



基于数据分割的受控模型自由变量选择方法

韩艺欣¹, 郭旭^{2*}, 邹长亮¹

1. 南开大学统计与数据科学学院, 天津 300071;

2. 北京师范大学统计学院, 北京 100875

E-mail: yuxiaohaihyx@gmail.com, xustat12@bnu.edu.cn, nk.chlzou@gmail.com

收稿日期: 2022-01-26; 接受日期: 2024-03-18; 网络出版日期: 2024-08-26; * 通信作者

国家自然科学基金 (批准号: 11925106, 12231011, 11931001, 11971247, 12322112 和 12071038) 和国家重点研发计划 (批准号: 2022YFA1003703 和 2022YFA1003800) 资助项目

摘要 从高维数据中识别重要变量并控制错误发现率是一个重要的统计问题. 本文在充分降维框架中, 通过数据分割的手段针对高维数据提出一种误差可控且模型自由的变量选择方法. 该方法首先通过一个响应变换函数将一般化的模型转化为求解最小二乘估计问题. 然后构造一系列边际对称的统计量和一个数据驱动的拒绝临界值, 来实现错误发现率可控的变量选择. 在一些温和的理论条件下, 本文证明该方法能精确控制有限样本下的错误发现率, 同时也能实现大样本下的错误发现率控制. 本文通过数值模拟和高维疾病基因识别的实例分析, 验证了该方法相较于其他现有方法能以较高的检验效率且更快更准确地控制错误发现率.

关键词 数据分割 错误发现率 模型自由 充分降维 对称性

MSC (2020) 主题分类 62J02, 62H15

1 引言

充分降维 (sufficient dimension reduction, SDR) 是一种从高维数据中提取有用信息的降维技术 (参见文献 [10, 21, 22, 32]). 用 Y 表示响应变量, 其支撑集为 Ω_Y , $\mathbf{X} = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$ 表示 p 维协变量. SDR 的基本思想旨在用原始协变量投影子空间中的一系列线性组合来代替所有的原始变量, 同时不损失给定 $\mathbf{X}$ 后 Y 的条件分布的信息. 实际上, 高维数据中往往会收集到大量特征, 但通常其中仅有一小部分变量对响应变量真正起作用. 然而, 在挖掘数据中的重要特征或模式时, SDR 中所涉及的降维子空间通常由所有原始变量组成, 这使得模型的解释性不高, 尤其是在高维情形中. 因此, 本文的目标是开发一种模型自由的变量选择方法, 通过控制某种错误发现率来筛选出真正无贡献的变量或识别出重要的变量, 从而简化后续的模型构建使其简单可行, 并有助于降低由高维数据引起的计算成本.

英文引用格式: Han Y X, Guo X, Zou C L. Model-free controlled variable selection via data splitting (in Chinese). Sci Sin Math, 2025, 55: 1567–1592, doi: 10.1360/SCM-2022-0066

设 $F(Y | \mathbf{X})$ 是给定 $\mathbf{X}$ 后 Y 的条件分布函数. 当 p 很大时, 条件分布的模型构造和验证极其困难. 所以可用 SDR 解决“维数灾难”问题, 一般目标是研究 $\mathbf{X}$ 如何影响条件分布 $F(Y | \mathbf{X})$ 的. 信号变量和噪声变量的集合分别定义为

$$\begin{aligned}\mathcal{A} &= \{j : F(Y | \mathbf{X}) \text{ 依赖于 } X_j, j = 1, \dots, p\}, \\ \mathcal{A}^c &= \{j : F(Y | \mathbf{X}) \text{ 不依赖于 } X_j, j = 1, \dots, p\}.\end{aligned}$$

针对上述任务, 许多流行的变量选择方法是对线性模型或广义线性模型范式考虑正则化, 如 LASSO (least absolute shrinkage and selection operator)^[29]、SCAD (smoothly clipped absolute deviation)^[14], 或自适应 LASSO^[38], 可参见文献 [16] 和专著 [15] 以获得更完整的参考文献.

相较于正则化类型的方法, 模型自由的变量选择可通过 SDR 来实现, 它不需要真实模型的全部信息, 因此研究人员不需要处理有关模型指定错误的问题. 它旨在找到信号变量子集 $\mathcal{A}$ 使得

$$Y \perp\!\!\!\perp \mathbf{X}_{\mathcal{A}^c} | \mathbf{X}_{\mathcal{A}} \quad (1.1)$$

成立, 其中“ $\perp\!\!\!\perp$ ”表示统计独立性, $\mathbf{X}_{\mathcal{A}} = \{\mathbf{X}_j : j \in \mathcal{A}\}$ 表示由重要变量子集 $\mathcal{A}$ 组成的向量, 且 $\mathbf{X}_{\mathcal{A}^c}$ 表示 $\mathbf{X}_{\mathcal{A}}$ 的补集. 条件 (1.1) 表明 $\mathbf{X}_{\mathcal{A}}$ 中包含了所有与 Y 有关的协变量的相关信息. 文献 [23] 结合充分降维和变量选择提出了一种与模型无关的方法. 文献 [7] 提出了一种坐标无关的稀疏估计方法, 可以同时高效地实现稀疏 SDR 并筛除无关变量. 文献 [31] 则关注的是具有不断增多的预测变量数量的模型自由变量选择. 边际坐标假设检验由文献 [8] 提出, 用于低维情形下的模型无关的变量选择. 之后, 文献 [27, 34] 对文献 [8] 进行了改进和推广. 文献 [35] 将边际坐标检验扩展到了切片逆回归中, 文献 [36] 提出了用于超高维特征选择的边际 SIR (sliced inverse regression) 效用迹追踪算法, 但却未提供控制错误发现的方法. 可参阅综述 [24, 37] 以获得更全面的文献总结. 在与模型无关的变量选择方法中, 通常需要知道边际渐近分布指定一个临界值来控制信号变量和噪声变量之间的分界点. 然而, 这种方法并不直接适用于高维情形, 因为渐近分布此时可能很难计算甚至不存在, 故这些方法并不能直接考虑变量选择的不确定性量化, 即在高维情形下对重要协变量的选定子集进行全局错误率控制.

多重检验为大规模数据集识别稀疏信号提供了一种有效的方法. 通常采用控制错误发现率^[4] (false discovery rate, FDR) 的手段来避免在多重检验问题中的虚假发现. 文献 [1] 介绍了线性模型中控制 FDR 的 Knockoff 框架, 通过“Knockoff 副本”构建的统计量将满足 (或大致满足) 联合可交换性, 以此实现有限样本 FDR 控制. 在一般化的高维非线性情形中, 文献 [6] 进一步开发了 Model-X Knockoff 方法, 可在与模型无关的情形中控制 FDR. 然而, Model-X Knockoff 方法需要协变量的联合分布已知, 这在实际中通常很难获得. 最近, 文献 [18] 通过使用去相关的得分统计量改进了边际坐标检验^[8, 34], 进一步应用标准的 Benjamini-Hochberg (BH) 程序^[4] 对 p 值进行修正以控制 FDR, 但是去相关过程的密集计算限制其高维情形下的应用. 从另一个角度出发, 文献 [12] 提出了一种名为 SDA (symmetrized data aggregation) 过滤器的数据分割策略, 以构建具有全局对称性质的一系列统计量, 然后应用对称性推导出用于错误率控制的数据驱动阈值. 具体而言, 文献 [12] 将相依结构聚合到一个具有伪响应和固定协变量的线性模型中, 使相依结构成为提高功率的福音. 在众多的应用场景中, 与实际中的渐近正态性相比, 渐近对称性更容易满足. 因此 SDA 比其竞争对手更有实用价值, 这激发了我们将其应用于充分降维问题.

本文提出一种误差可控且模型自由的变量选择方法. 首先证明模型自由问题等效于对多个响应变换函数的一组线性回归系数进行统计推断. 随后分别针对低维和高维情形的错误发现率控制开发新的

变量选择方法. 主要贡献总结如下: (1) 本文提出一种新的数据驱动变量选择方法来控制 FDR, 该方法通过挑选不同的响应变换函数, 结合现有 SDR 方法可实现模型自由的变量选择; (2) 该方法不需要估计任何讨厌参数, 例如不需要估计 SDR 中的结构维数; (3) 本文方法仅涉及一次数据分割和从两部分子样本中获得降维矩阵的乘积的计算, 故计算效率高且易于实现; (4) 在不需要过多理论假设的前提下, 证明了所提出方法能在有限样本和大样本下实现 FDR 的有效控制; (5) 数值实验表明, 所提出方法比现有方法具有更令人满意的 FDR 控制和更高的检验功效.

本文余下内容的安排如下. 第 2 节展开介绍问题描述和模型构建. 第 3 节首先研究通过错误发现率控制的低维变量选择方法, 然后提出高维变量选择方法. FDR 控制的理论在第 4 节中讨论. 第 5 节展示模拟研究和实例分析, 以验证本节所提出方法的优越性能. 第 6 节对本文进行一些讨论, 并提出一些可改进的想法. 附录 A 给出主要理论结果的证明. 附录 B 描述理论证明和额外的数值结果.

记号 令 $\lambda_{\min}(\mathbf{B})$ 和 $\lambda_{\max}(\mathbf{B})$ 表示方阵 $\mathbf{B}$ 的最大和最小特征值. 记 $\|\mathbf{B}\|_2 = (\sum_i \sum_j a_{ij}^2)^{1/2}$ 和 $\|\mathbf{B}\|_\infty = \max_i \sum_j |a_{ij}|$. 令 $\|\boldsymbol{\mu}\|_1 = \sum_i |\mu_i|$ 和 $\|\boldsymbol{\mu}\|_2 = (\sum_i \mu_i^2)^{1/2}$ 分别表示向量 $\boldsymbol{\mu}$ 的 L_1 和 L_2 范数. $\mathbb{E}(\mathbf{X})$ 和 $\text{cov}(\mathbf{X})$ 分别表示随机向量 $\mathbf{X}$ 的均值和协方差阵. $A_n \approx B_n$ 表示两个量 A_n 和 B_n 渐近等价. 存在常数 $C > 1$ 使得 $B_n/C \leq A_n \leq B_n C$ 以趋于 1 的概率成立. “ $\gtrsim$ ” 和 “ $\lesssim$ ” 也类似定义.

2 问题描述和模型构建

模型 (1.1) 中关心的变量选择问题可转化为如下多重检验问题:

$$\mathbb{H}'_{0j} : j \in \mathcal{A}^c \quad \text{vs.} \quad \mathbb{H}'_{1j} : j \in \mathcal{A}. \quad (2.1)$$

这是文献 [8, 35] 提出的边际坐标检验, 相关研究还有文献 [23, 27, 34]. 这类变量选择方法一般是定义非负的边际效用统计量 W_j , 来衡量 X_j 在某种意义下对 Y 的重要性. 然而, 这些方法的拒绝临界值通常与误差分布有关, 当分布未知或很难求时, 无法直接在变量选择时控制错误发现率.

接下来为 (1.1) 提供一种新的形式. 令 $\boldsymbol{\Sigma} = \text{cov}(\mathbf{X}) > 0$ 并设 $\mathbb{E}(\mathbf{X}) = 0$. 记 $\mathcal{C}$ 是一个降维子空间. 对于满足 $\mathbb{E}\{f(Y)\} = 0$ 的任何函数 $f(Y)$, 根据文献 [31, 33], 可证明

$$\boldsymbol{\Sigma}^{-1} \text{cov}(\mathbf{X}, f(Y)) \in \mathcal{C},$$

这里要求线性条件^[22]成立. 已知线性条件通常当 $\mathbf{X}$ 是椭球分布时会满足. 响应变量变换函数 $f(\cdot)$ 的使用是为了提高模型拟合优度, 这与传统的变换函数承担的角色不同, 这里用作执行降维的中间工具. 因此, 可挑选一系列形式不依赖于数据的变换函数^[11], $f_1(Y), \dots, f_H(Y)$, 其中 $H (> d)$ 是预先指定的, 一般称 H 为工作维数, d 是真正的降维子空间结构维数. 在总体意义下, 可获得在 $\mathbf{X}$ 上回归 $f_h(Y)$ 的普通最小二乘估计

$$\boldsymbol{\beta}_h^0 = \arg \min_{\boldsymbol{\beta}_h} \mathbb{E}[\{f_h(Y) - \mathbf{X}^\top \boldsymbol{\beta}_h\}^2], \quad h = 1, \dots, H. \quad (2.2)$$

令 $\mathbf{B}_0 = (\boldsymbol{\beta}_1^0, \dots, \boldsymbol{\beta}_H^0)$, 则有 $\text{Span}(\mathbf{B}_0) \subseteq \mathcal{C}$, 其中 $\text{Span}(\mathbf{B}_0)$ 表示由 $\mathbf{B}_0$ 的列向量所张成的空间. 基于 SDR 文献中的常用协议, 每当有 $\text{Span}(\mathbf{B}_0) \subseteq \mathcal{C}$, 通常会进一步假设覆盖条件成立, 即 $\text{Span}(\mathbf{B}_0) = \mathcal{C}$. 该条件在实际中通常是成立的, 可参见文献 [9] 以获取更多讨论.

对于 $j = 1, \dots, p$, 令 β_{hj}^0 表示 $\boldsymbol{\beta}_h^0 \in \mathbb{R}^p$ 的第 j 个元素, $h = 1, \dots, H$. 如果第 j 个变量是噪声变量, 则 $\mathbf{B}_j = \mathbf{0} \in \mathbb{R}^H$, 其中 $\mathbf{B}_j = (\beta_{1j}^0, \beta_{2j}^0, \dots, \beta_{Hj}^0)^\top \in \mathbb{R}^H$ 表示 $\mathbf{B}_0$ 的第 j 行. 进一步地, $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{X}_{\mathcal{A}}$

意味着 $\sum_{h=1}^H |\beta_{hj}^0| > 0$ 对于 $j \in \mathcal{A}$ 成立且 $\sum_{h=1}^H |\beta_{hj}^0| = 0$ 对于 $j \in \mathcal{A}^c$ 成立. 换言之, 如果 j 属于信号子集 $\mathcal{A}$, 则响应变量 Y 必须通过 H 个线性组合中的至少一个依赖于变量 X_j . 如果 j 属于噪声子集 $\mathcal{A}^c$, 则 H 个线性组合中没有一个是与 X_j 有关 (参见文献 [35]). 这种联系启发我们可尝试通过恢复降维中心子空间来挑选信号子集. 因此, 检验问题 (2.1) 等价于以下多重假设检验问题:

$$\mathbb{H}_{0j} : \sum_{h=1}^H |\beta_{hj}^0| = 0 \quad \text{vs.} \quad \mathbb{H}_{1j} : \sum_{h=1}^H |\beta_{hj}^0| > 0. \quad (2.3)$$

基于上述讨论, 在模型自由框架中挑选信号变量相当于在多元响应线性模型中选择重要变量. 假设收集到独立同分布样本 $\mathcal{D} = \{\mathbf{X}_i, Y_i\}_{i=1}^{2n}$. 令 $\{W_j, 1 \leq j \leq p\}$ 表示基于样本 $\mathcal{D}$ 计算的边际统计量. 则基于临界值 L 的变量选择方法为

$$\hat{\mathcal{A}}(L) = \{j : W_j \geq L, 1 \leq j \leq p\}, \quad (2.4)$$

其中 $\hat{\mathcal{A}}(L)$ 是基于临界值 L 得到的 $\mathcal{A}$ 的估计. 显然, L 在变量选择中起着控制模型复杂度的重要作用. 下文将通过控制 FDR 来构建适当的临界值, 以实现 SDR 中模型自由的变量选择.

