

19

Principal Components Regression

主元回归

输入特征主成分分析，输出数据投影到选定主元超平面



大理石中我看到了天使，我拿起刻刀不停雕刻，直到还它自由。

I saw the angel in the marble and carved until I set him free.

—— 米开朗琪罗 (Michelangelo) | 文艺复兴三杰之一 | 1475 ~ 1564



- seaborn.relplot() 绘制散点图和曲线图
- seaborn.heatmap() 绘制数据热图
- seaborn.jointplot() 绘制联合分布和边际分布
- seaborn.kdeplot() 绘制 KDE 核概率密度估计曲线
- sklearn.decomposition.PCA() 主成分分析函数
- seaborn.lineplot() 绘制线图
- statsmodels.api.add_constant() 线性回归增加一列常数 1
- statsmodels.api.OLS() 最小二乘法函数



19.1 主元回归

本节讲解主元回归 (Principal Components Regression, PCR)。主元回归类似本章前文介绍的正交回归。多元正交回归中，自变量和因变量数据 $[X, y]$ 利用正交化，按照特征值从大小排列特征向量，用 $[v_1, v_2, \dots, v_D]$ 构造一个全新超平面， v_{D+1} 垂直于超平面关系求解出正交化回归系数。

而主元回归，因变量数据 y 完全不参与正交化，即仅仅 X 参与 PCA 分解，获得特征值由大到小排列 D 个主元 $V = (v_1, v_2, \dots, v_D)$ ；这 D 个主元方向 (v_1, v_2, \dots, v_D) 两两正交。选取其中 k ($k < D$) 个特征值较大主元 (v_1, v_2, \dots, v_k) ，构造超平面；最后一步，用最小二乘法将因变量 y 投影在超平面上。

图 1 提供一个例子， X 有三个维度数据， $X = [x_1, x_2, x_3]$ 。首先对 X 列向量 PCA 分解，获得正交化向量 $[v_1, v_2, v_3]$ 。然后，选取作为 v_1 和 v_2 主元，构造一个平面；用最小二乘法，将因变量 y 投影在平面上，获得回归方程。再次请大家注意，主元回归因变量 y 数据并不参与正交化；另外，主元回归选取前 P ($P < D$) 个特征值较大主元 $V_{D \times P} (v_1, v_2, \dots, v_P)$ ，构造一个超平面。

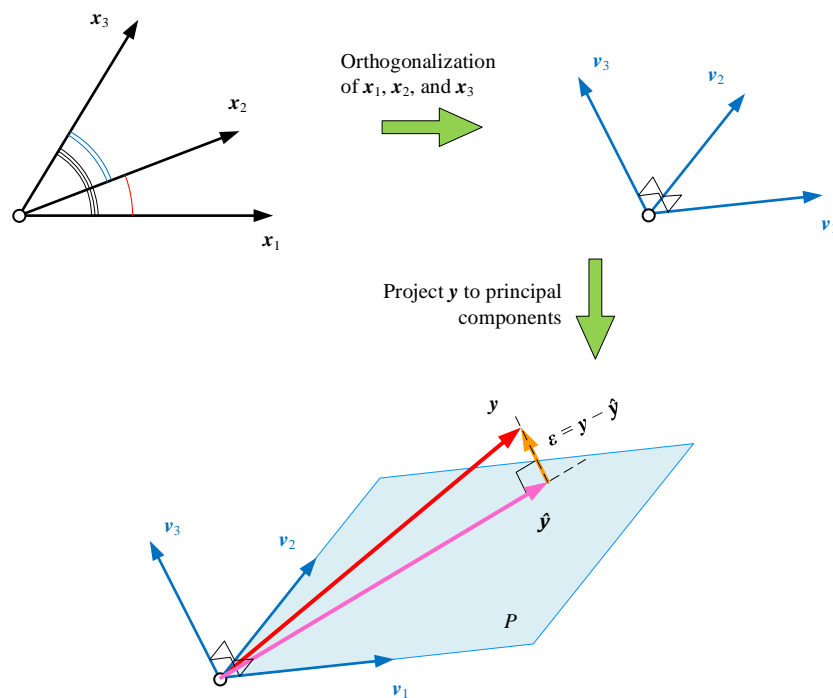


图 1. 主元回归原理

19.2 原始数据

下载如图 2 所示为归一化股价数据，将其转化为日收益率，作为数据 X 和 y ；其中 S&P 500 日收益率为数据 y ，其余股票日收益率作为数据 X 。图 3 所示为数据 X 和 y 的热图。

