中国科学: 数学 2020 年 第 50 卷 第 4 期: 447～490
SCIENTIA SINICA Mathematica
创刊 70 周年特邀综述

高维因子模型及其在统计机器学习中的应用

陈钊[1†], 范剑青[2*], 王丹[3†]

1. 复旦大学大数据学院, 上海 200433;
2. Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA;
3. 上海纽约大学商学部, 上海 200122

E-mail: zchen_fdu@fudan.edu.cn, jqfan@princeton.edu, christina.wang@nyu.edu

收稿日期: 2020-02-16; 接受日期: 2020-03-18; 网络出版日期: 2020-03-28; * 通信作者
国家自然科学基金 (批准号: 11690014, 11690015, 71991471, 71991475, U1811461 和 11901395) 和上海浦江计划 (批准号: 19PJ1408200 和 19PJ1400900) 资助项目

摘要 本文综述近年来因子模型研究的最新进展及其在统计机器学习中的应用. 因子模型通过较少的因子实现降维, 并为协方差矩阵提供了一种低秩加稀疏的结构, 不仅受到高维数据分析领域的关注, 也被广泛应用于计量经济学、数量金融学、基因组学、神经科学和图像处理等许多科学、工程及人文社科领域的研究中. 本文系统阐述利用主成分分析方法提取潜在因子、估计因子载荷、异质结构与整体协方差矩阵的统计推断方法, 这套方法被证明可以有效应对当前大数据所表现出的高维性、强相关性、厚尾性和异质性等重大挑战; 另外, 还重点介绍了高维因子模型在处理协方差矩阵估计、模型选择和多重检验等高维统计学习问题中的作用; 最后, 通过几个应用实例说明因子模型与现代机器学习问题之间的密切联系, 其中包括当下流行的网络分析和低秩矩阵还原等.

关键词 因子模型 主成分分析 结构化协方差矩阵 因子调节 模型选择 多重检验

MSC (2010) 主题分类 62H25, 68T05

1 引言

互联网的广泛普及与信息技术的快速发展将人们带入了大数据时代, 数据所展现出的规模与复杂性都是前所未见的, 使得对智能化分析方法的需求成为时代呼声. 这对以数据作为主要研究对象的统计学提出了全新的要求与巨大的挑战. 创新的数据思维、新颖的统计方法、高效的计算技术和深刻的数学理论都是必不可少的因素.

伴随大数据而来的是激增的变量维数, 高维化是当代工程、科学和人文社科等领域统计问题的主要特征. 例如, 在使用微阵列 (microarray) 或蛋白质组学数据进行疾病分类时, 成千上万的分子或蛋

† 陈钊和王丹同为第一作者, 作者顺序按照姓氏拼音排序.

英文引用格式: Chen Z, Fan J Q, Wang C D. High-dimensional factor model and its applications to statistical machine learning (in Chinese). Sci Sin Math, 2020, 50: 447–490, doi: 10.1360/SSM-2020-0041

白质表达是潜在的预测因子; 在全基因组关联分析 (genome-wide association study, GWAS) 中, 数以百万个单核苷酸多态性 (single nucleotide polymorphisms, SNPs) 是潜在的协变量; 在机器学习中, 数千万甚至数十亿的特征从文档、图像及其他素材中提取出来.

高维变量之间通常具有很强的相依性, 同一行业的股票表现出显著的相关收益, 基因表达经常受到细胞因子的刺激或受到生物过程的调控等. 忽视这种相依结构, 使得高维统计推断方法产生明显的系统性偏差, 导致效率降低. 以正则回归方法为例, 其模型选择相合性要求协变量满足不可表示条件 (irrepresentable condition [1–3]). 然而, 当所有变量由某种共同因素驱动时, 这个条件不再成立. 如何准确地刻画这种相依关系成为近期高维统计推断研究的热门方向.

因子模型常用来对变量之间的相依关系进行建模, 对于高维数据, 因子模型可以通过几个 "因子" 来捕捉整个变量间的相依结构 [4, 5]. 通常因子的个数会远少于变量的维数, 这不仅实现了变量降维, 还为高维协方差矩阵引入了 "低秩" 加 "稀疏" 的结构. 文献 [6, 7] 指出高维统计学习的主要目标是, 首先, 根据数据构建有效的预测方法; 然后, 深入了解响应变量与特征之间的科学关系; 最后, 再改进预测方法. 因子模型能够很好地实现这一目标, 强调预测结果的同时重视背后蕴含的科学原理, 有效降低维数建立更稳定的模型.

因子模型最早源于关于人类能力的测量研究 [8], 现在已经成为多元分析中最流行和最强大的工具之一, 并在过去一个世纪里对心理学 [9]、经济金融学 [5, 10–12] 和生物学 [13–15] 等学科产生了深远的影响. 高维因子模型 (即 $p \geqslant n$) 在计量经济学和统计学文献中得到了广泛的研究. 计量经济学主要聚焦在估计潜因子和因子载荷, 主要工作包括: 文献 [5, 16, 17] 估计了潜在因子并将其应用于预测宏观经济变量, 文献 [18] 建立了潜在因子、因子载荷和二者乘积估计的渐近正态性, 文献 [19] 进一步研究了极大似然估计, 文献 [20] 估计了弱因子模型的因子, 文献 [21] 研究了近似因子模型的惩罚似然方法, 文献 [22, 23] 考虑了半参数因子模型. 统计学更关注预测变量或其异质成分的协方差矩阵和精度矩阵估计.

本文综述近年来高维因子模型研究领域的代表性成果, 系统地介绍高维因子模型的理论及方法, 列举其在统计机器学习中的典型应用, 以期读者能够对这一热门研究方向有所了解. 在撰写过程中, 我们参照和翻译了文献 [24, 第 9–11 章] 的部分内容, 并参考文献 [25] 的综述文章. 第 2 节介绍高维因子模型, 并阐述主成分分析与高维因子模型的关系, 及其在提取潜在因子中的关键作用; 第 3 节介绍因子模型下高维协方差矩阵的估计理论和方法, 其中还包括了潜在因子、因子负荷, 以及异质项协方差矩阵的估计及理论结果, 还讨论了稳健初始协方差估计及因子个数选择这两个高维环境下的专题; 第 4 节讨论利用数据的因子结构解决许多统计学习中的问题, 展示如何将因子模型和主成分分析应用于高维回归、多重检验和模型选择; 第 5 节演示谱方法 (主成分分析的一般化方法) 与机器学习问题之间的联系, 包括高斯混合模型、社群发现和矩阵填补等内容. 另外, 我们特意在某些章节后增加了文献小结, 这样既可以方便读者查阅文献了解相关研究内容, 同时也不会使本文结构显得过于松散拖沓.

为叙述方便, 我们对文中常用记号做一点说明. $\boldsymbol{I}_p$ 记作 p 维单位阵, $\mathbf{1}_p = (1, \ldots, 1)^{\mathrm{T}}$, $I(\cdot)$ 为示性函数. 假设 $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^P$, 定义 ℓ_q $(1 \leqslant q < \infty)$ 范数为 $\|\boldsymbol{x}\|_q = (\sum_{j=1}^{p} |x_j|^q)^{1/q}$. 特别定义 ℓ_0 范数为向量 $\boldsymbol{x}$ 中非零元素的个数, 即 $\|\boldsymbol{x}\|_0 = |\{j : x_j \neq 0\}|$, 以及 ℓ_∞ 范数为 $\|\boldsymbol{x}\|_\infty = \max_{1 \leqslant j \leqslant p} |x_j|$. 对于矩阵 $\boldsymbol{M} \in \mathbb{R}^{m \times n}$, 分别定义谱范数 (或者算子范数、二范数) $\|\boldsymbol{M}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})}$, 即 $\boldsymbol{M}$ 的 Gram 矩阵的最大特征值平方根; 核范数 $\|\boldsymbol{M}\|_* = \mathrm{tr}(\sqrt{\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}})$; Frobenius 范数 $\|\boldsymbol{M}\|_F = \sqrt{\sum_{i,j} m_{ij}^2} = \sqrt{\mathrm{tr}(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})}$; 最大范数 $\|\boldsymbol{M}\|_{\max} = \max_{i,j} |m_{ij}|$; 向量 ℓ_1 范数 $\|\boldsymbol{M}\|_1 = \sum_{i,j} |m_{ij}|$. 当 $\boldsymbol{M}$ 为方阵时, 令

$|\boldsymbol{M}|$ 为其行列式. 对于任意两个非负实数 a 和 b, 若存在常数 $C_1>0$ 使得 $a\leqslant C_1b$, 则记 $a=O(b)$ 或者 $a\lesssim b$; 若存在常数 $C_2>0$ 使得 $a\geqslant C_2b$, 则记 $a=\Omega(b)$ 或者 $a\gtrsim b$; 若 $a=O(b)$ 和 $a=\Omega(b)$ 同时成立, 则记 $a\asymp b$. 对于随机变量序列 $\{X_n\}_{n=1}^{\infty}$ 和一组给定的非负常数序列 $\{a_n\}_{n=1}^{\infty}$, 若存在常数 $N>0$ 和 $C>0$, 当 $n>N$ 时, 对任意的 $\varepsilon>0$, 有 $\mathrm{P}(|X_n|\geqslant Ca_n)\leqslant\varepsilon$, 则记 $X_n=O_P(a_n)$; 若对任意的 $\varepsilon>0$, $C>0$, 都存在常数 $N>0$, 当 $n>N$ 时, 使得 $\mathrm{P}(|X_n|\geqslant Ca_n)\leqslant\varepsilon$, 则记 $X_n=o_P(a_n)$. 记 $[p]=\{1,\dots,p\}$ 为一个指标集.

2 高维因子模型与因子提取

2.1 因子模型

因子模型是刻画多个变量之间公共相依关系的一种强有力工具, 在统计学、经济学、金融学、社会学、基因组学和计算生物学等领域有着广泛的应用. 下面给出因子模型的定义.

定义 1 设 $\boldsymbol{x}=(X_1,\dots,X_p)^{\mathrm{T}}$ 为 p 维观测值向量, $\boldsymbol{f}=(f_1,\dots,f_K)$ 为 K 维公共因子, 则因子模型满足如下假设:

$$\boldsymbol{x}=\boldsymbol{a}+\boldsymbol{B}\boldsymbol{f}+\boldsymbol{u},\quad \mathrm{E}\,\boldsymbol{u}=\boldsymbol{0},\quad \mathrm{Cov}(\boldsymbol{f},\boldsymbol{u})=\boldsymbol{0},\tag{2.1}$$

或者分量形式

$$X_j=a_j+\boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{f}+u_j=a_j+b_{j1}f_1+\cdots+b_{jK}f_K+u_j,\quad j\in[p].\tag{2.2}$$

模型 (2.1) 中, $\boldsymbol{a}$ 为公共截距项; $\boldsymbol{u}$ 为随机误差项, 也常称作异质成分 (idiosyncratic component), 它与公共因子 $\boldsymbol{f}$ 不相关; $\boldsymbol{B}=\{b_{jk}\}_{j\in[p],k\in[K]}$ 是 $p\times K$ 维因子载荷矩阵. 特别地, 当异质成分的协方差矩阵 $\boldsymbol{\Sigma}_u=\mathrm{Var}(\boldsymbol{u})$ 是一个对角矩阵时, 模型 (2.1) 被称为严格因子模型 (strict factor model); 当 $\boldsymbol{\Sigma}_u=\mathrm{Var}(\boldsymbol{u})$ 是一个稀疏矩阵时, 模型 (2.1) 被称为近似因子模型 (approximate factor model).

分量形式 (2.2) 清楚地表明原来 p 维变量现在由 K 个公共的因子来驱动, 载荷矩阵 $\boldsymbol{B}$ 的元素 b_{jk} 表示第 j 个分量 X_j 对第 k 个因子 f_k 的依赖程度. 特别当因子 $\boldsymbol{f}$ 已知时, 因子模型就是通常的线性回归模型. 因子模型的主要目的是通过很少的几个 “因子” 来准确地刻画多元变量间的相关性. 根据因子模型的定义, 多元变量 $\boldsymbol{x}$ 的协方差阵 $\boldsymbol{\Sigma}=\mathrm{Var}(\boldsymbol{x})$ 具有如下形式:

$$\boldsymbol{\Sigma}=\boldsymbol{B}\boldsymbol{\Sigma}_f\boldsymbol{B}^{\mathrm{T}}+\boldsymbol{\Sigma}_u,\quad \boldsymbol{\Sigma}_f=\mathrm{Var}(\boldsymbol{f}),\quad \boldsymbol{\Sigma}_u=\mathrm{Var}(\boldsymbol{u}),\tag{2.3}$$

其中 $\boldsymbol{\Sigma}_f$ 是一个 $K\times K$ 维矩阵. 因此在高维情形下, 通常假定因子的个数 K 远小于变量的维数 p, 且误差项的协方差阵 $\boldsymbol{\Sigma}_u$ 是稀疏的, 也就是说相关性主要由公共因子来刻画, 则误差项是弱相关的. 然而, 在很多实际问题中, 我们事先并不知道真实的因子和对应的因子载荷矩阵是什么, 只能通过观测数据去探究它们. 这时会遇到一个问题, 任意给定一个可逆矩阵 $\boldsymbol{H}$, 始终有

$$\boldsymbol{x}=\boldsymbol{a}+(\boldsymbol{B}\boldsymbol{H})(\boldsymbol{H}^{-1}\boldsymbol{f})+\boldsymbol{u}.\tag{2.4}$$

也就是说, 我们总可以用 $\boldsymbol{B}\boldsymbol{H}$ 和 $\boldsymbol{H}^{-1}\boldsymbol{f}$ 同时替换掉原来的 $\boldsymbol{B}$ 和 $\boldsymbol{f}$ 而不改变测量值 $\boldsymbol{x}$ 本身, 这就使得模型 (2.1) 存在识别性问题. 为此, 我们取 $\boldsymbol{H}$ 使得 $\mathrm{Var}(\boldsymbol{H}^{-1}\boldsymbol{f})=\boldsymbol{I}_p$, 同时 $\boldsymbol{B}\boldsymbol{H}$ 的各列之间相互正交. 另一方面, 对 (2.1) 取期望得到 $\mathrm{E}\,\boldsymbol{x}=\boldsymbol{a}+\boldsymbol{B}\,\mathrm{E}\,\boldsymbol{f}$, 当 $\boldsymbol{B}$ 和 $\boldsymbol{f}$ 未知时, 截距项 $\boldsymbol{a}$ 和 $\mathrm{E}\,\boldsymbol{f}$ 两者也不可识别. 类似经典线性模型中的中心化方法, 假设 $\mathrm{E}\,\boldsymbol{f}=\boldsymbol{0}$, 则 $\boldsymbol{a}=\mathrm{E}\,\boldsymbol{x}$. 这样就得到了因子模型的可识别性条件.

条件 1 (可识别性) $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$ 是对角矩阵, $\mathrm{E}\,\boldsymbol{f}=\boldsymbol{0}$ 且 $\mathrm{Cov}(\boldsymbol{f})=\boldsymbol{I}_p$.

其他可识别性条件不在本文中过多赘述, 具体内容可以参见文献 [19, 26]. 由可识别性条件 1, 多元变量 $\boldsymbol{x}$ 的协方差结构 (2.3) 变为

$$\boldsymbol{\Sigma}=\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}+\boldsymbol{\Sigma}_u, \tag{2.5}$$

其中 $\mathrm{rank}\,(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}})=K$. 当 $K\ll p$ 时, 近似因子模型导出一种 "低秩"+"稀疏" 的特殊结构. 利用惩罚最小二乘方法 (penalized least squares method) 可以估计这种结构. 记 $\boldsymbol{\Theta}=\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$, $\boldsymbol{\Gamma}=\boldsymbol{\Sigma}_u$, 则

$$(\widehat{\boldsymbol{\Theta}},\widehat{\boldsymbol{\Gamma}})=\mathrm{argmin}\,\|\boldsymbol{S}-\boldsymbol{\Theta}-\boldsymbol{\Gamma}\|_{\mathrm{F}}^2+\lambda\,\|\boldsymbol{\Theta}\|_*+\lambda\nu\sum_{i\neq j}|\gamma_{ij}|\,, \tag{2.6}$$

其中 $\boldsymbol{S}$ 是样本协方差阵, $\boldsymbol{\Theta}$ 和 $\boldsymbol{\Gamma}$ 是半正定阵, λ 和 ν 是调优参数. 核范数 $\|\boldsymbol{\Theta}\|_*$ 用于保证低秩性; 第二个惩罚项则用于保证 $\boldsymbol{\Gamma}=(\gamma_{ij})$ 的稀疏性. (2.6) 也被称为稳健主成分分析 (robust principal component analysis[27, 28]) 方法. 就如同一个实数可以分解为整数和小数部分, 我们根据因子模型将一个矩阵分解为 "低秩" 成分和 "稀疏" 成分.

2.2 主成分分析

主成分分析 (principal component analysis, PCA) 是一种常用的数据分析和降维技术. 主成分分析在数学上描述为一组使原始变量 $\boldsymbol{x}$ 的方差达到最大的正交线性变换. 例如, 在面试中, 对受试者多项能力进行测试后, 作为考官总希望用一个分数来进行排序, 同时这个分数还要具有极高的区分度, 即获得总分

$$Z_1=\sum_{j=1}^{p}\xi_{1j}X_j=\boldsymbol{\xi}_1^{\mathrm{T}}\boldsymbol{x},$$

且对于标准化的 $\|\boldsymbol{\xi}_1\|=1$, Z_1 的数值变化尽可能大. 翻译成数学语言, 这就是通过 $\boldsymbol{\xi}$ 的选择使 Z_1 的方差最大化,

$$\begin{aligned}&\boldsymbol{\xi}_1=\underset{\boldsymbol{\xi}\in\mathbb{R}^p}{\mathrm{argmax}}\,\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\xi},\\&\text{s.t. }\ \|\boldsymbol{\xi}\|=1.\end{aligned} \tag{2.7}$$

接下来, 希望构建一个新的度量 Z_2 来概括应试者不同方面的能力, Z_2 所包含的信息应与 Z_1 无关. 为此, 我们在 (2.7) 中依次加入关于不相关性的约束条件, 就可以序贯地定义出所有的 "主成分". 已知前 k 个主成分 $Z_l=\boldsymbol{\xi}_l^{\mathrm{T}}\boldsymbol{x}$, $1\leqslant l\leqslant k$, 第 $k+1$ 个主成分 $Z_{k+1}=\boldsymbol{\xi}_{k+1}^{\mathrm{T}}\boldsymbol{x}$ 就定义为

$$\begin{aligned}&\boldsymbol{\xi}_{k+1}=\underset{\boldsymbol{\xi}\in\mathbb{R}^p}{\mathrm{argmax}}\,\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\xi},\\&\text{s.t. }\ \|\boldsymbol{\xi}\|=1,\\&\qquad\boldsymbol{\xi}_l^{\mathrm{T}}\boldsymbol{\xi}_k=0,\quad\forall\,1\leqslant l\leqslant k.\end{aligned} \tag{2.8}$$

从几何角度看, 主成分定义了正交约束下最大化方差的投影方向. 假设协方差矩阵 $\boldsymbol{\Sigma}$ 的特征值 $\lambda_1,\ldots,\lambda_p$ 都不相同, 其中 λ_1 为最大 (第一) 特征值, 也是第一主成分 Z_1 的方差, 由 Rayleigh-Ritz 定理知, $\boldsymbol{\xi}_1$ 是 λ_1 对应的特征向量. 由归纳法可证, $\boldsymbol{\xi}_k$ 就是 $\boldsymbol{\Sigma}$ 的第 k 个特征向量. 因此, 我们可以通过协方差矩阵 $\boldsymbol{\Sigma}$ 的谱分解实现主成分分析:

$$\boldsymbol{\Sigma}=\sum_{j=1}^{p}\lambda_j\boldsymbol{\xi}_j\boldsymbol{\xi}_j^{\mathrm{T}}=\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{\mathrm{T}}, \tag{2.9}$$

其中 $\boldsymbol{\Gamma}=(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_p)$, $\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\ldots,\lambda_p)$, $\boldsymbol{\xi}_k$ 被称为第 k 个主成分 $Z_k=\boldsymbol{\xi}_k^{\mathrm{T}}\boldsymbol{x}$ 的载荷. 同时, 还有

$$\mathrm{cov}(Z_j,Z_k)=\boldsymbol{\xi}_j^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\xi}_k=\lambda_k\boldsymbol{\xi}_j^{\mathrm{T}}\boldsymbol{\xi}_k=\begin{cases}0, & \text{若 } j\neq k,\\ \lambda_k, & \text{若 } j=k.\end{cases}\tag{2.10}$$

主成分分析最早由 Pearson 和 Hotelling 分别在 1901 和 1933 年提出; 文献 [29] 在其经典著作中对主成分分析作了系统地阐述; 文献 [30] 提出了主曲线和主曲面; 文献 [31] 在再生核 Hilbert 空间中讨论了核主成分分析; 文献 [32] 证明了在高维 $p/n\to c>0$ 情形下, 主成分分析对于秩为 1 的协方差矩阵无法得到相合的第一特征向量估计. 除了上述之外, 还有大量文献研究超高维条件下主成分分析的渐近性质, 代表性工作包括文献 [33–35] 等. 下一节将着重介绍如何通过主成分分析方法提取未知因子.

2.3 高维因子模型的潜在因子提取

主成分分析被广泛用于高维统计学习和推断问题中的潜在因子提取, 如手写邮编分类 [36]、人脸识别 [37] 和基因表达数据分析 [38,39] 等. 假设观测到来自模型 (2.1) 的 n 个样本 $\{\boldsymbol{x}_i\}_{i=1}^n$, 其满足

$$\boldsymbol{x}_i=\boldsymbol{a}+\boldsymbol{B}\boldsymbol{f}_i+\boldsymbol{u}_i,\tag{2.11}$$

或者矩阵形式

$$\boldsymbol{X}=\boldsymbol{a}\boldsymbol{1}_n^{\mathrm{T}}+\boldsymbol{B}\boldsymbol{F}^{\mathrm{T}}+\boldsymbol{U},\tag{2.12}$$

其中 $\boldsymbol{X}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\in\mathbb{R}^{p\times n}$, $\boldsymbol{a}$、$\boldsymbol{B}$ 和 $\boldsymbol{F}=(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_n)^{\mathrm{T}}$ 均为未知. 在本文中除非特别说明, 模型 (2.11) 中只有 $\{\boldsymbol{x}_i\}_{i=1}^n$ 可观测, 目标是通过 $\{\boldsymbol{x}_i\}_{i=1}^n$ 推断出 $\boldsymbol{B}$ 和 $\{\boldsymbol{f}_i\}_{i=1}^n$, 进而研究 (2.5) 中的协方差 $\boldsymbol{\Sigma}$. 直观来看, 如果 $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ 能够充分控制 $\boldsymbol{\Sigma}_u$, 则 $\boldsymbol{\Sigma}$ 的前 K 个特征向量生成的特征空间与 $\boldsymbol{B}$ 的列空间 (列向量生成空间) 就会非常相近, 这对于通过主成分分析来估计 $\boldsymbol{B}$ 的列空间是很重要的. 数学上表现为 $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ 与 $\boldsymbol{\Sigma}_u$ 的特征值之间存在一个比较明显的差距, 如果这个差距相较于 $\boldsymbol{\Sigma}_u$ 的特征值不够大, 则主成分分析方法就会得到有偏的估计 [32]. 根据上述讨论, 我们可以通过主成分分析来提取潜在因子 (见算法 1). 在步骤 1 中, 我们可以通过样本均值估计 $\boldsymbol{a}$, 样本协方差阵或者其他稳健方法估计 $\boldsymbol{\Sigma}$. 我们将详细说明 (2.13) 中两个估计的由来. 正如之前讨论的那样, 我们用 $\sum_{j=1}^K\widehat{\lambda}_j\widehat{\boldsymbol{v}}_j\widehat{\boldsymbol{v}}_j^{\mathrm{T}}$ 估计 $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$, 则 (2.13) 中的估计 $\widehat{\boldsymbol{B}}$ 由尺度调整后的 $\widehat{\boldsymbol{\Sigma}}$ 的前 K 个特征向量组成. 由 (2.1) 可知 $\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{a})=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{f}_i+\boldsymbol{B}^{\mathrm{T}}\boldsymbol{u}_i$, 在高维情形下, 由于 $\boldsymbol{u}_i$ 各分量之间是弱相依的, $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{u}_i$ 被平均掉. 记 $\{\boldsymbol{b}_j\}_{j=1}^K$ 是 $\boldsymbol{B}$ 的列向量, 由可识

算法 1 潜在因子提取算法

输入: 观测数据 $\{\boldsymbol{x}_i\}_{i=1}^n$ 及潜在因子个数 K.

1. 建立 $\boldsymbol{a}$ 和 $\boldsymbol{\Sigma}$ 的估计 $\widehat{\boldsymbol{a}}$ 和 $\widehat{\boldsymbol{\Sigma}}$.

2. 计算协方差阵估计的特征分解

$$\widehat{\boldsymbol{\Sigma}}=\sum_{j=1}^p\widehat{\lambda}_j\widehat{\boldsymbol{v}}_j\widehat{\boldsymbol{v}}_j^{\mathrm{T}},$$

其中 $\{\widehat{\lambda}_k\}_{k=1}^K$ 为最大的 K 个特征值, $\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K$ 是对应的特征向量. 记 $\widehat{\boldsymbol{V}}=(\widehat{\boldsymbol{v}}_1,\ldots,\widehat{\boldsymbol{v}}_K)\in\mathbb{R}^{p\times K}$, $\widehat{\boldsymbol{\Lambda}}=\mathrm{diag}(\widehat{\lambda}_1,\ldots,\widehat{\lambda}_K)\in\mathbb{R}^{K\times K}$.

3. 基于主成分分析方法计算估计值

$$\widehat{\boldsymbol{B}}=\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Lambda}}^{1/2}\quad\text{和}\quad\widehat{\boldsymbol{F}}=(\boldsymbol{X}-\widehat{\boldsymbol{a}}\boldsymbol{1}^{\mathrm{T}})^{\mathrm{T}}\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Lambda}}^{-1/2}.\tag{2.13}$$

别性条件 1, 这时近似成立

$$\boldsymbol{f}_i \approx (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{a}) = \mathrm{diag}(\|\boldsymbol{b}_1\|^2, \ldots, \|\boldsymbol{b}_K\|^2)^{-1}\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{a}). \tag{2.14}$$

代入 (2.13) 中的估计 $\widehat{\boldsymbol{B}}$, 得到 $\widehat{\boldsymbol{f}}_i = \widehat{\boldsymbol{\Lambda}}^{-1/2}\widehat{\boldsymbol{V}}^{\mathrm{T}}(\boldsymbol{x}_i - \widehat{\boldsymbol{a}})$, 写成矩阵的形式, 即 (2.13) 中的第二个公式, 这个公式是 $\boldsymbol{x}_i - \widehat{\boldsymbol{a}}$ 在尺度调整后的特征子空间上的投影. 在解释 (2.13) 中的第一个公式前 (见 (2.16)), 我们需要引入一个重要的条件.

