17 主成分分析与回归

PCA and Regressions

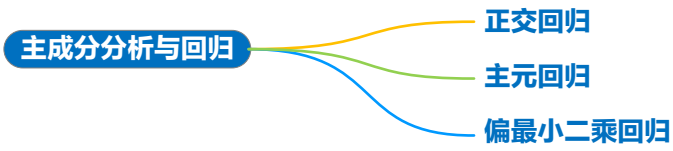
正交回归、主元回归、偏最小二乘回归

数学展现出秩序、对称和有限——这些都是美的极致形态。

The mathematical sciences particularly exhibit order, symmetry, and limitations; and these are the greatest forms of the beautiful.

—— 亚里士多德 (Aristotle) | 古希腊哲学家 | 384 ~ 322 BC

- `numpy.linalg.eig()` 特征值分解
- `numpy.linalg.svd()` 奇异值分解
- `numpy.mean()` 计算均值
- `numpy.std()` 计算均方差
- `numpy.var()` 计算方差
- `pandas_datareader.get_data_yahoo()` 下载股价数据
- `scipy.odr` 正交回归
- `scipy.odr.Model()` 构造正交回归模型
- `scipy.odr.ODR()` 设置正交回归数据、模型和初始自
- `scipy.odr.RealData()` 加载正交回归数据
- `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
- `seaborn.lineplot()` 绘制线图
- `seaborn.relplot()` 绘制散点图和曲线图
- `sklearn.decomposition.PCA()` 主成分分析函数
- `statsmodels.api.add_constant()` 线性回归增加一列常数 1
- `statsmodels.api.OLS` 最小二乘法线性回归



17.1 正交回归

本章将介绍三种与主成分分析有着千丝万缕联系的回归方法——正交回归、主元回归、偏最小二乘回归。让我们首先了解**正交回归** (orthogonal regression)。

正交回归，也叫做**正交距离回归** (Orthogonal Distance Regression, ODR)，又叫**全线性回归** (total linear regression)。正交回归通过将自变量通过主成分分析转换成互相正交的新变量，来消除自变量之间的多重共线性问题，从而提高回归分析的准确性和稳定性。

具体来说，正交回归通过以下步骤实现：1) 对自变量进行主成分分析，得到主成分变量，使它们互相正交。2) 对因变量和主成分变量进行回归分析，得到每个主成分变量的回归系数。3) 根据主成分变量的回归系数和主成分分析的结果，计算出每个自变量的回归系数和截距项。

正交回归的优点之一是消除自变量之间的多重共线性，提高回归分析的准确性和稳定性。正交回归可以在保证预测准确性的前提下，降低自变量的维度，提高回归模型的可解释性。

正交回归的缺点是计算复杂度较高，需要进行主成分分析和回归分析等多个步骤。此外，由于正交回归是基于主成分分析的，因此它可能会失去一些原始自变量的信息，因此需要在可接受的误差范围内进行权衡。

举个例子，平面上，最小二乘法线性回归 OLS 仅考虑纵坐标方向上误差，如图 1 (a) 所示；而正交回归 TLS 同时考虑横纵两个方向误差，如图 1 (b) 所示。

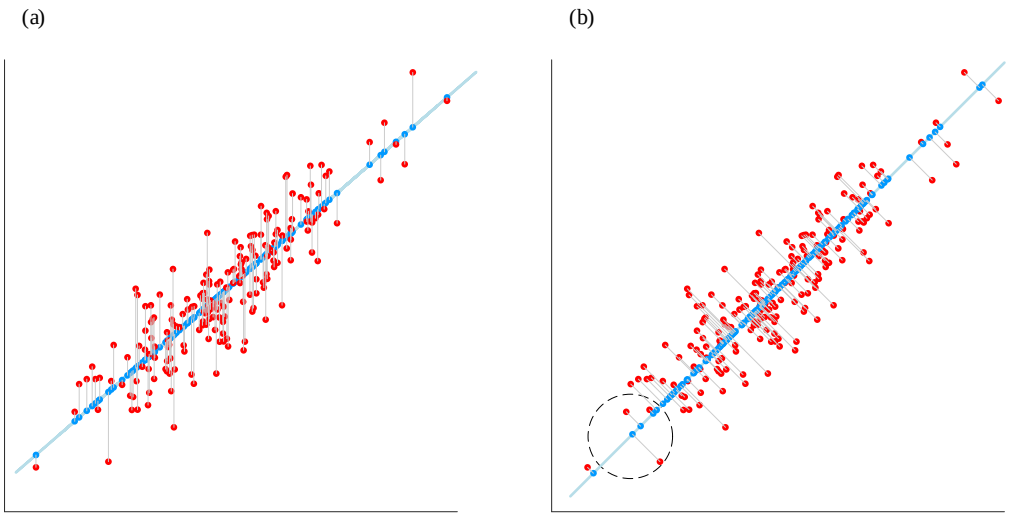


图 1. 对比 OLS 和 TLS 线性回归

从主成分分析角度，正交回归特点是输入数据 $\boldsymbol{X}$ 和输出数据 $\boldsymbol{y}$ 都参与主成分分析。按照特征值从大到小顺序排列特征向量 $[\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_D, \boldsymbol{v}_{D+1}]$，用其中前 D 个向量 $[\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_D]$ 构造一个全新超平面 H。利用 $\boldsymbol{v}_{D+1}$ 垂直于超平面 H 便可以求解出回归系数。

下面用两特征 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2]$ 数据作例子，聊一下主成分回归的思想。如图 2 所示，$\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 为输入数据，$\boldsymbol{y}$ 为输出数据；通过主成分分析，$\boldsymbol{x}_1$、$\boldsymbol{x}_2$ 和 $\boldsymbol{y}$ 正交化之后得到 $\boldsymbol{v}_1$、$\boldsymbol{v}_2$ 和 $\boldsymbol{v}_3$ (根据特征值从大到小排

列)；$\boldsymbol{v}_1$、$\boldsymbol{v}_2$和$\boldsymbol{v}_3$两两正交。第一主成分$\boldsymbol{v}_1$和第二主成分$\boldsymbol{v}_2$构造平面H。$\boldsymbol{v}_3$垂直于平面H，通过这层关系求解出正交回归系数。

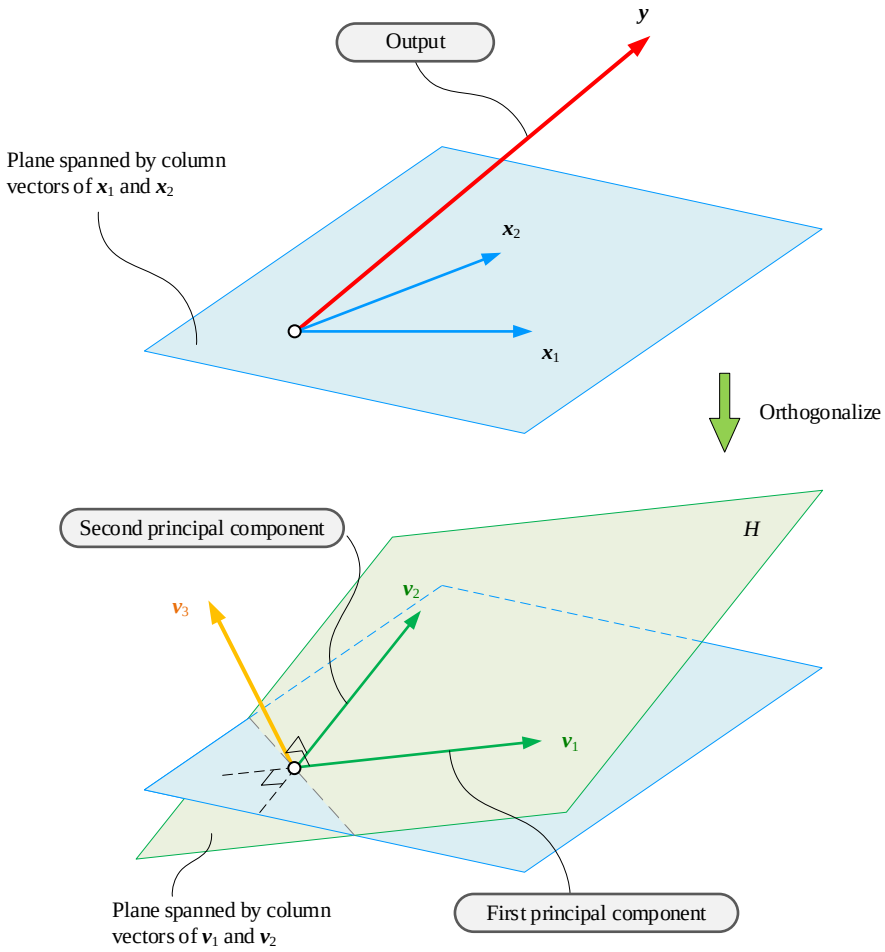


图 2. 通过主成分分析构造正交空间

前文介绍的线性回归采用算法叫做**普通最小二乘法** (Ordinary Least squares, OLS)；而正交回归采用的算法叫做**完全最小二乘法** (Total Least Squares, TLS)。

如图 3 所示，最小二乘回归，将$\boldsymbol{y}$投影到$\boldsymbol{x}_1$和$\boldsymbol{x}_2$构造的平面上。而对于正交回归，将$\boldsymbol{y}$投影到H，得到$\hat{\boldsymbol{y}}$。而残差，$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}$，平行于$\boldsymbol{v}_3$。再次强调，平面$H$是由第一主成分$\boldsymbol{v}_1$和第二主成分$\boldsymbol{v}_2$构造。

此外，建议读者完成本章学习之后，回过头来再比较图 3 和图 4。这样，相信大家会更清楚 OLS 和 TLS 之间的区别。

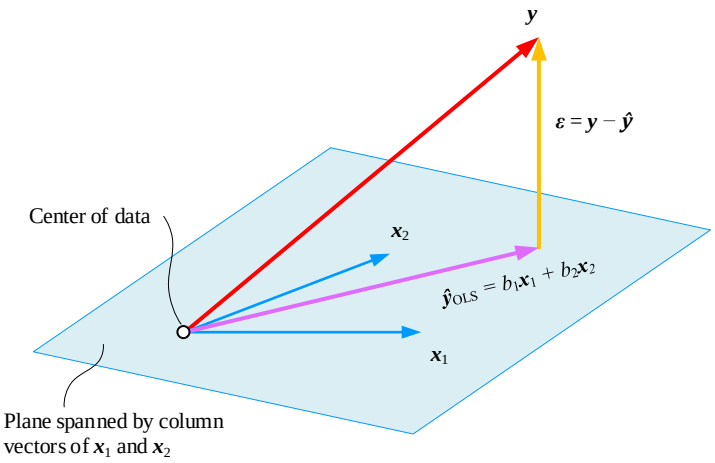


图 3. 最小二乘回归，将 $\boldsymbol{y}$ 投影到 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 构造的平面上

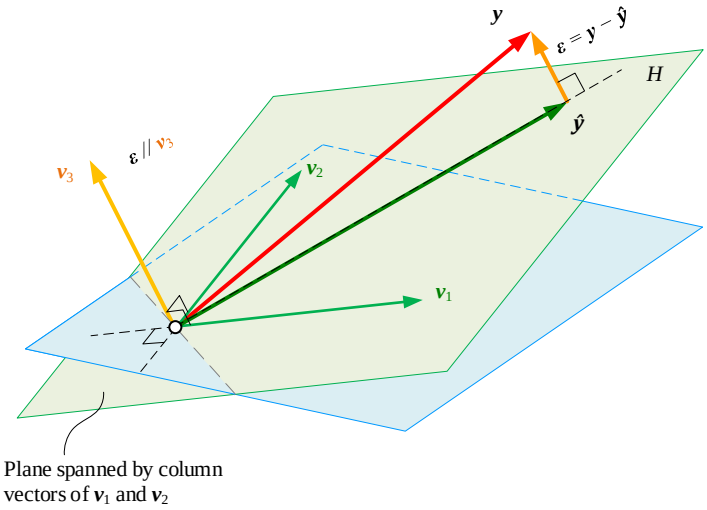


图 4. 正交回归，将输出数据 $\boldsymbol{y}$ 投影到 H

下一节首先用一元正交回归给大家建立正交回归的直观印象，本章后续将逐步扩展到二元回归和多元回归。

17.2 一元正交回归

设定一元正交回归解析式如下：

$$y = b_0 + b_1 x \tag{1}$$

其中，b_0 为截距项，b_1 为斜率。

如图 5 所示，x-y 平面上任意一点 $(x^{(i)}, y^{(i)})$ 和正交回归直线距离可以利用下式获得：

$$d_i = \frac{y^{(i)} - \left(b_0 + b_1 x^{(i)}\right)}{\sqrt{1 + b_1^2}} \tag{2}$$

当 $i = 1 \sim n$ 时，d_i 构成列向量为 $\boldsymbol{d}$：

$$\boldsymbol{d} = \frac{\boldsymbol{y} - \left(b_0 + b_1 \boldsymbol{x}\right)}{\sqrt{1 + b_1^2}} \tag{3}$$

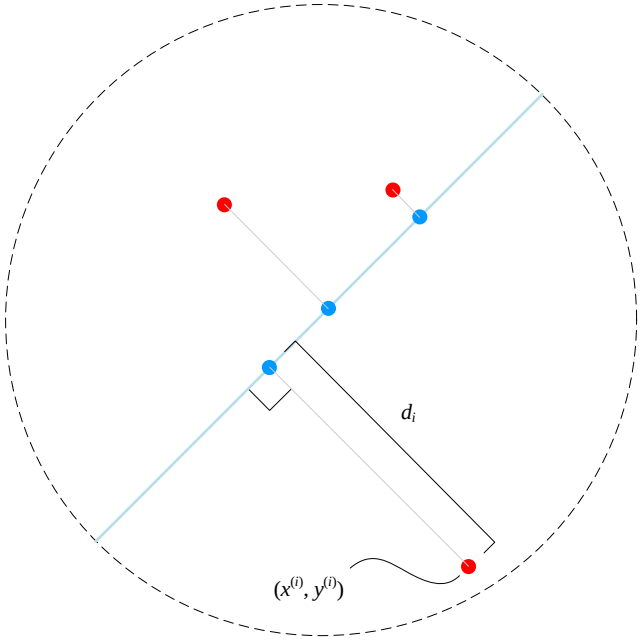


图 5. 正交投影几何关系

构造如下优化问题，b_0 和 b_1 为优化变量，优化目标为最小化欧氏距离平方和：

$$\underset{b0,b1}{\arg\min}\ f\left(b_0, b_1\right) = \left\|\boldsymbol{d}\right\|^2 = \boldsymbol{d}^{\mathrm{T}}\boldsymbol{d} \tag{4}$$

将 (3) 代入 $f(b_0, b_1)$ 得到：

$$f\left(b_0, b_1\right) = \frac{\left(\boldsymbol{y} - \left(b_0 + b_1 \boldsymbol{x}\right)\right)^{\mathrm{T}}\left(\boldsymbol{y} - \left(b_0 + b_1 \boldsymbol{x}\right)\right)}{1 + b_1^2} \tag{5}$$