记 $p_0 = |\mathcal{A}^c|$ 和 $p_1 = |\mathcal{A}|$. 假设 p_1 总是由 p 控制着, 即 $p_1 = o(p)$. 因此与选择过程 (2.4) 有关的错误发现比例 (false discovery proportion, FDP) 是

$$\text{FDP}(\hat{\mathcal{A}}(L)) = \frac{\#\{j : j \in \hat{\mathcal{A}}(L) \cap \mathcal{A}^c\}}{\#\{j : j \in \hat{\mathcal{A}}(L)\} \vee 1},$$

其中 $a \vee b = \max\{a, b\}$ 且 $\#\{\cdot\}$ 表示一个集合的基数. FDR 被定义为 FDP 的期望, 即 $\text{FDR}(L) = \mathbb{E}(\text{FDP}(L))$. 本文的主要目标是找到一个数据驱动临界值 L , 该临界值能够渐近地控制 FDR 在目标显著性水平 α 下, 即 $\limsup_{n \rightarrow \infty} \text{FDR}(\hat{\mathcal{A}}(L)) \leq \alpha$.

3 控制错误发现率的变量选择

本节首先在第 3.1 小节中提出数据驱动的临界值选取, 基于数据分割技术针对 $p < n$ 的情形来控制模型自由的变量选择过程中的错误发现率, 并在第 3.2 小节中讨论高维版本.

3.1 低维方法

令 $\hat{\beta}_1, \dots, \hat{\beta}_H$ 表示 (2.2) 中的估计. 首先将全样本数据 $\mathcal{D} = \{\mathbf{X}_i, Y_i\}_{i=1}^{2n}$ 随机划分为两个子样本量相同的独立子集 $\mathcal{D}_1 = \{\mathbf{X}_{1i}, Y_{1i}\}_{i=1}^n$ 和 $\mathcal{D}_2 = \{\mathbf{X}_{2i}, Y_{2i}\}_{i=1}^n$, 分别用于构造降维子空间估计量 $\hat{\mathbf{B}}_1$ 和 $\hat{\mathbf{B}}_2$, 其中 $\hat{\mathbf{B}}_1 = (\hat{\beta}_1^{(1)}, \dots, \hat{\beta}_H^{(1)})$ 且 $\hat{\mathbf{B}}_2 = (\hat{\beta}_1^{(2)}, \dots, \hat{\beta}_H^{(2)})$. 在分割 $\mathcal{D}_k$ 上, $\mathbf{B}_j$ 的最小二乘估计为

$$\hat{\mathbf{B}}_{kj}^\top = \mathbf{e}_j^\top \left(\sum_{i=1}^n \mathbf{X}_{ki} \mathbf{X}_{ki}^\top \right)^{-1} \sum_{i=1}^n \mathbf{X}_{ki} \mathbf{f}_{ki}^\top,$$

其中 $k = 1, 2$, $\mathbf{f} = (f_1(Y), \dots, f_H(Y))^\top$, $\mathbf{e}_j$ 表示第 j 维为 1 其余维数全为 0 的 p 维单位向量. 然后将两部分中的信息通过内积结合起来, 构造一个对称的排序统计量

$$W_j = \frac{\hat{\mathbf{B}}_{1j}^\top \hat{\mathbf{B}}_{2j}}{s_{1j} s_{2j}}, \quad j = 1, \dots, p, \quad (3.1)$$

其中 $s_{kj}^2 = \mathbf{e}_j^\top (\sum_{i=1}^n \mathbf{X}_{ki} \mathbf{X}_{ki}^\top)^{-1} \mathbf{e}_j$, $k = 1, 2$. 对于信号变量, 若 $\sum_{h=1}^H |\beta_{hj}^0|$ 数值较大 (在 $\mathbb{H}_{1j}$ 下), 则 $\hat{\mathbf{B}}_{1j}$ 和 $\hat{\mathbf{B}}_{2j}$ 具有相同的符号且倾向于有较大的绝对数值, 从而得到一个较大的正数 W_j . 对于一个噪声变量, W_j 对称地在零轴附近波动, 即 W_j 对于噪声变量具有边际对称性^[1,12], 以此可区分信号变量和噪声变量. 该想法启发我们, 挑选一个如下所示的临界值 L 便能控制错误发现率在给定的显著性水平 α 下:

$$L = \inf \left\{ t > 0 : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \alpha \right\}, \quad (3.2)$$

我们的决策规则由 $\hat{\mathcal{A}}(L) = \{j : W_j \geq L, 1 \leq j \leq p\}$ 给出. 若该集合为空集, 则可简单设置 $L = +\infty$. 在零假设下基于 W_j 的边际对称性, (3.2) 中的比例是 FDP 的一个很好的估计, 因为 $\#\{j : W_j \leq -t\}$ 是 $\#\{j : W_j \geq t, j \in \mathcal{A}^c\}$ 一个很好的近似. 该方法的核心是通过响应变换函数实现模型自由, 并用数据分割技术来构造具有对称性的边际统计量, 从而可得到一个数据驱动的临界值以实现变量选择. 因此我们命名该方法为基于数据分割的模型自由变量选择 (model-free selection via data splitting, MFSDS).

由于估计的 $\hat{\mathbf{B}}_1$ 和 $\hat{\mathbf{B}}_2$ 仅是 $\mathbf{B}$ 的两个近似, 它们不是由特征分解导出的^[10,22], 故不用担心 $\hat{\mathbf{B}}_1$ 和 $\hat{\mathbf{B}}_2$ 是否属于同一个降维子空间, 因为我们的主要目标是识别信号变量而非恢复降维子空间本身. 这意味着通过 (2.2) 实现的变量选择不需要估计降维基方向, 也不需要估计结构维数 d . 因此, 所提出方法可通过选择不同的 $f(Y)$ 来适应一系列逆切片回归估计量. 本文的方法由于其简单性、计算高效和通用性可被广泛使用. 总结其为算法 1.

算法 1 误差可控且模型自由的变量选择 (MFSDS)

步骤一 (初始化) 给定参数 H 、 $\mathbf{f}$ 和 α ;

步骤二 (数据分割) 随机将全数据分割为样本量相同且互不相交的两个子集 $\mathcal{D}_1$ 和 $\mathcal{D}_2$. 根据 (2.2) 获得降维估计 $\hat{\mathbf{B}}_1$ 和 $\hat{\mathbf{B}}_2$;

步骤三 (排序统计量) 由 (2.1) 构造检验统计量 W_j 并排序;

步骤四 (构造临界值) 基于 (3.2) 计算临界值 L , 并获得被挑选的信号子集 $\hat{\mathcal{A}}(L)$.

算法 1 的总计算复杂度为 $O(2nHp^2 + p \log p)$, 因此 MFSDS 很容易实现. 但需要注意 MFSDS 涉及数据分割, 这可能会导致全样本数据的信息有所损失^[12]. 幸运的是, 该方法基于零假设下 W_j 的边际对称性会得到一个数据驱动的临界值, 从而避免推导复杂的零渐近分布. 这里用一个演示模拟来说明数据分割的优势, 详细信息可参考第 5 节. 在图 1 中观察到, 数据分割方法 (图 1(a)) 将大多数信号变量置于零轴以上, 许多噪声变量围绕零轴对称分布. 这是选择算法的一个关键特性, 而基于全样本的估计 (图 1(b)) 和基于部分样本的估计 (图 1(c)) 方法都无法达到这种对称性.

3.2 高维方法

实际上, 当协变量的维数 p 非常大时, 上述 (2.2) 中的普通最小二乘方法就会失效. 但注意到, 我们的数据分割方法可灵活扩展到正则化形式的版本. 因此, 这里借鉴了文献 [12] 提出的 SDA, 针对高维数据提出了如下的变量选择方法.

为了从 $\mathcal{D}_1$ 中提取信息, 将 (2.2) 中的最小二乘解替换为 LASSO 变量选择方法^[29],

$$\beta_{\lambda_h} = \arg \min_{\beta_h} \mathbb{E}[\{f_h(Y) - \mathbf{X}^\top \beta_h\}^2] + \lambda_h \|\beta_h\|_1, \quad h = 1, \dots, H, \quad (3.3)$$

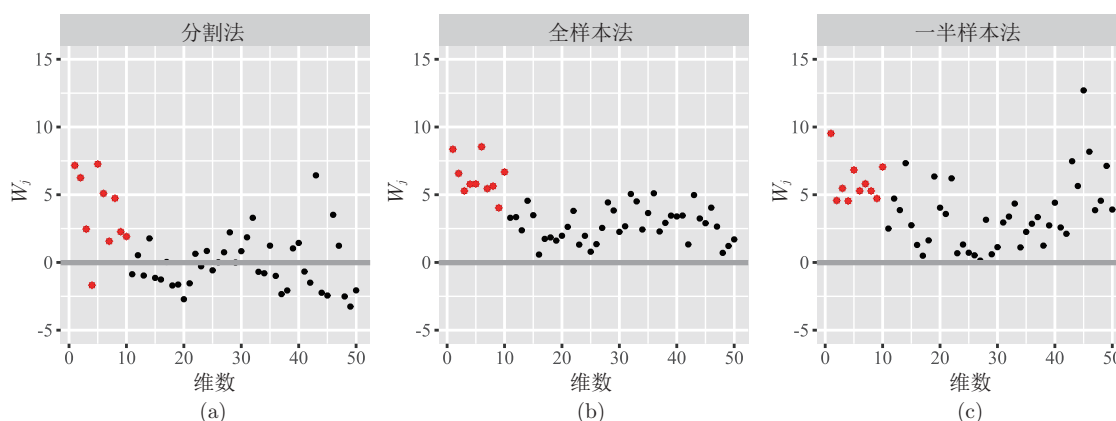


图 1 (网络版彩图) 统计量 W_j 的散点图, 其中红点和黑点分别表示信号变量和噪声变量. (a) 本文所提出的 (3.1) 中的统计量 W_j ; (b) $W_j = \sum_{h=1}^H |\beta_{hj}|$, 这里 $(B)_{hj} = \beta_{hj}$ 表示基于全样本对回归系数的最小二乘估计; (c) 在 (3.1) 中将 B_2 替换为 B_1

其中 $\lambda_h > 0$ 是调节参数. 虽然 LASSO 不能为所选出的重要变量提供错误率控制的保证, 但它同时考虑了文献 [12] 中描述的稀疏性和相依结构. 本文不需要同时惩罚 H 个切片中的系数来构造 B 的稀疏性, 因为本文的选择不是通过 B 本身实现的, 这与传统的模型自由变量选择方法大有不同, 如文献 [31].

令 $S = \{j : \sum_{h=1}^H |\hat{\beta}_{\lambda_h, j}^{(1)}| > 0\}$ 表示由 (3.3) 挑选的信号变量子集, 其中 $\hat{\beta}_{\lambda_h, j}^{(1)}$ 是 β_{λ_h} 基于 $\mathcal{D}_1$ 的第 j 个元素的估计. 然后, 在 $\mathcal{D}_2$ 上基于缩小子集 S 中的协变量来获得 (2.2) 中的最小二乘估计 $\hat{B}_{2S}$. 令 $\hat{B}_1$ 和 $\hat{B}_2$ 表示在 $\mathcal{D}_1$ 和 $\mathcal{D}_2$ 上对 B 估计: 当 $j \in S$ 时, $\hat{B}_{2j} = \hat{B}_{2Sj}$; 否则, $\hat{B}_{2j} = 0$. 因而, 高维中的统计量为

$$W_j = \frac{\hat{B}_{1j}^\top \hat{B}_{2j}}{s_{1Sj} s_{2Sj}}, \quad j = 1, \dots, p,$$

其中 $s_{kSj}^2 = e_j^\top (n^{-1} \sum_{i=1}^n \mathbf{X}_{kSi} \mathbf{X}_{kSi}^\top)^{-1} e_j$, $k = 1, 2$ 是关于子集 S 中变量的方差估计. 这里的统计量 W_j 具有与 (3.1) 中提出的统计量相似的性质, 它在零假设下 (渐近) 对称且均值为零, 在备择假设下不需要指定 Y 和 $\mathbf{X}$ 之间的关系且具有较大的数值. 因此, 本文建议构造临界值 L_+ 为

$$L_+ = \inf \left\{ t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \alpha \right\}, \quad (3.4)$$

并通过 $\hat{\mathcal{A}}(L_+) = \{j : W_j \geq L_+, 1 \leq j \leq p\}$ 来选择信号变量. (3.4) 中提出的 L_+ 与 Model-X Knockoff^[6] 和 SDA^[12] 具有类似的想法以获得精确的 FDR 控制. 然而, 在高维变量选择问题中, 通常无法收集到足够的 $(Y, \mathbf{X})$ 信息, 且当 $p > n$ 时 Knockoff 一般不可用. 幸运的是, MFSDS 不需要任何关于 $(Y, \mathbf{X})$ 的分布和渐近分布的先验信息, 因此更适合处理高维变量选择问题.

4 理论性质

本节主要研究控制错误发现率的理论. 首先, 需要对响应变换函数 $f(Y)$ 作一些常规的假设.

假设 4.1 (响应变换) 任意函数 $f(Y)$ 须满足 $\mathbb{E}\{f(Y)\} = 0$ 和 $\text{var}\{f(Y) - \mathbf{X}^\top \beta_0\} < \infty$.

条件 4.1 保证了我们的工作区别于大多数基于模型的变量选择方法. 本文方法将一般模型转化为多元响应线性模型, 进而实现了模型自由的变量选择^[31]. 基于该条件给出第一个理论结果如下.

定理 4.1 (有限样本下的 FDR 控制) 若假设 4.1 成立, 且统计量 W_j 是明确定义好的, 则对于任意 $\alpha \in (0, 1)$, 本文模型自由的变量选择方法 MFSDS 满足

$$\text{FDR} \leq \min_{\epsilon \geq 0} \left\{ \alpha(1 + 4\epsilon) + \left(\max_{j \in \mathcal{A}^c} \Delta_j > \epsilon \right) \right\},$$

其中 $\Delta_j = |(|W_j| - |W_j|, \mathbf{W}_{-j}) - 1/2|$ 且 $\mathbf{W}_{-j} = (W_1, \dots, W_p)^\top \setminus W_j$.

无论协变量 $\mathbf{X}$ 和响应 Y 之间的关系已知与否, 定理 4.1 均成立. 该结论可根据文献 [3] 提供的证明验证. 这里 Δ_j 可被解释为衡量 X_j 的非对称性以及 W_j 和 $\mathbf{W}_{-j}$ 之间的依赖关系对 FDR 的影响的度量. 即使是在非对称情形下, Δ_j 仍可以期望很小, 因为如果 n 不是太小的话, $\hat{\beta}_j^{(1)}$ 和 $\hat{\beta}_j^{(2)}$ 都会收敛到正态分布. 定理 4.1 意味着, 在不对称情形下, 对 Δ_j 进行严格控制也可得到有效的 FDR 控制.

当证明本文中渐近意义下的 FDR 控制时, 需要以下理论假设. 这些假设虽然不是最弱的, 但有助于为附录 B 中的理论证明提供技术手段. 令

$$\|\mathbf{A} - \Sigma_S\|_\infty = O_p(a_{np}),$$

其中 $a_{np} \rightarrow 0$, $\mathbf{A} = n^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \mathbf{X}_{2Si}^\top$ 且 $\Sigma_S = \mathbb{E}(\mathbf{X}_S \mathbf{X}_S^\top)$. 记 $v_n = \max\{\|\Sigma_S\|_\infty, \|\Sigma_S^{-1}\|_\infty\}$, $\mathbf{B}_{0S} = \{\mathbf{B}_j : j \in S\}$, $d_n = |\mathcal{A}|$, $q_n = |S|$, $q_{0n} = |S \cap \mathcal{A}^c|$. 假设 q_n 由一个非随机序列 $\bar{q}_n$ 一致有界控制.

假设 4.2 (确定筛选性) 当 $n \rightarrow \infty$ 时, 有 $\Pr(\mathcal{A} \subseteq S) \rightarrow 1$.

