第8章　主成分分析与因子模型

2025年11月

本讲介绍

本讲介绍机器学习中常用的降维方法——主成分分析(PCA)，并简要概述统计因子模型。

PCA是机器学习中的无监督学习，用于将现有变量压缩到更少变量，保留数据中的信息。

统计因子模型与常用金融或时间序列因子模型不同，它将数据分解为低秩且独立的因子形式，因子不可观测。而时间序列因子模型，如CAPM和Fama-French三因子模型等，是具有可观测自变量的回归模型。

本讲重难点：PCA的估计和原理，统计因子模型的性质和估计。

本讲关键问题

（1）如何衡量数据的变异程度？

（2）如何依次构建数据的主成分？

（3）如何理解PCA的数据压缩功能？

（4）什么是统计因子模型？

（5）统计因子模型与线性回归模型有何区别？

（6）统计因子模型怎么估计？

（7）理解统计因子模型的不唯一性。

（8）因子数量怎么选取？

关键词：统计因子模型、主成分分析、降维、奇异值分解

一、主成分分析 – 基本概念

设 $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ 为 p 维随机向量，例如 p 只股票的收益率。

假设$E(X) = 0$（方便计算，均值在主成分分析中不起作用。）

X 的方差矩阵为 $\Sigma = var(X) = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix}$，　　$\sigma_{kl} = cov(X_k, X_l)$

一、主成分分析 – 基本概念

奇异值分解 $\sum = e\Lambda e^T$，其中 $\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}, \lambda_1 \geq \cdots \geq \lambda_p > 0$

$$e = \begin{pmatrix} e_1 & \cdots & e_p \end{pmatrix} = \begin{pmatrix} e_{11} & \cdots & e_{1p} \\ \vdots & \ddots & \vdots \\ e_{p1} & \cdots & e_{pp} \end{pmatrix} \quad ,$$

e 是 $p \times p$ 的正交矩阵，$e^T e = I_p$。

令 p 维向量 e_k 表示 e 的第 k 列，$e_1, \ldots, e_p$ 之间相互正交，$e_k^T e_k = 1$, $e_i^T e_j = 0, \mathrm{i} \neq j$。

$(\lambda_k, e_k), k = 1, \ldots, p$，是矩阵 Σ 的 p 个特征值-特征向量对。

一、主成分分析 – 构建主成分

一维随机变量的变异程度用方差量化，p 维随机变量X的变异由方差矩阵 Σ 描述。量化 X 总“变异量”的方法是计算Σ的迹，$trace(\Sigma) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp}$。

我们希望用一个单一变量（一维）来尽可能多地量化所有 p 个变量的变异，例如通过构造 X 的线性组合来实现。

构造 $Y_1 = a_1^T X$使其方差最大，

$$var(Y_1) = a_1^T var(X) a_1 = a_1^T \Sigma a_1$$

Y_1的方差取决于a_1的大小。为了标准化，我们固定a_1的欧几里得范数为1，$\|a_1\| = \sqrt{a_1^T a_1} = 1$。问题变为

$$a_1 = \underset{||b||=1}{\text{argmax}} \quad b^T \Sigma b$$

一、主成分分析 – 构建主成分

回顾奇异值分解，$\sum = e\Lambda e^T$ ，e_i彼此正交且具有单位范数。e是p维线性空间的一组单位正交基，

则p维向量b总能表示为$b = z_1 e_1 + \cdots + z_p e_p = ez$，其中$z = \left(z_1, \ldots, z_p\right)^T$是范数为1的常数向量。

$$b^T \Sigma b = z^T e^T e \Lambda e^T e z = z^T \Lambda z = \lambda_1 z_1^2 + \cdots + \lambda_p z_p^2$$

当$z_1 = 1, z_2 = \cdots = \mathrm{z}_p = 0$时，$Y_1$达到最大方差 λ_1，解为$a_1 = e_1$，我们称$Y_1 = e_1^{\mathrm{T}} X$为X的第一个主成分。

