arXiv:2305.05141v2 [stat.ME] 3 Aug 2023

基于随机方法的最优稀疏切片逆回归投影

Jia Zhang* Runxiong Wu† Xin Chen‡2023年8月4日

摘要

我们提出了一种基于随机投影的新型稀疏切片逆回归方法，用于大 p 小 n 规模情形。该方法嵌入广义特征值框架，最终可归结为并行执行低维（广义）特征值分解，从而实现较高的计算效率。从理论上看，我们证明了在适当假设下，该方法达到了极小极大最优收敛速率。此外，我们的算法采用了精细的重新加权方案，能够显著提高协变量活跃集的可识别性。大量数值研究表明，与竞争方法相比，所提算法具有显著优势。

关键词：充分降维；稀疏切片逆回归；随机投影；高维统计学；极小极大

*统计学院，西南财经大学，中国成都611130。
†共同第一作者。工程学院，加州大学戴维斯分校，戴维斯，加利福尼亚州 95616，美国。
‡通讯作者。南方科技大学统计与数据科学系，中国深圳 518055；电子邮箱：chenx8@sustech.edu.cn。

1 引言

面对现代应用中大量的协变量，充分降维（SDR）通过寻找原始预测变量的低维线性组合，提供了一种统计框架，可在不损失信息的前提下降低问题的维度。在涉及单变量响应 $Y \in \mathbb{R}$ 和p维预测变量向量 $\boldsymbol{X} = (X_1, ..., X_p)^\top \in \mathbb{R}^p$的回归问题中，SDR 旨在找到 $\mathbb{R}^p$ 的降维子空间，其基为 $\boldsymbol{B}$，使得

$$Y \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{B}^\top \boldsymbol{X}\,, \tag{1.1}$$

其中 $\perp\!\!\!\perp$ 表示统计独立性。降维子空间通常并不唯一，因此，降维的主要关注点在于所有降维子空间的交集。该交集具有最小维度，并且在温和条件下本身也是一个降维子空间（Cook，1996）。我们称这个交集为中心子空间，并将其记作 $\mathcal{S}_{Y|\boldsymbol{X}}$。设 $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$是 $\mathcal{S}_{Y|\boldsymbol{X}}$的一组基，则 $\boldsymbol{\beta}^\top \boldsymbol{X}$能够刻画 Y 关于 $\boldsymbol{X}$的完整回归关系。由于中心子空间的维度 d 通常远小于 p，因此我们可以使用低维的 $\boldsymbol{\beta}^T \boldsymbol{X}$ 来预测 Y，而不会丢失任何信息。

目前已有许多充分降维方法被提出，用于估计中心子空间 $\mathcal{S}_{Y|\boldsymbol{X}}$ [1]。其中，开创性的切片逆回归（SIR）（Li，1991）因其简单性、通用性和计算效率而广受欢迎。与大多数充分降维方法一样，已有研究证明，SIR 在预测变量 p的维度固定或随样本量 n 缓慢增长的传统情形下表现良好（Li，1991；Hsing 和 Carroll，1992；Zhu 和 Ng，1995；Zhu 等人，2006）。然而，当 $p \asymp n$ 或 $p \gg n$ 时，这种情况在现代数据集中十分常见，SIR 在理论和计算两个方面都会失效（Hung 和 Huang，2019；Lin 等人，2018；Lin 等人，2019）。事实上，Lin 等人（2018）证明，在温和条件下，当且仅当 p/n 趋于零时，中心子空间的 SIR 估计量才具有一致性。

为解决这一问题，并进一步提升可解释性和模型简约性，[1]有关充分降维方法的全面综述，请参见Yin (2011) 和Ma 和 Zhu (2013)。

在回归中施加合理的稀疏性条件，以限制活跃预测变量的数量 s 。Lin 等人（2018）证明，在 $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I}_p$ 的情形下，中心子空间估计的最优速率应为 $(s\log p)/n$，相差一个常数因子，并提出了一种用于单指标模型的对角阈值算法，该算法在 $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I}_p$ 时可达到最优速率。在后续工作中，Lin 等人（2019）为具有一般协方差矩阵$\boldsymbol{X}$的多指标模型提出了一种高效的切片逆回归 Lasso 变体。然而，研究表明，当 p 的阶为 $o(n^2c^2)$ 时，该方法才具有速率最优性，其中 c 是广义信噪比。为弥补理论上的不足，并解决对协方差矩阵的限制，Tan 等人（2020）为稀疏切片逆回归提出了一种自适应估计方案；该方案计算上易于处理，即使在 $\log(p) = o(n)$ 的情况下也具有速率最优性。

[{"id": 0, "input": "In this paper, we propose a novelefficient Sparse SIR method via Random Projection(SSIRvRP) for large p small n problems. This approach can attain the minmax optimal ratewhen $\log(p) = o(n)$ under mild assumptions and enjoys high computational efficiency. Com-pared with existing algorithms, our SSIRvRPalgorithm exhibits marked simplicity, robust-ness to poor initialization, and scalability through easy parallelization. Indeed, SSIRvRPcanbe implemented through theparallel execution of low-dimensional (generalized) eigenvaluedecompositions, thereby avoiding directlycomputingthe inverse ofthesample covariancematrix or the eigenvectors of the sampleconditional covariance matrix and selecting multi-ple tuningparameters appeared in existing sparseSIR methods. Additionally, it does noteven requirecomputing the(conditional)samplecovariancematrixitself, sinceitsufficestoextractitsprincipal submatrices, which can be computed from the low-dimensional pro-jected data. Finally, thealgorithm can be readily extended to other SDR methods via t hegeneralizedeigenvalue formulation (Li, 2007, Hung and Huang, 2019). Extensive numericalexperiments demonstratethe superior performance of SSIRvRP in comparison to competingapproaches.", "layout_label": "plain text"}]

当 $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I}_p$时，所提算法将退化为 Gataric 等人 (2020年提出的稀疏主成分分析算法。然而，我们的算法绝非将特征值分解简单直接地扩展为广义特征值分解，这一点可从以下三个方面看出。首先，也是最重要的一点，

所提算法背后的理论证据与 Gataric 等人 (2020年截然不同。我们的工作弥补了标准特征值分解与广义特征值分解之间的理论空白，至少在基于随机投影的算法方面如此，这主要得益于对 Cholesky 变换的巧妙运用。其次，作为一种广义特征值问题，稀疏切片逆回归具有一些独特的特点，尤其是在计算方面。详情见第 2.2 节。最后，我们在所提算法中嵌入了重新加权方案，显著提升了识别活跃协变量的能力。

相关文献。随机投影技术在降维方法的设计中发挥了作用。举几个例子，Qi 和 Hughes (2012) 和 Anaraki 和 Hughes (2014) 提出通过对数据进行一系列低维随机投影，计算样本协方差矩阵的主特征向量。Gataric 等人(2020年) 探讨了在较大 p 小型 n 设置中通过轴对齐随机投影进行稀疏主成分分析，这启发了我们的方法。Hung 和 Huang(2019) 提出了一种所谓的集成随机划分充分降维方法，通过整合由协变量随机划分得到的中心子空间的多个草图来实现。最近，Liu 等人(2023年) 提出了一种针对高维单指标模型假设检验的随机投影方法。

我们注意到，Tan 等人 (2018) 建议通过截断 Rayleigh 流方法解决稀疏广义特征值问题。然而，该方法只能用于估计主广义特征向量，这限制了它作为充分降维方法的应用。关于稀疏充分降维方法的更多讨论，请参阅Li 等人 (2020年) 的全面综述。

记号。我们引入本文中通用的一些记号。对于整数$n > 0$，令 $[n] := \{1, 2, \ldots, n\}$ 且 $\mathbb{E}_n(\boldsymbol{X}) := n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i$，其中随机向量 $\boldsymbol{X}$ 有一个样本$\{\boldsymbol{X}_i\}_{i=1}^{n}$。对于向量 $\boldsymbol{\alpha} \in \mathbb{R}^p$，用 $\alpha^{(j)}$ 表示其第 j个分量；对于 $S \subset [p]$，令$\boldsymbol{\alpha}^{(S)}$ 表示 $\boldsymbol{\alpha}$ 的一个子向量，其分量索引属于 S。用 $\|\boldsymbol{\alpha}\|_2$表示 $\boldsymbol{\alpha}$的欧几里得范数。对于矩阵 $\boldsymbol{U} \in \mathbb{R}^{p \times d}$，令 $U^{(i,j)}$表示其第(i, j)个元素，$\boldsymbol{U}^{(i,\cdot)}$ 表示其第 i行，$\boldsymbol{U}^{(\cdot,j)}$表示其第 j列。对于 $S \subseteq [p]$和 $S' \subseteq [d]$，记 $\boldsymbol{U}^{(S,S')}$为行索引属于 S且列索引属于 S'的 $|S| \times |S'|$子矩阵，并分别将 $\boldsymbol{U}^{([p],S')}$和 $\boldsymbol{U}^{(S,[d])}$简记为 $\boldsymbol{U}^{(\cdot,S')}$和 $\boldsymbol{U}^{(S,\cdot)}$。令 $\|\boldsymbol{U}\|_{\mathrm{F}}$ 和 $\|\boldsymbol{U}\|_{\mathrm{op}}$ 分别表示其弗罗贝尼乌斯范数和算子范数，

分别。

对于任意指标集 $J \subseteq [p]$，P_J 表示投影矩阵，它是一个 $p \times p$ 对角矩阵，且第 j 个对角元素为 $\mathbf{1}_{\{j \in J\}}$。对于 $\mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$ 中的实对称矩阵对 $(\boldsymbol{A}, \boldsymbol{B})$，令 $\lambda_1(\boldsymbol{A}, \boldsymbol{B}) \geq \lambda_2(\boldsymbol{A}, \boldsymbol{B}) \geq \ldots \geq \lambda_p(\boldsymbol{A}, \boldsymbol{B})$ 表示按降序排列的广义特征值，令 $\boldsymbol{v}_1(\boldsymbol{A}, \boldsymbol{B}) \ldots, \boldsymbol{v}_p(\boldsymbol{A}, \boldsymbol{B})$ 表示对应的特征向量，使得

$$\boldsymbol{A}\boldsymbol{v}_i(\boldsymbol{A}, \boldsymbol{B}) = \lambda_i(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{B}\boldsymbol{v}_i(\boldsymbol{A}, \boldsymbol{B})$$

对于任意 $i \in [p]$。

论文组织结构。本文其余部分安排如下。第 2 节提出一种通过随机投影实现的稀疏切片逆回归估计量，并在第 3 节探讨其理论性质。第 4 节通过加入重新加权方案来提高所提算法的效率，第 5 节讨论改进算法的超参数选择。第 6 节进行数值实验，并在第 7 节将所提方法应用于基因表达数据分析。第8 节总结全文。所有技术性证明均置于附录。

2 方法

2.1 重新审视 SparseSIR

将 Y 的离散化版本定义为

$$\tilde{Y} = \sum_{h=1}^{H} h \cdot \mathbf{I}_{\{Y \in J_h\}},$$

其中，$\{J_1, J_2, \ldots, J_H\}$ 是 Y的样本空间的可测划分。如果 $H \geq d + 1$，˜我们知道，Y 可用于识别 $\mathcal{S}_{Y|\boldsymbol{X}}$，而非 Y，且不会造成信息损失（Bura 和 Cook，2001，Cook 和 Forzani，2009）。那么，切片逆回归过程实际上是一个˜针对核矩阵 $\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)} = \mathrm{Cov}\{\mathbb{E}(\boldsymbol{X}|Y)\}$

关于 $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$，即，

$$\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{\beta}_i = \lambda_i \boldsymbol{\Sigma}\boldsymbol{\beta}_i \text{ with } \boldsymbol{\beta}_i^\top \boldsymbol{\Sigma}\boldsymbol{\beta}_j = \mathbf{1}_{\{i=j\}}\,, \tag{2.1}$$

其中 $i,\ j = 1,\ \cdots,\ p$ 和 $\lambda_1 \geq \cdots \geq \lambda_d > 0 = \lambda_{d+1} = \cdots = \lambda_p$。前 d 个 $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d\}$ 对应于非零广义特征值 $\lambda_1 \geq \cdots \geq \lambda_d$ 的广义特征向量构成了 $\mathcal{S}_{Y|\boldsymbol{X}}$ 的一组基。因此，求解以下优化问题可得到 $\mathcal{S}_{Y|\boldsymbol{X}}$ 的一组基$\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\ \ldots\ldots,\ \boldsymbol{\beta}_d)$:

$$\boldsymbol{\beta} = \underset{\boldsymbol{B}\in\mathbb{R}^{p\times d}}{\mathrm{argmax}}\ \mathrm{tr}\left\{\boldsymbol{B}^\top \boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{B}\right\} \quad \text{s.t. } \boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{I}_d\,. \tag{2.2}$$

为了更好地解释提取出的成分，通常建议对切片逆回归进行变量选择。变量选择的目标是寻找预测变量$\boldsymbol{X}^{(\mathcal{A})}$中最小的子集，其划分为 $\boldsymbol{X} = \{(\boldsymbol{X}^{(\mathcal{A})})^\top,\ (\boldsymbol{X}^{(\mathcal{A}^c)})^\top\}^\top$，使得

$$Y \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{X}^{(\mathcal{A})}\,, \tag{2.3}$$

其中 $\mathcal{A} \subseteq [p]$ 表示真正相关的预测变量集，而 $\mathcal{A}^c$ 表示无关预测变量集（Li 等人，2005 年，Bondell 和 Li，2009）。根据 $\boldsymbol{X}$的划分，可以相应地将$\boldsymbol{\beta}$ 划分为

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^{(\mathcal{A},\cdot)} \\ \boldsymbol{\beta}^{(\mathcal{A}^c,\cdot)} \end{pmatrix}, \quad \boldsymbol{\beta}^{(\mathcal{A},\cdot)} \in \mathbb{R}^{|\mathcal{A}|\times d}, \quad \boldsymbol{\beta}^{(\mathcal{A}^c,\cdot)} \in \mathbb{R}^{(p-|\mathcal{A}|)\times d},$$

其中，$|\mathcal{A}|$ 是 $\mathcal{A}$的基数。于是，(2.3) 意味着 $\boldsymbol{\beta}^{(\mathcal{A}^c,\cdot)} = 0$ (Bondell 和 Li，2009)。令 $\mathrm{supp}(\boldsymbol{\beta}) = \{j \in [p] :\ \boldsymbol{\beta}^{(j,\cdot)} \neq \mathbf{0}^\top\}$ 表示 $\boldsymbol{\beta}$的行支撑集，则 $\mathrm{supp}(\boldsymbol{\beta}) = \mathcal{A}$ 属于上述划分。假设 $|\mathcal{A}| \leq s$，稀疏切片逆回归进一步基于 (2.2) 定义，即寻找满足以下条件的 $\boldsymbol{\beta}$

$$\begin{aligned} \boldsymbol{\beta} = &\ \underset{\boldsymbol{B}\in\mathbb{R}^{p\times d}}{\mathrm{argmax}}\ \mathrm{tr}\left\{\boldsymbol{B}^\top \boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{B}\right\}, \\ &\text{s.t. } \boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{I}_d \text{ and } |\mathrm{supp}(\boldsymbol{B})| \leq s\,. \end{aligned} \tag{2.4}$$

通过求解以下问题，可以得到一种自然的稀疏SIR估计量

$$\begin{aligned} \check{\boldsymbol{\beta}} = & \underset{\boldsymbol{B}\in\mathbb{R}^{p\times d}}{\operatorname{argmax}} \operatorname{tr}\left\{\boldsymbol{B}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{B}\right\}, \\ & \text{s.t. } \boldsymbol{B}^{\top}\widehat{\boldsymbol{\Sigma}}\boldsymbol{B} = \boldsymbol{I}_d \text{ and } |\operatorname{supp}(\boldsymbol{B})| \le s, \end{aligned} \tag{2.5}$$

其中 $\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}$ 和 $\widehat{\boldsymbol{\Sigma}}$ 分别是条件期望的样本协方差矩阵$\mathbb{E}(\boldsymbol{X}|Y)$ 和 $\boldsymbol{X}$。 Tan 等人 (2020年证明了，这种自然估计量在各种常用损失函数下都是速率最优的。然而，直接求解问题 (2.5)在计算上不可行，因为这需要在满足稀疏性约束的所有 $\boldsymbol{B}\in\mathbb{R}^{p\times d}$中进行穷举搜索。为解决这一问题，Tan 等人(2020年)进一步基于Gao 等人 (2017)的工作提出了改进的三步估计量。下文中，我们借鉴Gataric 等人 (2020年)的思路，提出另一种基于随机投影的估计量。该估计量计算可行、简单得多，并且能够达到最优统计速率。

2.2 通过随机投影实现的稀疏切片逆回归估计量

对于 $k\in[p]$，令 $\mathcal{P}_k := \{P_S\colon S\subseteq[p], |S| = k\}$ 为 k 维投影的集合。我们的方法描述如下。对于两个固定整数 $A, B\in\mathbb{N}$，我们从 $\mathcal{P}_k$ 中独立且均匀地生成 $A\times B$ 个投影 $\{P_{a,b}\colon a\in[A], b\in[B]\}$。我们可以将这些投影视为 A 组，每组的基数为 B。对于每个 $a\in[A]$，令

$$b^*(a) := \underset{b\in[B]}{\operatorname{sargmax}} \sum_{i=1}^{d} \lambda_i(P_{a,b}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{a,b}, P_{a,b}\widehat{\boldsymbol{\Sigma}}P_{a,b})$$

表示所选投影在第 a 组中的索引，其中，sargmax 表示在所有 argmax 值中按字典序排列最小的元素。

考虑 oracle 情形，此时 $P_{a,\ b^*(a)} = P_{\mathcal{A}}$ 其中 $\mathcal{A}$ =supp($\boldsymbol{\beta}$)。那么 (2.4) 可简化为寻找 $\boldsymbol{\beta}$ 使得

$$\begin{aligned} \boldsymbol{\beta} = & \underset{\boldsymbol{B}\in\mathbb{R}^{p\times d}}{\operatorname{argmax}} \operatorname{tr}\left\{\boldsymbol{B}^{\top}P_{\mathcal{A}}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{\mathcal{A}}\boldsymbol{B}\right\}, \\ & \text{s.t. } \boldsymbol{B}^{\top}P_{\mathcal{A}}\boldsymbol{\Sigma}P_{\mathcal{A}}\boldsymbol{B} = \boldsymbol{I}_d, \end{aligned} \tag{2.6}$$

这意味着 $\boldsymbol{\beta}$ 可以由矩阵对 $\hat{}\hat{}(P_{\mathcal{A}}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{\mathcal{A}},\ P_{\mathcal{A}}\boldsymbol{\Sigma}P_{\mathcal{A}})$ 的主广义特征向量构成。因此，基本思路是构造一个集合 S，使得 $\mathcal{A}\subseteq S$ 以高概率成立。从某种程度上说，这解释了为何 $b^*(a)$ 如上所述定义。

接下来，为了汇总所有 A 个估计量的信息，我们计算第 j 个变量的重要性得分 $\hat{}w^{(j)}$，其定义为

$$\hat{w}^{(j)} := \frac{1}{A}\sum_{a=1}^{A}\sum_{i=1}^{d}(\hat{\lambda}_{a,b^*(a);i} - \hat{\lambda}_{a,b^*(a);d+1})(\hat{v}^{(j)}_{a,b^*(a);i})^2,$$

$\hat{}\hat{}$其中 $\lambda_{a,b^*(a);i}$ 是矩阵对 $(P_{a,b^*(a)}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{a,b^*(a)},P_{a,b^*(a)}\widehat{\boldsymbol{\Sigma}}P_{a,b^*(a)})$的第 i 个广义特征值，而 $\hat{}v^{(j)}_{a\ b^*,\ (a);i}$ 是对应广义特征向量的第 j 个元素。这意味着，我们不仅考虑每个坐标被选中的频率，还考虑它们在所选特征向量中的相应大小，以及信号强度的估计。估计过程总结为以下算法，我们将其命名为 SSIRvRP（基于随机投影的稀疏切片逆回归）。

算法 1： 用于中心子空间估计的 SSIRvRP 算法伪代码。
tion

输入： $\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)},\ \widehat{\boldsymbol{\Sigma}},\ A, B \in \mathbb{N},\ k, l \in [p],\ d \in [k]$

生成 $\{P_{a,b}:\ a \in [A], b \in [B]\}$，从 $\mathcal{P}_k$ 中独立且均匀地抽取

对于 $a = 1, \cdots, A$ **do**

  对于 $b = 1, \cdots, B$ **do**

    对于 $i \in [d+1]$，计算 $\hat{\lambda}_{a,b;i} := \lambda_i\ (P_{a,b}\ \boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{a,b},\ P_{a,b}\,\widehat{\boldsymbol{\Sigma}}P_{a,\ b})$ 和对应的广义特征向量 $\hat{}\boldsymbol{v}_{a,\ b;i} = \boldsymbol{v}_i(P_{a,\ b}\qquad \boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{a,b},\ P_{a,b}\,\widehat{\boldsymbol{\Sigma}}P_{a,b})$ with $\hat{\lambda}_{a,b;k+1} = 0$

  end

  计算 $b^*(a) := \mathrm{sargmax}_{b\in[B]}\sum_{i=1}^{d}\hat{}\lambda_{a,b;i}$

末尾

计算 $\widehat{\boldsymbol{w}} = (\hat{}w^{(1)}, \cdots, \hat{}w^{(p)})^\top$ with

$$\hat{w}^{(j)} := \frac{1}{A}\sum_{a=1}^{A}\sum_{i=1}^{d}(\hat{\lambda}_{a,b^*(a);i} - \hat{\lambda}_{a,b^*(a);d+1})(\hat{v}^{(j)}_{a,b^*(a);i})^2,\ j \in [p]$$

令 $S \subseteq [p]$ 为 $\widehat{\boldsymbol{w}}$中最大的 l 个分量所对应的指标集。

输出： $\boldsymbol{\beta} = (\hat{}\boldsymbol{v}_1,\ \ldots\ldots,\ \hat{}\boldsymbol{v}_d)$，其中 $\hat{}\boldsymbol{v}_1,\ \ldots\ldots,\ \hat{}\boldsymbol{v}_d$ 是 $d\,(P_{\hat{}S}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{\hat{}S},$ $P_{\hat{S}}\boldsymbol{\Sigma}P_{\hat{}S}$的前 d 个广义特征向量

在算法1中， $\widehat{\boldsymbol{\Sigma}} = \mathbb{E}_n[\{\boldsymbol{X} - \mathbb{E}_n(\boldsymbol{X})\}\{\boldsymbol{X} - \mathbb{E}_n(\boldsymbol{X})\}^\top]$ 和

$$\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)} = \sum_{h=1}^{H} \hat{p}_h \{\mathbb{E}_h(\boldsymbol{X}|\tilde{Y} = h) - \mathbb{E}_n(\boldsymbol{X})\}\{\mathbb{E}_h(\boldsymbol{X}|\tilde{Y} = h) - \mathbb{E}_n(\boldsymbol{X})\}^\top, \tag{2.7}$$

其中，$\hat{p}_h = \mathbb{E}_n[\mathbf{1}\{\tilde{Y} = h\}]$ 和 $\mathbb{E}_h(\boldsymbol{X}|\tilde{Y} = h) = \hat{p}_h^{-1} n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i \mathbf{1}\{\tilde{Y}_i = h\}$。正整数 A、B、k、l 和 d 是所提算法的超参数，其取值将在接下来的理论与数值研究中进行分析。

注1。另一种估计 $\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}$ 的方法是利用恒等式 $\widehat{\mathrm{Cov}}\{\mathbb{E}(\boldsymbol{X}|Y)\} = \widehat{\mathrm{Cov}}(\boldsymbol{X}) - \mathbb{E}\{\widehat{\mathrm{Cov}}(\boldsymbol{X}|Y)\}$。于是，我们可以估计 $\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)} = \boldsymbol{\Sigma} - \boldsymbol{T}$，其中

$$\widehat{\boldsymbol{T}} = \frac{1}{H} \sum_{h=1}^{H} \left\{ \frac{1}{n_h} \sum_{i \in S_h} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{S_h})(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{S_h})^\top \right\}, \tag{2.8}$$

