**Statistics and Application** 统计学与应用, 2019, 8(1), 149-154
Published Online February 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.81017

Hans汉斯

# Simulations for Data of Stock Market in Low-Dimensions

**Zupeng Zhao, Weidong Fang**

School of Mathematics, South China University of Technology, Guangzhou Guangdong
Email: zhaozupeng950419@163.com

Received: Jan. 27th, 2019; accepted: Feb. 11th, 2019; published: Feb. 19th, 2019

## Abstract

**The situations of stock market are important references of a country's development in economy, and we often have to deal with the high-dimensional data when we analyze those data from stock market. It's a hard work to analyze those high-dimensional data directly, so we use the simulation methods by reducing the dimensions of variables to decrease the difficulty of analysis. SVD (Singular Value Decomposition), FA (Factor Analysis) and PCR (Principal Component Regression) are three most common simulation methods which are considered to reduce the dimensions of variables. In order to compare simulational effectiveness of the three methods, we used the method of theoretical deduction and demonstration, and got the result that the three methods had the same simulational effectiveness in some case. Hence it was able to draw the conclusion that those three methods of reducing the variable dimensions were coincident in some conditions.**

## Keywords

**Dimensionality Reduction, SVD, FA, PCR, Simulation**

# 股票市场数据的低维模拟

**赵祖鹏，方卫东**

华南理工大学数学学院，广东 广州
Email: zhaozupeng950419@163.com

收稿日期：2019年1月27日；录用日期：2019年2月11日；发布日期：2019年2月19日

## 摘　要

**股票市场的情况是一个国家在经济上发展水平的重要参考，在分析股票市场数据时往往需要处理高维度**

文章引用: 赵祖鹏, 方卫东. 股票市场数据的低维模拟[J]. 统计学与应用, 2019, 8(1): 149-154.
DOI: 10.12677/sa.2019.81017

**的变量数据。直接分析这些高维度的变量数据是一件困难的工作，因此在处理这些数据时会使用降低变量维度的模拟方法来减少分析的难度。奇异值分解、因子分析和主成分回归是三种最常见的被考虑用来降低变量维度的模拟方法。为了比较这三种方法的模拟效果，本文中使用理论推导和证明的方法，得到三种方法在一定条件下有相同模拟效果的结果。于是可以得出在一定的条件下这三种降低变量维度的模拟方法具有一致性的结论。**

## 关键词

**降维，奇异值分解，因子分解，主成分回归，模拟**



## 1. 引言

股票市场中包含各种各样的时间序列数据，例如股票收盘价、股票价格指数等。在处理多个时间序列的数据时，需要分析的变量维度有时会很高，这给分析和处理带来了一定的难度。尽管每个时间序列都有各自的变化特点，但它们之间有一定的相关性。利用这种相关性，就能够减少变量的数量，从而减少分析的难度和成本。

本文对奇异值分解[1]、因子分析[2] [3] [4]和主成分回归[5]三种常见的降维方法进行讨论。其中，奇异值分解方法通过提取的奇异值来构造模拟矩阵；因子分析通过提取因子来构造模拟矩阵；主成分回归通过提取主成分和线性回归的方法来构造模拟矩阵。最后通过理论推导和证明的方式说明在一定的条件下，三种方法得到的模拟结果是一致的。

## 2. 模拟方法

### 2.1. 基本假设

对于时间序列，这里有一定的条件限制。第一点，时间序列为平稳时间序列；第二点，时间序列的期望值为 0。本文中以行业价格指数作为例子，不再对假设条件做过多的讨论。设 $\{X_j(t), t\in\mathbb{N}\}$ 是第 $j$ $(j=1,2,\cdots,r)$ 个行业股票指数每日收益率的时间序列，其中

$$\text{每日收益率}=(\text{当日收盘价}-\text{昨日收盘价})/\text{昨日收盘价}$$

为了简化操作，将时间序列 $\{X_j(t), t\in\mathbb{N}\}$ 简化为随机变量 $X_j$ 产生的多个独立同分布的样本。设向量

$$\boldsymbol{x}_j=\left[x_{1j},x_{2j},\cdots,x_{nj}\right]^{\mathrm{T}}$$

为第 $j$ 个行业的时间序列对应的数值。那么对于所有的 $r$ 个行业，有数据矩阵(观测值矩阵)

$$\underset{(n\times r)}{\boldsymbol{A}}=\left[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_r\right] \tag{1}$$