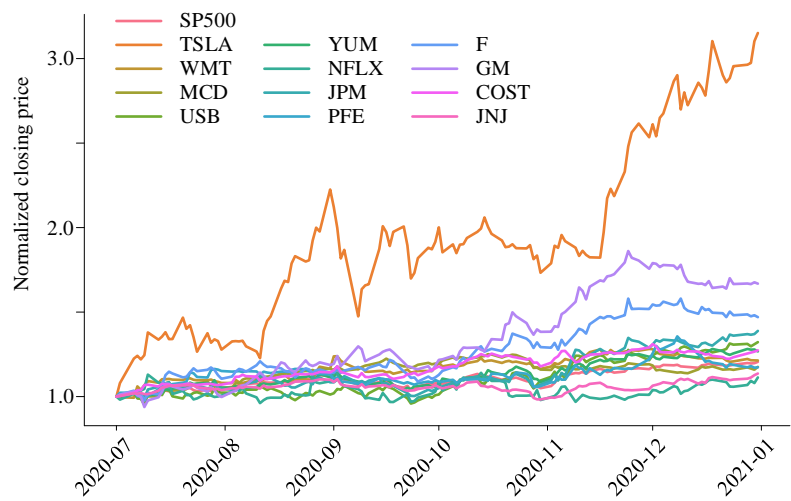


图 2. 股价走势，归一化数据

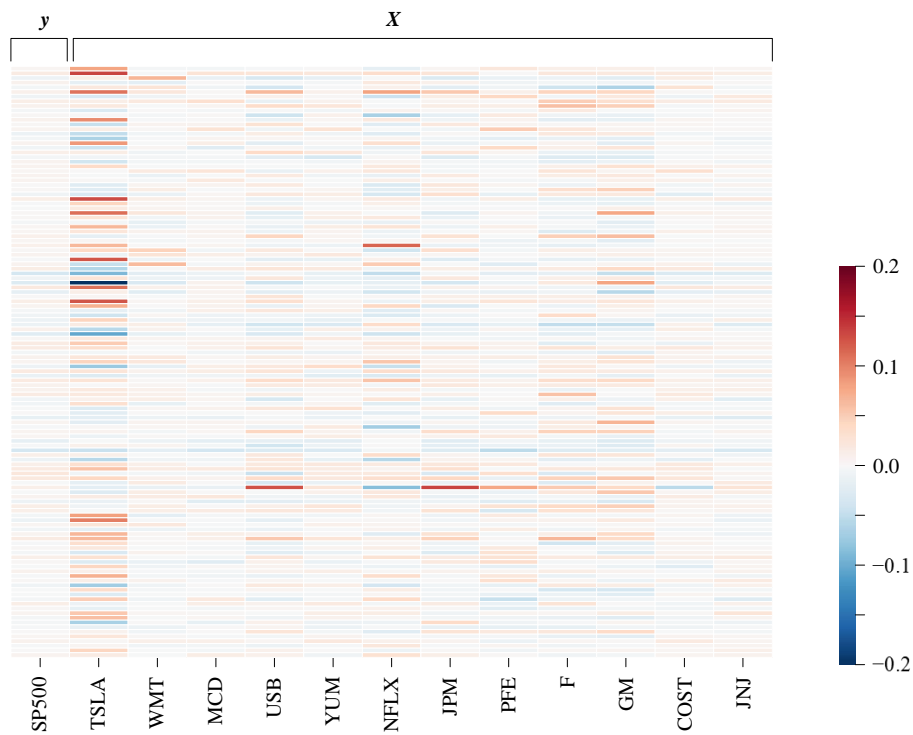


图 3. 数据 X 和 y 的热图

图 4 几个分图给出的是数据 X 和 y 的 KDE 分布。

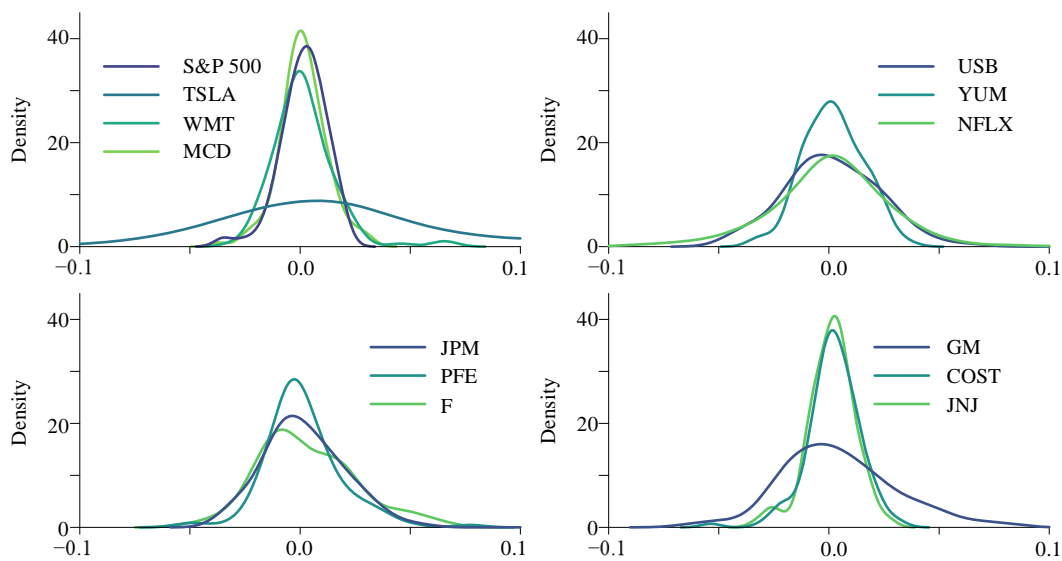


图 4. 数据 X 和 y 的 KDE 分布

19.3 主成分分析

对数据 X 进行主成分分析，可以获得如表 1 所示的前四个主成分 $V_{D \times p}$ 参数。可以利用热图和线图对 $V_{D \times p}$ 进行可视化，如图 5 所示。