其中，S_1、……、S_H 包含按照¯的尺度对 Y进行划分后所对应的样本索引，n_h 表示切片 S_h的样本量，$\boldsymbol{X}_{S_h}$ 表示该切片的样本均值。在我们的数值实验中，这个估计量的表现与上述估计量一样好。

注意，如果 $\boldsymbol{\Sigma} = \boldsymbol{I}_p$，算法1 就会退化为Gataric 等人（2020年）针对稀疏主成分分析所提出的算法。然而，我们的算法绝非将特征值分解简单、直接地扩展到广义特征值分解，这一点可从以下几个方面体现出来。首先，也是最重要的一点，所提算法背后的理论证据与Gataric 等人（2020年）的理论依据截然不同。众所周知，无论从理论还是计算角度来看，广义特征值问题原则上都比标准特征值问题更加困难。在标准特征值分解中一些常用的技术，例如谱分解，在广义特征值分解问题中并没有对应方法，这给算法1的理论分析带来了巨大挑战。我们的工作弥合了标准特征值分解与广义特征值分解之间的空白，至少对于基于随机投影的算法而言是如此。理论细节请参见第3.2 节和第9.2节。其次，作为一个广义特征值问题，稀疏切片逆回归具有一些独特特征，尤其是在计算方面。详见下文讨论。

算法的步骤 4　算法1 涉及求解一个非　-

$\widehat{\ }$ $\widehat{\ }$负定矩阵对$(P_{a,b}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{a,b}, P_{a,b}\boldsymbol{\Sigma}P_{a,b})$，这不同于常见的广义特征值问题，在这类问题中，矩阵对中的第二个矩阵是正定的（Li,2007，Tan 等人，2018，Hung 和 Huang，2019）。第二个矩阵的非负性$\frown P_{a,b}\boldsymbol{\Sigma}P_{a,b}$ 会导致广义特征向量存在多个解。为了清楚地说明这一点，令 S 表示与$\frown$ $\frown P_{a,b}$的非零对角元素相对应的 k 行索引集合。那么，不失一般性，矩阵对可表示为 $(P_S\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_S, P_S\boldsymbol{\Sigma}P_S)$，其中

$$P_S\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_S = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad P_S\widehat{\boldsymbol{\Sigma}}P_S = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}^{(S,S)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

以及广　步骤 4 中关于此对的广义特征值问题可简化为　　o

$$\widehat{\boldsymbol{\Sigma}}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)}\hat{\boldsymbol{v}}^{(S)}_i = \hat{\lambda}_i\widehat{\boldsymbol{\Sigma}}^{(S,S)}\hat{\boldsymbol{v}}^{(S)}_i \text{ with } \hat{\boldsymbol{v}}^{(S),\top}_i\widehat{\boldsymbol{\Sigma}}^{(S,S)}\hat{\boldsymbol{v}}^{(S)}_j = \mathbf{1}_{\{i=j\}} \tag{2.9}$$

针对 $i,j = 1,\ldots,k$，其中 $\hat{}\ \lambda_i = \lambda_i(P_S\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_S, P_S\widehat{\boldsymbol{\Sigma}}P_S)$ 和 $\hat{}\boldsymbol{v}_i = (\hat{}\boldsymbol{v}^{(S)}_{,i}{}^{\top}, \hat{}\boldsymbol{v}^{(S^c)}_{,i}{}^{\top})^{\top} = \boldsymbol{v}_i(P_S\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_S, P_S\widehat{\boldsymbol{\Sigma}}P_S)$。由于 (2.9) 未对 $\hat{}\boldsymbol{v}^{(S^c)}{}_i$ 施加任何限制，因此原问题的前 k 个广义特征向量有无穷多个解。为解决这一问题，我们选择 $\hat{}\boldsymbol{v}_i = (\hat{}\boldsymbol{v}^{(S)}_{,i}{}^{\top}, \mathbf{0}^{\top})^{\top}$，作为步骤 4 中广义特征值$\widehat{\ }$问题的 $i \in [k]$。这一点与 $P_S\boldsymbol{\Sigma}P_S$ 的特征值分解有很大不同，后者自然会得到稀疏特征向量，如 Gataric 等人（2020）所示。

接下来需要求解降维后的低维广义特征值问题（2.9），可通过传统算法快速求解。注意，若子矩阵 $\widehat{\boldsymbol{\Sigma}}^{(S,S)}$可逆，则（2.9）可进一步化为低维特征值分解问题。事实上，对于适当选择的k，　$\widehat{\boldsymbol{\Sigma}}^{(S,S)}$ 以高概率可逆。回顾 $\widehat{\boldsymbol{\Sigma}}^{(S,S)} = \mathbb{E}_n[\{\boldsymbol{X}^{(S)} - \mathbb{E}_n(\boldsymbol{X}^{(S)})\}\{\boldsymbol{X}^{(S)} - \mathbb{E}_n(\boldsymbol{X}^{(S)})\}^{\top}]$。那么，只要1 且 $k \ll n$ 满秩，　$\boldsymbol{\Sigma}$ 就以趋近于的概率可逆，这有助于提高所提算法的计算效率。因此，我们提出的 SSIRvRP 算法通过执行低维特征值分解，可高效求解计算上难以处理的稀疏切片逆回归问题（2.5）。此外，由于 $\{P_{a,b} : a \in [A], b \in [b]\}$ 是从 $\mathcal{P}_k$中随机生成的，　$A \times B$ 低-

维度特征值分解可以并行执行，从而进一步加快计算。最后，矩阵对 $(\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}^{(S,S)}, \widehat{\boldsymbol{\Sigma}}^{(S,S)})$ 可以从矩阵对 $(\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}, \widehat{\boldsymbol{\Sigma}})$ 中提取，也可以通过投影协变量 $\boldsymbol{X}^{(S)}$ 进行估计。当 p 足够大时，后一种方案更可取。

值得注意的是，不同于针对稀疏切片逆回归或充分降维的凸松弛算法（Tan 等人，2018，Lin 等人，2019，Tan 等人，2020年），所提算法不是迭代式的，因此对较差的初始化具有鲁棒性。

我们方法的另一个显著优势在于其可扩展至其他充分降维方法。众所周知，包括切片平均方差估计（Cook 和 Weisberg，1991）、主海森方向（Li，1992，Cook，1996）、方向回归（Li 和 Wang，2007）等在内的逆回归类充分降维方法，都可以表述为广义特征值问题。因此，作为稀疏广义特征值问题的解，所提 SSIRvRP 算法可以直接应用于这些方法，以获得中心子空间的稀疏估计量。

3 理论分析

3.1 各向同性协方差的上界

~我们考虑如下设定：$(\boldsymbol{X}_i, Y_i)_{i=1}^n$ 独立同分布，满足 $\boldsymbol{X}_i|(Y_i = h) \sim \mathcal{N}_p(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$，其中 $h \in [H]$；Cook 和 Yin（2001）、Cook（2007）、Cook 和 Forzani（2009）以及 Tan 等人（2020）也作出了这一假设。为简化理论分析，我们首先假设 $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I}_p$。此时，广义特征值问题（2.1）化为以下特征值问题：

$$\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{\beta}_i = \lambda_i \boldsymbol{\beta}_i, \quad \text{with } \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = \mathbf{1}_{\{i=j\}},$$

~其中，$\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}$ 是条件期望 $\mathbb{E}(\boldsymbol{X}|Y)$ 的协方差矩阵，$i, j = 1, \cdots, p$，且 $\lambda_1 \geq \cdots \geq \lambda_d > 0 = \lambda_{d+1} = \cdots = \lambda_p$。令 $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots\ldots, \lambda_d)$，并回顾一下，$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots\ldots, \boldsymbol{\beta}_d)$ 收集了对应于非零特征值 $\lambda_1 \geq \cdots \geq \lambda_d$ 的前 d 个特征向量。因此，我们可以将 $\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)} = \boldsymbol{\beta}\boldsymbol{\Lambda}\boldsymbol{\beta}^\top$ 分解。

需要满足以下技术条件。

(A1) $\kappa\lambda \geq \lambda_1 \geq \cdots \geq \lambda_d \geq \lambda > 0$ 对于某个常数 $\kappa > 1$。

(A2) $\boldsymbol{\beta} \in \boldsymbol{\Theta}_{p,\ d,\ s}(\mu)$，其中

$$\boldsymbol{\Theta}_{p,d,s}(\mu) := \left\{ \boldsymbol{V} \in \boldsymbol{\Theta}_{p,d}, \mathrm{supp}(\boldsymbol{V}) \leq s, \frac{\max_{j:\|\boldsymbol{V}^{(j,\cdot)}\|_2 \neq 0} \|\boldsymbol{V}^{(j,\cdot)}\|_2}{\min_{j:\|\boldsymbol{V}^{(j,\cdot)}\|_2 \neq 0} \|\boldsymbol{V}^{(j,\cdot)}\|_2} \leq \mu \right\},$$

并且 $\boldsymbol{\Theta}_{p,\ d}$ 表示由具有正交列的实数 $p \times d$ 矩阵构成的集合 s.

假设 (A1) 可视为一种覆盖条件（Cook，2004；Yu 等人，2016），文献中也有类似假设（Cai 等人，2013；Gao 等人，2015；Tan 等人，2020年）。假设 (A2) 是一种非相干条件，要求感兴趣的参数 $\boldsymbol{\beta}$ 非零行数不能过多，且非零行的欧几里得范数应大致相当。对于任意 $\boldsymbol{\beta} \in \boldsymbol{\Theta}_{p,\ d,\ s}(\mu)$，由于 $\sum_{j\in\mathcal{A}} \|\boldsymbol{\beta}^{(j,\cdot)}\|_2^2 = \|\boldsymbol{\beta}\|_{\mathrm{F}}^2 = d$，假设 (A2) 蕴含

$$\frac{d}{s\mu^2} \leq \|\boldsymbol{\beta}^{(j,\cdot)}\|_2^2 \leq \frac{d\mu^2}{s}, \quad \forall\, j \in \mathcal{A}. \tag{3.1}$$

对于 $\boldsymbol{U}$，$\boldsymbol{V} \in \boldsymbol{\Theta}_{p,\ d}$，遵循 Gataric 等人 (2020年) 的做法，我们使用损失函数

$$L(\boldsymbol{U}, \boldsymbol{V}) := \|\sin\{\boldsymbol{D}(\boldsymbol{U}, \boldsymbol{V})\}\|_{\mathrm{F}} \tag{3.2}$$

用于评估 $\boldsymbol{U}$ 与 $\boldsymbol{V}$之间的距离，其中正弦函数逐元素作用，$\boldsymbol{D}(\boldsymbol{U},\ \boldsymbol{V})$ 是 $d \times d$ 对角矩阵，其第 j 个对角元素为 $\boldsymbol{U}$ 与 $\boldsymbol{V}$之间第 j 个主角，即 $\cos^{-1}(\sigma_j)$，其中 σ_j 是 $\boldsymbol{U}^\top\boldsymbol{V}$的第 j 个奇异值。

在以下理论分析中，s 和 p 可以依赖于样本量 n，$\widehat{\ }$而 λ，μ，d 和 H 被视为固定常数。回顾一下，$\boldsymbol{\beta}$ 是算法$\widehat{\ }$ $\widehat{\ }$1 以输入 $\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}$，$\boldsymbol{\Sigma}$，$A$，$B$，$d$，$k$ 和 l得到的输出。定理1 给出了所提出的 SSIRvRP 估计量的上界。

定理 1. 假设假设 (A1)-(A2) 成立。如果 $k \geq \max\{d+1, s\}$, $l \geq s$, 并且

$$16K\sqrt{\frac{k\log p}{n}} \leq \frac{\lambda_d}{s\mu^2}, \tag{3.3}$$

那么有】 【。

$$\mathbb{P}\left\{L(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}) \leq 2K\sqrt{\frac{dl\log(p)}{n\lambda_d^2}}\right\} \geq 1 - cp^{-3} - p\exp\left(-\frac{A\lambda_d^2}{50p^2\mu^8\lambda_1^2}\right),$$

其中， K、 $c > 0$ 为一些常数。

定理 1 是 Gataric 等人 (2020年中定理 1 的推广，他们在受限协方差集中条件下研究了主子空间的估计问题。该条件由 Wang 等人 (2016a引入。我们将其结果推广到中心子空间估计问题，并且定理 1 证明中的引理 1的结果与受限协方差集中条件相仿。 Tan 等人 (2020年也研究了中心子空间估计问题，并针对稀疏切片逆回归提出了一种自适应估计方案，其在损失函数 $\sqrt{s\log(p)/n}$ 下达到了由式（3.2）定义的上界。如果 $\ell \asymp s$，我们的 SSIRvRP 估计量也能在相差一些常数的意义下达到相同的上界。

3.2 一般协方差下的上界

我们进一步研究式（2.1）中规定的稀疏切片逆回归的真实广义特征值问题，其中协变量的协方差 $\boldsymbol{\Sigma}$ $\boldsymbol{X}$ 不受特殊结构的限制。众所周知，广义特征值问题原则上比标准特征值问题更为困难，尤其是在理论分析方面。因此，我们将施加若干高层假设，以突出本节所处理的广义特征值问题与上文针对 $\boldsymbol{\Sigma} = \boldsymbol{I}_p$~这一简化情形所处理的标准特征值问题之间的技术差异。数据 $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ 也假设为独立同分布。

为了与算法背后的逻辑保持一致1并利用特征值分解的标准性质，出于理论分析的目的，我们希望找到一种适当的变换，将广义特征值分解（2.1）化为标准特征值分解，同时保留基的行稀疏结构 $\boldsymbol{\beta}$。事实证明， $\boldsymbol{\Sigma}$ 的 Cholesky 分解发挥着至关重要的作用。对于正定协方差矩阵 $\boldsymbol{\Sigma}$，利用 Cholesky 分解，我们可以将 $\boldsymbol{\Sigma}$ 分解为

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top,$$

其中 $\boldsymbol{L}$ 是一个对角元素为正的实下三角矩阵。令 $\boldsymbol{\beta}_i = (\boldsymbol{L}^\top)^{-1}\boldsymbol{\gamma}_i$对于 $i \in [p]$。那么，(2.1) 意味着

$$\{\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}(\boldsymbol{L}^{-1})^\top\}\boldsymbol{\gamma}_i = \lambda_i\boldsymbol{\gamma}_i \text{ with } \boldsymbol{\gamma}_i^\top\boldsymbol{\gamma}_j = \mathbf{1}_{\{i=j\}}\,, \tag{3.4}$$

其中 i、$j \in [p]$ 和 $\lambda_1 \geq \cdots \geq \lambda_d > 0 = \lambda_{d+1} = \cdots = \lambda_p$。记 $\boldsymbol{M} = \boldsymbol{L}^{-1}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}(\boldsymbol{L}^{-1})^\top$ 和 $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d)$，则 $\boldsymbol{M}$ 是一个实对称矩阵，可直接进行谱分解，即 $\boldsymbol{M} = \boldsymbol{\gamma}\boldsymbol{\Lambda}\boldsymbol{\gamma}^\top$，其中 $\boldsymbol{\Lambda} =$ 为 $\mathrm{diag}(\lambda_1, \ldots, \lambda_d)$。此外，$\boldsymbol{M}$ 的特征值恰好是矩阵对 $(\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}, \boldsymbol{\Sigma})$ 的广义特征值；更重要的是，$\boldsymbol{M}$ 的主特征向量 $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d)$ 完整保留了 $(\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}, \boldsymbol{\Sigma})$ 的广义特征向量 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ 的行稀疏结构。为清楚说明这一点，回顾 $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}^{(\mathcal{A},\cdot)^\top}, \mathbf{0}^\top)$，且 $\boldsymbol{\beta}_i = (\boldsymbol{L}^\top)^{-1}\boldsymbol{\gamma}_i$，其中 $\mathcal{A}$ 表示活跃协变量的集合，$\boldsymbol{L}$ 是一个下三角矩阵。于是有 [2]

$$\boldsymbol{\gamma} = \boldsymbol{L}^\top\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{L}^{(\mathcal{A},\mathcal{A}),\top} & \boldsymbol{L}^{(\mathcal{A}^c,\mathcal{A}),\top} \\ \mathbf{0}^\top & \boldsymbol{L}^{(\mathcal{A}^c,\mathcal{A}^c),\top} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{(\mathcal{A},\cdot)} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}^{(\mathcal{A},\mathcal{A}),\top}\boldsymbol{\beta}^{(\mathcal{A},\cdot)} \\ \mathbf{0} \end{pmatrix},$$

其中

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{L}^{(\mathcal{A},\mathcal{A})} & \mathbf{0} \\ \boldsymbol{L}^{(\mathcal{A}^c,\mathcal{A})} & \boldsymbol{L}^{(\mathcal{A}^c,\mathcal{A}^c)} \end{pmatrix}.$$

将广义特征值问题（2.1）转化为标准特征值问题（3.4），同时保持主特征向量的行稀疏结构不变，从而使一般协方差下算法 1 的理论分析变得可行。

需要以下假设。

(A1') $\iota\theta \geq \theta_1 \geq \cdots \geq \theta_p \geq \theta > 0$ 对于某个常数 $\iota > 1$，其中 θ_i 是 $\boldsymbol{\Sigma}$的特征值。

(A2') $\boldsymbol{\gamma} \in \boldsymbol{\Theta}_{p,\,d,\,s}(\mu)$，其中 $\boldsymbol{\Theta}_{p,\,d,\,s}(\mu)$ 在假设 (A2) 中定义。

[2]Cholesky 分解取决于随机向量$\boldsymbol{X}$中变量的排列顺序，并且仅适用于变量具有自然排序的情况。

（A3）对任意 $k \in [p]$，定义 $\mathcal{S}_k := \{S \subset [p] : |S| = k\}$。然后，

$$\mathbb{P}\left[\sup_{S \in \mathcal{S}_k} \max\{\|\widehat{\boldsymbol{L}}^{(S,S)} - \boldsymbol{L}^{(S,S)}\|_{\mathrm{op}}, \|\widehat{\boldsymbol{M}}^{(S,S)} - \boldsymbol{M}^{(S,S)}\|_{\mathrm{op}}\} \le K\sqrt{\frac{k \log p}{n}} \right] \ge 1 - c_1 p^{-c_2},$$

其中 $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top$ 是 $\widehat{\boldsymbol{\Sigma}}$的 Cholesky 分解，$\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{L}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}(\widehat{\boldsymbol{L}}^{-1})^\top$和$K$，$c_1$，$c_2 > 0$均为常数。

(A4) 对于任意的 $j \in [p]$ 以及 $\boldsymbol{X}$的任意排序，都存在某个 $\tau \in (0, 1]$，使得 $\tau \le (L^{(j,j)})^{-2} \le \tau^{-1}$。

假设 (A1') 是关于 $\boldsymbol{\Sigma}$的正则性条件，其特征值被要求与零和无穷远保持一定距离。假设 (A2') 是关于变换后基 $\boldsymbol{\gamma}$ 的不相干条件，要求 $\boldsymbol{\gamma}$ 的非零行数不能过多，且非零行的欧几里得范数应当相近。回顾一下，$\boldsymbol{\gamma} = \boldsymbol{L}^\top\boldsymbol{\beta}$ 和 $\boldsymbol{L}$ 是下三角矩阵。因此，假设 (A2') 等价于要求 $\boldsymbol{\beta}$ 的非零行数不能过多，且 $\boldsymbol{L}^{(\mathcal{A},\mathcal{A})\top}\boldsymbol{\beta}^{(\mathcal{A},\cdot)}$ 的各行具有相近的欧几里得范数。对于任意 $\boldsymbol{\gamma} \in \boldsymbol{\Theta}_{p,\,d,\,s}(\mu)$，由于 $\sum_{j\in\mathcal{A}} \|\boldsymbol{\gamma}^{(j,\cdot)}\|_2^2 = \|\boldsymbol{\gamma}\|_{\mathrm{F}}^2 = d$，假设 (A2') 推出

$$\frac{d}{s\mu^2} \le \|\boldsymbol{\gamma}^{(j,\cdot)}\|_2^2 \le \frac{d\mu^2}{s}, \quad \forall\, j \in \mathcal{A}. \tag{3.5}$$

假设 (A3) 确实是针对 $\boldsymbol{\Sigma}$和 $\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}$的样本估计值所施加的一项高层次条件。当 $\boldsymbol{\Sigma}$ 为对角矩阵时，在与 Tan 等人 (2020年中类似的温和条件下，假设 (A3) 成立；详情请参见他们的引理 1。假设 (A4) 是一项技术条件，旨在保证在总体层面上，活跃变量的权重高于非活跃变量。由于 $\boldsymbol{L}$ 的所有对角元素均为正，假设 (A4) 似乎相当宽松。此外，只要 $\boldsymbol{\Sigma}$ 是对角矩阵，该假设便可由假设 (A1') 推出。特别是，对于 $\boldsymbol{\Sigma} = \boldsymbol{I}_p$这一简化情形，上述假设自然成立。

对于 $\boldsymbol{U}$，$\boldsymbol{V} \in \boldsymbol{\Theta}_{p,\,d}(\boldsymbol{\Sigma}) := \{\boldsymbol{V} \in \mathbb{R}^{p\times d} : \boldsymbol{V}^\top\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{I}_d\}$，我们采用常用的广义损失函数

$$L(\boldsymbol{U}, \boldsymbol{V}) := \|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{V}\boldsymbol{V}^\top\|_{\mathrm{F}} \tag{3.6}$$

来评估线性子空间之间的距离；参见Cai 等人(2013）、Gao 等人(2015）和 Tan 等人(2020年）。注意，当 $\mathbf{\Sigma} = \boldsymbol{I}_p$时，该损失函数在相差一个常数 $\sqrt{2}$的意义下，化为式（3.2）中定义的损失函数。

在下文中，s 和 p 允许依赖于样本量 n，而 θ、λ、μ、τ、d 和 H 均视为固定常数。定理2 给出了所提出的适用于一般协方差的 SSIRvRP 估计量的上界。

定理 2。若满足假设(A1)-(A4)，且 $k \geq \max\{d+1, s\}$, $l \geq s$, $K\sqrt{k\log(p)/n} \leq \min\{\sqrt{\lambda_1/(4d)}\theta_1, \theta_p/(6\sqrt{\theta_1})\}$, $K\sqrt{l\log(p)/n} \leq \min\{\sqrt{\lambda_d/(2\sqrt{2})}\theta_1, \theta_p/(6\sqrt{\theta_1})\}$ 且

$$C\sqrt{\frac{k\log p}{n}} \leq \frac{\tau\lambda_d}{s\mu^2} \tag{3.7}$$

对于某个充分大的常数 $C > 0$，则有

$$\mathbb{P}\left\{L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq C'\sqrt{\frac{l\log(p)}{n}}\right\} \geq 1 - c'p^{-c_2} - p\exp\left(-\frac{A\tau^4\lambda_d^2}{50p^2\mu^8\lambda_1^2}\right),$$

其中 $C', c' > 0$ 为某些常数。

定理2 将定理1 推广到一般协方差情形，给出了上界$\sqrt{s\log(p)/n}$，适用于 $l \asymp s$，这与中心子空间估计的最新研究结果相呼应（Lin 等人，2019，Tan 等人，2020年）。尽管它看起来与定理1相似，但定理2 的证明与定理1的证明截然不同，它巧妙地利用了 Cholesky 分解以及三角矩阵的若干良好性质；详见第9.2 节。

3.3 下界

定理 1 和定理 2，连同下述定理 3，表明在满足 $l \asymp s$ 的条件下，SSIRvRP 估计量在所有可能的稀疏切片逆回归估计量中，除对数因子外达到了极小极大最优。定理 3 建立了所有可能的稀疏切片逆回归估计量之间的极小极大下界$\widetilde{\ }$，$\boldsymbol{\beta}$这与 Tan 等人 (2020年) 建立的下界相似。与他们的方法不同，我们要求中心子空间满足非相干性条件。然而，我们证明，对参数施加这类限制