一、主成分分析 – 构建主成分

找到第一个主成分后，寻找“第二重要”的变量$Y_2 = a_2^T X$，使得

$$a_2 = \underset{\|b\|=1, b \perp e_1}{argmax}\, b^T \sum b$$

$b \perp e_1$等价于Y_1和Y_2之间的零相关性，我们寻找的是试图解释 X 中未被Y_1解释的变异量的“第二重要”变量。以此类推，可以依次找到第二、第三，直到第p个主成分。

第 k 个主成分的解为：

$$Y_k = e_k^T X, \quad var(Y_k) = \lambda_k$$

对于$1 \le k \ne l \le p$，

$$cov(Y_k, Y_l) = e_k^T \Sigma e_l = e_k^T e \Lambda e^T e_l = 0$$

主成分间彼此正交，“重要性”通过方差衡量，第一个最重要，第二个次重要，依此类推。

一、主成分分析 – 总结

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} = e^T(X-\mu) = \begin{pmatrix} e_1^T \\ \vdots \\ e_p^T \end{pmatrix}(X-\mu)$$

$$\mathrm{var}(Y) = e^T var(X) e = \Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

第 k 个主成分 Y_k 解释了未被前 $k-1$ 个主成分解释的最多方差，$var(Y_k) = \lambda_k$，解释的总方差比例为

$$\frac{var(Y_k)}{var(Y_1)+\cdots+var(Y_p)} = \frac{\lambda_k}{\lambda_1+\cdots+\lambda_p}$$

PCA提供了一种理想的降维办法。保留前k个主成分则解释了 $\dfrac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\ldots+\lambda_p}$ 部分的信息。

一、主成分分析 – 分析视角

通过最大化 X 中的方差信息来简化维度：在 U是 $r \times p$ 的正交矩阵且 $UU^T = I_r$ 的约束下，最大化 $trace(UX)$。可以证明，此时

$$U = \begin{pmatrix} e_1^T \\ \vdots \\ e_r^T \end{pmatrix}$$

且 $UX = (Y_1, \dots, Y_r)^T$。这意味着通过较低维度的变量最大化X的方差信息。

一、主成分分析 – 压缩合成视角

试图找到秩为r的 $p \times p$矩阵A，最小化$trace(X - AX)$。可以证明

$$A = \begin{pmatrix} e_1 & \cdots & e_r \end{pmatrix} \begin{pmatrix} e_1^T \\ \vdots \\ e_r^T \end{pmatrix}$$

因此

$$AX = \begin{pmatrix} e_1 & \cdots & e_r \end{pmatrix} \begin{pmatrix} e_1^T \\ \vdots \\ e_r^T \end{pmatrix} X = \sum_{j=1}^{r} Y_j e_j$$

其背后的解释是：如果试图用一个r维（低维度）变量的线性组合来重建X，那么该变量必须是前r个主成分的线性组合。

一、主成分分析 – 自动编码器视角

从自动编码器角度理解，将 X 转化为 Y（主成分）的过程是编码步骤，从 Y 回到 X 的过程是解码步骤：

$$X \quad \Rightarrow^{\text{编码}} \quad Y = e^T X \quad \Rightarrow^{\text{解码}} \quad X = eY$$

或

$$X \quad \Rightarrow^{\text{编码}} \quad Y_k = e_k^T X, \quad k = 1, \dots, p \quad \Rightarrow^{\text{解码}} \quad X = \sum_{k=1}^{p} Y_k e_k$$

在实际应用中，通常只保留前r个最重要的主成分。此时过程变为：

$$X \quad \Rightarrow^{\text{编码}} \quad Y_k = e_k^T X, \quad k = 1, \dots, r \quad \Rightarrow^{\text{解码}} \quad X^* = \sum_{j=1}^{r} Y_j e_j$$

编码相当于压缩原始变量或数据，而解码相当于解压缩编码后的变量或数据。

一、主成分分析 – 样本主成分

假设有p维变量的n次观测，表示为矩阵

$$X = \begin{pmatrix} X_{(1)} & \cdots & X_{(p)} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$X_{(k)}$是第k个变量的n维观测向量。

设S为样本方差矩阵， $S = \frac{1}{n}(X - \bar{X})^T(X - \bar{X})$，奇异值分解$S = \hat{e}\,\hat{\Lambda}\,\hat{e}^T$。