假设 4.3 (矩条件) 令 $\boldsymbol{\varepsilon} = \mathbf{f} - \mathbf{B}_{0S}^\top \mathbf{X}_S \in \mathbb{R}^H$. 在给定 S 时, 存在一个正数发散序列 K_n 和一个常数 $\varpi > 2$, 对于 $i \in \mathcal{D}_2$, 满足

$$\max_{1 \leq h \leq H} \max_{1 \leq i \leq n} \mathbb{E}(\|\Sigma_S^{-1} \mathbf{X}_{2Si} \varepsilon_{ih}\|_\infty^\varpi) \leq K_n^\varpi.$$

假定随 $n \rightarrow \infty$, 对于某个较小的 $\gamma > 0$, 有 $\bar{q}_n^{1/\varpi + \gamma + 1/2} K_n / n^{1/2 - \gamma - 1/\varpi} \rightarrow 0$.

假设 4.4 (设计阵) 存在两个正常数 $\underline{\kappa}$ 和 $\bar{\kappa}$ 以趋于 1 的概率满足

$$\underline{\kappa} \leq \liminf_{n \rightarrow \infty} \lambda_{\min}(\mathbf{X}_{2S}^\top \mathbf{X}_{2S} / n) < \limsup_{n \rightarrow \infty} \lambda_{\max}(\mathbf{X}_{2S}^\top \mathbf{X}_{2S} / n) \leq \bar{\kappa}.$$

假设 4.5 (估计精度) 在 $\mathcal{D}_1$ 上构造的估计 $\hat{\mathbf{B}}_1$ 对 $j \in S$ 一致满足 $\|\hat{\mathbf{B}}_{1j} - \mathbf{B}_j\|_\infty = O_p(c_{np})$, 其中 c_{np} 是一个序列满足 $c_{np} \rightarrow 0$ 且 $1/(\sqrt{n}c_{np}) = O(1)$.

假设 4.6 (信号) 定义集合 $\mathcal{C}_B = \{j \in \mathcal{A} : \|\mathbf{B}_j\|_2^2 / \max\{c_{np}^2, \bar{q}_n \log \bar{q}_n / n\} \rightarrow \infty\}$. 当 $(n, p) \rightarrow \infty$ 时, 假设 $\eta_n := |\mathcal{C}_B| \rightarrow \infty$.

假设 4.7 (相关性) 令 ρ_{jl} 表示在给定 $\mathcal{D}_1$ 时 W_j 和 W_l 之间的条件相关系数. 存在常数 $C > 0$, 对于每个 j , 有 $\#\{l \in \mathcal{A}^c : |\rho_{jl}| \geq C(\log n)^{-2-\nu}\} \leq r_p$, 其中 $\nu > 0$ 是某个较小的常数, 且随着 $(n, p) \rightarrow \infty$, 有 $r_p / \eta_n \rightarrow 0$.

注 4.1 假设 4.2 在文献 [2, 12, 26] 中也被使用, 该条件能够保证当 $j \in S$ 时 $\hat{\mathbf{B}}_{2j}$ 是无偏估计. 假设 4.3–4.4 常被用于变量选择的问题中. 假设 4.5 确保在 $\mathcal{D}_1$ 上 $\hat{\mathbf{B}}_1$ 是 $\mathbf{B}$ 一个合理的估计. 若采用 LASSO 进行变量筛选, 取 $c_{np} = d_n \sqrt{\log p / n}$ 可满足假设 4.5 的要求. 假设 4.6 表示随着 $p \rightarrow \infty$, 具有可识别的有效信号变量个数不能太少. 假设 4.7 允许 W_j 与其他所有变量相关, 但要求相关系数需要以对数速率收敛到 0. 这个条件类似于文献 [13] 中给出的弱相依结构.

定理 4.2 (渐近 FDR 控制) 若假设 4.1–4.7 和 $c_{np} a_{np} v_n \sqrt{n \bar{q}_n} (\log \bar{q}_n)^{3/2 + \gamma} \rightarrow 0$ 对于较小的 $\gamma > 0$ 成立, 则对于任意 $\alpha \in (0, 1)$, 由临界值 L 构建的 MFSDS 的 FDP 满足以下结论:

$$\text{FDP}(L) := \frac{\#\{j : W_j \geq L, j \in \mathcal{A}^c\}}{\#\{j : W_j \geq L\} \vee 1} \leq \alpha + o_p(1), \quad \limsup_{(n,p) \rightarrow \infty} \text{FDR} \leq \alpha.$$

定理 4.2 表明, 本文基于数据驱动临界值 L 所构造的变量选择方法可渐近控制 FDR 在给定的显著性水平 α 下. 我们需要对条件 $c_{np}a_{np}v_n\sqrt{n\bar{q}_n}(\log \bar{q}_n)^{3/2+\gamma} \rightarrow 0$ 作进一步的讨论来加深理解. 常规的估计 $\|\mathbf{A} - \boldsymbol{\Sigma}_S\|_\infty$ 表明 $a_{np} = O_p(v_n\sqrt{\log \bar{q}_n/n})$. 对于 LASSO 估计, 可知 $c_{np} = d_n\sqrt{\log p/n}$. 因此当 p 是 n 的多项式阶时, 上述条件可退化为 $d_nv_n^2\bar{q}_n/\sqrt{n} \rightarrow 0$, 且此条件仅由速度 d_n 、 v_n 和 $\bar{q}_n$ 控制着. 若进一步假设 d_n 和 v_n 是有界的, 则在 $\mathcal{D}_1$ 上需要满足 $\bar{q}_n = o(n^{1/2})$. 再者由于潜在的假设 $d_n \leq \bar{q}_n$ 成立, 所以 $\bar{q}_n = o(n^{1/2})$ 可以满足定理 4.2 中的条件. 该条件在有关参数数量发散的问题中通常是合理的^[17,31].

5 数值研究

我们分别在低维和高维情形下评估 MFSDS 在模拟数据集和真实数据集上的性能.

5.1 算法细节

考虑将以下方法作为基准与我们的 MFSDS 进行比较. 第一个是文献 [8] 中有关 SIR 的边际坐标检验 (marginal SIR), 该方法的目标是控制每个坐标上的错误发现率. 然后将 BH 方法^[4] 应用到所得的 p 值上以进行全局错误率控制 (MSIR-BH). 此方法是使用 R 软件包 `dr` 中的函数 “dr” 和 “drop1” 实现的. 第二种方法是 Model-X Knockoff (MX-Knockoff)^[6], 该方法也是一种能控制错误率且数据驱动的变量选择方法. 此方法由 R 软件包 `knockoff` 中的函数 “create.gaussian” 使用 LSM (LASSO signed maximum) 特征重要程度统计量实现. 这两种方法分别记为 MSIR-BH 和 MX-Knockoff.

FDR 控制的显著性水平设为 $\alpha = 0.2$. 所有的模拟结果均进行 500 次重复实验. 本文方法 MFSDS 与上述方法进行对比的度量标准包括 FDR、真阳率 (true positive rate, TPR) $\text{TPR} = \#\{j : j \in \hat{\mathcal{A}}(L) \cap \mathcal{A}\}/|\mathcal{A}|$ 、 $P_a = \Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}(L))$ 和平均计算时间.

5.2 数值模拟

5.2.1 低维研究

这里按照 3 个分布生成协变量 $\mathbf{X}$: 具有协方差阵 $\boldsymbol{\Sigma} = (\sigma_{ij}) = \rho^{|ij|}$, $1 \leq i, j \leq p$ 的多元正态分布 $\mathcal{N}(0, \boldsymbol{\Sigma})$; 具有协方差 $\boldsymbol{\Sigma}$ 的多元 $t(5)$ 分布; 由 3 个分布组成的混合分布, 其中 $\{X_j\}_{j=1}^{[p/3]}$ 来自 $\mathcal{N}(0, \boldsymbol{\Sigma}_{[p/3]})$, $\{X_j\}_{j=[p/3]+1}^{[2p/3]}$ 来自 $\mathcal{N}(0, [p/3])$, $\{X_j\}_{j=[2p/3]+1}^p$ 来自 $t(5)$ 分布. 误差项 η 是与 $\mathbf{X}$ 独立的标准正态分布. 固定 $(p, p_1) = (20, 10)$. 对于标量 c , 将 $\mathbf{c}_p = (c, \dots, c)$ 写成 c 的 p 维行向量. 考虑以下 5 个模型:

- 情景 1a: $Y = \boldsymbol{\beta}^\top \mathbf{X} + 3\eta$, 其中 $\boldsymbol{\beta} = (\mathbf{1}_{p_1}, \mathbf{0}_{p-p_1})^\top$.
- 情景 1b: $Y = |\boldsymbol{\beta}_1^\top \mathbf{X}| + \exp(3 + \boldsymbol{\beta}_2^\top \mathbf{X}) + \eta$, 其中 $\boldsymbol{\beta}_1 = (\mathbf{1}_5, \mathbf{0}_{p-5})^\top$, $\boldsymbol{\beta}_2 = (\mathbf{0}_5, \mathbf{1}_5, \mathbf{0}_{p-p_1})^\top$.
- 情景 1c: $Y = \boldsymbol{\beta}_1^\top \mathbf{X} + (\boldsymbol{\beta}_2^\top \mathbf{X} + 3)^2 + \exp(\boldsymbol{\beta}_3^\top \mathbf{X}) + \eta$, 其中 $\boldsymbol{\beta}_1 = (\mathbf{1}_3, \mathbf{0}_{p-3})^\top$, $\boldsymbol{\beta}_2 = (\mathbf{0}_3, \mathbf{1}_3, \mathbf{0}_{p-6})^\top$, $\boldsymbol{\beta}_3 = (\mathbf{0}_6, \mathbf{1}_4, \mathbf{0}_{p-p_1})^\top$.
- 情景 1d: $Y = \boldsymbol{\beta}_1^\top \mathbf{X} / \{0.5 + (1.5 + \boldsymbol{\beta}_2^\top \mathbf{X})^2\} + (\boldsymbol{\beta}_3^\top \mathbf{X})^2 + \exp(\boldsymbol{\beta}_4^\top \mathbf{X}) + \eta$, 其中 $\boldsymbol{\beta}_1 = (\mathbf{1}_2, \mathbf{0}_{p-2})^\top$, $\boldsymbol{\beta}_2 = (\mathbf{0}_2, \mathbf{1}_2, \mathbf{0}_{p-4})^\top$, $\boldsymbol{\beta}_3 = (\mathbf{0}_4, \mathbf{1}_2, \mathbf{0}_{p-6})^\top$, $\boldsymbol{\beta}_4 = (\mathbf{0}_6, \mathbf{1}_4, \mathbf{0}_{p-p_1})^\top$.
- 情景 1e: $Y = 2(\boldsymbol{\beta}_1^\top \mathbf{X})^2 \sin(\boldsymbol{\beta}_2^\top \mathbf{X}) + 3(\boldsymbol{\beta}_3^\top \mathbf{X})^3 \exp(\boldsymbol{\beta}_4^\top \mathbf{X}) + |\boldsymbol{\beta}_5^\top \mathbf{X}| \eta$, 其中 $\boldsymbol{\beta}_1 = (\mathbf{1}_2, \mathbf{0}_{p-2})^\top$, $\boldsymbol{\beta}_2 = (\mathbf{0}_2, \mathbf{1}_2, \mathbf{0}_{p-4})^\top$, $\boldsymbol{\beta}_3 = (\mathbf{0}_4, \mathbf{1}_2, \mathbf{0}_{p-6})^\top$, $\boldsymbol{\beta}_4 = (\mathbf{0}_6, \mathbf{1}_2, \mathbf{0}_{p-8})^\top$, $\boldsymbol{\beta}_5 = (\mathbf{0}_8, \mathbf{1}_2, \mathbf{0}_{p-p_1})^\top$.

对于本文的方法 MFSDS, 考虑以下 3 种响应变换函数 $f_h(Y)$, $h = 1, \dots, H$: (1) 文献 [22] 中的切片示性函数, 若 Y 在切片 h 中 $f_h(Y) = 1$, 否则为 0; (2) 文献 [9] 中的 CIRE (covariance inverse

regression estimation) 类型的响应转换, 若 Y 在切片 h 中 $f_h(Y) = Y$, 否则为 0; (3) 文献 [33] 中的标准化的多项式响应变换, 若 Y 在切片 h 中 $f_h(Y) = Y^m$, 否则为 0. 这里固定 $m = 2$. 本文将这 3 个函数分别称为 Indicator (示性函数)、CIRE (协方差逆回归) 和 Poly (多项式).

图 2 表明, 无论工作维数 H 和响应变换函数 $f(\cdot)$ 如何选取, MFSDS 都能成功将 FDR 控制在目标水平的可接受范围内. 因为 3 种响应变换函数在控制 FDR 时表现出相似的模式, 故本文中并未具体给出哪个 $f(Y)$ 效果是最好的. 本文的方法不需要对给定 H 的 d 进行估计. 因此, 在余下的模拟中, 对于 MFSDS 方法, 将固定 $f(Y)$ 为 Indicator 和 $H = 4$.

接下来在低维情形下, 在表 1 和 2 中比较了本文的 MFSDS 方法与 MSIR-BH 和 MX-Knockoff 的 FDR 和 TPR. 表 1 研究了所提出的 MFSDS 和基准方法如何受到协变量分布和样本量大小的影响. 表 2 展示了 3 种方法在不同协变量相关性下 FDR 和 TPR 的比较. 在所有模型中, MFSDS 的 FDR 始终保持在目标水平附近. 并且在大多数情形下, MFSDS 的 TPR 高于 MSIR-BH 和 MX-Knockoff. 我们还给出以下讨论.

(1) MFSDS vs. MSIR-BH. 基于 SIR 的边际坐标检验方法或者 MSIR, 在所有模型下都能够控制 FDR 但却非常保守. 这是因为 BH 方法在低维情形下控制的 FDR 的水平为 $\alpha p_0/p$. 相比之下, MFSDS 在线性和非线性模型 [6] 中能够得到准确的 FDR 控制和更高的 TPR. 在极少数情形下, 所提出的 MFSDS 方法的检验功效似乎略低于 MSIR-BH, 这是因为 MFSDS 涉及数据分割用于构建统计量的对称性. 但随着样本量 n 的增加, 本文的 MFSDS 在低维情形中比其他方法更有效.

(2) MFSDS vs. MX-Knockoff. MFSDS 和 MX-Knockoff 都是误差可控且模型自由的变量选择方法. 但随着样本量 n 的增加, MFSDS 能更准确地控制 FDR 接近给定的显著性水平. 而 MX-Knockoff 在 $t(5)$ 分布时的 FDR 控制失败, 如表 1 的第 2 列所示. 表 1 还表明 MX-Knockoff 在非线性情形下的



图 2 (网络版彩图) 给定 $(n, \rho) = (500, 0.5)$ 且 X 服从多元正态分布. 在情景 1a–1c 中, MFSDS 在不同 H 和不同 $f_h(Y)$ 时的 FDR 和 TPR (%). 灰实线表示给定的显著性水平 $\alpha = 0.2$

TPR 不如本文的方法. 表 2 表明 MFSDS 在给定的显著性水平下能够稳健地控制 FDR. 但是当所有模型的相关性变得更高时, MX-Knockoff 却表现出更保守的 FDR 和较低的 TPR.

