doi:10.3969/j.issn.2095-0411.2020.02.004

面向高维缺失数据集的线性判别分析方法

刘 鹏,叶 宾

(中国矿业大学 地下空间智能控制教育部工程研究中心,江苏 徐州 221116;中国矿业大学 信息与控制工程学院,江苏 徐州 221116)

摘要:线性判别分析尽管在许多实际应用中表现良好,但是它在处理含有缺失数据的高维数据集时,效果却很不理想。这一方面是由于线性判别分析方法无法准确地预测或填充缺失数据,另一方面是由于在高维情况下,线性判别分析使用的样本协方差矩阵不再是总体协方差矩阵的一个良好估计。因此导致计算出的判别函数值产生很大的偏差。基于随机矩阵理论,采用总体协方差矩阵的 Lasso 估计,提出了一种处理高维缺失数据集的线性判别分析改进方法。在多种人造及真实数据集上的仿真结果表明,所提方法的分类正确率优于其他同类算法。

关键词:线性判别分析;缺失数据;高维数据;Lasso 估计

中图分类号:TP 181 **文献标志码**:A **文章编号**:2095-0411(2020)02-0031-07

Linear Discriminant Analysis for High-Dimensional Dataset with Missing Observations

LIU Peng, YE Bin

(Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China; School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: Although it performs well in many applications, Linear discriminant analysis (LDA) is impractical for high-dimensional datasets with missing observations. One of the reasons for it is that most of the classification methods cannot predict or impute the missing values correctly, the other reason is that the sample covariance matrix used in LDA is no longer a good estimator of the population covariance matrix in high dimensions. Therefore, there will be a relatively large deviation for the discriminant function values. Based on the results from random matrix theory and by exploiting a Lasso estimator of the population covariance matrix, an improved LDA classifier for high-dimensional

收稿日期:2019-11-21。

基金项目:徐州市应用基础研究计划资助项目(KC18069)。

作者简介:刘鹏(1992—),男,江苏徐州人,硕士生。通信联系人:叶宾(1980—),E-mail:yebin@cumt.edu.cn

引用本文:刘鹏,叶宾.面向高维缺失数据集的线性判别分析方法[J]. 常州大学学报(自然科学版),2020,32(2):31-37.

dataset with missing observations is proposed. Simulation results show that our proposed method is superior to the competitors for a wide variety of synthetic and real data sets.

Key words: linear discriminant analysis; missing data; high-dimensional data; Lasso estimation

随着大数据时代的到来,高维数据普遍产生和存在于工业制造、生物信息、金融证券以及电子商务等各个领域[1-3]。高维数据的特点在于其变量(特征)数目 p 接近甚至超过样本数目 n 。因此高维数据不再符合经典多元统计分析中,特征向量的维度 p 固定而样本数目 n 趋于无穷的假设。数据维度的增加给一些经典的统计方法和机器学习算法带来巨大挑战。线性判别分析(Linear discriminant analysis, LDA)作为一种常用的有监督分类算法,在处理高维数据时却效率很低或者根本不适用。其中一个主要的原因是在高维条件下,样本协方差矩阵是病态的或奇异的,它不再是总体协方差矩阵的一个良好估计。当使用该样本协方差矩阵来代替真实协方差矩阵计算判别函数数值时,无疑将产生非常大的偏差,最终造成错误的分类以及正确率的降低。

目前,针对高维数据的判别分类问题,已经在 LDA 的基础上,相继提出了一些改进的算法。其中一类方法主要是直接对 LDA 算法进行改进,例如 FRIEDMAN 通过提高 LDA 中总体协方差矩阵 $\boldsymbol{\Sigma}$ 估计的精度[4],提出了一种正则化的 LDA 算法,该算法在一定范围内可用于高维数据的分类;另一类方法主要是把降维算法和 LDA 结合来对高维数据进行分类,例如 SUN 等通过主成分分析对燃油数据先进行维数约减,根据样本正确率的变化趋势,选择 5 个主成分进行线性判别分析[5]。但上述这些方法都要求输入完整的观测数据集。随着数据收集方式越来越多,含有缺失值的数据集大量存在,这将导致上述方法无法正常使用。