为了方便计算，也引入全 1 向量 $\boldsymbol{1}$，它和 $\boldsymbol{x}$ 形状一样为 n 行 1 列向量；$f(b_0, b_1)$ 展开整理为下式：

$$f\left(b_0, b_1\right) = \frac{n b_0^2 + 2 b_0 b_1 \boldsymbol{x}^{\mathrm{T}}\boldsymbol{1} + b_1^2 \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} - 2 b_0 \boldsymbol{y}^{\mathrm{T}}\boldsymbol{1} - 2 b_1 \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}{1 + b_1^2} \tag{6}$$

$f(b_0, b_1)$ 对 b_0 偏导为 0，构造如下等式：

$$\frac{\partial f\left(b_0, b_1\right)}{\partial b_0} = \frac{2 n b_0 + 2 b_1 \boldsymbol{x}^{\mathrm{T}}\boldsymbol{1} - 2 \boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}}{1 + b_1^2} = 0 \tag{7}$$

$f(b_0, b_1)$ 对 b_1 偏导为 0，构造如下等式：

$$\frac{\partial f\left(b_0,b_1\right)}{\partial b_1}=\frac{2b_1\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}+2b_0\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}-2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{1+b_1^2}-\frac{\left(nb_0^2+2b_0b_1\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}+b_1^2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}-2b_0\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}-2b_1\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}+\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}\right)2b_1}{\left(1+b_1^2\right)^2}=0 \tag{8}$$

观察 (7)，容易用 b_1 表达 b_0：

$$b_0=\frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}-b_1\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}}{n}=\mathrm{E}\left(\boldsymbol{y}\right)-b_1\mathrm{E}\left(\boldsymbol{x}\right) \tag{9}$$

其中，

$$\begin{cases}\mathrm{E}\left(\boldsymbol{x}\right)=\dfrac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}}{n}=\dfrac{\sum_{i=1}^{n}x^{(i)}}{n}\\ \mathrm{E}\left(\boldsymbol{y}\right)=\dfrac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}}{n}=\dfrac{\sum_{i=1}^{n}y^{(i)}}{n}\end{cases} \tag{10}$$

将 (9) 给出 b_0 解析式代入 (8) 获得仅含有 b_1 的一元二次方程：

$$b_1^2+kb_1-1=0 \tag{11}$$

其中，

$$\begin{aligned}k&=\frac{n\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}-n\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}+\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}}{n\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}}\\&=\frac{\left(\dfrac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}{n}-\dfrac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}}{n^2}\right)-\left(\dfrac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}{n}-\dfrac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}}{n^2}\right)}{\dfrac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{n}-\dfrac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}}{n^2}}\\&=\frac{\mathrm{var}\left(\boldsymbol{x}\right)-\mathrm{var}\left(\boldsymbol{y}\right)}{\mathrm{cov}\left(\boldsymbol{x},\boldsymbol{y}\right)}=\frac{\sigma_x^2-\sigma_y^2}{\rho_{xy}\sigma_x\sigma_y}\end{aligned} \tag{12}$$

上式，不区分求解方差协方差时，$1/(n-1)$和 $1/n$ 之间差别。

求解 (11) 一元二次方程，得到 b_1 解如下：

$$b_1=\frac{-k\pm\sqrt{k^2+4}}{2} \tag{13}$$

将 (12) 给出的 k，代入 (13)，整理得到 b_1 解：

$$b_1=\frac{\left(\sigma_y^2-\sigma_x^2\right)\pm\sqrt{\left(\sigma_x^2-\sigma_y^2\right)^2+4\left(\rho_{xy}\sigma_x\sigma_y\right)^2}}{2\rho_{xy}\sigma_x\sigma_y} \tag{14}$$

发现 b_1 两个解即**主成分分析** (principal component analysis, PCA) 主元方向。

构造 $[\boldsymbol{x},\boldsymbol{y}]$ 数据矩阵，它的协方差矩阵 $\boldsymbol{\Sigma}$ 可以记做：

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_x^2 & \rho_{xy}\sigma_x\sigma_y\\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2\end{bmatrix} \tag{15}$$

对 $\boldsymbol{\Sigma}$ 进行特征值分解，得到两个特征向量：

$$\boldsymbol{v}_1 = \begin{bmatrix} \dfrac{\left(\sigma_y^2 - \sigma_x^2\right) + \sqrt{\left(\sigma_x^2 - \sigma_y^2\right)^2 + 4\left(\rho_{xy}\sigma_x\sigma_y\right)^2}}{2\rho_{xy}\sigma_x\sigma_y} \\ 1 \end{bmatrix}$$
$$\boldsymbol{v}_2 = \begin{bmatrix} \dfrac{\left(\sigma_y^2 - \sigma_x^2\right) - \sqrt{\left(\sigma_x^2 - \sigma_y^2\right)^2 + 4\left(\rho_{xy}\sigma_x\sigma_y\right)^2}}{2\rho_{xy}\sigma_x\sigma_y} \\ 1 \end{bmatrix} \tag{16}$$

$\boldsymbol{\Sigma}$ 两个特征值，从大到小排列：

$$\lambda_1 = \frac{\sigma_x^2 + \sigma_y^2}{2} + \sqrt{\left(\rho_{xy}\sigma_x\sigma_y\right)^2 + \left(\frac{\sigma_x^2 - \sigma_y^2}{2}\right)^2}$$
$$\lambda_2 = \frac{\sigma_x^2 + \sigma_y^2}{2} - \sqrt{\left(\rho_{xy}\sigma_x\sigma_y\right)^2 + \left(\frac{\sigma_x^2 - \sigma_y^2}{2}\right)^2} \tag{17}$$

特征值较大的特征向量为正交回归直线切线向量；特征值较小特征向量对应直线法线向量，这样求得 b_1 斜率。有了上述思路，便可以用 PCA 分解来获得正交回归系数，这是下一节要讲解的内容。

如下代码首先介绍如何利用 scipy.odr 可以求解得到正交回归系数。构造线性函数 linear_func(b, x)，利用 scipy.odr.Model(linear_func) 创建线性模型；然后，采用 scipy.odr.RealData() 加载数据，再用 scipy.odr.ODR() 整合数据、模型和初始值，输出为 odr。odr.run() 求解回归问题。然后，用 pprint()打印结果。

```
Beta: [0.00157414 1.43773257]
Beta Std Error: [0.00112548 0.05617699]
Beta Covariance: [[ 1.21904872e-02 -2.43641786e-02]
 [-2.43641786e-02  3.03712371e+01]]
Residual Variance: 0.00010390932459480641
Inverse Condition #: 0.22899877744275976
Reason(s) for Halting:
Sum of squares convergence
```

一元正交回归的解析式为：

$$y = 1.4377x + 0.00157 \tag{18}$$

下一节将介绍如下采用主成分分析来求解一元正交回归系数，并比较正交回归和最小二乘法线性回归。

17.3 几何角度看正交回归

图 6 所示为正交回归和 PCA 分解关系，发现主元回归直线通过数据中心 (E($\boldsymbol{x}$), E($\boldsymbol{y}$))，回归直线方向和主元方向 $\boldsymbol{v}_1$ 平行，垂直于次元 $\boldsymbol{v}_2$ 方向。即，次元方向 $\boldsymbol{v}_2$ 和直线法向量 $\boldsymbol{n}$ 平行。

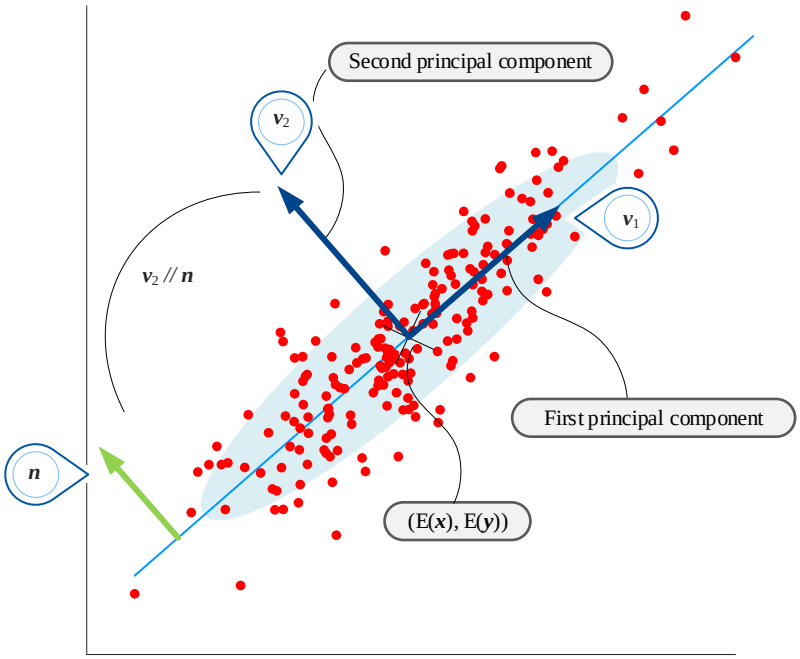


图 6. 正交回归和 PCA 分解关系

对于 (1) 所示一元一次函数，构造二元 $F(x, y)$ 函数如下：

$$F(x,y) = b_0 + b_1 x - y \tag{19}$$

$F(x, y)$ 法向量，即平面上形如 (1) 直线法向量 $\boldsymbol{n}$ 可以通过下式求解：

$$\boldsymbol{n} = \left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}\right)^{\mathrm{T}} = \begin{bmatrix} b_1 \\ -1 \end{bmatrix} \tag{20}$$

如前文所示，$\boldsymbol{n}$ 方向即 PCA 分解第二主元方向，即次元方向。

为了方便计算，假设数据已经经过中心化处理，即已经完成如下运算：

$$\boldsymbol{x} = \boldsymbol{x} - \mathrm{E}(\boldsymbol{x}), \quad \boldsymbol{y} = \boldsymbol{y} - \mathrm{E}(\boldsymbol{y}) \tag{21}$$

由于 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 已经是中心化向量，协方差矩阵可以通过下式运算得到：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{x} & \boldsymbol{y} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{x} & \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}^{\mathrm{T}} \\ \boldsymbol{y}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} & \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} & \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \tag{22}$$

为了方便计算，本节计算协方差矩阵不考虑系数 $1/(n-1)$。

由于 $\boldsymbol{n}$ 为 $\boldsymbol{\Sigma}$ 次元方向：

$$\boldsymbol{\Sigma}\boldsymbol{n} = \lambda_2 \boldsymbol{n} \quad \Rightarrow \quad \begin{bmatrix} \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} & \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \boldsymbol{n} = \lambda_2 \boldsymbol{n} \tag{23}$$

将 (20) 代入 (23)，整理得到如下两个等式：

$$\begin{bmatrix} \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} & \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \begin{bmatrix} b_1 \\ -1 \end{bmatrix} = \lambda_2 \begin{bmatrix} b_1 \\ -1 \end{bmatrix} \quad \Rightarrow \quad \begin{cases} \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} b_1 - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \lambda_2 b_1 \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} b_1 - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} = -\lambda_2 \end{cases} \tag{24}$$

联立 (24) 两个等式，用 λ_2 表示 b_1：

$$b_{1\_TLS} = \left(\boldsymbol{x}^T\boldsymbol{x} - \lambda_2\right)^{-1}\boldsymbol{x}^T\boldsymbol{y} \tag{25}$$

下式为本书前文获得的一元线性回归 OLS 中 b_1 解：

$$b_{1\_OLS} = \left(\boldsymbol{x}^T\boldsymbol{x}\right)^{-1}\boldsymbol{x}^T\boldsymbol{y} \tag{26}$$

对比 OLS 和 TLS；当 (25) 中 λ_2 为 0 时，两种回归方法得到斜率完全一致。$\lambda_2 = 0$ 时，$\boldsymbol{y}$ 和 $\boldsymbol{x}$ 完全线性相关。

数据中心化前后，回归直线梯度向量不变；中心化之前的回归直线通过 (E($\boldsymbol{x}$), E($\boldsymbol{y}$)) 一点，即：

$$\mathrm{E}(\boldsymbol{y}) = b_0 + b_1\mathrm{E}(\boldsymbol{x}) \tag{27}$$

获得回归式截距项 b_0 表达式：

$$b_0 = \mathrm{E}(\boldsymbol{y}) - b_1\mathrm{E}(\boldsymbol{x}) \tag{28}$$

图 7 所示为一元正交回归数据之间关系。发现自变量 $\boldsymbol{x}$ 列向量和因变量 $\boldsymbol{y}$ 列向量数据都参与 PCA 分解得到正交化向量 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$，然后用特征值中较大值对应特征向量 $\boldsymbol{v}_1$ 作为一元正交回归直线切线向量。更为简单计算方法是，用特征值较小值对应特征向量 $\boldsymbol{v}_2$ 作为一元正交回归直线法向量。

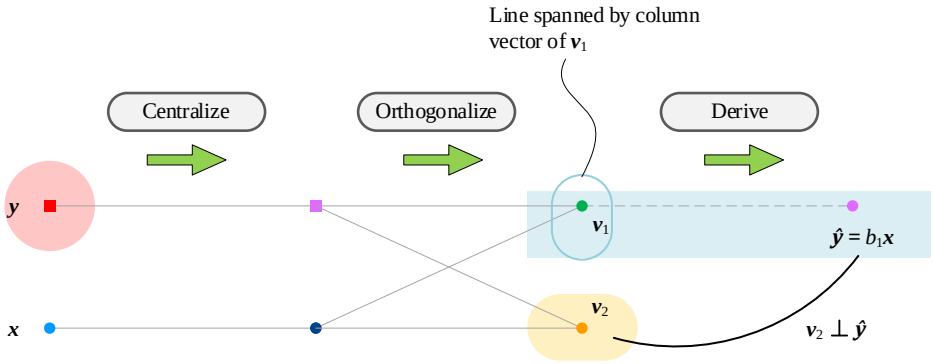


图 7. 一元正交回归 TLS 数据关系

图 8 所示为最小二乘法 OLS 一元线性回归系数，对应的一元 OLS 解析式为：

$$y = 1.1225x + 0.0018 \tag{29}$$

图 9 比较 OLS 和 TLS 结果。

OLS Regression Results

Dep. Variable:	AAPL	R-squared:	0.687
Model:	OLS	Adj. R-squared:	0.686
Method:	Least Squares	F-statistic:	549.7
Date:	Thu, 07 Oct 2021	Prob (F-statistic):	4.55e-65
Time:	07:08:46	Log-Likelihood:	678.03
No. Observations:	252	AIC:	-1352.
Df Residuals:	250	BIC:	-1345.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0018	0.001	1.759	0.080	-0.000	0.004
SP500	1.1225	0.048	23.446	0.000	1.028	1.217

Omnibus:	52.424	Durbin-Watson:	1.864
Prob(Omnibus):	0.000	Jarque-Bera (JB):	210.804
Skew:	0.777	Prob(JB):	1.68e-46
Kurtosis:	7.203	Cond. No.	46.1

