

# 16

## Orthogonal Distance Regression

# 正交回归

输入和输出数据都参与主成分分析，构造正交空间



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
- ◀ `statsmodels.api.add_constant()` 增加 OLS 常数项
- ◀ `statsmodels.api.OLS` 最小二乘法线性回归

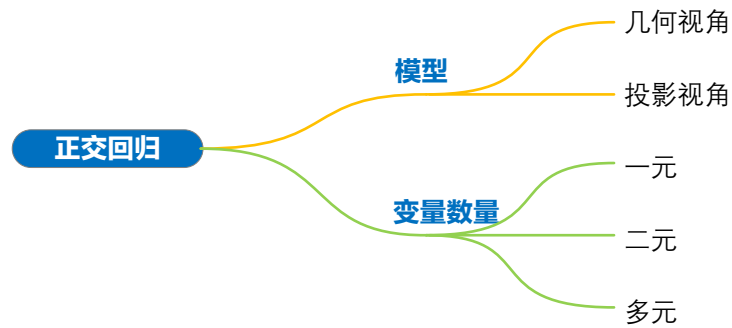
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)



## 16.1 主成分与回归

本章主要介绍一种和主成分分析息息相关的回归方法——**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。

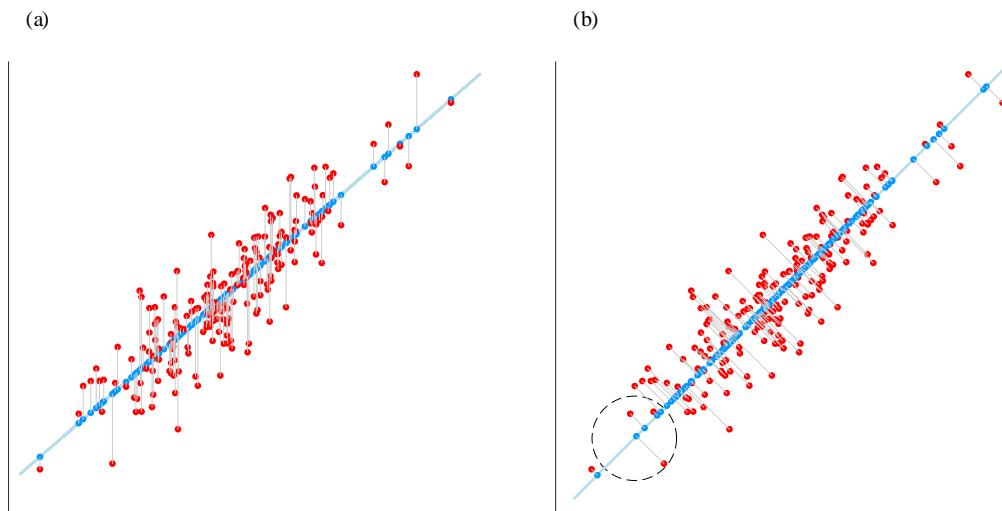


图 1. 对比 OLS 和 TLS 线性回归

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

下面用两特征  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2]$  数据作例子，聊一下主成分回归的思想。如图 2 所示， $\mathbf{x}_1$  和  $\mathbf{x}_2$  为输入数据， $\mathbf{y}$  为输出数据；通过主成分分析， $\mathbf{x}_1$ 、 $\mathbf{x}_2$  和  $\mathbf{y}$  正交化之后得到  $\mathbf{v}_1$ 、 $\mathbf{v}_2$  和  $\mathbf{v}_3$  (根据特征值从小到大排列)； $\mathbf{v}_1$ 、 $\mathbf{v}_2$  和  $\mathbf{v}_3$  两两正交。第一主成分  $\mathbf{v}_1$  和第二主成分  $\mathbf{v}_2$  构造平面  $H$ 。 $\mathbf{v}_3$  垂直于平面  $H$ ，通过这层关系求解出正交回归系数。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

