中国科学: 数学 2026 年 第 56 卷 第 ? 期: 1 ∼ 40
SCIENTIA SINICA Mathematica
论 文

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

基于超高维分类数据的无模型特征筛选

邹丰[1], 张虎[1,2], 崔恒建[3*]

1. 中南财经政法大学统计与数学学院, 武汉 430073;
2. 中南财经政法大学大数据研究院, 武汉 430073;
3. 首都师范大学数学科学学院, 北京 100048
E-mail: zoufengjiayou@163.com, zhh11497@sina.com, hjcui@bnu.edu.cn

收稿日期: 2024-10-19; 接受日期: 2025-07-20; 网络出版日期: 2025-11-03; * 通信作者
国家自然科学基金 (批准号: 12201317, 12031016 和 12371261)、国家社会科学基金 (批准号: 23&ZD057)、国家资助博士后研究人员计划 C 档 (批准号: GZC20233143) 和中南财经政法大学中央高校基本科研业务费专项资金 (批准号: 2722024BQ065, 2722024EJ002 和 2722024EJ026) 资助项目

摘要 本文提出一种稳健的半距离相关 (robust semi-distance correlation, RSDCor) 去度量一个分类随机变量 $Y \in \mathbb{R}^1$ 和一个连续型随机向量 $\boldsymbol{z} \in \mathbb{R}^d$ ($d \geqslant 1$) 的相关性. RSDCor 可以看作是 Zhong 等 (2024) 提出的半距离相关 (semi-distance correlation, SDCor) 的一个推广. 与 SDCor 相比, RSDCor 的显著优势是对厚尾分布或者异常值比较稳健. 此外, RSDCor 满足 $0 \leqslant \mathrm{RSDCor}(\boldsymbol{z}, Y) \leqslant 1$ 且当且仅当 $\boldsymbol{z}$ 与 Y 独立时, $\mathrm{RSDCor}(\boldsymbol{z}, Y) = 0$. 进一步地, 本文基于 RSDCor 提出一种适用于超高维分类数据的特征筛选程序, 简称为 RSD-SIS. 理论上, 本文建立了 RSD-SIS 的确定筛选性质和秩相合性. RSD-SIS 具有以下几个优点: 第一, 它不依赖于具体的模型假设以及预测变量的有限矩假设; 第二, 它也可以直接用于组预测变量与分类响应变量的组特征筛选; 第三, 它允许响应变量的类别个数以 $O(n^\kappa)$ 的速度发散, 其中 $\kappa \geqslant 0$. Monte Carlo 模拟和实际数据分析均表明, RSD-SIS 与现有的筛选程序相比具有非常优秀的表现.

关键词 相关性度量 超高维分类 特征筛选 稳健 筛选性质

MSC (2020) 主题分类 62H12, 62H20, 62H30

1 引言

在信息化和数字化迅猛发展的今天, 越来越多的超高维稀疏分类数据如生物医学成像数据 [31]、肿瘤分类数据 [1] 和图像分类数据 [19] 等涌现于人们的实际生活, 其中超高维指的是预测变量的个数 p 远大于样本量 n, 甚至 p 随着 n 呈指数增长, 而稀疏性意味着只有少数的预测变量对分类响应变量真正有贡献. 由文献 [8] 可知, 该类数据的超高维属性给一些经典的分类方法如线性判别分析 (LDA) 和二次判别分析 (QDA) 等带来了巨大的挑战, 其主要包括如下内容. (1) 由于 p 相对于 n 较大, 故估计得

英文引用格式: Zou F, Zhang H, Cui H J. Model-free feature screening for ultrahigh-dimensional classification data (in Chinese). Sci Sin Math, 2026, 56: 1–40, doi: 10.1360/SSM-2024-0320

© 2025《中国科学》杂志社

到的样本协方差矩阵将遭受奇异性问题. 此外, 如文献 [2] 所述, 即使假设总体协方差矩阵已知, 其条件数也可能随着 p 的发散而趋于无穷, 进而导致其求逆过程不稳定并造成误差累积. (2) 当对总体均值向量进行估计时, 由于维数较高, 故即使总体均值向量的每个部分可以准确估计, 但其累积后的估计误差可能非常大, 从而增加误分类率. 为了应对上述挑战, 对超高维稀疏分类数据进行降维似乎是一个不可缺少的环节. 通过降维, 可以去掉一些冗余的预测变量并保留相关重要变量, 进而增强分类模型的可解释性并提高预测精度. 虽然目前已有一些降维方法如 DALASS [32], PenLDA [34], SDA [5], LDAL1-S 和 2DLDAL1-S [16] 等可直接用于分类数据的降维, 但是文献 [10] 指出, 以上基于惩罚的降维方法在超高维数据中可能面临计算复杂度大、估计精度不高和算法不稳定等问题.

为了更好地应对超高维数据中的降维问题, 文献 [9] 基于超高维线性模型提出了确定性独立筛选 (sure independence screening, SIS) 的思想, 并根据单个预测变量与响应变量的 Pearson 相关系数的绝对值大小对变量的重要性进行排序. 通过过滤掉相关性较弱的变量, 便可快速地将原来的超高维 p 降到一个相对温和的维数 d_n. 文献 [9] 在一定条件下证明了 SIS 程序具有确定筛选性质 (sure screening property), 即筛选后得到的变量集合包含全部重要变量的概率随着样本量的增大而趋于 1. 通过该文献不难得到以下两点启发: (1) SIS 可以过滤掉多数与响应变量不相关的变量从而达到快速降维的目的, 随后一些基于惩罚的降维方法可以更好地应用于筛选后的模型; (2) 文献 [9] 中的响应变量虽然是连续型变量, 但是 SIS 的思想也适用于响应变量为分类变量的超高维数据.

由于文献 [9] 的开创性工作, 特征/变量筛选受到了学者们的广泛关注. 对于连续型响应变量的超高维特征筛选问题, 已有大量的文献对此进行了研究 (参见文献 [11,17,18,36]). 与此同时, 考虑到现实中不少超高维数据的响应变量为分类变量, 故一些基于超高维分类数据的特征筛选方法近年来也得到了发展. 例如, 文献 [8] 提出了基于两样本 t- 统计量的筛选方法, 但该方法只适用于二元分类问题且对预测变量的分布较为敏感. 文献 [20] 针对二元分类问题提出了一种基于 Kolmogorov-Smirnov 统计量的特征筛选程序, 并且文献 [21] 将该方法进一步推广到多类别响应. 文献 [24] 基于正态性假设构建了一种成对确定筛选程序. 然而, 由于该筛选程序基于不同类别下的均值差异, 故其对于厚尾数据表现不佳. 文献 [6] 利用边际分布函数和条件分布函数之间的差异构建了一个特征筛选指标, 且基于该指标建立的筛选程序对厚尾分布和异常值具有稳健性. 文献 [13] 通过赋予分布尾部的观测值更大的权重, 进而提出了一种 Anderson-Darling 型筛选指标. 文献 [22] 利用边际信息熵和条件信息熵的差异提出了一种基于熵的特征筛选程序. 文献 [4] 构建了一致性指数筛选程序来处理具有分类响应的超高维数据. 不难发现, 以上提到的筛选程序均是基于单个预测变量与分类响应变量的特征筛选, 故其无法较好地处理组预测变量与分类响应的组筛选问题. 然而, 文献 [3] 指出, 在许多超高维数据中预测变量自然而然地含有组结构. 因此, 发展出一些基于组预测变量与分类响应的组特征筛选程序也十分重要.

在现有的文献中, 已有一些基于响应变量为连续型变量的组特征筛选程序, 如文献 [15,18,25] 等. 而对于分类响应变量的组筛选问题, 最近几年也陆续有学者提出了一些相应的方法. 例如, 文献 [33] 利用基尼杂质系数与条件基尼杂质系数的差异构建了一个分类筛选指标. 然而, 由于其对应的筛选程序是为分类预测变量而设计, 故当预测变量为数值型时, 常需要利用切片手段将数值型预测变量转化为类别变量, 但这可能意味着丢失变量的大量信息. 此外, 如何进行切片也是应用该方法时遇到的一个挑战. 文献 [35] 提出了一种基于半距离相关的组筛选程序, 其不依赖未知调节参数但要求预测变量满足次指数假设. 这意味着当预测变量服从厚尾分布时, 该筛选程序的表现可能会大打折扣. 文献 [27] 提出了一种基于基尼距离相关的组筛选程序, 但由于受次指数假设的限制, 故其在厚尾数据下亦表现不佳. 事实上, 文献 [27,33,35] 不仅适用于组特征筛选, 还可直接用于单个预测变量与分类响应变量的筛选问题. 经过进一步分析发现, 无论是文献 [35] 提出的半距离还是文献 [27] 提出的基尼距离, 这两种

新定义的距离的本质都是基于特征函数和条件特征函数差值的加权积分. 不同的是, 由于文献 [27,35] 对这两种距离所采用的标准化方法不一样, 从而分别定义了半距离相关和基尼距离相关.

针对以上筛选程序的不足, 本文提出利用一种稳健的半距离相关 (robust semi-distance correlation, RSDCor) 去度量一个分类变量和连续型随机向量的相关性, 其中 RSDCor 不依赖于变量的有限矩假设且当且仅当该两变量独立时, RSDCor 等于 0. 与半距离相关不同, RSDCor 含有一个未知的窗宽参数. 当 RSDCor 中的窗宽参数趋于无穷时, 此时的 RSDCor 等价于文献 [35] 提出的半距离相关. 随着窗宽参数取不同的正数, 我们可以得到一类 RSDCor. 从这个角度来说, RSDCor 可以看作是半距离相关的一种推广. 受文献 [12] 的启发, 本文通过利用中值启发式 (median heuristic) 方法选择窗宽, 此时的 RSDCor 具有位置尺度不变性. 更进一步地, 本文利用基于中值启发式的 RSDCor 发展出了一种适用于超高维分类数据的特征筛选程序, 简称为 RSD-SIS. 后续的理论分析和数值模拟表明, RSD-SIS 不仅具有确定筛选性质和秩相合性, 而且还具有以下优点: (1) 它不依赖具体的模型假设, 也不要求预测变量满足有限矩假设; (2) 它既适用于单个预测变量的筛选, 也适用于组变量筛选; (3) 它允许分类响应变量的类别数以 $O(n^\kappa)$ 的速度发散, 其中 $\kappa \geqslant 0$.

本文余下内容的安排如下. 第 2 节介绍 RSDCor 的定义及其相关的理论性质. 第 3 节介绍筛选程序 RSD-SIS 以及其确定筛选性质和秩相合性. 第 4 节利用数值模拟和一个实际数据分析去评价 RSD-SIS 在有限样本下的表现. 第 5 节对稳健的协方差距离 RSDCov 以及筛选程序 RSD-SIS 进行讨论和总结. 附录 A 和 B 给出相关公式、性质、定理的证明.

2 稳健的半距离相关

2.1 稳健的半距离协方差及其性质

为了介绍 RSDCor, 本文先给出稳健的半距离协方差 (robust semi-distance covariance, RSDCov) 的定义. 令 $\boldsymbol{z} \in \mathbb{R}^d$ ($d \geqslant 1$) 为 d 维连续型随机向量 (注意, 在后文中, 若无特别说明, 则均默认 $\boldsymbol{z}$ 连续), 分类变量 Y 满足 $P(Y = L_r) = p_r > 0$, $r = 1, \ldots, R$. 令 $\psi_{\boldsymbol{z}}(\boldsymbol{t})$ 和 $\psi_{\boldsymbol{z}|r}(\boldsymbol{t})$ 分别表示 $\boldsymbol{z}$ 的特征函数以及给定 $Y = L_r$ 和 $\boldsymbol{z}$ 的条件特征函数. 根据特征函数的性质, 易知 $\boldsymbol{z}$ 与 Y 独立当且仅当对于任意的 $Y = L_r$ 和 $\boldsymbol{t} \in \mathbb{R}^d$, 均有

$$E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}} \mid Y = L_r), \tag{2.1}$$

其中 i 表示虚数单位且满足 $\mathrm{i}^2 = -1$, T 表示转置. 进一步地, 不难发现 (2.1) 等价于

$$\sum_{r=1}^{R} p_r \int_{\mathbb{R}^d} |E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}}) - E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}} \mid Y = L_r)|^2 \omega(\boldsymbol{t}) d\boldsymbol{t} = 0, \tag{2.2}$$

其中当 ψ 为复函数时, $|\psi|^2 = \psi\bar{\psi}$, $\omega(\boldsymbol{t}) > 0$ 为权重函数. 在 (2.2) 中, 允许 $\omega(\boldsymbol{t})$ 有不同的选择. 而本文采用的 $\omega(\boldsymbol{t})$ 为随机向量 $\mathcal{T} = (T_1, \ldots, T_d)^{\mathrm{T}}$ 的概率密度函数, 且 $\mathcal{T}$ 的特征函数满足

$$\psi_{\mathcal{T}}(\boldsymbol{\theta}) = E\mathrm{e}^{\mathrm{i}\boldsymbol{\theta}^{\mathrm{T}}\mathcal{T}} = \mathrm{e}^{-(\frac{\|\boldsymbol{\theta}\|}{\sigma})^{\alpha}}, \quad 0 < \alpha \leqslant 2, \quad \sigma > 0, \tag{2.3}$$

其中 $\|\cdot\|$ 表示 Euclid 范数, 窗宽参数 σ 和指数参数 α 均为未知参数且依赖于 $\mathcal{T}$ 的具体分布. 从本质上说, 特征函数满足 (2.3) 的分布是一类特殊的关于原点对称的 α 稳定 (symmetric α-stable) 分布, 欲知更多细节可参见文献 [26,38]. 为了叙述方便, 不妨称满足 (2.3) 的分布为 $\mathcal{S}\alpha\mathcal{S}$ 分布. 如文献 [23] 所述, $\mathcal{S}\alpha\mathcal{S}$ 分布的密度函数在绝大多数情形下均没有显示表达且非常复杂, 故文献 [26] 利用特征函数来定义 d 元 $\mathcal{S}\alpha\mathcal{S}$ 分布. 综合 (2.1)–(2.3), 为度量 $\boldsymbol{z}$ 和 Y 的相关性, 本文给出如下定义.

定义 2.1 对于任意的 $\boldsymbol{z}=(Z_1,\ldots,Z_d)^{\mathrm{T}}$, $Y\in\{L_1,\ldots,L_R\}$, 令 (2.3) 中 $\psi_{\mathcal{T}}(\boldsymbol{\theta})$ 的指数参数 $\alpha=1$, $\omega(\boldsymbol{t})$ 为变量 $\mathcal{T}$ 的概率密度函数, 则定义 $\boldsymbol{z}$ 和 Y 的稳健的半距离协方差为

$$\mathrm{RSDCov}(\boldsymbol{z},Y)=\sum_{r=1}^{R}p_r\int_{\mathbb{R}^d}|E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}})-E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}}\mid Y=L_r)|^2\omega(\boldsymbol{t})d\boldsymbol{t}. \tag{2.4}$$

由 (2.4) 易知, $\mathrm{RSDCov}(\boldsymbol{z},Y)\geqslant 0$ 当且仅当 $\boldsymbol{z}\perp\!\!\!\perp Y$ 时, $\mathrm{RSDCov}(\boldsymbol{z},Y)=0$. 令 $(\boldsymbol{z}_1,Y_1)$ 和 $(\boldsymbol{z}_2,Y_2)$ 独立同分布于 $(\boldsymbol{z},Y)$, 则 $\mathrm{RSDCov}(\boldsymbol{z},Y)$ 有如下简单的显式表达:

$$\mathrm{RSDCov}(\boldsymbol{z},Y)=E[\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}D(Y_1,Y_2)], \tag{2.5}$$

其中 $\sigma>0$, $D(Y_1,Y_2)=\sum_{r=1}^{R}I(Y_1=L_r,Y_2=L_r)/p_r-1$ 且 $I(\cdot)$ 为指示函数. (2.5) 表明, 所谓的 RSDCov 实际上是两部分乘积的期望, 其中一部分是关于 $\boldsymbol{z}$ 之间的以 1/e 为底的指数距离, 另一部分是关于分类变量 Y 的散度度量. 而稳健指的是 $\mathrm{RSDCov}(\boldsymbol{z},Y)$ 不依赖于变量 $\boldsymbol{z}$ 的有限矩假设.

注 2.1 在定义 2.1 中, 本文只考虑了 $\alpha=1$ 的情形. 若令 $\alpha\in(0,2]$, 则有 $\mathrm{RSDCov}(\boldsymbol{z},Y)=E[\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})^{\alpha}}D(Y_1,Y_2)]$. 这一点实际上与文献 [30] 提出的距离协方差 (distance covariance, DC) 类似, 只不过在使用 DC 时通常默认距离 $|\cdot|^{\delta}$ $(0<\delta<2)$ 中的指数参数 $\delta=1$. 通过进一步分析可以发现, 当 $\alpha=1$ 时, 变量 $\mathcal{T}$ 服从位置参数为 0、散度矩阵为 $\boldsymbol{I}/\sigma^2$、自由度为 1 的 d 元 t 分布, 其中 $\boldsymbol{I}$ 表示单位阵. 此时, 其对应的概率密度函数 $\omega(\boldsymbol{t})=\frac{\Gamma((d+1)/2)}{\pi^{(d+1)/2}}\frac{1/\sigma}{(1/\sigma^2+\|\boldsymbol{t}\|^2)^{(d+1)/2}}$, 其中 $\Gamma(\cdot)$ 表示伽马函数; 当 $\alpha=2$ 时, 变量 $\mathcal{T}$ 服从位置参数为 0、协方差矩阵为 $2\boldsymbol{I}/\sigma^2$ 的 d 元 Gauss 分布, 其对应的密度函数 $\omega(\boldsymbol{t})=\frac{\sigma^d}{2^{3d/2}\pi^{d/2}}\mathrm{e}^{-\sigma^2\|\boldsymbol{t}\|^2/4}$. 此外, 如果在 (2.4) 中令 $\omega(\boldsymbol{t})=\frac{\Gamma((p+1)/2)}{|\boldsymbol{t}|^{p+1}\pi^{(p+1)/2}}$, 则有 $\mathrm{RSDCov}(\boldsymbol{z},Y)=E[\|\boldsymbol{z}_1-\boldsymbol{z}_2\|(1-\sum_{r=1}^{R}I(Y_1=L_r,Y_2=L_r)/p_r)]$, 即文献 [35] 提出的半距离协方差 (semi-distance covariance, SDCov). 易知, SDCov 和 RSDCov 的本质都是基于特征函数和条件特征函数差值的加权积分. 但是, RSDCov 相对于 SDCov 的一个显著优势在于 RSDCov 不依赖于变量 $\boldsymbol{z}$ 的有限矩假设.

注 2.2 由 (2.5) 可知, $\mathrm{RSDCov}(\boldsymbol{z},Y)$ 中含有未知窗宽参数 σ. 当 σ 的取值在 $(0,+\infty)$ 上变化时, 实际上可以得到一类 RSDCov. 若令 $\sigma\to 0$, 则易知 $\mathrm{RSDCov}(\boldsymbol{z},Y)\to 0$; 若令 $\sigma\to+\infty$, 由于 $\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}\approx 1-\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma}$, 故 $\mathrm{RSDCov}(\boldsymbol{z},Y)=E[\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}D(Y_1,Y_2)]\approx\frac{1}{\sigma}E[-\|\boldsymbol{z}_1-\boldsymbol{z}_2\|D(Y_1,Y_2)]$. 令 $\mathrm{SDCov}(\boldsymbol{z},Y)=E[-\|\boldsymbol{z}_1-\boldsymbol{z}_2\|D(Y_1,Y_2)]$, 则 $\mathrm{RSDCov}(\boldsymbol{z},Y)\approx\frac{1}{\sigma}\mathrm{SDCov}(\boldsymbol{z},Y)$, 其中 $\mathrm{SDCov}(\boldsymbol{z},Y)$ 即为文献 [35] 提出的半距离协方差. 换而言之, 当 σ 趋于正无穷时, RSDCov 与 SDCov 成倍数关系. 从后续的第 2.3 小节可以看到, 将 RSDCov 标准化后倍数因子 $1/\sigma$ 可以去掉. 基于这个角度, 本文提出的 RSDCov 在某种程度上可以看作是 SDCov 的推广.

性质 2.1 令 $(\boldsymbol{z}_1,Y_1)$, $(\boldsymbol{z}_2,Y_2)$, $(\boldsymbol{z}_3,Y_3)$, $(\boldsymbol{z}_4,Y_4)$ 独立且与 $(\boldsymbol{z},Y)$ 同分布, 则有

(a) $\mathrm{RSDCov}(\boldsymbol{z},Y)=E[U_{\boldsymbol{z}}D(Y_1,Y_2)]$, 其中 $U_{\boldsymbol{z}}=\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}-E_{\boldsymbol{z}_1}[\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}]-E_{\boldsymbol{z}_2}[\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}]+E[\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}]$ 且 $E_{\boldsymbol{z}_1}\varphi(\boldsymbol{z}_1,\boldsymbol{z}_2)$ 表示函数 $\varphi(\boldsymbol{z}_1,\boldsymbol{z}_2)$ 关于 $\boldsymbol{z}_1$ 取期望, $D(Y_1,Y_2)$ 的定义见 (2.5);

(b) $0\leqslant\mathrm{RSDCov}(\boldsymbol{z},Y)\leqslant\mathrm{dvar}(\boldsymbol{z})\sqrt{R-1}<\sqrt{2(R-1)}$, 其中 $\mathrm{dvar}(\boldsymbol{z})$ 满足

$$\{\mathrm{dvar}(\boldsymbol{z})\}^2=E[U_{\boldsymbol{z}}^2]=E[\mathrm{e}^{-2(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}]-2E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)-(\|\boldsymbol{z}_1-\boldsymbol{z}_3\|/\sigma)}]+E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)-(\|\boldsymbol{z}_3-\boldsymbol{z}_4\|/\sigma)}].$$

(c) $\mathrm{RSDCov}(\boldsymbol{z},Y)=\sum_{r>s}p_rp_s(\mathcal{G}_{rr}+\mathcal{G}_{ss}-2\mathcal{G}_{rs})$, 其中 $\mathcal{G}_{rs}=E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}\mid Y_1=L_r,Y_2=L_s]$. 此外, 对于任意的 $r\neq s$, 有 $\mathcal{G}_{rr}+\mathcal{G}_{ss}-2\mathcal{G}_{rs}=\int_{\mathbb{R}^d}|E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}}\mid Y=L_r]-E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}}\mid Y=L_s]|^2\omega(\boldsymbol{t})d\boldsymbol{t}\geqslant 0$, 其中 $\omega(\boldsymbol{t})$ 的定义同 (2.4).

性质 2.1(a) 给出了 $\mathrm{RSDCov}(\boldsymbol{z},Y)$ 的另一种表达形式, 其中 $U_{\boldsymbol{z}}$ 是 $\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}$ 的双中心化版本且满足 $E_{\boldsymbol{z}_1}(U_{\boldsymbol{z}})=E_{\boldsymbol{z}_2}(U_{\boldsymbol{z}})=E(U_{\boldsymbol{z}})=0$. 性质 2.1(b) 给出了 $\mathrm{RSDCov}(\boldsymbol{z},Y)$ 的上下界. 相比于 (2.5) 和性质 2.1(a), 根据性质 2.1(c) 更容易看出 $\mathrm{RSDCov}(\boldsymbol{z},Y)\geqslant 0$. 当 $\mathrm{RSDCov}(\boldsymbol{z},Y)=0$ 时, 意味着对于任意的 $r\neq s$, 有 $E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}}\mid Y=r]=E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}}\mid Y=s]$, 即 $\boldsymbol{z}\perp\!\!\!\perp Y$.

2.2 RSDCov 的样本估计及其性质

令 $(\boldsymbol{z}_1,Y_1),\ldots,(\boldsymbol{z}_n,Y_n)$ 为 n 个独立同分布于 $(\boldsymbol{z},Y)$ 的随机样本. 给定 σ, 对于 $k,l=1,\ldots,n$, 定义 $a_{kl}=\mathrm{e}^{-(\|\boldsymbol{z}_k-\boldsymbol{z}_l\|/\sigma)}$, $\bar{a}_{k\cdot}=\sum_{l=1}^n a_{kl}/n$, $\bar{a}_{\cdot l}=\sum_{k=1}^n a_{kl}/n$, $\bar{a}_{\cdot\cdot}=\sum_{k,l=1}^n a_{kl}/n^2$, $A_{kl}=a_{kl}-\bar{a}_{k\cdot}-\bar{a}_{\cdot l}+\bar{a}_{\cdot\cdot}$ 以及 $\hat{p}_r=\sum_{k=1}^n I(Y_k=L_r)/n=:n_r/n$, $B_{kl}=\sum_{r=1}^R I(Y_k=L_r,Y_l=L_r)/\hat{p}_r-1$.