图 8. 最小二乘法 OLS 一元线性回归结果

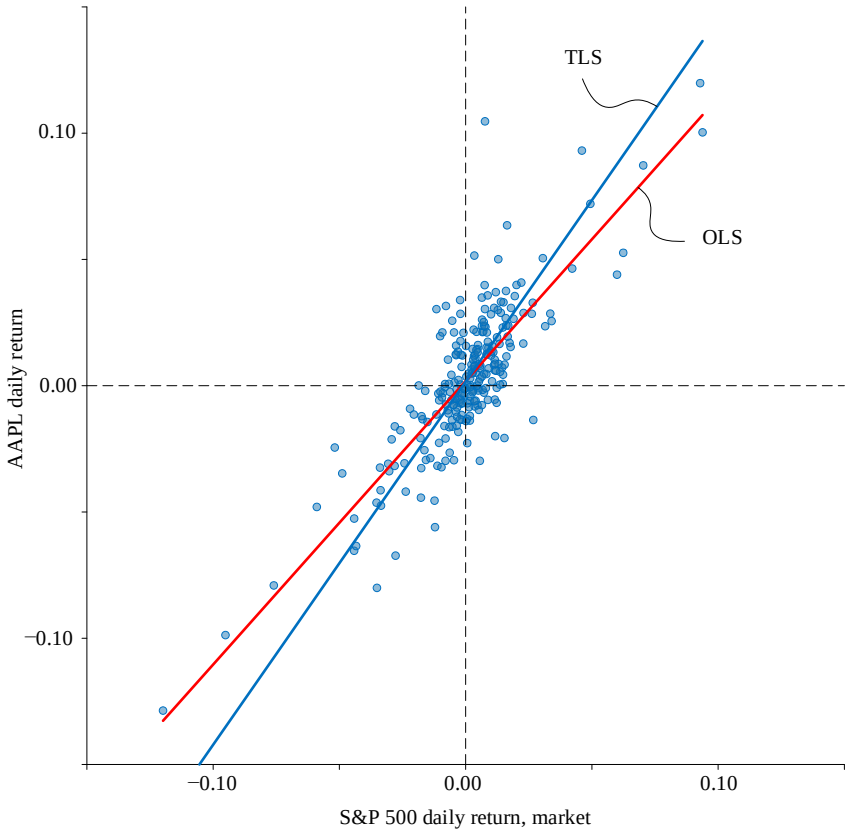


图 9. 比较 OLS 和 TLS 结果

Bk7_Ch17_01.ipynb 绘制本节图像。

17.4 二元正交回归

这一节用主成分分析讨论二元正交回归。

首先也是对数据进行中心化处理：

$$\boldsymbol{x}_1 = \boldsymbol{x}_1 - \mathrm{E}(\boldsymbol{x}_1), \quad \boldsymbol{x}_2 = \boldsymbol{x}_2 - \mathrm{E}(\boldsymbol{x}_2), \quad \boldsymbol{y} = \boldsymbol{y} - \mathrm{E}(\boldsymbol{y}) \tag{30}$$

根据 PCA 计算法则，首先求解协方差矩阵。由于 $\boldsymbol{x}_1$、$\boldsymbol{x}_2$ 和 $\boldsymbol{y}$ 已经为中心化矩阵，因此协方差矩阵 $\boldsymbol{\Sigma}$ 通过下式计算获得。

$$\begin{aligned}\boldsymbol{\Sigma} &= \begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{y}\end{bmatrix}^{\mathrm{T}} \begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{y}\end{bmatrix} \\ &= \begin{bmatrix}\boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \boldsymbol{y}^{\mathrm{T}}\end{bmatrix} \begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{y}\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2 & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_2 & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_2 & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}\end{bmatrix}\end{aligned} \tag{31}$$

为了方便计算，本节也计算不考虑系数 $1/(n-1)$。

正交回归解析式表达：

$$y = b_0 + b_1 x_1 + b_2 x_2 \tag{32}$$

构造二元 $F(x_1, x_2, y)$ 函数如下：

$$F(x_1, x_2, y) = b_0 + b_1 x_1 + b_2 x_2 - y \tag{33}$$

$F(x_1, x_2, y)$ 法向量即平面 $f(x_1, x_2)$ 法向量 $\boldsymbol{n}$ 通过下式求解：

$$\boldsymbol{n} = \left(\frac{\partial F}{\partial x_1}, \frac{\partial F}{\partial x_2}, \frac{\partial F}{\partial y}\right)^{\mathrm{T}} = \begin{bmatrix} b_1 & b_2 & -1\end{bmatrix}^{\mathrm{T}} \tag{34}$$

$\boldsymbol{n}$ 平行于 $\boldsymbol{\Sigma}$ 矩阵 PCA 分解特征值最小特征向量，即：

$$\boldsymbol{\Sigma}\boldsymbol{v}_3 = \lambda_3 \boldsymbol{v}_3 \quad \Rightarrow \quad \begin{bmatrix}\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2 & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_2 & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_2 & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}\end{bmatrix} \boldsymbol{n} = \lambda_3 \boldsymbol{n} \tag{35}$$

整理得到：

$$\begin{bmatrix}\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2 & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_2 & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_2 & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}\end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ -1\end{bmatrix} = \lambda_3 \begin{bmatrix} b_1 \\ b_2 \\ -1\end{bmatrix} \quad \Rightarrow \quad \begin{cases}\left(\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1 - \lambda_3\right) b_1 + \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2 b_2 = \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_1 b_1 + \left(\boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_2 - \lambda_3\right) b_2 = \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{y}\end{cases} \tag{36}$$

$\boldsymbol{n}$ 平行于 $\boldsymbol{\Sigma}$ 矩阵 PCA 分解特征值最小特征向量 $\boldsymbol{v}_3$，构造如下等式并求解 b_1 和 b_2：

$$\begin{bmatrix} b_1 \\ b_2 \\ -1 \end{bmatrix} = k\boldsymbol{v}_3 \quad \Rightarrow \quad \begin{bmatrix} b_1 \\ b_2 \\ -1 \end{bmatrix} = k \begin{bmatrix} v_{1,3} \\ v_{2,3} \\ v_{3,3} \end{bmatrix} \tag{37}$$

根据 (37) 最后一行，可以求得 k

$$k = \frac{-1}{v_{3,3}} \tag{38}$$

b_1 和 b_2 构成的列向量为：

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \frac{-1}{v_{3,3}} \begin{bmatrix} v_{1,3} \\ v_{2,3} \end{bmatrix} \tag{39}$$

回归方程常数项通过下式获得：

$$b_0 = \mathrm{E}(\boldsymbol{y}) - \begin{bmatrix} \mathrm{E}(\boldsymbol{x}_1) & \mathrm{E}(\boldsymbol{x}_2) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{40}$$

为了方便多元正交回归运算，令：

$$\begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{y} \end{bmatrix} \tag{41}$$

协方差矩阵 $\boldsymbol{\Sigma}$ 为：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \tag{42}$$

上式 $\boldsymbol{\Sigma}$ 也不考虑系数 $1/(n-1)$：

$$\boldsymbol{\Sigma}\boldsymbol{v}_3 = \lambda_3\boldsymbol{v}_3 \quad \Rightarrow \quad \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \boldsymbol{n} = \lambda_3\boldsymbol{n} \tag{43}$$

构造 $\boldsymbol{b} = [b_1, b_2]^{\mathrm{T}}$ 这样重新构造特征值和特征向量以及 $\boldsymbol{\Sigma}$ 之间关系：

$$\boldsymbol{n} = \begin{bmatrix} b_1 \\ b_2 \\ -1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ -1 \end{bmatrix} \tag{44}$$

将 (44) 代入 (43)，整理得到 $\boldsymbol{b}$：

$$\begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ -1 \end{bmatrix} = \lambda_3 \begin{bmatrix} \boldsymbol{b} \\ -1 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{b} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \lambda_3\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{45}$$

下一节将使用 (45) 这一解析式计算正交回归解析式系数。

图 10 回顾本章第一节介绍的二元正交回归坐标转换过程。

数据 $[\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}]$ 中心化后，用 PCA 正交化获得正交系 $[\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]$。$\boldsymbol{v}_1$, $\boldsymbol{v}_2$ 和 $\boldsymbol{v}_3$ 对应特征值由大到小。前两个主元向量 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 相互垂直，构成了一个平面 H，特征值最小主元 $\boldsymbol{v}_3$ 垂直于该平面。$\boldsymbol{n}$ 为 H 平面法向量，$\boldsymbol{n}$ 和 $\boldsymbol{v}_3$ 两者平行。

图 10 还比较了 OLS 和 TLS 回归结果。值得大家注意的是，如图 10 上半部分所示，对于最小二乘回归 OLS，$\hat{\boldsymbol{y}}$ 在 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 构造的平面上；而如图 10 下半部分，正交回归 TLS 中，$\hat{\boldsymbol{y}}$ 在 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 构造平面 H 上。

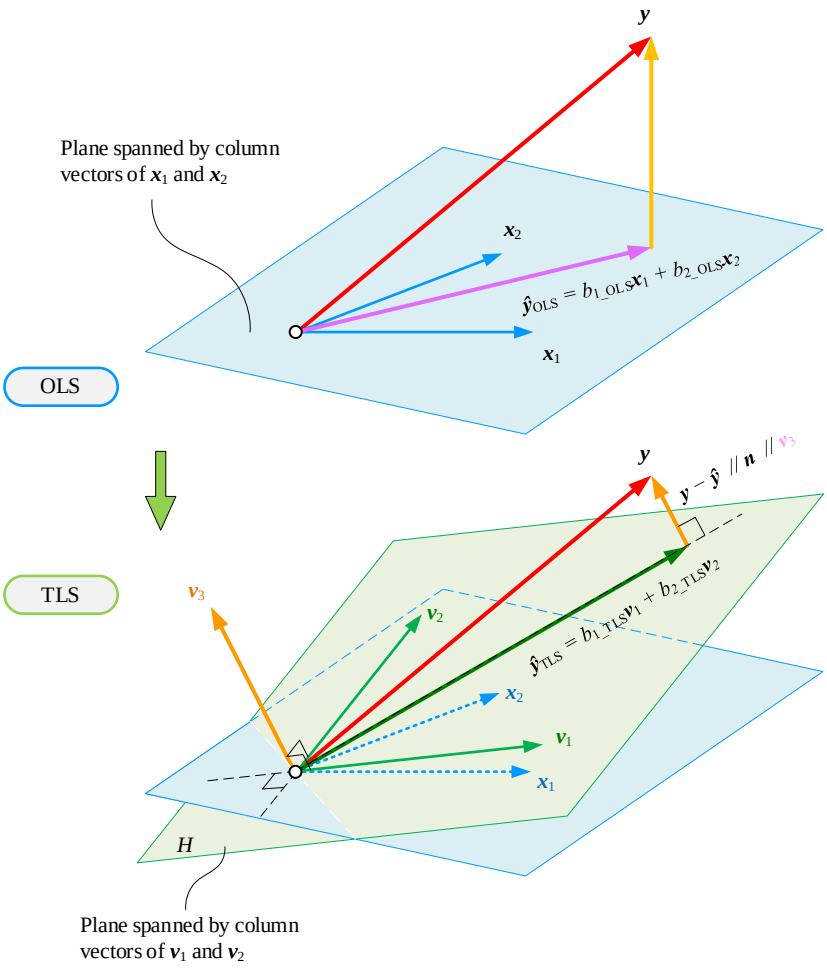


图 10. 几何角度解释二元正交回归坐标转换

图 11 解释二元正交回归数据关系。如前文反复强调，输入数据和输出数据都参与主成分分析，也就是正交化过程，因此特征向量既有“输入”成分，也有“输出”成分，呈现“你中有我，我中有你”。

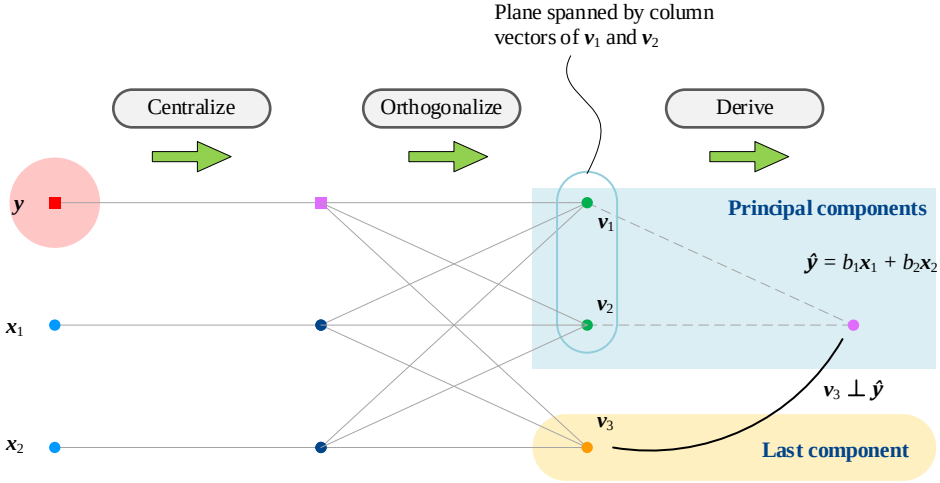


图 11. 二元正交回归数据关系

利用上一节介绍的 scipy.odr，可以求解一个二元正交回归的结果如下。利用主成分分析，我们可以获得相同正交回归的系数。

```
Beta: [-0.00061177  0.40795725  0.44382723]
Beta Std Error: [0.00057372 0.02454606 0.02864744]
Beta Covariance: [[ 5.46486647e-03 -2.24817813e-02  1.00466594e-02]
 [-2.24817813e-02  1.00032390e+01 -7.07446738e+00]
 [ 1.00466594e-02 -7.07446738e+00  1.36253753e+01]]
Residual Variance: 6.02314210079386e-05
Inverse Condition #: 0.16900716799896934
Reason(s) for Halting:
Sum of squares convergence
```

二元正交回归的平面解析式为：

$$y = 0.4079x_1 + 0.4438x_2 - 0.00061 \tag{46}$$

图 12 所示为最小二乘法 OLS 二元线性回归结果，对应的平面解析式如下：

$$y = 0.3977x_1 + 0.4096x_2 - 0.006 \tag{47}$$

OLS Regression Results

Dep. Variable:	SP500	R-squared:	0.830
Model:	OLS	Adj. R-squared:	0.829
Method:	Least Squares	F-statistic:	607.4
Date:	Thu, 07 Oct 2021	Prob (F-statistic):	1.69e-96
Time:	07:31:57	Log-Likelihood:	831.06
No. Observations:	252	AIC:	-1656.
Df Residuals:	249	BIC:	-1646.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0006	0.001	-0.984	0.326	-0.002	0.001
AAPL	0.3977	0.024	16.326	0.000	0.350	0.446
MCD	0.4096	0.028	14.442	0.000	0.354	0.465

Omnibus:	37.744	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	157.710
Skew:	0.492	Prob(JB):	5.67e-35
Kurtosis:	6.749	Cond. No.	59.4