以上的启发式推导说明这种基于主成分分析的因子提取方法的成败在于因子信噪比, 换言之, $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ (信号) 的最小非零特征值是否远大于 $\boldsymbol{\Sigma}_u$ (噪声) 的最大特征值, 使得 $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ 的特征空间与 $\boldsymbol{\Sigma}$ 的特征空间近似相等. 这一特征直观上表现为特征值的尖峰性现象 [26–28, 40]. 为了更加精确地描述尖峰性, 我们引入如下条件. 记 $\Omega(1)$ 为一个大于 0 小于 ∞ 的量.

条件 2 (泛在性, pervasiveness assumption) $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ 最大的 K 个特征值的阶数为 $\Omega(p)$, 然而 $\|\boldsymbol{\Sigma}_u\|_2 = O(1)$.

文献 [29] 证明了因子模型与主成分分析并不完全相同, 尤其是对于有限维问题. 但在泛在性条件下, 文献 [26] 指出两者是近似相同的, 可以通过主成分分析构造因子及载荷的相合估计, 这使得基于主成分分析的因子模型的统计推断方法成为研究的主流.

由识别性条件可知 $\mathrm{Cov}(\boldsymbol{f}_i) = \boldsymbol{I}_K$, 对于任意 $k \in [K]$, 第 k 个因子的载荷的均方则需要满足 $p^{-1}\sum_{j=1}^{p} B_{jk}^2 = \Omega(1)$, 特别当 $\{B_{jk}\}_{j=1}^{p}$ 服从独立同分布的非退化分布时, 假设的第一部分满足. 假设的第二部分则表明, 当公共因子被去除后, 横截面相关性就应该变得很弱, 即 $\boldsymbol{\Sigma}_u$ 稀疏, 进一步则有 $\|\boldsymbol{\Sigma}_u\|_2 = O(1)$. $K = 2$ 的情形如图 1 所示. 在泛在性条件下, $\boldsymbol{\Sigma}$ 的前 K 个特征值可以清楚地与其余特征值分离开. 由 Davis-Kahan 定理 [41], 我们可以用 $\boldsymbol{\Sigma}$ 的前 K 个特征向量生成的特征空间作为 $\boldsymbol{B}$ 列空间的相合估计.

泛在性假设可能不是最弱的条件 [20], 但却是最方便的. 它意味着在 $\boldsymbol{x}$ 的测量值中公共因子的影响占比很大. 若因子载荷 $\{\boldsymbol{b}_j\}_{j=1}^{p}$ 是一组来自 K 维非退化总体的现实, 则由大数定律可知,

$$\frac{1}{p}\sum_{j=1}^{p} \boldsymbol{b}_j\boldsymbol{b}_j^{\mathrm{T}} \to \mathrm{E}\,\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}.$$

利用 Weyl 定理可证 $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$ 和 $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ 的非零特征值均为 p 阶. 由条件 1 可知, $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ 的非零特征值和对

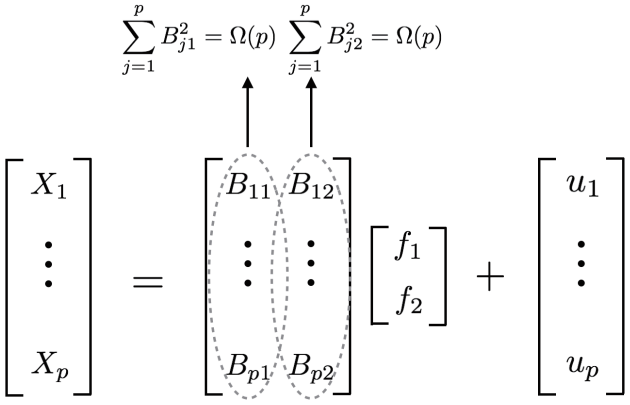

图 1 泛在性假设示意图. $K = 2$ 的情形

应的特征向量分别为

$$\lambda_j(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}) = \|\widetilde{\boldsymbol{b}}_j\|^2, \quad \boldsymbol{v}_j(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}) = \frac{\widetilde{\boldsymbol{b}}_j}{\|\widetilde{\boldsymbol{b}}_j\|}, \tag{2.15}$$

其中 $\widetilde{\boldsymbol{b}}_j$ 是按照 $\|\widetilde{\boldsymbol{b}}_j\|$ 大小递减排序后 $\boldsymbol{B}$ 的第 j 列. 再次利用 Wely 和 Davis-Kahan 定理, 得到如下性质.

命题 1 令 $\lambda_j = \lambda_j(\boldsymbol{\Sigma}), \boldsymbol{v}_j = \boldsymbol{v}_j(\boldsymbol{\Sigma})$. 由可识别性条件 1 和泛在性条件 2, 有

$$\begin{aligned} &|\lambda_j - \|\widetilde{\boldsymbol{b}}_j\|^2| \leqslant \|\boldsymbol{\Sigma}_u\|_2, \quad \text{当 } j \leqslant K, \\ &|\lambda_j| \leqslant \|\boldsymbol{\Sigma}_u\|_2, \quad \text{当 } j > K. \end{aligned}$$

此外还有

$$\left\| \boldsymbol{v}_j - \frac{\widetilde{\boldsymbol{b}}_j}{\|\widetilde{\boldsymbol{b}}_j\|} \right\| = O(p^{-1}\|\boldsymbol{\Sigma}_u\|_2), \quad \text{当 } j \leqslant K.$$

命题的第一部分表明, 当 $j \leqslant K$ 时, $\lambda_j = \Omega(p)$; 当 $j > K$ 时, $\lambda_j = O(1)$. 很显然, 以 K 为界, 特征值就被分离开. 另外, 命题的第二部分则说明, 当 $j \leqslant K$ 时, 主成分的方向 $\boldsymbol{v}_j$ 能够很好地逼近 $\boldsymbol{B}$ 的第 j 个标准化列向量. 换言之就是

$$\left| \frac{\lambda_j}{\|\widetilde{\boldsymbol{b}}_j\|^2} - 1 \right| = O(p^{-1}), \quad \widetilde{\boldsymbol{b}}_j = \lambda_j^{1/2}\boldsymbol{v}_j + O(p^{-1/2}), \quad \text{当 } j \leqslant K. \tag{2.16}$$

把这个结果用矩阵的形式来表达就导出了 (2.13) 中因子载荷的估计.

对于模型 (2.11), 我们总可以不失一般性地假设模型的截距项满足 $\boldsymbol{a} = \boldsymbol{0}$. 当用样本协方差阵作为估计时, 可以证明上述估计与最小二乘估计是相同的. 还是回到模型 (2.11), 最小化均方损失来寻找 $\boldsymbol{B}$ 和 $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^{\mathrm{T}}$,

$$\begin{aligned} &\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{f}_i\|^2 = \|\boldsymbol{X} - \boldsymbol{F}\boldsymbol{B}\|_F^2, \\ &\text{s.t. } n^{-1}\sum_{i=1}^{n} \boldsymbol{f}_i\boldsymbol{f}_i^{\mathrm{T}} = \boldsymbol{I}_K, \\ &\quad \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} \text{ 为对角阵}. \end{aligned} \tag{2.17}$$

若给定 $\boldsymbol{B}$, 可得最小二乘估计 $\widehat{\boldsymbol{f}}_i = \mathrm{diag}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}_i$, 或者 $\widehat{\boldsymbol{F}} = \boldsymbol{X}\boldsymbol{B}\boldsymbol{D}$, 其中 $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}$. 将估计代回 (2.17), 得到一个新的目标函数

$$\left\|\boldsymbol{X}(\boldsymbol{I}_p - \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^{\mathrm{T}})\right\|_F^2 = \mathrm{tr}[(\boldsymbol{I}_p - \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^{\mathrm{T}})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}].$$

可以证明使得上式达到最小的解 $\boldsymbol{B}_0 = \mathrm{span}(\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_K)$. 同理, 我们可以讨论 $\boldsymbol{F}$ 给定的情形, 由于设计阵 $\boldsymbol{F}$ 列正交, 得到最小二乘估计 $\widehat{\boldsymbol{B}} = n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{F}$ 和新目标函数

$$\left\| \boldsymbol{X} - \frac{\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{X}}{n} \right\|_F^2 = \mathrm{tr}\left[\left(\boldsymbol{I}_n - \frac{\boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}}{n} \right) \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \right].$$

类似地, $\widehat{\boldsymbol{F}}/\sqrt{n}$ 的列是 $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 最大的 K 个特征值对应的特征向量, $\widehat{\boldsymbol{B}} = n^{-1}\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{F}}$ [5]. 这时通过计算 $n \times n$ 维矩阵 $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 的特征分解代替计算 $p \times p$ 维矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的特征分解. 当 $p \gg n$ 时, 这会大大减少计算时间. 综上分析, 我们给出如下结论.

命题 2 对于中心化数据 ($\overline{\boldsymbol{x}} = 0$), (2.17) 定义的最小二乘估计为 $\widehat{\boldsymbol{F}} = \sqrt{n} \times (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$ 的前 K 个特征向量, $\widehat{\boldsymbol{B}} = n^{-1}\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{F}}$ 与 (2.13) 中的给出的估计是完全相同的.

本小节所有的讨论都是在已知因子个数 K 的前提下, 然而, 究竟应该如何选择因子 K 的个数, 或者以何种准则来决定 K 的大小仍然是一个十分困扰应用者的重要问题. 我们将在下一节系统地讨论这一问题.

2.4 因子个数的选择方法

在实际应用中, 对于因子模型, 特别是高维因子模型, 在估计公共因子和因子载荷矩阵之前需要首先确定因子的个数 K. 个数 K 通常是利用协方差阵的初步估计 $\widehat{\boldsymbol{\Sigma}}$ 的特征值来获得, 其中经典方法包括似然比检验 (likelihood ratio test [42])、平行分析 (parallel analysis [43]) 和碎石图 (scree plot [44]) 等. 经典的方法主要考虑前 K 个特征所能解释的方差占比, 即观察

$$\widehat{p}_k = \frac{\sum_{j=1}^{k}\lambda_j(\widehat{\boldsymbol{\Sigma}})}{\sum_{j=1}^{p}\lambda_j(\widehat{\boldsymbol{\Sigma}})}$$

的变化, 当选到某个 k 之后, $\widehat{p}_k$ 不再显著增加, 此时的 k 值即为所求. 类似地, 也可以考查相关系数矩阵 $\widehat{\boldsymbol{R}} = \mathrm{diag}(\widehat{\boldsymbol{\Sigma}})^{-1/2}\widehat{\boldsymbol{\Sigma}}\,\mathrm{diag}(\widehat{\boldsymbol{\Sigma}})^{-1/2}$, 选择

$$\widehat{K}_1 = \#\{j : \lambda_j(\widehat{\boldsymbol{R}}) > 1\}$$

作为因子个数. 当 $p/n \to c > 0$ 时, 文献 [45] 建议使用

$$\widehat{K}_1 = \#\left\{j : \lambda_j(\widehat{\boldsymbol{R}}) > 1 + \sqrt{\frac{p}{n}}\right\}.$$

这种方法的好处是不包含任何调优参数. 接下来介绍其他一些方法. 第一, 基于特征值的比值; 第二, 基于特征值的差; 第三, 基于信息准则. 为了便于表述, 假定初始估计为样本协方差矩阵, 其非零特征值依次按照降序排列, $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{n\wedge p}$, 其中 $n \wedge p = \min\{n, p\}$.

文献 [46–48] 提出了基于相邻特征值比值的因子个数估计方法. 具体如下: 给定因子个数的上限 $k_{\max}$, 基于特征值比值的因子个数估计为

$$\widehat{K}_2 = \underset{j\leqslant k_{\max}}{\mathrm{argmax}}\,\frac{\lambda_j}{\lambda_{j+1}}.$$

直观上看, 当显著的那些主成分与其余的主成分完全分离时, 上述比值在 $k = K$ 时达到最大. 在一定的条件下, 不需要涉及任何调优参数, 就可以建立估计值 $\widehat{K}_2$ 的相合性.

文献 [49] 提出了基于相邻特征值之差的因子个数估计方法. 对于给定的阈值 $\delta > 0$ 和因子个数的上限 $k_{\max}$, 估计定义为

$$\widehat{K}_3(\delta) = \max\{j \leqslant k_{\max} : \lambda_j - \lambda_{j+1} \geqslant \delta\}.$$

利用随机矩阵经验谱分布的相关理论和泛在性条件 2, 文献 [49] 证明了 $\widehat{K}_3(\delta)$ 的相合性, 并提出了一种从样本协方差矩阵经验谱分布来确定 δ 的数据驱动方法.

最后是基于信息准则的选择方法. 注意到

$$V(k) = \frac{1}{np}\min_{\boldsymbol{B}\in\mathbb{R}^{p\times k},\boldsymbol{F}\in\mathbb{R}^{n\times k}}\|\boldsymbol{X} - \boldsymbol{1}_n\overline{\boldsymbol{x}}^{\mathrm{T}} - \boldsymbol{F}\boldsymbol{B}^{\mathrm{T}}\|_F^2 = \frac{1}{p}\sum_{j>k}\lambda_j.$$

对于给定的 k, $V(k)$ 可以看作用 k 个因子拟合数据得到的残差平方和. 类似模型变量选择, 文献 [12] 提出了通过最优化带有惩罚项的残差平方和来估计因子个数 $\widehat{K} \leqslant k_{\max}$, 即最小化

$$PC(k) = V(k) + k\widehat{\sigma}^2 g(n,p), \quad g(n,p) := \frac{n+p}{np}\log\left(\frac{n+p}{np}\right),$$

其中 $\widehat{\sigma}^2$ 是 $(np)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{d}\mathrm{E}\,u_{ji}^2$ 的相合估计. 文献 [12] 建立了对于更加一般的 $g(n,p)$ 的相合性结论.

作为本节的结束, 我们指出对于无结构假设的高维数据可以通过因子模型学习出一种潜在的结构, 主成分分析是因子学习的重要途径, 具有降维和可解释等多种优点, 但是当 $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$ 的最小非零特征值远小于 $\|\boldsymbol{\Sigma}_u\|_2$ 时, 因为 "信号" 与 "噪声" 无法区分, 得到的估计不相合. 近来主成分分析的相合性受到很多关注, 例如, 文献 [50] 使用随机矩阵理论对平行分析进行了分析.

3 结构化协方差矩阵与精度矩阵学习

协方差和精度矩阵的统计推断在统计学、应用数学和机器学习领域无处不在, 如 Fisher 判别分析、最优投资组合、高斯图模型、Hotelling T^2 检验和伪发现率控制等. 它们也被广泛应用在诸如经济学、金融学、基因组学、心理学和计算社会科学等学科中. 协方差矩阵学习从数据推断出维数平方阶的矩阵未知元素, 是一个典型的超高维问题, 通常需要加入稀疏或带状等结构性假设. 因子模型建立了结构学习与协方差矩阵学习之间的桥梁, 使我们可以推断出驱动变量相关性的潜在因子, 通过这些潜因子实现高维协方差矩阵的统计推断. 本小节将着重介绍因子模型诱导的协方差和精度矩阵估计.

3.1 协方差矩阵的统计推断

在有些实际应用中, 因子是可以被观测到的, 如著名的 Fama-French 三 (五) 因子模型 [11, 51], 这时因子模型就变为一个多元线性模型. 假设观测数据为 $\{(\boldsymbol{f}_i, \boldsymbol{x}_i)\}_{i=1}^n$, 通过最小二乘法估计模型 (2.1) 中回归系数

$$(\widehat{a}_j, \widehat{\boldsymbol{b}}_j^{\mathrm{T}})^{\mathrm{T}} = \underset{a_j\in\mathbb{R},\boldsymbol{b}_j\in\mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{n}(X_{ij} - a_j - \boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{f}_i)^2, \quad j = 1,\ldots,p,$$

得到载荷矩阵的估计 $\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{b}}_1^{\mathrm{T}},\ldots,\widehat{\boldsymbol{b}}_p^{\mathrm{T}})$ 和残差 $\widehat{\boldsymbol{u}}_i = (\widehat{u}_{i1},\ldots,\widehat{u}_{ip})^{\mathrm{T}}$, $\widehat{u}_{ij} = X_{ij} - \widehat{a}_j - \widehat{\boldsymbol{b}}_j^{\mathrm{T}}\boldsymbol{f}_i$, 并记 $\widehat{\boldsymbol{\Sigma}}_u$ 是残差序列 $\{\widehat{\boldsymbol{u}}_i\}_{i=1}^n$ 的初始协方差阵估计. 由于 $\boldsymbol{\Sigma}_u$ 具有稀疏性, 对 $\widehat{\boldsymbol{\Sigma}}_u$ 进行正则化得到稀疏估计 $\widehat{\boldsymbol{\Sigma}}_{u,\lambda}$ (具体参见第 3.2 小节), λ 是调优参数, 则有整体的协方差矩阵估计

$$\widehat{\boldsymbol{\Sigma}}_\lambda = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^{\mathrm{T}} + \widehat{\boldsymbol{\Sigma}}_{u,\lambda}, \tag{3.1}$$

其中 $\widehat{\boldsymbol{\Sigma}}_f$ 是可观测因子 $\{\boldsymbol{f}_i\}_{i=1}^n$ 的样本协方差阵. 若对 $\widehat{\boldsymbol{\Sigma}}_u$ 的相关系数矩阵做阈值为 1 的截断, 则得到 $\widehat{\boldsymbol{\Sigma}}_{u,\lambda} = \mathrm{diag}(\widehat{\boldsymbol{\Sigma}}_u)$ 和总体估计

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{B}}^{\mathrm{T}} + \mathrm{diag}(\widehat{\boldsymbol{\Sigma}}_u). \tag{3.2}$$

(3.2) 能保持矩阵的正定性且适用于严格因子模型. 文献 [52] 指出相较于样本协方差阵, 基于因子模型的正则估计 (3.1) 和 (3.2) 在估计 $\boldsymbol{\Sigma}$ 时有相同的收敛速度, 在估计 $\boldsymbol{\Sigma}^{-1}$ 时有更快的收敛速度. 其他大量研究也充分证明了这一点.

例 1 [52] 假设单因子模型满足 $\boldsymbol{B}=\mathbf{1}_p$, $\boldsymbol{\Sigma}_u=\boldsymbol{I}_p$, 则 $\boldsymbol{\Sigma}=\sigma_f^2\mathbf{1}_p\mathbf{1}_p^{\mathrm{T}}+\boldsymbol{I}_p$, $\sigma_f^2:=\mathrm{Var}(f)$ 的估计为 $\widehat{\sigma}_f^2=n^{-1}\sum_{i=1}^n(f_i-\bar{f})^2$. 下列结论成立:

$$\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_2=p\left|\widehat{\sigma}_f^2-\sigma_f^2\right|.$$

由中心极限定理可知 $\sqrt{n}p^{-1}\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_2=\sqrt{n}\|\widehat{\sigma}_f^2-\sigma_f^2\|$ 渐近服从 (非负) 正态分布, 因此, 当 $p\asymp\sqrt{n}$ 时, $\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_2=O_P(p/\sqrt{n})$ 是不相合的.

受到此例的启发, 文献 [52] 考虑了关于矩阵尺度不变的两种损失函数—二次损失

$$\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_\Sigma:=p^{-1/2}\|\boldsymbol{\Sigma}^{-1/2}(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}\|_F=p^{-1/2}\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}-\boldsymbol{I}\|_F,$$

以及熵损失 [53]

$$\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})-\log|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}|-p.$$

通过 $\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}$ 的特征值可以很容易地发现两种损失的关系, 记其特征值依次为 $\widetilde{\lambda}_1,\ldots,\widetilde{\lambda}_p$, 则有

$$\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}-\boldsymbol{I}\|_F^2=\sum_{j=1}^p(\widetilde{\lambda}_j-1)^2,$$

以及熵损失等于

$$\sum_{j=1}^p(\widetilde{\lambda}_j-1-\log\widetilde{\lambda}_j)\approx\frac{1}{2}\sum_{j=1}^p(\widetilde{\lambda}_j-1)^2.$$

约等号由在 $\widetilde{\lambda}_j=1$ 处进行 Taylor 展开可得, 即当 $\widehat{\boldsymbol{\Sigma}}$ 与 $\boldsymbol{\Sigma}$ 非常接近时, 两种损失近似相等. 文献 [54] 建立了正则化估计 (3.1) 在不同损失下的收敛速度.

3.2 协方差矩阵稀疏性与门限方法

协方差结构假设是估计高维协方差矩阵和精度矩阵的重要基础. 上节中由因子模型导出的 "低秩"+"稀疏" 的协方差结构 [27, 28] 被文献 [52] 称为条件稀疏性, 文献 [55] 提出的稀疏协方差矩阵可以看成是其中一种特例. 本节系统介绍稀疏协方差矩阵的估计方法.

设 p 维观测样本 $\{\boldsymbol{x}_i\}_{i=1}^n$ 的矩估计为

$$\widehat{\boldsymbol{\mu}}=n^{-1}\sum_{i=1}^n\boldsymbol{x}_i,\quad \boldsymbol{S}=n^{-1}\sum_{i=1}^n(\boldsymbol{x}_i-\widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\widehat{\boldsymbol{\mu}})^{\mathrm{T}}.$$

当 $p\geqslant n$ 时, 估计 $O(p^2)$ 个元素所产生的累积误差会导致样本协方差阵与总体协方差阵之间存在明显的偏差, 且 $\boldsymbol{S}$ 退化. 为此, 通常会考虑协方差矩阵具有某种特殊结构, 最简单的假设是 $\boldsymbol{\Sigma}$ 存在大量非对角线元素为 0, 且可以通过 ℓ_1 范数控制矩阵每一行的有效参数个数, 即非零元素的个数, 定义度量

$$m_{p,0}(\boldsymbol{\Sigma})=\max_{i\leqslant p}\sum_{j=1}^p I(\sigma_{ij}\neq 0). \tag{3.3}$$

将这种精确的度量推广就得到了近似稀疏度

$$m_{p,q}(\boldsymbol{\Sigma})=\max_{i\leqslant p}\sum_{j=1}^p|\sigma_{ij}|^q. \tag{3.4}$$

对于给定的水平 $m_p>0$, 文献 [55] 考虑了稀疏矩阵空间

$$\{\boldsymbol{\Sigma}\geqslant 0:\sigma_{ii}\leqslant C,\ m_{p,q}(\boldsymbol{\Sigma})\leqslant m_p\}.$$

其好处是有

$$\|\boldsymbol{\Sigma}\|_2\leqslant m_{p,q}\leqslant \max_i\sum_j(\sigma_{ii}\sigma_{jj})^{(1-q)/2}\,|\sigma_{ij}|^q\leqslant C^{1-q}m_p.$$

由此可以得到更一般的稀疏参数空间 [56]

$$\mathcal{C}_q(m_p)=\left\{\boldsymbol{\Sigma}\geqslant 0:\max_i\sum_j(\sigma_{ii}\sigma_{jj})^{(1-q)/2}\,|\sigma_{ij}|^q\leqslant C^{1-q}m_p\right\}. \tag{3.5}$$

回到估计问题, 有对数似然函数 $-n\log|\boldsymbol{\Sigma}|-\sum_{i=1}^n(\boldsymbol{x}_i-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}}$. 代入 $\boldsymbol{\mu}$ 的极大似然估计 $\bar{\boldsymbol{x}}$, 并加入正则项 (惩罚项), 就可以得到关于 $\boldsymbol{\Sigma}$ 的惩罚似然估计

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{PMLE}}=\operatorname{argmin}-\log\left|\boldsymbol{\Sigma}^{-1}\right|+\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S})+\sum_{i\neq j}p_{\lambda_{ij}}(|\sigma_{ij}|). \tag{3.6}$$

因为通常假设对角线元素非零, 所以这里不对其进行惩罚. 然而即使罚函数是凸的, (3.6) 也仍是一个复杂难解的优化问题. 为了克服这一困难, 文献 [55] 开发了一种硬门限方法来获得稀疏估计, 给定门限参数 $\lambda>0$, 对初始估计 $\boldsymbol{S}=(s_{ij})=(\widehat{\sigma}_{ij})$ 进行逐元素阈值转化,

$$\widehat{\boldsymbol{\Sigma}}_\lambda=(\widehat{\sigma}_{ij}I(|\widehat{\sigma}_{ij}|\geqslant\lambda)),\quad\forall\, i\neq j. \tag{3.7}$$

这种方法将样本协方差阵中所有小于阈值的非对角线元素压缩为 0, 其他保持不变. (3.7) 也可以表示成惩罚最小二乘的形式:

$$\sum_{i,j}(\sigma_{ij}-s_{ij})^2+\sum_{i\neq j}p_\lambda(|\sigma_{ij}|), \tag{3.8}$$

其中惩罚函数取为 $p_\lambda(x)=2^{-1}\lambda^2-2^{-1}(\lambda-t)^2I(t<\lambda)$. 通过下述广义门限法则可以灵活地选择 (3.8) 中的惩罚函数.

定义 2 函数 $\tau_\lambda(\cdot)$ 被称为广义门限函数, 若其满足

(1) $|\tau_\lambda(z)|\leqslant a\,|y|$, 对于所有满足 $|z-y|\leqslant\lambda/2$ 的 z 和 y, 以及某个常数 $a>0$;

(2) $|\tau_\lambda(z)-z|\leqslant\lambda$, $\forall\, z\in\mathbb{R}$.

当 $y=0$ 时, $\tau_\lambda(z)=0$, $|z|\leqslant\lambda/2$, 这满足门限函数的要求, 软门限函数 (soft-thresholding) 对应 $a=1$, 硬门限函数 (hard-thresholding) 对应 $a=2$.

