基于秩能量距离的超高维特征筛选研究*

何胜美 李高荣 许王莉

内容提要: 特征筛选是超高维数据分析中常用的快速降维方法。本文首先基于秩能量距离提出了一种新的适用于超高维判别分析的特征筛选方法(RED-SIS)。该方法无需假定模型结构和有限矩条件,对厚尾协变量数据具有较好的稳健性。其次,本文研究了该方法的理论性质,并在几个较为宽松的正则条件下,证明了确定筛选性质和排序相合性。结果表明,RED-SIS 能有效处理变量维数 p 和样本量 n 满足 $\log p = O(n^{\alpha})$ 的超高维判别分析特征筛选问题,且随着样本量的增加,筛选出的特征集合包含全部真实重要特征集合的概率趋近于 1。最后,蒙特卡罗模拟研究该方法的有限样本性质,并和现有的超高维特征筛选方法进行比较。数值模拟结果表明,该方法在厚尾数据情况下具有明显的优越性,同时,实际数据分析的研究结果也说明 RED-SIS 方法的有效性。

关键词: 超高维数据; 特征筛选; 秩能量距离; 确定筛选性质
DOI: 10.19343/j.cnki.11-1302/c.2020.08.009
中图分类号: O212 **文献标识码:** A **文章编号:** 1002-4565(2020)08-0117-12

A Feature Screening for Ultra-high Dimensional Discriminant Analysis Using Rank-based Energy Distance

He Shengmei Li Gaorong Xu Wangli

Abstract: Feature screening is a common method for dimensionality reduction in ultra-high dimensional data analysis. In this paper, a new feature screening procedure, named RED-SIS, is first proposed based on rank-based energy distance. This procedure does not need to assume model structure and finite moment conditions, and is robust for heavy-tailed covariate. Secondly, the asymptotical properties of the proposed method are studied, the sure screening property and ranking consistency property are proved under some mild regularity conditions. It shows that the proposed RED-SIS can effectively deal with the ultra-high dimensional discriminant analysis with the sample size n and the dimension number p satisfying $\log p = O(n^{\alpha})$. Also, as the sample size increases, the screened set contains all true important feature sets with the probability tending to 1. Last, we present the finite sample performance of the proposed method by numerical analysis, and compare the proposed method with the existing methods for the feature screening in ultra-high dimensional discriminant analysis. Both simulation and real data analysis shows that RED-SIS can be more competitive for feature screening with heavy-tailed distribution.

* 基金项目: 国家自然科学基金项目"非参数独立性检验及其应用研究"(11971478); 国家自然科学基金"高维回归模型的大规模统计学习和推断"(11871001); 国家自然科学基金"不完全数据下半参数混合效应模型的研究"(11971001); 北京市自然科学基金"大规模因果推断模型的统计学习及其应用"(1182003); 中央高校基本科研业务费专项资金"高维半参数因果推断模型的统计学习研究"(2019NTSS18); 北京市自然科学基金资助项目"监督学习的降维理论与优化方法"(Z190002); 教育部人文社会科学重点研究基地重大项目"基于大数据的精准医学生物统计分析方法及其应用研究"(16JJD910002)。

Key words: Ultra-high Dimensional Data; Feature Screening; Rank-based Energy Distance; Sure Screening Property

一、引言

随着数据收集和数据存储技术的提高,超高维数据出现在生物影像,经济预测和肿瘤分类等各个研究领域。例如,随着基因芯片技术的应用,可以应用基因表达谱数据进行肿瘤分类和诊断。人类肺癌数据集(Bhattacharjee 等,2001)分成了 5 个不同的子类,包含了 12600 个 mRNA 表达水平,203 个样本。不同类型的肺癌病人,需要采用不同的治疗方案。因此,如何利用该数据集构建模型,正确识别肺癌类型,及识别癌症类型的生物标志物,在临床医学上显得意义重大。但由于数据集中特征个数远远超过样本数,给统计建模带来了很大挑战。

为了解决上述超高维变量选择遇到的挑战,Fan 和 Lv (2008) 首次提出了变量筛选的概念。该文献在超高维线性回归问题中,尝试利用边际 Pearson 相关系数对协变量重要性排序,筛选出相关系数较大的部分变量,提出了超高维线性模型下的确定性独立筛选方法 SIS (Sure Independence Screening),并证明在一定正则条件下该方法满足确定筛选性质,即模型所选变量集合包含全部重要变量的概率趋向于 1。此后,变量筛选方法得到了学者的广泛关注,大量以边际相关系数为基础的变量筛选方法相继涌现。针对广义线性模型,Fan 和 Song (2010) 提出了基于边际极大似然的筛选方法。Chang 等(2013) 利用边际经验对数似然比对变量进行选择。但这些方法依赖于模型的线性或者广义线性结构假设,当模型结构假设出现偏差时,筛选效果较差。Fan 等(2011) 利用 B 样条估计边际相关效应,提出了非参数独立筛选方法 NIS (Nonparametric Independence Screening),去掉了模型假设的限制。其他非参数筛选方法还有 Li 等(2012) 利用 Kendall τ 相关系数代替 Pearson 相关系数建立变量筛选方法 RRCS (Robust rank correlation screening); Li 等(2012) 基于距离相关性,提出了距离相关筛选方法 DC-SIS (Distance correlation SIS); Pan 等(2019) 定义了球相关并以此提出了球相关变量筛选方法 BCor-SIS (Ball correlation SIS)。但是 DC-SIS 受限于一阶矩的有限条件,对异常值比较敏感。BCor 的估计需要构造 6 阶统计量,使得 BCor-SIS 的计算复杂度达到了 $O(n^6p)$,计算复杂性较高。针对超高维判别分类问题,Ma 和 Zou (2013) 基于 Kolmogorov-Smirnov 统计量提出了特征筛选方法 KF (Kolmogorov Filter),但该方法只适用于二分类的情况。Ma 和 Zou (2015) 提出了 FKF (Fused KF) 方法将 KF 变量筛选推广到因变量为多类和连续的情形。Cui 等(2015) 基于条件分布与无条件分布的差异提出了适用于多分类的稳健特征筛选方法 MV-SIS,筛选指标是条件分布函数与无条件分布函数的距离加权,也就是 Cramer-von-Mises 距离。Cheng 等(2017) 利用条件秩和无条件秩的差异,提出了基于秩的变量筛选方法 RRS (Robust Rank Screening),该方法避免估计样本分布函数,对小样本数据具有一定的稳健性。He 等(2019) 通过对 MV-SIS 进行标准化修正,提出了特征筛选方法 AD-SIS,明显提升了厚尾数据的变量筛选效果。

考虑超高维判别分析,如果预测变量 X 对分类变量 Y 有显著影响,则不同类别的 Y 值对应的 X 的分布有显著性差异。因此,超高维判别分析的变量筛选,就是将类别间分布有显著性差异的预测变量筛选出来。分布差异的刻画通常采用分布函数的 Kolmogorov-Smirnov 距离和 Cramer-von-Mises 距离。KF-SIS 和 MV-SIS 分别基于这两种距离提出了相应的变量筛选方法,而 AD-SIS 则是采用加权 Cramer-von-Mises 距离。Curry 等 (2018) 结合能量距离的思想(Székelya 和 Rizzo,2017),提出了基于秩的能量距离来刻画两类分布间的差异,并用于两类样本分布函数相等的检验。相对于能量距离,基于秩的能量距离不受限于一阶矩的条件,具有更好的稳健性。