样本主成分为

$$\hat{Y}_{n\times p} = \begin{pmatrix} \hat{Y}_{(1)} & \cdots & \hat{Y}_{(p)} \end{pmatrix} = \begin{pmatrix} X_{(1)} - \bar{X}_1 & \cdots & X_{(p)} - \bar{X}_p \end{pmatrix} \hat{e}$$

其中 $\bar{X}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$ 是第k个变量的样本平均值。

一、主成分分析 - 标准化

主成分的组成不仅与 X 的相关结构有关系，与每个分量X_k的方差也有关系，方差大的分量在第一主成分中贡献更大。

在各分量可比的情况下不存在问题，比如，X_k量纲相同或 X_k表示各科考试成绩且满分均为100。
但如果分量之间不可比，此时量纲会对主成分结果造成很大影响。比如，某分量的量纲从千克变成了克，其方差就变大到了原来的一百万倍，在第一主成分中的贡献就比原来大很多。

因此主成分分析经常事先对变量进行标准化

$$Z_{(k)} = \frac{X_{(k)} - \bar{X}_k}{\sigma(X_k)}$$

并对变量Z进行主成分分析。此时相当于用X的相关系数矩阵代替协方差矩阵作特征值分解。
对相关系数矩阵与对协方差矩阵作特征值分解的结果一般是不同的。选择相关系数矩阵作主成分分析默认了所有分量具有相同的重要性。

二、统计因子模型 – 基本概念

因子分析是主成分分析的进一步发展，其目标是把多个变量的信息压缩到少数几个变量中，用少数几个潜在变量（因子）描述相关变量。

设$X = (X_1, \cdots, X_p^T)$ 为随机变量，

$$E(X) = \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, var(X) = \sum = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix}$$

二、统计因子模型 – 基本概念

正交因子模型定义为

$$X_{p\times 1} - \mu_{p\times 1} = L_{p\times m}F_{m\times 1} + \epsilon_{p\times 1}$$

其中$m \le p, \mu = E(X)$，$L = \left(l_{ij}\right)_{p\times m}$是因子载荷矩阵（非随机），$F = \begin{pmatrix} F_1 & \cdots & F_m \end{pmatrix}^T$是公共因子（随机变量），$\epsilon = \begin{pmatrix} \epsilon_1 & \cdots & \epsilon_p \end{pmatrix}^T$是特异性误差（随机变量）。

该模型也可表示为：

$$X_i - \mu_i = \sum_{j=1}^{m} l_{ij}F_j + \epsilon_i, \quad i = 1, \dots, p$$

其中l_{ij}被称为X_i在因子F_j上的载荷。

二、统计因子模型 - 模型假设

正交因子模型的假设：

1. $E(F) = 0_{m\times 1}$，且$var(F) = I_m$。
2. $E(\epsilon) = 0_{p\times 1}$，且$var(\epsilon) = \Psi, \Psi$是对角矩阵，对角元素为$\psi_1, \dots, \psi_p$。
3. $cov(F, \epsilon) = 0_{m\times p}$。

这些假设意味着：

$$cov\left(F_i, F_j\right) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

$$cov\left(F_i, \epsilon_j\right) = 0, cov\left(\epsilon_i, \epsilon_j\right) = \begin{cases} \psi_i, & i = j \\ 0, & i \neq j \end{cases}$$

$$cov(X, F) = L, \quad cov(X_i, F_j) = l_{ij}$$

二、统计因子模型 - 方差分解

在正交因子模型下，随机变量 X 的方差矩阵 Σ 可以写成：

$$\Sigma = \mathrm{var}(LF + \epsilon) = \mathrm{var}(LF) + \mathrm{var}(\epsilon) = LL^T + \Psi$$

$$\sigma_{ii} = \mathrm{var}(X_i) = \sum_{j=1}^{m} l_{ij}^2 + \psi_i \equiv h_i^2 + \psi_i$$

$$\sigma_{ij} = \mathrm{cov}(X_i, X_j) = \sum_{k=1}^{m} l_{ik} l_{jk}, \quad i \neq j$$