表 1. 前四个主成分

	PC1	PC2	PC3	PC4
TSLA	-0.947	-0.004	0.256	0.121
WMT	-0.073	0.016	-0.193	0.066
MCD	-0.056	0.076	-0.111	0.115
USB	-0.021	0.503	0.122	-0.502
YUM	-0.044	0.188	-0.037	0.057
NFLX	-0.281	-0.133	-0.776	-0.448
JPM	-0.019	0.442	0.167	-0.425
PFE	-0.045	0.174	0.187	0.118
F	-0.004	0.457	-0.179	0.178
GM	0.007	0.491	-0.360	0.518
COST	-0.096	-0.027	-0.203	0.114
JNJ	-0.042	0.108	0.021	0.066

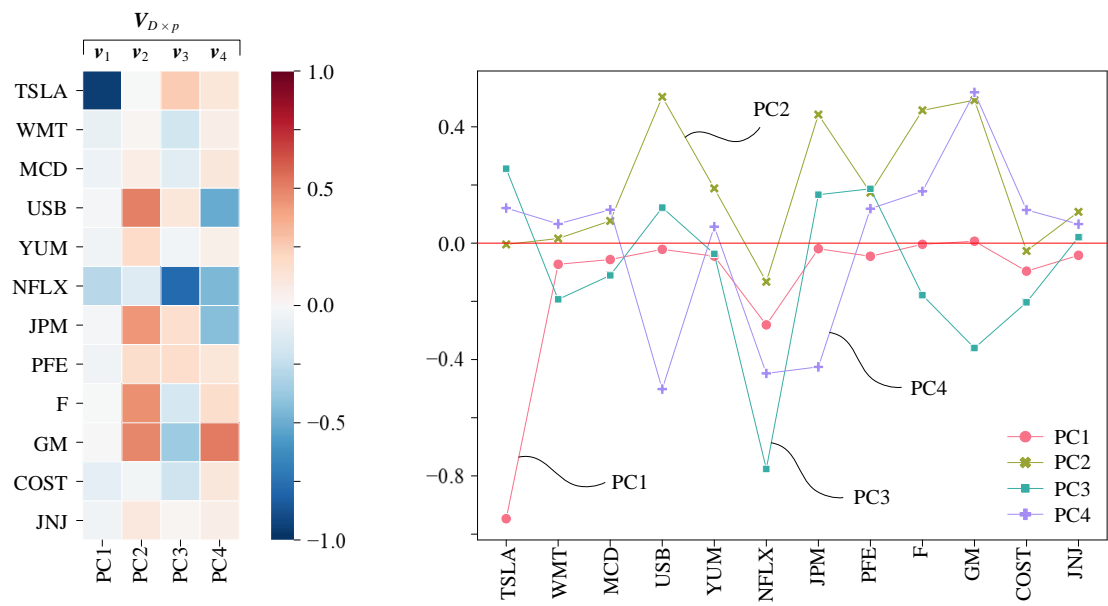


图 5. 前四个主成分可视化

图 5 所示 $V_{D \times p}$ 两两正交，具有如下性质：

$$V_{D \times p}^T V_{D \times p} = I_{p \times p}$$

(1)

图 6 所示为 (1) 计算热图。

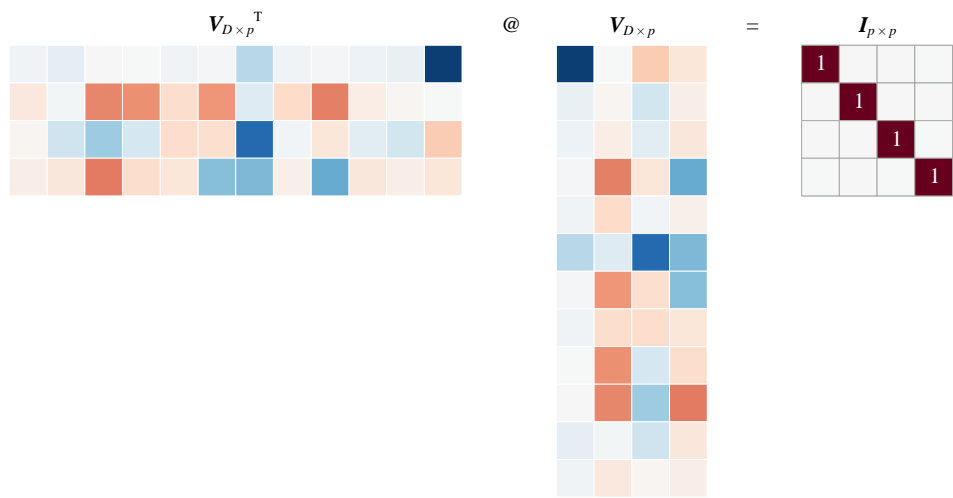


图 6. $V_{D \times p}$ 两两正交

19.4 数据投影

如图 7 所示，原始数据 X 在 p 维正交空间 $(v_1, v_2, ..., v_p)$ 投影得到数据 $Z_{n \times p}$ ：

$$Z_{n \times p} = X_{n \times D} V_{D \times p}$$
(2)

