ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

High dimensional discrimination analysis via a semiparametric model

Binyan Jiang [a,*], Chenlei Leng [b]

[a] Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong
[b] Department of Statistics, University of Warwick, United Kingdom

ARTICLE INFO

Article history:
Received 19 November 2013
Received in revised form 7 November 2015
Accepted 7 November 2015
Available online xxxx

Keywords:
Bayes rule
Linear discrimination analysis
Monotone transformation
Semiparametric discriminant analysis
Sparsity

ABSTRACT

We propose a semiparametric linear programming discriminant (SLPD) rule for high dimensional discriminant analysis under a semiparametric model. As an extension, we further propose a two-stage SLPD (TSLPD) rule, which can have better classification performance under mild sparsity assumptions.

1. Introduction

High dimension low sample size data sets are frequently encountered nowadays in different fields. However it is known that the statistical analysis of these data sets is very challenging and possibly intractable in some instances. For example, in high dimensional classification, the classical linear discriminant analysis is asymptotically equivalent to random guess even when the Gaussian assumptions are satisfied (Bickel and Levina, 2004). Fortunately, in many situations the data can be assumed to be sparse in that many parameters are close or equal to zero. Motivated by this observation, many approaches are proposed to exploit this sparsity assumption.

Let $X = (x_1, \ldots, x_p)^T$ and $Y = (y_1, \ldots, y_p)^T$ be random variables from two different classes. We shall call these two classes class X and class Y throughout this paper. Assume the Gaussian model where $X \sim N(\mu_x, \Sigma)$ and $Y \sim N(\mu_y, \Sigma)$. Given a random observation Z from class X or class Y, the well known Bayes rule classifies Z into class X if $[Z - (\mu_x + \mu_y)/2]$ $\Sigma^{-1}(\mu_y - \mu_x) \leq 0$ and into class Y otherwise.

Practically, μ_x, μ_y and Σ are unknown and it is a standard technique to separately estimate μ_x, μ_y and Σ or Σ^{-1} from the sample and plug them into the above Bayes rule. Assuming that both Σ and $\mu = \mu_y - \mu_x$ are sparse, Shao et al. (2011) used thresholding procedures for estimating Σ and μ. By noticing that the Bayes rule depends on Σ and μ only through $\beta = \Sigma^{-1}\mu$, instead of estimating Σ^{-1} and μ separately, Cai and Liu (2011) obtained sparse estimators for β directly. Other approaches for sparse linear discriminant analysis under multivariate normal assumptions can be found in Fan et al. (2012) and Mai et al. (2012) and the references therein.

A limitation of the linear discriminant rules is the normality assumption. When p is fixed, Lin and Jeon (2003) considered the so-called transnormal or nonparanormal distribution to allow the marginal distributions unspecified, as discussed in the

* Corresponding author.
E-mail address: by.jiang@polyu.edu.hk (B. Jiang).

http://dx.doi.org/10.1016/j.spl.2015.11.012
0167-7152/

Please cite this article in press as: Jiang, B., Leng, C., High dimensional discrimination analysis via a semiparametric model. Statistics and Probability Letters (2015), http://dx.doi.org/10.1016/j.spl.2015.11.012

next subsection; see also Kon and Nikolaev (2011). In this paper, we consider discriminant analysis under this generalized distribution when the dimension p far exceeds the sample size n but grows slower than $\exp(n^{1/2})$. We derive the Bayes rule under this semiparametric model and propose estimators for its components. We show that the risk of our classification rule tends to the Bayes risk in probability.

1.1. A semiparametric model

We begin by introducing some notations. For any matrix M, write M^T as the transpose of M. Let $v = (v_1, \ldots, v_p)^T \in \mathcal{R}^p$ be a p-dimensional vector. Define $|v|_0 = \sum_{i=1}^p I_{\{v_i \neq 0\}}$ and $|v|_\infty = \max_{1\leq i\leq p} |v_i|$. For any $1 \leq q < \infty$, the l_q norm of v is defined as $|v|_q = (\sum_{i=1}^q |v_i|^q)^{1/q}$. We denote the p-dimensional vector of ones as 1_p and the p-dimensional vector of zeros as 0_p.

Following Lin and Jeon (2003), we say a random vector $V = (V_1, \ldots, V_p)^T$ has a transnormal distribution $TN(h, \mu, 1_p, \Gamma)$ if there exists a set of univariate strictly monotone and differentiable functions $h = (h_1, \ldots, h_p)^T$ such that $h(V) = (h_1(V_1), \ldots, h_p(V_p))^T$ is multivariate normal with mean $\mu = (\mu_1, \ldots, \mu_p)^T$ and correlation matrix $\Gamma = (\gamma_{ij})_{p\times p}$.

