已提交至《应用统计学年刊》

arXiv:2602.04178v1 [stat.ME] 4 Feb 2026

通过双重阈值化进行稀疏组主成分分析及其在多细胞程序中的应用

BY QI XU[1,a], JING LEI[1,b] AND KATHRYN ROEDER[1,c]

[1]统计与数据科学系，卡内基梅隆大学a*qixu@andrew.cmu.edu*；
b*jinglei@andrew.cmu.edu*；c*roeder@andrew.cmu.edu*

多细胞程序（多细胞程序）是跨相互作用细胞类型的协调基因表达模式，共同驱动组织发育和免疫反应等复杂生物过程。尽管多细胞程序通常使用稀疏主成分分析或潜因子模型等方法从高维基因表达数据中进行估计，但这些方法往往计算成本高昂，且统计功效有限。在本研究中，我们提出稀疏组主成分分析（SGPCA），通过利用多细胞程序固有的组稀疏性和个体稀疏性来估计多细胞程序。我们引入了一种基于幂迭代的高效双重阈值算法。在每次迭代中，组阈值化步骤首先识别相关基因群，随后通过个体阈值化步骤选择活跃的细胞类型。该算法的计算复杂度为 $O(np)$，因而对于大规模基因组分析具有很高的效率和可扩展性。我们为 SGPCA 建立了理论保证，包括统计相合性，以及优于竞争方法的收敛速率。通过广泛的模拟，我们证明 SGPCA 能够实现更高的估计准确度，并提高信号检测的统计功效。此外，我们将 SGPCA 应用于一项 Lupus 研究，发现了可区分 Lupus 患者与正常受试者的差异表达多细胞程序。

1. 引言。 单细胞RNA测序（scRNA-seq）使研究人员能够以单个细胞为分辨率分析基因表达，彻底改变了现代生物学。这种高分辨率视角揭示了复杂组织中显著的细胞异质性，并支撑细胞类型注释：细胞根据不同的转录特征进行分组（Abdelaal等人，2019年；Ji等人，2023年）。除了识别细胞类型之外，一个新兴的挑战是理解不同细胞类型如何在组织内协调其基因表达。这推动了对多细胞程序（MCPs）的研究；多细胞程序是指在多种细胞类型中同时上调或下调的一组基因，并参与关键生物过程（Jerby-Arnon和Regev，2022年；Mitchel等人，2024年）。多细胞程序不仅具有生物学意义，也与临床相关。例如，Ling等人（2024年）发现了突触神经元与星形胶质细胞程序（SNAP）。这是一种在神经元和星形胶质细胞中观察到的协调表达程序，其活性降低与精神分裂症或高龄有关。

MCP发现采用了多种统计方法，包括潜因子模型（Stegle等人，2012；Ling等人，2024）和稀疏主成分分析（稀疏主成分分析）（Witten、Tibshirani和Hastie，2009；Jerby-Arnon和Regev，2022）。主成分分析和因子模型能够捕捉基因共表达的共有模式，而稀疏主成分分析通过纳入稀疏结构，尤其适用于基因组数据。从生物学角度看，基因表达本质上具有稀疏性，因为在任何给定细胞中，只有一部分基因处于活跃表达状态。从统计学角度看，在高维情形下，稀疏性至关重要 $(p > n)$，因为众所周知，若无结构性假设，经典主成分分析并不一致（Johnstone 和 Lu，2009）。

关键词和短语：数据整合、层次稀疏性、单细胞RNA测序、稀疏组套索。

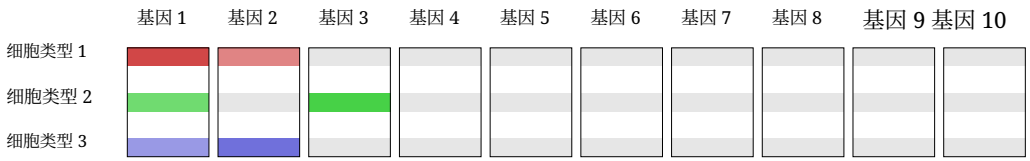

图1 1. 三种细胞类型中基因共表达模式的示意图。在此示例中，基因*1*在所有细胞类型中均有表达；基因*2*在细胞类型*1*和*3*中表达；基因*3*仅在细胞类型*2*中表达。所有其他基因在所有细胞类型中均不表达。灰色方块表示不表达，彩色方块表示活跃表达，透明度的变化则表示不同的表达水平。

从高维、多细胞类型基因表达数据中估计多细胞程序时，必须考虑更细致的层次稀疏结构。在我们的设定中，每个基因构成一个组，包含其在所有细胞类型中的表达。通常，只有少数基因参与任一给定的多细胞程序，因此形成组级稀疏性。在每个活跃基因内部，其表达模式可能因细胞类型而异：一个基因可能在多个、但不一定是全部参与的细胞类型中共同表达。这使得对应于不同细胞类型的各项呈现组内稀疏性。图1 展示了这种层次稀疏性；自闭症谱系障碍（Wamsley等人，2024）、衰老以及精神分裂症（Ling等人，2024）中的生物学发现也支持这一点。因此，准确估计多细胞程序需要能够明确捕捉这两个层次稀疏性的统计方法。

一种实现层次稀疏性的自然方法源自稀疏组套索（Friedman、Hastie和Tibshirani，2010年；Simon等人，2013；Cai、Zhang 和 Zhou，2022年），该方法结合使用 ℓ_1 和 ℓ_2 惩罚项，以选择活跃组以及每个组内的活跃变量。近期，Xiao 和 Xiao（2024）基于 Fantope 公式中的凸松弛，将这一思路应用于稀疏主成分分析（sparsePCA）（Vu等人，2013）。然而，由此得到的交替方向乘子法（ADMM）算法需要执行代价高昂的 Fantope投影步骤，其每次迭代的复杂度为 $O(p^3)$，限制了其在高维基因表达数据上的可扩展性。此外，Xiao 和 Xiao（2024）提出的理论误差率 $O(\sqrt{\log(p)}/n)$未利用潜在的组稀疏性，因此在存在层次结构的情形下并非最优。

本文提出**稀疏组主成分分析**(SGPCA)，这是一种可扩展算法，用于在层次稀疏性条件下估计前导主成分（主成分）。我们的算法以经典幂迭代（GolubandVanLoan，2013）为基础，在每次迭代中加入双重阈值化步骤：先按组阈值化以识别活跃组，再逐元素阈值化以促进个体稀疏性。值得注意的是，Ma（2013）提出的迭代阈值化过程是我们方法的一个特例，其中仅执行逐元素阈值化。为估计多个主成分，我们采用去膨胀过程（Jolliffe，2002）。一项关键的实际挑战是为每个主成分选择稀疏参数。我们借鉴稳定性选择（Meinshausen and Bühlmann，2010；Sun、Wang和Fang，2013），采用一种基于重抽样的准则来解决这一问题。该准则评估不同重采样数据集上估计主成分的对齐度，并在变量选择方面展现出优异的经验性能。

我们的方法为相关文献作出了多项贡献。首先，SGPCA计算效率高且可扩展，非常适用于高维基因表达数据。每次迭代在矩阵乘法上需要 $O(np)$ 次运算，在归一化和阈值化上仅需 $O(p)$ 次运算，显著优于（ $O(p^3)$ ）中的计算复杂度（Xiao 和 Xiao，2024年）。其次，在具有层次稀疏性的尖峰协方差模型下，我们证明了估计的主要主成分具有一致性，并推导出比（Xiao 和 Xiao，2024年）和（Ma，2013）更快的收敛速率。第三，与常用方法相比，我们基于重抽样的调参策略在变量选择方面表现更优，

例如数据拟合（Witten、Tibshirani和Hastie，2009年）或稀疏程度（Zou、Hastie和Tibshirani，2006），在信号检测中实现了更低的第一类和第二类错误率，这对MCP发现问题尤其有益。最后，我们提供了高效且易于使用的 R 包实现，并配备交互式调参流程，便于实际应用和结果复现。该软件包以 R 包SGPCA的形式发布，并附有可复现的数值实验，网址为 https://github.com/statsqixu/SGPCA_EXP。

1.1. 其他相关研究。 稀疏性一直是高维主成分分析的核心主题。早期方法，例如（Jolliffe、Trendafilov和UDDIN，2003），对载荷向量施加了 ℓ_1约束，但计算成本过高。后续方法将稀疏主成分分析重新表述为正则化回归或矩阵分解问题（Zou、Hastie和Tibshirani，2006；Shen and Huang，2008；Witten、Tibshirani和Hastie，2009年），通过阈值化步骤实现了高效的迭代算法。另一类研究聚焦于稀疏主成分分析的凸松弛，并使用半定规划估计主成分（d'Aspremont等人，2004；Vu等人，2013）。这些方法具有强有力的理论保证，但计算成本仍然很高。全面综述请参见（Zou和Xue，2018年）。

张量分解提供了另一种视角，将受试者、基因和细胞类型视为不同的模。Tucker分解已用于MCP发现（MITCHEL等人，2024年），但相合估计通常要求张量各模的维度相当（ZHANG和Xia，2018年），而在典型的单细胞RNA测序数据集中，这一条件并不成立，因为细胞类型的数量远少于基因或受试者的数量。因此，张量方法在这种情形下缺乏理论支持。

另一个相关方向是多视图学习，它通过整合多个数据集来识别共享成分和视图特异成分。相关例子包括加性分解（Lock 等人，2013）和乘性结构（Tang 和 Allen，2021）。近期研究则探讨了部分共享成分（Gaynanova 和 Li，2019年；Lock、Park 和 Hoadley，2022年；Xiao 和 Xiao，2024年；Xu 和 Qu，2025年），能够灵活地刻画视图间的异质关系。从这一角度看，MCP估计可视为对多视图数据进行稀疏主成分分析，其中各视图中主成分的支撑集呈现部分共享结构。