处理缺失数据的通常方法是首先进行数据预处理,例如简单地剔除含有缺失数据的样本或者对缺失数据进行填充处理等[6]。但是当数据集中缺失数据过多时,剔除这些缺失数据样本可能会造成数据样本过少而无法使用[7]。处理缺失数据集的其他常用方法为首先使用最近邻填充或者多重填充等方法对缺失数据进行填充,得到一个完整的数据集,然后再应用 LDA 或其他分类算法,这样的处理步骤相对较为繁琐,效率较低[8]。最近,LOUNICI 基于随机矩阵理论分析,采用总体协方差矩阵的 Lasso 估计,提出了一种含有缺失数据的高维协方差矩阵无偏估计方法[9]。课题组把这种高维协方差矩阵的无偏估计与 LDA 相结合,提出了一种面向高维缺失数据集的线性判别分析方法。对高维缺失数据集的仿真实验表明,该方法能够有效地处理含有缺失数据的高维数据分类问题,而且具有较高的分类正确率。

1 线性判别分析

线性判别分析是一种有监督的分类方法,其基本原理是利用类标签已知的样本,计算判别函数的参数模型,然后根据判别函数把未知样本分类到已知类别中。

假设 p 维随机向量 $\boldsymbol{x}$ 的 n 个观测样本分别为 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$,由其构成的数据矩阵为 $\boldsymbol{X}=(\boldsymbol{x}_1;\cdots;\boldsymbol{x}_n)\in\mathbb{R}^{n\times p}$ 。假如这些观测数据分为 K 类,并且第 $k\in\{1,2,\cdots,K\}$ 类观测数据中的 p 个变量服从多元正态分布 $N(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ 时,其概率密度函数可以表示为

$$f_k(\boldsymbol{x})=\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \tag{1}$$

式中 $\boldsymbol{\mu}_k$ 和 $\boldsymbol{\Sigma}_k$ 分别表示第 k 类观测数据的均值向量和总体协方差矩阵。在 LDA 中,通常假设每类样本都有相同的协方差矩阵,即 $\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}$ 。对于一个新的观测数据向量 $\boldsymbol{x}\in\mathbb{R}^{1\times p}$,它属于第 k 类的后验概率 $P(Y=k\mid\boldsymbol{x})$可以由贝叶斯定理得到

$$P(Y=k \mid \boldsymbol{x})=\frac{f_k(\boldsymbol{x})\pi_k}{\sum_{l=1}^{k} f_l(\boldsymbol{x})\pi_l} \tag{2}$$

式中 π_k 是第 k 类的先验概率。通过选择 k 使后验概率 $P(Y=k \mid \boldsymbol{x})$ 最大，得到新样本 $\boldsymbol{x}$ 的最优分类结果为

$$G_k(\boldsymbol{x})=\arg\max_k P(Y=k \mid \boldsymbol{x})=\arg\max_k \left\{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k-\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k+\log(\pi_k)\right\} \tag{3}$$

式中均值向量 $\boldsymbol{\mu}_k$，协方差矩阵 $\boldsymbol{\Sigma}$ 以及先验概率 π_k 都是由观测数据矩阵 $\boldsymbol{X}$ 估算得到。

2　高维缺失数据的线性判别分析

对于高维数据集，其样本数目 n 接近甚至超过数据维数 p，此时样本协方差矩阵不再是总体协方差矩阵 $\boldsymbol{\Sigma}$ 的一个良好估计。如果观测数据中还存在有一些缺失值，总体协方差矩阵 $\boldsymbol{\Sigma}$ 将更难以被正确估计。这些都是导致 LDA 在处理高维数据分类问题时效果不佳的一些重要原因。本节通过利用高维缺失数据的总体协方差矩阵无偏估计方法，提出一种改进的线性判别分析算法。

2.1　含缺失值的高维数据协方差矩阵的无偏估计

假设第 $j(j=1,\cdots,p)$ 个变量 $x^{(j)}$ 在第 i 次观测中可获得的概率为 $\delta\in(0,1]$，即数据矩阵 $\boldsymbol{X}$ 的第 (i,j) 个观测值 X_{ij} 被观测到的概率为 δ。当 $\delta=1$ 时，表明数据矩阵 $\boldsymbol{X}$ 中不存在缺失观测值，它是一个完整的数据矩阵。样本协方差矩阵可以表示为

$$\boldsymbol{\Sigma}_n=\frac{1}{n-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \tag{4}$$

基于随机矩阵理论分析，可以建立关于总体协方差矩阵 $\boldsymbol{\Sigma}$ 的 Lasso 估计的 Oracle 不等式[9-11]，为

$$\hat{\boldsymbol{\Sigma}}_n=\arg\min_{\boldsymbol{S}\in\boldsymbol{S}_{p\times p}} \|\boldsymbol{\Sigma}_n-\boldsymbol{S}\|_{\mathrm{F}}^2+\lambda\|\boldsymbol{S}\|_1 \tag{5}$$

