18 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

18.1 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 $[\boldsymbol{X}, \boldsymbol{y}]$ 利用正交化，按照特征值从大小排列特征向量，用 $[\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_D]$ 构造一个全新超平面，$\boldsymbol{v}_{D+1}$ 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 $\boldsymbol{y}$ 完全不参与正交化，即仅仅 $\boldsymbol{X}$ 参与 PCA 分解，获得特征值由大到小排列 D 个主元 $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_D)$；这 D 个主元方向 $(\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_D)$ 两两正交。选取其中 k ($k < D$) 个特征值较大主元 $(\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_k)$，构造超平面；最后一步，用最小二乘法将因变量 $\boldsymbol{y}$ 投影在超平面上。

图 1 提供一个例子，$\boldsymbol{X}$ 有三个维度数据，$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3]$。首先对 $\boldsymbol{X}$ 列向量 PCA 分解，获得正交化向量 $[\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]$。然后，选取作为 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 主元，构造一个平面；用最小二乘法，将因变量 $\boldsymbol{y}$ 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 $\boldsymbol{y}$ 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 $\boldsymbol{V}_{D\times P}$ $(\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_P)$，构造一个超平面。

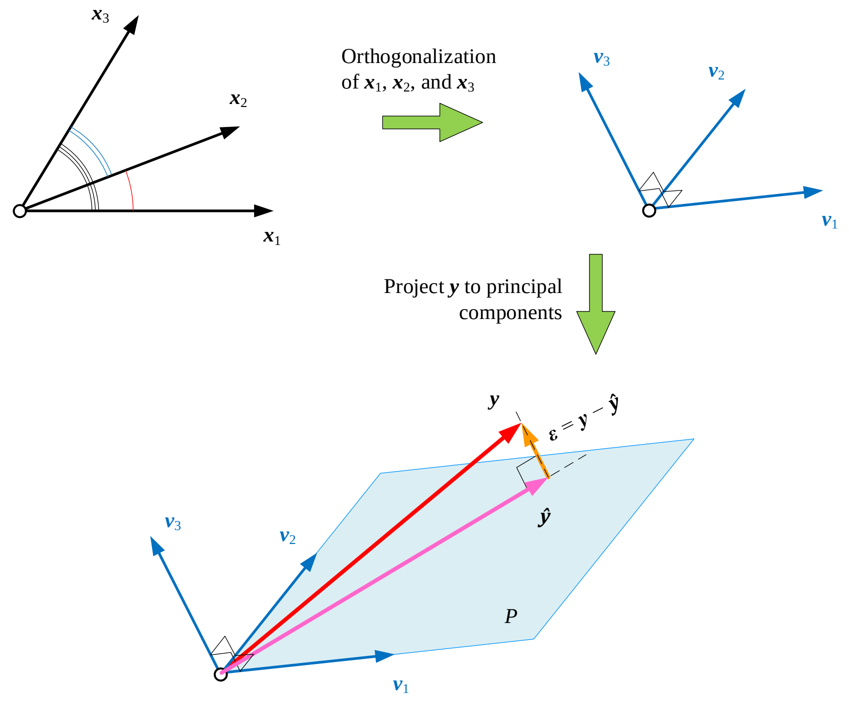

图 1. 主元回归原理

18.2 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 $\boldsymbol{X}$ 和 $\boldsymbol{y}$；其中 S&P 500 日收益率为数据 $\boldsymbol{y}$，其余股票日收益率作为数据 $\boldsymbol{X}$。图 3 所示为数据 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 的热图。