空间并未从最小最大化角度让估计变得更容易。为简单起见，
~ ~我们假设 $(\boldsymbol{X}_i,\ Y_i)_{i=1}^n$ 独立同分布，且 $\boldsymbol{X}_i|(Y_i = h) \sim \mathcal{N}_p($ $\boldsymbol{\mu}_h,\ \boldsymbol{\Sigma}_h)$，其中 $h \in [H]$ 和 $\boldsymbol{\Sigma} = \boldsymbol{I}_p$。**定理 3**。假设 $4(s-1) \leq p-1,\ (s-1)\log\{($ $p-1)/(s-1)\} \geq 6$ 并且 $5n \geq 4s(s-1)\log\{(p-1)/(s-1)\}$。那么

$$\inf_{\widetilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Theta}_{p,d,s}(3)} \mathbb{E}_{P_{\boldsymbol{\beta}}}\{L(\widetilde{\boldsymbol{\beta}},\boldsymbol{\beta})\} \gtrsim \sqrt{\frac{s\log(p/s)}{n}},$$

其中期望是相对于（$\boldsymbol{X}_i$、Y_i）$\sim_{i.i.d.} P_{\boldsymbol{\beta}}$而言的。

定理3 中的界与主特征子空间估计中所建立的界相似（Gataric 等人，2020年）。我们将他们的结论推广到了中心子空间估计的情形。尽管结论相似，定理3 的技术证明与他们的证明有很大不同。此外，本文建立的下界~实际上不仅适用于 $\boldsymbol{X}|Y$ 的正态性假设和协方差矩阵的各向同性假设。

4 改进算法

本节中，为增强对协变量活跃集的识别能力，我们在算法 1中增加一个重新加权步骤，以细化对应于 $\widehat{\boldsymbol{w}}$最大值的权重。具体而言，令^S' 为算法 1产生的 $\widehat{\boldsymbol{w}}$ 中 l' 个最大分量的指标集。我们建议通过使用子矩阵对（$\widehat{\boldsymbol{\Sigma}}$(^$s'$、^$s'$）$\mathbb{E}$（$\boldsymbol{x}_{|Y}$）$\widehat{\boldsymbol{\Sigma}}$(^$s'$、^$s'$）重复算法 1中的步骤 1–8，重新计算^$S'$ 中这些变量的权重。然后，我们选择对应于 $\widehat{\boldsymbol{w}}'$ 最大值的 l个指标，构成集合^S，并将估计~ ^ ~$\boldsymbol{\beta}$输出为（P^$_S\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{x}|Y)}$）$P$^$_S$、$P_{\hat{S}}\boldsymbol{\Sigma}P$^$_S$）的前 d个广义特征向量。该修改后的 SSIRvRP 算法的伪代码见算法 2。我们发现，带有重新加权步骤的新算法确实表现良好，如数值研究所示。

运用定理2的类似技巧，可以证明算法2的估计具有与定理 $\sqrt{l\log(p)/n}$ 相同的上界2。然而，算法 2 第 9 步中的重新加权方案似乎有助于在有限样本中更准确地识别 $\mathcal{A}$，即 $\boldsymbol{\beta}$ 的行支撑集^。考虑到 S 的构成，这并不令人意外，因为它是一个

算法 2： 带重新加权的 SSIRvRP 算法伪代码

输入： $\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}, \widehat{\boldsymbol{\Sigma}}, A_1, A_2, B_1, B_2 \in \mathbb{N}, k, l, l' \in [p], d \in [k]$**1** 生成 $\{P_{a,b}$： 从 $\mathcal{P}_k$ 中独立且均匀地生成 $a \in [A_1], b \in [B_1]\}$ **2** 对于 $a = 1, \cdots, A_1$， 执行 **3** 对于 $b = 1, \cdots, B_1$， 执行 ^^ ^**4** 对于 $i \in [d+1]$， 计算 $\lambda_{a,b;i} := \lambda_i(P_{a,b}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{a,b}, P_{a,b}\boldsymbol{\Sigma}P_{a,b})$ 及其对应的广义特征向量 $\hat{\boldsymbol{v}}_{a,b;i} = \boldsymbol{v}_i(P_{a,b}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{a,b}, P_{a,b}\boldsymbol{\Sigma}P_{a,b})$ ^， 其中 $\lambda_{a,b;k+1} = 0$**5** 结束 **6** 计算 $b^*(a) := \text{sargmax}\sum_{i=1}^{d} \hat{\lambda}_{a,b;i}{}_{b\in[B_1]}$

7 结束

8 计算 $\widehat{\boldsymbol{w}} = (\hat{w}^{(1)}, \cdots, \hat{w}^{(p)})^\top$， 其中

$$\hat{w}^{(j)} := \frac{1}{A_1}\sum_{a=1}^{A_1}\sum_{i=1}^{d}(\hat{\lambda}_{a,b^*(a);i} - \hat{\lambda}_{a,b^*(a);d+1})(\hat{v}^{(j)}_{a,b^*(a);i})^2, \ j \in [p]$$

9 令 $\hat{S}'$ 为 $\widehat{\boldsymbol{w}}$ 中 l' 个最大分量的指标集。通过重复步骤 $1-8$， 使用子矩阵对 $(\widehat{\boldsymbol{\Sigma}}^{(\hat{S}', \hat{S}')}_{\mathbb{E}(\boldsymbol{X}|Y)}, \widehat{\boldsymbol{\Sigma}}^{(\hat{S}', \hat{S}')})$、投影参数 A_2， B_2 以及新定义的$\mathcal{P}_k := \{P_S : S \subseteq [l'], |S| = k\}$， 重新计算 l' 维权重向量 $\widehat{\boldsymbol{w}}'$

10 记 S $\widehat{\boldsymbol{w}}'$ 中 l 个最大分量的指标集^**输出：** $\boldsymbol{\beta} = (\hat{\boldsymbol{v}}_1, \ldots\ldots, \hat{\boldsymbol{v}}_d)$， 其中 $\hat{\boldsymbol{v}}_1, \ldots$ $\ldots, \hat{\boldsymbol{v}}_d$ 为^ ^$(P_{\hat{S}}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{\hat{S}}, P_{\hat{S}}\boldsymbol{\Sigma}P_{\hat{S}})$ 的前 d 个广义特征向量

变量选择过程。从这个角度来看，算法 2 模拟了两轮选择，这为重新评估变量的重要性并找回第一轮中误删的变量提供了机会，这些变量通常伴随着弱信号。

注释2。请注意，SSIRvRP 的估计量自然满足正交约束 $\widehat{\boldsymbol{\beta}}^\top\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}} = \boldsymbol{I}_d$。然而，改进的三步估计量（Tan 等人，2020年）通过放宽原始约束进行优化，因此若不经过归一化步骤，便无法满足正交约束。Lasso-SIR（Lin 等人，2019）同样不满足这一约束。这可以看作所提算法的另一项优势。

5 超参数的选择

在 SSIRvRP 算法中，实施算法前需要选择若干超参数。我们发现，对于$(A,\ B)$、$(A_1,\ B_1)$、k 和 l'的多种组合，所提算法都表现出相当稳健的性能，如以下数值实验所示。对于中心子空间的维度 d，已有若干方法可用于稀疏切片逆回归场景（Chen 等人，2010、Lin 等人，2019），因此我们将 d 视为已知常数。

稀疏度 l 是所提出方法中的一个关键调优参数。为选择该调优参数，我们最小化以下准则，该准则曾被Chen 等人（2010）使用：

$$-\log\left\{\mathrm{tr}(\widehat{\boldsymbol{\beta}}_l^\top \widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}\widehat{\boldsymbol{\beta}}_l)\right\}+\delta\cdot \mathrm{df}_l\,,$$

其中，$\widehat{\boldsymbol{\beta}}_l$ 表示给定稀疏度 l时 $\boldsymbol{\beta}$ 的解，df_l 表示有效参数个数，$\delta=2/n$ 对应 AIC 型准则，$\delta=\log(n)/n$ 对应 BIC 型准则。由于 $\widehat{\boldsymbol{\beta}}_l$ 的非零行数恰好为 l，因此我们可以用 $(l-d)\cdot d$估计 df_l。因此，我们通过最小化以下表达式来选择最佳的 l

$$-\log\left\{\mathrm{tr}(\widehat{\boldsymbol{\beta}}_l^\top \widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}\widehat{\boldsymbol{\beta}}_l)\right\}+\delta\cdot(l-d)\cdot d\,.$$

调优过程 l 的一项优势在于，它仅在算法2 的第 10 步中进行，而第 1 至 9 步只需计算一次。此外，由于 l 是整数，且其参数空间是可数且有限的，因此与在连续参数空间上进行调优的方法相比，该调优过程具有更高的计算效率。

6 模拟研究

在本节中，我们进行了大量数值实验，将所提方法与若干具有竞争力的方法进行比较，展示算法2中重新加权步骤的效果，并为所提算法中超参数的选择提供一些经验指导。

6.1 与现有方法的比较

我们将所提方法（算法 2）与 Tan 等人（2020年）提出的改进的三步估计量（RTS，以下简称 RTS）进行比较，该估计量已被证明在 $\log p = o(n)$ 时具有最优收敛率；此外，还与 Lin 等人（2019）提出的 Lasso-SIR 以及 Lin 等人（2017）提出的 DT-SIR 进行比较。已有证明表明，当 $p = o(n^2)$ 时，Lasso-SIR 具有最优收敛率；当 $\boldsymbol{X}$ 的协方差矩阵为单位矩阵时，DT-SIR 具有最优收敛率。我们将所提方法分为三种：第一种使用真实稀疏度（SSIRvRP），第二种使用由 BIC 准则调优的稀疏度（SSIRvRP-BIC），第三种使用由 AIC 准则调优的稀疏度（SSIRvRP-AIC）。

为公平起见，我们沿用 Tan 等人 (2020年中的模拟设置，并回顾其中的结果。考虑了五种模型，即，

模型 I： $Y = \boldsymbol{\beta}^\top \boldsymbol{X} + \sin(\boldsymbol{\beta}^\top \boldsymbol{X}) + \epsilon,$

模型 II： $Y = 2\arctan(\boldsymbol{\beta}^\top \boldsymbol{X}) + \epsilon,$

模型 III： $Y = (\boldsymbol{\beta}^\top \boldsymbol{X})^3 + \epsilon,$

模型 IV： $Y = \sinh(\boldsymbol{\beta}^\top \boldsymbol{X}) + \epsilon,$

模型 V： $Y = \exp(\boldsymbol{\beta}_1^\top \boldsymbol{X}) \cdot \text{sign}(\boldsymbol{\beta}_2^\top \boldsymbol{X}) + 0.2\epsilon,$

其中，$\boldsymbol{X} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$、$\epsilon \sim \mathcal{N}(0, 1)$ 以及 $\boldsymbol{X}$ 和 ϵ 相互独立。协方差矩阵 $\boldsymbol{\Sigma}$ 遵循以下四种结构：

(1) 单位协方差矩阵：$\boldsymbol{\Sigma} = \boldsymbol{I}_p$，(2) 稠密协方差矩阵：$\boldsymbol{\Sigma} = (\sigma_{ij})_{1 \le i,\ j \le p}$，其中 $\sigma_{ii} = 1$ 和 $\sigma_{ij} = 0.6$ 对于 $i \neq j$，(3) 托普利茨协方差矩阵：$\boldsymbol{\Sigma} = (\sigma_{ij})_{1 \le i,\ j \le p}$，其中 $\sigma_{ij} = 0.5^{|i-j|}$，(4) 稀疏逆协方差矩阵：$\boldsymbol{\Sigma}$ 是 $\boldsymbol{\Sigma}_0$ 的相关矩阵，且 $\boldsymbol{\Sigma}_0^{-1} = (w_{ij})_{1 \le i,\ j \le p}$，其中 $w_{ij} = \mathbf{1}_{\{i=j\}} + 0.5 \times \mathbf{I}_{\{|i-j|=1\}} + 0.4 \times \mathbf{I}\{|i-j|=2\}$。

回归系数 $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ 和 $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathbb{R}^{p \times 2}$ 设为有 5 个非零行，其索引从集合 $[p]$ 中随机选取。这些非零行中的元素是从有限集合 $\{-2, -1, 0, 1, 2\}$ 上的均匀分布中抽取的随机数。

对于所提算法的超参数，我们将 $(A_1, B_1) = (900, 300)$，$(A_2, B_2) = (600, 200)$，并将 $k = 20, l' = 50$ 和 d 设为其真实值。由于真实参数不满足约束 $\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} = \boldsymbol{I}_d$，我们无法直接根据 (3.2) 或 (3.6) 中给出的损失函数比较这些稀疏SIR估计。因此，我们使用相关性损失（Li, 1991, Tan 等人, 2020），其定义如下：

$$L_\rho(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = 1 - \frac{1}{d} \operatorname{tr} \left\{ (\widehat{\boldsymbol{\beta}}^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{\beta}})^{-1} (\widehat{\boldsymbol{\beta}}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}) (\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta})^{-1} (\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{\beta}}) \right\},$$

为衡量估计的效率，并将我们的方法与竞争方法进行比较。每次模拟重复 100 次，我们汇总了在各种(n, p) 配置下，五种模型与四种协方差结构所有组合的平均相关性损失，结果见表1-4。对于每种模型设置，最优算法的平均损失以粗体突出显示，次优算法的平均损失以斜体突出显示。

显然，在各种模型设置下，所提出的方法总体上优于其他方法。具体而言，在模型 I、III 和 IV 中，SSIRvRP、SSIRvRP-BIC 和 SSIRvRP-AIC 的表现远优于其他方法，平均损失降低了近一个数量级。然而，在模型 II 和 V 中，当样本量和维度都较低时，RTS 有时表现最佳，即 $(n, p) = (100,200)$ 和 $(100,400)$。随着样本量和维度增加，我们的 SSIRvRP 往往很快迎头赶上，平均损失显著降低。尽管已有证明，当 $\log(p) = o(n)$ 时 RTS 是最优的，这与我们的方法相同，但其有限样本表现通常不如我们的方法。原因可能在于，我们的方法在实现时不涉及初始点的选择，而其他方法基于梯度，因此容易受到初始化选择不当的影响。

此外，在比较 SSIRvRP-AIC 和 SSIRvRP-BIC 时，发现 SSIRvRP-BIC 在模型 II 和 III 中表现略好。然而，综合考察全部五个模型后，并未发现哪种方法更优。总体而言，两者表现相近。因此，任一种方法都可作为合适的工作算法。

最后，平均相关性损失会随着 n 的增加而降低，这一点可以从以下内容看出。

表 1：具有单位协方差结构时相关性损失的平均值。

$(n,\ p)$	模型	DT-SIR	Lasso-SIR	RTS	SSIRvRP	SSIRvRP-BIC	SSIRvRP-AIC
(100,200)	I	0.939	0.230	0.070	**0.015**	*0.024*	0.027
	II	0.926	0.389	**0.046**	*0.052*	0.053	0.122
	III	0.955	0.163	0.021	**0.002**	0.009	**0.002**
	IV	0.923	0.395	0.085	**0.002**	0.008	**0.002**
	V	0.769	0.404	**0.066**	*0.090*	0.120	0.112
(100,400)	I	0.954	0.435	0.033	**0.011**	*0.017*	0.032
	II	0.949	0.448	**0.015**	0.092	*0.081*	0.162
	III	0.863	0.355	0.039	**0.002**	0.007	*0.004*
	IV	0.888	0.611	0.120	**0.002**	0.005	*0.003*
	V	0.683	0.527	**0.091**	*0.138*	0.150	0.170
(100,600)	I	0.902	0.587	0.109	*0.023*	**0.022**	0.039
	II	0.879	0.675	0.175	*0.109*	**0.099**	0.196
	III	0.968	0.489	0.021	*0.006*	0.009	**0.005**
	IV	0.875	0.756	0.215	**0.003**	0.009	**0.003**
	V	0.701	0.561	0.179	**0.140**	*0.142*	0.179
(200,600)	I	0.997	0.094	0.012	**0.003**	**0.003**	0.018
	II	0.930	0.233	0.039	**0.010**	*0.014*	0.091
	III	0.944	0.058	0.008	**0.001**	**0.001**	**0.001**
	IV	0.970	0.212	0.020	**0.001**	**0.001**	**0.001**
	V	0.851	0.242	0.046	**0.032**	**0.039**	0.100
(400,600)	I	0.014	0.014	0.009	**0.002**	**0.002**	0.010
	II	0.924	0.027	0.013	**0.004**	*0.005*	0.047
	III	0.971	0.004	0.004	**0.000**	**0.000**	**0.000**
	IV	0.989	0.021	0.009	**0.000**	**0.000**	0.001
	V	0.874	0.045	0.019	**0.008**	*0.009*	0.051

表2：稠密协方差结构下相关性损失的平均值。

$(n,\ p)$	模型	DT-SIR	Lasso-SIR	RTS	SSIRvRP	SSIRvRP-BIC	SSIRvRP-AIC
(100,200)	I	0.472	0.106	0.143	**0.045**	0.054	*0.053*
	II	0.766	0.283	0.160	**0.130**	*0.129*	0.175
	III	0.904	0.483	0.074	**0.003**	0.024	*0.005*
	IV	0.902	0.697	0.298	**0.004**	0.028	*0.010*
	V	0.439	0.338	**0.129**	*0.173*	0.199	0.184
(100,400)	I	0.856	0.509	0.074	*0.057*	*0.057*	**0.054**
	II	0.828	0.887	**0.019**	0.135	*0.124*	0.194
	III	0.256	0.026	0.012	**0.004**	0.023	*0.008*
	IV	0.245	0.104	0.081	**0.006**	0.033	*0.010*
	V	0.432	0.410	**0.139**	*0.207*	0.208	0.218
（100,600）	I	0.883	0.600	0.301	*0.062*	**0.059**	0.080
	II	0.340	*0.159*	**0.131**	0.170	0.160	0.235
	III	0.928	0.926	0.284	**0.008**	0.027	*0.019*
	IV	0.849	0.544	0.453	**0.007**	0.028	*0.018*
	V	0.612	0.425	0.245	**0.195**	*0.197*	0.215
(200,600)	I	0.854	0.187	0.023	**0.007**	*0.012*	0.024
	II	0.507	0.085	0.096	*0.036*	**0.034**	0.088
	III	0.637	0.028	0.031	**0.002**	0.007	*0.004*
	IV	0.946	0.944	0.045	**0.002**	0.009	*0.005*
	V	0.548	0.307	0.086	*0.078*	**0.073**	0.101
(400,600)	I	0.822	0.050	0.005	**0.002**	*0.003*	0.010
	II	0.287	0.027	0.020	**0.010**	**0.010**	0.038
	III	0.472	0.006	0.019	**0.001**	**0.001**	**0.001**
	IV	0.591	0.035	0.038	**0.001**	**0.001**	0.002
	V	0.363	0.126	0.027	**0.024**	*0.026*	0.054

案例（n，p）=（100，600），（200，600），以及（400，600）），并且随着 p的增加而增大，这可从案例（n，p）=（100，200），（100，400），以及（100，600））中看出。这些结果与我们的理论发现相当一致。

表 5 报告了与表 3类似的结果，其中我们使用备注1 中的公式（2.8）来˜计算 $\mathbb{E}$（$\boldsymbol{X}|Y$）的样本协方差矩阵。两种估计方法的效果相似，因此您可以选择自己偏好的任意一种。

6.2 重新加权步骤

与算法 1相比，我们在算法 2 中加入了一个重新加权步骤，以帮助找回真实信号。图1展示了重新加权步骤对平均相关性

表 3：Toplitz 协方差结构下相关性损失的平均值。

$(n,\ p)$	模型	DT-SIR	Lasso-SIR	RTS	SSIRvRP	SSIRvRP-BIC	SSIRvRP-AIC
(100,200)	I	0.945	0.186	0.073	**0.013**	*0.019*	0.022
	II	0.943	0.297	**0.046**	*0.056*	0.058	0.099
	III	0.951	0.088	0.022	**0.002**	0.007	**0.002**
	IV	0.915	0.323	0.092	**0.002**	0.006	*0.002*
	V	0.794	0.345	**0.061**	*0.109*	0.119	0.117
(100,400)	I	0.952	0.283	0.033	**0.013**	*0.024*	0.034
	II	0.919	0.368	**0.017**	0.089	*0.081*	0.155
	III	0.917	0.254	0.042	*0.003*	0.005	**0.002**
	IV	0.907	0.477	0.141	**0.002**	0.006	**0.002**
	V	0.763	0.496	**0.099**	*0.127*	0.133	0.152
(100,600)	I	0.927	0.474	0.105	**0.022**	*0.026*	0.036
	II	0.874	0.581	0.187	*0.114*	**0.103**	0.188
	III	0.935	0.340	0.021	**0.005**	0.014	*0.006*
	IV	0.868	0.692	0.202	**0.003**	0.010	**0.003**
	V	0.765	0.528	0.201	*0.142*	**0.140**	0.178
(200,600)	I	0.932	0.070	0.010	**0.004**	**0.004**	0.018
	II	0.926	0.150	0.034	**0.012**	*0.016*	0.084
	III	0.831	0.030	0.008	**0.001**	**0.001**	**0.001**
	IV	0.938	0.151	0.017	**0.001**	**0.001**	0.002
	V	0.603	0.179	0.040	**0.028**	*0.035*	0.079
(400,600)	I	0.372	0.014	0.008	**0.002**	**0.002**	0.009
	II	0.289	0.021	0.012	**0.004**	*0.005*	0.043
	III	0.176	0.004	0.003	**0.000**	**0.000**	**0.000**
	IV	0.577	0.021	0.010	**0.001**	**0.001**	**0.001**
	V	0.199	0.039	0.018	**0.007**	*0.009*	0.045

损失和信号率的影响。这里，信号率概括了算法准确识别所有真实信号的频率。显然，在各种模型设置下，重新加权步骤显著降低了相关性损失并提高了信号率。因此，我们建议使用算法 2 来解决稀疏切片逆回归问题。

6.3 超参数的选择

在本小节中，我们通过数值实验，为算法2中超参数的选择提供一些指导。总的来说，所提出的方法在较大范围的超参数组合下似乎都具有相当稳健的性能。

图 2 展示了相关性损失与 A、 B以及它们的组合之间的关系

表4：采用稀疏逆协方差结构时的平均相关性损失。

$(n,\ p)$	模型	DT-SIR	Lasso-SIR	RTS	SSIRvRP	SSIRvRP-BIC	SSIRvRP-AIC
(100,200)	I	0.876	0.153	0.084	**0.015**	*0.023*	*0.023*
	II	0.906	0.265	**0.060**	0.100	*0.086*	0.127
	III	0.848	0.056	0.024	**0.002**	0.004	**0.002**
	IV	0.895	0.280	0.108	**0.002**	0.010	**0.002**
	V	0.557	0.267	**0.068**	*0.129*	0.142	*0.129*
(100,400)	I	0.930	0.203	0.040	*0.029*	**0.026**	0.031
	II	0.943	0.238	**0.016**	0.108	*0.088*	0.159
	III	0.851	0.133	0.039	**0.003**	0.009	**0.003**
	IV	0.873	0.327	0.118	**0.003**	0.004	**0.003**
	V	0.685	0.550	**0.087**	*0.151*	0.154	0.170
(100,600)	I	0.930	0.388	0.147	*0.036*	**0.032**	0.046
	II	0.930	0.431	0.183	*0.119*	**0.105**	0.172
	III	0.926	0.192	0.023	*0.003*	0.008	**0.002**
	IV	0.893	0.576	0.241	**0.002**	*0.006*	*0.006*
	V	0.778	0.433	**0.159**	0.191	*0.184*	0.201
(200,600)	I	0.452	0.033	0.011	*0.005*	**0.004**	0.017
	II	0.704	0.082	0.038	**0.012**	*0.016*	0.069
	III	0.255	0.010	0.007	**0.001**	**0.001**	**0.001**
	IV	0.730	0.086	0.021	**0.001**	**0.001**	**0.001**
	V	0.319	0.146	0.048	**0.031**	*0.038*	0.073
(400,600)	I	0.328	0.016	0.012	**0.002**	**0.002**	0.006
	II	0.228	0.023	0.012	**0.003**	*0.004*	0.034
	III	0.334	0.007	0.011	**0.001**	**0.001**	**0.001**
	IV	0.248	0.017	0.008	**0.001**	**0.001**	**0.001**
	V	0.257	0.051	0.018	**0.008**	*0.010*	0.045