本文尝试将秩能量距离引入分类因变量的筛选问题。基于秩能量距离,提出了一种新的适用于多分类因变量的特征筛选方法。与现存的超高维判别分析特征筛选方法相比,本文所提方法具有以下优势: 首先,该方法无需特定的模型假设和相关的矩条件; 其次,该方法对厚尾分布预测变量具有较好的稳健性。与 Curry 等(2018) 不同的是,该方法能用于多分类因变量 Y 的情况。

二、特征筛选方法及其理论性质

本节主要研究了超高维判别分析的变量筛选问题。对预测变量 $X_j(j=1,2,\cdots,p)$,根据 Y 的不同类别,预测变量对应分成不同的类别。变量筛选就是将不同类别间有显著性差异的预测变量筛选出来。本文提出新的衡量不同类别间有显著性差异的方法 RED-SIS,基于此方法刻画不同类分布函数之间的差异,进而筛选出对 Y 有显著影响的预测变量。本节首先介绍秩能量距离,然后基于秩能量距离提出新的特征筛选方法,并探讨该方法的理论性质。

(一) 能量距离、秩能量距离及其估计

能量距离是 Székelya 和 Rizzo (2017) 提出的刻画两个随机向量的距离。

定义 1(能量距离) 设 X,Y 是两相互独立的 d 维随机向量,且满足条件 $E|X|_d<\infty$ 和 $E|Y|_d<\infty$。X' 是 X 的独立复制,Y' 是 Y 的独立复制,则 X,Y 的能量距离为

$$\varepsilon(X,Y)=2E|X-Y|_d-E|X-X'|_d-E|Y-Y'|_d \tag{1}$$

其中,$|X|_d=(X^TX)^{1/2}$ 是 d 维空间中的欧氏距离。当 $d=1$ 时,$\varepsilon(X,Y)$ 是两个随机变量间的能量距离。定义 1 中,$\varepsilon(X,Y)\geqslant 0$,等号成立当且仅当 X 与 Y 的分布函数相等。因此,$\varepsilon(X,Y)$ 可用于刻画两个随机变量分布函数的差异,但受到有限一阶矩条件 $E|X|<\infty$ 和 $E|Y|<\infty$ 的限制。Curry 等(2018) 基于随机变量的标准秩,提出了如下秩能量距离。

设 $H(y)$ 是分布函数,给定数值 y,$R(y,H)$ 表示分布函数 $H(y)$ 下 y 的标准秩,即 $R(y,H)=H(Y)$。假设 X,Y 分别来自总体分布 F 和 G 的独立连续随机变量,其混合分布函数为 $H=\pi F+(1-\pi)G$,其中 $0\leqslant\pi\leqslant 1$。令 X',X'' 是 X 的独立复制,Y',Y'' 是 Y 的独立复制。则 X,X',X'' 和 Y,Y',Y'' 在分布 H 下的标准秩为 $R(X,H)$,$R(X',H)$,$R(X'',H)$ 和 $R(Y,H)$,$R(Y',H)$,$R(Y'',H)$,Curry 等(2018) 定义标准秩的能量距离如下

$$RED(X,Y)=E|R(X,H)-R(Y,H)|-\frac{1}{2}E|R(X',H)-R(X'',H)|$$
$$-\frac{1}{2}E|R(Y',H)-R(Y'',H)| \tag{2}$$

同时,Curry 等(2018) 给出了如下结论,说明秩能量距离 $RED(X,Y)$ 的性质。

定理 1: 设 X,X',X'' 和 Y,Y',Y'' 是分别来自总体分布 F 和 G 的独立连续随机变量,$H=\pi F+(1-\pi)G$ 是 F 和 G 的混合分布,其中 $0\leqslant\pi\leqslant 1$。则 $RED(X,Y)\geqslant 0$ 恒成立,等式成立当且仅当 $F=G$。

设 $X_1,X_2,\cdots,X_{n_1}$ 是来自总体分布 F 的样本,$Y_1,Y_2,\cdots,Y_{n_2}$ 是来自总体分布 G 的样本,$H_n=\frac{n_1}{n}F+\frac{n_2}{n}G$ 是混合样本经验分布函数,$n=n_1+n_2$,$R(X_i,H_n)$,$i=1,2,\cdots,n_1$,$R(Y_i,H_n)$,$i=1,2,\cdots,n_2$ 分别是样本 X_i 和 Y_i 在混合经验分布函数 H_n 下的标准化秩,即:

$$R(X_i,H_n)=R(X_i)/n=n^{-1}\Big(\sum\nolimits_{j=1}^{n_1}I(X_j\leqslant X_i)+\sum\nolimits_{j=1}^{n_2}I(Y_j\leqslant X_i)\Big) \tag{3}$$

$$R(Y_i,H_n)=R(Y_i)/n=n^{-1}\Big(\sum\nolimits_{j=1}^{n_1}I(X_j\leqslant Y_i)+\sum\nolimits_{j=1}^{n_2}I(Y_j\leqslant Y_i)\Big) \tag{4}$$

那么，式(2)中 $RED(X,Y)$ 的估计为

$$\widehat{RED}(X,Y)=\frac{1}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}|R(X_i,H_n)-R(Y_j,H_n)| \tag{5}$$
$$-\frac{1}{2n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|R(X_i,H_n)-R(X_j,H_n)|$$
$$-\frac{1}{2n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}|R(Y_i,H_n)-R(Y_j,H_n)|$$

(二) 变量筛选方法

本节考虑超高维判别分析中的变量筛选。设离散因变量 Y 取 $K(K\geqslant 2)$ 个可能值 $R_y=\{y_1,y_2,\cdots,y_K\}$，且设 $p_k=\Pr(Y=y_k)$，$k=1,2,\cdots,K$，$X=(X_1,X_2,\cdots,X_p)^T$ 是 p 维连续型随机向量。定义重要变量集合为

$$\mathcal{A}=\{j:\text{存在 } y\in R_y,F(y\mid X)\text{ 与 }X_j\text{ 有关}\} \tag{6}$$

其补集 $\mathcal{A}^c=\{j:\text{对任意 } y\in R_y,F(y\mid X)\text{ 与 }X_j\text{ 无关}\}$，则若 $j\in\mathcal{A}$，对应协变量 X_j 为重要变量，反之，若 $j\in\mathcal{A}^c$，则认为 X_j 为不重要变量。

引入变量 $S_k=I(Y=y_k)$，$k=1,2,\cdots,K$，其中 $I(\cdot)$ 为示性函数。显然 $Y=\sum_{k=1}^{K}y_kS_k$ 是 S_k 的线性组合。设 $Y=y_k$，$S_k=1$ 时，X 的取值为 $X=X^{k1}=(X_1^{k1},X_2^{k1},\cdots,X_p^{k1})^T$，而 $Y\neq y_k$，$S_k=0$ 时，X 的取值为 $X=X^{k0}=(X_1^{k0},X_2^{k0},\cdots,X_p^{k0})^T$。考虑 X_j^{k1} 与 X_j^{k0} 的分布函数，根据秩能量距离定义有：$RED(X_j^{k1},X_j^{k0})\geqslant 0$，等号成立当且仅当 X_j^{k1} 与 X_j^{k0} 的分布相同，此时 X_j 与 S_k 相互独立。基于该性质，对于第 j 个特征，构建如下特征筛选指标 ω_j，来衡量 X_j 与 Y 的相关程度

$$\omega_j=\sup_{1\leqslant k\leqslant K}RED(X_j^{k1},X_j^{k0}),\ j=1,2,\cdots,p \tag{7}$$