对应的样本协方差矩阵为

$$\underset{(r\times r)}{\boldsymbol{S}}=\begin{bmatrix} s_{11}^2 & \cdots & s_{1r}^2 \\ \vdots & \ddots & \vdots \\ s_{r1}^2 & \cdots & s_{rr}^2 \end{bmatrix}$$

假设 $n>r$，$\frac{1}{n}\sum_{i=1}^{n}x_{ij}=0\left(j=1,2,\cdots,r\right)$ 且 $\text{rank}\left(\boldsymbol{A}\right)=r$。

## 2.2. 奇异值分解模型

对式(1)定义的矩阵 $\boldsymbol{A}$，由奇异值分解定理，存在正交矩阵

$$\underset{(n\times n)}{\boldsymbol{U}}=\left[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_n\right] \tag{2}$$

$$\underset{(r\times r)}{\boldsymbol{V}}=\left[\boldsymbol{v}_1,\boldsymbol{v}_1,\cdots,\boldsymbol{v}_r\right] \tag{3}$$

和矩阵

$$\underset{(n\times r)}{\boldsymbol{M}}=\left[\begin{array}{ccc|c}\mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \underset{(r\times(n-r))}{\boldsymbol{0}} \\ & & \mu_r & \end{array}\right]^{\mathrm{T}} \tag{4}$$

使

$$\boldsymbol{A}=\boldsymbol{UMV}^{\mathrm{T}}$$

其中 $\mu_i\left(i=1,2,\cdots,r\right)$ 为 $\boldsymbol{A}$ 的奇异值(默认 $\mu_1\geq\mu_2\geq\cdots\geq\mu_r\geq0$，下文不再提及)。则

$$\boldsymbol{B}_{\text{SVD}}=\sum_{i=1}^{s}\mu_i\boldsymbol{u}_i\boldsymbol{v}_i^{\mathrm{T}} \tag{5}$$

是矩阵 $\boldsymbol{A}$ 的一个秩为 $s\left(s<r\right)$ 的同阶模拟矩阵，均方误差

$$MSE_{\text{SVD}}=\frac{1}{(n-1)r}tr\left[\left(\boldsymbol{A}-\boldsymbol{B}\right)\left(\boldsymbol{A}-\boldsymbol{B}\right)^{\mathrm{T}}\right]=\frac{1}{(n-1)r}\sum_{i=s+1}^{p}\mu_i^2$$

## 2.3. 因子分析模拟

假设 $X_j$ 由 $s\left(s<r\right)$ 个公共因子 $F_i$ 组成，即

$$\begin{aligned}X_1&=l_{11}F_1+l_{12}F_2+\cdots+l_{1s}F_s+\varepsilon_1\\X_2&=l_{21}F_1+l_{22}F_2+\cdots+l_{2s}F_s+\varepsilon_2\\&\vdots\\X_r&=l_{r1}F_1+l_{r2}F_2+\cdots+l_{rs}F_s+\varepsilon_r\end{aligned}$$

或者写为矩阵形式

$$\underset{(r\times1)}{\boldsymbol{X}}=\underset{(r\times s)}{\boldsymbol{L}}\underset{(s\times1)}{\boldsymbol{F}}+\underset{(r\times1)}{\boldsymbol{\varepsilon}}$$

设 $\left(\hat{\lambda}_k,\hat{\boldsymbol{e}}_k\right),k=1,2,\cdots,r$ 为样本协方差矩阵 $\boldsymbol{S}$ 的特征值－特征向量对(默认 $\hat{\boldsymbol{e}}_k$ 为单位向量且 $\hat{\lambda}_1\geq\hat{\lambda}_2\geq\cdots\geq\hat{\lambda}_r\geq0$，下文不再提及)，则