图 12. 最小二乘法 OLS 二元线性回归结果

图 13 比较 OLS 和 TLS 二元回归结果。

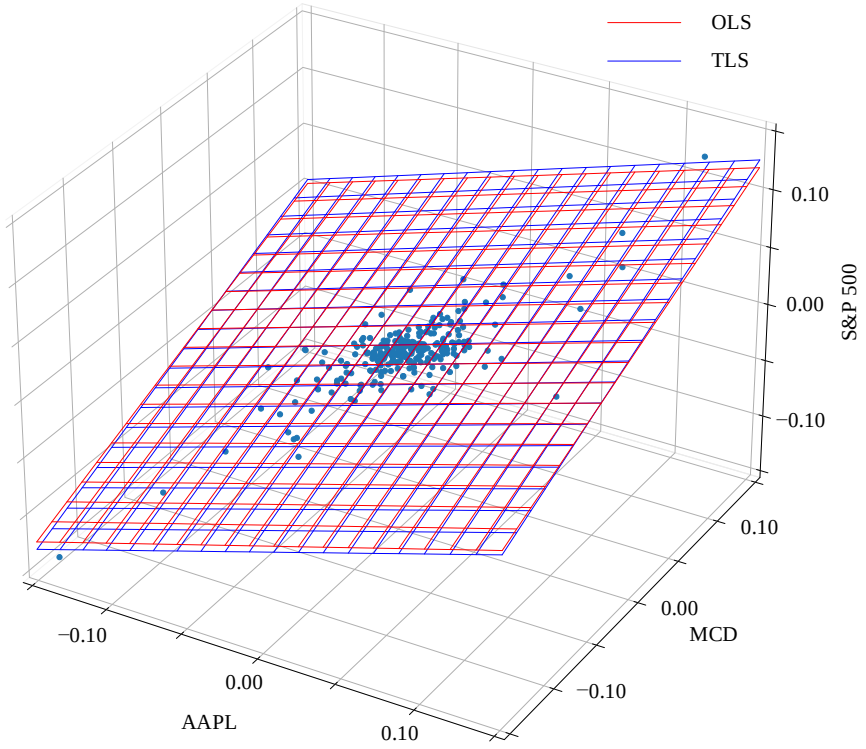


图 13. 比较 OLS 和 TLS 二元回归结果

Bk7_Ch17_02.ipynb 完成本节回归运算。

17.5 多元正交回归

下面，把上述思路推广到 D 维度 $\boldsymbol{X}$ 矩阵。首先中心化数据，获得如下两个中心化 $\boldsymbol{X}, \boldsymbol{y}$ 向量：

$$\boldsymbol{X}_{n\times D} = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^{T}\right)\boldsymbol{X}, \quad \boldsymbol{y} = \boldsymbol{y} - \mathrm{E}(\boldsymbol{y}) \tag{48}$$

为了表达方便，假设 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 已经为中心化数据；这样，构造回归方程式时，不必考虑常数项 b_0，即回归方程中没有截距项：

$$y = b_1 x_1 + b_2 x_2 + \cdots + b_{D-1} x_{D-1} + b_D x_D \tag{49}$$

为了进行 PCA 分解，首先计算 $[\boldsymbol{X}, \boldsymbol{y}]$ 矩阵协方差矩阵。

$\boldsymbol{X}$ 和 $\boldsymbol{y}$ 均是中心化数据，不考虑系数 $1/(n-1)$，协方差矩阵通过下式简单运算获得：

$$\boldsymbol{\Sigma}_{(D+1)\times(D+1)} = [\boldsymbol{X}, \boldsymbol{y}]^{\mathrm{T}}[\boldsymbol{X}, \boldsymbol{y}] = \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}} \\ \boldsymbol{y}^{\mathrm{T}} \end{bmatrix}[\boldsymbol{X}, \boldsymbol{y}] = \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \tag{50}$$

上述协方差矩阵行列宽度均为 $D+1$。对它进行特征值分解得到：

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1} \tag{51}$$

其中，

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \ddots & & \\ & & & \lambda_D & \\ & & & & \lambda_{D+1} \end{bmatrix}, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \geq \lambda_{D+1} \tag{52}$$

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_D & \boldsymbol{v}_{D+1} \end{bmatrix}$$

特征值矩阵对角线特征值从左到右，由大到小。有了本章之前内容铺垫，相信读者已经清楚正交回归的矩阵运算过程，具体如图 14 所示。

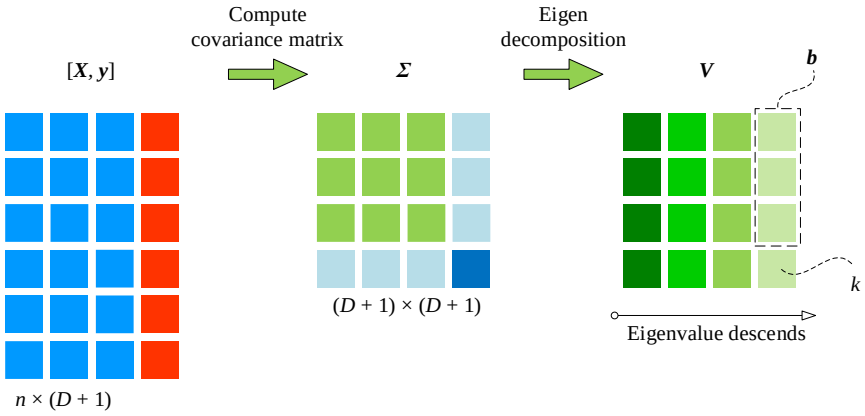


图 14. 多元正交回归矩阵运算过程

$\boldsymbol{V}$ 中第 1 到第 D 个列向量 $[\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_D]$ 构造超平面 H，而 $\boldsymbol{v}_{D+1}$ 垂直于该超平面。

构造 $F(x_1, x_2, ..., x_D, y)$ 函数：

$$F\left(x_1, x_2, ..., x_D, y\right) = b_1 x_1 + b_2 x_2 + \cdots + b_{D-1} x_{D-1} + b_D x_D - y \tag{53}$$

$F(x_1, x_2, ..., x_D, y)$ 法向量即平面上 $f(x_1, x_2, ..., x_D)$ 法向量 $\boldsymbol{n}$ 通过下式求解：

$$\boldsymbol{n} = \left(\frac{\partial F}{\partial x_1}, ..., \frac{\partial F}{\partial x_D}, \frac{\partial F}{\partial y} \right)^{\mathrm{T}} = \begin{bmatrix} b_1 & b_2 & \cdots & b_D & -1 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{b} \\ -1 \end{bmatrix} \tag{54}$$

这样重新构造特征值 λ_{D+1} 和特征向量 $\boldsymbol{v}_{D+1}$ 以及 $\boldsymbol{\Sigma}$ 之间关系。注意，$\boldsymbol{n}$ 平行 $\boldsymbol{v}_{D+1}$。$\boldsymbol{n}$ 对应 $\boldsymbol{\Sigma}$ 矩阵 PCA 分解特征值最小特征向量，即：

$$\boldsymbol{\Sigma}\boldsymbol{v}_{D+1} = \lambda_{D+1}\boldsymbol{v}_{D+1} \quad \Rightarrow \quad \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \boldsymbol{n} = \lambda_{D+1}\boldsymbol{n} \tag{55}$$

求解获得多元正交回归系数列向量 $\boldsymbol{b}$ 解：

$$\begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ -1 \end{bmatrix} = \lambda_{D+1} \begin{bmatrix} \boldsymbol{b} \\ -1 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{b}_{\mathrm{TLS}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \lambda_{D+1}\boldsymbol{I} \right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{56}$$

对比多元线性最小二乘系数向量结果：

$$\boldsymbol{b}_{\text{OLS}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{57}$$

发现当 λ_{D+1} 等于 0 时，$\boldsymbol{y}$ 完全被 $\boldsymbol{X}$ 列向量解释，即两个共线性。

这里我们再次区分一下最小二乘法和正交回归。最小二乘法寻找因变量和自变量之间残差平方和最小超平面；几何角度上讲，将因变量投影在自变量构成超平面 H，使得残差向量垂直 H。正交回归则通过正交化自变量和因变量，构造一个新正交空间；这个新正交空间基底向量为分解得到主元向量，具体如图 15 所示。

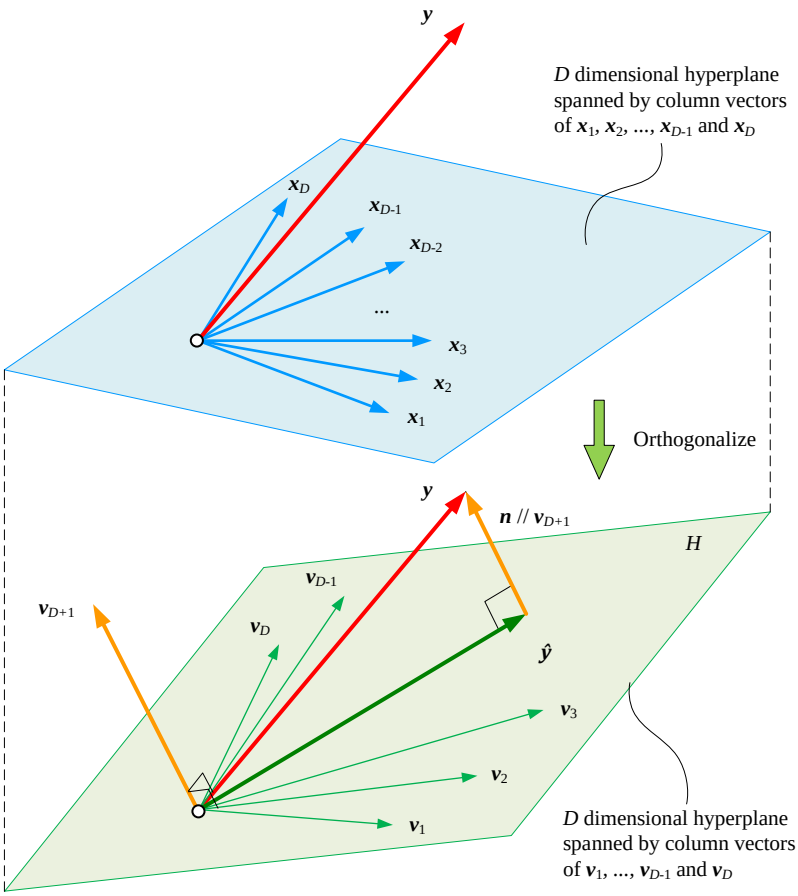


图 15. 几何角度解释多元正交回归

$\boldsymbol{n}$ 平行于数据 $[\boldsymbol{X}, \boldsymbol{y}]$ PCA 分解特征值最小特征向量 $\boldsymbol{v}_{D+1}$，构造如下等式并求解 $b_1, \ldots, b_D$：

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_D \\ -1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ -1 \end{bmatrix} = k\boldsymbol{v}_{D+1} \quad \Rightarrow \quad \begin{bmatrix} \boldsymbol{b} \\ -1 \end{bmatrix} = k \begin{bmatrix} v_{1,D+1} \\ v_{2,D+1} \\ \vdots \\ v_{D,D+1} \\ v_{D+1,D+1} \end{bmatrix} \tag{58}$$

求解 k 得到：

$$k = \frac{-1}{v_{D+1,D+1}} \tag{59}$$

求解 $\boldsymbol{b}$ 得到：

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_D \end{bmatrix} = \frac{-1}{v_{D+1,D+1}} \begin{bmatrix} v_{1,D+1} \\ v_{2,D+1} \\ \vdots \\ v_{D,D+1} \end{bmatrix} \tag{60}$$

b_0 通过下式求得。

$$b_0 = \mathrm{E}(\boldsymbol{y}) - \begin{bmatrix} \mathrm{E}(\boldsymbol{x}_1) & \mathrm{E}(\boldsymbol{x}_2) & \cdots & \mathrm{E}(\boldsymbol{x}_D) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_D \end{bmatrix} \tag{61}$$

图 16 展示多元正交回归运算数据关系。看到数据 $[\boldsymbol{X}, \boldsymbol{y}]$ 均参与到了正交化中；正交化结果为 $D + 1$ 个正交向量 $[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_D, \boldsymbol{v}_{D+1}]$。通过向量 $\boldsymbol{v}_{D+1}$ 垂直 $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_D$ 构成超平面，推导出多元正交回归解析式。

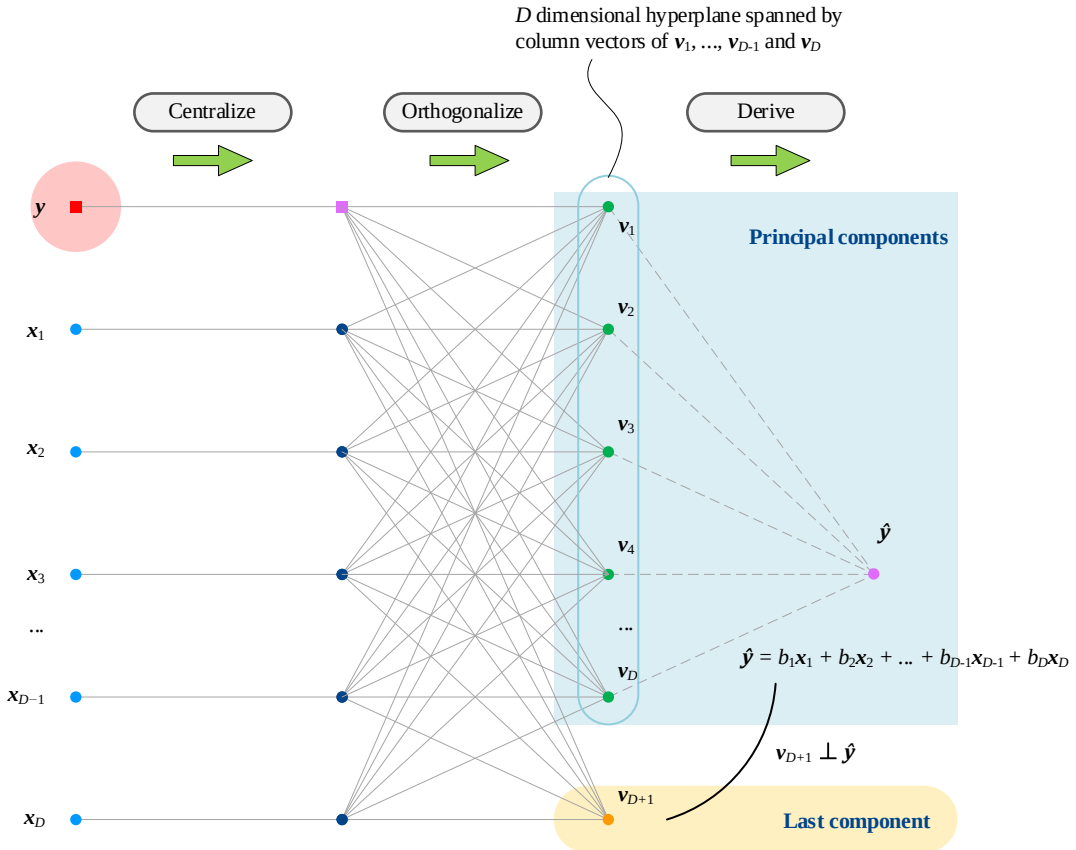


图 16. 多元正交回归运算数据关系

图 17 所示直方图，比较多元 TLS 回归和多元 OLS 回归系数。

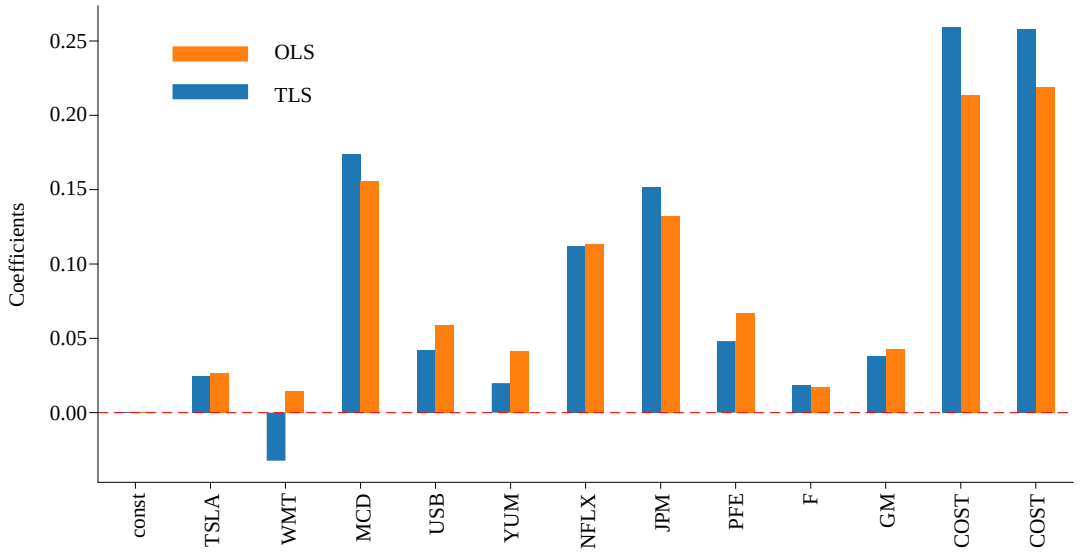


图 17. 比较多元 TLS 回归和多元 OLS 回归系数

Bk7_Ch17_03.ipynb 完成本节回归运算。

17.6 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，按照特征值从大到小排列特征向量，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 18 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。