式中 $\boldsymbol{S}_{p\times p}$ 是一组 $p\times p$ 的半正定对称矩阵；$\|\boldsymbol{S}\|_{\mathrm{F}}$ 和 $\|\boldsymbol{S}\|_1$ 分别表示矩阵 $\boldsymbol{S}$ 的 Frobenius 范数和迹范数。

对于含有缺失值的高维数据 $\boldsymbol{X}^{(\delta)}$，为了能较好地估计出总体协方差矩阵 $\boldsymbol{\Sigma}$，需要基于已观测到的数据，使用简单的均值填充方法，对缺失数据进行填充，得到完整的观测数据矩阵 $\boldsymbol{X}_{\mathrm{c}}^{(\delta)}$。对于填充后的数据矩阵 $\boldsymbol{X}_{\mathrm{c}}^{(\delta)}$，其样本协方差矩阵为

$$\boldsymbol{\Sigma}^{(\delta)}=\frac{1}{n-1}(\boldsymbol{X}_{\mathrm{c}}^{(\delta)})^{\mathrm{T}}\boldsymbol{X}_{\mathrm{c}}^{(\delta)} \tag{6}$$

在这种情况下，可以利用 $\boldsymbol{\Sigma}^{(\delta)}$ 代替式(5)中的 $\boldsymbol{\Sigma}_n$ 得到 Oracle 不等式。然而，当数据集含有缺失值时，$\boldsymbol{\Sigma}^{(\delta)}$ 不再是总体协方差矩阵 $\boldsymbol{\Sigma}$ 的良好估计。基于简单的观测，可以定义矩阵 $\widetilde{\boldsymbol{\Sigma}}$ 作为缺失数据总体协方差矩阵的无偏估计

$$\widetilde{\boldsymbol{\Sigma}}=(\delta^{-1}-\delta^{-2})\operatorname{diag}(\boldsymbol{\Sigma}^{(\delta)})+\delta^{-2}\boldsymbol{\Sigma}^{(\delta)} \tag{7}$$

因此对于不完整的数据矩阵 $\boldsymbol{X}^{(\delta)}$，需要首先估算出式(7)中 δ 的值(它近似等于 $\boldsymbol{X}^{(\delta)}$ 中已观测到的数据占总数据量的比值)。总体协方差 $\boldsymbol{\Sigma}$ 的 Lasso 估计可通过求解以下凸优化问题得到[12-14]

$$\hat{\boldsymbol{\Sigma}}=\arg\min_{\boldsymbol{S}\in\boldsymbol{S}_{p\times p}}\|\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{S}\|_{\mathrm{F}}^2+\lambda\|\boldsymbol{S}\|_1 \tag{8}$$

其中 λ 的最优参数取值为

$$\lambda^*=C\frac{\sqrt{\operatorname{tr}(\widetilde{\boldsymbol{\Sigma}})\,\|\widetilde{\boldsymbol{\Sigma}}\|_\infty}}{\delta}\sqrt{\frac{\log(2p)}{n}} \tag{9}$$

式中 C 是一个较大的正常数。从式(8)的 Lasso 估计可以看出，这种估计方法是矩阵回归框架下 Lasso

估计在协方差矩阵中的一个应用;并且已经理论证明,式(8)中的 Lasso 估计 $\hat{\boldsymbol{\Sigma}}$ 对任意的 n,p,δ 来说,都是 minimax 最优的[9]。

2.2 高维缺失数据集的线性判别分析算法

将上述总体协方差矩阵的无偏估计和线性判别分析相结合,提出如下处理高维缺失数据集的分类算法。

算法:高维缺失数据集的线性判别分析(Linear discriminant analysis for high-dimensional dataset with missing observations, LHDMO)

输入:含有缺失值的训练数据矩阵 $\boldsymbol{X}_{\mathrm{tr}}^{(\delta)}$ 和测试数据矩阵 $\boldsymbol{X}_{\mathrm{te}}$

输出:分类正确率

1—将 $\boldsymbol{X}_{\mathrm{tr}}^{(\delta)}$ 中被标记的样本分为 K 类;

2—计算 $\boldsymbol{X}_{\mathrm{tr}}^{(\delta)}$ 中已观测到的数据比例,即可估算出 δ 值;

3—计算公式(6)中的样本协方差矩阵 $\boldsymbol{\Sigma}^{(\delta)}$;