采用统一的阈值 λ 虽然很简便, 但是忽略了协方差矩阵的尺度变化. 解决途径有两种, 首先是考虑自适应门限估计 [56], 例如, 基于 t- 统计量构造估计 $\widehat{\boldsymbol{\Sigma}}_\lambda=(\widehat{\sigma}_{ij}I(|\widehat{\sigma}_{ij}|/\mathrm{SE}(\widehat{\sigma}_{ij})\geqslant\lambda)$, 其中 $\mathrm{SE}(\widehat{\sigma}_{ij})$ 是估计的标准误; 另外是对相关系数矩阵 $\boldsymbol{\Psi}$ 进行阈值变换, 得到 $\widehat{\boldsymbol{\Psi}}_\lambda$ ($\lambda\in[0,1]$) 和协方差矩阵估计 $\widehat{\boldsymbol{\Sigma}}^*_\lambda=\operatorname{diag}(\widehat{\boldsymbol{\Sigma}})^{1/2}\widehat{\boldsymbol{\Psi}}_\lambda\operatorname{diag}(\widehat{\boldsymbol{\Sigma}})^{1/2}$. 两种特殊情形: 当 $\lambda=0$ 退化为样本协方差阵, 当 $\lambda=1$ 退化为由分量方差构成的对角阵. 综合以上, 给出一般化的门限方法

$$\widehat{\boldsymbol{\Sigma}}^\tau_\lambda=(\tau_{\lambda_{ij}}(\widehat{\sigma}_{ij}))\coloneqq(\widehat{\sigma}^\tau_{ij}),\quad\lambda_{ij}=\lambda\sqrt{\widehat{\sigma}_{ii}\widehat{\sigma}_{jj}\frac{\log p}{n}}. \tag{3.9}$$

$\tau_{\lambda_{ij}}(\cdot)$ 是广义门限函数, $\widehat{\sigma}_{ij}$ 是 $\boldsymbol{\Sigma}$ 的某种初始估计 (如样本协方差阵, 或者其他稳健化估计等), $\sqrt{\log p/n}$ 代表了一致收敛率.

门限方法由于将数值过小的元素直接压缩为 0, 从而大大减少了累计误差, 可同时也带了估计偏差, 但决定一个元素是否应被估计总比把这个元素估计准确来得容易, 再加上矩阵稀疏性的假设, 我们总是可以把偏差控制在可接受的范围.

3.3 精度矩阵与图模型

精度矩阵, 即协方差矩阵的逆, 在统计推断问题中扮演着核心的角色, 如 Fisher 判别分析、Hotelling T^2 检验, 尤其在高斯图模型中占有重要地位. 假设 $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 令 $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ 为精度矩阵. 考虑 $\boldsymbol{x} = (\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}})^{\mathrm{T}}$, 则协方差矩阵满足

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}.$$

由多元正态性质可知 $(\boldsymbol{x}_1 \mid \boldsymbol{x}_2) \sim N(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11.2}), \boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, $\boldsymbol{\Omega}_{11} = \boldsymbol{\Sigma}_{11.2}^{-1}$. 对于一种特殊情形 $\boldsymbol{x}_1 = (X_1, X_2)^{\mathrm{T}}$, $\boldsymbol{x}_2 = (X_3, \ldots, \boldsymbol{X}_p)^{\mathrm{T}}$, $\boldsymbol{\Sigma}_{11.2}^{-1}$ 是一个具有显式表达 2×2 矩阵, 可以证明

$$\omega_{12} = 0 \text{ 当且仅当 } \boldsymbol{\Sigma}_{11.2} \text{ 的 } (1,2) \text{ 位置上的元素为 } 0.$$

我们把这一结论推广为如下结论.

命题 3 对于多元正态随机向量 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$, 记其精度矩阵 $\boldsymbol{\Omega} = (\omega_{ij})$. $\omega_{ij} = 0$ 当且仅当给定其他变量 X_i 与 X_j 条件独立.

命题给出了稀疏精度矩阵在高斯图模型中的重要解释, 如果把 $\boldsymbol{x}$ 中的每个元素看作图中的一个顶点, 顶点之间的连边代表随机变量之间的条件相依性, 也就是说顶点 i 与 j 之间存在连边当且仅当 $\omega_{ij} \neq 0$. 图 2 给出了一个简单图模型例子. 惩罚似然 (3.6) 给出了一个稀疏精度矩阵的估计, 惩罚函数的奇异性和 $\boldsymbol{\Omega}$ 的正定性约束使得这个优化问题难于求解. 文献中提出了各种算法来解决 (3.6) 与 ℓ_1 惩罚. 文献 [57] 提出了 maxdet 算法, 文献 [58,59] 建立了一种更加有效的顺时针下降算法 (clockwise descent algorithm), 文献 [60] 提出了投影下降法, 文献 [61] 运用了 Nesterov 光滑优化技术来进行求解, 文献 [62] 应用了交叉方向乘子法 (alternating direction method of multipliers, ADMM) 求解 LASSO (least absolute shrinkage and selection operator) 惩罚正态似然估计.

文献 [3] 最早研究了高维图模型. 惩罚似然是估计精度矩阵的主要方法, 参见文献 [57–59, 63–67]. 还有其他一些估计方法, 例如, 文献 [68] 提出了惩罚散度估计 (penalized log-determinant divergence), 文献 [69] 提出了 D- 迹估计 (D-trace estimator), 文献 [70, 71] 提出了基于约束 ℓ_1 优化的精度矩阵估

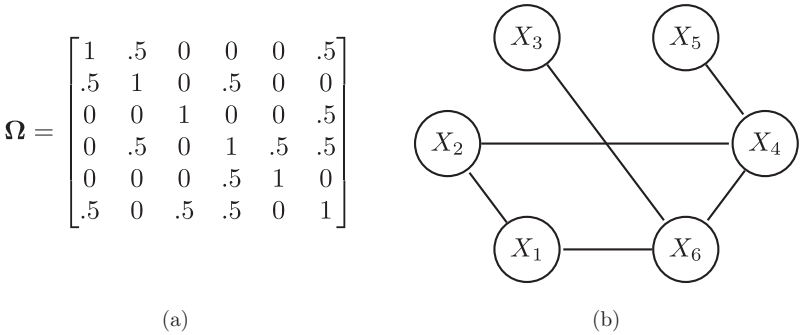

图 2 (a) 精度矩阵; (b) 对应的图表示

计, 文献 [72] 提出了一种对调优参数不敏感的估计方法.

3.4 协方差矩阵的初始稳健估计

对于高维数据, 协方差矩阵的估计需要对观测样本进行数万次甚至数百万次的计算才能得到, 只有在特殊的尾概率分布假设下, 累计误差才会被控制住. 高维统计理论大都建立在次高斯分布假设上, 一个一维随机变量 X 满足次高斯性是指, 该随机变量对应的次高斯范数 $\|X\|_{\psi_2}=\sup_{q\geqslant 1}q^{-1/2}(\mathrm{E}\,|X|^q)^{1/q}$ 是有界的. 常见的满足次高斯性的例子包括正态随机变量、有界随机变量、薄尾分布随机变量及其他尾分布与正态分布相似的随机变量等. 对于多维随机向量, 采用诱导范数 $\|\boldsymbol{x}\|_{\psi_2}=\sup_{\|\boldsymbol{v}\|_2=1}\left\|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{v}\right\|_{\psi_2}$ 定义. 然而, 要求成千上万的变量都满足这种特殊的薄尾假设只是一种数学化的理想, 现实的需求使稳健方法应运而生.

文献 [73–75] 指出对于独立同分布样本 $X_1,\dots,X_n\sim(\mu,\sigma^2)$, 如果没有次高斯假设, 则样本均值 $\bar{X}$ 不具有指数收缩性, 由 Markov 不等式可知,

$$\mathrm{P}\left(|\bar{X}-\mu|\geqslant\frac{t\sigma}{\sqrt{n}}\right)\leqslant t^{-2},$$

在不附加更多条件的情形下, 样本均值通常只具有 Cauchy 尾. 文献 [74] 对数据做截尾处理 $\widetilde{X}_i=\mathrm{sgn}(X_i)\min\{|X_i|,\tau\}$, $\tau\asymp\sigma\sqrt{n}$, 再次计算均值得到

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i-\mu\right|\geqslant\frac{t\sigma}{\sqrt{n}}\right)\leqslant 2\exp\{-ct^2\},\quad c>0. \tag{3.10}$$

可以说, 只要对二阶矩有限的数据做截尾处理后, 其均值都表现出次高斯收缩性.

将这一技术推广到协方差矩阵的估计. 由于 $\mathrm{Cov}(X_i,X_j)=\mathrm{E}\,X_iX_j-\mathrm{E}\,X_i\,\mathrm{E}\,X_j$, 所以高维协方差矩阵 $\boldsymbol{\Sigma}=(\sigma_{ij})_{i,j=1}^n$ 的估计涉及 $O(p^2)$ 个单变量均值的估计. 首先回顾样本协方差阵估计的上界问题 [76–78], 有如下结果.

定理 1 假设 $\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\}_{i=1}^n$ 为独立同分布随机向量且服从次高斯分布. 记 $\kappa=\|\boldsymbol{x}_i\|_{\psi_2},\delta=C\sqrt{p/n}+t/\sqrt{n}$, 则存在只依赖于 κ 的常数 c 和 C 使得不等式

$$\mathrm{P}(\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{sam}}-\boldsymbol{\Sigma}\|_2\geqslant\max\{\delta,\delta^2\}\|\boldsymbol{\Sigma}\|_2)\leqslant 2\exp\{-ct^2\} \tag{3.11}$$

对任意 $t\geqslant 0$ 都成立.

可以发现样本协方差阵估计的上界依赖于原矩阵的维数 p, 在高维情形下, 这个上界会控制得过松; 另外, 结果中的收缩程度还取决于观测随机向量的次高斯性. 当存在成千上万个变量时, 这种假设几乎无法验证 (参见文献 [74]). 尤其对于厚尾数据, 样本协方差的谱范数通常不满足次高斯性 [73,79,80]. 因此, 不宜对厚尾数据的样本协方差阵做主成分分析, 而应该选择其他稳健的协方差矩阵初始估计. 考虑逐项稳健的协方差估计

$$(\widehat{\boldsymbol{\Sigma}}_{\mathcal{E}})_{ij}=\widehat{(\mathrm{E}\,X_iX_j)}^{\mathcal{A}}-\widehat{(\mathrm{E}\,X_i)}^{\mathcal{A}}\widehat{(\mathrm{E}\,X_j)}^{\mathcal{A}}, \tag{3.12}$$

其中上角标 $\mathcal{A}$ 代表不同的稳健方法, $\mathcal{A}=\mathcal{T}$ 为截尾稳健估计, $\mathcal{A}=\mathcal{H}$ 为自适应 Huber 稳健估计, 其定义为

$$\widehat{\mu}_\tau=\operatorname*{argmin}_{\mu}\sum_{i=1}^{n}\rho_\tau(X_i-\mu),\quad \rho_\tau(x)=\begin{cases}x^2, & \text{若 } |x|\leqslant\tau,\\ \tau(2|x|-\tau), & \text{若 } |x|>\tau.\end{cases} \tag{3.13}$$

令 $\widetilde{\boldsymbol{x}}_i^{\tau}$ 为水平 τ 的逐项截尾数据, 截尾稳健协方差矩阵估计为

$$\widetilde{\boldsymbol{\Sigma}}_{\mathcal{T}}(\tau)=\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i^{\tau}\widetilde{\boldsymbol{x}}_i^{\tau^{\mathrm{T}}}-\overline{\boldsymbol{x}}_{\mathcal{T}}\overline{\boldsymbol{x}}_{\mathcal{T}}^{\mathrm{T}},\quad \overline{\boldsymbol{x}}_{\mathcal{T}}=\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i^{\tau}. \tag{3.14}$$

在 4 阶矩有限条件下, 与样本协方差阵估计在次高斯条件下有着相同的收敛率 (对于给定的范数而言), 即

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{E}}-\boldsymbol{\Sigma}\|_{\max}=O_P\left(\sqrt{\frac{\log p}{n}}\right). \tag{3.15}$$

一言概之, 只要数据有有限的 4 阶矩, 我们就可以通过稳健化方法获得与高斯数据样本协方差矩阵相同的估计速率, 同时适用于各种矩阵正则化方法 [55,56]. 最后给出一个严格的定理.

定理 2 假设 $u^2=\max_{i,j}\operatorname{var}(X_iX_j)$, $v^2=\max_i\operatorname{var}(X_i)$, $\sigma=\max\{|u|,|v|\}$ 都有界, 且 $\log p=o(n)$, 则有

(1) 对水平 $\tau\asymp\sigma\sqrt{n}$ 的截断估计 $\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}}(\tau)$, 收缩不等式

$$\mathrm{P}\left(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{T}}(\tau)-\boldsymbol{\Sigma}\|_{\max}\geqslant\sqrt{\frac{a\log p}{cn}}\right)\leqslant 4p^{2-a}$$

对任意的正数 a 和正的常数 c 均成立;

(2) 对水平 $\tau\asymp\sigma\sqrt{n/(a\log p)}$ 的自适应 Huber 估计 $\widehat{\boldsymbol{\Sigma}}_{\mathcal{H}}(\tau)$, 收缩不等式

$$\mathrm{P}\left(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{H}}(\tau)-\boldsymbol{\Sigma}\|_{\max}\geqslant\sqrt{\frac{a\log p}{cn}}\right)\leqslant 4p^{2-a/16}$$

对任意的正数 a 和正的常数 c 均成立.

我们也可以将逐项截尾替换为整体截尾, 考虑一种不必计算均值的协方差计算方法

$$\boldsymbol{\Sigma}=\frac{1}{2}\mathrm{E}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}=\frac{1}{2}\mathrm{E}\,\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\,\frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}}{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2},$$

直接对二次项 $\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2$ 做水平 τ 的截尾处理就得到了稳健 U- 协方差 (robust U-covariance) 估计

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\mathcal{U}}(\tau)&=\left(2\binom{n}{2}\right)^{-1}\sum_{i\neq j}\min\{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2,\tau\}\frac{(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}}{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}\\&=\left(2\binom{n}{2}\right)^{-1}\sum_{i\neq j}\min\left\{1,\frac{\tau}{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}\right\}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}.\end{aligned} \tag{3.16}$$

由于 $\boldsymbol{x}_i-\boldsymbol{x}_j$ 有这种对称结构, 这就回避了逐项计算均值. 文献 [81] 给出了整体截尾方法的理论结果.

定理 3 令 $v^2=2^{-1}\|\mathrm{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}]^2+\operatorname{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^2\|_2$. 当截尾水平满足 $\tau\geqslant|v|\,n^{1/2}/(2t)$, 对任意正数 t 时, 如下收缩不等式成立:

$$\mathrm{P}(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{U}}(\tau)-\boldsymbol{\Sigma}\|_2\geqslant 4\,|v|\,tn^{-1/2})\leqslant 2\,p\exp\{-t^2\}.$$

这一结果也可以拓展到逐项的 ℓ_∞ 范数, 即 $\|\cdot\|_{\max}$. 在稳健 U- 协方差估计提出之前, 文献 [74,82] 分别利用 4 阶矩提出了样本协方差的压缩变换. 具体而言, 当 $\mathrm{E}\,\boldsymbol{x}=0$ 时, 他们提出如下估计:

$$\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}}(\tau)=\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i^{\tau}\widetilde{\boldsymbol{x}}_i^{\tau^{\mathrm{T}}},\quad \widetilde{\boldsymbol{x}}_i=\frac{\min\{\|\boldsymbol{x}_i\|_4,\tau\}\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_4}. \tag{3.17}$$

只要 $\boldsymbol{x}$ 的 4 阶矩有限, 这种压缩估计在谱范数意义下就具有次高斯性.

定理 4 假设 p 维单位向量 $\boldsymbol{v}\in\mathcal{S}^{p-1}$ 满足 $\mathrm{E}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}_i)^4\leqslant R$, 当截尾水平 $\tau\asymp(nR/(\delta\log p))^{1/4}$ 时, 对任意正数 c 和 $\delta>0$, 如下收缩不等式成立:

$$\mathrm{P}\left(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{S}}(\tau)-\boldsymbol{\Sigma}\|_2\geqslant\sqrt{\frac{\delta Rp\log p}{n}}\right)\leqslant p^{1-c\delta}.$$

另外, 对于椭球分布族, 文献 [83,84] 分别基于两种秩相关系数: Kendall-τ 相关系数和 Spearman 相关系数构造了稳健的相关系数矩阵估计. 基于这两种稳健的相关系数矩阵估计, 利用文献 [73] 的方法就可以得到单个变量方差的稳健估计. 以上方法得到的估计也同样满足定理 2 的逐项一致收敛性.

3.5 POET 方法

文献 [26] 提出了 POET (principal orthogonal complement threshholding), 用来估计协方差. 这种方法通过高维主成分分析估计潜在因子, 再利用门限方法得到稀疏的异质项协方差矩阵估计. 具体算法见算法 2.

算法 2 POET 方法

输入: 观测数据 $\{\boldsymbol{x}_i\}_{i=1}^n$, 初始协方差估计 $\widehat{\boldsymbol{\Sigma}}$, 潜在因子个数 K.

1. 利用算法 1 得到因子载荷矩阵估计 $\widehat{\boldsymbol{B}}$ 和潜在因子估计 $\widehat{\boldsymbol{f}}_i$;
2. 计算 $\boldsymbol{\Sigma}_u$ 的初始估计 $\widehat{\boldsymbol{\Sigma}}_u=\sum_{j=K+1}^p\widehat{\lambda}_j\widehat{\boldsymbol{v}}_j\widehat{\boldsymbol{v}}_j^{\mathrm{T}}:=(\widehat{\sigma}_{u,ij})$ 和门限估计 $\widehat{\boldsymbol{\Sigma}}_{u,\lambda}^{\tau}:=(\tau_{\lambda_{ij}}(\widehat{\sigma}_{u,ij}))$, 其中 $\lambda_{ij}=\lambda\sqrt{\widehat{\sigma}_{u,ii}\widehat{\sigma}_{u,jj}\log p/n}$;
3. 合并得到协方差阵 $\boldsymbol{\Sigma}=\mathrm{Cov}(\boldsymbol{x})$ 的估计

$$\widehat{\boldsymbol{\Sigma}}_{\lambda}=\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^{\mathrm{T}}+\widehat{\boldsymbol{\Sigma}}_{u,\lambda}^{\tau}=\sum_{j=1}^K\widehat{\lambda}_j\widehat{\boldsymbol{v}}_j\widehat{\boldsymbol{v}}_j^{\mathrm{T}}+\widehat{\boldsymbol{\Sigma}}_{u,\lambda}^{\tau}.\tag{3.18}$$

步骤 2 中的 $\tau_{\lambda_{ij}}(\cdot)$ 为广义门限函数, $\lambda\sqrt{\log p/n}$ 为门限水平. 门限函数有着灵活的选择, 例如, 文献 [85] 研究时空数据时选择了 $\tau_{\lambda_{ij}}(\widehat{\sigma}_{ij})=\widehat{\sigma}_{ij}I(|i-j|\leqslant k)$, 即时空上相隔较远的两个变量之间应该具有较弱的相关性. 其他相关文献还包括了文献 [86–90] 等.

另一种应用中常见的情形是部分因子已知, 部分因子未知, 这种情形在资产组合理论经常遇到. 我们可以先通过回归方法将已知因子的效应去掉, 得到残差 $\{\widehat{\boldsymbol{u}}_i\}_{i=1}^n$ 和其协方差阵估计 $\widehat{\boldsymbol{\Sigma}}_u$, 然后对 $\widehat{\boldsymbol{\Sigma}}_u$ 应用 POET 方法即可. 由回归的性质可知这些新提取的因子与已知因子间也是正交的, 因此得到了一个满足可识别性条件的增广因子模型.

3.6 理论结果综述

3.6.1 扰动上界

特征空间的扰动理论是研究因子模型及相关学习问题的基本技术工具. 对于具体问题, 只需要对号入座即可, 例如, 在协方差阵分解 (2.3) 中, 可以将 $\boldsymbol{\Sigma}$ 看作因子载荷矩阵 $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ 加上一个扰动量 $\boldsymbol{\Sigma}_u$, 抑或将协方差矩阵的估计 $\widetilde{\boldsymbol{\Sigma}}$ 看作其真值 $\boldsymbol{\Sigma}$ 伴随扰动量 $\widetilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}$.

处理特征空间扰动的关键是 Davis-Kahan 定理, 它通常用于推导对称矩阵的 ℓ_2 型上界 (包括谱范数界). 对于 $\mathbb{R}^p$ 中任意两个 K 维子空间 $\mathcal{S}$ 和 $\widetilde{\mathcal{S}}$ 及各自的一组标准正交基 $\boldsymbol{V},\widetilde{\boldsymbol{V}}\in\mathbb{R}^{p\times K}$, 定义两个子空间的距离为

$$d_2(\mathcal{S},\widetilde{\mathcal{S}})=\|\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^{\mathrm{T}}-\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\|_2,\quad d_F(\mathcal{S},\widetilde{\mathcal{S}})=\|\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^{\mathrm{T}}-\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\|_F.$$

因为 $\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^{\mathrm{T}}$ 和 $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}$ 是两个投影算子, 所以, 这两个距离是良定义的, 且不依赖于基向量的选取. 另外, 这两个距离与典则角 (主角) 有着密切联系, 具体来说, 若记 $\widetilde{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{V}$ 的奇异值为 $\{\sigma_k\}_{k=1}^{K}$, 对应的典则角定义为 $\theta_k=\cos^{-1}\sigma_k,\ k=1,\ldots,K$, 令 $\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{V}},\boldsymbol{V})=\mathrm{diag}(\sin\theta_1,\ldots,\sin\theta_K)\in\mathbb{R}^{K\times K}$, 则有下列被大家熟知的等式:

$$\|\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{V}},\boldsymbol{V})\|_2=d_2(\mathcal{S},\widetilde{\mathcal{S}}),\quad \sqrt{2}\|\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{V}},\boldsymbol{V})\|_F=d_F(\mathcal{S},\widetilde{\mathcal{S}}),$$

以及给定的 $\widetilde{\boldsymbol{V}}$ 和 $\boldsymbol{V}$, 对 Frobenius 范数和谱范数均有

$$\|\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{V}},\boldsymbol{V})\|\leqslant\min_{\boldsymbol{R}\in\mathcal{O}(K)}\|\widetilde{\boldsymbol{V}}\boldsymbol{R}-\boldsymbol{V}\|\leqslant\sqrt{2}\|\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{V}},\boldsymbol{V})\|,$$

其中 $\mathcal{O}(K)$ 是由所有 $K\times K$ 正交阵构成的空间, 不等式左侧的最小值在 $\widetilde{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{V}$ 的奇异值处取得, 详情可参见文献 [91].

定理 5 (Davis-Kahan $\sin\theta$ 定理[41]) 假设 $\boldsymbol{A},\widetilde{\boldsymbol{A}}\in\mathbb{R}^{n\times n}$ 对称, $\boldsymbol{V},\widetilde{\boldsymbol{V}}\in\mathbb{R}^{p\times K}$ 分别是 $\boldsymbol{A}$ 和 $\widetilde{\boldsymbol{A}}$ 的特征向量组成的标准正交阵. 令 $\mathcal{L}(\boldsymbol{V})$ 是 $\boldsymbol{V}$ 中特征向量对应的特征值集合, $\mathcal{L}(\boldsymbol{V}^{\perp})$ 是不在 $\boldsymbol{V}$ 中特征向量对应的特征值集合. 若存在区间 $[\alpha,\beta]$ 和 $\delta>0$ 使得 $\mathcal{L}(\boldsymbol{V})\subset[\alpha,\beta]$, $\mathcal{L}(\widetilde{\boldsymbol{V}}^{\perp})\subset(-\infty,\alpha-\delta]\cup[\beta+\delta,+\infty)$, 则对于任意正交不变范数有

$$\|\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{V}},\boldsymbol{V})\|\leqslant\delta^{-1}\|(\widetilde{\boldsymbol{A}}-\boldsymbol{A})\boldsymbol{V}\|.$$

基于这一经典结果, 在结构性假设下可以得到更精确的元素形式 (ℓ_∞) 的估计上限, 也可以推导出类似 Wedin 定理[92] 的矩形矩阵的上界. 关于这方面的成果可以参见文献 [91,93–98]. 然而, 定理中常数 δ 同时依赖 $\boldsymbol{A}$ 和 $\widetilde{\boldsymbol{A}}$ 的特征值, 这给实际使用带来了不小的困难, 为此利用 Weyl 定理 $\max_{1\leqslant j\leqslant n}|\lambda_j(\widetilde{\boldsymbol{A}})-\lambda_j(\boldsymbol{A})|\leqslant\|\widetilde{\boldsymbol{A}}-\boldsymbol{A}\|_2$ 就可以得到下面这个更加实用的推论.

推论 1 假设与定理 5 条件相同, 且 $\mathcal{L}(\widetilde{\boldsymbol{V}})$ 中的特征值与 $\mathcal{L}(\boldsymbol{V})$ 中的特征值具有相同的位次 (即特征值由大到小的排位次序). 若 $\mathcal{L}(\boldsymbol{V})\subset[\alpha,\beta]$, $\mathcal{L}(\boldsymbol{V}^{\perp})\subset(-\infty,\alpha-\delta_0]\cup[\beta+\delta_0,+\infty)$, $\delta_0>0$, 则

$$\|\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{V}},\boldsymbol{V})\|_2\leqslant2\,\delta_0^{-1}\|\widetilde{\boldsymbol{A}}-\boldsymbol{A}\|_2.$$

利用范数不等式 $\|\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{V}},\boldsymbol{V})\|_F\leqslant\sqrt{K}\|\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{V}},\boldsymbol{V})\|_2$ 可推导出 Frobenius 范数下的结论. 考虑最特殊的情形 $\mathcal{L}=\{\lambda\}$, $\boldsymbol{V}=\boldsymbol{v}$, $\widetilde{\boldsymbol{V}}=\widetilde{\boldsymbol{v}}$ 退化为两个向量, 令 $\alpha=\beta=\lambda$, 则有

$$\min_{s=\pm1}\|\widetilde{\boldsymbol{v}}-s\boldsymbol{v}\|_2\leqslant\sqrt{2}\sin\theta(\widetilde{\boldsymbol{v}},\boldsymbol{v})\leqslant2\sqrt{2}\,\delta_0^{-1}\|\widetilde{\boldsymbol{A}}-\boldsymbol{A}\|_2.$$

回到因子模型上, 当总体协方差阵存在足够大的谱隙 (eigengap) 时, 我们可以确定地说主成分分析与因子模型是近似相同的. 由可识别性条件 1 得到 $\boldsymbol{\Sigma}=\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}+\boldsymbol{\Sigma}_u$, 令 Weyl 定理和推论 1 中的 $\boldsymbol{A}=\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$, $\widetilde{\boldsymbol{A}}=\boldsymbol{\Sigma}$, 则特征值 (向量) 之间差异的上界为 $\|\boldsymbol{\Sigma}_u\|_2$, 这种差异可以解释为主成分分析对近似因子模型的偏倚; 另外, 谱隙在泛在性条件 2 下也相对较小.

