相关矩阵因子分析方法及其在税收数据中的应用

朱德伟

重庆工商大学数学与统计学院，重庆市南岸区，400067；

摘要：税收是国家向社会提供公共产品、满足社会共同需要、参与社会产品分配，按照法律的规定，强制、无偿取得财政收入的一种规范形式，是国家不可或缺的一部分。本文研究了税收数据的样本相关矩阵因子分析方法，通过选取 2004-2016 年中国财政月度税收数据（行指标为月份，列指标为 82 项税收指标），利用样本相关矩阵的调整相关阈值方法（ACT）得到了影响税收的 6 个公共因素，比样本协方差矩阵因子分析得到的 9 个公共因素更精准，进一步拓宽了样本相关矩阵因子分析的应用范畴。

关键词：税收；估计因子数；因子分析；相关矩阵

DOI：10.69979/3029-2700.24.7.041

引言

税收是国家向社会强制、无偿取得财政收入的一种规范形式,挖掘税收指标之间潜在的公共因子数有重要的意义。例如，梁利（2023）利用主成分分析法建立国内生产总值等指标和税收收入之间的线性回归方程以对税收收入进行预测；胡洪曙和吴纤媚（2023）从总税收实际税负等角度出发，采用准最大似然估计对税收竞争反应函数进行估计，但是上述研究都没有考虑税收指标的公共影响因素。

因子模型中公因子数目估计是一个重要问题。例如，Bai 和 Ng（2002）提出两个模型选择准则函数 PC 和 IC 来估计公因子数目；Onatski（2010）提出了一种替代估计量，其使用差分特征值估计公因子数目，并将其命名为“边缘分布”估计量。

公因子数目的正确界定是因子模型理论和实证研究的核心。现有研究主要基于样本协方差矩阵的特征值进行估计，且估计的因子数目大于真实的因子数，而 Fan,Guo&Zheng(2022)提出的自由调整刻度不变的调整相关阈值方法(ACT)能正确估计因子数目，其证实了总体相关矩阵大于 1 的特征值个数与公因子数目相同，因此本文使用样本相关矩阵因子分析方法来分析 2004 年至 2016 年中国财政月度税收数据。

本文其余部分如下。第 2 章介绍了向量因子模型；第 3 章为假设条件和相关矩阵因子分析；第 4 章为样本特征根和对因子数目估计；第 5 章为样本相关矩阵因子分析在税收数据中的应用；第 6 章为结论。

1 向量因子模型

因子模型可以将向量 y 分解为潜在因子和异质性部分，如下所示：

$$y = \alpha + Bf + \epsilon, \#(1)$$

其中 $y = (y_1, \ldots, y_p)^T$ 为 p 维可观测向量，$f = (f_1, \ldots, f_K)^T$ 为 K 维潜在因子矩阵，$\epsilon = (\epsilon_1, \ldots, \epsilon_p)^T$ 为与潜在因子不相关的 p 维误差向量，α 为 p 维拦截向量，B 为 p×K 维负载矩阵。假设不失一般性的令 cov（f）= IK，其中 IK 为 K×K 维单位矩阵，则由因子模型（1）表示，y 的协方差矩阵为：

$$\Sigma = (\sigma_{ij}) = BB^T + \Psi, \#(2)$$

其中 cov（ϵ）= Ψ。本文将因子数量定义为负载矩阵 B 的秩，并将其假设为 K。

Σ 的 p×p 维相关矩阵为：

$$R = [\mathrm{diag}(\Sigma)]^{-1/2}\Sigma[\mathrm{diag}(\Sigma)]^{-1/2} \#(3)$$

其中 $\mathrm{diag}(\Sigma)$是对角矩阵。我们的目标是通过随机矩阵理论推导出一种无调谐和尺度不变的选择 K 的方法。

上述方法主要分为以下三个方面。首先，建立总体相关矩阵的特征值和公因子个数之间的关系：

$$K = \max\{j: \lambda_j(R) > 1, j \in [p]\}, \#(4)$$

其中 $\lambda_1(R) \geq \lambda_2(R) \geq \ldots \geq \lambda_p(R)$为相关矩阵 R 的特征值，K 是真实的公因子个数。然后，建立 $\lambda_i(R)$ 的偏差矫正估计量 $\widehat{\lambda_i^C}$：

$$\hat{K}^C = \max\left\{j: \widehat{\lambda_i^C} > s\right\}, s = 1 + \sqrt{\frac{p}{n-1}} \#(5)$$