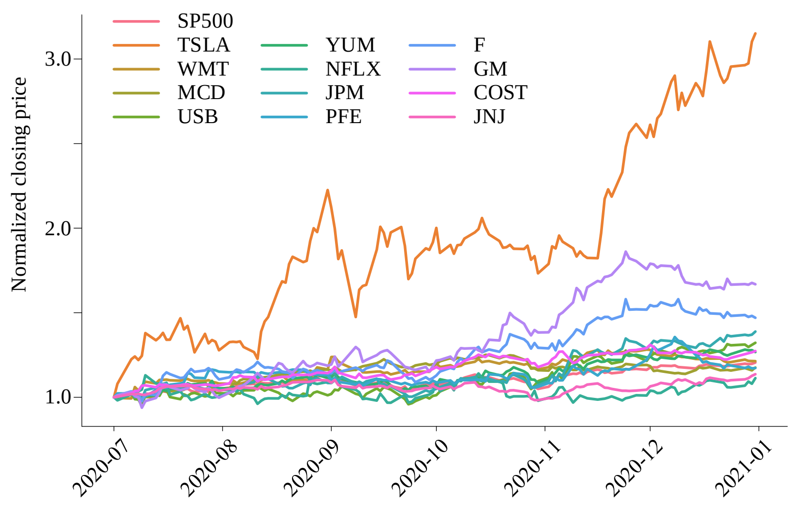

图 2. 股价走势，归一化数据

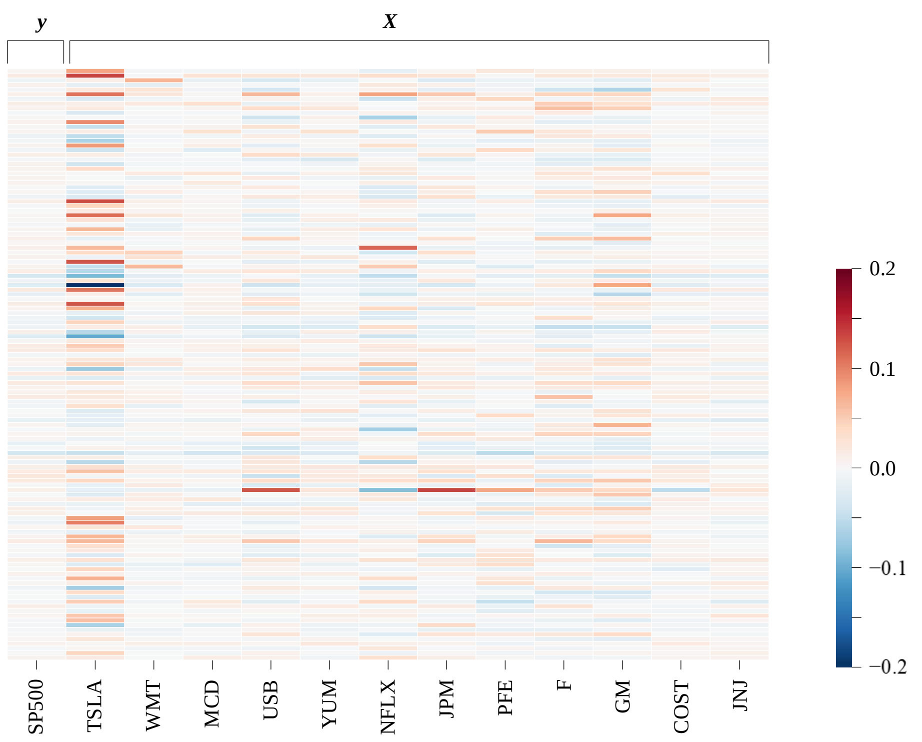

图 3. 数据 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 的热图

图 4 几个分图给出的是数据 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 的 KDE 分布。

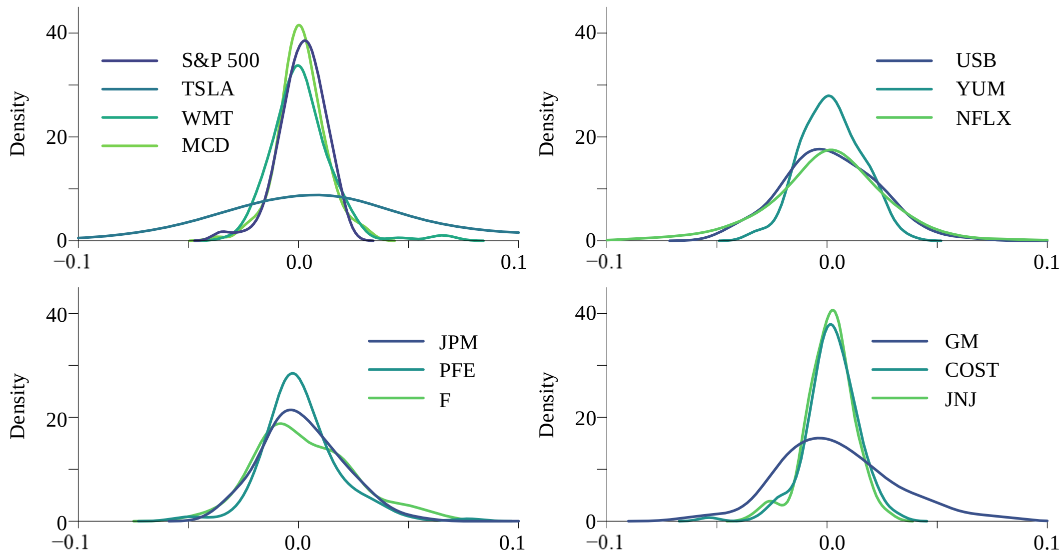

图 4. 数据 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 的 KDE 分布

18.3 主成分分析

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

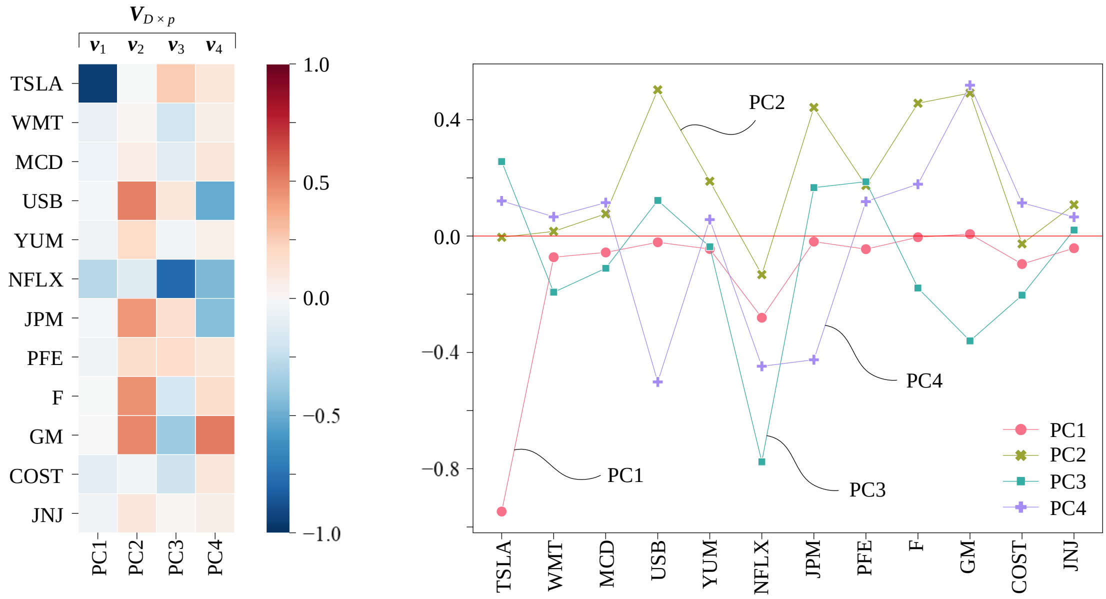

图 5. 前四个主成分可视化

图 5 所示 $\boldsymbol{V}_{D\times p}$ 两两正交，具有如下性质：

$$\boldsymbol{V}_{D\times p}{}^{\mathrm{T}}\boldsymbol{V}_{D\times p} = \boldsymbol{I}_{p\times p} \tag{1}$$

图 6 所示为 (1) 计算热图。

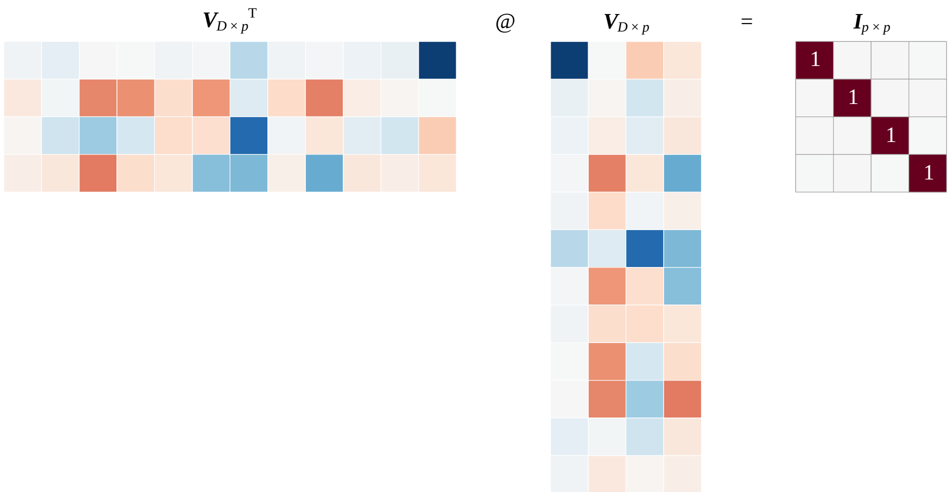

图 6. $\boldsymbol{V}_{D\times p}$ 两两正交

18.4 数据投影

如图 7 所示，原始数据 $\boldsymbol{X}$ 在 p 维正交空间 $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ 投影得到数据 $\boldsymbol{Z}_{n\times p}$：

$$\boldsymbol{Z}_{n\times p} = \boldsymbol{X}_{n\times D}\boldsymbol{V}_{D\times p} \tag{2}$$

图 8 所示为 $\boldsymbol{Z}_{n\times p}$ 数据热图。

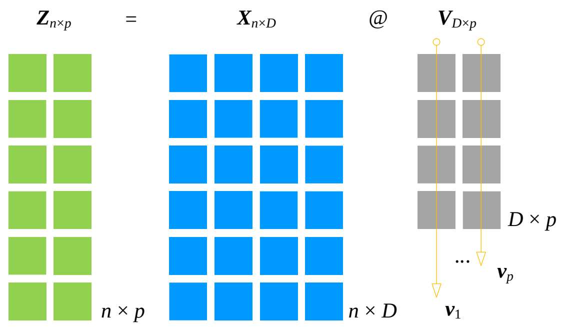

图 7. PCA 分解部分数据关系

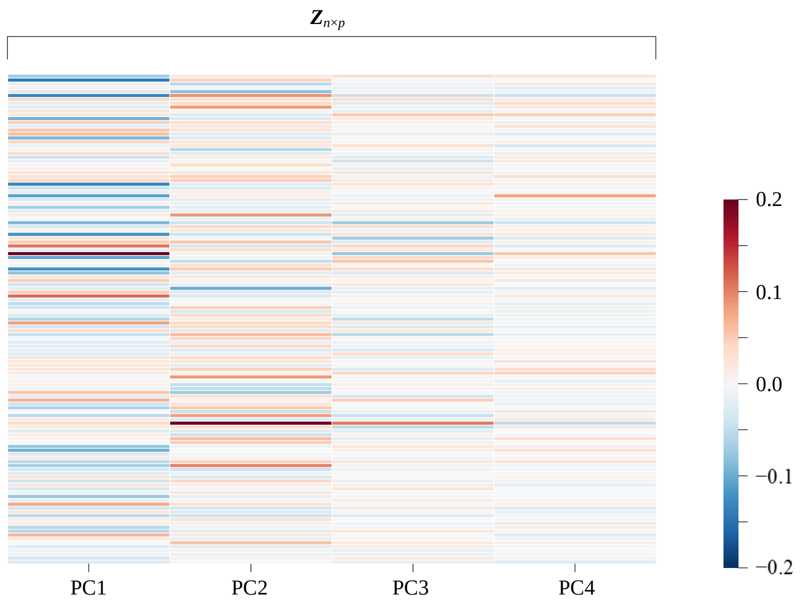

图 8. 前四个主成分数据

图 9 所示为 $\boldsymbol{Z}_{n\times p}$ 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。