进一步, 对于任何协方差矩阵的估计, 包括我们之前提到的各种稳健化估计, 同样也可以运用上述方法对特征值 (向量) 的估计进行分析, 即令 $\boldsymbol{A}=\boldsymbol{\Sigma}$, $\widetilde{\boldsymbol{A}}=\widehat{\boldsymbol{\Sigma}}$, 得到子空间估计误差的上界 $\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_2/\delta_0$. 为简单起见, 接下来介绍特征向量之差 $\widetilde{\boldsymbol{v}}-\boldsymbol{v}$ 的逐元素上界, 而非一般的特征空间. 在许多情形下, 特征向量中并不存在主扰动项, 利用朴素不等式 $\|\cdot\|_\infty\leqslant\|\cdot\|_2$ 和 Davis-Kahan 定理只能得到次最优的结果. 文献 [96] 通过如下形式的逐项上界解决了这一问题:

$$|[\widetilde{\boldsymbol{v}}-\boldsymbol{v}]_m|\lesssim\mu\,\delta_0^{-1}\|\widetilde{\boldsymbol{A}}-\boldsymbol{A}\|_2+\text{小项},\quad\forall\,m\in[n],$$

这里 $[\cdot]_m$ 表示向量的第 m 个分量, $\mu \in [0,1]$ 是一个与对应统计问题相关的常数, 在高维设定下通常具有 $O(1/\sqrt{n})$ 的阶. 最后的小项通常与数据的独立性模式有关, 在较弱的独立性条件下, 其数值一般很小.

严格来说, 假设 $\widetilde{\boldsymbol{A}}$、$\boldsymbol{A}$ 和 $\boldsymbol{W}$ 均为 $n \times n$ 对称阵, 满足 $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{W}$, $\operatorname{rank}(\boldsymbol{A}) = K < n$, 将其特征值按绝对值从大到小排列 (允许特征值为负) 得到特征分解

$$\boldsymbol{A} = \sum_{k=1}^{K} \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k^{\mathrm{T}}, \quad \widetilde{\boldsymbol{A}} = \sum_{k=1}^{K} \widetilde{\lambda}_k \widetilde{\boldsymbol{v}}_k \widetilde{\boldsymbol{v}}_k^{\mathrm{T}} + \sum_{k=K+1}^{n} \widetilde{\lambda}_k \widetilde{\boldsymbol{v}}_k \widetilde{\boldsymbol{v}}_k^{\mathrm{T}},$$

$\{\boldsymbol{v}_k\}_{k=1}^K$ 和 $\{\widetilde{\boldsymbol{v}}_k\}_{k=1}^n$ 为标准化特征向量. 当扰动 $\boldsymbol{W}$ 不太大时, 由 Weyl 不等式, $\{\widetilde{\boldsymbol{v}}_k\}_{k=1}^K$ 和 $\{\widetilde{\boldsymbol{v}}_k\}_{k=K+1}^n$ 可以被分离开. 令 $\lambda_0 = +\infty, \lambda_{K+1} = -\infty$, 定义 λ_k 与其他特征值之间的最小距离为谱隙

$$\delta_k = \min\{\lambda_{k-1} - \lambda_k, \lambda_k - \lambda_{k+1}, |\lambda_k|\}, \quad \forall\, k \in [K].$$

这一定义在 $\mathcal{L} = \{\lambda_k\}$ 时与推论 1 中的谱隙定义是一致的, 即只关心单个特征值及其对应的特征向量. 当扰动无偏 $\mathrm{E}\,\boldsymbol{W} = 0$ 时, 文献 [96] 严格建立了随机向量 $\widetilde{\boldsymbol{v}}_k$ 的一阶近似

$$\widetilde{\boldsymbol{v}}_k = \widetilde{\lambda}_k^{-1} \widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{v}}_k \approx \lambda_k^{-1} \widetilde{\boldsymbol{A}} \boldsymbol{v}_k = \boldsymbol{v}_k + \lambda_k^{-1} (\widetilde{\boldsymbol{A}} - \boldsymbol{A}) \boldsymbol{v}_k.$$

文献 [25] 简化了文献 [96] 的方法, 在更一般的条件下建立了同样的结论. 在叙述前引入一些新的记号. 对于任一 $m \in [n]$, 设 $\boldsymbol{W}^{(m)} = \{W_{ij}^{(m)}\}_{i,j=1}^n, W_{ij}^{(m)} = W_{ij} I_{\{i \neq m\}} I_{\{i \neq m\}}, i, j \in [n]$, $\widetilde{\boldsymbol{A}}^{(m)} = \boldsymbol{A} + \boldsymbol{W}^{(m)}$, 以及对应的特征值为 $\{\widetilde{\lambda}_k^{(m)}\}_{k=1}^n$ 和特征向量为 $\{\widetilde{\boldsymbol{v}}_k^{(m)}\}_{k=1}^n$. 这种构造与概率论和统计中的留一技术 (leave-one-out) 有关, 近期关于此方法的使用参见文献 [96, 99, 100].

定理 6 给定 $\ell \in [K]$, 假设 $|\lambda_\ell| \asymp \max_{k \in [K]} |\lambda_k|$, 谱隙 $\delta_\ell \geqslant 5 \|\boldsymbol{W}\|_2$. 记 $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$, 则通过调整特征向量各分量的正负, 总有

$$|[\widetilde{\boldsymbol{v}}_\ell - \boldsymbol{v}_\ell]_m| \lesssim \frac{\|\boldsymbol{W}\|_2}{\delta_\ell} \left(\sum_{k=1}^{K} [\boldsymbol{v}_k]_m^2 \right)^{1/2} + \frac{|\langle \boldsymbol{w}_m, \widetilde{\boldsymbol{v}}_\ell^{(m)} \rangle|}{\delta_\ell}, \quad \forall\, m \in [n]. \tag{3.19}$$

与定理 5 中标准的 ℓ_2-上界 $\|\widetilde{\boldsymbol{v}}_\ell - \boldsymbol{v}_\ell\|_2 \lesssim \delta_\ell^{-1} \|\boldsymbol{W}\|_2$ 相比, 上界 (3.19) 表明第 m 个分量的扰动会更加小, 第一项中的因子 $(\sum_{k=1}^K [\boldsymbol{v}_k]_m^2)^{1/2}$ 通常远小于 1, 以单位球面的均匀分布为例, 其阶数为 $O(\sqrt{K \log n / n})$. (3.19) 中的第二项通常也远小于 $\delta_\ell^{-1} \|\boldsymbol{W}\|_2$, 例如, $\boldsymbol{w}_m$ 服从独立同分布标准正态分布, 则 $|\langle \boldsymbol{w}_m, \widetilde{\boldsymbol{v}}_\ell^{(m)} \rangle| = O_P(1)$, 然而 $\|\boldsymbol{W}\|_2$ 通常有 $\sqrt{n}$ 的阶. 尽管这里只给出了第 m 个元素的上界, 只要对所有 $m \in [n]$ 加入独立性假设 (如随机图), 结论就可以拓展到 ℓ_∞ 范数. 文献 [96] 将这一结果拓展为特征空间的扰动上界, 文献 [95, 97, 101] 利用一定的随机矩阵的理论进一步放宽关于特征值的条件.

最后介绍奇异值向量的扰动结果, 采用类似的记号 $\widetilde{\boldsymbol{L}}, \boldsymbol{L}, \boldsymbol{E} \in \mathbb{R}^{n \times p}$, $\widetilde{\boldsymbol{L}} = \boldsymbol{L} + \boldsymbol{E}, \operatorname{rank}(\boldsymbol{L}) = K < \min\{n, p\}$, 有奇异值分解

$$\boldsymbol{L} = \sum_{k=1}^{K} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^{\mathrm{T}}, \quad \widetilde{\boldsymbol{L}} = \sum_{k=1}^{K} \widetilde{\sigma}_k \widetilde{\boldsymbol{u}}_k \widetilde{\boldsymbol{v}}_k^{\mathrm{T}} + \sum_{k=K+1}^{\min\{n,p\}} \widetilde{\sigma}_k \widetilde{\boldsymbol{u}}_k \widetilde{\boldsymbol{v}}_k^{\mathrm{T}},$$

$\widetilde{\sigma}_k$ 和 σ_k 为降序排列奇异值, $\boldsymbol{u}_k$ 和 $\boldsymbol{v}_k$ 为标准化奇异值向量, 定义 $\sigma_0 = +\infty$, $\sigma_{K+1} = 0$, 谱隙为

$$\gamma_k = \min\{\sigma_{k-1} - \sigma_k, \sigma_k - \sigma_{k+1}\}, \quad \forall\, k \in [K].$$

对于任意 $i \in [n]$, $j \in [p]$, 记 $\{\widetilde{\boldsymbol{u}}_k^{(j)}\}_{k=1}^{\min\{n,p\}} \subseteq \mathbb{R}^n$ 和 $\{\widetilde{\boldsymbol{v}}_k^{(i)}\}_{k=1}^{\min\{n,p\}} \subseteq \mathbb{R}^p$ 分别为用零向量替换掉对应的矩阵 $\boldsymbol{E}$ 中的第 j 列 (记为 $\boldsymbol{e}_j^{\mathrm{col}}$) 与第 i 行 (记为 $\boldsymbol{e}_i^{\mathrm{row}}$) 得到的单位向量 (先替换, 后分解, 再标准化).

推论 2 给定 $\ell \in [K]$, 假设 $\sigma_\ell \asymp \max_{k\in[K]}\sigma_k$, 谱隙 $\gamma_\ell \geqslant 5\|\boldsymbol{E}\|_2$, 则通过调整奇异值向量各分量的正负, 总有

$$\begin{aligned} |[\widetilde{\boldsymbol{u}}_\ell - \boldsymbol{u}_\ell]_i| &\lesssim \frac{\|\boldsymbol{E}\|_2}{\gamma_\ell}\bigg(\sum_{k=1}^K [\boldsymbol{u}_k]_i^2\bigg)^{1/2} + \frac{|\langle (\boldsymbol{e}_i^{\mathrm{row}})^{\mathrm{T}}, \widetilde{\boldsymbol{v}}_\ell^{(i)}\rangle|}{\gamma_\ell}, \quad \forall\, i\in[n], \\ |[\widetilde{\boldsymbol{v}}_\ell - \boldsymbol{v}_\ell]_j| &\lesssim \frac{\|\boldsymbol{E}\|_2}{\gamma_\ell}\bigg(\sum_{k=1}^K [\boldsymbol{v}_k]_j^2\bigg)^{1/2} + \frac{|\langle \boldsymbol{e}_j^{\mathrm{col}}, \widetilde{\boldsymbol{u}}_\ell^{(j)}\rangle|}{\gamma_\ell}, \qquad \forall\, j\in[p]. \end{aligned} \tag{3.20}$$

这个推论为分析低秩矩阵的奇异空间估计误差提供了有力的工具. 对应到模型中, $\widetilde{\boldsymbol{L}}$ 为观测数据矩阵 ($\boldsymbol{X}$), 低秩矩阵 $\boldsymbol{L}$ 为 $\boldsymbol{B}\boldsymbol{F}^{\mathrm{T}}$. 以 $K=1$ 为例, 假设 $\boldsymbol{E}$ 服从独立同分布标准多元正态, 得到具有噪声的观测 $\widetilde{\boldsymbol{L}} = \boldsymbol{L} + \boldsymbol{E}$, $\boldsymbol{L} = \sigma_1 \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$. 这是一个只有单一尖峰的尖峰矩阵模型. 由 $\boldsymbol{e}_i^{\mathrm{row}}$ 和 $\widetilde{\boldsymbol{v}}_\ell^{(i)}$ 的独立性, 推论 2 指出以 $1-o(1)$ 的概率, 至少存在一组可能的奇异值向量正负号选择, 使得

$$\|\widetilde{\boldsymbol{u}}_1 - \boldsymbol{u}_1\|_\infty \leqslant \sigma_1^{-1}\|\boldsymbol{E}\|_2\|\boldsymbol{u}_1\|_\infty + \sigma_1^{-1}O(\sqrt{\log n}). \tag{3.21}$$

由随机矩阵结论 $\|\boldsymbol{E}\|_2 \asymp \sqrt{n}+\sqrt{p}$, 则有 $\|\widetilde{\boldsymbol{u}}_1 - \boldsymbol{u}_1\|_\infty \leqslant \sigma_1^{-1}\|\boldsymbol{E}\|_2$. 由于 $\|\boldsymbol{u}_1\|_\infty$ 在高维设定下远小于 1, 这一上界比 (3.21) 中的大得多, 因此相较于 ℓ_2 范数, (3.21) 给出了更精确的逐元素控制上界. 除此之外, 推论 2 还有许多优点: 首先, 允许 K 适当大一些; 其次, 结果是确定性的, 因此对随机矩阵也适用; 最后, 结果对于每一个 $i\in[n]$, $j\in[p]$ 都保持不变, 因此即使 $\boldsymbol{E}$ 各项不独立 (例如, 协变量的一个子集是相关的), 结果仍然成立.

3.6.2 渐近性质

本小节将介绍因子模型中参数估计的一些渐近性质, 主要讨论因子载荷矩阵估计 $\widehat{\boldsymbol{B}}$、潜因子估计 $\{\widehat{\boldsymbol{f}}_i\}_{i=1}^n$、POET 估计 $\widehat{\boldsymbol{\Sigma}}_\lambda$ 和 $\widehat{\boldsymbol{\Sigma}}_\lambda^\tau$ 的渐近性质, 这些性质主要由文献 [18,26] 建立.

(1) 因子载荷矩阵的估计性质

POET 方法步骤 1 中 $\widehat{\boldsymbol{B}}$ 由 $\widetilde{\lambda}_j = \|\widetilde{\boldsymbol{b}}_j\|^2$ 和 $\widetilde{\boldsymbol{v}}_j = \widetilde{\boldsymbol{b}}_j/\|\widetilde{\boldsymbol{b}}_j\|$ 两部分组成. 由命题 1 和条件 2, 若 $\|\boldsymbol{\Sigma}_u\| = o(p)$, 则有 $\widetilde{\lambda}_j = \Omega(p)$; 又若 $\|\boldsymbol{B}\|_{\max}$ 有界, 则 $\|\widetilde{\boldsymbol{v}}_j\|_{\max} = \Omega(p^{-1/2})$, 即 $\{\widetilde{\boldsymbol{v}}_j\}_{j=1}^K$ 不存在尖峰分量, 进而 $\{\boldsymbol{v}_j\}_{j=1}^K$ 也不存在尖峰分量. 使用文献 [102] 中的记号, 记协方差阵 $\boldsymbol{\Sigma}$ 及其主特征值 $\boldsymbol{\Lambda}$ 和对应的主特征向量 $\boldsymbol{\Gamma}$ 的初始估计分别为 $\widehat{\boldsymbol{\Sigma}}$、$\widehat{\boldsymbol{\Lambda}} = \mathrm{diag}(\widehat{\lambda}_1,\ldots,\widehat{\lambda}_K)$ 和 $\widehat{\boldsymbol{\Gamma}}_{p\times K} = (\widehat{\boldsymbol{v}}_1,\ldots,\widehat{\boldsymbol{v}}_K)$, 其满足

$$\begin{aligned} \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} &= O_P\bigg(\sqrt{\frac{\log p}{n}}\bigg), \\ \|(\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda})\boldsymbol{\Lambda}^{-1}\|_{\max} &= O_P\bigg(\sqrt{\frac{\log p}{n}}\bigg), \\ \|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_{\max} &= O_P\bigg(\sqrt{\frac{\log p}{np}}\bigg). \end{aligned} \tag{3.22}$$

为简单起见, 计算 $\widehat{\boldsymbol{\Sigma}}_u = \widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}$.

定理 7 假设前 K 个特征值可区分且满足泛在性条件 2, $\|\boldsymbol{B}\|_{\max}$ 和 $\|\boldsymbol{\Sigma}_u\|_2$ 有界. 若 (3.22) 成立, 取 $w_n = K^2(\sqrt{\log p/n} + 1/\sqrt{p})$, 则得到

$$\begin{aligned} \|\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}} - \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\|_{\max} &= O_P(w_n), \\ \|\widehat{\boldsymbol{B}} - \boldsymbol{B}\|_{\max} &= O_P(K^{-2}w_n). \end{aligned} \tag{3.23}$$

这里 w_n 中的 K^2 反映了学习 K 个潜因子所付出的代价, $1/\sqrt{p}$ 反映了通过主成分分析近似因子载荷的偏差, $\sqrt{\log p/n}$ 反映了主成分估计的随机误差. 定理 7 成立的关键是条件 (3.22) 能否满足. 文献 [26] 证明了将样本协方差阵作为初始估计, 在次高斯和弱相依条件下, 条件 (3.22) 满足; 另外, 文献 [25] 证明了对于边际和空间 Kendall-τ 估计, 条件 (3.22) 均成立; 文献 [94] 还证明了对于逐项自适应 Huber 估计, 条件 (3.22) 依然成立.

(2) 协方差矩阵的估计性质

下述定理是文献 [26,102] 中关于 $\boldsymbol{\Sigma}_u$ 和 $\boldsymbol{\Sigma}$ 估计的推广.

定理 8 假设定理 7 中的条件仍然成立, 若 $\boldsymbol{\Sigma}_u \in \mathcal{C}_q(m_p)$ (3.5), $m_p\, w_n^{1-q} = o(1)$, $\|\boldsymbol{\Sigma}^{-1}\|_2 = O(1)$, 则对于 $\boldsymbol{\Sigma}_u$ 的估计, 有

$$\begin{aligned} &\|\widehat{\boldsymbol{\Sigma}}_u^{\tau} - \boldsymbol{\Sigma}_u\|_{\max} = O_P(w_n), \\ &\|\widehat{\boldsymbol{\Sigma}}_u^{\tau} - \boldsymbol{\Sigma}_u\|_2 = O_P(m_p\, w_n^{1-q}), \\ &\|(\widehat{\boldsymbol{\Sigma}}_u^{\tau})^{-1} - \boldsymbol{\Sigma}_u^{-1}\|_2 = O_P(m_p\, w_n^{1-q}), \end{aligned} \tag{3.24}$$

对于 $\boldsymbol{\Sigma}$ 的估计, 有

$$\begin{aligned} &\|\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma}\|_{\max} = O_P(w_n), \\ &\|\widehat{\boldsymbol{\Sigma}}^{\tau} - \boldsymbol{\Sigma}\|_{\Sigma} = O_P\bigg(K^{3/2}p^{1/2}\frac{\log p}{n} + m_p\, w_n^{1-q} + \frac{Kw_n}{p}\bigg), \\ &\|(\widehat{\boldsymbol{\Sigma}}^{\tau})^{-1} - \boldsymbol{\Sigma}^{-1}\|_2 = O_P(K^2 m_p\, w_n^{1-q}). \end{aligned} \tag{3.25}$$

条件中的 $\|\boldsymbol{\Sigma}^{-1}\|_2 = O(1)$ 只是用来证明逆矩阵相关的结论, 并无它用. 当 $K = O(p/m_p)$ 时, (3.25) 中的第三项会被第二项控制. 当 $K < \infty$ 时, 定理 8 指出为了学习潜因子需要付出额外 $1/\sqrt{p}$ 的代价, 当 $p \gg n\log p$ 时, 这种代价则小到可以被忽略.

(3) 已实现潜因子的估计性质

定理 9 假设定理 7 中的条件仍然成立, 若 $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| = O_P(\sqrt{p\log p/n})$, 则对于已实现潜因子 $\boldsymbol{f}_i$ 的估计, 有

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\|\widehat{\boldsymbol{f}}_i - \boldsymbol{f}_i\| = O_P\bigg(\frac{K}{\sqrt{p}} + K\sqrt{\frac{\log p}{n}}\bigg), \\ &\frac{1}{n}\sum_{i=1}^{n}\|\widehat{\boldsymbol{f}}_i - \boldsymbol{f}_i\|^2 = O_P\bigg(\frac{K^2}{p} + \frac{K^2\log p}{n}\bigg), \\ &\max_{i\leqslant n}\|\widehat{\boldsymbol{f}}_i - \boldsymbol{f}_i\| = O_P\Big(K^{-3/2}p^{-1/2}w_n \max_{i\leqslant n}\|\boldsymbol{x}_i - \boldsymbol{\mu}\| + p^{-1}\max_{i\leqslant n}\|\boldsymbol{B}^{\mathrm{T}}\boldsymbol{u}_i\|\Big). \end{aligned} \tag{3.26}$$

当总体中包含 K 个 $O(p)$ 阶的尖峰特征值时, $\mathrm{E}\,\|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 = \mathrm{tr}(\boldsymbol{\Sigma}) = O(Kp)$, $\mathrm{E}\,\big\|\boldsymbol{B}^{\mathrm{T}}\boldsymbol{u}_i\big\|^2 = O(K^2p)$. 因此, 若 $\|\boldsymbol{x}_i - \boldsymbol{\mu}\|$ 和 $\|\boldsymbol{B}^{\mathrm{T}}\boldsymbol{u}_i\|$ 满足次高斯性, 则有

$$\begin{aligned} &\max_{i\leqslant n}\|\boldsymbol{x}_i - \boldsymbol{\mu}\| = O_P(\sqrt{Kp\log n}), \\ &\max_{i\leqslant n}\|\boldsymbol{B}^{\mathrm{T}}\boldsymbol{u}_i\| = O_P(\sqrt{K^2p\log n}), \end{aligned} \tag{3.27}$$

对于 (3.26) 中第三项, 则得到

$$\max_{i\leqslant n}\|\widehat{\boldsymbol{f}}_i - \boldsymbol{f}_i\| = O_P\bigg(K\sqrt{\frac{\log n}{p}} + K\sqrt{\frac{\log p\log n}{n}}\bigg).$$

进一步有如下推论.

推论 3 假设定理 9 中的条件和 (3.27) 成立, 若 $\max_{i\leqslant n}\|\boldsymbol{f}_i\| = O_P(\sqrt{K\log n})$, 则有

$$\max_{i\leqslant n,\, j\leqslant p} |\widehat{\boldsymbol{b}}_j^{\mathrm{T}}\widehat{\boldsymbol{f}}_i - \boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{f}_i| = O_P\bigg(K^{3/2}\sqrt{\frac{\log n}{p}} + K^{3/2}\sqrt{\frac{\log p\log n}{n}}\bigg).$$

(4) 个体项的估计性质

结合上述结果, 我们可以进一步建立异质项估计 $\widehat{\boldsymbol{u}}_i = \boldsymbol{x}_i - \widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{f}}_i$ 的渐近性质.

推论 4 假设推论 3 中的条件成立, 则有

$$\max_{i\leqslant n}\|\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i\|_{\max} = O_P\bigg(K^{3/2}\sqrt{\frac{\log n}{p}} + K^{3/2}\sqrt{\frac{\log p\log n}{n}}\bigg).$$

由于 $\|\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i\| \leqslant \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| + \|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{f}}_i - \boldsymbol{B}\boldsymbol{f}_i\| \leqslant \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| + \|\widehat{\boldsymbol{B}}\|\|\widehat{\boldsymbol{f}}_i - \boldsymbol{f}_i\| + \|\widehat{\boldsymbol{B}} - \boldsymbol{B}\|\,\|\boldsymbol{f}_i\|$, 再利用定理 7 和 9 可知,

$$\frac{1}{np}\sum_{i=1}^{n}\|\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i\|^2 = O_P\bigg(\frac{K^3}{p} + \frac{K^3\log p}{n}\bigg).$$

4 因子模型在统计学习中的应用

本节将利用因子模型解决一些统计学习中的相关问题, 包括因子调整方法在高维模型选择和多重检验中的应用, 以及增广因子模型在高维回归中的应用. 主成分分析是解决这类因子学习问题的重要方法, 即通过数据决定未知的潜在因子. 可以说, 主成分分析之于因子学习的意义就如同于回归分析之于统计学的意义.

4.1 带因子调节的正则化模型选择 (FarmSelect)

首先, 介绍基于因子学习的模型选择方法. 模型选择是高维回归分析中的核心问题, 在过去的二十余年间, 针对模型选择问题提出了各种各样的方法, 其中包括 LASSO、SCAD (smoothly clipped absolute deviation)、弹性网 (elastic net) 和 Dantzig 选择器等. 但是, 这些方法都要求协变量满足一定的正则条件, 如不可表示条件 (irrepresentable condition) 等, 而这些条件在协变量具有因子结构 (2.1) 时无法满足, 模型选择的相合性无法得到保证. 举一个简单例子, 模拟 100 组来自于稀疏回归模型的随机样本, 每组样本的样本量 $n = 100$, 协变量维数 $p = 250$, 考虑高维线性模型

$$Y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \quad \boldsymbol{\beta} = (\overbrace{3,\ldots,3}^{10}, \overbrace{0,\ldots,0}^{p-10})^{\mathrm{T}}, \quad \varepsilon \sim N(0, 0.3^2), \tag{4.1}$$

其中 $\boldsymbol{x} \sim N(0, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ 是一个复合对称 (compound symmetry) 的相关系数矩阵, 相关系数为 ρ, 应用 LASSO 方法并记录平均有效模型维数和模型选择的正确率. 分析结果 (图 3) 显示应用 LASSO 方法在 $\rho \geqslant 0.2$ 时的模型选择正确率很低, 这是对模型维数过估计导致的结果.

