Statistics and Probability Letters 117 (2016) 100–112

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Discriminant analysis on high dimensional Gaussian copula model

Yong He [a,*], Xinsheng Zhang [b], Pingping Wang [c]

[a] School of Management, Department of Statistics, Fudan University, Shanghai, 200433, China
[b] School of Management, Fudan University, Shanghai, China
[c] School of Statistics, East China Normal University, Shanghai, China

ARTICLE INFO

Article history:
Received 18 January 2016
Received in revised form 26 April 2016
Accepted 23 May 2016
Available online 31 May 2016

Keywords:
Classification
Gaussian copula model
Rotate-and-Solve

ABSTRACT

In this paper we propose a classifier for the high dimensional Gaussian copula model. Besides, a Rotate-and-Solve procedure is proposed to tackle with the non-sparse case. Both theoretical analysis and simulation study show that the classifier performs better than some state-of-the-art methods.

1. Introduction

Linear discriminant analysis (LDA) is of great interest to statisticians for classification. Since the work of Bickel and Levina (2004) which shows that the classical low dimensional normal-based LDA performs as poorly as random guess when the dimension p increases fast compared to the sample size n ($p/n \to \infty$), many follow-up works have been proposed based on sparsity conditions, including the nearest shrunken centroids (Wang and Zhu, 2007). Fan and Fan (2008) proposed the features annealed independence rule (FAIR) which is based on a working independence assumption. Thereafter, large number of alternative methods have been proposed by finding the sparse discriminant affine space directly, such as Trendafilov and Jolliffe (2007), Wu et al. (2009), Cai and Liu (2011), Fan et al. (2012) and Mai et al. (2012).

Consider a binary classification problem: suppose we have two p-dimensional normal distributions with the same covariance matrix, $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ for class 0 and $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ for class 1. Given a random vector $\boldsymbol{X}$ which comes from either class 0 or class 1 with equal prior probabilities $\pi_0 = \pi_1$. The well known Bayes rule classifier is

$$\psi(\boldsymbol{X}) := I((\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Omega}\boldsymbol{\delta} > 0), \tag{1.1}$$

where $\boldsymbol{\mu} := (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)/2$, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ and $I(\cdot)$ is the indicator function. Therefore, the main idea of a series of LDA-based methods for the high dimensional classification problem is to find efficient methods which can work well under sparsity condition so that $\boldsymbol{\mu}$ and $\boldsymbol{\beta} = \boldsymbol{\Omega}\boldsymbol{\delta}$ in Eq. (1.1) can be well estimated. Fan et al. (2012) proposed the Regularized Optimal Affine Discriminant (ROAD) approach to estimate $\boldsymbol{\beta}$, which minimizes $\boldsymbol{\omega}^\top \widehat{\boldsymbol{\Sigma}}_n \boldsymbol{\omega}$ with $\boldsymbol{\omega}^\top \widehat{\boldsymbol{\delta}}$ restricted to be a constant

* Corresponding author.
E-mail address: hey12@fudan.edu.cn (Y. He).

http://dx.doi.org/10.1016/j.spl.2016.05.018
0167-7152/

value, that is,

$$\min_{\boldsymbol{\omega}^\top \widehat{\boldsymbol{\delta}}=1} \{\|\boldsymbol{\omega}^\top \widehat{\boldsymbol{\Sigma}}_n \boldsymbol{\omega}\|_1, \text{ subject to } \|\boldsymbol{\omega}\|_1 \le c\}, \tag{1.2}$$

where $\widehat{\boldsymbol{\Sigma}}_n$ is the pooled sample covariance matrix and $\widehat{\boldsymbol{\delta}} := (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)$ with $\widehat{\boldsymbol{\mu}}_0$ and $\widehat{\boldsymbol{\mu}}_1$ being the sample mean vectors from class 0 and class 1 respectively. Cai and Liu (2011) proposed a direct estimation approach to estimate $\boldsymbol{\beta}$, which endeavors to exploit the sparsity of $\boldsymbol{\beta}$ in the ℓ_∞ norm, that is,

$$\min_{\boldsymbol{\beta}} \{\|\boldsymbol{\beta}\|_1, \text{ subject to } \|\widehat{\boldsymbol{\Sigma}}_n \boldsymbol{\beta} - \widehat{\boldsymbol{\mu}}\|_\infty \le \lambda_n\}, \tag{1.3}$$

where $\widehat{\boldsymbol{\mu}} := (\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_0)/2$. Optimization problem (1.3) can be solved by linear programming and Cai and Liu (2011) named the resulting classification rule as linear programming discriminant (LPD) rule. The method turns out to be highly related to the Dantzig selector (refer to Candes and Tao, 2007; Yuan, 2010). Mai et al. (2012) proposed an approach of the sparse linear discriminant analysis based on an equivalent least square formulation of the LDA. In brief, an ℓ_1 penalty is added in the method to exploit a sparsity pattern of $\boldsymbol{\beta}$, and hence nice theoretical properties have been obtained under certain regularity conditions. However, all these methods referred above require the normality condition which can be very restrictive in realistic application.

