



基于自适应阈值选取的超高维生存数据特征筛选与 FDR 控制

张婧¹, 崔恒建^{2*}

1. 中南财经政法大学统计与数学学院, 武汉 430073;

2. 首都师范大学数学科学学院, 北京 100875

E-mail: jing66@zuel.edu.cn, hjcui@bnu.edu.cn

收稿日期: 2025-02-06; 接受日期: 2025-12-31; 网络出版日期: 2026-01-29; * 通信作者

国家社会科学基金 (批准号: 24CTJ036)、国家自然科学基金 (批准号: 12031016, 12531012, 11971324 和 11901581) 和首都师范大学交叉研究院生物信息学与统计学交叉建设资助项目

摘要 在超高维生存数据分析中, 识别对患者生存时间具有显著影响的关键因素, 对于优化治疗策略和推动精准医疗具有重要意义. 特征筛选是解决此类问题的有效方法, 但其中一个关键问题是, 如何建立合适的阈值方法来区分重要变量和不重要变量. 本文利用数据分割策略和重组技术建立一种数据驱动的阈值选取方法, 使其在不同情境下能选取合适的阈值, 有效区分重要变量和不重要变量. 进一步地, 基于所选阈值提出适用于超高维右删失生存数据的特征筛选方法. 该方法的显著优点是, 不仅能够在快速有效降维的同时准确保留所有重要变量, 而且可以把错误发现率 (FDR) 控制在一定范围内. 在一定的正则化条件下, 本文建立该方法的理论框架, 包括确定筛选性质、排序秩相合性、有限样本下的 FDR 界限理论和渐近 FDR 控制理论. 数值模拟结果显示, 此方法在变量选择和 FDR 控制方面均有优良表现, 且效果明显优于现有的其他方法. 最后, 本文将该方法应用于弥漫性大 B 细胞淋巴瘤数据集的分析中, 进一步验证了其有效性和实用性.

关键词 错误发现率控制 特征筛选 阈值选择 超高维右删失生存数据

MSC (2020) 主题分类 62N01, 62P10, 62R07

1 引言

随着科技的迅速发展, 超高维数据广泛出现在经济学、生物医学、基因组学等众多科学领域. 例如, 在基因组学中, 核糖核酸 (RNA) 测序或全基因组测序通常生成包含数万个特征的大规模数据. 生物医学影像技术, 如磁共振成像 (MRI) 和计算机断层扫描 (CT) 扫描, 也会产生复杂的高维数据集. 超高维数据的一个显著特点是, 协变量个数 p 远远大于样本量 n , 且可以随着 n 指数阶增长. 在稀疏性假设下, 在众多协变量中只有少数协变量对感兴趣事件有显著影响. 因此, 如何准确识别这些重

英文引用格式: Zhang J, Cui H J. Feature screening and false discovery rate control for ultrahigh-dimensional survival data based on adaptive threshold selection (in Chinese). Sci Sin Math, 2026, 56: 1–22, doi: [10.1360/SSM-2025-0025](https://doi.org/10.1360/SSM-2025-0025)

要变量, 成为超高维数据分析中的一个重要且核心问题. 为了解决这一问题, Fan 和 Lv [8] 首次提出了“独立特征筛选”的思想, 并在线性回归模型的框架下建立了确定独立特征筛选 (SIS) 方法. 此方法可以在快速有效降维的同时准确保留所有重要变量. 此后, SIS 方法被推广到不同的模型和数据结构中. 例如, Zhao 和 Li [23] 提出了 Cox 比例风险模型下处理超高维生存数据的特征筛选方法, Gorst-Rasmussen 和 Scheike [9] 提出了单指标危险率模型下超高维生存数据的特征筛选方法. Song 等 [18]、Zhang 等 [21]、Zhou 和 Zhu [24]、Liu 等 [14]、Zhang 等 [22]、Pan 等 [16] 和 Zhang 等 [20] 进一步提出了不依赖于任何模型假设的特征筛选方法. 这些方法突破了模型的限制, 应用范围更加广泛.

特征筛选方法的核心思想是评估每个协变量对响应变量的边际影响, 将那些影响显著的协变量视为重要变量, 而将影响较小的协变量视为不重要变量. 为了量化每个协变量的影响, 需要定义一个非负统计量指标 ω_j , 并根据其取值大小决定是否保留或舍弃对应的协变量 Z_j . 假设 $\hat{\omega}_j$ 为基于观测数据得到的 ω_j 的估计值, 则所选重要变量指标集可定义为

$$\hat{A}(L) = \{j : \hat{\omega}_j \geq L, j = 1, \dots, p\},$$

其中 L 是用于区分重要变量和不重要变量的阈值, 它的选取至关重要. 然而, L 的理论值通常依赖于协变量与响应变量联合分布中的一些未知参数, 这些参数往往很难获取. 因此, 现有文献通常采用 Fan 和 Lv [8] 提出的硬阈值准则, 即根据影响程度从大到小对协变量进行排序, 选取前 $[n/\log(n)]$ 个变量作为重要变量. 此准则虽然应用简单, 但这种通用的方式在许多情况下效果并不理想, 经常会误选许多不重要变量, 导致错误发现率 (FDR) 过高. 在特征筛选方法中, 阈值的选择是一个关键问题. 过大的阈值可能导致一些重要变量的漏选, 无法满足确定筛选性质; 而过小的阈值则可能保留过多不重要变量, 进而增加 FDR. 因此, 如何选择合适的阈值 L , 既能保留所有重要变量, 又能实现 FDR 控制, 成为特征筛选方法中的一个重要挑战. 现有关于 FDR 控制的阈值选取方法大致可以分为两类. 第一类方法通常会建立一个基准模型, 与此模型相比具有较高边际效应的协变量会被选出. 例如, Zhu 等 [25] 提出了一种基于辅助变量的阈值选择方法, Fan 等 [7]、Barut 等 [3] 和 Pan 等 [16] 也借助辅助变量提出了类似的阈值选择方法. 然而, 这些研究未深入探讨 FDR 控制的理论性质. 第二类方法则利用边际效应的特殊性质来构造 FDR 的估计量, 并选择符合预期 FDR 的协变量. 例如, Zhao 和 Li [23] 以及 Barut 等 [3] 借助 t 检验来实现 FDR 控制, Liu 等 [13] 利用 knockoff 特征来确定阈值并实现 FDR 控制, Guo 等 [10]、Du 等 [6] 和 He 等 [11] 采用样本分割策略, 构造了具有对称性质的统计量来选取阈值. 目前, 关于 FDR 控制的阈值选取方法的研究已取得显著进展, 但大多数研究集中在完整数据或线性模型上, 对删失生存数据的研究很少. 在生物医学和临床试验中, 人们经常收集到超高维的删失生存数据. 除了维度极高的特点外, 这类数据的响应变量 (如患者的生存时间 T) 由于存在删失, 通常无法被完全观测, 这使得对其统计分析更加具有挑战性. 现有对超高维删失生存数据进行降维的特征筛选方法 [9, 14, 16, 21, 22, 24] 大多采用硬阈值规则, 并未考虑 FDR 控制的问题. 因此, 如何在超高维删失生存数据中设计有效的阈值选择方法, 并在此基础上建立高效的特征筛选方法, 以确保在快速降维的同时既能保留所有重要变量又能实现 FDR 控制, 成为一个重要且具有实际意义的研究问题.

本文在 Cox 比例风险模型的框架下, 研究超高维右删失生存数据的特征筛选与 FDR 控制问题. 首先, 建立数据驱动的阈值选取方法, 使其在不同情境下能选取合适的阈值, 更好地区分重要变量和不重要变量. 其核心思想是利用数据分割策略和重组技术构造具有对称性质的统计量, 然后利用此统计量的对称性得到错误发现数量的近似值, 进而获取合适的阈值. 其次, 基于所选阈值提出适用于超高维右删失生存数据且能实现 FDR 控制的变量筛选方法. 此方法的突出优点是, 不仅能够在快速有效降维的同时保留所有重要变量, 而且可以将 FDR 控制在预设的范围内. 进一步地, 在一定的正则化

条件下建立所提方法的理论框架. 本文不仅证明了特征筛选的理论性质, 如确定筛选性质和排序秩相容性; 还证明了 FDR 控制的理论性质, 包括有限样本下的 FDR 界限理论和渐近 FDR 控制理论. 接下来, 通过数值模拟实验验证该方法在有限样本下的表现, 并与已有的方法进行对比, 结果发现此方法在变量筛选和 FDR 控制方面的性能明显优于现有的其他方法. 最后, 将提出的方法应用到弥漫性大 B 细胞淋巴瘤 (DLBCL) 数据集的分析中, 可以有效识别与患者生存时间密切相关的基因, 并预测患者未来的生存风险. 该方法计算过程简单且高效, 为其在实际问题中的应用提供了极大的便利.

本文的余下内容安排如下. 第 2 节首先构建一种数据驱动的阈值选取方法, 并在此基础上提出适用于超高维右删失生存数据且能实现 FDR 控制的变量筛选方法. 第 3 节在适当的正则化条件下建立所提方法的理论框架. 第 4 节通过大量数值模拟, 系统评估所提方法在有限样本下的表现. 第 5 节将此方法应用于弥漫性大 B 细胞淋巴瘤研究的真实数据分析中, 验证其实用性. 第 6 节总结全文, 并探讨本文所提方法的潜在拓展方向.

2 阈值选取方法的提出和特征筛选方法的建立

2.1 模型和数据

令 T 表示失效时间, C 表示删失时间, $\mathbf{Z} = (Z_1, \dots, Z_p)^T$ 表示 p 维协变量向量, $X = \min(T, C)$ 为实际观测时间, $\Delta = I(T \leq C)$ 为删失示性变量, $I(\cdot)$ 为示性函数. 不失一般性, 本文假设给定协变量 $\mathbf{Z}$ 时 T 和 C 是相互独立的. Cox 比例风险模型^[4] 假设给定协变量 $\mathbf{Z}$ 时生存时间 T 的条件危险率函数形式如下:

$$\lambda(t | \mathbf{Z}) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}),$$

其中 $\lambda_0(t)$ 为基准危险率函数, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ 是 p 维未知参数. 在超高维数据中, 协变量的维数 p 远远大于样本量 n , 且可以允许它随着 n 指数阶增长. 根据稀疏性假设, 在众多协变量中只有少部分协变量对失效时间 T 有显著影响, 这些协变量被称为重要协变量. 令 $\mathcal{A} = \{j : \beta_j \neq 0, j = 1, \dots, p\}$ 表示重要变量指标集, $\mathcal{A}^c = \{1, \dots, p\} \setminus \mathcal{A}$ 表示不重要变量指标集. 本文的研究目的是找到协变量的一个指标集 $\hat{\mathcal{A}}$, 使其能够包含所有重要变量且满足

$$\lim_{n \rightarrow \infty} \sup E \left\{ \frac{\#\{j : j \in \hat{\mathcal{A}} \cap \mathcal{A}^c\}}{\#\{j : j \in \hat{\mathcal{A}}\} \vee 1} \right\} \leq \alpha,$$

其中 α 为事先设定的阈值, $\#\{\cdot\}$ 表示集合中元素个数.

2.2 特征筛选方法的建立

为了实现这个目标, 本文利用样本拆分策略和重组技术构建具有对称性质的新统计量, 然后基于该统计量的对称性建立数据自适应阈值选择方法, 最后基于所选阈值提出一种处理超高维右删失生存数据的变量筛选方法. 该方法在数据降维过程中可以有效保留所有重要变量且实现 FDR 控制. 假设此研究包含 n 个独立个体, 观测数据可表示为

$$\{X_i, \Delta_i, \mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})^T : i = 1, \dots, n\}.$$

首先, 将数据分为两个独立的子集 D_1 和 D_2 , 分别表示为

$$(X_{D_1}, \Delta_{D_1}, \mathbf{Z}_{D_1}) \text{ 和 } (X_{D_2}, \Delta_{D_2}, \mathbf{Z}_{D_2}),$$

样本量分别为 n_1 和 $n_2 = n - n_1$. 接下来, 利用第一部分数据集 D_1 进行初步变量选择, 在所选变量中拟合 Cox 模型, 并构造具有对称性质的统计量以量化每个协变量的影响效应. 通过数据驱动的方法选取合适的阈值, 以实现目标 FDR 水平的控制. 具体而言, 所提方法包括以下步骤.

