Statistics and Application 统计学与应用, 2024, 13(2), 496-503
Published Online April 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.132049

Hans汉斯

基于比值法对高维时间序列因子模型中因子个数的实证分析

——以美国宏观经济数据为例

宁晓霞

华南农业大学数学与信息学院，广东 广州

收稿日期：2024年3月28日；录用日期：2024年4月18日；发布日期：2024年4月26日

摘　要

经济关系民生，代表一个国家的生产力水平。宏观经济是国家经济的总体表现和指导方向，直接影响着国家的发展和稳定。宏观经济数据分析是理解经济运行机制、指导政策制定和实施、支持企业战略规化以及风险管理的重要工具，对于实现经济稳定增长和可持续增长至关重要。本文采用高维时间序列因子模型，对美国宏观经济数据进行降维分析，在ER、CR、TCR、GR四种比值估计器下，估计公共因子的个数。估计结果显示，ER、GR估计器识别出的因子个数为2，CR、TCR估计器识别出的因子个数为3。通过AIC和BIC准则对估计结果进行评估，发现CR、TCR估计器识别出的因子个数结果更为准确，将3个公共因子分别解释为GDP、就业与失业、消费价格指数和信心指数，能够更好地对宏观经济数据进行解释，进而了解国家的经济状况。

关键词

宏观经济，因子模型，大维度，比值估计，因子个数

Empirical Analysis of the Number of Factors in High-Dimensional Time Series Factor Models Based on the Ratio Method

—Taking US Macroeconomic Data as an Example

Xiaoxia Ning

College of Mathematics and Information, South China Agricultural University, Guangzhou Guangdong

Received: Mar. 28th, 2024; accepted: Apr. 18th, 2024; published: Apr. 26th, 2024

文章引用：宁晓霞. 基于比值法对高维时间序列因子模型中因子个数的实证分析[J]. 统计学与应用, 2024, 13(2): 496-503. DOI: 10.12677/sa.2024.132049

Abstract

The economy directly affects people's livelihoods and represents a country's level of productivity. Macroeconomics reflects the overall performance and guides principles of a nation's economy, directly impacting its development and stability. Analyzing macroeconomic data is a crucial tool for understanding economic mechanisms, guiding policy formulation and implementation, supporting strategic planning for businesses, and managing risks. It is vital for achieving both stable and sustainable economic growth. In this study, a high-dimensional time series factor model is employed to conduct dimensionality reduction analysis on macroeconomic data from the United States. Using four ratio estimators, ER, CR, TCR, and GR, the number of common factors is estimated. The results indicate that the ER and GR estimators identify two common factors, while the CR and TCR estimators identify three. Evaluation based on AIC and BIC criteria suggests that the CR and TCR estimators provide more accurate results in identifying the number of factors. These three common factors are interpreted as GDP, employment and unemployment, and consumer price and confidence indices, offering better insights into macroeconomic data and understanding the country's economic conditions.

Keywords

Macroeconomics, Factor Model, Large Dimension, Ratio Estimate, The Number of Factors

1. 引言

理论的发展离不开数据的支撑。在大数据时代来临之前，理论发展受阻于数据获取的困难，这需要大量的时间和精力。然而，随着现代技术的不断进步，数据呈现爆炸性增长，在为理论分析带来便捷的同时，也给数据分析带来了挑战。在进行实际问题分析时，常常会遇到相关变量过多，且变量间存在相依性等问题。传统的一些模型更适用于处理相关变量较少的情况，因此一些新的处理方法应运而生。

现实世界中，每时每刻都会产生高维时间序列数据，比如金融市场上的股票数据，环境监测数据以及医疗数据等。然而，经典的多元时间序列处理技术在处理这些高维时间序列数据时往往束手无策，运用因子模型进行降维能够更有效地处理这些数据。以宏观经济为例，经济是一个国家的命脉，了解一个国家经济的健康状况，可以帮助决策者制定更适合的经济政策，也能让消费者在生活和职业方面做出更明智的决定。例如，陆晓明通过相关系数，发现美联储货币政策会影响消费者的持有资产，进而影响消费者的消费信心[1]；邵延晟对货币政策的影响进行理论总结，发现货币政策会影响投资和消费，即货币政策不确定性上升时会削弱投资，同时会使消费者自行改变储蓄、消费和投资的分配[2]。反之，政府也可以通过财政政策和货币政策对宏观经济进行调控，引导国家的经济增长[3]。因此，对宏观经济数据进行分析至关重要。然而，宏观经济涉及的衡量指标众多，使用因子模型进行降维后，只需要使用少量的公共因子就能对经济数据进行解释，即可以找到影响宏观经济的最大的几个因子，从而帮助人们进行分析，且由于公共因子个数较少，数据分析的工作量会大幅下降。