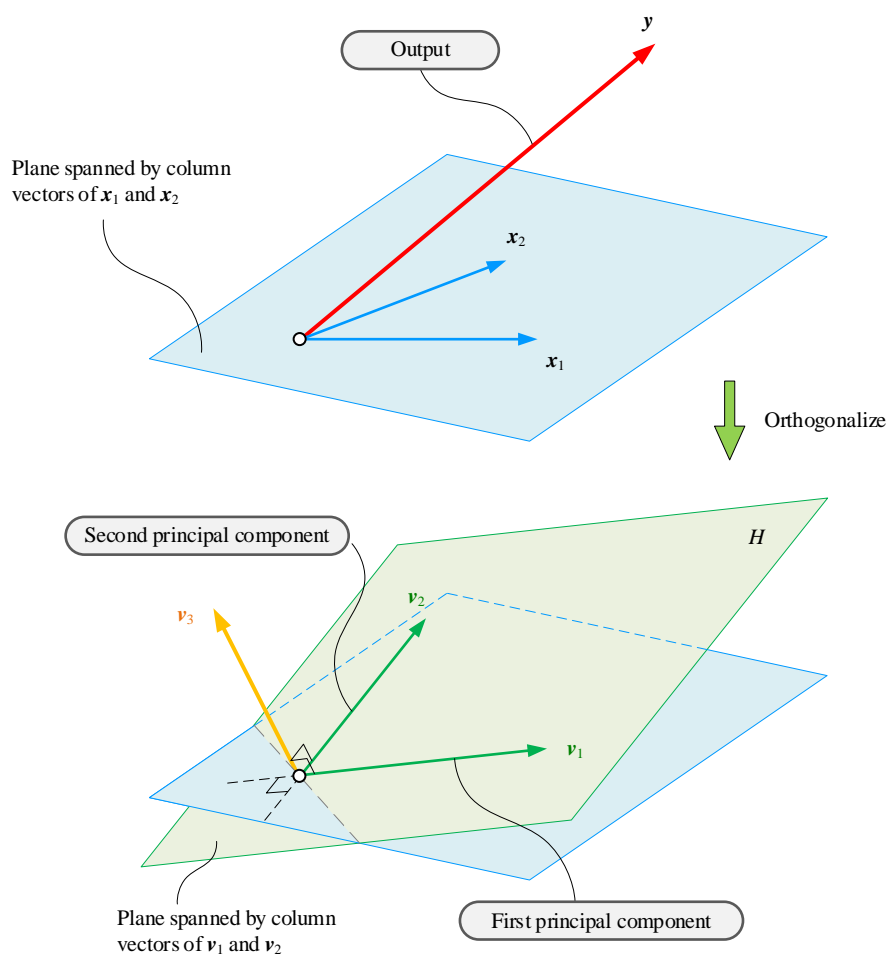
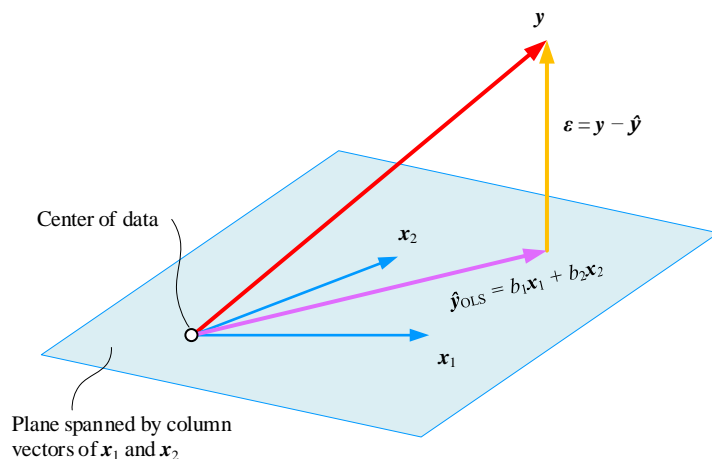
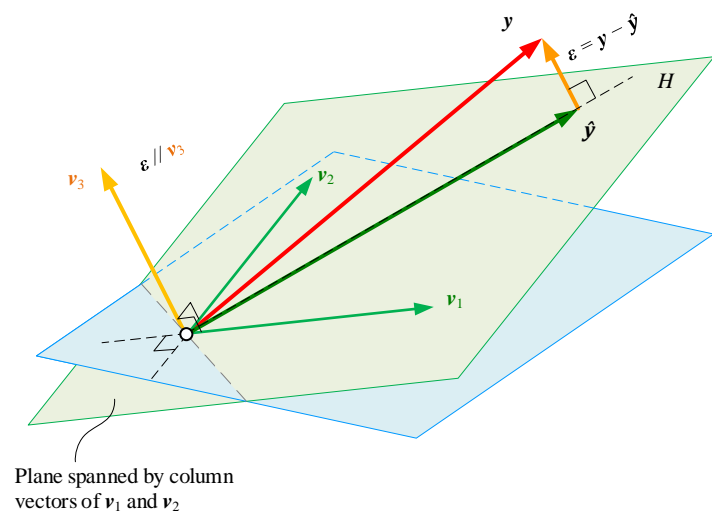


图 2. 通过主成分分析构造正交空间

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  $y$  投影到  $x_1$  和  $x_2$  构造的平面上。而对于正交回归, 将  $y$  投影到  $H$ , 得到  $\hat{y}$ 。而残差,  $\varepsilon = y - \hat{y}$ , 平行于  $v_3$ 。再次强调, 平面  $H$  是由第一主成分  $v_1$  和第二主成分  $v_2$  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  $y$  投影到  $x_1$  和  $x_2$  构造的平面上图 4. 正交回归，将输出数据  $y$  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。

## 16.2 一元正交回归

设定一元正交回归解析式如下：

$$y = b_0 + b_1 x \quad (1)$$

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：

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

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

$$\mathbf{d} = \frac{\mathbf{y} - (b_0 + b_1 \mathbf{x})}{\sqrt{1 + b_1^2}} \quad (3)$$

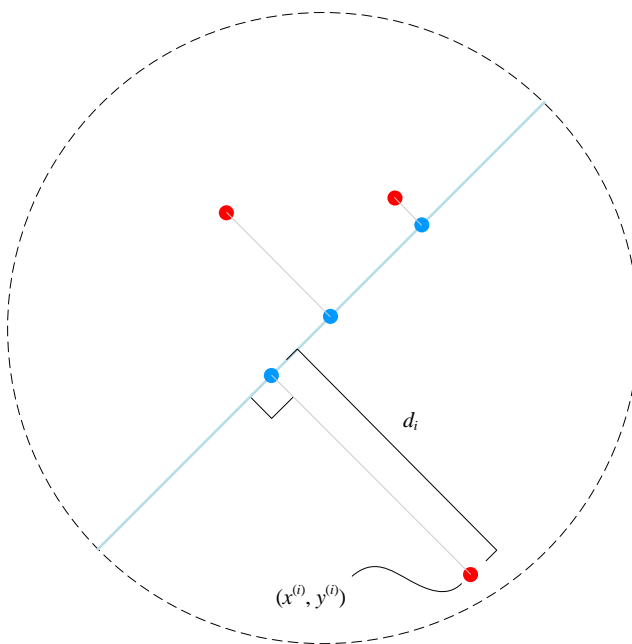


图 5. 正交投影几何关系

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:

$$\arg \min_{b_0, b_1} f(b_0, b_1) = \|\mathbf{d}\|^2 = \mathbf{d}^T \mathbf{d} \quad (4)$$

将 (3) 代入  $f(b_0, b_1)$  得到:

$$f(b_0, b_1) = \frac{(\mathbf{y} - (b_0 + b_1 \mathbf{x}))^T (\mathbf{y} - (b_0 + b_1 \mathbf{x}))}{1 + b_1^2} \quad (5)$$

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

$$f(b_0, b_1) = \frac{nb_0^2 + 2b_0 b_1 \mathbf{x}^T \mathbf{I} + b_1^2 \mathbf{x}^T \mathbf{x} - 2b_0 \mathbf{y}^T \mathbf{I} - 2b_1 \mathbf{x}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}}{1 + b_1^2} \quad (6)$$

$f(b_0, b_1)$  对  $b_0$  偏导为 0, 构造如下等式:

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \frac{2nb_0 + 2b_1 \mathbf{x}^T \mathbf{I} - 2\mathbf{y}^T \mathbf{I}}{1 + b_1^2} = 0 \quad (7)$$