下面定理 2 说明了用 ω_j 衡量 X_j 与 Y 相关程度的合理性。

定理 2： $\omega_j\geqslant 0$，等式成立当且仅当 X_j 与 Y 相互独立。

定理 2 表明，ω_j 可衡量随机变量 X_j 与 Y 的独立性，X_j 与 Y 不独立时，$\omega_j>0$，并且 ω_j 越大，X_j 对 Y 的影响越大。因此，可以采用指标 ω_j 构建特征筛选，记为 RED-SIS。

假定 $\{(Y_i,X_i),i=1,2,\cdots,n\}$ 是总体 (Y,X) 的随机样本，$X_i=(X_{i1},X_{i2},\cdots,X_{ip})^T$，则 ω_j 可估计为

$$\hat{\omega}_j=\sup_{1\leqslant k\leqslant K}\widehat{RED}(X_j^{k1},X_j^{k0}) \tag{8}$$

其中，$\widehat{RED}(X_j^{k1},X_j^{k0})$ 由式(5)给出。进一步，利用指标 $\hat{\omega}_j$，可以给出重要变量集合 $\mathcal{A}$ 的估计

$$\hat{\mathcal{A}}=\{j:\hat{\omega}_j\geqslant cn^{-\tau},1\leqslant j\leqslant p\} \tag{9}$$

这里，c,τ 是下述条件(C2)中给定常数。在实际应用中，通过设定具体的模型大小 d_n 来估计 $\mathcal{A}$：$\hat{\mathcal{A}}=\{j:\hat{\omega}_j\text{ 属于前 }d_n\text{ 最大值}\}$，在数值模拟部分，类似于文献 He 等(2019)，取 $d_n=[n/\log(n)]$。

(三) 特征筛选性质

为研究上述特征筛选方法的性质，给出如下假设：

(C1) 存在常数 $c_1,c_2>0$，使 $\frac{c_1}{K}\leqslant\min_{1\leqslant k\leqslant K}p_k\leqslant\max_{1\leqslant k\leqslant K}p_k\leqslant\frac{c_2}{K}$ 成立。其中，$p_k=\Pr(Y=y_k)$，$K=O(n^\gamma)$，$\gamma>0$。

(C2) 存在任一正的常数 $c>0$ 和 $0\leqslant\tau\leqslant\frac{1}{2}$，满足：$\min_{j\in A}\omega_j\geqslant 2cn^{-\tau}$。

(C3) 存在常数 $c_3 > 0$, 满足 $\lim\limits_{p\to\infty} inf\{\min\limits_{j\in\mathcal{A}}\omega_j - \min\limits_{j\notin\mathcal{A}}\omega_j\} > c_3$。

条件(C1)要求因变量 Y 取各个值的概率不能太大,也不能太小,确保样本中各类比例保持在一定的平衡程度之内。类别数 $K = O(n^{\gamma})$,说明类别数可以随着样本量 n 按照一定速度趋向于无穷。条件(C2)给出了能正确识别的重要变量的最小信号强度。重要变量的信号不能太小,但随着样本增加,对最小信号的要求逐步以 $n^{-\tau}$ 的速度逐步降低。

在条件(C1)和(C2)下,RED-SIS 的确定筛选性质如下:

定理 3 (确定筛选性质) 在条件(C1)下,对任意 $0 \leqslant \gamma \leqslant 1 - 2\tau$,存在依赖于 c, c_1, c_2 的正常数 b,使得下式成立:

$$\Pr\left(\max_{1\leqslant j\leqslant p} |\hat{\omega}_j - \omega_j| > cn^{-\tau}\right) \leqslant O\left(p\exp\{-bn^{1-2\tau-\gamma} + (1+\gamma)\log n\}\right) \tag{10}$$

在条件(C1)和(C2)下,有:

$$\Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}) \geqslant 1 - O\left(s_n\exp\{-bn^{1-2\tau-\gamma} + (1+\gamma)\log(n)\}\right) \tag{11}$$

其中,s_n 集合 $\mathcal{A}$ 中元素的个数。

由定理 3 可以看出,在比较宽松的正则条件下,RED-SIS 可以处理变量维数 p 和样本量 n 满足 $\log p = O(n^{\alpha})$ 的超高维特征筛选问题,并且随着样本量 n 的增大,选出的变量子集 $\hat{\mathcal{A}}$ 包含真实重要变量集合 $\mathcal{A}$ 的概率趋向于 1。由于不需要具体模型结构和相应矩条件的假设,所以模型 RED-SIS 是自由的,且对厚尾数据和异常值有较好的稳健性。同时,注意到确定筛选性质在响应变量类别个数 K 随 n 趋向于无穷时同样成立。

定理 4 (排序相合性) 如果当 $K\log(n)/n = o(1)$ 和 $K\log(p)/n = o(1)$ 时,条件(C1)和(C3)成立,则有

$$\lim_{n\to\infty} inf\{\min_{j\in\mathcal{A}}\hat{\omega}_j - \min_{j\notin\mathcal{A}}\hat{\omega}_j\} > 0, a.s. \tag{12}$$

定理 4 说明 RED-SIS 特征筛选方法大概率把重要的变量排在不重要的变量前面,确保变量筛选结果的有效性。定理 3 和定理 4 的相关证明见附录。

注 1: 基于 $RED(X_j^{k1}, X_j^{k0})$, $k = 1,2,\cdots,K$ 构造变量筛选指标,另外一种方法可以采用如下概率加权平均的方式

$$\omega_j' = \sum_{k=1}^{K} p_k RED(X_j^{k1}, X_j^{k0}) \tag{13}$$

其中,$p_k = \Pr(Y = y_k)$, $k = 1,2,\cdots,K$。不难证明 $\omega_j' \geqslant 0$,等号成立当且仅当 X_j 与 Y 相互独立。因此可以用 ω_j' 作为变量筛选指标,其估计值为

$$\hat{\omega}_j' = \sum_{k=1}^{K} \hat{p}_k \widehat{RED}(X_j^{k1}, X_j^{k0}) \tag{14}$$

用 ω_j' 进行变量筛选,在一定条件下同样满足确定筛选性质和排序相合性。限于篇幅,本文不再予以证明。比较而言,本文所提出的最大化 $RED(X_j^{k1}, X_j^{k0})$, $k = 1,2,\cdots,K$ 得到筛选指标 ω_j 更容易捕捉重要变量的信号。下面例子能很好说明这一点。