定义 2.2 对于给定的窗宽参数 σ, 定义 RSDCov 的样本估计量为

$$\mathrm{RSDCov}_n(\boldsymbol{z},Y)=\frac{1}{n^2}\sum_{k=1}^n\sum_{l=1}^n A_{kl}B_{kl}. \tag{2.6}$$

令 $\boldsymbol{K}_{\boldsymbol{z}}$ 和 $\boldsymbol{K}_Y$ 分别表示两个 $n\times n$ 矩阵, 且 $\boldsymbol{K}_{\boldsymbol{z}}$ 和 $\boldsymbol{K}_Y$ 的第 k 行第 l 列个元素分别满足 $(\boldsymbol{K}_{\boldsymbol{z}})_{k,l}=\mathrm{e}^{-(\|\boldsymbol{z}_k-\boldsymbol{z}_l\|/\sigma)}$, $(\boldsymbol{K}_Y)_{k,l}=\sum_{r=1}^R I(Y_k=L_r,Y_l=L_r)/\hat{p}_r-1$, 则 $\mathrm{RSDCov}_n(\boldsymbol{z},Y)$ 有如下的矩阵表达形式:

$$\mathrm{RSDCov}_n(\boldsymbol{z},Y)=\frac{1}{n^2}\mathrm{tr}(\widetilde{\boldsymbol{K}}_{\boldsymbol{z}}\boldsymbol{K}_Y), \tag{2.7}$$

其中 $\mathrm{tr}(\cdot)$ 表示求迹函数, 矩阵 $\widetilde{\boldsymbol{K}}_{\boldsymbol{z}}$ 满足

$$\widetilde{\boldsymbol{K}}_{\boldsymbol{z}}=\boldsymbol{K}_{\boldsymbol{z}}-\mathrm{L}\mathrm{L}^{\mathrm{T}}\boldsymbol{K}_{\boldsymbol{z}}/n-\boldsymbol{K}_{\boldsymbol{z}}\mathrm{L}\mathrm{L}^{\mathrm{T}}/n+\mathrm{L}^{\mathrm{T}}\boldsymbol{K}_{\boldsymbol{z}}\mathrm{L}\mathrm{L}\mathrm{L}^{\mathrm{T}}/n^2,\quad \mathrm{L}=(1,1,\ldots,1)^{\mathrm{T}}.$$

性质 2.2 根据 (2.6) 中 $\mathrm{RSDCov}_n(\boldsymbol{z},Y)$ 的定义, 在给定 σ 下, 有

(a) $\lim_{n\to\infty}\mathrm{RSDCov}_n(\boldsymbol{z},Y)=\mathrm{RSDCov}(\boldsymbol{z},Y)$, a.s.;

(b) $\mathrm{RSDCov}_n(\boldsymbol{z},Y)=\sum_{r=1}^R\hat{p}_r\int_{\mathbb{R}^d}[E_n(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}}\mid Y=L_r)-E_n(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}})]^2\omega(\boldsymbol{t})d\boldsymbol{t}$, 其中 $E_n(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{x}})$ 表示经验特征函数, $E_n(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{x}}\mid Y=L_r)$ 表示条件经验特征函数;

(c) $0\leqslant\mathrm{RSDCov}_n(\boldsymbol{z},Y)\leqslant\mathrm{dvar}_n(\boldsymbol{z})\sqrt{R-1}<\sqrt{2(R-1)}$, 其中 $\mathrm{dvar}_n(\boldsymbol{z})$ 满足

$$\{\mathrm{dvar}_n(\boldsymbol{z})\}^2=\frac{1}{n^2}\sum_{i,j=1}^n\mathrm{e}^{-2(\frac{\|\boldsymbol{z}_i-\boldsymbol{z}_j\|}{\sigma})}-\frac{2}{n^3}\sum_{i,j,k=1}^n\mathrm{e}^{-(\frac{\|\boldsymbol{z}_i-\boldsymbol{z}_j\|+\|\boldsymbol{z}_i-\boldsymbol{z}_k\|}{\sigma})}+\frac{1}{n^4}\sum_{i,j,k,l=1}^n\mathrm{e}^{-(\frac{\|\boldsymbol{z}_i-\boldsymbol{z}_j\|+\|\boldsymbol{z}_k-\boldsymbol{z}_l\|}{\sigma})}.$$

性质 2.2(a) 刻画了 $\mathrm{RSDCov}_n(\boldsymbol{z},Y)$ 的强相合性. 性质 2.2(b) 基于 (2.4) 给出了 $\mathrm{RSDCov}(\boldsymbol{z},Y)$ 的样本估计量, 其本质上与 (2.6) 一样. 性质 2.2(c) 给出了 $\mathrm{RSDCov}_n(\boldsymbol{z},Y)$ 的上下界. 此外, 根据性质 2.2(b) 中 $\mathrm{RSDCov}_n(\boldsymbol{z},Y)$ 的数学表达式, 易知 $\mathrm{RSDCov}_n(\boldsymbol{z},Y)\geqslant 0$.

2.3 RSDCov 与稳健的鞅差散度的关系

本小节讨论 RSDCov 与稳健的鞅差散度 (robust martingale difference divergence, RMDD) 之间的关系. 令随机变量 $V\in\mathbb{R}^1$, 随机向量 $\boldsymbol{u}\in\mathbb{R}^q$ $(q\geqslant 1)$, $(V_1,\boldsymbol{u}_1),\ldots,(V_n,\boldsymbol{u}_n)$ 为来自总体 $(V,\boldsymbol{u})$ 的 n 个独立同分布的随机样本. 文献 [29] 定义给定 $\boldsymbol{u}$ 和 V 的鞅差散度 (martingale difference divergence, MDD) 如下:

$$[\mathrm{MDD}(V\mid\boldsymbol{u})]^2=\int_{\mathbb{R}^q}|E(V\mathrm{e}^{\mathrm{i}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{u}})-E(V)E(\mathrm{e}^{\mathrm{i}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{u}})|^2\omega(\boldsymbol{s})d\boldsymbol{s}$$

$$= -E[(V_1 - EV_1)(V_2 - EV_2)\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|], \tag{2.8}$$

其中 $\omega(\boldsymbol{s}) = \Gamma((q+1)/2)/(|\boldsymbol{s}|^{q+1}\pi^{(q+1)/2})$.

定义 2.3 若令 (2.8) 中的 $\omega(\boldsymbol{s})$ 为满足 (2.3) 的 q 维随机变量 $S = (s_1, \ldots, s_q)^{\mathrm{T}}$ 的概率密度函数且其对应的指数参数 $\alpha = 1$, 则给定 $\boldsymbol{u}$ 和 V 的稳健的鞅差散度 (RMDD) 可定义为

$$[\mathrm{RMDD}(V \mid \boldsymbol{u})]^2 = E[(V_1 - EV_1)(V_2 - EV_2)\mathrm{e}^{-(\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|/\sigma)}], \tag{2.9}$$

其中稳健指的是与定义在 (2.8) 中的 MDD 相比, RMDD 不依赖于变量 $\boldsymbol{u}$ 的有限矩假设. 与性质 2.1(a) 类似, 通过将 $\mathrm{e}^{-(\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|/\sigma)}$ 双中心化, 可以得到 $\mathrm{RMDD}(V \mid \boldsymbol{u})$ 的另一种表达, 即

$$[\mathrm{RMDD}(V \mid \boldsymbol{u})]^2 = E[D(V_1, V_2)U_{\boldsymbol{u}}],$$

其中 $U_{\boldsymbol{u}} = \mathrm{e}^{-(\frac{\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|}{\sigma})} - E_{\boldsymbol{u}_1}[\mathrm{e}^{-(\frac{\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|}{\sigma})}] - E_{\boldsymbol{u}_2}[\mathrm{e}^{-(\frac{\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|}{\sigma})}] + E[\mathrm{e}^{-(\frac{\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|}{\sigma})}]$, $D(V_1, V_2) = (V_1 - EV_1)(V_2 - EV_2)$. 给定 σ, 令 $c_{ij} = \mathrm{e}^{-(\|\boldsymbol{u}_i - \boldsymbol{u}_j\|/\sigma)}$, $\bar{c}_{i\cdot} = \sum_{j=1}^n c_{ij}/n$, $\bar{c}_{\cdot j} = \sum_{i=1}^n c_{ij}/n$, $\bar{c}_{\cdot\cdot} = \sum_{i,j=1}^n c_{ij}/n^2$, $C_{ij} = c_{ij} - \bar{c}_{i\cdot} - \bar{c}_{\cdot j} + \bar{c}_{\cdot\cdot}$ 以及 $\bar{V}_n = \sum_{i=1}^n V_i/n$, $D_{ij} = (V_i - \bar{V}_n)(V_j - \bar{V}_n)$, 定义 $\mathrm{RMDD}(V \mid \boldsymbol{u})$ 的样本估计量为下式的算术平方根:

$$[\mathrm{RMDD}_n(V \mid \boldsymbol{u})]^2 = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n C_{ij}D_{ij}. \tag{2.10}$$

性质 2.3 对于第 2.1 小节的分类变量 $Y \in \mathbb{R}^1$, $\boldsymbol{z} \in \mathbb{R}^d$, 有

(a) $\mathrm{RSDCov}(\boldsymbol{z}, Y) = \sum_{r=1}^R \{\mathrm{RMDD}[I(Y = L_r) \mid \boldsymbol{z}]\}^2 p_r$, 其中 $\mathrm{RMDD}[I(Y = L_r) \mid \boldsymbol{z}]$ 为给定 $\boldsymbol{z}$ 和 $I(Y = L_r)$ 的稳健的鞅差散度;

(b) $\mathrm{RSDCov}_n(\boldsymbol{z}, Y) = \sum_{r=1}^R \{\mathrm{RMDD}_n[I(Y = L_r) \mid \boldsymbol{z}]\}^2 \hat{p}_r$, 其中 $\mathrm{RMDD}_n[I(Y = L_r) \mid \boldsymbol{z}]$ 表示在样本水平下, $I(Y = L_r)$ 在给定 $\boldsymbol{z}$ 下的稳健的鞅差散度, $\hat{p}_r$ 的定义见第 2.2 小节.

2.4 稳健的半距离相关及其性质

基于第 2.1–2.3 小节的分析, 本节给出稳健的半距离相关的定义及其样本估计. 利用性质 2.1(b), 定义 $\boldsymbol{z}$ 和 Y 的稳健的半距离相关 (robust semi-distance correlation, RSDCor) 为

$$\mathrm{RSDCor}_\sigma(\boldsymbol{z}, Y) = \frac{\mathrm{RSDCov}(\boldsymbol{z}, Y)}{\mathrm{dvar}(\boldsymbol{z})\sqrt{R-1}}, \tag{2.11}$$

其中 $\mathrm{RSDCor}_\sigma(\boldsymbol{z}, Y)$ 中的下标 σ 表示其依赖于给定的窗宽参数. 结合 (2.6) 和性质 2.2(c), RSDCor_σ 的样本估计量可定义为

$$\mathrm{RSDCor}_{\sigma n}(\boldsymbol{z}, Y) = \frac{\mathrm{RSDCov}_n(\boldsymbol{z}, Y)}{\mathrm{dvar}_n(\boldsymbol{z})\sqrt{R-1}}, \tag{2.12}$$

其中 $0 \leqslant \mathrm{RSDCor}_{\sigma n}(\boldsymbol{z}, Y) \leqslant 1$. 与 Pearson 相关系数、距离相关系数 [30] 和投影相关系数 [37] 等不同, RSDCor_σ 作为一种相关性度量系数不具有位置尺度不变性, 即

$$\mathrm{RSDCor}_\sigma(\boldsymbol{a} + b\boldsymbol{H}\boldsymbol{z}, Y) \neq \mathrm{RSDCor}_\sigma(\boldsymbol{z}, Y), \tag{2.13}$$

其中 $\boldsymbol{a} \in \mathbb{R}^d$, $b \neq 0$, $\boldsymbol{H}$ 为正交矩阵. 更准确地说, RSDCor_σ 不具备尺度不变性. 受文献 [12] 的启发, 本文利用中值启发式方法去选择 σ, 如下所示.

定义 2.4 令 $T_{\boldsymbol{z}} = \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|$, $\sigma = \mathrm{Med}(T_{\boldsymbol{z}}) = F_{\boldsymbol{z}}^{-1}(1/2)$, 其中 $F_{\boldsymbol{z}}(\cdot)$ 表示 $T_{\boldsymbol{z}}$ 的累积分布函数. 此时, $\boldsymbol{z}$ 和 Y 的 RSDCor 满足位置尺度不变性, 即

$$\mathrm{RSDCor}(\boldsymbol{z}, Y) = \frac{\mathrm{RSDCov}_m(\boldsymbol{z}, Y)}{\mathrm{dvar}_m(\boldsymbol{z})\sqrt{R-1}}, \tag{2.14}$$

其中下标 m 仅表示此时 $\mathrm{RSDCov}(\boldsymbol{z}, Y)$ 以及 $\mathrm{dvar}(\boldsymbol{z})$ 中的窗宽参数为 $T_{\boldsymbol{z}}$ 的中位数 (median). 此外, 不难知道定义在 (2.14) 中的 $\mathrm{RSDCor}(\boldsymbol{z}, Y)$ 具有位置尺度不变性, 详细的证明可参见附录 B.3. 令 σ 的样本估计量为 $\hat{\sigma} = \mathrm{Med}\{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|, 1 \leqslant i < j \leqslant n\}$, 定义 $\mathrm{RSDCor}(\boldsymbol{z}, Y)$ 的样本估计为

$$\mathrm{RSDCor}_n(\boldsymbol{z}, Y) = \frac{\mathrm{RSDCov}_{m,n}(\boldsymbol{z}, Y)}{\mathrm{dvar}_{m,n}(\boldsymbol{z})\sqrt{R-1}}. \tag{2.15}$$

这里, 只需要将 (2.6) 的 $\mathrm{RSDCov}_n(\boldsymbol{z}, Y)$ 和性质 (2.2)(c) 中的 $\mathrm{dvar}_n(\boldsymbol{z})$ 的固定参数 σ 均改成 $\hat{\sigma}$, 便可分别得到 $\mathrm{RSDCov}_{m,n}(\boldsymbol{z}, Y)$ 和 $\mathrm{dvar}_{m,n}(\boldsymbol{z})$ 的表达.

注 2.3 通过对 (2.14) 和 (2.15) 进行分析, 可以得到以下几点. (1) 与 $\mathrm{RSDCov}(\boldsymbol{z}, Y)$ 和 $\mathrm{dvar}(\boldsymbol{z})$ 的固定窗宽相比, $\mathrm{RSDCov}_m(\boldsymbol{z}, Y)$ 和 $\mathrm{dvar}_m(\boldsymbol{z})$ 中的窗宽参数依赖于变量 $T_{\boldsymbol{z}}$ 的分布. 由于估计 σ 时使用了变量 $\boldsymbol{z}$ 的信息, 故 $\mathrm{RSDCov}_{m,n}(\boldsymbol{z}, Y)$ 和 $\mathrm{dvar}_{m,n}(\boldsymbol{z})$ 均是基于数据驱动的窗宽而得到的估计量. (2) 对于 $\mathrm{RSDCov}_m(\boldsymbol{z}, Y)$ 和 $\mathrm{dvar}_m(\boldsymbol{z})$ 以及 $\mathrm{RSDCov}_{m,n}(\boldsymbol{z}, Y)$ 和 $\mathrm{dvar}_{m,n}(\boldsymbol{z})$ 而言, 性质 2.1–2.3 依然成立. 例如, 欲证明 $\lim_{n\to\infty} \mathrm{RSDCov}_{m,n}(\boldsymbol{z}, Y) = \mathrm{RSDCov}_m(\boldsymbol{z}, Y)$, a.s., 可以基于 (A.13) 和 (A.14) 先将 $\mathrm{RSDCov}_{m,n}(\boldsymbol{z}, Y)$ 拆分成 3 个部分. 然后, 再结合 (A.35)–(A.38) 的证明思路以及 Borel-Cantelli 引理便可得证. (3) 当 σ 在 $(0, +\infty)$ 上变化时, 可以得到不同的 $\mathrm{RSDCor}_\sigma(\boldsymbol{z}, Y)$. 而 $\sigma = \mathrm{Med}(T_{\boldsymbol{z}})$ 可以看作是一种基于数据驱动选择的窗宽. 第 4 节的数值模拟研究表明, RSD-SIS 在数据驱动的窗宽下有比较优秀的表现. 此外, 为了进一步说明 $\sigma = \mathrm{Med}(T_{\boldsymbol{z}})$ 确实是一种不错的窗宽选择方式, 本文也模拟了 RSD-SIS 在某些给定的窗宽参数下的表现, 具体参见附录 B.3.

为了进一步了解本文在 (2.14) 所定义的 RSDCor, 考虑如下 4 种简单情形. (a) 令 Z 服从标准正态分布 $N(0,1)$, $\mathcal{V} = aZ + \epsilon$, 其中 $\epsilon \sim N(0,1)$ 且与 Z 独立, a 表示信噪比强度; (b) 令 Z 和 ϵ 的设置同 (a), $\mathcal{V} = aZ^2 + \epsilon$; (c) 令 $\boldsymbol{z} = (Z_1, Z_2)^{\mathrm{T}}$ 服从二元标准正态分布, $\mathcal{V} = a(Z_1 - Z_2) + \epsilon$, 其中 $\epsilon \sim N(0,1)$ 且与 $\boldsymbol{z}$ 独立; (d) 令 $\boldsymbol{z}$ 和 ϵ 的设置同 (c), $\mathcal{V} = a(Z_1^2 + Z_2^2) + \epsilon$. 在情形 (a)–(d) 中, 分别令 $Y = I(\mathcal{V} \leqslant qt_1) + 2I(qt_1 < \mathcal{V} \leqslant qt_2) + 3I(qt_2 < \mathcal{V} \leqslant qt_3) + 4I(\mathcal{V} > qt_3)$, 其中 qt_1, qt_2 和 qt_3 分别表示 $\mathcal{V}$ 的 25%, 50% 和 75% 的分位数. 令样本量 $n = 200$, $a = 0, 0.5, 1, 2$, 其中 $a = 0$ 意味着 Y 和 Z (或 $\boldsymbol{z}$) 独立. 基于 600 次独立重复实验, 可以得到 $\mathrm{RSDCor}_n(Z, Y)$ 和 $\mathrm{RSDCor}_n(\boldsymbol{z}, Y)$ 在不同信噪比下的箱线图, 如图 1 所示. 该图表明, 本文提出的稳健的半距离相关 RSDCor 可以有效地度量一个连续型随机变量和一个分类变量的线性以及非线性相关强度.

3 基于 RSDCor 的特征筛选程序及其性质

本节利用 (2.14) 的 RSDCor 构建一种与模型无关且适用于超高维分类数据的特征筛选程序, 并进一步研究该筛选程序的渐近性质.

3.1 基于 RSDCor 的特征筛选程序

令 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 为 p 维预测向量, 分类变量 Y 为定义在第 2.1 小节的分类变量. 假设根据先验知识, $\boldsymbol{x}$ 可以分成 G 个组 $\{X_{(1)}, \ldots, X_{(G)}\}$, 其中 $X_{(g)} = (X_{g1}, \ldots, X_{gq_g})^{\mathrm{T}}$, 每个组变量的维数 q_g

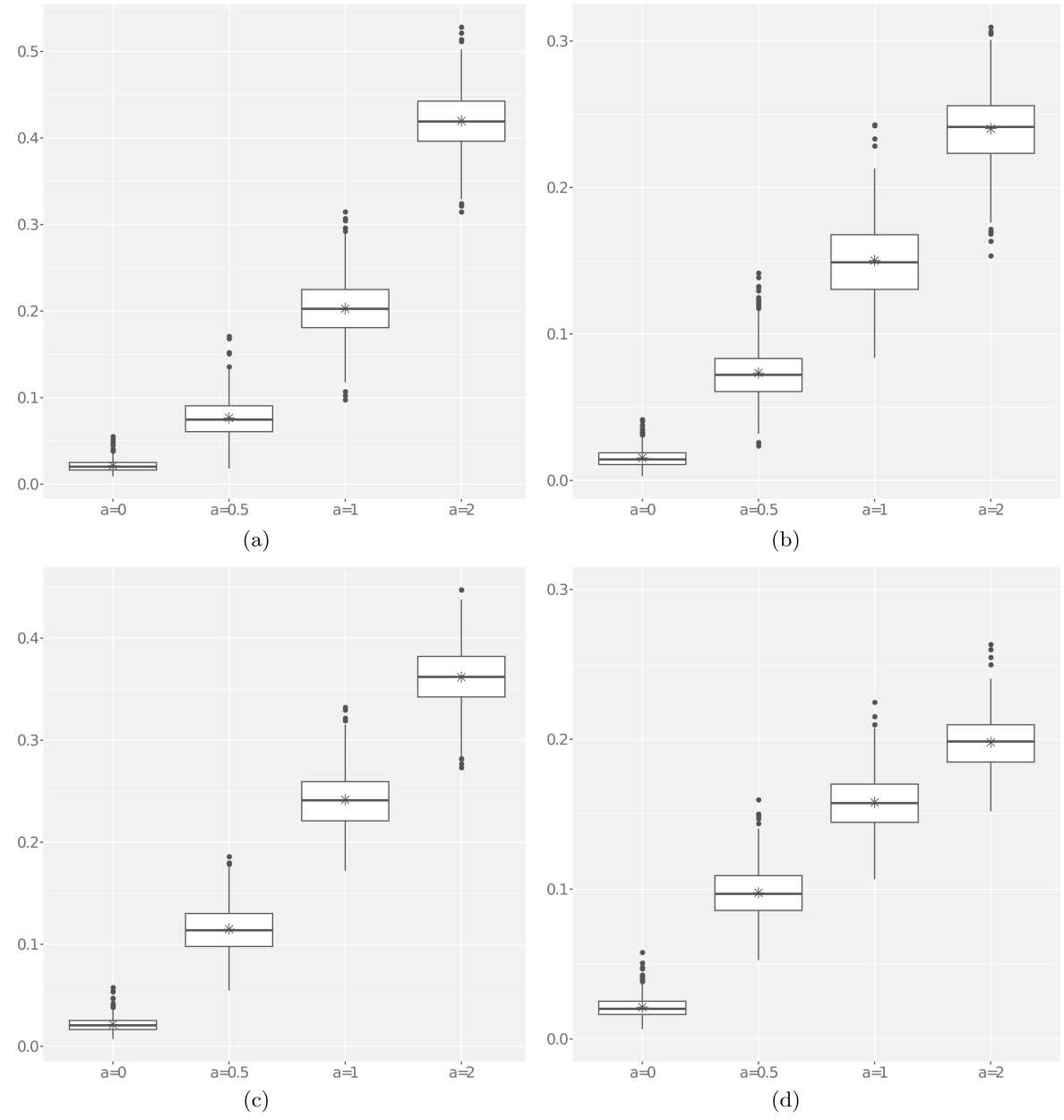


图 1 **(a)–(d) 展示基于 600 次重复模拟, $\mathbf{RSDCor}_n(\boldsymbol{Z},\boldsymbol{Y})$ 和 $\mathbf{RSDCor}_n(z,\boldsymbol{Y})$ 在不同信噪比下的箱线图, 其中纵坐标表示 $\mathbf{RSDCor}_n(\boldsymbol{Z},\boldsymbol{Y})$ 或 $\mathbf{RSDCor}_n(z,\boldsymbol{Y})$ 的大小, 横坐标表示 a 的 4 种取值, 而箱线图中的星号表示平均值. 在 (a) 和 (c) 中, $\boldsymbol{Y}$ 和 $\boldsymbol{Z}$ (或者 z) 的相关性以线性的形式体现. 而在 (b) 和 (d) 中, $\boldsymbol{Y}$ 和 $\boldsymbol{Z}$ (或者 z) 的相关性以非线性的形式体现. 当 $a=0$ 时, 变量 $\boldsymbol{Y}$ 与 $\boldsymbol{Z}$ (或者 z) 独立. 此时, 容易看出 $\mathbf{RSDCor}_n(\boldsymbol{Z},\boldsymbol{Y})$ (或者 $\mathbf{RSDCor}_n(z,\boldsymbol{Y})$) 均与 0 非常接近. 然而, 随着 a 的增大, $\boldsymbol{Y}$ 和 $\boldsymbol{Z}$ (或者 z) 的相关性越来越强, 故 $\mathbf{RSDCor}_n(\boldsymbol{Z},\boldsymbol{Y})$ (或者 $\mathbf{RSDCor}_n(z,\boldsymbol{Y})$) 明显远离 0**

满足 $\sum_{g=1}^{G} q_g = p$, $g = 1,\ldots,G$. 本文假定 q_g 有限, 维数 p 和组数 G 大于或远大于样本量 n, 甚至随着 n 呈指数增长. 显然, 当 $q_1 = \cdots = q_G = 1$ 时, 每一个预测变量单独成 “组”. 不失一般性, 当组变量只含有一个变量时, 亦称其为 “组变量”. 令 $\mathcal{F}(Y = L_r \mid \boldsymbol{x})$ 表示给定 $\boldsymbol{x}$ 和 Y 的条件分布函数, 定义重

要组变量和非重要组变量的指标集合分别为

$$\begin{aligned}\mathcal{A}&=\{g:\text{对于某些 } Y=L_r,\mathcal{F}(Y=L_r\mid\boldsymbol{x})\text{ 与 }X_{(g)}\text{ 有关}\},\\ \mathcal{I}&=\{g:\text{对于任意的 } Y=L_r,\mathcal{F}(Y=L_r\mid\boldsymbol{x})\text{ 与 }X_{(g)}\text{ 无关}\},\end{aligned}\tag{3.1}$$

其中 $\mathcal{A}$ 中元素的个数 $|\mathcal{A}|=s_n$. 若 $g\in\mathcal{A}$, 则 $X_{(g)}$ 为重要组变量; 反之, 若 $g\in\mathcal{I}$, 则 $X_{(g)}$ 为非重要组变量. 令 $T_g=\|X_{1(g)}-X_{2(g)}\|$, $\sigma_g=\mathrm{Med}(T_g)=F_g^{-1}(1/2)$, 其中 $X_{1(g)}$ 和 $X_{2(g)}$ 独立且与 $X_{(g)}$ 同分布, $F_g(\cdot)$ 表示 T_g 的累积分布函数. 为了鉴定出重要的组变量指标集合, 定义如下的边际筛选指标:

$$W_g=\mathrm{RSDCor}(X_{(g)},Y)=\frac{\mathrm{RSDCov}_m(X_{(g)},Y)}{\mathrm{dvar}_m(X_{(g)})\sqrt{R-1}},\tag{3.2}$$

其中 $\mathrm{RSDCov}_m(X_{(g)},Y)$ 和 $\mathrm{dvar}_m(X_{(g)})$ 的表达可参考 (2.14). 根据第 2 节的阐述, 易知 $W_{(g)}=0$ 当且仅当 $X_{(g)}\perp\!\!\!\perp Y$. 若进一步假设部分正交性条件 [11, 14] 成立, 即 $\{X_{(g)}:g\in\mathcal{A}\}\perp\!\!\!\perp\{X_{(g)}:g\in\mathcal{I}\}$, 则当 $g\in\mathcal{A}$ 时, $W_g>0$; 当 $g\in\mathcal{I}$ 时, $W_g=0$. 此时, W_g 自然而然地能够有效地分离出重要组变量和非重要组变量. 这也是本文将 W_g 作为变量筛选指标的重要动机. 此外, 根据后续的条件 (C_4) 也容易看到部分正交性假设在本文并不是必要条件.

令 $(\boldsymbol{x}_1,Y_1),\ldots,(\boldsymbol{x}_n,Y_n)$ 为 n 个独立同分布于 $(\boldsymbol{x},Y)$ 的随机样本, $\hat{\sigma}_g=\mathrm{Med}\{\|X_{i(g)}-X_{j(g)}\|,1\leqslant i<j\leqslant n\}$ 为 σ_g 的样本估计, 定义 W_g 的样本估计量为

$$\widehat{W}_g=\mathrm{RSDCor}_n(X_{(g)},Y)=\frac{\mathrm{RSDCov}_{m,n}(X_{(g)},Y)}{\mathrm{dvar}_{m,n}(X_{(g)})\sqrt{R-1}},\tag{3.3}$$

其中 $\mathrm{RSDCov}_{m,n}(X_{(g)},Y)$ 和 $\mathrm{dvar}_{m,n}(X_{(g)})$ 的表达可参考 (2.15).

令 $\widehat{W}=(\widehat{W}_1,\ldots,\widehat{W}_G)$ 表示通过边际 RSDCor 得到的 G 维向量, 其中 $\widehat{W}_g$ 越大意味着 $X_{(g)}$ 与 Y 的相关性可能越强. 根据 $\{\widehat{W}_g\}_{g=1}^G$ 的大小, 定义如下的组变量指标集合:

$$\hat{\mathcal{A}}=\{1\leqslant g\leqslant G:\widehat{W}_g\geqslant cn^{-\nu}\},\tag{3.4}$$

其中 c 和 ν 为预定阈值, 具体可参见条件 (C_3). 由于在实际应用中 c 和 ν 往往未知, 故常采用硬阈值的方式来选择组变量指标集合, 即令

$$\hat{\mathcal{A}}_{d_n}=\{g:\widehat{W}_g\text{ 是 }\{\widehat{W}_1,\ldots,\widehat{W}_G\}\text{ 中前 }d_n\text{ 个最大的}\},\tag{3.5}$$

其中 d_n 表示截断值. 根据文献 [9], 令 $d_n=[n/\log n]$, 其中 $[\cdot]$ 表示取整. 为了方便, 将上述基于 (3.2)–(3.5) 的特征筛选方法简称为 RSD-SIS.