$(A,\ B)$，同时保持其他参数不变。当 B 固定时，增大 A 不会显著降低相关性损失；当 A 固定时，也会出现类似的趋势。当同时增大 A 和 B，同时保持 $B = A/3$ 不变时，图中显示，对于较小的 A，相关性损失呈现明显的下降趋势。当 A 增大至 400 时，相关性损失的下降趋势趋于平缓。此外，当样本量增加至 400 和 600 时，对于 A、B 和 $(A,\ B)$ 的所有取值，损失似乎都具有较好的稳健性。

对于 k的选择，图 3 显示，在样本量较小时，可能存在某个最优的 k；而当 n 增大时，则没有明显趋势。图 4 表明，当 $n = 200$时，损失对 l' 的选择具有稳健性。对于 $n = 100$的情况，图中显示，当 l' 较小时呈下降趋势，而随着 l' 增大，趋势逐渐趋于平缓。

表5：采用 2.8 对 $\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}$ 进行估计时，Toplitz 协方差结构下相关性损失的平均值。

$(n,\ p)$	模型	切片逆回归	Lasso-SIR	RTS	SSIRvRP	SSIRvRP-BIC	SSIRvRP-AIC
(100, 200)	I	0.945	0.186	0.073	0.011	0.017	0.022
	II	0.943	0.297	0.046	0.060	0.055	0.120
	III	0.951	0.088	0.022	0.002	0.009	0.002
	IV	0.915	0.323	0.092	0.002	0.008	0.003
	V	0.794	0.345	0.061	0.125	0.135	0.157
(100, 400)	I	0.952	0.283	0.033	0.016	0.018	0.036
	II	0.919	0.368	0.017	0.119	0.111	0.166
	III	0.917	0.254	0.042	0.003	0.006	0.003
	IV	0.907	0.477	0.141	0.003	0.008	0.003
	V	0.763	0.496	0.099	0.142	0.157	0.177
(100, 600)	I	0.927	0.474	0.105	0.024	0.026	0.038
	II	0.874	0.581	0.187	0.104	0.098	0.186
	III	0.935	0.340	0.021	0.003	0.004	0.003
	IV	0.868	0.692	0.202	0.002	0.007	0.006
	V	0.765	0.528	0.201	0.125	0.135	0.157
(200, 600)	I	0.932	0.070	0.010	0.004	0.003	0.018
	II	0.926	0.150	0.034	0.012	0.016	0.081
	III	0.831	0.030	0.008	0.001	0.001	0.002
	IV	0.938	0.151	0.017	0.001	0.001	0.001
	V	0.603	0.179	0.040	0.031	0.038	0.093
(400, 600)	I	0.372	0.014	0.008	0.002	0.002	0.008
	II	0.289	0.021	0.012	0.003	0.005	0.039
	III	0.176	0.004	0.003	0.000	0.000	0.001
	IV	0.577	0.021	0.010	0.000	0.000	0.000
	V	0.199	0.039	0.019	0.009	0.012	0.047

总而言之，当收集到数量较多的样本时，无需担心超参数选择。然而，当样本量有限时，我们建议为 (A, B) 和 l' 选择较大的值，并为 k 选择中等尺度。

7 实证数据研究

在本节中，我们考虑来自国际“HapMap”项目的基因表达数据（Thorisson 等人，2005 年）。该数据包含90个样本，其中45个中国人、45个日本人，并报告了47293个探针的基因表达水平。基因CHRNA6是许多尼古丁成瘾研究的对象（Thorgeirsson 等人，2010），我们将其

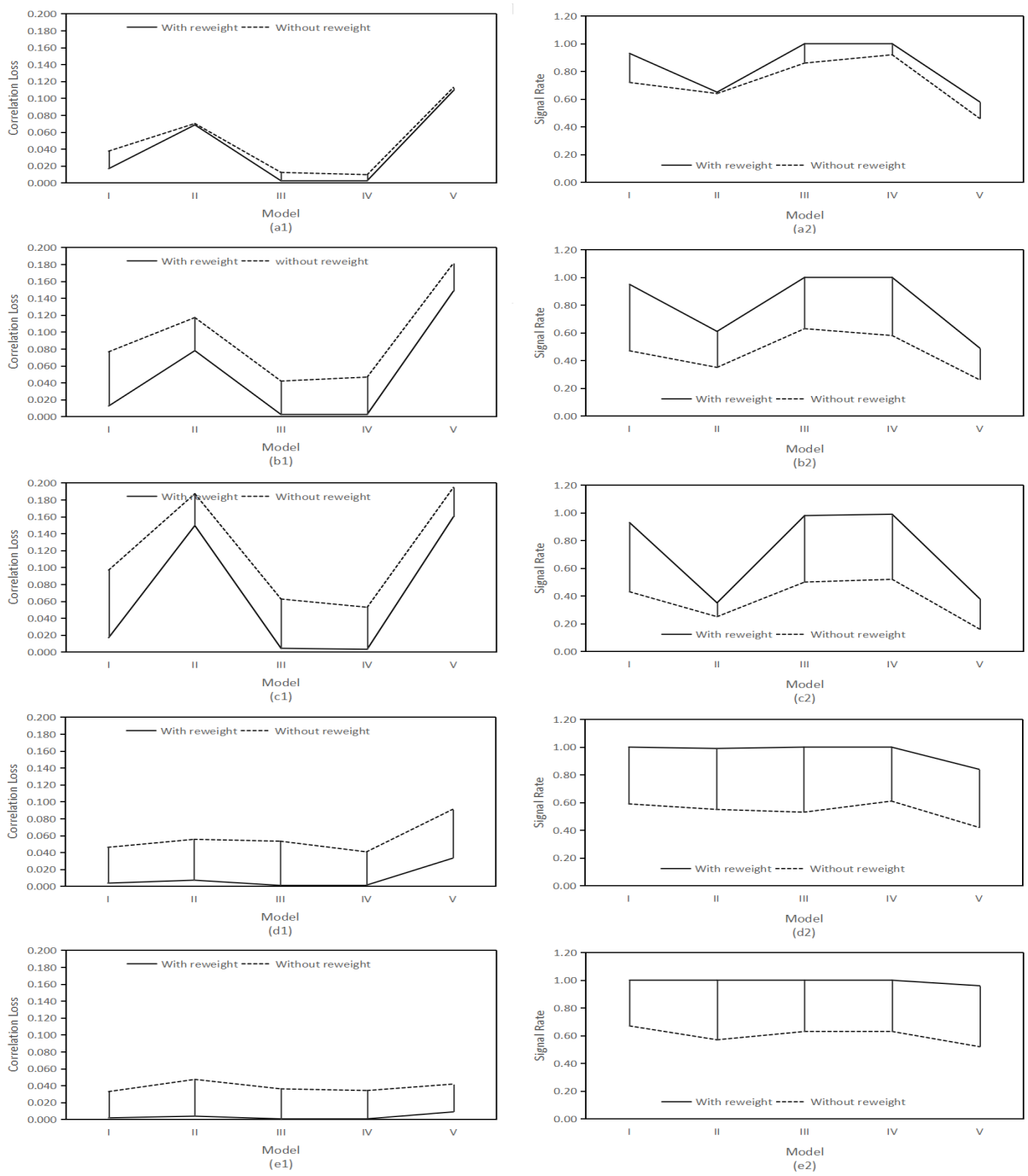


图1：Toplitz 协方差结构下重新加权步骤的效果。(a1)-(e1) 和 (a2)-(e2) 对应于 (n， p) 设置：(100，200)-(400，600)。

mRNA表达作为响应变量， Y 与Fan 等人(2018)和Lin 等人(2019)中的做法相同。其他47292个基因的表达量作为协变量。因此，我们现在面临一个 $p = 47292 \gg n = 90$。

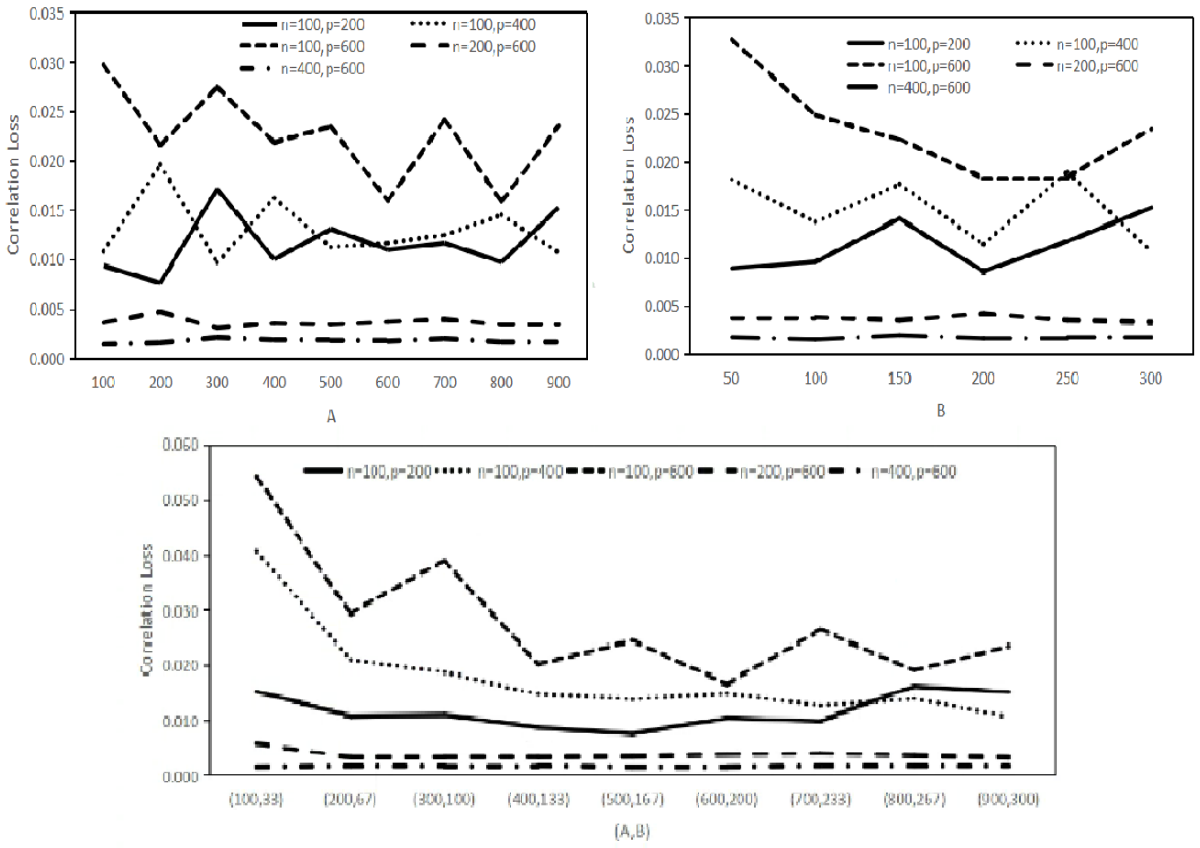


图2：单位协方差矩阵下模型 I 的 A 和 B 的选择。

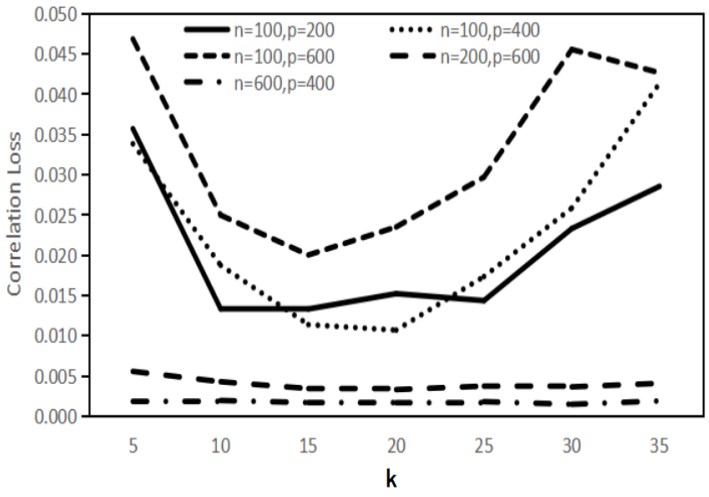


图3： k 的选择。

我们的目标是找出几个由少数基因组成的组合，以表征 Y 与 $\boldsymbol{X}$之间的回归关系，而所提出的 SSIRvRP 方法非常适合这一任务。

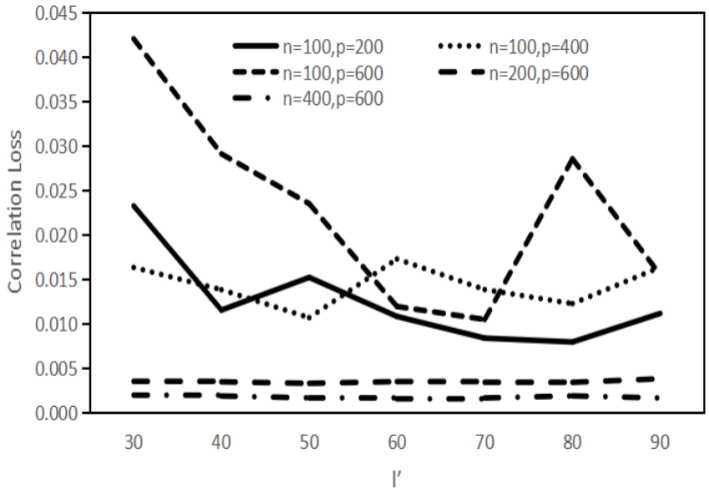


图4： l' 的选择。

按照 Lin 等人 (2019) 的做法，我们将 $d = 1$ 和 $l = 13$ 以及其他超参数设定为与模拟中相同的值。根据估计得到的系数 $\widehat{\boldsymbol{\beta}}$，我们定义 $Z = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}$ 并在图Y中绘制其相对于 Z 的图像，见图 5。随后我们发现Y与 Z之间存在中等程度的二次关系。受此发现启发，我们进一步拟合 Y 与Z之间的回归模型， Z^2得到调整后的 R 平方为 0.620，p 值为 0.000。拟合模型的均方误差为 0.040。这些结果似乎略优于Lin 等人 (2019) 的结果，他们得到的 R 平方为 0.578，均方误差为 0.044。

<escape>作为比较，我们还采用了 Gataric 等人~(2020年 提出的稀疏主成分分析算法来估计系数 $\boldsymbol{\beta}$ ，其中缺少 Y 的信息，并计算$\sim Z' = \boldsymbol{\beta}^\top \boldsymbol{X}$。 Y 与 Z' 的散点图并未显示这两个变量之间存在任何有趣的模式。对 Y 和 Z' 进行线性回归分析，得到的 R 平方值为 0.001，p 值为 0.792，这表明这两个变量之间不存在线性关系。与由 Y监督的所提出方法所取得的结果相比，这一结果似乎意义较小。</escape>

8 讨论

本文提出了一种随机投影方法，用于在 $p \gg n$ 时估计稀疏切片逆回归中的中心子空间。与现有方法相比，所提算法的计算

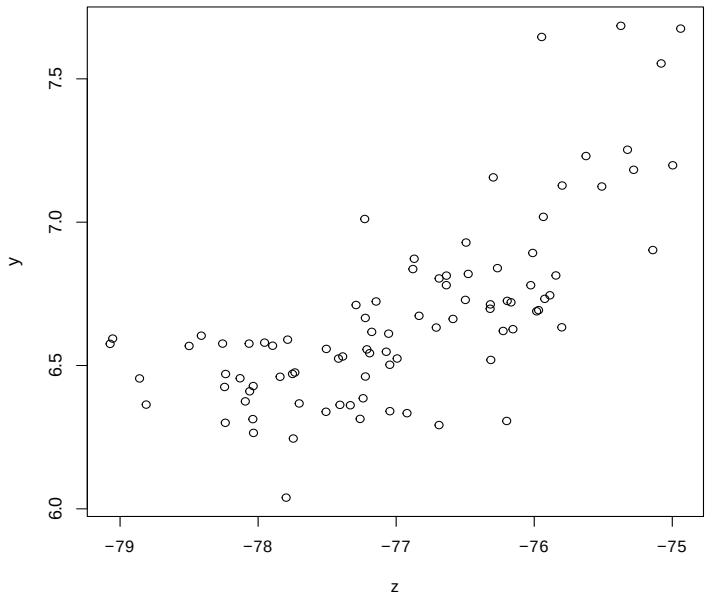


图 5：二次关系。

开销更低、效率更高。理论上，我们证明了在适当假设下，所提估计量达到极小极大最优速率。值得注意的是，引入随机投影技术是为了解决一个广义特征值—特征向量问题。因此，该技术也可应用于稀疏费希尔判别分析进行分类，以及稀疏典型相关分析，以探索两个高维随机向量之间的关系。我们将其留待进一步研究。

9 附录

9.1 定理1的证明

下面的引理是必需的，其证明见第9.4.1节。

引理 1. 在假设 (A1)下，存在一个事件Ω，其概率至少为 $1 - cp^{-3}$，使得在Ω上，以下结论成立：

$$\sup_{\boldsymbol{u}\in\mathcal{B}_0^{p-1}(k)}|\boldsymbol{u}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}-\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)})\boldsymbol{u}|\leq K\sqrt{\frac{k\log(p)}{n}}$$

条件是 k对数 $p \ll n$，其中 $\mathcal{B}_0^{p-1}(k) = \{\boldsymbol{v} = (v^{(1)}, \ldots, v^{(p)})^\top \in \mathbb{R}^p : \quad \|\boldsymbol{v}\|_2 = 1, \sum_{j=1}^p \mathbf{1}\{v^{(j)} \neq 0\} \leq k\}$ 对于 $k \in [p]$，且 K 是一个依赖于 H的常数。

我们开始证明。该证明与 Gataric 等人 (2020年) 的定理 1 证明类似，但有些细节需要调整，因为现在我们有协方差的切片估计量^矩阵。为简化记号，我们省略下标 $\mathbb{E}(\boldsymbol{X}|Y)$ 的 $\boldsymbol{\Sigma}$，并且 $\boldsymbol{\Sigma}$。定义 $\mathcal{S}_k := \{S \subset [p]: \quad |S| = k\}$ 并回顾 $\mathcal{A} = \{j \in [p]: \boldsymbol{\beta}^{(j,\cdot)} \neq \mathbf{0}\}$ 它表示 $\boldsymbol{\beta}$的非零行集合。对于任意 $S \in \mathcal{S}_k$，注意到 $\boldsymbol{\Sigma}^{(S,S)} = \boldsymbol{\beta}^{(S,\cdot)} \boldsymbol{\Lambda} (\boldsymbol{\beta}^{(S,\cdot)})^\top$。因此，

$$\sum_{i=1}^d \lambda_i(\boldsymbol{\Sigma}^{(S,S)}) = \sum_{i=1}^k \lambda_i(\boldsymbol{\Sigma}^{(S,S)}) - \sum_{i=d+1}^k \lambda_i(\boldsymbol{\Sigma}^{(S,S)}) = \operatorname{tr}(\boldsymbol{\Sigma}^{(S,S)}) = \sum_{i=1}^d \sum_{j \in S \cap \mathcal{A}} \lambda_i(\beta^{(j,i)})^2. \quad (9.1)$$

接着，根据引理 1，我们知道存在一个事件 Ω，其概率至少为 $1 - cp^3$，使得在 Ω 上，对任意 $S \in \mathcal{S}_k$,

$$\|\widehat{\boldsymbol{\Sigma}}^{(S,S)} - \boldsymbol{\Sigma}^{(S,S)}\|_{\mathrm{op}} \leq K \sqrt{\frac{k \log p}{n}},$$

其中 $K > 0$ 是一个与 S无关的常数。在上述定义的事件 Ω 上，根据 Weyl 不等式，我们有

$$\begin{aligned} \left| \sum_{i=1}^d \lambda_i(\widehat{\boldsymbol{\Sigma}}^{(S,S)}) - \sum_{i=1}^d \sum_{j \in S \cap \mathcal{A}} \lambda_i(\beta^{(j,i)})^2 \right| &\overset{(9.1)}{=} \left| \sum_{i=1}^d \{\lambda_i(\widehat{\boldsymbol{\Sigma}}^{(S,S)}) - \lambda_i(\boldsymbol{\Sigma}^{(S,S)})\} \right| \\ &\leq K d \sqrt{\frac{k \log p}{n}} \overset{(3.3)}{\leq} \frac{d\lambda_d}{16 s \mu^2}. \end{aligned}$$

因此，在事件 Ω 上，如果 $(S \cap \mathcal{A}) \subsetneq (S' \cap \mathcal{A})$，那么

$$\begin{aligned} &\sum_{i=1}^d \lambda_i(\widehat{\boldsymbol{\Sigma}}^{(S,S)}) - \sum_{i=1}^d \lambda_i(\widehat{\boldsymbol{\Sigma}}^{(S',S')}) \\ \leq &\sum_{i=1}^d \sum_{j \in S \cap \mathcal{A}} \lambda_i(\beta^{(j,i)})^2 - \sum_{i=1}^d \sum_{j \in S' \cap \mathcal{A}} \lambda_i(\beta^{(j,i)})^2 + \frac{d\lambda_d}{8 s \mu^2} \\ \leq &-\sum_{i=1}^d \sum_{j \in (S' \setminus S) \cap \mathcal{A}} \lambda_i(\beta^{(j,i)})^2 + \frac{d\lambda_d}{8 s \mu^2}. \end{aligned} \quad (9.2)$$

注意，表达式 (3.1) 意味着

$$\sum_{i=1}^{d} \lambda_i (\beta^{(j,i)})^2 \geq \lambda_d \|\boldsymbol{\beta}^{(j,\cdot)}\|_2^2 \geq \frac{d\lambda_d}{s\mu^2} \tag{9.3}$$

对任意 $j \in \mathcal{A}$。然后，结合式（9.2）和（9.3），可得

$$\sum_{i=1}^{d} \lambda_i (\widehat{\boldsymbol{\Sigma}}^{(S,S)}) < \sum_{i=1}^{d} \lambda_i (\widehat{\boldsymbol{\Sigma}}^{(S',S')}) \,. \tag{9.4}$$

根据上述表达式，我们现在证明，在事件 Ω 上，对某些固定的 $j \in \mathcal{A}$ 和$j' \notin \mathcal{A}$，有

$$q_j \geq q_{j'} \,, \tag{9.5}$$

其中，对任意 $k \in [p]$，$q_k = \mathbb{P}(k \in S_{a,b^*(a)} | \boldsymbol{X})$ 且某个固定的 $a \in [A]$。我们开始证明。~ 对 $j \in \{j, j'\}$ 和 $b \in [B]$，定义以下集合:

$$\mathcal{S}_{b,\tilde{j}} := \{(S_{a,1}, \ldots, S_{a,B}) : b^*(a) = b, \tilde{j} \in S_{a,b}\}, \quad \text{and}$$
$$\mathcal{S}_b := \{(S_{a,1}, \ldots, S_{a,B}) : b^*(a) = b\} \,.$$

定义映射 $\psi : \mathcal{S}_k \to \mathcal{S}_k$，使得

$$\psi(S) := \begin{cases} (S \setminus \{j'\}) \cup \{j\} & \text{if } j' \in S \text{ and } j \notin S, \\ S & \text{otherwise.} \end{cases}$$

于是，对每个 $S \in \mathcal{S}_k$，要么 $\psi(S) = S$，要么 $S \cap \mathcal{A} \subsetneq \psi(S) \cap \mathcal{A}$。因此，不等式（9.4）意味着

$$\sum_{i=1}^{d} \lambda_i (\widehat{\boldsymbol{\Sigma}}^{(S,S)}) \leq \sum_{i=1}^{d} \lambda_i (\widehat{\boldsymbol{\Sigma}}^{(\psi(S),\psi(S))}) \,. \tag{9.6}$$