载荷矩阵 $\boldsymbol{L}$ 的估计值为

$$\underset{(r\times s)}{\hat{\boldsymbol{L}}}=\left[\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1\ \middle|\ \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{e}}_2\ \middle|\ \cdots\ \middle|\ \sqrt{\hat{\lambda}_s}\hat{\boldsymbol{e}}_s\right]=\left\{\hat{l}_{ij}\right\}_{r\times s}$$

矩阵 $\boldsymbol{A}$ 的近似估计

$$\boldsymbol{B}_{\text{FA}}=\boldsymbol{AS}^{-1}\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^{\mathrm{T}} \tag{6}$$

均方误差

$$MSE_{\mathrm{FA}}=\frac{1}{(n-1)r}tr\left[(\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}-\boldsymbol{B})^{\mathrm{T}}\right]=\frac{1}{r}\sum_{i=1}^{r}\hat{\psi}_i$$

其中 $\hat{\psi}_i=s_{ii}^2-\hat{l}_{i1}^2-\cdots-\hat{l}_{is}^2=\mathrm{Var}(\hat{\varepsilon}_i)$。

(这里载荷矩阵的估计使用的是主成分法，因子得分使用的是回归法。)

## 2.4. 主成分回归模拟

对于 $X_j\ (j=1,2,\cdots,p)$，通过样本协方差矩阵 $\boldsymbol{S}$ 提取其前 $s$ 个主成分的估计值

$$\hat{\boldsymbol{Y}}_k=\left[\hat{y}_{k1},\hat{y}_{k2},\cdots,\hat{y}_{kn}\right]^{\mathrm{T}}=\boldsymbol{A}\hat{\boldsymbol{e}}_k,\ k=1,2,\cdots,s$$

其中 $\left(\hat{\lambda}_k,\hat{\boldsymbol{e}}_k\right)$ 为 $\boldsymbol{S}$ 的特征值-特征向量对，于是有回归函数

$$\boldsymbol{x}_j=\beta_{j1}\hat{\boldsymbol{Y}}_1+\beta_{j2}\hat{\boldsymbol{Y}}_2+\cdots+\beta_{js}\hat{\boldsymbol{Y}}_s+\boldsymbol{\varepsilon}_j,\ j=1,2,\cdots,r$$

(这里令常数项为 0。)

由多元线性回归结果为：

$$\hat{\boldsymbol{\beta}}_j=\left[\beta_{j1},\beta_{j2},\cdots,\beta_{js}\right]^{\mathrm{T}}=\left(\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{Y}}\right)^{-1}\hat{\boldsymbol{Y}}^{\mathrm{T}}\boldsymbol{x}_j$$

其中

$$\underset{(n\times s)}{\hat{\boldsymbol{Y}}}=\left[\hat{\boldsymbol{Y}}_1\mid\hat{\boldsymbol{Y}}_2\mid\cdots\mid\hat{\boldsymbol{Y}}_s\right]$$

于是矩阵 $\boldsymbol{A}$ 有近似估计

$$\boldsymbol{B}_{\mathrm{PCR}}=\hat{\boldsymbol{Y}}\left(\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{Y}}\right)^{-1}\hat{\boldsymbol{Y}}^{\mathrm{T}}\boldsymbol{A}\tag{7}$$

均方误差

$$MSE_{\mathrm{PCR}}=\frac{1}{(n-1)r}tr\left[(\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}-\boldsymbol{B})^{\mathrm{T}}\right]$$

## 3. 三种模拟方法的一致性

**引理 3.1** 设 $\boldsymbol{A}$ 是 $n\times r(n>r)$ 阶实矩阵，$s<r=\mathrm{rank}(\boldsymbol{A})$，并且有奇异值分解 $\boldsymbol{UMV}^{\mathrm{T}}$，具体形式见式(2) (3) (4)，则

$$\boldsymbol{B}^{*}=\sum_{i=1}^{s}\mu_i\boldsymbol{u}_i\boldsymbol{v}_i^{\mathrm{T}}$$

是 $\boldsymbol{A}$ 的秩-$s$ 最小二乘逼近，使得在所有秩小于等于 $s$ 的 $n\times r$ 阶矩阵 $\boldsymbol{B}$ 中，平方误差和 $tr\left[(\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}-\boldsymbol{B})^{\mathrm{T}}\right]$ 最小，且最小值为 $\sum_{i=s+1}^{r}\mu_i^2$ (见文献[6])。