图 8 所示为 $Z_{n \times p}$ 数据热图。

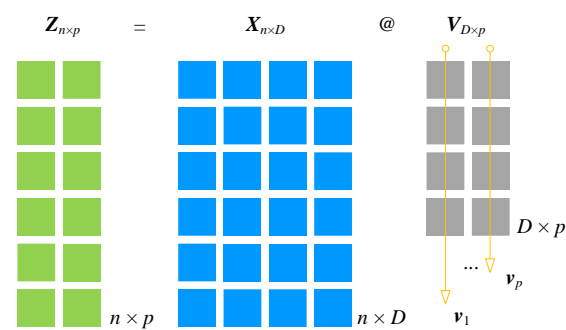


图 7. PCA 分解部分数据关系

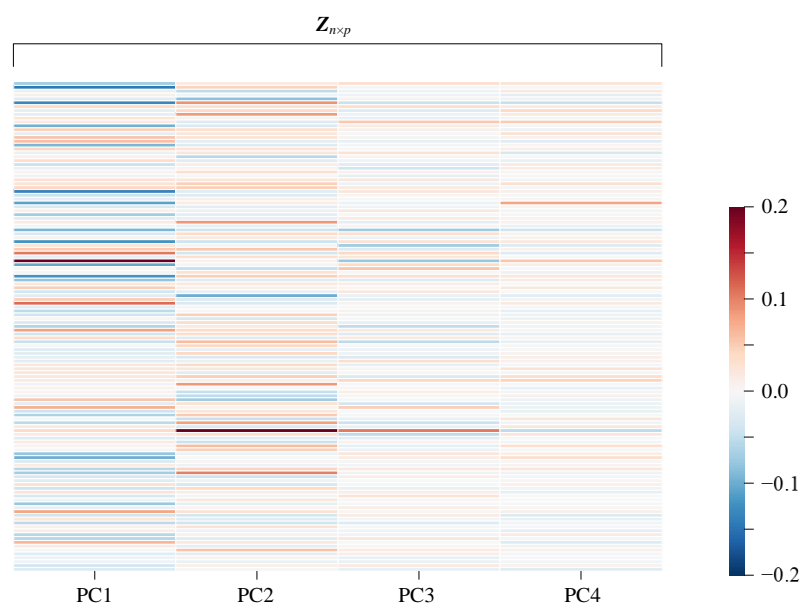


图 8. 前四个主成分数据

图 9 所示为 $Z_{n \times p}$ 每列主成分数据的分布情况。容易注意到，第一主成分数据解释最大方差。

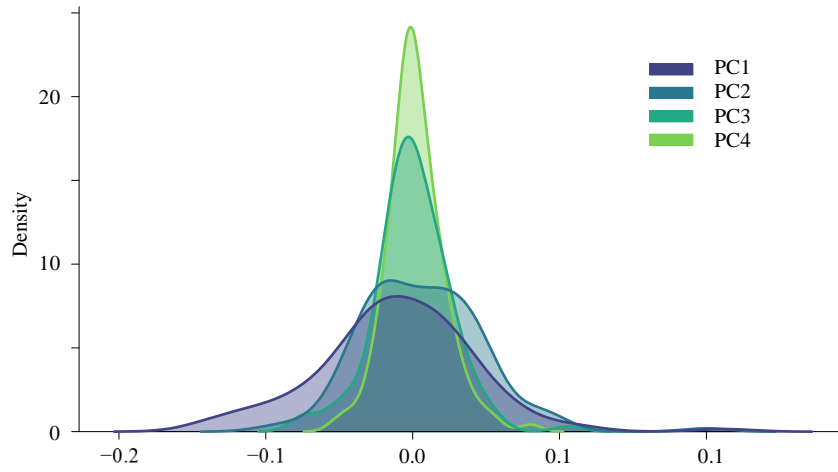


图 9. 前四个主成分数据分布

图 10 所示为 $Z_{n \times p}$ 数协方差矩阵热图。

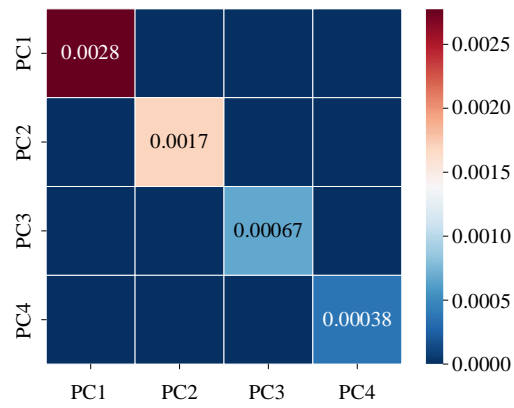


图 10. 前四个主元的协方差矩阵

前四个主成分对应的奇异值分别为：

$$s_1 = 0.5915, \quad s_2 = 0.4624, \quad s_3 = 0.2911, \quad s_4 = 0.2179 \quad (3)$$