1.2. 符号与文章结构。本文采用以下符号体系。随机变量或向量用大写字母表示，例如 X；粗体符号 $\mathbf{X}$、$\mathbf{x}$ x分别表示矩阵、向量和标量。对于任意正整数n，我们使用 $[n]=\{1,2,\ldots,n\}$，而 $|\cdot|$表示实数的绝对值或集合的基数。我们记 $a\wedge b=\min(a,b)$ 和 $a\vee b=\max(a,b)$。对于非负数列 $\{a_n\}$ 和 $\{b_n\}$，若存在与 n无关的常数$C>0$，使得 $a_n\le Cb_n$，则记作 $a_n\lesssim b_n$ 或 $a_n=O(b_n)$；若 $a_n\lesssim b_n$ 和 $b_n\lesssim a_n$ 均成立，则记作 $a_n\asymp b_n$。此外，$a_n=o(b_n)$表示存在某个 $c_n\to 0$，使得 $a_n=c_nb_n$。对于任意向量 $\mathbf{x}$，$\|\mathbf{x}\|_q$ 表示其 ℓ_q 范数。对于任意矩阵 $\mathbf{X}$，$\|\mathbf{X}\|_2=\sup_{\|\mathbf{v}\|_2=1}\|\mathbf{X}\mathbf{v}\|_2$ 表示谱范数，而 $\|\mathbf{X}\|_F$表示弗罗贝尼乌斯范数。默认情况下，$\|\cdot\|$ 对向量表示 ℓ_2 范数，对矩阵表示谱范数。

我们形式化问题设定，并介绍双阈值算法及调节参数程序，见第2节。所提出估计量的一致性与收敛速率在第3节中建立。在第4节中，我们给出模拟研究，并评估估计与信号检测性能。此外，我们用所提出的估计量展示了在Lupus数据中发现多细胞程序的结果，见第5节。结论与进一步讨论见第6节。

2. 方法论。 假设我们的数据由 n 个独立同分布随机样本$\mathbf{x}_1, \cdots, \mathbf{x}_n$组成，这些样本来自 $\mathcal{P}_X$的分布，其中每个观测$\mathbf{x}_i \in \mathbb{R}^p$服从低秩因子模型：

$$\mathbf{x}_i = \sum_{j=1}^{\bar{J}} \lambda_j q_{ij} \mathbf{v}_j + \epsilon_i, \quad i \in [n], \quad p \gg n, \tag{2.1}$$

其中，向量 $\mathbf{v}1, \cdots, \mathbf{v}_{\bar{J}} \in \mathbb{R}^p$ 是确定性的正交归一载荷向量，满足 $\mathbf{v}_i \mathbf{v}_j = \delta_{ij}$, $\delta_{ij} = 1$（当 $i = j$ 时），否则满足 $\delta_{ij} = 0$。常数 $\lambda_1 > \lambda_2 > \cdots > \lambda_{\bar{J}} > 0$表示相应载荷向量的信号强度。对于每个样本 i，随机向量 $q_i = (q_{i1}, \cdots, q_{i\bar{J}})$包含随机得分，假设其满足 $\mathbb{E}[q_{ij}] = 0$¯，且对所有 i, $i' \in [n], j, j' \in [J]$均有 $\mathbb{E}[q_{ij} q_{i'j'}] = \delta_{ii'} \delta_{jj'}$。也就是说，这些得分在不同载荷向量之间以及不同样本之间互不相关。噪声向量 $\epsilon_i \in \mathbb{R}^p$ 满足 $\mathbb{E}[\epsilon_i] = \mathbf{0}, \mathbb{E}[\epsilon_i \epsilon_i^T] = \sigma^2 \mathbf{I}$，且与 q_i不相关。最后，我们假设每个随机向量 X都存在一个已知的划分，划分为 G 个互不相交的组：$[p] = \cup_{g=1}^{G} I_g$，其中 $I_g = \{k : k$属于第 g 个组$\}$。相应地，$X = (X_{(I_1)} X_{(I_2)} \cdots, X_{(I_G)}), X_{(I_g)} \in \mathbb{R}^{p_g}$和 $p = \sum_{g \in [G]} p_g$。此外，每个载荷向量 $\mathbf{v}j$ 也按相同方式划分为 $\mathbf{v}j = (\mathbf{v}_{j,(I_1)}, \mathbf{v}_{j,(I_2)} \cdots, \mathbf{v}_{j,(I_G)})$。进一步地，我们将每个组中的元素写作 $\mathbf{v}_{j,(I_g)} = (v_{j,(I_g)\,1}, \cdots, v_{j,(I_g)\,t}, \cdots, v_{j,(I_g)\,p_g})$。

基于上述设定，X 的协方差矩阵为

$$\Sigma = \sum_{j=1}^{\bar{J}} \lambda_j^2 \mathbf{v}_j \mathbf{v}_j^T + \sigma^2 \mathbf{I}. \tag{2.2}$$

这是经典的尖峰协方差模型（Johnstone，2001年）：每个$\mathbf{v}j$ 都是 j阶特征向量，对应特征值为 $\lambda_j^2 + \sigma^2$，其余 $p - \bar{J}$ 个特征值均为 σ^2。我们的目标是在这种高维、具有组结构的设定下，估计特征向量$\mathbf{v}j$，即主成分。

众所周知，在高维设置中，除非施加额外的结构假设，否则经典主成分分析无法估计主特征向量（Johnstone 和 Lu，2009年）。本文对载荷向量$\mathbf{v}j$采用层次稀疏性假设：对于每个成分 j，只有少数预先定义的 G 组处于活跃状态，且在每个活跃组内，$\mathbf{v}j$ 在坐标层面进一步具有稀疏性。第3节将对层次稀疏性作严格刻画。为利用这种双层稀疏结构，我们提出一种双重阈值化算法，并在第2.1节介绍。

在多细胞程序发现的背景下，低秩因子模型在我们所用的统计对象与其底层生物学机制之间建立了直接且易于解释的联系。随机向量 $X = (X_{(I_1)} \cdots, X_{(I_G)})$ 表示 G 个基因在所有细胞中的表达水平，其中每个基因对应一个组，其组级区块 $X_{(I_g)} \in \mathbb{R}^T$ 包含该基因在所研究的 T 种细胞类型中的表达。在尖峰协方差模型中，每个载荷向量 $\mathbf{v}j = (\mathbf{v}_{j,(I_1)} \cdots, \mathbf{v}_{j,(I_G)})$ 定义第 j 个多细胞程序：区块 $\mathbf{v}_{j,(I_g)}$ 描述基因 g 对该多细胞程序的贡献强度（涵盖所有细胞类型），而 $\mathbf{v}_{j,(I_g)}$ 中的非零项则指出基因 g 参与该程序的具体细胞类型。$\mathbf{v}j$ 的基因层面组稀疏性对应于这样一种生物学事实：每个多细胞程序仅涉及一部分基因；而组内稀疏性则反映了参与其中的基因可能仅在部分相关细胞类型中活跃。按照这种解释，协方差结构的主特征向量编码了组织中存在的主要多细胞程序，而估计稀疏的 $\mathbf{v}j$ 相当于恢复这些程序在不同细胞类型中的底层基因共表达模式。

2.1. 双重阈值算法。 令 $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ 表示样本协方差矩阵。在没有稀疏性约束的情况下，可通过经典幂迭代算法（Golub和Van Loan，2013）求得主成分（PC）；该算法交替执行矩阵乘法和向量归一化。随后，可对压缩协方差矩阵应用幂迭代来估计后续主成分。

为纳入上述分层稀疏结构，我们通过引入双重阈值化步骤来扩展幂迭代。在每次迭代中，我们首先对各组应用块（按组）软阈值算子，按照各组的 ℓ_2 范数对其进行收缩。对于此步骤后保留的组，我们再逐个应用（逐元素）软阈值化，以在每个活跃组内的各元素之间进一步引入稀疏性。由此得到的算法总结于算法1中。

算法 1： 用于稀疏组主成分分析的双重阈值算法

输入： 样本协方差矩阵 $\hat{\mathbf{\Sigma}}$；主成分数量 J；按组阈值水平 η_1、……、η_J；逐元素阈值水平 τ_1、……、τ_J；

初始估计 $\hat{\mathbf{v}}_1^{(0)}, \ldots, \hat{\mathbf{v}}_J^{(0)}$，$\mathbf{S} = \hat{\mathbf{\Sigma}}$.

