中国科学: 数学 2024 年 第 54 卷 第 2 期: 211 ∼ 230
SCIENTIA SINICA Mathematica
论 文

《中国科学》杂志社
SCIENCE CHINA PRESS
CrossMark
click for updates

基于投影相关的超高维生存数据的特征筛选新方法

潘莹丽[1], 葛翔宇[2*], 周艳丽[3]

1. 湖北大学数学与统计学学院, 应用数学湖北省重点实验室, 武汉 430062;
2. 中南财经政法大学统计与数学学院, 武汉 430073;
3. 中南财经政法大学金融学院, 武汉 430073
E-mail: ylpan@hubu.edu.cn, xiangyu_ge@zuel.edu.cn, yanli.zhou1517@zuel.edu.cn

收稿日期: 2023-02-01; 接受日期: 2023-05-18; 网络出版日期: 2023-11-28; * 通信作者
国家自然科学基金 (批准号: 11901175, 71974204 和 71901222)、国家社会科学基金 (批准号: 20&ZD132) 和中南财经政法大学中央高校基本科研业务费专项资金 (编号: 2722022AK001) 资助项目

摘要 本文对超高维右删失生存数据的特征筛选提出一种基于投影相关且具有确定独立筛选 (projection correlation sure independent screening, PC-SIS) 的新方法. 一方面, PC-SIS 方法并不需要指定任何模型, 也不需要对生存函数进行非参数估计, 且对矩条件和次指数条件不敏感, 适用于对异常值或厚尾数据的分析. 另一方面, 在一定的正则化条件下, PC-SIS 方法具有确定筛选性和秩相合性. 模拟和实证研究表明, PC-SIS 方法能在保留所有重要特征的前提下剔除与响应变量相关程度较弱的特征, 以实现降维的目的.

关键词 投影相关 秩相合性 确定筛选 生存数据 超高维

MSC (2020) 主题分类 62H12, 62H20

1 引言

随着科学技术的快速发展, 大数据降维已逐渐成为当今机器学习、计算机科学和统计学等领域的热点话题. Fan 等[7] 指出, 分析大数据集受到计算成本、统计准确性和高维算法稳定性的挑战. 因此, 为简化计算过程, 一个有效的策略是在进行统计分析之前筛选出不重要的特征. 目前, 采用变量选择的方法降低数据的维数备受研究者们的青睐. 该方法主要分为两类: 第一类是子集选择方法, 包括逐步回归、向前选择和向后选择等. 然而, 由于子集选择方法具有不稳定性及计算量大等缺点, 它的应用范围受到限制. 第二类是惩罚估计方法, 包括最小绝对收缩和选择算子 (least absolute shrinkage and selection operator, LASSO) [23]、平滑剪裁绝对偏差 (smoothly clipped absolute deviation, SCAD) 惩罚 [4]、自适应 LASSO [34]、Dantzig 选择 [1] 和最小最大凹度惩罚 (minimax concave penalty, MCP) [26] 等. 然

英文引用格式: Pan Y L, Ge X Y, Zhou Y L. A new feature screening method for ultra-high-dimensional survival data based on projection correlation (in Chinese). Sci Sin Math, 2024, 54: 211–230, doi: 10.1360/SCM-2023-0067

© 2023《中国科学》杂志社

而, 由于超高维数据具有变量维数 p 随样本量 n 增大呈指数阶增长的特征, 即满足 $p=\exp(n^{\alpha})$, 其中常数 $\alpha>0$, 这直接导致现有的惩罚估计方法无法处理这些数据[6]. 因此, 针对超高维数据带来的挑战, 引入基于模型或无模型的特征筛选方法来筛选掉超高维数据中不重要的特征显得尤为重要. 例如, 对线性回归模型和广义线性模型, Fan 和 Lv[5] 及 Fan 和 Song[8] 分别提出了一种基于边际 Pearson 相关的确定独立筛选 (sure independence screening, SIS) 方法和基于最大边际似然估计量的确定独立筛选 (SIS with maximum marginal likelihood estimator, SIS-MMLE) 方法. 对超高维可加性模型, Fan 等[2] 探索了一种基于 B 样条近似的非参数独立筛选 (nonparametric independence screening, NIS) 方法. 对非参数异构误差回归模型, Wu 和 Yin[25] 提出了一种条件分位数筛选 Q-SIS (condition quantile SIS) 方法. 对多元响应变量超高维线性模型, Lu 等[17] 建立了一种经验似然特征筛选 MELSIS (multi-response empirical likelihood SIS) 方法. 以上这些变量筛选方法都是基于某个特定模型提出的, 所以可通过交叉验证或信息准则的方法来选择阈值参数. 然而, 由于超高维数据的复杂程度高, 在没有排除冗余协变量的前提下, 想要确定一个正确的模型是非常具有挑战性的. 为了避免模型的设定错误, 研究者们提出了一些不依赖于模型的变量筛选方法. 例如, Zhu 等[32] 在未明确模型结构的情形下, 采用无模型筛选 SIRS (sure independent ranking and screening) 方法进行特征筛选; Li 等[11] 开发了 Kendall τ 相关筛选 RRCS (robust rank correlation screening) 方法; Li 等[14] 引入了距离相关筛选 DCSIS (distance correlation SIS) 方法; Mai 和 Zou[18] 构建了基于融合 Kolmogorov 滤波器的特征筛选 FK-SIS (fused Kolmogorov SIS) 方法; He 和 Xie[10] 提出了超高维多类判别分析的特征筛选 MVSIS (multi-category SIS) 方法等.

生存分析将感兴趣的事件发生时间及其相关因素的关系作为研究对象, 是统计学研究中的一个重要的分支领域. 但由于数据存在删失情形, 即在特定时间点后的部分事件结果变得不可观察或某些实例不可观察, 因此, 传统的处理超高维完全数据的变量筛选方法均不再适用, 迫切需要探究超高维生存数据的变量筛选方法. 为此, 许多研究者都在探寻更优的超高维生存数据的变量筛选新的方法方面开展研究. 一方面, Tibshirani[24]、Fan 等[3] 及 Zhao 和 Li[31] 均基于 Cox 模型分别通过使用单变量 Cox 评分统计量、单变量偏似然估计量和标准边际极大偏似然估计量对超高维生存时间数据进行了降维处理. Gorst-Rasmussen 和 Scheike[9] 基于 FAST (feature aberration at survival times) 统计量提出了适用于所有单指标危险率模型的变量筛选 FAST-SIS 方法. Zhang 等[30] 在比例模型的框架下, 提出了一种针对超高维病例队列生存数据的加权的变量筛选 WSIS (weighted SIS) 方法, 并用这种方法的一个扩展的迭代版本处理一些具有协变量共同重要性且响应变量略微不相关或弱相关的问题. 另一方面, 为了放宽对模型的假设, 也出现了大量处理超高维生存数据的无模型筛选方法. Song 等[21] 基于删失逆概率加权 Kendall τ 提出了一种对异常值点稳健且不依赖于模型的变量筛选 CRIS (censored rank independent screening) 方法. Zhang 等[28] 和 Pan[19] 研究了一种可有效处理超高维生存数据的确定独立筛选 CR-SIS (correlation rank SIS) 方法, CR-SIS 方法除 Kaplan-Meier 估计外, 并不需要任何非参数近似. Li 等[13] 指出了生存影响指数 (survival impact index, SII) 可捕获协变量对生存时间分布的总体影响, 且可使用非参数方法对未知函数进行估计, 并基于 SII 提出了一个无模型特征筛选 SII (SII-based screener) 方法. Zhang 等[29] 还提出了一种具有确定独立筛选特性的无模型的删失累积残差独立筛选 (censored cumulative residual independent screening, CCRIS) 方法. Lin 等[16] 提出了基于逆概率加权的稳健无模型特征筛选 RCSIS (robust censored SIS) 方法. 然而, 上述这些超高维生存数据的无模型特征筛选方法都需要涉及生存函数的 Kaplan-Meier 估计, 从而导致在高删失率的情形下, 使用上述方法进行特征筛选时结果并不太理想. 另外, 为避免在特征筛选中引入非参数估计, Zhang 等[27] 提出了一种新的无模型筛选 DC-SIS (distance correlation SIS) 方法, DC-SIS 方法只涉及超高维生存数

据的距离相关性, 并不涉及生存函数的 Kaplan-Meier 估计, 在生存数据删失率较高和存在异常值或厚尾的情形下, 使用 DC-SIS 方法依然能较好地处理数据. 然而, DC-SIS 方法却存在较强依赖特征和响应变量的矩条件, 还需要协变量、失效时间和删失时间均满足次指数假设等条件才能确定良好的筛选性能的问题.

本文余下内容的安排如下. 第 2 节介绍基于投影相关的特征筛选方法, 首先引入投影相关的概念. Zhu 等 [33] 指出在概率意义下投影相关具有一些优良特性, 例如, 当且仅当两个随机向量独立时, 投影相关值记为 0, 且对两个随机向量的维数不作要求, 并对正交变换群投影相关为不变量. 然后, 提出一种无模型特征筛选 PC-SIS (projection correlation SIS) 方法, 对特征和响应变量之间的投影相关大小进行排序, 剔除那些与响应变量投影相关较小的特征, 以达到降低维数的目的. 再次, 还将探究 PC-SIS 方法的理论性质, 与已有的特征筛选方法相比, PC-SIS 方法具有两个显著的优势: 第一个优势是在方法上, PC-SIS 方法不依赖于任何模型, 不需要对生存函数进行 Kaplan-Meier 估计, 对矩条件和次指数假设均不敏感, 故与现有的其他筛选方法相比具有更强的稳健性, 特别在生存数据删失率相对较高、存在异常值或厚尾的情形下, PC-SIS 方法占绝对优势; 第二个优势是在理论上, 在相对比较弱的正则化条件下, PC-SIS 方法拥有确定筛选性和秩相合性. 第 3 节展示基于 PC-SIS 方法的模拟研究, 首先, 通过转换模型、Cox 比例风险模型和非线性模型, 与 Python 中的 Random 包随机产生满足某种设置要求的实验数据, 并模拟随机和非随机两种删失机制和协变量存在异常值或厚尾的各种情形; 其次, 将依据模拟结果对不同筛选方法的筛选性能进行比较分析, 试图用模拟出的实验结果来说明 PC-SIS 方法在超高维生存数据特征筛选中的表现将会优于已有的特征筛选方法. 第 4 节是基于生存数据的实证研究, 将 PC-SIS 方法应用于弥漫大 B 细胞淋巴瘤患者生存结果的基因筛选中, 实证分析验证 PC-SIS 方法在实际应用中是可行的, 可以较好地实现降维的目标.

