第53卷 第1期
2023年1月

中国海洋大学学报
PERIODICAL OF OCEAN UNIVERSITY OF CHINA

53(1):147~156
Jan., 2023

超高维缺失响应数据的特征筛选*

邹丽英，刘 祎**

（中国海洋大学数学科学学院，山东 青岛 266100）

摘 要： 本文提出了一种解决超高维缺失响应数据的特征筛选的新方法。首先，通过插补技术，补全缺失响应变量值，构造插补响应变量与协变量分布函数之间的距离相关系数，以此作为筛选指标进行特征筛选。所提出的筛选方法具有以下优势：第一，它是一个非参数无模型假设的方法，可以处理变量间的非线性关系；第二，对协变量异常值稳健；第三，可以直接处理多维响应变量情形。然后，通过数值模拟展示了所提方法的性能与表现，并与现有的筛选方法进行了比较。最后，将所提方法应用于弥漫性大B细胞淋巴瘤的数据分析，分析结果表明基于该方法拟合后的模型具有更好的风险分离效果。

关键词： 超高维数据；随机缺失；特征筛选；稳健距离相关；无模型假设

中图法分类号： O212 **文献标志码：** A **文章编号：** 1672-5174(2023)01-147-10

DOI： 10.16441/j.cnki.hdxb.20210350

引用格式： 邹丽英，刘祎. 超高维缺失响应数据的特征筛选[J]. 中国海洋大学学报(自然科学版)，2023，53(1)：147-156.
Zou Liying，Liu Yi. Feature screening for ultra-high dimensional missing response[J]. Periodical of Ocean University of China，2023，53(1)：147-156.

随着科技水平的飞速发展和数据收集能力的大幅提升，超高维数据已经越来越频繁地出现在包括金融学、基因学、医学等各领域中。在高维数据中，为了选出对响应变量有重要影响的少数预测变量，许多惩罚方法被提出，例如LASSO[1]、SCAD[2]、自适应LASSO[3]和Dantzing[4]等。但在超高维数据中，即当变量个数相对于样本量呈指数增长时，这些方法面临着计算复杂性、统计准确性和算法稳定性的共同挑战[5]。为此许多学者提出超高维数据的特征筛选方法，用于解决面临的困难。Fan和Lv[6]基于线性模型提出了一种依赖于边际皮尔逊相关系数的确定独立筛选(Sure independent screening，SIS)，它可以将超高维数据降维至合适的大小，并且通过对相关系数绝对值大小进行排序，将对响应变量有重要影响的预测因子选中的概率趋于1。Zhu等[7]用一种针对一般多指标模型确定独立排序和筛选技术(Sure independent ranking and screening，SIRS)来排序重要变量。Li等[8]基于距离相关系数发展了SIS程序，但它对重尾数据不具有稳健性。Zhong等[9]在稳健距离相关的基础上提出了一种应用于单指标模型的特征筛选方法。然而上述特征筛选方法都是针对完全可观测数据的。

在实际案例中，例如成本效益分析、教育调查、全基因组关联研究以及基因表达研究等领域，由于部分受试者不愿意回答敏感问题，或者不可控因素导致的信息丢失，出现响应变量随机缺失(Missing at random，MAR)的情况是较为常见的。有关于缺失数据的统计分析，已有许多文献进行研究，例如：Wang和Rao[10]提出了处理缺失响应问题的经验似然方法。Qin、Shao和Zhang[11]将逆概率加权方法用于处理协变量相关的缺失响应数据。Hu、Follmann和Qin[12]研究了缺失数据平均响应的半参数降维估计。近年来，一系列研究开始集中于处理缺失响应变量的超高维数据，并提出高效、准确的特征筛选方法。Lai等[13]调整了Zhu等[7]的SIRS方法并结合逆概率加权技术，提出了一种无模型特征筛选方法。逆概率加权方法对缺失概率较敏感，因此非参数插补作为一种处理缺失数据的方法得到广泛应用。Fang[14]基于非参数插补技术提出了无模型的特征筛选方法，并说明其比逆概率加权方法具有更好的筛选效果。

本文提出了一种新的特征筛选程序(Imputed distance correlation，IDC)，构造插补响应变量与协变量分布函数之间的距离相关系数作为筛选指标进行特征筛选。所提方法不依赖于模型假设且对协变量异常值稳健，而且可以直接处理响应变量是多维的情形。通过

* 基金项目：国家自然科学基金项目(11801567)资助
Supported by the National Nature Science Foundation of China(11801567)
收稿日期：2021-09-27；修订日期：2021-11-18
作者简介：邹丽英(1996—)，女，硕士生。E-mail：zouly@stu.ouc.edu.cn
** 通讯作者：E-mail：liuyi@ouc.edu.cn

数值模拟和微阵列弥漫性大 B 细胞淋巴瘤(Diffuse Large-B-Cell Lymphoma,DLBCL)数据分析,对方法的有限样本性质进行了验证。

1 方法

记 Y 为连续的响应变量,$X=(X_1,\cdots,X_p)^{\mathrm{T}}$ 为 $p\times 1$ 维的连续预测变量。假设维数 p 相对于样本容量 n 呈指数增长,即 $\log(p)=O(n^{\alpha})$,常数 $\alpha>0$。其中 X 是完全可观测的,而 Y 可能缺失。由于观测数据是不完全的,记(X_i,Y_i,δ_i)为样本数据,$i=1,\cdots,n$,其中δ_i为缺失响应的指示变量,即如果Y_i缺失,则$\delta_i=0$,如果Y_i可观测,则$\delta_i=1$。我们假设 δ 只依赖于 X 使得倾向得分函数有 $\pi(X)=P(\delta=1|X)$的形式。由 Little 和 Rubin[15]可知,上述关于缺失机制的定义为随机缺失。Y 是随机缺失的,简单来说,就是假设

$$P(\delta=1|X,Y)=P(\delta=1|X)。$$

基于稀疏假设,只有少数的预测变量与 Y 有关。我们定义活跃预测变量的索引集:

$\mathcal{A}=\{k:F(y|X)$依赖于X_k,对某个 $y\in\psi_Y$,$k=1,\cdots,p\}$,其中 $F(y|X)=P(Y\leqslant y|X)$为给定变量 X 条件下变量 Y 的条件分布函数,ψ_Y为 Y 的支撑集,且$|\mathcal{A}|$为集合 $\mathcal{A}$ 的势。在超高维数据分析中我们假设 $p\gg n$ 且 $p\gg|\mathcal{A}|$。记 $\mathcal{A}^c=\{1,2,\cdots,p\}\setminus\mathcal{A}$ 为非活跃预测变量的索引集,即 $\mathcal{A}^c=\{k:F(y|X)$不依赖于 X_k,对 $\forall y\in\psi_Y,k=1,\cdots,p\}$。令$X_{\mathcal{A}}=\{X_k:k\in\mathcal{A}\}$且 $X_{\mathcal{A}^c}=\{X_k:k\in\mathcal{A}^c\}$分别为预测变量的活跃集和非活跃集。

首先回顾一下距离相关系数[16]的定义。假设 U 和 V 为两个随机向量,维度分别为d_U和d_V。令 $\varphi_U(u)$ 和 $\varphi_V(v)$分别为 U 和 V 的特征函数,$\varphi_{U,V}(u,v)$为 U 和 V 的联合特征函数。距离协方差(Distance covariance)定义为非负数dcov(U,V),即

$$\mathrm{dcov}^2(U,V)=\int_{R^{d_U+d_V}}\|\varphi_{U,V}(u,v)-\varphi_U(u)\varphi_V(v)\|^2 w(u,v)\,\mathrm{d}u\mathrm{d}v,$$

其中 $w(u,v)=\{C_{d_U}C_{d_V}\|u\|_{d_U}^{1+d_U}\|v\|_{d_V}^{1+d_V}\}^{-1}$且 $C_d=\pi^{(1+d)/2}/\Gamma(1+d)$。

U 和 V 的距离相关系数(Distance correlation,DC)定义为

$$\mathrm{dcorr}(U,V)=\frac{\mathrm{dcov}(U,V)}{\sqrt{\mathrm{dcov}(U,U)\,\mathrm{dcov}(V,V)}}。$$

DC 作为相关关系的一种度量,具有良好的特性,即 $\mathrm{dcorr}(U,V)=0$ 当且仅当 U 和 V 是相互独立的。这一性质使得距离相关尤其适用于超高维数据的变量筛选。并且对于超高维的完整数据集,Zhong 等[9]在稳健距离相关的基础上提出了一种应用于单指标模型的特征筛选程序,这启发我们将稳健的距离相关应用于具有缺失响应的超高维数据中。

本文采用的稳健距离相关筛选指标是指

$$\omega_k=\mathrm{dcorr}^2(F_k(X_k),Y)=\frac{\mathrm{dcov}^2(F_k(X_k),Y)}{\mathrm{dcov}(F_k(X_k),F_k(X_k))\,\mathrm{dcov}(Y,Y)},$$

其中$F_k(x)=P(X_k\leqslant x)$。在样本中,我们可以通过矩估计方法来估计$F_k(x)$,即 $\hat{F}_{nk}(x)=n^{-1}\sum_{i=1}^{n}I(X_{ki}\leqslant x)$。然而,当响应变量 Y 随机缺失时,由于响应变量 Y 不能完全观测,使得上述提出的筛选指标不再适用。为解决这一问题,我们采用插补技术得到完整的数据集。

处理缺失数据的插补方法是通过非参数回归得到的,设 $m(x)=E(Y|X=x)$为给定 X 时 Y 的回归函数,则由核回归[17-18]可得到 $m(x)$的估计为

$$\hat{m}(x)=\frac{\sum_{j=1}^{n}\delta_j Y_j K_b(X_j,x)}{\sum_{j=1}^{n}\delta_j K_b(X_j,x)},$$

其中 $K_b(X_j,x)=b^{-1}K((X_j-x)/b)$,$K(\cdot)$为高斯核函数;$b=b(n)$为当 n 趋于∞时向 0 减小的带宽序列。由上述公式得到的$\hat{m}(X_i)$对缺失的Y_i进行插补。定义插补的响应变量$\hat{Y}$为

$$\hat{Y}_{ki}=\begin{cases}\dfrac{\sum_{j=1}^{n}\delta_j Y_j K_b(X_{kj},X_{ki})}{\sum_{j=1}^{n}\delta_j K_b(X_{kj},X_{ki})}, & \text{当 } \delta_i=0 \text{ 时}\\ Y_i, & \text{当 } \delta_i=1 \text{ 时}\end{cases}。$$

上式等价于 $\hat{Y}_{ki}=\delta_i Y_i+(1-\delta_i)\hat{m}_k(X_{ki})$,其中

$$\hat{m}_k(X_{ki})=\frac{\sum_{j=1}^{n}\delta_j Y_j K_b(X_{kj},X_{ki})}{\sum_{j=1}^{n}\delta_j K_b(X_{kj},X_{ki})},k=1,\cdots,p \text{ 且 } i,j=1,\cdots,n。$$

根据插补后的数据集,我们得到筛选指标的估计为

$$\hat{\omega}_k=\mathrm{dcorr}^2(\hat{F}_{nk}(X_{ki}),\hat{Y}_{ki})=\frac{\mathrm{dcov}^2(\hat{F}_{nk}(X_{ki}),\hat{Y}_{ki})}{\mathrm{dcov}(\hat{F}_{nk}(X_{ki}),\hat{F}_{nk}(X_{ki}))\,\mathrm{dcov}(\hat{Y}_{ki},\hat{Y}_{ki})}。$$

其中 $\mathrm{dcov}^2(\hat{F}_{nk}(X_{ki}),\hat{Y}_{ki})=\hat{S}_{k1}+\hat{S}_{k2}-2\hat{S}_{k3}$,$\hat{S}_{k1}=n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}|\hat{F}_{nk}(X_{ki})-\hat{F}_{nk}(X_{kj})||\hat{Y}_{ki}-\hat{Y}_{kj}|$,$\hat{S}_{k2}=n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}|\hat{F}_{nk}(X_{ki})-\hat{F}_{nk}(X_{kj})|\,n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}|\hat{Y}_{ki}-\hat{Y}_{kj}|$,$\hat{S}_{k3}=n^{-3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|\hat{F}_{nk}(X_{ki})-\hat{F}_{nk}(X_{kj})|\cdot|\hat{Y}_{ki}-\hat{Y}_{kl}|$;$\mathrm{dcov}(\hat{F}_{nk}(X_{ki}),\hat{F}_{nk}(X_{ki}))$和 $\mathrm{dcov}(\hat{Y}_{ki},$