3.2 RSD-SIS 的渐近性质

本小节主要研究筛选程序 RSD-SIS 的确定筛选性质和秩相合性, 相关的假设条件如下.

(C_1) 对于任意 $g=1,\ldots,G$, 存在正数 $\tilde{m}$ 使得变量 T_g 的中位数 $\sigma_g\geqslant\tilde{m}$. 令 $f_g(\sigma_g)$ 表示变量 T_g 的分布函数 $F_g(\cdot)$ 在 σ_g 处的导数, 存在正数 $\breve{m}$ 使得 $f_g(\sigma_g)\geqslant\breve{m}$.

(C_2) 存在两个正数 a_1 和 a_2 满足 $a_1/R_n\leqslant\min_{1\leqslant r\leqslant R_n}p_r\leqslant\max_{1\leqslant r\leqslant R_n}p_r\leqslant a_2/R_n$, 其中 R_n 与前文中的 R 同义, 下标 n 仅表示类别个数 R 允许随着样本量 n 发散, 即 $R_n=O(n^\kappa)$, $\kappa\geqslant0$.

(C_3) 存在常数 $c>0$ 和 $0\leqslant\nu<1/2$, 使得 $\min_{g\in\mathcal{A}}W_g\geqslant2cn^{-\nu}$.

(C_4) 存在常数 $\eta>0$, 使得 $\min_{g\in\mathcal{A}}W_g-\max_{g\in\mathcal{I}}W_g\geqslant\eta$.

注 3.1 条件 (C_1) 是一个比较温和的假设, 当 $X_{(g)}$ 服从 Gauss 分布或 Cauchy 分布时, 该假设均成立. 此外, 附录 A 的证明部分也表明, 我们允许 $\tilde{m}$ 和 $\check{m}$ 随着 $n \to \infty$ 以某种速度趋于 0. 条件 (C_2) 要求响应变量 Y 的每一个类别所占的比例不能太小或者太大, 且允许总类别数 R_n 随着 n 以一定的速度发散. 条件 (C_3) 给出了重要组变量的最小信号强度, 这一假设在与特征筛选有关的文献中经常使用, 如文献 [6,27,35] 等. 条件 (C_4) 也是一个常见假设, 其表明重要组变量和不重要组变量的信号强度之间有一个正向差距. 当部分正交性条件成立时, 该条件自然成立. 因此, 条件 (C_4) 是一个比部分正交性条件更弱的假设. 实际上, 本文也允许 η 随着 $n \to \infty$ 而趋于 0, 详见定理 3.2.

定理 3.1 (确定筛选性质) 假设条件 (C_1) 和 (C_2) 成立, 则对于任意的 $\nu \in [0, 1/2)$, $0 \leqslant \kappa < 1/4 - \nu/2$, 存在常数 $\xi > 0$ 使得

$$P\Big(\max_{1 \leqslant g \leqslant G} |\widehat{W}_g - W_g| \geqslant cn^{-\nu}\Big) \leqslant O(Gn^{\kappa} \exp\{-\xi n^{1-2\nu-4\kappa}\}).$$

若进一步有条件 (C_3) 成立, 则

$$P(\mathcal{A} \subset \hat{\mathcal{A}}) \geqslant 1 - O(s_n n^{\kappa} \exp\{-\xi n^{1-2\nu-4\kappa}\}),$$

其中 $s_n = |\mathcal{A}|$, 即集合 $\mathcal{A}$ 中元素的个数.

定理 3.1 表明, 随着样本量 n 趋于无穷, 本文选出的组变量指标集合 $\hat{\mathcal{A}}$ 包含真实组变量指标集合 $\mathcal{A}$ 的概率将趋于 1. 此外, RSD-SIS 能够处理组变量个数 G (或者维数 p) 和样本量 n 满足 $\log G = o(n^{1-2\nu-4\kappa})$ 的超高维特征筛选问题. 若假设类别个数 R_n 固定, 即 $\kappa = 0$, 则有 $\log G = o(n^{1-2\nu})$ 且 $0 \leqslant \nu < 1/2$. 同时, 注意到文献 [35] 提出的 SD-SIS 要求变量 $\boldsymbol{x}$ 满足次指数假设, 故其在 $\kappa = 0$ 的情形下处理的变量维数满足 $\log G = o(n^{(1-2\nu)/3})$, $0 \leqslant \nu < 1/2$.

定理 3.2 (秩相合性) 若条件 (C_1), (C_2) 和 (C_4) 成立, 则存在常数 $\zeta > 0$ 使得

$$P\Big(\min_{g \in \mathcal{A}} \widehat{W}_g > \max_{g \in \mathcal{I}} \widehat{W}_g\Big) \geqslant 1 - O\bigg(GR \exp\Big\{-\frac{n\zeta\eta^2}{R^4}\Big\}\bigg).$$

定理 3.2 表明, 若 $GR\exp\{-n\zeta\eta^2/R^4\} = o(1)$, 则重要组变量总是排在非重要组变量前面的概率随着样本量 n 的增大趋于 1, 这保证了筛选结果的有效性. 此外, 条件 (C_4) 也可以进一步放松. 具体而言, 令 $\eta = O(n^{-\gamma})$, $0 < \gamma < 1/2$ 且 $\log G = o(n^{1-2\gamma-4\kappa})$, 则秩相合性依然成立.

4 数值模拟与实际数据分析

本文通过 5 个模拟实验 (即例 4.1–4.3 以及附录 B 的例 B.1 和 B.2) 和一个实例来考察本文提出的特征筛选程序 RSD-SIS 在有限样本下的表现, 其中例 4.1 和 4.2 关心的是单个预测变量与分类响应变量的特征筛选问题, 即此时的组变量个数 $G = p$. 而例 4.3 对应的是组预测变量与分类响应变量的组特征筛选. 类似于文献 [6], 令 MMS (minimum model size) 表示在包含所有重要组变量的模型中, 最小的模型大小, 本节定义以下 4 个考察指标: (1) $\mathcal{M}$, 即 MMS 的中位数; (2) $\mathcal{S}$=(MMS 的 75% 分位数 − MMS 的 25% 分位数)/1.349, 即 MMS 的稳健的标准差; (3) $\mathcal{P}_j^s$, 即给定 (3.5) 中的截断值 d_n, 第 j 个重要组变量被选到的概率; (4) $\mathcal{P}_a$, 即给定截断值 d_n, 所有重要组变量均被选到的概率. 根据文献 [8] 中的建议, 令 $d_n = [n/\log n]$. 由 (1)–(4) 可知, 一种优秀的筛选方法应使得 $\mathcal{M}$ 的值尽量接近重要的组变量的个数, $\mathcal{S}$ 尽量接近 0, 而 $\mathcal{P}_j^s$ 和 $\mathcal{P}_a$ 的值均尽量接近 1.

为了更好地评价 RSD-SIS, 本节将其与现有的一些特征筛选方法如 CI-SIS [4], GN-SIS [27], FKF-SIS [21], MV-SIS [6], PSIS [24], SD-SIS [35] 进行对比, 其中 CI-SIS, FKF-SIS, MV-SIS, PSIS 更适用于单个预测变量的筛选. 而 GN-SIS, SD-SIS 与 RSD-SIS 一样, 同时适用于单变量和组变量筛选. 此外, 为了使得 DC-SIS [18] 能够适用于本文的筛选问题, 本文采取与文献 [6, 21] 相同的做法将分类响应变量转换为多个虚拟向量. 具体而言, 假设 Y 有 R 个类别, 当 $Y = L_r$ 时, 构造对应的 R 维虚拟向量 $\boldsymbol{y}^{dv} = (0, \ldots, \underbrace{1}_{r}, \ldots, 0)^{\mathrm{T}}$, $r = 1, \ldots, R$, 即 $\boldsymbol{y}^{dv}$ 的第 r 个元素为 1, 其他为 0. 上述提及的 8 种特征筛选程序均通过 R4.3.2 软件来执行完成.

4.1 数值模拟学习

例 4.1 该例来自文献 [20] 的模型 (3). 令 Y 为二元分类变量, $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \sim N(0, \Sigma)$, 其中 $\Sigma = (\sigma_{ij})_{p\times p}$ 且 $\sigma_{ij} = 0.8^{|i-j|}$. 此外, Y 来自如下的逻辑回归模型:

$$\log\left\{\frac{P(Y=1 \mid \boldsymbol{x})}{P(Y=-1 \mid \boldsymbol{x})}\right\} = 2X_1 + 2X_2 + 2X_3 + 3\sin(X_4) + 4X_5^2 - 3.$$

在例 4.1 中, 每个组变量的个数均为 1, 即组类别个数 $G = p$. 令样本量 $n = 50$ 或者 $n = 80$, $p = 2000$, 根据 MMS, $\mathcal{S}$, $\mathcal{P}_j^s$ 和 $\mathcal{P}_a$ 来评价 CI-SIS, DC-SIS, GN-SIS, FKF-SIS, MV-SIS, PSIS, SD-SIS, RSD-SIS 这 8 种筛选方法的表现, 其中基于 600 次重复实验的模拟结果如表 1 所示. 表 1 表明, 所有筛选程序的表现均随着样本量 n 的增加而提高. 此外, RSD-SIS, SD-SIS, GN-SIS 和 DC-SIS 的表现基本差不多, 且优于 MV-SIS, FKF-SIS, CI-SIS 和 PSIS.

例 4.2 该例主要来自文献 [6, 例 1]. 令分类变量 Y 的分布来自如下两种不同的设计: (1) 平衡设计, $P(Y = r) = 1/R$, $r = 1, \ldots, R$; (2) 非平衡设计, $p_r = P(Y = r) = 2[1 + (r-1)/(R-1)]/3R$, 其

表 1 在例 4.1 中, $\mathcal{M}$, $\mathcal{S}$, $\mathcal{P}_a$, $\mathcal{P}_1^s(\boldsymbol{X_1})$, $\mathcal{P}_2^s(\boldsymbol{X_2})$, $\mathcal{P}_3^s(\boldsymbol{X_3})$, $\mathcal{P}_4^s(\boldsymbol{X_4})$ 和 $\mathcal{P}_5^s(\boldsymbol{X_5})$ 的大小

n, p	方法	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{P}_3^s(X_3)$	$\mathcal{P}_4^s(X_4)$	$\mathcal{P}_5^s(X_5)$
$n = 50$ $p = 2000$	CI-SIS	19.00	55.04	0.420	0.950	0.988	0.978	0.888	0.433
	DC-SIS	9.00	14.83	0.602	0.930	0.992	0.988	0.923	0.652
	GN-SIS	9.00	17.05	0.585	0.948	0.980	0.987	0.937	0.638
	FKF-SIS	14.00	25.57	0.477	0.828	0.962	0.958	0.900	0.618
	MV-SIS	12.00	21.50	0.502	0.902	0.985	0.977	0.920	0.570
	PSIS	36.50	111.56	0.312	0.890	0.938	0.942	0.777	0.333
	SD-SIS	9.00	17.79	0.606	0.942	0.987	0.985	0.932	0.652
	RSD-SIS	10.00	14.83	0.573	0.833	0.957	0.970	0.925	0.747
$n = 80$ $p = 2000$	CI-SIS	6.00	9.64	0.753	0.998	1.000	1.000	0.997	0.753
	DC-SIS	5.00	0.74	0.937	0.998	1.000	1.000	0.998	0.938
	GN-SIS	5.00	0.74	0.932	0.998	1.000	1.000	0.998	0.935
	FKF-SIS	5.00	1.48	0.905	0.993	1.000	1.000	0.995	0.913
	MV-SIS	5.00	1.48	0.920	0.997	1.000	1.000	1.000	0.923
	PSIS	10.00	23.72	0.650	0.995	1.000	1.000	0.967	0.652
	SD-SIS	5.00	0.74	0.942	1.000	1.000	1.000	1.000	0.942
	RSD-SIS	5.00	0.74	0.970	0.995	0.998	1.000	0.992	0.978

满足 $\max_{1\leqslant r\leqslant R} p_r = 2\min_{1\leqslant r\leqslant R} p_r$. 当 Y 为二元分类变量 (即 $R=2$) 时, $p_1=1/3$, $p_2=2/3$. 给定 $Y=r$, 令 $\boldsymbol{x}=(X_1,\ldots,X_p)^{\mathrm{T}}=\boldsymbol{u}+\boldsymbol{\epsilon}$, 其中 $\boldsymbol{u}=(U_1,\ldots,U_r,\ldots,U_p)^{\mathrm{T}}$ 且满足第 r 个元素 $U_r=3$, 其他元素为 0, 而 $\boldsymbol{\epsilon}=(\epsilon_1,\ldots,\epsilon_p)^{\mathrm{T}}$ 是一个与 $\boldsymbol{x}$ 独立的 p 维误差项, $\{\epsilon_j\}_{j=1}^p$ 相互独立且与 ϵ 同分布. 这里考虑 ϵ 的 3 种不同分布: (a) $\epsilon\sim N(0,1)$; (b) $\epsilon\sim t(2)$; (c) $\epsilon\sim C(0,1)$, 其中 $t(2)$ 表示自由度为 2 的 t 分布, $C(0,1)$ 表示标准的 Cauchy 分布. 在情形 (b) 和 (c) 下, 易知 $X_1,\ldots,X_p$ 均服从厚尾分布且其均不满足二阶矩有限. 因此, 通过情形 (b) 和 (c) 可以检验各筛选方法的稳健性.

令 $(R,n,p)=(2,40,2000)$, $(R,n,p)=(10,200,2000)$, 本文检验了 RSD-SIS 以及其他 7 种对比方法在不同情形下的表现, 其中基于 600 次重复实验的模拟结果如表 2–6 所示. 从这些表格中可以清楚地看到, 当 $\epsilon\sim N(0,1)$, 即预测变量的有限矩假设成立时, 所有筛选方法均表现优秀. 但是, 当 $\epsilon\sim C(0,1)$ 时, PSIS, DC-SIS, SD-SIS, GN-SIS 的表现大打折扣. 然而, CI-SIS, FKF-SIS, MV-SIS, RSD-SIS 的表现却相对较好, 尤其是 RSD-SIS. 造成该现象的主要原因在于 DC-SIS, SD-SIS, GN-SIS 要求预测变量满足的次指数假设不成立, 而 CI-SIS, FKF-SIS, MV-SIS, RSD-SIS 与预测变量的有限矩条件无关. 同时容易看到, 当 $\epsilon\sim t(2)$ 和 $\epsilon\sim C(0,1)$ 时, PSIS 的表现均不是很理想. 这很大程度上归因于 PSIS 的

表 2　在例 4.2 中, 对于 $(R,n,p)=(2,40,2000)$, $\mathcal{M}$, $\mathcal{S}$, $\mathcal{P}_a$, $\mathcal{P}_1^s(X_1)$ 和 $\mathcal{P}_2^s(X_2)$ 的大小

ϵ	方法	平衡设计					非平衡设计				
		$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$
$N(0,1)$	CI-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	1.000	1.000	1.000
	DC-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	1.000	1.000	1.000
	GN-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	1.000	1.000	1.000
	FKF-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	0.998	0.998	1.000
	MV-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	1.000	1.000	1.000
	PSIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	1.000	1.000	1.000
	SD-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	1.000	1.000	1.000
	RSD-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	1.000	1.000	1.000
$t(2)$	CI-SIS	2.00	0.0	0.973	0.992	0.982	2.00	1.48	0.910	0.952	0.952
	DC-SIS	2.00	0.0	0.963	0.988	0.975	2.00	0.74	0.935	0.960	0.972
	GN-SIS	2.00	0.0	0.980	0.985	0.995	2.00	0.0	0.950	0.973	0.975
	FKF-SIS	2.00	0.0	0.985	0.992	0.993	2.00	0.74	0.967	0.982	0.983
	MV-SIS	2.00	0.0	0.985	0.995	0.990	2.00	0.74	0.930	0.967	0.962
	PSIS	43.00	27.43	0.002	0.045	0.032	48.00	35.03	0.002	0.015	0.023
	SD-SIS	2.00	0.0	0.980	0.993	0.987	2.00	0.0	0.932	0.960	0.970
	RSD-SIS	2.00	0.0	0.995	0.998	0.997	2.00	0.0	0.980	0.983	0.992
$C(0,1)$	CI-SIS	7.00	20.20	0.572	0.760	0.747	15.00	40.03	0.453	0.662	0.685
	DC-SIS	31.50	152.15	0.390	0.618	0.640	46.00	207.19	0.332	0.557	0.558
	GN-SIS	15.00	69.13	0.458	0.690	0.675	21.00	73.57	0.402	0.627	0.625
	FKF-SIS	2.50	3.34	0.837	0.915	0.912	3.50	8.15	0.722	0.835	0.850
	MV-SIS	3.00	5.19	0.780	0.870	0.902	5.50	14.27	0.617	0.777	0.800
	PSIS	847.00	435.32	0.0	0.005	0.002	931.50	494.63	0.0	0.008	0.005
	SD-SIS	30.50	168.83	0.357	0.607	0.622	66.50	242.96	0.275	0.552	0.535
	RSD-SIS	2.00	1.48	0.882	0.945	0.935	2.00	2.22	0.855	0.928	0.918

表 3 在例 4.2 的平衡设计中, 对于 $(R,n,p)=(10,200,2000)$, $\mathcal{P}_a$ 和 $\{\mathcal{P}_j^s(X_j)\}_{j=1}^{10}$ 的大小

ϵ	方法	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{P}_3^s(X_3)$	$\mathcal{P}_4^s(X_4)$	$\mathcal{P}_5^s(X_5)$	$\mathcal{P}_6^s(X_6)$	$\mathcal{P}_7^s(X_7)$	$\mathcal{P}_8^s(X_8)$	$\mathcal{P}_9^s(X_9)$	$\mathcal{P}_{10}^s(X_{10})$
$N(0,1)$	CI-SIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	DC-SIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GN-SIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	FKF-SIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	MV-SIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	PSIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SD-SIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	RSD-SIS	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000
$t(2)$	CI-SIS	0.990	0.998	1.000	1.000	0.995	1.000	1.000	0.998	1.000	1.000	0.998
	DC-SIS	0.830	0.980	0.985	0.978	0.985	0.992	0.973	0.982	0.983	0.985	0.973
	GN-SIS	0.953	0.998	0.995	0.990	0.995	0.995	0.993	0.995	0.998	0.997	0.997
	FKF-SIS	0.957	0.993	0.998	0.993	0.998	0.995	0.995	0.995	0.993	0.997	0.998
	MV-SIS	0.938	0.987	0.997	0.995	0.995	0.988	0.992	0.990	0.998	0.997	0.998
	PSIS	0.0	0.037	0.047	0.028	0.023	0.040	0.038	0.042	0.038	0.052	0.043
	SD-SIS	0.948	1.000	0.997	0.993	0.993	0.998	0.995	0.995	0.987	0.993	0.997
	RSD-SIS	0.952	0.993	0.992	0.995	0.998	1.000	0.988	0.993	1.000	0.998	0.993
$C(0,1)$	CI-SIS	0.465	0.927	0.925	0.923	0.918	0.923	0.918	0.917	0.917	0.940	0.937
	DC-SIS	0.0	0.528	0.537	0.532	0.477	0.557	0.527	0.532	0.543	0.525	0.522
	GN-SIS	0.002	0.585	0.607	0.593	0.572	0.600	0.550	0.582	0.583	0.567	0.637
	FKF-SIS	0.595	0.943	0.955	0.948	0.942	0.958	0.930	0.950	0.955	0.940	0.953
	MV-SIS	0.405	0.908	0.918	0.920	0.910	0.935	0.895	0.912	0.925	0.902	0.923
	PSIS	0.0	0.030	0.015	0.010	0.020	0.013	0.012	0.0167	0.022	0.017	0.018
	SD-SIS	0.003	0.505	0.557	0.483	0.517	0.515	0.488	0.527	0.472	0.502	0.501
	RSD-SIS	0.665	0.960	0.970	0.970	0.960	0.963	0.960	0.955	0.947	0.968	0.953

表 4 在例 4.2 的平衡设计中, 对于 $(R,n,p)=(10,200,2000)$, $\mathcal{M}$ 和 $\mathcal{S}$ 的大小

ϵ	$\mathcal{M}/\mathcal{S}$	CI-SIS	DC-SIS	GN-SIS	FKF-SIS	MV-SIS	PSIS	SD-SIS	RSD-SIS
$N(0,1)$	$\mathcal{M}$	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
	$\mathcal{S}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$t(2)$	$\mathcal{M}$	10.00	12.00	10.00	11.00	11.00	300.00	10.00	10.00
	$\mathcal{S}$	0.19	7.41	0.74	2.97	2.97	129.54	0.74	1.48
$C(0,1)$	$\mathcal{M}$	43.00	1735.00	469.00	28.00	56.00	1780.00	1728.00	21.00
	$\mathcal{S}$	95.07	406.22	305.41	30.39	84.69	149.00	502.96	31.13

筛选指标的构建是基于不同类别下的均值差异, 故在预测变量服从 $t(2)$ 型和 $C(0,1)$ 型分布的情形下, 样本均值不是总体均值的相合估计. 为了更好地说明这一点, 令 ϵ 服从二阶矩有限的轻度厚尾分布, 如 $t(3)$, $t(4)$ 和 $t(5)$, 并进一步分析 PSIS 的表现, 如附录 B 的表 B4–B6 所示. 模拟结果表明, 当预测变量服从 $t(3)$, $t(4)$ 或 $t(5)$ 型分布时, PSIS 表现较好.

例 4.3 该例主要来自文献 [35, 例 5.3]. 令分类变量 $Y\in\{1,\ldots,R\}$, 考虑如下两种设计: (1) 平衡

表 5 在例 4.2 的非平衡设计中, 对于 $(R, n, p) = (10, 200, 2000)$, $\mathcal{P}_a$ 和 $\{\mathcal{P}_j^s(\boldsymbol{X}_j)\}_{j=1}^{10}$ 的大小

ϵ	方法	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{P}_3^s(X_3)$	$\mathcal{P}_4^s(X_4)$	$\mathcal{P}_5^s(X_5)$	$\mathcal{P}_6^s(X_6)$	$\mathcal{P}_7^s(X_7)$	$\mathcal{P}_8^s(X_8)$	$\mathcal{P}_9^s(X_9)$	$\mathcal{P}_{10}^s(X_{10})$
$N(0,1)$	CI-SIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	DC-SIS	0.945	0.967	0.985	0.993	0.998	1.000	1.000	1.000	1.000	1.000	1.000
	GN-SIS	0.998	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	FKF-SIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	MV-SIS	0.985	0.990	0.997	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	PSIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SD-SIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	RSD-SIS	0.982	0.990	0.995	0.998	0.998	1.000	1.000	1.000	1.000	1.000	1.000
$t(2)$	CI-SIS	0.960	0.985	0.986	0.997	0.998	0.995	0.998	1.000	0.998	1.000	0.998
	DC-SIS	0.578	0.797	0.878	0.935	0.962	0.970	0.983	0.987	0.997	0.998	1.000
	GN-SIS	0.943	0.978	0.980	0.998	0.993	0.998	0.995	1.000	1.000	1.000	1.000
	FKF-SIS	0.882	0.985	0.975	0.967	0.983	0.978	0.992	0.998	0.993	0.993	0.997
	MV-SIS	0.830	0.925	0.960	0.965	0.987	0.995	0.995	0.997	1.000	1.000	1.000
	PSIS	0.0	0.027	0.028	0.042	0.030	0.040	0.025	0.022	0.040	0.033	0.040
	SD-SIS	0.897	0.957	0.980	0.982	0.988	0.995	0.997	0.998	1.000	1.000	0.998
	RSD-SIS	0.827	0.923	0.960	0.970	0.982	0.993	0.993	0.998	1.000	1.000	0.998
$C(0,1)$	CI-SIS	0.322	0.863	0.855	0.857	0.858	0.877	0.883	0.903	0.903	0.912	0.910
	DC-SIS	0.0	0.250	0.320	0.325	0.425	0.522	0.553	0.615	0.623	0.643	0.720
	GN-SIS	0.0	0.407	0.418	0.512	0.527	0.570	0.585	0.657	0.665	0.655	0.687
	FKF-SIS	0.380	0.892	0.885	0.907	0.865	0.877	0.900	0.900	0.905	0.913	0.922
	MV-SIS	0.222	0.690	0.768	0.810	0.878	0.898	0.923	0.950	0.958	0.970	0.972
	PSIS	0.0	0.020	0.017	0.017	0.027	0.015	0.017	0.018	0.018	0.022	0.020
	SD-SIS	0.002	0.353	0.372	0.413	0.427	0.493	0.532	0.528	0.537	0.570	0.593
	RSD-SIS	0.575	0.842	0.893	0.915	0.940	0.945	0.982	0.987	0.978	0.990	0.993

表 6 在例 4.2 的非平衡设计中, 对于 $(R, n, p) = (10, 200, 2000)$, $\mathcal{M}$ 和 $\mathcal{S}$ 的大小

ϵ	$\mathcal{M}/\mathcal{S}$	CI-SIS	DC-SIS	GN-SIS	FKF-SIS	MV-SIS	PSIS	SD-SIS	RSD-SIS
$N(0,1)$	$\mathcal{M}$	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
	$\mathcal{S}$	0.0	0.74	0.0	0.0	0.0	0.0	0.0	0.0
$t(2)$	$\mathcal{M}$	10.00	27.00	10.00	13.25	14.00	320.00	10.00	11.00
	$\mathcal{S}$	1.48	65.42	1.48	8.15	11.86	142.33	2.97	8.15
$C(0,1)$	$\mathcal{M}$	75.50	1722.00	526.50	52.50	87.00	1781.50	1670.00	29.00
	$\mathcal{S}$	141.59	490.73	333.77	76.07	142.51	157.89	549.67	56.52

设计, $p_r = P(Y = r) = 1/R$; (2) 非平衡设计, $p_r = P(Y = r) = 2[1 + (r-1)/(R-1)]/3R$. 令 p 维预测向量 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}} = (X_{(1)}^{\mathrm{T}}, \ldots, X_{(G)}^{\mathrm{T}})^{\mathrm{T}}$, 其中每一个组预测变量 $X_{(g)}$ 的维数 $q_g = 4$ 且 $G = p/4$, $g = 1, \ldots, G$.