所对应的特征值：

$$\begin{aligned} \lambda_1 &= \frac{s_1^2}{n-1} = \frac{0.5915^2}{126} = 0.0028 \\ \lambda_2 &= \frac{s_2^2}{n-1} = \frac{0.4624^2}{126} = 0.0017 \\ \lambda_3 &= \frac{s_3^2}{n-1} = \frac{0.2911^2}{126} = 0.00067 \\ \lambda_4 &= \frac{s_4^2}{n-1} = \frac{0.2179^2}{126} = 0.00038 \end{aligned} \quad (4)$$

这四个特征值对应图 10 热图对角线元素。如图 11 所示陡坡图，前四个主元解释了 84.87% 方差。

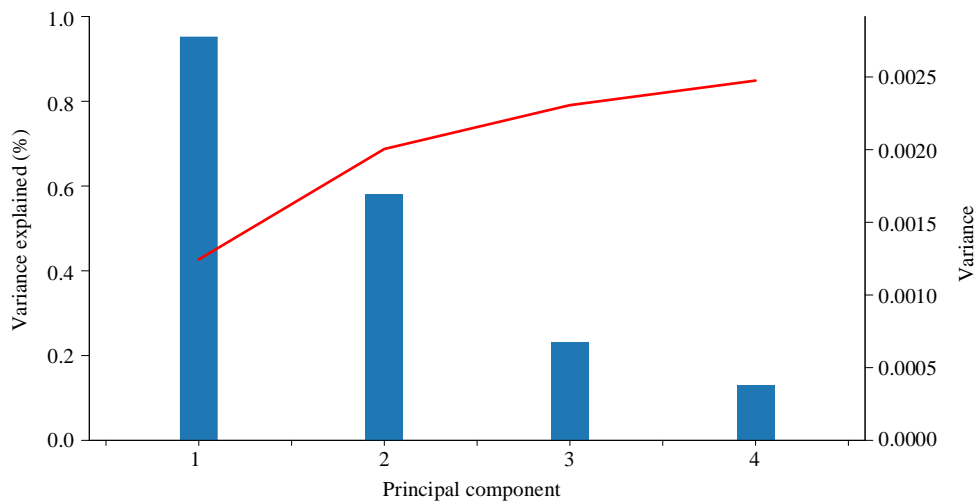


图 11. 陡坡图

转化矩阵 $Z_{n \times P}$ 仅包含 X 部分信息，两者信息之间差距通过下式计算获得，如图 12：

$$X_{n \times D} = Z_{n \times P} (V_{D \times P})^T + E_{n \times D} \tag{5}$$



图 12. $Z_{n \times P}$ 还原数据和 X 信息差距

19.5 最小二乘法

主元回归最后一步，用最小二乘法把因变量 y 投影在数据 $Z_{n \times P}$ 构造空间中：

$$\hat{y} = b_{z,1} z_1 + b_{z,2} z_2 + \dots + b_{z,p} z_p \tag{6}$$

写成矩阵运算：

$$\hat{\mathbf{y}} = \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix} \begin{bmatrix} b_{Z,1} \\ b_{Z,2} \\ \vdots \\ b_{Z,p} \end{bmatrix} = \mathbf{Z}_{n \times P} \mathbf{b}_Z \quad (7)$$

图 13 所示为上述运算过程。

$$\mathbf{y} = \mathbf{Z}_{n \times P} \times \mathbf{b}_Z + \boldsymbol{\varepsilon}$$

图 13. 最小二乘法回归获得 $\mathbf{y} = \mathbf{Z}_{n \times P} \mathbf{b}_Z + \boldsymbol{\varepsilon}$

根据本书前文讲解内容最小二乘法解，获得 \mathbf{b}_Z ：

$$\begin{aligned} \mathbf{b}_Z &= (\mathbf{Z}_{n \times P}^T \mathbf{Z}_{n \times P})^{-1} \mathbf{Z}_{n \times P}^T \mathbf{y} \\ &= \left((\mathbf{X}_{n \times D} \mathbf{V}_{D \times P})^T (\mathbf{X}_{n \times D} \mathbf{V}_{D \times P}) \right)^{-1} (\mathbf{X}_{n \times D} \mathbf{V}_{D \times P})^T \mathbf{y} \end{aligned} \quad (8)$$

如图 13 所示， \mathbf{y} 、拟合数据 $\hat{\mathbf{y}}$ 和数据 $\mathbf{Z}_{n \times P}$ 关系如下：

$$\begin{cases} \mathbf{y} = \mathbf{Z}_{n \times P} \mathbf{b}_Z + \boldsymbol{\varepsilon} \\ \hat{\mathbf{y}} = \mathbf{Z}_{n \times P} \mathbf{b}_Z \\ \boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} \end{cases} \quad (9)$$