其中 $h_i^2 \equiv l_{i1}^2 + \cdots + l_{im}^2$ 被称为communality（公共度），它表示X_i的方差中由公共因子解释的部分。

ψ_i被称为特异性方差idiosyncratic variance，它表示X_i的方差中由特定因子解释的部分。

二、统计因子模型 – 与线性回归模型的比较

尽管正交因子模型与线性回归模型在形式上有一定相似性

$$Y = X\beta + \epsilon$$

但存在本质差别：

- 因子模型中的因子载荷矩阵L是未知的，线性回归模型中的自变量X是已知的；
- 因子模型中公共因子F是未知的随机向量，线性回归模型中的回归系数β是未知的非随机参数向量。

二、统计因子模型 - 旋转不变性

正交因子模型中的 L 和 F 是不可识别的，仅有旋转不变性，

$$X - \mu = LF + \epsilon = L^*F^* + \epsilon$$

其中$L^* = LT$，$F^* = T^TF$且 T是任意的$m \times m$正交矩阵。 F*和ϵ仍满足假设1-3。

公共度不改变 $LL^T = L^*L^{*T}$ 。

二、统计因子模型 - 估计方法

统计因子模型的估计主要是估计载荷矩阵 L，可以采用极大似然法或主成分分析法。

- 主成分分析法：

$$\Sigma \approx \sum_{i=1}^{m} \lambda_i e_i e_i^T + \Psi$$

其中Ψ为 $\Sigma - \sum_{i=1}^{m} \lambda_i e_i e_i^T$ 主对角线元素组成的对角阵。因此

$$L = \begin{pmatrix} e_1 & \cdots & e_m \end{pmatrix} \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m \end{pmatrix}^{1/2}$$

- 极大似然法：假设$X_1, \ldots, X_n$是来自 p 维正态分布$N_p(\mu, \sum)$ 的独立同分布样本，$\sum = LL^T + \Psi$。通过最大化似然函数来估计参数 μ、L和 Ψ。

$$l(\mu, \Sigma) \equiv (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) \right\}$$

μ的极大似然估计为$\hat{\mu} = \bar{X}$。L和Ψ则可在$\Sigma = LL^T + \Psi$的条件下通过最大化 $-\frac{1}{2} n \log\det(\Sigma) - \frac{1}{2} \mathrm{tr}(W\Sigma^{-1})$ 得到，$W = \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^T$。最优$\hat{\sum}$，$\hat{L}$和$\hat{\Psi}$可通过EM迭代算法得到。

二、统计因子模型 - 因子旋转

矩阵L仅在正交变换下是唯一的，通常的做法是将L乘以一个适当的正交矩阵Q（称为因子旋转矩阵），使得LQ具有简单易解释的结构。令 $L^* = LQ$，通常选择Q使得如下varimax准则最大化：

$$C = \sum_{j=1}^{m}\left[\frac{1}{p}\sum_{i=1}^{p} l_{ij}^{*4} - \left(\frac{\sum_{i=1}^{p} l_{ij}^{*2}}{p}\right)^2\right]$$

$$\propto \sum_{j=1}^{m} \mathrm{Var}(\text{squared loadings of the jth factor})$$

最大化C等价于尽可能地分散每个因子上的载荷平方。

二、统计因子模型 - 因子得分

由于因子$F_i, 1 \le i \le n$是不可观测的，对模型诊断通常需要估算这些值。当 L、Ψ 和μ 已知时，因子的广义最小二乘估计为：

$$\hat{F} = arg\min_{F} \epsilon^T \Psi^{-1} \epsilon$$

$$= \left(L^T \Psi^{-1} L\right)^{-1} L^T \Psi^{-1} (X - \mu)$$

用$\hat{L}, \hat{\Psi}, \hat{\mu}$替代真实值得到估计

$$\hat{F} = \left(\hat{L}^T \hat{\Psi}^{-1} \hat{L}\right)^{-1} \hat{L}^T \hat{\Psi}^{-1} (X - \hat{\mu})$$