2 基于投影相关的特征筛选方法

2.1 投影相关

Zhu 等 [33] 提出的投影相关的一些背景信息为研究本文涉及的超高维生存数据的特征筛选问题奠定了基础. 以下将简述投影相关的一些背景, 为提出 PC-SIS 方法作准备. 设 $U \in \mathbb{R}^p$ 和 $V \in \mathbb{R}^q$ 为两个随机向量, 投影相关由以下独立性检验问题引出:

$$H_0\text{: } U \text{ 和 } V \text{ 是独立的} \quad \text{vs.} \quad H_1\text{: 其他.}$$

为检验 U 和 V 的独立性, 即对任意单位向量 α 和 β, 检验 $S = \alpha^{\mathrm{T}}U$ 与 $T = \beta^{\mathrm{T}}V$ 是否相互独立. 若记 $\mathrm{F}_{S,T}(u,v)$ 为多元随机变量 (S,T) 的联合概率分布函数, 记 $\mathrm{F}_S(u)$ 和 $\mathrm{F}_T(v)$ 分别为随机变量 S 和 T 的边际概率分布函数, 则对于给定的 α 和 β, 随机变量 S 与 T 独立的充分必要条件是对于所有的 $u, v \in \mathbb{R}$ 有

$$\mathrm{F}_{S,T}(u,v) - \mathrm{F}_S(u)\mathrm{F}_T(v) = \mathrm{cov}[I(\alpha^{\mathrm{T}}U \leqslant u), I(\beta^{\mathrm{T}}V \leqslant v)] = 0, \tag{2.1}$$

其中 $I(\cdot)$ 为示性函数.

基于 (2.1), 投影协方差的平方定义为

$$\mathrm{P}^2_{\mathrm{cov}}(U,V) = \iiint (\mathrm{F}_{S,T}(u,v) - \mathrm{F}_S(u)\mathrm{F}_T(v))^2 d\mathrm{F}_{S,T}(u,v) d\alpha d\beta$$

$$= \iiint \mathrm{cov}^2[\mathrm{I}(\alpha^{\mathrm{T}}U \leqslant u), \mathrm{I}(\beta^{\mathrm{T}}V \leqslant v)]d\mathrm{F}_{S,T}(u,v)d\alpha d\beta. \tag{2.2}$$

基于 (2.2), 定义 U 和 V 之间的投影相关的平方 $\mathrm{PC}^2(U,V)$ 为

$$\mathrm{PC}^2(U,V) = \frac{\mathrm{P}^2_{\mathrm{cov}}(U,V)}{\mathrm{P}_{\mathrm{cov}}(U,U)\mathrm{P}_{\mathrm{cov}}(V,V)}, \tag{2.3}$$

其中, 如果 $\mathrm{P}_{\mathrm{cov}}(U,U)=0$ 或者 $\mathrm{P}_{\mathrm{cov}}(V,V)=0$, 则 $\mathrm{PC}(U,V)=0$. 命题 2.1 给出投影相关的一些性质 (参见文献 [33]).

命题 2.1 (1) $0 \leqslant \mathrm{PC}(U,V) \leqslant 1$. 特别地, 当且仅当 U 和 V 相互独立时, $\mathrm{PC}(U,V)=0$ 成立; 当且仅当 $U=\mathrm{E}(U)$ 时, $\mathrm{PC}(U,U)=0$ 成立.

(2) $\mathrm{PC}(U,V)=\mathrm{PC}(a_1+b_1A_1U, a_2+b_2A_2V)$, 其中, $A_1 \in \mathbb{R}^{p\times p}$ 和 $A_2 \in \mathbb{R}^{q\times q}$ 为正交矩阵, $a_1 \in \mathbb{R}^p$ 和 $a_2 \in \mathbb{R}^q$ 为向量, b_1 和 b_2 为标量.

命题 2.1 中的性质 (1) 表明投影相关可以作为度量两个随机向量是否独立的指标. 性质 (2) 表明投影相关对于正交变换群是不变量.

通过简单计算, 可简化由 (2.2) 定义的投影协方差的平方 $\mathrm{P}^2_{\mathrm{cov}}(U,V)$ 为

$$\mathrm{P}^2_{\mathrm{cov}}(U,V) = J_1 + J_2 - 2J_3, \tag{2.4}$$

其中,

$$\begin{aligned}
J_1 &= \mathrm{E}\left[\arccos\left\{\frac{(U_1-U_3)^{\mathrm{T}}(U_4-U_3)}{\|U_1-U_3\|\|U_4-U_3\|}\right\}\arccos\left\{\frac{(V_1-V_3)^{\mathrm{T}}(V_4-V_3)}{\|V_1-V_3\|\|V_4-V_3\|}\right\}\right],\\
J_2 &= \mathrm{E}\left[\arccos\left\{\frac{(U_1-U_3)^{\mathrm{T}}(U_4-U_3)}{\|U_1-U_3\|\|U_4-U_3\|}\right\}\arccos\left\{\frac{(V_2-V_3)^{\mathrm{T}}(V_5-V_3)}{\|V_2-V_3\|\|V_5-V_3\|}\right\}\right],\\
J_3 &= \mathrm{E}\left[\arccos\left\{\frac{(U_1-U_3)^{\mathrm{T}}(U_4-U_3)}{\|U_1-U_3\|\|U_4-U_3\|}\right\}\arccos\left\{\frac{(V_2-V_3)^{\mathrm{T}}(V_4-V_3)}{\|V_2-V_3\|\|V_4-V_3\|}\right\}\right],
\end{aligned}$$

且 $(U_1,V_1),\ldots,(U_5,V_5)$ 在随机变量 (U,V) 中独立同分布, $\|\cdot\|$ 表示 L_2 范数. 从 (2.4) 可以看出, 投影协方差的显著特征是由 $(U_k-U_l)/\|U_k-U_l\|$ 和 $(V_k-V_l)/\|V_k-V_l\|$ 的函数表示的, 这两项的二阶矩均存在且为 1, 从一个侧面反映投影协方差消除了对随机向量 (U,V) 需要存在二阶矩的要求. 然而从距离协方差的定义上看, 随机变量存在二阶矩的假设是必不可少的.

假设 $\{(U_i,V_i): i=1,\ldots,n\}$ 是一个随机样本, 与随机向量 (U,V) 同分布, 受 Serfling [20] 的启发, 可采用矩估计方法来估计 J_k $(k=1,2,3)$ 的值, 其具体估计量如下:

$$\begin{aligned}
\widehat{J}_1 &= n^{-3}\sum_{i,k,l=1}^{n}\left[\arccos\left\{\frac{(U_i-U_k)^{\mathrm{T}}(U_l-U_k)}{\|U_i-U_k\|\|U_l-U_k\|}\right\}\arccos\left\{\frac{(V_i-V_k)^{\mathrm{T}}(V_l-V_k)}{\|V_i-V_k\|\|V_l-V_k\|}\right\}\right],\\
\widehat{J}_2 &= n^{-5}\sum_{i,j,k,l,r=1}^{n}\left[\arccos\left\{\frac{(U_i-U_k)^{\mathrm{T}}(U_l-U_k)}{\|U_i-U_k\|\|U_l-U_k\|}\right\}\arccos\left\{\frac{(V_j-V_k)^{\mathrm{T}}(V_r-V_k)}{\|V_j-V_k\|\|V_r-V_k\|}\right\}\right],\\
\widehat{J}_3 &= n^{-4}\sum_{i,j,k,l=1}^{n}\left[\arccos\left\{\frac{(U_i-U_k)^{\mathrm{T}}(U_l-U_k)}{\|U_i-U_k\|\|U_l-U_k\|}\right\}\arccos\left\{\frac{(V_j-V_k)^{\mathrm{T}}(V_l-V_k)}{\|V_j-V_k\|\|V_l-V_k\|}\right\}\right],
\end{aligned}$$

这里, 不妨定义 $0/0=1$. 因此, 投影协方差 $\mathrm{P}_{\mathrm{cov}}(U,V)$ 的平方估计可表示为

$$\widetilde{\mathrm{P}}^2_{\mathrm{cov}}(U,V) = \widehat{J}_1 + \widehat{J}_2 - 2\widehat{J}_3. \tag{2.5}$$

Székely 和 Rizzo [22] 指出 $\widetilde{\mathrm{P}}_{\mathrm{cov}}(U,V)$ 是 $\mathrm{P}_{\mathrm{cov}}(U,V)$ 的非常自然的估计量, 然而计算起来却很困难. Zhu 等 [33] 给出了估计量 $\widetilde{\mathrm{P}}^2_{\mathrm{cov}}(U,V)$ 的等价形式, 即 U 和 V 的投影协方差的平方可以通过如下的 $\widehat{\mathrm{P}}^2_{\mathrm{cov}}(U,V)$ 来估计

$$\widehat{\mathrm{P}}^2_{\mathrm{cov}}(U,V)=n^{-3}\sum_{k,l,r=1}^{n}A_{klr}B_{klr}, \tag{2.6}$$

其中, 对于 $k,l,r=1,\ldots,n$, 有

$$\begin{aligned}
&A_{klr}=a_{klr}-\overline{a}_{k\cdot r}-\overline{a}_{\cdot lr}+\overline{a}_{\cdot\cdot r},\quad B_{klr}=b_{klr}-\overline{b}_{k\cdot r}-\overline{b}_{\cdot lr}+\overline{b}_{\cdot\cdot r},\\
&a_{klr}=\arccos\left\{\frac{(U_k-U_r)^{\mathrm{T}}(U_l-U_r)}{\|U_k-U_r\|\|U_l-U_r\|}\right\},\\
&a_{klr}=0,\quad \text{如果 } k=r \text{ 或 } l=r,\\
&\overline{a}_{k\cdot r}=n^{-1}\sum_{l=1}^{n}a_{klr},\quad \overline{a}_{\cdot lr}=n^{-1}\sum_{k=1}^{n}a_{klr},\quad \overline{a}_{\cdot\cdot r}=n^{-2}\sum_{k=1}^{n}\sum_{l=1}^{n}a_{klr},\\
&b_{klr}=\arccos\left\{\frac{(V_k-V_r)^{\mathrm{T}}(V_l-V_r)}{\|V_k-V_r\|\|V_l-V_r\|}\right\},\\
&b_{klr}=0,\quad \text{如果 } k=r \text{ 或 } l=r,\\
&\overline{b}_{k\cdot r}=n^{-1}\sum_{l=1}^{n}b_{klr},\quad \overline{b}_{\cdot lr}=n^{-1}\sum_{k=1}^{n}b_{klr},\quad \overline{b}_{\cdot\cdot r}=n^{-2}\sum_{k=1}^{n}\sum_{l=1}^{n}b_{klr}.
\end{aligned}$$

同理, 定义样本投影方差 $\widehat{\mathrm{P}}_{\mathrm{cov}}(U,U)$ 和 $\widehat{\mathrm{P}}_{\mathrm{cov}}(V,V)$ 分别为

$$\widehat{\mathrm{P}}_{\mathrm{cov}}(U,U)=\sqrt{n^{-3}\sum_{k,l,r=1}^{n}A^2_{klr}},\quad \widehat{\mathrm{P}}_{\mathrm{cov}}(V,V)=\sqrt{n^{-3}\sum_{k,l,r=1}^{n}B^2_{klr}}. \tag{2.7}$$

因此, U 和 V 之间的样本投影相关的平方 $\widehat{\mathrm{PC}}^2(U,V)$ 为

$$\widehat{\mathrm{PC}}^2(U,V)=\frac{\widehat{\mathrm{P}}^2_{\mathrm{cov}}(U,V)}{\widehat{\mathrm{P}}_{\mathrm{cov}}(U,U)\widehat{\mathrm{P}}_{\mathrm{cov}}(V,V)}. \tag{2.8}$$

命题 2.2 给出样本投影相关的指数偏差不等式 (参见文献 [33]).