表 1 给定 $(n, \rho) = (500, 0.5)$. 情景 1a–1e, 不同变量选择方法不同协变量分布下的 FDR 和 TPR (%)

情景	方法	正态分布		$t(5)$		混合分布	
		FDR	TPR	FDR	TPR	FDR	TPR
1a	MFSDS	20.7	87.1	22.0	84.4	22.2	96.4
	MSIR-BH	10.4	91.4	10.1	84.6	10.8	98.2
	MX-Knockoff	21.0	99.9	22.4	100.0	22.8	100.0
1b	MFSDS	21.8	85.6	19.1	70.4	21.5	89.7
	MSIR-BH	10.0	83.1	13.1	68.3	10.5	88.9
	MX-Knockoff	13.3	39.9	39.4	50.6	16.3	58.9
1c	MFSDS	21.3	78.5	20.9	68.9	20.9	84.2
	MSIR-BH	10.5	74.8	14.1	68.1	11.2	80.9
	MX-Knockoff	13.6	46.0	38.8	52.4	19.1	82.2
1d	MFSDS	18.6	61.9	20.0	51.9	20.0	61.9
	MSIR-BH	10.6	57.4	14.1	48.5	11.6	56.0
	MX-Knockoff	14.7	40.2	38.9	52.5	18.6	51.9
1e	MFSDS	20.2	57.1	18.7	49.5	18.0	58.2
	MSIR-BH	11.7	51.1	12.9	46.2	10.9	52.8
	MX-Knockoff	14.5	40.0	38.8	52.6	16.0	33.4

表 2 给定 $n = 500$ 且 X 服从多元正态分布. 在情景 1a–1e 中, 不同变量选择方法在不同相关系数下的 FDR 和 TPR (%)

情景	方法	$\rho = 0.2$		$\rho = 0.5$		$\rho = 0.8$	
		FDR	TPR	FDR	TPR	FDR	TPR
1a	MFSDS	22.6	99.1	20.7	87.1	14.6	27.7
	MSIR-BH	11.0	99.9	10.4	91.4	14.0	10.0
	MX-Knockoff	22.8	100.0	21.0	99.9	20.6	95.6
1b	MFSDS	22.9	99.5	21.8	85.6	21.6	30.2
	MSIR-BH	11.2	99.8	10.0	83.1	12.0	16.5
	MX-Knockoff	15.9	54.9	13.3	39.9	10.1	20.7
1c	MFSDS	21.6	89.9	21.3	78.5	20.2	34.9
	MSIR-BH	11.2	85.9	10.5	74.8	11.7	18.4
	MX-Knockoff	17.5	69.3	13.6	46.0	11.1	22.6
1d	MFSDS	18.2	64.8	18.6	61.9	20.0	40.9
	MSIR-BH	10.2	61.6	10.6	57.4	12.5	31.2
	MX-Knockoff	16.7	48.0	14.7	40.2	11.5	22.5
1e	MFSDS	19.1	64.7	20.2	57.1	18.1	45.6
	MSIR-BH	11.2	61.2	11.7	51.1	9.6	40.2
	MX-Knockoff	17.0	38.3	14.5	40.0	10.6	25.1

5.2.2 高维研究

在高维情形下, 考虑以下方法作为基准. 首先因为无法直接在高维模型中获得 p 值, 低维时的对比方法 MSIR-BH 在高维中无法使用. 因此, 这里考虑数据分割方法^[30], 该方法先使用 LASSO 进行变量筛选, 然后将 BH 应用于通过边际 SIR^[8] 计算的 p 值上. 这里采用交叉验证准则 (cross-validation, CV) 用于变量筛选阶段时 LASSO 中调节参数的选取以得到过拟合的模型. MX-Knockoff 由 R 包 `knockoff` 中的函数 `create.second_order` 执行, 来近似高维情形^[6] 中的精度矩阵的有效估计. 下一个方法被命名为 MFSDS-DB^[19] (debiasing), 该方法主要是将 (2.2) 中的最小二乘解扩展到 (3.3) 中带有 L_1 惩罚的正则化版本, 并使用 R 软件包 `selectiveInference` 进行偏差校正. 第 4 个方法是文献 [35] 提出的边际独立 SIR. 这里挑选两个模型大小 $\lfloor cn/\log(n) \rfloor$, $c = 0.05, 0.50$, 分别记为 MI-SIR1 和 MI-SIR2. 记 a 为信号强度. 当 $p > n$ 时, 考虑以下模型:

- 情景 2a: $Y = a \cdot \exp(5 + \beta^\top X) + \eta$, 其中 $\beta = (\mathbf{1}_{p_1}, \mathbf{0}_{p-p_1})^\top$.
- 情景 2b: $Y = a \cdot \{2\beta_1^\top X + 3\exp(\beta_2^\top X)\} + \eta$, 其中 $\beta_1 = (\mathbf{1}_5, \mathbf{0}_{p-5})^\top$, $\beta_2 = (\mathbf{0}_5, \mathbf{1}_5, \mathbf{0}_{p-p_1})^\top$.
- 情景 2c: $Y = a \cdot \{\beta_1^\top X + |\beta_2^\top X + 5| + \exp(\beta_3^\top X)\} + \eta$, 其中 $\beta_1 = (\mathbf{1}_3, \mathbf{0}_{p-3})^\top$, $\beta_2 = (\mathbf{0}_3, \mathbf{1}_3, \mathbf{0}_{p-6})^\top$, $\beta_3 = (\mathbf{0}_6, \mathbf{1}_4, \mathbf{0}_{p-p_1})^\top$.

为了研究高维情形下的错误率控制和变量选择的一致性, 表 3 给出了不同协变量分布下的比较结果. 本文方法 MFSDS 的 FDR 以更高的概率近似控制在目标水平 α , 这与我们的理论结果一致. 同时, 表 3 显示, MFSDS 的 P_α (选择出所有信号变量的比例) 大于其他方法. 这意味着当 n 较大时, MFSDS 可以实现变量选择的一致性. 正如我们所期待的, 虽然 MFSDS 涉及数据分割可能会丢失一些数据信息但其功效也可以很高, 因为变量筛选阶段会显著提高信噪比. 此外, 本文的方法没有像边际坐标检验^[8] 中为每个协变量构建渐近分布, 也没有像 MX-Knockoff^[6] 中生成变量副本, 故本文的方法耗时应该最短. 我们还进行以下探讨.

(1) MFSDS vs. MFSDS-DB. MFSDS-DB 方法与本文方法 MFSDS 的 FDR 都控制得相当好, 但它的检验功效却相对较低, 这是因为 MFSDS-DB 中用纠偏过程代替了变量筛选阶段导致无法提高信噪比. 已知 MFSDS-DB 是 MFSDS 低维方法基于纠偏 LASSO 估计的高维扩展, 该方法涉及精度矩阵的估计从而引入更多计算成本.

(2) MFSDS vs. MSIR-BH. MSIR-BH 方法实现了比 MFSDS 更加保守的 FDR 控制. 该方法也采用了数据分割的思想^[26], 但仅在 $\mathcal{D}_2$ 上构建检验统计量, 从而导致检验功效和选择一致性严重受损. 文献 [18] 基于全样本提供了对 MSIR-BH 的一个很好的改进, 但求解过程与 MFSDS-DB 中的去偏过程类似且复杂.

(3) MFSDS vs. MI-SIR. 仅有当模型大小 $\lfloor cn/\log(n) \rfloor$ 大于信号个数 p_1 时, 硬临界值 MI-SIR 方法才能检测到更多的信号变量. 表 3 表明硬临界值方法无法通过用户人为指定的临界值控制 FDR 并保证模型的选择一致性.

(4) MFSDS vs. MX-Knockoff. MX-Knockoff 是专门为模型自由的高维数据变量选择而设计的. 但 MX-Knockoff 需要假设 X 的分布已知或可准确估计, 而且无法控制混合分布下的 FDR, 这与低维时的结论一致. 相比之下, MFSDS 在混合分布和 $t(5)$ 下能更准确地控制 FDR. 此外, MFSDS 的对称性属性源自样本分割, 而 MX-Knockoff 则需要扩大变量维数, 进而需求更大的计算量.

图 3 进一步研究了 MFSDS 在高维情形下针对不同协变量维数 p 、信号个数 p_1 、协变量相关性 ρ 和信号强度 a 的效果. 图 3 表明, 无论维数、信号数量、相关性及信号强度如何在合理的变化范围内变动, MFSDS 的 FDR 在目标水平的可接受范围内都是稳健的. 由于 MI-SIR2 无法实现有效的 FDR 控

表 3 给定 $(n, p, p_1, \rho, a) = (500, 1000, 10, 0.5, 1.0)$. 在情景 2a–2c 中, 不同变量选择方法在不同协变量分布下的 FDR、TPR、 P_a (%) 和计算时间 (秒)

情景	方法	正态分布				混合分布			
		FDR	TPR	P_a	计算时间	FDR	TPR	P_a	计算时间
2a	MFSDS	18.3	98.7	90.2	12.9	17.5	98.3	86.6	14.1
	MFSDS-DB	17.1	57.9	7.0	75.1	17.0	77.3	19.8	65.5
	MSIR-BH	4.5	32.8	0.0	11.4	4.2	32.9	0.0	12.6
	MI-SIR1	0.0	40.0	0.0	26.3	0.0	40.0	0.0	34.8
	MI-SIR2	75.0	100.0	100.0	26.3	75.0	100.0	100.0	34.5
	MX-Knockoff	6.5	4.3	0.0	31.8	26.4	9.2	0.0	31.6
2b	MFSDS	17.0	94.7	62.0	12.8	17.5	94.6	58.8	14.7
	MFSDS-DB	15.6	71.4	4.2	74.8	16.9	80.5	14.4	75.3
	MSIR-BH	5.3	34.9	0.0	11.5	4.6	34.6	0.0	13.3
	MI-SIR1	0.0	40.0	0.0	26.3	0.0	40.0	0.0	27.3
	MI-SIR2	75.0	100.0	100.0	26.3	75.0	100.0	100.0	27.1
	MX-Knockoff	10.3	11.6	0.0	31.9	29.0	19.8	0.0	31.4
2c	MFSDS	17.9	92.7	50.6	12.4	19.2	92.8	49.8	22.8
	MFSDS-DB	17.0	62.1	8.0	102.3	17.4	73.2	6.0	126.8
	MSIR-BH	6.5	33.4	0.0	12.2	6.3	31.1	0.0	21.8
	MI-SIR1	0.0	40.0	0.0	22.7	0.0	40.0	0.0	36.2
	MI-SIR2	75.0	100.0	100.0	22.9	75.0	100.0	100.0	35.2
	MX-Knockoff	12.1	12.2	0.0	28.4	28.8	19.1	0.0	42.0

制, 所以本文所提出方法 MFSDS 总是比除了 MI-SIR2 以外的其他方法实现最高的检验功效. 从图 3 可知, 虽然较大的 p_1 将 FDR 控制在目标水平以下, 但 MX-Knockoff 遭受更大的检验功效损失, 因为随着信号数 p_1 和相关性 ρ 的增加检验功效会被破坏. 可在图 B1 中找到具有正态协变量分布的补充数值模拟结果. 图 B1 表示虽然 MX-Knockoff 的 FDR 在不同 p 、 ρ 、 p_1 以及 a 在正态分布下控制与 MFSDS 一样好, 但也与 MFSDS 的检验功效存在较大差距.

5.3 实例应用

本小节将本文方法 MFSDS 应用于儿童癌症数据对小圆蓝细胞肿瘤 (small round blue cell tumors, SRBCT) 进行研究. 该数据已在文献 [20, 35] 分析过. SRBCT 数据集着眼于常规组织学中具有相似视觉特征的 4 类不同儿童肿瘤进行分类. 数据是从 83 个肿瘤样本中收集而来, 并为每个样本的 2,308 个基因的提供了表达测量值. 本文的目标是基于这个数据探究 MFSDS 与其他方法的基因识别和 FDR 控制性能. 图 4 表明, MFSDS 和 MX-Knockoff 的排序统计量 W_j 的散点图正如我们所期待, 具有较大 W_j 值的基因可能会被选为有影响力的基因 (红色点), 而未被选中的基因 (黑色点) 即使许多 W_j 的值确实为零也大致上呈对称分布在零线附近.

接下来, 通过引入一些模拟变量作为不重要的基因, 以进一步研究本文提出的方法的性能. 具体而言, 这里增加了 2,000 个独立生成的噪声变量, 其中有 1,000 个噪声变量 Z_1 来自 $\mathcal{N}(0, 1)$, 另外 1,000 个噪声变量 Z_2 来自 $t(3)$. 将加入噪声变量后的数据集命名为 SRBCT&Noise, 主要目标是从整个 4,328

个变量中识别出对引发癌症比较重要的基因. 由于 Z_1 和 Z_2 在实际构造时就是噪声变量, 因此它们可直接被视为真正的错误发现. W_j 的散点图和被识别的基因数量如图 4 和表 4 所示. 文献 [35] 在

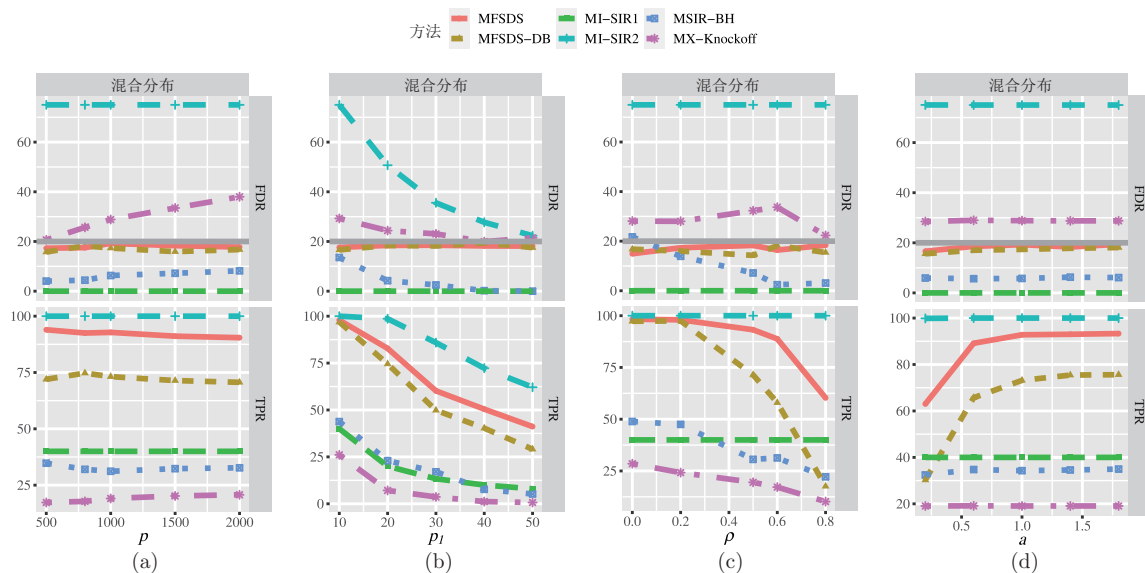


图 3 (网络版彩图) 给定 $n = 500$ 且 X 服从混合分布. 在情景 2c 中, 不同检验方法随不同协变量维数 p 、不同信号个数 p_1 不同协变量相关性 ρ 和不同的信号强度 a 的 FDR 和 TPR (%) 曲线. (a) $(p_1, \rho, a) = (10, 0.5, 1.0)$; (b) $(p, \rho, a) = (1000, 0.5, 1.0)$; (c) $(p, p_1, a) = (1000, 10, 1.0)$; (d) $(p, p_1, \rho) = (1000, 10, 0.5)$. 灰实线表示给定的显著性水平 $\alpha = 0.2$