为了生成 $\boldsymbol{x}$, 先按照如下方式生成 $\boldsymbol{z} = (Z_1, \ldots, Z_G)^{\mathrm{T}} \in R^G$. 令 $\{\boldsymbol{z} \mid Y = r\} \sim N(\boldsymbol{\mu}_r, \boldsymbol{I})$, 其中 $\boldsymbol{\mu}_r = (\mu_{r1}, \ldots, \mu_{rr}, \ldots, \mu_{rG})^{\mathrm{T}}$ 的第 r 个元素 $\mu_{rr} = 2$, 其他元素为 0. 进一步地, 令 $k = 1, 2, 3, 4$, 并

基于以下 3 种情形来生成 $\boldsymbol{x}$: (a) $\{X_{4(g-1)+k} \mid Z_g = z_g\} \sim N(z_g, 1)$; (b) $\{X_{4(g-1)+k} \mid Z_g = z_g\} \sim C(z_g, 1)$, 其中 $C(z_g, 1)$ 表示位置参数为 z_g、尺度参数为 1 的 Cauchy 分布; (c) $\{X_{4(g-1)+k} \mid Z_g = z_g\} \sim 0.8N(z_g, 1) + 0.2C(z_g, 1)$, 即 $\{X_{4(g-1)+k} \mid Z_g = z_g\}$ 以 0.8 的概率来自分布 $N(z_g, 1)$, 以 0.2 的概率来自分布 $C(z_g, 1)$. 在例 4.3 中不难看出, 重要的组预测变量指标为 $\{1, \ldots, R\}$, 且其对应的重要预测变量指标为 $\{1, 2, 3, 4, \ldots, (4R-3), (4R-2), (4R-1), (4R)\}$. 具体地, 当 $R = 3$ 时, 令 $(n, p) = (75, 2000)$ 或者 $(n, p) = (120, 2000)$; 当 $R = 10$ 时, 令 $(n, p) = (250, 3000)$ 或者 $(n, p) = (300, 3000)$. 基于 600 次重复实验, 本文检验了 RSD-SIS 在不同情形下的表现. 与例 4.1 和 4.2 不同, 在该例中只有 4 种筛选程序 DC-SIS, GN-SIS, SD-SIS, RSD-SIS 能更好地处理组筛选问题, 模拟结果如表 7 和 8 中的情形 (a), (b) 和 (c) 所示. 表 7 和 8 表明, 在预测向量的有限矩假设成立的情形下, RSD-SIS 的表现与 GN-SIS 和 SD-SIS 差不多, 且稍优于 DC-SIS. 而当预测向量的有限矩假设不成立时, RSD-SIS 的表现明显优于 GN-SIS, SD-SIS 和 DC-SIS.

基于对例 4.1–4.3 的分析, 易知 RSD-SIS 具有如下优点: (1) 与 CI-SIS, FKF-SIS, MV-SIS 和 PSIS 相比, RSD-SIS 既可用于单变量特征筛选, 又可用于组变量特征筛选; (2) 当预测变量/向量的有限距假设成立时, RSD-SIS 与其他 7 种对比方法相比具有非常强大的竞争力; (3) 当预测变量/向量不满足有限矩条件时, 如预测变量服从 $C(0, 1)$ 时, RSD-SIS 与其对比方法 SD-SIS, GN-SIS, DC-SIS 相比具有明显的优势, 这一点在组筛选情形中尤为突出.

4.2 实际数据分析

本小节通过一个实例来说明本文提出的 RSD-SIS 在实际中的应用, 并将其表现与其他 7 种筛选方法 CI-SIS, DC-SIS, GN-SIS, FKF-SIS, MV-SIS, PSIS, SD-SIS 进行对比.

该数据集来自文献 [1], 其目的是通过 mRNA (messenger ribonucleic acid) 的表达谱对人类肺癌数据进行归类. 在 203 个快速冷冻的肿瘤肺和正常肺中, 共有 12600 个 mRNA 表达水平, 即在该数据集中, 样本量 $n = 203$, 变量个数 $p = 12600$. 该 203 例样本可以分为 5 个类别: 139 例肺腺癌 (ADEN)、21 例鳞状细胞肺癌 (SQUA)、6 例小细胞肺癌 (SCLC)、20 例肺类癌 (COID) 和其余 17 例正常肺样本 (NORMAL). 在分类之前, 先将这 12600 个 mRNA 表达水平标准化, 使其分别满足均值为 0, 方差为 1.

按照文献 [27] 的做法, 本文从每个类别中随机选择大约 $100\tau\%$ 的观测值作为训练样本, 其余 $100(1-\tau)\%$ 的观测结果作为测试样本, 其中 τ 设为 0.8 或者 0.9. 将每种筛选方法与 LDA 相结合并应

表 7 在例 4.3 中, $\mathcal{M}$, $\mathcal{S}$, $\mathcal{P}_a$ 和 $\{\mathcal{P}_j^s(\boldsymbol{X}_{(j)})\}_{j=1}^3$ 的大小, 其中 $\mathcal{P}_1^s, \mathcal{P}_2^s$ 和 $\mathcal{P}_3^s$ 分别表示 $\mathcal{P}_1^s(\boldsymbol{X}_{(1)}), \mathcal{P}_2^s(\boldsymbol{X}_{(2)})$ 和 $\mathcal{P}_3^s(\boldsymbol{X}_{(3)})$

情形 R, n, p	方法	平衡设计						非平衡设计					
		$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$
(a)	DC-SIS	3.00	0.0	1.000	1.000	1.000	1.000	3.00	0.0	0.985	0.985	1.000	1.000
$R = 3$	GN-SIS	3.00	0.0	1.000	1.000	1.000	1.000	3.00	0.0	1.000	1.000	1.000	1.000
$n = 75$	SD-SIS	3.00	0.0	1.000	1.000	1.000	1.000	3.00	0.0	0.997	0.997	1.000	1.000
$p = 2000$	RSD-SIS	3.00	0.0	0.998	1.000	1.000	0.998	3.00	0.0	0.995	0.995	1.000	1.000
(b)	DC-SIS	222.50	233.14	0.125	0.485	0.537	0.473	258.50	225.91	0.090	0.363	0.488	0.593
$R = 3$	GN-SIS	50.00	76.54	0.298	0.675	0.707	0.688	60.00	85.43	0.277	0.583	0.677	0.710
$n = 120$	SD-SIS	246.00	232.77	0.130	0.492	0.518	0.503	260.00	228.32	0.098	0.372	0.468	0.528
$p = 2000$	RSD-SIS	3.00	2.97	0.948	0.987	0.983	0.978	4.00	3.71	0.923	0.933	0.995	0.995

表 8　在例 4.3 中, $\mathcal{M}$, $\mathcal{S}$, $\mathcal{P}_a$ 和 $\{\mathcal{P}_j^s\}_{j=1}^{10}$ 的大小, 其中 $\{\mathcal{P}_j^s\}_{j=1}^{10}$ 分别表示 $\{\mathcal{P}_j^s(\boldsymbol{X}_{(j)})\}_{j=1}^{10}$

情形	方法	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$	$\mathcal{P}_4^s$	$\mathcal{P}_5^s$	$\mathcal{P}_6^s$	$\mathcal{P}_7^s$	$\mathcal{P}_8^s$	$\mathcal{P}_9^s$	$\mathcal{P}_{10}^s$
							平衡设计							
	DC-SIS	10.00	0.74	0.973	0.997	0.995	0.998	1.000	0.998	0.997	0.995	1.000	0.998	0.995
	GN-SIS	10.00	0.0	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000
	SD-SIS	10.00	0.0	0.998	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(a)	RSD-SIS	10.00	0.74	0.983	1.000	0.998	0.998	0.998	0.995	0.998	0.998	1.000	0.998	0.998
$R=10$							非平衡设计							
$n=250$	DC-SIS	12.00	11.12	0.867	0.932	0.975	0.975	0.992	0.997	0.997	1.000	1.000	1.000	1.000
$p=3000$	GN-SIS	10.00	0.0	0.995	0.998	0.998	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	SD-SIS	10.00	0.0	0.993	0.995	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	RSD-SIS	11.00	2.41	0.945	0.967	0.990	0.997	0.992	1.000	1.000	0.998	1.000	1.000	1.000
							平衡设计							
	DC-SIS	648.50	154.37	0.053	0.765	0.758	0.735	0.740	0.752	0.737	0.727	0.730	0.757	0.717
	GN-SIS	20.00	44.11	0.713	0.965	0.968	0.968	0.962	0.975	0.967	0.960	0.957	0.963	0.962
	SD-SIS	645.00	172.35	0.052	0.743	0.753	0.727	0.720	0.695	0.765	0.718	0.762	0.765	0.780
(c)	RSD-SIS	10.00	0.0	0.992	1.000	0.998	0.998	1.000	1.000	0.998	0.998	1.000	1.000	0.998
$R=10$							非平衡设计							
$n=300$	DC-SIS	642.00	147.33	0.040	0.548	0.610	0.688	0.717	0.723	0.748	0.798	0.820	0.793	0.815
$p=3000$	GN-SIS	28.00	56.52	0.627	0.915	0.945	0.938	0.952	0.963	0.967	0.970	0.965	0.978	0.970
	SD-SIS	645.00	137.88	0.040	0.628	0.672	0.685	0.703	0.735	0.763	0.742	0.747	0.775	0.770
	RSD-SIS	10.00	2.22	0.962	0.982	0.997	0.990	0.995	0.998	1.000	0.998	1.000	1.000	1.000

用于训练集, 然后通过测试集来考察对应的筛选方法的表现. 具体的步骤如下: (1) 首先基于训练集利用特征筛选方法将维数由 $p=12600$ 降低到 $d_n=[n/\log n]=38$, 随后将筛选出的 d_n 个相关变量按由强到弱的顺序排列; (2) 利用 LDA 中的留一法交叉验证 (leave-one-out cross-validation), 从 2 到 d_n 中选择出最佳的模型或者模型大小, 然后再将该模型应用于测试集; (3) 重复随机切分 400 次, 根据训练分类误差、测试分类误差以及模型大小的平均值和标准差来评价相应的筛选方法的表现, 具体的模拟结果如表 9 所示. 通过该表可以清楚地观察到, 在该肿瘤分类问题中, 基于筛选程序 RSD-SIS 的 LDA 方法在训练和测试分类误差以及所选基因数量方面与 MV-SIS 的表现基本差不多, 且明显优于其他方法. 简单来说, RSD-SIS+LDA 和 MVSIS+LDA 使用少数与分类响应变量相关性较强的基因变量实现了相对最佳的表现. 这在一定程度上意味着 RSD-SIS 在实际中有更广泛的应用.

总体来说, 结合第 4.1 和 4.2 小节, 不难发现 RSD-SIS 非常适用于超高维分类数据的特征筛选, 且具有非常优秀的表现.

5　讨论与结论

为了度量连续型随机向量 $\boldsymbol{z}\in\mathbb{R}^d$ 和分类变量 $Y\in\mathbb{R}^1$ 的相关性, 本文基于特征函数与条件特征函数差异的加权积分提出一种稳健的半距离协方差 $\mathrm{RSDCov}(\boldsymbol{z},Y)$, 其中权重函数为 $\mathcal{S}\alpha\mathcal{S}$ 分布的概率密度函数. 通过利用中值启发式方法选择窗宽参数 σ, 本文构建一种稳健的半距离相关 $\mathrm{RSDCor}(\boldsymbol{z},Y)$,

表 9 在该实际数据中, 具有 5 个类别的肺癌数据的分类误差以及最后选择的模型大小, 其中括号外的数据表示平均值, 括号内的数据表示标准差

方法	$\tau = 0.8$			$\tau = 0.9$		
	训练分类误差 (%)	测试分类误差 (%)	模型大小	训练分类误差 (%)	测试分类误差 (%)	模型大小
CI-SIS	5.70 (2.23)	12.78 (4.86)	25.27 (5.65)	4.18 (1.71)	10.82 (6.32)	29.07 (4.61)
DC-SIS	5.68 (2.10)	12.68 (4.78)	22.78 (6.55)	5.47 (1.95)	11.56 (6.34)	26.29 (6.95)
GN-SIS	7.34 (2.40)	15.03 (4.74)	23.21 (5.19)	6.84 (1.85)	15.58 (5.98)	26.49 (5.78)
FKF-SIS	6.41 (3.10)	18.97 (6.96)	24.97 (5.71)	4.89 (1.56)	16.10 (7.41)	29.18 (4.41)
MV-SIS	3.36 (1.18)	9.16 (4.25)	20.19 (6.21)	3.60 (1.12)	9.22 (5.95)	20.24 (6.33)
PSIS	6.14 (2.21)	15.70 (5.22)	27.32 (3.72)	5.74 (2.03)	14.83 (7.31)	30.41 (3.55)
SD-SIS	8.89 (2.53)	16.35 (4.12)	21.65 (5.74)	8.25 (2.21)	17.55 (6.08)	25.82 (6.08)
RSD-SIS	3.72 (1.48)	9.33 (4.23)	21.16 (6.47)	3.75 (1.21)	8.97 (5.65)	23.63 (6.49)

其中 $\mathrm{RSDCor}(\boldsymbol{z}, Y)$ 具有位置尺度不变性且满足 $0 \leqslant \mathrm{RSDCor}(\boldsymbol{z}, Y) \leqslant 1$. 当 $\mathrm{RSDCor}(\boldsymbol{z}, Y) = 0$ 时, 当且仅当 $\boldsymbol{z}$ 和 Y 独立; 而当 $\mathrm{RSDCor}(\boldsymbol{z}, Y) = 1$ 时, 则很大可能意味着 $\boldsymbol{z}$ 与 Y 要满足如下 3 个条件. (a) 类别决定性: $\boldsymbol{z}$ 完全由 Y 决定, 即存在映射 $f : \{1, \ldots, R\} \to \mathbb{R}^d$, 使得 $\boldsymbol{z} = f(Y)$; (b) 均匀分布: 分类变量 Y 满足 $p_r = P(Y = L_r) = 1/R$, $r = 1, \ldots, R$; (c) 距离的一致性: 对于任意的 $L_r \neq L_s$, $\|f(L_r) - f(L_s)\|$ 的值相同. 进一步地, 本文基于 RSDCor 构建一种适用于超高维分类数据的特征筛选程序 RSD-SIS. 理论上, 本文证明了 RSD-SIS 的确定筛选性质以及秩相合性. Monte Carlo 模拟和实际数据分析均表明, RSD-SIS 与其对比方法相比具有非常强大的竞争力. 此外, 注意到加权函数在 R^d 上的取值恒大于 0, 故当 $\boldsymbol{z}$ 为 d 维离散变量时, $\mathrm{RSDCov}(\boldsymbol{z}, Y)$ 依然可以刻画 $\boldsymbol{z}$ 和 Y 的相关性. 也就是说, RSD-SIS 也可用于 $\boldsymbol{x}$ 为离散变量的情形, 具体的数值模拟分析可参见附录 B.5.

本文令 RSD-SIS 中的窗宽参数 σ_g 为 T_g 的中位数. 故从本质上说, RSD-SIS 是一个基于数据驱动的特征筛选程序. 若令 σ_g 趋于无穷, 则 RSD-SIS 与文献 [35] 提出的 SD-SIS 等价. 虽然数值模拟研究表明 RSD-SIS 在基于中值启发式的窗宽选择下有非常好的表现, 但是如何从特征筛选的角度给出最优的窗宽 σ 有待未来进一步探索. 此外, 本文在构造稳健的半距离协方差 $\mathrm{RSDCov}(\boldsymbol{z}, Y)$ 时, 采用的权重函数为满足 (2.3) 的稳定分布的概率密度函数, 其中 $\mathrm{RSDCov}(\boldsymbol{z}, Y)$ 要求指数参数 $\alpha \in (0, 2]$. 当 $\alpha > 2$ 时, 虽然可以基于特征函数和密度函数的逆转公式来找到适当的加权函数, 但是该加权函数在 $\mathbb{R}^d$ 上并不能做到恒大于 0, 故其并不是严格意义上的概率密度函数. 这意味着 (2.5) 在 $\alpha > 2$ 的情形下无法用于刻画 $\boldsymbol{z}$ 和 Y 的独立性.

致谢 本文作者非常感谢审稿人对稿件的认真审阅并提出宝贵的意见和建议, 从而使得论文的质量得到进一步的提高.

参考文献

1 Bhattacharjee A, Richards W G, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA, 2001, 98: 13790–13795
2 Bickel P J, Levina E. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli, 2004, 10: 989–1010
3 Chen L S, Paul D, Prentice R L, et al. A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. J Amer Statist Assoc, 2011, 106: 1345–1360
4 Cheng X W, Wang H. A generic model-free feature screening procedure for ultrahigh-dimensional data with categorical response. Comput Methods Programs Biomed, 2023, 229: 107269
5 Clemmensen L, Hastie T, Witten D, et al. Sparse discriminant analysis. Technometrics, 2011, 53: 406–413

6 Cui H J, Li R Z, Zhong W. Model-free feature screening for ultrahigh dimensional discriminant analysis. J Amer Statist Assoc, 2015, 110: 630–641
7 Cui H J, Zou F, Ling L. Feature screening and error variance estimation for ultrahigh-dimensional linear model with measurement errors. Commun Math Stat, 2025, 13: 139–171
8 Fan J Q, Fan Y Y. High-dimensional classification using features annealed independence rules. Ann Statist, 2008, 36: 2605–2637
9 Fan J Q, Lv J C. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol, 2008, 70: 849–911
10 Fan J Q, Samworth R, Wu Y C. Ultrahigh dimensional feature selection: Beyond the linear model. J Mach Learn Res, 2009, 10: 2013–2038
11 Fan J Q, Song R. Sure independence screening in generalized linear models with NP-dimensionality. Ann Statist, 2010, 38: 3567–3604
12 Gretton A, Fukumizu K, Teo C H, et al. A kernel statistical test of independence. Adv Neural Inform Proc Syst, 2008, 20: 585–592
13 He S M, Ma S G, Xu W L. A modified mean-variance feature-screening procedure for ultrahigh-dimensional discriminant analysis. Comput Statist Data Anal, 2019, 137: 155–169
14 Huang J, Horowitz J, Ma S G. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann Statist, 2008, 36: 587–613
15 Kang J, Hong H G, Li Y. Partition-based ultrahigh-dimensional variable screening. Biometrika, 2017, 104: 785–800
16 Li C N, Shang M Q, Shao Y H, et al. Sparse L1-norm two dimensional linear discriminant analysis via the generalized elastic net regularization. Neurocomputing, 2019, 337: 80–96
17 Li G R, Peng H, Zhang J, et al. Robust rank correlation based screening. Ann Statist, 2012, 40: 1846–1877
18 Li R Z, Zhong W, Zhu L P. Feature screening via distance correlation learning. J Amer Statist Assoc, 2012, 107: 1129–1139
19 Lu D, Weng Q. A survey of image classification methods and techniques for improving classification performance. Int J Remote Sens, 2007, 28: 823–870
20 Mai Q, Zou H. The Kolmogorov filter for variable screening in high-dimensional binary classification. Biometrika, 2013, 100: 229–234
21 Mai Q, Zou H. The fused Kolmogorov filter: A nonparametric model-free screening method. Ann Statist, 2015, 43: 1471–1497
22 Ni L, Fang F. Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification. J Nonparametr Stat, 2016, 28: 515–530
23 Nolan J P. Multivariate elliptically contoured stable distributions: Theory and estimation. Comput Statist, 2013, 28: 2067–2089
24 Pan R, Wang H S, Li R Z. Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening. J Amer Statist Assoc, 2016, 111: 169–179
25 Qiu D, Ahn J. Grouped variable screening for ultrahigh-dimensional data for linear model. Comput Statist Data Anal, 2020, 178: 1–15
26 Samorodnitsky G, Taqqu M S. Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance. New York: Chapman and Hall, 1994
27 Sang Y L, Dang X. Grouped feature screening for ultrahigh-dimensional classification via Gini distance correlation. J Multivariate Anal, 2024, 204: 105360
28 Serfling R J. Approximation Theorems of Mathematical Statistics. New York: John Wiley & Sons, 1980
29 Shao X F, Zhang J S. Martingale difference correlation and its use in high-dimensional variable screening. J Amer Statist Assoc, 2014, 109: 1302–1318
30 Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. Ann Statist, 2007, 35: 2769–2794
31 Tchapga C T, Mih T A, Kouanou A T, et al. Biomedical image classification in a big data architecture using machine learning algorithms. J Healthcare Eng, 2021, 2: 1–11
32 Trendafilov N T, Jolliffe I T. DALASS: Variable selection in discriminant analysis via the LASSO. Comput Statist Data Anal, 2007, 51: 3718–3736
33 Wang Z Z, Deng G M, Xu H Y. Group feature screening based on Gini impurity for ultrahigh-dimensional multi-classification. AIMS Math, 2023, 8: 4342–4362
34 Witten D M, Tibshirani R. Penalized classification using Fisher's linear discriminant. J R Stat Soc Ser B Stat Methodol, 2011, 73: 753–772

35 Zhong W, Li Z X, Guo W W, et al. Semi-distance correlation and its applications. J Amer Statist Assoc, 2024, 119: 2919–2933
36 Zhu L P, Li L X, Li R Z, et al. Model-free feature screening for ultrahigh-dimensional data. J Amer Statist Assoc, 2011, 106: 1464–1475
37 Zhu L P, Xu K, Li R Z, et al. Projection correlation between two random vectors. Biometrika, 2017, 104: 829–843
38 Zou F, Zou C L, Cui H J. A class of robust independence tests based on weighted integrals of empirical characteristic functions. Acta Math Sin (Engl Ser), 2024, 40: 2921–2952

附录 A 主要证明

本附录主要证明 (2.5)、(2.9)、性质 2.1–2.3 以及定理 3.1 和 3.2.

(2.5) 的证明 令 $D(Y_1,Y_2)=\sum_{r=1}^{R} I(Y_1=L_r,Y_2=L_r)/p_r-1$, 由定义 2.1 中的 (2.4) 可知,

$$\begin{aligned}
\mathrm{RSDCov}(\boldsymbol{z},Y)&=\sum_{r=1}^{R}p_r\int_{\mathbb{R}^d}|E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}})-E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}}\mid Y=L_r)|^2\omega(\boldsymbol{t})d\boldsymbol{t}\\
&=\sum_{r=1}^{R}p_r\int_{\mathbb{R}^d}[E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_1})-E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_1}\mid Y_1=L_r)][E(\mathrm{e}^{-\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_2})-E(\mathrm{e}^{-\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_2}\mid Y_2=L_r)]\omega(\boldsymbol{t})d\boldsymbol{t}\\
&=\sum_{r=1}^{R}p_r\int_{\mathbb{R}^d}E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_1-\boldsymbol{z}_2)}]\omega(\boldsymbol{t})d\boldsymbol{t}+\sum_{r=1}^{R}p_r\int_{\mathbb{R}^d}E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_1-\boldsymbol{z}_2)}\mid Y_1=L_r,Y_2=L_r]\omega(\boldsymbol{t})d\boldsymbol{t}\\
&\quad-\sum_{r=1}^{R}p_r\int_{\mathbb{R}^d}E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_1-\boldsymbol{z}_2)}\mid Y_2=L_r]\omega(\boldsymbol{t})d\boldsymbol{t}-\sum_{r=1}^{R}p_r\int_{\mathbb{R}^d}E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_1-\boldsymbol{z}_2)}\mid Y_1=L_r]\omega(\boldsymbol{t})d\boldsymbol{t}\\
&=E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}]+\sum_{r=1}^{R}E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}I(Y_1=L_r,Y_2=L_r)]/p_r-2E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}]\\
&=E\Big\{\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}\Big[\sum_{r=1}^{R}I(Y_1=L_r,Y_2=L_r)/p_r-1\Big]\Big\},
\end{aligned}\tag{A.1}$$

其中根据 Fubini 定理和 (2.3), 第 4 个等号成立. 至此, (2.5) 证毕. □

(2.9) 的证明 由定义 2.3 可知,

$$\begin{aligned}
[\mathrm{RMDD}(V\mid\boldsymbol{u})]^2&=\int_{\mathbb{R}^q}\{[E(V_1\mathrm{e}^{\mathrm{i}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{u}_1})-E(V_1)E(\mathrm{e}^{\mathrm{i}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{u}_1})][E(V_2\mathrm{e}^{-\mathrm{i}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{u}_2})-E(V_2)E(\mathrm{e}^{-\mathrm{i}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{u}_2})]\}\omega(\boldsymbol{s})d\boldsymbol{s}\\
&=\int_{\mathbb{R}^q}E[V_1V_2\mathrm{e}^{\mathrm{i}\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{u}_1-\boldsymbol{u}_2)}]\omega(\boldsymbol{s})d\boldsymbol{s}-\int_{\mathbb{R}^q}E[V_2]E[V_1\mathrm{e}^{\mathrm{i}\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{u}_1-\boldsymbol{u}_2)}]\omega(\boldsymbol{s})d\boldsymbol{s}\\
&\quad-\int_{\mathbb{R}^q}E[V_1]E[V_2\mathrm{e}^{\mathrm{i}\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{u}_1-\boldsymbol{u}_2)}]\omega(\boldsymbol{s})d\boldsymbol{s}+\int_{\mathbb{R}^q}E[V_2]E[V_1]E[\mathrm{e}^{\mathrm{i}\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{u}_1-\boldsymbol{u}_2)}]\omega(\boldsymbol{s})d\boldsymbol{s}\\
&=E[V_1V_2\mathrm{e}^{-(\|\boldsymbol{u}_1-\boldsymbol{u}_2\|/\sigma)}]-E[V_2]E[V_1\mathrm{e}^{-(\|\boldsymbol{u}_1-\boldsymbol{u}_2\|/\sigma)}]\\
&\quad-E[V_1]E[V_2\mathrm{e}^{-(\|\boldsymbol{u}_1-\boldsymbol{u}_2\|/\sigma)}]+E[V_1]E[V_2]E[\mathrm{e}^{-(\|\boldsymbol{u}_1-\boldsymbol{u}_2\|/\sigma)}]\\
&=E[(V_1-EV_1)(V_2-EV_2)\mathrm{e}^{-(\|\boldsymbol{u}_1-\boldsymbol{u}_2\|/\sigma)}],
\end{aligned}\tag{A.2}$$

其中根据 Fubini 定理和 (2.3), 第 4 个等号成立. 故 (2.9) 证毕. □

性质 2.1 的证明 对于性质 2.1(a), 只需要证明等式的右边等于左边即可:

$$E[U_{\boldsymbol{z}}D(Y_1,Y_2)] = E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}D(Y_1,Y_2)] - E\{E_{\boldsymbol{z}_1}[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}]D(Y_1,Y_2)\} - E\{E_{\boldsymbol{z}_2}[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}]D(Y_1,Y_2)\} + E\{E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}]D(Y_1,Y_2)\}. \tag{A.3}$$

注意到

$$E\{E_{\boldsymbol{z}_1}[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}]D(Y_1,Y_2)\} = E\{E_{\boldsymbol{z}_2}[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}]D(Y_1,Y_2)\} = E\{E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}]D(Y_1,Y_2)\} = 0,$$