对于 $j \leftarrow 1$ **至** J，**执行**

- 设定 $k = 0$;
- **重复**
  - **乘法：** $\hat{\boldsymbol{\gamma}}_j^{(k)} = \mathbf{S}\hat{\mathbf{v}}_j^{(k)}$;
  - **对于** $g \leftarrow 1$ **到** G **执行**
    - **按组阈值化：**

$$\hat{\boldsymbol{\gamma}}_{j,(g)}^{(k+1/2)} = \mathtt{bst}\left(\hat{\boldsymbol{\gamma}}_{j,(g)}^{(k)}; \sqrt{p_g}\eta_j\right) = \left(1 - \frac{\sqrt{p_g}\eta_j}{\|\hat{\boldsymbol{\gamma}}_{j,(g)}^{(k)}\|_2}\right)_+ \hat{\boldsymbol{\gamma}}_{j,(g)}^{(k)}.$$

    - **逐元素阈值化：** **对于** $t \leftarrow 1$ **到** p_g **执行**

$$\hat{\gamma}_{j,(I_g),t}^{(k+1)} = \mathtt{st}\left(\hat{\gamma}_{j,(I_g),t}^{(k+1/2)}; \tau_j\right) = \left(1 - \frac{\tau_j}{|\hat{\gamma}_{j,(I_g),t}^{(k+1/2)}|}\right)_+ \hat{\gamma}_{j,(I_g),t}^{(k+1/2)}.$$

  - **归一化：** $\hat{\mathbf{v}}_j^{(k+1)} = \frac{\hat{\boldsymbol{\gamma}}_j^{(k+1)}}{\|\hat{\boldsymbol{\gamma}}_j^{(k+1)}\|_2}$;
  - $k \leftarrow k + 1$;
- **直到**收敛;
- **去膨胀：**

$$\mathbf{S} \leftarrow \mathbf{S} - \left(\hat{\mathbf{v}}_j^\top \mathbf{S}\hat{\mathbf{v}}_j\right)\hat{\mathbf{v}}_j\hat{\mathbf{v}}_j^\top.$$

输出： $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_J$

块软阈值算子 bst 和逐项软阈值算子 st 是以下凸惩罚最小二乘问题的唯一极小化解

$$\mathtt{bst}(x;\lambda) = \arg\min_y \left\{\frac{1}{2}\|y - x\|_2^2 + \lambda\|y\|_2\right\}, \qquad \mathtt{st}(x;\lambda) = \arg\min_y \left\{\frac{1}{2}(y - x)^2 + \lambda|y|\right\},$$

and 也可以使用其他惩罚方法（例如硬阈值、平滑剪切绝对偏差（SCAD）Fan 和 Li，2001）。 d.

良好的初始化和适当的停止规则对于理论保证和经验性能都至关重要。出于理论目的，我们将Johnstone 和 Lu (2009年)的对角阈值法加以扩展，构造一种同时实现组稀疏性和元素级稀疏性的初始化方法，如算法3所述。当估计误差达到与预言估计量相当的目标速率时，算法停止，详见第3节。在实践中，我们使用粗略调参的SPC（Witten、Tibshirani和Hastie，2009年）进行初始化，该方法在模拟中始终表现良好。对于每个$\hat{\mathbf{v}}j$，当距离

$$d\left(\hat{\mathbf{v}}_j^{(K_j)}, \hat{\mathbf{v}}_j^{(K_j-1)}\right) = \left\| \hat{\mathbf{v}}_j^{(K_j)} \hat{\mathbf{v}}_j^{(K_j)\top} - \hat{\mathbf{v}}_j^{(K_j-1)} \hat{\mathbf{v}}_j^{(K_j-1)\top} \right\|_F^2 \leq \mathtt{tol},$$

该指标衡量连续估计值所张成的一维子空间之间的 Frobenius距离，且不受符号翻转影响。我们在数值实验中始终使用 $\mathtt{tol} = 10^{-5}$ 。

该算法的计算效率很高。乘法步骤可计算为$\hat{\boldsymbol{\gamma}}_j^{(k)} = \frac{1}{n}\mathbf{X}^\top(\mathbf{X}\hat{\mathbf{v}}_j^{(k)})$，需要 $O(np)$ 次运算。阈值化和归一化各需要 $O(p)$ 次运算，按组阈值化则可在各组之间并行执行。因此，每次迭代的复杂度为 $O(np)$；而Xiao 和 Xiao （2024年）的方法由于需要进行 Fantope投影，成本为$O(p^3)$ 。这种计算效率优势使 SGPCA 能够扩展至高维基因表达数据，并非常适用于 MCP发现。

2.2. 调节参数的选择。 在任何稀疏主成分估计方法中，与稀疏性相关的调节参数对于平衡稀疏性与估计偏差之间的权衡至关重要。文献中提出了若干选择这些参数的原则，例如设定目标稀疏程度（Shen and Huang，2008；Zou、Hastie和Tibshirani，2006）或优化数据拟合精度（Witten、Tibshirani和Hastie，2009年）。然而，这些准则的设计重点并非变量选择，而变量选择正是 MCP发现的核心关注点；我们的目标是识别驱动每个 MCP 的基因和细胞类型。

受稳定性选择原则（Meinshausen and Bühlmann，2010年；Sun、Wang和Fang，2013）的启发，我们提出了一种基于重采样的程序，用于为双重阈值算法选择阈值化水平，旨在最大限度地提高估计主成分在重采样数据集间的稳定性。具体而言，给定重采样比例$\rho \in (0,\ 1)$ （$\rho = 1/2$ 在我们的数值实验中），我们从数据矩阵 $\mathbf{X}$ 的原始 n 行中无放回抽取$\lfloor n\rho \rfloor$ 个观测值，生成 B 个子样本。对于一组给定的调节参数$(\eta,\ \tau)$和固定的成分索引 j，令$\hat{\mathbf{v}}_{j,\ b,\ \eta,\ \tau}$表示使用第 b个子样本$\mathbf{X}_b$通过算法1得到的第 j个主成分的估计量。我们将第 j个主成分在$(\eta,\ \tau)$下的对齐得分定义为

$$\mathtt{align}(j,\eta,\tau) = \frac{2}{B(B-1)} \sum_{1 \leq b < b' \leq B} \left| \hat{\mathbf{v}}_{j,b,\tau,\eta}^\top \hat{\mathbf{v}}_{j,b',\tau,\eta} \right|.$$

由于每个$\hat{\mathbf{v}}_{j,b,\tau,\eta}$的范数均为单位范数，绝对内积 $\left|\hat{\mathbf{v}}_{j}^\top{}_{,b,\tau,\eta}\hat{\mathbf{v}}_{j,b',\tau,\eta}\right| \in [0,1]$ 衡量了两个估计之间的对齐度；取绝对值则消除了特征向量的符号不确定性。align(j,τ,η)的值越大、越接近1，表示重采样样本间的一致性越强，因此所选参数对（τ,η）对应的估计量越稳定；而值越接近0，则表示估计结果几乎正交，稳定性较差。

在算法 1中，主成分通过去膨胀序贯估计，而不是像 J中那样联合估计 Ma (2013) 所述的 - 维主子空间。这一选择部分源于

算法 2： 基于重采样的调参与主成分估计

输入： 数据矩阵 $\mathbf{X}$；主成分数量 J；候选集 $\mathcal{H}$（组级阈值）和 $\mathcal{T}$（元素级阈值）；初始化 $\hat{\mathbf{v}}(0)$；重采样比例 ρ；重采样样本数量 B。

对于 $j \leftarrow 1$ **到** J，**执行**

  对于每个 $\eta \in \mathcal{H}, \tau \in \mathcal{T}$，**执行**

    对于 $b \leftarrow 1$ **到** B **执行**

      // 重采样样本

      从 $\mathbf{X}$ 中无放回抽样，随机抽取 $\lfloor n\rho \rfloor$ 行，得到 $\mathbf{X}b$；

      // 在重采样样本上进行估计

      计算 $\hat{\mathbf{\Sigma}}_b = \frac{1}{\lfloor n\rho \rfloor} \mathbf{X}_b^\top \mathbf{X}_b$；

      应用算法 1 至 $\hat{\mathbf{\Sigma}}_b$ 使用调节参数（η、τ），仅从以下对象中提取第一个主成分与第 j 个分量对应的压缩协方差矩阵。将所得估计量记为 $\hat{\mathbf{v}}_{j,b,\tau,\eta}$；

    // 对齐得分

    计算

$$\text{align}(j, \tau, \eta) = \frac{2}{B(B-1)} \sum_{1 \le b < b' \le B} \left| \hat{\mathbf{v}}_{j,b,\tau,\eta}^\top \hat{\mathbf{v}}_{j,b',\tau,\eta} \right|.$$

  选择 $(\tau_j^*, \eta_j^*) = \arg\max_{\tau \in \mathcal{T}, \eta \in \mathcal{H}} \text{align}(j, \tau, \eta)$；

  // 在完整数据上进行估计

  计算 $\hat{\mathbf{\Sigma}} = \frac{1}{n} \mathbf{X}^\top \mathbf{X}$；

  应用算法 1 对 $\hat{\mathbf{\Sigma}}$ 使用调节参数 $\left(\sqrt{\lfloor n\rho \rfloor / n}\, \tau_j^*, \sqrt{\lfloor n\rho \rfloor / n}\, \eta_j^* \right)$ 估计 jth 主成分，得到 $\hat{\mathbf{v}}_j$；

  // 去膨胀

  更新 $\mathbf{X} \leftarrow \mathbf{X} - (\mathbf{X}\hat{\mathbf{v}}_j)\hat{\mathbf{v}}_j^\top$；

输出： $\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_J$

调优参数选择的计算负担。假设逐元素阈值水平 τ_1、……、τ_J从候选集 $\mathcal{T}$中选取，而按组阈值水平η_1、……、η_J从候选集 $\mathcal{H}$中选取。联合估计 J-维主子空间需要在 $(|\mathcal{T}||\mathcal{H}|)^J$ 个调优参数对组合中进行搜索；当搜索空间较大时，这在计算上将变得不可行 $\mathcal{T}$， $\mathcal{H}$ 或者当主成分数量较多时亦然 J。相比之下，算法 1 中的序贯消去法仅需 $J|\mathcal{T}||\mathcal{H}|$ 个组合，从而显著节省计算成本。

另一方面，如果实践者认为各 J主要主成分的稀疏度相近，也可以在正交迭代方案（Golub和Van Loan，2013）中应用双重阈值化步骤，直接估计主子空间，并对所有分量使用一组共同的阈值水平。在这种情况下，可以修改算法1，完全避免消去步骤。基于子空间的估计还有一个优点，即能够保持估计主成分之间的正交性；而消去与稀疏性相结合则可能导致其偏离严格正交。