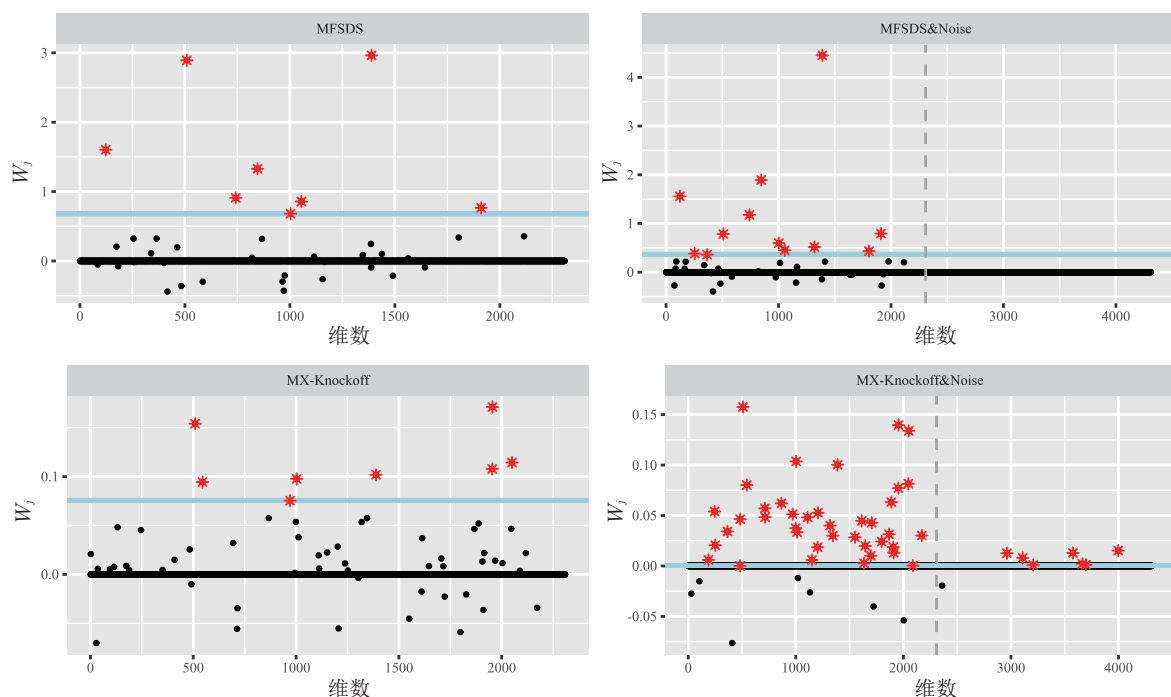


图 4 (网络版彩图) MFSDS 和 MX-Knockoff 的统计量 W_j 的散点图, 红点表示所识别的信号变量. 其中浅蓝线表示当 $\alpha = 0.2$ 时数据驱动临界值

表 4 不同变量选择方法在数据集 SRBCT 和 SRBCT & Noise 上拒绝的基因个数 (number of discoveries, ND) 和计算时间 (秒). 括号中数字表示在噪声变量中被识别出的信号个数

方法	SRBCT		SRBCT & Noise	
	ND	计算时间	ND	计算时间
MFSDS	8	1.52	12 (0)	1.92
MSIR-BH	5	0.50	5 (0)	0.72
MI-SIR ($c = 0.05$)	1	130.20	1 (0)	1,514.23
MI-SIR ($c = 1.00$)	18	128.31	18 (0)	1,516.59
MX-Knockoff	8	49.82	46 (7)	240.20

2,308 个基因中识别出 9 个重要基因, 这与本文的方法选择的变量数量非常接近. 我们观察到 SDS 和 MX-Knockoff 在没有加入噪声变量时识别出相同个数的重要基因, 但当加入噪声变量后 MX-Knockoff 会错误地选择出少量的噪声变量. 此外, MFSDS 比 MI-SIR 和 MX-Knockoff 需求相对较少的计算时间.

6 结论

识别真正有用的变量是统计分析中一项非常重要的任务. 本文通过数据分割技术在充分降维框架中构造了一种新的误差可控且模型自由的变量选择方法. 本文所提出的 MFSDS 在计算上是高效的, 并且不涉及任何高维情形中需要指定的讨厌参数. 理论和数值结果均表明, 所提出的 MFSDS 可以在目标水平上渐近地控制 FDR, 并在一些较弱条件下以很高的概率选出所有信号变量.

我们用几个值得继续深入研究的问题来总结本文. 第一, 研究人员也可将 MFSDS 的思想进行扩展以控制其他类型的错误率, 如家族错误率 (family-wise error rate) 和边际错误发现率 (marginal false discovery rate). 第二, 为简单起见, 本文的模拟和理论中使用了等样本量的单次数据分割. 可进一步在 MFSDS 中采用多次数据分割技术以提高误差控制的稳健性, 但这可能会引入更多的计算量. 另外, 我们已经测试过在 $\mathcal{D}_1$ 中分配较多的样本会在一定程度上提高检验功效. 但如何在没有数据分割或推导复杂的渐近分布的情形下构建对称性质可能是另一个有前景的研究课题. 第三, 得益于响应变换函数, 本文的方法同样适用于响应变量是离散、连续或分类变量的数据. 最后, 如果 H 个切片的重要性已经预先确定, 该方法或许还可以解决大多数模型中方向个数 d 的确定问题 (如文献 [25]). 但最优工作维数 H 的选择也值得进一步探索.

致谢 作者衷心感谢主编、副主编及两位审稿人提出的宝贵建议.

参考文献

1 Barber R F, Candès E J. Controlling the false discovery rate via knockoffs. *Ann Statist*, 2015, 43: 2055–2085
2 Barber R F, Candès E J. A knockoff filter for high-dimensional selective inference. *Ann Statist*, 2019, 47: 2504–2537
3 Barber R F, Candès E J, Samworth R J. Robust inference with knockoffs. *Ann Statist*, 2020, 48: 1409–1431
4 Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol*, 1995, 57: 289–300
5 Cai T T, Liu W D. Large-scale multiple testing of correlations. *J Amer Statist Assoc*, 2016, 111: 229–240
6 Candès E, Fan Y Y, Janson L, et al. Panning for gold: “Model-X” knockoffs for high dimensional controlled variable selection. *J R Stat Soc Ser B Stat Methodol*, 2018, 80: 551–577
7 Chen X, Zou C L, Cook R D. Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann Statist*, 2010, 38: 3696–3723

- 8 Cook R D. Testing predictor contributions in sufficient dimension reduction. *Ann Statist*, 2004, 32: 1062–1092
- 9 Cook R D, Ni L Q. Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika*, 2006, 93: 65–74
- 10 Cook R D, Weisberg S. Sliced inverse regression for dimension reduction: Comment. *J Amer Statist Assoc*, 1991, 86: 328–332
- 11 Dong Y X. A brief review of linear sufficient dimension reduction through optimization. *J Statist Plann Inference*, 2021, 211: 154–161
- 12 Du L L, Guo X, Sun W G, et al. False discovery rate control under general dependence by symmetrized data aggregation. *J Amer Statist Assoc*, 2023, 118: 607–621
- 13 Fan J Q, Han X, Gu W J. Estimating false discovery proportion under arbitrary covariance dependence. *J Amer Statist Assoc*, 2012, 107: 1019–1035
- 14 Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, 2001, 96: 1348–1360
- 15 Fan J Q, Li R Z, Zhang C-H, et al. *Statistical Foundations of Data Science*. Boca Raton: CRC Press, 2020
- 16 Fan J Q, Lv J C. A selective overview of variable selection in high dimensional feature space. *Statist Sinica*, 2010, 20: 101–148
- 17 Fan J Q, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann Statist*, 2004, 32: 928–961
- 18 Guo X, Li R Z, Zhang Z, et al. Model-free statistical inference on high-dimensional data. *J Amer Statist Assoc*, in press, 2024
- 19 Javanmar A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. *J Mach Learn Res*, 2014, 15: 2869–2909
- 20 Khan J, Wei J S, Ringnér M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, 2001, 7: 673–679
- 21 Li B, Wang S L. On directional regression for dimension reduction. *J Amer Statist Assoc*, 2007, 102: 997–1008
- 22 Li K C. Sliced inverse regression for dimension reduction. *J Amer Statist Assoc*, 1991, 86: 316–327
- 23 Li L X, Cook R D, Nachtshiem C J. Model-free variable selection. *J R Stat Soc Ser B Stat Methodol*, 2005, 67: 285–299
- 24 Li L X, Wen X M, Yu Z. A selective overview of sparse sufficient dimension reduction. *Statist Theory Related Fields*, 2020, 4: 121–133
- 25 Luo W, Li B. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 2016, 103: 875–887
- 26 Meinshausen N, Meier L, Bühlmann P. p -values for high-dimensional regression. *J Amer Statist Assoc*, 2009, 104: 1671–1681
- 27 Shao Y W, Cook R D, Weisberg S. Marginal tests with sliced average variance estimation. *Biometrika*, 2007, 94: 285–296
- 28 Svetulevičienė V. Multidimensional local limit theorems for probabilities of moderate deviations. *Lith Math J*, 1982, 22: 416–420
- 29 Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B Stat Methodol*, 1996, 58: 267–288
- 30 Wasserman L, Roeder K. High-dimensional variable selection. *Ann Statist*, 2009, 37: 2178–2201
- 31 Wu Y C, Li L X. Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statist Sinica*, 2011, 2011: 707–730
- 32 Xia Y C, Tong H, Li W K, et al. An adaptive estimation of dimension reduction space. *J R Stat Soc Ser B Stat Methodol*, 2002, 64: 363–410
- 33 Yin X R, Cook R D. Dimension reduction for the conditional k th moment in regression. *J R Stat Soc Ser B Stat Methodol*, 2002, 64: 159–175
- 34 Yu Z, Dong Y X. Model-free coordinate test and variable selection via directional regression. *Statist Sinica*, 2016, 26: 1159–1174
- 35 Yu Z, Dong Y X, Shao J. On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection. *Ann Statist*, 2016, 44: 2594–2623
- 36 Yu Z, Dong Y X, Zhu L X. Trace pursuit: A general framework for model-free variable selection. *J Amer Statist Assoc*, 2016, 111: 813–821
- 37 Zhu L P. Review of sparse sufficient dimension reduction: Comment. *Statist Theory Related Fields*, 2020, 4: 134–134
- 38 Zou H. The adaptive lasso and its oracle properties. *J Amer Statist Assoc*, 2006, 101: 1418–1429

附录 A

本附录给出一些证明中用到的关键性引理和定理的简要证明, 附录中使用的额外引理及其证明和一些补充的模拟结果都作为补充材料整理在附录 B. 我们仅考虑高维情形下的证明, 低维版本可类比验证.

对于任意向量 $\mathbf{v} \in \mathbb{R}^H$, 有

$$\begin{aligned} (\hat{\mathbf{B}}_{2j} - \mathbf{B}_j)^\top \mathbf{v} &= \mathbf{e}_j^\top \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_{2Si} \mathbf{X}_{2Si}^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{2Si} (f_{2i} - \mathbf{B}_0^\top \mathbf{X}_{2Si})^\top \mathbf{v} \\ &= \mathbf{e}_j^\top \left(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_{2Si} \mathbf{X}_{2Si}^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{2Si} \boldsymbol{\varepsilon}_i^\top \mathbf{v}, \end{aligned}$$

其中 $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_H)^\top = \mathbf{f} - \mathbf{B}_0^\top \mathbf{X}_S \in \mathbb{R}^H$ 且 $\mathbb{E}(\mathbf{X}_S \boldsymbol{\varepsilon}^\top) = 0$. 然而由于使用了响应变换函数, $\boldsymbol{\varepsilon}$ 与 $\mathbf{X}_S$ 并不独立或者 $\mathbb{E}(\boldsymbol{\varepsilon} | \mathbf{X}_S) = 0$ 并不成立, 这与通常的设置大有不同. 为了建立 W_j 的 FDR 控制, 构造以下修改版的统计量:

$$\widetilde{W}_j = \frac{\hat{\mathbf{B}}_{1j}^\top \widetilde{\mathbf{B}}_{2j}}{s_{1Sj} \widetilde{s}_{2Sj}}, \quad j = 1, \dots, p.$$

这里当 $j \in \mathcal{S}$ 时, $\widetilde{\mathbf{B}}_{2j} = \mathbf{e}_j^\top n^{-1} \boldsymbol{\Sigma}_S^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \boldsymbol{\varepsilon}_{2i}^\top$ 且 $\widetilde{s}_{2Sj}^2 = n^{-1} \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j$, 否则 $\widetilde{\mathbf{B}}_{2j} = 0$ 且 $\widetilde{s}_{2Sj}^2 = 0$. 令 $G(t) = q_{0n}^{-1} \sum_{j \in \mathcal{A}^c} \Pr(\widetilde{W}_j \geq t | \mathcal{D}_1)$, $G_-(t) = q_{0n}^{-1} \sum_{j \in \mathcal{A}^c} \Pr(\widetilde{W}_j \leq -t | \mathcal{D}_1)$ 且 $G^{-1}(y) = \inf\{t \geq 0 : G(t) \leq y\}$ 对于 $0 \leq y \leq 1$ 成立. 在给出主要定理的证明之前, 首先陈述两个关键引理. 这些引理的证明可参见附录 B. 第一个引理表征 $G(t)$ 和 $G_-(t)$ 之间的接近程度, 这在证明中起着重要作用.

引理 A.1 (对称性) 若假设 4.1–4.4 成立, 则对于任意 $0 \leq t \leq G^{-1}(\alpha\eta_n/q_{0n})$, 有

$$\frac{G(t)}{G_-(t)} - 1 \rightarrow 0.$$

引理 A.2 (一致收敛性) 若假设 4.1–4.4 和 4.7 成立, 则有

$$\sup_{0 \leq t \leq G^{-1}(\alpha\eta_n/q_{0n})} \left| \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(\widetilde{W}_j \geq t)}{q_{0n} G(t)} - 1 \right| = o_p(1), \quad (\text{A.1})$$

$$\sup_{0 \leq t \leq G^{-1}(\alpha\eta_n/q_{0n})} \left| \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(\widetilde{W}_j \leq -t)}{q_{0n} G_-(t)} - 1 \right| = o_p(1). \quad (\text{A.2})$$

引理 A.3 若假设 4.1–4.5 成立, 则 $W_j - \widetilde{W}_j = O_p(c_{np} a_{np} v_n \sqrt{n \bar{q}_n \log \bar{q}_n})$.