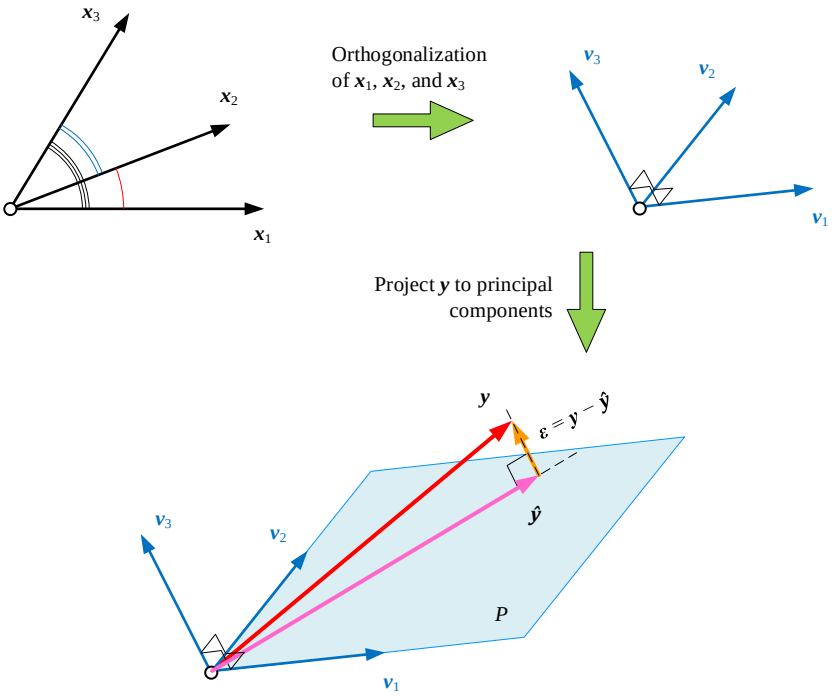


图 18. 主元回归原理

原始数据

图 19 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 20 所示为数据 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 的热图。

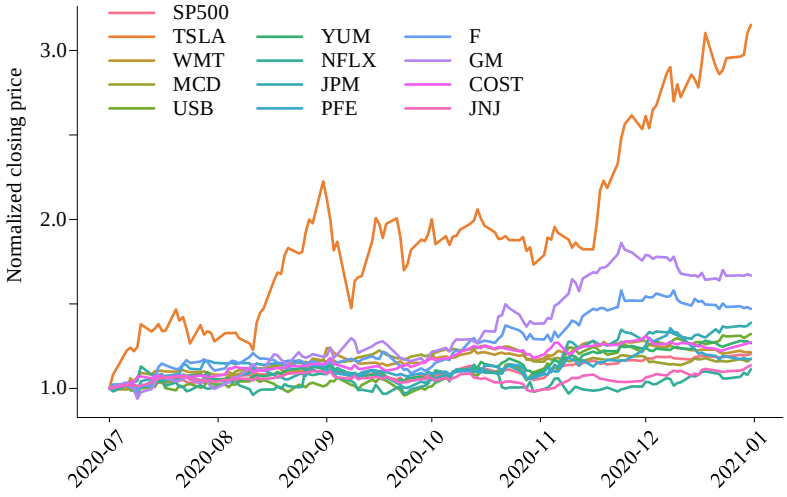


图 19. 股价走势，归一化数据

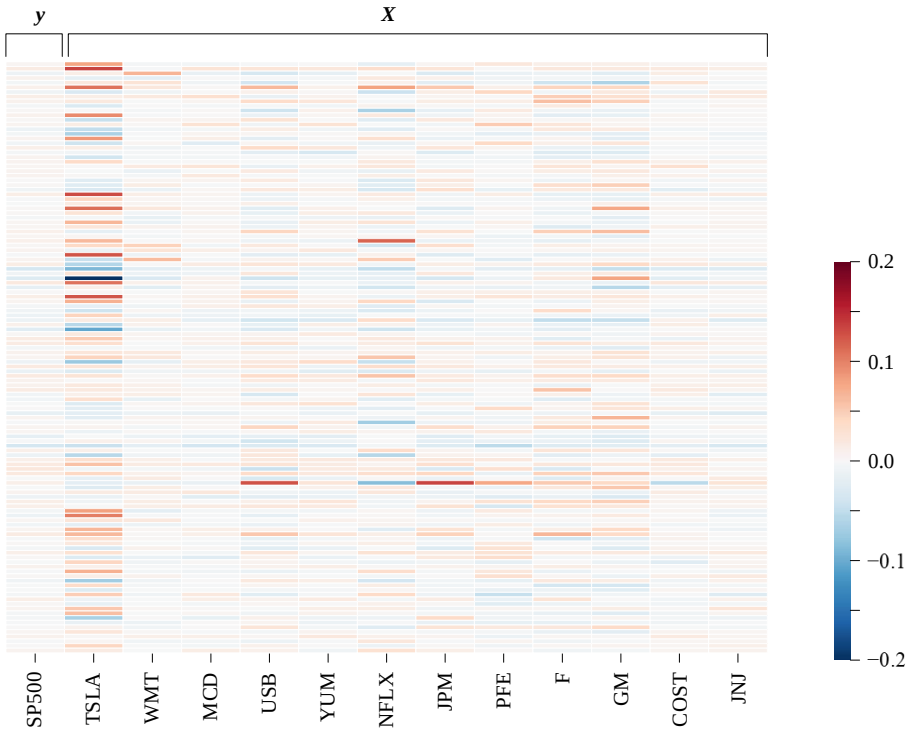


图 20. 数据 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 的热图

图 21 几个分图给出的是数据 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 的 KDE 分布。

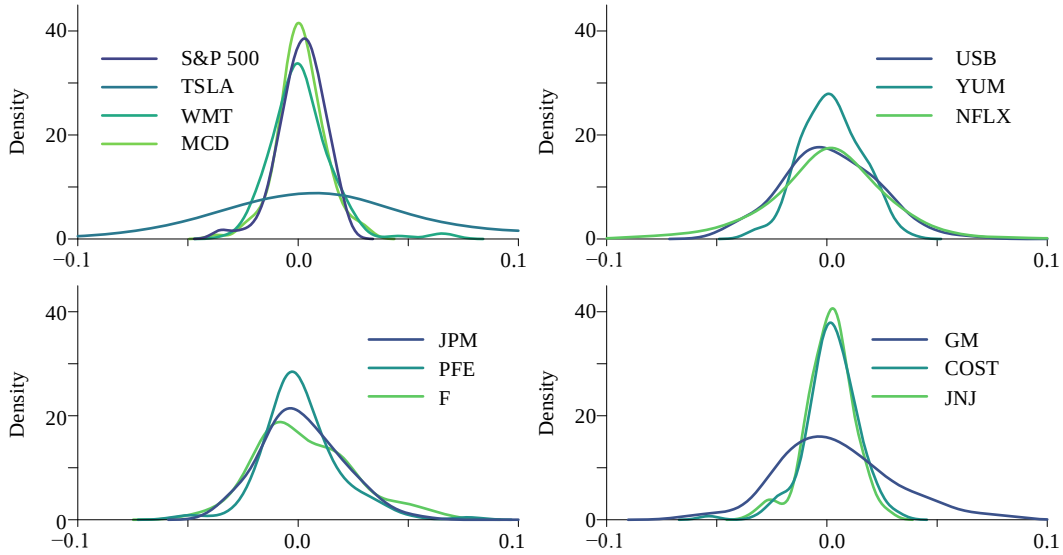


图 21. 数据 $\boldsymbol{X}$ 和 $\boldsymbol{y}$ 的 KDE 分布

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 22 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

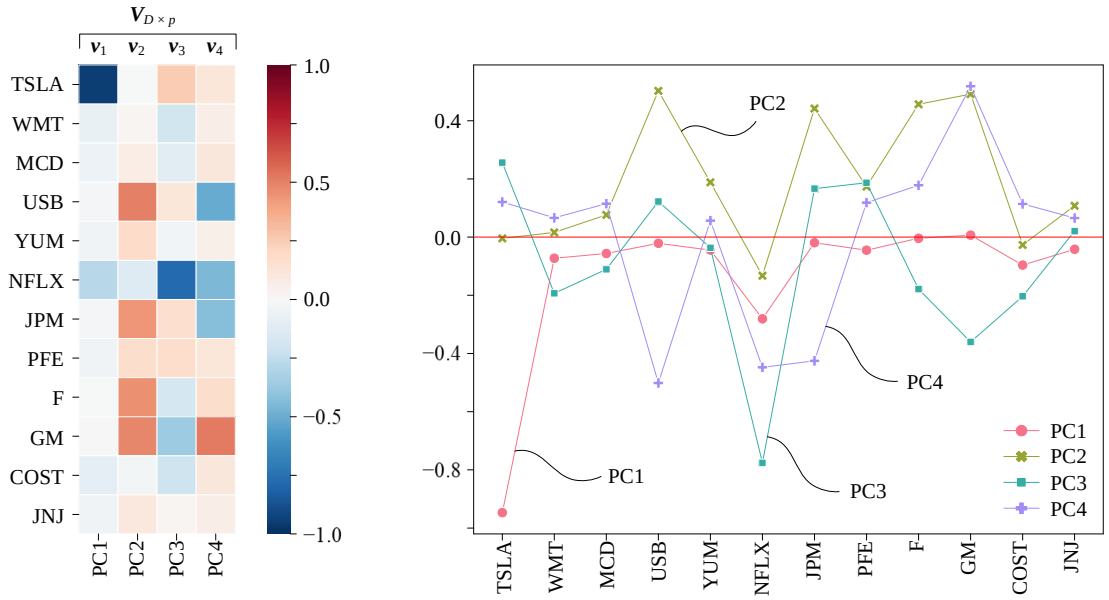


图 22. 前四个主成分可视化

图 22 所示 $\boldsymbol{V}_{D\times p}$ 两两正交，具有如下性质：

$$\boldsymbol{V}_{D\times p}{}^{\mathrm{T}}\boldsymbol{V}_{D\times p}=\boldsymbol{I}_{p\times p} \tag{62}$$