2.2.1 对称统计量的构造

首先, 基于数据集 D_1 来建立统计量. 分别考虑超高维数据和高维数据两种情形. 对于维数远大于样本量 n 的超高维数据 (例如, $p = O(\exp(n^\alpha))$ 且 $\alpha > 0$), 首先利用生存数据中超高维变量筛选方法, 如 Zhao 和 Li^[23] 提出的变量筛选方法 PSIS (principled sure independence screening procedure), 进行初步的数据降维, 再利用正则化方法 LASSO (least absolute shrinkage and selection operator) 进一步选择变量. 令 I 表示变量筛选后剩余协变量的指标集, $\mathbf{Z}_I = (\mathbf{Z}_j, j \in I)$, $p_0 = |I|$ 表示剩余协变量的数量. 基于观测数据 $(X_{D_1}, \Delta_{D_1}, \mathbf{Z}_{D_1, I})$, 似然函数可表示为

$$l_1(\boldsymbol{\beta}) = \prod_{i \in D_1} \left\{ \frac{\exp(\boldsymbol{\beta}^T \mathbf{Z}_{i, I})}{\sum_{j \in D_1} I(X_j > X_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_{j, I})} \right\}^{\Delta_i}.$$

正则化方法 LASSO 对应的目标函数表示为

$$L_1(\boldsymbol{\beta}) = \log(l_1(\boldsymbol{\beta})) - \lambda \|\boldsymbol{\beta}\|_1 = \sum_{i \in D_1} \Delta_i \left\{ \boldsymbol{\beta}^T \mathbf{Z}_{i, I} - \log \left(\sum_{j \in D_1} I(X_j > X_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_{j, I}) \right) \right\} - \lambda \|\boldsymbol{\beta}\|_1.$$

极大化上面的目标函数可以得到 LASSO 估计值

$$\hat{\boldsymbol{\beta}}_{1, \text{LASSO}} = (\hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \dots, \hat{\beta}_{1,p_0})^T = \arg\max_{\boldsymbol{\beta}} L_1(\boldsymbol{\beta}).$$

基于第一部分数据集 D_1 所建立的统计量表示为 $\hat{\boldsymbol{\mu}}_1 = (\hat{\mu}_{11}, \dots, \hat{\mu}_{1p})^T$, 其中

$$\hat{\mu}_{1j} = \begin{cases} \hat{\beta}_{1,j}, & j \in I, \\ 0, & j \in I^c. \end{cases} \quad (2.1)$$

对于高维数据 (例如, $p = O(n^\alpha)$ 且 $\alpha > 1$), 直接采用正则化方法 LASSO 来进行变量选择. 基于观测数据 $(X_{D_1}, \Delta_{D_1}, \mathbf{Z}_{D_1})$, 惩罚对数偏似然函数可表示为

$$L_1^*(\boldsymbol{\beta}) = \sum_{i \in D_1} \Delta_i \left\{ \boldsymbol{\beta}^T \mathbf{Z}_i - \log \left(\sum_{j \in D_1} I(X_j > X_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_j) \right) \right\} - \lambda \|\boldsymbol{\beta}\|_1.$$

通过极大化惩罚对数偏似然函数可得到 LASSO 估计值

$$\hat{\boldsymbol{\beta}}_{1, \text{LASSO}}^* = (\hat{\beta}_{1,1}^*, \hat{\beta}_{1,2}^*, \dots, \hat{\beta}_{1,p}^*)^T = \arg\max_{\boldsymbol{\beta}} L_1^*(\boldsymbol{\beta}).$$

此时, 基于第一部分数据集 D_1 所建立的统计量表示为

$$\hat{\boldsymbol{\mu}}_1 = (\hat{\mu}_{11}, \dots, \hat{\mu}_{1p})^T = \hat{\boldsymbol{\beta}}_{1, \text{LASSO}}^*. \quad (2.2)$$

接下来, 基于数据集 D_2 来建立统计量. 令

$$S = \{j : \hat{\beta}_j \neq 0, j = 1, \dots, p\}$$

表示基于数据集 D_1 所选出的重要变量指标集,

$$S^c = \{1, 2, \dots, p\} \setminus S$$

表示未选中的不重要变量指标集. 再基于 S 中变量拟合 Cox 模型, 利用数据集 D_2 来计算极大似然估计. 令

$$\mathbf{Z}_S = (Z_j, j \in S),$$

基于数据集 $(X_{D_2}, \Delta_{D_2}, \mathbf{Z}_{D_2, S})$, 极大似然估计值

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_s)^T$$

定义为

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = \operatorname{argmax}_{\boldsymbol{\beta}} l_2(\boldsymbol{\beta}),$$

其中

$$l_2(\boldsymbol{\beta}) = \prod_{i \in D_2} \left\{ \frac{\exp(\boldsymbol{\beta}^T \mathbf{Z}_{i, S})}{\sum_{j \in D_2} I(X_j > X_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_{j, S})} \right\}^{\Delta_i}.$$

因此, 基于第二部分数据集 D_2 可建立统计量 $\hat{\boldsymbol{\mu}}_2 = (\hat{\mu}_{21}, \dots, \hat{\mu}_{2p})^T$, 其中

$$\hat{\mu}_{2j} = \begin{cases} \hat{\beta}_j, & j \in S, \\ 0, & j \in S^c. \end{cases} \quad (2.3)$$

最后, 将基于数据集 D_1 和 D_2 所建立的统计量融合起来. 对于第 j 个协变量 Z_j , 基于 D_1 和 D_2 所建立的边际效应统计量分别记为 $\hat{\mu}_{1j}$ 和 $\hat{\mu}_{2j}$, 将二者融合建立新的统计量如下:

$$\widehat{W}_j = \hat{\mu}_{1j} \cdot \hat{\mu}_{2j}, \quad j = 1, \dots, p. \quad (2.4)$$

直观来看, 若 Z_j 是重要变量, 则 $\hat{\mu}_{1j}$ 和 $\hat{\mu}_{2j}$ 通常取值较大且符号相同, 因此 $\widehat{W}_j$ 的值较大; 若 Z_j 是不重要变量, 后面的引理 3.1 将证明当 $j \in \mathcal{A}^c$ 时 $\widehat{W}_j$ 满足对称性. 综上, 基于 $\widehat{W}_j$ 的取值可以有效区分重要变量和不重要变量.

注 2.1 (2.4) 中 $\widehat{W}_j$ 的构建方式并不唯一, 还可以选择其他方式来构建 $\widehat{W}_j$. 例如, $\widehat{W}_j = |\hat{\mu}_{1j} + \hat{\mu}_{2j}| - |\hat{\mu}_{1j} - \hat{\mu}_{2j}|$ 或 $\widehat{W}_j = \operatorname{sign}(\hat{\mu}_{1j} \cdot \hat{\mu}_{2j})(|\hat{\mu}_{1j}| + |\hat{\mu}_{2j}|)$. 注意, 在构造 $\widehat{W}_j$ 时必须保证, 当 $j \in \mathcal{A}^c$ 时, $\widehat{W}_j$ 满足对称性或近似对称性.

2.2.2 阈值选取方法的建立和变量筛选方法的提出

基于上面建立的新统计量 $\widehat{W}_j$, 给定阈值 $t > 0$, 所选重要变量指标集定义为

$$\widehat{\mathcal{A}}(t) = \{j : \widehat{W}_j \geq t, j = 1, \dots, p\}.$$

$\widehat{\mathcal{A}}(t)$ 中的错误发现比率 (FDP) 可表示为

$$\text{FDP}(t) = \frac{\#\{j \in \mathcal{A}^c : \widehat{W}_j \geq t\}}{\#\{j : \widehat{W}_j \geq t\} \vee 1}.$$

根据 $\widehat{W}_j$ 的定义可知 $\#\{j \in \mathcal{A} : \widehat{W}_j \leq -t\}$ 非常小. 利用 $\widehat{W}_j$ ($j \in \mathcal{A}^c$) 的对称性可知

$$\#\{j \in \mathcal{A}^c : \widehat{W}_j \geq t\} \approx \#\{j \in \mathcal{A}^c : \widehat{W}_j \leq -t\}.$$

因此, 可以得到 $\text{FDP}(t)$ 的一个保守估计如下:

$$\widehat{\text{FDP}}(t) = \frac{\#\{j : \widehat{W}_j \leq -t\}}{\#\{j : \widehat{W}_j \geq t\} \vee 1}.$$

为了将 FDR 控制在给定水平 α 之下, 本文提出如下数据驱动的阈值选取方法来选择合适的阈值 $L > 0$:

$$L = \inf \left\{ t > 0 : \frac{1 + \#\{j : \widehat{W}_j \leq -t\}}{\#\{j : \widehat{W}_j \geq t\} \vee 1} \leq \alpha \right\}. \quad (2.5)$$

进一步地, 所选重要变量指标集可定义为

$$\widehat{\mathcal{A}}(L) = \{j : \widehat{W}_j \geq L, j = 1, \dots, p\}. \quad (2.6)$$

称这种方法为基于数据划分和融合的特征筛选方法 (FSDS).

2.2.3 基于 FSDS 的硬阈值筛选方法

在上文所建立的 FSDS 方法中, 我们是基于数据分割后所建立的新统计量 $\widehat{W}_j$ 来进行变量筛选. 在实际应用中, 人们可能更倾向于使用原始全样本数据去建立筛选统计量以提高效率. 接下来, 将 FSDS 方法和硬阈值筛选思路相结合, 提出基于 FSDS 的硬阈值筛选方法.

首先, 利用 FSDS 方法确定筛选阈值 L . 具体地, 将数据集随机分为 D_1 和 D_2 两部分, 分别计算统计量 $\widehat{\mu}_1$ 和 $\widehat{\mu}_2$, 然后按照 (2.4) 将其融合得到新的统计量 $\widehat{W}_j$, 再根据 (2.5) 确定合适的阈值 L . 接下来, 定义所选重要变量的指标集为

$$\widehat{\mathcal{A}}(L) = \{j : \widehat{W}_j \geq L, j = 1, \dots, p\},$$

并计算该集合中的元素个数 $|\widehat{\mathcal{A}}(L)|$. 然后, 基于原始全样本数据构建新的边际效应统计量 $\widehat{\omega}_j$, 并从中选择前 $|\widehat{\mathcal{A}}(L)|$ 个具有最大取值的变量作为重要变量. 此处通过硬阈值 $|\widehat{\mathcal{A}}(L)|$ 进行特征筛选等价于使用如下定义的阈值 L^* :

$$L^* = \widehat{\omega}_{(p-|\widehat{\mathcal{A}}(L)|+1)},$$

其中 $\widehat{\omega}_{(1)} \leq \dots \leq \widehat{\omega}_{(p)}$ 表示 $\widehat{\omega}_j$ 的顺序统计量. 由于使用全样本信息来构建筛选统计量 $\widehat{\omega}_j$, 可以预期其能够更好地实现确定筛选性质. 称这个方法为基于 FSDS 的硬阈值筛选方法 (HT-FSDS). 这里可以使用已有文献 [9, 14, 16, 18, 21–24] 中的变量筛选方法来构建筛选统计量 $\widehat{\omega}_j$. HT-FSDS 方法能够有效处理 Cox 比例风险模型下超高维右删失数据中的变量筛选问题, 并实现对 FDR 的有效控制.