引理 A.4 若假设 4.1–4.5 成立, 则对较小的 $\gamma > 0$, 有 $c_{np} a_{np} v_n \sqrt{n \bar{q}_n (\log \bar{q}_n)^{3/2+\gamma}} \rightarrow 0$. 对 $M > 0$, 有

$$\begin{aligned} \sup_{M \leq t \leq G^{-1}(\alpha\eta_n/q_{0n})} \left| \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(\widetilde{W}_j \geq t)}{\sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \geq t)} - 1 \right| &= o_p(1), \\ \sup_{M \leq t \leq G^{-1}(\alpha\eta_n/q_{0n})} \left| \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(\widetilde{W}_j \leq -t)}{\sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \leq -t)} - 1 \right| &= o_p(1). \end{aligned}$$

定理 4.1 的证明 我们基于 L_+ 来证明有限样本下的 FDR 控制. 有关 L 的结论可以类似得出. 证明过程中的主要技巧来自于文献 [3, 12]. 固定 $\epsilon > 0$ 且对于任意的临界值 $t > 0$, 定义

$$R_\epsilon(t) = \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \geq t, \Delta_j \leq \epsilon)}{1 + \sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \leq t)}.$$

定义事件 $\mathcal{A} = \{\Delta \equiv \max_{j \in \mathcal{A}^c} \Delta_j \leq \epsilon\}$. 进一步考虑一个将统计量 $\mathbf{W}$ 映射到临界值 $L \geq 0$ 的阈值规则 $L = T(\mathbf{W})$. 定义 $L_j = T(W_1, \dots, W_{j-1}, |W_j|, W_{j+1}, \dots, W_p) \geq 0$. 则对 MFSDS 的临界值 L_+ , 有

$$\frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \geq L_+, \Delta_j \leq \epsilon)}{1 \vee \sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \leq L_+)} = \frac{1 + \sum_j \mathbb{I}(W_j \leq -L_+)}{1 \vee \sum_j \mathbb{I}(W_j \geq L_+)} \times \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \geq L_+, \Delta_j \leq \epsilon)}{1 + \sum_j \mathbb{I}(W_j \leq -L_+)} \leq \alpha \times R_\epsilon(L_+).$$

接下来, 将为 $\mathbb{E}\{R_\epsilon(L_+)\}$ 推导出一个上界. 注意到

$$\begin{aligned} \mathbb{E}\{R_\epsilon(L_+)\} &= \mathbb{E}\left\{\frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \geq L, \Delta_j \leq \epsilon)}{1 + \sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \leq -L_+)}\right\} \\ &= \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{\mathbb{I}(W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_j)}\right\} \\ &= \sum_{j \in \mathcal{A}^c} \mathbb{E}\left[\mathbb{E}\left\{\frac{\mathbb{I}(W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_j)} \mid |W_j|, \mathbf{W}_{-j}\right\}\right] \\ &= \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{\Pr(W_j > 0 \mid |W_j|, \mathbf{W}_{-j}) \mathbb{I}(|W_j| \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_j)}\right\}, \end{aligned}$$

其中最后一个等式成立是因为在给定 $(|W_j|, \mathbf{W}_{-j})$ 时, 唯一未知的量仅有 W_j 的符号. 回顾记号 $\Delta_j = |\Pr(W_j > 0 \mid |W_j|, \mathbf{W}_{-j}) - 1/2|$, 有 $\Pr(W_j > 0 \mid |W_j|, \mathbf{W}_{-j}) \leq 1/2 + \Delta_j$. 因此,

$$\begin{aligned} \mathbb{E}\{R_\epsilon(L_+)\} &\leq \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{(1/2 + \Delta_j) \mathbb{I}(|W_j| \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_j)}\right\} \\ &= \left(\frac{1}{2} + \Delta_j\right) \left[\sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{\mathbb{I}(W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_j)}\right\} \right. \\ &\quad \left. + \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{\mathbb{I}(W_j \leq -L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_j)}\right\} \right] \\ &\leq \left(\frac{1}{2} + \Delta_j\right) \left[\sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{\mathbb{I}(W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_j)}\right\} \right. \\ &\quad \left. + \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{\mathbb{I}(W_j \leq -L_j)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_j)}\right\} \right] \\ &= \left(\frac{1}{2} + \Delta_j\right) \left[\mathbb{E}\{R_\epsilon(L_+)\} + \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{\mathbb{I}(W_j \leq -L_j)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_j)}\right\} \right]. \quad (\text{A.3}) \end{aligned}$$

(A.3) 中最后一个求和可进一步简化. 如果在零假设下的 j , 满足 $W_j > -L_j$, 则该求和为零. 否则

$$\sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{\mathbb{I}(W_j \leq -L_j)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_j)}\right\} = \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{\mathbb{I}(W_j \leq -L_j)}{1 + \sum_{k \in \mathcal{A}^c, k \neq j} \mathbb{I}(W_k \leq -L_k)}\right\} = 1,$$

其中第一个等号成立是因为对于任意 j 和 k , 当 $W_j \leq \min(L_j, L_k)$ 且 $W_k \leq -\min(L_j, L_k)$, 有 $L_j = L_k$ (详情参见文献 [3]). 相应地, 有 $\mathbb{E}\{R_\epsilon(L_+)\} \leq (1/2 + \epsilon)[\mathbb{E}\{R_\epsilon(L_+)\} + 1]$. 据此可有

$$\mathbb{E}\{R_\epsilon(L_+)\} \leq \frac{1/2 + \epsilon}{1/2 - \epsilon} \leq 1 + 4\epsilon.$$

综上所述, 该定理证毕. $\square$

定理 4.2 的证明 根据定义, 本文的检验等价于挑选 $W_j \geq L$ 的变量, 其中

$$L = \inf \left\{ t \geq 0 : \sum_j \mathbb{I}(W_j \leq -t) \leq \alpha \left(1 \vee \sum_j \mathbb{I}(W_j \geq t) \right) \right\}.$$

为了应用引理 A.1–A.4, 需要为该临界值 L 建立渐近上界. 令 $t^* = G^{-1}(\alpha\eta_n/q_{0n})$. 根据引理 A.2 和 A.4, 它遵循

$$\frac{\alpha\eta_n}{q_{0n}} = G_-(t^*) = \frac{1}{q_{0n}} \sum_{j \in \mathcal{A}^c} \mathbb{I}(\widetilde{W}_j < -t^*) \{1 + o(1)\} = \frac{1}{q_{0n}} \sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j < -t^*) \{1 + o(1)\}.$$

另外, 对于任意 $j \in \mathcal{C}_B$ 和 $\eta_n = |\mathcal{C}_B|$, 能够证明 $\Pr(W_j < t^*, j \in \mathcal{C}_B) \rightarrow 0$. 其实, 容易得到

$$\begin{aligned} & \Pr(W_j < t^*, \text{ 对于某个 } j \in \mathcal{C}_B) \\ & \leq \eta_{n,p} \Pr(\widehat{\mathbf{B}}_{1j}^\top \widehat{\mathbf{B}}_{2j} / (s_{1j}s_{2j}) - \mathbf{B}_j^\top \mathbf{B}_j / (s_{1j}s_{2j}) < t^* - \mathbf{B}_j^\top \mathbf{B}_j / (s_{1j}s_{2j})) \\ & \leq \eta_{n,p} \Pr(|\widehat{\mathbf{B}}_{1j}^\top (\widehat{\mathbf{B}}_{1j} - \mathbf{B}_j + \widehat{\mathbf{B}}_{2j} - \mathbf{B}_j)| + |(\widehat{\mathbf{B}}_{1j} - \mathbf{B}_j)^\top (\widehat{\mathbf{B}}_{2j} - \mathbf{B}_j)| > \mathbf{B}_j^\top \mathbf{B}_j - t^* s_{1j}s_{2j}). \end{aligned}$$

为了证明最后一个不等式, 定义 $d_j = \|\mathbf{B}_j\|_2^2 - t^* s_{1j}s_{2j}$. 根据假设 4.5, 有 $d_j = \|\mathbf{B}_j\|_2^2 \{1 + o(1)\}$. 从而可得

$$\begin{aligned} & \Pr(|\widehat{\mathbf{B}}_{1j}^\top (\widehat{\mathbf{B}}_{1j} - \mathbf{B}_j + \widehat{\mathbf{B}}_{2j} - \mathbf{B}_j)| + |(\widehat{\mathbf{B}}_{1j} - \mathbf{B}_j)^\top (\widehat{\mathbf{B}}_{2j} - \mathbf{B}_j)| > d_j) \\ & \leq \Pr(|\widehat{\mathbf{B}}_{1j}^\top (\widehat{\mathbf{B}}_{1j} - \mathbf{B}_j + \widehat{\mathbf{B}}_{2j} - \mathbf{B}_j)| > d_j/2) + \Pr(|(\widehat{\mathbf{B}}_{1j} - \mathbf{B}_j)^\top (\widehat{\mathbf{B}}_{2j} - \mathbf{B}_j)| > d_j/2) \\ & =: \Lambda_1 + \Lambda_2. \end{aligned}$$

根据假设 4.5 和补充材料中的引理 B.4, 可以证明

$$\begin{aligned} \Lambda_1 & \leq \Pr(\|\widehat{\mathbf{B}}_{1j} - \mathbf{B}_j\|_2 > d_j/(4\|\mathbf{B}_j\|_2)) + \Pr(\|\widehat{\mathbf{B}}_{2j} - \mathbf{B}_j\|_2 > d_j/(4\|\mathbf{B}_j\|_2)) \rightarrow 0, \\ \Lambda_2 & \leq \Pr(\|\widehat{\mathbf{B}}_{1j} - \mathbf{B}_j\|_2 > c_{np}) + \Pr(\|\widehat{\mathbf{B}}_{2j} - \mathbf{B}_j\|_2 > C\sqrt{\bar{q}_n \log \bar{q}_n/n}) \rightarrow 0. \end{aligned}$$

因此, 有 $\Pr(\sum_j \mathbb{I}(W_j > t^*) \geq \eta_n) \rightarrow 1$. 从而可以证明 $\sum_j \mathbb{I}(W_j < -t^*) \lesssim \alpha\eta_n \leq \alpha \sum_j \mathbb{I}(W_j > t^*)$. 至此, 找到临界值 L 的一个上界, $L \lesssim t^*$. 这意味着所提出的方法至少可以检测出 $\sum_j \mathbb{I}(W_j > t^*)$ 个信号. 由引理 A.1、A.2 和 A.4, 可得

$$\begin{aligned} & \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \geq L)}{\sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \leq -L)} - 1 \xrightarrow{p} 0, \\ \text{FDP} & = \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \geq L)}{1 \vee \sum_j \mathbb{I}(W_j \geq L)} = \frac{\sum_j \mathbb{I}(W_j \leq -L)}{1 \vee \sum_j \mathbb{I}(W_j \geq L)} \times \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \geq L)}{\sum_j \mathbb{I}(W_j \leq -L)} \leq \alpha \times R(L). \end{aligned} \tag{A.4}$$

注意到 $R(L) \leq \sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \geq L) / \sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \leq -L)$, 且根据 (A.4), 有 $\limsup_{n \rightarrow \infty} \text{FDP} \leq \alpha$ 依概率收敛. 因此定理 4.2 中的第一个结论得到证明. 进一步地, 对于任意 $\epsilon > 0$, 有

$$\text{FDR} \leq (1 + \epsilon)\alpha R(L) + \Pr(\text{FDP} \geq (1 + \epsilon)\alpha R(L)),$$

至此该定理的第二部分得到证明. $\square$

附录 B 补充材料

补充材料主要包含若干引理的证明和补充的随机模拟. 为了记号简洁, 常数 c 和 C 会被重复使用.

附录 B.1 补充的引理

第一个补充的引理是标准的 Bernstein 不等式. 第二个引理是有关均值的中度大偏差的结论 (参见文献 [28, 定理 1]).

引理 B.1 (Bernstein 不等式) 假设独立且中心化的随机序列 $X_1, \dots, X_n$ 几乎处处绝对有界, $A < \infty$. 令 $\sigma^2 = n^{-1} \sum_{i=1}^n \mathbb{E}(X_i^2)$. 则对于 $x > 0$, 有

$$\Pr\left(\sum_{i=1}^n X_i \geq x\right) \leq \exp\left(-\frac{x^2}{2n\sigma^2 + 2Ax/3}\right).$$

引理 B.2 (独立和的中度偏差) 假设 $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^s$ 是独立且同分布的 s 维随机向量, 具有零均值和标准化方差. 令 $\mathbb{E}(\|\mathbf{X}\|_2^q) < \infty$ 对于某个 $q > 2 + c^2$ 和 $c^2 > 0$ 成立. 令 $p_n(\mathbf{x})$ 是 $\sum_{i=1}^n \mathbf{X}_i/\sqrt{n}$ 的密度函数, 则对于定义域中的 $\|\mathbf{x}\|_2 \leq c\sqrt{\log n}$, 当 $n \rightarrow \infty$ 时, 有以下结论一致成立:

$$\frac{p_n(\mathbf{x})}{(2\pi)^{-s/2} \exp(-\|\mathbf{x}\|_2^2/2)} \rightarrow 1.$$

附录 B.2 有用的引理

为了简化符号, 此后 c 和 C 均表示常数. 以下引理为 $\tilde{\mathbf{B}}_{2j}$ $j = 1, \dots, p$ 建立了一致上界.

引理 B.3 若假设 4.1–4.4 成立, 则对于较大的 $C > 0$, 有

$$\Pr(\tilde{s}_{2Sj}^{-1} \|\tilde{\mathbf{B}}_{2j}\|_2 > \sigma \sqrt{C \log \bar{q}_n} \mid \mathcal{D}_1) = o(1/\bar{q}_n)$$

对于 $j \in \mathcal{S} \cap \mathcal{A}^c$ 一致成立, 这里 $\sigma^2 = H \max_{h=1}^H \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbb{E}[\mathbf{X}_S \mathbf{X}_S^\top \varepsilon_h^2] \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j / \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j$.

证明 注意到

$$\begin{aligned} \Pr(\tilde{s}_{2Sj}^{-1} \|\tilde{\mathbf{B}}_{2j}\|_2 > x) &= \Pr\left(\sum_{h=1}^H \tilde{\mathbf{B}}_{2j,h}^2 > x^2 \tilde{s}_{2Sj}^2\right) \\ &\leq \Pr(\tilde{\mathbf{B}}_{2j,h}^2 > x^2 \tilde{s}_{2Sj}^2 / H, \exists h = 1, \dots, H) \\ &\leq H \max_h \Pr(|\tilde{\mathbf{B}}_{2j,h}| > x \tilde{s}_{2j} / \sqrt{H}). \end{aligned}$$

已知 $\tilde{\mathbf{B}}_{2j,h} = \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} n^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \varepsilon_{ih}$. 对任意较小的 $\gamma > 0$, 令 $m_n = (n \bar{q}_n)^{1/\varpi + \gamma} K_n$. 记

$$\mathcal{B} = \left\{ \max_{1 \leq h \leq H} \max_{1 \leq i \leq n} \|\boldsymbol{\Sigma}_S^{-1} \mathbf{X}_{2Si} \varepsilon_{ih}\|_\infty \leq m_n \right\}.$$

首先, 在事件 $\mathcal{B}$ 上进行讨论. 定义

$$\sigma_h^2 = \max_j \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbb{E}[\mathbf{X}_S \mathbf{X}_S^\top \varepsilon_h^2] \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j,$$

故 $\sigma^2 = \sigma_h^2 H / (\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j)$. 由引理 B.1 中的 Bernstein 不等式, 有

$$\Pr(|\tilde{\mathbf{B}}_{2j,h}| > x \tilde{s}_{2Sj} / \sqrt{H} \text{ 对于某些 } j \mid \mathcal{D}_1)$$

$$\begin{aligned}
 &\leq q_n \max_{j \in \mathcal{S}} \Pr \left(\left| \sum_{i=1}^n \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{X}_{2Si} \varepsilon_{ih} \right| > nx \tilde{s}_{2Sj} / \sqrt{H} \mid \mathcal{D}_1 \right) \\
 &\leq 2\bar{q}_n \max_{j \in \mathcal{S}} \exp \left\{ - \frac{n^2 x^2 \tilde{s}_{2Sj}^2 / H}{2n\sigma_h^2 + 2nx \tilde{s}_{2Sj} / \sqrt{H} \cdot m_n / 3} \right\} \\
 &\leq 2\bar{q}_n \max_{j \in \mathcal{S}} \exp \left\{ - \frac{x^2 \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j / H}{2\sigma_h^2 + 2x \sqrt{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j / nH} \cdot m_n / 3} \right\} \\
 &= 2\bar{q}_n \max_{j \in \mathcal{S}} \exp \left\{ - \frac{x^2}{2\sigma^2 + 2m_n x / (3\sqrt{n}) \sqrt{H / \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}} \right\} \\
 &= o(1/\bar{q}_n)
 \end{aligned}$$

对于 $j \in \mathcal{S}$ 一致成立, 其中使用条件 $m_n \sqrt{\log \bar{q}_n / n} = o(1)$, 当 $x = \sqrt{C \log \bar{q}_n}$ 时, 假设 4.3 保证该结论成立. 接下来, 在 $\mathcal{B}^c$ 上讨论. 由假设 4.3 和 Markov 不等式, 有

$$\Pr(\mathcal{B}^c) \leq nH \max_{i,h} \Pr(\|\boldsymbol{\Sigma}^{-1} \mathbf{X}_{2i} \varepsilon_{ih}\|_\infty^\varpi > m_n^\varpi) = o(1/\bar{q}_n).$$

综上所述, 该引理由此证毕. $\square$

引理 B.4 若假设 4.1–4.4 成立, 则对于一个较大的 $C > 0$, 以下结论对 $j \in \mathcal{S}$ 一致成立:

$$\Pr(s_{2Sj}^{-1} \|\hat{\mathbf{B}}_{2j} - \mathbf{B}_j\|_2 > \sigma \sqrt{C \bar{q}_n \log \bar{q}_n} \mid \mathcal{D}_1) = o(1/\bar{q}_n).$$