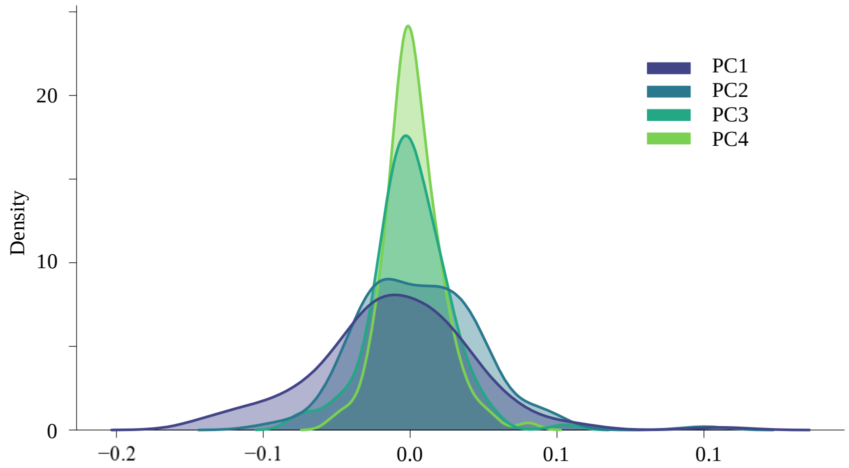

图 9. 前四个主成分数据分布

图 10 所示为 $\boldsymbol{Z}_{n\times p}$ 数协方差矩阵热图。

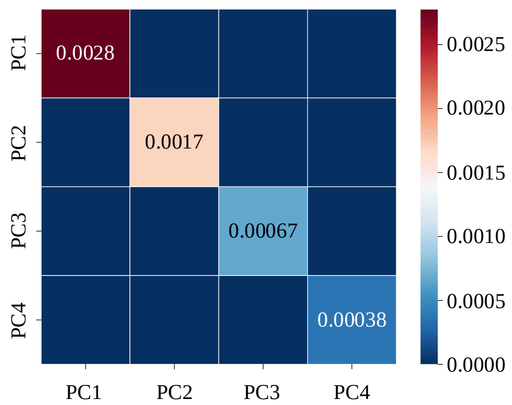

图 10. 前四个主元的协方差矩阵

前四个主成分对应的奇异值分别为：

$$s_1 = 0.5915, \quad s_2 = 0.4624, \quad s_3 = 0.2911, \quad s_4 = 0.2179 \tag{3}$$

所对应的特征值：

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

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。

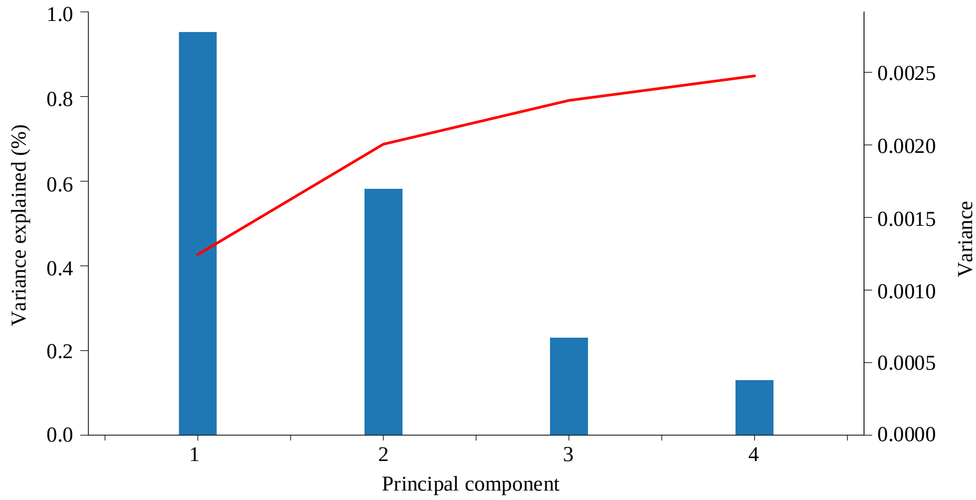

图 11. 陡坡图

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

$$\boldsymbol{X}_{n\times D}=\boldsymbol{Z}_{n\times P}\left(\boldsymbol{V}_{D\times P}\right)^{\mathrm{T}}+\boldsymbol{E}_{n\times D} \tag{5}$$

图 12. $\boldsymbol{Z}_{n\times P}$ 还原数据和 $\boldsymbol{X}$ 信息差距

18.5 最小二乘法

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

$$\hat{\boldsymbol{y}}=b_{Z,1}\boldsymbol{z}_1+b_{Z,2}\boldsymbol{z}_2+\ldots+b_{Z,P}\boldsymbol{z}_P \tag{6}$$

写成矩阵运算：

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_P \end{bmatrix} \begin{bmatrix} b_{Z,1} \\ b_{Z,2} \\ \vdots \\ b_{Z,P} \end{bmatrix} = \boldsymbol{Z}_{n\times P}\boldsymbol{b}_Z \tag{7}$$

图 13 所示为上述运算过程。

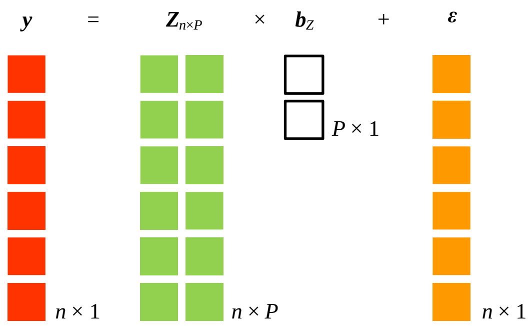

图 13. 最小二乘法回归获得 $\boldsymbol{y} = \boldsymbol{Z}_{n\times P}\boldsymbol{b}_Z + \boldsymbol{\varepsilon}$