3 FSDS 方法的理论性质

本节首先证明 FSDS 方法的变量筛选性质, 包括确定筛选性质和排序秩相合性. 然后, 建立 FSDS 方法的 FDR 控制理论, 包括 FDR 界的有限样本理论以及 FDR 控制的渐近理论. 假设 $\mathbf{a} =$

$(a_1, \dots, a_p)^T$ 和 $\mathbf{b} = (b_1, \dots, b_p)^T$ 均为 p 维向量, $\mathbf{a} \otimes \mathbf{b}$ 表示矩阵 $\mathbf{ab}^T$, $\mathbf{a}^{\otimes 2}$ 表示矩阵 $\mathbf{a} \otimes \mathbf{a}$. 令 $W_j = \beta_j^2$ ($j = 1, 2, \dots, p$), $N_i(t) = I(T_i \leq t, \Delta_i = 1)$, $Y_i(t) = I(T_i \geq t)$, τ 表示研究结束时间.

$$S = \{j : \hat{\beta}_j \neq 0, j = 1, \dots, p\}$$

表示在第一部分数据集 D_1 中基于正则化方法 LASSO 所选变量集合. 定义

$$\begin{aligned} S^{(l)}(\beta, t) &= \sum_{i=1}^{n_2} \mathbf{Z}_{i,S}^{\otimes l} Y_i(t) \exp(\beta^T \mathbf{Z}_{i,S}) / n_2 \quad (l = 0, 1, 2), \\ E(\beta, t) &= S^{(1)}(\beta, t) / S^{(0)}(\beta, t), \\ V(\beta, t) &= S^{(2)}(\beta, t) / S^{(0)}(\beta, t) - E(\beta, t)^{\otimes 2}. \end{aligned}$$

在整个证明过程中, 假设 FSDS 方法满足下面正则化条件.

$$(C1) \int_0^\tau \lambda_0(t) dt < \infty.$$

(C2) 存在 β 的一个邻域 $\mathcal{B}$, 以及定义在 $\mathcal{B} \times [0, \tau]$ 上的标量 $s^{(0)}$ 、向量 $s^{(1)}$ 和矩阵函数 $s^{(2)}$ 满足

$$\sup_{t \in [0, \tau], \beta \in \mathcal{B}} \|S^{(l)}(\beta, t) - s^{(l)}(\beta, t)\| \xrightarrow{P} 0 \quad (l = 0, 1, 2).$$

此外, 存在 $\delta > 0$ 使得 $n_2^{-1/2} \sup_{i,t} |\mathbf{Z}_{i,S}| Y_i(t) I\{\beta^T \mathbf{Z}_{i,S} > -\delta |\mathbf{Z}_{i,S}|\} \xrightarrow{P} 0$.

(C3) 令 $\mathcal{B}$, $s^{(0)}$, $s^{(1)}$ 和 $s^{(2)}$ 满足条件 (C2), 定义 $e(\beta, t) = s^{(1)}(\beta, t) / s^{(0)}(\beta, t)$, $v(\beta, t) = s^{(2)}(\beta, t) / s^{(0)}(\beta, t) - e(\beta, t)^{\otimes 2}$. 对于所有 $\beta \in \mathcal{B}$, $t \in [0, \tau]$ 满足 $s^{(1)}(\beta, t) = \partial s^{(0)}(\beta, t) / \partial \beta$, $s^{(2)}(\beta, t) = \partial^2 s^{(0)}(\beta, t) / \partial \beta^2$. 对于 $t \in [0, \tau]$, $s^{(0)}(\cdot, t)$, $s^{(1)}(\cdot, t)$ 和 $s^{(2)}(\cdot, t)$ 是 $\beta \in \mathcal{B}$ 上的连续函数, $s^{(0)}$, $s^{(1)}$ 和 $s^{(2)}$ 在 $\mathcal{B} \times [0, \tau]$ 上有界, $s^{(0)}$ 在 $\mathcal{B} \times [0, \tau]$ 上远离零, 并且矩阵 $\Sigma = \int_0^\tau v(\beta, t) s^{(0)}(\beta, t) \lambda_0(t) dt$ 是正定阵.

(C4) 当 $n \rightarrow \infty$ 时, $P(\mathcal{A} \subseteq S) \rightarrow 1$.

(C5) 估计量 $\hat{\mu}_1$ 满足 $\|\hat{\mu}_1 - \beta\|_\infty = O_p(c_{np})$, 其中 c_{np} 是一个序列, 满足 $c_{np} \rightarrow 0$ 和 $1/(\sqrt{n} \cdot c_{np}) = O(1)$.

(C6) $\min_{j \in \mathcal{A}} W_j \geq 2cn^{-\kappa}$, 其中 $c > 0$ 是常数, $\kappa \in [0, 1/2)$.

这些正则化条件比较温和, 在大多数情况下可以满足. 条件 (C1)–(C3) 主要用于推导 Cox 比例风险模型中极大似然估计的渐近性质^[1]. 条件 (C4) 确保在使用 LASSO 正则化方法降维时, 所有重要变量可以被保留. 通过选择合适的 λ 值可以实现这一点. 此条件已广泛用于很多文献中 (参见文献 [6]). 在实际应用中, 可以使用 Akaike 信息准则 (Akaike information criterion, AIC) 选择一个过拟合的模型来达到此条件. 条件 (C5) 假设 $\hat{\mu}_1$ 是 β 的一个合理估计量, Du 等^[6] 在证明理论性质时也假设了类似的条件. 条件 (C6) 要求重要变量对应的信号不能太弱, 这是一个很合理的假设, 已广泛用于变量筛选的相关文献中 (参见文献 [8, 18]).

引理 3.1 假设条件 (C1)–(C5) 成立, 无论使用何种数据分割方式, 对于 $j \in \mathcal{A}^c$, $\widehat{W}_j$ 的渐近分布关于零对称.

注 3.1 此引理证明了无论采用何种数据分割方式, $\widehat{W}_j$ ($j \in \mathcal{A}^c$) 的分布在渐近意义上关于零对称. 这为数据驱动阈值选取方法 (2.5) 的建立奠定了理论基础.

接下来证明估计值 $\widehat{W}_j$ 的相合性, 并进一步证明 FSDS 方法的变量筛选性质, 包括确定筛选性质和排序秩相合性.

定理 3.1 假设条件 (C1)–(C6) 成立, 可以得到下列结论.

(i) (估计的相合性) 对于任意 $c > 0$ 和 $\kappa \in [0, 1/2)$, 存在常数 $c_1 > 0$ 和 $c_2 > 0$ 满足

$$P\left(\max_{1 \leq j \leq p} |\widehat{W}_j - W_j| \geq cn^{-\kappa}\right) \leq O\left\{p \left[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})\right]\right\}.$$

(ii) (秩相合性) 存在常数 $c_1 > 0$ 和 $c_2 > 0$ 满足

$$P\left(\max_{j \notin \mathcal{A}} \widehat{W}_j < \min_{j \in \mathcal{A}} \widehat{W}_j\right) > 1 - O\left\{p \left[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})\right]\right\}.$$

(iii) (确定筛选性) 如果 $\alpha \geq 1/a$, 存在常数 $c_1 > 0$ 和 $c_2 > 0$ 使得

$$P(\mathcal{A} \subseteq \widehat{\mathcal{A}}(L)) > 1 - O\left\{p \left[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})\right]\right\},$$

其中 L 的值由 (2.5) 确定, $\widehat{\mathcal{A}}(L)$ 的定义见 (2.6).

注 3.2 定理 3.1 奠定了特征筛选方法 FSDS 的理论基础. 由定理 3.1 中的 (ii) 可以得到, 当 $n \rightarrow \infty$ 时, $P(\max_{j \notin \mathcal{A}} \widehat{W}_j < \min_{j \in \mathcal{A}} \widehat{W}_j) \rightarrow 1$. 因此, 根据 $\widehat{W}_j$ 的取值从大到小对协变量进行排序, 确实可以以趋近 1 的概率将所有重要变量排列在不重要变量之前. 由定理 3.1 中的 (iii) 可以得到, 当 $n \rightarrow \infty$ 时, $P(\mathcal{A} \subseteq \widehat{\mathcal{A}}(L)) \rightarrow 1$. 由此可见, 用 FSDS 方法对数据进行筛选时, 能以趋近 1 的概率保留所有重要变量.

接下来建立 FSDS 方法的 FDR 控制性质, 包括 FDR 界的有限样本理论以及 FDR 控制的渐近理论. 定义 $\widehat{\mathbf{W}}_S = (\widehat{W}_j : j \in S)$, $\widehat{\mathbf{W}}_{-j} = \widehat{\mathbf{W}}_S \setminus \widehat{W}_j$, $\theta_j = |P(\widehat{W}_j > 0 \mid |\widehat{W}_j|, \widehat{\mathbf{W}}_{-j}) - 1/2|$.

定理 3.2 假设条件 (C1)–(C5) 成立, 对于任意 $\alpha > 0$, 用 FSDS 方法进行变量筛选时, 其对应的 FDR 满足

$$\text{FDR} \leq \min_{\epsilon \geq 0} \left\{ \alpha \cdot \frac{1+2\epsilon}{1-2\epsilon} + P\left(\max_{j \in \mathcal{A}^c \cap S} \theta_j > \epsilon\right) \right\}. \quad (3.1)$$

定理 3.2 给出了 FSDS 方法的 FDR 上界. 它表明对 θ_j 的严格控制可以实现 FDR 的有效控制. 接下来, 建立该方法 FDR 控制的渐近理论. 对于任意 $j \in \mathcal{A}^c$, 定义 $\nu_j = I(\widehat{W}_j < 0)$, $S_j = \nu_1 + \nu_2 + \cdots + \nu_j$. 假设存在某个序列 $c_n \rightarrow \infty$, 当 $n \rightarrow \infty$ 时满足 $E(\nu_j) = 1/2 + o(c_n^{-1})$, 且当 $(n, p) \rightarrow \infty$ 时 $c_n/p \rightarrow 0$. 可以得到以下引理 3.2.

引理 3.2 假设条件 (C1)–(C5) 成立, 阈值 L 由 (2.5) 所确定, 可以得到

$$E\left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L)}{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -L)}\right] \leq 1 + o(1), \quad (3.2)$$

其中当 $(n, p) \rightarrow \infty$ 时, $o(1) \rightarrow 0$.

定理 3.3 假设条件 (C1)–(C5) 成立, 对于任意 $\alpha > 0$, 用 FSDS 方法进行变量筛选时, 其对应的 FDR 满足

$$\text{FDR} = E\left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L)}{1 \vee \sum_{j=1}^p I(\widehat{W}_j \geq L)}\right] \leq \alpha + o(1), \quad (3.3)$$

其中当 $(n, p) \rightarrow \infty$ 时, $o(1) \rightarrow 0$. 因此

$$\limsup_{(n, p) \rightarrow \infty} \text{FDR} \leq \alpha. \quad (3.4)$$

注 3.3 结合定理 3.1 和 3.3 的内容可知, 当 $\alpha \geq 1/a$ 时, 所建立的特征筛选方法 FSDS 既满足确定筛选性质, 又能将 FDR 有效地控制在给定水平 α 之下, 即本文所建立的特征筛选方法 FSDS 不仅能在快速有效降维的同时准确保留所有重要变量, 而且可以将 FDR 控制在一定范围内.