此外，根据 $\psi(\cdot)$ 的定义，我们知道，如果 $j' \in S$ 对于某个 $S \in \mathcal{S}_k$ 成立，那么 $j \in \psi(S)$。因此，对于任意 $(S_{a,1}, \ldots, S_{a,b}, \ldots, S_{a,B}) \in \mathcal{S}_{b,j'}$，都存在 $(S_{a,1}, \ldots, \psi(S_{a,b}) \ldots, S_{a,B}) \in \mathcal{S}_{b,j}$。

显然，由于 $(S_{a,1}, \ldots\ldots, S_{a,b}, \ldots\ldots, S_{a,B}) \in \mathcal{S}_{b,j'}$，那么 $b^*(a) = b$ 且 $j' \in S_{a,b}$。注意，$b^*(a) = b$ 意味着

$$\max_{b\in[B]} \sum_{i=1}^{d} \lambda_i(\widehat{\mathbf{\Sigma}}^{(S_{a,b},S_{a,b})}) \le \sum_{i=1}^{d} \lambda_i(\widehat{\mathbf{\Sigma}}^{(S_{a,b},S_{a,b})}) .$$

结合式（9.6），可知

$$\max_{b\in[B]} \sum_{i=1}^{d} \lambda_i(\widehat{\mathbf{\Sigma}}^{(S_{a,b},S_{a,b})}) \le \sum_{i=1}^{d} \lambda_i(\widehat{\mathbf{\Sigma}}^{(\psi(S_{a,b}),\psi(S_{a,b}))}) ,$$

这意味着 $(S_{a,1}, \ldots, \psi(S_{a,b}) \ldots, S_{a,B})$ 满足 $b^*(a) = b$。结合 $j \in \psi(S_{a,b})$ 这一事实，可得 $(S_{a,1}, \ldots, \psi(S_{a,b}) \ldots, S_{a,B}) \in \mathcal{S}_{b,j}$。因此，$|\mathcal{S}_{b,j'}| \le |\mathcal{S}_{b,j}|$。于是，在 Ω 上，对所有 $b \in [B]$，都有

$$\begin{aligned}
&\mathbb{P}\{j \in S_{a,b^*(a)} | \boldsymbol{X}, b^*(a) = b\} \\
=& \frac{\mathbb{P}\{j \in S_{a,b^*(a)}, b^*(a) = b | \boldsymbol{X}\}}{\mathbb{P}\{b^*(a) = b | \boldsymbol{X}\}} = \frac{|\mathcal{S}_{b,j}|}{|\mathcal{S}_b|} \\
\ge& \frac{|\mathcal{S}_{b,j'}|}{|\mathcal{S}_b|} = \frac{\mathbb{P}\{j' \in S_{a,b^*(a)}, b^*(a) = b | \boldsymbol{X}\}}{\mathbb{P}\{b^*(a) = b | \boldsymbol{X}\}} \\
=& \mathbb{P}\{j' \in S_{a,b^*(a)} | \boldsymbol{X}, b^*(a) = b\} .
\end{aligned}$$

因此，我们已证明 $q_j \ge q_{j'}$ 如式（9.5）所示。

注意到

$$\sum_{\tilde{j}\in[p]} q_{\tilde{j}} = \sum_{\tilde{j}\in[p]} \mathbb{P}\{\tilde{j} \in S_{a,b^*(a)} | \boldsymbol{X}\} = \frac{1}{B} \sum_{b\in[B]} \sum_{\tilde{j}\in[p]} \mathbb{P}\{\tilde{j} \in S_{a,b^*(a)} | b^*(a) = b, \boldsymbol{X}\} = \sum_{\tilde{j}\in[p]} \frac{k}{p} = k .$$

因此，由（9.5）可得，在Ω 上

$$q_j \ge \frac{\sum_{\tilde{j}\in([p]\setminus\mathcal{A}\cup\{j\})} q_{\tilde{j}}}{p-s+1} = \frac{k - \sum_{\tilde{j}\in\mathcal{A}\setminus\{j\}} q_{\tilde{j}}}{p-s+1} \ge \frac{k-s+1}{p-s+1} \ge \frac{1}{p} . \tag{9.7}$$

注 3。 由于 $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I}_p$, $b^*(a)$ 在算法 1 简化为

$$b^*(a) := \mathrm{sargmax}_{b\in[B]} \sum_{i=1}^{d} \hat{\lambda}_{a,b;i} \,,$$

其中 $\hat{\lambda}_{a,b;i} = \lambda_i(P_{a,b}\boldsymbol{\Sigma}P_{a,b})$ 是矩阵 $P_{a,b}\boldsymbol{\Sigma}P_{a,b}$ 的第 i 大特征值。

定义 $\lambda_{a,b;i} := \lambda_i(P_{a,b}\boldsymbol{\Sigma}P_{a,b})$，以及对应的特征向量 $\boldsymbol{v}_{a,b;i} := \boldsymbol{v}_i(P_{a,b}\boldsymbol{\Sigma}P_{a,b})$，分别对应于 $b \in [B]$ 和 $i \in [p]$。注意， $\boldsymbol{\Sigma}$ 有 d 个非零特征值，$P_{a,b}\boldsymbol{\Sigma}P_{a,b}$ 也如此。于是 $\hat{\lambda}_{a,b;d+1} = \ldots = \hat{\lambda}_{a,b;p} = 0$。记 $\boldsymbol{V}_{a,b} := (\boldsymbol{v}_{a,b;1}, \ldots, \boldsymbol{v}_{a,b;d})$, $\widehat{\boldsymbol{V}}_{a,b} := (\hat{\boldsymbol{v}}_{a,b;1}, \ldots, \hat{\boldsymbol{v}}_{a,b;d})$, $\boldsymbol{\Lambda}_{a,b} = \mathrm{diag}(\lambda_{a,b;1}, \ldots, \lambda_{a,b;d})$ 和 $\widehat{\boldsymbol{\Lambda}}_{a,b} = \mathrm{diag}(\hat{\lambda}_{a,b;1} - \hat{\lambda}_{a,b;d+1}, \ldots, \hat{\lambda}_{a,b;d} - \hat{\lambda}_{a,b;d+1})$。对于 $j \in \{j, j'\}$, 令

$$\hat{w}_a^{(\tilde{j})} := (\widehat{\boldsymbol{V}}_{a,b^*(a)}\widehat{\boldsymbol{\Lambda}}_{a,b^*(a)}\widehat{\boldsymbol{V}}_{a,b^*(a)}^\top)^{(\tilde{j},\tilde{j})} = \sum_{i=1}^{d}(\hat{\lambda}_{a,b^*(a);i} - \hat{\lambda}_{a,b^*(a);d+1})(\hat{v}_{a,b^*(a);i}^{(\tilde{j})})^2.$$

现在给出 $\hat{w}_a^{(\tilde{j})}$ 的上界和下界。

注意,

$$\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top = \sum_{i=1}^{p} \lambda_{a,b^*(a);i}\boldsymbol{v}_{a,b^*(a);i}\boldsymbol{v}_{a,b^*(a);i}^\top = P_{a,b^*(a)}\boldsymbol{\Sigma}P_{a,b^*(a)} \,,$$

这意味着，当 $\tilde{j} \in S_a$ 时，$(\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(\tilde{j},\tilde{j})} = \Sigma^{(\tilde{j},\tilde{j})}$；否则为 $(\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(\tilde{j},\tilde{j})} = 0$。根据 Gataric 等人（2020）附录 A.5 中的引理 2，在 Ω 上有

$$\begin{aligned}
&|(\widehat{\boldsymbol{V}}_{a,b^*(a)}\widehat{\boldsymbol{\Lambda}}_{a,b^*(a)}\widehat{\boldsymbol{V}}_{a,b^*(a)}^\top)^{(\tilde{j},\tilde{j})} - (\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(\tilde{j},\tilde{j})}| \\
&\overset{(1)}{\leq} \|\widehat{\boldsymbol{V}}_{a,b^*(a)}\widehat{\boldsymbol{\Lambda}}_{a,b^*(a)}\widehat{\boldsymbol{V}}_{a,b^*(a)}^\top - \boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top\|_{\mathrm{op}} \leq 4d\|P_{a,b^*(a)}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})P_{a,b^*(a)}\|_{\mathrm{op}} \\
&\overset{(2)}{\leq} 4Kd\sqrt{\frac{k\log(p)}{n}} \overset{(3.3)}{\leq} \frac{d\lambda_d}{4s\mu^2} \,,
\end{aligned} \tag{9.8}$$

其中，不等式 (1) 根据矩阵算子范数的定义得到，而 (2) 则由引理 1 推得。因此，在 $\Omega \cap \{j \in S_{a,b^*(a)}\}$ 上，我们有

$$\begin{aligned}
\hat{w}_a^{(j)} &= (\widehat{\boldsymbol{V}}_{a,b^*(a)}\widehat{\boldsymbol{\Lambda}}_{a,b^*(a)}\widehat{\boldsymbol{V}}_{a,b^*(a)}^\top)^{(j,j)} - (\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(j,j)} + (\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(j,j)} \\
&\geq \Sigma^{(j,j)} - |(\widehat{\boldsymbol{V}}_{a,b^*(a)}\widehat{\boldsymbol{\Lambda}}_{a,b^*(a)}\widehat{\boldsymbol{V}}_{a,b^*(a)}^\top)^{(j,j)} - (\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(j,j)}| \overset{(9.8)}{\geq} \Sigma^{(j,j)} - \frac{d\lambda_d}{4s\mu^2}
\end{aligned}$$

$$\geq \lambda_d \|\boldsymbol{V}^{(j,\cdot)}\|_2^2 - \frac{d\lambda_d}{4s\mu^2} \overset{(3.1)}{\geq} \frac{3d\lambda_d}{4s\mu^2}. \tag{9.9}$$

类似地，在 $\Omega \cap \{j \in S_{a,b^*(a)}\}$ 上，我们有

$$\hat{w}_a^{(j)} \leq \frac{5d\lambda_1\mu^2}{4s}. \tag{9.10}$$

此外，在 $\Omega \cap \{j' \in S_{a,b^*(a)}\}$ 上，有

$$-\frac{d\lambda_d}{4s\mu^2} \leq \hat{w}_a^{(j')} \leq \frac{d\lambda_d}{4s\mu^2} \tag{9.11}$$

注意到 $j' \in \mathcal{A}^c$ 且 $\Sigma^{(j',j')} = \sum_{i=1}^d \lambda_i (\boldsymbol{\beta}_i^{(j')})^2 = 0$。最后，对于所有 $j \in [p]$，若 $j \notin S_{a,b^*(a)}$，则根据 $\hat{\boldsymbol{v}}_a$ 的定义，$\hat{v}_{a,b^*(a),i}$ 对于 $i \in [d]$，我们知道 $(\hat{v}_{a,b^*(a),i})^{(j)} = 0$ 对于满足 $\hat{\lambda}_{a,b^*(a),i} > 0$ 的 i。因此，根据 $\hat{w}_a^{(j)}$ 的定义，有 $\hat{w}_a^{(j)} = 0$ 对于 $j \notin S_{a,b^*(a)}$。于是，我们给出了 $\hat{w}_a^{(\tilde{j})}$ 的上界和下界。利用上述给出的下界和上界，我们得到在 Ω 上

$$\begin{aligned}
&\mathbb{E}(\hat{w}_a^{(j)} - \hat{w}_a^{(j')} | \boldsymbol{X}) \\
&= \mathbb{E}[\hat{w}_a^{(j)}(\mathbf{1}\{j \in S_{a,b^*(a)}\} + \mathbf{1}\{j \notin S_{a,b^*(a)}\}) | \boldsymbol{X}] - \mathbb{E}[\hat{w}_a^{(j')}(\mathbf{1}\{j' \in S_{a,b^*(a)}\} + \mathbf{1}\{j' \notin S_{a,b^*(a)}\}) | \boldsymbol{X}] \\
&= \mathbb{E}(\hat{w}_a^{(j)} \mathbf{1}\{j \in S_{a,b^*(a)}\} - \hat{w}_a^{(j')} \mathbf{1}\{j' \in S_{a,b^*(a)}\} | \boldsymbol{X}) \\
&\geq \frac{3q_j d\lambda_d}{4s\mu^2} - \frac{q_{j'} d\lambda_d}{4s\mu^2} \overset{(9.5)}{\geq} \frac{q_j d\lambda_d}{2s\mu^2} \overset{(9.7)}{\geq} \frac{d\lambda_d}{2ps\mu^2}.
\end{aligned} \tag{9.12}$$

现在，我们再次令 a、j 和 j' 自由变化。定义 $\Omega_1 := \{\min_{j \in \mathcal{A}} \hat{w}^{(j)} > \max_{j \in \mathcal{A}^c} \hat{w}^{(j)}\}$。由于 $l \geq s$，在 Ω_1，上有 $\mathcal{A} \subseteq S$，这意味着 d 个主特征向量 of $P_{\hat{S}}^\top \boldsymbol{\Sigma} P_{\hat{S}}$ 与 $\boldsymbol{\Sigma}$ 的主特征向量相同。因此，根据 Yu 等人（2015）的引理 1 和定理 2，在 $\Omega \cap \Omega_1$，

$$L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq \frac{2d^{1/2}\|P_{\hat{S}}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})P_{\hat{S}}\|_{\mathrm{op}}}{\lambda_d} \leq 2K\sqrt{\frac{dl\log(p)}{n\lambda_d^2}}.$$

那么，只需对 $\mathbb{P}(\Omega \cap \Omega_1)$ 的下界求导。注意到 $\hat{w}^{(j)} = A^{-1}\sum_{a=1}^A \hat{w}_a^{(j)}$

对任意 $j \in [p]$ 均成立。于是，对于任意 $j \in \mathcal{A}$ 和 $j' \in \mathcal{A}^c$，在 Ω 上有

$$\begin{aligned}
&\hat{w}^{(j)} - \hat{w}^{(j')} \\
&= \{\hat{w}^{(j)} - \mathbb{E}(\hat{w}^{(j)}|\boldsymbol{X})\} - \{\hat{w}^{(j')} - \mathbb{E}(\hat{w}^{(j')}|\boldsymbol{X})\} + \mathbb{E}(\hat{w}^{(j)} - \hat{w}^{(j')}|\boldsymbol{X}) \\
&\overset{(9.12)}{\geq} \frac{d\lambda_d}{2ps\mu^2} + \{\hat{w}^{(j)} - \mathbb{E}(\hat{w}^{(j)}|\boldsymbol{X})\} - \{\hat{w}^{(j')} - \mathbb{E}(\hat{w}^{(j')}|\boldsymbol{X})\},
\end{aligned}$$

这意味着

$$\Omega_1^c \subseteq \left[\bigcup_{j\in\mathcal{A}} \left\{\hat{w}^{(j)} - \mathbb{E}(\hat{w}^{(j)}|\boldsymbol{X}) \leq \frac{d\lambda_d}{4ps\mu^2}\right\}\right] \cup \left[\bigcup_{j'\notin\mathcal{A}} \left\{\hat{w}^{(j')} - \mathbb{E}(\hat{w}^{(j')}|\boldsymbol{X}) \geq \frac{d\lambda_d}{4ps\mu^2}\right\}\right].$$

注意到，给定 $\boldsymbol{X}$，$\{\hat{w}_a^{(j)}\}_{a\in[A]}$ 是独立同分布随机变量，并且（9.9）、（9.10）和（9.11）表明，$\hat{w}_a^{(j)}$在 Ω 上有界，对所有 $j \in [p]$均成立。因此，根据并集界以及给定 $\boldsymbol{X}$时的 Hoeffding 不等式，在 Ω 上，我们得到

$$\mathbb{P}(\Omega_1^c|\boldsymbol{X}) \leq p\exp\left(-\frac{A\lambda_d^2}{50p^2\mu^8\lambda_1^2}\right).$$

我们通过注意到以下事实来完成证明

$$\mathbb{P}(\Omega \cap \Omega_1) \geq 1 - cp^{-3} - p\exp\left(-\frac{A\lambda_d^2}{50p^2\mu^8\lambda_1^2}\right).$$

□

9.2 定理的证明2

回顾 $\mathcal{S}_k := \{S \subset [p] : |S| = k\}$ 和 $\mathcal{A} = \{j \in [p] : \boldsymbol{\beta}^{(j,\cdot)} \neq \boldsymbol{0}\}$。对于任意 $S \in \mathcal{S}_k$，注意到 $\boldsymbol{M}^{(S,S)} = \boldsymbol{\gamma}^{(S,\cdot)}\boldsymbol{\Lambda}(\boldsymbol{\gamma}^{(S,\cdot)})^\top$。因此，

$$\sum_{i=1}^{d} \lambda_i(\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}^{(S,S)}, \boldsymbol{\Sigma}^{(S,S)}) = \sum_{i=1}^{d} \lambda_i(\boldsymbol{M}^{(S,S)}) \tag{9.13}$$

$$= \sum_{i=1}^{k} \lambda_i(\boldsymbol{M}^{(S,S)}) - \sum_{i=d+1}^{k} \lambda_i(\boldsymbol{M}^{(S,S)}) = \mathrm{tr}(\boldsymbol{M}^{(S,S)}) = \sum_{i=1}^{d} \sum_{j\in S\cap\mathcal{A}} \lambda_i(\gamma^{(j,i)})^2, \tag{9.14}$$

其中，公式（9.13）由以下引理得到。

引理 2. 对于任意 $S \in \mathcal{S}_k$，设 $\boldsymbol{\Sigma}^{(S,S)} = \boldsymbol{L}_{(S,S)}\boldsymbol{L}_{(S,S)}^\top$ 和 $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$ 分别为 $\boldsymbol{\Sigma}^{(S,S)}$ 和 $\boldsymbol{\Sigma}$ 的 *Cholesky* 分解。则有 $\boldsymbol{L}_{(S,S)} = \boldsymbol{L}^{(S,S)}$ 并且

$$\boldsymbol{L}_{(S,S)}^{-1}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}^{(S,S)}(\boldsymbol{L}_{(S,S)}^{-1})^\top = \{\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}(\boldsymbol{L}^{-1})^\top\}^{(S,S)} = \boldsymbol{M}^{(S,S)}$$

对于随机向量 $\boldsymbol{X}$ 的某种排列。

然后，通过与（3.4）类似的论证，我们得到

$$\lambda_i(\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}^{(S,S)}, \boldsymbol{\Sigma}^{(S,S)}) = \lambda_i(\boldsymbol{L}_{(S,S)}^{-1}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}^{(S,S)}(\boldsymbol{L}_{(S,S)}^{-1})^\top) = \lambda_i(\boldsymbol{M}^{(S,S)}),$$

这给出（9.13）。

根据假设 (A3)，我们知道存在一个概率至少为 $1 - c'p^{-c_2}$ 的事件 Ω，使得在 Ω 上有

$$\sup_{S\in\mathcal{S}_k} \max\{\|\widehat{\boldsymbol{L}}^{(S,S)} - \boldsymbol{L}^{(S,S)}\|_{\mathrm{op}}, \|\widehat{\boldsymbol{M}}^{(S,S)} - \boldsymbol{M}^{(S,S)}\|_{\mathrm{op}}\} \le K\sqrt{\frac{k\log p}{n}},$$

and

$$\sup_{S\in\mathcal{S}_l} \max\{\|\widehat{\boldsymbol{L}}^{(S,S)} - \boldsymbol{L}^{(S,S)}\|_{\mathrm{op}}, \|\widehat{\boldsymbol{M}}^{(S,S)} - \boldsymbol{M}^{(S,S)}\|_{\mathrm{op}}\} \le K\sqrt{\frac{l\log p}{n}},$$

其中，$c' > 0$ 是一个与 S 无关的常数。在上述定义的事件 Ω 上，根据 Weyl 不等式，我们有：

$$\begin{aligned}\left|\sum_{i=1}^d \lambda_i(\widehat{\boldsymbol{M}}^{(S,S)}) - \sum_{i=1}^d \sum_{j\in S\cap\mathcal{A}} \lambda_i(\gamma^{(j,i)})^2\right| &\overset{(9.14)}{=} \left|\sum_{i=1}^d \{\lambda_i(\widehat{\boldsymbol{M}}^{(S,S)}) - \lambda_i(\boldsymbol{M}^{(S,S)})\}\right| \\ &\le Kd\sqrt{\frac{k\log p}{n}} \overset{(3.7)}{\le} \frac{d\lambda_d}{16s\mu^2}.\end{aligned}$$

因此，在事件 Ω 上，若 $(S\cap\mathcal{A}) \subsetneq (S'\cap\mathcal{A})$，则

$$\sum_{i=1}^d \lambda_i(\widehat{\boldsymbol{M}}^{(S,S)}) - \sum_{i=1}^d \lambda_i(\widehat{\boldsymbol{M}}^{(S',S')})$$

$$\leq \sum_{i=1}^{d} \sum_{j \in S \cap \mathcal{A}} \lambda_i (\gamma^{(j,i)})^2 - \sum_{i=1}^{d} \sum_{j \in S' \cap \mathcal{A}} \lambda_i (\gamma^{(j,i)})^2 + \frac{d\lambda_d}{8s\mu^2}$$
$$\leq -\sum_{i=1}^{d} \sum_{j \in (S' \setminus S) \cap \mathcal{A}} \lambda_i (\gamma^{(j,i)})^2 + \frac{d\lambda_d}{8s\mu^2} \,. \tag{9.15}$$

注意，表达式（3.5）意味着

$$\sum_{i=1}^{d} \lambda_i (\gamma^{(j,i)})^2 \geq \lambda_d \|\boldsymbol{\gamma}^{(j,\cdot)}\|_2^2 \geq \frac{d\lambda_d}{s\mu^2} \tag{9.16}$$

对于任意 $j \in \mathcal{A}$。然后，结合（9.15）和（9.16），我们得到

$$\sum_{i=1}^{d} \lambda_i (\widehat{\boldsymbol{M}}^{(S,S)}) < \sum_{i=1}^{d} \lambda_i (\widehat{\boldsymbol{M}}^{(S',S')}) \,. \tag{9.17}$$

给定上述表达式，我们现在证明：在事件 Ω 上，对于某些固定的 $j \in \mathcal{A}$ 和 $j' \notin \mathcal{A}$，有以下结论。

$$q_j \geq q_{j'} \,, \tag{9.18}$$

其中，对于任意 $k \in [p]$，以及某个固定的 $a \in [A]$，有 $q_k = \mathbb{P}(k \in S_{a,b^*(a)} | \boldsymbol{X})$。我们开始证明。~对于 $j \in \{j, j'\}$ 和 $b \in [B]$，定义如下集合：

$$\mathcal{S}_{b,\tilde{j}} := \{(S_{a,1}, \ldots, S_{a,B}) : b^*(a) = b, \tilde{j} \in S_{a,b}\}, \quad \text{and}$$
$$\mathcal{S}_b := \{(S_{a,1}, \ldots, S_{a,B}) : b^*(a) = b\} \,.$$

令映射 $\psi : \mathcal{S}_k \to \mathcal{S}_k$ 定义为，使得

$$\psi(S) := \begin{cases} (S \setminus \{j'\}) \cup \{j\} & \text{if } j' \in S \text{ and } j \notin S \,, \\ S & \text{otherwise} \,. \end{cases}$$

那么，对每个 $S \in \mathcal{S}_k$，要么 $\psi(S) = S$，要么 $S \cap \mathcal{A} \subsetneq \psi(S) \cap \mathcal{A}$。因此，不等式 (9.17)

意味着

$$\sum_{i=1}^{d}\lambda_i(\widehat{\boldsymbol{M}}^{(S,S)}) \le \sum_{i=1}^{d}\lambda_i(\widehat{\boldsymbol{M}}^{(\psi(S),\psi(S))})\,. \tag{9.19}$$