证明 基于引理 B.3 的证明和假设 4.4, 有

$$\begin{aligned}
 \Pr(s_{2Sj}^{-1} \|\hat{\mathbf{B}}_{2j} - \mathbf{B}_j\|_2 > x) &\leq H \max_h \Pr(|\hat{\mathbf{B}}_{2j,h} - \mathbf{B}_{j,h}| > xs_{2Sj} / \sqrt{H}) \\
 &\leq H \max_h \Pr(|\hat{\mathbf{B}}_{2j,h} - \mathbf{B}_{j,h}| > x / (\bar{\kappa} \sqrt{nH})).
 \end{aligned}$$

给定 $\mathcal{D}_1$ 后, 可有 $\hat{\mathbf{B}}_{2j,h} - \mathbf{B}_{j,h} = \mathbf{e}_j^\top (n^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \mathbf{X}_{2Si}^\top)^{-1} n^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \varepsilon_{ih}$. 根据假设 4.4, 有

$$\max_{j \in \mathcal{S}} |\hat{\mathbf{B}}_{2j,h} - \mathbf{B}_{j,h}| \leq \bar{\kappa}^{-1} \left| \frac{1}{n} \sum_{i=1}^n \mathbf{e}_j^\top \mathbf{X}_{2Si} \varepsilon_{ih} \right| \leq \bar{\kappa}^{-1} \sqrt{\bar{q}_n} \max_{j \in \mathcal{S}} |\zeta_j|,$$

其中 $\boldsymbol{\zeta} = n^{-1} \sum_{i=1}^n \mathbf{e}_j^\top \mathbf{X}_{2Si} \varepsilon_{ih} = (\zeta_1, \dots, \zeta_p)^\top$. 令 $z = c \sqrt{\log \bar{q}_n / n}$ 且 c 非常大. 设 L 是一个正常数. 根据引理 B.1, 有

$$\begin{aligned}
 \Pr \left(\max_{j \in \mathcal{S}} |\zeta_j| > z \right) &\leq \bar{q}_n \max_{j \in \mathcal{S}} \Pr \left(\left| \sum_{i=1}^n \mathbf{e}_j^\top \mathbf{X}_{2Si} \varepsilon_{ih} \right| > nz \right) \\
 &\leq 2\bar{q}_n \max_{j \in \mathcal{S}} \exp \left\{ - \frac{n^2 z^2}{2n \mathbb{E}[(\mathbf{e}_j^\top \mathbf{X}_{2Si})^2 \varepsilon_{ih}^2] + 2(\max_{1 \leq h \leq H, 1 \leq i \leq n} |\mathbf{e}_j^\top \mathbf{X}_{2Si} \varepsilon_{ih}|) \times nz / 3} \right\} \\
 &\rightarrow 0,
 \end{aligned}$$

这里使用了结论 $\max_{1 \leq h \leq H, 1 \leq i \leq n} \mathbb{E} \|\mathbf{X}_{2Si} \varepsilon_{ih}\|_\infty^\varpi \leq L K_n^\varpi$ 和 $(n\bar{q}_n)^{1/\varpi+\gamma} \times \sqrt{\log \bar{q}_n / n} \rightarrow 0$. 这些结论可由条件 4.3 保证. 因此对于某个较大的 $C > 0$, 有

$$\Pr(s_{2Sj}^{-1} \|\hat{\mathbf{B}}_{2j} - \mathbf{B}_j\|_2 > x) \leq H \max_h \Pr(|\tilde{\mathbf{B}}_{2j,h} - \mathbf{B}_{j,h}| > x / (\bar{\kappa} \sqrt{nH}))$$

$$\begin{aligned}
&\leq H \max_h \Pr \left(\max_{j \in \mathcal{S}} \left| \frac{1}{n} \sum_{i=1}^n \mathbf{e}_j^\top \mathbf{X}_{2Si} \varepsilon_{ih} \right| > x \underline{\kappa} / (\bar{\kappa} \sqrt{n \bar{q}_n H}) \right) \\
&= H \max_h \Pr \left(\max_{j \in \mathcal{S}} |\zeta_j| > z \right) \\
&\rightarrow 0
\end{aligned}$$

当 $x = C\sqrt{\bar{q}_n \log \bar{q}_n}$ 时成立. □

附录 B.3 附录中引理的证明

引理 A.1 的证明 定义 $b_n = \sigma\sqrt{C \log \bar{q}_n}$, 其中 $C > 0$. 这里为了简洁, 记 $T_{1j} = \hat{\mathbf{B}}_{1j}/s_{1Sj}$ 和 $T_{2j} = \tilde{\mathbf{B}}_{2j}/\tilde{s}_{2Sj}$, 则有 $\tilde{W}_j = T_{1j}^\top T_{2j}$. 观察到

$$\begin{aligned}
\frac{G(t)}{G_-(t)} - 1 &= \frac{\sum_{j \in \mathcal{A}^c} \{\Pr(\tilde{W}_j \geq t \mid \mathcal{D}_1) - \Pr(\tilde{W}_j \leq -t \mid \mathcal{D}_1)\}}{\sum_{j \in \mathcal{A}^c} \Pr(\tilde{W}_j \leq -t \mid \mathcal{D}_1)} \\
&= \frac{\sum_{j \in \mathcal{A}^c} \{\Pr(T_{1j}^\top T_{2j} \geq t, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1) - \Pr(T_{1j}^\top T_{2j} \leq -t, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1)\}}{q_{0n} G_-(t)} \\
&\quad + \frac{\sum_{j \in \mathcal{A}^c} \{\Pr(T_{1j}^\top T_{2j} \geq t, \|T_{2j}\|_2 > b_n \mid \mathcal{D}_1) - \Pr(T_{1j}^\top T_{2j} \leq -t, \|T_{2j}\|_2 > b_n \mid \mathcal{D}_1)\}}{q_{0n} G_-(t)} \\
&=: \Lambda_1 + \Lambda_2.
\end{aligned}$$

根据定义, 已知 $G_-(t)$ 是一个单调递减函数. 首先对于第二项 Λ_2 , 由引理 B.3 可得

$$\frac{\sum_{j \in \mathcal{A}^c} \Pr(T_{1j}^\top T_{2j} \geq t, \|T_{2j}\|_2 > b_n \mid \mathcal{D}_1)}{q_{0n} G_-(t)} \leq \frac{\sum_{j \in \mathcal{A}^c} \Pr(\|T_{2j}\|_2 > b_n \mid \mathcal{D}_1)}{\alpha \eta_n} \lesssim \frac{\bar{q}_n \times o(1/\bar{q}_n)}{\eta_n},$$

从而可以证明 $\Lambda_2 = o(1)$. 回顾

$$T_{2j}^\top = \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \varepsilon_{2i}^\top / \sqrt{n \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}.$$

根据引理 B.2 和假设 4.3, 有

$$\frac{\Pr(T_{1j}^\top T_{2j} \geq t, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1)}{\Pr(T_{1j}^\top U \geq t, \|U\|_2 \leq b_n \mid \mathcal{D}_1)} \rightarrow 1,$$

其中 $U_j \sim \mathcal{N}_H(0, \tilde{\boldsymbol{\Sigma}}_j)$ 且独立于 T_{1j} . 这里 $\tilde{\boldsymbol{\Sigma}}_j = (\tilde{\sigma}_{lh})_{l,h=1}^H$ 且

$$\tilde{\sigma}_{lh} = \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbb{E}[\varepsilon_l \mathbf{X}_S^\top \mathbf{X}_S \varepsilon_h^\top] \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j / \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j.$$

类似可得

$$\frac{\Pr(T_{1j}^\top T_{2j} \leq -t, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1)}{\Pr(T_{1j}^\top U_j \leq -t, \|U_j\|_2 \leq b_n \mid \mathcal{D}_1)} \rightarrow 1.$$

注意到 $\Pr(T_{1j}^\top U_j \leq -t, \|U_j\|_2 \leq b_n \mid \mathcal{D}_1) = \Pr(T_{1j}^\top U_j \geq t, \|U_j\|_2 \leq b_n \mid \mathcal{D}_1)$, 进一步有 $\Lambda_1 = o(1)$. □

引理 A.2 的证明 这里以第一个公式为例, 第二个公式的证明可类比推导. 由引理 A.1 的证明和 $b_n = \sigma\sqrt{C \log \bar{q}_n}$, 可得

$$G(t) = \frac{1}{q_{0n}} \sum_{j \in \mathcal{A}^c} \Pr(\widetilde{W}_j \geq t, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1) \{1 + o(1)\} =: \bar{G}(t) \{1 + o(1)\}.$$

进一步地可有

$$\frac{1}{q_{0n}} \sum_{j \in \mathcal{A}^c} \mathbb{I}(\widetilde{W}_j \geq t) = \frac{1}{q_{0n}} \sum_{j \in \mathcal{A}^c} \mathbb{I}(\widetilde{W}_j \geq t, \|T_{2j}\|_2 \leq b_n) \{1 + o_p(1)\}.$$

因此, 仅需要证明如下结论:

$$\sup_{0 \leq t \leq G^{-1}(\alpha\eta_n/q_{0n})} \left| \frac{\sum_{j \in \mathcal{A}^c} \mathbb{I}(\widetilde{W}_j \geq t, \|T_{2j}\|_2 \leq b_n)}{q_{0n} \bar{G}(t)} - 1 \right| = o_p(1).$$

注意到 $\bar{G}(t)$ 是一个单调递减连续函数. 令 $a_p = \alpha\eta_n$, $z_0 < z_1 < \dots < z_{h_n} \leq 1$ 且 $t_i = \bar{G}^{-1}(z_i)$, 其中 $z_0 = a_p/q_{0n}$, $z_i = a_p/q_{0n} + b_p \exp(i^\zeta)/q_{0n}$, $h_n = \{\log((q_{0n} - a_p)/b_p)\}^{1/\zeta}$, 则有 $b_p/a_p \rightarrow 0$ 和 $0 < \zeta < 1$. 已知对于任意 i , $\bar{G}(t_i)/\bar{G}(t_{i+1}) = 1 + o(1)$ 一致成立, 足以推导下式的收敛速度:

$$D_n = \sup_{0 \leq i \leq h_n} \left| \frac{\sum_{j \in \mathcal{A}^c} \{\mathbb{I}(\widetilde{W}_j \geq t_i, \|T_{2j}\|_2 \leq b_n) - \Pr(\widetilde{W}_j \geq t_i, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1)\}}{q_{0n} \bar{G}(t_i)} \right|.$$

定义 $\mathcal{B} = \{\|T_{2j}\|_2 \leq b_n, j \in \mathcal{A}^c\}$, 则有

$$\begin{aligned} D(t) &= \mathbb{E} \left[\left(\sum_{j \in \mathcal{A}^c} \{\mathbb{I}(\widetilde{W}_j > t, \|T_{2j}\|_2 \leq b_n) - \Pr(\widetilde{W}_j > t, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1)\} \right)^2 \mid \mathcal{D}_1 \right] \\ &= \sum_{j \in \mathcal{A}^c} \sum_{l \in \mathcal{A}^c} \{\Pr(\widetilde{W}_j > t, \widetilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B}) - \Pr(\widetilde{W}_j > t \mid \mathcal{D}_1, \mathcal{B}) \Pr(\widetilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B})\} \{1 + o(1)\}. \end{aligned}$$

进一步定义事件

$$\mathcal{M}_j = \{l \in \mathcal{A}^c : |\rho_{jl}| \geq C(\log n)^{-2-\nu}\},$$

这里 ρ_{jl} 表示在给定 $\mathcal{D}_1$ 后 $\widetilde{W}_j$ 和 $\widetilde{W}_l$ 的条件相关系数. 回顾统计量的定义

$$\widetilde{W}_j = \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \boldsymbol{\varepsilon}_{2i}^\top T_{1j} / \sqrt{n \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}.$$

其条件方差和条件协方差分别为

$$\begin{aligned} \text{var}(\widetilde{W}_j \mid \mathcal{D}_1) &= \frac{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbb{E}[\mathbf{X}_S \mathbf{X}_S^\top (\boldsymbol{\varepsilon}^\top T_{1j})^2] \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}, \\ \text{cov}(\widetilde{W}_j, \widetilde{W}_l \mid \mathcal{D}_1) &= \frac{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbb{E}[\mathbf{X}_S \mathbf{X}_S^\top \boldsymbol{\varepsilon}^\top T_{1j} T_{1l}^\top \boldsymbol{\varepsilon}] \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_l}{\sqrt{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j} \cdot \sqrt{\mathbf{e}_l^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_l}}. \end{aligned}$$

因此可得

$$\rho_{jl} = \frac{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbb{E}[\mathbf{X}_S \mathbf{X}_S^\top \boldsymbol{\varepsilon}^\top T_{1j} T_{1l}^\top \boldsymbol{\varepsilon}] \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_l}{\sqrt{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbb{E}[\mathbf{X}_S \mathbf{X}_S^\top (\boldsymbol{\varepsilon}^\top T_{1j})^2] \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j} \cdot \sqrt{\mathbf{e}_l^\top \boldsymbol{\Sigma}_S^{-1} \mathbb{E}[\mathbf{X}_S \mathbf{X}_S^\top (\boldsymbol{\varepsilon}^\top T_{1l})^2] \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_l}}.$$

这里在给定 $\mathcal{D}_1$ 后, $\widetilde{W}_j$ 和 $\widetilde{W}_l$ 之间的条件相关性 ρ_{jl} 仍与 $\mathcal{D}_1$ 有关导致 ε 与 $\mathbf{X}_S$ 不独立. 由假设 4.7, 可得

$$\begin{aligned} D(t) &\leq \sum_{j \in \mathcal{A}^c} \sum_{j \in \mathcal{M}} \Pr(\widetilde{W}_j > t \mid \mathcal{D}_1) + \sum_{j \in \mathcal{A}^c} \sum_{l \in \mathcal{M}_j^c} \{\Pr(\widetilde{W}_j > t, \widetilde{W}_l > t \mid \mathcal{D}_1) \\ &\quad - \Pr(\widetilde{W}_j > t \mid \mathcal{D}_1) \Pr(\widetilde{W}_l > t \mid \mathcal{D}_1)\} \\ &\leq r_p q_{0n} G(t) + \sum_{j \in \mathcal{A}^c} \sum_{l \in \mathcal{M}_j^c} \{\Pr(\widetilde{W}_j > t, \widetilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B}) - \Pr(\widetilde{W}_j > t \mid \mathcal{D}_1, \mathcal{B}) \Pr(\widetilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B})\}. \end{aligned}$$

而在给定 $\mathcal{D}_1$ 后, 对于每个 $j \in \mathcal{A}^c$ 和 $l \in \mathcal{M}_j^c$, 根据文献 [5, 引理 1], 有

$$\begin{aligned} &\left| \frac{\Pr(\widetilde{W}_j > t, \widetilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B}) - \Pr(\widetilde{W}_j > t \mid \mathcal{D}_1, \mathcal{B}) \Pr(\widetilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B})}{\Pr(\widetilde{W}_j > t \mid \mathcal{D}_1, \mathcal{B}) \Pr(\widetilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B})} \right| \\ &= \left| \frac{\Pr(\widetilde{W}_j > t, \widetilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B}) - \Pr(\widetilde{W}_j > t \mid \mathcal{D}_1, \mathcal{B}) \Pr(\widetilde{W}_l > t \mid \mathcal{D}_1, \mathcal{B})}{\bar{G}^2(t)} \right| \\ &\leq A_n \end{aligned}$$