其中，n 为样本容量，$\hat{K}^C$不依赖于任何其他的调节参数。令$\frac{p}{n-1} \to \rho$ 且 $F(v_0) = \{R: R$ 为因子模型(1)中可观测向量的相关矩阵且 $\lambda_K(R) > v_0\}$，其中 v_0是一个正常数，表示信号强度。这里给出信号强度 v_0和阈

值 s 的最佳下界为：$v_0 = 1 + \sqrt{\rho}$，$s = 1 + \sqrt{p/(n-1)}$。最后，推导出高维因子模型中样本相关矩阵的最大 K 个样本特征值的渐进性质。

2 假设条件和相关矩阵因子分析

考虑因子模型（1）满足以下条件：

条件(1)：因子$f_1, \ldots, f_K$相互独立；因子向量$(f_1, \ldots, f_K)$与误差向量$(\epsilon_1, \ldots, \epsilon_p)$独立。

条件（2）：$E(f) = 0_K$，$cov(f) = I_K$。

条件（3）：$E(\epsilon) = 0_p$，$cov(\epsilon) = \Psi > 0_{p\times p}$其中$\Psi$是对于$q \in [0,1]$时满足稀疏条件$m_p = \max_{i\le p} \sum_{j\le p} |\sigma_{ij}|^q = o(p)$时为一个对角或者更一般的矩阵。

条件（4）：$p > K$并且负载矩阵 B 列满秩。

将 B 写为 $B = (b_1, \ldots, b_K)$，其中$b_j = (b_{1j}, \ldots, b_{pj})^T$为当$j \in [K]$时的一个 p 维列向量。对于$j \in [K]$，当$\ell \in [p]$时，至多有一个系数$b_{\ell j} \neq 0$。

公因子的定义：对于$j \in [K]$，当$\ell_1, \ell_2 \in [p]$时，至少有两个系数$b_{\ell_1 j}, b_{\ell_2 j} \neq 0$，那么我们称因子$f_j$为一个公因子。

条件（5）：对于任意的$j \in [K]$，当$\ell_1, \ell_2 \in [p]$时，至少有两个系数$b_{\ell_1 j}, b_{\ell_2 j} \neq 0$。

由（3）式，（1）中 y 的总体相关矩阵为：

$$R = QQ^T = Q_1 Q_1{}^T + Q_2 Q_2{}^T, \#(6)$$

$$Q = \left[\mathrm{diag}(\Sigma)\right]^{-1/2} (B, \Psi^{1/2}) = (Q_1, Q_2), \#(7)$$

$Q_1 = \left[\mathrm{diag}(\Sigma)\right]^{-1/2} B$，$Q_2 = \left[\mathrm{diag}(\Sigma)\right]^{-1/2} \Psi^{1/2}$。

令$||M|| = \sqrt{\lambda_1(MM^T)}$表示算子范数，在条件（1）-（5）的情况下，如果$\left|\left|\left[\mathrm{diag}(\Sigma)\right]^{-1} \Psi\right|\right| \le 1$，我们有：

$$\lambda_j(R) \le 1, j = K+1, \ldots, p,$$

此外，我们可得：

$$K = \max\{j: \lambda_j(R) > 1, j \in [p]\} \#(8)$$

如果因子负载$\{\tilde{b}_i\}_{i=1}^p$（$\tilde{b}_i{}^T$是 B 的第 i 行）是一个来自总体的随机样本且$E\left(\tilde{b}_i \tilde{b}_i{}^T\right) = \Sigma_b$，当 K 固定时，通过强大数定律，我们有：

$$p^{-1} B^T B = p^{-1} \sum_{i=1}^{p} \tilde{b}_i \tilde{b}_i{}^T \to \Sigma_b. \#(9)$$