故有

$$E[U_{\boldsymbol{z}}D(Y_1,Y_2)] = E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}D(Y_1,Y_2)] = \mathrm{RSDCov}(\boldsymbol{z},Y). \tag{A.4}$$

如此便证明了性质 2.1(a). 为证明性质 2.1(b), 根据 Cauchy 不等式, 有

$$\mathrm{RSDCov}(\boldsymbol{z},Y) = E[U_{\boldsymbol{z}}D(Y_1,Y_2)] \leqslant \sqrt{E[U_{\boldsymbol{z}}^2]E\{[D(Y_1,Y_2)]^2\}}. \tag{A.5}$$

而在 (A.5) 中, $E[U_{\boldsymbol{z}}^2]$ 和 $E\{[D(Y_1,Y_2)]^2\}$ 分别满足

$$\begin{aligned} E[U_{\boldsymbol{z}}^2] &= E\{[\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})} - E_{\boldsymbol{z}_1}(\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}) - E_{\boldsymbol{z}_2}(\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}) + E(\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})})]^2\} \\ &= E\{\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}[\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})} - E_{\boldsymbol{z}_1}(\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}) - E_{\boldsymbol{z}_2}(\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})}) + E(\mathrm{e}^{-(\frac{\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{\sigma})})]\} \\ &= E[\mathrm{e}^{-2(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}] - 2E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)-(\|\boldsymbol{z}_1-\boldsymbol{z}_3\|/\sigma)}] \\ &\quad + E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)-(\|\boldsymbol{z}_3-\boldsymbol{z}_4\|/\sigma)}], \end{aligned} \tag{A.6}$$

$$\begin{aligned} E\{[D(Y_1,Y_2)]^2\} &= E\left[\sum_{r=1}^{R} I(Y_1=L_r,Y_2=L_r)/p_r\right]^2 + 1 - 2E\left[\sum_{r=1}^{R} I(Y_1=L_r,Y_2=L_r)/p_r\right] \\ &= R-1. \end{aligned} \tag{A.7}$$

综合 (A.5)–(A.7), 性质 2.1(b) 证毕. 为证明性质 2.1(c), 只需要证明等式的右边等于左边:

$$\begin{aligned} &\sum_{r>s} p_r p_s(\mathcal{G}_{rr}+\mathcal{G}_{ss}-2\mathcal{G}_{rs}) \\ &= \sum_{r>s} p_r p_s \mathcal{G}_{rr} + \sum_{r>s} p_r p_s \mathcal{G}_{ss} - 2\sum_{r>s} p_r p_s \mathcal{G}_{rs} \\ &= \sum_{r>s}\left\{p_s \frac{E(\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}I(Y_1=L_r,Y_2=L_r))}{p_r}\right\} + \sum_{r>s}\left\{p_r \frac{E(\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}I(Y_1=L_s,Y_2=L_s))}{p_s}\right\} \\ &\quad - 2\sum_{r>s}\{E(\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}I(Y_1=L_r,Y_2=L_s))\} \\ &= \sum_{s=1}^{R}\sum_{r=s+1}^{R}\left\{\frac{p_s E(\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}I(Y_1=L_r,Y_2=L_r))}{p_r} + \frac{p_r E(\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}I(Y_1=L_s,Y_2=L_s))}{p_s}\right\} \\ &\quad + E\left\{\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}\left[\sum_{r=1}^{R} I(Y_1=L_r,Y_2=L_r)-1\right]\right\} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{r=1}^{R} \frac{E(\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)} I(Y_1=L_r, Y_2=L_r))}{p_r} - E\left[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)} \sum_{r=1}^{R} I(Y_1=L_r, Y_2=L_r)\right] \\
&\quad + E\left[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)} \sum_{r=1}^{R} I(Y_1=L_r, Y_2=L_r)\right] - E\left[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}\right] \\
&= E\left\{\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}\left[\sum_{r=1}^{R} \frac{I(Y_1=L_r, Y_2=L_r)}{p_r} - 1\right]\right\} \\
&= E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)} D(Y_1, Y_2)]. \qquad \text{(A.8)}
\end{aligned}
$$

此外, 有

$$
\begin{aligned}
&\int_{\mathbb{R}^d} |E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}} \mid Y=L_r] - E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}} \mid Y=L_s]|^2 \omega(\boldsymbol{t}) d\boldsymbol{t} \\
&= \int_{\mathbb{R}^d} \{E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_1} \mid Y_1=L_r] - E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_1} \mid Y_1=L_s]\}\{E[\mathrm{e}^{-\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_2} \mid Y_2=L_r] - E[\mathrm{e}^{-\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_2} \mid Y_2=L_s]\}\omega(\boldsymbol{t}) d\boldsymbol{t} \\
&= \int_{\mathbb{R}^d} E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_1-\boldsymbol{z}_2)} \mid Y_1=L_r, Y_2=L_r]\omega(\boldsymbol{t}) d\boldsymbol{t} - \int_{\mathbb{R}^d} E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_1-\boldsymbol{z}_2)} \mid Y_1=L_r, Y_2=L_s]\omega(\boldsymbol{t}) d\boldsymbol{t} \\
&\quad - \int_{\mathbb{R}^d} E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_1-\boldsymbol{z}_2)} \mid Y_1=L_s, Y_2=L_r]\omega(\boldsymbol{t}) d\boldsymbol{t} + \int_{\mathbb{R}^d} E[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_1-\boldsymbol{z}_2)} \mid Y_1=L_s, Y_2=L_s]\omega(\boldsymbol{t}) d\boldsymbol{t} \\
&= E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)} \mid Y_1=L_r, Y_2=L_r] - E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)} \mid Y_1=L_r, Y_2=L_s] \\
&\quad - E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)} \mid Y_1=L_s, Y_2=L_r] + E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)} \mid Y_1=L_s, Y_2=L_s] \\
&= \mathcal{G}_{rr} + \mathcal{G}_{ss} - 2\mathcal{G}_{rs}, \qquad \text{(A.9)}
\end{aligned}
$$

其中根据 Fubini 定理和 (2.3), (A.9) 的第 3 个等号成立. 故性质 2.1(c) 证毕. □

性质 2.3 的证明 为了更好地证明性质 2.2(a), 下面先证明性质 2.3. 根据 (2.9), 可得

$$
\begin{aligned}
\sum_{r=1}^{R} \{\mathrm{RMDD}[I(Y=L_r) \mid \boldsymbol{z}]\}^2/p_r &= \sum_{r=1}^{R} E\{[I(Y_1=L_r)-p_r][I(Y_2=L_r)-p_r]\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}\}/p_r \\
&= E\left[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}\left(\sum_{r=1}^{R} I(Y_1=L_r, Y_2=L_r)/p_r - 1\right)\right]. \qquad \text{(A.10)}
\end{aligned}
$$

如此便证明了性质 2.3(a). 令 $\hat{p}_r = \sum_{i=1}^{n} I(Y_i=L_r)/n$, $E_{ij} = (I(Y_i=L_r)-\hat{p}_r)(I(Y_j=L_r)-\hat{p}_r)$, 根据 (2.10), 可得

$$
\begin{aligned}
\sum_{r=1}^{R} \{\mathrm{RMDD}_n[I(Y=L_r) \mid \boldsymbol{z}]\}^2/\hat{p}_r &= \frac{1}{n^2} \sum_{r=1}^{R} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} E_{ij}/\hat{p}_r \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \left(\sum_{r=1}^{R} E_{ij}/\hat{p}_r\right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} B_{ij}, \qquad \text{(A.11)}
\end{aligned}
$$

其中 A_{ij} 和 B_{ij} 的定义见 (2.6). 故性质 2.3(b) 证毕. □

性质 2.2 的证明 为了证明性质 2.2(a), 首先不难发现 (2.9) 的 $[\mathrm{RMDD}(V \mid \boldsymbol{u})]^2$ 可改写成

$$[\mathrm{RMDD}(V \mid \boldsymbol{u})]^2 = E[V_1V_2\mathrm{e}^{-(\|\boldsymbol{u}_1-\boldsymbol{u}_2\|/\sigma)}] - 2E[V_1V_3\mathrm{e}^{-(\|\boldsymbol{u}_1-\boldsymbol{u}_2\|/\sigma)}] + E[V_3V_4\mathrm{e}^{-(\|\boldsymbol{u}_1-\boldsymbol{u}_2\|/\sigma)}]. \tag{A.12}$$

根据 (A.12) 和性质 2.3(a), 可得

$$\begin{aligned}\mathrm{RSDCov}(\boldsymbol{z}, Y) = \sum_{r=1}^{R}\{&E[I(Y_1 = L_r)I(Y_2 = L_r)\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}] - 2E[I(Y_1 = L_r)I(Y_3 = L_r)\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}] \\ &+ E[I(Y_3 = L_r)I(Y_4 = L_r)\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma)}]\}/p_r.\end{aligned} \tag{A.13}$$

根据 (A.13), 令

$$\begin{aligned}\mathcal{V}_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[\mathrm{e}^{-(\|\boldsymbol{z}_i-\boldsymbol{z}_j\|/\sigma)}I(Y_i = L_r)I(Y_j = L_r)], \\ \mathcal{V}_2 &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}[\mathrm{e}^{-(\|\boldsymbol{z}_i-\boldsymbol{z}_j\|/\sigma)}I(Y_i = L_r)I(Y_k = L_r)], \\ \mathcal{V}_3 &= \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}[\mathrm{e}^{-(\|\boldsymbol{z}_i-\boldsymbol{z}_j\|/\sigma)}I(Y_k = L_r)I(Y_l = L_r)],\end{aligned} \tag{A.14}$$

则定义在 (2.6) 的样本估计量可改写成

$$\mathrm{RSDCov}_n(\boldsymbol{z}, Y) = \sum_{r=1}^{R}\left\{\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}h[(\boldsymbol{z}_i, Y_i), (\boldsymbol{z}_j, Y_j), (\boldsymbol{z}_k, Y_k), (\boldsymbol{z}_l, Y_l)]\right\}/\hat{p}_r, \tag{A.15}$$

其中核函数 $h(\cdot,\cdot,\cdot,\cdot)$ 满足

$$\begin{aligned}&h((\boldsymbol{z}_i, Y_i), (\boldsymbol{z}_j, Y_j), (\boldsymbol{z}_k, Y_k), (\boldsymbol{z}_l, Y_l)) \\ &= \frac{1}{4!}\sum_{s,t,u,v}^{i,j,k,l}[\mathrm{e}^{-(\|\boldsymbol{z}_s-\boldsymbol{z}_t\|/\sigma)}I(Y_s = L_r)I(Y_t = L_r) - 2\mathrm{e}^{-(\|\boldsymbol{z}_s-\boldsymbol{z}_t\|/\sigma)}I(Y_s = L_r)I(Y_u = L_r) \\ &\quad + \mathrm{e}^{-(\|\boldsymbol{z}_s-\boldsymbol{z}_t\|/\sigma)}I(Y_u = L_r)I(Y_v = L_r)].\end{aligned} \tag{A.16}$$

注意到 $E(|h((\boldsymbol{z}_i, Y_i), (\boldsymbol{z}_j, Y_j), (\boldsymbol{z}_k, Y_k), (\boldsymbol{z}_l, Y_l))|) \leqslant 4$, 根据文献 [28, 第 5.4 小节, 定理 A], 易知

$$(\mathcal{V}_1 - 2\mathcal{V}_2 + \mathcal{V}_3) \xrightarrow{\text{a.s.}} \{\mathrm{RMDD}[I(Y = L_r) \mid \boldsymbol{z}]\}^2. \tag{A.17}$$

由强大数律, 可得 $\hat{p}_r \xrightarrow{\text{a.s.}} p_r$. 对于给定的 R, 根据连续映射定理, 可知

$$\begin{aligned}\mathrm{RSDCov}_n(\boldsymbol{z}, Y) = \sum_{r=1}^{R}[\mathcal{V}_1 - 2\mathcal{V}_2 + \mathcal{V}_3]/\hat{p}_r &\xrightarrow{\text{a.s.}} \sum_{r=1}^{R}\{\mathrm{RMDD}[I(Y = L_r) \mid \boldsymbol{z}]\}^2/p_r \\ &= \mathrm{RSDCov}(\boldsymbol{z}, Y).\end{aligned} \tag{A.18}$$

故性质 2.2(a) 证毕. 对于性质 2.2(b), 只需要证明等式的右边等于左边, 即

$$\sum_{r=1}^{R}\hat{p}_r\int_{\mathbb{R}^d}[E_n(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}} \mid Y = L_r) - E_n(\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}})]^2\omega(\boldsymbol{t})d\boldsymbol{t}$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{r=1}^{R}\hat{p}_r\int_{\mathbb{R}^d}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_i-\boldsymbol{z}_j)}\left(\frac{I(Y_i=r)}{\hat{p}_r}-1\right)\left(\frac{I(Y_j=r)}{\hat{p}_r}-1\right)\omega(\boldsymbol{t})d\boldsymbol{t}\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{z}_i-\boldsymbol{z}_j)}\left[\sum_{r=1}^{R}\frac{I(Y_i=L_r,Y_j=L_r)}{\hat{p}_r}-1\right]\omega(\boldsymbol{t})d\boldsymbol{t}\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{e}^{-(\|\boldsymbol{z}_i-\boldsymbol{z}_j\|/\sigma)}\left[\sum_{r=1}^{R}\frac{I(Y_i=L_r,Y_j=L_r)}{\hat{p}_r}-1\right]\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}B_{ij}. \qquad (A.19)
\end{aligned}
$$

故性质 2.2(b) 证毕. 为证明性质 2.2(c), 由 (2.6) 和 Cauchy 不等式可知

$$
\mathrm{RSDCov}_n(\boldsymbol{z},Y)=\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}A_{kl}B_{kl}\leqslant\sqrt{\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}A_{kl}^2}\sqrt{\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}B_{kl}^2}. \qquad (A.20)
$$

而在 (A.20) 中, $\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}A_{kl}^2$ 和 $\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}B_{kl}^2$ 分别满足

$$
\begin{aligned}
\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}A_{kl}^2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{e}^{-2(\|\boldsymbol{z}_i-\boldsymbol{z}_j\|/\sigma)}-\frac{2}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathrm{e}^{-(\|\boldsymbol{z}_i-\boldsymbol{z}_j\|/\sigma+\|\boldsymbol{z}_i-\boldsymbol{z}_k\|/\sigma)}\\
&\quad+\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\mathrm{e}^{-(\|\boldsymbol{z}_i-\boldsymbol{z}_j\|/\sigma+\|\boldsymbol{z}_k-\boldsymbol{z}_l\|/\sigma)}, \qquad (A.21)
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}B_{kl}^2 &= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}\left[\left(\sum_{r=1}^{R}\frac{I(Y_k=L_r,Y_l=L_r)}{\hat{p}_r}\right)^2+1-2\sum_{r=1}^{R}\frac{I(Y_k=L_r,Y_l=L_r)}{\hat{p}_r}\right]\\
&= R-1. \qquad (A.22)
\end{aligned}
$$

结合 (A.20)–(A.22), 性质 2.2(c) 证毕. □

为了更好地证明定理 3.1 和 3.2, 先给出如下引理.

引理 A.1 令 $h(Z_1,\ldots,Z_m)$ 表示 U- 统计量 T_n 的核. 若 $a\leqslant h(Z_1,\ldots,Z_m)\leqslant b$, 则对于任意的 $\varepsilon>0$ 和 $n\geqslant m$, 有

$$
P\{|T_n-E[T_n]|\geqslant\varepsilon\}\leqslant 2\exp\{-2[n/m]\varepsilon^2/(b-a)^2\},
$$

其中 $[n/m]$ 表示 n/m 的整数部分.

证明 根据文献 [28, 定理 5.6.1A] 和 U- 统计量的对称性, 该引理立刻得证. □

引理 A.2 假设 θ_1 和 θ_2 为两个有界参数, 即存在常数 $M_1>0$ 和 $M_2>0$, 使得 $|\theta_1|\leqslant M_1$ 和 $|\theta_2|\leqslant M_2$. 令 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 分别为基于 n 个样本的 θ_1 和 θ_2 的估计量. 如果对于任意的 $\varepsilon\in(0,1)$, 存在 $c_1>0$ 和 $c_2>0$ 使得

$$
P(|\hat{\theta}_1-\theta_1|\geqslant\varepsilon)\leqslant c_1\exp(-n\varepsilon^2/c_1),\quad P(|\hat{\theta}_2-\theta_2|\geqslant\varepsilon)\leqslant c_2\exp(-n\varepsilon^2/c_2),
$$

则有

$$
P(|\hat{\theta}_1\hat{\theta}_2-\theta_1\theta_2|\geqslant\varepsilon)\leqslant\mathcal{C}_1\exp(-n\varepsilon^2/\mathcal{C}_1),\quad P(|\hat{\theta}_1^2-\theta_1^2|\geqslant\varepsilon)\leqslant\mathcal{C}_2\exp(-n\varepsilon^2/\mathcal{C}_2),
$$

$$P(|(\hat{\theta}_1-\hat{\theta}_2)-(\theta_1-\theta_2)|\geqslant\varepsilon)\leqslant\mathcal{C}_3\exp(-n\varepsilon^2/\mathcal{C}_3).$$

其中 $\mathcal{C}_1,\mathcal{C}_2$ 和 $\mathcal{C}_3$ 均为某个大于 0 的常数. 此外, 若存在 $M_3>0$ 使得 $|\theta_2|\geqslant M_3$, 则有

$$P\left(\left|\frac{\hat{\theta}_1}{\hat{\theta}_2}-\frac{\theta_1}{\theta_2}\right|\geqslant\varepsilon\right)\leqslant\mathcal{C}_4\exp(-n\varepsilon^2/\mathcal{C}_4),$$

其中 $\mathcal{C}_4$ 为某个大于 0 的常数. 若进一步假设 $\theta_2\geqslant 0$, 则存在 $\mathcal{C}_5>0$ 使得

$$P(|\sqrt{\hat{\theta}_2}-\sqrt{\theta_2}|\geqslant\varepsilon)\leqslant\mathcal{C}_5\exp(-n\varepsilon^2/\mathcal{C}_5).$$

证明 证明可参见文献 [7, 引理 6.2]. □

引理 A.3 对于任意的 $\varepsilon\in(0,1)$ 和 $1\leqslant r\leqslant R$, 有

$$P(|\hat{p}_r-p_r|\geqslant\varepsilon)\leqslant 2\exp\left\{-\frac{n\varepsilon^2}{2(p_r+\varepsilon/3)}\right\}.$$

证明 证明可参见文献 [6, 引理 A.3]. □

引理 A.4 假设 $\{\boldsymbol{z}_k\}_{k=1}^n$ 为独立同分布于 $\boldsymbol{z}$ 的 n 个随机样本, 定义 $T_{\boldsymbol{z}}=\|\boldsymbol{z}_1-\boldsymbol{z}_2\|$, $F_{\boldsymbol{z}}$ 表示 $T_{\boldsymbol{z}}$ 的累积分布函数, $F_{\boldsymbol{z}n}(t)=\frac{2}{n(n-1)}\sum_{1\leqslant i<j\leqslant n}I(\|\boldsymbol{z}_i-\boldsymbol{z}_j\|\leqslant t)$ 表示 $T_{\boldsymbol{z}}$ 的经验分布函数. 进一步地, 假定 σ 和 $\hat{\sigma}$ 分别为 $F_{\boldsymbol{z}}(\cdot)$ 和 $F_{\boldsymbol{z}n}(\cdot)$ 的中位数, $F_{\boldsymbol{z}}(\cdot)$ 在 σ 处可导, 则对于任意的 $\varepsilon>0$, 有

$$P(|\hat{\sigma}-\sigma|\geqslant\varepsilon)\leqslant 2\exp\{-nf_{\boldsymbol{z}}^2(\sigma)\varepsilon^2/4\},$$

其中 $f_{\boldsymbol{z}}(\sigma)$ 表示 $F_{\boldsymbol{z}}(\cdot)$ 在 σ 处的导数.

证明 由 $\hat{\sigma}$ 是 $F_{\boldsymbol{z}n}(\cdot)$ 的中位数可知, $F_{\boldsymbol{z}n}(\hat{\sigma})\geqslant 1/2$ 且 $F_{\boldsymbol{z}n}(\hat{\sigma}^-)\leqslant 1/2$, 其中 $F_{\boldsymbol{z}n}(\hat{\sigma}^-)=\lim_{t\to\hat{\sigma}}F_{\boldsymbol{z}n}(t)$ 表示左极限. 若 $\hat{\sigma}\geqslant\sigma+\varepsilon$, 则有 $F_{\boldsymbol{z}n}(\sigma+\varepsilon)\leqslant 1/2$; 同理, 若 $\hat{\sigma}\leqslant\sigma-\varepsilon$, 则有 $F_{\boldsymbol{z}n}(\sigma-\varepsilon)\geqslant 1/2$. 因此, 事件 $|\hat{\sigma}-\sigma|\geqslant\varepsilon$ 包含在以下联合事件中, 即

$$\{|\hat{\sigma}-\sigma|\geqslant\varepsilon\}=[\{\hat{\sigma}\geqslant\sigma+\varepsilon\}\cup\{\hat{\sigma}\leqslant\sigma-\varepsilon\}]\subset[\{F_{\boldsymbol{z}n}(\sigma+\varepsilon)\leqslant 1/2\}\cup\{F_{\boldsymbol{z}n}(\sigma-\varepsilon)\geqslant 1/2\}]. \tag{A.23}$$

由于经验分布 $F_{\boldsymbol{z}n}(t)$ 是一个 U 统计量, 核函数 $h(\boldsymbol{z}_i,\boldsymbol{z}_j)=I(\|\boldsymbol{z}_i-\boldsymbol{z}_j\|\leqslant t)\in[0,1]$. 故根据引理 A.1, 对于给定的 t 和任意的 $\delta>0$, 有

$$P(|F_{\boldsymbol{z}n}(t)-F_{\boldsymbol{z}}(t)|\geqslant\delta)\leqslant 2\exp\{-n\delta^2\}. \tag{A.24}$$

注意到 (A.24) 的右边与 t 无关, 故其对任意的 t 均成立. 根据 Taylor 展开, 可得 $F_{\boldsymbol{z}}(\sigma+\varepsilon)=\frac{1}{2}+f_{\boldsymbol{z}}(\sigma)\varepsilon+O(\varepsilon^2)$. 令 $\Delta_1=F_{\boldsymbol{z}}(\sigma+\varepsilon)-\frac{1}{2}=f_{\boldsymbol{z}}(\sigma)\varepsilon+O(\varepsilon^2)$, 对于充分小的 ε, 有 $\Delta_1\geqslant\frac{f_{\boldsymbol{z}}(\sigma)\varepsilon}{2}$. 令 (A.24) 中的 $t=\sigma+\varepsilon$, 则

$$\{F_{\boldsymbol{z}n}(\sigma+\varepsilon)\leqslant 1/2\}=\{F_{\boldsymbol{z}n}(\sigma+\varepsilon)\leqslant F_{\boldsymbol{z}}(\sigma+\varepsilon)-\Delta_1\}=\{F_{\boldsymbol{z}n}(\sigma+\varepsilon)-F_{\boldsymbol{z}}(\sigma+\varepsilon)\leqslant-\Delta_1\}. \tag{A.25}$$

因此,

$$P(F_{\boldsymbol{z}n}(\sigma+\varepsilon)\leqslant 1/2)=P(F_{\boldsymbol{z}n}(\sigma+\varepsilon)-F_{\boldsymbol{z}}(\sigma+\varepsilon)\leqslant-\Delta_1)\leqslant\exp\{-nf_{\boldsymbol{z}}^2(\sigma)\varepsilon^2/4\}. \tag{A.26}$$

同理, $F_{\boldsymbol{z}}(\sigma-\varepsilon)=\frac{1}{2}-f_{\boldsymbol{z}}(\sigma)\varepsilon+O(\varepsilon^2)$. 令 $\Delta_2=\frac{1}{2}-F_{\boldsymbol{z}}(\sigma-\varepsilon)=f_{\boldsymbol{z}}(\sigma)\varepsilon-O(\varepsilon^2)$, 对于充分小的 ε, 有 $\Delta_2\geqslant\frac{f_{\boldsymbol{z}}(\sigma)\varepsilon}{2}$. 令 (A.24) 中的 $t=\sigma-\varepsilon$, 则

$$\{F_{\boldsymbol{z}n}(\sigma-\varepsilon)\geqslant 1/2\}=\{F_{\boldsymbol{z}n}(\sigma-\varepsilon)\geqslant F_{\boldsymbol{z}}(\sigma-\varepsilon)+\Delta_2\}=\{F_{\boldsymbol{z}n}(\sigma-\varepsilon)-F_{\boldsymbol{z}}(\sigma-\varepsilon)\geqslant\Delta_2\}. \tag{A.27}$$

进一步有

$$P(F_{\boldsymbol{z}n}(\sigma-\varepsilon)\geqslant 1/2)=P(F_{\boldsymbol{z}n}(\sigma-\varepsilon)-F_{\boldsymbol{z}}(\sigma-\varepsilon)\geqslant\Delta_2)\leqslant\exp\{-nf_{\boldsymbol{z}}^2(\sigma)\varepsilon^2/4\}. \tag{A.28}$$

最后, 综合 (A.23), (A.26) 和 (A.28), 有

$$P(|\hat{\sigma}-\sigma|\geqslant\varepsilon)\leqslant P(F_{\boldsymbol{z}n}(\sigma+\varepsilon)\leqslant 1/2)+P(F_{\boldsymbol{z}n}(\sigma-\varepsilon)\geqslant 1/2)\leqslant 2\exp\{-nf_{\boldsymbol{z}}^2(\sigma)\varepsilon^2/4\}. \tag{A.29}$$

证毕. □

引理 A.5 假设条件 (C_1) 成立, 对于任意的 $g=1,\ldots,G$, $r=1,\ldots,R$ 以及任意的 $\varepsilon\in(0,1)$, 存在某个常数 $\zeta>0$ 使得

$$P(|(\mathrm{RMDD}_{m,n}^{g,r})^2-(\mathrm{RMDD}_{m}^{g,r})^2|\geqslant\varepsilon)\leqslant 18\exp\{-n\zeta\varepsilon^2\},$$

其中 $(\mathrm{RMDD}_{m,n}^{g,r})^2=\{\mathrm{RMDD}_{m,n}[I(Y=L_r)\mid X_{(g)}]\}^2$, $(\mathrm{RMDD}_{m}^{g,r})^2=\{\mathrm{RMDD}_{m}[I(Y=L_r)\mid X_{(g)}]\}^2$, 下标 m 仅表示此时的窗宽参数 σ_g 为 $T_g=\|X_{1(g)}-X_{2(g)}\|$ 的中位数.