$\hat{Y}_{ki}$)可类似计算。

根据距离相关系数的性质，当预测因子 X_k 对响应变量 Y 无重要影响时，$\hat{\omega}_k$ 应该很小，当 X_k 对响应变量 Y 有重要影响时，$\hat{\omega}_k$ 越大。因此提出活跃因子指标集为

$$\hat{\mathcal{A}}=\{k:\hat{\omega}_k\geqslant c\,n^{-\alpha},k=1,\cdots,p\}。$$

为了更好地选择 c 和 α，我们可以将 $\hat{\omega}_k$ 从大到小排序，选择第 d_n 个最大值，其中 $d_n=[n/\log n]$ 为常数。

2　数值模拟

在本节中我们将本文提出的插补距离相关方法(Imputed distance correlation，IDC)与其他处理超高维缺失响应方法作比较。对比方法有：Lai 等[13]提出的基于逆概率加权的超高维特征筛选方法(Inverse probability weighted，IPW)，Fang[14]提出的基于插补技术的非参数特征筛选方法(Method of Imputation Technique，ITM)以及本文方法在完整数据下的情形(Distance correlation of full sample data，FDC)。在整个模拟研究中，我们将倾向得分函数 $P(\delta=1\mid X)$ 设置为逻辑回归模型，即 $P(\delta=1|X)=\exp(\theta X)/(1+\exp(\theta X))$，通过改变 θ 来改变缺失率(Missing rate，MR)，设置缺失率大约为 0.2 和 0.4。对于每种情形，设置 200 次重复。

为了评估所提出方法的性能，我们采用 Li 等[8]的三个评估准则。第一个评估准则是包括所有活跃的预测变量的最小模型数，用 S 表示。我们给出了 200 次重复模拟中 S 的 5%、25%、50%、75%和 95%分位数。第二个评估准则是每个活跃预测变量的覆盖率，用 P_i 表示，我们给出了当给定模型大小为 d_n 时 200 次重复模拟中 X_i 被选择的比例。第三个评估准则是所有活跃预测变量的覆盖率，用 P_a 表示。我们给出了在 200 次重复模拟中，对于给定模型大小 d_n，所有活跃预测变量均被选择的比例。我们选择 d_n 为 $d_n=[n/\log n]$，其中 $[x]$ 表示 x 的整数部分。

实例 1(线性模型)　考虑线性模型如下：

$$Y=\boldsymbol{X}^{\mathrm{T}}\beta+\varepsilon。$$

其中 $\beta=(1.5,1.8,2.1,2.4,0,\cdots,0)^{\mathrm{T}}\in\mathbf{R}^p$，即只有前四个变量是活跃的。$\boldsymbol{X}_{p\times n}=(X_1,\cdots,X_p)^{\mathrm{T}}$ 产生于均值为 0，协方差矩阵为 $\boldsymbol{\Sigma}=(\rho^{|i-j|})$ 的多元正态分布，其中 $i,j=1,\cdots,p$，ρ 分别取为 0.5 和 0.8，ε 服从标准柯西分布 $C(1,0)$。为获得重尾预测变量，我们将 X_3 替换为服从自由度为 2 的 t 分布产生的随机样本。核函数为高斯核函数，带宽为 $h=1.06\,\hat{\sigma}_X\,n^{-1/5}$，其中 $\hat{\sigma}_X$ 是 X_k 的样本标准差。在此情形下，我们设置 $(\theta_1,\theta_2,\theta_p)=(0.5,0.3,0.1)$ 和 $(3,0.5,6)$ 得到约 20%和 40%的缺失率，选取样本量 n 和协变量个数 p 为 $(n,p)=(200,2\,000)$ 和 $(200,1\,000)$，模拟结果如表 1 和 2 所示。

表 1　线性模型中单个活跃因子被选中的概率 P_s 及所有活跃因子被选中的概率 P_a

Table 1　In linear models selection proportions P_s for each active predictor and P_a for all active predictors

ρ	缺失率 Missing rate	方法 Method	$n=200,p=2\,000$					$n=200,p=1\,000$				
			P_1	P_2	P_3	P_4	P_a	P_1	P_2	P_3	P_4	P_a
0.5	0.2	IDC	0.825	0.905	0.945	0.925	0.775	0.855	0.935	0.925	0.865	0.775
		ITM	0.435	0.480	0.595	0.490	0.320	0.405	0.460	0.580	0.485	0.300
		IPW	0.480	0.580	0.680	0.575	0.300	0.460	0.540	0.665	0.535	0.280
	0.4	IDC	0.670	0.720	0.715	0.750	0.465	0.680	0.745	0.800	0.735	0.475
		ITM	0.450	0.495	0.540	0.490	0.280	0.425	0.515	0.635	0.505	0.280
		IPW	0.430	0.485	0.570	0.445	0.195	0.390	0.510	0.655	0.550	0.240
	0	FDC	0.990	0.995	1.000	0.995	0.990	0.990	0.995	1.000	0.985	0.985
0.8	0.2	IDC	0.975	0.975	0.905	0.970	0.890	0.975	0.980	0.905	0.970	0.880
		ITM	0.540	0.585	0.575	0.585	0.415	0.560	0.580	0.550	0.575	0.445
		IPW	0.605	0.695	0.610	0.665	0.425	0.645	0.680	0.645	0.680	0.470
	0.4	IDC	0.860	0.885	0.665	0.875	0.595	0.855	0.880	0.695	0.870	0.625
		ITM	0.570	0.590	0.575	0.595	0.435	0.585	0.610	0.565	0.615	0.450
		IPW	0.575	0.645	0.625	0.615	0.420	0.620	0.640	0.605	0.630	0.415
	0	FDC	1.000	1.000	0.995	1.000	0.995	1.000	1.000	0.995	1.000	0.995

从表 1 和 2 可以看出，在线性模型假设，以及协变量与响应变量都存在重尾数据下，本文提出的 IDC 方法相比于 IPW 和 ITM 两种方法而言，具有较为明显的优势。具体表现在活跃预测变量入选模型的比例更高，所需要的最小模型数更小。说明 IDC 方法可以将重要变量排在不重要变量的前面。完整数据下应用本文的 FDC 方法表现最好，因为其利用了样本的所有信息。