根据本书前文讲解内容最小二乘法解，获得 $\boldsymbol{b}_Z$：

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

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

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

图 14 所示为最小二乘法线性回归结果。

系数向量 $\boldsymbol{b}_Z$ 结果如下：

$$\boldsymbol{b}_Z = \begin{bmatrix} -0.1039 & 0.1182 & -0.0941 & -0.0418 \end{bmatrix}^{\mathrm{T}} \tag{10}$$

```
                            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 $\boldsymbol{b}_Z$ 利用 $(\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_P)$ 转换为 $\boldsymbol{b}_X$，具体过程图 15 所示：

$$\boldsymbol{b}_X = \boldsymbol{V}_{D\times P}\boldsymbol{b}_Z = \boldsymbol{V}_{D\times P}\left(\boldsymbol{Z}_{n\times P}{}^{\mathrm{T}}\boldsymbol{Z}_{n\times P}\right)^{-1}\boldsymbol{Z}_{n\times P}{}^{\mathrm{T}}\boldsymbol{y} \tag{11}$$

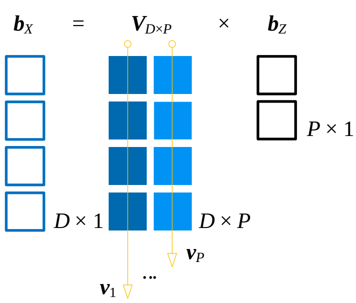

图 15. $\boldsymbol{b}_Z$ 和 $\boldsymbol{b}_X$ 之间转换关系

系数 $\boldsymbol{b}_X$ 可以通过下式计算得到：

$$\boldsymbol{b}_X = \boldsymbol{V}_{D\times P}\boldsymbol{b}_Z = \boldsymbol{V}_{D\times P}\begin{bmatrix}-0.1039 & 0.1182 & -0.0941 & -0.0418\end{bmatrix}^{\mathrm{T}} \tag{12}$$

图 16 所示为系数 $\boldsymbol{b}_X$ 直方图。

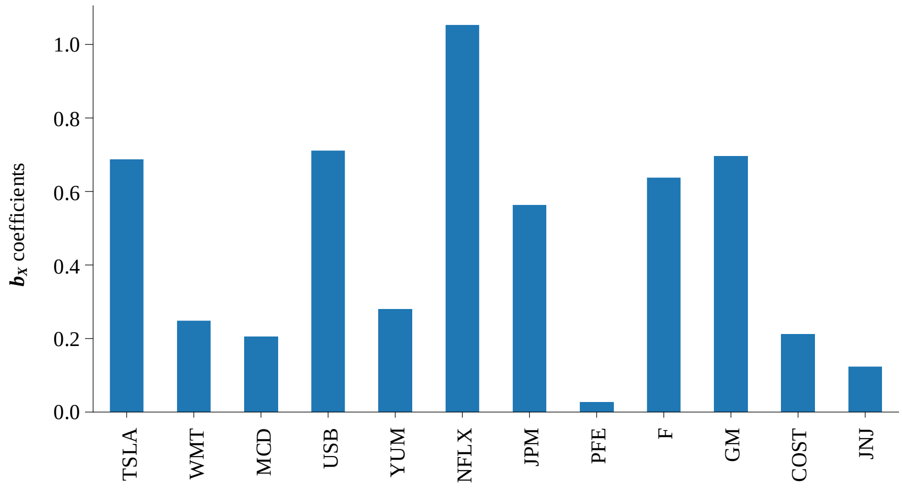

图 16. 系数 $\boldsymbol{b}_X$ 直方图

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

$$\begin{cases} \boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}_X + \boldsymbol{\varepsilon} \\ \hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}_X \\ \boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} \end{cases} \tag{13}$$

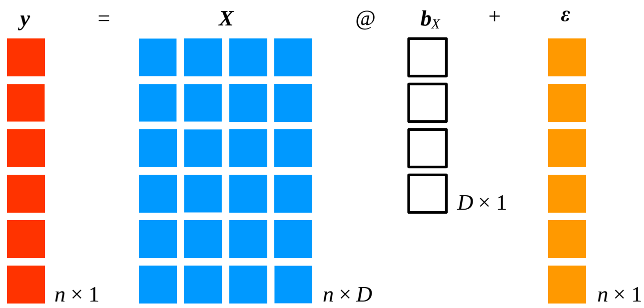

图 17. $\boldsymbol{y}$ 和数据 $\boldsymbol{X}$ 之间回归方程

计算截距项系数 b_0：

$$b_0 = \mathrm{E}(\boldsymbol{y}) - \begin{bmatrix} \mathrm{E}(\boldsymbol{x}_1) & \mathrm{E}(\boldsymbol{x}_2) & \cdots & \mathrm{E}(\boldsymbol{x}_D) \end{bmatrix} \boldsymbol{b}_X \tag{14}$$

计算截距项系数 b_0：

$$\begin{aligned} b_0 &= \mathrm{E}(\boldsymbol{y}) - \begin{bmatrix} \mathrm{E}(\boldsymbol{x}_1) & \mathrm{E}(\boldsymbol{x}_2) & \cdots & \mathrm{E}(\boldsymbol{x}_D) \end{bmatrix} \boldsymbol{b}_X \\ &= -0.00034057 \end{aligned} \tag{15}$$

最后主元回归函数可以通过下式计算得到：

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

图 18 展示主元回归计算过程数据关系。

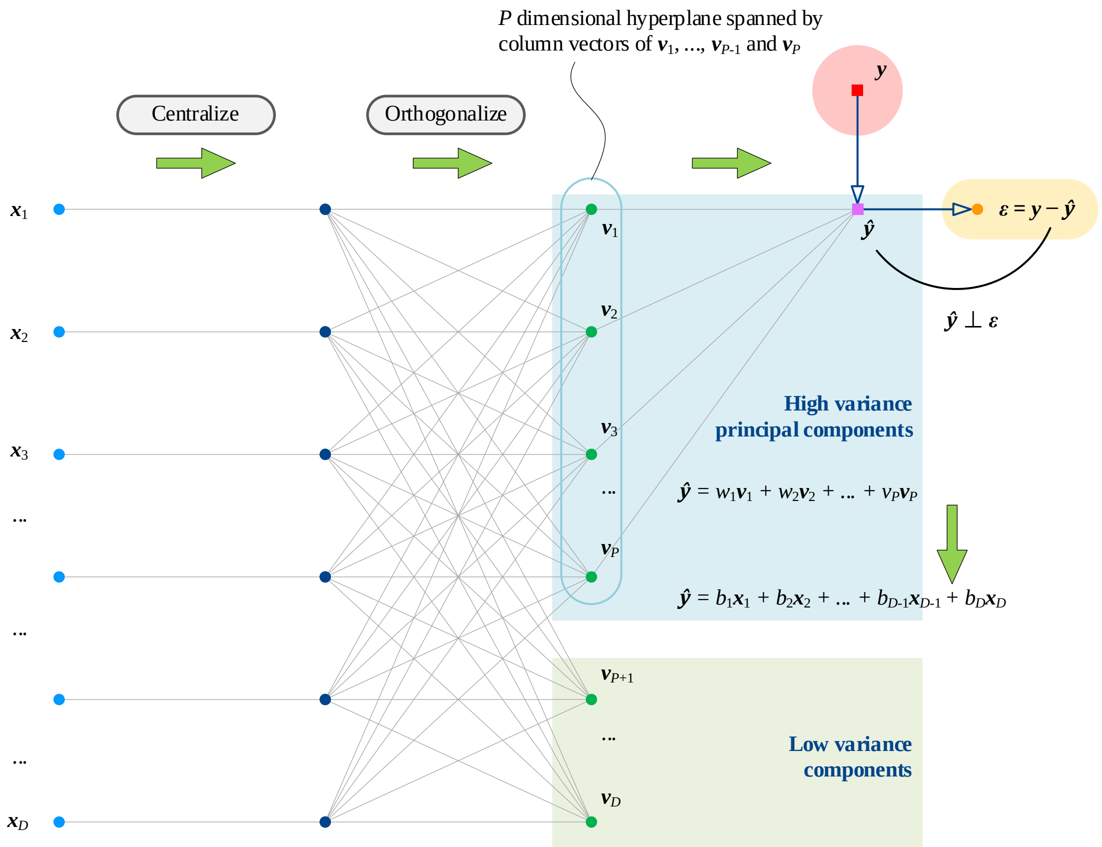

图 18. 主元回归数据关系

18.6 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。

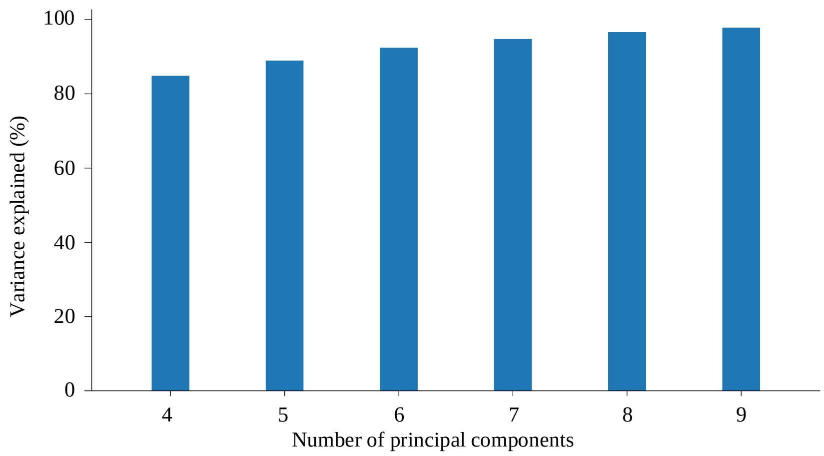

图 19. 主元数量对累计已释方差和百分比

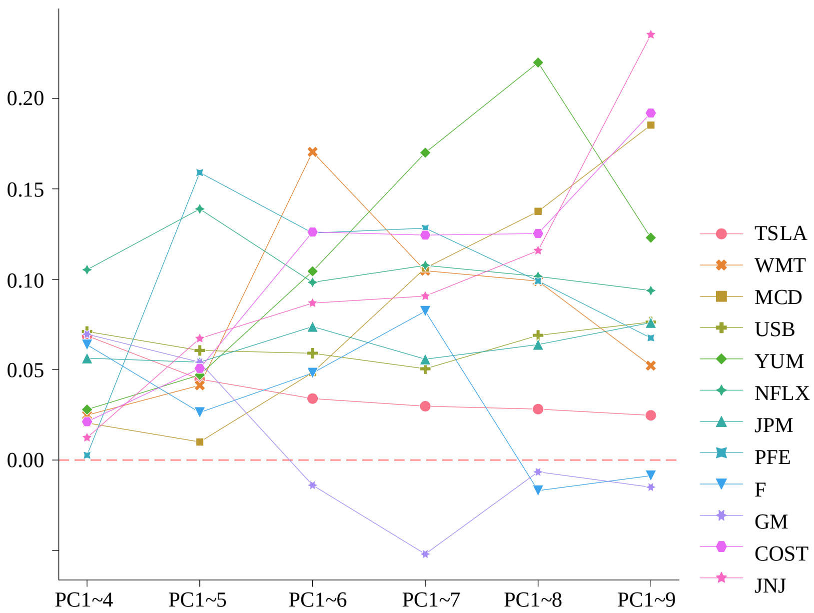

图 20. 参与主元回归主元数量对于系数的影响

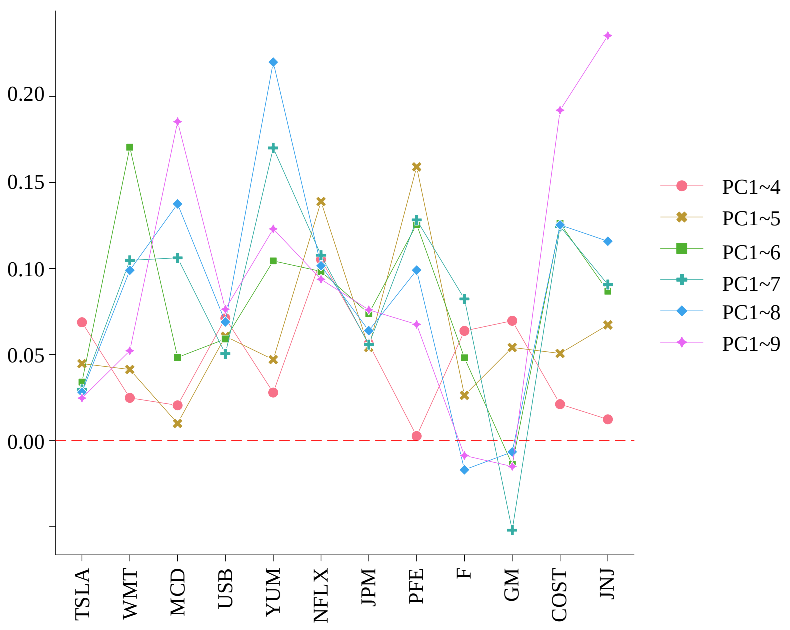

图 21. 参与主元回归主元数量对于系数的影响，第二视角

Bk6_Ch18_01.py 完成主元回归运算图像。

18.7 偏最小二乘回归

本章最后介绍偏最小二乘回归 (partial least squares regression, PLS)。类似主元回归，偏最小二乘回归也是一种降维回归方法。