证明 为了方便, 令 $Y_i^r:=I(Y_i=L_r)$, $\omega_g^{ij}=\mathrm{e}^{-(\|X_{i(g)}-X_{j(g)}\|/\sigma_g)}$, $\hat{\omega}_g^{ij}=\mathrm{e}^{-(\|X_{i(g)}-X_{j(g)}\|/\hat{\sigma}_g)}$, $i,j=1,\ldots,n$, 其中 $\hat{\sigma}_g=\mathrm{Med}\{\|X_{i(g)}-X_{j(g)}\|,1\leqslant i<j\leqslant n\}$. 根据 (A.12) 和 (A.14), 可得

$$(\mathrm{RMDD}_{m}^{g,r})^2=\psi_1^g-2\psi_2^g+\psi_3^g,\quad (\mathrm{RMDD}_{m,n}^{g,r})^2=\psi_{n1}^g-2\psi_{n2}^g+\psi_{n3}^g, \tag{A.30}$$

其中

$$\begin{aligned}&\psi_1^g=E[Y_1^rY_2^r\omega_g^{12}],\quad \psi_2^g=E[Y_1^rY_3^r\omega_g^{12}],\quad \psi_3^g=E[Y_3^rY_4^r\omega_g^{12}],\quad \psi_{n1}^g=\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n[\hat{\omega}_g^{ij}Y_i^rY_j^r],\\ &\psi_{n2}^g=\frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n[\hat{\omega}_g^{ij}Y_i^rY_k^r],\quad \psi_{n3}^g=\frac{1}{n^4}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n[\hat{\omega}_g^{ij}Y_k^rY_l^r].\end{aligned} \tag{A.31}$$

根据条件 (C_1), 存在正数 m 使得 $\sigma_g\geqslant\tilde{m}$, $\forall\, g=1,\ldots,G$. 不失一般性, 对于任意的 $0<\varepsilon<1$, 有

$$\begin{aligned}P(|(\mathrm{RMDD}_{m,n}^{g,r})^2-(\mathrm{RMDD}_{m}^{g,r})^2|\geqslant\varepsilon)&=P(|(\psi_{n1}^g-2\psi_{n2}^g+\psi_{n3}^g)-(\psi_1^g-2\psi_2^g+\psi_3^g)|\geqslant\varepsilon)\\ &\leqslant P(|\psi_{n1}^g-\psi_1^g|\geqslant\varepsilon/3)+P(|\psi_{n2}^g-\psi_2^g|\geqslant\varepsilon/6)\\ &\quad+P(|\psi_{n3}^g-\psi_3^g|\geqslant\varepsilon/3).\end{aligned} \tag{A.32}$$

对于 (A.32) 右边的第一部分, 有

$$\begin{aligned}P(|\psi_{n1}^g-\psi_1^g|\geqslant\varepsilon/3)&=P\left(\left|\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n[\hat{\omega}_g^{ij}Y_i^rY_j^r]-E[Y_1^rY_2^r\omega_g^{12}]\right|\geqslant\varepsilon/3\right)\\ &\leqslant P\left(\left|\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n[\hat{\omega}_g^{ij}Y_i^rY_j^r]-\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n[\omega_g^{ij}Y_i^rY_j^r]\right|\geqslant\varepsilon/6\right)\end{aligned}$$

$$+P\bigg(\bigg|\frac{1}{n^2}\sum_{i,j=1}^{n}[\omega_g^{ij}Y_i^rY_j^r]-E[Y_1^rY_2^r\omega_g^{12}]\bigg|\geqslant\varepsilon/6\bigg). \tag{A.33}$$

根据引理 A.1, 对于 (A.33) 右边的第二部分, 有

$$\begin{aligned}
&P\bigg(\bigg|\frac{1}{n^2}\sum_{i,j=1}^{n}[\omega_g^{ij}Y_i^rY_j^r]-E[Y_1^rY_2^r\omega_g^{12}]\bigg|\geqslant\varepsilon/6\bigg)\\
&\leqslant P\bigg(\bigg|\frac{n-1}{n}\frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}[\omega_g^{ij}Y_i^rY_j^r]-E[Y_1^rY_2^r\omega_g^{12}]\bigg|\geqslant\varepsilon/6-1/n\bigg)\\
&\leqslant P\bigg(\bigg|\frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}[\omega_g^{ij}Y_i^rY_j^r]-E[Y_1^rY_2^r\omega_g^{12}]\bigg|\geqslant\varepsilon/12\bigg)\\
&\leqslant 2\exp\{-n\varepsilon^2/144\},
\end{aligned} \tag{A.34}$$

其中 (A.34) 中第二个不等号成立是因为对于任意的 $\varepsilon>0$, 存在充分大的 n 使得 $1/n+|E[Y_1^rY_2^r\omega_g^{12}]|/n\leqslant\varepsilon/12$. 对于 (A.33) 右边的第一部分, 有

$$\begin{aligned}
&P\bigg(\bigg|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[\hat{\omega}_g^{ij}Y_i^rY_j^r]-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[\omega_g^{ij}Y_i^rY_j^r]\bigg|\geqslant\varepsilon/6\bigg)\\
&\leqslant P\bigg(\bigg|\frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}Y_i^rY_j^r[\hat{\omega}_g^{ij}-\omega_g^{ij}]\bigg|\geqslant\varepsilon/6\bigg).
\end{aligned} \tag{A.35}$$

为了方便, 令 $\mathcal{D}=|\frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}Y_i^rY_j^r[\hat{\omega}_g^{ij}-\omega_g^{ij}]|$, $\Delta_g=\hat{\sigma}_g-\sigma_g$, 根据 $\hat{\omega}_g^{ij}$ 的一阶 Taylor 展开, 有

$$\begin{aligned}
[\hat{\omega}_g^{ij}-\omega_g^{ij}]&=[\mathrm{e}^{-(\|X_{i(g)}-X_{j(g)}\|/\hat{\sigma}_g)}-\mathrm{e}^{-(\|X_{i(g)}-X_{j(g)}\|/\sigma_g)}]\\
&=\mathrm{e}^{-(\|X_{i(g)}-X_{j(g)}\|/\sigma_g)}(\|X_{i(g)}-X_{j(g)}\|/\sigma_g^2)\Delta_g+O(\Delta_g^2).
\end{aligned} \tag{A.36}$$

令 $\mathcal{S}=\frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}\mathrm{e}^{-(\|X_{i(g)}-X_{j(g)}\|/\sigma_g)}(\|X_{i(g)}-X_{j(g)}\|/\sigma_g^2)Y_i^rY_j^r$, $\mathcal{Q}=\frac{\varepsilon}{12(E\mathcal{S}+1)}$, 则 (A.35) 可写成

$$P(\mathcal{D}\geqslant\varepsilon/6)=P(\mathcal{D}\geqslant\varepsilon/6,|\Delta_g|\geqslant\mathcal{Q})+P(\mathcal{D}\geqslant\varepsilon/6,|\Delta_g|\leqslant\mathcal{Q}). \tag{A.37}$$

因为对于任意的 $x>0$, $x\mathrm{e}^{-x}<1$ 成立, 故 $0\leqslant\mathcal{S}\leqslant1/\sigma_g$. 此外, 对于充分小的 ε, 当 $|\Delta_g|\leqslant\mathcal{Q}$ 时, 有 $O(\Delta_g^2)\leqslant\varepsilon/12$. 这也意味着当 $|\Delta_g|\leqslant\mathcal{Q}$ 时, $\mathcal{D}\leqslant|\Delta_g|\mathcal{S}+\varepsilon/12$ 且 $\{\mathcal{D}\geqslant\varepsilon/6,|\Delta_g|\leqslant\mathcal{Q}\}\subset\{\mathcal{S}\geqslant E\mathcal{S}+1\}$. 基于假设 $0<\varepsilon<1$, 则进一步有

$$\begin{aligned}
P(\mathcal{D}\geqslant\varepsilon/6)&\leqslant P(|\Delta_g|\geqslant\mathcal{Q})+P(\mathcal{S}\geqslant E\mathcal{S}+1)\leqslant P(|\Delta_g|\geqslant\mathcal{Q})+P(\mathcal{S}-E\mathcal{S}\geqslant\varepsilon)\\
&\leqslant 2\exp\{-nf_g^2(\sigma_g)\varepsilon^2/(576(E\mathcal{S}+1)^2)\}+\exp\{-n\sigma_g^2\varepsilon^2\}\\
&\leqslant 2\exp\{-n\check{m}^2\varepsilon^2/(576(E\mathcal{S}+1)^2)\}+\exp\{-n\tilde{m}^2\varepsilon^2\},
\end{aligned} \tag{A.38}$$

其中基于引理 A.4 和 A.1, 第三个不等号的第一和第二部分分别成立, 而最后一个不等号成立是基于条件 (C_1). 令 $C_1=\min\{\check{m}^2/(576(E\mathcal{S}+1)^2),\tilde{m}^2\}$, 综合 (A.33)–(A.38), 可得

$$P(|\psi_{n1}^g-\psi_1^g|\geqslant\varepsilon/3)\leqslant 5\exp\{-nC_2\varepsilon^2\}, \tag{A.39}$$

其中 $C_2=\min\{1/144,C_1\}$. 对于 (A.32) 右边的第二部分, 有

$$\begin{aligned}P(|\psi_{n2}^g-\psi_2^g|\geqslant\varepsilon/6)&=P\bigg(\bigg|\frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n[\hat{\omega}_g^{ij}Y_i^rY_k^r]-E[Y_1^rY_3^r\omega_g^{12}]\bigg|\geqslant\varepsilon/6\bigg)\\&\leqslant\underbrace{P\bigg(\bigg|\frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n[\hat{\omega}_g^{ij}Y_i^rY_k^r]-\frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n[\omega_g^{ij}Y_i^rY_k^r]\bigg|\geqslant\varepsilon/12\bigg)}_{I_1}\\&\quad+\underbrace{P\bigg(\bigg|\frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n[\omega_g^{ij}Y_i^rY_k^r]-E[Y_1^rY_3^r\omega_g^{12}]\bigg|\geqslant\varepsilon/12\bigg)}_{I_2}.\end{aligned}\tag{A.40}$$

根据 (A.36) 和 (A.40), 令 $\mathcal{W}=\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k=1}^n\mathrm{e}^{-(\|X_{i(g)}-X_{j(g)}\|/\sigma_g)}(\|X_{i(g)}-X_{j(g)}\|/\sigma_g^2)Y_i^rY_k^r$, 则

$$\begin{aligned}I_1&\leqslant P\bigg(\bigg|\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k=1}^n[\hat{\omega}_g^{ij}Y_i^rY_k^r]-\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k=1}^n[\omega_g^{ij}Y_i^rY_k^r]\bigg|\geqslant\varepsilon/24\bigg)\\&\leqslant 2\exp\{-n\check{m}^2\varepsilon^2/(9216(E\mathcal{W}+1)^2)\}+\exp\{-2n\tilde{m}^2\varepsilon^2/3\}\\&\leqslant 3\exp\{-nC_3\varepsilon^2\},\end{aligned}\tag{A.41}$$

其中 $C_3=\min\{\check{m}^2/(9216(E\mathcal{W}+1)^2),2\tilde{m}^2/3\}$. 在 (A.41) 中, 第一个不等号成立是因为当 n 较大时, 非主阶项可以由 $\varepsilon/24$ 来控制; 而第二个不等号成立是基于 (A.38) 的证明思路以及引理 A.4 和 A.1. 为了更方便地证明 I_2, 令

$$h(i,j,k):=h((X_{i(g)},Y_i),(X_{j(g)},Y_j),(X_{k(g)},Y_k))=\sum_{u,v,w}^{i,j,k}[\omega_g^{ij}Y_i^rY_k^r]/6,$$

其中 (u,v,w) 表示 (i,j,k) 的一个排列, 共有 6 种情况. 进一步有

$$\begin{aligned}I_2&\leqslant P\bigg(\bigg|\frac{1}{n^3}\sum_{i\neq j\neq k=1}^n[\omega_g^{ij}Y_i^rY_k^r]-E[Y_1^rY_3^r\omega_g^{12}]\bigg|\geqslant\varepsilon/12-3/n-1/n^2\bigg)\\&\leqslant P\bigg(\bigg|\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k=1}^n[\omega_g^{ij}Y_i^rY_k^r]-E[Y_1^rY_3^r\omega_g^{12}]\bigg|\geqslant\varepsilon/12-6/n+1/n^2\bigg)\\&\leqslant P\bigg(\bigg|\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k=1}^n[\omega_g^{ij}Y_i^rY_k^r]-E[Y_1^rY_3^r\omega_g^{12}]\bigg|\geqslant\varepsilon/24\bigg)\\&=P\bigg(\bigg|\frac{6}{n(n-1)(n-2)}\sum_{1\leqslant i<j<k\leqslant n}h(i,j,k)-E[Y_1^rY_3^r\omega_g^{12}]\bigg|\geqslant\varepsilon/24\bigg)\\&\leqslant 2\exp\{-n\varepsilon^2/864\},\end{aligned}\tag{A.42}$$

其中根据引理 A.1, 最后一个不等号成立. 令 $C_4=\min\{C_3,1/864\}$, 综合 (A.40)–(A.42), 可以得到

$$P(|\psi_{n2}^g-\psi_2^g|\geqslant\epsilon/6)\leqslant 5\exp\{-nC_4\varepsilon^2\}.\tag{A.43}$$

为证明 (A.32) 右边的第三部分, 令 $\mathcal{U}=\frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}\mathrm{e}^{-(\|X_{i(g)}-X_{j(g)}\|/\sigma_g)}(\|X_{i(g)}-X_{j(g)}\|/\sigma_g^2)$, 则有

$$\begin{aligned}
&P\left(\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\omega}_g^{ij}-E\omega_g^{ij}\right|\geqslant\varepsilon/3\right)\\
&\leqslant P\left(\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{\omega}_g^{ij}-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_g^{ij}\right|\geqslant\varepsilon/6\right)+P\left(\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_g^{ij}-E\omega_g^{12}\right|\geqslant\varepsilon/6\right)\\
&\leqslant 3\exp\{-nC_5\varepsilon^2\}+2\exp\{-n\varepsilon^2/144\}\\
&\leqslant 5\exp\{-nC_6\varepsilon^2\},
\end{aligned}\tag{A.44}$$

其中 $C_5=\min\{\check{m}^2/(576(E\mathcal{U}+1)^2),\tilde{m}^2\}$, $C_6=\min\{C_5,1/144\}$. 此外,

$$\begin{aligned}
P\left(\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}Y_i^rY_j^r-E[Y_1^rY_2^r]\right|\geqslant\varepsilon/3\right)&\leqslant P\left(\left|\frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}Y_i^rY_j^r-E[Y_1^rY_2^r]\right|\geqslant\varepsilon/6\right)\\
&\leqslant 2\exp\{-n\varepsilon^2/36\},
\end{aligned}\tag{A.45}$$

结合 (A.44) 和 (A.45), 根据引理 A.2, 不难得到

$$\begin{aligned}
P\left(\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\hat{\omega}_g^{ij}Y_k^rY_l^r-E[Y_3^rY_4^r\omega_g^{12}]\right|\geqslant\varepsilon/3\right)&=P(|\psi_{n3}^g-\psi_3^g|\geqslant\varepsilon/3)\\
&\leqslant 15\exp\{-nC_7\varepsilon^2\},
\end{aligned}\tag{A.46}$$

其中 C_7 为某个大于 0 的数. 令 $\zeta=\min\{C_2,C_4,C_7\}$, 综合 (A.39), (A.43) 和 (A.46), 可得

$$P(|(\mathrm{RMDD}_{m,n}^{g,r})^2-(\mathrm{RMDD}_{m}^{g,r})^2|\geqslant\varepsilon)\leqslant 25\exp\{-n\zeta\varepsilon^2\}.\tag{A.47}$$

证毕. □

定理 3.1 的证明 为了证明定理 3.1 的第一部分, 由引理 A.2 可知, 只需要证明 $\widehat{W}_g$ 的分子、分母部分均分别满足指数不等式即可. 根据性质 2.3, 对于 $\widehat{W}_g$ 的分子部分, 可以得到

$$\mathrm{RSDCov}_m(X_{(g)},Y)=\sum_{r=1}^{R}(\mathrm{RMDD}_m^{g,r})^2/p_r,\quad \mathrm{RSDCov}_{m,n}(X_{(g)},Y)=\sum_{r=1}^{R}(\mathrm{RMDD}_{m,n}^{g,r})^2/\hat{p}_r.\tag{A.48}$$

对于任意的 $\varepsilon>0$, 有

$$\begin{aligned}
&P(|\mathrm{RSDCov}_{m,n}(X_{(g)},Y)-\mathrm{RSDCov}_m(X_{(g)},Y)|\geqslant\varepsilon)\\
&\leqslant\underbrace{P\left(\left|\sum_{r=1}^{R}(\mathrm{RMDD}_{m,n}^{g,r})^2/\hat{p}_r-\sum_{r=1}^{R}(\mathrm{RMDD}_{m}^{g,r})^2/\hat{p}_r\right|\geqslant\varepsilon/2\right)}_{I_3}\\
&\quad+\underbrace{P\left(\left|\sum_{r=1}^{R}(\mathrm{RMDD}_{m}^{g,r})^2/\hat{p}_r-\sum_{r=1}^{R}(\mathrm{RMDD}_{m}^{g,r})^2/p_r\right|\geqslant\varepsilon/2\right)}_{I_4}.
\end{aligned}\tag{A.49}$$

对于 I_3, 有

$$I_3=P\left(\left|\sum_{r=1}^{R}\frac{1}{\hat{p}_r}[(\mathrm{RMDD}_{m,n}^{g,r})^2-(\mathrm{RMDD}_{m}^{g,r})^2]\right|\geqslant\varepsilon/2\right)$$

$$
\begin{aligned}
&\leqslant P\bigg(\frac{R}{\min_r \hat{p}_r}\max_r|(\mathrm{RMDD}_{m,n}^{g,r})^2-(\mathrm{RMDD}_m^{g,r})^2|\geqslant \varepsilon/2\bigg)\\
&\leqslant P\bigg(\frac{R}{\min_r \hat{p}_r}\max_r|(\mathrm{RMDD}_{m,n}^{g,r})^2-(\mathrm{RMDD}_m^{g,r})^2|\geqslant \varepsilon/2,\min_r \hat{p}_r\geqslant\frac{a_1}{2R}\bigg)+P\bigg(\min_r \hat{p}_r<\frac{a_1}{2R}\bigg)\\
&\leqslant P\bigg(\max_r|(\mathrm{RMDD}_{m,n}^{g,r})^2-(\mathrm{RMDD}_m^{g,r})^2|\geqslant\frac{a_1\varepsilon}{4R^2}\bigg)+P\bigg(\max_r|\hat{p}_r-p_r|\geqslant\frac{a_1}{2R}\bigg)\\
&\leqslant 25R\exp\bigg\{-\frac{n\zeta a_1^2\varepsilon^2}{16R^4}\bigg\}+2R\exp\bigg\{-\frac{na_1^2}{8R(a_1+a_2)}\bigg\},
\end{aligned}\tag{A.50}
$$

其中根据条件 (C_2)、引理 A.3 和 A.5, 最后两个不等号成立. 对于 I_4, 有

$$
\begin{aligned}
I_4&=P\bigg(\bigg|\sum_{r=1}^R(\mathrm{RMDD}_m^{g,r})^2/\hat{p}_r-\sum_{r=1}^R(\mathrm{RMDD}_m^{g,r})^2/p_r\bigg|\geqslant\varepsilon/2\bigg)\\
&=P\bigg(\bigg|\sum_{r=1}^R\frac{p_r-\hat{p}_r}{p_r\hat{p}_r}(\mathrm{RMDD}_m^{g,r})^2\bigg|\geqslant\varepsilon/2\bigg)\leqslant P\bigg(R\frac{\max_r|p_r-\hat{p}_r|}{p_r\min_r\hat{p}_r}\max_r(\mathrm{RMDD}_m^{g,r})^2\geqslant\varepsilon/2\bigg)\\
&\leqslant P\bigg(R\frac{\max_r|p_r-\hat{p}_r|}{p_r\min_r\hat{p}_r}\max_r(\mathrm{RMDD}_m^{g,r})^2\geqslant\varepsilon/2,\min_r\hat{p}_r\geqslant\frac{a_1}{2R}\bigg)+P\bigg(\min_r\hat{p}_r<\frac{a_1}{2R}\bigg)\\
&\leqslant P\bigg(\max_r|\hat{p}_r-p_r|\geqslant\frac{a_1^2\varepsilon}{16R^3}\bigg)+P\bigg(\max_r|\hat{p}_r-p_r|\geqslant\frac{a_1}{2R}\bigg)\\
&\leqslant 2R\exp\bigg\{-\frac{na_1^4\varepsilon^2}{512a_2R^5+16a_1^2\varepsilon R^3}\bigg\}+2R\exp\bigg\{-\frac{na_1^2}{8R(a_1+a_2)}\bigg\},
\end{aligned}\tag{A.51}
$$

其中 $\max_r(\mathrm{RMDD}_m^{g,r})^2\leqslant 4$. 综合 (A.49)–(A.51), 则有

$$
\begin{aligned}
&P(|\mathrm{RSDCov}_{m,n}(X_{(g)},Y)-\mathrm{RSDCov}_m(X_{(g)},Y)|\geqslant\varepsilon)\\
&\leqslant 25R\exp\bigg\{-\frac{n\zeta a_1^2\varepsilon^2}{16R^4}\bigg\}+4R\exp\bigg\{-\frac{na_1^2}{8R(a_1+a_2)}\bigg\}+2R\exp\bigg\{-\frac{na_1^4\varepsilon^2}{512a_2R^5+16a_1^2\varepsilon R^3}\bigg\}\\
&=25R\exp\bigg\{-\frac{nc_3\varepsilon^2}{R^4}\bigg\}+4R\exp\bigg\{-\frac{nc_4}{R}\bigg\}+2R\exp\bigg\{-\frac{n\varepsilon^2}{c_5R^5+c_6\varepsilon R^3}\bigg\}\\
&\leqslant O\bigg(R\exp\bigg\{-\frac{nc_3\varepsilon^2}{R^4}\bigg\}+R\exp\bigg\{-\frac{nc_4}{R}\bigg\}+R\exp\bigg\{-\frac{n\varepsilon^2}{c_5R^5+c_6\varepsilon R^3}\bigg\}\bigg),
\end{aligned}\tag{A.52}
$$

其中 $c_3=\zeta a_1^2/16$, $c_4=a_1^2/(8(a_1+a_2))$, $c_5=512a_2/a_1^4$, $c_6=16/a_1^2$. 与证明 $\widehat{W}_g$ 的分子部分满足指数不等式类似, 结合引理 A.2, 不难证明 $\widehat{W}_g$ 的分母部分满足

$$
P(|\mathrm{dvar}_m(X_{(g)})-\mathrm{dvar}_{m,n}(X_{(g)})|\geqslant\varepsilon)\leqslant O(\exp\{-nc_7\varepsilon^2\}),\tag{A.53}
$$

其中 c_7 为某个大于 0 的常数. 结合 (A.52) 和 (A.53), 再次利用引理 A.2, 可以得到

$$
\begin{aligned}
&P\bigg(\bigg|\frac{\mathrm{RSDCov}_{m,n}(X_{(g)},Y)}{\mathrm{dvar}_{m,n}(X_{(g)})}-\frac{\mathrm{RSDCov}_m(X_{(g)},Y)}{\mathrm{dvar}_m(X_{(g)})}\bigg|\geqslant\varepsilon\bigg)\\
&\quad\leqslant O\bigg(\exp\bigg\{-nd_1\bigg\}+R\exp\bigg\{\frac{-nd_2\varepsilon^2}{R^4}\bigg\}+R\exp\bigg\{\frac{-nc_4}{R}\bigg\}
\end{aligned}
$$

$$+ R\exp\left\{\frac{-n\varepsilon^2}{d_3R^5+d_4\varepsilon R^3}\right\}+\exp\left\{\frac{-nd_5\varepsilon^2}{R}\right\}\Bigg),$$

其中 d_1,d_2,d_3,d_4 和 d_5 为某些大于 0 的常数. 进一步有

$$\begin{aligned}P(|\widehat{W}_g-W_g|\geqslant\varepsilon)&=P\left(\left|\frac{\mathrm{RSDCov}_{m,n}(X_{(g)},Y)}{\mathrm{dvar}_{m,n}(X_{(g)})\sqrt{R-1}}-\frac{\mathrm{RSDCov}_m(X_{(g)},Y)}{\mathrm{dvar}_m(X_{(g)})\sqrt{R-1}}\right|\geqslant\varepsilon\right)\\&\leqslant O\left(R\exp\left\{\frac{-nd_2\varepsilon^2}{R^3}\right\}+R\exp\left\{\frac{-n\varepsilon^2}{d_3R^4+d_4\varepsilon R^2}\right\}\right)\\&=O\left(R\exp\left\{-\frac{n\varsigma\varepsilon^2}{R^4}\right\}\right),\end{aligned}\tag{A.54}$$

其中 $\varsigma>0$ 为某个常数. 令 $\varepsilon=cn^{-\nu}$ 且 $0\leqslant\nu<1/2$, $R:=R_n=O(n^\kappa)$ 且 $\kappa\geqslant0$, $\xi=\varsigma c^2$, 则有

$$P(|\widehat{W}_g-W_g|\geqslant cn^{-\nu})\leqslant O(n^\kappa\exp\{-\varsigma c^2n^{1-2\nu-4\kappa}\})=O(n^\kappa\exp\{-\xi n^{1-2\nu-4\kappa}\}).\tag{A.55}$$

根据 (A.55), 可以进一步得到

$$P\Big(\max_{1\leqslant g\leqslant G}|\widehat{W}_g-W_g|\geqslant cn^{-\nu}\Big)\leqslant\sum_{g=1}^{G}P(|\widehat{W}_g-W_g|\geqslant cn^{-\nu})=O(Gn^\kappa\exp\{-\xi n^{1-2\nu-4\kappa}\}).\tag{A.56}$$

如此便证明了定理 3.1 的第一部分. 对于该定理的第二部分, 根据条件 (C_3), 有

$$\begin{aligned}P(\mathcal{A}\subset\hat{\mathcal{A}})&\geqslant P\Big(\min_{g\in\mathcal{A}}\widehat{W}_g\geqslant cn^{-\nu}\Big)=P\Big(\min_{g\in\mathcal{A}}W_g-\min_{g\in\mathcal{A}}\widehat{W}_g\leqslant\min_{g\in\mathcal{A}}W_g-cn^{-\nu}\Big)\\&\geqslant P\Big(|\min_{g\in\mathcal{A}}W_g-\min_{g\in\mathcal{A}}\widehat{W}_g|\leqslant cn^{-\nu}\Big)\geqslant P\Big(\max_{g\in\mathcal{A}}|\widehat{W}_g-W_g|\leqslant cn^{-\nu}\Big)\\&=1-P\Big(\bigcup_{g\in\mathcal{A}}\{|\widehat{W}_g-W_g|\geqslant cn^{-\nu}\}\Big)\geqslant1-\sum_{g\in\mathcal{A}}P(|\widehat{W}_g-W_g|\geqslant cn^{-\nu})\\&\geqslant1-O(s_nn^\kappa\exp\{-\xi n^{1-2\nu-4\kappa}\}),\end{aligned}\tag{A.57}$$