模型选择上的偏差源自原始变量 $\boldsymbol{x}$ 各分量间的强相关性, 若考虑 $\boldsymbol{x}$ 具有因子结构 (2.1), 模型 (4.1) 变为

$$Y = \alpha + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{f} + \varepsilon, \tag{4.2}$$

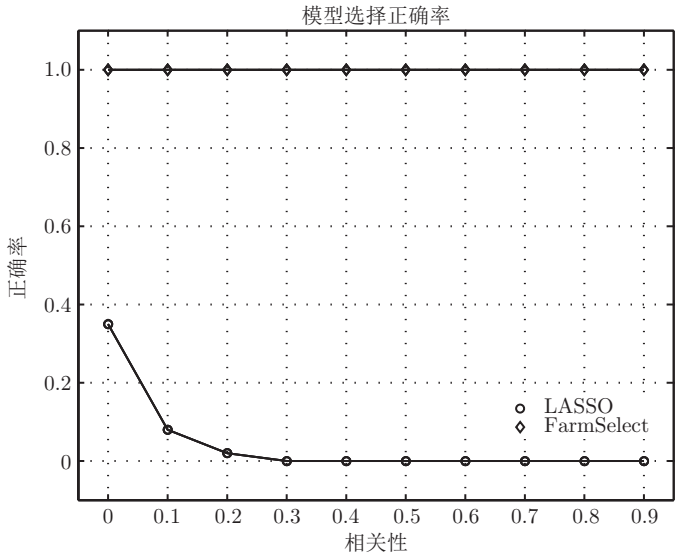

图 3 因子调节在模型选择中的有效性: 有因子调节和无因子调节对 LASSO 模型选择正确性的影响

其中 $\alpha = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{\beta}$, $\boldsymbol{\gamma} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\beta}$. 如果 $\boldsymbol{u}$ 和 $\boldsymbol{f}$ 可观测, 则 $\boldsymbol{\gamma}$ 和 α 为未知参数, 问题转换为一个 $p+1+K$ 维的回归问题, 且协变量 $\boldsymbol{f}$ 和 $\boldsymbol{u}$ 是弱相关的 (在 (4.1) 中, 它们是独立的), 进而可以利用正则化方法来拟合模型 (4.2), 得到的高维系数向量 $\boldsymbol{\beta}$ 与模型 (4.1) 中一样, 也是稀疏的. 然而在实际应用中, $\boldsymbol{u}$ 和 $\boldsymbol{f}$ 通常无法观测到, 可以通过第 3.5 小节中的方法估计, 将估计出的潜因子 $\{\boldsymbol{f}_i\}$ 和异质项 $\{\boldsymbol{u}_i\}$ 一同看作协变量做正则回归即可.

上述方法被称为带因子调节的正则化模型选择器 (factor-adjusted regularized model selector, FarmSelect). 利用因子调节来改进变量选择是为了弱化协变量之间的相关性, 图 3 展示了因子调节的功效: 无论变量间的相关性如何, 模型选择的正确率都能达到 100%. 类似的方法最早由文献 [103] 提出, 后由文献 [104] 完善并拓展. 因子个数的过估计并不会影响 FarmSelect 的效果, 因为因子较少的模型总是包含在因子更多的模型中, 新的协变量间的相关性也同样被弱化了. 因子调节方法还可以应用于更一般的稀疏模型上:

$$Y_i = g(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \varepsilon_i), \quad \text{其中 } \boldsymbol{x}_i = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{f}_i + \boldsymbol{u}_i \text{ 服从因子模型}, \tag{4.3}$$

$g(\cdot,\cdot)$ 是一个已知函数. 这类模型包含了广义线性模型和 Cox- 比例风险模型等. 可以利用损失函数 $L(Y_i, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ 来拟合模型 (4.3), 损失函数可以是对数似然函数或者 M- 估计等, 带因子调节的正则化模型选择 (FarmSelect) 由两个步骤组成, 见算法 3, 其中 $p_\lambda(\cdot)$ 为带有调优参数 λ 的折叠凹惩罚函数 (folded concave penalty function), 注意增广模型 (4.4) 中的 $\boldsymbol{\beta}$ 与模型 (4.3) 中是完全一样的, 且为稀疏

算法 3 FarmSelect

输入: 观测数据 $\{Y_i, \boldsymbol{x}_i\}_{i=1}^n$;

1. 因子分析: 对 $\{\boldsymbol{x}_i\}_{i=1}^n$ 应用算法 1 得到模型 (4.3) 中的因子估计 $\widehat{\boldsymbol{f}_i}$ 和残差 $\widehat{\boldsymbol{u}}_i$;
2. 增广正则化: 最小化惩罚经验风险函数

$$\sum_{i=1}^{n} L(Y_i, \alpha + \widehat{\boldsymbol{u}}_i^{\mathrm{T}}\boldsymbol{\beta} + \widehat{\boldsymbol{f}}_i^{\mathrm{T}}\boldsymbol{\gamma}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{4.4}$$

得到估计 $\widehat{\alpha}$、$\widehat{\boldsymbol{\beta}}$ 和 $\widehat{\boldsymbol{\gamma}}$.

的. 上述算法可以用 R 语言中的程序包 `FarmSelect` 来实现. 因子调节也可以应用于变量筛选 [104].

令 $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, $\nabla L(s,t) = \partial L(s,t)/\partial t$, 文献 [104] 证明了给定条件

$$\mathrm{E}\ \nabla L(Y, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{u} = 0 \quad \text{和} \quad \mathrm{E}\ \nabla L(Y, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{f} = 0, \tag{4.5}$$

$\mathrm{E}\ L(Y, \alpha + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{f}^{\mathrm{T}}\boldsymbol{\gamma})$ 在 $\boldsymbol{\theta}_0 = (\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\beta}^*, \boldsymbol{\beta}^{*\mathrm{T}}, (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\beta}^*)^{\mathrm{T}})^{\mathrm{T}}$ 处取到最小值; 当 $\{(\boldsymbol{u}_i^{\mathrm{T}}, \boldsymbol{f}_i)^{\mathrm{T}}\}_{i=1}^n$ 可观测时, 建立了惩罚 M- 估计的误差上界 $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\ell$ ($\ell = 1, 2, \infty$) 和符号相合性 $\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}^*)$; 最后还证明了对于广义线性模型, 增广回归 (4.4) 只受 $\{\widehat{\boldsymbol{f}}_i\}_{i=1}^n$ 所张成的特征空间估计的相合性影响, 即步骤 1 中 $\{(\widehat{\boldsymbol{u}}_i^{\mathrm{T}}, \widehat{\boldsymbol{f}}_i)^{\mathrm{T}}\}_{i=1}^n$ 的估计误差足够小, 不会影响惩罚 M- 估计 (4.4) 的相合性. 总地来说, 因子调节方法可以有效增强信噪比, 改进高维变量选择的效果.

4.2 带因子调节的多重检验 (FarmTest)

在高维统计中, 我们不仅要找到关键变量, 更要推断出这些变量的效果. 例如, 哪些基因在肿瘤与正常细胞中存在不同的表达, 或者哪些公募基金经理拥有正的 α-策略等. 这些问题都涉及高维统计中的一个热门方向—多重检验, 有关概述可参见文献 [105]. 由于基因的协同表达或基金管理中的羊群效应等因素, 不同个体检验统计量之间通常都是相关的, 这将导致无法有效控制多重检验的错误 (伪) 发现率 (参见文献 [106–108]). 因子调节方法为处理这一问题提供了有力工具.

4.2.1 多重检验与伪发现率控制

首先简要回顾多重检验及其常用准则—控制伪发现率, 这一概念类似于经典假设检验中的控制一类错误原则. 假设有 n 个来自模型的测量值,

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i, \quad \mathrm{E}\,\boldsymbol{\varepsilon}_i = 0, \quad i \in [n], \tag{4.6}$$

这里的 $\boldsymbol{x}_i$ 可以是一个 p-维向量, 可以是记录个体 i 的肿瘤与正常细胞基因表达之比, 或者是风险调整后的公募基金在第 i 期的收益等. 考虑如下多重假设检验问题:

$$H_{0j} : \mu_j = 0 \quad \text{vs.} \quad H_{1j} : \mu_j \neq 0, \quad \forall\, j \in [p]. \tag{4.7}$$

设 T_j 为 H_{0j} 的一个检验统计量, 给定临界值 $z > 0$, 有拒绝域 $|T_j| \geqslant z$. 假设检验统计量 T_j 在原假设下具有相同的分布, 否则临界值 z 应换为 z_j. 总发现数 $R(z)$ 和伪 (错误) 发现数 $V(z)$ 分别对应被拒绝的假设数和错误拒绝的假设数, 数学化定义为

$$R(z) = \#\{j : |T_j| \geqslant z\} \quad \text{和} \quad V(z) = \#\{j : |T_j| \geqslant z, \mu_j = 0\}. \tag{4.8}$$

需注意 $R(z)$ 是可以被观测的, 而 $V(z)$ 需要根据未知集合来估计, 这个未知集就是真 (正确) 假设集

$$\mathcal{S}_0 = \{j \in [p] : \mu_j = 0\}.$$

伪发现比例和伪发现率分别定义为

$$\mathrm{FDP}(z) = \frac{V(z)}{R(z)} \quad \text{和} \quad \mathrm{FDR}(z) = \mathrm{E}(\mathrm{FDP}(z)), \tag{4.9}$$

约定 $0/0=0$. 我们的目标是控制伪发现比例或伪发现率, 伪发现比例与当前的数据直接相关. 当检验统计数据相互独立, 且 p 很大时, 根据大数定律近似成立

$$\mathrm{FDP}(z)\approx\frac{\mathrm{E}\,V(z)}{\mathrm{E}\,R(z)}\approx\mathrm{FDR}(z).$$

但当观测值相依时, 这一结果就非常不可靠了. 设 P_j 是第 j 个假设的 p- 值, $P_{(1)}\leqslant\cdots\leqslant P_{(p)}$ 是排序后的 p- 值, 第 j 个假设的显著性 (或第 j 个变量的重要性) 可根据它们的 p- 值大小来衡量—越小越重要. 要将伪发现率控制在给定的 α-水平, 文献 [109] 建议选择

$$\widehat{k}=\max\left\{j:P_{(j)}\leqslant\frac{j}{p}\alpha\right\},\tag{4.10}$$

并且拒绝所有原假设 $H_{0,(j)}$, $j=1,\ldots,\widehat{k}$. 该文献证明了, 在 p- 值独立假设下, 这一过程能将伪发现率控制在 α-水平, 该过程也可以通过定义 Benjamini-Hochberg (B-H) 调整后 p- 值来实现:

$$\text{第 }j\text{ 个假设的调整后 p- 值}=\min_{k\geqslant j}P_{(k)}\frac{p}{k}.\tag{4.11}$$

取最小值为保持单调性, 当调整后 p- 值小于 α 时就拒绝原假设. R 语言中的函数 `p.adjust` 可实现 B-H 方法. 为了更好地理解上述过程, 令 p- 值表示检验过程: 给定置信水平 t, 当 $P_j\leqslant t$ 时拒绝 H_{0j}, 分别可以得到拒绝数和伪发现数分别为

$$r(t)=\sum_{j=1}^{p}I(P_j<t)\quad\text{和}\quad v(t)=\sum_{j\in\mathcal{S}_0}I(P_j<t),\tag{4.12}$$

以及检验的伪发现 (拒绝) 比例 $\mathrm{FDP}(t)=v(t)/r(t)$. 注意, 在原假设下, P_j 是服从均匀分布的. 若检验统计量 T_j 独立, 那么序列 $\{I(P_j<t)\}$ 为独立同分布成功概率为 t 的 Bernoulli 随机变量. 因此, $v(t)\approx p_0\,t$, 其中 $p_0=|\mathcal{S}_0|$ 是正确原假设的数目, 虽然未知, 但其存在上界 p, 则有估计为

$$\widehat{\mathrm{FDP}}(t)=\frac{pt}{r(t)}.\tag{4.13}$$

B-H 方法取 $t=P_{(\widehat{k})}$, 由 (4.10) 得到

$$\widehat{\mathrm{FDP}}(P_{(\widehat{k})})=\frac{pP_{(\widehat{k})}}{\widehat{k}}\leqslant\alpha,$$

这就解释了为什么 B-H 方法能将伪发现率控制在 α-水平.

如上所述, 将 $\pi_0=p_0/p$ 上界设为 1, 虽然能正确控制伪发现率, 但过于粗糙. 为此, 文献 [110] 提出了一种有效估计 π_0 的方法, 观察 p 个假设的 p- 值直方图 (参见图 4), 它混合了真假设 $\mathcal{S}_0$ 和伪 (错误) 假设 $\mathcal{S}_0^c$, 若伪假设的 p- 值都很小, 则超过给定阈值 $\eta\in(0,1)$ 的 p- 值就都来自真假设. 在这种情形下, $\pi_0(1-\eta)$ 应该与超过 η 的 p- 值百分比大致相同—这就引出了估计值: 对于给定的 η,

$$\widehat{\pi}_0(\eta)=\frac{1}{(1-\eta)p}\sum_{j=1}^{p}I(P_j>\eta).\tag{4.14}$$

在文献 [109] 关于 FDR 的开创性工作之后, 出现了大量关于独立检验 FDR 的文献. 重要的成果包括文献 [110–112] 等. 正确假设的比例估计该领域中另一个备受关注的问题, 可以参见文献 [110,113–115] 等.

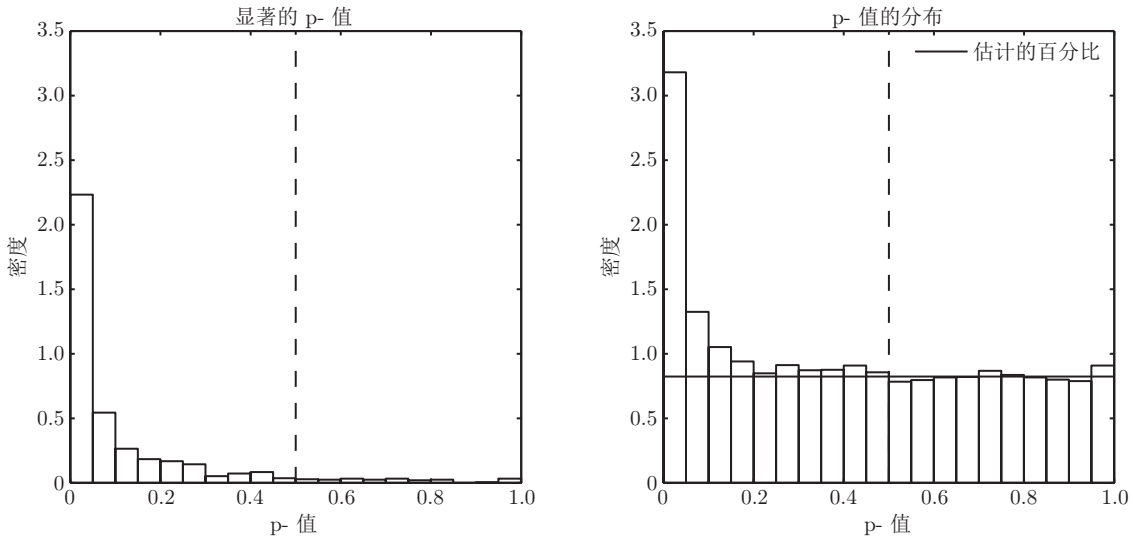

图 4　正确假设比例估计. 观察到的 p- 值 (b) 由显著变量 (通常较小) 和不显著变量 (均匀分布) 组成. 假设显著变量的 p- 值大多不超过 η (图中设为 0.5), 观察到的大于 η 的 p- 值的贡献多半来自正确假设, 这产生了一个自然的估计: p- 值大于 η 的直方图的平均高度 (实线). 红线以上的柱状图是从 (a) 的显著变量中估计出的 p- 值的分布

4.2.2　FarmTest

文献 [109,110] 处理了多重检验的方法, 在检验统计量间存在强相关性时, 无法很好地控制伪发现率和漏发现率. 文献 [106,107] 率先指出在高相关性情形下, FDR 技术的准确性会下降. 文献 [108,116] 考虑了近似因子模型下的 FDP 估计, 文献 [117] 研究了同时具有观测变量和潜因子的复杂模型, 文献 [118,119] 考虑了其他不同的因子模型, 所有这些文献都依赖于数据的联合正态性假设, 但这在实际应用中却难以保证. 最近, 文献 [81] 开发了一种带因子调节的稳健方法来处理重尾数据的 FDP 控制问题.

由于基因的协同表达, 基因表达向量 $\boldsymbol{x}_i$ 是相依的. 同样, 由于 "羊群效应", 基金的收益在第 i 期内也是相关的. 这需要对 (4.6) 里 ε_i 中的相依结构进一步建模. 如前所述, 因子模型是对相依性建模的一种常用技术, 通过因子模型对 (4.6) 中的 ε_i 进行刻画, 对于所有 $i = 1, \ldots, n$, 得到

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \underbrace{\boldsymbol{B}\boldsymbol{f}_i + \boldsymbol{u}_i}_{\varepsilon_i}, \quad \mathrm{E}\,\boldsymbol{f}_i = 0, \quad \mathrm{E}\,\boldsymbol{u}_i = 0, \quad \mathrm{E}\,\boldsymbol{f}_i\boldsymbol{u}_i^{\mathrm{T}} = 0. \tag{4.15}$$

由于存在很强相依性, FDR 和 FDP 可能相差很大, 这使得解释更加困难. 下面的例子改编自文献 [116].

例 2　为了深入了解相依性如何影响伪发现的数量, 简化模型 (4.15), 令单因素 $f_i \sim N(0,1)$, $\boldsymbol{B} = \rho\mathbf{1}$, $\boldsymbol{u}_i \sim N(0, (1-\rho^2)\boldsymbol{I}_p)$, 且 f_i 与 $\boldsymbol{u}_i$ 独立,

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \rho f_i \mathbf{1} + \boldsymbol{u}_i.$$

假设检验问题 (4.7) 的样本平均检验统计量为

$$\boldsymbol{Z} \equiv \sqrt{n}\bar{\boldsymbol{x}} = \sqrt{n}\boldsymbol{\mu} + \rho W\mathbf{1} + \sqrt{1-\rho^2}\,\boldsymbol{U},$$

其中 $W = \sqrt{n}\bar{f} \sim N(0,1)$ 与 $\boldsymbol{U} = \sqrt{n/(1-\rho^2)}\bar{\boldsymbol{u}} \sim N(0, \boldsymbol{I}_p)$ 独立. 令 $p_0 = |\mathcal{S}_0|$ 为正确假设的个数, 那

么伪发现数为

$$V(z)=\sum_{j\in\mathcal{S}_0} I(|Z_j|>z)=\sum_{j\in\mathcal{S}_0}[I(U_j>a(z-\rho W))+I(U_j<a(-z-\rho W))],$$

其中 $a=(1-\rho^2)^{-1/2}$. 根据大数定律, 在给定 W 时, 令 $p_0\to\infty$, 得到

$$p_0^{-1}V(z)=\Phi\bigg(\frac{-z+\rho W}{\sqrt{1-\rho^2}}\bigg)+\Phi\bigg(\frac{-z-\rho W}{\sqrt{1-\rho^2}}\bigg)+o_p(1).$$

因此, 伪发现数取决于 W 的实际值. 当 $\rho=0$ 时, $p_0^{-1}V(z)\approx 2\Phi(-z)$, 此时 FDP 和 FDR 大致相同. 若令 $p_0=1{,}000$, $z=2.236$ (标准正态分布的 99% 分位数), $\rho=0.8$, 那么

$$p_0^{-1}V(t)\approx\bigg[\Phi\bigg(\frac{(-2.236+0.8W)}{0.6}\bigg)+\Phi\bigg(\frac{(-2.236-0.8W)}{0.6}\bigg)\bigg].$$

当 W 分别等于 -3、-2、-1 和 0 时, $p_0^{-1}V(z)$ 分别约为 0.608、0.145、0.008 和 0, 与 $p_0^{-1}V(z)$ 始终近似为 0.02 的独立情形形成鲜明对比.

假设 $\boldsymbol{B}$ 和 $\boldsymbol{f}_i$ 在 (4.15) 中已知, 在调整后的数据上建立检验

$$\underbrace{\boldsymbol{x}_i-\boldsymbol{B}\boldsymbol{f}_i}_{\boldsymbol{y}_i}=\boldsymbol{\mu}+\boldsymbol{u}_i. \tag{4.16}$$

这样做有两点好处: 首先, 由于异质项 $\boldsymbol{u}_i$ 是弱相依的, 因此经过因子调节的数据也是弱相依的, 此时 FDP 和 FDR 大致相同, 换言之, 更容易控制一类错误; 更重要的是, $\boldsymbol{y}_i$ 的方差小于 $\boldsymbol{x}_i$ 的方差, 即因子调节在不抑制信号的情形下降低了噪声, 因而基于 $\{\boldsymbol{y}_i\}$ 的检验功效也会增加. 为便于直觉理解, 考虑如下数值模拟.

例 3 模拟一组来自三因素模型 (4.15) 的随机样本, 样本量 $n=100$, 维数 $p=500$,

$$\boldsymbol{f}_i\sim N(0,\boldsymbol{I}_3),\quad \boldsymbol{B}=(b_{jk}),\quad b_{jk}\overset{\text{i.i.d.}}{\sim}\text{Unif}(-1,1),\quad \boldsymbol{u}_i\sim t_3(0,\boldsymbol{I}_p).$$

当 $j\leqslant p/4$ 时, $\mu_j=0.6$, 否则取 0. 计算样本均值 $\bar{\boldsymbol{x}}$, 得到 p 个检验统计量, 其直方图如图 5(a). 由于估计中存在较大随机误差, 分布原本的双峰性消失, 然而基于因子调节数据 $\{\boldsymbol{x}_i-\boldsymbol{B}\boldsymbol{f}_i\}_{i=1}^n$ 的样本平均值直方图 (图 5(b)) 则清楚地显示了双峰性, 这是由于数据中的噪声被降低. 因此, 检验的功效也得到了增强. 与此同时, 由于去掉了公共因子, 调整后的数据现在也是不相关的, 进一步, 若误差 $\boldsymbol{u}$ 服从 $N(0,3\boldsymbol{I}_p)$, 则修正后数据独立, FDR 与 FDP 近似相等.

带因子调节的检验本质上就是对 (4.16) 中调整后的数据 $\boldsymbol{y}_i$ 进行多重检验. 首先还是通过主成分分析得到 $\boldsymbol{B}\boldsymbol{f}_i$, 对 $\boldsymbol{y}_i$ 再应用 t- 检验, 然后通过 B-H 方法控制伪发现率 FDR. 为实现上述基本概念的稳健化, 不再使用样本均值, 取而代之是使用自适应 Huber 稳健方法来估计 μ_j, $j\in[p]$. 给定稳健化参数 $\tau_j>0$, 考虑如下 μ_j 的 M-估计:

$$\widehat{\mu}_j=\arg\min_{\theta}\sum_{i=1}^{n}\rho_{\tau_j}(X_{ij}-\theta),$$

其中 $\rho_\tau(u)$ 为 Huber 损失函数. 文献 [81] 证明了

$$\sqrt{n}\,(\widehat{\mu}_j-\mu_j-\boldsymbol{b}_j^{\mathrm{T}}\bar{\boldsymbol{f}}\,)\to N(0,\sigma_{u,j}),\quad \text{对于 } j\in[p] \text{ 一致成立},$$

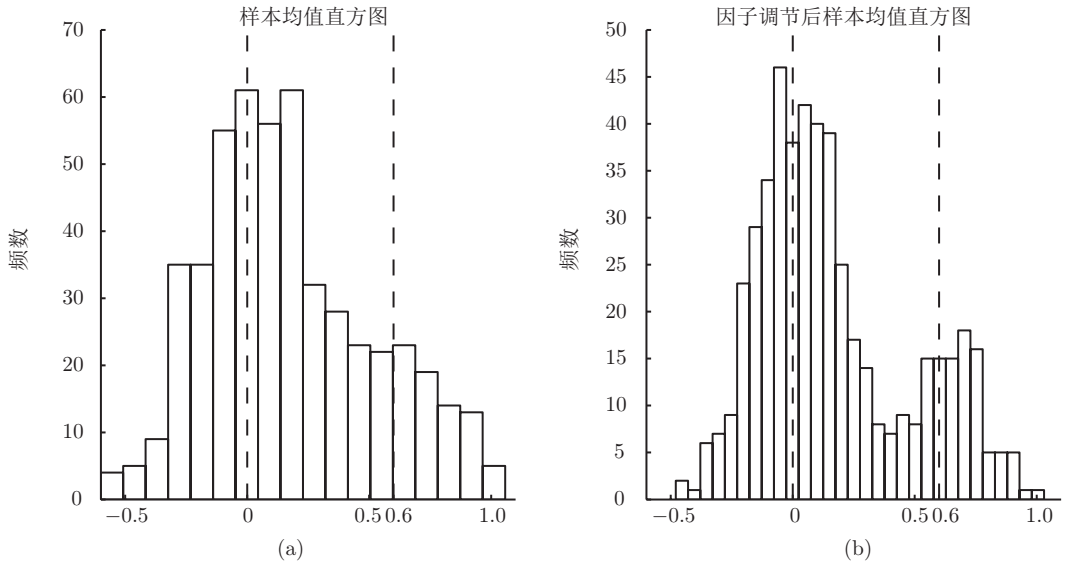

图 5　因子模型样本均值直方图. 样本的直方图来自一个合成的三因素模型, (a) 没有因子调节; (b) 加入因子调节

其中 $\bar{\boldsymbol{f}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{f}_i$, 并且 $\sigma_{u,j} = \mathrm{var}(u_j) = \mathrm{var}(X_{ij}) - \|\boldsymbol{b}_j\|_2^2$. 未知的 $\bar{\boldsymbol{f}}$ 可由如下回归问题稳健估计:

$$\bar{X}_j = \mu_j + \boldsymbol{b}_j^{\mathrm{T}}\bar{\boldsymbol{f}} + \bar{u}_j, \quad j = 1, \ldots, p. \tag{4.17}$$

如果给定 $\{\boldsymbol{b}_j\}$, 可将稀疏 $\{\mu_j\}_{j=1}^p$ 视为异常点. 因此, 给定 $\widehat{\boldsymbol{B}}$ 的一个稳健估计, 通过 (4.17) 得到 $\bar{\boldsymbol{f}}$ 的稳健估计 $\widehat{\boldsymbol{f}}$, 进而得到因子调节后的新统计量

$$T_j = \frac{\sqrt{n}(\widehat{\mu}_j - \widehat{\boldsymbol{b}}_j^{\mathrm{T}}\widehat{\boldsymbol{f}})}{\widehat{\sigma}_{u,j}},$$

$\widehat{\sigma}_{u,j}$ 也相应采用 $\sigma_{u,j}$ 的稳健估计. 这个检验统计量只是一个基于因子调节数据 $\{\boldsymbol{y}_i\}$ 的稳健化 z-检验.

由于假设异质项向量 $\boldsymbol{u}_i$ 的相关矩阵是稀疏的, 因此, $\{T_j\}$ 几乎独立. 根据大数定律可知, 伪发现数满足

$$V(z) = \sum_{j\in\mathcal{S}_0} I(|T_j| \geqslant z) \approx 2p_0\Phi(-z).$$

再由定义 (4.9), 有 $\mathrm{FDP}(z) \approx \mathrm{FDP}^A(z)$, 其中

$$\mathrm{FDP}^A(z) = \frac{2\pi_0 p\Phi(-z)}{R(z)}, \quad \pi_0 = \frac{p_0}{p}. \tag{4.18}$$

由于只有 π_0 未知, 因此, $\mathrm{FDP}^A(z)$ 几乎是已知的. 在许多应用中, 正确发现数的稀疏性使得 $\pi_0 \approx 1$, 当然也可以由 (4.14) 得到相合估计.