不同于主元回归，偏最小二乘回归利用因变量数据 $\boldsymbol{y}$ 和自变量数据 $\boldsymbol{X}$ (形状为 $n \times q$) 之间相关性构造一个全新空间。$\boldsymbol{y}$ 和 $\boldsymbol{X}$ 投影到新空间来确定一个线性回归模型。另外一个不同点，偏最小二乘回归采用迭代算法 (iterative algorithm)。

偏最小二乘法处理多元因变量，为方便区分，一元因变量被定义为 $\boldsymbol{y}$ (形状为 $n \times 1$)，多元因变量被定义为 $\boldsymbol{Y}$ (形状为 $n \times p$)。偏最小二乘回归迭代方法很多，本节介绍较为经典一元因变量对多元自变量迭代算法。迭代算法主要由七步构成；其中，第二步到第七步为循环。

第一步

获得中心化自变量数据矩阵 $\boldsymbol{X}^{(0)}$ 和因变量数据向量 $\boldsymbol{y}^{(0)}$：

$$\boldsymbol{X}^{(0)}=\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{l}\boldsymbol{l}^{\mathrm{T}}\right)\boldsymbol{X}=\begin{bmatrix}\boldsymbol{x}_1^{(0)} & \boldsymbol{x}_2^{(0)} & \cdots & \boldsymbol{x}_q^{(0)}\end{bmatrix}$$
$$\boldsymbol{y}^{(0)}=\boldsymbol{y}-\mathrm{E}(\boldsymbol{y})=\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{l}\boldsymbol{l}^{\mathrm{T}}\right)\boldsymbol{y} \tag{17}$$

偏最小二乘回归是迭代运算，上标 (0) 代表迭代代次。

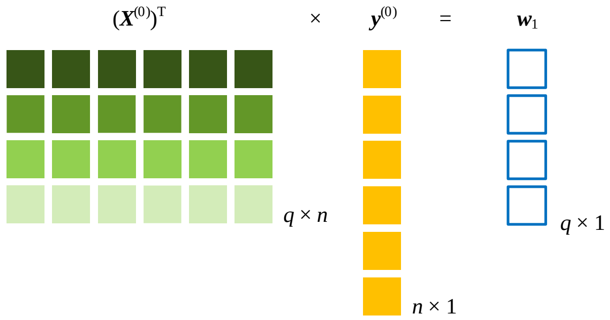

图 22. 计算权重系数列向量 $\boldsymbol{w}_1$

第二步

计算 $\boldsymbol{y}^{(0)}$ 和 $\boldsymbol{X}^{(0)}$ 列向量相关性，构建权重系数列向量 $\boldsymbol{w}_1$：

$$\boldsymbol{w}_1=\begin{bmatrix}\operatorname{cov}\left(\boldsymbol{x}_1^{(0)},\boldsymbol{y}^{(0)}\right)\\ \operatorname{cov}\left(\boldsymbol{x}_2^{(0)},\boldsymbol{y}^{(0)}\right)\\ \vdots \\ \operatorname{cov}\left(\boldsymbol{x}_q^{(0)},\boldsymbol{y}^{(0)}\right)\end{bmatrix}=\frac{1}{n}\begin{bmatrix}\left(\boldsymbol{x}_1^{(0)}\right)^{\mathrm{T}}\boldsymbol{y}^{(0)}\\ \left(\boldsymbol{x}_2^{(0)}\right)^{\mathrm{T}}\boldsymbol{y}^{(0)}\\ \vdots \\ \left(\boldsymbol{x}_q^{(0)}\right)^{\mathrm{T}}\boldsymbol{y}^{(0)}\end{bmatrix}=\left(\boldsymbol{X}^{(0)}\right)^{\mathrm{T}}\boldsymbol{y}^{(0)} \tag{18}$$

其中，列向量 $\boldsymbol{w}_1$ 行数为 q 行。

图 22 所示获得权重系数列向量计算过程；过程也可看做是一个投影运算，即将 $(\boldsymbol{X}^{(0)})^{\mathrm{T}}$ 投影到 $\boldsymbol{y}^{(0)}$。

为方便计算，将列向量 $\boldsymbol{w}_1$ 单位化：

$$\boldsymbol{w}_1=\frac{\boldsymbol{w}_1}{\left\|\boldsymbol{w}_1\right\|}=\begin{bmatrix}w_{1,1}\\ w_{2,1}\\ \vdots \\ w_{q,1}\end{bmatrix} \tag{19}$$

列向量 $\boldsymbol{w}_1$ 每个元素大小代表着 $\boldsymbol{y}^{(0)}$ 和 $\boldsymbol{X}^{(0)}$ 列向量相关性。

第三步，利用上一步获得权重系数列向量 $\boldsymbol{w}_1$ 和 $\boldsymbol{X}^{(0)}$ 构造偏最小二乘回归主元向量，$\boldsymbol{z}_1$：

$$\boldsymbol{z}_1=w_{1,1}\boldsymbol{x}_1+w_{2,1}\boldsymbol{x}_2+\cdots+w_{q,1}\boldsymbol{x}_q=\boldsymbol{X}^{(0)}\boldsymbol{w}_1 \tag{20}$$

图 23 所示为计算偏最小二程回归主元列向量 $\boldsymbol{z}_1$。这样理解，主元列向量 $\boldsymbol{z}_1$ 为 $\boldsymbol{X}^{(0)}$ 列向量通过加权构造；$\boldsymbol{y}^{(0)}$ 和 $\boldsymbol{X}^{(0)}$ 某一列向量相关性越高，这一列获得权重越高，在主元列向量 $\boldsymbol{z}_1$ 成分越高。同样，过程等价于投影过程，即 $\boldsymbol{X}^{(0)}$ 投影到 $\boldsymbol{w}_1$。

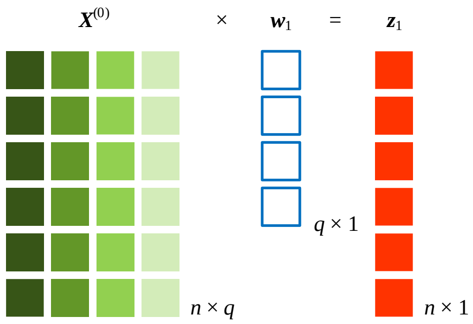

图 23. 计算偏最小二程回归主元列向量 $\boldsymbol{z}_1$