图 23 所示为 (62) 计算热图。

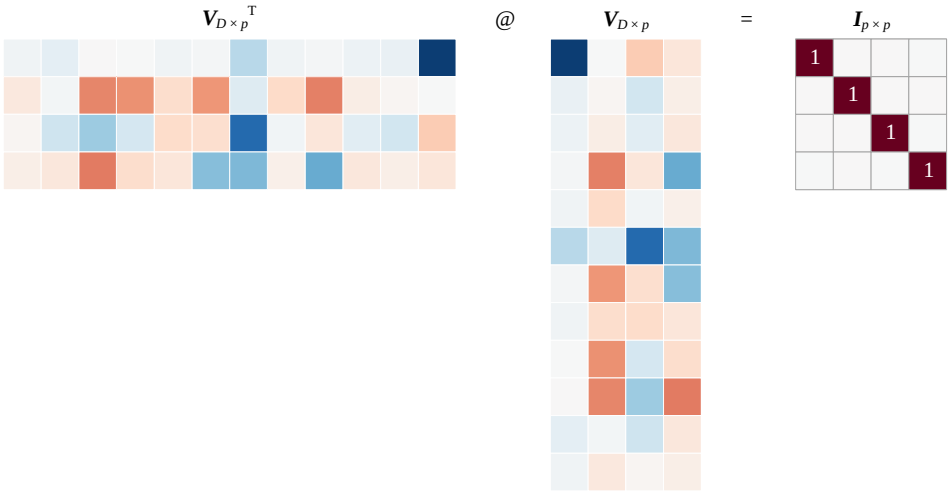


图 23. $V_{D\times p}$ 两两正交

数据投影

如图 24 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 $Z_{n\times p}$：

$$Z_{n\times p} = X_{n\times D} V_{D\times p} \tag{63}$$

图 25 所示为 $Z_{n\times p}$ 数据热图。

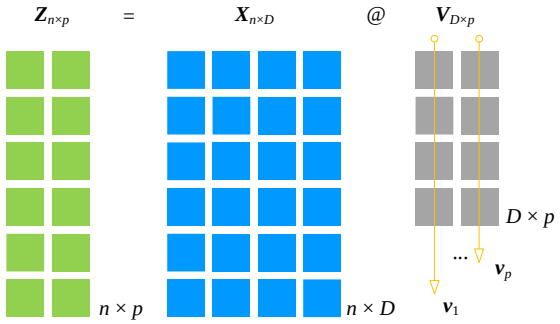


图 24. PCA 分解部分数据关系

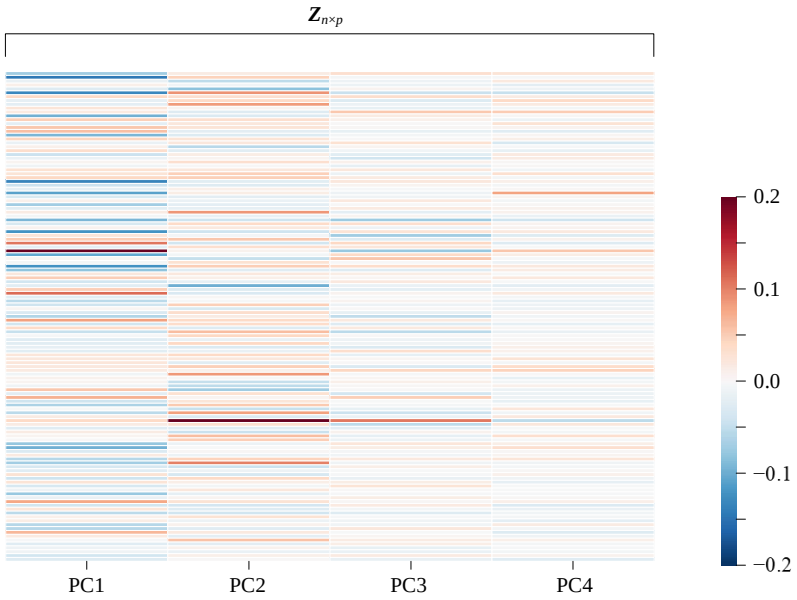


图 25. 前四个主成分数据

图 26 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。

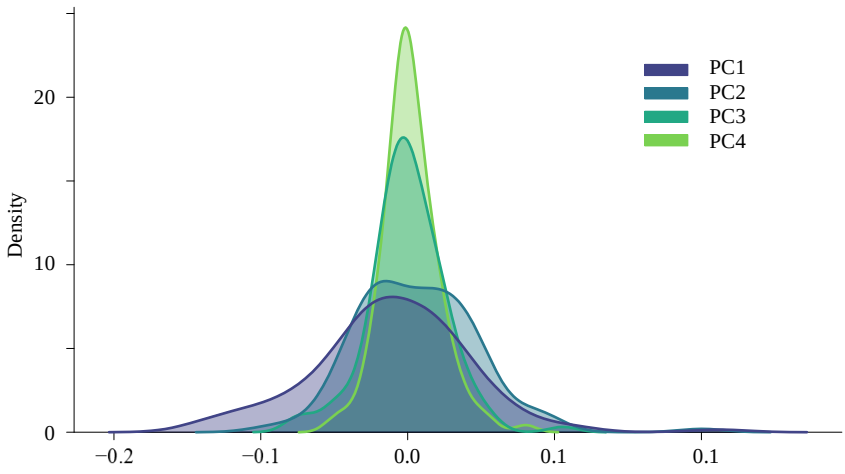


图 26. 前四个主成分数据分布

图 27 所示为 $Z_{n\times p}$ 数据协方差矩阵热图。

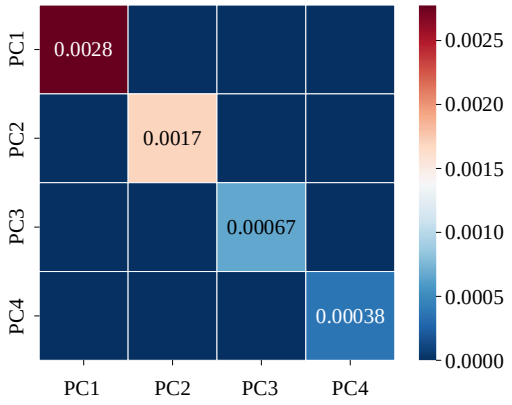


图 27. 前四个主元的协方差矩阵

前四个主成分对应的奇异值分别为：

$$s_1 = 0.5915, \quad s_2 = 0.4624, \quad s_3 = 0.2911, \quad s_4 = 0.2179 \tag{64}$$

所对应的特征值：

$$\begin{aligned} \lambda_1 &= \frac{s_1^2}{n-1} = \frac{0.5915^2}{126} = 0.0028 \\ \lambda_2 &= \frac{s_2^2}{n-1} = \frac{0.4624^2}{126} = 0.0017 \\ \lambda_3 &= \frac{s_3^2}{n-1} = \frac{0.2911^2}{126} = 0.00067 \\ \lambda_4 &= \frac{s_4^2}{n-1} = \frac{0.2179^2}{126} = 0.00038 \end{aligned} \tag{65}$$

这四个特征值对应图 27 热图对角线元素。如图 28 所示陡坡图，前四个主元解释了 84.87%方差。

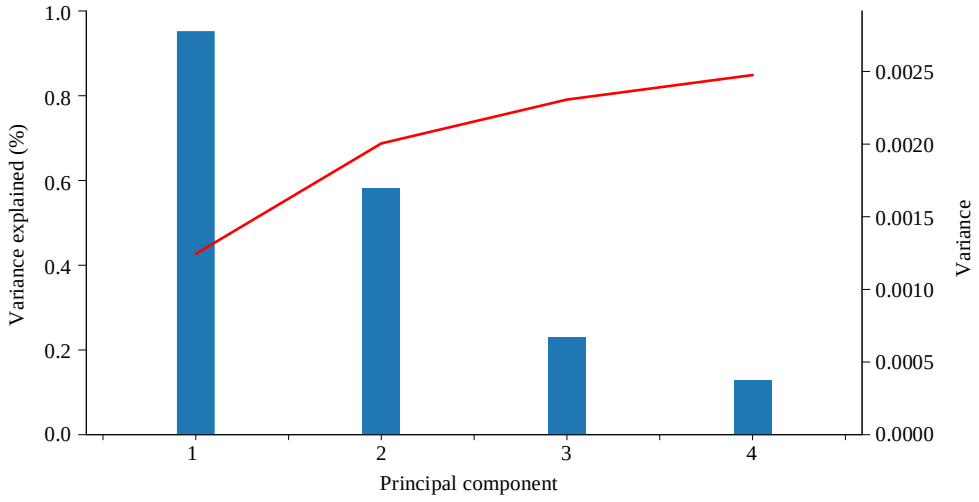


图 28. 陡坡图

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 29：

$$\boldsymbol{X}_{n\times D} = \boldsymbol{Z}_{n\times P}\left(\boldsymbol{V}_{D\times P}\right)^{\mathrm{T}} + \boldsymbol{E}_{n\times D} \tag{66}$$

图 29. $\boldsymbol{Z}_{n\times P}$ 还原数据和 $\boldsymbol{X}$ 信息差距

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：

$$\hat{\boldsymbol{y}} = b_{Z,1}\boldsymbol{z}_1 + b_{Z,2}\boldsymbol{z}_2 + \ldots + b_{Z,P}\boldsymbol{z}_P \tag{67}$$

写成矩阵运算：

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_P \end{bmatrix} \begin{bmatrix} b_{Z,1} \\ b_{Z,2} \\ \vdots \\ b_{Z,P} \end{bmatrix} = \boldsymbol{Z}_{n\times P}\boldsymbol{b}_Z \tag{68}$$

图 30 所示为上述运算过程。

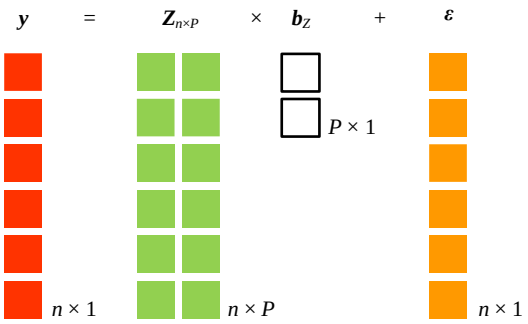


图 30. 最小二乘法回归获得 $\boldsymbol{y} = \boldsymbol{Z}_{n\times P}\boldsymbol{b}_Z + \boldsymbol{\varepsilon}$

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

$$\begin{aligned} \boldsymbol{b}_Z &= \left(\boldsymbol{Z}_{n\times P}^{\mathrm{T}}\boldsymbol{Z}_{n\times P}\right)^{-1}\boldsymbol{Z}_{n\times P}^{\mathrm{T}}\boldsymbol{y} \\ &= \left(\left(\boldsymbol{X}_{n\times D}\boldsymbol{V}_{D\times P}\right)^{\mathrm{T}}\left(\boldsymbol{X}_{n\times D}\boldsymbol{V}_{D\times P}\right)\right)^{-1}\left(\boldsymbol{X}_{n\times D}\boldsymbol{V}_{D\times P}\right)^{\mathrm{T}}\boldsymbol{y} \end{aligned} \tag{69}$$

如图 30 所示，$\boldsymbol{y}$、拟合数据 $\hat{\boldsymbol{y}}$ 和数据 $\boldsymbol{Z}_{n\times P}$ 关系如下：

$$\begin{cases} \boldsymbol{y} = \boldsymbol{Z}_{n\times P}\boldsymbol{b}_Z + \boldsymbol{\varepsilon} \\ \hat{\boldsymbol{y}} = \boldsymbol{Z}_{n\times P}\boldsymbol{b}_Z \\ \boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} \end{cases} \tag{70}$$

图 31 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：

$$\boldsymbol{b}_Z = \begin{bmatrix} -0.1039 & 0.1182 & -0.0941 & -0.0418 \end{bmatrix}^{\mathrm{T}} \tag{71}$$

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 31. 最小二乘法线性回归结果

下面将系数向量 $\boldsymbol{b}_Z$ 利用 $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_P)$ 转换为 $\boldsymbol{b}_X$，具体过程图 32 所示：

$$\boldsymbol{b}_X = \boldsymbol{V}_{D\times P}\boldsymbol{b}_Z = \boldsymbol{V}_{D\times P}\left(\boldsymbol{Z}_{n\times P}{}^{\mathrm{T}}\boldsymbol{Z}_{n\times P}\right)^{-1}\boldsymbol{Z}_{n\times P}{}^{\mathrm{T}}\boldsymbol{y} \tag{72}$$

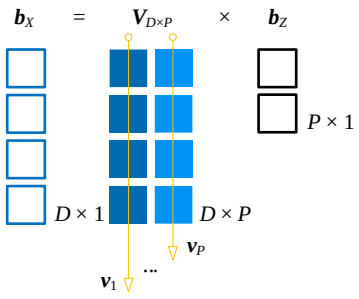


图 32. $\boldsymbol{b}_Z$ 和 $\boldsymbol{b}_X$ 之间转换关系

系数 $\boldsymbol{b}_X$ 可以通过下式计算得到：

$$\boldsymbol{b}_X = \boldsymbol{V}_{D\times P}\boldsymbol{b}_Z = \boldsymbol{V}_{D\times P}\begin{bmatrix}-0.1039 & 0.1182 & -0.0941 & -0.0418\end{bmatrix}^{\mathrm{T}} \tag{73}$$

图 33 所示为系数 $\boldsymbol{b}_X$ 直方图。

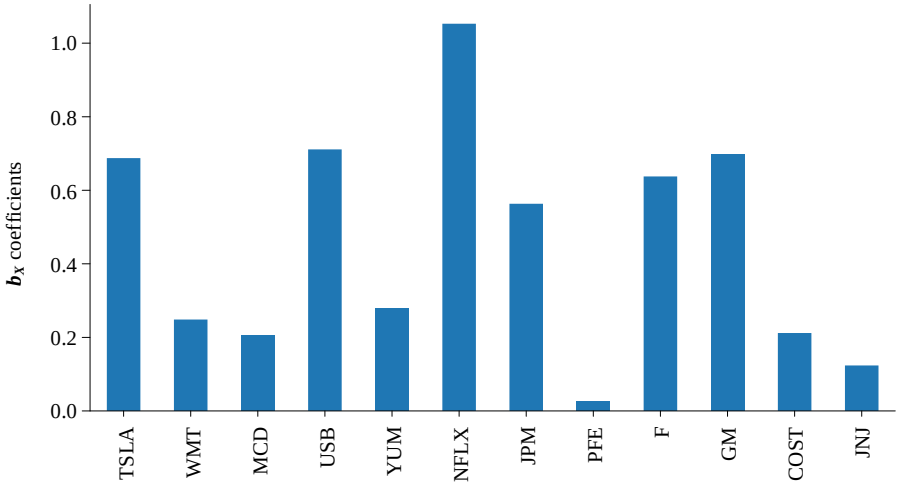


图 33. 系数 $\boldsymbol{b}_X$ 直方图

这样获得 $\boldsymbol{y}$、拟合数据 $\hat{\boldsymbol{y}}$ 和数据 $\boldsymbol{X}$ 之间关系，如图 34 所示：

$$\begin{cases}\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}_X + \boldsymbol{\varepsilon} \\ \hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}_X \\ \boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}\end{cases} \tag{74}$$

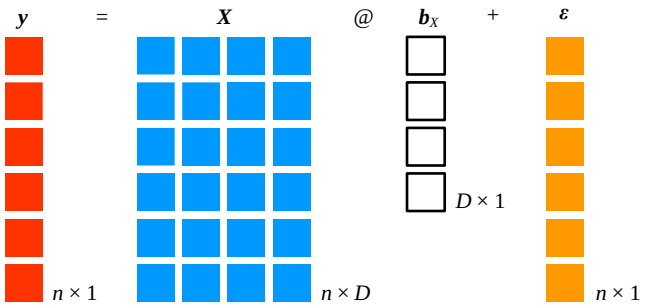


图 34. $\boldsymbol{y}$ 和数据 $\boldsymbol{X}$ 之间回归方程

计算截距项系数 b_0：

$$b_0 = \mathrm{E}(\boldsymbol{y}) - \begin{bmatrix}\mathrm{E}(\boldsymbol{x}_1) & \mathrm{E}(\boldsymbol{x}_2) & \cdots & \mathrm{E}(\boldsymbol{x}_D)\end{bmatrix}\boldsymbol{b}_X \tag{75}$$

计算截距项系数 b_0：

$$\begin{aligned} b_0 &= \mathrm{E}(\boldsymbol{y}) - \begin{bmatrix}\mathrm{E}(\boldsymbol{x}_1) & \mathrm{E}(\boldsymbol{x}_2) & \cdots & \mathrm{E}(\boldsymbol{x}_D)\end{bmatrix}\boldsymbol{b}_X \\ &= -0.00034057 \end{aligned} \tag{76}$$

最后主元回归函数可以通过下式计算得到：

$$\begin{aligned}\hat{y} &= b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_D x_D = b_0 + \begin{bmatrix} x_1 & x_2 & \cdots & x_D \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_D \end{bmatrix} = b_0 + \begin{bmatrix} x_1 & x_2 & \cdots & x_D \end{bmatrix} \boldsymbol{b}_X \\ &= b_0 + \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix} \boldsymbol{b}_Z = b_0 + \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix} \begin{bmatrix} b_{z1} \\ b_{z2} \\ \vdots \\ b_{zp} \end{bmatrix}\end{aligned} \tag{77}$$