实例 2(非线性模型 1)　考虑非线性模型如下：

$$Y=\beta_1 X_1+\beta_2 (X_2)^{\frac{2}{3}}+\beta_3 \sin(X_3)+\varepsilon。$$

其中 $\beta=(\beta_1,\beta_2,\beta_3,0,\cdots,0)^{\mathrm{T}}=(2,2.5,3,0,\cdots,0)^{\mathrm{T}}\in \mathbf{R}^p$，即只有前三个变量是活跃的，$\boldsymbol{X}=(X_1,\cdots,X_p)^{\mathrm{T}}$ 产生于均值为 0，协方差矩阵为 $\boldsymbol{\Sigma}$ 的多元正态分布，ε 服从 $N(0,1)$。另外为进一步获得重尾预测变量，我们类似于 Lai 等[13]的设置，替换 X_1 和 X_3 为服从自由度为 3 的 t 分布产生的随机样本，替换 X_2 为服从 $t(3)+1$ 分布产生的随机样本，核函数与实例 1 相同。在此情形下，我们设置$(\theta_1,\theta_2,\theta_p)=(1,0.3,0.1)$和$(3,-0.5,2)$得到约 20%和 40%的缺失率，选取$(n,p)=(100,1\ 000)$和$(200,2\ 000)$，模拟结果如表 3 和 4 所示。

从表 3 和 4 可以看出，IDC 方法对非线性关系的检测能力要明显优于 ITM 和 IPW 方法。这是由于不同于 ITM 和 IPW 筛选指标的线性表达形式，IDC 基于稳健距离相关指标，更适用于非线性模型的特征筛选。而且随着样本量的增大，方法的表现有明显改善。

表 2 线性模型中最小模型数 S 的各分位数

Table 2 In linear models the different quantiles of the minimum model size S

ρ	缺失率 Missing rate	方法 Method	$n=200,p=2\ 000$					$n=200,p=1\ 000$				
			5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.5	0.2	IDC	4.00	4.00	4.50	25.50	564.00	4.00	4.00	5.00	27.75	337.30
		ITM	4.00	16.25	447.50	1 449.25	1 956.15	4.00	16.50	368.00	789.75	982.10
		IPW	4.00	20.75	303.00	1 330.50	1 926.55	4.00	23.75	302.50	811.00	992.00
	0.4	IDC	4.00	8.00	56.00	598.25	1 553.95	4.00	8.00	42.50	360.50	783.20
		ITM	4.00	27.25	346.50	1 288.00	1 950.40	4.00	20.75	194.00	689.75	972.05
		IPW	5.00	79.75	646.00	1 563.75	1 968.30	4.00	46.00	393.00	794.50	948.40
	0	FDC	4.00	4.00	4.00	4.00	5.00	4.00	4.00	4.00	4.00	5.00
0.8	0.2	IDC	4.00	4.00	5.00	7.00	342.40	4.00	4.00	5.00	9.00	149.90
		ITM	4.00	5.00	107.00	1 135.75	1 916.55	4.00	6.00	84.50	587.25	983.20
		IPW	4.00	5.00	79.00	853.25	1 931.00	4.00	6.75	50.50	701.75	996.05
	0.4	IDC	4.00	5.00	13.00	141.75	1 040.10	4.00	5.00	15.00	182.00	682.90
		ITM	4.00	5.00	99.00	912.50	1 956.10	4.00	6.00	77.50	554.00	943.25
		IPW	4.00	7.00	118.00	889.00	1 901.20	4.00	7.00	108.50	656.00	995.05
	0	FDC	4.00	4.00	5.00	5.00	6.00	4.00	4.00	5.00	6.00	7.00

表 3 非线性模型 1 中单个活跃因子被选中的概率 P_s 及所有活跃因子被选中的概率 P_a

Table 3 In nonlinear models 1 selection proportions P_s for each active predictor and P_a for all active predictors

ρ	缺失率 Missing rate	方法 Method	$n=100,p=1\ 000$				$n=200,p=2\ 000$			
			P_1	P_2	P_3	P_a	P_1	P_2	P_3	P_a
0.5	0.2	IDC	1.000	0.995	0.920	0.915	1.000	1.000	1.000	1.000
		ITM	1.000	1.000	0.850	0.850	1.000	1.000	0.995	0.995
		IPW	1.000	1.000	0.855	0.855	1.000	0.995	0.985	0.985
	0.4	IDC	0.990	0.990	0.665	0.650	1.000	1.000	0.975	0.975
		ITM	1.000	0.950	0.600	0.575	1.000	1.000	0.930	0.930
		IPW	1.000	0.945	0.610	0.580	1.000	0.995	0.920	0.915
	0	FDC	1.000	1.000	0.940	0.940	1.000	1.000	1.000	1.000
0.8	0.2	IDC	1.000	0.995	0.905	0.900	1.000	1.000	1.000	1.000
		ITM	1.000	1.000	0.860	0.860	1.000	1.000	0.990	0.990
		IPW	1.000	1.000	0.830	0.830	1.000	0.995	0.985	0.985
	0.4	IDC	0.990	0.995	0.695	0.690	1.000	1.000	0.960	0.960
		ITM	1.000	0.940	0.645	0.595	1.000	1.000	0.925	0.925
		IPW	1.000	0.925	0.645	0.600	1.000	0.990	0.890	0.880
	0	FDC	1.000	1.000	0.935	1.000	1.000	1.000	1.000	1.000

表 4　非线性模型 1 中最小模型数 S 的各分位数

Table 4　In nonlinear models 1 the different quantiles of the minimum model size S