图 14 所示为最小二乘法线性回归结果。

系数向量 \mathbf{b}_Z 结果如下：

$$\mathbf{b}_Z = [-0.1039 \quad 0.1182 \quad -0.0941 \quad -0.0418]^T \quad (10)$$

OLS Regression Results						
Dep. Variable:	SP500	R-squared:	0.552			
Model:	OLS	Adj. R-squared:	0.537			
Method:	Least Squares	F-statistic:	37.60			
Date:	XXXXXXXXXX	Prob (F-statistic):	1.82e-20			
Time:	XXXXXXXXXX	Log-Likelihood:	450.53			
No. Observations:	127	AIC:	-891.1			
Df Residuals:	122	BIC:	-876.8			
Df Model:	4					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-0.0003	0.001	-0.520	0.604	-0.002	0.001
PC1	-0.1039	0.012	-8.647	0.000	-0.128	-0.080
PC2	0.1182	0.015	7.689	0.000	0.088	0.149
PC3	-0.0941	0.024	-3.854	0.000	-0.142	-0.046
PC4	-0.0418	0.033	-1.283	0.202	-0.106	0.023
Omnibus:	9.631	Durbin-Watson:	2.087			
Prob(Omnibus):	0.008	Jarque-Bera (JB) :	21.795			
Skew:	0.092	Prob(JB):	1.85e-05			
Kurtosis:	5.021	Cond. No.	51.7			

图 14. 最小二乘法线性回归结果

下面将系数向量 \mathbf{b}_Z 利用 $(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_P)$ 转换为 \mathbf{b}_X ，具体过程图 15 所示：

$$\mathbf{b}_X = \mathbf{V}_{D \times P} \mathbf{b}_Z = \mathbf{V}_{D \times P} \left(\mathbf{Z}_{n \times P}^T \mathbf{Z}_{n \times P} \right)^{-1} \mathbf{Z}_{n \times P}^T \mathbf{y}$$

(11)

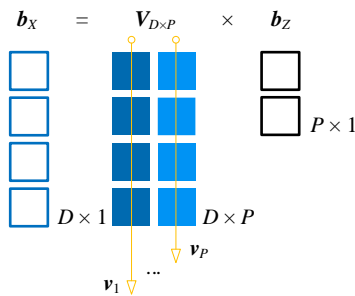


图 15. \mathbf{b}_Z 和 \mathbf{b}_X 之间转换关系

系数 \mathbf{b}_X 可以通过下式计算得到：

$$\mathbf{b}_X = \mathbf{V}_{D \times P} \mathbf{b}_Z = \mathbf{V}_{D \times P} [-0.1039 \quad 0.1182 \quad -0.0941 \quad -0.0418]^T$$

(12)

图 16 所示为系数 \mathbf{b}_X 直方图。

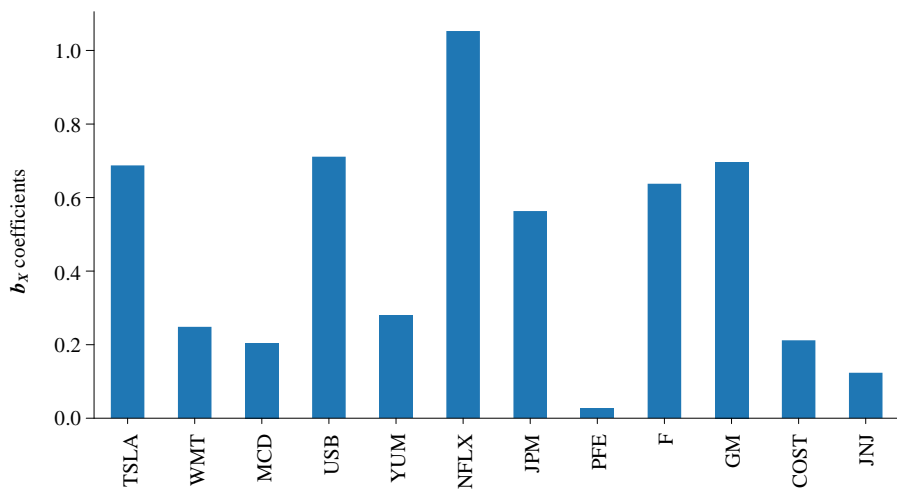


图 16. 系数 b_x 直方图

这样获得 y 、拟合数据 \hat{y} 和数据 X 之间关系，如图 17 所示：

$$\begin{cases} y = Xb_x + \varepsilon \\ \hat{y} = Xb_x \\ \varepsilon = y - \hat{y} \end{cases}$$

(13)