命题 2.2 对于任意 $0<\delta<1$, 当 $n\geqslant 10\pi^2/\delta$ 时, 有

$$\mathrm{P}\{|\widehat{\mathrm{PC}}^2(U,V)-\mathrm{PC}^2(U,V)|\geqslant\delta\}\leqslant 5d_1\exp\{-d_2\sigma n\delta^2\}, \tag{2.9}$$

其中, d_1 和 d_2 为两个大于 0 的常数, $\sigma=\min\{\sigma_U^3\sigma_V^3/(4M^4),\sigma_U^2\sigma_V^2/(4M^4)\}$, $\sigma_U=\mathrm{P}^2_{\mathrm{cov}}(U,U)$, $\sigma_V=\mathrm{P}^2_{\mathrm{cov}}(V,V)$ 且 $M=4\pi^2$.

注意到 (2.9) 中显示的指数不等式与两个随机向量 U 和 V 的维数及矩条件无关. 异常概率随样本量 n 的大小呈指数下降状态, 这可保证在有限样本条件下提出的估计 $\widehat{\mathrm{PC}}(U,V)$ 具有良好的估计性能.

2.2 基于投影相关的特征筛选 PC-SIS 方法

本小节将依据投影相关的基础研究, 提出超高维生存数据的独立特征筛选 PC-SIS 方法. 设 $\widetilde{T}$ 表示潜在失效时间, C 表示删失时间, $\boldsymbol{x}=(X_1,\ldots,X_p)^{\mathrm{T}}$ 为 p 维的协变量. $T=\min(\widetilde{T},C)$ 为可观测时

间, $\Delta = I(\widetilde{T} \leqslant C)$ 为删失指示变量. 对于给定的 $\boldsymbol{x}$, 记 $S(t \mid \boldsymbol{x}) = \mathrm{P}(\widetilde{T} > t \mid \boldsymbol{x})$ 为 $\widetilde{T}$ 的条件生存函数. 在不指定任何回归模型的情形下, 可定义重要预测变量的索引集为

$$\mathcal{I} = \{k : S(t \mid \boldsymbol{x}) \text{ 依赖于 } X_k, k = 1, \ldots, p\}, \tag{2.10}$$

其中, $s = |\mathcal{I}|$ 表示重要特征的个数, $|\mathcal{I}|$ 表示集合 $\mathcal{I}$ 中元素的个数. 在超高维定义中, 假定协变量的维数 p 远超过样本量 n 的大小, 而重要特征的数量 s 远小于 n 的数量. 记 $\mathcal{I}^c$ 为集合 $\mathcal{I}$ 的余集, 表示所有非重要特征的索引集. 本文的主要目的是希望在某些正则条件下, 找到协变量的一个子集 $\widehat{\mathcal{I}}$, 使得当 $n \to \infty$ 时, $\mathrm{P}(\mathcal{I} \subseteq \widehat{\mathcal{I}})$ 趋近于 1.

假设观测到的数据 $\{T_i, \Delta_i, \boldsymbol{x}_i = (X_{1i}, X_{2i}, \ldots, X_{pi})^{\mathrm{T}} : i = 1, 2, \ldots, n\}$ 是 $(T, \Delta, \boldsymbol{x})$ 中的独立同分布的样本. 定义 $y \triangleq (T, \Delta)$ 为二维响应向量, 本文通过验证 y 与每个协变量 X_k $(k = 1, 2, \ldots, p)$ 之间的投影相关的大小来筛选出重要的特征, 即对于给定的特征 X_k $(k = 1, \ldots, p)$, 如果 X_k 和 y 相互独立, 则它们之间的投影相关的值应该在 0 附近摆动; 而认为那些投影相关值较大的协变量是重要的变量. 由第 2.1 小节投影相关的概念可知, 对于 $k = 1, \ldots, p$, $\omega_k = \mathrm{PC}^2(X_k, y)$ 可测量 y 和 X_k 之间的投影相关强度, 并可定义

$$\widehat{\omega}_k = \widehat{\mathrm{PC}}^2(X_k, y) \tag{2.11}$$

为 ω_k 的估计量. 因此, 可定义重要的特征指标集的估计为

$$\widehat{\mathcal{I}}(\epsilon) = \{k : \widehat{\omega}_k \geqslant \epsilon, k = 1, \ldots, p\}, \tag{2.12}$$

其中 ϵ 为预先指定的阈值. 给定一个阈值 $d > 0$, 重要特征集合的估计 (2.12) 与 (2.13) 等价,

$$\widehat{\mathcal{I}}(d) = \{k : \widehat{\omega}_k \text{ 是 } p \text{ 个 } \widehat{\omega}_j \ (j = 1, 2, \ldots, p) \text{ 中最大的 } d \text{ 个}\}. \tag{2.13}$$

在模拟和实证研究中, 选择阈值 $d = \lfloor n/\log n \rfloor$ (参见文献 [5]), 其中 $\lfloor a \rfloor$ 表示对 a 向下取整数部分. 本文称这种由 (2.12) 或 (2.13) 定义的超高维生存数据的特征筛选方法为基于投影相关的独立筛选 PC-SIS 方法.

2.3 PC-SIS 方法的理论性质

本小节将证明在一定的正则化条件下独立特征筛选 PC-SIS 方法具有确定筛选性和秩相合性. 在整个证明过程中, 假设所提出的 PC-SIS 方法满足下面的正则条件 2.1.

条件 2.1 (C1) 存在常数 $\delta > 0$ 和 $\tau > 0$ 使得 $\mathrm{P}(C \geqslant \tau) = \mathrm{P}(C = \tau) \geqslant \delta$, 其中 τ 为实验终止时间.

(C2) 对于任意给定的常数 $c > 0$ 和 $0 \leqslant \kappa < 1/2$, 重要特征的最小投影相关条件满足

$$\min_{k \in \mathcal{I}} \omega_k \geqslant 2cn^{-\kappa}.$$

(C3) 对于任意给定的常数 $c > 0$ 和 $0 \leqslant \kappa < 1/2$, 重要特征的最小投影相关和非重要特征的最大投影相关条件满足

$$\min_{k \in \mathcal{I}} \omega_k - \max_{k \in \mathcal{I}^c} \omega_k \geqslant 2cn^{-\kappa}.$$

条件 2.1 中的 (C1) 为生存分析中常见的假设之一, 它意味着至少有一些个体在实验终止时间 τ 时还未发生死亡事件, 由定义可知这些个体在 τ 时刻被右删失. 条件 2.1 中的 (C2) 表明重要特征和响应变量之间的投影相关的平方为一致有界, 且随着样本量 n 趋于无穷, 这个投影相关的平方无法快速收敛到 0. 换而言之, 该假设意味着重要变量所对应的投影相关不能太弱. 条件 2.1 中的 (C3) 是对重要特征和非重要特征之间的信号强度差距的合理假设, 只有当重要特征的信号足够强时, 相应的协变量才变得重要.

独立特征筛选 PC-SIS 方法拥有的确定筛选性和秩相合性将分别由定理 2.1 和 2.2 给出.

定理 2.1 若条件 2.1 成立, 且选择 $\epsilon \leqslant \min_{k\in\mathcal{I}} \omega_k/2$, 则存在常数 $c_1 > 0$ 使得

$$\mathrm{P}(\mathcal{I} \subseteq \widehat{\mathcal{I}}(\epsilon)) \geqslant 1 - O(s\exp\{-c_1 n^{1-2\kappa}\}).$$

证明 由第 2.1 小节中的指数不等式 (2.9) 可知

$$\mathrm{P}(|\widehat{\mathrm{PC}}^2(X_k, y) - \mathrm{PC}^2(X_k, y)| \geqslant \epsilon) \leqslant O(\exp\{-c_1^* n\epsilon^2\}),$$

其中 c_1^* 是一个正常数. 注意到 $\omega_k = \mathrm{PC}^2(X_k, y)$ 和 $\widehat{\omega}_k = \widehat{\mathrm{PC}}^2(X_k, y)$, 则有

$$\mathrm{P}(|\widehat{\omega}_k - \omega_k| \geqslant \epsilon) \leqslant O(\exp\{-c_1^* n\epsilon^2\}). \tag{2.14}$$

接下来证明

$$\{\mathcal{I} \nsubseteq \widehat{\mathcal{I}}(\epsilon)\} \subseteq \{k \in \mathcal{I} : |\widehat{\omega}_k - \omega_k| > cn^{-\kappa}\}, \tag{2.15}$$

其中 $c > 0$ 和 $0 \leqslant \kappa < 1/2$ 为已知常数. 实际上, $\mathcal{I} \nsubseteq \widehat{\mathcal{I}}(\epsilon)$ 等价于存在某个 $k \in \mathcal{I}$, 使得 $\widehat{\omega}_k < \epsilon$.

为证明 (2.15), 只需要证明

$$\{\widehat{\omega}_k < \epsilon\} \subseteq \{|\widehat{\omega}_k - \omega_k| > cn^{-\kappa}\}$$

对于 $k \in \mathcal{I}$ 和 $c > 0$ 及 $0 \leqslant \kappa < 1/2$ 成立. 根据三角不等式, 可得

$$\omega_k - \widehat{\omega}_k = |\omega_k| - |\widehat{\omega}_k| \leqslant |\widehat{\omega}_k - \omega_k|. \tag{2.16}$$

根据条件 2.1 中的 (C2), 有 $\min_{k\in\mathcal{I}} \omega_k/2 \geqslant cn^{-\kappa}$, 其中 $c > 0$ 和 $0 \leqslant \kappa < 1/2$ 为常数. 结合定理 2.1 中的条件 $\epsilon \leqslant \min_{k\in\mathcal{I}} \omega_k/2$, 对于 $k \in \mathcal{I}$, 有

$$\omega_k - \widehat{\omega}_k > \omega_k - \epsilon \geqslant \omega_k - \min_{k\in\mathcal{I}} \frac{\omega_k}{2} \geqslant \frac{\omega_k}{2} \geqslant cn^{-\kappa}. \tag{2.17}$$

因此, 结合 (2.16) 和 (2.17), 对于 $k \in \mathcal{I}$, 有 $|\widehat{\omega}_k - \omega_k| > cn^{-\kappa}$.