图 35 展示主元回归计算过程数据关系。

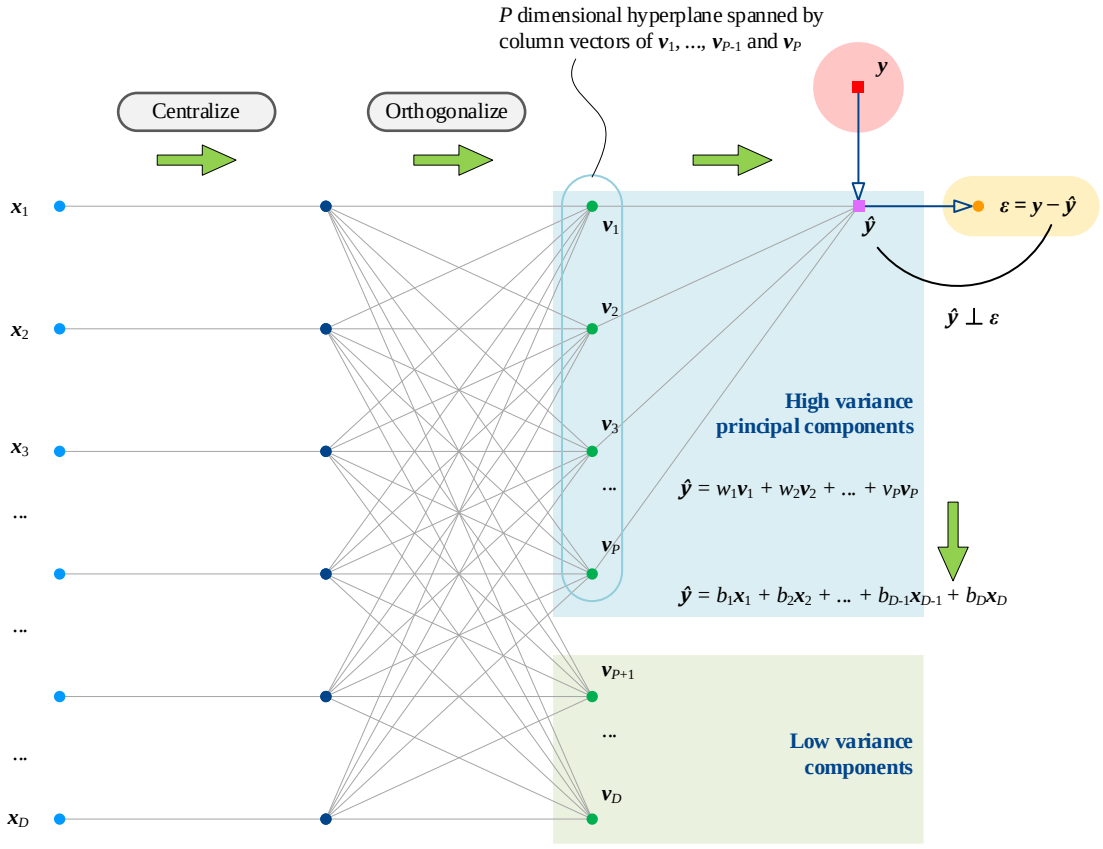


图 35. 主元回归数据关系

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 36 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 37 和图 38 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。

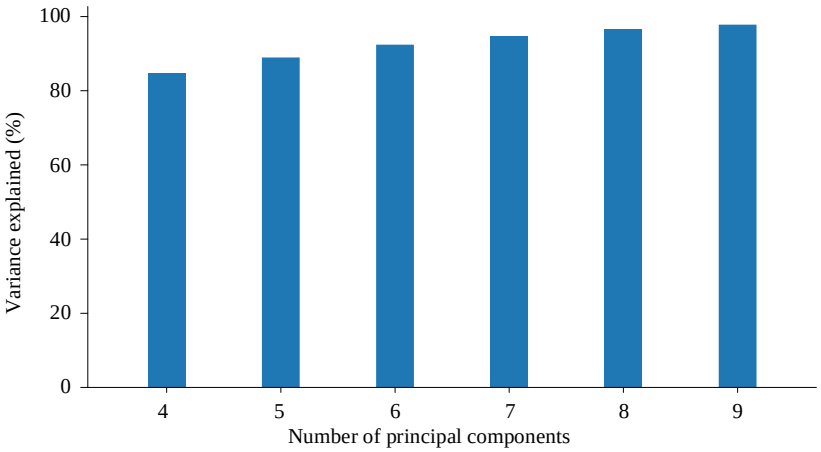


图 36. 主元数量对累计已释方差和百分比

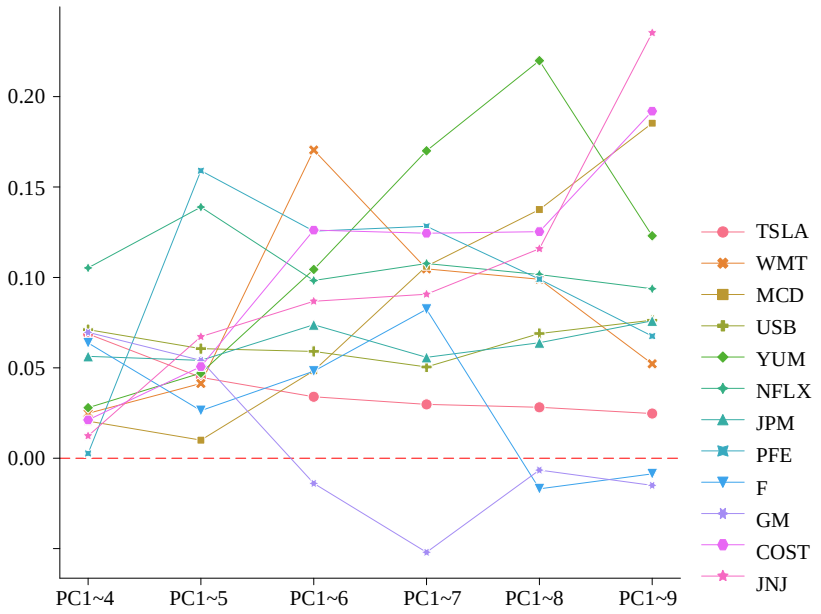


图 37. 参与主元回归主元数量对于系数的影响

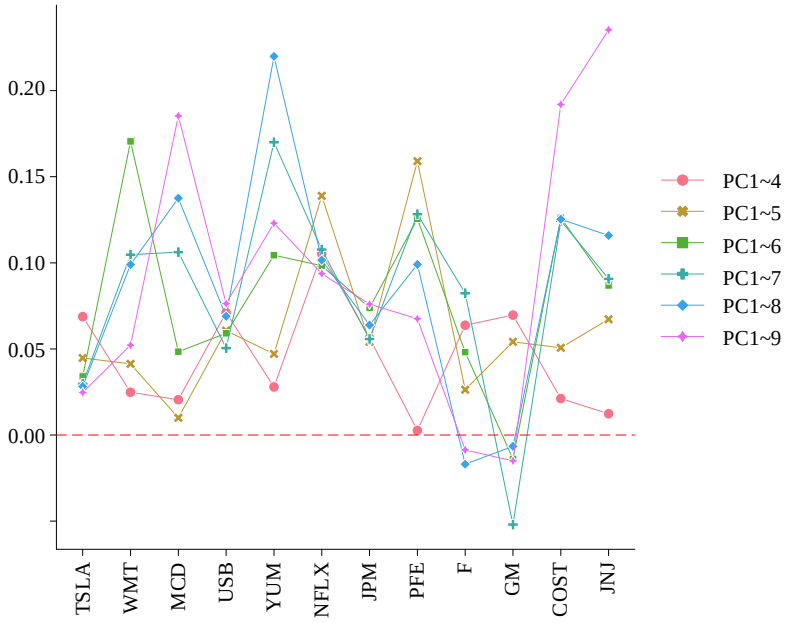


图 38. 参与主元回归主元数量对于系数的影响，第二视角

Bk7_Ch17_04.ipynb 完成主元回归运算图像。

17.7 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。PLS 在降低自变量维度的同时，建立自变量和因变量之间的线性关系模型，因此常被用于处理高维数据分析和建立多元回归模型。

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 $\boldsymbol{y}^{(0)}$：

$$\begin{aligned} \boldsymbol{X}^{(0)} &= \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{l}\boldsymbol{l}^{\mathrm{T}} \right) \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{(0)} & \boldsymbol{x}_2^{(0)} & \cdots & \boldsymbol{x}_q^{(0)} \end{bmatrix} \\ \boldsymbol{y}^{(0)} &= \boldsymbol{y} - \mathrm{E}(\boldsymbol{y}) = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{l}\boldsymbol{l}^{\mathrm{T}} \right) \boldsymbol{y} \end{aligned} \tag{78}$$

偏最小二乘回归是迭代运算，上标 (0) 代表迭代代次。

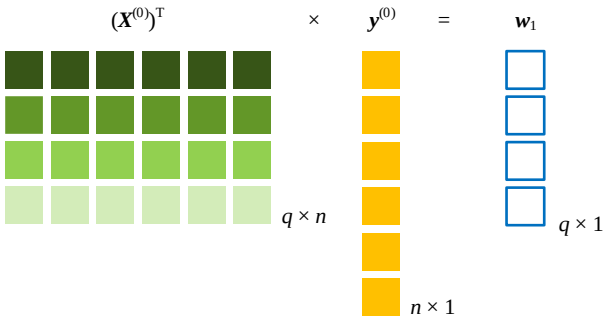


图 39. 计算权重系数列向量 $\boldsymbol{w}_1$

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 $\boldsymbol{w}_1$：

$$\boldsymbol{w}_1 = \begin{bmatrix} \mathrm{cov}\left(\boldsymbol{x}_1^{(0)}, \boldsymbol{y}^{(0)} \right) \\ \mathrm{cov}\left(\boldsymbol{x}_2^{(0)}, \boldsymbol{y}^{(0)} \right) \\ \vdots \\ \mathrm{cov}\left(\boldsymbol{x}_q^{(0)}, \boldsymbol{y}^{(0)} \right) \end{bmatrix} = \frac{1}{n} \begin{bmatrix} \left(\boldsymbol{x}_1^{(0)} \right)^{\mathrm{T}} \boldsymbol{y}^{(0)} \\ \left(\boldsymbol{x}_2^{(0)} \right)^{\mathrm{T}} \boldsymbol{y}^{(0)} \\ \vdots \\ \left(\boldsymbol{x}_q^{(0)} \right)^{\mathrm{T}} \boldsymbol{y}^{(0)} \end{bmatrix} = \left(\boldsymbol{X}^{(0)} \right)^{\mathrm{T}} \boldsymbol{y}^{(0)} \tag{79}$$

其中，列向量 $\boldsymbol{w}_1$ 行数为 q 行。

图 39 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：

$$\boldsymbol{w}_1 = \frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|} = \begin{bmatrix} w_{1,1} \\ w_{2,1} \\ \vdots \\ w_{q,1} \end{bmatrix} \tag{80}$$

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，$\boldsymbol{z}_1$：

$$\boldsymbol{z}_1 = w_{1,1}\boldsymbol{x}_1 + w_{2,1}\boldsymbol{x}_2 + \cdots + w_{q,1}\boldsymbol{x}_q = \boldsymbol{X}^{(0)}\boldsymbol{w}_1 \tag{81}$$

图 40 所示为计算偏最小二乘回归主元列向量 $\boldsymbol{z}_1$。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 $\boldsymbol{w}_1$。

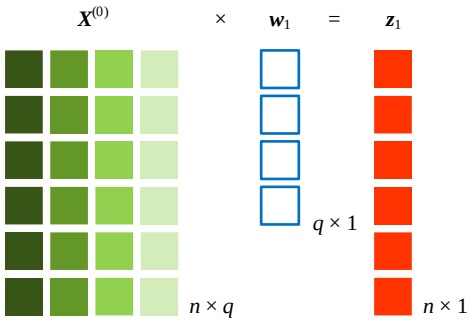


图 40. 计算偏最小二程回归主元列向量 z_1

将自变量数据矩阵 $X^{(0)}$ 和因变量数据向量 $y^{(0)}$投影到主元 z_1 方向上。

第四步

把自变量数据矩阵 $X^{(0)}$投影到主元列向量 z_1 上，获得系数向量 v_1。先以 $X^{(0)}$第一列解释投影过程。

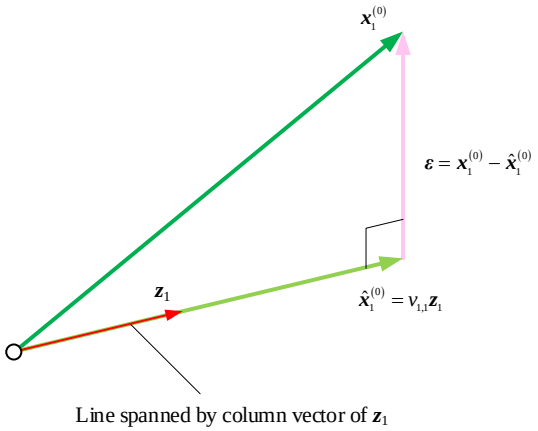


图 41. $X^{(0)}$第一列投影在主元列向量 z_1

如图 41 所示，将 $X^{(0)}$ 第一列投影到主元列向量 z_1，得到 $\hat{x}_1^{(0)}$：

$$\hat{x}_1^{(0)} = v_{1,1} z_1 \tag{82}$$

残差 ε 则垂直于主元列向量 z_1，计算获得系数 $v_{1,1}$:

$$\begin{aligned} &\varepsilon \perp z_1 \quad \Rightarrow \quad z_1^{\mathrm{T}} \varepsilon = z_1^{\mathrm{T}} \left(x_1^{(0)} - \hat{x}_1^{(0)} \right) = z_1^{\mathrm{T}} \left(x_1^{(0)} - v_{1,1} z_1 \right) = 0 \\ &\Rightarrow \quad v_{1,1} = \frac{z_1^{\mathrm{T}} x_1^{(0)}}{z_1^{\mathrm{T}} z_1} = \frac{\left(x_1^{(0)} \right)^{\mathrm{T}} z_1}{z_1^{\mathrm{T}} z_1} \end{aligned} \tag{83}$$

上式说明偏最小二乘法回归核心仍是 OLS。同样，把 $X^{(0)}$ 第二列投影在主元列向量 z_1，计算得到系数 $v_{2,1}$:

$$v_{2,1} = \frac{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{x}_2^{(0)}}{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{z}_1} = \frac{\left(\boldsymbol{x}_2^{(0)}\right)^{\mathrm{T}} \boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{z}_1} \tag{84}$$

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

$$\boldsymbol{v}_1 = \begin{bmatrix} v_{1,1} \\ v_{2,1} \\ \vdots \\ v_{q,1} \end{bmatrix} = \frac{\left(\boldsymbol{X}^{(0)}\right)^{\mathrm{T}} \boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{z}_1} = \frac{\left(\boldsymbol{X}^{(0)}\right)^{\mathrm{T}} \boldsymbol{X}^{(0)} \boldsymbol{w}_1}{\boldsymbol{w}_1^{\mathrm{T}} \left(\boldsymbol{X}^{(0)}\right)^{\mathrm{T}} \boldsymbol{X}^{(0)} \boldsymbol{w}_1} = \frac{\boldsymbol{\Sigma}^{(0)} \boldsymbol{w}_1}{\boldsymbol{w}_1^{\mathrm{T}} \boldsymbol{\Sigma}^{(0)} \boldsymbol{w}_1} \tag{85}$$

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 $\hat{\boldsymbol{X}}^{(0)}$：

$$\hat{\boldsymbol{X}}^{(0)} = \boldsymbol{z}_1 \boldsymbol{v}_1^{\mathrm{T}} = \boldsymbol{X}^{(0)} \boldsymbol{w}_1 \boldsymbol{v}_1^{\mathrm{T}} \tag{86}$$