4 FSDS 方法的模拟计算

本节通过数值模拟试验来验证 FSDS 在有限样本下的有效性. 为了对比不同方法的效果, 这里分别考虑 Zhao 和 Li [23] 在 Cox 比例风险模型下所提出的独立筛选方法 PSIS, Gorst-Rasmussen 和 Scheike [9] 提出的可以处理单指标危险率模型的变量筛选法 FAST, Song 等 [18] 提出的无模型筛选方法 CRIS, Zhou 和 Zhu [24] 提出的无模型筛选方法 CSIRS, Zhang 等 [20] 基于距离相关系数提出的无模型筛选方法 DCSIS. 在这五种方法中, 考虑 Fan 和 Lv [8] 提出的硬阈值准则, 即保留前 $d = [n/\log(n)]$ 个变量. 进一步地, 将这五种方法和本文提出的硬阈值筛选方法 HT-FSDS 相结合, 分别简记为 HT-FSDS-PSIS, HT-FSDS-FAST, HT-FSDS-CRIS, HT-FSDS-CSIRS 和 HT-FSDS-DCSIS. 此外, 为了更全面地评估 FSDS 方法在 FDR 控制方面的性能, 本文进一步引入了 He 等 [11] 提出的 SDA-C 方法作为对比. 此方法基于 Cox 比例风险模型构建, 在变量筛选的过程中考虑了 FDR 控制的问题. 为了评估不同方法的效果, 分别考虑以下指标.

- (1) 错误发现率 (FDR): 被错误识别为重要变量的非重要变量所占的比例.
- (2) 真阳性率 (TPR): 被正确识别为重要变量的重要变量所占的比例.
- (3) 模型大小 (MS): 最后所保留的变量的数目.
- (4) F1- 得分 (F1): 精确率和召回率的调和平均数, 计算公式为

$$F1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = \frac{TP}{TP + 0.5 \cdot (FP + FN)},$$

其中 TP 表示被正确识别为重要变量的变量数目, FP 表示被错误识别为重要变量的变量数目, FN 表示被错误识别为不重要变量的变量数目. F1- 得分越高, 表现越好.

- (5) $\mathcal{P}_a$: 所有重要变量同时被选出来的概率.

我们给出 200 次模拟实验所得到的 FDR, TPR, MS, F1 和 $\mathcal{P}_a$ 的平均数.

例 4.1 本文考虑从如下 Cox 比例风险模型中生成生存时间 T_i , 其条件危险率函数为

$$\lambda(t | \mathbf{Z}_i) = \lambda_0(t) \exp(\mathbf{Z}_i^T \boldsymbol{\beta}),$$

其中基准危险率函数 $\lambda_0(t) = (t - 0.5)^2$, $\boldsymbol{\beta} = (1, \dots, 1, 0, \dots, 0)$, 即前 14 个分量值为 1, 余下的分量值为 0. 因此, 只有前 14 个协变量是重要变量, 重要变量指标集为 $\mathcal{A} = \{1, 2, \dots, 14\}$. 协变量 $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})$ 来自多元正态分布 $N(\mathbf{0}, \boldsymbol{\Sigma})$, 这里分别考虑两种不同的协方差矩阵.

情况 (a) $\boldsymbol{\Sigma} = (\sigma_{ij})$, 当 $i \in \mathcal{A}$ 且 $j \in \mathcal{A}$ 时, $\sigma_{ij} = 0.5$; 当 $i \in \mathcal{A}^c$ 且 $j \in \mathcal{A}^c$ 时, $\sigma_{ij} = 0.2$; 余下的 $\sigma_{ij} = 0$.

情况 (b) $\boldsymbol{\Sigma} = (\sigma_{ij})$, 对于 $i, j = 1, \dots, p$, $\sigma_{ij} = 0.8^{|i-j|}$.

删失时间 C 来自均匀分布 $\text{Unif}(0, \tau)$, τ 的取值大小取决于所设定的删失率. 考虑样本量 $n = 400$, 协变量的维数 $p = 2000, 5000$, 删失率 (cr) 为 20% 和 40%, $\alpha = 0.2$, 对每种设置都重复 200 次实验.

模拟计算的结果见表 1 和 2. 通过对比分析可以看出, 所提出的 FSDS 方法在各种情况下均能将 FDR 控制在预设水平 $\alpha = 0.2$ 之下, 同时 TPR 和 $\mathcal{P}_a$ 的值接近 1. 基于硬阈值准则的变量筛选方法 PSIS, FAST, CSIRS, DCSIS 和 CRIS 在数据稀疏情况下虽然能将重要变量准确选出, 但 FDR 太高, 会导致很多不重要变量被错误保留. 例如, 在情况 (a) 中, 当 $n = 400$, $p = 2000$, cr = 20% 时, 所提出的 FSDS 方法的经验 FDR 为 0.180, 而 PSIS, FAST, CSIRS, DCSIS 和 CRIS 方法的经验 FDR 为 0.788. 这表明 PSIS, FAST, CSIRS, DCSIS 和 CRIS 方法无法控制 FDR. SDA-C 方法虽然能较好地控

表 1 例 1 情况 (a) 中 FDR、TPR、F1- 得分、 $\mathcal{P}_a$ 和 MS 的模拟计算结果

p	方法	cr = 20%					cr = 40%				
		FDR	TPR	F1	$\mathcal{P}_a$	MS	FDR	TPR	F1	$\mathcal{P}_a$	MS
2000	FSDS	0.180	1.000	0.895	0.995	17.585	0.185	0.999	0.892	0.990	17.675
	SDA-C	0.183	0.777	0.770	0.225	14.010	0.173	0.658	0.706	0.045	11.665
	HT-FSDS-PSIS	0.180	1.000	0.895	1.000	17.585	0.185	1.000	0.892	1.000	17.675
	HT-FSDS-FAST	0.180	1.000	0.895	1.000	17.585	0.185	1.000	0.892	1.000	17.675
	HT-FSDS-CSIRS	0.180	1.000	0.895	1.000	17.585	0.185	1.000	0.892	1.000	17.675
	HT-FSDS-DCSIS	0.180	1.000	0.895	1.000	17.585	0.185	1.000	0.892	1.000	17.675
	HT-FSDS-CRIS	0.180	1.000	0.895	1.000	17.585	0.185	1.000	0.892	1.000	17.675
	PSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	FAST	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CSIRS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	DCSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CRIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	FSDS	0.176	1.000	0.898	1.000	17.450	0.184	0.999	0.892	0.985	17.705
	SDA-C	0.144	0.600	0.676	0.040	10.250	0.192	0.535	0.618	0.020	9.835
	HT-FSDS-PSIS	0.176	1.000	0.898	1.000	17.450	0.183	1.000	0.892	0.995	17.705
5000	HT-FSDS-FAST	0.176	1.000	0.898	1.000	17.450	0.183	1.000	0.892	0.995	17.705
	HT-FSDS-CSIRS	0.176	1.000	0.898	1.000	17.450	0.183	1.000	0.892	0.995	17.705
	HT-FSDS-DCSIS	0.176	1.000	0.898	1.000	17.450	0.183	1.000	0.892	0.995	17.705
	HT-FSDS-CRIS	0.176	1.000	0.898	1.000	17.450	0.183	1.000	0.892	0.995	17.705
	PSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	FAST	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CSIRS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	DCSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CRIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000

制 FDR, 但 TPR 取值较低, 说明它会漏选一些重要变量. 此外, 该方法的 $\mathcal{P}_a$ 取值也较小, 说明其很少能在一次筛选中同时选出所有的重要变量. 此外, 可以观察到 HT-FSDS-PSIS, HT-FSDS-FAST, HT-FSDS-CSIRS, HT-FSDS-DCSIS 和 HT-FSDS-CRIS 也表现良好, 在某些情况下 FDR 值略小于 FSDS 方法. 这并不令人意外, 因为这些方法利用了全样本信息来构建统计量进行变量筛选, 因此避免了信息损失. 此外, 为了评估 FSDS 方法在不同 α 取值下的表现, 本文在情况 (a) 和 (b) 中分别考虑了 $\alpha = 0.1, 0.15, 0.20, 0.25, 0.30$. 当 $n = 400, p = 2000$ 时的模拟结果见表 3, 通过观察可见, 本文所提出的 FSDS 方法在不同的 α 取值下均能有效控制 FDR, 并且变量筛选效果良好. 这表明 FSDS 方法具有较强的稳健性.

例 4.2 为了评估 FSDS 方法在不同模型设定下的稳健性, 本文考虑采用如下转换模型生成生存时间:

$$H(T) = -\beta^T \mathbf{Z} + \epsilon,$$

其中 $H(t) = \log\{0.5(e^{2t} - 1)\}$, $\epsilon \sim N(0, 1)$, 协变量 $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})$ 来自多元正态分布 $N(\mathbf{0}, \Sigma)$, 协

表 2 例 1 情况 (b) 中 FDR、TPR、F1- 得分、 $\mathcal{P}_a$ 和 MS 的模拟计算结果

p	方法	cr = 20%					cr = 40%				
		FDR	TPR	F1	$\mathcal{P}_a$	MS	FDR	TPR	F1	$\mathcal{P}_a$	MS
2000	FSDS	0.182	0.999	0.895	0.990	17.535	0.182	0.996	0.893	0.965	17.490
	SDA-C	0.157	0.738	0.761	0.100	12.920	0.167	0.612	0.676	0.010	10.960
	HT-FSDS-PSIS	0.182	0.999	0.895	0.990	17.535	0.181	0.996	0.893	0.970	17.490
	HT-FSDS-FAST	0.182	0.999	0.894	0.985	17.535	0.181	0.996	0.893	0.970	17.490
	HT-FSDS-CSIRS	0.182	0.999	0.895	0.990	17.535	0.181	0.996	0.893	0.970	17.490
	HT-FSDS-DCSIS	0.183	0.999	0.894	0.980	17.535	0.181	0.996	0.893	0.970	17.490
	HT-FSDS-CRIS	0.183	0.998	0.893	0.970	17.535	0.183	0.995	0.892	0.950	17.490
	PSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	FAST	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CSIRS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	DCSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CRIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
5000	FSDS	0.187	0.999	0.891	0.985	17.670	0.177	0.996	0.895	0.955	17.510
	SDA-C	0.127	0.570	0.663	0.005	9.670	0.143	0.490	0.601	0.005	8.480
	HT-FSDS-PSIS	0.187	1.000	0.891	0.995	17.670	0.176	0.998	0.896	0.980	17.510
	HT-FSDS-FAST	0.187	1.000	0.891	0.995	17.670	0.175	0.998	0.897	0.985	17.510
	HT-FSDS-CSIRS	0.187	1.000	0.891	0.995	17.670	0.176	0.998	0.896	0.980	17.510
	HT-FSDS-DCSIS	0.187	0.999	0.891	0.990	17.670	0.176	0.998	0.896	0.975	17.510
	HT-FSDS-CRIS	0.189	0.998	0.890	0.970	17.670	0.178	0.995	0.893	0.940	17.510
	PSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	FAST	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CSIRS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	DCSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CRIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	0.995	66.000

方差矩阵按照情况 (b) 的方式生成. 其他设置和例 4.1 保持一致. 模拟计算的结果见表 4. 通过观察对比, 可以得到与例 4.1 类似的结论. 本文所提出的 FSDS 方法能够将 FDR 控制在预设水平 $\alpha = 0.2$ 之下, 且 TPR 和 $\mathcal{P}_a$ 的取值均接近 1. 基于硬阈值的筛选方法 PSIS, FAST, CSIRS, DCSIS 和 CRIS 无法控制 FDR. SDA-C 方法虽然能较好地控制 FDR, 但 TPR 取值较低, 会漏选一些重要变量. 综上所述, 即便数据生成过程偏离 Cox 模型假设, FSDS 方法依然在 FDR 控制和变量筛选精度方面展现出良好且稳健的性能.