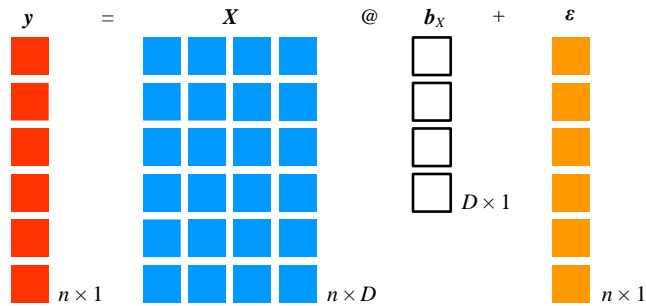


图 17. y 和数据 X 之间回归方程

计算截距项系数 b_0 :

$$b_0 = E(y) - [E(x_1) \ E(x_2) \ \cdots \ E(x_D)]b_x$$

(14)

计算截距项系数 b_0 :

$$\begin{aligned} b_0 &= E(y) - [E(x_1) \ E(x_2) \ \cdots \ E(x_D)]b_x \\ &= -0.00034057 \end{aligned}$$

(15)

最后主元回归函数可以通过下式计算得到：

$$\begin{aligned}
 \hat{y} &= b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_D x_D = b_0 + \begin{bmatrix} x_1 & x_2 & \cdots & x_D \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_D \end{bmatrix} = b_0 + \begin{bmatrix} x_1 & x_2 & \cdots & x_D \end{bmatrix} \mathbf{b}_x \\
 &= b_0 + \begin{bmatrix} z_1 & z_2 & z_3 & z_4 \end{bmatrix} \mathbf{V}_{D \times P} \mathbf{b}_Z \\
 &= b_0 + \begin{bmatrix} z_1 & z_2 & z_3 & z_4 \end{bmatrix} \begin{bmatrix} b_{z1} \\ b_{z2} \\ b_{z3} \\ b_{z4} \end{bmatrix}
 \end{aligned} \tag{16}$$