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 $\boldsymbol{E}^{(0)}$：

$$\boldsymbol{E}^{(0)} = \boldsymbol{X}^{(0)} - \hat{\boldsymbol{X}}^{(0)} = \boldsymbol{X}^{(0)} - \boldsymbol{X}^{(0)} \boldsymbol{w}_1 \boldsymbol{v}_1^{\mathrm{T}} = \boldsymbol{X}^{(0)} \left(\boldsymbol{I} - \boldsymbol{w}_1 \boldsymbol{v}_1^{\mathrm{T}}\right) \tag{87}$$

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

$$\boldsymbol{X}^{(1)} = \boldsymbol{E}^{(0)} = \boldsymbol{X}^{(0)} - \hat{\boldsymbol{X}}^{(0)} = \boldsymbol{X}^{(0)} \left(\boldsymbol{I} - \boldsymbol{w}_1 \boldsymbol{v}_1^{\mathrm{T}}\right) \tag{88}$$

数据矩阵 $\boldsymbol{X}^{(1)}$ 和原始数据 $\boldsymbol{X}^{(0)}$ 之间关系如图 42 所示。

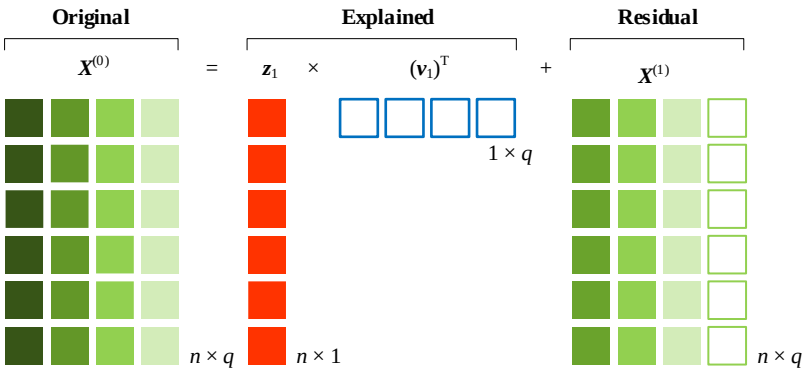


图 42. 计算得到数据矩阵 $\boldsymbol{X}^{(1)}$

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 43 所示，用最小二乘法计算获得系数 b_1：

$$\begin{aligned} &\boldsymbol{\varepsilon} \perp \boldsymbol{z}_1 \quad \Rightarrow \quad \boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{\varepsilon} = \boldsymbol{z}_1^{\mathrm{T}} \left(\boldsymbol{y}^{(0)} - \hat{\boldsymbol{y}}^{(0)}\right) = \boldsymbol{z}_1^{\mathrm{T}} \left(\boldsymbol{y}^{(0)} - b_1 \boldsymbol{z}_1\right) = 0 \\ &\Rightarrow \quad b_1 = \frac{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{y}^{(0)}}{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{z}_1} = \frac{\left(\boldsymbol{y}^{(0)}\right)^{\mathrm{T}} \boldsymbol{z}_1}{\boldsymbol{z}_1^{\mathrm{T}} \boldsymbol{z}_1} \end{aligned} \tag{89}$$

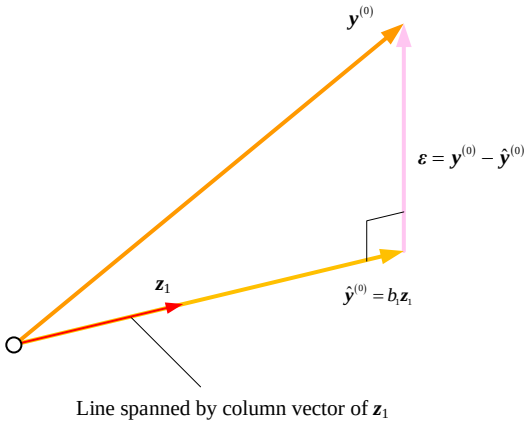


图 43. $y^{(0)}$ 向量投影在主元列向量 z_1

第七步

根据 OLS 原理，利用列向量 b_1 和 z_1 估算因变量列向量 y，并到拟合 $\hat{y}^{(0)}$：

$$\hat{y}^{(0)} = b_1 z_1 = \frac{z_1^T y^{(0)} z_1}{z_1^T z_1} = \frac{\left(y^{(0)}\right)^T z_1 z_1}{z_1^T z_1} \tag{90}$$

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

$$\varepsilon^{(0)} = y^{(1)} = y^{(0)} - \hat{y}^{(0)} = y^{(0)} - \frac{z_1^T y^{(0)} z_1}{z_1^T z_1} \tag{91}$$

而残差向量 $\varepsilon^{(0)}$ 便是进入迭代循环第二步数据向量 $y^{(1)}$。如图 44 所示，$\hat{y}^{(0)}$ 解释部分 $y^{(0)}$。

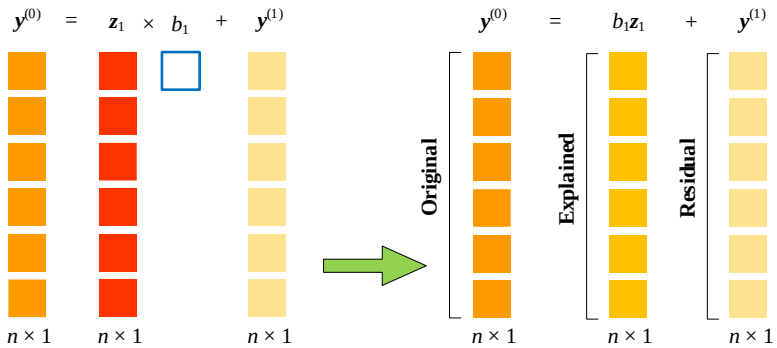


图 44. 估算 $y^{(0)}$

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：

$$\boldsymbol{w}_2 = \frac{\left(\boldsymbol{X}^{(1)}\right)^{\mathrm{T}} \boldsymbol{y}^{(1)}}{\left\| \left(\boldsymbol{X}^{(1)}\right)^{\mathrm{T}} \boldsymbol{y}^{(1)} \right\|} \tag{92}$$

重复第三步，利用权重系数列向量 $\boldsymbol{w}_2$ 和 $\boldsymbol{X}^{(1)}$ 构造偏最小二乘回归第二主元向量，$\boldsymbol{z}_2$：

$$\boldsymbol{z}_2 = \boldsymbol{X}^{(1)} \boldsymbol{w}_2 \tag{93}$$

重复第四步，把自变量数据残差矩阵 $\boldsymbol{X}^{(1)}$ 投影于第二主元列向量 $\boldsymbol{z}_2$ 上，获得系数向量 $\boldsymbol{v}_2$：

$$\boldsymbol{v}_2 = \begin{bmatrix} v_{1,2} \\ v_{2,2} \\ \vdots \\ v_{q,2} \end{bmatrix} = \frac{\left(\boldsymbol{X}^{(1)}\right)^{\mathrm{T}} \boldsymbol{z}_2}{\boldsymbol{z}_2^{\mathrm{T}} \boldsymbol{z}_2} = \frac{\left(\boldsymbol{X}^{(1)}\right)^{\mathrm{T}} \boldsymbol{X}^{(1)} \boldsymbol{w}_2}{\boldsymbol{w}_2^{\mathrm{T}} \left(\boldsymbol{X}^{(1)}\right)^{\mathrm{T}} \boldsymbol{X}^{(1)} \boldsymbol{w}_2} = \frac{\boldsymbol{\Sigma}^{(1)} \boldsymbol{w}_2}{\boldsymbol{w}_2^{\mathrm{T}} \boldsymbol{\Sigma}^{(1)} \boldsymbol{w}_2} \tag{94}$$

重复第五步，用列向量 $\boldsymbol{v}_2$ 和 $\boldsymbol{z}_2$ 估算，并到拟合矩阵 $\hat{\boldsymbol{X}}^{(1)}$：

$$\hat{\boldsymbol{X}}^{(1)} = \boldsymbol{z}_2 \boldsymbol{v}_2^{\mathrm{T}} = \boldsymbol{X}^{(1)} \boldsymbol{w}_2 \boldsymbol{v}_2^{\mathrm{T}} \tag{95}$$

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

$$\boldsymbol{X}^{(2)} = \boldsymbol{E}^{(1)} = \boldsymbol{X}^{(1)} - \hat{\boldsymbol{X}}^{(1)} = \boldsymbol{X}^{(1)} \left(\boldsymbol{I} - \boldsymbol{w}_2 \boldsymbol{v}_2^{\mathrm{T}}\right) \tag{96}$$

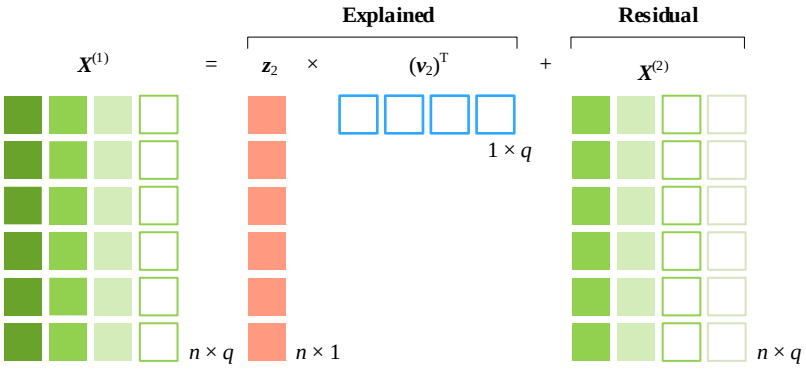


图 45. 计算得到数据矩阵 $\boldsymbol{X}^{(2)}$

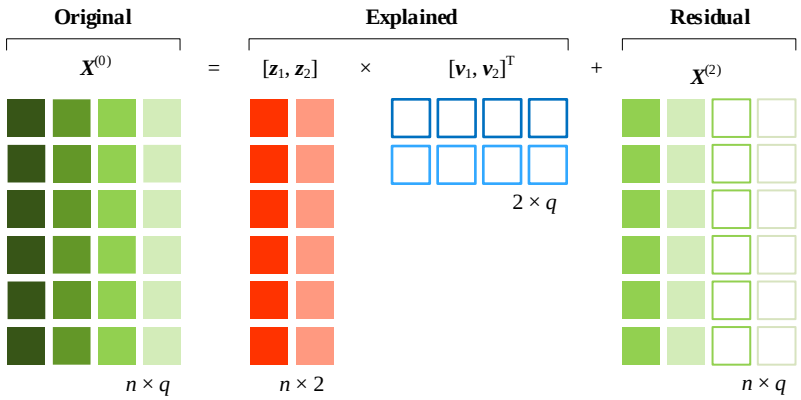


图 46. 前两个主元 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_1$ 还原数据矩阵 $\boldsymbol{X}^{(0)}$

图 42 和图 45 相结合获得图 46，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：

$$b_2 = \frac{\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{y}^{(1)}}{\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_2} = \frac{\left(\boldsymbol{y}^{(1)}\right)^{\mathrm{T}}\boldsymbol{z}_2}{\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_2} \tag{97}$$

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

$$\hat{\boldsymbol{y}}^{(1)} = b_2\boldsymbol{z}_2 \tag{98}$$

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 $\boldsymbol{\varepsilon}^{(1)}$：

$$\boldsymbol{\varepsilon}^{(0)} = \boldsymbol{y}^{(2)} = \boldsymbol{y}^{(1)} - \hat{\boldsymbol{y}}^{(1)} = \boldsymbol{y}^{(1)} - b_2\boldsymbol{z}_2 \tag{99}$$

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 $\boldsymbol{y}^{(2)}$。

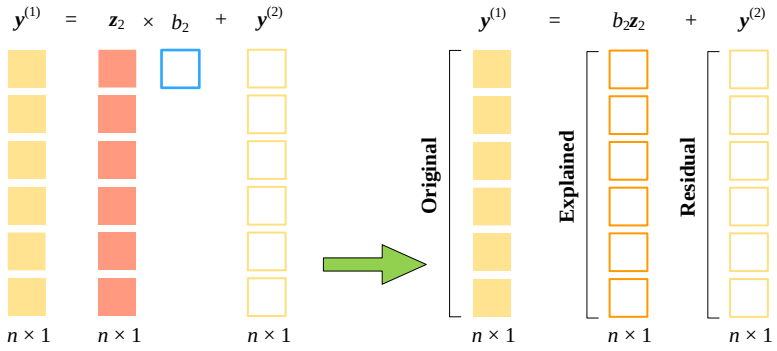


图 47. 估算 $\boldsymbol{y}^{(1)}$

图 48 结合图 44 和图 47，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。

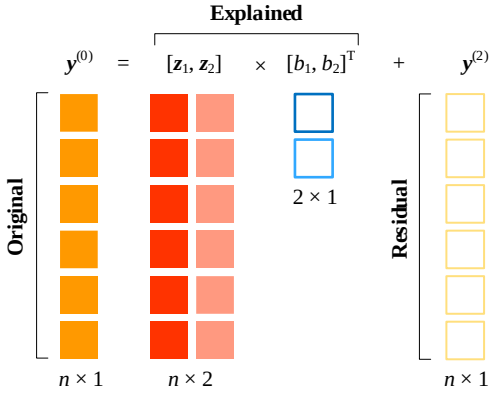
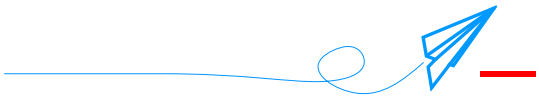


图 48. 前两个主元 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_1$ 还原部分数据列向量 $\boldsymbol{y}^{(0)}$

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



正交回归和最小二乘法回归都是回归分析中的方法，但它们之间有很大的区别。

OLS 通过最小化实际观测值与预测值之间的误差平方和，来确定回归系数。这种方法非常直观且易于理解，但存在一些缺点，例如当数据存在多重共线性时，OLS 的估计结果可能会变得不稳定，且估计结果受到极端值的影响较大。

与 OLS 不同，正交回归是一种基于主成分分析的回归方法。它通过将自变量通过主成分分析转换成互相正交的新变量，来消除自变量之间的多重共线性问题，从而提高回归分析的准确性和稳定性。

因此，正交回归方法相对于 OLS 方法更加鲁棒，适用于多重共线性较强的数据集，同时也能够在保证预测准确性的前提下，降低自变量的维度，提高回归模型的可解释性。

主元回归 PCR 是一种基于主成分分析的回归方法，它在回归建模之前，先对自变量进行主成分分析，将自变量降维成少量的主成分变量，然后再对这些主成分变量进行回归分析。

PCR 的基本思想是将自变量通过主成分分析转换成少数互相正交的主成分变量，从而消除自变量之间的多重共线性问题，提高回归分析的准确性和稳定性。在降维过程中，PCR 保留了自变量中最主要的信息，因此相比于直接使用全部自变量的回归分析，PCR 可以显著提高回归模型的准确性和可解释性。

偏最小二乘 PLS 也是一种基于主成分分析和回归分析的统计建模方法，它是对 PCR 的一种改进，主要用于解决多重共线性和高维数据分析问题。

与 PCR 不同的是，PLS 在主成分分析的过程中，不仅仅考虑了自变量之间的方差，还考虑了自变量和因变量之间的协方差，从而将主成分分析与回归分析相结合，得到了一组互相正交的主成分变量，每个主成分变量都包含了自变量和因变量的信息，可以用于回归分析。

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