf{x} - \bar{\mathbf{x}}_2)^\intercal\mathbf{S}_\delta^{-1}(\mathbf{x} - \bar{\mathbf{x}}_2), \tag{3}$$

where

$$\mathbf{S}_\delta = \mathbf{S} + \delta\mathbf{I}_p = \frac{1}{n}\sum_{i=1}^{2}\sum_{k=1}^{n_i}(\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)^\intercal + \delta\mathbf{I}_p. \tag{4}$$

Please cite this article in press as: Son, W., et al., Accuracy of regularized D-rule for binary classification. Journal of the Korean Statistical Society (2017), https://doi.org/10.1016/j.jkss.2017.11.002.