Computational Statistics and Data Analysis 82 (2015) 110–125

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Stein's method in high dimensional classification and applications

Junyong Park, DoHwan Park *

Department of Mathematics and Statistics, University of Maryland Baltimore County, 21250 MD, USA

ARTICLE INFO

Article history:
Received 8 October 2013
Received in revised form 17 April 2014
Accepted 15 August 2014
Available online 26 August 2014

Keywords:
Classification
Sparsity
High dimension
Stein's estimator
Shrinkage

ABSTRACT

In the context of classification, it is a common phenomenon that high-dimensional data such as micro-array data consist of only a few informative components. If one uses standard statistical modeling and estimation procedures with entire information, it tends to overfit the data due to noise information. Therefore, some regularization conditions are required to select important information. A class of regularization methods is proposed through various shrinkage estimators using Stein's identity. Since hard thresholding does not satisfy the condition of Stein's identity, the proposed methods consider linear classifiers with soft, firm and SCAD thresholdings incorporating Stein's identity and show some asymptotic properties. Simulation studies and applications to three different micro array data sets show that the proposed methods work well. Also the proposed methods are compared with some existing methods.

1. Introduction

The problem of deciding the membership of an observation occurs frequently in many scientific areas such as microarray cancer data, medical studies and pattern recognition in engineering. For example, one may be interested in assigning a class to a new gene expression based on a database of n_1 gene expressions from cancer tissues of different patients and n_2 gene expressions from normal tissues.

The main idea of classification problems is to allocate a new gene expression into one of two groups. More specifically, we consider the problem of finding a classifier for the response $Y \in \{-1, 1\}$ based on a vector $(X_1, X_2, \ldots, X_p) \in R^p$ of explanatory variables. Suppose there is a training set (or a sample) of $n_1 + n_2$ examples where n_1 examples are $Y_i = -1$, $(Y_i, X_{i1}, X_{i2}, \ldots, X_{ip})$, $i = 1, \ldots, n_1$, and n_2 examples are $Y_i = 1$, $(Y_i, X_{i1}, X_{i2}, \ldots, X_{ip})$, $i = n_1 + 1, \ldots, n_1 + n_2$. In particular, we assume a multivariate normal distribution of the vector $X = (X_1, \ldots, X_p)$ conditional on the value of Y, i.e., we have $X|Y = -1 \sim N(\boldsymbol{\mu}_1, \Sigma)$ and $X|Y = 1 \sim N(\boldsymbol{\mu}_2, \Sigma)$ where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip})$. One aim is to assign a new observation, namely X_{new} to one of $Y = 1$ and $Y = -1$. A linear rule has been commonly used for the classification. It is given by

$$Y = \text{sign}\left(\sum_{j=1}^{p} a_j X_j + a_0\right), \tag{1}$$

where $a_0, a_1, \ldots, a_p$ are constants. Some classical classification rules such as Fisher's rule are fashioned for large sample size problems, but those rules may not work well for extremely high dimensional situations, when $p \gg n$. See, for example, Dudoit et al. (2002) and Bickel and Levina (2004).

* Corresponding author. Tel.: +1 410 455 2408.
E-mail address: dhpark@umbc.edu (D. Park).

http://dx.doi.org/10.1016/j.csda.2014.08.009
0167-9473/

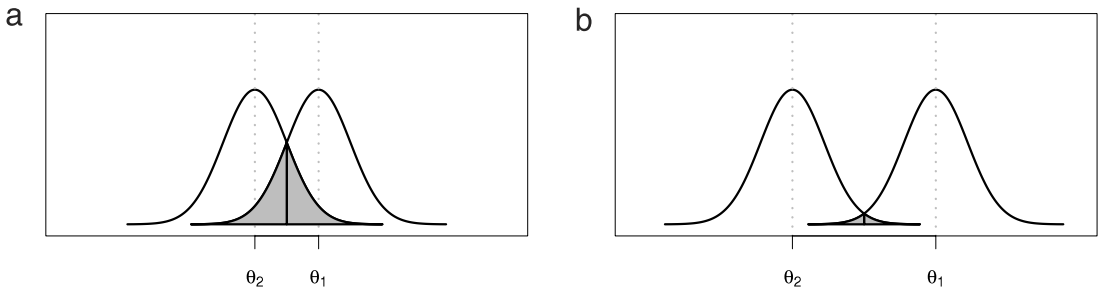

Fig. 1. Distribution of $\sum a_j X_j$ for different $\theta_1 - \theta_2$. Shaded areas represent misclassification probabilities.