一致成立, 其中 $A_n = (\log n)^{-1-\nu_1}$ 且 $\nu_1 = \min(\nu, 1/2)$. 基于以上结论, 对于任意 $\xi > 0$, 有

$$\begin{aligned} &\Pr(D_n \geq \xi \mid \mathcal{D}_1) \\ &\leq \sum_{i=0}^{h_n} \Pr\left(\left| \frac{\sum_{j \in \mathcal{A}^c} \{\mathbb{I}(W_j > t_i, \|T_{2j}\|_2 \leq b_n) - \Pr(W_j > t_i, \|T_{2j}\|_2 \leq b_n \mid \mathcal{D}_1)\}}{q_{0n} \bar{G}(t_i)} \right| \geq \xi \mid \mathcal{D}_1\right) \\ &\leq \frac{1}{\xi^2} \sum_{i=0}^{h_n} \frac{1}{q_{0n}^2 \bar{G}^2(t_i)} D(t_i) \\ &\leq \frac{1}{\xi^2} \sum_{i=0}^{h_n} \frac{1}{q_{0n}^2 \bar{G}^2(t_i)} \{r_p q_{0n} \bar{G}(t_i) + q_{0n}^2 \bar{G}^2(t_i) A_n\} \\ &\leq \frac{1}{\xi^2} \left\{ \sum_{i=0}^{h_n} \frac{r_p}{q_{0n} \bar{G}(t_i)} + h_n A_n \right\}. \end{aligned}$$

进一步地, 注意到

$$\sum_{i=0}^{h_n} \frac{1}{q_{0n} \bar{G}(t_i)} = \frac{1}{a_p} + \sum_{i=1}^{h_n} \frac{1}{a_p + b_p e^{i\zeta}} \lesssim b_p^{-1}.$$

因为 ζ 可以取得任意接近于 1 的值使得 $h_n A_n \rightarrow 0$ 成立. 同时 b_p 也可取得充分大保证 $b_p/a_p \rightarrow 0$ 成立, 则当 $r_p/b_p \rightarrow 0$ 时有 $D_n = o_p(1)$. $\square$

引理 A.3 的证明 根据定义, 有

$$W_j - \widetilde{W}_j = \frac{\widehat{\mathbf{B}}_{1j}^\top \widehat{\mathbf{B}}_{2j}}{s_{1Sj} s_{2Sj}} - \frac{\widehat{\mathbf{B}}_{1j}^\top \widetilde{\mathbf{B}}_{2j}}{s_{1Sj} \widetilde{s}_{2Sj}} = \frac{\widehat{\mathbf{B}}_{1j}^\top}{s_{1Sj}} \left(\frac{\widehat{\mathbf{B}}_{2j}}{s_{2Sj}} - \frac{\widetilde{\mathbf{B}}_{2j}}{\widetilde{s}_{2Sj}} \right).$$

由假设 4.5 可知 $\|\widehat{\mathbf{B}}_{1j}/s_{1Sj}\|_\infty = O_p(\sqrt{n}c_{np})$, $j \in \mathcal{S}$. 根据定义, 可得

$$\frac{\widehat{\mathbf{B}}_{2j}^\top}{s_{2Sj}} - \frac{\widetilde{\mathbf{B}}_{2j}^\top}{\widetilde{s}_{2Sj}} = \frac{\mathbf{e}_j^\top \mathbf{A}^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \varepsilon_i^\top}{\sqrt{n \mathbf{e}_j^\top \mathbf{A}^{-1} \mathbf{e}_j}} - \frac{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \varepsilon_i^\top}{\sqrt{n \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}}$$

$$\begin{aligned}
 &= \frac{1}{\sqrt{n}} \mathbf{e}_j^\top (\mathbf{A}^{-1} - \boldsymbol{\Sigma}_S^{-1}) \sum_{i=1}^n \mathbf{X}_{2Si} \boldsymbol{\epsilon}_i^\top \frac{1}{\sqrt{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}} \\
 &\quad + \left(\frac{1}{\sqrt{\mathbf{e}_j^\top \mathbf{A}^{-1} \mathbf{e}_j}} - \frac{1}{\sqrt{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}} \right) \frac{1}{\sqrt{n}} \mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \boldsymbol{\epsilon}_i^\top \\
 &\quad + \frac{1}{\sqrt{n}} \mathbf{e}_j^\top (\mathbf{A}^{-1} - \boldsymbol{\Sigma}_S^{-1}) \sum_{i=1}^n \mathbf{X}_{2Si} \boldsymbol{\epsilon}_i^\top \left(\frac{1}{\sqrt{\mathbf{e}_j^\top \mathbf{A}^{-1} \mathbf{e}_j}} - \frac{1}{\sqrt{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}} \right) \\
 &=: \Lambda_1 + \Lambda_2 + \Lambda_3,
 \end{aligned}$$

其中

$$\mathbf{A} = n^{-1} \sum_{i=1}^n \mathbf{X}_{2Si} \mathbf{X}_{2Si}^\top.$$

首先, 注意到

$$\|\mathbf{A}^{-1} - \boldsymbol{\Sigma}_S^{-1}\|_\infty \leq \|\boldsymbol{\Sigma}_S^{-1}\|_\infty \|\mathbf{A}^{-1}\|_\infty \|\boldsymbol{\Sigma}_S - \mathbf{A}\|_\infty = O_p(v_n \sqrt{\bar{q}_n} a_{np}).$$

进一步地有

$$\begin{aligned}
 \left| \frac{1}{\sqrt{n}} \mathbf{e}_j^\top (\mathbf{A}^{-1} - \boldsymbol{\Sigma}_S^{-1}) \sum_{i=1}^n \mathbf{X}_{2Si} \boldsymbol{\epsilon}_i^\top \right| &\leq \|\mathbf{A}^{-1} - \boldsymbol{\Sigma}_S^{-1}\|_\infty \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{X}_{2Si} \boldsymbol{\epsilon}_i^\top \right\|_\infty = O_p(v_n \sqrt{\bar{q}_n} a_{np} \sqrt{\log \bar{q}_n}), \\
 \left| \frac{1}{\sqrt{\mathbf{e}_j^\top \mathbf{A}^{-1} \mathbf{e}_j}} - \frac{1}{\sqrt{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}} \right| &= \frac{|\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j - \mathbf{e}_j^\top \mathbf{A}^{-1} \mathbf{e}_j|}{|\sqrt{\mathbf{e}_j^\top \mathbf{A}^{-1} \mathbf{e}_j} + \sqrt{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}|} \left| \frac{1}{\sqrt{\mathbf{e}_j^\top \mathbf{A}^{-1} \mathbf{e}_j} \sqrt{\mathbf{e}_j^\top \boldsymbol{\Sigma}_S^{-1} \mathbf{e}_j}} \right| \\
 &\leq \|\boldsymbol{\Sigma}_S^{-1} - \mathbf{A}^{-1}\|_\infty = O_p(v_n \sqrt{\bar{q}_n} a_{np}).
 \end{aligned}$$

故有 $\Lambda_1 = O_p(v_n \sqrt{\bar{q}_n} a_{np} \sqrt{\log \bar{q}_n})$, $\Lambda_2 = O_p(v_n \sqrt{\bar{q}_n} a_{np} \sqrt{\log \bar{q}_n})$, 且 $\Lambda_3 = O_p(v_n^2 \bar{q}_n a_{np}^2 \sqrt{\log \bar{q}_n})$ 是 Λ_1 和 Λ_2 的小项. 由三角不等式和引理 B.4 可得

$$W_j - \widetilde{W}_j = O_p(\sqrt{n} c_{np}) O_p(v_n \sqrt{\bar{q}_n} a_{np} \sqrt{\log \bar{q}_n}) = O_p(c_{np} a_{np} v_n \sqrt{n \bar{q}_n \log \bar{q}_n}).$$

证毕. □

引理 A.4 的证明 根据引理 A.3, 可知下式以概率 1 成立:

$$\begin{aligned}
 \left| \sum_{j \in \mathcal{A}^c} \mathbb{I}(W_j \geq t) - \sum_{j \in \mathcal{A}^c} \mathbb{I}(\widetilde{W}_j \geq t) \right| &\leq \left| \sum_{j \in \mathcal{A}^c} \{\mathbb{I}(\widetilde{W}_j \geq t + l_n) - \mathbb{I}(\widetilde{W}_j \geq t)\} \right| \\
 &\quad + \left| \sum_{j \in \mathcal{A}^c} \{\mathbb{I}(\widetilde{W}_j \geq t - l_n) - \mathbb{I}(\widetilde{W}_j \geq t)\} \right| \\
 &=: \Lambda_1 + \Lambda_2,
 \end{aligned}$$

其中随 $(n, p) \rightarrow \infty$, 可得

$$\frac{l_n}{c_{np} a_{np} v_n \sqrt{n \bar{q}_n \log \bar{q}_n}} \rightarrow \infty.$$

这里仅讨论 Λ_1 的证明, 有关 Λ_2 的证明可类似推导. 经简单计算, 可得

$$\mathbb{E}(\Lambda_1) = \mathbb{E} \left\{ \sum_{j \in \mathcal{A}^c} \mathbb{I}(t \leq \widetilde{W}_j \leq t + l_n) \right\}$$

$$\begin{aligned}
&= \sum_{j \in \mathcal{A}^c} \Pr(t \leq \widetilde{W}_j \leq t + l_n, \|T_{2j}\|_2 \leq b_n) + \sum_{j \in \mathcal{A}^c} \Pr(t \leq \widetilde{W}_j \leq t + l_n, \|T_{2j}\|_2 > b_n) \\
&\leq \sum_{j \in \mathcal{A}^c} \Pr(t \leq \widetilde{W}_j \leq t + l_n, \|T_{2j}\|_2 \leq b_n) + o(1),
\end{aligned}$$

其中, 由引理 B.3 可得

$$\Pr(t \leq \widetilde{W}_j \leq t + l_n, \|T_{2j}\|_2 > b_n) = o(\bar{q}_n^{-1}).$$

回顾 $U_j \sim N_H(0, \widetilde{\Sigma}_j)$. 由引理 B.2 可得

$$\begin{aligned}
&\sum_{j \in \mathcal{A}^c} \Pr(t \leq T_{1j}^\top U_j \leq t + l_n) \\
&= \sum_{j \in \mathcal{A}^c} \Pr\left(\frac{t}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}} \leq \frac{T_{1j}^\top U_j}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}} \leq \frac{t + l_n}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}}\right) \\
&= \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\Phi\left(\frac{t + l_n}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}}\right) - \Phi\left(\frac{t}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}}\right)\right\} \\
&\leq \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\frac{l_n}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}} \phi\left(\frac{t}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}}\right)\right\} \\
&\leq l_n \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\left[\frac{1}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}} \left(\frac{t}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}} + \frac{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}}{t}\right)\right] \widetilde{\Phi}\left(\frac{t}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}}\right)\right\} \\
&\lesssim l_n M^{-1} \log \bar{q}_n \sum_{j \in \mathcal{A}^c} \mathbb{E}\left\{\widetilde{\Phi}\left(\frac{t}{\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}}}\right)\right\},
\end{aligned}$$

其中

$$\widetilde{\Phi}(x) = 1 - \Phi(x),$$

$\phi(x)$ 和 $\Phi(x)$ 分别表示标准正态分布的密度函数和分布函数. 上式第二个不等式成立是因为

$$\phi(x) < \frac{x^2 + 1}{x} \widetilde{\Phi}(x), \quad \text{对于任意的 } x > 0.$$

另外, 由定义可知

$$\sum_{j \in \mathcal{A}^c} \Pr(\widetilde{W}_j > t) = \sum_{j \in \mathcal{A}^c} \mathbb{E}\{\widetilde{\Phi}(t/\sqrt{T_{1j}^\top \widetilde{\Sigma}_j T_{1j}})\} \{1 + o(1)\}.$$

注意到在引理 A.2 中, 仅需要保证当 $n \rightarrow \infty$ 时 h_n 可任意小. 基于引理 A.2 证明和条件, 可得

$$c_{np} a_{np} v_n \sqrt{n \bar{q}_n \log \bar{q}_n} \log \bar{q}_n h_n \rightarrow 0.$$

证毕. □

附录 B.4 补充的模拟结果

补充的模拟结果如表 B1 所示.

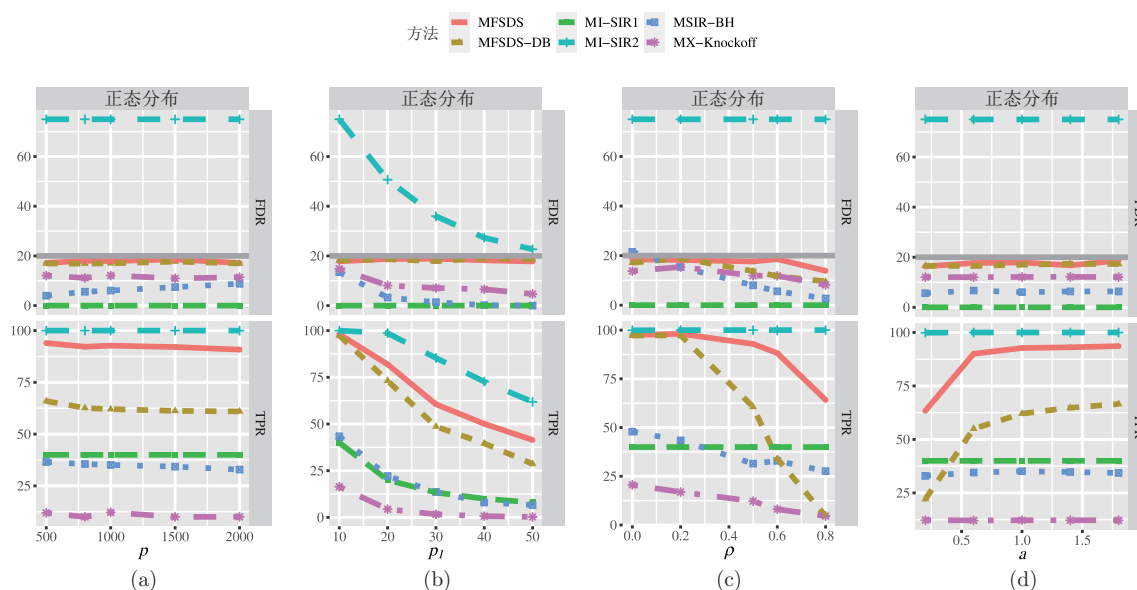


图 B1 (网络版彩图) 给定 $n = 500$ 且 X 服从多元正态分布. 在情景 2c 中, 不同检验方法随不同协变量维数 p 、协变量相关性 ρ 信号个数 p_1 和信号强度 a 的 FDR 和 TPR (%) 曲线. (a) $(p_1, \rho, a) = (10, 0.5, 1.0)$; (b) $(p, \rho, a) = (1000, 0.5, 1.0)$; (c) $(p, p_1, a) = (1000, 10, 1.0)$; (d) $(p, p_1, \rho) = (1000, 10, 0.5)$ 灰实线表示给定的显著性水平 $\alpha = 0.2$

Model-free controlled variable selection via data splitting

Yixin Han, Xu Guo & Changliang Zou

Abstract Addressing the simultaneous identification of contributory variables while controlling the false discovery rate (FDR) in high-dimensional data is a crucial statistical challenge. In this paper, we propose a novel model-free variable selection procedure in a sufficient dimension reduction framework via a data splitting technique. The variable selection problem is first converted to a least squares procedure with several response transformations. We construct a series of statistics with global symmetry property and leverage the symmetry to derive a data-driven threshold aimed at error rate control. Our approach demonstrates the capability for achieving finite-sample and asymptotic FDR control under mild theoretical conditions. Numerical experiments confirm that our procedure has satisfactory FDR control and higher power compared with existing methods.

Keywords data splitting, false discovery rate, model-free, sufficient dimension reduction, symmetry

MSC(2020) 62J02, 62H15

doi: 10.1360/SCM-2022-0066