我们来总结 FarmTest (在算法 4 中实现): 输入包括 $\{\boldsymbol{x}_i\}_{i=1}^n$, 一个从数据中得到的稳健协方差矩阵估计 $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{p\times p}$, 一个预先指定的 FDP 控制水平 $\alpha \in (0,1)$, 因子数 K, 以及稳健化参数 γ 和 $\{\tau_j\}_{j=1}^p$. 有些参数在计算过程中可以简化. 例如, K 可由第 2.4 小节中的方法确定, 稳健化参数 $\{\tau_j\}$ 可由五分组交叉验证结合理论最优阶数确定.

算法 4 FarmTest

输入: 观测数据 $\{\boldsymbol{x}_i\}_{i=1}^n$.

1. 因子分析: 对 $\{\boldsymbol{x}_i\}_{i=1}^n$ 应用算法 1 得到因子载荷的估计 $\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{b}}_1, \ldots, \widehat{\boldsymbol{b}}_p)^{\mathrm{T}} \in \mathbb{R}^{p\times K}$.

2. 假设检验: 令 $\bar{x}_j = n^{-1}\sum_{i=1}^n x_{ij}$, $j = 1, \ldots, p$, $\widehat{\boldsymbol{f}} = \mathrm{argmax}_{\boldsymbol{f}\in\mathbb{R}^K} \sum_{j=1}^p \ell_\gamma(\bar{x}_j - \widehat{\boldsymbol{b}}_j^{\mathrm{T}}\boldsymbol{f})$, 计算检验统计量

$$T_j = \sqrt{\frac{n}{\widehat{\sigma}_{u,jj}}}(\widehat{\mu}_j - \widehat{\boldsymbol{b}}_j^{\mathrm{T}}\widehat{\boldsymbol{f}}), \quad j = 1, \ldots, p, \tag{4.19}$$

其中 $\widehat{\sigma}_{u,jj} = \widehat{\theta}_j - \widehat{\mu}_j^2 - \|\widehat{\boldsymbol{b}}_j\|_2^2$, $\widehat{\theta}_j = \mathrm{argmin}_{\theta\geqslant\widehat{\mu}_j^2+\|\widehat{\boldsymbol{b}}_j\|_2^2} \ell_{\tau_j}(x_{ij}^2 - \theta)$.

3. B-H 方法: 计算临界值 $z_\alpha = \inf\{z \geqslant 0 : \mathrm{FDP}^{\mathrm{A}}(z) \leqslant \alpha\}$, 其中 $\mathrm{FDP}^{\mathrm{A}}(z) = 2\widehat{\pi}_0\, p\, \Phi(-z)/R(z)$ (参见 (4.18)). 当 $|T_j| \geqslant z_\alpha$ 时拒绝原假设 H_{0j}.

如第 4.2.1 小节所述, FarmTest 步骤 3 与 B-H 方法控制 FDR 类似. R 语言的软件包 `FarmTest` 可实现上述过程. 尽管我们用比较直观的方式介绍了 FarmTest 方法, 但其统计有效性已被文献 [81] 严格证明, 建议读者参阅论文的细节. 在实际使用中, 文献 [81] 发现, 与不进行因子调节的方法相比, FarmTest 能够更准确地控制伪发现率, 并有效改进漏发现率; 更值得注意是, 当误差项 $\boldsymbol{u}_i$ 的分布是厚尾时, FarmTest 比其他已有的方法更胜一筹.

4.3 主成分回归

主成分回归 (principal component regression, PCR) 是利用主成分建立回归模型达到有效降维的一种方法, 而因子模型则为理解其在统计建模中的意义提供了一种新的视角. 假设原始数据中的潜因子同时驱动因变量和自变量, 如图 6 所示, 自然考虑从自变量中提取潜因子, 并用主成分估计因子作为回归变量, 因此, 这种方法被称为主成分回归 (PCR).

为了便于理解, 假设数据 $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^n$ 满足模型

$$\begin{cases} Y_i = g(\boldsymbol{f}_i) + \varepsilon_i, \\ \boldsymbol{x}_i = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{f}_i + \boldsymbol{u}_i. \end{cases} \tag{4.20}$$

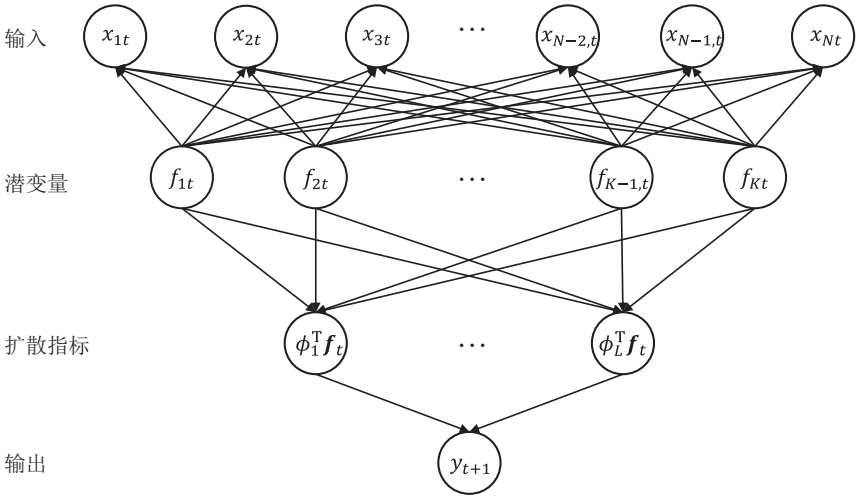

图 6 主成分回归中的数据产生机制. 预测因子 $\boldsymbol{x}_i$ 和响应变量 $\boldsymbol{y}_i$ 都受潜在因子 $\boldsymbol{f}_i$ 的驱动. 主成分回归通过 $\boldsymbol{x}$ 的主成分提取潜因子, 并将得到的估计结果 $\widehat{\boldsymbol{f}}$ 作为新的预测变量. 用 $\boldsymbol{Y}$ 对 $\widehat{\boldsymbol{f}}$ 进行回归得到主成分回归的估计值 $\boldsymbol{\theta} \in \mathbb{R}^{\boldsymbol{K}}$. 由于降维的原因, 这个估计通常具有较小的方差

该模型也适用于时间序列预测. 利用主成分分析方法得到潜因子估计 $\{\widehat{\boldsymbol{f}}_i\}_{i=1}^n$, 再拟合回归模型

$$Y_i = g(\widehat{\boldsymbol{f}}_i) + \varepsilon_i, \quad i \in [n]. \tag{4.21}$$

主成分回归通常指在模型 (4.21) 中假定 $g(\cdot)$ 为线性函数得到的多元回归

$$Y_i = \alpha + \boldsymbol{\beta}^{\mathrm{T}} \widehat{\boldsymbol{f}}_i + \varepsilon_i, \quad i \in [n].$$

由于潜因子要满足可识别性条件, 通常对 $\{\boldsymbol{f}_i\}_{i=1}^n$ 进行标准化处理, 则有最小二乘估计

$$\widehat{\alpha} = n^{-1}\sum_{i=1}^n Y_i, \quad \widehat{\boldsymbol{\beta}} = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})\widehat{\boldsymbol{f}}_i.$$

主成分回归与因子模型的出发点不尽相同, 对设计阵进行奇异值分解得 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Delta}\mathbf{V}^{\mathrm{T}}$, 其中 $\boldsymbol{\Delta} = \mathrm{diag}(\delta_1, \ldots, \delta_p)$ 为奇异值组成的对角阵, $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)$ 和 $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ 分别是 $n \times p$ 和 $p \times p$ 维正交矩阵. 主成分回归是在考虑约束最小二乘问题 [120]

$$\min_{\beta} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2, \quad \text{s.t.} \quad \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{v}_j = 0, \quad j > K. \tag{4.22}$$

在许多实用场景中, 还存在一组额外的低维变量 $\boldsymbol{w}_i$, 使得观测数据变为 $\{(Y_i, \boldsymbol{x}_i, \boldsymbol{w}_i)\}_{i=1}^n$. 首先从 $\{\boldsymbol{x}_i\}$ 中提取因子 $\{\boldsymbol{f}_i\}$, 接着使用因子和额外变量一同拟合模型

$$Y_i = g(\widehat{\boldsymbol{f}}_i, \boldsymbol{w}_i) + \varepsilon_i, \quad i \in [n]. \tag{4.23}$$

将这一类预测模型称作增广因子回归预测模型或者 FarmPredict. 当 $g(\cdot)$ 为线性函数时, 模型 (4.23) 变为多元线性回归模型

$$Y_i = \alpha + \boldsymbol{\beta}^{\mathrm{T}} \widehat{\boldsymbol{f}}_i + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{w}_i + \varepsilon_i, \quad i \in [n], \tag{4.24}$$

也被称作增广主成分回归模型. (4.23) 中的回归函数既可以通过核或样条等非参数方法进行估计, 也可以用当下时髦的深度神经网络方法, 还可以通过如图 6 所示的多指标模型去估计. 多指标模型通过多个因子线性组合 (文献 [16] 称它为扩散指标, diffusion indices) 去预测 Y. 文献 [121] 详细研究了多指标模型

$$Y_i = g(\phi_1^{\mathrm{T}}\boldsymbol{x}_i^*, \ldots, \phi_L^{\mathrm{T}}\boldsymbol{x}_i^*) + \varepsilon_i, \tag{4.25}$$

其中 $\boldsymbol{x}_i^* = (\boldsymbol{f}_i^{\mathrm{T}}, \boldsymbol{w}_i^{\mathrm{T}})^{\mathrm{T}}$ 为增广协变量, 特别当 $\boldsymbol{w}_i$ 不存在, 模型退化为图 6 中的情形. 从文献 [122,123] 对多指标模型的研究开始, 至今有关扩散指标的研究已经有了很大发展, 详情参见文献 [124,125] 及其参考文献.

4.4 增广因子模型与投影主成分分析

虽然可以通过主成分分析从数据中推断出完全未知的 $\boldsymbol{B}$ 和 $\boldsymbol{f}_i$, 然而在许多应用程序中, 我们还有可用的边带信息. 例如, 第 j 只股票对市场风险因子的因子载荷应该取决于对应公司的规模、价值、动量和波动率等属性, 我们希望利用增加的信息提取潜在因子及其相关的载荷矩阵. 为此, 文献 [126] 假设第 j 个公司的因子载荷可以部分地被额外协变量向量 $\boldsymbol{w}_j$ 解释, 并通过半参数模型对其建模

$$b_{jk} = g_k(\boldsymbol{w}_j) + \gamma_{jk}, \quad \mathrm{E}(\gamma_{jk} \mid \boldsymbol{w}_j) = 0, \tag{4.26}$$

其中 $g_k(\boldsymbol{w}_j)$ 表示第 j 个公司在第 k 个因子上可以被协变量解释的部分. 将系数 $\{b_{jk}\}$ 看作模型 (4.26) 的观测样本, 这就是一个随机效应模型, 由模型 (2.12) 得到矩阵表达

$$\boldsymbol{x}_i = [\boldsymbol{G}(\boldsymbol{W}) + \boldsymbol{\Gamma}]\boldsymbol{f}_i + \boldsymbol{u}_i. \tag{4.27}$$

类似地, 应用中也存在潜因子包含其他已知信息的情形. 例如, 虽然不能确定 Fama-French 因子 $\boldsymbol{w}_i$ 是否为真因子, 但它们至少可以解释其中的一部分, 即可以考虑模型

$$\boldsymbol{f}_i = \boldsymbol{g}(\boldsymbol{w}_i) + \boldsymbol{\gamma}_i, \quad \mathrm{E}(\boldsymbol{\gamma}_i \mid \boldsymbol{w}_i) = 0. \tag{4.28}$$

这里就将 Fama-French 因子作为了额外增加的信息 $\boldsymbol{w}_i$. 同样也有矩阵形式

$$\boldsymbol{x}_i = \boldsymbol{B}[\boldsymbol{g}(\boldsymbol{w}_i) + \boldsymbol{\gamma}_i] + \boldsymbol{u}_i.$$

若假设 $\boldsymbol{u}_i$ 具有外生性 $\mathrm{E}(\boldsymbol{u}_i \mid \boldsymbol{w}_i) = 0$, 对上式等号两边求条件期望, 得到无噪声因子模型

$$\underbrace{\mathrm{E}(\boldsymbol{x}_i \mid \boldsymbol{w}_i)}_{\widetilde{\boldsymbol{x}}_i} = \boldsymbol{B}\,\underbrace{\boldsymbol{g}(\boldsymbol{w}_i)}_{\widetilde{\boldsymbol{f}}_i}. \tag{4.29}$$

当 $n > K$ 时, $\boldsymbol{B}$ 可以通过线性方程组进行精确求解.

投影主成分分析法由文献 [126] 首先提出用来处理模型 (4.27), 又被文献 [127] 用来处理模型 (4.28). 通过将 $\boldsymbol{x}_i$ 的每一个主成分回归到 $\boldsymbol{w}_i$ 上得到拟合值 $\widehat{\boldsymbol{x}}_i$, 这一个过程可以用线性回归模型, 也可以用非参数可加模型或者核机器, 即将回归过程视为一种投影, 再对投影后的数据 $\{\widehat{\boldsymbol{x}}_i\}_{i=1}^n$ 施行主成分分解得到载荷矩阵的估计 $\widehat{\boldsymbol{B}}$ 以及 (4.29) 中的 $\widetilde{\boldsymbol{f}}_i = \boldsymbol{g}(\boldsymbol{w}_i)$ 的估计 $\widehat{\boldsymbol{g}}(\boldsymbol{w}_i)$. 进一步, 可根据 (2.14) 计算潜因子 $\widehat{\boldsymbol{f}}_i$、残差项 $\widehat{\boldsymbol{\gamma}}_i = \widehat{\boldsymbol{f}}_i - \widehat{\boldsymbol{g}}(\boldsymbol{w}_i)$ 以及 $\boldsymbol{f}_i$ 中可被附加信息 $\boldsymbol{w}_i$ 解释的百分比

$$1 - \frac{\sum_{i=1}^n \|\widehat{\boldsymbol{\gamma}}_i\|^2}{\sum_{i=1}^n \|\widehat{\boldsymbol{f}}_i\|^2}.$$

再次考虑模型 (4.27), 按照通常的表达习惯对使用的符号做少许调整,

$$X_{ji} = (\boldsymbol{g}(\boldsymbol{w}_j) + \boldsymbol{\gamma}_j)^{\mathrm{T}}\boldsymbol{f}_i + u_{ji},$$

其中 $\boldsymbol{g}(\boldsymbol{W}_j)$ 和 $\boldsymbol{\gamma}_j$ 分别表示矩阵 $\boldsymbol{G}(\boldsymbol{W})$ 和 $\boldsymbol{\Gamma}$ 的第 j 行. 同样也加入 u_{ji} 与 $\boldsymbol{W}_j$ 的外生性假设, 则得到

$$\underbrace{\mathrm{E}(X_{ji} \mid \boldsymbol{w}_j)}_{\widetilde{X}_{ji}} = \underbrace{\boldsymbol{g}(\boldsymbol{w}_j)^{\mathrm{T}}}_{\widetilde{\boldsymbol{b}}_j}\boldsymbol{f}_i. \tag{4.30}$$

与模型 (4.29) 类似, 这同样是一个关于 "新数据" $\widetilde{X}_{ji}$ 的无噪声因子模型, 但不同之处是, 潜因子 $\boldsymbol{f}_i$ 保持不变, 且回归是对于维数 j 进行的. 对于任意给定的 i, 令 $\boldsymbol{P}$ 是对应回归 $(\{X_{ji}\}_{j=1}^p, \{\boldsymbol{W}_j\}_{j=1}^p)$ 的 $p \times p$ 投影阵, 在模型 (4.27) 的等号两边同时乘以 $\boldsymbol{P}$, 若有

$$\boldsymbol{P}\boldsymbol{\Gamma} \approx 0, \quad \boldsymbol{P}\boldsymbol{u}_i \approx 0, \tag{4.31}$$

则可以得到近似无噪声模型

$$\boldsymbol{P}\boldsymbol{x}_i = [\boldsymbol{P}\boldsymbol{G}(\boldsymbol{W}) + \boldsymbol{P}\boldsymbol{\Gamma}]\boldsymbol{f}_i + \boldsymbol{P}\boldsymbol{u}_i \approx \boldsymbol{P}\boldsymbol{G}(\boldsymbol{W})\boldsymbol{f}_i.$$

这就是为什么潜因子 $\boldsymbol{f}_i$ 在模型 (4.30) 中不变的原因. 另外, 我们之前提到 $\widehat{\boldsymbol{F}}/\sqrt{n}$ 是由 $n\times n$ 矩阵 $(\boldsymbol{PX})^{\mathrm{T}}\boldsymbol{PX}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{PX}$ 的前 K 个特征值所确定, 进而得到代入估计 $\widehat{\boldsymbol{B}}=n^{-1}\boldsymbol{X}\widehat{\boldsymbol{F}}$. 根据之前的推断, 只要找到一个投影矩阵 $\boldsymbol{P}$ 能够平滑掉噪声项 $\boldsymbol{u}_i$ ($i\in[n]$) 和 $\boldsymbol{\gamma}_j$ ($j\in[p]$), 即满足条件 (4.31), 上述投影方法就成立, 其好处通过文献 [126] 中的简单例子可见一斑.

例 4 与前例类似, 考虑最简单的单因子模型, 则 (4.26) 退化为

$$X_{ij}=(g(W_j)+\gamma_j)f_i+u_{ij},$$

用局部平均 (local averaging) 估计函数 g 就得到对应的投影矩阵. 为了便于理解投影方法的益处, 不妨假设 $g(\cdot)=\beta>0$ 为一个常数, 则有

$$X_{ij}=(\beta+\gamma_j)f_i+u_{ij},$$

这时的局部平均就变成了对指标 j 做全局平均, 得到 $\bar{X}_i=(\beta+\bar{\gamma})f_i+\bar{u}_i\approx\beta f_i$. 因此, 很自然地通过等式 $\widehat{f}_i=\bar{X}_i/\widehat{\beta}$ 和标准化条件 $n^{-1}\sum_{i=1}^n\widehat{f}_i^2=1$, 得到估计

$$\widehat{\beta}=\left(n^{-1}\sum_{i=1}^n\bar{X}_i^2\right)^{1/2},\quad \widehat{f}_i=\frac{\bar{X}_i}{\widehat{\beta}}.$$

因为 $\bar{X}_i$ 是 p 个随机变量的均值, 所以不管 n 是多少, 由中心极限定理可知, $\widehat{f}_i$ 的收敛速度为 $O(1/\sqrt{p})$.

本节的最后依旧以模型 (4.26) 为例介绍投影矩阵的构造方法, 类似的方法可以推广到其他模型. 为避免 "维数灾难" 问题, 通常将 $g_k(\boldsymbol{w}_j)$ 考虑为一个非参数可加模型, 然后利用筛分基 (sieve basis) 函数去逼近这些未知函数,

$$g_k(\boldsymbol{w}_j)=\sum_{l=1}^L g_{kl}(W_{jl}),\quad g_{kl}(W_{jl})\approx\sum_{m=1}^M a_{m,kl}\phi_m(W_{jl}),$$

其中 L 是 $\boldsymbol{w}_j$ 的分量个数, $\{\phi_k(\cdot)\}_{m=1}^M$ 是筛分基函数, 其对于每一组 (k,l) 都是相同的. 为简便起见, 采用矩阵表达 $g_k(\boldsymbol{w}_j)\approx\boldsymbol{\Phi}(\boldsymbol{w}_j)^{\mathrm{T}}\boldsymbol{a}_k$, 记 $\boldsymbol{a}_k$ 为 LM 维的回归系数向量,

$$\begin{aligned}&\boldsymbol{\phi}(\boldsymbol{w}_j)=(\phi_1(W_{j1}),\ldots,\phi_M(W_{j1}),\ldots,\phi_1(W_{jL}),\ldots,\phi_M(W_{jL}))^{\mathrm{T}},\\&\boldsymbol{\Phi}(\boldsymbol{W})=(\phi(\boldsymbol{w}_1),\ldots,\phi(\boldsymbol{w}_p))^{\mathrm{T}}.\end{aligned}$$

筛分基展开, 就能够将某个给定的因变量 $\{Z_j\}_{j=1}^p$ (在应用中, 对每个给定的 i, 令 $Z_j=\boldsymbol{X}_{ij}$) 关于 $\{\boldsymbol{w}_j\}_{j=1}^p$ 的可加模型转化为一个线性模型

$$Z_j=\sum_{l=1}^L\sum_{m=1}^M a_{ml}\phi_m(W_{jl})+\varepsilon_j,\quad j=1,\ldots,p,$$

其对应的投影矩阵为 $\boldsymbol{P}=\boldsymbol{\Phi}(\boldsymbol{W})(\boldsymbol{\Phi}(\boldsymbol{W})^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{W}))^{-1}\boldsymbol{\Phi}(\boldsymbol{W})^{\mathrm{T}}$, 由 Hilbert 空间投影定理可知, 投影矩阵满足条件 (4.31), 且线性预报为 $\widehat{\boldsymbol{Z}}=\boldsymbol{PZ}$.

综上所述, 投影主成分分析用 $\sqrt{n}\boldsymbol{XPX}^{\mathrm{T}}$ 的前 K 个特征向量估计潜在因子 $\boldsymbol{F}$, 同时用 $\widehat{\boldsymbol{B}}=n^{-1}\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{F}}$ 估计因子载荷 $\boldsymbol{B}$.

5 谱方法在机器学习中的应用

因子建模和主成分是密切相关的, 主成分分析被广泛应用于统计机器学习和应用数学当中. 事实上, 主成分分析, 常被称为谱分解, 是一种特殊的谱学习方法, 用来在协方差阵估计上提取潜因子. 在接下来讨论到的应用中, 谱学习将被应用到 Wigner 矩阵类上, 这些矩阵具有随机对称性 (或者 Hermite 性), 且上三角阵中的元素相互独立. 通常假定 Wigner 矩阵中的独立元素都是均值为 0 方差有限的. 对比之下, 样本协方差矩阵中各个元素却是相关的.

5.1 社群发现

社群发现本质上是一个基于网络数据的聚类问题, 它在社交网络 (social networks)、引文网络 (citation networks)、基因网络 (genomic networks) 和经济与金融网络 (economics and financial networks) 中有着广泛的应用. 观测到的网络数据通常以 "图" 的抽象结构表示, 图中的 "结点" 代表网络中的个体, 图中的 "边" 代表个体间存在某种联系, 在数学上通常用 "邻接矩阵" $\boldsymbol{A}=(a_{ij})$ 描述 "图", 若第 i 与 j 个节点之间有联系, 则 $a_{ij}=1$, 否则 $a_{ij}=0$.

将社群发现置于概率模型框架下进行研究始于文献 [128–130] 中提出的随机分块模型 (stochastic block model). 关于这一领域最新进展可参见文献 [131].

定义 3 假设存在 n 个顶点 $\{1,\dots,n\}$ 可以被划分成 K 个互斥 (无交集) 的社群 $\mathcal{C}_1,\dots,\mathcal{C}_K$. 对任意两个顶点 $k\in\mathcal{C}_i$ 和 $l\in\mathcal{C}_j$, 随机分块模型假定它们以概率 p_{ij} 相连且与其他连接独立. 换言之, 邻接矩阵 $\boldsymbol{A}=(A_{kl})$ 可视为一系列独立 Bernoulli 随机变量的现实:

$$\mathrm{P}(A_{kl}=1)=p_{ij},\quad k\in\mathcal{C}_i,\quad l\in\mathcal{C}_j. \tag{5.1}$$

$K\times K$ 对称矩阵 $\boldsymbol{P}=(p_{ij})$ 则称为连边概率矩阵.

在随机分块模型中, 连边概率 p_{ij} 描述了两个社群 $\mathcal{C}_i$ 和 $\mathcal{C}_j$ 连通性强弱, 当在一个社群内时, 连接的概率为 p_{ii}. 当对所有 i 和 j, 都有 $p_{ij}=p$, 这个最简单的模型被称为 Erdös-Rényi 图模型 [132]. 这一模型退化为随机图, 如何分割变得不再重要, 我们通常用 $G(n,p)$ 表示它.

社群发现的目标是检测 "图" 中是否有潜在的结构, 并将其重现, 图 7 展示了一个两社群网络及其社群发现的结果. 若给定邻接矩阵 $\boldsymbol{A}=(a_{ij})$, 很自然考虑使用极大似然法. $\mathcal{C}(i)$ 表示 i 所在的社群 (未知), 观测数据 $\{a_{ij}\}_{i>j}$ 的似然函数即为 Bernoulli 似然的乘积:

$$\prod_{i>j}p_{\mathcal{C}(i)\mathcal{C}(j)}^{a_{ij}}(1-p_{\mathcal{C}(i)\mathcal{C}(j)})^{1-a_{ij}},$$

其中 $\{\mathcal{C}(i)\}_{i=1}^{K}$, $\boldsymbol{P}=(p_{ij})\in\mathbb{R}^{K\times K}$ 为未知参数. 在计算上, 这是一个 NP 难问题, 包括半正定规划和谱方法在内的一系列松弛算法被用于解决此问题, 具体内容可参见文献 [131].

谱方法基于矩方法, 设 $\boldsymbol{\Gamma}$ 表示一个 $n\times K$ 维的成员矩阵. 矩阵 $\boldsymbol{\Gamma}$ 将每个节点归属到某一个社群, 其第 i 行代表节点 i 所属的社群, 即在 $\mathcal{C}(i)$ 位置记为 1, 其他位置记为 0. 易知 $\mathrm{E}\,A_{ij}=p_{\mathcal{C}(i)\mathcal{C}(j)}$ 及低秩分解

$$\mathrm{E}\,\boldsymbol{A}=\boldsymbol{\Gamma}\boldsymbol{P}\boldsymbol{\Gamma}^{\mathrm{T}},\quad \boldsymbol{A}=\boldsymbol{\Gamma}\boldsymbol{P}\boldsymbol{\Gamma}^{\mathrm{T}}+(\boldsymbol{A}-\mathrm{E}\,\boldsymbol{A}). \tag{5.2}$$

大致而言, 矩阵 $\boldsymbol{\Gamma}$ 扮演着与因子或载荷矩阵 (非标准化) 类似的角色, $\boldsymbol{A}-\mathrm{E}\,\boldsymbol{A}$ 类似于噪声项 (异质成分). 由此可得示性矩阵 $\boldsymbol{\Gamma}$ 将落入由 $\mathrm{E}\,\boldsymbol{A}$ 前 K 个特征向量张成的特征空间中, 而且由于同个社群内成

图 7　两社群网络

员的可换性, $\mathrm{E}\boldsymbol{A}$ 的前 K 个特征向量只有 K 行不同的值. 事实上, 如果 $\boldsymbol{P}$ 非退化, $\boldsymbol{\Gamma}$ 的列向量空间与 $\mathrm{E}\boldsymbol{A}$ 的前 K 特征子空间是相同的.