根据 (2.15), 可得

$$\Big\{\max_{k\in\mathcal{I}} |\widehat{\omega}_k - \omega_k| \leqslant cn^{-\kappa}\Big\} \subseteq \{\mathcal{I} \subseteq \widehat{\mathcal{I}}(\epsilon)\}.$$

因此, 根据 (2.14) 并通过简单的运算, 可得

$$\begin{aligned}
\mathrm{P}(\mathcal{I} \subseteq \widehat{\mathcal{I}}(\epsilon)) &\geqslant \mathrm{P}\Big(\max_{k\in\mathcal{I}} |\widehat{\omega}_k - \omega_k| \leqslant cn^{-\kappa}\Big) \\
&\geqslant 1 - \mathrm{P}\Big(\max_{k\in\mathcal{I}} |\widehat{\omega}_k - \omega_k| > cn^{-\kappa}\Big) \\
&\geqslant 1 - s\max_{k\in\mathcal{I}} \mathrm{P}(|\widehat{\omega}_k - \omega_k| > cn^{-\kappa}) \\
&\geqslant 1 - O(s\exp\{-c_1 n^{1-2\kappa}\}),
\end{aligned}$$

其中 c_1 为一个常数. □

注 2.1 由定理 2.1 可知, 如果设 $\epsilon = cn^{-\kappa}$, ϵ 满足条件 $\epsilon \leqslant \min_{k\in\mathcal{I}} \omega_k/2$, 则明显有

$$\mathrm{P}(\mathcal{I} \subseteq \widehat{\mathcal{I}}(cn^{-\kappa})) \geqslant 1 - O(s\exp\{-c_1 n^{1-2\kappa}\}). \tag{2.18}$$

从 (2.18) 可知, 当 $\epsilon = cn^{-\kappa}$ 且 $n \to \infty$ 时, 依概率 1 可全部选择出所有真实的重要特征. 类似地, 当 $\epsilon = cn^{-\kappa}$ 时, Zhang 等 [27] 指出了基于距离相关的特征筛选 DC-SIS 方法满足

$$\mathrm{P}(\mathcal{I} \subseteq \widehat{\mathcal{I}}(cn^{-\kappa})) \geqslant 1 - O(s\exp\{-c_1' n^{1-2(\kappa+\gamma)}\} + n\exp\{-c_1'' n^{\gamma}\}), \tag{2.19}$$

其中 c_1'、c_1'' 和 $\gamma \in (0, 1/2-\kappa)$ 为给定的常数. 通过比较 (2.18) 和 (2.19) 可知, 基于 (2.18) 的结论更加简化.

定理 2.2 若条件 2.1 成立, 则存在常数 $c_2 > 0$ 使得

$$\mathrm{P}\Big(\min_{k\in\mathcal{I}} \widehat{\omega}_k - \max_{k\in\mathcal{I}^c}\widehat{\omega}_k > 0\Big) > 1 - O(p\exp\{-c_2 n^{1-2\kappa}\}).$$

此外, 对于 $0 \leqslant \kappa < 1/2$, 若 $\log(p) = o(n^{1-2\kappa})$, 则

$$\mathrm{P}\Big(\liminf_{n\to\infty}\Big\{\min_{k\in\mathcal{I}} \widehat{\omega}_k - \max_{k\in\mathcal{I}^c} \widehat{\omega}_k > 0\Big\}\Big) = 1.$$

证明 记 $l = \mathrm{argmin}_{k\in\mathcal{I}}\widehat{\omega}_k$, $m = \mathrm{argmin}_{k\in\mathcal{I}^c}\widehat{\omega}_k$, 通过简单运算, 可得

$$\begin{aligned}
\mathrm{P}\Big(\min_{k\in\mathcal{I}} \widehat{\omega}_k - \max_{k\in\mathcal{I}^c} \widehat{\omega}_k \leqslant 0\Big) &\leqslant \mathrm{P}\Big(\min_{k\in\mathcal{I}} \widehat{\omega}_k - \max_{k\in\mathcal{I}^c} \widehat{\omega}_k \leqslant \min_{k\in\mathcal{I}} \omega_k - \max_{k\in\mathcal{I}^c} \omega_k - 2cn^{-\kappa}\Big) \\
&= \mathrm{P}\Big(\min_{k\in\mathcal{I}} \omega_k - \min_{k\in\mathcal{I}} \widehat{\omega}_k + \max_{k\in\mathcal{I}^c} \widehat{\omega}_k - \max_{k\in\mathcal{I}^c} \omega_k \geqslant 2cn^{-\kappa}\Big) \\
&\leqslant \mathrm{P}(\omega_l - \widehat{\omega}_l + \widehat{\omega}_m - \omega_m \geqslant 2cn^{-\kappa}) \\
&\leqslant \mathrm{P}(|\widehat{\omega}_l - \omega_l| \geqslant cn^{-\kappa}) + \mathrm{P}(|\widehat{\omega}_m - \omega_m| \geqslant cn^{-\kappa}) \\
&\leqslant 2\mathrm{P}\Big(\max_{1\leqslant k\leqslant p} |\widehat{\omega}_k - \omega_k| \geqslant cn^{-\kappa}\Big) \\
&\leqslant c_2' p\,\exp\{-c_2 n^{1-2\kappa}\},
\end{aligned} \tag{2.20}$$

其中 c_2' 和 c_2 为正的常数. (2.20) 中的第一个不等式可由条件 2.1 中的 (C3) 得到, (2.20) 中的最后一个不等式可由命题 2.2 推导出. 因此, 可得

$$\mathrm{P}\Big(\min_{k\in\mathcal{I}} \widehat{\omega}_k - \max_{k\in\mathcal{I}^c} \widehat{\omega}_k > 0\Big) > 1 - O(p\exp\{-c_2 n^{1-2\kappa}\}).$$

此外, 如果选择 $\log p = o(n^{1-2\kappa})$, 其中 $0 \leqslant \kappa < 1/2$, 则对于足够大的 n, 有 $p \leqslant \exp\{c_2 n^{1-2\kappa}/2\}$. 通过运算, 可得

$$p\exp\{-c_2 n^{1-2\kappa}\} \leqslant \exp\Big\{-\frac{c_2 n^{1-2\kappa}}{2}\Big\} \leqslant \exp\{-2\log n\} = n^{-2}.$$

进而, 对于足够大的 n_1, 可得

$$\sum_{n=n_1}^{\infty} c_2' p\,\exp\{-c_2 n^{1-2\kappa}\} \leqslant c_2' \sum_{n=n_1}^{\infty} n^{-2} \leqslant \infty. \tag{2.21}$$

将 (2.20) 和 (2.21) 结合起来, 可得

$$\sum_{n=n_1}^{\infty} \mathrm{P}\Big(\min_{k\in\mathcal{I}} \widehat{\omega}_k - \max_{k\in\mathcal{I}^c} \widehat{\omega}_k \leqslant 0\Big) \leqslant \infty.$$

然后根据 Borel-Cantelli 引理, 可得

$$\mathrm{P}\Big(\limsup_{n\to\infty}\Big\{\min_{k\in\mathcal{I}} \widehat{\omega}_k - \max_{k\in\mathcal{I}^c} \widehat{\omega}_k \leqslant 0\Big\}\Big) = 0.$$

通过运算, 可得

$$\mathrm{P}\Big(\liminf_{n\to\infty}\Big\{\min_{k\in\mathcal{I}} \widehat{\omega}_k - \max_{k\in\mathcal{I}^c} \widehat{\omega}_k > 0\Big\}\Big) = \mathrm{P}\Big(\Big[\limsup_{n\to\infty}\Big\{\min_{k\in\mathcal{I}} \widehat{\omega}_k - \max_{k\in\mathcal{I}^c} \widehat{\omega}_k \leqslant 0\Big\}\Big]^c\Big) = 1.$$

证毕. □

注 2.2 由定理 2.2 可知, 如果重要特征和非重要特征之间的信号间隙满足条件 2.1 中的 (C3), 则重要特征可能排在前面, 非重要特征可能排在后面. 换而言之, 存在一种 ϵ 选择能完美地将重要特征与非重要特征分开, 而且还是一种大概率事件.

3 基于 PC-SIS 方法的模拟研究

本节将通过 Monte Carlo 模拟评估 PC-SIS 方法的有限样本性能, 并将其与另外 5 种已有的超高维生存数据的特征筛选方法进行比较, 即

- Zhao 和 Li [31] 提出的基于 Cox 模型的特征筛选 P-SIS (principled SIS) 方法.
- Gorst-Rasmussen 和 Scheike [9] 提出的基于单指数风险率模型的特征筛选 FAST-SIS 方法.
- Song 等 [21] 提出的基于删失逆概率加权 Kendall τ 的无模型独立筛选 CRIS 方法.
- Zhang 等 [28] 提出的基于相关秩排序的无模型独立筛选 CR-SIS 方法.
- Zhang 等 [27] 提出的基于距离相关的无模型特征筛选 DC-SIS 方法.

在以下 3 个例子中, 设置样本量大小为 $n = 200$, 维数为 $p = 2{,}000$, 重复每个实验 200 次, 根据以下 3 个指标评估方法进行有限样本性能的模型分析.

- Size: 最小模型变量维数大小, 即在所选模型中能够保留所有重要变量的变量个数取最小值. 通过 200 次模拟实验所得值的 5%、25%、50%、75% 和 95% 分位数来衡量此变量筛选方法所选出模型的复杂程度, 其模拟值越接近真实模型变量维数大小, 筛选的结果越好.
- P_e: 对于给定的模型变量维数大小 $\lfloor n/\log n\rfloor$, 进行 200 次数值模拟分别选出每个重要变量的概率. P_e 越接近 1, 数值模拟结果越好.
- P_a: 对于给定的模型变量维数大小 $\lfloor n/\log n\rfloor$, 进行 200 次数值模拟, 同时选出所有重要变量的概率. P_a 越接近 1, 数值模拟结果越好.

例 3.1 假设生存时间 $\widetilde{T}$ 由以下转换模型 [21] 生成:

$$H(\widetilde{T}) = -\beta^{\mathrm{T}}\boldsymbol{x} + \varepsilon,$$

其中, $H(t) = \log\{0.5(\mathrm{e}^{2t} - 1)\}$, 真实参数为 $\beta = (-1, -0.9, 0_6, 0.8, 1.0, 0_{p-10})^{\mathrm{T}}$, 剩余设置考虑 3 种情形. 第一种情形: 协变量 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 由均值为 0、方差矩阵为 $\Sigma = (0.5^{|i-j|})_{p\times p}$ $(i, j = 1, \ldots, p)$ 的多元正态分布生成, 误差项 ε 由标准正态分布 $N(0, 1)$ 或标准的 Cauchy 分布 Cauchy$(0, 1)$ 生成. 删失

时间 C 由 0 到 c 上的均匀分布 Unif$(0,c)$ 生成, 其中通过调整常数 c 使得删失率 Cr 被设定约为 20% 和 40%. 第二种情形: 删失时间 C 由 0 到 $c\cdot|X_1-X_2|$ 上的均匀分布 Unif$(0,c\cdot|X_1-X_2|)$ 生成, 协变量、误差项和删失率的设置与第一种情形相同. 第三种情形: 协变量 $\boldsymbol{x}=(X_1,\ldots,X_p)^{\mathrm{T}}$ 由混合分布 $0.9N_p(0,\Sigma)+0.1\widetilde{X}$ 生成, 其中, $\widetilde{X}=(\widetilde{X}_1,\ldots,\widetilde{X}_p)^{\mathrm{T}}$ 为一个 p 维随机向量, 且每个分量都独立同分布于自由度为 1 的 t 分布. 方差矩阵 Σ、误差项和删失时间以及删失率的设置与第一种情形相同. 第一种情形和第二种情形旨在验证本文提出的 PC-SIS 方法在完全随机删失机制和随机删失机制下的有限样本性能. 第三种情形旨在验证当某些协变量受到异常值污染的情形下, 本文提出的 PC-SIS 方法具有稳健性.