其中 s_n 表示集合 $\mathcal{A}$ 中元素的个数. 至此, 定理 3.1 证毕. □

定理 3.2 的证明 结合 (A.55) 和条件 (C_4), 则有

$$\begin{aligned}P\Big(\min_{g\in\mathcal{A}}\widehat{W}_g>\max_{g\in\mathcal{I}}\widehat{W}_g\Big)&\geqslant1-P\Big(\max_{g\in\mathcal{I}}\widehat{W}_g-\min_{g\in\mathcal{A}}\widehat{W}_g+\min_{g\in\mathcal{A}}W_g-\max_{g\in\mathcal{I}}W_g\geqslant\eta\Big)\\&\geqslant1-P\Big(\max_{g\in\mathcal{I}}|\widehat{W}_g-W_g|+\max_{g\in\mathcal{A}}|\widehat{W}_g-W_g|\geqslant\eta\Big)\\&\geqslant1-P\Big(\max_{1\leqslant g\leqslant G}|\widehat{W}_g-W_g|\geqslant\eta/2\Big)\geqslant1-\sum_{g=1}^{G}P\Big(|\widehat{W}_g-W_g|\geqslant\eta/2\Big)\\&\geqslant1-O\left(GR\exp\left\{-\frac{n\varsigma\eta^2}{4R^4}\right\}\right)\overset{令\ \zeta=\varsigma/4}{=\!=\!=}1-O\left(GR\exp\left\{-\frac{n\zeta\eta^2}{R^4}\right\}\right).\end{aligned}\tag{A.58}$$

至此, 定理 3.2 证毕. □

附录 B 额外的数值模拟以及理论证明

附录 B.1 RSDSIS 等筛选方法在例 B.1 和 B.2 中的表现

例 B.1 该例来自文献 [27, 例 3.3], 利用 RSD-SIS 去处理预测变量为分类变量、响应变量为连续变量的超高维特征筛选问题. 该问题的研究背景与全基因组关联学习有关, 关于其更详细的阐述可参见文献 [6]. 令 Z_{ij} 表示为第 i 位受试者第 j 个单核苷酸多态性 (SNP) 显性效应的指标, $i=1,\ldots,n$, $j=1,\ldots,p$, 其生成方式如下:

$$Z_{ij}=\begin{cases}1, & X_{ij}<qt_1,\\ 0, & qt_1\leqslant X_{ij}<qt_3,\\ -1, & X_{ij}\geqslant qt_3,\end{cases}$$

其中 $\boldsymbol{x}_i=(X_{i1},\ldots,X_{ip})^{\mathrm{T}}\sim N(0,\Sigma)$ 且 Σ 的第 k 行第 l 列个元素 $\sigma_{kl}=0.5^{|k-l|}, k,l=1,\ldots,p$, (qt_1,qt_3) 分别为标准正态分布 $N(0,1)$ 的第 25% 和 75% 分位数. 响应变量 Y 表示某种特征或疾病, 其满足

$$Y=\beta_1Z_1+\beta_2Z_2+2\beta_3Z_{10}+2\beta_4Z_{20}-2\beta_5|Z_{100}|+\epsilon,$$

其中 $\beta_s=(-1)^U(a+|\mathcal{Z}_s|)$, $s=1,\ldots,5$, $a=2\log n/\sqrt{n}, U\sim$ Bernoulli(0.4), $\mathcal{Z}_1,\mathcal{Z}_2,\mathcal{Z}_3,\mathcal{Z}_4$ 和 $\mathcal{Z}_5$ 独立同分布于 $N(0,1)$, 误差项 $\epsilon\sim N(0,1)$ 或者 $\epsilon\sim C(0,1)$. 在例 B.1 中, 只有 5 个 SNP (即重要变量 Z_1,Z_2,Z_{10},Z_{20} 和 Z_{100}) 对响应变量 Y 真正有贡献. 令 $n=200$ 或者 $n=300$, $p=2000$, 基于 600 次重复实验得到的模拟结果如表 B1 和 B2 所示. 容易看到, 随着样本量由 200 增加到 300, 所有筛选方法的表现均有显著提高. 当 $n=300$, $p=2000$, $\epsilon\sim C(0,1)$ 时, MV-SIS 的表现最佳, 其次是 RSD-SIS 和

表 B1 在例 B.1 中, 对于 $n=200$, $p=2000$, $\mathcal{M}$, $\mathcal{S}$, $\mathcal{P}_a$ 以及 $\mathcal{P}_1^s(Z_1)$, $\mathcal{P}_2^s(Z_2)$, $\mathcal{P}_3^s(Z_{10})$, $\mathcal{P}_4^s(Z_{20})$ 和 $\mathcal{P}_5^s(Z_{100})$ 的大小

ϵ	方法	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(Z_1)$	$\mathcal{P}_2^s(Z_2)$	$\mathcal{P}_3^s(Z_{10})$	$\mathcal{P}_4^s(Z_{20})$	$\mathcal{P}_5^s(Z_{100})$
$N(0,1)$	CI-SIS	13.00	40.31	0.683	0.910	0.898	0.982	0.983	0.845
	DC-SIS	15.00	34.28	0.702	0.875	0.892	0.975	0.980	0.895
	GN-SIS	9.00	31.50	0.722	0.915	0.900	0.988	0.972	0.878
	FKF-SIS	24.50	68.38	0.573	0.865	0.890	0.973	0.982	0.733
	MV-SIS	14.00	49.11	0.682	0.897	0.910	0.973	0.982	0.830
	PSIS	8.00	17.05	0.772	0.952	0.942	0.987	0.993	0.848
	SD-SIS	11.00	22.24	0.762	0.928	0.930	0.980	0.985	0.890
	RSD-SIS	25.50	107.12	0.557	0.850	0.873	0.962	0.955	0.797
$C(0,1)$	CI-SIS	45.00	127.13	0.452	0.812	0.805	0.937	0.945	0.732
	DC-SIS	124.50	359.15	0.285	0.657	0.665	0.868	0.870	0.717
	GN-SIS	58.00	217.94	0.430	0.762	0.765	0.902	0.937	0.753
	FKF-SIS	82.00	183.28	0.357	0.780	0.820	0.933	0.942	0.617
	MV-SIS	55.50	162.71	0.438	0.815	0.792	0.953	0.937	0.732
	PSIS	1375.50	706.45	0.030	0.233	0.277	0.363	0.372	0.167
	SD-SIS	103.50	307.82	0.343	0.692	0.720	0.903	0.898	0.655
	RSD-SIS	107.00	269.27	0.342	0.742	0.748	0.918	0.918	0.690

表 B2 在例 B.1 中, 对于 $n = 300$, $p = 2000$, $\mathcal{M}$, $\mathcal{S}$, $\mathcal{P}_a$ 以及 $\mathcal{P}_1^s(Z_1)$, $\mathcal{P}_2^s(Z_2)$, $\mathcal{P}_3^s(Z_{10})$, $\mathcal{P}_4^s(Z_{20})$ 和 $\mathcal{P}_5^s(Z_{100})$ 的大小

ϵ	方法	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(Z_1)$	$\mathcal{P}_2^s(Z_2)$	$\mathcal{P}_3^s(Z_{10})$	$\mathcal{P}_4^s(Z_{20})$	$\mathcal{P}_5^s(Z_{100})$
$N(0,1)$	CI-SIS	7.00	11.12	0.853	0.948	0.977	0.995	0.997	0.925
	DC-SIS	7.00	10.38	0.855	0.948	0.953	0.992	0.987	0.948
	GN-SIS	7.00	8.15	0.882	0.963	0.957	0.993	0.995	0.958
	FKF-SIS	10.00	22.61	0.795	0.952	0.972	0.998	0.998	0.852
	MV-SIS	7.00	8.90	0.857	0.960	0.965	0.998	0.993	0.927
	PSIS	7.00	6.67	0.908	0.992	0.985	0.998	1.000	0.932
	SD-SIS	7.00	7.41	0.878	0.963	0.967	0.990	0.995	0.945
	RSD-SIS	8.00	20.20	0.782	0.935	0.948	0.985	0.980	0.893
$C(0,1)$	CI-SIS	19.00	67.46	0.647	0.927	0.900	0.975	0.968	0.810
	DC-SIS	36.00	163.83	0.543	0.825	0.818	0.920	0.918	0.840
	GN-SIS	22.00	104.15	0.620	0.868	0.885	0.958	0.955	0.813
	FKF-SIS	23.50	83.21	0.617	0.898	0.888	0.980	0.962	0.803
	MV-SIS	12.50	49.85	0.700	0.912	0.902	0.970	0.985	0.873
	PSIS	1378.50	641.40	0.035	0.282	0.275	0.400	0.395	0.215
	SD-SIS	34.00	172.91	0.562	0.835	0.847	0.933	0.938	0.810
	RSD-SIS	20.00	87.29	0.647	0.888	0.903	0.973	0.970	0.832

CI-SIS. 此外, 不难注意到当 $n = 200$ 或者 300, $p = 2000$, $\epsilon \sim N(0,1)$ 时, PSIS 的表现最佳, 而 RSD-SIS 的表现稍弱于其对比方法. 这可能很大程度上与 RSD-SIS 中窗宽参数的选择有关, 即基于中值启发式窗宽下的 RSD-SIS 虽然在很多情形下均有很好的表现, 但并不一定总是相对最优的.

例 B.2 该例类似文献 [27, 例 3.4], 其中分类响应变量 Y 的类别数 $R = 3$, $\boldsymbol{x} = (X_{(1)}^{\mathrm{T}}, \ldots, X_{(G)}^{\mathrm{T}})^{\mathrm{T}}$ 且 $X_{(g)} = (X_{3(g-1)+1}, X_{3(g-1)+2}, X_{3g})^{\mathrm{T}}$, $g = 1, \ldots, G$. 该例假定只有 2 个重要的组预测变量 $X_{(1)} = (X_1, X_2, X_3)^{\mathrm{T}}$ 和 $X_{(2)} = (X_4, X_5, X_6)^{\mathrm{T}}$. 对于第 i 个观测, 令 $p_r = P(Y_i = L_r)$ 表示 $Y_i \in L_r$ 的概率, $r = 1, 2, 3$. 这里, 分别考虑 Y 的 3 种不同分布设计: (1) 平衡设计, $p = (p_1, p_2, p_3) = (1/3, 1/3, 1/3)$; (2) 轻度不平衡设计, $p = (p_1, p_2, p_3) = (3/12, 4/12, 5/12)$; (3) 极度不平衡设计, $p = (p_1, p_2, p_3) = (0.1, 0.3, 0.6)$. 给定 $Y_i = L_r$, 基于位置模型 $\boldsymbol{x}_i = \boldsymbol{\mu}_G^r + \boldsymbol{\epsilon}_i$ 来生成第 i 个预测向量 $\boldsymbol{x}_i$, 其中 $\boldsymbol{\epsilon}_i \in R^{3G}$ 且其每个分量独立同分布于变量 ϵ, $\boldsymbol{\mu}_G^r = (\boldsymbol{\mu}_1^{(r)}, \boldsymbol{\mu}_2^{(r)}, 0, \ldots, 0)^{\mathrm{T}} \in R^{3G}$. 令 $\boldsymbol{\mu}^{(1)} = (1, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\mu}^{(2)} = (0, 1, 0)^{\mathrm{T}}$, $\boldsymbol{\mu}^{(3)} = (0, 0, 1)^{\mathrm{T}}$, 考虑如下两种情形: (a)$\boldsymbol{\mu}_1^{(r)} = \boldsymbol{\mu}_2^{(r)} = 1.5\boldsymbol{\mu}^{(r)}$; (b) $\boldsymbol{\mu}_1^{(r)} = 2\boldsymbol{\mu}^{(r)}$, $\boldsymbol{\mu}_2^{(r)} = \boldsymbol{\mu}^{(r)}$. 不难看出, 在情形 (a) 中, $X_{(1)}$ 和 $X_{(2)}$ 的信号强度一样; 而在情形 (b) 中, $X_{(2)}$ 的信号强度弱于 $X_{(1)}$. 令 $n = 200, p = 3000$ (即 $G = 1000$), $\epsilon \sim N(0,1)$ 或者 $\epsilon \sim C(0,1)$, 基于 600 次重复实验, RSD-SIS 以及其对比方法在不同情形下的模拟结果见表 B3. 容易看到, 当预测向量的有限矩假设满足时, RSD-SIS 的表现与 GN-SIS, SD-SIS 和 DC-SIS 差不多; 而当预测向量的有限矩假设不成立时, RSD-SIS 的表现依然优秀, 且明显优于 GN-SIS, SD-SIS 和 DC-SIS.

附录 B.2 当 ϵ 服从 $t(3)$, $t(4)$ 或 $t(5)$ 分布时, PSIS 在例 4.2 和 B.1 中的表现

当预测变量的二阶矩有限时, 为了进一步评价 PSIS 的表现, 本文在例 4.2 和 B.1 中令 ϵ 分别服从 $t(3)$, $t(4)$ 或 $t(5)$ 分布, 具体的模拟结果如表 B4–B8 所示. 例 4.2 中表 B4–B6 表明, 与预测变量服

表 B3 在例 B.2 中, 对于 $(R,n,p)=(3,200,3000)$, $\mathcal{M}$, $\mathcal{S}$, $\mathcal{P}_a$, $\mathcal{P}_1^s$ 和 $\mathcal{P}_2^s$ 的大小, 其中 $\mathcal{P}_1^s$ 和 $\mathcal{P}_2^s$ 分别表示 $\mathcal{P}_1^s(\boldsymbol{X}_{(1)})$ 和 $\mathcal{P}_2^s(\boldsymbol{X}_{(2)})$

情形	p_r	方法	$\epsilon\sim N(0,1)$					$\epsilon\sim C(0,1)$				
			$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$
(a)	平衡设计	DC-SIS	2.00	0.0	1.000	1.000	1.000	131.50	472.57	0.332	0.585	0.558
		GN-SIS	2.00	0.0	1.000	1.000	1.000	26.50	71.16	0.553	0.765	0.742
		SD-SIS	2.00	0.0	1.000	1.000	1.000	150.50	425.13	0.333	0.558	0.595
		RSD-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	1.000	1.000	1.000
	轻度不平衡设计	DC-SIS	2.00	0.0	1.000	1.000	1.000	113.00	338.40	0.343	0.605	0.587
		GN-SIS	2.00	0.0	1.000	1.000	1.000	22.00	56.52	0.598	0.775	0.762
		SD-SIS	2.00	0.0	1.000	1.000	1.000	179.50	333.70	0.300	0.570	0.542
		RSD-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	1.000	1.000	1.000
	极度不平衡设计	DC-SIS	2.00	0.0	1.000	1.000	1.000	236.00	466.27	0.235	0.495	0.488
		GN-SIS	2.00	0.0	1.000	1.000	1.000	30.00	77.28	0.483	0.678	0.682
		SD-SIS	2.00	0.0	1.000	1.000	1.000	270.50	433.84	0.110	0.360	0.298
		RSD-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	0.998	1.000	0.998
(b)	平衡设计	DC-SIS	2.00	0.0	1.000	1.000	1.000	192.00	462.38	0.273	0.747	0.370
		GN-SIS	2.00	0.0	1.000	1.000	1.000	96.00	151.41	0.302	0.905	0.337
		SD-SIS	2.00	0.0	1.000	1.000	1.000	223.00	418.09	0.228	0.728	0.305
		RSD-SIS	2.00	0.0	1.000	1.000	1.000	2.00	0.0	0.962	1.000	0.962
	轻度不平衡设计	DC-SIS	2.00	0.0	1.000	1.000	1.000	261.50	439.40	0.200	0.745	0.292
		GN-SIS	2.00	0.0	1.000	1.000	1.000	84.00	140.29	0.320	0.917	0.350
		SD-SIS	2.00	0.0	1.000	1.000	1.000	264.50	470.90	0.223	0.697	0.327
		RSD-SIS	2.00	0.0	1.000	1.000	1.000	3.00	4.48	0.937	1.000	0.937
	极度不平衡设计	DC-SIS	2.00	0.0	1.000	1.000	1.000	307.50	439.77	0.170	0.683	0.250
		GN-SIS	2.00	0.0	1.000	1.000	1.000	109.50	173.09	0.290	0.898	0.318
		SD-SIS	2.00	0.0	1.000	1.000	1.000	381.00	441.25	0.083	0.582	0.135
		RSD-SIS	2.00	0.0	1.000	1.000	1.000	5.00	11.87	0.882	1.000	0.882

表 B4 在例 4.2 中, 当 $(R,n,p)=(2,40,2000)$ 时, PSIS 的表现

方法	ϵ 的分布	平衡设计					非平衡设计				
		$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$
PSIS	$t(5)$	2.00	0.0	1.000	1.000	1.000	2.00	0.0	0.998	0.998	1.000
	$t(4)$	2.00	0.0	0.997	1.000	0.997	2.00	0.0	0.993	0.997	0.997
	$t(3)$	2.00	0.0	0.957	0.983	0.972	4.00	2.22	0.913	0.955	0.953

从 $t(2)$ 或 $C(0,1)$ 型分布相比, PSIS 在预测变量服从 $t(3)$, $t(4)$ 或 $t(5)$ 型分布时的表现显然更好. 而例 B.1 中表 B7 和 B8 再次表明, 虽然 PSIS 在一些非常厚尾的预测变量 (如 $C(0,1)$) 下表现不理想, 但是其在一些轻度厚尾 (如 $t(3)$, $t(4)$, $t(5)$) 或者轻尾 (如 $N(0,1)$) 的预测变量下表现较好.

表 B5 在例 4.2 的平衡设计中, 当 $(R, n, p) = (10, 200, 2000)$ 时, PSIS 的表现

方法	ϵ 的分布	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{P}_3^s(X_3)$	$\mathcal{P}_4^s(X_4)$	$\mathcal{P}_5^s(X_5)$	$\mathcal{P}_6^s(X_6)$	$\mathcal{P}_7^s(X_7)$	$\mathcal{P}_8^s(X_8)$	$\mathcal{P}_9^s(X_9)$	$\mathcal{P}_{10}^s(X_{10})$
PSIS	$t(5)$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$t(4)$	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000
	$t(3)$	0.890	0.987	0.993	0.987	0.988	0.987	0.993	0.990	0.982	0.992	0.982

表 B6 在例 4.2 的非平衡设计中, 当 $(R, n, p) = (10, 200, 2000)$ 时, PSIS 的表现

方法	ϵ 的分布	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{P}_3^s(X_3)$	$\mathcal{P}_4^s(X_4)$	$\mathcal{P}_5^s(X_5)$	$\mathcal{P}_6^s(X_6)$	$\mathcal{P}_7^s(X_7)$	$\mathcal{P}_8^s(X_8)$	$\mathcal{P}_9^s(X_9)$	$\mathcal{P}_{10}^s(X_{10})$
PSIS	$t(5)$	0.993	0.998	0.997	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$t(4)$	0.995	0.997	1.000	0.998	1.000	1.000	0.998	1.000	1.000	1.000	1.000
	$t(3)$	0.780	0.962	0.967	0.972	0.977	0.968	0.978	0.973	0.970	0.987	0.988

表 B7 在例 B.1 中, 当 $n = 200$, $p = 2000$ 时, PSIS 的表现

方法	ϵ 的分布	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(Z_1)$	$\mathcal{P}_2^s(Z_2)$	$\mathcal{P}_3^s(Z_{10})$	$\mathcal{P}_4^s(Z_{20})$	$\mathcal{P}_5^s(Z_{100})$
PSIS	$t(5)$	20.01	10.0	0.768	0.955	0.948	0.993	0.993	0.847
	$t(4)$	30.39	10.0	0.725	0.933	0.937	0.992	0.990	0.822
	$t(3)$	30.02	11.0	0.722	0.945	0.940	0.990	0.990	0.813

表 B8 在例 B.1 中, 当 $n = 300$, $p = 2000$ 时, PSIS 的表现

方法	ϵ 的分布	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(Z_1)$	$\mathcal{P}_2^s(Z_2)$	$\mathcal{P}_3^s(Z_{10})$	$\mathcal{P}_4^s(Z_{20})$	$\mathcal{P}_5^s(Z_{100})$
PSIS	$t(5)$	6.12	7.0	0.893	0.987	0.978	1.000	1.000	0.922
	$t(4)$	6.67	7.0	0.882	0.980	0.980	0.995	0.998	0.917
	$t(3)$	8.34	7.0	0.875	0.978	0.977	0.995	0.998	0.912

附录 B.3 在定义 2.4 中, RSDCor$(\boldsymbol{z}, Y)$ 的位置尺度不变性证明以及不同的 $\boldsymbol{\sigma}$ 对 RSD-SIS 的影响

欲证明 RSDCor$(\boldsymbol{z}, Y)$ 具有位置尺度不变性, 即证明对于任意的常向量 $\boldsymbol{a} \in \mathbb{R}^d$、非零常数 b 和正交矩阵 $\boldsymbol{H} \in \mathbb{R}^{d\times d}$, 有

$$\text{RSDCor}(\boldsymbol{a} + b\boldsymbol{H}\boldsymbol{z}, Y) = \text{RSDCor}(\boldsymbol{z}, Y). \tag{B.1}$$

(B.1) 的证明 对于 (B.1) 的左边, 根据 (2.14), 有

$$\text{RSDCor}(\boldsymbol{a} + b\boldsymbol{H}\boldsymbol{z}, Y) = \frac{\text{RSDCov}_m(\boldsymbol{a} + b\boldsymbol{H}\boldsymbol{z}, Y)}{\text{dvar}_m(\boldsymbol{a} + b\boldsymbol{H}\boldsymbol{z})\sqrt{R-1}}. \tag{B.2}$$

令 $\mathcal{E}_{\boldsymbol{z}} = \|\boldsymbol{a} + b\boldsymbol{H}\boldsymbol{z}_1 - (\boldsymbol{a} + b\boldsymbol{H}\boldsymbol{z}_2)\| = |b|\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|$, 则 $\text{RSDCov}_m(\boldsymbol{a} + b\boldsymbol{H}\boldsymbol{z}, Y)$ 和 $\text{dvar}_m(\boldsymbol{a} + b\boldsymbol{H}\boldsymbol{z})$ 的窗宽参数 $\sigma_{\mathcal{E}_{\boldsymbol{z}}}$ 满足

$$\sigma_{\mathcal{E}_{\boldsymbol{z}}} = \text{Med}(\mathcal{E}_{\boldsymbol{z}}) = \text{Med}(|b|T_{\boldsymbol{z}}) = |b|F^{-1}(1/2), \tag{B.3}$$

其中 $T_{\boldsymbol{z}} = \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|$ 且 $F(\cdot)$ 表示 $T_{\boldsymbol{z}}$ 的累积分布函数. 为了叙述简单, 令 $\sigma_{\boldsymbol{z}} = F^{-1}(1/2)$, 则 $\sigma_{\mathcal{E}_{\boldsymbol{z}}} = |b|\sigma_{\boldsymbol{z}}$

且根据 (B.2) 和 (2.14) 有

$$\mathrm{RSDCov}_m(\boldsymbol{a}+b\boldsymbol{H}\boldsymbol{z},Y)=E[\mathrm{e}^{-(\frac{|b|\|\boldsymbol{z}_1-\boldsymbol{z}_2\|}{|b|\sigma_z})}D(Y_1,Y_2)]=\mathrm{RSDCov}_m(\boldsymbol{z},Y), \tag{B.4}$$

$$\begin{aligned}\{\mathrm{dvar}_m(\boldsymbol{a}+b\boldsymbol{H}\boldsymbol{z})\}^2=&E\{\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma_{\boldsymbol{z}})-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma_{\boldsymbol{z}})}\}-2E\{\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma_{\boldsymbol{z}})-(\|\boldsymbol{z}_1-\boldsymbol{z}_3\|/\sigma_{\boldsymbol{z}})}\}\\&+E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma_{\boldsymbol{z}})}]E[\mathrm{e}^{-(\|\boldsymbol{z}_1-\boldsymbol{z}_2\|/\sigma_{\boldsymbol{z}})}]=\{\mathrm{dvar}_m(\boldsymbol{z})\}^2.\end{aligned} \tag{B.5}$$

由于 $\{\mathrm{dvar}_m(\boldsymbol{a}+b\boldsymbol{H}\boldsymbol{z})\}$ 和 $\{\mathrm{dvar}_m(\boldsymbol{z})\}$ 均非负, 故由 (B.5) 可知 $\{\mathrm{dvar}_m(\boldsymbol{a}+b\boldsymbol{H}\boldsymbol{z})\}=\{\mathrm{dvar}_m(\boldsymbol{z})\}$. 因此, 结合 (2.14) 和 (B.2)–(B.5), (B.1) 成立. □

回到 (2.11) 中的 $\mathrm{RSDCor}_\sigma(\boldsymbol{z},Y)$, 其中下标 σ 表示给定的窗宽参数. 与 $\mathrm{RSDCor}_\sigma(\boldsymbol{z},Y)$ 相比, 定义在 (2.14) 的 $\mathrm{RSDCor}(\boldsymbol{z},Y)$ 对于 $\boldsymbol{z}$ 保持位置尺度不变性, 其本质原因在于令窗宽参数 $\sigma=\mathrm{Med}(T_{\boldsymbol{z}})$. 因此, $\mathrm{RSDCor}(\boldsymbol{z},Y)$ 不受变量 $\boldsymbol{z}$ 的量纲影响, 这一点在实际数据处理中比较重要. 此外, 与固定窗宽参数相比, 这种基于成对样本中值距离的窗宽选择, 也使得 $\mathrm{RSDCor}(\boldsymbol{z},Y)$ 利用了相对更多的数据信息. 从这个角度看, 本文提出的 RSD-SIS 本质上是一种基于数据驱动的特征筛选程序. 下面, 通过例 4.1–4.3 来评价 RSD-SIS 在某些给定窗宽 (如 $\sigma=1,2,4,16,64,128,1024$) 以及基于数据驱动的窗宽 (即窗宽 σ 为成对样本的中值距离, 简记为 $\sigma=\sigma_m$) 下的表现, 具体的模拟结果如表 B9–B12 所示. 通过这些表可以发现: (1) RSD-SIS 的表现容易遭受窗宽参数的影响, 这一点在表 B10–B12 中更为明显; (2) RSD-SIS 在 σ_m 下的表现虽然并不一定总是最好, 但是表 B9–B12 均表明令 $\sigma=\sigma_m$ 确实是一种较好的窗宽选择. 这也是本文利用定义在 (2.14) 的 RSDCor 来构造特征筛选程序的重要动机.