上述讨论揭示了 $\boldsymbol{\Gamma}$ 的列向量空间可以通过 $\boldsymbol{A}$ 的前 K 个特征值 (按照绝对值由大到小排序) 张成的特征空间进行估计. 这是许多方法的基础 [133–135]. 为了找到 $\boldsymbol{\Gamma}$ 的相合估计, 需要做一个旋转, 但这个旋转通常是未知的且比较难寻找. 作为替代, 步骤 2 经常使用聚类算法, 如 K-means 算法, 这是因为 $\mathrm{E}\boldsymbol{A}$ 的前 K 个特征向量只有 K 行不同的值. 将典型谱聚类算法总结如下 (见算法 5).

算法 5　社群发现的谱聚类算法

1. 由邻接矩阵 $\boldsymbol{A}$ 的前 K 个 (绝对值) 大的特征值对应的特征向量组成 $n \times K$ 阶矩阵 $\widehat{\boldsymbol{\Gamma}}$.
2. 将 $\widehat{\boldsymbol{\Gamma}}$ 中的每一行作为输入数据, 使用聚类算法 (如 K-means 聚类算法), 将数据分成 K 个组, 如此将 n 个成员分成 K 个社群.

上述谱方法还有很多不同的变型和改进型. 记 $d_i = \sum_{j \neq i} a_{ij}$ 为节点 i 的 "度", 即节点 i 的连边数量, $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_n)$ 为 "度" 矩阵. 定义 Laplace 矩阵, 对称标准化 Laplace 矩阵和随机游走标准化 Laplace 矩阵如下:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}, \quad \boldsymbol{L}^{\mathrm{sym}} = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}, \quad \boldsymbol{L}^{\mathrm{rw}} = \boldsymbol{D}^{-1}\boldsymbol{L}. \tag{5.3}$$

在算法 5 步骤 1 中的邻接矩阵 $\boldsymbol{A}$ 可被 (5.3) 中的任意一个 Laplace 矩阵代替. 步骤 2 里 $\boldsymbol{\Gamma}$ 的构造中, 对原始特征向量还有许多其他的改进. 例如, 将 $\boldsymbol{\Gamma}$ 的行向量投影到单位球面 [136], 计算基于特征值比率的得分 [137] 等.

随机分块模型根据不同的应用场景也存在着许多变体. 混合社群模型被文献 [138] 提出以处理一个成员属于多个社群的情形, 文献 [139] 引入了带度修正的随机分块模型来描述社群内成员间连接是变化的情形.

定义 4　对任意节点 $i \in [n]$, 令 $\boldsymbol{\pi}_i$ 为该顶点所属社群的 K 维概率向量, $\boldsymbol{\pi}_i(k) = \mathrm{P}(i \in \mathcal{C}_k),\ k \in [K]$. $\theta_i > 0$ 表示节点 i 亲和力的异质性程度. 在带度修正的混合社群模型中, 假设

$$\mathrm{P}(A_{kl} = 1 \mid k \in \mathcal{C}_i, l \in \mathcal{C}_j) = \theta_k \theta_l\, p_{ij}, \quad \forall i, j, k, l, \tag{5.4}$$

且 $0 \leqslant \theta_k\theta_l p_{ij} \leqslant 1$. 因此, 在节点 k 和 l 中存在连接的概率为

$$\mathrm{P}(A_{kl}=1)=\theta_k\theta_l\sum_{i=1}^{K}\sum_{j=1}^{K}\boldsymbol{\pi}_k(i)\boldsymbol{\pi}_l(j)p_{ij}. \tag{5.5}$$

当 $\theta_i=1$ 且 $\boldsymbol{\pi}_i$ 为节点 i 所属社群的示性向量时, 模型 (5.5) 退化成随机分块模型 (5.1). 若 $\{\theta_i\}$ 是变化的, $\boldsymbol{\pi}_i$ 仍是节点 i 所属社群的示性函数, 则得到带度修正的随机分块模型

$$\mathrm{P}(A_{kl}=1)=\theta_k\theta_l p_{ij},\quad k\in\mathcal{C}_i,\quad l\in\mathcal{C}_j. \tag{5.6}$$

回到一般模型 (5.5), 令 $\boldsymbol{\Theta}=\mathrm{diag}(\theta_1,\ldots,\theta_n)$, 则有

$$\mathrm{E}\,\boldsymbol{A}=\boldsymbol{\Theta\Pi P\Pi}^{\mathrm{T}}\boldsymbol{\Theta}, \tag{5.7}$$

其中 $\boldsymbol{\Pi}$ 为 $n\times K$ 维矩阵, $\boldsymbol{\pi}_i^{\mathrm{T}}$ 为其第 i 行. 因此, $\boldsymbol{\Theta\Pi}$ 落入 $\mathrm{E}\,\boldsymbol{A}$ 前 K 个特征值对应特征向量所张成的空间. 为了消除 $\boldsymbol{\Theta}$ 的影响, 文献 [137] 先计算 $\boldsymbol{A}$ 的第 j ($j\leqslant K$) 与 1 列特征向量的比值, 然后对这 $K-1$ 列特征向量比值按行进行聚类, 从而得到社群的分类, 详情参见文献 [140] 关于社群分类的推断问题.

5.2 矩阵填补

矩阵填补问题互联网时代有着广泛的应用, 最具代表性的例子是在网飞 (Netflix) 上的电影评分问题: 当观众 i 看过电影 j 时, 就会给出该电影的一个评分, 并填写在评分矩阵的第 (i,j) 位置上, 否则这个元素就是缺失的. 在网飞上有成千上万部电影以及上千万的观众, 它们所形成的庞大的评分矩阵中有相当一部分都是缺失的. 矩阵填补的任务就是根据已有数据对这些缺失值进行填补, 从而通过评分向观众推荐影片. 相似地还有书籍、音乐和电器等各种网购产品的推荐, 这些都是推荐系统 (recommender system) 中协同过滤的基础. 矩阵填补的另一个重要应用是针对词汇和文本矩阵: 将文本集合中的词汇使用频率表示成一个矩阵, 将这些矩阵填补过后, 便可以以词频为特征, 对不同的文本进行分类.

令 $\boldsymbol{\Theta}$ 代表 $n_1\times n_2$ 维真实的偏好矩阵, 其第 (i,j) 元素代表第 i 个观众对第 j 个电影的偏好. 实际数据 $\boldsymbol{X}$ 只有一小部分可以被观测到, 所有被观测到的位置记录在坐标集 $\boldsymbol{\Omega}$ 中, 这些观测数据还可能存在噪声污染 (打分的不确定性), 假设观测数据服从模型:

$$X_{ij}=\theta_{ij}+\varepsilon_{ij},\quad \text{对 } (i,j)\in\boldsymbol{\Omega}. \tag{5.8}$$

不妨假设 ε_{ij} 是独立同分布的 $N(0,\sigma^2)$ 随机变量. 进一步假设数据是随机缺失的, 即 $\boldsymbol{\Omega}$ 是随机决定的, 第 (i,j) 个元素被观测到的概率是 p, 且与其他元素独立. 令 I_{ij} 为独立同分布 Bernoulli 随机变量, 成功概率为 p, 对观测数据进行建模 $\boldsymbol{\Omega}=\{(i,j):I_{ij}=1\}$. 即使不存在噪声, 上述问题仍然是欠定的, 可行的解决方案是对 $\boldsymbol{\Theta}$ 引入低秩假设, 即得到一种特殊的降秩回归 (reduced-rank regression),

$$\min_{\Theta\in\mathbb{R}^{n_1\times n_2}}\frac{1}{2}\left\|\mathcal{P}_\Omega(\boldsymbol{X})-\mathcal{P}_\Omega(\boldsymbol{\Theta})\right\|_F^2,\quad \text{s.t.}\quad \mathrm{rank}\,(\boldsymbol{\Theta})=K, \tag{5.9}$$

其中抽样算子 $\mathcal{P}_\Omega:\mathbb{R}^{n_1\times n_2}\to\mathbb{R}^{n_1\times n_2}$ 定义为 $(\mathcal{P}_\Omega(\boldsymbol{\Theta}))_{ij}=\theta_{ij}I_{ij}$. 由于非凸约束 $\mathrm{rank}\,(\boldsymbol{\Theta})=K$ 的存在, 通常考虑如下凸的惩罚最小二乘得到估计 $\widehat{\boldsymbol{\Theta}}$:

$$\sum_{(i,j)\in\boldsymbol{\Omega}}(X_{ij}-\theta_{ij})^2+\lambda\left\|\boldsymbol{\Theta}\right\|_*, \tag{5.10}$$

其中核范数惩罚是对低秩约束的一种松弛. 文献 [141] 证明了在低噪声情形下, 松弛优化问题的最优解依大概率也是带秩约束的非凸优化问题的最优解. 文献 [142] 进一步建立了统计最优理论并给出了推断方法.

低秩假设可以更好地通过因子模型来理解, 以电影评分为例, 每部电影有一系列公共特征 (因子), 每个观众对这些特征有不同的偏好再加上基于个人品味的一些得分. 这就形成了因子模型 (2.11) 和 (5.8) 中的低秩矩阵 $\boldsymbol{\Theta}$, 即 $\theta_{ij} = \boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{f}_j$, $\boldsymbol{f}_j \in \mathbb{R}^K$ 是第 j 件商品所具有的特征, $\boldsymbol{b}_i \in \mathbb{R}^K$ 是第 i 个用户的偏好. 与之前不同的是, 这里只观测到因子模型中的一小部分数据. 现在从谱方法的角度提供一个快速解法. 首先, 通过逆概率加权填补出一个完整的数据矩阵:

$$\boldsymbol{Y} = \frac{X_{ij} I_{(i,j)\in\boldsymbol{\Omega}}}{\widehat{p}}, \tag{5.11}$$

其中 $\widehat{p}$ 是观测到的数据所占的比例. 当观测足够多时, $\widehat{p}$ 则为 p 的精确估计. 忽略 p 的估计误差, 得到低秩矩阵:

$$\mathrm{E}\,\boldsymbol{Y} = \boldsymbol{\Theta}.$$

这为我们提供了如下谱方法: 假定 $\boldsymbol{\Theta}$ 的秩已知为 K, 令

$$\boldsymbol{Y} = \sum_{i=1}^{\min(n_1,n_2)} \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$$

为其奇异值分解, 给出谱估计如下:

$$\widehat{\boldsymbol{\Theta}} = \sum_{i=1}^{K} \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}. \tag{5.12}$$

文献 [96] 建立了该方法的逐元素估计最大误差上界, 并证明了在 Frobenius 范数意义下其复原的收敛速率达到对数最优 [143].

5.3 排序问题

排序是一个古老的问题, 最早可追溯至 17 世纪. 相比起同时对多个个体排序, 人们往往更擅长对两个个体进行比较. 因此, 一个重要的任务便是通过对 K 个个体两两比较对他们进行排序. 类似的排序问题在很多领域存在应用, 如网页检索 [144]、推荐系统 [145] 和体育竞赛 [146] 等.

如何通过统计方法达到上述目标呢? 假设个体 i 存在潜在分数 $w_i^* > 0$, 那么对任意的一对个体 (i, j), 有

$$\mathrm{P}(\text{个体 } j \text{ 优于个体 } i) = \frac{w_j^*}{w_i^* + w_j^*}. \tag{5.13}$$

该模型被称为 Bradley-Terry-Luce 模型 [147, 148]. 因此, 我们的目标就是估计潜在分数 $\boldsymbol{w}^* = (w_1^*, \ldots, w_n^*)$, 并通过这种偏好得分来对个体进行排序. 由于 $\boldsymbol{w}^*$ 放缩常数倍都可以被识别, 所以假定它为概率向量. 文献 [149] 提出了用 MM (minorization-maximization) 算法来拟合 Bradlye-Terry-Luce 模型.

事实上, 并不是所有配对比较都是可观测的, 可比较的只有一小部分. 引入图结构 $\mathcal{G}$, 其顶点和连边用来表示两个个体是否进行了两两比较 (如图 8 所示), 令 $\mathcal{E}$ 作为图 $\mathcal{G}$ 的连边集合. 对 $(i, j) \in \mathcal{E}$, 得到 L_{ij} 个独立比较结果:

$$Y_{ij}^{(l)} \underset{\sim}{\text{ ind. }} \mathrm{Bernoulli}\left(\frac{w_j^*}{w_i^* + w_j^*}\right), \quad 1 \leqslant l \leqslant L_{ij}. \tag{5.14}$$

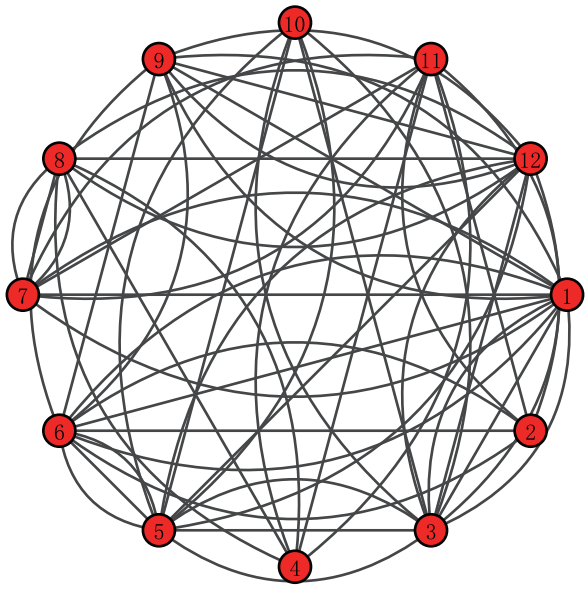

图 8　前 K 名比较图. 每两个点之间都会比较 L_{ij} 次

令 $\widehat{p}_{ij}$ 表示第 j 个个体优于第 i 个个体的概率估计. 为了理论研究, 假定图 $\mathcal{G}$ 是 Erdös-Rényi 图模型的一个现实.

一个简单的排序方法是计算所有逊于个体 i 的总比率 $\widehat{p}_i$, 再根据 $\{\widehat{p}_i\}_{i=1}^n$ 进行排序. 该方法忽略了其他竞争对手的次序, 在概率模型框架下, 一个自然的方法便是极大似然. 给定图 $\mathcal{G}$, 观测数据的似然函数为一系列 Bernoulli 概率的乘积:

$$L(\boldsymbol{w}) = \prod_{(i,j)\in\mathcal{E}} \left(\frac{w_j^*}{w_i^*+w_j^*}\right)^{L_{ij}\widehat{p}_{ij}} \left(\frac{w_i^*}{w_i^*+w_j^*}\right)^{L_{ij}(1-\widehat{p}_{ij})},$$

对数似然函数如下:

$$\ell(\boldsymbol{w}) = \sum_{(i,j)\in\mathcal{E}} L_{ij}\widehat{p}_{ij}\log\left(\frac{w_j}{w_i+w_j}\right) + L_{ij}(1-\widehat{p}_{ij})\log\left(\frac{w_i}{w_i+w_j}\right). \tag{5.15}$$

令 $\theta_i = \log w_i$ 产生某种类似逻辑回归的模型且对数似然函数是凸函数. 在高维情形时, 也可以通过正则项对似然函数进行惩罚:

$$\sum_{(i,j)\in\mathcal{E}} [L_{ij}\widehat{p}_{ij}\theta_j + L_{ij}(1-\widehat{p}_{ij})\theta_i - L_{ij}\log(\exp(\theta_i)+\exp(\theta_j))] + \lambda\sum_{i=1}^n \theta_i^2. \tag{5.16}$$

下面介绍谱排序法 (也称为秩中心化方法, 见算法 6). 不失一般性, 假定 $\boldsymbol{w}^*$ 为标准化概率向量, 目标

算法 6　Top-K 谱排序算法 (秩中心化算法)

1. 输入对比图 $\mathcal{G}$, 充分统计量 $\{\widehat{p}_{ij}, (i,j)\in\mathcal{E}\}$ 和标准化因子 d;
2. 如 (5.19) 定义概率转移矩阵 $\widehat{\boldsymbol{P}} = (\widehat{P}_{i,j})_{1\leqslant i,j\leqslant n}$;
3. 计算 $\boldsymbol{P}$ 的第一左特征向量 $\widehat{\boldsymbol{w}}$;
4. 输出 $\widehat{\boldsymbol{w}}$ 中最大的 K 个分量对应的个体.

是从图 $\mathcal{G}$ 构建一个 Markov 链, 使得 $\boldsymbol{w}^*$ 为此 Markov 链的不变分布 (也称为稳定分布). 定义转移矩阵 $\boldsymbol{P}^*$ 为

$$p_{ij}^* = \begin{cases} \dfrac{1}{d}\dfrac{w_j^*}{w_i^*+w_j^*}, & \text{若 } (i,j)\in\mathcal{E}, \\ 1-\dfrac{1}{d}\displaystyle\sum_{k:(i,k)\in\mathcal{E}}\dfrac{w_k^*}{w_i^*+w_k^*}, & \text{若 } i=j, \\ 0, & \text{否则}, \end{cases} \tag{5.17}$$

其中 d 为参数, 当 d 足够大 (例如, 最大的度) 使得对角元素非负时, 矩阵 $\boldsymbol{P}^*$ 是转移概率矩阵, 即每行加和为 1. 这个 Markov 链倾向于转移到分数更高的状态, 容易验证转移矩阵 $\boldsymbol{P}^*$ 存在如下具体均衡:

$$w_i^* P_{i,j}^* = w_j^* P_{j,i}^*, \quad \forall (i,j).$$

因此,

$$\sum_{i=1}^n w_i^* P_{i,j}^* = w_j^* \sum_{i=1}^n P_{j,i}^* = w_j^*$$

用矩阵归纳为

$$\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{P}^* = \boldsymbol{w}^{*\mathrm{T}}. \tag{5.18}$$

换句话说, $\boldsymbol{w}^*$ 是 $\boldsymbol{P}^*$ 的左特征向量, 对应特征值为 1, 也是转移矩阵为 $\boldsymbol{P}^*$ 的 Markov 链的稳定分布. 谱方法中, 使用转移矩阵的样本估计 $\widehat{\boldsymbol{P}}$ 替代未知的转移矩阵, 估计方法如下:

$$\widehat{P}_{ij} = \begin{cases} \dfrac{1}{d}\widehat{p}_{i,j}, & \text{若 } (i,j)\in\mathcal{E}, \\ 1-\dfrac{1}{d}\displaystyle\sum_{k:(i,k)\in\mathcal{E}}\widehat{p}_{i,k}, & \text{若 } i=j, \\ 0, & \text{否则}. \end{cases} \tag{5.19}$$

如果 $d > d_{\max}$ (图中最大的度), 那么 $\widehat{\boldsymbol{P}}$ 为转移矩阵. 在实际应用中, 通常取 $d = 2d_{\max}$.

文献 [150] 建立了谱估计 $\widehat{\boldsymbol{w}}$ 的 L_2 收敛速度. 文献 [142] 对其结果进行了优化并大大降低了逐项估计的最大误差, 因此可以同时获得谱方法和正则化极大似然方法模型选择的相合性. 此时, 抽样复杂度达到了文献 [151] 给出的信息量下界, 具体内容可以参见文献 [142,150]. 特别地, 后者还给出了该领域近期研究进展的综述.

5.4 高斯混合模型

作为谱方法应用于非凸优化问题的典型例子, 考虑高斯混合模型

$$\boldsymbol{x} \sim w_1 N(\boldsymbol{\mu}_1, \sigma_1^2\boldsymbol{I}_p) + \cdots + w_K N(\boldsymbol{\mu}_K, \sigma_K^2\boldsymbol{I}_p), \tag{5.20}$$

其中 $\{w_k, \boldsymbol{\mu}_k, \sigma_k^2\}_{k=1}^K$ 为未知参数. 由于随机样本 $\{\boldsymbol{x}_i\}_{i=1}^n$ 的似然函数是非凸的, 这给计算其极大似然估计带来了极大的挑战, 选取一个性质良好的初值是解决这类问题的关键.

主成分分析可以帮助我们找到一个较好的非凸优化问题初始值. 事实上, 文献 [152] 中的经典结果告诉我们, 对于有限维参数问题, 如果初始估计是 $\sqrt{n}$-相合的, 则一步 Newton-Raphson 迭代给出的结果统计意义就足够好. 换句话说, 优化误差 (与全局最优解的距离) 是统计误差 (估计量的标准差) 的低阶无穷小, 继续迭代只能提高优化误差, 详情参见文献 [153].

矩方法是一种有效寻找 $\sqrt{n}$-相合估计的简单方法. 对于混合模型 (5.20), 一、二阶矩为

$$\mathrm{E}\,\boldsymbol{x}=\sum_{i=1}^{K}w_k\,\boldsymbol{\mu}_k,\quad \mathrm{E}[\boldsymbol{x}\otimes\boldsymbol{x}]=\sum_{i=1}^{K}w_k\,\boldsymbol{\mu}_k\otimes\boldsymbol{\mu}_k+\sigma_w^2\boldsymbol{I}_p, \tag{5.21}$$

其中 $\otimes$ 表示 Kronecker 乘积, $\sigma_w^2=\sum_{i=1}^{K}w_k\sigma_k^2$ 表示加权平均方差. 因此, 由 $\{\boldsymbol{\mu}_k\}_{k=1}^{p}$ 张成的列 (向量) 空间与前 K 个主成分对应的特征空间相同, 并且 σ_w^2 是 $\mathrm{E}[\boldsymbol{x}\otimes\boldsymbol{x}]$ 的最小特征值.

为了进一步确定旋转方式, 文献 [154] 提出了三阶张量法. 假设 $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ 是线性无关的, 且 $w_k>0$, $k\in[K]$, 使得问题非退化, 则 σ_w^2 是 $\boldsymbol{\Sigma}=\mathrm{Cov}(\boldsymbol{x})$ 的最小特征值. 再令 $\boldsymbol{v}$ 是 $\boldsymbol{\Sigma}$ 的任意的对应特征值 σ_w^2 的特征向量, 定义

$$\begin{aligned}
&\boldsymbol{M}_1=\mathrm{E}[(\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{x}-\mathrm{E}\,\boldsymbol{x}))^2\boldsymbol{x}]\in\mathbb{R}^p,\\
&\boldsymbol{M}_2=\mathrm{E}[\boldsymbol{x}\otimes\boldsymbol{x}]-\sigma_w^2\boldsymbol{I}_p\in\mathbb{R}^{p\times p},\\
&\boldsymbol{M}_3=\mathrm{E}[\boldsymbol{x}\otimes\boldsymbol{x}\otimes\boldsymbol{x}]-\sum_{j=1}^{p}\sum_{\mathrm{cyc}}\boldsymbol{M}_1\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_j\in\mathbb{R}^{p\times p\times p},
\end{aligned}$$

其中

$$\sum_{\mathrm{cyc}}\boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{c}:=\boldsymbol{a}\otimes\boldsymbol{b}\otimes\boldsymbol{c}+\boldsymbol{b}\otimes\boldsymbol{c}\otimes\boldsymbol{a}+\boldsymbol{c}\otimes\boldsymbol{a}\otimes\boldsymbol{b},$$

$\boldsymbol{e}_j$ 为第 j 个位置为 1 的单位向量. 文献 [154] 证明了

$$\begin{aligned}
\boldsymbol{M}_1&=\sum_{k=1}^{K}w_k\sigma_k^2\boldsymbol{\mu}_k,\\
\boldsymbol{M}_2&=\sum_{k=1}^{K}w_k\boldsymbol{\mu}_k\otimes\boldsymbol{\mu}_k,\\
\boldsymbol{M}_3&=\sum_{k=1}^{K}w_k\boldsymbol{\mu}_k\otimes\boldsymbol{\mu}_k\otimes\boldsymbol{\mu}_k.
\end{aligned}\tag{5.22}$$

$\{\boldsymbol{M}_i\}_{i=1}^{3}$ 可由矩估计得到, 剩下就是从 (5.22) 中反解出参数. 文献 [154,155] 提出了一种稳健张量幂方法 (robust tensor power method), 主要思想类似于计算主成分的幂迭代法, 通过 $\boldsymbol{M}_2$ 正交化 $\boldsymbol{M}_3$ 中的 $\{\boldsymbol{\mu}_k\}$, 再利用幂方法计算张量分解.

若 $\boldsymbol{M}_2=\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}$ 为 $\boldsymbol{M}_2$ 的谱分解, 记

$$\boldsymbol{W}=\boldsymbol{U}\boldsymbol{D}^{-1/2}\in\mathbb{R}^{p\times K},\quad \widetilde{\boldsymbol{\mu}}_k=\sqrt{w_k}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\mu}_k\in\mathbb{R}^K. \tag{5.23}$$

注意到 $\boldsymbol{W}$ 是 $\boldsymbol{M}_2^{1/2}$ 的广义逆, 故有 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_2\boldsymbol{W}=\boldsymbol{I}_K$, 代入 (5.22) 中可得

$$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_2\boldsymbol{W}=\sum_{i=1}^{K}w_k\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{W}=\sum_{i=1}^{K}\widetilde{\mu}_k\widetilde{\mu}_k^{\mathrm{T}}=\boldsymbol{I}_K.$$

因此, $\{\widetilde{\boldsymbol{\mu}}_k\}_{k=1}^{K}$ 是标准正交的且

$$\boldsymbol{\mu}_k=\boldsymbol{U}\boldsymbol{D}^{1/2}\frac{\widetilde{\mu}_k}{\sqrt{w_k}}. \tag{5.24}$$

将二次型 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_2\boldsymbol{W}$ 记成 $\boldsymbol{M}_2(\boldsymbol{W},\boldsymbol{W})$,

$$\boldsymbol{M}_2(\boldsymbol{W},\boldsymbol{W})=\sum_{k=1}^{K}w_k(\boldsymbol{W}^{\mathrm{T}}\mu_k)^{\otimes 2}\in\mathbb{R}^{K\times K},$$

其中 $\boldsymbol{a}^{\otimes 2}=\boldsymbol{a}\otimes\boldsymbol{a}$. 类似定义

$$\widetilde{\boldsymbol{M}}_3:=\boldsymbol{M}_3(\boldsymbol{W},\boldsymbol{W},\boldsymbol{W}):=\sum_{k=1}^{K}w_k(\boldsymbol{W}^{\mathrm{T}}\mu_k)^{\otimes 3}=\sum_{k=1}^{K}\frac{1}{\sqrt{w_k}}\tilde{\boldsymbol{\mu}}_k^{\otimes 3}\ \in\mathbb{R}^{K\times K\times K},\tag{5.25}$$

其中 $\boldsymbol{a}^{\otimes 3}=\boldsymbol{a}\otimes\boldsymbol{a}\otimes\boldsymbol{a}$. 因此, $\widetilde{M}_3$ 存在一个类似对称矩阵谱分解的正交张量分解, 且其能通过幂方法快速计算. 可以验证

$$\widetilde{\boldsymbol{M}}_3=\mathrm{E}[(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x})^{\otimes 3}]-\sum_{j=1}^{p}\sum_{\mathrm{cyc}}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_1)\otimes(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{e}_j)\otimes(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{e}_j),\tag{5.26}$$

这是 (5.22) 经过旋转后的形式.