图 18 展示主元回归计算过程数据关系。

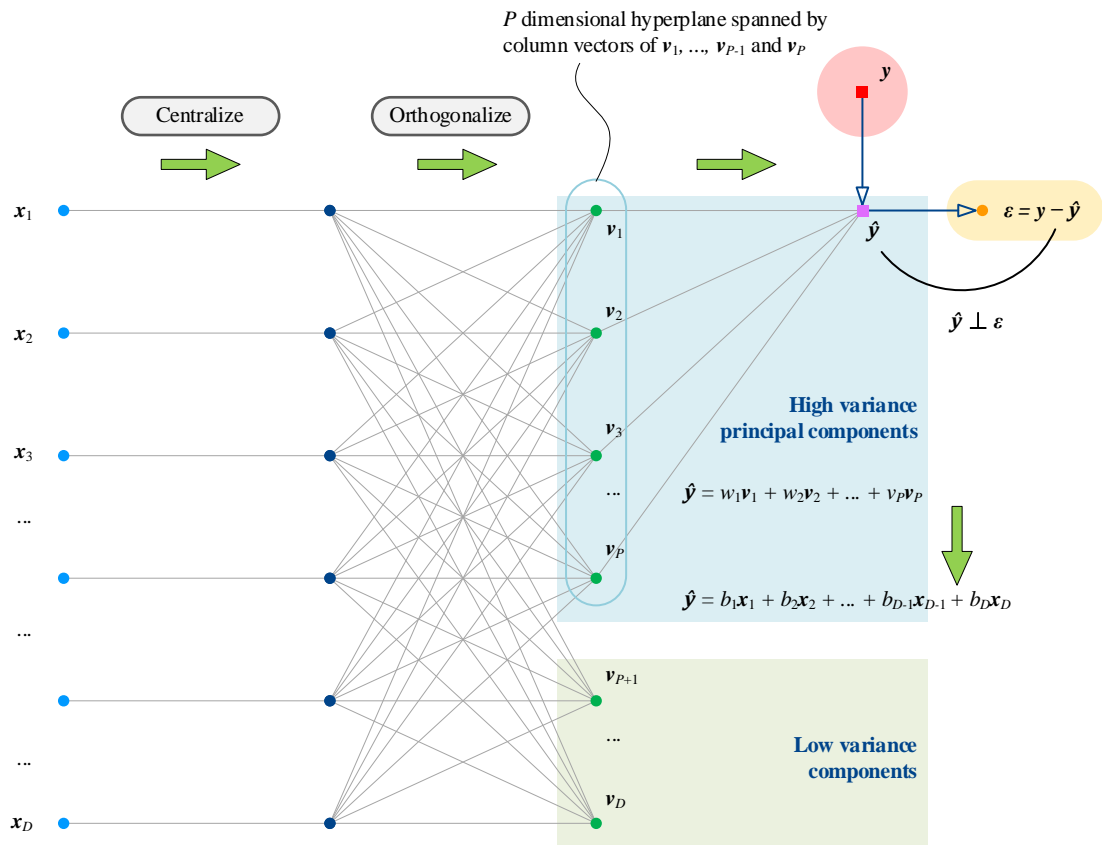


图 18. 主元回归数据关系

19.6 改变主元数量

对于主元回归，当改变参与最小二乘法线性回归的主元数量时，线性回归结果会有很大变化；本节将重点介绍主元数量对主元回归的影响。

图 19 所示为主元数量从 4 增加到 9 时，累计已释方差和百分比变化情况。图 20 和图 21 展示两个视角观察参与主元回归主元数量对于系数的影响。

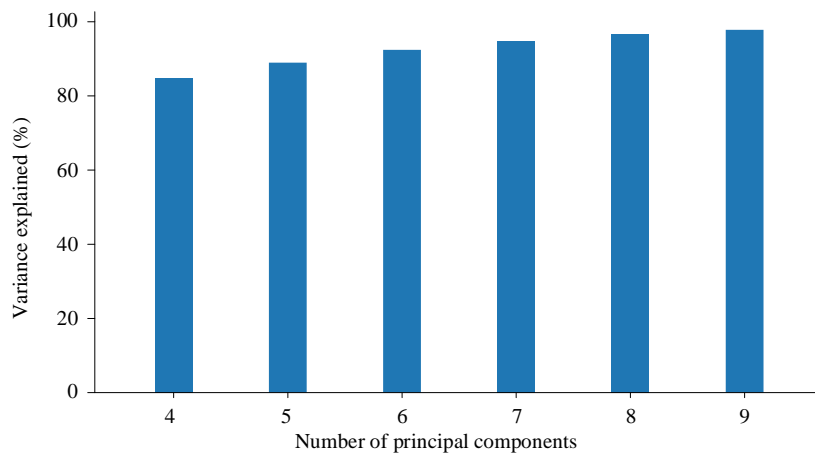


图 19. 主元数量对累计已释方差和百分比

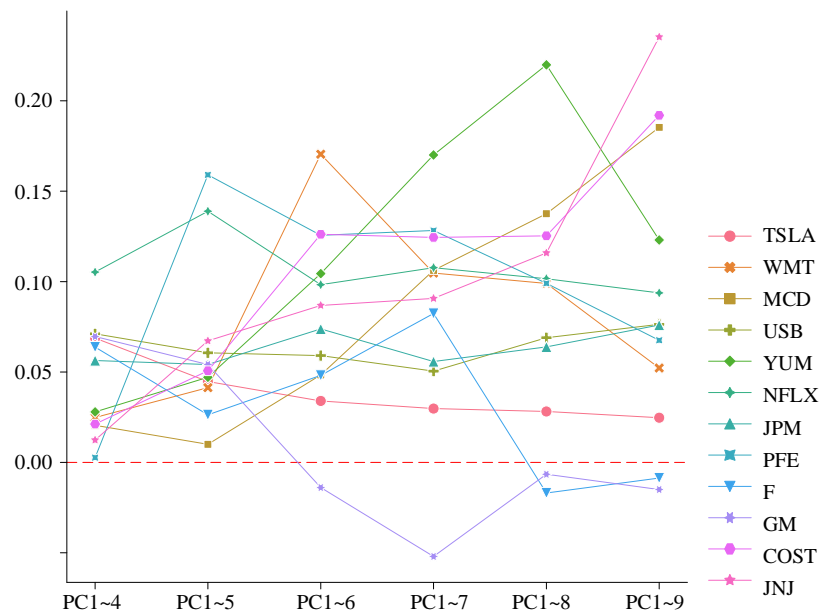


图 20. 参与主元回归主元数量对于系数的影响

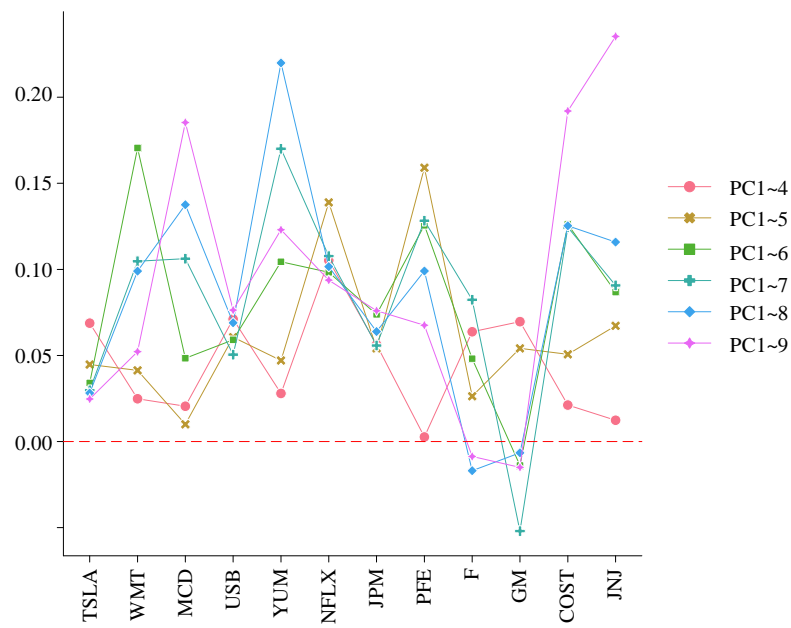


图 21. 参与主元回归主元数量对于系数的影响，第二视角



Bk6_Ch19_01.py 完成本章主元回归运算图像。