5 DLBCL 数据的分析

本节将所提出的 FSDS 方法应用于 DLBCL 患者数据集的分析. DLBCL 是美国成人中最常见的淋巴瘤类型之一, 约 60% 的 DLBCL 患者在接受化疗后未能存活. 因此, 深入了解哪些遗传因子对患者的生存时间具有显著影响, 对于制定有效的治疗策略至关重要. Rosenwald 等^[17] 收集了 240 名患者的生存时间 (T) 和 7399 个基因表达水平的数据, 并对这些患者的特征进行了总结. 在该数据的探索

表 3 例 1 不同 α 取值下 FDR、TPR、F1- 得分、 $\mathcal{P}_a$ 和 MS 的模拟计算结果

情况	α	cr = 20%					cr = 40%				
		FDR	TPR	F1	$\mathcal{P}_a$	MS	FDR	TPR	F1	$\mathcal{P}_a$	MS
(a)	0.10	0.068	1.000	0.962	0.995	15.185	0.070	0.999	0.961	0.990	15.200
	0.15	0.126	1.000	0.929	0.995	16.305	0.129	0.999	0.927	0.990	16.375
	0.20	0.180	1.000	0.895	0.995	17.585	0.185	0.999	0.892	0.990	17.675
	0.25	0.242	1.000	0.854	0.995	19.275	0.235	0.999	0.859	0.990	18.995
	0.30	0.279	1.000	0.829	0.995	20.350	0.278	0.999	0.831	0.990	20.185
(b)	0.10	0.066	0.999	0.963	0.990	15.170	0.065	0.998	0.963	0.965	15.080
	0.15	0.127	0.999	0.928	0.990	16.315	0.119	0.998	0.932	0.970	16.110
	0.20	0.182	0.999	0.895	0.990	17.535	0.183	0.998	0.893	0.970	17.545
	0.25	0.230	0.999	0.864	0.990	18.760	0.240	0.998	0.857	0.975	19.030
	0.30	0.270	0.999	0.836	0.990	19.880	0.283	0.999	0.827	0.980	20.310

性分析中, 所有基因的表达水平均进行了标准化处理, 即均值为零, 标准差为 1. 在 240 名患者中, 102 名患者在最后一次随访时仍然存活, 因此该数据集的删失率为 42.5%. 该数据集可表示为 $\{X_i, \Delta_i, \mathbf{Z}_i\}$ ($i = 1, \dots, 240$), 其中 $\mathbf{Z} = (Z_1, \dots, Z_{7399})^T$ 表示 $p = 7399$ 个基因的表达水平, X_i 为第 i 名患者的生存时间, Δ_i 为删失指示变量. 研究的主要目标是从 $p = 7399$ 个基因表达水平中识别出与患者生存时间密切相关的基因.

此处, 采用本文所提出的 FSDS 方法对数据进行分析. 首先, 将 240 名患者随机均匀地分为两个数据集. 然后, 构建具有对称性质的统计量, 并根据 (2.5) 确定合适的阈值. 基于该阈值, 从 7399 个基因中筛选出重要基因. 在设定 FDR 水平 $\alpha = 0.2$ 的条件下, FSDS 方法成功识别出五个重要基因, 其对应的唯一标识符 (UNIQID) 分别为 31981, 17154, 28178, 33358 和 32679. 接着, 基于这些选定的基因变量拟合了 Cox 模型, 得到的系数估计值分别为 0.543, -0.186 , -0.321 , 0.416 和 -0.503 . 分析结果表明, 这五个基因可能对患者的生存风险具有重要影响. 这一发现与 Zhang 等 [20] 的分析结果一致, 此文献中同样识别出基因 31981 和 32679 对患者的生存风险有显著影响. 此外, 通过观察估计系数, 可以发现基因 31981 对患者生存风险的影响最大, 这一结论与 Li 和 Luan [12] 的分析结果一致, 他们的研究也表明该基因对患者生存风险的影响最为显著.

进一步地, 为了评估所提出方法的预测性能, 类似于 Rosenwald 等 [17] 的方法, 将 240 名患者随机分为训练集 (160 名患者) 和测试集 (80 名患者). 训练集用于建立预测模型, 测试集用于评估模型的性能. 首先, 将所提出的 FSDS 方法应用于训练集, 以筛选重要基因. 基于这些选定的基因, 拟合 Cox 比例风险模型, 并获得未知参数的估计值. 接下来, 计算测试集中患者的风险得分. 将训练集中患者风险得分的均值设为阈值, 并根据该阈值将测试集患者划分为低风险组和高风险组, 随后绘制 Kaplan-Meier 生存曲线. 为了进行比较, 这里对比了另外五种方法, 分别是 Zhao 和 Li [23] 提出的 PSIS, Gorst-Rasmussen 和 Scheike [9] 提出的 FAST, Song 等 [18] 提出的 CRIS, Zhou 和 Zhu [24] 提出的 CSIRS, 以及 Zhang 等 [20] 提出的 DCSIS. 具体而言, 首先将这五种变量筛选方法应用于训练集, 选取前 $\lceil 160/\log(160) \rceil = 31$ 个基因. 然后, 基于所选基因拟合 Cox 比例风险模型, 并使用正则化方法 LASSO 从这 31 个基因中进一步筛选出重要基因, 借助五折交叉验证方法选择合适的惩罚参数. 同样地, 基于得到的未知参数估计值计算测试集中患者的风险得分, 并将其划分为低风险组和高风险组, 接着绘制 Kaplan-Meier 生存曲线. 图 1 展示了不同方法对应的 Kaplan-Meier 生存曲线. 通过观察图 1,

表 4 例 2 情况 (b) 中 FDR、TPR、F1- 得分、 $\mathcal{P}_a$ 和 MS 的模拟计算结果

p	方法	cr = 20%					= 40%				
		FDR	TPR	F1	$\mathcal{P}_a$	MS	FDR	TPR	F1	$\mathcal{P}_a$	MS
2000	FSDS	0.178	0.999	0.896	0.990	17.580	0.175	0.999	0.898	0.985	17.405
	SDA-C	0.144	0.737	0.767	0.080	12.745	0.128	0.611	0.692	0.010	10.325
	HT-FSDS-PSIS	0.178	1.000	0.896	0.995	17.580	0.174	1.000	0.899	1.000	17.405
	HT-FSDS-FAST	0.178	1.000	0.896	0.995	17.580	0.174	1.000	0.899	1.000	17.405
	HT-FSDS-CSIRS	0.178	1.000	0.896	0.995	17.580	0.174	1.000	0.899	1.000	17.405
	HT-FSDS-DCSIS	0.178	0.999	0.896	0.990	17.580	0.174	1.000	0.899	1.000	17.405
	HT-FSDS-CRIS	0.179	0.999	0.895	0.980	17.580	0.178	0.996	0.895	0.950	17.405
	PSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	FAST	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CSIRS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	DCSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CRIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
5000	FSDS	0.167	1.000	0.904	1.000	17.220	0.177	0.998	0.897	0.975	17.430
	SDA-C	0.135	0.596	0.674	0.030	10.325	0.160	0.485	0.590	0.000	8.655
	HT-FSDS-PSIS	0.168	0.998	0.902	0.975	17.220	0.176	1.000	0.898	0.995	17.430
	HT-FSDS-FAST	0.169	0.998	0.901	0.965	17.220	0.176	1.000	0.898	0.995	17.430
	HT-FSDS-CSIRS	0.168	0.999	0.903	0.980	17.220	0.176	0.999	0.898	0.990	17.430
	HT-FSDS-DCSIS	0.168	0.998	0.902	0.975	17.220	0.175	1.000	0.898	1.000	17.430
	HT-FSDS-CRIS	0.168	0.998	0.902	0.975	17.220	0.179	0.996	0.895	0.945	17.430
	PSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	FAST	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CSIRS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	DCSIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	1.000	66.000
	CRIS	0.788	1.000	0.350	1.000	66.000	0.788	1.000	0.350	0.995	66.000

可以看出, 基于 FSDS 方法所绘制的 Kaplan-Meier 生存曲线明显分开, 而 FAST, PSIS, CRIS, CSIRS 和 DCSIS 方法所对应的生存曲线区分较不明显. 为了客观评估这些生存曲线之间的差异, 我们对各方法进行了 Log-rank 检验. 结果显示, FSDS 方法的 P 值为 0.005, PSIS 方法的 P 值为 0.135, FAST 方法的 P 值为 0.422, CSIRS 方法的 P 值为 0.478, DCSIS 方法的 P 值为 0.234, CRIS 方法的 P 值为 0.598. 这一结果进一步验证了 FSDS 方法在预测性能上的有效性, 且其表现明显优于其他方法. 综上所述, 本文提出的 FSDS 方法能够有效处理超高维右删失生存数据的降维问题, 并且基于该方法所建立的预测模型能有效预测患者未来的生存风险.

6 结论

本文聚焦于 Cox 比例风险模型下超高维右删失生存数据的降维问题, 首先提出了一种数据自适应阈值选择方法, 该方法能够在不同情境下选择合适的阈值, 准确地区分重要变量与不重要变量. 基于该阈值, 进一步提出了适用于超高维右删失生存数据的变量筛选方法 FSDS. 在构建 FSDS 方法时,

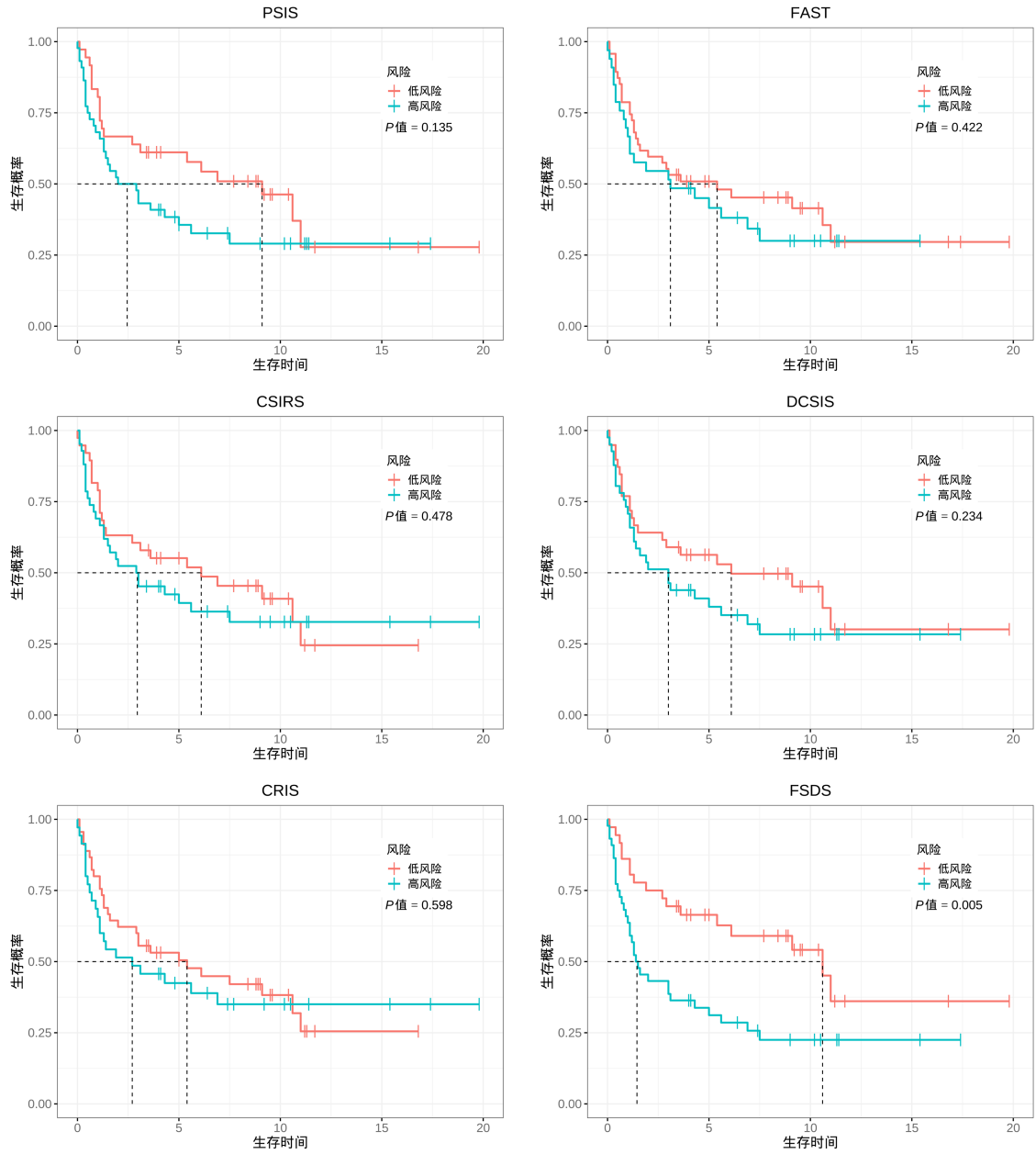


图 1 (网络版彩图) 测试集中高风险组和低风险组对应的 Kaplan-Meier 生存曲线

我们充分考虑了 FDR 控制问题. 因此, 此方法不仅能在快速有效降维的同时保留所有重要变量, 还可以将 FDR 控制在一定的范围之内. 在一定的正则化条件下, 本文证明了 FSDS 方法具有良好的变量筛选性质, 包括确定筛选性质和排序秩相合性. 此外, 本文还证明了该方法拥有 FDR 控制的相关性质, 包括有限样本下的 FDR 界限理论和渐近 FDR 控制理论. 结合定理 3.1 和 3.3 的内容可知, 当 $\alpha \geq 1/a$ 时, 所建立的特征筛选方法 FSDS 既满足确定筛选性质, 又能将 FDR 有效地控制在给定水平 α 之下. 数值模拟结果显示, FSDS 方法在变量选择和 FDR 控制方面均有优良表现, 且效果明显优于现有的其他方法. 实际数据分析进一步证明了 FSDS 方法能够有效处理超高维右删失生存数据的降维问题, 并