自二十一世纪以来，高维时间序列因子模型在理论方面持续发展和完善，特别是与两个核心问题：因子个数和因子载荷的估计有关的理论方面。在这两个基本问题中，因子个数的确定又会影响因子载荷矩

阵估计，因此确定因子个数的任务极其重要。估计因子个数的经典方法有主成分法、似然推理和比值估计法等。在比值估计法确定因子个数上，2012 年 Lam 和 Yao 基于样本滞后信息协方差矩阵进行特征分解，提出了通过相邻特征值之比来确定因子个数的特征值之比法(Eigenvalue Ratio, ER) [4]，2013 年 Ahn 和 Horenstein 通过后 $i+1$ 个特征值在后 $i+2$ 个特征值的占比，提出了生长比估计法(Growth Ratio, GR) [5]，2017 年 Xia 等人在 GR 估计器的基础上进行改进，提出了转换贡献率之比(Transformed Contribution Ratio, TCR) [6]，2018 年 Xia 等人通过相邻特征值对于包含该特征值在内的所有更小特征值的贡献值比值，提出了贡献率之比法(Contribution Ratio, CR) [7]。其中，TCR 估计器在理论方面已经证明具有相合性，其余估计器还不具有该性质。

2. 模型介绍及估计方法

2.1. 模型介绍

近似静态因子模型[8]

近似静态因子模型的结构如下：

$$y_t = Ax_t + \varepsilon_t \left(t = 1, \cdots, T\right),$$

上式中，$y_t\left(N\times 1\right)$ 是可观测数据；$A\left(N\times r\right)$ 是因子载荷矩阵，用于描述可观测数据与潜在因子之间的线性关系；$x_t\left(r\times 1\right)$ 是公共因子，是影响所有可观测数据的共同因素；$\varepsilon_t\left(N\times 1\right)$ 是特殊因子，表示可观测序列中未被公共因子解释的部分。近似静态因子模型与传统因子模型的区别在于，近似静态因子模型允许特殊因子截面相关。

动态因子模型[9]

动态因子模型的结构为：

$$y_t = A\left(L\right)x_t + \varepsilon_t \left(t = 1, \cdots, T\right),$$

$$A\left(L\right) = A_1 L + A_2 L^2 + \cdots + A_q L^q \text{。}$$

模型中，$A\left(L\right)$ 是由 q 阶滞后算子多项式构成的动态因子载荷矩阵，且因子载荷矩阵可能随时间变化，L 是滞后算子，公共因子 x_t 为——ARMA 过程，特殊因子 ε_t 为未被动态因子解释的部分。

高维时间序列因子模型[4]

高维时间序列因子模型的结构如下：

$$y_t = Ax_t + \varepsilon_t \left(t = 1, \cdots, T\right)\text{。}$$

假设高维时间序列的维度为 N，因子个数为 r，则上式中，$y_t\left(N\times 1\right)$ 是可观测时间序列数据，$A\left(N\times r\right)$ 是因子载荷矩阵，$x_t\left(r\times 1\right)$ 是公共因子，$\varepsilon_t\left(N\times 1\right)$ 是特殊因子并且假设为白噪声序列。

在因子模型中，只有 y_t 是可观测数据，因子载荷矩阵、公共因子、特殊因子都是不可观测数据，其中因子个数和因子载荷矩阵都是待估计对象。高维时间序列因子模型是本文所关注的模型。

2.2. 预备知识

随机向量的协方差矩阵[10]

假设 x、y 分别为 p、q 维随机向量，则随机向量 x 和 y 的协方差矩阵定义为：

$$\operatorname{cov}\left(x, y\right) = \begin{pmatrix} \operatorname{cov}\left(x_1, y_1\right) & \operatorname{cov}\left(x_1, y_2\right) & \cdots & \operatorname{cov}\left(x_1, y_q\right) \\ \operatorname{cov}\left(x_2, y_1\right) & \operatorname{cov}\left(x_2, y_2\right) & \cdots & \operatorname{cov}\left(x_2, y_q\right) \\ \vdots & \vdots & & \vdots \\ \operatorname{cov}\left(x_p, y_1\right) & \operatorname{cov}\left(x_p, y_2\right) & \cdots & \operatorname{cov}\left(x_p, y_q\right) \end{pmatrix},$$