此外，根据 $\psi(\cdot)$ 的定义，我们知道，如果 $j' \in S$ 对某个 $S \in \mathcal{S}_k$ 成立，那么 $j \in \psi(S)$。因此，对于任意 $(S_{a,1},\ldots,S_{a,b},\ldots,S_{a,B}) \in \mathcal{S}_{b,j'}$，存在 $(S_{a,1},\ldots,\psi(S_{a,b})\ldots,S_{a,B}) \in \mathcal{S}_{b,j}$。为清楚起见，由于 $(S_{a,1},\ldots,S_{a,b},\ldots,S_{a,B}) \in \mathcal{S}_{b,j'}$，因此 $b^*(a) = b$ 且 $j' \in S_{a,b}$。注意，$b^*(a) = b$ 意味着

$$\begin{aligned}
&\max_{b\in[B]}\sum_{i=1}^{d}\lambda_i(\widehat{\boldsymbol{\Sigma}}^{(S_{a,b},S_{a,b})}_{\mathbb{E}(\boldsymbol{X}|Y)},\widehat{\boldsymbol{\Sigma}}^{(S_{a,b},S_{a,b})}) \overset{\text{Lemma 2}}{=} \max_{b\in[B]}\sum_{i=1}^{d}\lambda_i(\widehat{\boldsymbol{M}}^{(S_{a,b},S_{a,b})})\\
\le &\sum_{i=1}^{d}\lambda_i(\widehat{\boldsymbol{\Sigma}}^{(S_{a,b},S_{a,b})}_{\mathbb{E}(\boldsymbol{X}|Y)},\widehat{\boldsymbol{\Sigma}}^{(S_{a,b},S_{a,b})}) = \sum_{i=1}^{d}\lambda_i(\widehat{\boldsymbol{M}}^{(S_{a,b},S_{a,b})})\,.
\end{aligned}$$

结合（9.19），我们知道

$$\max_{b\in[B]}\sum_{i=1}^{d}\lambda_i(\widehat{\boldsymbol{M}}^{(S_{a,b},S_{a,b})}) \le \sum_{i=1}^{d}\lambda_i(\widehat{\boldsymbol{M}}^{(\psi(S_{a,b}),\psi(S_{a,b}))})\,,$$

这意味着 $(S_{a,1},\ldots,\psi(S_{a,b})\ldots,S_{a,B})$ 满足 $b^*(a) = b$。结合 $j \in \psi(S_{a,b})$ 这一事实，我们得到 $(S_{a,1},\ldots,\psi(S_{a,b})\ldots,S_{a,B}) \in \mathcal{S}_{b,j}$。因此，$|\mathcal{S}_{b,j'}| \le |\mathcal{S}_{b,j}|$。于是，在 Ω 上，对所有 $b \in [B]$，都有

$$\begin{aligned}
&\mathbb{P}\{j \in S_{a,b^*(a)}|\boldsymbol{X}, b^*(a) = b\}\\
=\ &\frac{\mathbb{P}\{j \in S_{a,b^*(a)}, b^*(a) = b|\boldsymbol{X}\}}{\mathbb{P}\{b^*(a) = b|\boldsymbol{X}\}} = \frac{|\mathcal{S}_{b,j}|}{|\mathcal{S}_b|}\\
\ge\ &\frac{|\mathcal{S}_{b,j'}|}{|\mathcal{S}_b|} = \frac{\mathbb{P}\{j' \in S_{a,b^*(a)}, b^*(a) = b|\boldsymbol{X}\}}{\mathbb{P}\{b^*(a) = b|\boldsymbol{X}\}}\\
=\ &\mathbb{P}\{j' \in S_{a,b^*(a)}|\boldsymbol{X}, b^*(a) = b\}\,.
\end{aligned}$$

因此，我们已证明 $q_j \ge q_{j'}$ 如（9.18）所示。

注意到

$$\sum_{\tilde{j}\in[p]} q_{\tilde{j}} = \sum_{\tilde{j}\in[p]} \mathbb{P}\{\tilde{j}\in S_{a,b^*(a)}|\boldsymbol{X}\} = \frac{1}{B}\sum_{b\in[B]}\sum_{\tilde{j}\in[p]} \mathbb{P}\{\tilde{j}\in S_{a,b^*(a)}|b^*(a)=b,\boldsymbol{X}\} = \sum_{\tilde{j}\in[p]}\frac{k}{p} = k\,.$$

因此，由（9.18）可得，在Ω 上有

$$q_j \geq \frac{\sum_{\tilde{j}\in([p]\setminus\mathcal{A}\cup\{j\})} q_{\tilde{j}}}{p-s+1} = \frac{k-\sum_{\tilde{j}\in\mathcal{A}\setminus\{j\}} q_{\tilde{j}}}{p-s+1} \geq \frac{k-s+1}{p-s+1} \geq \frac{1}{p}\,. \tag{9.20}$$

定义 $\lambda_{a,b;i} := \lambda_i(\boldsymbol{\Sigma}^{(S_{a,b},S_{a,b})}\,\mathbb{E}(\boldsymbol{X}|Y)\,\boldsymbol{\Sigma}^{(S_{a,b},S_{a,b})})$ 和 $\hat{}\lambda_{a,b;i} := \lambda_i(\widehat{\boldsymbol{\Sigma}}^{(S_{a,b},S_{a,b})}\,\mathbb{E}(\boldsymbol{X}|Y)\,\widehat{\boldsymbol{\Sigma}}^{(S_{a,b},S_{a,b})})$，以及 $\boldsymbol{v}_{a,b;i} := \boldsymbol{v}_i(\boldsymbol{\Sigma}^{(S_{a,b},S_{a,b})}\,\mathbb{E}(\boldsymbol{X}|Y)\,\boldsymbol{\Sigma}^{(S_{a,b},S_{a,b})})$ 和 $\hat{}\boldsymbol{v}_{a,b;i} := \boldsymbol{v}_i(\widehat{\boldsymbol{\Sigma}}^{(S_{a,b},S_{a,b})}\,\mathbb{E}(\boldsymbol{X}|Y)\,\widehat{\boldsymbol{\Sigma}}^{(S_{a,b},S_{a,b})})$，分别对应 $b\in[B]$ 和 $i\in S_{a,b}$。注意，矩阵对 $(\boldsymbol{\Sigma}^{(S_{a,b},S_{a,b})}\,\mathbb{E}(\boldsymbol{X}|Y)\,\boldsymbol{\Sigma}^{(S_{a,b},S_{a,b})})$ 有 d 个非零广义特征值。于是$\backsim\lambda_{a,b;d+1} = \ldots = \lambda_{a,b;k} = 0$。记 $\boldsymbol{V}_{a,b} := (\boldsymbol{v}_{a,b;1},\ldots,\boldsymbol{v}_{a,b;d})$，$\boldsymbol{V}_{a,b} := (\hat{}\boldsymbol{v}_{a,b;1},\ldots,\hat{}\boldsymbol{v}_{a,b;d})$，$\hat{}\hat{}\hat{}\hat{}\boldsymbol{\Lambda}_{a,b} = \mathrm{diag}(\lambda_{a,b;1},\ldots,\lambda_{a,b;d})$ 和 $\boldsymbol{\Lambda}_{a,b} = \mathrm{diag}(\lambda_{a,b;1}-\lambda_{a,b;d+1},\ldots,\lambda_{a,b;d}-\lambda_{a,b;d+1})$。对于$\backsim j\in S_{a,b^*(a)}$，令

$$\hat{w}_a^{(\tilde{j})} := (\widehat{\boldsymbol{V}}_{a,b^*(a)}\widehat{\boldsymbol{\Lambda}}_{a,b^*(a)}\widehat{\boldsymbol{V}}_{a,b^*(a)}^\top)^{(\tilde{j},\tilde{j})} = \sum_{i=1}^{d}(\hat{\lambda}_{a,b^*(a);i}-\hat{\lambda}_{a,b^*(a);d+1})(\hat{v}_{a,b^*(a);i}^{(\tilde{j})})^2.$$

现在我们给出 $\hat{}w_a^{(\tilde{j})}$ 的上界和下界。

注意，

$$\begin{aligned}
&\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top \\
=\;& \sum_{i=1}^{d}\lambda_{a,b^*(a);i}\boldsymbol{v}_{a,b^*(a);i}\boldsymbol{v}_{a,b^*(a);i}^\top \overset{(a)}{=} \sum_{i=1}^{k}\lambda_{a,b^*(a);i}\boldsymbol{v}_{a,b^*(a);i}\boldsymbol{v}_{a,b^*(a);i}^\top \\
\overset{(b)}{=}\;& \boldsymbol{L}^{-1,\top}_{(S_{a,b^*(a)},S_{a,b^*(a)})}\left(\sum_{i=1}^{k}\lambda_{a,b^*(a);i}\boldsymbol{\gamma}_{a,b^*(a);i}\boldsymbol{\gamma}_{a,b^*(a);i}^\top\right)\boldsymbol{L}^{-1}_{(S_{a,b^*(a)},S_{a,b^*(a)})} \\
\overset{(c)}{=}\;& \boldsymbol{L}^{-1,\top}_{(S_{a,b^*(a)},S_{a,b^*(a)})}(\boldsymbol{L}^{-1}_{(S_{a,b^*(a)},S_{a,b^*(a)})}\boldsymbol{\Sigma}^{(S_{a,b^*(a)},S_{a,b^*(a)})}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{L}^{-1,\top}_{(S_{a,b^*(a)},S_{a,b^*(a)})})\boldsymbol{L}^{-1}_{(S_{a,b^*(a)},S_{a,b^*(a)})} \\
\overset{(d)}{=}\;& (\boldsymbol{L}^{(S_{a,b^*(a)},S_{a,b^*(a)})})^{-1,\top}\boldsymbol{M}^{(S_{a,b^*(a)},S_{a,b^*(a)})}(\boldsymbol{L}^{(S_{a,b^*(a)},S_{a,b^*(a)})})^{-1} \\
\overset{(e)}{=}\;& (\boldsymbol{L}^{(S_{a,b^*(a)},S_{a,b^*(a)})})^{-1,\top}(\boldsymbol{\gamma}^{(S_{a,b^*(a)},\cdot)}\boldsymbol{\Lambda}\boldsymbol{\gamma}^{(S_{a,b^*(a)},\cdot),\top})(\boldsymbol{L}^{(S_{a,b^*(a)},S_{a,b^*(a)})})^{-1}\,,
\end{aligned} \tag{9.21}$$

其中 $\boldsymbol{\Sigma}^{(S_{a,b^*(a)},S_{a,b^*(a)})} = \boldsymbol{L}_{(S_{a,b^*(a)},S_{a,b^*(a)})}\boldsymbol{L}^\top_{(S_{a,b^*(a)},S_{a,b^*(a)})}$ 是.....的 Cholesky 分解

$\boldsymbol{\Sigma}^{(S_{a,b^*(a)},S_{a,b^*(a)})}$ 和 $\boldsymbol{\gamma}_{a,b^*(a);i} = \boldsymbol{v}_i(\boldsymbol{L}^{-1\ (S_{a,b^*(a)},S_{a,b^*(a)})}\boldsymbol{\Sigma}^{(S_{a,b^*(a)},S_{a,b^*(a)})}\mathbb{E}(\boldsymbol{x}|Y)\ \boldsymbol{L}^{-1,\top\ (S_{a,b^*(a)},S_{a,b^*(a)})})$. 等式（$a$）可由以下事实推出：$\lambda_{a,b^*(a);d+1} = \ldots = \lambda_{a,b^*(a);k} = 0$；(b) 通过类似于（3.4）的变换得到，(c) 由谱分解给出，(d) 用到了引理 2，而 (e) 可由以下事实推出：$\boldsymbol{M} = \boldsymbol{\gamma}\boldsymbol{\Lambda}\boldsymbol{\gamma}^\top$。

令 $S_0 = S_{a,\ b^*\ (a)} \cap \mathcal{A}$ 以及 $S_0^c = S_{a,\ b^*\ (a)} \cap \mathcal{A}^c$。不失一般性，我们记作

$$\boldsymbol{L}^{(S_{a,b^*(a)},S_{a,b^*(a)})} = \begin{pmatrix} \boldsymbol{L}^{(S_0,S_0)} & \boldsymbol{0} \\ \boldsymbol{L}^{(S_0^c,S_0)} & \boldsymbol{L}^{(S_0^c,S_0^c)} \end{pmatrix}.$$

接着，回顾变换后的基 $\boldsymbol{\gamma}$ 保留了 $\boldsymbol{\beta}$的行稀疏性，由式（9.21）可得

$$\begin{aligned}
\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top &= \begin{pmatrix} \boldsymbol{L}^{(S_0,S_0),-1,\top} & * \\ \boldsymbol{0}^\top & * \end{pmatrix}\begin{pmatrix} \boldsymbol{\gamma}^{(S_0,\cdot)}\boldsymbol{\Lambda}\boldsymbol{\gamma}^{(S_0,\cdot),\top} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{L}^{(S_0,S_0),-1} & \boldsymbol{0} \\ * & * \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{L}^{(S_0,S_0),-1,\top}(\boldsymbol{\gamma}^{(S_0,\cdot)}\boldsymbol{\Lambda}\boldsymbol{\gamma}^{(S_0,\cdot),\top}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{L}^{(S_0,S_0),-1} & \boldsymbol{0} \\ * & * \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{L}^{(S_0,S_0),-1,\top}(\boldsymbol{\gamma}^{(S_0,\cdot)}\boldsymbol{\Lambda}\boldsymbol{\gamma}^{(S_0,\cdot),\top})\boldsymbol{L}^{(S_0,S_0),-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},
\end{aligned}$$

这意味着，在对 $\boldsymbol{X}$ 进行某种排序时，$(\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(\tilde{j},\tilde{j})} = (L^{(\tilde{j},\tilde{j})})^{-2}M^{(\tilde{j},\tilde{j})}$ for $\tilde{j} \in S_0$；否则，$(\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a}^\top{}_{,b^*(a)})^{(\tilde{j},\tilde{j})} = 0$。注意到 $K\sqrt{k\log(p)/n} \le \min\{\lambda_1/(4d)\sqrt{\theta_1}, \theta_p/(6\sqrt{\theta_1})\}$，在 Ω 上有

$$\begin{aligned}
&|(\widehat{\boldsymbol{V}}_{a,b^*(a)}\widehat{\boldsymbol{\Lambda}}_{a,b^*(a)}\widehat{\boldsymbol{V}}_{a,b^*(a)}^\top)^{(\tilde{j},\tilde{j})} - (\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(\tilde{j},\tilde{j})}| \\
\overset{(1)}{\le}\ & \|\widehat{\boldsymbol{V}}_{a,b^*(a)}\widehat{\boldsymbol{\Lambda}}_{a,b^*(a)}\widehat{\boldsymbol{V}}_{a,b^*(a)}^\top - \boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top\|_{\mathrm{op}} \\
=\ & \Bigg\|(\widehat{\boldsymbol{L}}^{(S_{a,b^*(a)},S_{a,b^*(a)})})^{-1,\top}\Bigg\{\sum_{i=1}^d(\hat{\lambda}_{a,b^*(a);i} - \hat{\lambda}_{a,b^*(a);d+1})\widehat{\boldsymbol{\gamma}}_{a,b^*(a);i}\widehat{\boldsymbol{\gamma}}_{a,b^*(a);i}^\top\Bigg\}(\widehat{\boldsymbol{L}}^{(S_{a,b^*(a)},S_{a,b^*(a)})})^{-1} - \\
& (\boldsymbol{L}^{(S_{a,b^*(a)},S_{a,b^*(a)})})^{-1,\top}\Bigg\{\sum_{i=1}^d(\lambda_{a,b^*(a);i} - \lambda_{a,b^*(a);d+1})\boldsymbol{\gamma}_{a,b^*(a);i}\boldsymbol{\gamma}_{a,b^*(a);i}^\top\Bigg\}(\boldsymbol{L}^{(S_{a,b^*(a)},S_{a,b^*(a)})})^{-1}\Bigg\|_{\mathrm{op}} \\
\overset{(2)}{\le}\ & 4^{-1}Cd\sqrt{\frac{k\log(p)}{n}} \overset{(3.7)}{\le} \frac{d\tau\lambda_d}{4s\mu^2},
\end{aligned} \tag{9.22}$$

其中 $\widehat{\gamma}_{a,b^*(a);i} = \boldsymbol{v}_i((\widehat{\boldsymbol{L}}^{(S_{a,b^*(a)},S_{a,b^*(a)})})^{-1}\widehat{\boldsymbol{\Sigma}}^{(S_{a,b^*(a)},S_{a,b^*(a)})}\mathbb{E}(\boldsymbol{X}|Y)(\widehat{\boldsymbol{L}}^{(S_{a,b^*(a)},S_{a,b^*(a)})})^{-1,\top})$，$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top$ 是在对 $\boldsymbol{X}$ 进行某种排序时 $\boldsymbol{\Sigma}$ 的 Cholesky 分解。不等式 (1) 由矩阵算子范数的定义得到，而 (2) 则由假设 (A1)、(A1')、(A3)、三角不等式以及 Gataric 等人（2020）的引理 2 推出。因此，在 $\Omega \cap \{j \in S_{a,b^*(a)}\}$ 上，根据假设 (A4)，我们有

$$\begin{aligned}\hat{w}_a^{(j)} &= (\widehat{\boldsymbol{V}}_{a,b^*(a)}\widehat{\boldsymbol{\Lambda}}_{a,b^*(a)}\widehat{\boldsymbol{V}}_{a,b^*(a)}^\top)^{(j,j)} - (\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(j,j)} + (\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(j,j)} \\ &\geq (L^{(j,j)})^{-2}M^{(j,j)} - |(\widehat{\boldsymbol{V}}_{a,b^*(a)}\widehat{\boldsymbol{\Lambda}}_{a,b^*(a)}\widehat{\boldsymbol{V}}_{a,b^*(a)}^\top)^{(j,j)} - (\boldsymbol{V}_{a,b^*(a)}\boldsymbol{\Lambda}_{a,b^*(a)}\boldsymbol{V}_{a,b^*(a)}^\top)^{(j,j)}| \\ &\overset{(9.22)}{\geq} (L^{(j,j)})^{-2}M^{(j,j)} - \frac{d\tau\lambda_d}{4s\mu^2} \geq \tau\lambda_d\|\boldsymbol{\gamma}^{(j,\cdot)}\|_2^2 - \frac{d\tau\lambda_d}{4s\mu^2} \overset{(3.5)}{\geq} \frac{3d\tau\lambda_d}{4s\mu^2}. \end{aligned} \tag{9.23}$$

类似地，在 $\Omega \cap \{j \in S_{a,b^*(a)}\}$ 上，我们有

$$\hat{w}_a^{(j)} \leq \frac{5d\lambda_1\mu^2}{4\tau s}. \tag{9.24}$$

此外，在 $\Omega \cap \{j' \in S_{a,b^*(a)}\}$ 上，有

$$-\frac{d\tau\lambda_d}{4s\mu^2} \leq \hat{w}_a^{(j')} \leq \frac{d\tau\lambda_d}{4s\mu^2}. \tag{9.25}$$

最后，对于所有 $j \in [p]$，如果 $j \notin S_{a,b^*(a)}$，那么根据 $\hat{\boldsymbol{v}}_{a,b^*(a);i}$ 中算法1的定义，我们知道 $(\hat{v}_{a,b^*(a);i})^{(j)} = 0$。因此，根据 $\hat{w}_a^{(j)}$ 的定义，对于 $j \notin S_{a,b^*(a)}$，有 $\hat{w}_a^{(j)} = 0$。因此，我们给出了 $\hat{w}_a^{(j)}$ 的上界和下界。利用上述给出的下界和上界，我们在 Ω 上得到

$$\begin{aligned}&\mathbb{E}(\hat{w}_a^{(j)} - \hat{w}_a^{(j')}|\boldsymbol{X}) \\ &= \mathbb{E}[\hat{w}_a^{(j)}(\mathbf{1}\{j \in S_{a,b^*(a)}\} + \mathbf{1}\{j \notin S_{a,b^*(a)}\})|\boldsymbol{X}] - \mathbb{E}[\hat{w}_a^{(j')}(\mathbf{1}\{j' \in S_{a,b^*(a)}\} + \mathbf{1}\{j' \notin S_{a,b^*(a)}\})|\boldsymbol{X}] \\ &= \mathbb{E}(\hat{w}_a^{(j)}\mathbf{1}\{j \in S_{a,b^*(a)}\} - \hat{w}_a^{(j')}\mathbf{1}\{j' \in S_{a,b^*(a)}\}|\boldsymbol{X}) \\ &\geq \frac{3q_jd\tau\lambda_d}{4s\mu^2} - \frac{q_{j'}d\tau\lambda_d}{4s\mu^2} \overset{(9.18)}{\geq} \frac{q_jd\tau\lambda_d}{2s\mu^2} \overset{(9.20)}{\geq} \frac{d\tau\lambda_d}{2ps\mu^2}. \end{aligned} \tag{9.26}$$

现在我们再次令 a、j 和 j' 任意变化。定义 $\Omega_1 := \{\min_{j\in\mathcal{A}} \hat{w}^{(j)} > \max_{j\in\mathcal{A}^c} \hat{w}^{(j)}\}$。令 $\boldsymbol{\gamma}^{(\hat{S},\cdot)} = \boldsymbol{V}_d(\boldsymbol{M}^{(\hat{S},\hat{S})})$ 和 $\widehat{\boldsymbol{\gamma}}^{(\hat{S},\cdot)} = \widehat{\boldsymbol{V}}_d(\widehat{\boldsymbol{M}}^{(\hat{S},\hat{S})})$ 分别表示 d $\boldsymbol{M}^{(\hat{S},\hat{S})}$ 的主特征向量

和 $\widehat{\boldsymbol{M}}^{(\hat{S},\hat{S})}$。然后，根据假设（A3）以及 Yu 等人 （2015）的定理 2，在$\Omega \cap \Omega_1$,

$$L(\widehat{\boldsymbol{\gamma}}^{(\hat{S},\cdot)}, \boldsymbol{\gamma}^{(\hat{S},\cdot)}) \leq \frac{2d^{1/2}\|\widehat{\boldsymbol{M}}^{(\hat{S},\hat{S})} - \boldsymbol{M}^{(\hat{S},\hat{S})}\|_{\mathrm{op}}}{\lambda_d} \leq 2K\sqrt{\frac{dl\log(p)}{n\lambda_d^2}}.$$

由于 $l \geq s$，在 Ω_1，上有 $\mathcal{A} \subseteq \hat{S}$，这意味着 $\boldsymbol{\gamma} = (\boldsymbol{\gamma}^{(\hat{S},\cdot),\top}, \mathbf{0}^\top)^\top$。令 $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}^{(\hat{S},\cdot)\top}, \mathbf{0}^\top)^\top$。注意，对于 $\boldsymbol{X}$ 的某种排列，真实参数 $\boldsymbol{\beta} = (\boldsymbol{\gamma}^{(\hat{S},\cdot)\top}\boldsymbol{L}^{(\hat{S},\hat{S})-1}, \mathbf{0}^\top)$ 和 $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\gamma}}^{(\hat{S},\cdot)\top}\widehat{\boldsymbol{L}}^{(\hat{S},\hat{S})-1}, \mathbf{0}^\top)$。然后，根据假设 (A1)、(A1') 和 (A3)，注意到 $K\sqrt{l\log(p)/n} \leq \min\{\lambda_d/(2\sqrt{2}\sqrt{\theta_1}), \theta_p/(6\sqrt{\theta_1})\}$，在 Ω 上有

$$L(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}) = L(\boldsymbol{\beta}^{(\hat{S},\cdot)}, \widehat{\boldsymbol{\beta}}^{(\hat{S},\cdot)}) = \|\boldsymbol{\beta}^{(\hat{S},\cdot)}\boldsymbol{\beta}^{(\hat{S},\cdot),\top} - \widehat{\boldsymbol{\beta}}^{(\hat{S},\cdot)}\widehat{\boldsymbol{\beta}}^{(\hat{S},\cdot),\top}\|_{\mathrm{F}} \leq C\sqrt{\frac{dl\log(p)}{n\lambda_d^2}}$$