$f(b_0, b_1)$  对  $b_1$  偏导为 0, 构造如下等式:

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \frac{2b_1 \mathbf{x}^T \mathbf{x} + 2b_0 \mathbf{x}^T \mathbf{I} - 2\mathbf{x}^T \mathbf{y}}{1 + b_1^2} - \frac{(nb_0^2 + 2b_0 b_1 \mathbf{x}^T \mathbf{I} + b_1^2 \mathbf{x}^T \mathbf{x} - 2b_0 \mathbf{y}^T \mathbf{I} - 2b_1 \mathbf{x}^T \mathbf{y} + \mathbf{y}^T \mathbf{y})2b_1}{(1 + b_1^2)^2} = 0 \quad (8)$$

观察 (7)，容易用  $b_1$  表达  $b_0$ ：

$$b_0 = \frac{\mathbf{y}^T \mathbf{I} - b_1 \mathbf{x}^T \mathbf{I}}{n} = E(\mathbf{y}) - b_1 E(\mathbf{x}) \quad (9)$$

其中，

$$\begin{cases} E(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{I}}{n} = \frac{\sum_{i=1}^n x^{(i)}}{n} \\ E(\mathbf{y}) = \frac{\mathbf{y}^T \mathbf{I}}{n} = \frac{\sum_{i=1}^n y^{(i)}}{n} \end{cases} \quad (10)$$

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：

$$b_1^2 + kb_1 - 1 = 0 \quad (11)$$

其中，

$$\begin{aligned} k &= \frac{n\mathbf{x}^T \mathbf{x} - \mathbf{x}^T \mathbf{I} \mathbf{x}^T \mathbf{I} - n\mathbf{y}^T \mathbf{y} + \mathbf{y}^T \mathbf{I} \mathbf{y}^T \mathbf{I}}{n\mathbf{x}^T \mathbf{y} - \mathbf{x}^T \mathbf{I} \mathbf{y}^T \mathbf{I}} \\ &= \left( \frac{\mathbf{x}^T \mathbf{x}}{n} - \frac{\mathbf{x}^T \mathbf{I} \mathbf{x}^T \mathbf{I}}{n^2} \right) - \left( \frac{\mathbf{y}^T \mathbf{y}}{n} - \frac{\mathbf{y}^T \mathbf{I} \mathbf{y}^T \mathbf{I}}{n^2} \right) \\ &= \frac{\mathbf{x}^T \mathbf{y}}{n} - \frac{\mathbf{x}^T \mathbf{I} \mathbf{y}^T \mathbf{I}}{n^2} \\ &= \frac{\text{var}(\mathbf{x}) - \text{var}(\mathbf{y})}{\text{cov}(\mathbf{x}, \mathbf{y})} = \frac{\sigma_x^2 - \sigma_y^2}{\rho_{xy} \sigma_x \sigma_y} \end{aligned} \quad (12)$$

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：

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

将 (12) 给出的  $k$ ，代入 (13)，整理得到  $b_1$  解：

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

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

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

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho_{xy} \sigma_x \sigma_y \\ \rho_{xy} \sigma_x \sigma_y & \sigma_y^2 \end{bmatrix} \quad (15)$$

对  $\Sigma$  进行特征值分解，得到两个特征向量：

$$\begin{aligned} \mathbf{v}_1 &= \begin{bmatrix} \frac{(\sigma_y^2 - \sigma_x^2) + \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4(\rho_{xy}\sigma_x\sigma_y)^2}}{2\rho_{xy}\sigma_x\sigma_y} \\ 1 \end{bmatrix} \\ \mathbf{v}_2 &= \begin{bmatrix} \frac{(\sigma_y^2 - \sigma_x^2) - \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4(\rho_{xy}\sigma_x\sigma_y)^2}}{2\rho_{xy}\sigma_x\sigma_y} \\ 1 \end{bmatrix} \end{aligned} \quad (16)$$

$\Sigma$  两个特征值，从大到小排列：

$$\begin{aligned} \lambda_1 &= \frac{\sigma_x^2 + \sigma_y^2}{2} + \sqrt{\left(\rho_{xy}\sigma_x\sigma_y\right)^2 + \left(\frac{\sigma_x^2 - \sigma_y^2}{2}\right)^2} \\ \lambda_2 &= \frac{\sigma_x^2 + \sigma_y^2}{2} - \sqrt{\left(\rho_{xy}\sigma_x\sigma_y\right)^2 + \left(\frac{\sigma_x^2 - \sigma_y^2}{2}\right)^2} \end{aligned} \quad (17)$$

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：

$$y = 1.4377x + 0.00157 \quad (18)$$

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。

## 16.3 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 ( $E(\mathbf{x})$ ,  $E(\mathbf{y})$ )，回归直线方向与主元方向  $\mathbf{v}_1$  平行，垂直于次元  $\mathbf{v}_2$  方向。即，次元方向  $\mathbf{v}_2$  和直线法向量  $\mathbf{n}$  平行。