表 1 和 2 分别给出在上述 3 种情形下的误差项为 $\varepsilon\sim\mathrm{N}(0,1)$ 和 $\varepsilon\sim\mathrm{Cauchy}(0,1)$ 的模拟结果. 由表 1 的结果可以看出, 在所有 3 种情形下 PC-SIS 方法的表现都很好, 并优于其他 5 种方法. 特别是在第三种情形中, 当一些协变量受到异常值污染时, 由 PC-SIS 方法得到的最小模型变量的维数大小 Size 接近或等于 4, 而由其他方法如 P-SIS 方法、FAST-SIS 方法和 CR-SIS 方法在 75% 和 95% 分位数处的表现可能与随机猜测的结果一样不理想, 即为了筛选出所有真实的重要特征, 几乎选出所有全部特征. 此外, 比较 P_e 和 P_a 的结果, 也可看到 PC-SIS 方法的表现更好, 保留个别活跃预测变量或所有重要变量的概率比其他 5 种方法的概率更大. 通过比较第一种情形和第二种情形下的结果可知 PC-SIS 方法在完全随机删失机制或随机删失机制下均可有效地筛选特征. 由表 2 与 1 中相应的结果相比较, FAST-SIS 方法、CRIS 方法和 CR-SIS 方法的表现变差. 确切地, 在使用 FAST-SIS 方法、CRIS 方法和 CR-SIS 方法时, 较难在 75% 和 95% 分位数处选择一个合理的最小模型变量维数. 此外, 在第三种情形下, 当数据中存在异常值或厚尾时, P-SIS 方法、FAST-SIS 方法、CRIS 方法、CR-SIS 方法和 DC-SIS 方法在 95% 分位数处的表现都不好. 相比之下, PC-SIS 方法的表现仍然良好, 并遥遥领先其他 5 种方法的表现.

例 3.2 假设生存时间 $\widetilde{T}$ 由以下的 Cox 比例风险回归模型 [28] 生成:

$$\lambda(\widetilde{T}\mid\boldsymbol{x})=\lambda_0(\widetilde{T})\exp(\boldsymbol{x}^{\mathrm{T}}\beta_0),$$

其中, 设置基准风险函数为 $\lambda_0(t)=(t-0.5)^2$, 真实参数为 $\beta_0=(0.6,0.6,0.6,0.6,0.6,0,\ldots,0)^{\mathrm{T}}$, 剩余设置与例 3.1 保持一致. 表 3 汇总了在 Cox 比例风险回归模型下的特征筛选结果. 从表 3 的结果可以观察到, 在上述 3 种情形下, PC-SIS 方法的表现都很好. 在第一种情形和第二种情形中, P-SIS 方法和 DC-SIS 方法的表现与 PC-SIS 方法的表现相当, 但在第三种情形中的表现都较差. 其余结果与表 1 和 2 中的结果类似.

例 3.3 假设生存时间 $\widetilde{T}$ 由以下的非线性模型 [27] 生成:

$$\log\widetilde{T}=(2+\sin X_1)^2+(1+X_5)^3+3X_{10}^2+X_1X_{10}+\varepsilon,$$

其中, 协变量 $\boldsymbol{x}=(X_1,\ldots,X_p)^{\mathrm{T}}$、删失时间和删失率与例 3.1 相同, 误差项 $\varepsilon\sim\mathrm{N}(0,1)$. 本部分的模拟结果在表 4 中展示. 同样地, 与其他 5 种方法相比, PC-SIS 方法表现更出色. 在上述 3 种情形下, 使用 P-SIS 方法、FAST-SIS 方法、CRIS 方法和 CR-SIS 方法得到的最小模型变量维数大小的 95% 分位数几乎接近 2,000. 此外, 在第三种情形中 DC-SIS 方法得到的最小模型变量维数大小的 95% 分位数也接近 2,000, 即在最坏的情形下, P-SIS 方法、FAST-SIS 方法、CRIS 方法、CR-SIS 方法和 DC-SIS 方法在不漏掉任何重要特征的前提下都不能有效地降低数据的维数. 然而, 本文提出的 PC-SIS 方法却能相对较好地达到降维的目的.

表 1 转换模型下, 当 $\epsilon\sim N(0,1)$ 且删失率分别等于 20% 和 40% 时在 3 种情形下 6 种筛选方法的模拟结果

Cr		筛选方法	Size					P_e				P_a
			5%	25%	50%	75%	95%	X_1	X_2	X_9	X_{10}	
20%	第一种情形	P-SIS	4.0	4.0	4.0	4.0	4.0	1.000	1.000	1.000	1.000	1.000
		FAST-SIS	4.0	5.0	10.0	25.5	128.0	0.965	0.965	0.905	0.975	0.835
		CRIS	4.0	4.0	4.0	8.0	96.5	0.950	0.960	0.980	1.000	0.910
		CR-SIS	4.0	11.0	23.0	61.5	490.5	0.885	0.895	0.855	0.885	0.645
		DC-SIS	4.0	4.0	4.0	4.0	4.0	1.000	1.000	1.000	1.000	1.000
		PC-SIS	4.0	4.0	4.0	4.0	4.0	1.000	1.000	1.000	1.000	1.000
	第二种情形	P-SIS	4.0	4.0	4.0	4.0	4.0	1.000	1.000	1.000	1.000	1.000
		FAST-SIS	4.0	5.0	10.0	25.0	121.0	0.980	0.950	0.935	0.970	0.840
		CRIS	4.0	4.0	4.5	7.0	166.0	0.950	0.940	0.980	0.995	0.905
		CR-SIS	4.0	11.0	23.0	66.0	555.0	0.885	0.880	0.840	0.890	0.615
		DC-SIS	4.0	4.0	4.0	4.0	4.0	1.000	1.000	1.000	1.000	1.000
		PC-SIS	4.0	4.0	4.0	4.0	4.0	1.000	1.000	1.000	1.000	1.000
	第三种情形	P-SIS	4.0	4.5	216.5	1,532.0	1,958.5	0.745	0.705	0.665	0.710	0.370
		FAST-SIS	33.0	449.5	1,188.0	1,639.0	1,935.0	0.285	0.240	0.230	0.230	0.050
		CRIS	4.0	4.0	5.0	18.0	269.5	0.930	0.920	0.940	0.995	0.830
		CR-SIS	25.0	227.5	739	1,532.0	1,942.5	0.415	0.415	0.305	0.415	0.080
		DC-SIS	4.0	4.0	4.0	212.5	1,496.0	0.810	0.765	0.750	0.825	0.680
		PC-SIS	4.0	4.0	4.0	4.0	4.0	1.000	1.000	1.000	1.000	1.000
40%	第一种情形	P-SIS	4.0	4.0	4.0	4.0	5.0	1.000	1.000	1.000	1.000	1.000
		FAST-SIS	4.0	11.0	33.5	101.5	306.0	0.890	0.880	0.790	0.860	0.545
		CRIS	4.0	4.0	6.0	24.5	276.5	0.885	0.870	0.990	0.990	0.800
		CR-SIS	6.5	26.5	60.0	147.0	803.5	0.780	0.760	0.700	0.800	0.385
		DC-SIS	4.0	4.0	4.0	4.0	4.0	1.000	1.000	1.000	1.000	1.000
		PC-SIS	4.0	4.0	4.0	4.0	4.0	1.000	1.000	1.000	1.000	1.000
	第二种情形	P-SIS	4.0	4.0	4.0	4.0	5.0	1.000	1.000	1.000	1.000	1.000
		FAST-SIS	4.5	12.0	34.0	113.5	434.5	0.915	0.850	0.805	0.850	0.550
		CRIS	4.0	4.0	6.0	17.0	378.5	0.910	0.885	0.970	0.980	0.820
		CR-SIS	8.0	33.5	75.0	195.5	993.0	0.740	0.715	0.680	0.760	0.290
		DC-SIS	4.0	4.0	4.0	4.0	4.5	1.000	1.000	1.000	1.000	1.000
		PC-SIS	4.0	4.0	4.0	4.0	5.0	1.000	1.000	1.000	1.000	1.000
	第三种情形	P-SIS	4.0	6.0	267.0	1,502.5	1,941.5	0.730	0.670	0.665	0.725	0.380
		FAST-SIS	132.5	730.0	1,261.5	1,685.0	1,956.0	0.255	0.170	0.170	0.190	0.020
		CRIS	4.0	5.0	11.5	67.0	590.0	0.885	0.830	0.885	0.960	0.660
		CR-SIS	69.0	386.5	1,087.0	1,705.5	1,936.5	0.285	0.295	0.225	0.330	0.020
		DC-SIS	4.0	4.0	4.0	302.0	1,546.5	0.780	0.760	0.725	0.800	0.650
		PC-SIS	4.0	4.0	4.0	4.0	5.0	1.000	1.000	0.995	1.000	0.995

表 2　转换模型下, 当 $\varepsilon \sim \mathbf{Cauchy}(0,1)$ 且删失率分别等于 20% 和 40% 时在 3 种情形下 6 种筛选方法的模拟结果

			Size					P_e				P_a
Cr		筛选方法	5%	25%	50%	75%	95%	X_1	X_2	X_9	X_{10}	
20%	第一种情形	P-SIS	4.0	4.0	4.0	6.0	23.0	1.000	0.995	0.980	1.000	0.975
		FAST-SIS	13.0	89.5	284.5	741.0	1,680.5	0.575	0.560	0.465	0.525	0.145
		CRIS	4.0	13.0	41.5	156.5	930.5	0.810	0.785	0.760	0.870	0.470
		CR-SIS	26.0	99.0	245.5	516.0	1,515.0	0.555	0.565	0.445	0.515	0.080
		DC-SIS	4.0	4.0	4.0	5.0	12.5	1.000	1.000	1.000	1.000	1.000
		PC-SIS	4.0	4.0	4.0	4.0	7.5	1.000	1.000	1.000	1.000	1.000
	第二种情形	P-SIS	4.0	4.0	4.0	7.0	37.5	0.990	0.985	0.985	0.990	0.950
		FAST-SIS	16.0	109.0	445.5	1,039.5	1,762.5	0.545	0.510	0.400	0.480	0.090
		CRIS	4.0	14.5	51.0	191.0	851.5	0.800	0.760	0.750	0.875	0.455
		CR-SIS	26.5	99.5	270.0	545.0	1,616.0	0.555	0.555	0.420	0.500	0.080
		DC-SIS	4.0	4.0	4.0	5.0	11.0	0.995	1.000	0.995	1.000	0.990
		PC-SIS	4.0	4.0	4.0	5.0	8.5	1.000	1.000	0.995	1.000	0.995
	第三种情形	P-SIS	6.0	16.0	64.5	277.5	989.5	0.665	0.555	0.520	0.610	0.150
		FAST-SIS	207	788.5	1,352.5	1,730.0	1,941.0	0.185	0.130	0.165	0.205	0.015
		CRIS	5.0	16.0	64.5	277.5	989.5	0.770	0.730	0.780	0.825	0.385
		CR-SIS	142.5	538.0	1,090.5	1,637.5	1,884.0	0.200	0.175	0.185	0.255	0.000
		DC-SIS	4.0	5.0	50.5	715.5	1,748.5	0.725	0.665	0.620	0.700	0.465
		PC-SIS	4.0	4.0	4.0	6.0	17.0	0.990	1.000	0.985	1.000	0.975
40%	第一种情形	P-SIS	4.0	4.0	4.0	4.0	9.5	1.000	1.000	0.995	1.000	0.995
		FAST-SIS	8.5	73.0	189.0	466.5	1,375.0	0.650	0.570	0.530	0.585	0.165
		CRIS	4.0	14.5	95.0	433.0	1,657.5	0.635	0.600	0.815	0.890	0.400
		CR-SIS	59.5	229.5	421.5	899.0	1,720.5	0.425	0.440	0.305	0.360	0.030
		DC-SIS	4.0	4.0	4.0	5.0	9.0	1.000	1.000	1.000	1.000	1.000
		PC-SIS	4.0	4.0	4.0	5.0	7.5	1.000	1.000	1.000	1.000	1.000
	第二种情形	P-SIS	4.0	4.0	4.0	5.0	18.5	1.000	1.000	0.985	0.995	0.980
		FAST-SIS	20.0	105.0	310.5	840.0	1,631.0	0.550	0.590	0.455	0.545	0.105
		CRIS	5.0	20.5	106.5	301.0	1,276.0	0.690	0.680	0.710	0.815	0.320
		CR-SIS	71.0	262.0	520.5	1,074.5	1,766.0	0.380	0.415	0.285	0.350	0.015
		DC-SIS	4.0	4.0	4.0	5.0	16.0	1.000	1.000	0.985	0.995	0.980
		PC-SIS	4.0	4.0	4.0	5.0	14.0	1.000	1.000	0.990	0.995	1.000
	第三种情形	P-SIS	6.5	81.0	723.5	1,577.5	1,883.5	0.645	0.530	0.545	0.620	0.170
		FAST-SIS	291.5	794.5	1,351.5	1,714.0	1,955.5	0.145	0.115	0.135	0.175	0.000
		CRIS	7.0	36.0	134.5	584.0	1,476.0	0.625	0.595	0.710	0.800	0.250
		CR-SIS	345.5	756.0	1,306.5	1,668.5	1,966.5	0.150	0.120	0.130	0.170	0.000
		DC-SIS	4.0	6.0	78.5	712.0	1,725.5	0.665	0.640	0.570	0.640	0.395
		PC-SIS	4.0	4.0	5.0	8.0	47.0	0.990	0.985	0.965	0.995	0.935