To overcome the restrictiveness of the Gaussian assumption, Han et al. (2013) proposed a classifier named the Copula Discriminant Analysis (CODA), which extends the underlying conditional distributions from Gaussian to the Nonparanormal family (refer to Liu et al., 2009). Han et al. (2013) showed that the optimization problem (1.2) is connected to the Lasso in the sense that they are equivalent if fixing the second tuning parameter. Han et al. (2013) also proved that the classifier CODA is variable selection consistent.

The classification methods mentioned above are all based on the sparsity conditions and are efficient when these sparsity assumptions hold. However, they may not work well when the sparsity conditions are violated. Although the sparsity assumptions make sense in some applications, they can be very restrictive in many scenarios. Hao et al. (2015) proposed a family of rotations to create the required sparsity. The basic idea is to use the principal components of the sample covariance matrix of the pooled samples and its variants to rotate the data first and then to apply an existing high dimensional classifier. The Rotate-and-Solve procedure can be combined with any existing classifier and is robust against the sparsity level of the true model.

In this paper, we also consider the classification problem under the Gaussian copula model. The classifier proposed by Han et al. (2013) is based on optimization problem (1.2) while our classifier is mainly based on optimization problem (1.3). The main contribution of this paper is that we propose to estimate $\boldsymbol{\Omega\delta}$ directly by virtue of the rank-based covariance matrix estimator, which does not require the precision matrix $\boldsymbol{\Omega}$ to be sparse and enjoys significant computational advantages over existing methods that require separate estimation of $\boldsymbol{\Omega}$ and $\boldsymbol{\delta}$. The advantage of the method also lies in that it is robust to noisy data (e.g., contaminated by outliers) and can be implemented efficiently by linear programming. Theoretical analysis shows that the classifier is misclassification consistent. Besides, we propose a Rotate-and-Solve procedure based on our rank-based covariance matrix estimator to handle the scenarios where the sparsity conditions are violated. The advantage of the Rotate-and-Solve procedure also lies in that it achieves dimension reduction. Therefore, the reduction of computation cost is significant when $p \gg n$, since the singular value decomposition of the rank-based covariance matrix estimator is much faster. We also conduct thorough simulation study to show that the classification error is improved by our procedure.

The rest of this paper is organized as follows. In the next section we give some basic notations and briefly review the Gaussian copula model. In Section 3, we propose our LPD based classifier for Gaussian copula model. We also present the Rotate-and-Solve procedure when the original sparsity conditions do not hold in this section. In Section 4 we present a theoretical analysis of the classifier and prove that the proposed classifier is misclassification consistent, with more detailed proofs collected in the Appendix. In Section 5, we present numerical results on both synthetic and real data. Discussions and future work are presented in the last section.

2. Background

2.1. Notation

For any vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d$, $\|\boldsymbol{\mu}\|_0 = \sum_{i=1}^d I(\mu_i \neq 0)$, $\|\boldsymbol{\mu}\|_1 = \sum_{i=1}^d |\mu_i|$, $\|\boldsymbol{\mu}\|_2 = \sqrt{\sum_{i=1}^d \mu_i^2}$ and $\|\boldsymbol{\mu}\|_\infty = \max_i |\mu_i|$, let $\boldsymbol{\mu}_{\setminus i}$ denote the $(d-1) \times 1$ vector by removing the ith entry from $\boldsymbol{\mu}$. Let $A = [a_{ij}] \in \mathbb{R}^{d\times d}$. $\|A\|_{\ell_1} = \max_{1\le j\le d} \sum_{i=1}^d |a_{ij}|$, $\|A\|_\infty = \max_{i,j} |a_{ij}|$ and $\|A\|_1 = \sum_{i=1}^d \sum_{j=1}^d |a_{ij}|$. We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the smallest and largest eigenvalues of A respectively. we say A is k-sparse if each row/column has at most k nonzero entries. For a set $\mathcal{H}$, denote by $|\mathcal{H}|$ the cardinality of $\mathcal{H}$. For a matrix $B = [b_{ij}] \in \mathbb{R}^{q\times d}$, $B_{i,\setminus j}$ denote the ith row of B with its jth entry removed and $B_{\setminus i,j}$ denote the jth column of B with its ith entry removed. $B_{\setminus i,\setminus j}$ denotes a $(q-1)\times(d-1)$ matrix obtained by removing the ith row and jth column of B. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \le C|b_n|$ holds for all n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, and write $a_n \asymp b_n$ if there exist constants c and C such that $c \le a_n/b_n \le C$ for all n.