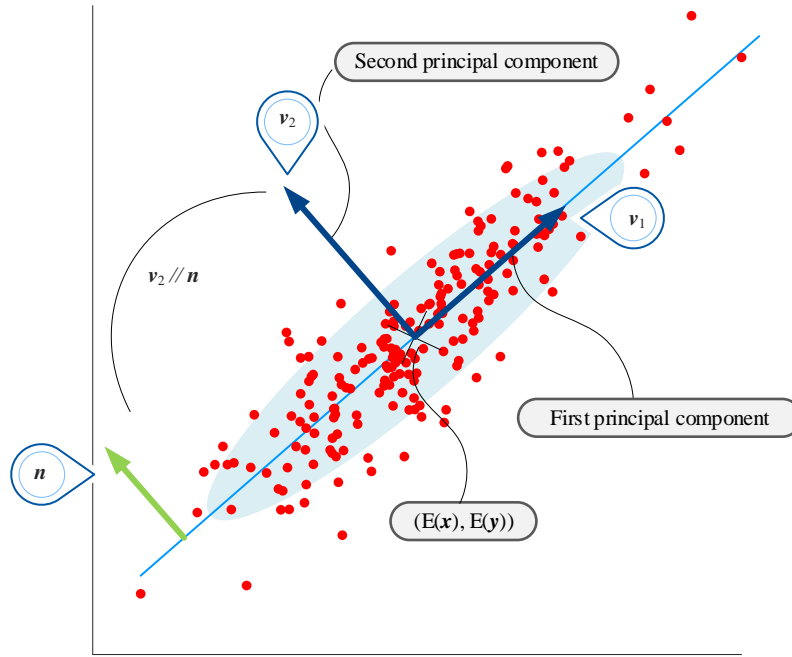


图 6. 正交回归和 PCA 分解关系

对于 (1) 所示一元一次函数，构造二元  $F(x, y)$  函数如下：

$$F(x, y) = b_0 + b_1 x - y \quad (19)$$

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

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

如前文所示， $\mathbf{n}$  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：

$$\mathbf{x} = \mathbf{x} - E(\mathbf{x}), \quad \mathbf{y} = \mathbf{y} - E(\mathbf{y}) \quad (21)$$

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

$$\Sigma = [\mathbf{x} \quad \mathbf{y}]^T [\mathbf{x} \quad \mathbf{y}] = \begin{bmatrix} \mathbf{x}^T \\ \mathbf{y}^T \end{bmatrix} [\mathbf{x} \quad \mathbf{y}] = \begin{bmatrix} \mathbf{x}^T \mathbf{x} & \mathbf{x}^T \mathbf{y} \\ \mathbf{y}^T \mathbf{x} & \mathbf{y}^T \mathbf{y} \end{bmatrix} \quad (22)$$

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  $\mathbf{n}$  为  $\Sigma$  次元方向：

$$\Sigma \mathbf{n} = \lambda_2 \mathbf{n} \Rightarrow \begin{bmatrix} \mathbf{x}^T \mathbf{x} & \mathbf{x}^T \mathbf{y} \\ \mathbf{y}^T \mathbf{x} & \mathbf{y}^T \mathbf{y} \end{bmatrix} \mathbf{n} = \lambda_2 \mathbf{n} \quad (23)$$

将 (20) 代入 (23)，整理得到如下两个等式：

$$\begin{bmatrix} \mathbf{x}^T \mathbf{x} & \mathbf{x}^T \mathbf{y} \\ \mathbf{y}^T \mathbf{x} & \mathbf{y}^T \mathbf{y} \end{bmatrix} \begin{bmatrix} b_1 \\ -1 \end{bmatrix} = \lambda_2 \begin{bmatrix} b_1 \\ -1 \end{bmatrix} \Rightarrow \begin{cases} \mathbf{x}^T \mathbf{x} b_1 - \mathbf{x}^T \mathbf{y} = \lambda_2 b_1 \\ \mathbf{y}^T \mathbf{x} b_1 - \mathbf{y}^T \mathbf{y} = -\lambda_2 \end{cases} \quad (24)$$

联立 (24) 两个等式，用  $\lambda_2$  表示  $b_1$ ：

$$b_{1\_TLS} = (\mathbf{x}^T \mathbf{x} - \lambda_2)^{-1} \mathbf{x}^T \mathbf{y} \quad (25)$$

下式为本书前文获得的一元线性回归 OLS 中  $b_1$  解：

$$b_{1\_OLS} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{y} \quad (26)$$

对比 OLS 和 TLS；当 (25) 中  $\lambda_2$  为 0 时，两种回归方法得到斜率完全一致。 $\lambda_2 = 0$  时， $\mathbf{y}$  和  $\mathbf{x}$  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  $(E(\mathbf{x}), E(\mathbf{y}))$  一点，即：