表 3 Cox 比例风险回归模型下, 当删失率分别等于 20% 和 40% 时在 3 种情形下 6 种筛选方法的模拟结果

			Size					P_e					P_a
Cr		筛选方法	5%	25%	50%	75%	95%	X_1	X_2	X_3	X_4	X_5	
20%	第一种情形	P-SIS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		FAST-SIS	5.0	5.0	7.0	15.0	63	0.960	1.000	1.000	1.000	0.950	0.910
		CRIS	5.0	5.0	5.0	7.0	87.5	0.950	1.000	0.995	0.995	0.950	0.910
		CR-SIS	5.0	12.0	33.0	122.5	456.0	0.750	0.925	0.960	0.930	0.760	0.520
		DC-SIS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		PC-SIS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
	第二种情形	P-SIS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		FAST-SIS	5.0	5.0	8.0	15.5	84.5	0.925	0.995	1.000	0.995	0.955	0.870
		CRIS	5.0	5.0	5.0	10.0	130.0	0.930	0.990	1.000	1.000	0.950	0.885
		CR-SIS	5.0	12.0	40.5	144.5	429.5	0.740	0.935	0.960	0.925	0.755	0.485
		DC-SIS	5.0	5.0	5.0	5.0	5.5	1.000	1.000	1.000	1.000	1.000	1.000
		PC-SIS	5.0	5.0	5.0	5.0	6.0	1.000	1.000	1.000	1.000	1.000	1.000
	第三种情形	P-SIS	5.0	11.0	546.0	1,739.0	1,975.5	0.665	0.730	0.780	0.730	0.635	0.315
		FAST-SIS	7.5	227.5	1,192.0	1,728.5	1,945.5	0.310	0.375	0.370	0.335	0.320	0.160
		CRIS	5.0	5.0	7.0	21.0	167.0	0.895	0.975	0.995	0.995	0.925	0.835
		CR-SIS	33.0	226.0	899.5	1,616.5	1,955.0	0.270	0.420	0.475	0.395	0.315	0.050
		DC-SIS	5.0	5.0	110.0	1,454.0	1,899.0	0.525	0.560	0.590	0.565	0.535	0.470
		PC-SIS	5.0	5.0	5.0	5.0	7.0	0.990	1.000	1.000	1.000	1.000	0.990
40%	第一种情形	P-SIS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		FAST-SIS	5.0	10.5	27.0	67.5	285.5	0.830	0.930	0.980	0.980	0.790	0.600
		CRIS	5.0	5.0	6.0	10.0	150.0	0.945	0.995	0.995	0.985	0.945	0.905
		CR-SIS	8.0	32.0	119.0	334.0	817.5	0.620	0.855	0.915	0.820	0.640	0.265
		DC-SIS	5.0	5.0	5.0	5.0	6.0	1.000	1.000	1.000	1.000	1.000	1.000
		PC-SIS	5.0	5.0	5.0	5.0	7.5	1.000	1.000	1.000	1.000	1.000	1.000
	第二种情形	P-SIS	5.0	5.0	5.0	5.0	5.0	1.000	1.000	1.000	1.000	1.000	1.000
		FAST-SIS	5.0	9.5	28.5	79.5	306.0	0.720	0.985	0.995	0.965	0.840	0.550
		CRIS	5.0	5.0	8.0	28.5	331.5	0.870	0.980	0.995	0.995	0.915	0.795
		CR-SIS	8.5	35.0	128.0	377.5	941.0	0.570	0.870	0.920	0.845	0.605	0.260
		DC-SIS	5.0	5.0	5.0	5.0	12.0	0.995	1.000	1.000	1.000	1.000	0.995
		PC-SIS	5.0	5.0	5.0	5.0	12.5	0.980	1.000	1.000	1.000	0.995	0.975
	第三种情形	P-SIS	5.0	37.0	737.0	1,797.0	1,953.5	0.615	0.695	0.760	0.700	0.610	0.250
		FAST-SIS	26.0	518.0	1,347.5	1,729.5	1,949.0	0.235	0.315	0.335	0.300	0.210	0.080
		CRIS	5.0	6.0	16.0	66.0	470.5	0.795	0.955	0.965	0.970	0.845	0.675
		CR-SIS	79.5	482.0	1,223.5	1,697.5	1,947.5	0.215	0.300	0.345	0.265	0.205	0.030
		DC-SIS	5.0	6.0	214.5	1,378.0	1,902.5	0.495	0.575	0.580	0.580	0.540	0.420
		PC-SIS	5.0	5.0	5.0	7.0	31.0	0.980	0.995	1.000	1.000	0.990	0.965

表 4　非线性模型下, 当 $\epsilon \sim N(0,1)$ 且删失率分别等于 20% 和 40% 时在 3 种情形下 6 种筛选方法的模拟结果

Cr		筛选方法	Size					P_e			P_a
			5%	25%	50%	75%	95%	X_1	X_5	X_{10}	
20%	第一种情形	P-SIS	104.5	443.0	959.5	1,535.0	1,849.5	0.985	1.000	0.015	0.015
		FAST-SIS	95.0	482.5	963.5	1,477.0	1,920.5	0.185	1.000	0.130	0.025
		CRIS	99.0	432.0	789.5	1,462.5	1,880.5	0.825	0.970	0.030	0.015
		CR-SIS	67.0	430.0	794.5	1,366.5	1,873.5	0.695	0.950	0.045	0.025
		DC-SIS	5.0	13.5	33.5	122.0	637.0	0.620	1.000	0.840	0.520
		PC-SIS	3.0	5.0	6.0	9.0	18.0	1.000	1.000	0.990	0.990
	第二种情形	P-SIS	117.0	520.0	1,031.0	1,561.0	1,857.0	0.995	1.000	0.010	0.010
		FAST-SIS	75.5	410.0	925.0	1,438.0	1,889.5	0.200	1.000	0.085	0.025
		CRIS	75.5	393.5	862.5	1,409.5	1,904.0	0.845	0.995	0.025	0.025
		CR-SIS	60.5	410.0	845.0	1,410.0	1,881.0	0.645	0.910	0.055	0.025
		DC-SIS	6.0	21.0	61.0	181.5	740.0	0.475	1.000	0.720	0.350
		PC-SIS	3.0	5.0	7.0	10.0	22.5	1.000	1.000	0.985	0.985
	第三种情形	P-SIS	97.5	747.5	1,333.0	1,670.0	1,940.5	0.420	0.880	0.020	0.015
		FAST-SIS	462.0	941.0	1,455.0	1,724.0	1,943.5	0.035	0.275	0.045	0.005
		CRIS	102.5	458.5	949.5	1,597.5	1,897.5	0.730	0.945	0.030	0.020
		CR-SIS	182.0	631.0	1,300.5	1,696.5	1,934.0	0.175	0.485	0.060	0.000
		DC-SIS	40.0	415.0	980.5	1,478.0	1,921.5	0.115	0.485	0.205	0.050
		PC-SIS	3.0	4.0	5.0	8.0	22.5	1.000	1.000	0.995	0.995
40%	第一种情形	P-SIS	111.0	422.5	904.0	1,390.0	1,852.5	0.990	1.000	0.010	0.010
		FAST-SIS	96.5	462.5	926.5	1,461.5	1,934.0	0.320	0.960	0.080	0.015
		CRIS	112.5	511.5	1,143.5	1,587.0	1,976.0	0.335	0.630	0.035	0.020
		CR-SIS	81.0	403.5	866.0	1,477.0	1,861.0	0.615	0.745	0.040	0.015
		DC-SIS	3.0	7.0	14.5	41.0	164.0	0.960	1.000	0.770	0.735
		PC-SIS	3.0	6.0	8.0	13.0	48.0	1.000	1.000	0.930	0.930
	第二种情形	P-SIS	173.0	666.0	1,146.0	1,519.0	1,866.0	1.000	1.000	0.005	0.005
		FAST-SIS	121.5	447.5	1,007.0	1,520.5	1,867.5	0.350	0.930	0.065	0.015
		CRIS	136.0	498.0	1,065.5	1,595.0	1,914.0	0.415	0.750	0.040	0.015
		CR-SIS	67.5	501.0	854.0	1,451.5	1,874.5	0.610	0.770	0.065	0.020
		DC-SIS	3.0	8.0	18.0	43.5	156.5	0.940	1.000	0.765	0.735
		PC-SIS	3.5	6.0	10.0	17.0	49.5	1.000	1.000	0.905	0.905
	第三种情形	P-SIS	76.0	757.0	1,373.0	1,651.0	1,921.5	0.515	0.890	0.025	0.025
		FAST-SIS	455.5	988.5	1,421.0	1,768.0	1,953.5	0.045	0.200	0.030	0.005
		CRIS	163.5	601.0	1,163.0	1,738.5	1,984.5	0.360	0.620	0.030	0.000
		CR-SIS	236.0	763.0	1,323.5	1,687.5	1,946.0	0.150	0.360	0.050	0.000
		DC-SIS	9.0	81.5	624.0	1,289.0	1,904.5	0.330	0.515	0.235	0.150
		PC-SIS	3.0	5.0	8.5	16.5	53.0	1.000	1.000	0.915	0.915