且基于该方法所选变量构建的模型具有较好的预测性能. FSDS 方法计算简单且快速, 具有广泛的应用前景. 例如, 在生物医学和临床试验中, FSDS 方法能够帮助研究人员揭示影响患者生存时间的关键因素, 辅助疾病预测及个性化治疗方案的制定. 此外, FSDS 方法还可应用于医疗健康领域的早期诊断、风险评估等实际问题中, 在公共卫生监测、金融风险识别、精准医疗和个性化治疗等领域也具有重要的应用价值.

尽管 FSDS 方法是在 Cox 比例风险模型的框架下提出的, 其核心思想可平行推广至其他类型的生存分析模型, 如加速失效时间模型或加性风险模型等. 具体而言, 在推广过程中, 只需要根据目标模型的具体结构对统计量 $\hat{\mu}_1$ 和 $\hat{\mu}_2$ 进行相应调整. 对称统计量 $\widehat{W}_j$ 的构建和阈值的选取均可采取与本文类似的方法. 为增强方法的通用性和模型稳健性, 可以进一步探索不依赖于特定模型的阈值选择方法. 例如, 可考虑利用协变量与生存时间之间的相关性指标 (如相关系数、非参数相关测度等) 来构造具有对称结构的统计量 $\widehat{W}_j$. 我们将在后续研究中进一步系统地开展这些方法和理论的探索.

致谢 谨此感谢主编、副主编、编委及各位审稿人对本文提出的宝贵意见和悉心指导, 这些建议对提升本文质量具有重要作用.

参考文献

- 1 Andersen P K, Gill R D. Cox's regression model for counting processes: A large sample study. *Ann Statist*, 1982, 10: 1100–1120
- 2 Barber R F, Candès E J, Samworth R J. Robust inference with knockoffs. *Ann Statist*, 2020, 48: 1409–1431
- 3 Barut E, Fan J, Verhasselt A. Conditional sure independence screening. *J Amer Statist Assoc*, 2016, 111: 1266–1277
- 4 Cox D R. Regression models and life-tables. *J R Stat Soc Ser B Stat Methodol*, 1972, 34: 187–202
- 5 Dai C, Lin B, Xing X, et al. False discovery rate control via data splitting. *J Amer Statist Assoc*, 2023, 118: 2503–2520
- 6 Du L, Guo X, Sun W, et al. False discovery rate control under general dependence by symmetrized data aggregation. *J Amer Statist Assoc*, 2023, 118: 607–621
- 7 Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J Amer Statist Assoc*, 2011, 106: 544–557
- 8 Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B Stat Methodol*, 2008, 70: 849–911
- 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 Guo X, Ren H, Zou C, et al. Threshold selection in feature screening for error rate control. *J Amer Statist Assoc*, 2023, 118: 1773–1785
- 11 He B, Shi H, Guo X, et al. Variable selection for high-dimensional Cox model with error rate control. *J Syst Sci Complex*, 2025, 38: 1162–1185
- 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 Liu W, Ke Y, Liu J, et al. Model-free feature screening and FDR control with knockoff features. *J Amer Statist Assoc*, 2022, 117: 428–443
- 14 Liu Y, Zhang J, Zhao X. A new nonparametric screening method for ultrahigh-dimensional survival data. *Comput Statist Data Anal*, 2018, 119: 74–85
- 15 Massart P. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann Probab*, 2000, 28: 863–884
- 16 Pan W, Wang X, Xiao W, et al. A generic sure independence screening procedure. *J Amer Statist Assoc*, 2019, 114: 928–937
- 17 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. *N Engl J Med*, 2002, 346: 1937–1947
- 18 Song R, Lu W, Ma S, et al. Censored rank independence screening for high-dimensional survival data. *Biometrika*, 2014, 101: 799–814
- 19 Tong Z, Cai Z, Yang S, et al. Model-free conditional feature screening with FDR control. *J Amer Statist Assoc*, 2023, 118: 2575–2587

- 20 Zhang J, Liu Y, Cui H. Model-free feature screening via distance correlation for ultrahigh dimensional survival data. *Statist Papers*, 2021, 62: 2711–2738
- 21 Zhang J, Liu Y, Wu Y. Correlation rank screening for ultrahigh-dimensional survival data. *Comput Statist Data Anal*, 2017, 108: 121–132
- 22 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
- 23 Zhao S D, Li Y. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J Multivariate Anal*, 2012, 105: 397–411
- 24 Zhou T, Zhu L. Model-free feature screening for ultrahigh dimensional censored regression. *Stat Comput*, 2017, 27: 947–961
- 25 Zhu L P, Li L, Li R, et al. Model-free feature screening for ultrahigh-dimensional data. *J Amer Statist Assoc*, 2011, 106: 1464–1475

附录 A 附录

附录 A.1 引理 3.1 的证明

根据文献 [1, 定理 3.2], 当条件 (C1)–(C3) 成立时, 极大似然估计值 $\widehat{\beta}_{\text{MLE}} = (\widehat{\beta}_1, \widehat{\beta}_2, \dots, \widehat{\beta}_s)^T$ 渐近服从正态分布. 具体地, 可得到 $\sqrt{n_2}(\widehat{\beta}_{\text{MLE}} - \beta) \xrightarrow{d} N(\mathbf{0}, \Sigma^{-1})$. 基于这个结论, 结合 $\widehat{\mu}_{2j}$ 的定义, 可知对于任意 $j \in \mathcal{A}^c \cap S$, $\widehat{\mu}_{2j}$ 渐近服从均值为零的正态分布. 因此, $\widehat{\mu}_{2j}$ 的分布在渐近意义上关于零对称. Dai 等在文献 [5, 引理 2.1] 中指出, 对于任意 $j \in \mathcal{A}^c$, 若 $\widehat{\mu}_{1j}$ 和 $\widehat{\mu}_{2j}$ 中至少有一个的分布是关于零渐近对称的, 则无论采取何种数据分割形式, $\widehat{W}_j$ 的分布在渐近意义下也关于零对称. 基于此结论和 $\widehat{\mu}_{2j}$ 的对称性, 可以直接推断出, 对于任意 $j \in \mathcal{A}^c$, $\widehat{W}_j$ 的渐近分布关于零对称.

附录 A.2 定理 3.1 的证明

(i) 由 $\widehat{W}_j$ 和 W_j 的定义可知, 存在常数 $C_1 > 0$ 使得

$$\begin{aligned} |\widehat{W}_j - W_j| &= |\widehat{\mu}_{1j} \cdot \widehat{\mu}_{2j} - \beta_j \cdot \beta_j| = |\widehat{\mu}_{1j} \cdot \widehat{\mu}_{2j} - \widehat{\mu}_{1j} \cdot \beta_j + \widehat{\mu}_{1j} \cdot \beta_j - \beta_j \cdot \beta_j| \\ &\leq |\widehat{\mu}_{1j}| \cdot |\widehat{\mu}_{2j} - \beta_j| + |\beta_j| \cdot |\widehat{\mu}_{1j} - \beta_j| \leq C_1 \cdot (|\widehat{\mu}_{2j} - \beta_j| + |\widehat{\mu}_{1j} - \beta_j|). \end{aligned}$$

因此, 对于任意 $\epsilon > 0$, 可以得到

$$P(|\widehat{W}_j - W_j| > \epsilon) \leq P(|\widehat{\mu}_{1j} - \beta_j| > \epsilon/(2C_1)) + P(|\widehat{\mu}_{2j} - \beta_j| > \epsilon/(2C_1)). \quad (\text{A.1})$$

首先, 计算 $P(|\widehat{\mu}_{1j} - \beta_j| > \epsilon/(2C_1))$. 根据 Markov 不等式可得

$$P(|\widehat{\mu}_{1j} - \beta_j| > \epsilon/(2C_1)) = P(|\widehat{\mu}_{1j} - \beta_j|^2 > \epsilon^2/(4C_1^2)) \leq \frac{E(|\widehat{\mu}_{1j} - \beta_j|^2)}{\epsilon^2/(4C_1^2)}.$$

由条件 (C5) 可知, 统计量 $\widehat{\mu}_1$ 满足 $\|\widehat{\mu}_1 - \beta\|_\infty = O_p(c_{np})$, 其中 c_{np} 满足 $c_{np} \rightarrow 0$ 和 $1/(\sqrt{n} \cdot c_{np}) = O(1)$. 因此, 对于任意 j , 有 $|\widehat{\mu}_{1j} - \beta_j| = O_p(c_{np})$. 由 $1/(\sqrt{n} \cdot c_{np}) = O(1)$, 可得 $E(|\widehat{\mu}_{1j} - \beta_j|^2) = O(1/n)$. 因此,

$$P(|\widehat{\mu}_{1j} - \beta_j| > \epsilon/(2C_1)) \leq O(1/(n\epsilon^2)). \quad (\text{A.2})$$

接下来计算 $P(|\widehat{\mu}_{2j} - \beta_j| > \epsilon/(2C_1))$. 根据 Andersen 和 Gill [1] 的讨论和 $\widehat{\beta}_{\text{MLE}}$ 的定义, 可得到

$$n_2^{-1/2}U(\beta, 1) = n_2^{-1}I(\beta^*, 1) \cdot \sqrt{n_2}(\widehat{\beta}_{\text{MLE}} - \beta),$$

其中

$$U(\beta, 1) = \sum_{i \in D_2} \int_0^1 \left[Z_{i,S} - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right] dM_i(t) = \sum_{i \in D_2} \int_0^1 [Z_{i,S} - E(\beta, t)] dM_i(t)$$

和

$$I(\beta^*, 1) = \sum_{i \in D_2} \int_0^1 V(\beta^*, t) dN_i(t),$$

$M_i(t) = N_i(t) - \int_0^t \lambda_i(u) du$ ($i \in D_2, t \in [0, 1]$) 是定义在 $[0, 1]$ 上的局部鞅, β^* 取值落于 β 与 $\hat{\beta}_{1,MLE}$ 之间. 令 $U_j(\beta, 1)$ 表示 $U(\beta, 1)$ 的第 j 个分量, $I_{jj}(\beta^*, 1)$ 表示矩阵 $I(\beta^*, 1)$ 主对角线上的第 j 个元素. 根据 $\hat{\mu}_{2j}$ ($j \in S$) 的定义可知

$$n_2^{-1/2} U_j(\beta, 1) = n_2^{-1} I_{jj}(\beta^*, 1) \cdot \sqrt{n_2} (\hat{\mu}_{2j} - \beta_j).$$