In such a high dimensional data, one trick is ignoring off-diagonal elements in the covariance matrix and use only the diagonal elements, which is known as naive Bayes rule or independent rule (IR) in the linear classification. Bickel and Levina (2004) investigated the performance of IR for multivariate normal variables whereas Park and Ghosh (2007) and Park (2009) studied the performance of IR for classification of high dimensional multivariate binary variables. One may avoid the noise from estimating many unnecessary correlation coefficients by assuming independence of variables. Although the independence rule has been successfully used in various practical problems, it can easily fail when only a small number of variables are meaningful and the majority of variables are noise, namely sparsity situation. Therefore, to improve the performance of the IR under sparsity, it is necessary to remove many of redundant variables. Recently, Fan and Fan (2008) proposed Feature Annealed Independence Rule (henceforth FAIR) which performs variables selection in linear classifier based on hard thresholding. They sorted t-statistics from two classes by descending order and applied hard thresholding to select variables. However, except the hard thresholding, not much attention has been paid to various shrinkages such as soft, firm and SCAD in high dimensional classification problems whereas different types of shrinkages have widely used in regression and wavelet shrinkages.

We propose purely data dependent procedure to determine shrinkage estimators incorporating Stein's identity (Stein, 1981) in linear classifiers in high dimension. Shrinkage estimators with Stein's identity have been widely used in various model selection problems, see, for example, Li (1985) and Donoho and Johnstone (1995), however, to the best of our knowledge, the idea of shrinkage estimators incorporating Stein's identity has not been applied to classification problem. Since hard thresholding does not satisfy the conditions to use Stein's identity, we consider three types of thresholdings such as soft, firm and SCAD while FAIR in Fan and Fan (2008) used the hard thresholding to select variables in linear classifier. We present some asymptotic results of proposed classifiers and compare them with some currently existing procedures such as FAIR, IR and support vector machine (SVM).

The rest of the paper is organized as follows. Section 2 briefly presents the idea of shrinkage estimators in classification. We introduce various shrinkage estimators and propose a new classifier based on Stein's methods using shrinkage estimators in Section 3. In Section 4, we establish the asymptotic theories for suggested approaches for independent variables and correlated variables. In Section 5 we present a comprehensive simulation comparison of SCAD, firm and soft thresholdings with existing methods such as SVM (Support Vector Machine), FAIR and the IR. In Section 6, we demonstrate the practical performance of our methods in three microarray data sets and compare the results with those from existing methods. Concluding remark is provided in Section 7.

2. Regularization through shrinkage

As mentioned in the introduction, we consider a linear classification rule in (1) which is one of the most popular classifiers in high dimensional classification. For simplicity, let us first assume $X|Y = 1 \sim N(\boldsymbol{\mu}_1, I)$ and $X|Y = -1 \sim N(\boldsymbol{\mu}_2, I)$ where $X = (X_1, X_2, \ldots, X_p)'$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})'$. In addition, let I be $p \times p$ identity matrix, i.e., all the variables are independent with unit variance. First step is searching coefficients $a_1, \ldots, a_p$ in linear rule with constraint $\|\mathbf{a}\|^2 = 1$ where $\|\cdot\|$ is l^2-norm, i.e., $\sum_{j=1}^p a_j^2 = 1$ and then find a_0 based on estimators of a_js, $1 \le j \le p$.

One naive method to choose a_j's, $j = 1, \ldots, p$, is the case when a_j's separate two class conditional distributions of $\sum_{j=1}^p a_j X_j$ given $Y = 1$ or $Y = -1$ as much as possible. Equivalently, coefficients a_js must satisfy maximizing the difference of two corresponding mean values, denoted by $\theta_i = \mathbf{a}'\boldsymbol{\mu}_i$ for $i = 1, 2$, i.e., we maximize

$$V \equiv \theta_1 - \theta_2 = \sum_{j=1}^p a_j \mu_{1j} - \sum_{j=1}^p a_j \mu_{2j} \equiv \sum_{j=1}^p a_j \Delta_j \tag{2}$$

subject to $\|\mathbf{a}\| = 1$ under assuming $\theta_1 > \theta_2$. Fig. 1 demonstrates how two groups are separated depending on $\theta_1 - \theta_2$. As $\theta_1 - \theta_2$ increases, class conditional distributions of $\sum a_j X_j$ are separated better and it leads to smaller misclassification probability. Therefore, the optimal choice of $a_1, \ldots, a_p$ is one which maximizes V. Note that under 0–1 loss, given $(a_1, a_2, \ldots, a_p)$, the natural choice of a_0 is $a_0 = -\sum_{j=1}^p a_j(\mu_{1j} + \mu_{2j})/2$.