$$E(\mathbf{y}) = b_0 + b_1 E(\mathbf{x}) \quad (27)$$

获得回归式截距项  $b_0$  表达式：

$$b_0 = E(\mathbf{y}) - b_1 E(\mathbf{x}) \quad (28)$$

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

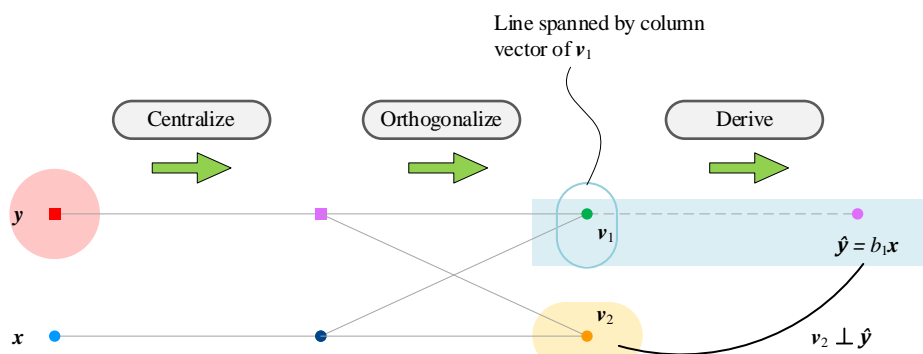


图 7. 一元正交回归 TLS 数据关系

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：

$$y = 1.1225x + 0.0018 \quad (29)$$

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

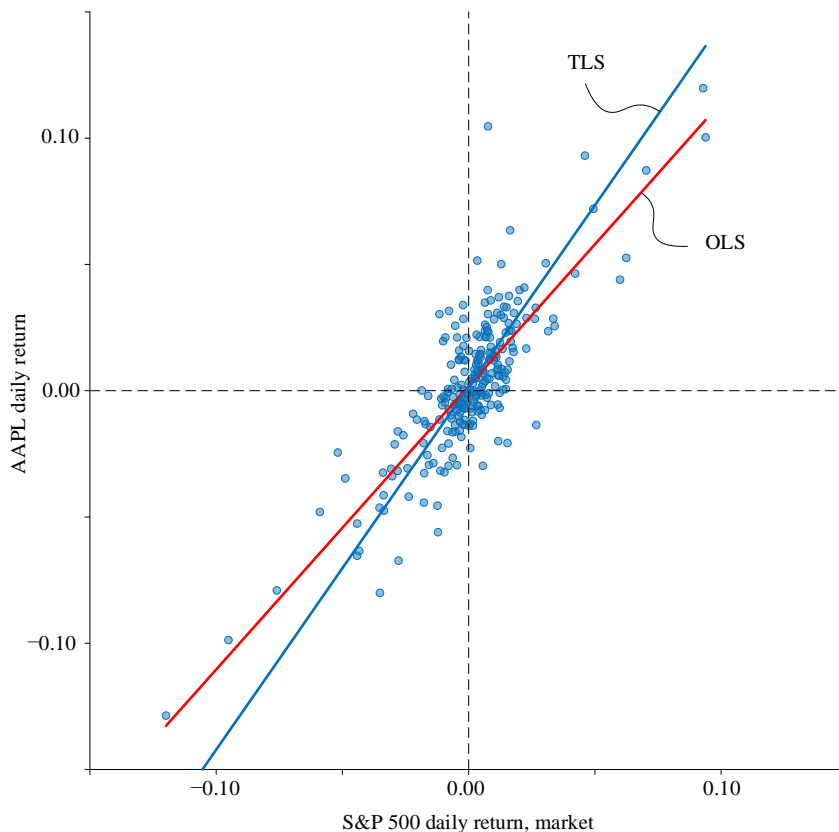


图 9. 比较 OLS 和 TLS 结果



Bk6\_Ch16\_01.py 绘制本节图像。

## 16.4 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：

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

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

$$\begin{aligned} \Sigma &= [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{y}]^T [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{y}] \\ &= \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \mathbf{y}^T \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \mathbf{x}_1 & \mathbf{x}_1^T \mathbf{x}_2 & \mathbf{x}_1^T \mathbf{y} \\ \mathbf{x}_2^T \mathbf{x}_1 & \mathbf{x}_2^T \mathbf{x}_2 & \mathbf{x}_2^T \mathbf{y} \\ \mathbf{y}^T \mathbf{x}_1 & \mathbf{y}^T \mathbf{x}_2 & \mathbf{y}^T \mathbf{y} \end{bmatrix} \end{aligned} \quad (31)$$

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：

$$y = b_0 + b_1 x_1 + b_2 x_2 \quad (32)$$

构造二元  $F(x_1, x_2, y)$  函数如下：

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

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

$$\mathbf{n} = \left( \frac{\partial F}{\partial x_1}, \frac{\partial F}{\partial x_2}, \frac{\partial F}{\partial y} \right)^T = [b_1 \quad b_2 \quad -1]^T \quad (34)$$

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

$$\Sigma \mathbf{v}_3 = \lambda_3 \mathbf{v}_3 \Rightarrow \begin{bmatrix} \mathbf{x}_1^T \mathbf{x}_1 & \mathbf{x}_1^T \mathbf{x}_2 & \mathbf{x}_1^T \mathbf{y} \\ \mathbf{x}_2^T \mathbf{x}_1 & \mathbf{x}_2^T \mathbf{x}_2 & \mathbf{x}_2^T \mathbf{y} \\ \mathbf{y}^T \mathbf{x}_1 & \mathbf{y}^T \mathbf{x}_2 & \mathbf{y}^T \mathbf{y} \end{bmatrix} \mathbf{n} = \lambda_3 \mathbf{n} \quad (35)$$

整理得到：

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

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

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

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

根据 (37) 最后一行，可以求得  $k$

$$k = \frac{-1}{v_{3,3}} \quad (38)$$

$b_1$  和  $b_2$  构成的列向量为：

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

回归方程常数项通过下式获得：

$$b_0 = E(\mathbf{y}) - [E(\mathbf{x}_1) \ E(\mathbf{x}_2)] \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad (40)$$

为了方便多元正交回归运算，令：

$$\begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} = [\mathbf{X}] \Rightarrow \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{y} \end{bmatrix} = [\mathbf{X} \ \mathbf{y}] \quad (41)$$

协方差矩阵  $\Sigma$  为：

$$\Sigma = \begin{bmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{y} \\ \mathbf{y}^T \mathbf{X} & \mathbf{y}^T \mathbf{y} \end{bmatrix} \quad (42)$$

上式  $\Sigma$  也不考虑系数  $1/(n-1)$ ：

$$\Sigma \mathbf{v}_3 = \lambda_3 \mathbf{v}_3 \Rightarrow \begin{bmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{y} \\ \mathbf{y}^T \mathbf{X} & \mathbf{y}^T \mathbf{y} \end{bmatrix} \mathbf{n} = \lambda_3 \mathbf{n} \quad (43)$$

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

$$\mathbf{n} = \begin{bmatrix} b_1 \\ b_2 \\ -1 \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ -1 \end{bmatrix} \quad (44)$$

将 (44) 代入 (43)，整理得到  $\mathbf{b}$ ：

$$\begin{bmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{y} \\ \mathbf{y}^T \mathbf{X} & \mathbf{y}^T \mathbf{y} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ -1 \end{bmatrix} = \lambda_3 \begin{bmatrix} \mathbf{b} \\ -1 \end{bmatrix} \Rightarrow \mathbf{b} = (\mathbf{X}^T \mathbf{X} - \lambda_3 \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \quad (45)$$

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

图 10 还比较了 OLS 和 TLS 回归结果。值得注意的是，如图 10 上半部分所示，对于最小二乘回归 OLS， $\hat{y}$  在  $x_1$  和  $x_2$  构造的平面上；而如图 10 下半部分，正交回归 TLS 中， $\hat{y}$  在  $v_1$  和  $v_2$  构造平面  $H$  上。