ρ	缺失率 Missing rate	方法 Method	$n=100, p=1\,000$					$n=200, p=1\,000$				
			5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.5	0.2	IDC	3.00	3.00	4.00	8.00	35.15	3.00	3.00	3.00	3.00	4.00
		ITM	3.00	3.00	4.00	11.25	84.30	3.00	3.00	3.00	3.00	5.00
		IPW	3.00	3.00	4.00	11.00	73.80	3.00	3.00	3.00	3.00	4.00
	0.4	IDC	3.00	4.00	9.00	32.00	120.05	3.00	3.00	3.00	4.00	23.00
		ITM	3.00	5.00	15.00	53.25	183.40	3.00	3.00	3.00	7.00	67.00
		IPW	3.00	4.00	14.50	43.25	262.10	3.00	3.00	3.00	8.00	100.25
	0	FDC	3.00	3.00	3.00	5.00	26.10	3.00	3.00	3.00	3.00	3.00
0.8	0.2	IDC	3.00	3.00	4.00	9.00	46.25	3.00	3.00	3.00	3.00	4.00
		ITM	3.00	3.00	4.00	11.00	83.10	3.00	3.00	3.00	3.00	6.05
		IPW	3.00	3.00	3.00	10.00	91.00	3.00	3.00	3.00	3.00	5.05
	0.4	IDC	3.00	4.00	10.00	32.00	158.00	3.00	3.00	3.00	4.00	24.10
		ITM	3.00	4.00	12.00	60.25	255.00	3.00	3.00	3.00	7.00	56.05
		IPW	3.00	4.00	13.00	71.50	391.15	3.00	3.00	3.00	10.00	209.85
	0	FDC	3.00	3.00	3.00	4.00	26.10	3.00	3.00	3.00	3.00	3.00

实例 3(非线性模型 2)　为了进一步验证方法在非线性模型上的性能，考虑非线性模型如下：

$$Y=\frac{1}{2}(\beta_{11}X_2 I(X_1<0)+\beta_{12}X_2+\beta_{13}X_3)^2+\cos(\beta_{21}X_1+\beta_{22}X_2+\beta_{23}\tanh X_3)+\varepsilon,$$

其中$(\beta_{11},\beta_{12},\beta_{13},\beta_{21},\beta_{22},\beta_{23})=(0.8,1,1.2,1.6,1.4,1.2)$，即只有前三个变量是活跃的，其余设置与实例 2 中的设置相同。在此情形下，我们选取样本量 n 和协变量个数 p 为$(n,p)=(100,1\,000)$和$(200,1\,000)$，模拟结果见表 5 和 6。

表 5　非线性模型 2 中单个活跃因子被选中的概率 P_s 及所有活跃因子被选中的概率 P_a

Table 5　In nonlinear models 2 selection proportions P_s for each active predictor and P_a for all active predictors

ρ	缺失率 Missing rate	方法 Method	$n=100, p=1\,000$				$n=200, p=1\,000$			
			P_1	P_2	P_3	P_a	P_1	P_2	P_3	P_a
0.5	0.2	IDC	0.535	0.770	0.630	0.300	0.870	0.985	0.925	0.795
		ITM	0.000	0.810	0.480	0.000	0.005	0.925	0.740	0.000
		IPW	0.035	0.810	0.505	0.015	0.165	0.965	0.740	0.140
	0.4	IDC	0.770	0.485	0.130	0.055	0.655	0.695	0.580	0.340
		ITM	0.040	0.245	0.175	0.000	0.385	0.240	0.490	0.090
		IPW	0.020	0.310	0.135	0.000	0.140	0.310	0.490	0.025
	0	FDC	0.835	1.000	0.365	0.305	0.695	1.000	1.000	0.695
0.8	0.2	IDC	0.875	0.930	0.240	0.210	0.865	0.980	0.925	0.795
		ITM	0.000	0.915	0.285	0.000	0.005	0.925	0.725	0.005
		IPW	0.320	0.930	0.235	0.080	0.190	0.950	0.745	0.145
	0.4	IDC	0.675	0.440	0.135	0.050	0.640	0.725	0.550	0.325
		ITM	0.045	0.265	0.220	0.005	0.410	0.250	0.520	0.105
		IPW	0.010	0.285	0.180	0.000	0.125	0.350	0.510	0.040
	0	FDC	0.845	1.000	0.375	0.330	0.700	1.000	1.000	0.700

表 6 非线性模型 2 中最小模型数 S 的各分位数

Table 6 In nonlinear models 2 the different quantiles of the minimum model size S

ρ	缺失率 Missing rate	方法 Method	$n=100, p=1\,000$					$n=200, p=1\,000$				
			5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.5	0.2	IDC	4.00	17.75	58.50	175.00	633.60	3.00	3.00	8.00	27.75	168.55
		ITM	216.05	576.00	764.50	897.25	982.10	204.90	526.00	734.00	892.00	982.10
		IPW	45.85	159.00	354.50	618.75	926.10	14.95	73.75	268.00	582.25	858.95
	0.4	IDC	20.00	96.50	294.50	511.50	881.15	4.00	24.00	78.50	210.00	534.55
		ITM	119.90	389.75	579.50	823.50	982.05	18.85	120.75	373.50	732.75	944.20
		IPW	161.00	370.00	517.00	727.50	934.40	67.70	252.00	441.50	720.75	973.25
	0	FDC	3.95	15.00	59.50	206.25	637.10	3.00	4.00	11.50	53.25	374.80
0.8	0.2	IDC	3.95	30.75	124.50	379.50	791.50	3.00	3.00	8.00	27.75	168.55
		ITM	537.00	799.75	923.00	976.50	1 000.00	204.90	526.00	734.00	892.00	982.10
		IPW	18.85	107.50	298.00	580.25	897.20	14.95	73.75	268.00	582.25	858.95
	0.4	IDC	24.80	102.75	286.50	582.50	928.20	4.00	24.00	83.00	230.25	587.30
		ITM	122.30	392.75	641.50	805.50	973.00	19.95	122.50	324.00	684.00	964.20
		IPW	139.25	343.00	547.00	807.00	976.30	44.90	206.50	437.00	673.25	943.25
	0	FDC	3.00	13.75	54.50	205.50	639.55	3.00	4.00	11.00	52.25	375.60

从表 5 和 6 可以看出，在更复杂的非线性模型中，IDC 对比 ITM 和 IPW 两种方法的优势更为明显。ITM 和 IPW 两种方法的接近于，即基本不可能选出所有的活跃预测变量，这可能是由于模型中存在示性函数及双曲函数，使得非线性关系复杂，变量筛选也变得更加困难。而且可以看到，随着样本量的增大，ITM 和 IPW 两种方法并没有明显的改善。