4—计算公式(7)中 $\widetilde{\boldsymbol{\Sigma}}$ 以及公式(9)中参数 λ^{*} ;

5—求解公式(8)中的凸优化问题,得到 $\boldsymbol{\Sigma}$ 的无偏估计 $\hat{\boldsymbol{\Sigma}}$;

6—对每一类,计算公式(3)中的先验概率 π_k;基于 $\boldsymbol{X}_{\mathrm{tr}}^{(\delta)}$ 中已观测到的数据,计算公式(3)中的均值向量 $\boldsymbol{\mu}_k$;

7—把 $\boldsymbol{X}_{\mathrm{te}}$ 中的样本向量 $\boldsymbol{x}$ 逐个代入公式(3)中,计算 $G_k(\boldsymbol{x})$;将 $\boldsymbol{X}$ 分类到使 $G_k(\boldsymbol{X})$ 最大的那一类中;

8—计算出分类正确率。

如果测试数据矩阵 $\boldsymbol{X}_{\mathrm{te}}$ 含有缺失数据,要先用数据填充的方法对数据填充,然后再进行分类。对于算法中的步骤 5 的凸优化问题,可以通过调用 Matlab 中 CVX 程序包进行求解[15]。

3 仿真实验结果及分析

3.1 仿真数据模型

在仿真实验中,模拟数据集由 3 类样本数据构成。这 3 类数据分别对应 3 种均值不同、协方差矩阵相同的多元高斯分布:$N_1(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$;$N_2(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$ 和 $N_3(\boldsymbol{\mu}_3,\boldsymbol{\Sigma})$。其中第 1 类样本数据的均值向量 $\boldsymbol{\mu}_1=\boldsymbol{0}$,第 2 类的均值向量 $\boldsymbol{\mu}_2=\boldsymbol{0.5}$,第 3 类的均值向量 $\boldsymbol{\mu}_3=-\boldsymbol{\mu}_2$。总体协方差矩阵 $\boldsymbol{\Sigma}$ 利用式(10)对应的块对角矩阵生成(此种形式的协方差矩阵已广泛应用于分类问题的研究[16-17]),为

$$\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{\rho} & 0 & 0 & \cdots & 0\\ 0 & \boldsymbol{\Sigma}_{-\rho} & 0 & \cdots & 0\\ 0 & 0 & \boldsymbol{\Sigma}_{\rho} & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & 0\\ 0 & 0 & 0 & \cdots & \boldsymbol{\Sigma}_{-\rho}\end{bmatrix}_{p\times p} \tag{10}$$

其中

$$\boldsymbol{\Sigma}_{\rho}=\begin{bmatrix}1 & \rho & \cdots & \rho^{9}\\ \rho & 1 & \cdots & \rho^{8}\\ \vdots & \vdots & \ddots & \vdots\\ \rho^{9} & \rho^{8} & \cdots & 1\end{bmatrix}_{10\times 10} \tag{11}$$

由于矩阵块 $\boldsymbol{\Sigma}_\rho$ 中的第 (i,j) 个元素为

$$\sigma_{ij}=\rho^{|i-j|},1\leqslant i,j\leqslant 10 \tag{12}$$

因此,$\boldsymbol{\Sigma}_\rho$ 中的变量相关性随着任何一对变量之间的距离的变化而变化。实验中选取 $\rho=0.8$。

为了构造缺失数据集,根据不同的 δ 值,在这 3 类样本数据中随机地剔除一些数据点。

3.2　实验结果与分析

为了将本文所提算法 LHDMO 与其他处理缺失数据的分类方法相比较,对含有缺失值的高维数据集,选取两种广泛应用的判别分类方法——正则化的高维线性判别分析(HDRDA)[17]和最小距离经验贝叶斯估计器(MDEB)[18]进行分类。这些分类算法可以从 R 语言中的 sparsediscrim 包中获得。在分类之前,首先使用 K-最近邻(KNN)[19]以及主成分分析(PCA)[20]方法对缺失数据集进行填充预处理。

仿真实验中,变量数目固定为 $p=50$,样本数目 n 依次为 24,30,60,并且另外生成 150 个样本作为测试数据集。实验结果见表 1 和表 2,每种算法的平均分类正确率都是对 50 次实验结果取平均得到。从表 1 和表 2 可以看出,LHDMO 在高维情况下(即 $p\geq n$ 时)优于其他高维线性判别分析算法,并始终保持着较高的分类正确率。而且,随着缺失数据的增多(从表 1 中的 $\delta=0.9$ 变化到表 2 中的 $\delta=0.7$),LHDMO 的分类正确率虽然有了降低,但仍高于其他算法。此外,当样本数目 n 大于变量数目 p 时,LHDMO 算法的分类效果也是这几种算法中最好的。