令 $E_j(\beta, t)$ 表示 $E(\beta, t)$ 的第 j 个分量, $e_j(\beta, t)$ 表示 $e(\beta, t)$ 的第 j 个分量. 根据鞅的性质和条件 (C2), 可得

$$\begin{aligned} n_2^{-1/2} U_j(\beta, 1) &= n_2^{-1/2} \sum_{i \in D_2} \int_0^1 [Z_{ij} - E_j(\beta, t)] dM_i(t) \\ &= n_2^{-1/2} \sum_{i \in D_2} \left\{ \int_0^1 [Z_{ij} - e_j(\beta, t)] dM_i(t) + \int_0^1 [e_j(\beta, t) - E_j(\beta, t)] dM_i(t) \right\} \\ &= n_2^{-1/2} \sum_{i \in D_2} \int_0^1 [Z_{ij} - e_j(\beta, t)] dM_i(t) + O_p(1/\sqrt{n_2}). \end{aligned}$$

令 $f_{ij}(\beta) = \int_0^1 [Z_{ij} - e_j(\beta, t)] dM_i(t)$, 则 $n_2^{-1/2} U_j(\beta, 1) = n_2^{-1/2} \sum_{i \in D_2} f_{ij}(\beta) + O_p(1/\sqrt{n_2})$. 显然, $f_{ij}(\beta)$ ($i \in D_2$) 相互独立且满足 $E\{f_{ij}(\beta)\} = 0$. 令 E_n 表示经验测度, 则

$$n_2^{-1/2} |U_j(\beta, 1)| \leq n_2^{1/2} \sup_{\beta} |(E_n - E)f_{ij}(\beta)| + O_p(1/\sqrt{n_2}).$$

由条件 (C1)–(C3) 可知, 存在常数 $C_2 > 0$ 使得对于任意 $j \in S$ 和 $i \in D_2$, 有 $|f_{ij}(\beta)| \leq C_2$. 根据文献 [15, 定理 9] 可得, 对于任意 $\epsilon > 0$, 存在常数 $C_3 > 0$ 满足

$$\begin{aligned} P\left(n_2^{1/2} \sup_{\beta} |(E_n - E)f_{ij}(\beta)| \geq \epsilon\right) &= P\left(n_2^{-1/2} \sup_{\beta} |n_2 \cdot (E_n - E)f_{ij}(\beta)| \geq \epsilon\right) \\ &= P\left(\sup_{\beta} |n_2 \cdot (E_n - E)f_{ij}(\beta)| \geq n_2^{1/2} \epsilon\right) \\ &\leq 2 \exp(-C_3 n_2 \epsilon^2). \end{aligned}$$

接下来, 将上式和 $|\hat{\mu}_{2j} - \beta_j|$ 关联. 根据前面的讨论可知 $n_2^{-1/2} U_j(\beta, 1) = n_2^{-1} I_{jj}(\beta^*, 1) \cdot \sqrt{n_2} (\hat{\mu}_{2j} - \beta_j)$. 由条件 (C1)–(C3) 可知, 存在常数 $C_4 > 0$ 使得 $\sqrt{n_2} |\hat{\mu}_{2j} - \beta_j| = |n_2 I_{jj}(\beta^*, 1)^{-1}| \cdot |n_2^{-1/2} U_j(\beta, 1)| \leq C_4 \cdot |n_2^{-1/2} U_j(\beta, 1)| \leq C_4 \cdot n_2^{1/2} \sup_{\beta} |(E_n - E)f_{ij}(\beta)| + O_p(1/\sqrt{n_2})$. 因此

$$P(\sqrt{n_2} |\hat{\mu}_{2j} - \beta_j| \geq \epsilon) \leq P\left(C_4 \cdot n_2^{1/2} \sup_{\beta} |(E_n - E)f_{ij}(\beta)| + O_p(1/\sqrt{n_2}) \geq \epsilon\right).$$

对于任意 $\zeta > 0$, 有 $P(C_4 \cdot n_2^{1/2} \sup_{\beta} |(E_n - E)f_{ij}(\beta)| + O_p(1/\sqrt{n_2}) \geq \epsilon) \leq P(n_2^{1/2} \sup_{\beta} |(E_n - E)f_{ij}(\beta)| \geq C_4^{-1}(\epsilon - \zeta)) + P(O_p(1/\sqrt{n_2}) \geq \zeta)$. 因此, 存在常数 $C_5 > 0$ 使得 $P(\sqrt{n_2} |\hat{\mu}_{2j} - \beta_j| \geq \epsilon) \leq 2 \exp(-C_5 n_2 \epsilon^2)$.

进一步地, 存在常数 $C_6 > 0$ 使得

$$P(|\hat{\mu}_{2j} - \beta_j| > \epsilon/2C_1) \leq O(\exp(-C_6 n^2 \epsilon^2)). \quad (\text{A.3})$$

由 (A.1)–(A.3) 可得, 对于任意 $\epsilon > 0$, 有

$$P(|\widehat{W}_j - W_j| > \epsilon) \leq O(1/n\epsilon^2) + O(\exp(-C_6 n^2 \epsilon^2)). \quad (\text{A.4})$$

令 $\epsilon = cn^{-\kappa}$, 其中 $c > 0$, $\kappa \in [0, 1/2)$, 则存在常数 $c_1 > 0$ 和 $c_2 > 0$ 使得

$$P(|\widehat{W}_j - W_j| > cn^{-\kappa}) \leq O\left[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})\right]. \quad (\text{A.5})$$

从而可得

$$\begin{aligned} P\left(\max_{1 \leq j \leq p} |\widehat{W}_j - W_j| > cn^{-\kappa}\right) &\leq p \max_{1 \leq j \leq p} P(|\widehat{W}_j - W_j| > cn^{-\kappa}) \\ &\leq O\left\{p \left[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})\right]\right\}. \end{aligned} \quad (\text{A.6})$$

(ii) 根据条件 (C6) 和 $\min_{j \in \mathcal{A}} W_j \geq 2cn^{-\kappa}$, 结合 (A.6) 可得

$$\begin{aligned} P\left(\min_{j \in \mathcal{A}} \widehat{W}_j \leq \max_{j \notin \mathcal{A}} \widehat{W}_j\right) &= P\left(\max_{j \notin \mathcal{A}} \widehat{W}_j - \max_{j \notin \mathcal{A}} W_j - \min_{j \in \mathcal{A}} \widehat{W}_j + \min_{j \in \mathcal{A}} W_j \geq \min_{j \in \mathcal{A}} W_j\right) \\ &\leq P\left(\max_{j \notin \mathcal{A}} \widehat{W}_j - \max_{j \notin \mathcal{A}} W_j - \min_{j \in \mathcal{A}} \widehat{W}_j + \min_{j \in \mathcal{A}} W_j \geq 2cn^{-\kappa}\right) \\ &\leq P\left(\max_{j \notin \mathcal{A}} |\widehat{W}_j - W_j| \geq cn^{-\kappa}\right) + P\left(\max_{j \in \mathcal{A}} |\widehat{W}_j - W_j| \geq cn^{-\kappa}\right) \\ &\leq 2P\left(\max_{1 \leq j \leq p} |\widehat{W}_j - W_j| \geq cn^{-\kappa}\right) \\ &\leq O\left\{p \left[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})\right]\right\}. \end{aligned}$$

因此,

$$P\left(\max_{j \notin \mathcal{A}} \widehat{W}_j < \min_{j \in \mathcal{A}} \widehat{W}_j\right) > 1 - O\left\{p \left[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})\right]\right\}.$$

(iii) 由 (ii) 可知 $\min_{j \in \mathcal{A}} \widehat{W}_j > \max_{j \notin \mathcal{A}} \widehat{W}_j$ 成立的概率大于 $1 - O\{p[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})]\}$, 即重要变量排在不重要变量之前的概率大于 $1 - O\{p[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})]\}$. 考虑寻找 (2.5) 中所定义的阈值 L 的过程, 从最小的值 $|\widehat{W}_{d+1}| = 0$ 开始, 然后移动到第二小的值 $|\widehat{W}_d|$, 接着移动到 $|\widehat{W}_{d-1}|$, 依此类推. 当找到一个值满足 (2.5) 时, 过程停止. 考虑事件 $B = \{\min_{j \in \mathcal{A}} \widehat{W}_j > \max_{j \notin \mathcal{A}} |\widehat{W}_j|\}$, 可知事件 B 发生的概率大于等于 $1 - O\{p[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})]\}$. 令 $L_{\min} = \min_{j \in \mathcal{A}} |\widehat{W}_j| = \min_{j \in \mathcal{A}} \widehat{W}_j$. 当 $\alpha \geq 1/a$ 时,

$$\frac{1 + \#\{j : \widehat{W}_j \leq -L_{\min}\}}{\#\{j : \widehat{W}_j \geq L_{\min}\} \vee 1} = \frac{1 + 0}{a} \leq \alpha.$$

根据 (2.5) 中 L 的定义, 可知寻找 L 的过程最迟在 t 达到 $L_{\min}$ 时停止, 因此 $L \leq L_{\min}$. 可以得到

$$\widehat{\mathcal{A}}(L) = \{j : \widehat{W}_j \geq L, j = 1, \dots, p\} \supseteq \{j : \widehat{W}_j \geq L_{\min}, j = 1, \dots, p\} = \mathcal{A}.$$

因此,

$$P(\mathcal{A} \subseteq \widehat{\mathcal{A}}(L)) > 1 - O\left\{p \left[\frac{c_1}{n^{1-2\kappa}} + \exp(-c_2 n^{2(1-\kappa)})\right]\right\}.$$

附录 A.3 定理 3.2 的证明

对于任意 $t > 0$ 和 $\epsilon > 0$, 定义

$$R_\epsilon(t) = \frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq t, \theta_j \leq \epsilon)}{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -t)}.$$

考虑事件 $\zeta = \{\theta = \max_{j \in \mathcal{A}^c} \theta_j \leq \epsilon\}$ 和从 $\widehat{\mathbf{W}}$ 映射到 $L \geq 0$ 的阈值准则 $L = T(\widehat{\mathbf{W}})$. 对 $j = 1, \dots, p$, 定义

$$L_j = T(\widehat{W}_1, \dots, \widehat{W}_{j-1}, |\widehat{W}_j|, \widehat{W}_{j+1}, \dots, \widehat{W}_p) \geq 0.$$

由阈值 L 的定义可得

$$\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L, \theta_j \leq \epsilon)}{1 \vee \sum_{j=1}^p I(\widehat{W}_j \geq L)} = \frac{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -L)}{1 \vee \sum_{j=1}^p I(\widehat{W}_j \geq L)} \cdot \frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L, \theta_j \leq \epsilon)}{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -L)} \leq \alpha \cdot R_\epsilon(L).$$

接下来, 推导 $E[R_\epsilon(L)]$ 的上界. 根据 $R_\epsilon(L)$ 的定义可得

$$\begin{aligned} E[R_\epsilon(L)] &= E \left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L, \theta_j \leq \epsilon)}{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -L)} \right] \\ &= \sum_{j \in \mathcal{A}^c} E \left[\frac{I(\widehat{W}_j \geq L_j, \theta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_j)} \right] \\ &= \sum_{j \in \mathcal{A}^c} E \left\{ E \left[\frac{I(\widehat{W}_j \geq L_j, \theta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_j)} \mid |\widehat{W}_j|, \widehat{\mathbf{W}}_{-j} \right] \right\} \\ &= \sum_{j \in \mathcal{A}^c} E \left\{ P(\widehat{W}_j > 0 \mid |\widehat{W}_j|, \widehat{\mathbf{W}}_{-j}) \cdot \frac{I(|\widehat{W}_j| \geq L_j, \theta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_j)} \right\}. \end{aligned}$$