实例 4(双响应模型) 考虑双响应变量模型如下：

$$Y_1=\beta_{11}X_1^2I(X_2<0)+\beta_{12}X_2+\beta_{13}X_3^3+\varepsilon,$$

$$Y_2=\beta_{21}X_1+\beta_{22}|X_2|+\beta_{23}\tanh X_3^2+\varepsilon。$$

其中 $\beta_1=(\beta_{11},\beta_{12},\beta_{13},0,\cdots,0)=(1,0.8,0.6,0,\cdots,0)^{\mathrm{T}}\in\mathbf{R}^p$，$\beta_2=(\beta_{21},\beta_{22},\beta_{23},0,\cdots,0)=(1,1.5,2,0,\cdots,0)^{\mathrm{T}}\in\mathbf{R}^p$，即只有前三个变量是活跃的，其余设置与实例 2 中的设置相同。另外，我们对 ITM 方法中的筛选指标采取类似于文献[13]中公式(7)的二范数处理方式，使其适用于双响应模型的筛选指标计算。在此情形下，我们选取样本量 n 和协变量个数 p 为$(n,p)=(100,1\,000)$和$(200,1\,000)$，模拟结果见表 7 和 8。

表 7 双响应模型中单个活跃因子被选中的概率 P_s 及所有活跃因子被选中的概率 P_a

Table 7 In bivariate response models selection proportions P_s for each active predictor and P_a for all active predictors

ρ	缺失率 Missing rate	方法 Method	$n=100, p=1\,000$				$n=200, p=1\,000$			
			P_1	P_2	P_3	P_a	P_1	P_2	P_3	P_a
0.5	0.2	IDC	0.915	0.990	0.940	0.845	1.000	1.000	0.985	0.985
		ITM	0.395	0.370	0.750	0.235	0.460	0.470	0.835	0.370
		IPW	0.975	0.965	0.875	0.820	1.000	1.000	0.975	0.975
	0.4	IDC	0.855	0.990	0.915	0.775	0.995	1.000	0.955	0.955
		ITM	0.605	0.050	0.645	0.050	0.600	0.165	0.750	0.115
		IPW	0.965	0.205	0.860	0.180	0.980	0.575	0.970	0.565
	0	FDC	1.000	0.995	0.985	0.980	1.000	1.000	1.000	1.000
0.8	0.2	IDC	0.905	0.990	0.945	0.840	1.000	1.000	0.985	0.985
		ITM	0.410	0.385	0.740	0.265	0.445	0.425	0.820	0.330
		IPW	0.975	0.965	0.880	0.825	1.000	1.000	0.980	0.980
	0.4	IDC	0.845	0.995	0.910	0.755	1.000	1.000	0.955	0.955
		ITM	0.625	0.095	0.660	0.070	0.590	0.140	0.790	0.115
		IPW	0.970	0.255	0.830	0.215	0.945	0.525	0.955	0.500
	0	FDC	1.000	0.995	0.990	0.985	1.000	1.000	1.000	1.000

表 8　双响应模型中最小模型数 S 的各分位数

Table 8　In bivariate response models the different quantiles of the minimum model size S

ρ	缺失率 Missing rate	方法 Method	$n=100, p=1\,000$					$n=200, p=1\,000$				
			5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.5	0.2	IDC	3.00	3.00	4.00	12.00	75.15	3.00	3.00	3.00	3.00	8.05
		ITM	3.00	26.00	233.00	570.50	941.30	3.00	8.00	161.50	476.75	862.20
		IPW	3.00	3.00	4.00	12.00	103.25	3.00	3.00	3.00	4.00	21.10
	0.4	IDC	3.00	3.00	6.00	17.00	137.75	3.00	3.00	3.00	3.00	16.05
		ITM	21.90	210.25	503.00	761.00	970.10	11.95	139.75	369.50	669.00	944.15
		IPW	4.00	31.00	123.50	324.25	695.30	3.00	8.00	30.00	107.50	475.55
	0	FDC	3.0	3.00	3.00	3.00	8.05	3.00	3.00	3.00	3.00	3.00
0.8	0.2	IDC	3.00	3.00	4.00	11.25	77.10	3.00	3.00	3.00	3.00	9.05
		ITM	3.00	19.75	243.00	597.25	952.15	3.00	9.00	166.00	494.50	860.50
		IPW	3.00	3.00	4.00	14.00	71.10	3.00	3.00	3.00	3.00	18.05
	0.4	IDC	3.00	3.00	6.00	19.25	164.20	3.00	3.00	3.00	3.00	25.15
		ITM	12.85	174.50	472.00	775.25	973.25	9.90	118.75	364.00	623.25	950.35
		IPW	4.00	32.75	132.50	398.25	780.45	3.00	7.00	37.50	144.25	588.30
	0	FDC	3.00	3.00	3.00	3.00	9.10	3.00	3.00	3.00	3.00	3.00

从表 7 和 8 可以看出，在多响应变量的非线性模型中，对比 ITM 和 IPW 两种方法，当缺失率提高时，插补技术较于逆概率加权方法表现较好，且在一定程度上也说明了 IDC 方法的指标对于复杂模型更具稳健性。

除了上述模拟，我们还验证了方法在小样本下的表现。在实例 1 的模型设置下，我们考虑了$(n, p)=(50,500)$和$(n, p)=(50,1\,000)$两种情况，模拟结果见表 9 和 10。从中可以看出各种变量筛选方法的表现趋势与大样本结果大致相同。

另外，在实例 1 线性模型$(n, p)=(100,1\,000)$的情况下，我们增加模拟了基于多重插补[19]的 IDC 方法，其中多重插补方法选取为 10 重插补(MI10-DC)，从结果可以看出，IDC 和 MI10-DC 方法都优于 IPW 和 ITM 方法，这说明稳健距离相关的指标优势。但 MI10-DC 方法在 P_a 上表现劣于 IDC 方法，结合结果来看，可能是由于 X_3 重尾分布对多重插补的抽样过程产生了影响。模拟结果见表 11。

表 9　在小样本的线性模型中单个活跃因子被选中的概率 P_s 及所有活跃因子被选中的概率 P_a

Table 9　In linear models of small sample selection proportions P_s for each active predictor and P_a for all active predictors