表 1　不同算法随样本数变化的平均分类正确率($p=50$,$\delta=0.9$)

Table 1　The average correct classification rate for different algorithms with different sample sizes ($p=50$, $\delta=0.9$)

算法	$n=24$	$n=30$	$n=60$
PCA+ HDRDA	0.613	0.537	0.787
KNN+HDRDA	0.603	0.757	0.713
PCA+MDEB	0.750	0.877	0.907
KNN+MDEB	0.740	0.843	0.867
PCA+LHDMO	0.860	0.937	0.953
KNN+LHDMO	0.857	0.937	0.947

表 2　不同算法随样本数变化的平均分类正确率($p=50$,$\delta=0.7$)

Table 2　The average correct classification rate for different algorithms with different sample sizes ($p=50$, $\delta=0.7$)

算法	$n=24$	$n=30$	$n=60$
PCA+ HDRDA	0.607	0.507	0.637
KNN+HDRDA	0.550	0.487	0.520
PCA+MDEB	0.717	0.770	0.803
KNN+MDEB	0.660	0.690	0.780
PCA+LHDMO	0.703	0.787	0.857
KNN+LHDMO	0.643	0.730	0.817

4　真实数据实验结果及分析

本节所用的 2 个真实数据集分别是乳腺癌数据集和小白鼠蛋白质表达数据集[21]。乳腺癌数据集不含有缺失数据,共有 212 个恶性肿瘤样本和 357 个良性肿瘤样本,并且每个样本都包含 $p=30$ 个特征变量。根据不同的 δ 值任意地去除样本中的数据点,并随机选取 n 个样本作为训练数据,其余样本作为测试数据,每种算法的平均分类正确率都是由 50 次实验结果取平均数得到。对于不同的数据缺失率,实验结果见表 3 和表 4。虽然在个别情况下,LHDMO 的分类正确率略低于 MDEB 方法,但它总体上仍然获得了较高的分类正确率。

表 3 乳腺癌数据集的平均分类正确率($\delta=0.9$)

Table 3 The average correct classification rate for the breast cancer dataset($\delta=0.9$)

算法	$n=20$	$n=30$	$n=50$
PCA+ HDRDA	0.739	0.790	0.850
KNN+HDRDA	0.751	0.793	0.866
PCA+MDEB	0.883	0.899	0.899
KNN+MDEB	0.881	0.899	0.899
PCA+LHDMO	0.881	0.902	0.902
KNN+LHDMO	0.874	0.902	0.898

表 4 乳腺癌数据集的平均分类正确率($\delta=0.7$)

Table 4 The average correct classification rate for the breast cancer dataset($\delta=0.7$)

算法	$n=20$	$n=30$	$n=50$
PCA+ HDRDA	0.703	0.776	0.857
KNN+HDRDA	0.654	0.760	0.837
PCA+MDEB	0.906	0.902	0.895
KNN+MDEB	0.888	0.882	0.873
PCA+LHDMO	0.913	0.904	0.900
KNN+LHDMO	0.893	0.886	0.873

小白鼠蛋白质表达数据集共含有 1 080 个样本,分为 8 个类别,其中前 4 类样本是通过测量 38 个对照组小鼠大脑皮层中的 77 种蛋白质的表达水平而得到,另外 4 类样本是关于 34 个唐氏综合症小鼠的 77 种蛋白质水平的测量结果。每组小鼠分别独立测量 15 次。该数据集本身含有缺失值,经计算,其中已观测到的数据比例为 $\delta=0.983\ 2$。实验结果见表 5。尽管在高维情况下,每种算法的分类正确率都不高,但是 LHDMO 算法始终要高于其他 2 种分类算法。而且随着样本数 n 的增加,各种分类算法的正确率也都有了明显的提高。

表 5 小白鼠蛋白质表达数据集的平均分类正确率($p=77$)

Table 5 The average correct classification rate for the mice protein expression dataset($p=77$)

算法	$n=40$	$n=64$	$n=128$
PCA+ HDRDA	0.636	0.724	0.856
KNN+HDRDA	0.621	0.673	0.852
PCA+MDEB	0.560	0.631	0.729
KNN+MDEB	0.561	0.620	0.730
PCA+LHDMO	0.690	0.744	0.870
KNN+LHDMO	0.667	0.712	0.874