二、统计因子模型 - 因子数量的选择

选择合适的因子数量m是因子分析中的重要问题，以下是几种常见方法：

- 特征值准则：保留特征值大于1的主成分作为因子。特征值小于1表示该因子解释的方差小于原始变量的平均方差，可能包含的信息较少。
- 计算方差贡献率：当累计方差贡献率达到一定的阈值时，例如80%、85%或90%等，就认为选取的因子已经能够足够好地代表原始数据的信息。
- 碎石图（Scree Plot）：以主成分的序号为横坐标，以特征值为纵坐标的折线图。通常在某个点之后，折线的下降趋势会变得平缓，这个点对应的主成分数量就是一个比较合适的选择。这种方法相对较为直观，但也具有一定的主观性，需要根据实际情况进行判断。
- 信息准则：如赤池信息准则（AIC）、贝叶斯信息准则（BIC）等。选择使信息准则值最小的因子数量m。
- 似然比检验：假设$X_1, \ldots, X_p$独立且服从$N(\mu, \sum)$。原假设H_0是m-因子模型成立，$\sum = LL^T + \Psi$，L是$p \times m$的矩阵。检验该原假设与非约束 Σ的广义似然比统计量为：$\Lambda = n\left\{\log\det\left(\hat{L}\hat{L}^T + \hat{\Psi}\right) - \log\det\left(\hat{\Sigma}\right)\right\}$，
  其中 $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$ 是Σ的非约束最大似然估计。在H_0下，Λ近似服从χ^2分布，自由度为：

$$\frac{1}{2}\{(p-m)^2 - p - m\}$$

二、统计因子模型 - 为什么需要因子模型？

在许多统计问题中，与样本大小相当的高维数很常见。

Fan et. al (2008, JOE) 研究了在样本量n增加时，维数p趋于无穷的情况下的协方差矩阵估计。他们发现因子模型在估计协方差矩阵的逆时比传统的样本协方差矩阵更具优势，利用因子结构减少了参数数量，降低了估计复杂性。在投资组合相关应用中，若涉及协方差矩阵的逆，因子模型的估计优势就会凸显，使得估计结果更准确。

二、统计因子模型 - 主成分分析 vs 因子分析

主成分分析和因子分析都是降维技术，即它们可以用来用一组较低维度的新变量描述高维度变量。它们也常常会给出相似的结果。

然而，这两种方法在目标和底层模型上有所不同。如果只需要使用较少的维度来近似数据，应该使用主成分分析；当需要一个解释数据之间相关性的模型时，应该使用因子分析。

三、实证分析

选取中国十只上市公司股票（贵州茅台600519、格力电器000651、中国平安601318、美的集团000333、招商银行600036、长江电力600900、五粮液000858、紫金矿业601899、比亚迪 002594、中信证券600030）的日收益率。在Matlab中将收益率序列定义为$n \times p$的矩阵X。