The transnormal distribution is also called the nonparanormal distribution in some recent literature and is also related to the Gaussian copula model; see for example Liu et al. (2009). Denote the density functions of X and Y as f_X and g_Y respectively. In this paper we assume that $X \sim TN(h, \mu_x, 1_p, \Gamma)$ and $Y \sim TN(h, \mu_y, 1_p, \Gamma)$. Without loss of generality we assume that $\mu_x = (0, \ldots, 0)^T$, $\mu_y = \mu = (\mu_1, \ldots, \mu_p)^T$. Therefore $h_i(x_i) \sim N(0, 1)$, $h_i(y_i) \sim N(\mu_i, 1)$, and we immediately have

$$h_i = \Phi^{-1} \circ F_i = (\Phi^{-1} \circ G_i) + \mu_i, \quad 1 \leq i \leq p, \tag{1}$$

where $\circ$ denotes the composition of functions, Φ is the univariate standard Gaussian cumulative distribution function, F_i is the cumulative distribution function of x_i and G_i is the cumulative distribution function of y_i. This is a sub model of the functional analysis of variance model; see for example Lin and Jeon (2003) for more discussion. In addition, when $X \sim N(\mu_x, \Sigma)$ and $Y \sim N(\mu_y, \Sigma)$, model (1) is satisfied with $\mu = \mu_y - \mu_x$.

1.2. Discriminant analysis through the semiparametric model

Suppose h, μ and Γ are known and let $Z = (z_1, \ldots, z_p)^T$ be an independent observation from class X or class Y. Under the semiparametric model introduced in the last subsection, the well known Bayes procedure yields a classification rule that classifies Z to class X if and only if $D_L(Z) \leq 0$ where

$$D_L(Z) = \{h(Z) - \mu/2\}^T \Gamma^{-1} \mu. \tag{2}$$

This is in fact equivalent to applying Fisher's LDA to the transformed data $h(Z)$, $h(X)$ and $h(Y)$ and the misclassification rate of this rule is seen as

$$R = \Phi(-\Delta_p/2), \quad \text{where } \Delta_p = \sqrt{\mu^T \Gamma^{-1} \mu}. \tag{3}$$

When p is bounded, what we introduced above is similar to Case 1 in Lin and Jeon (2003). We now discuss the estimation of the components in $D_L(Z)$ when p is very large. Noting that the discrimination rule $D_L(Z)$ depends on Γ and μ only through the product $\Gamma^{-1}\mu$, we propose to estimate $\beta := \Gamma^{-1}\mu$ by the Dantzig selector in Candes and Tao (2007) and Cai and Liu (2011) as

$$\hat{\beta} = \arg\min_{\beta\in\mathcal{R}^p} \{|\beta|_1 \text{ subject to } |\hat{\Gamma}\beta - \hat{\mu}|_\infty \leq \lambda_n\}, \tag{4}$$

where λ_n is a tuning parameter, $\hat{\Gamma}$ and $\hat{\mu}$ are estimators of Γ and μ defined in Section 2. On the other hand, we estimate $h(Z) - \mu/2$ using $\tilde{h}_Z$ as in (7). We then classify Z to class X if $\tilde{h}_Z^T\hat{\beta} \leq 0$, and to class Y if $\tilde{h}_Z^T\hat{\beta} > 0$. We shall call this the *Semiparametric Linear Programming Discriminant* (SLPD) rule. Note from (3) that the Bayes risk is independent of h. Consistent to this, the SLPD rule is invariant about h; see Proposition 1.

While we are finishing this paper, we found that the semiparametric model in this paper is also studied in Han et al. (2013) and Mai and Zou (2013), but with key differences. Our method and assumptions are different from those in Han et al. (2013) and Mai and Zou (2013). Under the semiparametric model, we directly estimate the Bayes rule, while Mai and Zou (2013) made use of an equivalent least square formulation for estimating β and Han et al. (2013) is based on the regularized optimal affine discriminant analysis in Fan et al. (2012). In terms of estimation method, we use median in estimating μ and use Dantzig selector in estimating β. In terms of assumptions, we do not assume the irrepresentable condition (Zhao and Yu, 2006); see for example Definition 8 of Han et al. (2013) and (18) of Mai and Zou (2013). This condition is known to be sufficient for selecting the zero entries in β consistently in theory, but can be easily violated in practice (Zhao and Yu, 2006). What is more, our sparsity assumption on β is more general; see (12) in Theorem 1. More specifically, we do not require the number of nonzero elements of β to be relatively small, while Han et al. (2013) and Mai and Zou (2013) considered the case that the number of nonzero elements of β is much smaller than n. Last but not least, we allow the logarithm of the dimension to grow slower than the square root of sample size while Mai and Zou (2013) requires that the logarithm dimension grows slower than the cube root of n.

Please cite this article in press as: Jiang, B., Leng, C., High dimensional discrimination analysis via a semiparametric model. Statistics and Probability Letters (2015), http://dx.doi.org/10.1016/j.spl.2015.11.012