调节参数选择和主成分估计的完整流程总结于算法 2。从理论角度来看，如第 τ_j 和 η_j节所示，为了达到一致性，阈值水平应以 $\sqrt{1/n}$ 的速率衰减。当使用大小为 $\lfloor n\rho \rfloor$的子样本时，因此，在完整数据集上重新估计主成分时，我们会将选定的调节参数乘以 $\sqrt{\lfloor n\rho \rfloor / n}$，以考虑样本量的影响。算法 1和 2 已在 R 包 SGPCA中实现。该软件包还提供重采样数据集中对齐得分与平均支撑集大小的关系图，可帮助人工检查稀疏模式和阈值水平。作为一个副产品，

¯这些图有助于选择有效的尖峰分量数量 J。直观而言，如果某个主成分呈现稀疏结构，那么随着平均支撑集大小接近真实支撑集大小，对齐得分往往会趋近于 1；而纳入更多含噪¯变量后，对齐得分则会下降。相比之下，对于尖峰协方差模型（2.2）中的 $j > J$，其余特征向量是稠密的，对齐得分会随支撑集大小单调增加。附录 A 中提供了一个示例。

3. 统计性质。本节为算法1生成的估计量建立理论保证，该算法以算法3初始化。我们证明了在能够刻画组级和组内稀疏性的层次稀疏模型下的一致性，并给出收敛速率。与针对纯个体稀疏性设计的迭代阈值法（Ma，2013）相比，我们的分析适用于分组变量。由于算法1每次估计一个领先主成分（PC），因此我们聚焦于单尖峰协方差模型。

形式化设定。假设$\mathbf{x}_1$、……、$\mathbf{x}_n \in \mathbb{R}^p$由单尖峰模型生成

$$\mathbf{x}_i = \lambda q_i \mathbf{v} + \epsilon_i, \qquad i \in [n], \tag{3.1}$$

这是（2.1）在 $\bar{J} = 1$ 和 $\sigma^2 = 1$ 条件下的特例。我们假设组大小相等：$p_1 = \cdots = p_G = T$，因此 $p = GT$。我们将 n 作为问题索引，允许 $p = p(n)$、$G = G(n)$ 和 $T = T(n)$ 随 n 增大，而将尖峰强度 $\lambda > 0$ 视为常数（与 n 无关）。我们施加以下分布条件。

条件 3.1. 对于模型 (3.1)，我们作如下假设：

a. 潜在得分 $\{q_i\}_{i=1}^n$ 是独立同分布的、均值为零、方差为 *1* 的次高斯变量，其参数为 K_q，即对所有 $t > 0$，$\mathbb{P}(|q_i| \geq t) \leq 2\exp(-t^2/K_q^2)$。b. 噪声向量 $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})$ 的各坐标相互独立，其中每个 ϵ_{ij} 都是均值为零、方差为 *1* 的次高斯变量，其参数为 K_ϵ，即对所有 $t > 0$，$\mathbb{P}(|\epsilon_{ij}| \geq t) \leq 2\exp(-t^2/K_\epsilon^2)$。c. $\{q_i\}$ 与 $\{\epsilon_{ij}\}$相互独立。

在模型（3.1）和条件3.1下，总体协方差为

$$\Sigma = \lambda^2 \mathbf{v}\mathbf{v}^\top + \mathbf{I}.$$

对于单位向量 $\mathbf{v}_1$,$\mathbf{v}_2$,，定义子空间距离平方

$$d(\mathbf{v}_1, \mathbf{v}_2) := \|\mathbf{v}_1\mathbf{v}_1^\top - \mathbf{v}_2\mathbf{v}_2^\top\|_F^2.$$

几何上，$d(\mathbf{v}_1, \mathbf{v}_2) = 2\sin^2\theta(\mathbf{v}_1,\mathbf{v}_2)$，其中 $\theta(\mathbf{v}_1,\mathbf{v}_2) = \arccos(|\mathbf{v}_1^\top\mathbf{v}_2|)$ 是这两个一维子空间之间的主角。对于非零向量，我们使用尺度不变形式

$$d(\mathbf{v}_1, \mathbf{v}_2) := \left\| \frac{\mathbf{v}_1\mathbf{v}_1^\top}{\mathbf{v}_1^\top\mathbf{v}_1} - \frac{\mathbf{v}_2\mathbf{v}_2^\top}{\mathbf{v}_2^\top\mathbf{v}_2} \right\|_F^2,$$

其含义相同。此外，我们考虑以下高维情形：

条件 3.2。当 $n \to \infty$时，我们假设

a. $\log(G) = o(n)$ 且 $G \to \infty$；

b. $T = o(\log G)$ 且 $T \geq 1$。

条件 3.2(a) 允许组数以 n 中的次指数速率增长。条件 3.2(b) 确保 $T = o(n)$。这些条件在多细胞程序建模中很自然，在这类建模中，通常 $T \ll n \ll G$。

层次稀疏性假设。 我们考虑载荷向量 $\mathbf{v}$ 的分层稀疏结构。令 $\mathbf{v}_{(I_g)} \in \mathbb{R}^T$ 表示 $\mathbf{v}$ 在组 g 中的子向量。不失一般性，我们假设各组按其大小排序：

$$\|\mathbf{v}_{(I_1)}\|_2 \geq \cdots \geq \|\mathbf{v}_{(I_G)}\|_2.$$

我们称$\mathbf{v}$ 属于半径为 m_G 的弱-$\ell_{G,r}$ 球，记作 $\mathbf{v} \in w\ell_{G,r}(m_G)$，如果

$$\|\mathbf{v}_{(I_g)}\|_2 \leq m_G\, g^{-1/r}, \qquad \forall g \in [G], \tag{3.2}$$

对于某个固定的 $0 < r < 2$。为描述领先组中的组内稀疏性，定义

$$\mathcal{V}(g) := \{\, v_j : \; j \in g' \text{ for some } g' \leq g \,\},$$

即按 ℓ_2 范数排序后，前 g 组中包含的所有坐标。按大小排列 $\mathcal{V}(g)$ 中的各项： $|\mathbf{v}_{(1)}| \geq \cdots \geq |\mathbf{v}_{(gT)}|$。我们假设，对于每个 g，都存在一个（可能依赖于 g 的）常数 $m_{(g)}$，使得

$$|\mathbf{v}_{(j)}| \leq m_{(g)}\, j^{-1/r}, \qquad \forall j \in [gT]. \tag{3.3}$$

在全文中，我们假设 $r \in (0,2)$ 为固定值，且 m_G 和 $\{m_{(g)}\}$ 均由普适常数从上界限定。换言之，各组的幅度迅速衰减；在领先组内，各项幅度也迅速衰减。这种层次稀疏性比 ℓ_0型约束更为一般；这类约束在Cai、Zhang 和 Zhou (2022年); Xiao 和 Xiao(2024年) 中使用，并且与应用场景十分契合：在这些场景中，只有少数组处于激活状态，而且每组中也只有一部分坐标具有较强信号。

将改进的对角阈值法用作初始化。 为进行理论分析，我们通过扩展Johnstone 和 Lu (2009年) 提出的对角阈值法来构造初始化，以适应上述层级稀疏性假设，如算法 3 所示，并将调节参数设定为

$$\pi_n = \pi\sqrt{\frac{\log G}{n}}, \qquad \omega_n = \omega\sqrt{\frac{\log(T|\mathcal{B}_G|)}{n}},$$

取足够大的常数 π， $\omega > 0$。

算法 3： 用于初始化的改进对角阈值算法

输入： 样本协方差矩阵 $\hat{\Sigma}$；组阈值 π_n；个体阈值 ω_n。**输出：** 初始化向量 $\hat{\mathbf{v}}(0)$

1. **组选择**：

$$\mathcal{B}_G = \Big\{ g : \; \frac{1}{n}\sum_{i=1}^{n}\sum_{c\in g} x_{ic}^2 \geq T + \pi_n \Big\}.$$

2.**个体选择**：

$$\mathcal{B} = \Big\{ c \in \Psi(\mathcal{B}_G) : \; \frac{1}{n}\sum_{i=1}^{n} x_{ic}^2 \geq 1 + \omega_n \Big\}.$$

3. **降维** $\mathbf{C}^{\mathbf{A}}$：计算 pute the leadin g eig Σ的特征向量 $_{\mathcal{B}\mathcal{B}}$.

；4. **零填充**：将此特征向量嵌入 $\mathbb{R}^p$，方法是对所有不在 $\mathcal{B}$中的坐标赋零。

关键量与预言信号集。首先，我们定义

$$h(x) := \frac{x^2}{1+x}, \qquad x > 0,$$

它刻画了有效的信噪比尺度。特别地，当$x \ll 1$ 时，它的表现类似于 x^2 ；当 $x \gg 1$ 时，它的表现类似于 x 。基于 $h(\cdot)$，我们将组级和元素级载荷阈值定义为

$$\alpha_n = \alpha\Big(\frac{\log G}{n\,h(\lambda^2)}\Big)^{1/2}, \qquad \beta_n = \beta\Big(\frac{\log(T|\mathcal{G}|)}{n\,h(\lambda^2)}\Big)^{1/2}, \tag{3.4}$$

其中 $\alpha, \beta > 0$ 为固定常数。相应的预言信号集为

$$\mathcal{G} = \{g \in [G] : \|\mathbf{v}_{(I_g)}\|_2 \geq \alpha_n\}, \qquad \mathcal{S} = \{j \in \Psi(\mathcal{G}) : |v_j| \geq \beta_n\}. \tag{3.5}$$

在下面的引理中，我们给出预言信号集基数的上界。

引理 3.1. 当 n充分大时，预言集满足

$$|\mathcal{G}| \leq \Big(\frac{m_G}{\alpha_n}\Big)^r \wedge G, \qquad |\mathcal{S}| \leq T|\mathcal{G}| \wedge \Big(\frac{m_{(|\mathcal{G}|)}}{\beta_n}\Big)^r.$$