可简洁地表达为

$$\mathrm{cov}(x,y)=E\left[x-E(x)\right]\left[y-E(y)\right]',$$

$\left[y-E(y)\right]'$ 表示 $y-E(y)$ 的转置。

特征值和特征向量[11]

假设一个 n 阶矩阵 A 以及实数 λ，如果可以找到一个非零向量 α，满足：

$$A\alpha=\lambda\alpha,$$

则称 λ 是矩阵 A 的特征值，α 是矩阵 A 的属于特征值 λ 的特征向量。

一阶差分[12]

当自变量从 t 变到 $t+1$ 时，函数 $x=x(t)$ 的改变量 $\Delta x_t=x_{t+1}-x_t$，$t=0,1,2,\cdots$ 称为函数 $x(t)$ 在点 t 的一阶差分。

二阶差分[12]

对一阶差分后序列再进行一次差分称为二阶差分，记 $\Delta^2 x_t$ 为 x_t 的二阶差分，表示为：

$$\begin{aligned}\Delta^2 x_t&=\Delta x_t-\Delta x_{t-1}\\&=(x_t-x_{t-1})-(x_{t-1}-x_{t-2})。\\&=x_t-2x_{t-1}+x_{t-2}\end{aligned}$$

2.3. 估计方法

首先介绍本文使用的比值估计法相关的估计器，分别为 ER、CR、TCR、GR 四种估计器。令 $\Sigma_y(k)=\mathrm{cov}(y_{t+k},y_t)$ 为可观测时间序列 y_t 的滞后 k 阶协方差矩阵，$L=\sum_{k=1}^{k_0}\frac{\Sigma_y(k)\Sigma_y(k)'}{N^2}$，$\lambda_1>\lambda_2>\cdots>\lambda_N$ 为 L 的特征值，则 ER、CR、TCR、GR 估计器的估计原理为：

ER：$\hat{r}_1=\arg\min_{1\le i\le R}\frac{\hat{\lambda}_{i+1}}{\hat{\lambda}_i}$

CR：$\hat{r}_2=\arg\min_{1\le i\le R}\frac{\hat{\lambda}_{i+1}\Big/\sum_{l=i+1}^{m}\hat{\lambda}_l}{\hat{\lambda}_i\Big/\sum_{l=i}^{m}\hat{\lambda}_l}$

TCR：$\hat{r}_3=\arg\min_{1\le i\le R}\frac{\ln\left(1+\lambda_{i+1}\Big/\sum_{l=i+1}^{m}\lambda_l\right)}{\ln\left(1+\lambda_i\Big/\sum_{l=i}^{m}\lambda_l\right)}$

GR：$\hat{r}_4=\arg\min_{1\le i\le R}\frac{\ln\left[\sum_{l=i+1}^{m}\hat{\lambda}_l\Big/\sum_{l=i+2}^{m}\hat{\lambda}_l\right]}{\ln\left[\sum_{l=i}^{m}\hat{\lambda}_l\Big/\sum_{l=i+1}^{m}\hat{\lambda}_l\right]}$

其中，$r<R<N$ 是一个常数，通常取 $R=\frac{N}{2}$，否则会出现对因子个数高估的情况。

下面介绍具体的估计步骤：

1) 对 $\hat{L}$ 进行特征分析，得出 $\hat{L}$ 的特征值和特征向量，其中 $\hat{L}=\sum_{k=1}^{k_0}\frac{\hat{\Sigma}_y(k)\hat{\Sigma}_y(k)'}{N^2}$，

$\hat{\Sigma}_y(k)=\frac{1}{T-k}\sum_{t=1}^{T-k}(y_{t+k}-\bar{y})(y_t-\bar{y})'$， $\bar{y}=\frac{1}{T}\sum_{t=1}^{T}y_t$；

2) 将 $\hat{L}$ 的特征值按照降序排列为 $\hat{\lambda}_1\geq\hat{\lambda}_2\geq\cdots\geq\hat{\lambda}_N$；