设 Y 是取值于 $R_y = \{1,2,3,4,5,6,7,8\}$ 的离散随机变量,且 $p_1 = \Pr(Y = 1) = 0.05$, $p_k = \Pr(Y = k) = 0.95/7$, $k = 2,\cdots,8$。$X = (X_1, X_2, \cdots, X_{20})^T$ 是 20 维协变量,满足: 当 $Y = 1$ 时 $X_1 = X_2 = a \cdot t(1)$,其中 $a = 1$,5 和 10, $t(1)$ 表示自由度为 1 的 t 分布。$(X_3, \cdots, X_{20})^T \sim N(0_{18}, I_{18})$ 是 18 维零均值单位协方差阵的正态分布随机向量; 当 $Y \neq 1$ 时, $X = (X_1, X_2, \cdots, X_{20})^T \sim N(0_{20}, I_{20})$。显然, X_1, X_2 是重要预测变量,重要变量集合 $\mathcal{A} = \{1,2\}$,不重要变量集合 $\mathcal{A}^c = \{3,4,\cdots,20\}$。按照该设定,生成样本容量为 400 的样本,分别用 $\hat{\omega}_j$ 和 $\hat{\omega}_j'$ 来衡量协变量 X_j 的重要性。方便比较,类似于 He 等

(2019)[15]，分别计算 $\hat{\mathcal{A}}(\min_{j\in\mathcal{A}}\hat{\omega}_j > \max_{j\in\mathcal{A}^c}\hat{\omega}_j)$ 和 $\hat{\mathcal{A}}(\min_{j\in\mathcal{A}}\hat{\omega}_j' > \max_{j\in\mathcal{A}^c}\hat{\omega}_j')$，通过 1000 次重复模拟，得到：当 $a=1,5$ 和 10 时，$\hat{\mathcal{A}}(\min_{j\in\mathcal{A}}\hat{\omega}_j' > \max_{j\in\mathcal{A}^c}\hat{\omega}_j')$ 分别 0.004，0.060 和 0.108，而相应的 $\hat{\mathcal{A}}(\min_{j\in\mathcal{A}}\hat{\omega}_j > \max_{j\in\mathcal{A}^c}\hat{\omega}_j)$ 为 0.014，0.640 和 0.855。显然，$\hat{\omega}_j$ 比 $\hat{\omega}_j'$ 更有效。其原因在于：对于变量 X_1，他们在 $Y=k$ 时的分布函数与 $Y\neq k(k=2,3,\cdots,8)$ 时对应的分布函数差异较小，也就是 $RED(X_1^{k1},X_1^{k0})$，$k=2,\cdots,8$ 较小，虽然 $RED(X_1^{11},X_1^{10})$ 较大，但 p_1 较小，概率加权平均导致相对应的指标 ω_1' 较小，而取最大值的 ω_1 并不受影响。对于 X_2，也有类似的结论。

注 2：多类别判别分析中的变量筛选方法的构造着眼点有差别。Mai 和 Zhou(2015) 从任意两类的分布函数的 Kolmogorov 距离出发，给出了变量筛选方法 FKF，FKF 以任意两类的条件分布函数的最大差异作为变量筛选指标信号，对稀疏信号更有效果，但容易受到异常值的影响。MV-SIS(Cui 等，2015) 是各类的条件概率分布函数与总体无条件分布函数的概率加权 Cramer-von-Mises 距离，比 FKF 更为稳健，但它没能很好的反映分布函数在尾部的差异。He 等(2019) 利用权重函数对 MV-SIS 进行修正，提出了基于双重加权的 Cramer-von-Mises 距离的特征筛选方法 AD-SIS，对厚尾数据具有更好的稳健性。本文提出的方法 RED-SIS 能从标准秩的能量距离角度刻画 X_j^{k1} 与 X_j^{k0} 的分布函数的差异，凸显出协变量 X_j 对 $S_k=I(Y=y_k)$ 的不同影响。异质性是大数据的重要特征，揭示不同类别特征的重要变量往往不同(Xie 等，2019)。如果 X_j 是第 k 类的重要特征变量，则 X_j 能较好刻画 $Y=y_k$ 和 $Y\neq y_k$ 的差异。因此 RED-SIS 对异质性数据有较好的适应性。同时，最大化 $RED(X_j^{k1},X_j^{k0})$，$k=1,2,\cdots,K$ 得到筛选指标 $\omega_j=\sup_{1\leqslant k\leqslant K} RED(X_j^{k1},X_j^{k0})$，$j=1,2,\cdots,p$. 更容易捕捉到重要变量的信号。在以下数值模拟的例 2 中显示，各类别数据样本比例不平衡时，RED-SIS 的变量筛选效果明显好。

三、数值模拟与实例分析

(一) 数值模拟

本节通过蒙特卡洛模拟研究本文所提方法的有限样本性质。本文分别研究了该方法在两类判别分析和多类判别分析中的效果，并比较常用的多类别变量筛选方法 AD-SIS、MV-SIS、FKF 和 RRS 的变量筛选效果。我们考虑以下评价指标：(1) 最小模型大小(MMS) 的 5%，25%，50%，75% 和 95% 分位数，以及 75% 分位数和 25% 分位数差(IQR) 和标准差(SD)，这些结果基于 500 次模拟结果得到。同标准差 SD 一样，IQR 也是反映数据变异程度的指标；(2) P_i：在给定模型大小 $d=[n/\log(n)]$ 的情况下，500 次重复模拟中第 i 个重要变量在重复模拟中被选中的频率；(3) P_α：在给定模型大小 $d=[n/\log(n)]$ 的情况下，500 次重复模拟中所有重要变量在重复模拟中同时被选中的频率。一般而言，指标(1) 中 MMS 各分位数指标越接近模型真实大小说明筛选指标越好；IQR 和 SD 越小，说明筛选方法越稳定；而指标(2) 和(3) 中的各频率越大，说明各重要变量能被正确筛选的概率越高，筛选方法的效果越好。

例 1. 两类线性判别分析中的变量筛选。

设因变量 Y 是取值于 $\{1,2\}$ 的二值随机变量。对于第 i 个样本，给定 $Y_i=k$，$k=1,2$，对应预测向量 $X_i=\mu_k+\varepsilon_i$，其中 $\mu_k=(\mu_{k1},\mu_{k2},\cdots,\mu_{kp})^T\in R^p$ 是第 k 个分量 $\mu_{kk}\neq 0$，其他分量全为 0 的 p 维向量 $(k=1,2)$；$\varepsilon_i=(\varepsilon_{i1},\varepsilon_{i2},\cdots,\varepsilon_{ip})^T$ 是 p 维独立分量的误差向量。记 $p_k=\Pr(Y_i=k)$，取 $\mu_{kk}=3$，$n=40$，$p=2000$，考虑如下设定：

Case 1.1: $p_1=p_2=1/2$，$\varepsilon_{ij}\sim N(0,1)$.

Case 1.2: $p_1=1/3$，$p_2=2/3$，$\varepsilon_{ij}\sim N(0,1)$.

Case 1.3: $p_1=p_2=1/2$，$\varepsilon_{ij}\sim t(2)$.

Case 1.4: $p_1 = 1/3, p_2 = 2/3, \varepsilon_{ij} \sim t(2)$.