至此, 估计方法已经明了. 首先, 使用矩方法通过 (5.26) 估计 $\widetilde{\boldsymbol{M}}_3$, 再通过张量幂方法 [155] 找到 (5.25) 中的 $\{\widetilde{\boldsymbol{\mu}}_k\}$ 的正交张量分解和对应的特征值 $\{1/\sqrt{w_k}\}_{k=1}^{K}$. 在此过程中, 我们将原问题从 p-维转化为 K-维, 通过 (5.23) 和 (5.24) 计算 $\boldsymbol{\mu}_k$ 的估计 (见算法 7).

算法 7 混合高斯模型参数估计的张量幂算法

1. 计算样本协方差矩阵和它的最小特征值 $\widehat{\sigma}_w^2$ 及对应的特征向量 $\widehat{\boldsymbol{v}}$;
2. 通过 $\boldsymbol{x}$ 的经验矩、$\widehat{\boldsymbol{v}}$ 和 $\widehat{\sigma}_w^2$ 得到估计 $\widehat{\boldsymbol{M}}_1$ 和 $\widehat{\boldsymbol{M}}_2$;
3. 计算谱分解 $\widehat{\boldsymbol{M}}_2=\widehat{\boldsymbol{U}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{U}}^{\mathrm{T}}$ 和 $\widehat{\boldsymbol{W}}=\widehat{\boldsymbol{U}}\widehat{\boldsymbol{D}}^{-1/2}$, 将估计 $\widehat{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{X}$、$\widehat{\boldsymbol{W}}$ 和 $\widehat{\boldsymbol{M}}_1$ 代入 (5.26) 得到 $\widetilde{\boldsymbol{M}}_3$ 的估计 $\widehat{\boldsymbol{M}}_3$;
4. 对 $\widehat{\boldsymbol{M}}_3$ 应用稳健张量幂方法得到 $\{\widehat{\widetilde{\boldsymbol{\mu}}}_k\}_{k=1}^{K}$ 和 $\{\widehat{w}_k\}_{k=1}^{K}$;
5. 计算 $\widehat{\boldsymbol{\mu}}_k=\widehat{\boldsymbol{U}}\widehat{\boldsymbol{D}}^{1/2}\widehat{\widetilde{\boldsymbol{\mu}}}_k/\sqrt{\widehat{w}_k}$, 求解线性方差组 $\widehat{\boldsymbol{M}}_1=\sum_{k=1}^{K}\widehat{w}_k\widehat{\sigma}_k^2\widehat{\boldsymbol{\mu}}_k$ 得到 $\{\widehat{\sigma}_k\}_{k=1}^{K}$.

算法 7 中样本协方差阵也可替换成样本二阶矩矩阵 $n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$. 由 (5.21) 和 (5.22) 可知,

$$\mathrm{E}\,\boldsymbol{x}\otimes\boldsymbol{x}=\boldsymbol{M}_2+\sigma_w^2\boldsymbol{I}_p=\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}+\sigma_w^2\boldsymbol{I}_p,$$

其前 K 个特征值为 $\{d_1+\sigma_w^2,\ldots,d_K+\sigma_w^2\}$, 剩下的 $p-K$ 个特征值为 σ_w^2. 前 K 个特征向量落在 $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ 张成的空间中. 令 $\boldsymbol{v}$ 为 $\mathrm{E}[\boldsymbol{x}\otimes\boldsymbol{x}]$ 的最小特征值对应的特征向量, $\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{v}=0$, $k\in[K]$. 由总体协方差 $\boldsymbol{\Sigma}=\mathrm{E}[\boldsymbol{x}\otimes\boldsymbol{x}]-(\mathrm{E}\,\boldsymbol{x})\otimes(\mathrm{E}\,\boldsymbol{x})$ 可得

$$\boldsymbol{\Sigma}\boldsymbol{v}=\sigma_w^2\boldsymbol{v}-(\mathrm{E}\,\boldsymbol{x})(\mathrm{E}\,\boldsymbol{x})^{\mathrm{T}}\boldsymbol{v}=\sigma_w^2\boldsymbol{v}.$$

因此, σ_w^2 是 $\boldsymbol{\Sigma}$ 的特征值, $\boldsymbol{v}$ 是对应的特征向量, σ_w^2 是最小特征值.

至此, 上述内容提供了一个如何用谱方法来获得有效初始估计的思路. 文献 [155] 使用张量方法解决了一系列统计学习问题, 特别对于之前提到的异方差球形高斯混合问题, 引入了隐 Markov 模型 (Markov 和隐 Dirichlet 分配模型, 也称为三层 Bayes 概率模型) 来解决问题. 文献 [156] 则将张量方法用于学习混合广义线性模型.

参考文献

1 Zhao P, Yu B. On model selection consistency of Lasso. J Mach Learn Res, 2006, 7: 2541–2563

2 Zou H. The adaptive Lasso and its oracle properties. J Amer Statist Assoc, 2006, 101: 1418–1429
3 Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. Ann Statist, 2006, 34: 1436–1462
4 Lawley D N, Maxwell A E. Factor analysis as a statistical method. Statistician, 1962, 12: 209–229
5 Stock J H, Watson M W. Forecasting using principal components from a large number of predictors. J Amer Statist Assoc, 2002, 97: 1167–1179
6 Fan J, Li R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. ArXiv: math/0602133, 2006
7 Bickel P J. Discussion on the paper by Fan and Lv. J R Stat Soc Ser B Stat Methodol, 2008, 70: 883–884
8 Spearman C. The Abilities of Man, vol. 89. New York: Macmillan, 1927
9 Bartlett M S. Methods of estimating mental factors. Nature, 1938, 141: 609–610
10 Chamberlain G, Rothschild M. Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. Cambridge: National Bureau of Economic Research, 1982
11 Fama E F, French K R. Common risk factors in the returns on stocks and bonds. J Financial Economics, 1993, 33: 3–56
12 Bai J, Ng S. Determining the number of factors in approximate factor models. Econometrica, 2002, 70: 191–221
13 Hirzel A H, Hausser J, Chessel D, et al. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology, 2002, 83: 2027–2036
14 Hochreiter S, Clevert D A, Obermayer K. A new summarization method for Affymetrix probe level data. Bioinformatics, 2006, 22: 943–949
15 Leek J T, Storey J D. A general framework for multiple testing dependence. Proc Natl Acad Sci USA, 2008, 105: 18718–18723
16 Stock J H, Watson M W. Macroeconomic forecasting using diffusion indexes. J Bus Econom Statist, 2002, 20: 147–162
17 Boivin J, Ng S. Are more data always better for factor analysis? J Econometrics, 2006, 132: 169–194
18 Bai J. Inferential theory for factor models of large dimensions. Econometrica, 2003, 71: 135–171
19 Bai J, Li K. Statistical analysis of factor models of high dimension. Ann Statist, 2012, 40: 436–465
20 Onatski A. Asymptotics of the principal components estimator of large factor models with weakly influential factors. J Econometrics, 2012, 168: 244–258
21 Bai J, Liao Y. Efficient estimation of approximate factor models via penalized maximum likelihood. J Econometrics, 2016, 191: 1–18
22 Connor G, Linton O. Semiparametric estimation of a characteristic-based factor model of common stock returns. J Empirical Finance, 2007, 14: 694–717
23 Connor G, Hagmann M, Linton O. Efficient semiparametric estimation of the Fama-French model and extensions. Econometrica, 2012, 80: 713–754
24 Fan J, Li R, Zhang C H, et al. Statistical Foundations of Data Science, vol. 778. Boca Raton: CRC Press, 2020
25 Fan J, Wang K, Zhong Y, et al. Robust high dimensional factor models with applications to statistical machine learning. Statist Sci, 2020, in press
26 Fan J, Liao Y, Mincheva M. Large covariance estimation by thresholding principal orthogonal complements. J R Stat Soc Ser B Stat Methodol, 2013, 75: 603–680
27 Wright J, Ganesh A, Rao S, et al. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In: Advances in Neural Information Processing Systems. New York: Neural Information Processing Systems Foundation, 2009, 2080–2088
28 Candés E J, Li X, Ma Y, et al. Robust principal component analysis? J ACM, 2011, 58: 11
29 Jolliffe I T. Principal component analysis and factor analysis. In: Principal Component Analysis. New York: Springer, 1986, 115–128
30 Hastie T, Stuetzle W. Principal curves. J Amer Statist Assoc, 1989, 84: 502–516
31 Schölkopf B, Smola A J, Müller K R. Advances in Kernel Methods-Support Vector Learning. Cambridge: MIT Press, 1999
32 Johnstone I M, Lu A Y. On consistency and sparsity for principal components analysis in high dimensions. J Amer

Statist Assoc, 2009, 104: 682–693
33 Paul D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statist Sinica, 2007, 17: 1617–1642
34 Shen D, Shen H, Zhu H, et al. The statistics and mathematics of high dimension low sample size asymptotics. Statist Sinica, 2016, 26: 1747–1770
35 Wang W, Fan J. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. Ann Statist, 2017, 45: 1342–1374
36 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2009
37 Hancock P J B, Burton A M, Bruce V. Face processing: Human perception and principal components analysis. Memory Cognition, 1996, 24: 26–40
38 Misra J. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. Genome Res, 2002, 12: 1112–1120
39 Hastie T, Tibshirani R, Eisen M B, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol, 2000, 1: research0003.1
40 Agarwal A, Negahban S, Wainwright M J. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. Ann Statist, 2012, 40: 1171–1197
41 Davis C, Kahan W M. The rotation of eigenvectors by a perturbation, III. SIAM J Numer Anal, 1970, 7: 1–46
42 Bartlett M S. Tests of significance in factor analysis. British J Statistical Psychology, 1950, 3: 77–85
43 Horn J L. A rationale and test for the number of factors in factor analysis. Psychometrika, 1965, 30: 179–185
44 Cattell R B. The scree test for the number of factors. Multivariate Behaval Res, 1966, 1: 245–276
45 Fan J, Guo J, Zheng S. Estimating number of factors by adjusted eigenvalues thresholding. J Amer Statist Assoc, 2020, in press
46 Luo R, Wang H, Tsai C L. Contour projected dimension reduction. Ann Statist, 2009, 37: 3743–3778
47 Lam C, Yao Q. Factor modeling for high-dimensional time series: Inference for the number of factors. Ann Statist, 2012, 40: 694–726
48 Ahn S C, Horenstein A R. Eigenvalue ratio test for the number of factors. Econometrica, 2013, 81: 1203–1227
49 Onatski A. Determining the number of factors from empirical distribution of eigenvalues. Rev Economics Stat, 2010, 92: 1004–1016
50 Dobriban E. Factor selection by permutation. ArXiv:171000479, 2017
51 Fama E F, French K R. A five-factor asset pricing model. J Math Finance, 2015, 116: 1–22
52 Fan J, Fan Y, Lv J. High dimensional covariance matrix estimation using a factor model. J Econometrics, 2008, 147: 186–197
53 James W, Stein C. Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. Berkeley: University California Press, 1961, 361–379
54 Fan J, Liao Y, Mincheva M. High dimensional covariance matrix estimation in approximate factor models. Ann Statist, 2011, 39: 3320–3356
55 Bickel P J, Levina E. Covariance regularization by thresholding. Ann Statist, 2008, 36: 2577–2604
56 Cai T, Liu W. Adaptive thresholding for sparse covariance matrix estimation. J Amer Statist Assoc, 2011, 106: 672–684
57 Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika, 2007, 94: 19–35
58 Banerjee O, Ghaoui L E, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J Mach Learn Res, 2008, 9: 485–516
59 Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical Lasso. Biostatistics, 2008, 9: 432–441
60 Duchi J, Gould S, Koller D. Projected subgradient methods for learning sparse Gaussians. In: Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence. Corvallis: AUAI Press, 2008, 153–160
61 Lu Z. Smooth optimization approach for sparse covariance selection. SIAM J Optim, 2009, 19: 1807–1827
62 Scheinberg K, Ma S, Goldfarb D. Sparse inverse covariance selection via alternating linearization methods. In:

Advances in Neural Information Processing Systems. New York: Neural Information Processing Systems Foundation, 2010, 2101–2109
63 Levina E, Rothman A, Zhu J. Sparse estimation of large covariance matrices via a nested Lasso penalty. Ann Appl Stat, 2008, 2: 245–263
64 Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. Ann Statist, 2009, 37: 4254–4278
65 Fan J, Feng Y, Wu Y. Network exploration via the adaptive LASSO and SCAD penalties. Ann Appl Stat, 2009, 3: 521–541
66 Guo G, Jin H. A switching system approach to actuator assignment with limited channels. Internat J Robust Nonlinear Control, 2009, 20: 1407–1426
67 Shen X, Pan W, Zhu Y. Likelihood-based selection and sharp parameter estimation. J Amer Statist Assoc, 2012, 107: 223–232
68 Ravikumar P, Wainwright M J, Raskutti G, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electron J Stat, 2011, 5: 935–980
69 Zhang T, Zou H. Sparse precision matrix estimation via Lasso penalized D-trace loss. Biometrika, 2014, 101: 103–120
70 Cai T, Liu W, Luo X. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. J Amer Statist Assoc, 2011, 106: 594–607
71 Cai T T, Liu W, Zhou H H. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. Ann Statist, 2016, 44: 455–488
72 Liu H, Wang L. TIGER: A tuning-insensitive approach for optimally estimating Gaussian graphical models. Electron J Stat, 2017, 11: 241–294
73 Catoni O. Challenging the empirical mean and empirical variance: A deviation study. Ann Inst H Poincaré Probab Statist, 2012, 48: 1148–1185
74 Fan J, Wang W, Zhu Z. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. ArXiv:160308315, 2016
75 Fan J, Li Q, Wang Y. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. J R Stat Soc Ser B Stat Methodol, 2017, 79: 247–265
76 Vershynin R. Introduction to the non-asymptotic analysis of random matrices. ArXiv:10113027, 2010
77 Tropp J A. User-friendly tail bounds for sums of random matrices. Found Comput Math, 2012, 12: 389–434
78 Koltchinskii V, Lounici K. Concentration inequalities and moment bounds for sample covariance operators. ArXiv:14052468, 2014
79 Vershynin R. How close is the sample covariance matrix to the actual covariance matrix? J Theoret Probab, 2012, 25: 655–686
80 Srivastava N, Vershynin R. Covariance estimation for distributions with $2+\varepsilon$ moments. Ann Probab, 2013, 41: 3081–3111
81 Fan J, Ke Y, Sun Q, et al. FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. J Amer Statist Assoc, 2019, 114: 1880–1893
82 Minsker S. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. Ann Statist, 2018, 46: 2871–2903
83 Liu H, Han F, Zhang C H. Transelliptical graphical models. Adv Neural Inform Process Systems, 2012, 1: 800–808
84 Xue L, Zou H. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. Ann Statist, 2012, 40: 2541–2571
85 Bickel P J, Levina E. Regularized estimation of large covariance matrices. Ann Statist, 2008, 36: 199–227
86 Cai T T, Zhang C H, Zhou H H. Optimal rates of convergence for covariance matrix estimation. Ann Statist, 2010, 38: 2118–2144
87 Bien J, Bunea F, Xiao L. Convex banding of the covariance matrix. J Amer Statist Assoc, 2016, 111: 834–845
88 Yu G, Bien J. Learning local dependence in ordered data. J Mach Learn Res, 2017, 18: 1354–1413
89 Bien J. Graph-guided banding of the covariance matrix. J Amer Statist Assoc, 2019, 114: 782–792
90 Li D, Zou H. SURE information criteria for large covariance matrix estimation and their asymptotic properties. IEEE Trans Inform Theory, 2016, 62: 2153–2169

91 Cape J, Tang M, Priebe C E. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. Ann Statist, 2019, 47: 2405–2439

92 Wedin P. Perturbation bounds in connection with singular value decomposition. BIT, 1972, 12: 99–111

93 Yu Y, Wang T, Samworth R J. A useful variant of the Davis-Kahan theorem for statisticians. Biometrika, 2014, 102: 315–323

94 Fan J, Wang W, Zhong Y. An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. J Mach Learn Res, 2018, 18: 1–42

95 Koltchinskii V, Xia D. Perturbation of linear forms of singular vectors under Gaussian noise. In: High Dimensional Probability VII. New York: Springer, 2016, 397–423

96 Abbe E, Fan J, Wang K, et al. Entrywise eigenvector analysis of random matrices with low expected rank. ArXiv:170909565, 2017

97 Zhong Y. Eigenvector under random perturbation: A nonasymptotic Rayleigh-Schrödinger theory. ArXiv:170200139, 2017

98 Eldridge J, Belkin M, Wang Y. Unperturbed: Spectral analysis beyond Davis-Kahan. ArXiv:170606516, 2017

99 Bean D, Bickel P J, El Karoui N, et al. Optimal M-estimation in high-dimensional regression. Proc Natl Acad Sci USA, 2013, 110: 14563–14568

100 Zhong Y, Boumal N. Near-optimal bounds for phase synchronization. SIAM J Optim, 2018, 28: 989–1016

101 O'Rourke S, Vu V, Wang K. Random perturbation of low rank matrices: Improving classical bounds. Linear Algebra Appl, 2018, 540: 26–59

102 Fan J, Liu H, Wang W. Large covariance estimation through elliptical factor models. Ann Statist, 2018, 46: 1383–1414

103 Kneip A, Sarda P. Factor models and variable selection in high-dimensional regression analysis. Ann Statist, 2011, 39: 2410–2447

104 Fan J, Ke Y, Wang K. Factor-adjusted regularized model selection. Available at SSRN 3248047, 2018

105 Efron B. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, vol. 1. Cambridge: Cambridge University Press, 2012

106 Efron B. Correlation and large-scale simultaneous significance testing. J Amer Statist Assoc, 2007, 102: 93–103

107 Efron B. Correlated z-values and the accuracy of large-scale statistical estimates. J Amer Statist Assoc, 2010, 105: 1042–1055

108 Fan J, Han X, Gu W. Estimating false discovery proportion under arbitrary covariance dependence. J Amer Statist Assoc, 2012, 107: 1019–1035

109 Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol, 1995, 57: 289–300

110 Storey J D. A direct approach to false discovery rates. J R Stat Soc Ser B Stat Methodol, 2002, 64: 479–498

111 Genovese C, Wasserman L. A stochastic process approach to false discovery control. Ann Statist, 2004, 32: 1035–1061

112 Lehmann E L, Romano J P. Generalizations of the familywise error rate. In: Selected Works of EL Lehmann. New York: Springer, 2012, 719–735

113 Langaas M, Lindqvist B H, Ferkingstad E. Estimating the proportion of true null hypotheses, with application to DNA microarray data. J R Stat Soc Ser B Stat Methodol, 2005, 67: 555–572

114 Meinshausen N, Rice J. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. Ann Statist, 2006, 34: 373–393

115 Jin J, Cai T T. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. J Amer Statist Assoc, 2007, 102: 495–506

116 Fan J, Han X. Estimation of the false discovery proportion with unknown dependence. J R Stat Soc Ser B Stat Methodol, 2017, 79: 1143–1164

117 Wang J, Zhao Q, Hastie T, et al. Confounder adjustment in multiple hypothesis testing. Ann Statist, 2017, 45: 1863–1894

118 Friguet C, Kloareg M, Causeur D. A factor model approach to multiple testing under dependence. J Amer Statist Assoc, 2009, 104: 1406–1415

119 Desai K H, Storey J D. Cross-dimensional inference of dependent high-dimensional data. J Amer Statist Assoc, 2012, 107: 135–151

120 Park S H. Collinearity and optimal restrictions on regression parameters for estimating responses. Technometrics, 1981, 23: 289–295
121 Fan J, Xue L, Yao J. Sufficient forecasting using factor models. J Econometrics, 2017, 201: 292–306
122 Li K C. Sliced inverse regression for dimension reduction. J Amer Statist Assoc, 1991, 86: 316–327
123 Li K C. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. J Amer Statist Assoc, 1992, 87: 1025–1039
124 Cook R D. Fisher lecture: Dimension reduction in regression. Statist Sci, 2007, 22: 1–26
125 Cook R D. Regression Graphics: Ideas for Studying Regressions through Graphics, vol. 482. New York: John Wiley & Sons, 2009
126 Fan J, Liao Y, Wang W. Projected principal component analysis in factor models. Ann Statist, 2016, 44: 219–254
127 Fan J, Ke Y, Wang K. Factor-adjusted regularized model selection. J Econometrics, 2020, in press
128 Holland P W, Leinhardt S. An exponential family of probability distributions for directed graphs. J Amer Statist Assoc, 1981, 76: 33–50
129 Holland P W, Laskey K B, Leinhardt S. Stochastic blockmodels: First steps. Social Networks, 1983, 5: 109–137
130 Wang Y J, Wong G Y. Stochastic blockmodels for directed graphs. J Amer Statist Assoc, 1987, 82: 8–19
131 Abbe E. Community detection and stochastic block models. FNT Commun Inf Theory, 2017, 14: 1–162
132 Erdős Paul, Rényi A. On the evolution of random graphs. Publ Math Inst Hung Acad Sci, 1960, 5: 17–60
133 Rohe K, Chatterjee S, Yu B. Spectral clustering and the high-dimensional stochastic blockmodel. Ann Statist, 2011, 39: 1878–1915
134 Gao C, Ma Z, Zhang A Y, et al. Achieving optimal misclassification proportion in stochastic block models. J Mach Learn Res, 2017, 18: 1980–2024
135 Abbe E, Sandon C. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In: Proceedings of the IEEE 56th Annual Symposium on Foundations of Computer Science. Piscataway: IEEE, 2015, 670–688
136 Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems. New York: Neural Information Processing Systems Foundation, 2002, 849–856
137 Jin J. Fast community detection by SCORE. Ann Statist, 2015, 43: 57–89
138 Airoldi E M, Blei D M, Fienberg S E, et al. Mixed membership stochastic blockmodels. J Mach Learn Res, 2008, 9: 1981–2014
139 Karrer B, Newman M E J. Stochastic blockmodels and community structure in networks. Phys Rev E, 2011, 83: 016107
140 Fan J, Fan Y, Han X, et al. SIMPLE: Statistical inference on membership profiles in large networks. ArXiv:191001734, 2019
141 Candés E J, Recht B. Exact matrix completion via convex optimization. Found Comput Math, 2009, 9: 717–772
142 Chen Y, Fan J, Ma C, et al. Spectral method and regularized MLE are both optimal for top-K ranking. Ann Statist, 2019, 47: 2204–2235
143 Keshavan R H, Montanari A, Oh S. Matrix completion from noisy entries. J Mach Learn Res, 2010, 11: 2057–2078
144 Dwork C, Kumar R, Naor M, et al. Rank aggregation methods for the web. In: Proceedings of the 10th International Conference on World Wide Web. New York: ACM, 2001, 613–622
145 Baltrunas L, Makcinskas T, Ricci F. Group recommendations with rank aggregation and collaborative filtering. In: Proceedings of the 4th ACM Conference on Recommender Systems. New York: ACM, 2010, 119–126
146 Massey K. Statistical models applied to the rating of sports teams. Https://www.masseyratings.com/theory/massey97.pdf, 1997
147 Bradley R A, Terry M E. Rank analysis of incomplete block designs, I: The method of paired comparisons. Biometrika, 1952, 39: 324–345
148 Luce R D. Individual Choice Behavior: A Theoretical Analysis. New York: John Wiley & Sons, 1959
149 Hunter D R. MM algorithms for generalized Bradley-Terry models. Ann Statist, 2003, 32: 384–406
150 Negahban S, Oh S, Shah D. Rank centrality: Ranking from pairwise comparisons. Oper Res, 2017, 65: 266–287
151 Chen Y, Suh C. Spectral MLE: Top-K rank aggregation from pairwise comparisons. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37. Lille: MIT Press, 2015, 371–380

152 Bickel P J. One-step Huber estimates in the linear model. J Amer Statist Assoc, 1975, 70: 428–434
153 Robinson P M. The stochastic difference between econometric statistics. Econometrica, 1988, 56: 531–548
154 Hsu D, Kakade S M. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In: Proceedings of the 4th conference on Innovations in Theoretical Computer Science. New York: ACM, 2013, 11–20
155 Anandkumar A, Ge R, Hsu D, et al. Tensor decompositions for learning latent variable models. J Mach Learn Res, 2014, 15: 2773–2832
156 Sedghi H, Janzamin M, Anandkumar A. Provable tensor methods for learning mixtures of generalized linear models. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, vol. 51. Cadiz, 2016, 1223–1231

High-dimensional factor model and its applications to statistical machine learning

Zhao Chen, Jianqing Fan & Christina Dan Wang

Abstract This paper reviews the recent developments on factor model and its applications to statistical machine learning. The factor model reduces the dimensionality of variables, and provides a low-rank plus sparse structure for the high-dimensional covariance matrices. Therefore, it attracts much attention in high-dimensional data analysis, and has been widely applied in many fields of sciences, engineering, humanities and social sciences, including economics, finance, genomics, neuroscience, machine learning, and so on. We elaborate how to use principal component analysis method to extract latent factors, estimate their associated factor loadings, idiosyncratic components, and their associated covariance matrices. These methods have been proven to effectively cope with the challenges of big data, such as high dimensionality, strong dependence, heavy-tailed variables, and heterogeneity. In addition, we also focus on the role of the factor model in dealing with high-dimensional statistical learning problems such as covariance matrix estimation, model selection, multiple testing, and prediction. Finally, we illustrate the innate relationships between factor models and modern machine learning problems through several applications, including network analysis, matrix completion, ranking, and mixture models.

Keywords **factor model, PCA, structural covariance matrix, factor-adjusted method, model selection, multiple testing**

MSC(2010) **62H25, 68T05**

doi: 10.1360/SSM-2020-0041