将自变量数据矩阵 $\boldsymbol{X}^{(0)}$ 和因变量数据向量 $\boldsymbol{y}^{(0)}$投影到主元 $\boldsymbol{z}_1$ 方向上。

第四步

把自变量数据矩阵 $\boldsymbol{X}^{(0)}$投影到主元列向量 $\boldsymbol{z}_1$ 上，获得系数向量 $\boldsymbol{v}_1$。先以 $\boldsymbol{X}^{(0)}$第一列解释投影过程。

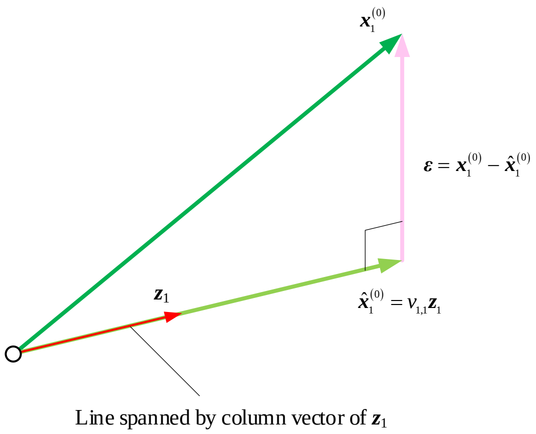

图 24. $\boldsymbol{X}^{(0)}$第一列投影在主元列向量 $\boldsymbol{z}_1$

如图 24 所示，将 $\boldsymbol{X}^{(0)}$ 第一列投影到主元列向量 $\boldsymbol{z}_1$，得到 $\hat{\boldsymbol{x}}_1^{(0)}$：

$$\hat{\boldsymbol{x}}_1^{(0)} = v_{1,1}\boldsymbol{z}_1 \tag{21}$$

残差 $\boldsymbol{\varepsilon}$ 则垂直于主元列向量 $\boldsymbol{z}_1$，计算获得系数 $v_{1,1}$：

$$
\begin{aligned}
&\boldsymbol{\varepsilon} \perp \boldsymbol{z}_1 \quad \Rightarrow \quad \boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\varepsilon} = \boldsymbol{z}_1^{\mathrm{T}}\left(\boldsymbol{x}_1^{(0)} - \hat{\boldsymbol{x}}_1^{(0)}\right) = \boldsymbol{z}_1^{\mathrm{T}}\left(\boldsymbol{x}_1^{(0)} - v_{1,1}\boldsymbol{z}_1\right) = 0 \\
&\Rightarrow \quad v_{1,1} = \frac{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{x}_1^{(0)}}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1} = \frac{\left(\boldsymbol{x}_1^{(0)}\right)^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}
\end{aligned}
\tag{22}
$$

上式说明偏最小二乘法回归核心仍是 OLS。同样，把 $\boldsymbol{X}^{(0)}$ 第二列投影在主元列向量 $\boldsymbol{z}_2$，计算得到系数 $v_{2,1}$：

$$
v_{2,1} = \frac{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{x}_2^{(0)}}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1} = \frac{\left(\boldsymbol{x}_2^{(0)}\right)^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}
\tag{23}
$$

类似，获得 $\boldsymbol{X}^{(0)}$ 每列投影在主元列向量 $\boldsymbol{z}_2$ 系数，这些系数一个列向量 $\boldsymbol{v}_1$。下式计算列向量 $\boldsymbol{v}_1$：

$$
\boldsymbol{v}_1 = \begin{bmatrix} v_{1,1} \\ v_{2,1} \\ \vdots \\ v_{q,1} \end{bmatrix} = \frac{\left(\boldsymbol{X}^{(0)}\right)^{\mathrm{T}}\boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1} = \frac{\left(\boldsymbol{X}^{(0)}\right)^{\mathrm{T}}\boldsymbol{X}^{(0)}\boldsymbol{w}_1}{\boldsymbol{w}_1^{\mathrm{T}}\left(\boldsymbol{X}^{(0)}\right)^{\mathrm{T}}\boldsymbol{X}^{(0)}\boldsymbol{w}_1} = \frac{\boldsymbol{\Sigma}^{(0)}\boldsymbol{w}_1}{\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{(0)}\boldsymbol{w}_1}
\tag{24}
$$

第五步

根据最小二乘回归原理，利用列向量 $\boldsymbol{v}_1$ 和 $\boldsymbol{z}_1$ 估算，并到拟合矩阵 $\hat{\boldsymbol{X}}^{(0)}$：

$$
\hat{\boldsymbol{X}}^{(0)} = \boldsymbol{z}_1\boldsymbol{v}_1^{\mathrm{T}} = \boldsymbol{X}^{(0)}\boldsymbol{w}_1\boldsymbol{v}_1^{\mathrm{T}}
\tag{25}
$$

原始数据矩阵 $\boldsymbol{X}$ 和拟合数据矩阵 $\hat{\boldsymbol{X}}^{(0)}$ 之差便是残差矩阵 $\boldsymbol{E}^{(0)}$：

$$