表 5 3 种模型下, 协变量维数 $p = 8{,}000$ 且删失率 20% 时 6 种筛选方法的模拟结果

		Size					P_e							P_a
	筛选方法	5%	25%	50%	75%	95%	X_1	X_2	X_3	X_4	X_5	X_9	X_{10}	
转换模型	P-SIS	4.0	44.0	2,237.5	6,430.5	7,732.0	0.8	0.5	–	–	–	0.6	0.7	0.3
	FAST-SIS	315.0	2,044.0	5,321.5	7,171.5	7,650.0	0.1	0.1	–	–	–	0.0	0.1	0.0
	CRIS	4.0	10.5	12.0	570.0	2,679.0	0.9	0.8	–	–	–	0.9	1.0	0.6
	CR-SIS	379.5	845.0	1,233.5	5,850.5	7,332.0	0.5	0.3	–	–	–	0.1	0.2	0.0
	DC-SIS	4.0	5.0	10.0	2,692.5	7,066.5	0.9	0.6	–	–	–	0.7	0.8	0.6
	PC-SIS	4.0	4.0	4.0	4.0	4.0	1.0	1.0	–	–	–	1.0	1.0	1.0
Cox 模型	P-SIS	5.0	45.0	3,258.0	6,258.5	7,136.0	0.6	0.7	0.5	0.3	0.5	–	–	0.4
	FAST-SIS	562.5	2,681.5	5,039.0	6,890.5	7,380.0	0.0	0.0	0.1	0.0	0.1	–	–	0.0
	CRIS	5.0	15.0	25.0	1,081.0	2,758.0	0.9	0.9	0.8	0.8	0.9	–	–	0.6
	CR-SIS	326.0	1,025.0	2,110.0	5,639.5	7,230.5	0.3	0.1	0.2	0.3	0.5	–	–	0.0
	DC-SIS	5.0	5.0	25.0	2,005.0	6,958.5	0.9	0.7	0.7	0.8	0.6	–	–	0.7
	PC-SIS	5.0	5.0	5.0	5.0	5.0	1.0	1.0	1.0	1.0	1.0	–	–	1.0
非线性模型	P-SIS	4,093.0	5,083.0	5,293.0	6,222.0	6,887.0	0.4	–	–	–	1.0	–	0.0	0.0
	FAST-SIS	2,027.5	2,625.0	2,906.0	3,240.0	3,721.0	0.0	–	–	–	0.2	–	0.0	0.0
	CRIS	3,281.5	3,894.0	4,396.0	5,058.0	5,542.5	0.8	–	–	–	1.0	–	0.0	0.0
	CR-SIS	1,333.5	1,944.0	2,322.0	3,305.0	4,052.0	0.2	–	–	–	0.4	–	0.0	0.0
	DC-SIS	1,347.0	1,362.0	3,116.0	4,498.0	4,620.0	0.0	–	–	–	0.6	–	0.2	0.0
	PC-SIS	4.0	5.0	5.0	7.0	10.0	1.0	–	–	–	1.0	–	1.0	1.0

为了进一步验证本文提出的 PC-SIS 方法的优势, 本文附加了一个模拟研究. 考虑在前述转换模型、Cox 模型和非线性模型中的第三种情形下, 误差项 ϵ 设置为 $N(0,1)$, 协变量维数设置为 $p = 8{,}000$, 删失率 Cr = 20%. 由于协变量维数 p 很大, 每次模拟生成随机数的时间很长, 考虑时间成本, 将模拟次数设置为 100, 其余设置不变. 模拟结果见表 5, 表 5 中的 “ – ” 代表相应的值不需要汇总. 删失率 Cr = 40% 的结果与 Cr = 20% 的结果类似, 其模拟结果类似表 5. 由表 5 可知, 当 $p = 8{,}000$ 时, 在所有 3 种不同模型下, PC-SIS 方法在特征筛选中仍然非常有效, 然而作为对比的 P-SIS 方法、FAST-SIS 方法、CRIS 方法、CR-SIS 方法和 DC-SIS 方法均在这种设置下没有优势.

4 基于生存数据的实证研究

本节将所提出的基于投影相关的超高维生存数据的特征筛选 PC-SIS 方法应用于弥漫大 B 细胞淋巴瘤生存数据 (DLBCL 数据集) 的实证研究中, 拟筛选出对 DLBCL 数据集中的患者最终生存结果有重大影响的基因. DLBCL 基因表达数据集和临床数据结果可查询 https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0020108. DLBCL 是在成年人群中最容易患的淋巴瘤之一, 即使患者接受化疗, 仍有高达 60% 的死亡率, 因此找出对影响患者最终生存结果有重大影响的基因十分必要. DLBCL 数据集包括美国 240 名患者在接受一定治疗后的生存时间 $\widetilde{T}$ 和 $p = 7{,}399$ 基因表达值. 在整个数据集中, 138 例患者死于 DLBCL, 102 例患者在最后一次随访时仍然存活, 删失率为 42.5%. 以观测到的生存时间和删失指示作为响应变量, 以 7,399 个基因表达水平为协变量, 分别采用前述已有模

拟中讨论过的 P-SIS 方法、FAST-SIS 方法、CRIS 方法、CR-SIS 方法、DC-SIS 方法和 PC-SIS 方法筛选出 DLBCL 数据集中对最终生存结果有重要影响的基因.

从 7,399 个基因中初步筛选出 DLBCL 数据集中 $\lfloor 240/\log 240 \rfloor = 43$ 个重要基因. 通过选出的结果可以看出, 本文提出的 PC-SIS 方法的重要基因筛选结果与其他 5 种方法在很大程度上有相似之处. 例如, 采用 PC-SIS 方法和 DC-SIS 方法可筛选出 28 个共同基因, 采用 PC-SIS 方法和 P-SIS 方法可筛选出 16 个共同基因, 采用 PC-SIS 方法和 FAST-SIS 方法可筛选出 5 个共同基因, 采用 PC-SIS 方法和 CRIS 方法可筛选出 6 个共同基因, 采用 PC-SIS 方法和 CR-SIS 方法仅筛选出 1 个共同基因. 为进一步筛选出重要基因, 本文通过 10 倍交叉验证选择调优参数分别拟合带 LASSO 惩罚、SCAD 惩罚和 MCP 的 Cox 比例风险模型, 对依据 PC-SIS 方法筛选出的 43 个重要基因继续进行进一步筛选. 最终将筛选出的重要基因编号、基因名称和相应回归系数估计列入表 6 中. 从表 6 的结果可知, 运用 LASSO 惩罚、SCAD 惩罚和 MCP 的正则化方法可共同选出 7 个重要基因, 这 7 个基因名称的描述如表 7 所示. 由于这 7 个基因被 3 种正则化方法同时选定, 表明这 7 个基因与患者的生存风险显著相关. 此外, 由表 6 中的系数估计结果可知, 基因编号 31981, 即基因 *BC012161* (隔膜蛋白 1 抗体) 对患者生存风险的影响最大, 这与 Li 和 Luan [12] 的分析结果一致.

为了评估本文所提出的 PC-SIS 方法的预测性能, 接下来将 240 名患者随机分为样本容量为 160

表 6　基于 LASSO 惩罚、SCAD 惩罚和 MCP 的正则化方法最终筛选出的基因编号、基因名称及其系数估计

LASSO			SCAD			MCP		
基因编号	基因名称	系数估计	基因编号	基因名称	系数估计	基因编号	基因名称	系数估计
31981	*BC012161*	0.370	31981	*BC012161*	0.380	31981	*BC012161*	0.400
24376	*BF129543*	−0.123	24376	*BF129543*	−0.145	24376	*BF129543*	−0.215
27592	*J03040*	−0.083	27592	*J03040*	−0.083	27592	*J03040*	−0.130
28641	*D13666*	−0.109	28641	*D13666*	−0.070	28641	*D13666*	−0.057
24432	*LC_24432*	0.076	24432	*LC_24432*	0.049	24432	*LC_24432*	0.053
27184	*S69790*	0.166	27184	*S69790*	0.067	27184	*S69790*	0.009
33358	*AA830781*	0.192	33358	*AA830781*	0.113	33358	*AA830781*	0.140
25116	*NM_014456*	0.066	25116	*NM_014456*	0.069	19373	*X00452*	−0.244
24203	*AI281624*	0.028	24203	*AI281624*	0.017			
31242	*U15552*	0.092	31242	*U15552*	0.036			
17646	*M14745*	0.057	17646	*M14745*	0.009			
19255	*AA283087*	−0.110	19255	*U70426*	−0.056			
28532	*M38690*	−0.081	28532	*M38690*	−0.006			
34364	*AI246189*	0.068	34364	*AI246189*	0.044			
24396	*M20430*	−0.131	24396	*M20430*	−0.288			
27612	*M25393*	0.032						
28325	*M16276*	−0.148						
29775	*AA214553*	−0.035						
26229	*LC_26229*	−0.040						
26146	*LC_26146*	0.029						
34729	*AF032885*	−0.076						

的训练集和样本容量为 80 的测试集. 应用上述 6 种筛选方法到训练集上, 选取前 $\lfloor 160/\log(160) \rfloor = 31$ 个基因, 并运用选出的这 31 个基因拟合 Cox 比例风险模型估计协变量系数. 在测试集中计算患者的风险评分, 将训练集中患者的风险评分的均值设定为阈值, 根据测试集中的风险得分和训练集所得的阈值将测试集中的患者分为高风险组和低风险组. 测试集中低风险组和高风险组的 Kaplan-Meier 生存曲线如图 1 所示. 从图 1 中的曲线可看出 PC-SIS 方法能有效分离两条曲线. 为使结果更具有说服