显然,Case 1.1 和 1.3 中,两类样本比例均衡,而 Case1.2 和 1.4 不均衡。同时,Case 1.3 和 1.4 具有厚尾误差项。该模型中,X_1 和 X_2 为重要变量。重复上述模拟 500 次,分别利用 RED-SIS 进行变量筛选,并和文献中存在的变量筛选方法 AD-SIS、MV-SIS、FKF 和 RRS 进行比较,模拟结果见表 1。

表 1　　两类线性判别分析的变量筛选 MMS 各分位数和两重要变量被正确选择的概率

情形	方法	5%	25%	50%	75%	95%	IQR	SD	P_1	P_2	P_α
Case 1.1	AD-SIS	2.000	2.000	2.000	2.000	2.000	0.000	0.000	1.000	1.000	1.000
	MV-SIS	2.000	2.000	2.000	2.000	2.000	0.000	0.000	1.000	1.000	1.000
	FKF	2.000	2.000	2.000	2.000	2.000	0.000	0.045	1.000	1.000	1.000
	RRS	2.000	2.000	2.000	2.000	2.000	0.000	0.000	1.000	1.000	1.000
	RED-SIS	2.000	2.000	2.000	2.000	2.000	0.000	0.000	1.000	1.000	1.000
Case 1.2	AD-SIS	2.000	2.000	2.000	2.000	2.000	0.000	0.045	1.000	1.000	1.000
	MV-SIS	2.000	2.000	2.000	2.000	2.000	0.000	0.184	1.000	1.000	1.000
	FKF	2.000	2.000	2.000	2.000	2.000	0.000	0.450	1.000	1.000	1.000
	RRS	2.000	2.000	2.000	2.000	2.000	0.000	0.141	1.000	1.000	1.000
	RED-SIS	2.000	2.000	2.000	2.000	2.000	0.000	0.184	1.000	1.000	1.000
Case 1.3	AD-SIS	2.000	2.000	2.000	2.000	5.000	0.000	6.047	0.984	0.988	0.972
	MV-SIS	2.000	2.000	2.000	2.000	4.000	0.000	4.532	0.988	0.988	0.976
	FKF	2.000	2.000	2.000	2.000	6.000	0.000	7.073	0.988	0.988	0.976
	RRS	2.000	2.000	2.000	2.000	8.000	0.000	7.927	0.978	0.988	0.960
	RED-SIS	2.000	2.000	2.000	2.000	4.000	0.000	4.522	0.986	0.988	0.974
Case 1.4	AD-SIS	2.000	2.000	2.000	3.000	12.050	1.000	12.028	0.968	0.976	0.944
	MV-SIS	2.000	2.000	2.000	2.000	11.000	0.000	10.659	0.970	0.980	0.950
	FKF	2.000	2.000	2.000	3.000	12.000	1.000	11.229	0.976	0.976	0.952
	RRS	2.000	2.000	2.000	3.000	22.000	1.000	23.841	0.962	0.958	0.920
	RED-SIS	2.000	2.000	2.000	2.000	11.000	0.000	10.655	0.970	0.980	0.950

注: 表中 5%,25%,50%,75%和 95%指 500 次重复中 MMS 的相应分位数,所有数据均四舍五入保留 3 位小数。

从表 1 中结果可以看出,在两类别情况下,样本比例不平衡对各方法变量筛选效果影响较小,从重要变量选入的频率 P_α 来看,Case 1.3 相比 Case 1.1 没有变化,而 Case 1.4 各结果也只比 Case 1.2 下降 3%左右。误差项的厚尾分布情况对结果有一定影响,无论是样本平衡情况还是不平衡情况下,误差项 $\varepsilon_{ij} \sim t(2)$ 时的变量筛选的结果都要稍差于 $\varepsilon_{ij} \sim N(0,1)$ 的情况。总体上看,所使用的五种变量筛选方法都能达到满意的效果,MMS 的 50%分位数均等于类别个数,反映结果稳定性的 IQR 都等于零,且在 Case1.3 和 1.4 中,RED-SIS 的标准差 SD 最小,同时包含所有重要变量的概率都在 0.900 以上,且各方法结果之间的差异较小。在两类别情况下,RED-SIS 取得了较为理想的效果,和 AD-SIS、MV-SIS 以及 FKF 差异不大,但 RRS 筛选效果最差,在 Case 1.4 中,样本比例不平衡且误差呈厚尾分布时,500 次模拟中包含全部重要变量的频率只有 0.920。

例 2. 多类线性判别分析中的变量筛选。

本例中考察变量筛选方法在多类别判别分析中的效果,选取类别数 $K = 10$,其他设置类似例 1。给定 $Y_i = k$,向量 $X_i = \mu_k + \varepsilon_i$,其中,$p$ 维向量 $\mu_k = (\mu_{k1}, \mu_{k2}, \cdots, \mu_{kp})^T$ 的第 k 个分量 $\mu_{kk} \neq 0$,其余分量全为 $0(k = 1, 2, \cdots, K)$,$\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{ip})^T$ 是 p 维独立分量的误差向量。误差变量 ε_{ij} 的分布通常设定为标准正态分布或者 t 分布,如 Cui 等(2015)、Cheng 等(2017)和 He 等(2019)。相对于正态分布,t 分布具有厚尾特性,能检验变量筛选方法在厚尾数据上的效果。拉普拉斯分布广泛存在于金融数据当中,具有尖峰、厚尾特征,特别非对称拉普拉斯分布,还具有偏态特性。为考察新特征筛选方法 RED-SIS 在这种复杂情况下的变量筛选效果,下面考虑误差项服从拉普拉斯分布的情况。记 $p_k = \Pr(Y_i = k)$,取 $\mu_{kk} = 3, n = 200, p = 2000$,考虑如下设定:

Case 2.1: $p_k = 1/K, k = 1,2,\cdots,K, \varepsilon_{ij} \sim N(0,1)$

Case 2.2: $p_k = 2[1 + (k-1)/(K-1)]/(3K), k = 1,2,\cdots,K, \varepsilon_{ij} \sim N(0,1)$

Case 2.3: $p_k = 1/K, k = 1,2,\cdots,K, \varepsilon_{ij} \sim Lp(0,1.5,1.5)$

Case 2.4: $p_k = 2[1 + (k-1)/(K-1)]/(3K), k = 1,2,\cdots,K, \varepsilon_{ij} \sim Lp(0,1.5,1.5)$

Case 2.5: $p_k = 1/K, k = 1,2,\cdots,K, \varepsilon_{ij} \sim Lp(0,1,2)$

Case 2.6: $p_k = 2[1 + (k-1)/(K-1)]/(3K), k = 1,2,\cdots,K, \varepsilon_{ij} \sim Lp(0,1,2)$

其中，$Lp(0,1.5,1.5)$ 是位置参数为 0，尺度参数为 1.5 的对称拉普拉斯分布，而 $Lp(0,1,2)$ 是未知参数为 0，尺度参数为 1 和 2 的非对称拉普拉斯分布。Case 2.3、2.4 和 2.6 中，$p_k = 2[1 + (k-1)/(K-1)]/(3K)$，得 $p_1 = 2/(3K)$，$p_K = 4/(3K)$，最后一个类别的概率是第一个类别概率的 2 倍，类别比例严重不平衡。Case 2.3 和 2.4 中存在尖峰厚尾情况，Case2.5 和 Case 2.6 更具有偏态特征。显然，$X_1, X_2, \cdots, X_{10}$ 为重要变量，分别利用 AD-SIS、MV-SIS、FKF、RRS 和 RED-SIS 进行变量筛选，得到相应的 MMS 值和每个重要变量选择的结果。重复模拟 500 次，结果见表 2 和表 3。