ρ	缺失率 Missing rate	方法 Method	$n=50, p=500$					$n=50, p=1\,000$				
			P_1	P_2	P_3	P_4	P_a	P_1	P_2	P_3	P_4	P_a
0.5	0.2	IDC	0.405	0.535	0.595	0.525	0.150	0.330	0.495	0.455	0.415	0.065
		ITM	0.245	0.350	0.450	0.355	0.120	0.200	0.255	0.360	0.235	0.060
		IPW	0.255	0.380	0.560	0.380	0.075	0.195	0.295	0.415	0.290	0.035
	0	FDC	0.620	0.750	0.805	0.725	0.365	0.455	0.660	0.720	0.610	0.170
0.8	0.2	IDC	0.685	0.770	0.480	0.740	0.270	0.690	0.795	0.385	0.715	0.230
		ITM	0.415	0.470	0.415	0.460	0.205	0.435	0.500	0.375	0.435	0.160
		IPW	0.480	0.585	0.515	0.515	0.205	0.485	0.555	0.390	0.480	0.140
	0	FDC	0.870	0.935	0.765	0.905	0.630	0.855	0.900	0.660	0.910	0.545

表 10 在小样本的线性模型中最小模型数 S 的各分位数

Table 10 In linear models of small sample the different quantiles of the minimum model size S

ρ	缺失率 Missing rate	方法 Method	$n=50, p=500$					$n=50, p=1\,000$				
			5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.5	0.2	IDC	5.95	24.25	83.50	216.25	402.70	11.95	56.75	172.00	485.25	777.50
		ITM	6.95	37.00	201.00	410.50	495.10	9.00	133.75	504.50	887.75	992.00
		IPW	7.00	50.50	210.50	389.50	479.10	16.90	162.50	486.00	842.50	980.15
	0	FDC	4.00	8.75	28.00	75.25	347.40	5.00	19.75	59.50	213.75	615.00
0.8	0.2	IDC	4.95	11.00	34.50	105.25	386.05	5.95	15.50	45.00	193.25	678.15
		ITM	4.95	17.00	87.00	363.00	497.10	5.95	22.75	147.50	777.00	983.10
		IPW	5.00	18.75	89.00	285.25	470.00	7.00	30.00	167.50	710.75	972.45
	0	FDC	4.00	5.00	8.00	29.25	196.35	4.00	5.00	10.50	43.50	334.15

表 11 在线性模型中单个活跃因子被选中的概率 P_s，所有活跃因子被选中的概率 P_a 及最小模型数 S 的各分位数

Table 11 In linear models selection proportions P_s for each active predictor, P_a for all active predictors and the different quantiles of the minimum model size S

ρ	缺失率 Missing rate	方法 Method	P_1	P_2	P_3	P_4	P_a	5%	25%	50%	75%	95%
0.5	0.2	IDC	0.620	0.770	0.745	0.745	0.430	4.00	8.00	29.50	171.00	650.90
		MI10-DC	0.865	0.835	0.515	0.510	0.320	4.95	15.75	57.00	202.50	610.10
		ITM	0.310	0.370	0.440	0.370	0.180	5.00	69.50	412.50	817.00	986.10
		IPW	0.365	0.475	0.545	0.440	0.155	5.00	50.75	429.00	832.50	978.20
	0	FDC	0.860	0.940	0.950	0.910	0.805	4.00	4.00	5.00	14.00	295.55
0.8	0.2	IDC	0.825	0.875	0.720	0.850	0.655	4.00	5.00	9.50	62.75	549.25
		MI10-DC	0.950	0.960	0.400	0.890	0.390	4.00	10.00	45.00	127.25	485.55
		ITM	0.500	0.510	0.490	0.515	0.345	4.00	7.00	202.50	688.75	968.30
		IPW	0.560	0.580	0.530	0.540	0.350	4.00	7.00	166.00	680.25	975.40
	0	FDC	0.950	0.965	0.915	0.975	0.895	4.00	4.00	5.00	7.00	142.10

3 实际案例

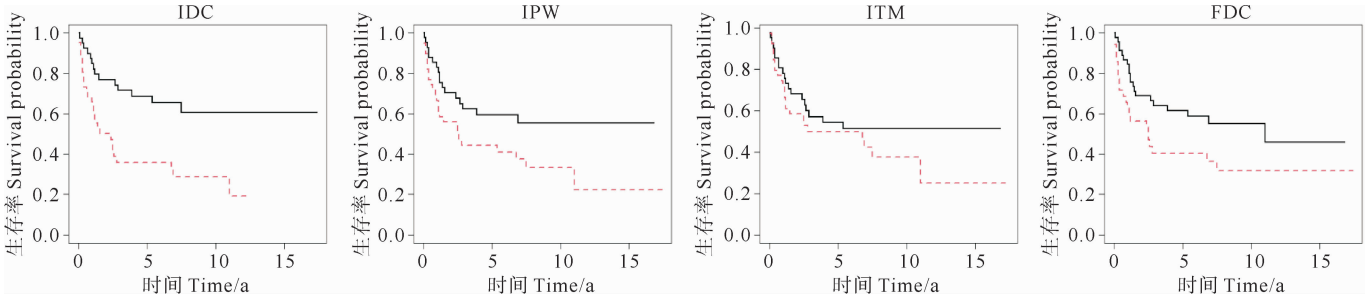
我们应用提出的方法来分析微阵列弥漫性大 B 细胞淋巴瘤(DLBCL)数据，见 Rosenwald 等[17]及 Zhu 等[7]。为了对比，我们也采用 IPW 方法、ITM 方法和 FDC 方法进行分析。DLBCL 数据集共包含 240 例患者，其中随访期间死亡 138 例。因此，相应的响应变量由观察到的生存时间和删失指标组成，从每个患者的 cDNA 微阵列中获得的 $p=7\ 399$ 个基因值是预测因子。另外，我们将数据划分为包含 $n_1=160$ 名患者的训练集和包含 $n_2=80$ 名患者的测试集。由于预测因子数量 p 远大于样本量 n，为拟合原始模型，进行特征筛选是有必要的。

由于该真实数据中不存在缺失响应，因此我们设置了一个类似于第 2 节的缺失机制。利用 FDC 方法，我们通过排序 $\hat{\omega}_k$ 得到前三个最大值为 $\hat{\omega}_{5\,778}$，$\hat{\omega}_{4\,131}$，$\hat{\omega}_{5\,777}$ 从而我们设置缺失机制的协变量为 $X_\delta=(X_{5\,778}, X_{4\,131}, X_{5\,777})^{\mathrm{T}}$。我们建立逻辑回归模型 $P(\delta=1|X_\delta, \theta_0)=\exp(\theta_0 X_\delta)/(1+\exp(\theta_0 X_\delta))$，其中 $\theta_0=(1,3,1)^{\mathrm{T}}$，可得到响应缺失率约 40%。