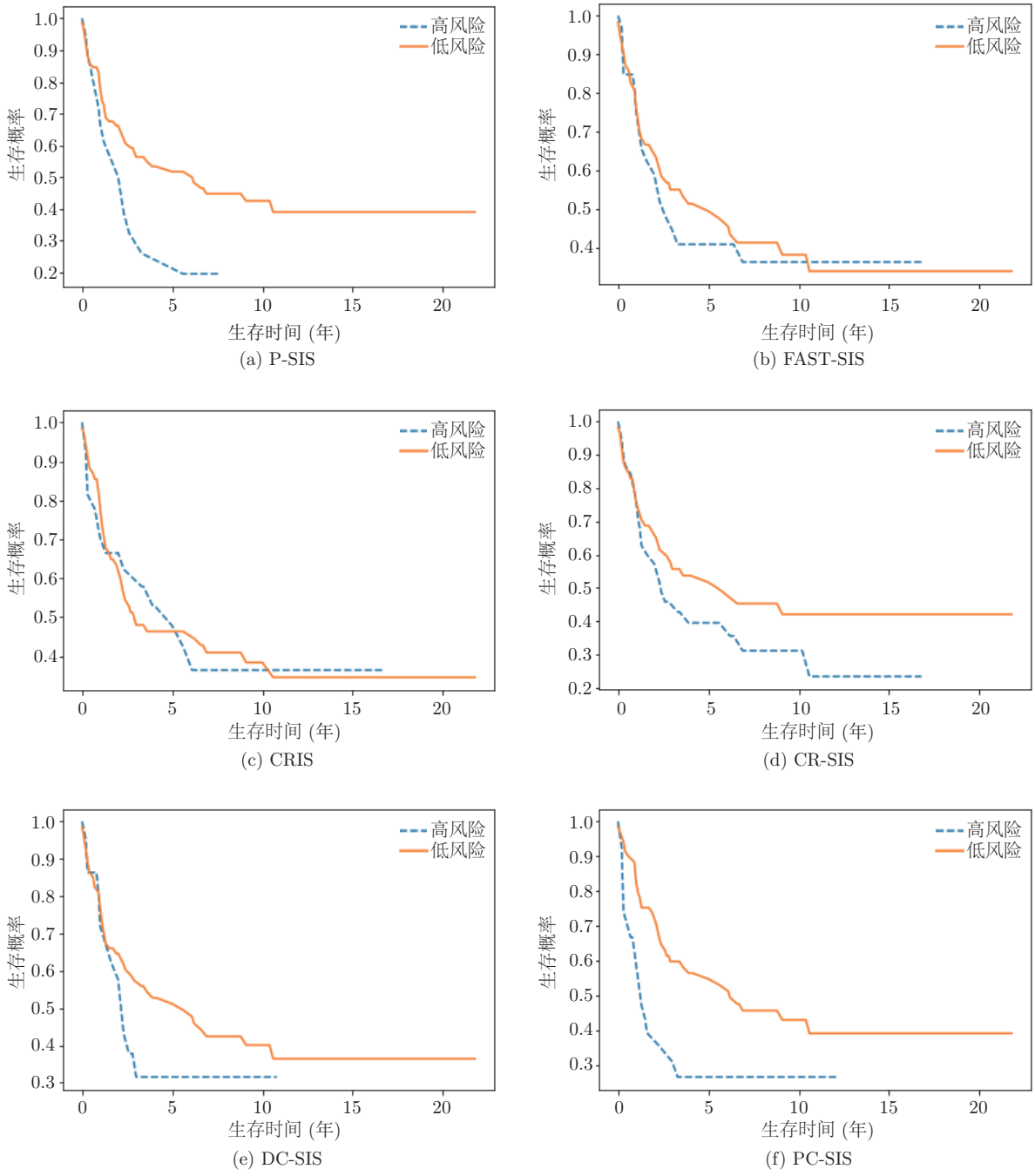

图 1 (网络版彩图) 测试数据集中 6 种筛选方法的低风险组和高风险组的 Kaplan-Meier 生存曲线

表 7 基于 LASSO 惩罚、SCAD 惩罚和 MCP 的正则化方法共同筛选出的 7 个基因名称的描述

基因名称	基因描述
BC012161	隔膜蛋白 1 抗体
BF129543	ESTs, 弱类似于 A47224 甲状腺素结合球蛋白前体
J03040	分泌蛋白、酸性, 富含半日光氨酸 (骨连接素)
D13666	成骨细胞特异性因子 2 (成束类蛋白 1)
LC_24432	没有描述
S69790	Hs.82318—WAS 蛋白族成员 3
AA830781	没有描述

力, 本文还进一步进行了 log-rank 检验, 分别采用 P-SIS 方法、FAST-SIS 方法、CRIS 方法、CR-SIS 方法、DC-SIS 方法和 PC-SIS 方法对得到的两条生存曲线是否存在显著差异进行检验. 上述 6 种方法的基于测试集的 log-rank 检验的 p 值分别为 0.0356、0.7490、0.9777、0.2075、0.2975 和 0.0067, 显然, PC-SIS 方法的 log-rank 检验的 p 值最小且小于 0.05, 表明 PC-SIS 方法具有良好的预测效果.

5 结论与展望

本文提出了一种新的无模型筛选 PC-SIS 方法. 首先, PC-SIS 筛选方法可以快速有效地降低超高维生存数据的维数, 同时保证以很大概率保留所有重要的特征. 其次, PC-SIS 方法不依赖于任何模型假设, 可适用于各种生存模型. 当数据存在异常值或厚尾的情形时, PC-SIS 方法具有很好的稳健性能. 另外, PC-SIS 方法不需要任何非参数估计和复杂的数值计算, 具有高效且易于实现的特点. 再次, 在较弱的正则化条件下, PC-SIS 方法拥有确定筛选性和秩相合性. 模拟研究验证表明, 本文给出的 PC-SIS 方法比其他已有的 5 种特征筛选方法—如基于模型的特征筛选 P-SIS 方法和 FAST-SIS 方法, 基于删失逆概率加权 Kendall τ 的特征筛选 CRIS 方法, 基于相关性排序的特征筛选 CR-SIS 方法, 以及基于距离相关的特征筛选 DC-SIS 方法等—对变量特征筛选更加稳健和有效. 最后, 基于 DLBCL 数据集中的临床数据结果的实证研究进一步验证了本文提出的 PC-SIS 方法在实际生活中具有可操作性.

PC-SIS 方法也有一定的局限性. PC-SIS 方法是在特征数量 p 较大而样本量 n 大小适中的情形下开发出的统计降维方法, 然而, 在科学研究中, 越来越常见 "大 p 大 n" 的数据集. 例如, 在现代全基因组遗传学研究中, 成千上万的参与者被分为数百万个单核苷酸多态性 (single nucleotide polymorphism, SNP). 在互联网研究中, 杀毒软件每分钟可能扫描数百万个统一资源定位 (uniform resource locator, URL) 中的数万个关键词. 当面对 "大 p 大 n" 的数据集时, 由于存储瓶颈和算法的可行性, 无模型特征筛选 PC-SIS 方法的效率可能较低, 因此, 如何对 "大 p 大 n" 的生存数据进行有效的特征筛选将是重要且有前景的方向 (参见文献 [15]).

致谢 感谢吴远山教授给予的建设性意见和建议. 感谢编委和各位审稿人给予的指导和建议.

参考文献

1 Candés E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. Ann Statist, 2007, 35: 2313–2351

2 Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. J Amer Statist Assoc, 2011, 106: 544–557

3 Fan J, Feng Y, Wu Y. High-dimensional variable selection for Cox's proportional hazards model. Inst Math Stat Collect, 2010, 6: 70–86
4 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc, 2001, 96: 1348–1360
5 Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol, 2008, 70: 849–911
6 Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Statist Sinica, 2010, 20: 101–148
7 Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection: Beyond the linear model. J Mach Learn Res, 2009, 10: 2013–2038
8 Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. Ann Statist, 2010, 38: 3567–3604
9 Gorst-Rasmussen A, Scheike T. Independent screening for single-index hazard rate models with ultrahigh dimensional features. J R Stat Soc Ser B Stat Methodol, 2013, 75: 217–245
10 He S M, Xie J Q. A feature screening procedure for ultra-high dimensional multi-category discriminant analysis. J Appl Stat Manag, 2021, 40: 679–691 [何胜美, 谢家泉. 超高维多类判别分析的特征筛选方法研究. 数理统计与管理, 2021, 40: 679–691]
11 Li G, Peng H, Zhang J, et al. Robust rank correlation based screening. Ann Statist, 2012, 40: 1846–1877
12 Li H, Luan Y. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. Bioinformatics, 2005, 21: 2403–2409
13 Li J, Zheng Q, Peng L, et al. Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. Biometrics, 2016, 72: 1145–1154
14 Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. J Amer Statist Assoc, 2012, 107: 1129–1139
15 Li X X, Li R Z, Xia Z M, et al. Distributed feature screening via componentwise debiasing. J Mach Learn Res, 2020, 21: 852–883
16 Lin Y, Liu X, Hao M. Model-free feature screening for high-dimensional survival data. Sci China Math, 2018, 61: 1617–1636
17 Lu J, Hu Q Q, Lin L. Feature screening for multi-response ultrahigh-dimensional linear models by empirical likelihood. Sci Sin Math, 2023, 53: 499–522 [陆军, 胡琴琴, 林路. 多元响应变量超高维线性模型的经验似然特征筛选方法. 中国科学: 数学, 2023, 53: 499–522]
18 Mai Q, Zou H. The fused Kolmogorov filter: A nonparametric model-free screening method. Ann Statist, 2015, 43: 1471–1497
19 Pan Y. Feature screening and FDR control with knockoff features for ultrahigh-dimensional right-censored data. Comput Statist Data Anal, 2022, 173: 107504
20 Serfling R J. Approximation Theorems of Mathematical Statistics. New York: John Wiley & Sons, 1980
21 Song R, Lu W, Ma S, et al. Censored rank independence screening for high-dimensional survival data. Biometrika, 2014, 101: 799–814
22 Székely G J, Rizzo M L. Brownian distance covariance. Ann Appl Stat, 2009, 3: 1236–1265
23 Tibshirani R J. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Stat Methodol, 1996, 58: 267–288
24 Tibshirani R J. Univariate shrinkage in the Cox model for high dimensional data. Stat Appl Genet Mol Biol, 2009, 8: 1–18
25 Wu Y, Yin G. Conditional quantile screening in ultrahigh-dimensional heterogeneous data. Biometrika, 2015, 102: 65–76
26 Zhang C H. Nearly unbiased variable selection under minimax concave penalty. Ann Statist, 2010, 38: 894–942
27 Zhang J, Liu Y, Cui H. Model-free feature screening via distance correlation for ultrahigh dimensional survival data. Statist Papers, 2021, 62: 2711–2738
28 Zhang J, Liu Y, Wu Y. Correlation rank screening for ultrahigh-dimensional survival data. Comput Statist Data Anal, 2017, 108: 121–132
29 Zhang J, Yin G, Liu Y, et al. Censored cumulative residual independent screening for ultrahigh-dimensional survival data. Lifetime Data Anal, 2018, 24: 273–292
30 Zhang J, Zhou H, Liu Y, et al. Feature screening for case-cohort studies with failure time outcome. Scand J Stat, 2021, 48: 349–370
31 Zhao S D, Li Y. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. J Multivariate Anal, 2012, 105: 397–411
32 Zhu L, Li L, Li R, et al. Model-free feature screening for ultrahigh-dimensional data. J Amer Statist Assoc, 2011, 106:

1464–1475
33 Zhu L, Xu K, Li R, et al. Projection correlation between two random vectors. Biometrika, 2017, 104: 829–843
34 Zou H. The adaptive Lasso and its oracle properties. J Amer Statist Assoc, 2006, 101: 1418–1429

A new feature screening method for ultra-high-dimensional survival data based on projection correlation

Yingli Pan, Xiangyu Ge & Yanli Zhou

Abstract In this paper, a projective correlation method with sure independent screening (abbreviated to the PC-SIS method) is proposed for feature screening of ultra-high-dimensional right-censored survival data. On the one hand, the PC-SIS method does not require any model to be specified, nor does it require non-parametric estimation of the survival function, and it is insensitive to moment conditions and sub-exponential conditions, so it is applicable to analyze the data with outliers or heavy-tailed. On the other hand, under certain regularization conditions, the PC-SIS method has sure screening and rank consistency properties. A simulation and an empirical study show that the PC-SIS method can eliminate features with weak correlation with response variables on the premise of preserving all important features to achieve the purpose of dimension reduction.

Keywords **projection correlation, rank consistency, sure screening, survival data, ultra-high dimensionality**
MSC(2020) **62H12, 62H20**
doi: 10.1360/SCM-2023-0067