表 2　　多分类判别分析变量筛选的 MMS 各指标结果

方法	5%	25%	50%	75%	95%	IQR	SD	5%	25%	50%	75%	95%	IQR	SD
	Case 2.1							Case 2.2						
AD-SIS	10	10	10	10	10	0	0	10	10	10	10	10	0	0
MV-SIS	10	10	10	10	10	0	1	10	10	10	10	14	0	13
FKF	10	10	10	10	10	0	0	10	10	10	10	12	0	5
RRS	10	10	10	10	10	0	0	10	10	10	10	14	0	16
RED-SIS	10	10	10	10	10	0	0	10	10	10	10	10	0	7
	Case 2.3							Case 2.4						
AD-SIS	10	11	14	26	115	15	78	10	13	24	67	309	54	170
MV-SIS	10	12	17	42	170	30	94	11	17	36	102	425	85	192
FKF	10	13	19	35	125	22	97	12	22	41	88	332	66	152
RRS	10	13	21	49	234	36	104	11	19	46	120	482	101	190
RED-SIS	10	10	10	12	55	2	71	10	10	12	26	268	16	160
	Case 2.5							Case 2.6						
AD-SIS	10	12	16	31	123	19	47	11	15	26	67	330	53	152
MV-SIS	10	12	16	28	103	16	41	10	15	25	67	356	52	142
FKF	10	12	15	23	56	11	33	11	14	22	42	129	28	65
RRS	10	13	19	34	107	21	43	11	17	30	69	353	52	130
RED-SIS	10	10	10	11	23	1	24	10	10	11	16	109	6	46

注：表中 5%，25%，50%，75%和 95%指 500 次模拟实验中 MMS 的相应分位数，表中所有数据均按四舍五入保留整数部分。

表 2 给出了多分类判别分析变量筛选中包含全部重要变量的最小模型大小 MMS 的 5%，25%，50%，75%和 95%分位数以及相应的 IQR 值。给定模型大小 $d = [n/\log(n)]$ 的情况下，500 次模拟中包含全部重要变量的频率结果见表 3。结果表明：

（1）误差项 $\varepsilon_{ij} \sim N(0,1)$ 时，所有变量筛选方法都取得了满意的筛选效果，MMS，50%分位数都等于 1，对应的 IQR 值等于 0。同时，类别比例不平衡对变量筛选几乎没有影响；

（2）Case 2.3-2.6 中，误差项 ε_{ij} 服从拉普拉斯分布具有尖峰、厚尾以及偏态等复杂特性，使得变量筛选更具有挑战性。相对误差项服从正态分布的情况，各方法的变量筛选效果都有一定幅度的降低。例如 Case 2.4 中的 RRS 筛选结果，其 MMS 的 50%分位数高达 46，已经严重偏离真实的模型大小 10，且对应的 IQR 为 101，标准差 SD 为 190，说明筛选结果变异幅度较大，非常不稳定。表 3 中变量筛选的频率结果同样可以看出，在误差项服从拉普拉斯分布情况下，包含全部真实重要变量的频率 P_α 同样都有大幅度下降；

（3）拉普拉斯分布误差的情况下，类别比例不平衡也给变量筛选带来了较大影响。相对于平衡情况，不平衡时各方法的 MMS 的 50%分位数都有一定的增长。MMS 的 IQR 值和标准差 SD 都

表 3　　多分类判别分析变量筛选的各重要变量正确筛选的频率结果

方法	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_α
	Case 2. 1										
AD-SIS	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000
MV-SIS	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000
FKF	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000
RRS	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000
RED-SIS	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000
	Case 2. 2										
AD-SIS	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000
MV-SIS	1. 000	0. 998	1. 000	0. 998	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	0. 996
FKF	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000
RRS	1. 000	0. 998	0. 998	0. 998	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	0. 994
RED-SIS	0. 998	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	1. 000	0. 998
	Case 2. 3										
AD-SIS	0. 968	0. 984	0. 980	0. 976	0. 976	0. 972	0. 988	0. 984	0. 990	0. 974	0. 814
MV-SIS	0. 962	0. 978	0. 970	0. 964	0. 970	0. 958	0. 976	0. 964	0. 966	0. 968	0. 718
FKF	0. 958	0. 972	0. 972	0. 972	0. 966	0. 978	0. 962	0. 970	0. 980	0. 968	0. 764
RRS	0. 960	0. 970	0. 968	0. 954	0. 960	0. 952	0. 974	0. 966	0. 968	0. 958	0. 684
RED-SIS	0. 990	0. 990	0. 996	0. 998	0. 996	0. 992	0. 994	0. 992	0. 992	0. 994	0. 936
	Case 2. 4										
AD-SIS	0. 874	0. 908	0. 890	0. 962	0. 980	0. 992	0. 990	0. 988	0. 996	0. 998	0. 640
MV-SIS	0. 806	0. 858	0. 874	0. 936	0. 974	0. 976	0. 982	0. 988	0. 990	0. 996	0. 510
FKF	0. 904	0. 888	0. 884	0. 938	0. 916	0. 936	0. 948	0. 934	0. 934	0. 938	0. 474
RRS	0. 780	0. 844	0. 864	0. 934	0. 970	0. 976	0. 976	0. 980	0. 982	0. 988	0. 458
RED-SIS	0. 930	0. 954	0. 946	0. 982	0. 994	1. 000	0. 994	1. 000	0. 998	0. 998	0. 808
	Case 2. 5										
AD-SIS	0. 982	0. 968	0. 982	0. 978	0. 970	0. 968	0. 988	0. 970	0. 988	0. 982	0. 800
MV-SIS	0. 986	0. 966	0. 982	0. 980	0. 974	0. 978	0. 992	0. 978	0. 990	0. 982	0. 818
FKF	0. 988	0. 984	0. 984	0. 992	0. 992	0. 988	0. 988	0. 988	0. 992	0. 992	0. 896
RRS	0. 982	0. 966	0. 976	0. 978	0. 960	0. 964	0. 982	0. 976	0. 984	0. 982	0. 780
RED-SIS	0. 994	0. 996	0. 996	1. 000	0. 996	0. 998	1. 000	1. 000	0. 998	0. 998	0. 978
	Case 2. 6										
AD-SIS	0. 866	0. 896	0. 924	0. 964	0. 964	0. 982	0. 994	0. 998	0. 994	0. 998	0. 642
MV-SIS	0. 852	0. 892	0. 912	0. 966	0. 966	0. 990	0. 992	1. 000	0. 994	0. 998	0. 624
FKF	0. 948	0. 956	0. 942	0. 944	0. 946	0. 964	0. 962	0. 978	0. 956	0. 964	0. 716
RRS	0. 836	0. 886	0. 906	0. 952	0. 962	0. 990	0. 986	0. 998	0. 996	0. 996	0. 580
RED-SIS	0. 938	0. 974	0. 976	0. 992	0. 996	0. 996	0. 998	1. 000	0. 998	1. 000	0. 876

有大幅增加。同时,包含全部重要变量的频率 P_α 都有一定的下滑;