对于某个足够大的常数 $C > 0$。

只需对 $\mathbb{P}(\Omega \cap \Omega_1)$ 的下界求导。注意，对于任意 $j \in [p]$，$\hat{w}^{(j)} = A^{-1}\sum_{a=1}^{A}\hat{w}_a^{(j)}$。于是，对于任意 $j \in \mathcal{A}$ 和 $j' \in \mathcal{A}^c$，在 Ω 上，有在线a

$$\begin{aligned}
&\hat{w}^{(j)} - \hat{w}^{(j')} \\
&= \{\hat{w}^{(j)} - \mathbb{E}(\hat{w}^{(j)}|\boldsymbol{X})\} - \{\hat{w}^{(j')} - \mathbb{E}(\hat{w}^{(j')}|\boldsymbol{X})\} + \mathbb{E}(\hat{w}^{(j)} - \hat{w}^{(j')}|\boldsymbol{X}) \\
&\overset{(9.26)}{\geq} \frac{d\tau\lambda_d}{2ps\mu^2} + \{\hat{w}^{(j)} - \mathbb{E}(\hat{w}^{(j)}|\boldsymbol{X})\} - \{\hat{w}^{(j')} - \mathbb{E}(\hat{w}^{(j')}|\boldsymbol{X})\},
\end{aligned}$$

这意味着

$$\Omega_1^c \subseteq \left[\bigcup_{j\in\mathcal{A}}\left\{\hat{w}^{(j)} - \mathbb{E}(\hat{w}^{(j)}|\boldsymbol{X}) \leq \frac{d\tau\lambda_d}{4ps\mu^2}\right\}\right] \cup \left[\bigcup_{j'\notin\mathcal{A}}\left\{\hat{w}^{(j')} - \mathbb{E}(\hat{w}^{(j')}|\boldsymbol{X}) \geq \frac{d\tau\lambda_d}{4ps\mu^2}\right\}\right].$$

注意，给定 $\boldsymbol{X}$，$\{\hat{w}_a^{(j)}\}_{a\in[A]}$ 是独立同分布的随机变量，并且（9.23）、（9.24）和（9.25）意味着 $\hat{w}_a^{(j)}$在 Ω 上有界，对于所有 $j \in [p]$。因此，根据并集界以及给定 $\boldsymbol{X}$条件下的霍夫丁不等式，在 Ω 上我们得到

$$\mathbb{P}(\Omega_1^c|\boldsymbol{X}) \leq p\exp\left(-\frac{A\tau^4\lambda_d^2}{50p^2\mu^8\lambda_1^2}\right).$$

我们通过指出以下事实来完成证明。

$$\mathbb{P}(\Omega \cap \Omega_1) \geq 1 - c' p^{-c_2} - p \exp\left(- \frac{A\tau^4 \lambda_d^2}{50 p^2 \mu^8 \lambda_1^2} \right).$$

□

9.3 定理3的证明

考虑一种特定情形，其中 $d = 1$ 和 $H = 2$。在这种情况下，感兴趣的参数$\boldsymbol{\beta} \in \boldsymbol{\Theta}_{p,\ d,\ s}$(3) 化简为一个一维向量 $\boldsymbol{\beta}_1 \in \boldsymbol{\Theta}_{p,\ 1,\ s}$(3)。我们考虑以下结构：

$$\begin{aligned} \boldsymbol{X}|(\tilde{Y} = 1) &\sim \mathcal{N}_p((1-\alpha)\boldsymbol{\beta}_1, \boldsymbol{I}_p - \boldsymbol{M}), \quad \mathbb{P}(\tilde{Y} = 1) = \alpha, \\ \boldsymbol{X}|(\tilde{Y} = 2) &\sim \mathcal{N}_p(-\alpha\boldsymbol{\beta}_1, \boldsymbol{I}_p - \boldsymbol{M}), \quad \mathbb{P}(\tilde{Y} = 2) = 1 - \alpha, \end{aligned} \tag{9.27}$$

其中 $\alpha \in (0,\ 1)$，$\boldsymbol{M}$ 是一个将在后文定义的正定矩阵。在此结构下，我们有以下三个结果：

˜(i) $\mathbb{E}\boldsymbol{X} = \mathbb{E}\{\mathbb{E}(\boldsymbol{X}|Y)\} = 0$;

˜ ˜ ˜(ii) 定义 $\boldsymbol{\mu}_h = \mathbb{E}\{\boldsymbol{X}|(Y = h)\}$ 和 $p_h = \mathbb{P}(Y = h)$，其中 $h \in [H]$。那么 $\mathrm{Cov}\{\mathbb{E}(\boldsymbol{X}|Y)\} = \sum_{h=1}^{H} p_h \boldsymbol{\mu}_h \boldsymbol{\mu}_h^\top - (\sum_{h=1}^{H} p_h \boldsymbol{\mu}_h)(\sum_{h=1}^{H} p_h \boldsymbol{\mu}_h)^\top = \sum_{h=1}^{H} p_h \boldsymbol{\mu}_h \boldsymbol{\mu}_h^\top = \alpha(1-\alpha)\boldsymbol{\beta}_1\boldsymbol{\beta}_1^\top$;

(iii) 因此，$\mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top) = \mathrm{Cov}\{\mathbb{E}(\boldsymbol{X}|\tilde{Y})\} + \mathbb{E}\{\mathrm{Cov}(\boldsymbol{X}|\tilde{Y})\} = \alpha(1-\alpha)\boldsymbol{\beta}_1\boldsymbol{\beta}_1^\top + \boldsymbol{I}_p - \boldsymbol{M}$。

令 $\boldsymbol{M} = \mathrm{Cov}\{\mathbb{E}(\boldsymbol{X}|\tilde{Y})\} = \alpha(1-\alpha)\boldsymbol{\beta}_1\boldsymbol{\beta}_1^\top$，则有 $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I}_p$。因此，$(\mathrm{Cov}\{\mathbb{E}(\boldsymbol{X}|Y)\}, \mathrm{Cov}(\boldsymbol{X}))$ 这一对矩阵的˜非零广义特征值可化为 $\mathrm{Cov}\{\mathbb{E}(\boldsymbol{X}|\tilde{Y})\}$ 的非零特征值。回顾 $\mathrm{Cov}\{\mathbb{E}(\boldsymbol{X}|\tilde{Y})\} = \alpha(1-\alpha)\boldsymbol{\beta}_1\boldsymbol{\beta}_1^\top$，这意味着 ˜$\mathrm{Cov}\{\mathbb{E}(\boldsymbol{X}|Y)\}$ 仅有一个非零特征值 $\lambda_1 = \alpha(1-\alpha)$，而作为我们关注的参数的主特征向量对应于 $\boldsymbol{\beta}_1$。

我们将使用 Fano 引理（Yu,1997）推导下界。应用该引理分为三个步骤：(i) 构造一些 $\boldsymbol{\beta} \in \boldsymbol{\Theta}_{p,1,s}$(3)；(ii) 推导 Kullback-Leiber

感兴趣的数据分布之间的散度，即，构造出的混合高斯分布；(iii) 在 $\boldsymbol{\Theta}_{p,1,s}$(3) 中找到一个具有合适装箱数的子集。

步骤证明(i)。我们考虑 Gataric 等人 (2020年)（命题1）中的一个特定设定，其中感兴趣的参数现在是一个向量，而不是矩阵。定义 $\boldsymbol{u} = (s^{-1/2}\mathbf{1}_s^\top, \mathbf{0}^\top)^\top \in \mathbb{R}^p$，并令 $\epsilon \in \big(0, \sqrt{1/(16s)}\big)$。对于任意 $J \in \boldsymbol{\Theta}_{p-1,1,s-1} := \{\boldsymbol{V} \in \boldsymbol{\Theta}_{p-1,1} : \mathrm{supp}(\boldsymbol{V}) \le s-1\}$，由于 $\boldsymbol{u}^\top\boldsymbol{u} = 1$，我们可以选取 $\boldsymbol{U} \in \boldsymbol{\Theta}_{p,p}$，使得 $\boldsymbol{U}$ 的第一列为 $\boldsymbol{u}$，且 $\boldsymbol{\beta}_J = \widetilde{\boldsymbol{U}}(\sqrt{1-\epsilon^2}, \epsilon J^\top)^\top$ 的非零分量索引是 $[s]$ 的子集。注意

$$\boldsymbol{\beta}_J = \widetilde{\boldsymbol{U}} \begin{pmatrix} \sqrt{1-\epsilon^2} \\ \epsilon J \end{pmatrix} = \boldsymbol{u} + \widetilde{\boldsymbol{U}} \begin{pmatrix} \sqrt{1-\epsilon^2} - 1 \\ \epsilon J \end{pmatrix} =: \boldsymbol{u} + \widetilde{\boldsymbol{U}}\Delta_J \,.$$

由于 $\|\widetilde{\boldsymbol{U}}\Delta_J\|_\infty \le \|\widetilde{\boldsymbol{U}}\Delta_J\|_2 \le \sqrt{2}\epsilon$ 和 $\epsilon \le \sqrt{1/(16s)}$，对于任意 $k \in [s]$，$|\beta_J^{(k)}| \in [0.64/\sqrt{s}, 1.36/\sqrt{s}]$，这意味着 $\boldsymbol{\beta}_J \in \boldsymbol{\Theta}_{p,1,s}$（3）。

步骤(ii)的证明。对于任意 J，$J' \in \boldsymbol{\Theta}_{p-1,1,s-1}$，令 $\boldsymbol{M}_J = \lambda_1 \boldsymbol{\beta}_J\boldsymbol{\beta}_J^\top$ 和 $\boldsymbol{M}_{J'} = \lambda_1 \boldsymbol{\beta}_{J'}\boldsymbol{\beta}_{J'}^\top$，然后记 $\mathbb{P}(\boldsymbol{\beta}_J, \boldsymbol{M}_J)$ $\mathbb{P}(\boldsymbol{\beta}_{J'}, \boldsymbol{M}_{J'})$ 为（9.27）中指定的相应混合高斯分布。记 $D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})$ 表示分布 $\mathbb{P}$ 到 $\mathbb{Q}$的KL散度。由KL散度的凸性可得

$$\begin{aligned}
&D_{\mathrm{KL}}\big(\mathbb{P}(\boldsymbol{\beta}_J, \boldsymbol{M}_J) \,\|\, \mathbb{P}(\boldsymbol{\beta}_{J'}, \boldsymbol{M}_{J'})\big) \\
&\le \alpha D_{\mathrm{KL}}\big(\mathcal{N}_p((1-\alpha)\boldsymbol{\beta}_J, \boldsymbol{I}_p - \boldsymbol{M}_J) \,\|\, \mathcal{N}_p((1-\alpha)\boldsymbol{\beta}_{J'}, \boldsymbol{I}_p - \boldsymbol{M}_{J'})\big) \qquad (9.28) \\
&\quad + (1-\alpha) D_{\mathrm{KL}}\big(\mathcal{N}_p(-\alpha\boldsymbol{\beta}_J, \boldsymbol{I}_p - \boldsymbol{M}_J) \,\|\, \mathcal{N}_p(-\alpha\boldsymbol{\beta}_{J'}, \boldsymbol{I}_p - \boldsymbol{M}_{J'})\big) \,. \qquad (9.29)
\end{aligned}$$

因此，只需界定两个高斯分布之间的 KL散度。对于（9.28）中的第一个 KL散度，有

$$\begin{aligned}
&D_{\mathrm{KL}}\big(\mathcal{N}_p((1-\alpha)\boldsymbol{\beta}_J, \boldsymbol{I}_p - \boldsymbol{M}_J) \,\|\, \mathcal{N}_p((1-\alpha)\boldsymbol{\beta}_{J'}, \boldsymbol{I}_p - \boldsymbol{M}_{J'})\big) \\
&= \frac{1}{2}\Bigg(\underbrace{[\mathrm{Tr}\{(\boldsymbol{I}_p - \boldsymbol{M}_{J'})^{-1}(\boldsymbol{I}_p - \boldsymbol{M}_J)\} - p]}_{T_1} + \underbrace{\log\left\{\frac{\det(\boldsymbol{I}_p - \boldsymbol{M}_{J'})}{\det(\boldsymbol{I}_p - \boldsymbol{M}_J)}\right\}}_{T_2} \\
&\quad + \underbrace{(1-\alpha)^2(\boldsymbol{\beta}_{J'} - \boldsymbol{\beta}_J)^\top(\boldsymbol{I}_p - \boldsymbol{M}_{J'})^{-1}(\boldsymbol{\beta}_{J'} - \boldsymbol{\beta}_J)}_{T_3} \Bigg) \,.
\end{aligned}$$

对于第一项，有

$$\begin{aligned} T_1 &= \mathrm{Tr}\{(\boldsymbol{I}_p - \boldsymbol{M}_{J'})^{-1}(\boldsymbol{M}_{J'} - \boldsymbol{M}_J)\} \\ &= \mathrm{Tr}\left\{\left(\boldsymbol{I}_p + \frac{\lambda_1}{1-\lambda_1}\boldsymbol{\beta}_{J'}\boldsymbol{\beta}_{J'}^\top\right)(\lambda_1\boldsymbol{\beta}_{J'}\boldsymbol{\beta}_{J'}^\top - \lambda_1\boldsymbol{\beta}_J\boldsymbol{\beta}_J^\top)\right\} \\ &= \frac{\lambda_1^2}{1-\lambda_1}\{1 - \mathrm{Tr}(\boldsymbol{\beta}_{J'}\boldsymbol{\beta}_{J'}^\top\boldsymbol{\beta}_J\boldsymbol{\beta}_J^\top)\} \\ &\le \frac{2\lambda_1^2}{1-\lambda_1}\|\boldsymbol{\beta}_J - \boldsymbol{\beta}_{J'}\|_2^2 . \end{aligned}$$

对于项 T_2，注意到 $\boldsymbol{I}_p - \boldsymbol{M}_J$ 和 $\boldsymbol{I}_p - \boldsymbol{M}_{J'}$ 具有相同的特征值，因此可得 $T_2 = 0$。对于项 T_3，有 $T_3 \le (1-\alpha)^2\lambda_{\max}\{(\boldsymbol{I}_p - \boldsymbol{M}_{J'})^{-1}\}\|\boldsymbol{\beta}_J - \boldsymbol{\beta}_{J'}\|_2^2 = (1-\alpha)^2(1-\lambda_1)^{-1}\|\boldsymbol{\beta}_J - \boldsymbol{\beta}_{J'}\|_2^2$。结合三项的上界，我们最终得到

$$D_{\mathrm{KL}}\big(\mathcal{N}_p((1-\alpha)\boldsymbol{\beta}_J, \boldsymbol{I}_p - \boldsymbol{M}_J)\,\|\,\mathcal{N}_p((1-\alpha)\boldsymbol{\beta}_{J'}, \boldsymbol{I}_p - \boldsymbol{M}_{J'})\big) \le \frac{2\lambda_1^2 + (1-\alpha)^2}{2(1-\lambda_1)}\|\boldsymbol{\beta}_J - \boldsymbol{\beta}_{J'}\|_2^2 . \tag{9.30}$$

式（9.29）中的第二个 KL散度具有类似的上界

$$D_{\mathrm{KL}}\big(\mathcal{N}_p(-\alpha\boldsymbol{\beta}_J, \boldsymbol{I}_p - \boldsymbol{M}_J)\,\|\,\mathcal{N}_p(-\alpha\boldsymbol{\beta}_{J'}, \boldsymbol{I}_p - \boldsymbol{M}_{J'})\big) \le \frac{2\lambda_1^2 + \alpha^2}{2(1-\lambda_1)}\|\boldsymbol{\beta}_J - \boldsymbol{\beta}_{J'}\|_2^2 . \tag{9.31}$$

结合式（9.28）-（9.31），有

$$D_{\mathrm{KL}}\big(\mathbb{P}(\boldsymbol{\beta}_J, \boldsymbol{M}_J)\,\|\,\mathbb{P}(\boldsymbol{\beta}_{J'}, \boldsymbol{M}_{J'})\big) \le \frac{2\lambda_1^2 + \lambda_1}{2(1-\lambda_1)}\|\boldsymbol{\beta}_J - \boldsymbol{\beta}_{J'}\|_2^2 = \frac{2\lambda_1^2 + \lambda_1}{1-\lambda_1}(1 - \boldsymbol{\beta}_J^\top\boldsymbol{\beta}_{J'}) .$$

回忆 $\boldsymbol{\beta}_J = \widetilde{\boldsymbol{U}}(\sqrt{1-\epsilon^2},\ \epsilon J^\top)^\top$ 对任意 $J \in \boldsymbol{\Theta}_{p-1,\ 1,\ s-1}$ 均成立。于是 $\boldsymbol{\beta}_J\boldsymbol{\beta}_{J'}^\top = 1 - \epsilon^2 + \epsilon^2 J^\top J' \ge 1 - \epsilon^2 - \epsilon^2\|J\|_2\|J'\|_2 = 1 - 2\epsilon^2$。因此，

$$D_{\mathrm{KL}}\big(\mathbb{P}(\boldsymbol{\beta}_J, \boldsymbol{M}_J)\,\|\,\mathbb{P}(\boldsymbol{\beta}_{J'}, \boldsymbol{M}_{J'})\big) \le \frac{2\epsilon^2(2\lambda_1^2 + \lambda_1)}{1-\lambda_1} . \tag{9.32}$$

步骤的证明 (iii)。令 $\mathcal{J} \subset \boldsymbol{\Theta}_{p-1,1,s-1}$ 满足 $\min_{J,J'\in\mathcal{J}} L(\boldsymbol{\beta}_J, \boldsymbol{\beta}_{J'}) \ge c\epsilon$，其中˜为某个 $c > 0$ 通用常数，且 $3 \le |\mathcal{J}| \le \exp\{n/(4s)\}$，具体取值将在后文给出。令 $\boldsymbol{\beta}$ 为

的任意可能稀疏SIR估计量 $\boldsymbol{\beta}$。然后由（9.32）和 Fano 引理（Yu，1997），我们得到

$$\begin{aligned}\inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Theta}_{p-1,1,s-1}(3)} \mathbb{E}_{P_{\boldsymbol{\beta}}}\{L(\tilde{\boldsymbol{\beta}},\boldsymbol{\beta})\} &\geq \inf_{\tilde{\boldsymbol{\beta}}} \max_{J\in\mathcal{J}} \mathbb{E}_{P_{\boldsymbol{\beta}_J}}\{L(\tilde{\boldsymbol{\beta}},\boldsymbol{\beta}_J)\} \\ &\geq \frac{c\epsilon}{2}\left\{1-\frac{2n\epsilon^2(2\lambda_1^2+\lambda_1)/(1-\lambda_1)+\log 2}{\log|\mathcal{J}|}\right\} \\ &\geq \frac{c\epsilon}{2}\left(\frac{1}{3}-\frac{n\epsilon^2}{\log|\mathcal{J}|}\right), \end{aligned} \tag{9.33}$$

这里我们利用了以下事实：$|\mathcal{J}|\geq 3$ 以及 $\lambda_1=\alpha(1-\alpha)\leq 1/4$。注意到$\log|\mathcal{J}|\leq n/(4s)$，我们可以选取 $\epsilon^2=\log(|\mathcal{J}|)/(4n)\leq 1/(16s)$。结合（9.33），得到

$$\inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Theta}_{p,1,s}(3)} \mathbb{E}_{P_{\boldsymbol{\beta}}}\{L(\tilde{\boldsymbol{\beta}},\boldsymbol{\beta})\} \gtrsim \sqrt{\frac{\log|\mathcal{J}|}{n}}. \tag{9.34}$$

接下来需要构造 $\mathcal{J}$。令 $\mathcal{S}=\{S\in[p-1]:\ |S|=s-1\}$。对于任意 $S\in\mathcal{S}$，定义 $J_S\in\mathbb{R}^{p-1}$ 使得 $J_S^{(S)}=(s-1)^{-1/2}\mathbf{1}_{s-1}$ 并且 $J^{(S^c)}{}_S=\mathbf{0}$。那么$J_S\in\boldsymbol{\Theta}_{p-1,\,1,\,s-1}$。根据 Varshamov-Gilbert 引理（Massart（2007）中的引理 4.10），由于Massart (2007)，$4(s-1)\leq p-1$，存在子集 $\mathcal{T}$属于 $\mathcal{S}$，使得

$$\log|\mathcal{T}|=\left\lfloor\frac{1}{5}(s-1)\log\left(\frac{p-1}{s-1}\right)\right\rfloor \tag{9.35}$$

且对于任意 $T,\ T'\in\mathcal{T}$，都有 $|T\cap T'|\leq(s-1)/2$，其中 $T\neq T'$。令 $\mathcal{J}=\{J_T:\ T\in\mathcal{T}\}$，则$|\mathcal{J}|=|\mathcal{T}|$。此外，事实 $|T\cap T'|\leq(s-1)/2$ 对于任意 $T,\ T'\in\mathcal{T}$，其中 $T\neq T'$，均成立，这意味着

$$\min_{J,J'\in\mathcal{J}:J\neq J'} L(J,J')=\min_{J,J'\in\mathcal{J}:J\neq J'}\sqrt{1-(J^\top J')^2}\geq\sqrt{1-\left(\frac{1}{2}\right)^2}\geq\frac{\sqrt{3}}{2}. \tag{9.36}$$

注意到

$$L(\boldsymbol{\beta}_J,\boldsymbol{\beta}_{J'})=\frac{1}{\sqrt{2}}\|\boldsymbol{\beta}_J\boldsymbol{\beta}_J^\top-\boldsymbol{\beta}_{J'}\boldsymbol{\beta}_{J'}^\top\|_{\mathrm{F}}=\sqrt{\epsilon^4L^2(J,J')+\epsilon^2(1-\epsilon^2)\|J-J'\|_2^2}\geq\epsilon L(J,J')$$

对于任意 $J \in \mathbf{\Theta}_{p-1,\ 1,\ s-1}$。结合式（9.36），可得

$$\min_{J,J' \in \mathcal{J}: J \neq J'} L(\boldsymbol{\beta}_J, \boldsymbol{\beta}_{J'}) \geq \frac{\sqrt{3}}{2}\epsilon\,.$$

最后，由于 $(s-1)\log\{(p-1)/(s-1)\} \geq 6$ 且 $5n \geq 4s(s-1)\log\{(p-1)/(s-1)\}$，我们有 $3 \leq |\mathcal{J}| \leq \exp\{n/(4s)\}$。因此，(9.34) 和 (9.35) 可推出

$$\inf_{\tilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta} \in \mathbf{\Theta}_{p,1,s}(3)} \mathbb{E}_{P_{\boldsymbol{\beta}}}\{L(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta})\} \gtrsim \sqrt{\frac{s\log(p/s)}{n}}\,.$$

证明完毕。 □

9.4 辅助引理的证明

9.4.1 引理 1 的证明

不失一般性，假设 $\mathbb{E}\boldsymbol{X} = \mathbf{0}$。那么，根据定义

$$\begin{aligned}\boldsymbol{\Sigma} &:= \boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)} = \mathrm{Cov}\{\mathbb{E}(\boldsymbol{X}|\tilde{Y})\} = \mathbb{E}[\{\mathbb{E}(\boldsymbol{X}|\tilde{Y})\}\{\mathbb{E}(\boldsymbol{X}|\tilde{Y})\}^\top] \\ &= \sum_{h=1}^{H} \mathbb{P}(\tilde{Y}=h)\{\mathbb{E}(\boldsymbol{X}|\tilde{Y}=h)\}\{\mathbb{E}(\boldsymbol{X}|\tilde{Y}=h)\}^\top.\end{aligned}$$

注意到

$$\begin{aligned}\mathbb{E}[\boldsymbol{X}\mathbf{1}\{\tilde{Y}=h\}] &= \mathbb{E}(\mathbb{E}[\boldsymbol{X}\mathbf{1}\{\tilde{Y}=h\}|\mathbf{1}\{\tilde{Y}=h\}]) \\ &= \mathbb{E}(\mathbf{1}\{\tilde{Y}=h\}\mathbb{E}[\boldsymbol{X}|\mathbf{1}\{\tilde{Y}=h\}]) \\ &= \mathbb{P}[\mathbf{1}\{\tilde{Y}=h\}=1]\cdot\mathbb{E}(\boldsymbol{X}|\mathbf{1}\{\tilde{Y}=h\}=1) \\ &= \mathbb{P}(\tilde{Y}=h)]\cdot\mathbb{E}(\boldsymbol{X}|\tilde{Y}=h) \\ &=: p_h\cdot\mathbb{E}(\boldsymbol{X}|\tilde{Y}=h)\,,\end{aligned}$$