#

**引理 3.2** 设 $\boldsymbol{A}$ 是 $n\times r(n>r)$ 阶实矩阵，$r=\mathrm{rank}(\boldsymbol{A})$，并且有奇异值分解 $\boldsymbol{UMV}^{\mathrm{T}}$，具体形式见式(2) (3) (4)，则

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}_i=\mu_i^2\boldsymbol{v}_i,\ i=1,2,\cdots,r$$

即 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 有特征值-特征向量对 $\left(\mu_i^2,\boldsymbol{v}_i\right)$ (见文献[6])。

#

利用之前的三个模型结果和引理 3.1、引理 3.2，可以证明下面的定理。

**定理 3.1** 对于在 2.2、2.3 和 2.4 中三种使用同阶的低维度矩阵 $\boldsymbol{B}\left(\mathrm{rank}\left(\boldsymbol{B}\right)=s\right)$ 来模拟原数据矩阵 $\boldsymbol{A}\left(\mathrm{rank}\left(\boldsymbol{A}\right)=r,r>s\right)$ 的方法中(见式(5) (6) (7)，且矩阵 $\boldsymbol{A}$ 满足 2.1.中的假设条件)，并且都使用样本协方差矩阵 $\boldsymbol{S}$ 进行操作时，三种方法的模拟结果相同，即模拟矩阵

$$\boldsymbol{B}_{\mathrm{FA}}=\boldsymbol{B}_{\mathrm{PCR}}=\boldsymbol{B}_{\mathrm{SVD}}=\boldsymbol{A}\boldsymbol{V}_s\boldsymbol{V}_s^{\mathrm{T}}$$

其中

$$\underset{(r\times s)}{\boldsymbol{V}_s}=\left[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_s\right]=\left[\hat{\boldsymbol{e}}_1,\hat{\boldsymbol{e}}_2,\cdots,\hat{\boldsymbol{e}}_s\right]$$

且均方误差

$$MSE_{\mathrm{SVD}}=MSE_{\mathrm{FA}}=MSE_{\mathrm{PCR}}=\frac{1}{r}\sum_{i=s+1}^{r}\hat{\lambda}_i=\frac{1}{(n-1)r}\sum_{i=s+1}^{r}\mu_i^2$$

达到最小值。

($\left(\hat{\lambda}_i,\hat{\boldsymbol{e}}_i\right)$ 为矩阵 $\boldsymbol{S}$ 的特征值-特征向量对，$\boldsymbol{v}_i$ 的定义见式(3)，$\mu_i$ 的定义见式(4)。)

#

**证明：**根据 $\frac{1}{n}\sum_{i=1}^{n}x_{ij}=0\left(j=1,2,\cdots,r\right)$ 有

$$\frac{1}{n-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{S} \tag{8}$$

于是由引理 3.2 有

$$\hat{\lambda}_i=\frac{\mu_i^2}{n-1},\ i=1,2,\cdots,r \tag{9}$$

$$\boldsymbol{v}_i=\hat{\boldsymbol{e}}_i,\ i=1,2,\cdots,r \tag{10}$$

(使(10)式成立有时需要做一定的调整，这里我们不多做考虑。)

令

$$\underset{(s\times s)}{\hat{\boldsymbol{\Lambda}}}=\begin{bmatrix}\hat{\lambda}_1 & 0 & \cdots & 0\\ 0 & \hat{\lambda}_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \hat{\lambda}_s\end{bmatrix}$$

对于特征值分解方法，有

$$\boldsymbol{B}_{\mathrm{SVD}}=\sum_{i=1}^{s}\mu_i\boldsymbol{u}_i\boldsymbol{v}_i'=\boldsymbol{U}\boldsymbol{M}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}_s\boldsymbol{V}_s^{\mathrm{T}}=\boldsymbol{A}\boldsymbol{V}_s\boldsymbol{V}_s^{\mathrm{T}} \tag{11}$$

对于因子分解方法，由特征值和特征向量的定义，有

$$\boldsymbol{S}\boldsymbol{V}_s=\boldsymbol{V}_s\hat{\boldsymbol{\Lambda}} \tag{12}$$