附录 B.4 基于边际 Hilbert-Schmidt 独立性准则 (Hilbert-Schmidt independence criterion, HSIC) 的特征筛选程序 HSIC-SIS

虽然 HSIC 无法直接处理分类变量, 但可以对 Y 进行独热编码, 然后再对编码后的变量进行相关性度量. 以基于 Laplace 核的 HSIC 为例, 令 $\boldsymbol{z}\in\mathbb{R}^d$ $(d\geqslant 1)$ 为数值型变量, Y 为具有 R 个类别的分类变量. 通过对 Y 进行独热编码, 当 Y 为第 r 个类别时, 令 $\boldsymbol{y}=(0,\ldots,1,\ldots,0)^{\mathrm{T}}$, 即 $\boldsymbol{y}$ 的第 r 个分量为 1, 其他分量为 0. 进一步地, 定义

$$\begin{aligned}\mathrm{HSIC}(\boldsymbol{z},Y)=&E[\mathrm{e}^{-(|\boldsymbol{z}_1-\boldsymbol{z}_2|/\sigma_1)-(|\boldsymbol{y}_1-\boldsymbol{y}_2|/\sigma_2)}]-2E[\mathrm{e}^{-(|\boldsymbol{z}_1-\boldsymbol{z}_2|/\sigma_1)-(|\boldsymbol{y}_1-\boldsymbol{y}_3|/\sigma_2)}]\\&+E[\mathrm{e}^{-(|\boldsymbol{z}_1-\boldsymbol{z}_2|/\sigma_1)}]E[\mathrm{e}^{-(|\boldsymbol{y}_1-\boldsymbol{y}_2|/\sigma_2)}],\end{aligned} \tag{B.6}$$

表 B9 在例 4.1 中, RSD-SIS 在不同窗宽 σ 下的表现

n,p	RSD-SIS	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{P}_3^s(X_3)$	$\mathcal{P}_4^s(X_4)$	$\mathcal{P}_5^s(X_5)$
$n=80$ $p=2000$	$\sigma=1$	5.00	0.74	0.973	0.995	0.998	1.000	0.998	0.980
	$\sigma=2$	5.00	0.00	0.968	0.997	1.000	1.000	0.997	0.975
	$\sigma=4$	5.00	0.74	0.958	0.998	1.000	1.000	0.998	0.960
	$\sigma=16$	5.00	0.74	0.957	1.000	1.000	1.000	0.993	0.960
	$\sigma=64$	5.00	1.48	0.922	0.995	1.000	1.000	0.998	0.927
	$\sigma=128$	5.00	0.74	0.925	0.998	1.000	1.000	0.998	0.927
	$\sigma=1024$	5.00	1.48	0.920	1.000	1.000	1.000	1.000	0.920
	$\sigma=\sigma_m$	5.00	0.74	0.970	0.995	0.998	1.000	0.992	0.978

表 B10　在例 4.2 中, 当 $(R, n, p) = (2, 40, 2000)$ 时, RSD-SIS 在不同窗宽 σ 下的表现

ϵ	RSD-SIS	平衡设计					非平衡设计				
		$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$
$C(0,1)$	$\sigma = 1$	2.00	2.22	0.862	0.932	0.923	3.00	3.71	0.788	0.887	0.892
	$\sigma = 2$	2.00	0.74	0.925	0.960	0.963	2.00	2.22	0.842	0.907	0.925
	$\sigma = 4$	2.00	0.74	0.912	0.965	0.943	2.00	2.97	0.815	0.905	0.895
	$\sigma = 16$	2.00	2.22	0.865	0.927	0.927	4.00	8.90	0.710	0.838	0.845
	$\sigma = 64$	6.00	17.05	0.617	0.780	0.782	10.00	37.81	0.508	0.695	0.693
	$\sigma = 128$	12.50	40.77	0.455	0.647	0.712	24.00	73.57	0.358	0.615	0.585
	$\sigma = 1024$	23.00	135.10	0.413	0.667	0.627	36.00	158.64	0.323	0.570	0.538
	$\sigma = \sigma_m$	2.00	1.48	0.882	0.945	0.935	2.00	2.22	0.855	0.928	0.918

表 B11　在例 4.3 的情形 (b) 中, RSD-SIS 在不同窗宽 σ 下的表现

情形 R, n, p	RSD-SIS	平衡设计						非平衡设计					
		$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$
(b) $R = 3$ $n = 120$ $p = 2000$	$\sigma = 1$	12.00	21.50	0.678	0.878	0.868	0.887	17.00	26.87	0.638	0.763	0.897	0.932
	$\sigma = 2$	4.00	2.97	0.947	0.973	0.987	0.987	5.00	6.67	0.868	0.900	0.977	0.990
	$\sigma = 4$	3.00	0.74	0.980	0.992	0.993	0.995	4.00	2.22	0.942	0.952	0.993	0.997
	$\sigma = 16$	4.00	2.22	0.957	0.982	0.990	0.985	5.00	5.19	0.923	0.940	0.987	0.992
	$\sigma = 64$	8.00	10.38	0.813	0.925	0.930	0.940	11.00	15.57	0.733	0.833	0.923	0.960
	$\sigma = 128$	14.00	22.24	0.665	0.862	0.878	0.882	22.00	29.65	0.552	0.715	0.842	0.912
	$\sigma = 1024$	125.50	135.29	0.172	0.563	0.582	0.585	135.00	147.70	0.122	0.428	0.533	0.577
	$\sigma = \sigma_m$	3.00	2.97	0.948	0.987	0.983	0.978	4.00	3.71	0.923	0.933	0.995	0.995

表 B12　在例 4.3 的情形 (c) 中, RSD-SIS 在不同窗宽 σ 下的表现

情形 R, n, p	RSD-SIS	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$	$\mathcal{P}_4^s$	$\mathcal{P}_5^s$	$\mathcal{P}_6^s$	$\mathcal{P}_7^s$	$\mathcal{P}_8^s$	$\mathcal{P}_9^s$	$\mathcal{P}_{10}^s$
		非平衡设计												
(c) $R = 10$ $n = 300$ $p = 3000$	$\sigma = 1$	62.00	77.65	0.460	0.798	0.878	0.883	0.910	0.948	0.952	0.970	0.963	0.985	0.990
	$\sigma = 2$	13.00	10.56	0.895	0.960	0.968	0.982	0.988	0.997	1.000	0.998	1.000	1.000	1.000
	$\sigma = 4$	10.00	1.48	0.962	0.987	0.987	0.997	0.998	0.995	0.998	1.000	1.000	1.000	1.000
	$\sigma = 16$	10.00	0.74	0.992	0.995	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\sigma = 64$	10.00	1.48	0.975	0.987	0.993	0.997	1.000	0.998	1.000	1.000	1.000	1.000	1.000
	$\sigma = 128$	11.00	4.45	0.933	0.973	0.985	0.992	0.992	0.997	0.997	1.000	0.998	1.000	0.998
	$\sigma = 1024$	463.00	236.66	0.083	0.693	0.745	0.753	0.778	0.792	0.8000	0.797	0.810	0.838	0.837
	$\sigma = \sigma_m$	10.00	2.22	0.962	0.982	0.997	0.990	0.995	0.998	1.000	0.998	1.000	1.000	1.000

其中 $(\boldsymbol{z}_1, \boldsymbol{y}_1)$, $(\boldsymbol{z}_2, \boldsymbol{y}_2)$ 和 $(\boldsymbol{z}_3, \boldsymbol{y}_3)$ 独立同分布于 $(\boldsymbol{z}, \boldsymbol{y})$, σ_1 和 σ_2 为相应的窗宽参数. 基于 (B.6), 定义基于 HSIC 的相关系数为

$$\mathrm{HSICCor}(\boldsymbol{z}, Y) = \frac{\mathrm{HSIC}(\boldsymbol{z}, \boldsymbol{y})}{\sqrt{\mathrm{HSIC}(\boldsymbol{z}, \boldsymbol{z})}\sqrt{\mathrm{HSIC}(\boldsymbol{y}, \boldsymbol{y})}}. \tag{B.7}$$

令 $(\boldsymbol{z}_1,\boldsymbol{y}_1),(\boldsymbol{z}_2,\boldsymbol{y}_2),\ldots,(\boldsymbol{z}_n,\boldsymbol{y}_n)$ 为独立同分布于 $(\boldsymbol{z},\boldsymbol{y})$ 的随机样本, 则 $\mathrm{HSIC}(\boldsymbol{z},Y)$ 的样本估计为

$$\mathrm{HSIC}_n(\boldsymbol{z},Y)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega(\boldsymbol{z}_i,\boldsymbol{z}_j)\omega(\boldsymbol{y}_i,\boldsymbol{y}_j)-\frac{2}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\omega(\boldsymbol{z}_i,\boldsymbol{z}_j)\omega(\boldsymbol{y}_i,\boldsymbol{y}_k)$$
$$+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega(\boldsymbol{z}_i,\boldsymbol{z}_j)\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega(\boldsymbol{y}_i,\boldsymbol{y}_j),\tag{B.8}$$

其中

$$\omega(\boldsymbol{z}_i,\boldsymbol{z}_j)=\mathrm{e}^{-(|\boldsymbol{z}_i-\boldsymbol{z}_j|/\sigma_1)},\quad \omega(\boldsymbol{y}_i,\boldsymbol{y}_j)=\mathrm{e}^{-(|\boldsymbol{y}_i-\boldsymbol{y}_j|/\sigma_2)}.$$

结合 (B.7) 和 (B.8), 则有

$$\mathrm{HSICCor}_n(\boldsymbol{z},Y)=\frac{\mathrm{HSIC}_n(\boldsymbol{z},\boldsymbol{y})}{\sqrt{\mathrm{HSIC}_n(\boldsymbol{z},\boldsymbol{z})}\sqrt{\mathrm{HSIC}_n(\boldsymbol{y},\boldsymbol{y})}}.\tag{B.9}$$

根据文献 [12], 令 $\mathrm{HSIC}(\boldsymbol{z},Y)$ 中的两个未知窗宽参数 σ_1 和 σ_2 分别为 $T_{\boldsymbol{z}}=|\boldsymbol{z}_1-\boldsymbol{z}_2|$ 和 $T_{\boldsymbol{y}}=|\boldsymbol{y}_1-\boldsymbol{y}_2|$ 的中位数, 并利用基于边际的 HSIC 来构造变量筛选程序, 简记为 HSIC-SIS. 然而, 在利用 HSIC-SIS 进行特征筛选时, 由于 $\boldsymbol{y}$ 是基于对分类变量 Y 进行独热编码得到, 故 σ_2 有可能为 0. 因此, 对于 HSIC-SIS, 我们对其窗宽给出了两种选择方式: (1) 令 $\sigma_1=\sigma_2=$ 某个固定常数, 如令 $\sigma_1=\sigma_2=1$, 并记此时的筛选程序为 HSIC_1-SIS; (2) 令 σ_1 为 $T_{\boldsymbol{z}}$ 的中位数, 并简记为 $\sigma_1=\mathrm{Med}(T_{\boldsymbol{z}})$. 此外, 如果 $T_{\boldsymbol{y}}$ 的中位数大于 0, 则令 $\sigma_2=\mathrm{Med}(T_{\boldsymbol{y}})$; 否则, 令 $\sigma_2=1\times10^{-5}$. 为了简单, 记此时的筛选程序为 HSIC_m-SIS. 下面通过例 4.1–4.3 来评价 HSIC_1-SIS, HSIC_m-SIS 和 RSD-SIS 这 3 种筛选方法的表现, 具体的模拟结果如表 B13–B17 所示. 通过这些表格容易看到, RSD-SIS 的表现优于 HSIC_1-SIS 和 HSIC_m-SIS. 此外, 对于 HSIC-SIS, 虽然可以对分类变量 Y 进行独热编码, 但如何寻找更好的窗宽 σ_1 和 σ_2 以提高 HSIC-SIS 的表现值得进一步探索.

表 B13 在例 4.1 中, HSIC-SIS 和 RSD-SIS 的表现

n,p	方法	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{P}_3^s(X_3)$	$\mathcal{P}_4^s(X_4)$	$\mathcal{P}_5^s(X_5)$
$n=50$ $p=2000$	HSIC_1-SIS	10.00	14.83	0.583	0.862	0.970	0.973	0.915	0.745
	HSIC_m-SIS	11.00	16.31	0.557	0.833	0.958	0.968	0.917	0.747
	RSD-SIS	10.00	14.83	0.573	0.833	0.957	0.970	0.925	0.747
$n=80$ $p=2000$	HSIC_1-SIS	5.00	0.74	0.973	0.997	1.000	1.000	0.995	0.977
	HSIC_m-SIS	5.00	0.74	0.973	0.995	1.000	1.000	0.998	0.980
	RSD-SIS	5.00	0.74	0.970	0.995	0.998	1.000	0.992	0.978

表 B14 在例 4.2 中, 当 $(R,n,p)=(2,40,2000)$ 时, HSIC-SIS 和 RSD-SIS 的表现

ϵ	方法	平衡设计					非平衡设计				
		$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$
$C(0,1)$	HSIC_1-SIS	2.00	1.48	0.877	0.947	0.927	3.00	4.45	0.785	0.880	0.885
	HSIC_m-SIS	2.00	0.74	0.905	0.947	0.958	2.00	2.97	0.837	0.918	0.907
	RSD-SIS	2.00	1.48	0.882	0.945	0.935	2.00	2.22	0.855	0.928	0.918

表 B15 在例 4.2 的非平衡设计中, 当 $(\boldsymbol{R}, n, p) = (10, 200, 2000)$ 时, HSIC-SIS 和 RSD-SIS 的表现

ϵ	方法	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{P}_3^s(X_3)$	$\mathcal{P}_4^s(X_4)$	$\mathcal{P}_5^s(X_5)$	$\mathcal{P}_6^s(X_6)$	$\mathcal{P}_7^s(X_7)$	$\mathcal{P}_8^s(X_8)$	$\mathcal{P}_9^s(X_9)$	$\mathcal{P}_{10}^s(X_{10})$
$C(0,1)$	HSIC_1-SIS	0.09	0.498	0.607	0.725	0.790	0.858	0.927	0.943	0.958	0.987	0.990
	HSIC_m-SIS	0.172	0.552	0.675	0.760	0.838	0.898	0.932	0.958	0.977	0.987	0.992
	RSD-SIS	0.575	0.842	0.893	0.915	0.940	0.945	0.982	0.987	0.978	0.990	0.993

表 B16 在例 4.3 的情形 (b) 中, HSIC-SIS 和 RSD-SIS 的表现

情形 (b) R, n, p	方法	平衡设计						非平衡设计					
		$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$
$R = 3$	HSIC_1-SIS	12.00	17.79	0.713	0.882	0.903	0.907	22.00	40.77	0.532	0.628	0.902	0.955
$n = 120$	HSIC_m-SIS	4.00	2.41	0.947	0.995	0.972	0.980	6.00	7.41	0.843	0.863	0.983	0.995
$p = 2000$	RSD-SIS	3.00	2.97	0.948	0.987	0.983	0.978	4.00	3.71	0.923	0.933	0.995	0.995

表 B17 在例 4.3 的情形 (c) 中, HSIC-SIS 和 RSD-SIS 的表现

情形 (c) R, n, p	方法	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$	$\mathcal{P}_4^s$	$\mathcal{P}_5^s$	$\mathcal{P}_6^s$	$\mathcal{P}_7^s$	$\mathcal{P}_8^s$	$\mathcal{P}_9^s$	$\mathcal{P}_{10}^s$
		非平衡设计												
$R = 10$	HSIC_1-SIS	138.50	153.26	0.217	0.587	0.678	0.812	0.830	0.905	0.947	0.955	0.988	0.990	0.997
$n = 300$	HSIC_m-SIS	17.00	25.39	0.778	0.873	0.948	0.963	0.988	0.992	0.997	0.997	1.000	1.000	1.000
$p = 3000$	RSD-SIS	10.00	2.22	0.962	0.982	0.997	0.990	0.995	0.998	1.000	0.998	1.000	1.000	1.000

附录 B.5 当预测变量 $\boldsymbol{x} = (\boldsymbol{X_1}, \ldots, \boldsymbol{X_p})^{\mathrm{T}}$ 为离散变量时, 特征筛选程序 RSD-SIS 的表现

在定义 2.1 中, 对于连续的 d 维变量 $\boldsymbol{z} = (Z_1, \ldots, Z_d)^{\mathrm{T}}$ 和分类变量 $Y \in \{L_1, \ldots, L_R\}$, 本文定义了稳健的半距离协方差 $\text{RSDCov}(\boldsymbol{z}, Y)$ 来刻画 $\boldsymbol{z}$ 和 Y 的相关性, 其中 $\text{RSDCov}(\boldsymbol{z}, Y)$ 的构造是基于 $\boldsymbol{z}$ 在给定 Y 下的条件特征函数与 $\boldsymbol{z}$ 的无条件特征函数差异的平方的加权积分, 且加权函数为满足 (2.3) 的 d 维 $\mathcal{S}\alpha\mathcal{S}$ 分布的概率密度函数. 注意到加权函数在 $\mathbb{R}^d$ 上的取值恒大于 0, 故当 $\boldsymbol{z}$ 为 d 维离散变量时, $\text{RSDCov}(\boldsymbol{z}, Y)$ 依然可以来刻画 $\boldsymbol{z}$ 和 Y 的相关性. 简单来说, 无论 $\boldsymbol{z}$ 是连续变量还是离散变量, 均有 $\boldsymbol{z} \perp\!\!\!\perp Y \Leftrightarrow \text{RSDCov}(\boldsymbol{z}, Y) = 0$. 从这个角度来看, RSD-SIS 亦可用于预测变量 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 为离散变量的情形. 下面分析 RSD-SIS 在例 B.3–B.5 中的表现, 其中 RSD-SIS 中的窗宽均为成对样本距离的中位数, 详见 (3.2). 与此同时, 本文也给出了 RSD-SIS 在一些给定窗宽 σ=1, 6, 32, 256 下的模拟结果以便对 RSD-SIS 有一个相对更全面的评价, 并分别简记为 RSD_1-SIS, RSD_6-SIS, RSD_{32}-SIS 和 RSD_{256}-SIS.

例 B.3 与例 4.1 类似, 不同的是令 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 的每个分量独立同分布且 $X_1 \sim \text{Poisson}(2) - 2$, 其中 Poisson(2) 表示参数为 2 的 Poisson 分布. 此外, Y 来自如下的逻辑回归模型:

$$\log\left\{\frac{P(Y = 1 \mid \boldsymbol{x})}{P(Y = -1 \mid \boldsymbol{x})}\right\} = 8X_1 + 8X_2 + 8X_3 + 8X_4 + 10X_5 - 3.$$

例 B.4 与例 4.2 类似, 不同的是 $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^{\mathrm{T}}$ 的每个分量独立同分布且 $\epsilon_1 \sim \text{Poisson}(2) - 2$. 此外, 分类变量 Y 的分布来自非平衡设计,

$$p_r = P(Y = r) = 2[1 + (r - 1)/(R - 1)]/3R,$$

其满足 $\max_{1\leqslant r\leqslant R} p_r = 2\min_{1\leqslant r\leqslant R} p_r$.

例 B.5 与例 4.3 类似, 不同的是此处基于以下方法生成:

$$\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}} : \{X_{4(g-1)+k} \mid Z_g = z_g\} \sim \text{Poisson}(|z_g| + 1),$$

其中 $k = 1, 2, 3, 4$, $g = 1, \ldots, G$.

从表 B18–B20 不难看出, 当预测变量 $\boldsymbol{x}$ 为离散变量时, RSD-SIS 的表现与其竞争者相比并不让人十分满意, 这很大可能归因于窗宽参数的选择. 与此同时, 可以看到在部分固定窗宽 σ 下, RSD$_\sigma$-SIS 有相对较好的表现. 因此, 如何统一给出 RSD-SIS 在预测变量 $\boldsymbol{x}$ 为离散情形下的窗宽选择有待未来进一步探索.

表 B18 在例 B.3 中, $\mathcal{M}$, $\mathcal{S}$, $\mathcal{P}_a$, $\mathcal{P}_1^s(\boldsymbol{X}_1)$, $\mathcal{P}_2^s(\boldsymbol{X}_2)$, $\mathcal{P}_3^s(\boldsymbol{X}_3)$, $\mathcal{P}_4^s(\boldsymbol{X}_4)$ 和 $\mathcal{P}_5^s(\boldsymbol{X}_5)$ 的大小

n, p	方法	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{P}_3^s(X_3)$	$\mathcal{P}_4^s(X_4)$	$\mathcal{P}_5^s(X_5)$
$n = 160$ $p = 2000$	CI-SIS	8.00	7.41	0.878	0.972	0.963	0.973	0.970	0.998
	DC-SIS	7.00	7.41	0.885	0.963	0.980	0.973	0.963	1.000
	GN-SIS	7.00	7.41	0.878	0.982	0.957	0.982	0.958	0.998
	FKF-SIS	18.00	24.46	0.677	0.920	0.917	0.917	0.900	0.990
	MV-SIS	8.00	10.38	0.833	0.943	0.953	0.963	0.958	1.000
	PSIS	7.00	6.67	0.870	0.973	0.968	0.955	0.972	0.997
	SD-SIS	8.00	8.34	0.867	0.967	0.965	0.970	0.960	1.000
	RSD-SIS	29.50	46.70	0.515	0.862	0.857	0.880	0.860	0.987
	RSD$_1$-SIS	32.00	60.79	0.497	0.855	0.843	0.842	0.873	0.990
	RSD$_6$-SIS	9.00	11.12	0.848	0.968	0.967	0.963	0.947	1.000
	RSD$_{32}$-SIS	8.00	9.64	0.850	0.967	0.957	0.967	0.957	1.000
	RSD$_{256}$-SIS	7.00	7.41	0.878	0.982	0.963	0.962	0.973	0.997

表 B19 在例 B.4 中, 对于 $(R, n, p) = (10, 200, 2000)$, $\mathcal{P}_a$ 和 $\{\mathcal{P}_j^s(\boldsymbol{X}_j)\}_{j=1}^{10}$ 的大小

方法	$\mathcal{P}_a$	$\mathcal{P}_1^s(X_1)$	$\mathcal{P}_2^s(X_2)$	$\mathcal{P}_3^s(X_3)$	$\mathcal{P}_4^s(X_4)$	$\mathcal{P}_5^s(X_5)$	$\mathcal{P}_6^s(X_6)$	$\mathcal{P}_7^s(X_7)$	$\mathcal{P}_8^s(X_8)$	$\mathcal{P}_9^s(X_9)$	$\mathcal{P}_{10}^s(X_{10})$
CI-SIS	0.867	0.967	0.982	0.977	0.978	0.982	0.982	0.992	0.983	0.992	0.995
DC-SIS	0.725	0.847	0.933	0.963	0.978	0.995	0.990	1.000	1.000	1.000	1.000
GN-SIS	0.982	0.988	0.997	0.998	1.000	0.998	1.000	1.000	1.000	1.000	1.000
FKF-SIS	0.993	0.998	0.998	0.995	0.998	0.997	0.997	0.998	0.997	0.997	0.997
MV-SIS	0.978	0.988	0.997	0.998	0.998	0.997	1.000	1.000	1.000	1.000	1.000
PSIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SD-SIS	0.988	0.993	0.997	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
RSD-SIS	0.728	0.863	0.927	0.965	0.970	0.985	0.997	0.995	1.000	1.000	1.000
RSD$_1$-SIS	0.708	0.888	0.908	0.942	0.975	0.972	0.990	0.995	0.998	1.000	1.000
RSD$_6$-SIS	0.972	0.982	0.995	1.000	0.997	0.998	1.000	1.000	1.000	1.000	1.000
RSD$_{32}$-SIS	0.978	0.988	0.995	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000
RSD$_{256}$-SIS	0.988	0.992	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

表 B20 在例 B.5 中, $\mathcal{M}$, $\mathcal{S}$, $\mathcal{P}_a$ 和 $\{\mathcal{P}_j^s(\boldsymbol{X}_{(j)})\}_{j=1}^3$ 的大小, 其中 $\mathcal{P}_1^s$, $\mathcal{P}_2^s$ 和 $\mathcal{P}_3^s$ 分别表示 $\mathcal{P}_1^s(\boldsymbol{X}_{(1)})$, $\mathcal{P}_2^s(\boldsymbol{X}_{(2)})$ 和 $\mathcal{P}_3^s(\boldsymbol{X}_{(3)})$

情形 R, n, p	方法	平衡设计						非平衡设计					
		$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{P}_a$	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$
(a) $R=3$ $n=75$ $p=2000$	DC-SIS	4.00	5.93	0.805	0.928	0.925	0.938	9.00	16.31	0.668	0.812	0.893	0.937
	GN-SIS	4.00	5.19	0.838	0.930	0.957	0.940	5.00	9.08	0.765	0.848	0.942	0.960
	SD-SIS	4.00	5.93	0.823	0.948	0.942	0.918	5.00	9.08	0.762	0.865	0.920	0.960
	RSD-SIS	6.00	11.86	0.738	0.918	0.887	0.905	9.00	16.31	0.668	0.812	0.893	0.937
	RSD_1-SIS	37.00	55.78	0.313	0.690	0.668	0.687	50.00	60.04	0.198	0.475	0.655	0.747
	RSD_6-SIS	5.00	8.90	0.775	0.910	0.930	0.925	6.00	9.64	0.773	0.877	0.927	0.948
	RSD_{32}-SIS	4.00	5.93	0.828	0.935	0.938	0.948	5.00	8.15	0.773	0.863	0.945	0.957
	RSD_{256}-SIS	4.00	5.19	0.823	0.925	0.942	0.953	6.00	10.38	0.752	0.860	0.930	0.940

Model-free feature screening for ultrahigh-dimensional classification data

Feng Zou, Hu Zhang & Hengjian Cui

Abstract In this paper, we propose a robust semi-distance correlation (RSDCor for short) to measure the dependence between a categorical random variable $Y \in \mathbb{R}^1$ and a continuous random vector $\boldsymbol{z} \in \mathbb{R}^d$ ($d \geqslant 1$). RSDCor can be regarded as a generalization of semi-distance correlation (for short SDCor) proposed by Zhong et al. (2024). Compared with the SDCor, the significant advantage of the RSDCor is that it is relatively robust to heavy-tailed distributions or outliers. In addition, RSDCor satisfies $0 \leqslant \text{RSDCor}(\boldsymbol{z}, Y) \leqslant 1$ and $\text{RSDCor}(\boldsymbol{z}, Y) = 0$ if and only if $\boldsymbol{z}$ is independent of Y. Furthermore, we propose a feature screening procedure for ultrahigh-dimensional classification data based on RSDCor, which is abbreviated as RSD-SIS. In theory, we establish the sure screening property and ranking consistency of the RSD-SIS. RSD-SIS enjoys several appealing merits. First, it does not require specific model assumptions or finite moment assumptions of predictor variables; second, it can also be directly used to screen grouped predictor variables and a categorical response variable; third, it allows the categorical response variable to have a diverging number of classes at a rate of $O(n^\kappa)$ with some $\kappa \geqslant 0$. Monte Carlo simulation and a real data analysis both indicate that RSD-SIS performs very well compared with the existing screening procedures.

Keywords **measures of dependence, ultrahigh-dimensional classification, feature screening, robustness, screening properties**

MSC(2020) **62H12, 62H20, 62H30**

doi: 10.1360/SSM-2024-0320