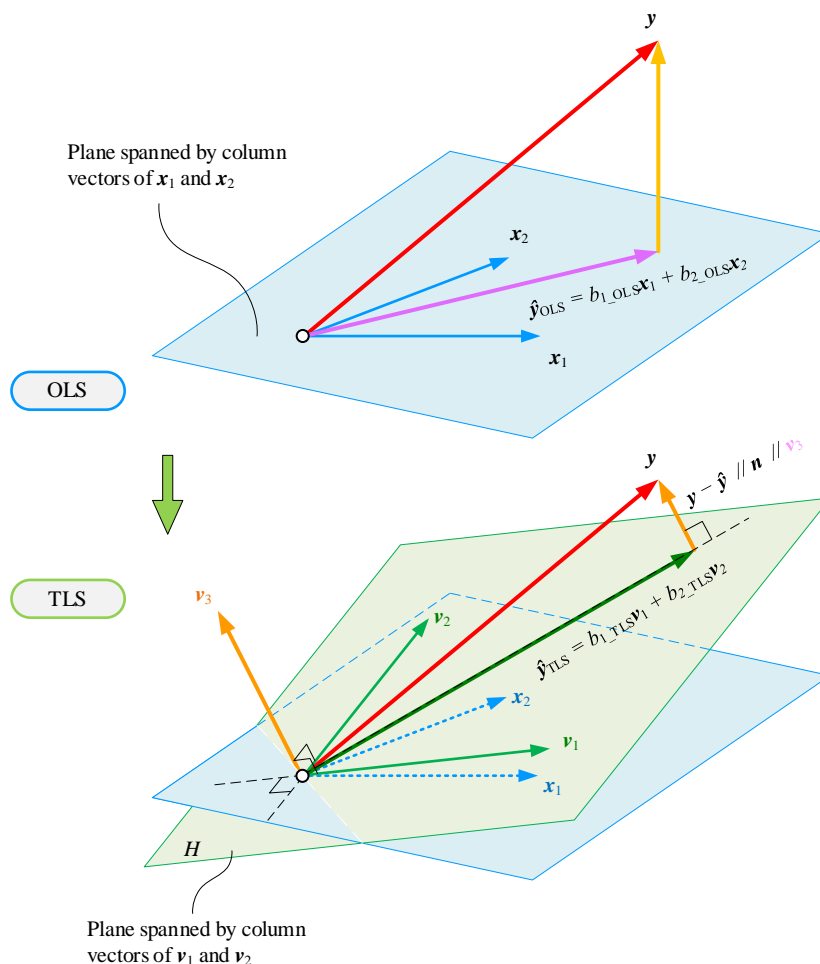
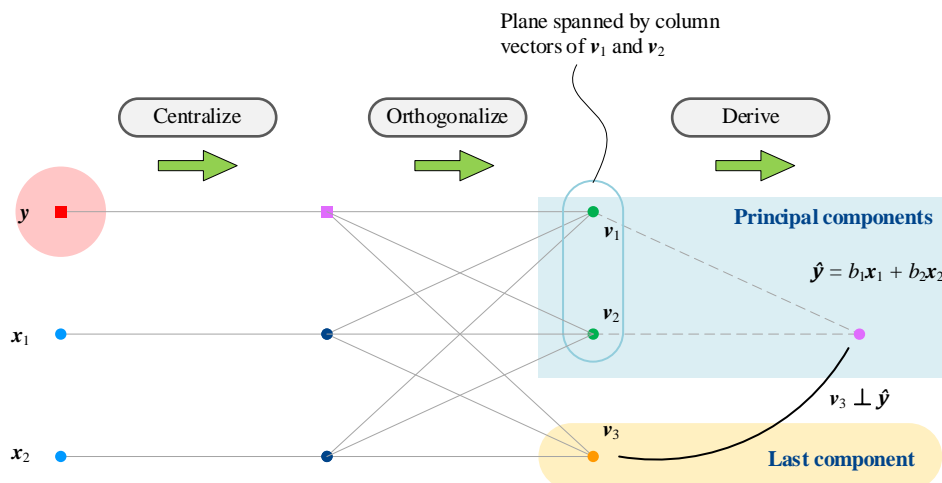