\boldsymbol{E}^{(0)} = \boldsymbol{X}^{(0)} - \hat{\boldsymbol{X}}^{(0)} = \boldsymbol{X}^{(0)} - \boldsymbol{X}^{(0)}\boldsymbol{w}_1\boldsymbol{v}_1^{\mathrm{T}} = \boldsymbol{X}^{(0)}\left(\boldsymbol{I} - \boldsymbol{w}_1\boldsymbol{v}_1^{\mathrm{T}}\right)
\tag{26}
$$

而残差矩阵 $\boldsymbol{E}^{(0)}$ 便是进入迭代过程第二步数据矩阵 $\boldsymbol{X}^{(1)}$：

$$
\boldsymbol{X}^{(1)} = \boldsymbol{E}^{(0)} = \boldsymbol{X}^{(0)} - \hat{\boldsymbol{X}}^{(0)} = \boldsymbol{X}^{(0)}\left(\boldsymbol{I} - \boldsymbol{w}_1\boldsymbol{v}_1^{\mathrm{T}}\right)
\tag{27}
$$

数据矩阵 $\boldsymbol{X}^{(1)}$ 和原始数据 $\boldsymbol{X}^{(0)}$ 之间关系如图 25 所示。

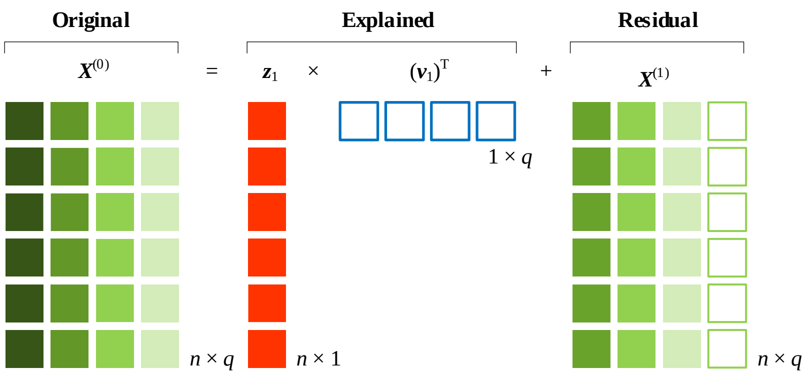

图 25. 计算得到数据矩阵 $\boldsymbol{X}^{(1)}$

第六步

把因变量数据列向量 $\boldsymbol{y}^{(0)}$ 投影于主元列向量 $\boldsymbol{z}_1$ 上，获得系数 b_1。类似第四步，如图 26 所示，用最小二乘法计算获得系数 b_1：

$$\begin{aligned} & \boldsymbol{\varepsilon} \perp \boldsymbol{z}_1 \quad \Rightarrow \quad \boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{\varepsilon} = \boldsymbol{z}_1^{\mathrm{T}} \left(\boldsymbol{y}^{(0)} - \hat{\boldsymbol{y}}^{(0)} \right) = \boldsymbol{z}_1^{\mathrm{T}} \left(\boldsymbol{y}^{(0)} - b_1 \boldsymbol{z}_1 \right) = 0 \\ & \Rightarrow \quad b_1 = \frac{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{y}^{(0)}}{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{z}_1} = \frac{\left(\boldsymbol{y}^{(0)} \right)^{\mathrm{T}} \boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{z}_1} \end{aligned} \tag{28}$$

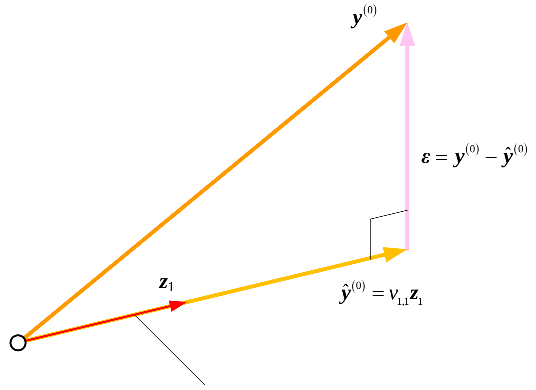

图 26. $\boldsymbol{y}^{(0)}$ 向量投影在主元列向量 $\boldsymbol{z}_1$

第七步

根据 OLS 原理，利用列向量 b_1 和 $\boldsymbol{z}_1$ 估算因变量列向量 $\boldsymbol{y}$，并到拟合 $\hat{\boldsymbol{y}}^{(0)}$：

$$\hat{\boldsymbol{y}}^{(0)} = b_1 \boldsymbol{z}_1 = \frac{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{y}^{(0)} \boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{z}_1} = \frac{\left(\boldsymbol{y}^{(0)} \right)^{\mathrm{T}} \boldsymbol{z}_1 \boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{z}_1} \tag{29}$$

原始因变量列向量 $\boldsymbol{y}^{(0)}$ 和拟合列向量 $\hat{\boldsymbol{y}}^{(0)}$ 之差便是残差向量 $\boldsymbol{\varepsilon}^{(0)}$：

$$\boldsymbol{\varepsilon}^{(0)} = \boldsymbol{y}^{(1)} = \boldsymbol{y}^{(0)} - \hat{\boldsymbol{y}}^{(0)} = \boldsymbol{y}^{(0)} - \frac{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{y}^{(0)} \boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{z}_1} \tag{30}$$

而残差向量 $\boldsymbol{\varepsilon}^{(0)}$ 便是进入迭代循环第二步数据向量 $\boldsymbol{y}^{(1)}$。如图 27 所示，$\hat{\boldsymbol{y}}^{(0)}$ 解释部分 $\boldsymbol{y}^{(0)}$。

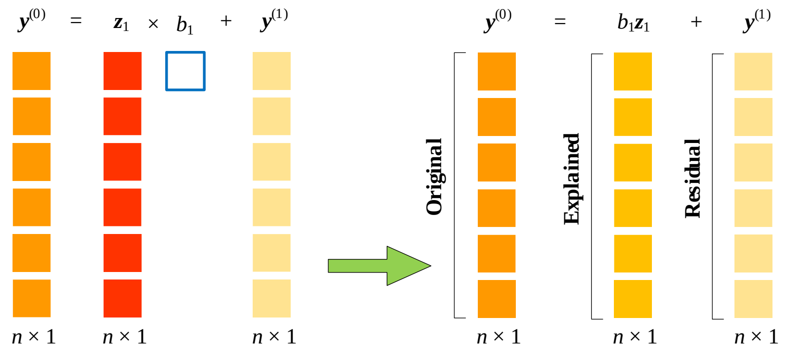

图 27. 估算 $y^{(0)}$

重复迭代

将数据矩阵 $X^{(1)}$ 和数据向量 $y^{(1)}$ 带入如上迭代运算第二步到第七步。

重复第二步得到权重系数列向量 w_2：

$$w_2 = \frac{\left(X^{(1)}\right)^{\mathrm{T}} y^{(1)}}{\left\|\left(X^{(1)}\right)^{\mathrm{T}} y^{(1)}\right\|} \tag{31}$$