3 样本特征根和因子数目估计

令$y_1, \ldots, y_n$为来自模型（1）的容量为 n 的独立同分布样本，有：

$$y_i = \alpha + B f_i + \epsilon_i, i \in [n] = \{1, \ldots, n\},$$

其样本相关矩阵和样本协方差矩阵为：

$$\widehat{\Sigma}_n = n^{-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T, \#(10)$$

$$\widehat{R} = \left[\mathrm{diag}\left(\widehat{\Sigma}_n\right)\right]^{-1/2} \widehat{\Sigma}_n \left[\mathrm{diag}\left(\widehat{\Sigma}_n\right)\right]^{-1/2}, \#(11)$$

其中$\bar{y} = n^{-1}\sum_{i=1}^n y_i$为样本均值。令$\widehat{R}$和 R 的经验光谱分布（ESDs）分别为$F_n(t)$和$H_{p-K}(t)$，如下所示：

$$F_n(t) = \frac{1}{p-K} \sum_{j=K+1}^{p} 1\left(\lambda_j(\widehat{R}) \le t\right), \#(12)$$

$$H_{p-K}(t) = \frac{1}{p-K} \sum_{j=K+1}^{p} 1\left(\lambda_j(R) \le t\right), \#(13)$$

对于任意实数 t 成立，其中$1(\cdot)$为一个指示函数。

为了估计高维因子模型中的公因子个数，我们需要作一些额外的假设。

假设（1）：令$x_i = (x_{1i}, \ldots, x_{p+K,i})^T = (f_{1i}, \ldots, f_{Ki}, e_{1i}, \ldots, e_{pi})^T$并且$(e_{1i}, \ldots, e_{pi}) = (\epsilon_{1i}, \ldots, \epsilon_{pi})\Psi^{-1/2}$，$\{x_{ji}, j \in [p+K], i \in [n]\}$为独立随机变量满足：

$$\frac{1}{n(p+K)\eta_n^4} \sum_{j=1}^{p+K} \sum_{i=1}^{n} E\left|x_{ji}^4\right| 1\left(|x_{ji}| > \eta_n \sqrt{n}\right) \to 0, \#(14)$$

其中$\{\eta_n\}$为一个给定的正数数列满足$\eta_n \to 0$且$\eta_n \log n \to +\infty$。

假设(2)：对于$\delta_0 > 0$和任意 p，$\sup_{j\in[p+K]} E(|x_{j1}|^{6+\delta_0})$是有界的。

假设(3)：当$n \to \infty$时，$\rho_n = p/n \to \rho \in (0, \infty)$。

假设（4）：公因子个数 K 是固定的。

假设（5）：$\left|\left|\left[\mathrm{diag}(\Sigma)\right]^{-1} \Psi\right|\right| \le 1$且从 R 的特征值$\lambda_{K+1}(R), \ldots, \lambda_p(R)$得到的经验谱分布$H_{p-K}(t)$的极限谱分布 H(t)存在。

对于$j \in [p]$，令$\widehat{\lambda}_j = \widehat{\lambda}_j(\widehat{R})$和$\lambda_j = \lambda_j(R)$。对于任意给定的 j，定义：

$$m_{n,j}(z) = (p-j)^{-1} \left[\sum_{\ell=j+1}^{p} \left(\widehat{\lambda}_\ell - z\right)^{-1} + \left(\left(3\widehat{\lambda}_j + \widehat{\lambda}_{j+1}\right)/4 - z\right)^{-1}\right], \#(15)$$

$$\underline{m}_{n,j}(z) = -\left(1 - \rho_{j,n-1}\right) z^{-1} + \rho_{j,n-1} m_{n,j}(z), \#(16)$$

其中$\rho_{j,n-1} = (p-j)/(n-1)$。令$\widehat{\lambda}_j$的修正后的特征值为：

$$\widehat{\lambda}_j^{\,c} = -\frac{1}{\underline{m}_{n,j}\left(\widehat{\lambda}_j\right)}, \#(17)$$

修正后的特征值$\widehat{\lambda}_j^{\,c}$是$\lambda_j$的一致估计量。

对于高维因子模型（1）满足条件（1）-（5）和假设(1)-(5)，我们可以得出最小信号强度v_0：$v_0 = 1 + \sqrt{\rho}$。因此，在本文的剩余部分，我们将在相关矩阵 R 的集合中考虑如下估计方法：

$$F\left(1+\sqrt{\rho}\right)=\{R:R\text{ 是因子模型(1)中可观测向量的相关矩阵且 }\lambda_K(R)>1+\sqrt{\rho}\}$$