于是

$$\boldsymbol{B}_{\mathrm{FA}}=\boldsymbol{A}\boldsymbol{S}^{-1}\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^{\mathrm{T}}=\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{V}_s\hat{\boldsymbol{\Lambda}}^{1/2}\left(\boldsymbol{V}_s\hat{\boldsymbol{\Lambda}}^{1/2}\right)^{\mathrm{T}}=\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{V}_s\hat{\boldsymbol{\Lambda}}\boldsymbol{V}_s^{\mathrm{T}}=\boldsymbol{A}\boldsymbol{V}_s\boldsymbol{V}_s^{\mathrm{T}} \tag{13}$$

对于主成分回归方法，有

$$\hat{\boldsymbol{Y}}=\boldsymbol{A}\boldsymbol{V}_s$$

再使用式(8)，(12)得到

$$\boldsymbol{B}_{\mathrm{PCR}}=\hat{\boldsymbol{Y}}\left(\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{Y}}\right)^{-1}\hat{\boldsymbol{Y}}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{A}\boldsymbol{V}_s\left(\boldsymbol{V}_s^{\mathrm{T}}\boldsymbol{S}\boldsymbol{V}_s\right)^{-1}\boldsymbol{V}_s^{\mathrm{T}}\boldsymbol{S}=\boldsymbol{A}\boldsymbol{V}_s\boldsymbol{V}_s^{\mathrm{T}} \tag{14}$$

综合式(11)，(13)和(14)得到

$$\boldsymbol{B}_{\mathrm{FA}}=\boldsymbol{B}_{\mathrm{PCR}}=\boldsymbol{B}_{\mathrm{SVD}}=\boldsymbol{A}\boldsymbol{V}_s\boldsymbol{V}_s^{\mathrm{T}}$$

于是

$$MSE_{\mathrm{SVD}}=MSE_{\mathrm{FA}}=MSE_{\mathrm{PCR}}$$

最后，根据引理 3.1 和式(9)得到均方误差

$$MSE_{\mathrm{SVD}}=MSE_{\mathrm{FA}}=MSE_{\mathrm{PCR}}=\frac{1}{p}\sum_{i=s+1}^{p}\hat{\lambda}_i=\frac{1}{(n-1)p}\sum_{i=s+1}^{p}\mu_i^2$$

达到最小值。

#

## 4. 结论

当由多个时间序列构成的数据矩阵满足对应时间序列的期望为零，且特征值和特征向量均由对应的样本协方差矩阵提取时，奇异值分解、因子分析和主成分回归构造的降维模拟方法具有一致性(这里的一致性仅限于上文提到的构造方法)。其中，模拟矩阵的结果仅依赖于所提取的特征向量(或奇异值分解的其中一个正交矩阵)，模拟矩阵均方误差的结果由所提取的特征值(或奇异值)完全决定。

## 参考文献

[1] 罗小桂. 矩阵奇异值分解(SVD)的应用[J]. 井冈山医专学报, 2005, 12(4): 133-135.

[2] 范龙振, 余世典. 中国股票市场的三因子模型[J]. 系统工程学报, 2002, 17(6): 537-546.

[3] 任福匀. 因子分析法在我国股票市场行业投资价值评价中的应用[D]: [硕士学位论文]. 长沙: 中南大学, 2005.

[4] Zhang, W. (2011) APT Model Based on Factor Analysis and an Empirical Study in China's Growth Enterprise Market. 2011 2*nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce* (*AIMSEC*), Dengleng, 8-10 August 2011, 994-997. https://doi.org/10.1109/AIMSEC.2011.6010620

[5] 张亚梅. 基于主成分回归分析科技创新对金融业的影响——以甘肃省为例[J]. 甘肃科技纵横, 2018, 47(9): 81-85.

[6] 理查德·A.约翰逊, 迪安·W.威克恩. 实用多元统计分析[M]. 第 6 版. 陆璇, 叶俊, 译. 北京: 清华大学出版社, 2008: 76-78.

**知网检索的两种方式：**

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2325-2251，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：sa@hanspub.org