上述第二步是基于 L 和 L_j 的定义; 第三步是基于条件期望的性质; 最后一步成立是因为给定 $(|\widehat{W}_j|, \widehat{\mathbf{W}}_{-j})$ 后, 唯一未知的量是 $\widehat{W}_j$ 的符号. 由 $\theta_j = |P(\widehat{W}_j > 0 \mid |\widehat{W}_j|, \widehat{\mathbf{W}}_{-j}) - 1/2|$, 可得 $P(\widehat{W}_j > 0 \mid |\widehat{W}_j|, \widehat{\mathbf{W}}_{-j}) - 1/2 \leq \theta_j$, $P(\widehat{W}_j > 0 \mid |\widehat{W}_j|, \widehat{\mathbf{W}}_{-j}) \leq 1/2 + \theta_j$. 因此,

$$\begin{aligned} E[R_\epsilon(L)] &\leq \sum_{j \in \mathcal{A}^c} E \left\{ \left(\frac{1}{2} + \theta_j \right) \cdot \frac{I(|\widehat{W}_j| \geq L_j, \theta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_j)} \right\} \\ &\leq \left(\frac{1}{2} + \theta_j \right) \cdot \sum_{j \in \mathcal{A}^c} E \left\{ \frac{I(|\widehat{W}_j| \geq L_j, \theta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_j)} \right\} \\ &\leq \left(\frac{1}{2} + \theta_j \right) \cdot \left[\sum_{j \in \mathcal{A}^c} E \left\{ \frac{I(\widehat{W}_j \geq L_j, \theta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_j)} \right\} \right. \\ &\quad \left. + \sum_{j \in \mathcal{A}^c} E \left\{ \frac{I(\widehat{W}_j \leq -L_j)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_j)} \right\} \right] \\ &= \left(\frac{1}{2} + \theta_j \right) \cdot \left[E[R_\epsilon(L)] + \sum_{j \in \mathcal{A}^c} E \left\{ \frac{I(\widehat{W}_j \leq -L_j)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_j)} \right\} \right]. \end{aligned}$$

由文献 [2, 引理 6] 可知, 对于任意 j 和 k , 如果 $\widehat{W}_j \leq -\min(L_j, L_k)$ 和 $\widehat{W}_k \leq -\min(L_j, L_k)$, 则 $L_j = L_k$. 如果对所有 $j \in \mathcal{A}^c$, 有 $\widehat{W}_j > -L_j$, 则

$$\sum_{j \in \mathcal{A}^c} E \left\{ \frac{I(\widehat{W}_j \leq -L_j)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_j)} \right\} = 0;$$

否则,

$$\sum_{j \in \mathcal{A}^c} E \left\{ \frac{I(\widehat{W}_j \leq -L_j)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_j)} \right\} = \sum_{j \in \mathcal{A}^c} E \left\{ \frac{I(\widehat{W}_j \leq -L_j)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} I(\widehat{W}_k \leq -L_k)} \right\} = 1.$$

因此,

$$E[R_\epsilon(L)] \leq \left(\frac{1}{2} + \theta_j \right) \cdot [E[R_\epsilon(L)] + 1] \leq \left(\frac{1}{2} + \epsilon \right) \cdot [E[R_\epsilon(L)] + 1].$$

通过简单的计算可得 $E[R_\epsilon(L)] \leq \frac{1+2\epsilon}{1-2\epsilon}$.

因此, FSDS 方法对应的 FDR 可表示为

$$\begin{aligned} \text{FDR} &= E \left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L)}{1 \vee \sum_{j=1}^p I(\widehat{W}_j \geq L)} \right] \\ &= E \left[\frac{\sum_{j \in \mathcal{A}^c} \{I(\widehat{W}_j \geq L, \theta_j \leq \epsilon) + I(\widehat{W}_j \geq L, \theta_j > \epsilon)\}}{1 \vee \sum_{j=1}^p I(\widehat{W}_j \geq L)} \right] \\ &\leq \alpha \cdot E[R_\epsilon(L)] + E \left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L, \theta_j > \epsilon)}{1 \vee \sum_{j=1}^p I(\widehat{W}_j \geq L)} \right]. \end{aligned}$$

因此,

$$\text{FDR} \leq \min_{\epsilon \geq 0} \left\{ \alpha \cdot \frac{1+2\epsilon}{1-2\epsilon} + P \left(\max_{j \in \mathcal{A}^c \cap S} \theta_j > \epsilon \right) \right\}.$$

附录 A.4 引理 3.2 的证明

考虑寻找 (2.5) 所定义的阈值 L 的过程. 不失一般性, 假设对于 $j \in \mathcal{A}^c$, $\widehat{W}_j$ 已按顺序排列为 $|\widehat{W}_1| \geq |\widehat{W}_2| \geq \dots \geq |\widehat{W}_d| > 0$, 进一步定义 $|\widehat{W}_{d+1}| = 0$. 为了找到阈值 L , 可以从最小的值 $|\widehat{W}_{d+1}| = 0$ 开始尝试, 然后依次尝试第二小的值 $|\widehat{W}_d|$, 接着尝试 $|\widehat{W}_{d-1}|$, 以此类推. 直到找到满足 (2.5) 所定义的阈值 L 时停止. 在这个过程中, L 可以看作一个停止时间.

特别地, 对于 $k = d+1, d, d-1, \dots, 1$, 定义

$$M_k = \frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq |\widehat{W}_k|)}{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -|\widehat{W}_k|)} = \frac{\sum_{j \in \mathcal{A}^c} I(j \leq k, \widehat{W}_j \geq 0)}{1 + \sum_{j \in \mathcal{A}^c} I(j \leq k, \widehat{W}_j < 0)}.$$

根据 ν_k 和 S_k 的定义, 可得 $\nu_k = I(\widehat{W}_k < 0)$, $S_k = \nu_1 + \nu_2 + \dots + \nu_k$, 因此

$$M_k = \frac{\sum_{j \in \mathcal{A}^c} I(j \leq k, \widehat{W}_j \geq 0)}{1 + \sum_{j \in \mathcal{A}^c} I(j \leq k, \widehat{W}_j < 0)} = \frac{(1 - \nu_1) + (1 - \nu_2) + \dots + (1 - \nu_k)}{1 + \nu_1 + \nu_2 + \dots + \nu_k} = \frac{k+1}{1 + S_k} - 1.$$

令 $\mathcal{F}_k$ 表示 $\nu_1 + \nu_2 + \cdots + \nu_k, \nu_{k+1}, \dots, \nu_{d+1}$ 生成的 σ 代数. 给定 $\mathcal{F}_k$, 可以判断 k 是否在重要变量指标集 $\mathcal{A}$ 中. 因此, L 是关于过程 $M_{d+1}, M_d, \dots, M_1$ 及其滤波序列 $\mathcal{F}_{d+1} \subset \mathcal{F}_d \subset \cdots \subset \mathcal{F}_1$ 的反向停止时间 (从 $d+1$ 到 1).

接下来, 证明过程 $M_{d+1}, M_d, \dots, M_1$ 是关于滤波序列 $\mathcal{F}_{d+1} \subset \mathcal{F}_d \subset \cdots \subset \mathcal{F}_1$ 的反向上鞅. 由文献 [19, 引理 8] 可知 $P(\nu_k = 1 | \mathcal{F}_k) = S_k/k$. 如果 $S_k = 0$, 则 $S_{k-1} = 0$, 从而 $M_{k-1} = \frac{k-1+1}{1+S_{k-1}} - 1 = k-1 < k = M_k$. 如果 $S_k > 0$, 则

$$\begin{aligned} E(M_{k-1} | \mathcal{F}_k) &= \left(\frac{k}{1+S_k} - 1 \right) \cdot P(\nu_k = 0 | \mathcal{F}_k) + \left(\frac{k}{1+S_k-1} - 1 \right) \cdot P(\nu_k = 1 | \mathcal{F}_k) \\ &= \left(\frac{k}{1+S_k} - 1 \right) \cdot \left(1 - \frac{S_k}{k} \right) + \left(\frac{k}{1+S_k-1} - 1 \right) \cdot \frac{S_k}{k} \\ &= \frac{k-S_k}{1+S_k} \\ &= M_k. \end{aligned}$$

因此, $E(M_{k-1} | \mathcal{F}_k) \leq M_k$, 意味着 $M_{d+1}, M_d, \dots, M_1$ 是关于滤波序列 $\mathcal{F}_{d+1} \subset \mathcal{F}_d \subset \cdots \subset \mathcal{F}_1$ 的反向上鞅. 因为 L 是停止时间, 所以由上鞅的可选停止定理可得

$$E \left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L)}{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -L)} \right] \leq E(M_d) = E \left(\frac{d+1}{1+S_d} - 1 \right) = E \left(\frac{1-S_d/d}{1/d+S_d/d} \right).$$

由文献 [19, 引理 11] 可知, 当 $(n, p) \rightarrow \infty$ 时, $S_d/d \xrightarrow{p} 1/2$. 定义函数 $f(x) = (1-x)/(1/d+x)$, 因为 $f(x)$ 有界且连续, 可以得到

$$E \left(\frac{1-S_d/d}{1/d+S_d/d} \right) \rightarrow E \left(\frac{1-1/2}{0+1/2} \right) = 1.$$

因此

$$E \left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L)}{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -L)} \right] \leq 1 + o(1), \quad (\text{A.7})$$

其中当 $(n, p) \rightarrow \infty$ 时, $o(1) \rightarrow 0$.

附录 A.5 定理 3.3 的证明

根据 (2.5) 中阈值 L 的定义、FDR 的定义和引理 3.2 的结果可得

$$\begin{aligned} \text{FDR} &= E \left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L)}{1 \vee \sum_{j=1}^p I(\widehat{W}_j \geq L)} \right] \\ &= E \left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L)}{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -L)} \cdot \frac{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -L)}{1 \vee \sum_{j=1}^p I(\widehat{W}_j \geq L)} \right] \\ &\leq E \left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L)}{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -L)} \cdot \alpha \right] \\ &= \alpha \cdot \left[\frac{\sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \geq L)}{1 + \sum_{j \in \mathcal{A}^c} I(\widehat{W}_j \leq -L)} \right] \\ &\leq \alpha + o(1). \end{aligned}$$

证毕.

Feature screening and false discovery rate control for ultrahigh-dimensional survival data based on adaptive threshold selection

Jing Zhang & Hengjian Cui

Abstract In the analysis of ultrahigh-dimensional survival data, identifying key factors associated with patients' survival times from a large set of covariates is essential for optimizing treatment strategies and advancing precision medicine. Feature screening is an effective approach to address this issue, but one key and challenging problem is how to choose an appropriate threshold to distinguish important variables from irrelevant ones. In this paper, we propose a data-driven threshold selection method based on a data-splitting strategy and a reassembling technique, enabling the selection of an appropriate threshold in different scenarios to better distinguish between active and inactive covariates. Based on the selected threshold, we introduce a feature screening method for ultrahigh-dimensional right-censored survival data. The significant advantage of this method is that it not only allows for rapid and effective dimensionality reduction while accurately retaining all important variables but also controls the false discovery rate (FDR) within a certain range. Under certain regularization conditions, we establish the theoretical framework for the proposed procedure, including the sure screening property, the rank consistency property, the finite-sample theory for FDR control, and the asymptotic theories for FDR control. Numerical simulations show that the proposed method performs excellently in both variable selection and FDR control, and outperforms existing methods significantly. Finally, we apply the proposed method to analyze the diffuse large B-cell lymphoma dataset, further validating its effectiveness and practical applicability.

Keywords false discovery rate control, feature screening, threshold selection, ultrahigh-dimensional right-censored survival data

MSC(2020) 62N01, 62P10, 62R07

doi: 10.1360/SSM-2025-0025