主成分分析指令

$$[coeff, score, latent, tsquared, explained, mu] = pca(X);$$

coeff：主成分系数$\hat{e}$

score：主成分$\hat{Y}$

latent：主成分方差$\hat{\Lambda}$

explained：$\lambda_k/(\lambda_1 + \cdots + \lambda_p)$

mu：$\bar{X}$

三、实证分析

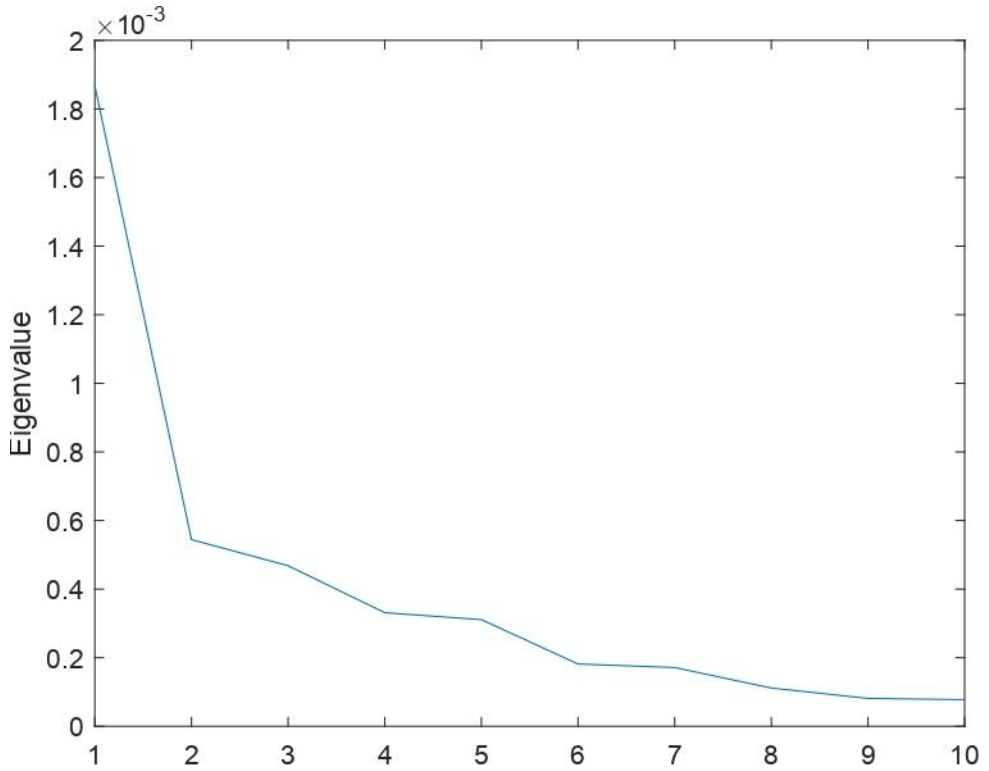


图9.1 主成分方差

表9.1 PCA结果总结

	方差($\times 10^{-4}$)	方差解释比率%	累积方差解释比率%
PC1	18.67	45.04	45.04
PC2	5.44	13.13	58.18
PC3	4.68	11.29	69.47
PC4	3.31	7.99	77.46
PC5	3.11	7.50	84.96
PC6	1.82	4.38	89.34
PC7	1.71	4.13	93.48
PC8	1.12	2.69	96.17
PC9	0.81	1.96	98.13
PC10	0.78	1.87	100.00