在算法1中，我们选择迭代阈值化水平，使其与层级信号尺度相匹配：

$$\eta_n = \eta\,\frac{\alpha_n}{\sqrt{T}}, \qquad \tau_n = \tau\,\beta_n,$$

其中常数 $\eta, \tau > 0$充分大。

评注 3.1. 对数因子 $|\mathcal{G}|$ 中的 $\log(T|\mathcal{G}|)$ 依赖于预言分析中的理论精确性 。在实践中，可以将 $\log(T|\mathcal{G}|)$ 替换为 $\log p$（或通过第 *2.2* 节所述的重抽样程序调整阈值），而不影响算法实现。

3.1. 主要结果。 下面给出主成分估计的主要结果，技术细节见附录 C–E。

定理 3.1. 假设条件 *3.1* 和 *3.2* 成立，且 $\mathbf{v}$ 满足层次稀疏性假设 (3.2)–(3.3)。对于足够大的常数 η, τ, π, ω，且当 $K \asymp \log n$ 时，由算法 $\hat{}\mathbf{v}^{(K)}$ 得到的估计量满足：对所有足够大的 n，

$$d(\mathbf{v}, \hat{\mathbf{v}}^{(K)}) \lesssim \Big(\Big(\frac{m_G}{\alpha_n}\Big)^r \wedge G\Big)\alpha_n^2 + \Big(\Big(\frac{m_{(|\mathcal{G}|)}}{\beta_n}\Big)^r \wedge T\Big(\frac{m_G}{\alpha_n}\Big)^r\Big)\beta_n^2 + \frac{\lambda^4+\lambda^2+1}{\lambda^4}\,\frac{\log G}{n}, \tag{3.6}$$

其概率至少为 $1 - C_0 G^{-2}$，其中 $C_0 > 0$ 为常数。特别地，在上述条件下，右侧为 $o(1)$，因此 $\hat{}\mathbf{v}^{(K)}$ 是一致的。

误差项的解释。界（3.6）包含三项，各自有不同的含义。第一项是（组级）非参数误差，源于忽略 $\mathcal{G}$ 之外的弱组，并在有效模型规模为 $|\mathcal{G}|$量级的模型上进行估计。第二项是（组内）非参数误差，源于忽略 $\Psi(\mathcal{G})$ 内部的弱坐标，并在大小为 $|\mathcal{S}|$的有效坐标集上进行估计。第三项是参数项，反映了将主尖峰与主体谱区分开来的难度；即使不存在稀疏性，这一项也会出现（Paul和Johnstone，2012）。

注3.2（化归为个体稀疏情形）。考虑特殊情形 $T=1$ 且 $G=p$，即每个组都只包含一个坐标。此时（3.2）化为对$\mathbf{v}$中各元素的弱-ℓ_r条件，而组内稀疏条件则变得多余。在这种情况下，我们的预言机阈值满足

$$\alpha_n = \alpha\Big(\frac{\log p}{n\,h(\lambda^2)}\Big)^{1/2}, \qquad \beta_n \equiv 0,$$

和算法3退化为*Ma (2013)*中的对角阈值初始化。此外，由于 $h(\lambda^2)$ 在 λ 固定时为常数，因此所得速率与（*Ma，2013*年）中的经典稀疏主成分分析速率仅相差常数因子。

备注 *3.3*（与基于半定规划的方法的比较）。利用不等式 $\|v_1 - v_2\|_2^2 \le d(v_1,v_2) \le 2\|v_1 - v_2\|_2^2$，我们可以将结果与以 ℓ_2误差表示的速率进行比较。在 ℓ_0 型同时稀疏性约束下，*Xiao* 和 *Xiao*（*2024*）得到的 $\|\hat{v} - v\|_2^2$ 速率阶为 $\log(p)/n$ （忽略对数因子）。相比之下，在层次弱-ℓ_r结构下，*(3.6)*中的前两项的量级为

$$\Big(\frac{\log G}{n\,h(\lambda^2)}\Big)^{1-\frac{r}{2}} \quad \textit{and} \quad \Big(\frac{\log(T|\mathcal{G}|)}{n\,h(\lambda^2)}\Big)^{1-\frac{r}{2}}$$

当层次稀疏性较强时（例如 r 和 $|\mathcal{G}| \ll G$ 较小），该值可以严格小于 $\log(p)/n$ 。这体现了利用组级和组内稀疏性所带来的统计优势。

4. 模拟研究。我们考察所提出方法（SGPCA）的有限样本性能，并将其与两种广泛使用的基线方法进行比较。第一种基线方法是标准主成分分析（PCA），它忽略稀疏性，并且已知在许多高维情形下表现不佳。第二种基线方法是稀疏主成分（SPC； Witten、Tibshirani 和 Hastie（2009））， 通过PMA 软件包实现。SPC 已被广泛应用于高维基因表达分析；例如，Jerby-Arnon 和 Regev (2022年) 使用 SPC 从单细胞和空间转录组数据中识别多细胞程序。由于其交替方向乘子法（ADMM）求解器的计算开销较大，我们没有在数值研究中纳入稀疏整合主成分分析（SIPCA； Xiao 和 Xiao 2024）；具体而言，Fantope投影会产生 $O(p^3)$ 开销，使 SIPCA 在基因表达规模下不切实际。正如 Xiao 和 Xiao (2024年) 所报告，对于 $n=100$ 和 $p=350$，估计单个主成分每次运行大约需要 30 分钟。

模拟设计。我们考虑三种设置。在每种设置中，我们从 $\mathcal{N}(\mathbf{0},\Sigma)$生成 $n=100$ 独立同分布样本，其中 Σ如下所述。

设置 *1*（组大小变化；组 + 个体稀疏性）。我们设定 $p = G \times T$，其中有 $G = 300$ 个组，每组大小为 $T \in \{3,5,10\}$。协方差为 $\Sigma = 5\,\mathbf{v}\mathbf{v}^\top + I_p$，其中 $\mathbf{v}$ 在组和坐标层面均是稀疏的：只有前1% 的组是活跃组，并且在每个活跃组内，80% 的坐标非零。非零项为独立同分布的拉ademacher变量。

设置 *2*（不同的稀疏结构）。我们固定 $G=300$ 和 $T=10$，并再次使用 $\Sigma = 5\mathbf{v}\mathbf{v}^\top + I_p$。我们比较两种稀疏度情形：(i) 仅组稀疏，即活跃组中的所有坐标均非零；以及 (ii) 组 + 个体稀疏性，即活跃组中80% 的坐标非零。非零项为独立同分布的拉ademacher变量。

设置 *3*（多个主成分）。我们模拟一个多尖峰协方差，

$$\Sigma = 20\,\mathbf{v}_1\mathbf{v}_1^\top + 10\,\mathbf{v}_2\mathbf{v}_2^\top + 5\,\mathbf{v}_3\mathbf{v}_3^\top + I_p,$$

其中，每个$\mathbf{v}_j$ 都遵循与设置 2(ii)中相同的层次稀疏模式。$\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3$的支撑集彼此不重叠，非零元素则独立地从拉ademacher分布中抽样。

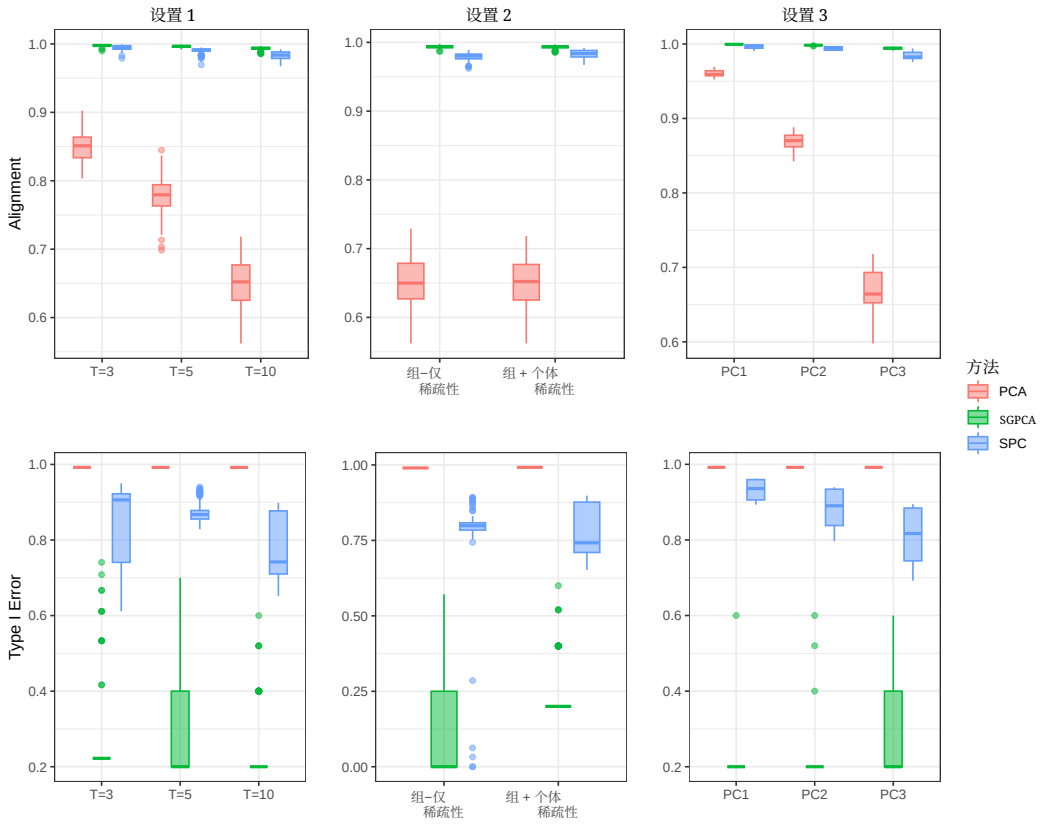

FIG 2。三种设置下的模拟结果。顶部：对齐度。底部：第一类错误。

评估指标和结果。。图2 总结了性能表现。上排报告估计主成分与真实主成分之间的对齐度（估计准确度），下排则报告第一类错误（假阳性率；信号检测准确度）。在所有设置中，SGPCA 始终实现较高的对齐度，表明它能准确恢复潜在的稀疏主成分。在设置 1 中，随着组大小增加，SGPCA 仍保持稳定；而当 C 增大时，SPC 和主成分分析的性能则显著下降——这一趋势与已知的主成分分析在高维设置中不一致的结论相符（Johnstone 和 Lu，2009年）。在设置2和设置3中，无论在两种稀疏度情形下，还是在多个尖峰下，SGPCA 都取得了最佳估计性能。