3) 第 i 个最大特征值所对应的特征向量列排即得因子模型中待估计的因子载荷矩阵；

4) 根据四个比值估计器，分别得出特征值之比最小值处所对应的 i 即为因子个数的估计值。

3. 实证分析

3.1. 数据的选择

本文选取 1959 年 1 月~2023 年 12 月的美国宏观经济数据——FRED-MD 数据。FRED-MD 数据从 8 个组别、共 134 个指标来描述美国的宏观经济。这 8 个组别分别是：收入和支出、劳动力市场、消费和订单、订单和库存、货币和信贷、利率和汇率、价格、股票市场。134 个指标具体为实际个人收入、扣除转账收据的实际个人数据、实际个人消费支出等。FRED-MD 数据来源见 http://research.stlouisfed.org/econ/mccracken/sel/。

3.2. 数据的处理

在对数据进行分析降维之前，需要先对数据进行预处理。以实际个人收入数据(Real Personal Income, RPI)为例，做出关于 RPI 的时序图见图 1。

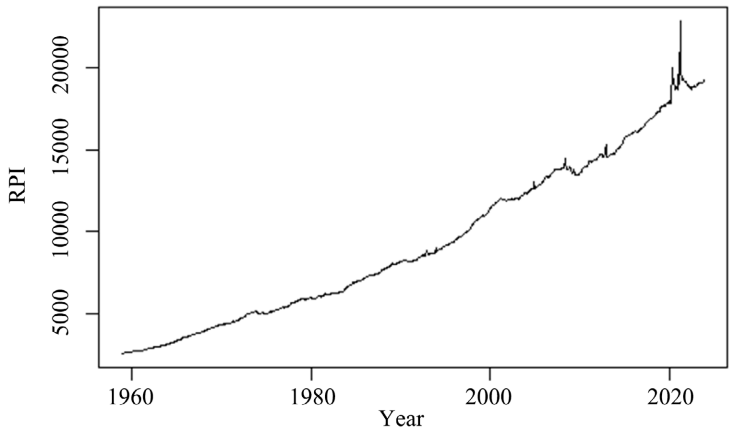

Figure 1. Diagram of RPI
图 1. RPI 的时序图

观察时序图发现，随着年份的增长，实际个人收入也在不断增加，即 RPI 序列不是平稳时间序列，因此需要先通过差分等方法使数据变得平稳。而不同指标的数据可能需要不同的方法来变得平稳，将需要相同方法而变得平稳的数据归为一类，则所有指标的数据需要 7 种方法来成为平稳时间序列。

假设 X_i 表示指标，X_{it} 表示该指标随时间的变化，则分为 7 个准则来对数据转换使数据平稳，具体准则见表 1。

Table 1. Criteria for the transformation of data
表 1. 数据的转换标准

转换代码	指标的变化
1	无变换

续表

2	ΔX_{it}
3	$\Delta^2 X_{it}$
4	$\log(X_{it})$
5	$\Delta\log(X_{it})$
6	$\Delta^2\log(X_{it})$
7	$\Delta(X_{it}/X_{i,t-1}-1)$

表 1 中，Δ、Δ^2 分别表示一阶差分和二阶差分。

在对原始数据进行转换之后，需要对数据进行异常值检验并处理，我们将满足下列不等式的称为异常值：

$$\left|X_{it}-\bar{X}_i\right|>10\times\text{四分位距}，$$

式中，$\bar{X}_i$ 表示第 i 个指标的平均值，X_{it} 表示第 i 个指标第 t 个时间点的数据。将满足异常值条件的异常值剔除，但经过异常值剔除后的数据集存在缺失值，本文用该指标下的均值来填补缺失值，以得到最终的数据集。

3.3. 因子个数的估计

本文针对 FRED-MD 数据，使用 R 软件对可观测时间序列数据分别采用 ER、CR、TCR、GR 四种方法进行因子个数的估计，本文采用折线图来表示每种比值法对因子个数的估计，折线图中的最低点对应的横坐标即为因子个数的估计值，估计结果见图 2。