图 10. 几何角度解释二元正交回归坐标转换

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”。



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

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：

$$y = 0.4079x_1 + 0.4438x_2 - 0.00061 \quad (46)$$

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：

$$y = 0.3977x_1 + 0.4096x_2 - 0.006 \quad (47)$$

```
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。

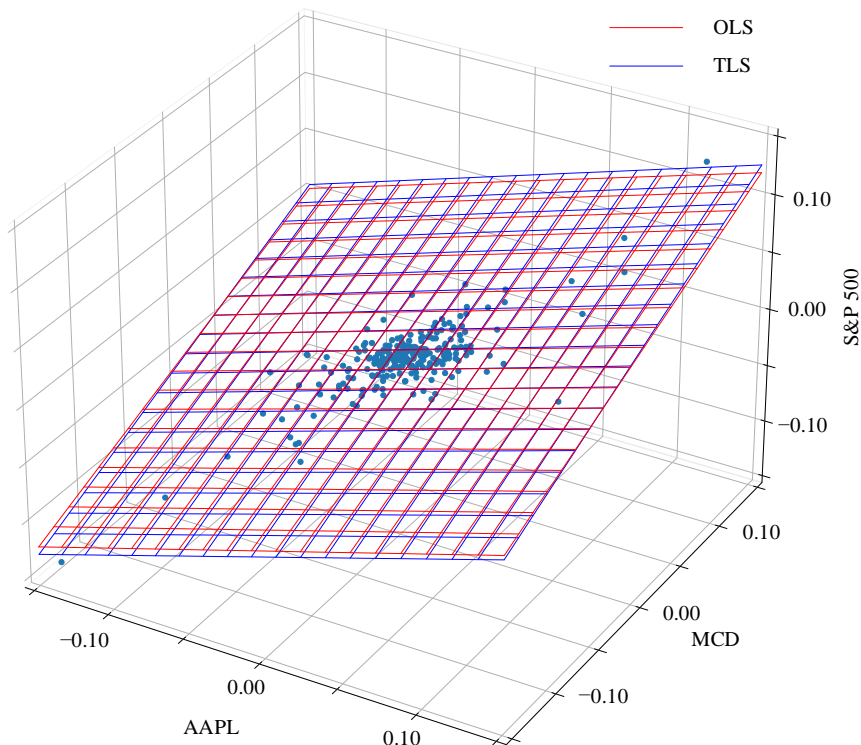


图 13. 比较 OLS 和 TLS 二元回归结果



Bk6\_Ch16\_02.py 完成本节回归运算。

## 16.5 多元正交回归

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

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

为了表达方便，假设  $X$  和  $y$  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：

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

为了进行 PCA 分解，首先计算  $[X, y]$  矩阵协方差矩阵。

$X$  和  $y$  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：

$$\Sigma_{(D+1) \times (D+1)} = [X, y]^T [X, y] = \begin{bmatrix} X^T \\ y^T \end{bmatrix} \begin{bmatrix} X \\ y \end{bmatrix} = \begin{bmatrix} X^T X & X^T y \\ y^T X & y^T y \end{bmatrix} \quad (50)$$

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：

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)



$$\Sigma = V\Lambda V^{-1} \quad (51)$$

其中,

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \\ & & & & \lambda_{D+1} \end{bmatrix}, \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D \geq \lambda_{D+1} \quad (52)$$

$$V = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_D \quad \mathbf{v}_{D+1}]$$

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。

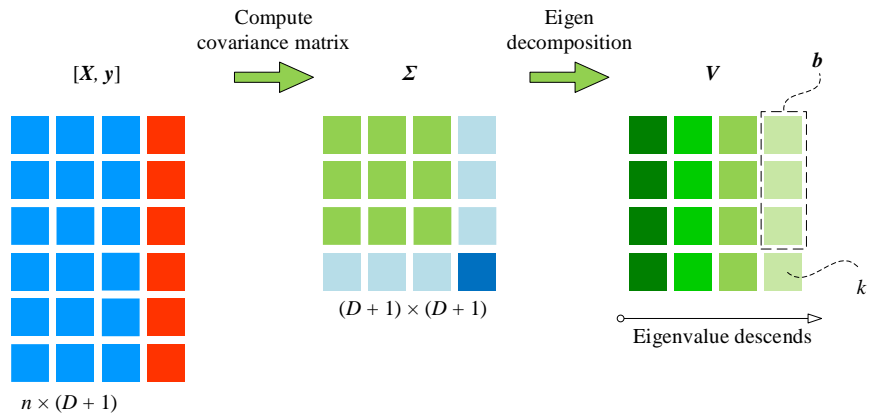


图 14. 多元正交回归矩阵运算过程

$V$  中第 1 到第  $D$  个行向量  $[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_D]$  构造超平面  $H$ , 而  $\mathbf{v}_{D+1}$  垂直于该超平面。

构造  $F(x_1, x_2, \dots, x_D, y)$  函数:

$$F(x_1, x_2, \dots, x_D, y) = b_1 x_1 + b_2 x_2 + \dots + b_{D-1} x_{D-1} + b_D x_D - y \quad (53)$$

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