在信号检测方面，SGPCA 在每种设置下也都具有最低的第一类错误。相比之下，SPC 和主成分分析的第一类错误都接近 1，表明它们保留了许多噪声坐标，且无法控制假阳性。第二类错误见附录 A；SGPCA 的第二类错误与主成分分析和 SPC 相当，这表明其对假阳性的改进控制并不是以漏检真实信号为代价的。

计算..... 运行时间见附录 A。尽管 SGPCA 比主成分分析和 SPC 慢，但仍适用于高维问题。此外，报告的时间包含基于SPC的初始化；实际上，后续的双重阈值迭代所需时间比总实际运行时间所暗示的更短。总体而言，SGPCA 的计算效率显著高于基于交替方向乘子法的 SIPCA 求解器（Xiao 和 Xiao，2024年），同时在恢复组稀疏模式和个体稀疏模式方面表现明显更优。

最后，针对不同组数 G 和组大小 T 开展的其他实验进一步证实，SGPCA 的表现始终优于 SPC 和主成分分析（附录 A）。

5. 狼疮数据示例。 我们分析系统性红斑狼疮（SLE）的单细胞RNA测序数据集。系统性红斑狼疮在女性以及亚裔、非裔和西裔血统人群中的患病率较高（Perez等人，2022年）。在4种主要免疫细胞亚型中均可获得基因表达谱：CD4+ T（T4细胞）、CD8+ T（T8细胞）、经典单核细胞（CM）和自然杀伤细胞（NK）。因此，我们主要关注发现与系统性红斑狼疮相关、存在于这4种细胞亚型中的多细胞程序。该数据集包括158名狼疮患者和98名未受影响的受试者，使我们能够检验估计得到的多细胞程序在狼疮患者与正常组群体中是否存在差异表达。

从数学上讲，发现多细胞程序等价于根据$\mathbf{v}_1, \cdots, \mathbf{v}_J$'的协方差矩阵$\mathbf{X}$估计主成分，其中 J 是由数据本身决定的估计多细胞程序数量。第 j个多细胞程序的差异表达检验表述如下：

$$H_0(\mathbf{v}_j) : (\mu_{\text{lupus}} - \mu_{\text{normal}})^T \mathbf{v}_j = 0 \quad \text{v.s.} \quad H_1(\mathbf{v}_j) : (\mu_{\text{lupus}} - \mu_{\text{normal}})^T \mathbf{v}_j \neq 0, \tag{5.1}$$

其中 μ_{lupus} 和 μ_{normal} 分别表示 Lupus 受试者和正常受试者的平均基因表达水平。量 $(\mu_{\text{lupus},i} - \mu_{\text{normal},i})^T \mathbf{v}_j$ 实际上是 Lupus 受试者与正常受试者在第 j 个主成分（或主成分得分）上的投影均值差。 H_0 表示 Lupus 受试者和正常受试者的平均基因表达水平在第 j 个多细胞程序上没有差异，而H_1 则表示两个群体在该多细胞程序上存在差异表达。我们提出的 SGPCA 旨在准确估计 $\mathbf{v}_j$，而提出的去偏检验程序 Zhang、Lei和Roeder (2024年) 旨在解决双重使用数据的问题。接下来，我们将 SGPCA 与去偏检验程序相结合，以检测 Lupus 受试者与正常受试者之间差异表达的多细胞程序。

在估计多细胞程序并检验差异表达之前，我们使用 Python 软件包 `scanpy`对数据进行预处理。对于每种细胞类型，我们选取 2,000 个高变异基因，并进一步汇总同一受试者、同一细胞亚型中的细胞数据，得到每个基因的伪批量计数，然后剔除在少于 10 名受试者中表达的基因。接下来，我们应用标准变换，利用大小因子将原始表达计数转换为对数归一化值（例如，Ahlmann-Eltze和Huber 2023年中的公式 (2)）；此变换可稳定方差，将高度偏斜、重尾的分布转换为更接近正态分布的形式。此外，我们对文库大小、性别、群体和处理批次进行回归，以消除潜在混杂因素。最后，我们按列拼接各细胞亚型矩阵，作为算法 1的输入。预处理后的数据包含四种细胞亚型中的 3,667 个不同基因。在这 3,667 个基因中，95 个基因在所有四种细胞亚型中均有观测，而大多数基因仅在一种细胞亚型中有记录。

5.1. **主成分估计。** 第一步，我们使用 SGPCA，从处理后的伪批量基因表达矩阵中估计主成分。根据超参数调优中对齐分数与平均支撑集大小之间关系所呈现的图形模式（见附录 B 中的图**??** ），我们识别出 $J = 4$ 个显著偏离噪声的主成分。在估计的主成分中，每个主成分检测到的非零项数量各不相同，代表共表达基因。值得注意的是，许多基因在多个甚至所有细胞亚型中共表达。具体而言，四个主成分中，分别有 15、32、18 和 45 个基因在四种免疫细胞亚型中共表达（完整列表见附录 B 中的表 6）。此外，还有其他基因在两种或三种细胞亚型中共表达，这凸显了在多细胞程序的估计过程中同时纳入组稀疏性和个体稀疏性的必要性。

第一主成分（PC1）主要由 CD8+T 细胞的基因表达驱动，表明它们对样本间转录组变异具有主要贡献。

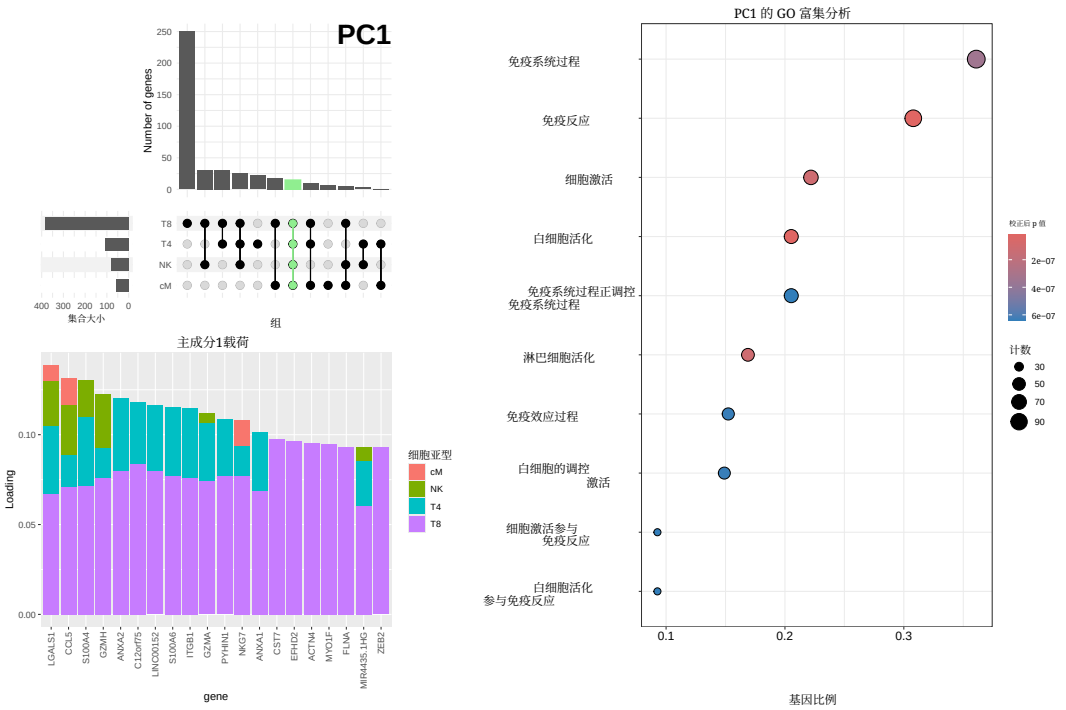

图3. *PC1* 的估计结果。左上面板：一个 *UpSet* 图，展示了不同细胞亚型之间共同表达的基因数量。左下面板：第一主成分中各细胞亚型载荷绝对值最大的前*20*个基因。右侧面板：第一主成分中富集程度最高的前*10*个生物过程。

PC1 上载荷较高的基因包括 LGALS1、CCL5、GZMH 和 NKG7 等关键免疫调节因子及细胞毒性效应因子，表明效应 T 细胞程序发挥了重要作用。对 PC1 相关基因进行 GO 富集分析发现，与白细胞和淋巴细胞活化、免疫效应过程及免疫系统调节相关的条目显著富集。附录 B 提供了另外三个多细胞程序的分析和可视化结果。