且在此基础上，对于高维因子模型（1）满足条件（1）-（5）和假设（1）-（5），$R\in F\left(1+\sqrt{\rho}\right)$，对于 δ 为一个非常小的正数时当 $s=1+\sqrt{\rho}+\delta<\lambda_K(R)$ 有：

$$P(\hat{K}^C(s)=K)\to 1,\#(18)$$

因此，$s=1+\sqrt{\rho}$ 为最佳阈值。

4 样本相关矩阵因子分析在税收数据中的应用

本节应用 ACT 方法来分析 2004 年至 2016 年中国财政月度税收数据，数据来源自 EPS 数据平台的中国财政税收数据库（olap.epsnet.com.cn/#/datas_data?cubId=1201）。

该数据有 156 个行指标，分别为 2004 年至 2016 年的月份；有 82 个列指标，为国内增值税等相关的 82 个税收指标。

首先去除偏离样本均值超过 10 个四分位数范围时的异常值。对数据做好处理后，数据的维度为 $p=82$，样本容量为 $n=156$。

在 R 语言中对 ACT 方法进行程序设计。将 2004 年至2016年中国财政月度税收数据带入程序中计算可得：通过样本协方差矩阵的方法选择了 9 个因子，9 个最大特征值为：3.75×107，8.05×105，6.39×105，1.64×105，7.42×104，5.25×104，4.44×104，3.99×104，2.62×104；使用 ACT 方法通过样本相关矩阵选择了 6 个因子， 9 个最大特征值为：35.78，9.66，5.46，4.43，3.19，2.79，2.16，1.57，1.49。在这 82 个时间序列中，使用样本协方差矩阵选择的 9 个因子的边际方差在 8.82×101 至 1.40×107 之间变化很大，使得基于协方差矩阵的方法的保真度受到影响，而使用 ACT 方法估计出了真实的公因子数，为高维数据因子分析提供了有效的指导。

从所选因子解释的方差百分比来看，所选择的 9 个因子解释了总异质性的 99.79%，选择的 6 个因子解释了总异质性的 99.52%；在标准化变量情况下，所选择的 9 个因子解释了总异质性的 81.09%，选择的 6 个因子解释了总异质性的 74.77%。

5 结论

本文以 2004 年至 2016 年中国财政月度税收数据为研究对象，构建了一个行指标为月份，列指标为 82 项税收指标的高维时间序列模型。应用 ACT 方法，正确估计出了高维时间序列模型中的真实公因子数，优于文献中的估计方法。

该方法的核心优势在于从相关矩阵出发，揭示了总体相关矩阵大于 1 的特征值个数与公因子数目的相等关系，借助利用样本相关矩阵的随机矩阵理论，纠正了估计顶部特征值时的偏差，并在估计特征值时考虑估计误差，克服了使用样本协方差矩阵使得可观测变量在尺度上具有高度异质性的缺点。

参考文献

[1]梁利.基于主成分视角对我国税收收入影响因素的计量分析[J].内蒙古科技与经济,2023,(21):73-79.

[2]胡洪曙,吴纤媚.我国市级地方政府间的税收竞争形式研究[J].华中师范大学学报,2023,62(06):74-93.DOI:10.19992/j.cnki.1000-2456.2023.06.006.

[3 Bai, J., and S. Ng (2022). Determining the Number of Factors in Approximate Factor Models. Econometrica, 70,191-221.

[4]Onatski, A. (2010). Determining the Number of Factors From Empirical Distribution of Eigenvalues. Review of Economic and Statistics, 92, 1004-1016.

[5]Fan, J.Q., Guo, J.H. and Zheng,S.R.(2022). Estimating Number of Factors by Adjusted Eigenvalues Thresholding. Journal of the American Statistical Association, 117:538, 852-861, DOI: 10.1080/01621459.2020.1825448.

作者简介：朱德伟（2000—），性别：男，民族：汉，籍贯：重庆市荣昌区，学生，学历：硕士研究生，重庆工商大学数学与统计学院;研究方向：数理统计。