(4) 误差项服从拉普拉斯分布时,RED-SIS 变量筛选的效果明显优于其他方法 Case 2. 3-2. 6 中,RED-SIS 的 MMS 的 50%分位数最大为 12,非常接近实际模型大小 10,相应的 IQR 指标是各种方法里最小,除了 Case2. 4,RED-SIS 的标准差 SD 为 160,大于 FKF 的 152,其他三种情况中,RED 的标准差 SD 明显小于其他四种方法,反映 RED-SIS 比其他四种方法更稳健。表 3 的频率结果更显示,对于包含全部真实重要变量的频率 P_α, RED-SIS 至少高出其他四种方法 12%。

(5) 由于每一类都只有一个重要变量,第 i 个类别的重要变量是 $X_i, i = 1,2,\cdots,10$; 因此,第 i 个类别重要变量筛选效果由每一个重要变量 $X_i, i = 1,2,\cdots,10$ 的被选中概率 $P_i, i = 1,2,\cdots,10$ 来体现。表 3 结果表明,RED-SIS 的效果均优于其他 4 种方法,对异质性数据具有更好的效果。

(二) 实际数据分析

人类肺癌数据集(Bhattacharjee 等,2001) 分成 5 个不同的子类,包含了 12600 个 mRNA 表达水平,203 个样本: 139 个肺腺癌 (ADEN) ,21 个肺鳞状细胞癌(ADEN) ,6 个小细胞肺癌(SCLC) ,20 个肺类癌(COID) 和 17 个正常肺(NORMAL) 。由于数据维数高,直接进行建模分析比较困难,考虑将新变量筛选方法 RED-SIS 用于该分析过程,同时也与其他变量筛选方法 AD-SIS、MV-SIS、RRS 和 FKF 进行比较。首先将数据集标准化,然后随机抽取 160 个样本组成训练集,余下的 43 个样本作

为测试集; 由于小细胞肺癌(SCLC) 样本量偏少,因此抽样时优先随机抽取 5 个样本放入训练集,剩下一个样本放入测试集。然后分三步处理: 第一步是利用变量筛选方法在训练集上进行变量筛选,选择前 93($3\times[n/\log(n)]$) 个对分类重要的变量。第二步是利用筛选出来的数据,基于惩罚逻辑回归模型(Zhu,2004) 建立分类模型。第三步是利用建立的分类模型对测试集分类。惩罚采用 LASSO,其中最优惩罚参数 λ 利用 10-折交叉验证方法确定。由于训练集和测试集的抽取具有随机性,将上述分析过程重复进行 100 次,计算每一次的训练误差和测试误差。统计 100 次重复试验的训练误差和测试误差的均值和标准差如表 4。

表 4　　不同变量筛选方法在人类肺癌数据集上的平均结果(括号内为对应的标准差)

方法	RED-SIS	AD-SIS	MV-SIS	RRS	FKF
训练误差(%)	2. 64(1. 78)	3. 10(1. 79)	2. 26(1. 36)	1. 96(1. 23)	3. 32(2. 83)
测试误差(%)	8. 02(3. 75)	8. 60(3. 96)	7. 60(3. 56)	8. 09(3. 16)	13. 00(5. 50)

表 4 结果表明,除了 FKF 外,其他四种方法在人类肺癌数据集上都取得了较好的效果,独立测试集的误差均小于 10%,方差都小 5%。虽然效果略低于 MV-SIS,但与 AD-SIS、RRS 和 FKF 比较,新的变量筛选方法 RED-SIS 具有一定优势。

四、基于 RED-SIS 的迭代特征筛选方法

在超高维判别分析中,RED-SIS 能够快速筛选对分类有效的特征变量。但是,类似其他边际筛选方法一样,该方法忽略了协变量间可能存在的相关性,进而漏选重要变量。针对具强相关性协变量的特征筛选,部分学者提出了迭代特征筛选方法,例如 Fan 和 Lv(2008) 提出了 SIS 的迭代特征筛选方法 ISIS; Zhu 等(2011) 提出了迭代的 SIRS 筛选方法; Zhong 和 Zhu(2015) 提出了 DC-SIS 的迭代版本 DC-ISIS。也有部分学者借鉴向前迭代的思想,提出了向前筛选的方法,例如 Wang (2009) 的向前筛选方法 FR 和 Zhou 等(2019) 的 C-FR 方法。本节将基于特征筛选方法 RED-SIS 构造相应的迭代特征筛选方法 RED-ISIS。

借鉴 Zhu 等(2011) 和 Zhong 和 Zhu(2015) 的思路,构造迭代筛选步骤如下:

第一步是用变量筛选方法 RED-SIS 对给定的数据 X 和 Y 进行变量筛选,其中 $X=(X_1,X_2,\cdots,X_n)^T$ 是 $n\times p$ 的数据矩阵,$Y=(Y_1,Y_2,\cdots,Y_n)^T$ 是类别向量。设该次筛选得到 $p_1<N=[n/\log(n)]$ 个协变量集合为 $\mathcal{A}_1$,对应的样本数据是 $n\times p_1$ 的矩阵 $X_{\mathcal{A}_1}$。

第二步是设 $\mathcal{A}_1^c$ 表示 $\mathcal{A}_1$ 的补集,$X_{\mathcal{A}_1^c}$ 表示余下 $p-p_1$ 个协变量的样本数据构成的 $n\times(p-p_1)$ 维的矩阵。定义 $X_{new}=(I_n-X_{\mathcal{A}_1}(X_{\mathcal{A}_1}^TX_{\mathcal{A}_1})^{-1}X_{\mathcal{A}_1}^T)X_{\mathcal{A}_1^c}$,应用 RED-SIS 对新数据 X_{new} 和 Y 进行变量筛选,得到 $p_2<N=[n/\log(n)]$ 个协变量,其集合为: $\mathcal{A}_2$。更新总的被选中变量集合为 $\mathcal{A}_1\cup\mathcal{A}_2$。

第三步是重复第二步,直到找到一个 k,使得所选择变量之和 $p_1+p_1+\cdots+p_k=[n/\log(n)]$。最终选择的变量集合为 $\mathcal{A}_1\cup\mathcal{A}_2\cup\cdots\cup\mathcal{A}_k$。

迭代的 RED-SIS 变量筛选方法记为 RED-ISIS。类似于 Zhong 和 Zhu(2015) ,采取两步迭代,且 $p_1=5$,$p_2=[n/\log(n)]-p_1$。Zhong 和 Zhu(2015) 指出从理论上进行论证是富有挑战的工作。因此,本文通过数值模拟来说明这种方法的有效性。

例 3. 迭代特征筛选方法 RED-ISIS

借鉴 Zhong 和 Zhu(2015) ,考虑 $Y^*=5X_1+5X_2+5X_3-15\sqrt{\rho}X_4+\varepsilon$,其中协变量 $X=(X_1,X_2,\cdots,X_p)^T$ 是服从多元正态分布 $N(0_p,\Sigma)$ 的随机向量,0_p 是 p 维零向量,$\Sigma=(\sigma_{ij})_{p\times p}$ 为协方差矩阵,满足: (1) $\sigma_{ii}=1,i=1,2,\cdots,p$; (2) $\sigma_{4i}=\sigma_{i4}=\sqrt{\rho},i=1,2,\cdots,p,i\neq4$; (3) $\sigma_{ij}=\rho,i\neq j,i\neq4,j\neq$