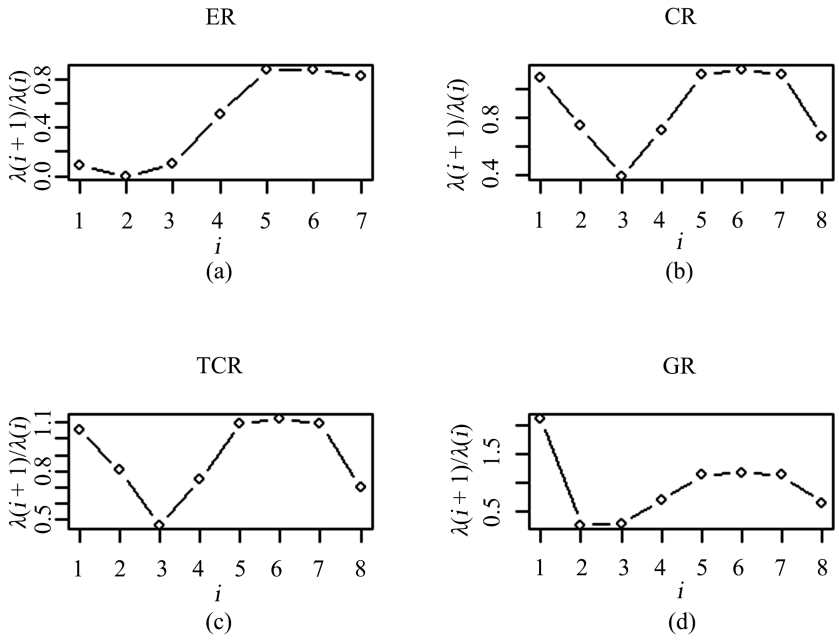

Figure 2. Estimation of the number of factors under ER, CR, TCR, GR estimators
图 2. ER、CR、TCR、GR 估计器下对因子个数的估计

观察图 2 可知，ER 估计器和 GR 估计器得到的因子个数为 2，CR 估计器和 TCR 估计器得到的因子个数为 3。

为了判断估计效果的准确性，本文使用 2002 年 Bai 和 Ng 提出的 AIC 准则和 BIC 准则来进行比较[13]。每种估计器下的 AIC 值和 BIC 值见表 2。AIC 准则和 BIC 准则如下：

Table 2. The AIC values and BIC values of ER, CR, TCR, GR estimators
表 2. ER、CR、TCR、GR 估计器下的 AIC 值和 BIC 值

	AIC1	BIC1	AIC2	BIC2	AIC3	BIC3
ER	0.00038	0.00039	0.00038	0.00038	0.00038	0.00039
CR	0.00013	0.00013	0.00013	0.00013	0.00013	0.00014
TCR	0.00013	0.00013	0.00013	0.00013	0.00013	0.00014
GR	0.00038	0.00039	0.00038	0.00038	0.00038	0.00039

$$\mathrm{AIC1}(k)=V+k\hat{\sigma}^2\left(\frac{2}{T}\right),\ \mathrm{BIC1}(k)=V+k\hat{\sigma}^2\left(\frac{\ln T}{T}\right);$$

$$\mathrm{AIC2}(k)=V+k\hat{\sigma}^2\left(\frac{2}{N}\right),\ \mathrm{BIC2}(k)=V+k\hat{\sigma}^2\left(\frac{\ln N}{N}\right);$$

$$\mathrm{AIC3}(k)=V+k\hat{\sigma}^2\left(2\frac{(N+T-k)}{NT}\right),\ \mathrm{BIC3}(k)=V+k\hat{\sigma}^2\left(\frac{(N+T-k)\ln(NT)}{NT}\right).$$

式中，$V=V\left(k,\hat{F}^k\right)=\frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_i^2$，$\hat{\sigma}_i^2=\frac{\hat{\varepsilon}_i'\hat{\varepsilon}_i}{T}$，$\hat{\sigma}^2=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left(\varepsilon_{it}\right)^2$，$k=\min\{N,T\}$。