为进一步研究各种方法的表现，我们首先在训练集中分别利用 IDC、ITM、IPW 和 FDC 方法筛选基因并拟合 Cox 比例风险模型。然后，在测试集评估各种方法的预测表现，计算测试集的风险得分，将其分为低风险组和高风险组，其中，区分值是由训练集中估计得分的中值得到的。图 1 展示了四种方法下两个风险组病人的生存曲线的 Kaplan-Meier 估计[21]。

从图 1 我们可以看出，在 MAR 情况下，运用 IDC

方法的曲线分离得最好，且其对数秩检验产生的 p 值为 0.001 2，而 IPW 方法和 ITM 方法对应的 p 值分别为 0.037 9 和 0.261 2。在完整数据集的情况下，FDC 方法的 p 值为 0.039 8。



（虚线代表高风险组，实线代表低风险组。The dotted line is high-risk group, and the solid line is low-risk group.）

图 1　测试集中两个风险组对应生存曲线的 Kaplan-Meier 估计

Fig.1　The Kaplan-Meier estimate of survival curves for two risk groups in the testing data

4　结语

本文提出了一种新的特征筛选程序（IDC），首先通过插补技术，补全缺失响应变量值，再构造插补响应变量与协变量分布函数之间的距离相关系数，作为筛选指标。与传统的参数方法相比，所提出的非参数方法在稳健性、可扩展性和非线性重要变量检测能力方面具有一定的优势。本文提出的 IDC 方法基于插补技术创建了一个完整的数据集，便于使用标准的完整数据算法。相比于逆概率加权方法，由于一次插补只为每个具有缺失值的变量创建一个伪观察，因此其影响远小于基于缺失概率的逆概率加权方法的特征筛选程序，即我们的方法仅依赖于插补，减少了极端缺失概率的不利影响。通过数值模拟和实例可以看出，IDC 方法较于对比方法在应用上具有一定的优势，但在确定筛选理论性质方面，基于稳健距离相关发展缺失响应变量的插补方法具有一定的难度，这也启发我们进一步改进筛选指标，优化处理缺失响应数据的超高维筛选方法。

参考文献：

[1] Tibshirani R J. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B, 1996, 58(1): 267-288.

[2] Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1348-1360.

[3] Zou H. The adaptive lasso and its oracle properties[J]. Journal of the American Statistical Association, 2006, 101(476): 1418-1429.

[4] Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n[J]. The Annals of Statistics, 2007, 35(6): 2313-2351.

[5] Fan J Q, Samworth R, Wu Y C. Ultrahigh dimensional feature selection: Beyond the linear model[J]. Journal of Machine Learning Research, 2009, 10: 2013-2038.

[6] Fan J Q, Lv J C. Sure Independence screening for ultrahigh dimensional feature space[J]. Journal of the Royal Statistical Society Series B, 2008, 70(5): 849-911.

[7] Zhu L P, Li L X, Li R Z, et al. Model-free feature screening for ultrahigh-dimensional data[J]. Journal of the American Statistical Association, 2011, 106(496): 1464-1475.

[8] Li R Z, Zhong W, Zhu L P. Feature screening via distance correlation learning[J]. Journal of the American Statistical Association, 2012b, 107(499): 1129-1139.

[9] Zhong W, Zhu L P, Li R Z, et al. Regularized quantile regression and robust feature screening for single index models[J]. Statistica Sinica, 2016, 26(1): 69-95.

[10] Wang Q H, Rao J N K. Empirical Likelihood-Based Inference under Imputation for Missing Response Data[J]. The Annals of Statistics, 2002, 30(3): 896-924.

[11] Qin J, Shao J, Zhang B. Efficient and doubly robust imputation for covariate-dependent missing responses [J]. Journal of the American Statistical Association, 2008, 103(482): 797-810.

[12] Hu Z H, Follmann D A, Qin J. Semiparametric dimension reduction estimation for mean response with missing data[J]. Biometrika, 2010, 97(2): 305-319.

[13] Lai P, Liu Y M, Liu Z, et al. Model free feature screening for ultrahigh dimensional data with responses missing at random[J]. Computational Statistics and Data Analysis, 2017, 105: 201-216.

[14] Fang J L. Nonparametric independence feature screening for ultra-high-dimensional missing data[J]. Communications in Statistics-Simulation and Computation, 2022, 51(10): 5670-5689. DOI: 10.1080/03610918.2020.1779292.

[15] Little R J A, Rubin D B. Statistical analysis with Missing Data [M]. Hoboken, N J, USA: John Wiley and Sons, Inc, 2002.

[16] Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances[J]. The Annals of Statistics, 2007, 35(6): 2769-2794.

[17] Cheng P E. Applications of kernel regression estimation: Survey [J]. Communications in Statistics-Theory and Methods, 1990, 19 (11): 4103-4134.

[18] Cheng P E. Nonparametric estimation of mean functionals with data missing at random[J]. Journal of the American Statistical

Association, 1994, 89(425): 81-87.
[19] Marc A, Gerda C, Niel H, et al. Local multiple imputation[J]. Biometrika, 2002, 89(2): 375-388.
[20] Rosenwald A, Wright G, Chan W C, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-Cell lymphoma[J]. New England Journal of Medicine, 2002, 346(25): 193.
[21] Kaplan E L, Meier P. Nonparametric estimation from incomplete observations[J]. Journal of the American Statistical Association, 1958, 53: 457-481.

Feature Screening for Ultra-High Dimensional Missing Response

Zou Liying, Liu Yi
(School of Mathematical Sciences, Ocean University of China, Qingdao 266100, China)

Abstract: This paper presents a new method to solve the feature screening of ultra-high dimensional data with responses missing at random. The values of the missing response are completed by imputation technology, and then the distance correlation between the imputed response and the distribution function of covariate is used as an index for feature screening. The proposed method has the following advantages. First, it is a nonparametric model-free method, which can detect the nonlinear relationship between variables. Second, it is robust to covariates with outliers. Third, it can deal with multi-dimensional response variables directly. Simulation studies were conducted to examine the performance of the proposed procedure and to compare with existing methods. Finally, our method was applied to the data analysis of diffuse large B-cell lymphoma.

Key words: ultra-high dimensional data; missing at random; feature screening; robust distance correlation; model-free
AMS Subject Classifications: 62D10; 62G99

责任编辑 朱宝象