4。$\varepsilon \sim N(0,1)$ 为误差项。按照如下规则离散化: $Y=1, Y^* \in (-\infty,-3)$; $Y=2, Y^* \in (-3,0)$; $Y=3, Y^* \in (0,3)$; $Y=4, Y^* \in (3,+\infty)$。

该模型中, Y 依赖于协变量 X_1, X_2, X_3 和 X_4, 同时容易计算得 $Cov(Y^*, X_4)=0$。因此, Y^* 与 X_4 是边际独立的, 进而 Y 与 X_4 是边际独立的。取 $n=200, p=2000$, 考虑如下两种情况: $\rho=0.5$ 和 $\rho=0.8$。重复模拟 500 次, 分别应用 RED-SIS 和 RED-ISIS 进行变量筛选, 结果见表 5。

表 5　　两种方法 RED-SIS 和 RED-ISIS 的变量筛选结果

方法	$\rho=0.5$					$\rho=0.8$				
	P_1	P_2	P_3	P_4	P_a	P_1	P_2	P_3	P_4	P_a
RED-SIS	0.984	0.968	0.982	0.000	0.000	0.766	0.778	0.790	0.000	0.000
RED-ISIS	0.992	0.984	0.974	0.966	0.920	0.944	0.938	0.952	0.922	0.760

表 5 结果表明, RED-SIS 无法将重要变量 X_4 筛选出来, 随着 ρ 从 0.5 增加到 0.8, X_1, X_2, X_3 被正确筛选的概率都有显著的下降。迭代特征筛选方法 RED-ISIS 有效的改进了这一劣势。在 $\rho=0.8$ 时, 对 X_1, X_2, X_3 和 X_4 的筛选效果都得到明显提升。另外, 对于 $\rho=0.5$ 和 $\rho=0.8$, 变量 X_4 被正确筛选的概率由 0.000 和 0.000, 分别提高到了 0.966 和 0.922。

五、研究结果与讨论

本文研究了超高维判别分析的特征筛选问题, 并基于秩能量距离提出了一种新的特征筛选方法 RED-SIS。证明了该方法在较为宽松的正则条件下满足确定筛选性质和排序的相合性。由于 RED-SIS 无需特定模型结构的假设, 也没有相关的矩条件限制, 故可以适用于线性和非线性判别分析的特征筛选, 同时对厚尾数据具有较好的稳健性。有限样本数值模拟研究表明, 相对于现有的几种典型超高维判别分析变量筛选方法 AD-SIS、MV-SIS、FKF、RRS 和 RED-SIS 具有较好的筛选效果, 特别是对尖峰、厚尾以及偏态数据的特征筛选。另外, 通过对人类肺癌数据集进行实证分析, 结果表明新的特征筛选方法 RED-SIS 具有较好的效果。最后, 构建了基于 RED-SIS 的迭代特征筛选方法 RED-ISIS, 模拟结果表明, RED-ISIS 能有效解决协变量相关性所带来的影响。

参考文献

[1] Bhattacharjee A, Richards W G, Staunton J, et al. Classification of Human Lung Carcinomas by Mrna Expression Profiling Reveals Distinct Adenocarcinoma Subclasses [J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(24): 13790-13795.

[2] Chang J Y, Tang C Y, Wu Y. Marginal Empirical Likelihood and Sure Independence Feature Screening [J]. The Annals of Statistics, 2013, 41(4): 2123-2148.

[3] Cheng G S, Li X X, Lai P, Song F L, Yu J. Robust Rank Screening for Ultrahigh Dimensional Discriminant Analysis [J]. Statistics and Computing, 2017, 27(2): 535-545.

[4] Cui H J, Li R Z, Zhong W. Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis [J]. Journal of the American Statistical Association, 2015, 110(510): 630-641.

[5] Curry J, Dang X, Sang H L. A Rank-Based Cramer-von-Mises-Type Test for Two Samples [J]. 2018, arXiv: 1802.06332v2.

[6] Fan J Q, Feng Y Y, Song R. Nonparametric Independence Screening in Sparse Ultra-High Dimensional Additive Models [J]. Journal of the American Statistical Association, 2009, 106(494): 544-557.

[7] Fan J Q, Lv J C. Sure Independence Screening for Ultrahigh Dimensional Feature Space [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008, 70(5): 849-911.

[8] Fan J Q, Song R. Sure Independence Screening in Generalized Linear Models with NP-Dimensionality [J]. The Annals of Statistics, 2010, 38(6): 3567-3604.

[9] He S M, Ma S G, Xu W L. A Modified Mean-Variance Feature-Screening Procedure for Ultrahigh-Dimensional Discriminant Analysis

[J]. Computational Statistics and Data Analysis, 2019, 137: 155-169.
[10] Li G R, Peng H, Zhang J, Zhu L X. Robust Rank Correlation Based Screening [J]. The Annals of Statistics, 2012, 40(3): 1846-1877.
[11] Li R Z, Zhong W, Zhu L P. Feature Screening via Distance Correlation Learning [J]. Journal of the American Statistical Association, 2012, 107(499): 1129-1139.
[12] Mai Q, Zou H. The Fused Kolmogorov Filter: A Nonparametric Model-Free Screening Method [J]. The Annals of Statistics, 2015, 43(4): 1471-1497.
[13] Mai Q, Zou H. The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification [J]. Biometrika, 2013, 1(1): 229-234.
[14] Pan W L, Wang X Q, Xiao W N, Zhu H T. A Generic Sure Independence Screening Procedure [J]. Journal of the American Statistical Association, 2019, 114(526): 928-937.
[15] Székely G J, Rizzo M L. The Energy of Data [J]. Annual Review of Statistics and Its Application, 2017, 4(1): 447-479.
[16] Wang H S. Forward Regression for Ultra-high Dimensional Variable Screening [J]. Journal of the American Statistical Association, 104: 1512-1524.
[17] Xie J H, Lin Y Y, Yan X D, Tang N S. Category-Adaptive Variable Screening for Ultra-High Dimensional Heterogeneous Categorical Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1573734.
[18] Zhong W, Zhu L P. An Iterative Approach to Distance Correlation-Based Sure Independence Screening [J]. Journal of Statistical Computation and Simulation, 2015, 85(11): 2331-2345.
[19] Zhou T Y, Zhu L P, X C, Li R Z. Model-Free Forward Screening via Cumulative Divergence [J]. Journal of the American Statistical Association, 2019, DOI: 10.1080/01621459.2019.1632078.
[20] Zhu J. Classification of Gene Microarrays by Penalized Logistic Regression [J]. Biostatistics, 2004, 5(3): 427-443.
[21] Zhu L P, Li L X, Li R Z, Zhu L X. Model-Free Feature Screening for Ultrahigh-Dimensional Data [J]. Journal of the American Statistical Association, 2011, 106(496): 1464-1475.

作者简介

何胜美,广东金融学院金融数学与统计学院讲师。研究方向为高维数据分析、机器学习等。

李高荣,北京师范大学统计学院教授、博士生导师。研究方向为非参数统计、高维统计、统计学习、纵向数据分析、测量误差和因果推断等。

许王莉(通讯作者),中国人民大学应用统计科学研究中心,中国人民大学统计学院教授、博士生导师。研究方向为模型检验、缺失数据分析、纵向数据分析、高维数据分析、抽样设计等。电子邮箱: xwlbnu@163.com。

(责任编辑:董 倩)