三、实证分析

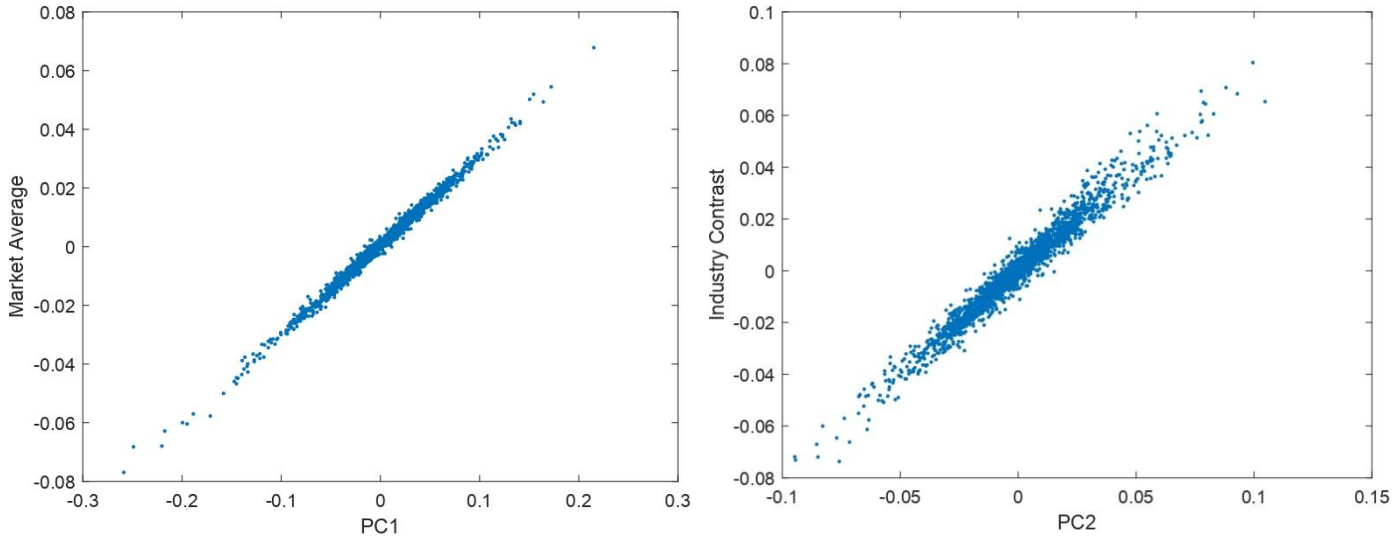


图9.2 主成分与市场平均收益，行业对比配对图

- 第一个主成分与十只股票的平均收益高度相关，也被称为市场成分，其方差比例为45.04%。这十只股票收益率的大部分方差是由市场收益率驱动。
- 第二个主成分代表了紫金矿业和比亚迪与其他股票之间的对比，前两个主成分解释了58.18%的方差。
- 其余的主成分不太容易解释，可能是单只股票的个体特征。然而，它们只解释了方差中的很小一部分。

三、实证分析

Matlab里用极大似然法做因子分析的指令是

$$[lambda, psi, T, stats, F] = factoran(X, m)$$

m：因子数量

$lambda$：因子载荷矩

psi：特异性方差

T：因子旋转矩阵$\hat{Q}$，默认varimax准则

$stats$：因子数量为m的似然比检验结果

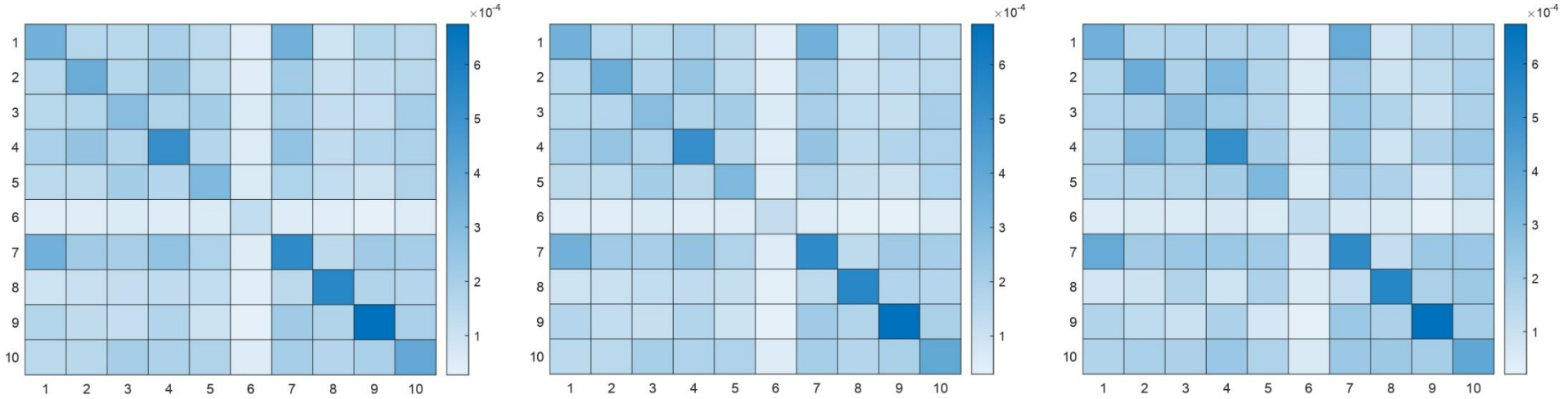
F：因子得分$\hat{F}$

Matlab在运行该指令时会先将原始数据进行标准化，输出结果是针对 X的相关系数矩阵所作的因子分解，所以我们需要将因子载荷矩阵 $lambda$ 乘以 X 的标准差以得到X方差矩阵的因子分解。

表9.2 第一因子载荷

股票代码	PCA	MLE
600519	1.41	1.70
000651	1.32	1.23
601318	1.25	1.18
000333	1.63	1.35
600036	1.17	1.07
600900	0.34	0.31
000858	1.87	1.97
601899	1.19	0.74
002594	1.54	1.06
600030	1.40	1.13

三、实证分析



(a) 样本　　(b) MLE因子模型　　(c) PCA因子模型

图9.3 方差协方差矩阵：样本vs因子模型

谢 谢!