5.2. 差异表达检验。 给定四种细胞亚型的估计主成分，我们希望了解与正常受试者相比，狼疮患者是否对这些多细胞程序具有上调或下调的表 达。具体而言，我们对 H_0 $(\hat{}\mathbf{v}j)$ 相对于 $H_1(\hat{}\mathbf{v}j)$ 进行假设检验，其中 $j = 1, \cdots, 4$，其中 H_0 和 H_1 的定义见（5.1）。结果显示， H_0 PC1和PC4的检验在0.05阈值下被拒绝，这表明狼疮患者与正常受试者之间的基因表达谱沿PC1和PC4方向存在显著差异。在图 4中，分别针对每个主成分可视化了狼疮患者和正常受试者的主成分得分，其中PC1和PC4得分显示出明显的均值差异，而PC2和PC3得分的分布相似。回顾可知，PC1主要由$CD8^+$ T细胞中的基因表达主导，这与（Perez等人，2022年）的结果一致，即$CD8^+$ 细胞在狼疮患者中存在功能失调。

此外，我们将每个主成分中识别出的基因与通过修正的广义线性模型（GCATE（Du、Wasserman 和 Roeder，2025年））发现的差异表达基因（DE）进行比较。从数学上看，去偏检验表明，对于 PC1 和 PC4，$(\mu_{\text{lupus}} - \mu_{\text{normal}})^T\mathbf{v} \neq 0$，而 GCATE 识别的是均值不同的基因，即 $I = \{i\colon \mu_{\text{lupus},\, i} \neq \mu_{\text{normal},\, i}\}$。因此，这两种检验策略的目标存在根本差异。

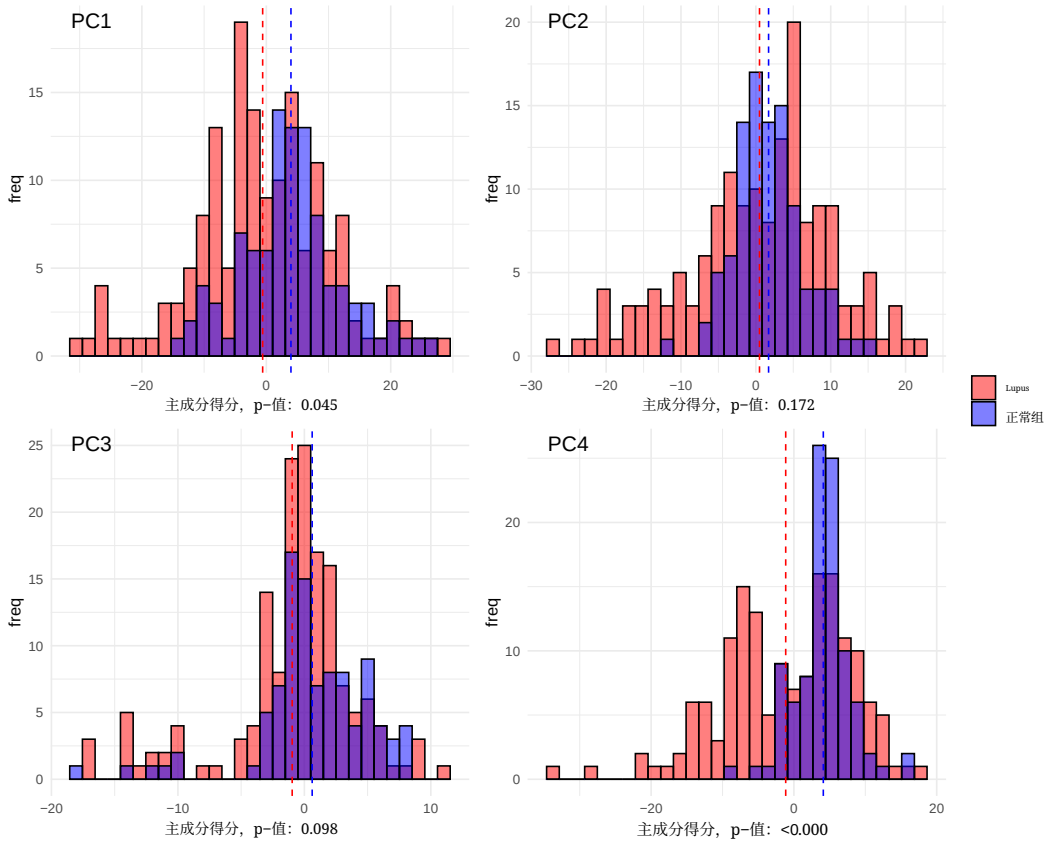

图4. *4* 个已识别主成分的得分直方图。垂直线表示各组的得分均值。*p*值由主成分得分的去偏检验得出。

投影检验旨在找出一组协同作用、在病例组和对照组受试者之间表达存在差异的基因。边际检验旨在识别表达存在差异的单个基因。对于后者，基因在队列之间必须存在显著差异才能达到统计显著性；而 MCP 中许多共表达基因的差异可能微乎其微，但这些效应在整个程序中累积起来便可能产生影响。值得注意的是，MCP 中也达到边际显著性的基因。这些基因可能是 MCP 的关键驱动因子，决定了病例组和对照组受试者之间的差异。因此，我们在 PC1 和 PC4 中识别出这一小组 I，由此得到 $(\mu_{\text{lupus},\ I} - \mu_{\text{normal},\ I})^T \mathbf{v}_I \neq 0$。在图 5中，我们按细胞亚型比较了每个主成分中的非零载荷与差异表达基因。PC1、PC4 与差异表达基因的交集基因列于表1。

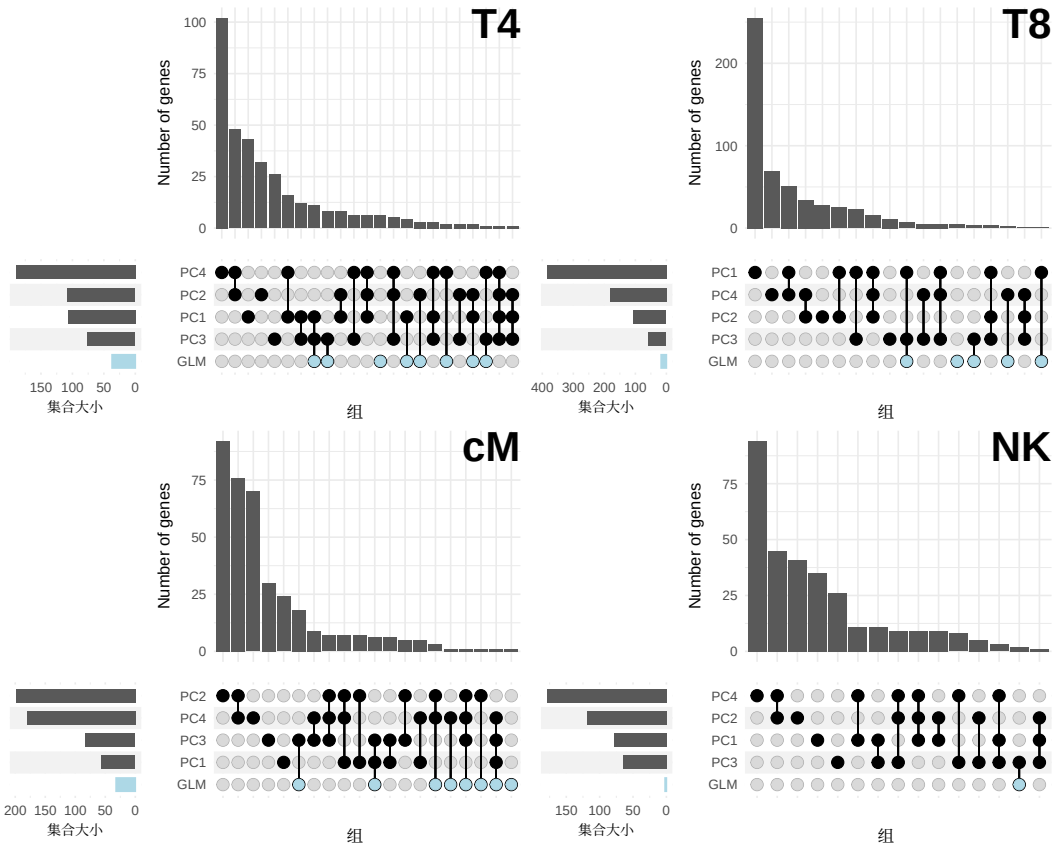

FIG 5。各细胞亚型中每个主成分的鉴定基因与*GCATE*方法鉴定的差异表达基因的*UpSet*图。

表1

细胞亚型	PC1	PC4
T4	*TNFRSF4* ，*STAT1* ，*SPATS2L* ，*TIGIT* ，*TMEM244* ，*FOXP3* ，*DUSP4* ，[illegible] ，*IFIT3* ，*LAG3*，[illegible] ，*IFI27* ，膜联蛋白*A2* ，*AC002331.1* ，*PMAIP1* ，*CD70*，*IGFL2*	*RTKN2* ，*ACVR2B* ，*PRDM6*
T8	*STAT1* ，*IFIT3* ，*LAG3*，*OASL*，*EPSTI1* ，*IFI27* ，*PATL2* ，*MT1E*	*SLC4A10* ，*PRDM6*
CM	*GBP1*，*STAT1* ，*PLAC8* ，*IFIT2* ，*IFIT3* ，*EPSTI1* ，*IFI27*	*IFIT2* ，*TNFSF10* ，*AREG*，*GABRE* ，*EPPK1* ，*NBPF15*
NK	-	-

6. 讨论。本文提出了*SGPCA*，一种用于估计主成分的全新方法，可处理分层 组级和条目级稀疏性。通过明确利用这两个层面的稀疏性，SGPCA 提高了估计准确度和信号检测能力，模拟研究也证实了这一点。在理论方面，我们证明了其一致性，并在高维渐近情形下推导出更快的收敛速率。