这意味着

$$\boldsymbol{\Sigma} = \sum_{h=1}^{H} \frac{\mathbb{E}[\boldsymbol{X}\mathbf{1}\{\tilde{Y}=h\}](\mathbb{E}[\boldsymbol{X}\mathbf{1}\{\tilde{Y}=h\}])^\top}{p_h} =: \sum_{h=1}^{H}\frac{\boldsymbol{u}_h\boldsymbol{u}_h^\top}{p_h}\,.$$

根据定义，我们还有

$$\widehat{\boldsymbol{\Sigma}} := \widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)} = \sum_{h=1}^{H} \frac{\hat{\boldsymbol{u}}_h \hat{\boldsymbol{u}}_h^\top}{\hat{p}_h},$$

其中 $\hat{}p_h = n^{-1}\sum_{i=1}^n \mathbf{1}\{\tilde{Y}_i = h\}$ 和 $\hat{}\boldsymbol{u}_h = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i \mathbf{1}\{\tilde{Y}_i = h\}$。因此，我们得到

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} &= \sum_{h=1}^{H} \left(\frac{\hat{\boldsymbol{u}}_h \hat{\boldsymbol{u}}_h^\top}{\hat{p}_h} - \frac{\boldsymbol{u}_h \boldsymbol{u}_h^\top}{p_h} \right) \\
&= \sum_{h=1}^{H} \left(\frac{\hat{\boldsymbol{u}}_h \hat{\boldsymbol{u}}_h^\top}{\hat{p}_h} - \frac{\boldsymbol{u}_h \boldsymbol{u}_h^\top}{\hat{p}_h} + \frac{\boldsymbol{u}_h \boldsymbol{u}_h^\top}{\hat{p}_h} - \frac{\boldsymbol{u}_h \boldsymbol{u}_h^\top}{p_h} \right) \\
&= \sum_{h=1}^{H} \{\hat{p}_h^{-1}(\hat{\boldsymbol{u}}_h \hat{\boldsymbol{u}}_h^\top - \boldsymbol{u}_h \boldsymbol{u}_h^\top) + (\hat{p}_h^{-1} - p_h^{-1}) \boldsymbol{u}_h \boldsymbol{u}_h^\top\} \\
&= \sum_{h=1}^{H} \{p_h^{-1}(\hat{\boldsymbol{u}}_h \hat{\boldsymbol{u}}_h^\top - \boldsymbol{u}_h \boldsymbol{u}_h^\top) + (\hat{p}_h^{-1} - p_h^{-1})(\hat{\boldsymbol{u}}_h \hat{\boldsymbol{u}}_h^\top - \boldsymbol{u}_h \boldsymbol{u}_h^\top) + (\hat{p}_h^{-1} - p_h^{-1}) \boldsymbol{u}_h \boldsymbol{u}_h^\top\}.
\end{aligned}$$

对于任意单位向量 $\boldsymbol{\alpha} \in \mathbb{R}^p$，有成酬

$$\begin{aligned}
&\boldsymbol{\alpha}^\top (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \boldsymbol{\alpha} \\
&= \sum_{h=1}^{H} \{p_h^{-1} |\boldsymbol{\alpha}^\top (\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)|^2 + 2 p_h^{-1} \boldsymbol{\alpha}^\top \boldsymbol{u}_h (\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)^\top \boldsymbol{\alpha} + (\hat{p}_h^{-1} - p_h^{-1}) |\boldsymbol{\alpha}^\top (\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)|^2 \\
&\quad + 2(\hat{p}_h^{-1} - p_h^{-1}) \boldsymbol{\alpha}^\top \boldsymbol{u}_h (\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)^\top \boldsymbol{\alpha} + (\hat{p}_h^{-1} - p_h^{-1}) |\boldsymbol{\alpha}^\top \boldsymbol{u}_h|^2\}. \qquad (9.37)
\end{aligned}$$

现在我们声称 对于所有 $h \in [H]$, $\|\boldsymbol{u}_h\|_2$ 都可以被一个常数所界定。根据定义。 n,

$$\boldsymbol{\Sigma} = \sum_{h=1}^{H} \frac{\boldsymbol{u}_h \boldsymbol{u}_h^\top}{p_h} = \boldsymbol{\beta} \boldsymbol{\Lambda} \boldsymbol{\beta}^\top,$$

其中 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)^\top$ 且 $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$。对等式两边取迹，得到 $\sum_{h=1}^H p_h^{-1} \|\boldsymbol{u}_h\|_2^2 = \sum_{i=1}^d \lambda_i \le C$ 根据假设 (A1)，其中常数 $C > 0$ 取决于d, κ 和 λ。注意到 $0 < p_h < 1$, (9.37) 可得

$$|\boldsymbol{\alpha}^\top (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \boldsymbol{\alpha}|$$

$$
\begin{aligned}
&\leq \sum_{h=1}^{H} p_h^{-1} |\boldsymbol{\alpha}^\top (\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)|^2 + 2 \sum_{h=1}^{H} p_h^{-1} |(\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)^\top \boldsymbol{\alpha}| \cdot |\boldsymbol{u}_h|_2 + \sum_{h=1}^{H} |\hat{p}_h^{-1} - p_h^{-1}| \cdot |\boldsymbol{\alpha}^\top (\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)|^2 \\
&\quad + 2 \sum_{h=1}^{H} |\hat{p}_h^{-1} - p_h^{-1}| \cdot |(\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)^\top \boldsymbol{\alpha}| \cdot |\boldsymbol{u}_h|_2 + \sum_{h=1}^{H} |\hat{p}_h^{-1} - p_h^{-1}| \cdot |\boldsymbol{u}_h|_2^2 \\
&\lesssim \sum_{h=1}^{H} |\boldsymbol{\alpha}^\top (\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)|^2 + \sum_{h=1}^{H} |\hat{p}_h^{-1} - p_h^{-1}| \cdot |\boldsymbol{\alpha}^\top (\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)|^2 \\
&\quad + \sum_{h=1}^{H} |(\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)^\top \boldsymbol{\alpha}| + \sum_{h=1}^{H} |\hat{p}_h^{-1} - p_h^{-1}| \cdot |(\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)^\top \boldsymbol{\alpha}| + \sum_{h=1}^{H} |\hat{p}_h^{-1} - p_h^{-1}| \,. \qquad (9.38)
\end{aligned}
$$

考虑项 $\boldsymbol{\alpha}^\top(\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)$。令 $Z_i(h) = \boldsymbol{\alpha}^\top \boldsymbol{X}_i \mathbf{1}\{\tilde{Y}_i = h\} - \mathbb{E}[\boldsymbol{\alpha}^\top \boldsymbol{X}_i \mathbf{1}\{\tilde{Y}_i = h\}]$ 对于任意单位向量 $\boldsymbol{\alpha} \in \mathbb{R}^p$ 和 $h \in [H]$。于是，$\boldsymbol{\alpha}^\top(\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h) = n^{-1} \sum_{i=1}^{n} Z_i(h)$。注意到 $\boldsymbol{X}_i$ 服从混合高斯分布，这意味着 $\{Z_i(h)\}_{1=1}^{n}$ 是独立同分布的次指数随机变量。由于 $\{\boldsymbol{X}_i, Y_i\}$ 是独立同分布的随机变量，且 $Z_i(h)$ 是 $\boldsymbol{X}_i$ 和 Y_i 的函数，因此我们知道 $\{Z_i(h)\}_{1=1}^{n}$ 也是独立同分布的随机变量。根据 Bernstein 不等式（Vershynin，2010），我们有

$$
\mathbb{P}\{|\boldsymbol{\alpha}^\top (\hat{\boldsymbol{u}}_h - \boldsymbol{u}_h)| \geq t\} \leq 2 \exp\{-Cn(t^2 \wedge t)\} \,. \qquad (9.39)
$$

对于项 $\hat{p}_h^{-1} - p_h^{-1}$，容易验证 $\hat{p}_h - p_h = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{\tilde{Y}_i = h\} - \mathbb{E}[\mathbf{1}\{\tilde{Y}_i = h\}]$，$|\mathbf{1}\{Y_i = h\} - \mathbb{E}[\mathbf{1}\{Y_i = h\}]| \leq 1$ 以及 $\mathrm{Var}(\mathbf{1}\{Y_i = h\} - \mathbb{E}[\mathbf{1}\{Y_i = h\}]) \leq 1/4$。根据 Bernstein 不等式（van der Vaart 和 Wellner，1996），我们有 $\mathbb{P}(|\hat{p}_h - p_h| \geq t) \leq 2\exp\{-Cn(t^2 \wedge t)\}$。令 $p_{\min} = \min\{p_1, \ldots, p_H\}$。由于 $|\hat{p}_h^{-1} - p_h^{-1}| = |\hat{p}_h - p_h|/(p_h \hat{p}_h)$，可得

$$
\begin{aligned}
&\mathbb{P}(|\hat{p}_h^{-1} - p_h^{-1}| \geq 2p_{\min}^{-2} t) \\
&\leq \mathbb{P}(|\hat{p}_h - p_h| \geq t) + \mathbb{P}(p_h \hat{p}_h \leq p_{\min}^2/2) \\
&\leq \mathbb{P}(|\hat{p}_h - p_h| \geq t) + \mathbb{P}(|\hat{p}_h - p_h| \geq p_{\min}/2) \\
&\leq 4 \exp\{-Cn(t^2 \wedge t)\} \qquad (9.40)
\end{aligned}
$$

前提是 $p_{\min} \geq 2t$。

结合（9.38）-（9.40）并使用并集界，我们有

$$\mathbb{P}(|\boldsymbol{\alpha}^\top(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\alpha}|\geq t)\lesssim C\exp\{-C'n(t^2\wedge t^{1/4})\}.$$

根据 Wang 等人 (2016b) 的引理 2，若 $\eta>0$ 满足 $k\log(p/\eta)/n\leq 1$，则有

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\boldsymbol{u}\in\mathcal{B}_0^{p-1}(k)}|\boldsymbol{\alpha}^\top(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\alpha}|\geq 2C\sqrt{\frac{k\log(p/\eta)}{n}}\right\}\\
&\lesssim k^{1/2}\binom{p}{k}\left(\frac{128}{\sqrt{255}}\right)^{k-1}\mathbb{P}\left\{|\boldsymbol{\alpha}^\top(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\alpha}|\geq C\sqrt{\frac{k\log(p/\eta)}{n}}\right\}\\
&\lesssim k^{1/2}\binom{p}{k}\left(\frac{128}{\sqrt{255}}\right)^{k-1}\exp\{-C'C^2k\log(p/\eta)\}\\
&\lesssim k^{1/2}\binom{p}{k}\left(\frac{128}{\sqrt{255}}\right)^{k-1}p^{-k}\eta^k\quad(\text{choose } C^2=(C')^{-1})\\
&\lesssim \eta\,.
\end{aligned}$$

同样地，如果 $k\log(p/\eta)/n>1$，则有

$$\mathbb{P}\left[\sup_{\boldsymbol{u}\in\mathcal{B}_0^{p-1}(k)}|\boldsymbol{\alpha}^\top(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\alpha}|\geq 2C\left\{\frac{k\log(p/\eta)}{n}\right\}^4\right]\lesssim\eta\,.$$

结合上述两个方程，我们得到

$$\mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathcal{B}_0^{p-1}(k)}|\boldsymbol{\alpha}^\top(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\alpha}|\geq 2C\max\left[\left\{\frac{k\log(p/\eta)}{n}\right\}^4,\sqrt{\frac{k\log(p/\eta)}{n}}\right]\right)\leq C'\eta\,.$$

选择 $\eta=p^{-3}$。注意到 $k\log p\ll n$ 并且（9.40）对于充分大的 n成立，证明完毕。□

9.4.2 引理证明 2

不失一般性，记

$$\boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}^{(S,S)}&\boldsymbol{\Sigma}^{(S,S^c)}\\ \boldsymbol{\Sigma}^{(S^c,S)}&\boldsymbol{\Sigma}^{(S^c,S^c)}\end{pmatrix}\text{ and }\boldsymbol{L}=\begin{pmatrix}\boldsymbol{L}^{(S,S)}&\boldsymbol{0}\\ \boldsymbol{L}^{(S^c,S)}&\boldsymbol{L}^{(S^c,S^c)}\end{pmatrix}.$$

于是，

$$\begin{pmatrix} \boldsymbol{\Sigma}^{(S,S)} & \boldsymbol{\Sigma}^{(S,S^c)} \\ \boldsymbol{\Sigma}^{(S^c,S)} & \boldsymbol{\Sigma}^{(S^c,S^c)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}^{(S,S)} & \mathbf{0} \\ \boldsymbol{L}^{(S^c,S)} & \boldsymbol{L}^{(S^c,S^c)} \end{pmatrix} \begin{pmatrix} \boldsymbol{L}^{(S,S),\top} & \boldsymbol{L}^{(S^c,S),\top} \\ \mathbf{0}^\top & \boldsymbol{L}^{(S^c,S^c),\top} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}^{(S,S)}\boldsymbol{L}^{(S,S),\top} & * \\ * & * \end{pmatrix},$$

这意味着 $\boldsymbol{\Sigma}^{(S,S)} = \boldsymbol{L}^{(S,S)}\boldsymbol{L}^{(S,S),\top}$。由于 $\boldsymbol{\Sigma}^{(S,S)} = \boldsymbol{L}_{(S,S)}\boldsymbol{L}_{(S,S)}^\top$ 和 $\boldsymbol{L}_{(S,S)}$ 唯一，我们得到 $\boldsymbol{L}_{(S,S)} = \boldsymbol{L}^{(S,S)}$。

根据分块矩阵的逆矩阵公式，有

$$\begin{aligned}
\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}(\boldsymbol{L}^{-1})^\top &= \begin{pmatrix} \boldsymbol{L}^{(S,S)} & \mathbf{0} \\ \boldsymbol{L}^{(S^c,S)} & \boldsymbol{L}^{(S^c,S^c)} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Sigma}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)} & \boldsymbol{\Sigma}^{(S,S^c)}_{\mathbb{E}(\boldsymbol{X}|Y)} \\ \boldsymbol{\Sigma}^{(S^c,S)}_{\mathbb{E}(\boldsymbol{X}|Y)} & \boldsymbol{\Sigma}^{(S^c,S^c)}_{\mathbb{E}(\boldsymbol{X}|Y)} \end{pmatrix} \begin{pmatrix} \boldsymbol{L}^{(S,S)} & \mathbf{0} \\ \boldsymbol{L}^{(S^c,S)} & \boldsymbol{L}^{(S^c,S^c)} \end{pmatrix}^{-1,\top} \\
&= \begin{pmatrix} \boldsymbol{L}^{(S,S),-1} & \mathbf{0} \\ -\boldsymbol{L}^{(S^c,S^c),-1}\boldsymbol{L}^{(S^c,S)}\boldsymbol{L}^{(S,S),-1} & \boldsymbol{L}^{(S^c,S^c),-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)} & \boldsymbol{\Sigma}^{(S,S^c)}_{\mathbb{E}(\boldsymbol{X}|Y)} \\ \boldsymbol{\Sigma}^{(S^c,S)}_{\mathbb{E}(\boldsymbol{X}|Y)} & \boldsymbol{\Sigma}^{(S^c,S^c)}_{\mathbb{E}(\boldsymbol{X}|Y)} \end{pmatrix} \\
&\quad \begin{pmatrix} \boldsymbol{L}^{(S,S),-1,\top} & -(\boldsymbol{L}^{(S^c,S^c),-1}\boldsymbol{L}^{(S^c,S)}\boldsymbol{L}^{(S,S),-1})^\top \\ \mathbf{0}^\top & \boldsymbol{L}^{(S^c,S^c),-1,\top} \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{L}^{(S,S),-1}\boldsymbol{\Sigma}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)} & \boldsymbol{L}^{(S,S),-1}\boldsymbol{\Sigma}^{(S,S^c)}_{\mathbb{E}(\boldsymbol{X}|Y)} \\ * & * \end{pmatrix} \begin{pmatrix} \boldsymbol{L}^{(S,S),-1,\top} & * \\ \mathbf{0}^\top & * \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{L}^{(S,S),-1}\boldsymbol{\Sigma}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{L}^{(S,S),-1,\top} & * \\ * & * \end{pmatrix}.
\end{aligned}$$

显然，$\{\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}(\boldsymbol{L}^{-1})^\top\}^{(S,S)} = \boldsymbol{L}^{(S,S),-1}\boldsymbol{\Sigma}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{L}^{(S,S),-1,\top}$。结合 $\boldsymbol{L}_{(S,S)} = \boldsymbol{L}^{(S,S)}$，我们有 $\{\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}(\boldsymbol{L}^{-1})^\top\}^{(S,S)} = \boldsymbol{L}^{-1}_{S(:,S)}\boldsymbol{\Sigma}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)}(\boldsymbol{L}^{-1}{}_{(S,:S)})^\top = \boldsymbol{M}^{(S,S)}$。证明完毕。□

参考文献

Anaraki, F. P. 和 Hughes, S.（2014），“通过极稀疏随机投影实现内存与计算高效的主成分分析”，载于国际机器学习大会，PMLR，第 1341–1349 页。

Bondell, H. D. 和 Li, L. (2009)，“用于无模型变量选择的收缩逆回归估计”，英国皇家统计学会学报：*B*辑（统计方法学），71，287–299。

Bura, E. 和 Cook, R. D.（2001），《扩展切片逆回归：加权卡方检验》，《美国统计学会会刊》，96，996–1003。Cai, T. T.、Ma, Z. 和 Wu, Y.（2013），《稀疏主成分分析：最优收敛速率与自适应估计》，《统计年刊》，41，3074–3110。Chen, X.、Zou, C. 和 Cook, R.（2010），《坐标无关的稀疏充分降维与变量选择》，《统计年刊》，38，3696–3723。Cook, R.（1996），《二元响应回归的图形方法》，《美国统计学会会刊》，91，983–992。Cook, R. D.（2004），《充分降维中的预测变量贡献检验》，《统计年刊》，32，1062–1092。—（2007），《费希尔讲座：回归中的降维》，《统计科学》，22，1–26。Cook, R. D. 和 Forzani, L.（2009），《基于似然的充分降维》，《美国统计学会会刊》，104，197–208。Cook, R. D. 和 Weisberg, S.（1991），《关于用于降维的切片逆回归的讨论》，《美国统计学会会刊》，86，328–332。Cook, R. D. 和 Yin, X.（2001），《理论与方法：特别邀请论文：判别分析中的降维与可视化（附讨论）》，《澳大利亚与新西兰统计学杂志》，43，147–199。Fan, J.、Shao, Q.-M. 和 Zhou, W.-X.（2018），《发现是否为虚假发现？最大虚假相关的分布及其应用》，《统计年刊》，46，989。Gao, C.、Ma, Z.、Ren, Z. 和 Zhou, H. H.（2015），《稀疏典型相关分析中的极小极大估计》，《统计年刊》，43，2168–2197。Gao, C.、Ma, Z.（2017），《稀疏典型相关分析：自适应估计与计算障碍》，《统计年刊》，45，2074–2101。Gataric, M.、Wang, T. 和 Samworth, R. J.（2020），《基于轴对齐随机投影的稀疏主成分分析》，《皇家统计学会杂志：*B*辑（统计方法论）》，82，329–359。Hsing, T. 和 Carroll, R. J.（1992），《切片逆回归的渐近理论》，《统计年刊》，1040–1061。Hung, H. 和 Huang, S.-Y.（2019），《针对大p小n问题的基于随机划分的充分降维》，《生物统计学》，75，245–255。Li, B. 和 Wang, S.（2007），《关于用于降维的方向回归》，《美国统计学会会刊》，102，997–1008。

Li, K. (1991), 《用于降维的切片逆回归》（含讨论），《美国统计学会会刊》，86，316–327。Li, K.-C. (1992), 《关于用于数据可视化和降维的主海森方向：Stein引理的另一种应用》，《美国统计学会会刊》，87，1025–1039。Li, L. (2007), 《稀疏充分降维》，《生物统计学》，94，603–613。Li, L.、Dennis Cook, R.和Nachtsheim, C. J. (2005), 《无模型变量选择》，《皇家统计学会杂志：*B*辑（统计方法论）》，67，285–299。Li, L.、Wen, X. M.和Yu, Z. (2020), 《稀疏充分降维概览》（含讨论），《统计理论与相关领域》，4，121–133。Lin, Q.、Li, X.、Huang, D.和Liu, J. S. (2017), 《高维情形下切片逆回归的最优性》，*arXiv*预印本 *arXiv:1701.06009*。Lin, Q.、Zhao, Z.和Liu, J. (2018), 《高维情形下切片逆回归的一致性与稀疏性》，《统计年刊》，46，580–610。Lin, Q.、Zhao, Z.和Liu, J. S. (2019), 《通过Lasso实现稀疏切片逆回归》，《美国统计学会会刊》，114，1726–1739。Liu, C.、Zhao, X.和Huang, J. (2023), 《高维单指标模型中假设检验的随机投影方法》，《美国统计学会会刊》，1–11。Ma, Y.和Zhu, L. (2013), 《降维综述》，《国际统计评论》，81，134–150。Massart, P. (2007), 《集中不等式与模型选择：*Saint-Flour*概率论暑期学校*XXXIII-2003*》，施普林格。Qi, H.和Hughes, S. M. (2012), 《数据低维随机投影下主成分的不变性》，载于《*2012*年第*19*届*IEEE*国际图像处理会议》，IEEE，第937–940页。Tan, K.、Shi, L.和Yu, Z. (2020), 《稀疏切片逆回归：最优速率与自适应估计》，《统计年刊》，48，64–85。Tan, K. M.、Wang, Z.、Liu, H.和Zhang, T. (2018), 《稀疏广义特征值问题》，《皇家统计学会杂志：*B*辑（统计方法论）》，80，1057–1086。Thorgeirsson, T. E.、Gudbjartsson, D. F.、Surakka, I.、Vink, J. M.、Amin, N.、Geller, F.、Sulem, P.、Rafnar, T.、Esko, T.、Walter, S.等 (2010), 《CHRNA6和CYP2A6上的序列变异影响吸烟行为》，《自然遗传学》，42，448–453。

Thorisson, G. A.、Smith, A. V.、Krishnan, L.和Stein, L. D. (2005), 《国际HapMap项目网站》, 《基因组研究》, 15, 1592–1593。van der Vaart, A. W.和Wellner, J. (1996), 《弱收敛与经验过程：及其在统计学中的应用》, 施普林格科学与商业媒体。Vershynin, R. (2010), 《随机矩阵的非渐近分析导论》, *arXiv*预印本 *arXiv:1011.3027*。Wang, T.、Berthet, Q.和Samworth, R. J. (2016a), 《稀疏主成分估计中的统计与计算权衡》, 《统计年刊》, 44, 1896–1930。Wang, T.、Berthet, Q.— (2016b), 《稀疏主成分估计中的统计与计算权衡》, 《统计年刊》, 44, 1896–1930。Yin, X. (2011), 《回归中的充分降维》, 载于《高维数据分析》, 世界科学出版社, 第257–273页。Yu, B. (1997), 《Assouad、Fano与Le Cam》, 载于《献给*Lucien Le Cam*的纪念文集》, 施普林格, 第423–435页。Yu, Y.、Wang, T.和Samworth, R. J. (2015), 《适用于统计学家的Davis–Kahan定理的一个有用变体》, 《生物统计学》, 102, 315–323。Yu, Z.、Dong, Y.和Shao, J. (2016), 《用于超高维无模型特征选择的边际切片逆回归》, 《统计年刊》, 44, 2594–2623。Zhu, L.、Miao, B.和Peng, H. (2006), 《带有高维协变量的切片逆回归》, 《美国统计学会会刊》, 101, 630–643。Zhu, L.-X.和Ng, K. W. (1995), 《切片逆回归的渐近理论》, 《中华统计学志》, 727–736。