重复第三步，利用权重系数列向量 w_2 和 $X^{(1)}$ 构造偏最小二乘回归第二主元向量，z_2：

$$z_2 = X^{(1)} w_2 \tag{32}$$

重复第四步，把自变量数据残差矩阵 $X^{(1)}$ 投影于第二主元列向量 z_2 上，获得系数向量 v_2：

$$v_2 = \begin{bmatrix} v_{1,2} \\ v_{2,2} \\ \vdots \\ v_{q,2} \end{bmatrix} = \frac{\left(X^{(1)}\right)^{\mathrm{T}} z_2}{z_2^{\mathrm{T}} z_2} = \frac{\left(X^{(1)}\right)^{\mathrm{T}} X^{(1)} w_2}{w_2^{\mathrm{T}} \left(X^{(1)}\right)^{\mathrm{T}} X^{(1)} w_2} = \frac{\Sigma^{(1)} w_2}{w_2^{\mathrm{T}} \Sigma^{(1)} w_2} \tag{33}$$

重复第五步，用列向量 v_2 和 z_2 估算，并到拟合矩阵 $\hat{X}^{(1)}$：

$$\hat{X}^{(1)} = z_2 v_2^{\mathrm{T}} = X^{(1)} w_2 v_2^{\mathrm{T}} \tag{34}$$

$X^{(1)}$ 和拟合数据矩阵 $\hat{X}^{(1)}$ 之差便是残差矩阵 $E^{(1)}$，$E^{(1)}$便是再次进入迭代过程第二步数据矩阵 $X^{(2)}$：

$$X^{(2)} = E^{(1)} = X^{(1)} - \hat{X}^{(1)} = X^{(1)} \left(I - w_2 v_2^{\mathrm{T}}\right) \tag{35}$$

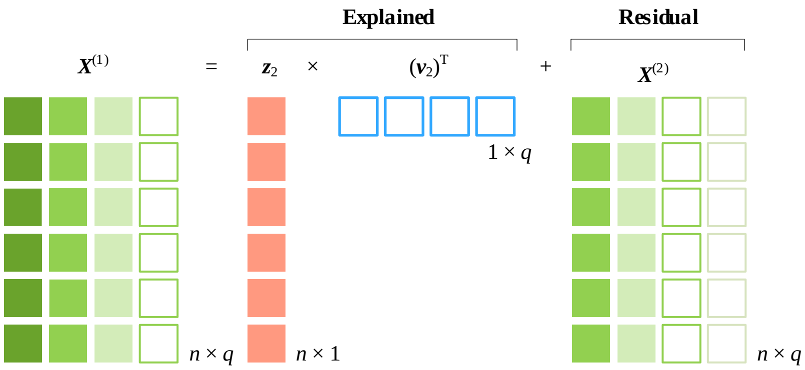

图 28. 计算得到数据矩阵 $\boldsymbol{X}^{(2)}$

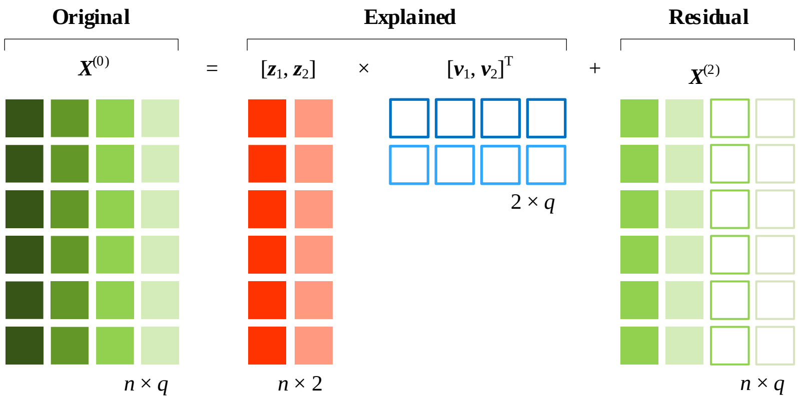

图 29. 前两个主元 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_1$ 还原数据矩阵 $\boldsymbol{X}^{(0)}$

图 25 和图 28 相结合获得图 29，这即前两个主元 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_1$ 还原数据矩阵 $\boldsymbol{X}^{(0)}$。随着主元数量不断增多，偏最小二乘回归更精确地还原原始数据 $\boldsymbol{X}^{(0)}$；即说，对数据 $\boldsymbol{X}^{(0)}$方差解释力度越强。

重复第六步，把因变量数据列向量 $\boldsymbol{y}^{(1)}$ 投影在主元列向量 $\boldsymbol{z}_2$ 上，获得系数 b_2：

$$b_2 = \frac{\boldsymbol{z}_2^T \boldsymbol{y}^{(1)}}{\boldsymbol{z}_2^T \boldsymbol{z}_2} = \frac{\left(\boldsymbol{y}^{(1)}\right)^T \boldsymbol{z}_2}{\boldsymbol{z}_2^T \boldsymbol{z}_2} \tag{36}$$

重复第七步，利用 b_2 和 $\boldsymbol{z}_2$ 得到拟合列向量 $\hat{\boldsymbol{y}}^{(1)}$：

$$\hat{\boldsymbol{y}}^{(1)} = b_2 \boldsymbol{z}_2 \tag{37}$$

列向量 $\boldsymbol{y}^{(1)}$ 和拟合数据列向量 $\hat{\boldsymbol{y}}^{(1)}$ 之差便是残差向量 $\boldsymbol{\varepsilon}^{(1)}$：

$$\boldsymbol{\varepsilon}^{(0)} = \boldsymbol{y}^{(2)} = \boldsymbol{y}^{(1)} - \hat{\boldsymbol{y}}^{(1)} = \boldsymbol{y}^{(1)} - b_2 \boldsymbol{z}_2 \tag{38}$$

而残差向量 $\boldsymbol{\varepsilon}^{(1)}$ 也是进入下一次迭代过程第二步数据向量 $\boldsymbol{y}^{(2)}$。

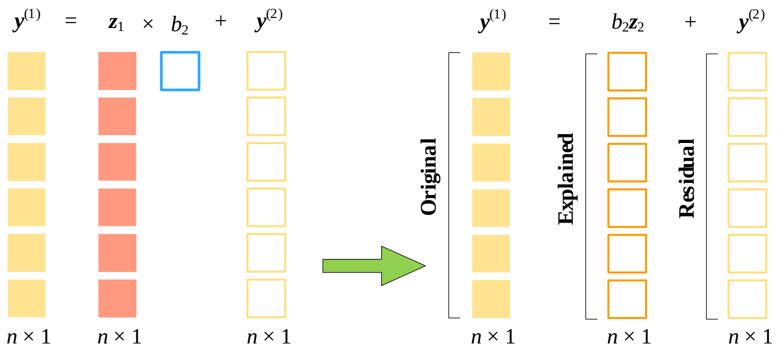

图 30. 估算 $\boldsymbol{y}^{(1)}$

图 31 结合图 27 和图 30，这幅图中前两个主元 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_1$ 还原部分数据列向量 $\boldsymbol{y}^{(0)}$。同理，随着主元数量不断增多，偏最小二乘回归更精确地还原原始因变量列向量 $\boldsymbol{y}^{(0)}$；即，对 $\boldsymbol{y}^{(0)}$ 方差解释力度越强。截止目前，迭代循环已经完成两次。

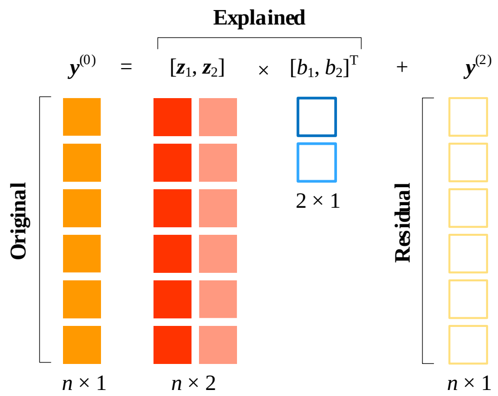

图 31. 前两个主元 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_1$ 还原部分数据列向量 $\boldsymbol{y}^{(0)}$

Scikit-learn 中 PLS 回归的函数为 sklearn.cross_decomposition.PLSRegression()。

下例展示如何使用偏最小二乘回归。这个例子还比较了本书最后一章要介绍的典型相关分析。请大家自行阅读学习：

https://scikit-learn.org/stable/auto_examples/cross_decomposition/plot_compare_cross_decomposition.html