其核心直觉很简单：当载荷向量呈现层次结构时，分两个阶段去除噪声在统计上是高效的。第一步（组）筛选在组内聚合

信息，并以与 $\log(G)$ 成正比的多重检验复杂度，过滤掉整个噪声组。在我们的增长条件 $T = o(\log G)$ 下，这一步的代价很小，却能显著降低环境维数。随后，第二步（组内）筛选在保留组中进一步消除残余噪声坐标，其复杂度为 $\log(T|\mathcal{G}|)$，而不是更为保守的 $\log(p)$。正是这种分离带来了误差界中改进的非参数分量。

从计算角度来看，双重阈值化细化的每次迭代仅需要矩阵-向量乘法和阈值操作，因此计算成本为 $O(np)$ 时间复杂度。这比基于半正定松弛的方法便宜得多，因为后者的核心子程序（例如 Fantope 投影）的复杂度可达 $O(p^3)$ （Xiao 和 Xiao, 2024年）。在实践中，SGPCA 的主要计算成本来自超参数选择，因为每个主成分都涉及两个阈值水平。当不同主成分之间的稀疏性模式相似时，如第2.2节所述，这一成本可以降低。总体而言，SGPCA 提供了一种具有统计学原理依据且计算高效的方法，并表现出良好的经验性能。我们还在 R 包 `SGPCA`中提供了高效实现。

在我们的应用中，我们使用 SGPCA 在一项狼疮研究中识别多细胞程序（多细胞程序），随后进行下游的基于投影的差异表达检验，以检测系统性红斑狼疮患者与健康对照组之间存在差异的多细胞程序。所得基因集及富集的基因本体术语与涉及系统性红斑狼疮的具有生物学意义的通路相一致。更广泛地说，本文所考虑的同时稀疏结构并非基因表达所特有。例如，在在线零售数据中，数百万种产品被组织到不同类别中（例如电子产品、服装和食品杂货）。将客户视为独立样本时，购买行为通常在两个层面上都具有稀疏性：大多数客户只与少数类别发生互动，而在这些类别中，他们也只涉及少数产品。这会自然地在潜在因子中诱导出组稀疏性和元素级稀疏性。因此，我们预计，在其他具有有意义组结构的领域中，SGPCA 也将具有广泛的实用价值。

致谢。 第一作者感谢 Jin-Hong Du 对狼疮数据集进行预处理，并感谢 Tianyu Zhang 提供宝贵的讨论意见。

经费支持。 本项目由美国国家精神卫生研究所（NIMH）项目资助 R01MH123184、美国国家科学基金会（NSF）项目资助 DMS-2310764 和美国国家科学基金会项目资助 DMS-2515687 提供经费支持。

补充材料

附录 A：更多模拟详情
更多模拟结果：模拟设置 4 下的第二类错误和运行时间，以及其他模拟设置。

附录 B：真实数据分析详情
真实数据分析的调参流程及所提取主成分 2 至主成分 4 的解读。

附录 C–E：理论分析的技术细节第3节中理论结果的技术工具与证明细节。

参考文献

ABDELAAL, T., MICHIELSEN, L., CATS, D., HOOGDUIN, D., MEI, H., REINDERS, M. J. and MAHFOUZ, A.（2019）。单细胞RNA测序数据自动细胞鉴定方法的比较。基因组生物学**201**–19。
AHLMANN-ELTZE, C. and HUBER, W.（2023）。单细胞RNA测序数据转换方法的比较。自然方法 **20** 665–672。CAI, T. T., ZHANG, A. R. and ZHOU, Y.（2022）。稀疏组套索：最优样本复杂度、收敛速率与统计推断。《*IEEE*信息论汇刊》**68**5975–6002。

D’ASPREMONT, A., GHAOUI, L., JORDAN, M. and LANCKRIET, G.（2004）。使用半定规划求解稀疏主成分分析的直接表述。神经信息处理系统进展 17。DU, J.-H., WASSERMAN, L. and ROEDER, K.（2025）。含未测量混杂因素的广义线性模型的同时推断。《美国统计学会会刊》。https://doi.org/10.1080/01621459. 2025.2485379 FAN, J. and LI, R.（2001年）。通过非凹惩罚似然进行变量选择及其轨道性质。《美国统计学会会刊》96 1348–1360。FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R.（2010年）。关于组套索和稀疏组套索的说明。arXiv预印本 arXiv:1001.0736。GAYNANOVA, I. and LI, G.（2019年）。多视图数据的结构学习与整合分解。《生物计量学》75 1121–1132。GOLUB, G. H. and VAN LOAN, C. F.（2013）。矩阵计算。约翰斯·霍普金斯大学出版社。JERBY-ARNON, L. and REGEV, A.（2022年）。DIALOGUE利用单细胞或空间转录组数据描绘组织中的多细胞程序。《自然·生物技术》40 1467–1477。JI, X., TSAO, D., BAI, K., TSAO, M., XING, L. and ZHANG, X.（2023）。scAnnotate：一种用于单细胞RNA测序数据的自动细胞类型注释工具。《生物信息学进展》3 vbad030。JOHNSTONE, I. M.（2001年）。主成分分析中最大特征值的分布。《统计学年刊》29 295–327。JOHNSTONE, I. M. and LU, A. Y.（2009年）。高维主成分分析的一致性与稀疏性。美国统计学会会刊 104 682–693。JOLLIFFE, I. T.（2002年）。特殊类型数据的主成分分析。施普林格。JOLLIFFE, I. T., TRENDAFILOV, N. T. and UDDIN, M.（2003）。一种基于LASSO的改进主成分技术。《计算与图形统计学杂志》12 531–547。LING, E., NEMESH, J., GOLDMAN, M., KAMITAKI, N., REED, N., HANDSAKER, R. E., GENOVESE, G., VOGELGSANG, J. S., GERGES, S., KASHIN, S. et al.（2024年）。一种协调的神经元-星形胶质细胞程序在衰老和精神分裂症中减弱。《自然》627 604–611。LOCK, E. F., PARK, J. Y. and HOADLEY, K. A.（2022年）。用于泛组学泛癌分析的双维链接矩阵分解。《应用统计学年刊》16 193。LOCK, E. F., HOADLEY, K. A., MARRON, J. S. and NOBEL, A. B.（2013）。用于多种数据类型整合分析的联合与个体变异解释（JIVE）。《应用统计学年刊》7 523。MA, Z.（2013）。稀疏主成分分析与迭代阈值化。MEINSHAUSEN, N. and BÜHLMANN, P.（2010年）。稳定性选择。《英国皇家统计学会杂志B辑：统计方法论》72 417–473。MITCHEL, J., GORDON, M. G., PEREZ, R. K., BIEDERSTEDT, E., BUENO, R., YE, C. J. and KHARCHENKO, P. V.（2024年）。张量分解揭示了将患者队列分层的协调性多细胞转录变异模式。《自然·生物技术》1–10。PAUL, D. and JOHNSTONE, I. M.（2012）。用于高维数据的增广稀疏主成分分析。arXiv预印本 arXiv:1202.1242。PEREZ, R. K., GORDON, M. G., SUBRAMANIAM, M., KIM, M. C., HARTOULAROS, G. C., TARG, S., SUN, Y., OGORODNIKOV, A., BUENO, R., LU, A. et al.（2022）。单细胞RNA测序揭示了狼疮中细胞类型特异性的分子和遗传关联。《科学》376 eabf1970。SHEN, H. and HUANG, J. Z.（2008）。通过正则化低秩矩阵近似实现稀疏主成分分析。《多元分析杂志》99 1015–1034。SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R.（2013）。稀疏组套索。《计算与图形统计学杂志》22 231–245。STEGLE, O., PARTS, L., PIIPARI, M., WINN, J. and DURBIN, R.（2012）。利用表达残差的概率估计（PEER）提升基因表达分析的效能和可解释性。《自然方法》7 500–507。SUN, W., WANG, J. and FANG, Y.（2013）。通过变量选择稳定性一致地选择调节参数。《机器学习研究杂志》14 3419–3440。TANG, T. M. and ALLEN, G. I.（2021）。整合主成分分析。《机器学习研究杂志》22 1–71。VU, V. Q., CHO, J., LEI, J. and ROHE, K.（2013）。Fantope投影与选择：稀疏主成分分析的一种近乎最优的凸松弛。神经信息处理系统进展 26。WAMSLEY, B., BICKS, L., CHENG, Y., KAWAGUCHI, R., QUINTERO, D., MARGOLIS, M., GRUNDMAN, J., LIU, J., XIAO, S., HAWKEN, N., MAZARIEGOS, S. and GESCHWIND, D. H.（2024）。单细胞基因组学揭示了自闭症谱系障碍中的分子级联和细胞类型特异性特征。《科学》384 eadh2602。https: // doi.org/10.1126/science.adh2602 WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T.（2009）。一种惩罚矩阵分解及其在稀疏主成分和典型相关分析中的应用。《生物统计学》10 515–534。

XIAO, L. 和 XIAO, L.（2024年）。面向多视图数据的稀疏整合主成分分析。《电子统计学杂志》**18** 3774–3824。XU, Q. 和 QU, A.（2025年）。异构数据整合的表征检索学习。*arXiv* 预印本 *arXiv:2503.09494*。ZHANG, T.、LEI, J. 和 ROEDER, K.（2024年）。单细胞表达数据的去偏投影双样本比较。*arXiv* 预印本 *arXiv:2403.05679*。ZHANG, A. 和 XIA, D.（2018年）。张量奇异值分解：统计与计算极限。《*IEEE*信息论汇刊》**64** 7311–7338。ZOU, H.、HASTIE, T. 和 TIBSHIRANI, R.（2006）。稀疏主成分分析。《计算与图形统计学杂志》**15** 265–286。ZOU, H. 和 XUE, L.（2018年）。稀疏主成分分析的选择性综述。《*IEEE*会议录》**106** 1311–1320。