5 结 论

针对线性判别分析方法在处理高维缺失数据集时效率较低的问题,提出了一种基于 Lasso 协方差矩阵估计的线性判别分析算法 LHDMO。该算法通过求解一个凸优化问题得到高维协方差矩阵的无偏估计,然后结合线性判别分析,实现对高维缺失数据集的分类。基于数据集和真实数据集的仿真实验表明,所提算法能够有效处理含有缺失数据的高维数据集,并且得到了较高的分类正确率。但是对于缺失程度较高的数据集,由于部分变量的均值偏差太大甚至无法求出,导致算法 LHDMO 的分类效率明显降低,这也是课题组将来需要解决的问题之一。

参考文献:

[1]张瑞, 蒋晨之, 苏剑波. 基于稀疏特征挑选和概率线性判别分析的表情识别研究[J]. 电子学报,2018,46(7):1710-1718.

[2]张靖,胡学钢,李培培,等. 基于迭代 Lasso 的肿瘤分类信息基因选择方法研究[J]. 模式识别与人工智能,2014,27(1): 49-59.

[3]刘丽萍.大维数据背景下金融协方差阵的估计及应用[J]. 系统工程理论与实践,2017,37(3):597-606.

[4]FRIEDMAN J H. Regularized discriminant analysis[J]. Journal of the American Statistical Association, 1989, 84(405): 165-175.

[5]SUN P Y, BAO K W, LI H H, et al. An efficient classification method for fuel and crude oil types based on m/z 256 mass chromatography by COW-PCA-LDA[J]. Fuel, 2018, 222: 416-423.

[6]GANTAYAT S S, MISRA A, PANDA B S. A study of incomplete data-a review[C]//Proceedings of the Internation-

al Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013. Cham: Springer, 2014: 401-408.

[7]ENDERS C K. Applied missing data analysis[M]. New York: Guilford Press, 2010.

[8]OUNPRASEUTH S, MOORE P C, YOUNG D M. Imputation techniques for incomplete data in quadratic discriminant analysis[J]. Journal of Statistical Computation and Simulation, 2012, 82(6): 863-877.

[9]EL KAROUI N. Spectrum estimation for large dimensional covariance matrices using random matrix theory[J]. The Annals of Statistics, 2008, 36(6): 2757-2790.

[10]JOHNSTONE I M, MA Z. Fast approach to the Tracy-Widom law at the edge of GOE and GUE[J]. Annals of Applied Probability, 2012, 22(5): 1962-1988.

[11] VERSHYNIN R. Introduction to the non-asymptotic analysis of random matrices[R]. Paris: Institut Henri Poincaré, 2011.

[12]LOUNICI K. High-dimensional covariance matrix estimation with missing observations[J]. Bernoulli, 2014, 20(3): 1029-1058.

[13]LOUNICI K. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators[J]. Electronic Journal of Statistics, 2008, 2: 90-102.

[14]BICKEL P J, RITOV Y, TSYBAKOV A B. Simultaneous analysis of Lasso and Dantzig selector[J]. The Annals of Statistics, 2009, 37(4): 1705-1732.

[15]GRANT M, BOYD S, YE Y. CVX: matlab software for disciplined convex programming[EB/OL].(2013-09-01)[2018-04-10].http://cvxy.com/cvx/.

[16]GUO Y, HASTIE T, TIBSHIRANI R. Regularized discriminant analysis and its application in microarrays[J]. Biostatistics, 2005, 1(1): 1-18.

[17]RAMEY J A, STEIN C K, YOUNG P D, et al. High-dimensional regularized discriminant analysis[R]. New York: arXiv, 2016:1602.01182.

[18]SRIVASTAVA M S, KUBOKAWA T. Comparison of discrimination methods for high dimensional data[J]. Journal of the Japan Statistical Society, 2007, 37(1): 123-134.

[19]ALEXANDER K, MATTHIAS T. Imputation with the R package VIM [J]. Journal of Statistical Software, 2016, 74(7): 1-16.

[20]JOSSE J, HUSSON F. missMDA: a package for handling missing values in multivariate data analysis[J]. Journal of Statistical Software, 2016, 70(1): 1-31.

[21]DUA D, KARRA T E. UCI machine learning repository[EB/OL].(1995-11-01)[2018-03-20]. http://archive.ics.uci.edu/ml.

(责任编辑:李艳)