根据表 2 可知，不论选择哪组 AIC 和 BIC，在 CR 和 TCR 估计器下的 AIC 值和 BIC 值都比 ER 和 GR 估计器得到的值小，根据 AIC 和 BIC 准则的含义可知，在 CR 和 TCR 估计器下，因子个数的估计结果更为准确。在因子个数估计值为 2 的情况下，影响美国宏观经济的因子解释为国内生产总值(Gross Domestic Product, GDP)、就业与失业；在因子个数估计值为 3 的情况下，将影响美国宏观经济的因子解释为 GDP、就业与失业、消费价格指数和信心指数。公共因子 GDP 可以直观地反映出一个国家的经济在某段时间的强弱情况，如果 GDP 不断增长，则表明国家的经济强盛，反之则表明国家的经济有衰减趋势。公共因子就业与失业可以了解一个国家的就业市场，若失业率持续走高(就业率持续下降)，则表明国家的经济下行。而公共因子消费价格指数和信心指数是从通货膨胀和消费者对未来的消费趋势去解释宏观经济。其中，消费价格指数(Consumer Price Index, CPI)可以反映通胀或通缩的程度，CPI 为正数表示通胀，负数表示通缩。通胀或通缩会影响消费者的购买、企业的投资以及国家的进出口贸易；消费信心指数(Consumer Confidence Index, CCI)反映消费者对未来的消费趋势，CCI 指数大于 100 表明消费者信心较足，指数在 100 以下表明消费者信心不足。当比值估计法识别出的因子个数为 3 个时，多了从消费价格指数和信息指数这一方面去解释宏观经济，而消费又是主导经济发展的核心，因此对宏观经济的解释更为准确。

4. 结论

本文选取了美国宏观经济数据，这是典型的高维时间序列数据，具有一定的代表性。由于宏观经济受到多种指标的影响，不同的指标之间会存在相关性，逐个分析会忽视它们之间的相关性，并且给数据分析带来很大的困难。因此，本文采用了因子模型对这些高维数据进行降维处理，在 ER、CR、TCR、

GR 四种比值估计器下对因子个数进行识别，结合 AIC 准则和 BIC 准则，发现 CR、TCR 估计器对因子个数的识别效果要优于 ER、GR 估计器。因此，针对本文的数据，CR 和 TCR 估计器识别出的公共因子能够更好地解释宏观经济的变化情况，即 GDP、就业与失业、消费价格指数和信心指数对美国宏观经济的解释更加准确。

参考文献

[1] 陆晓明. 美联储货币政策对家庭财富和分配的影响及其宏观经济意义[J]. 开发性金融研究, 2023(6): 14-26.

[2] 邵延晟. 货币政策冲击对中国宏观经济的影响研究[D]: [硕士学位论文]. 沈阳: 沈阳工业大学, 2022.

[3] 蔡应艳. 宏观之力: 政策如何影响经济增长的轨迹[J]. 中国商人, 2024(2): 120-121.

[4] Lam, C. and Yao, Q. (2012) Factor Modeling for High-Dimensional Time Series: Inference for the Number of Factors. *The Annals of Statistics*, **40**, 694-726. https://doi.org/10.1214/12-AOS970

[5] Ahn, S.C. and Horenstein, A.R. (2013) Eigenvalue Ratio Test for the Number of Factors. *Econometrica*, **81**, 1203-1227. https://doi.org/10.3982/ECTA8968

[6] Xia, Q., Liang, R. and Wu, J. (2017) Transformed Contribution Ratio Test for the Number of Factors in Static Approximate Factor Models. *Computational Statistics & Data Analysis*, **112**, 235-241. https://doi.org/10.1016/j.csda.2017.03.005

[7] Xia, Q., Liang, R., Wu, J., *et al*. (2018) Determining the Number of Factors for High-Dimensional Time Series. *Statistics and Its Interface*, **11**, 307-316. https://doi.org/10.4310/SII.2018.v11.n2.a8

[8] Chamberlain, G. and Rothschild, M. (1983) Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica*: *Journal of the Econometric Society*, **51**, 1281-1304. https://doi.org/10.2307/1912275

[9] Forni, M., Hallin, M., Lippi, M., *et al*. (2004) THE Generalized Dynamic Factor Model Consistency and Rates. *Journal of Econometrics*, **119**, 231-255. https://doi.org/10.1016/S0304-4076(03)00196-9

[10] 王学民. 应用多元分析[M]. 上海: 上海财经大学出版社, 2004: 27-28.

[11] 张贤达. 矩阵分析与应用[M]. 北京: 清华大学出版社有限公司, 2004: 47-48.

[12] 易丹辉, 王燕. 应用时间序列分析[M]. 北京: 中国人民大学出版社, 2019: 25-26.

[13] Bai, J. and Ng, S. (2002) Determining the Number of Factors in Approximate Factor Models. *Econometrica*, **70**, 191-221. https://doi.org/10.1111/1468-0262.00273