$$\mathbf{n} = \left( \frac{\partial F}{\partial x_1}, \dots, \frac{\partial F}{\partial x_D}, \frac{\partial F}{\partial y} \right)^T = [b_1 \quad b_2 \quad \dots \quad b_D \quad -1]^T = \begin{bmatrix} \mathbf{b} \\ -1 \end{bmatrix} \quad (54)$$

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

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

求解获得多元正交回归系数列向量  $\mathbf{b}$  解:

$$\begin{bmatrix} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{y} \\ \mathbf{y}^T \mathbf{X} & \mathbf{y}^T \mathbf{y} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ -1 \end{bmatrix} = \lambda_{D+1} \begin{bmatrix} \mathbf{b} \\ -1 \end{bmatrix} \Rightarrow \mathbf{b}_{\text{OLS}} = (\mathbf{X}^T \mathbf{X} - \lambda_{D+1})^{-1} \mathbf{X}^T \mathbf{y} \quad (56)$$

对比多元线性最小二乘系数向量结果：

$$\mathbf{b}_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (57)$$

发现当  $\lambda_{D+1}$  等于 0 时， $\mathbf{y}$  完全被  $\mathbf{X}$  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。

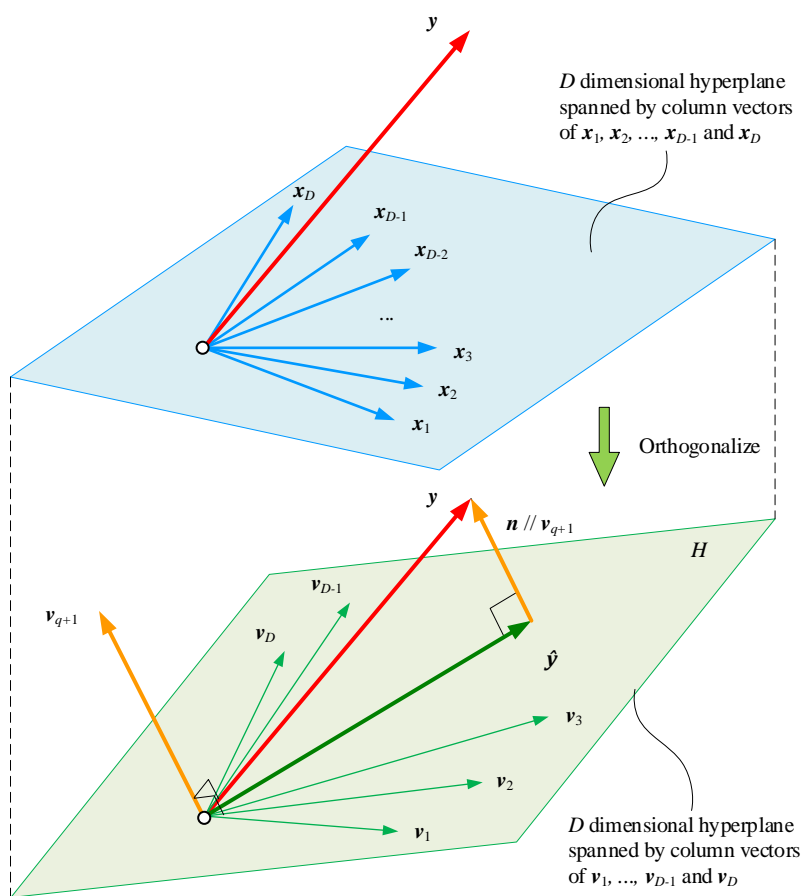


图 15. 几何角度解释多元正交回归

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

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

求解  $k$  得到：

$$k = \frac{-1}{v_{D+1,D+1}} \quad (59)$$

求解  $\mathbf{b}$  得到：

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_D \end{bmatrix} = \frac{-1}{v_{D+1,D+1}} \begin{bmatrix} v_{1,D+1} \\ v_{2,D+1} \\ \vdots \\ v_{D,D+1} \end{bmatrix} \quad (60)$$

$b_0$  通过下式求得。

$$b_0 = E(y) - [E(x_1) \ E(x_2) \ \cdots \ E(x_D)] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_D \end{bmatrix} \quad (61)$$

图 16 展示多元正交回归运算数据关系。看到数据  $[\mathbf{X}, \mathbf{y}]$  均参与到了正交化中；正交化结果为  $D + 1$  个正交向量  $[v_1, v_2, \dots, v_D, v_{D+1}]$ 。通过向量  $v_{D+1}$  垂直  $v_1, v_2, \dots, v_D$  构成超平面，推导出多元正交回归解析式。

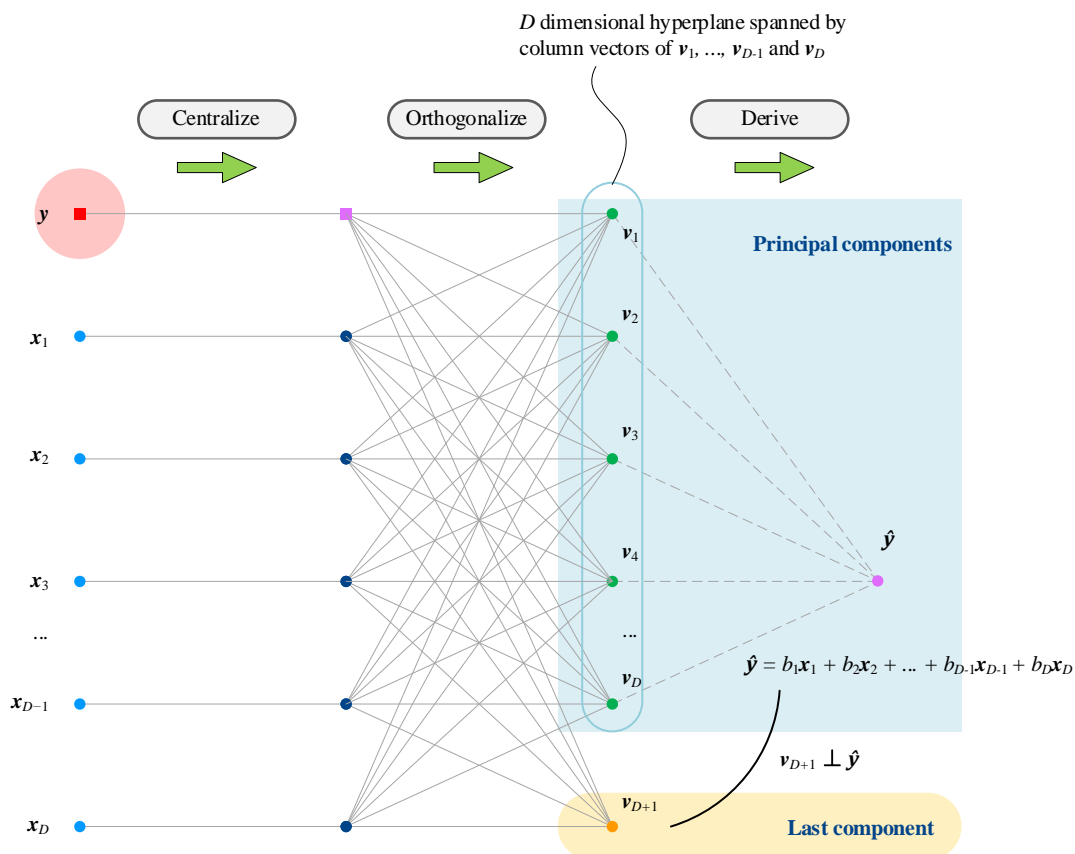


图 16. 多元正交回归运算数据关系

图 17 所示直方图，比较多元 TLS 回归和多元 OLS 回归系数。

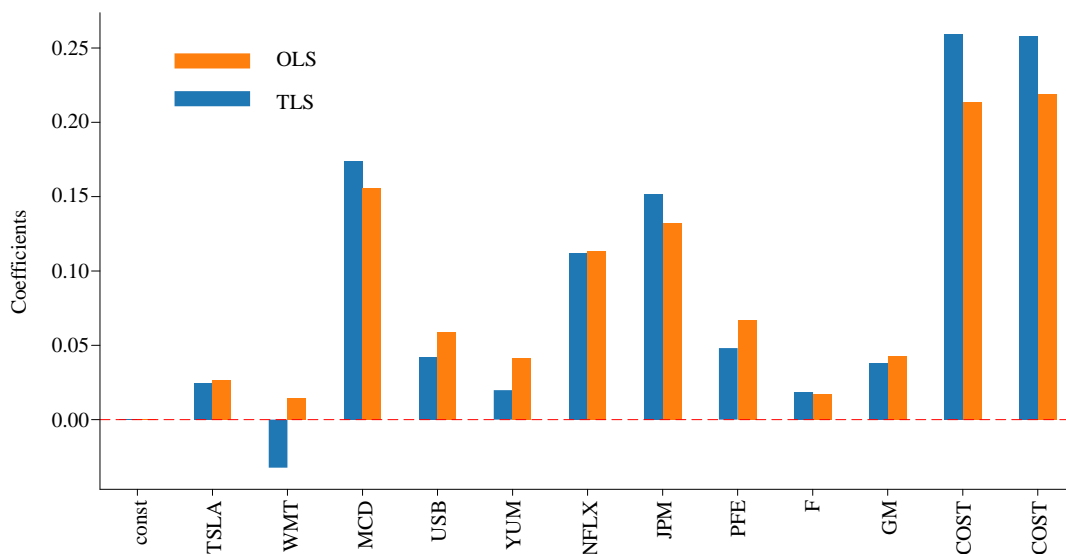


图 17. 比较多元 TLS 回归和多元 OLS 回归系数



Bk6\_Ch16\_03.py 完成本节回归运算。



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。