

回归分析的今生

偏最小二乘回归 (*PLS*)

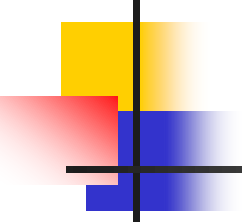
授课教师: 赵春晖

联系方式:

Email: chhzhao@zju.edu.cn

Phone: 13588312064

Room: 工控新楼308室

- 
-
- 偏最小二乘回归
 - 原理
 - 回归模型



偏最小二乘法(parital least squares) 简介

- PLS最先产生于化学领域，在利用分光镜来预测化学样本的组成时，作为解释变量的**红外区反射光谱的波长常有成百上千**，往往超过化学样本的个数，所造成的多重相关性使得人们很难利用传统的最小二乘法。
- 基于这个应用的需要，**1983**年首次提出了**PLS**回归方法并首先在化工领域取得了广泛的应用。



偏最小二乘法的基本原理

设有包含 p 个自变量 n 个样本点的数据表 $X = \{x_1, x_2, \dots, x_p\}_{n \times p}$ 和包含 q 个因变量 n 个样本点的数据表 $Y = \{y_1, y_2, \dots, y_q\}_{n \times q}$ 。偏最小二乘回归分别在 X 与 Y 中提取出成分 t_1 和 u_1 。在提取这两个成分时， t_1 和 u_1 必须满足下面两个条件：

- (1) t_1 和 u_1 应尽可能多地携带它们各自数据表中的信息；
- (2) t_1 和 u_1 的相关程度能够达到最大。

这两个要求表明： t_1 和 u_1 应尽可能好地代表数据表 X 和 Y ，同时，自变量的成分 t_1 对因变量的成分 u_1 又有很强的解释能力



偏最小二乘法的算法推导

为了数据推导方便，首先将数据做标准化处理， X ， Y 经过标准化处理的数据矩阵分别记为 $E_0 = \{E_{01}, E_{02}, \dots, E_{0p}\}_{n \times p}$ ，和 $F_0 = \{F_{01}, F_{02}, \dots, F_{0q}\}_{n \times q}$ 。

数据的标准化处理是指对数据进行中心化—压缩处理，即

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, p)$$

其中， \bar{x}_j 为第 j 个变量的样本均值， s_j 为第 j 个变量的样本标准差。经过这样的变换后，每个变量对应的数据均值都为0，方差都为1。



偏最小二乘法的算法推导

Step1

t_1 : E_0 的第一个成分, $t_1 = E_0 w_1$, w_1 是 E_0 的第一个坐标轴, 并且 $\|w_1\| = 1$ 。

u_1 : F_0 的第一个成分, $u_1 = F_0 c_1$, c_1 是 F_0 的第一个坐标轴, 并且 $\|c_1\| = 1$ 。

如果要 t_1 和 u_1 能分别很好地代表X与Y中的数据变异信息, 根据主成分分析原理, 应该有

$$\text{Var}(t_1) \rightarrow \max$$

$$\text{Var}(u_1) \rightarrow \max$$

由于回归建模的需要, 又要求 t_1 对 u_1 有最大的解释能力, 由典型相关分析的思路, t_1 与 u_1 的相关度应达到最大值, 即

$$r(t_1, u_1) \rightarrow \max$$

综合起来, 在偏最小二乘回归中, 我们要求 t_1 与 u_1 的协方差达到最大, 即

$$\text{Cov}(t_1, u_1) = \sqrt{\text{Var}(t_1)} \sqrt{\text{Var}(u_1)} r(t_1, u_1) \rightarrow \max$$

偏最小二乘法的算法推导

正规的数学表达式应该是求解下列优化问题。

$$\max \langle E_0 w_1, F_0 c_1 \rangle$$

$$s.t. \begin{cases} w_1^T w_1 = 1 \\ c_1^T c_1 = 1 \end{cases}$$



在 $\|w_1\| = 1$ 和 $\|c_1\| = 1$ 的约束条件下，去求 $w_1^T E_0^T F_0 c_1$ 的最大值。

采用拉格朗日算法，记

$$s = w_1^T E_0^T F_0 c_1 - \lambda_1 (w_1^T w_1 - 1) - \lambda_2 (c_1^T c_1 - 1)$$

对于 s 分别求关于 w_1 、 c_1 、 λ_1 、 λ_2 的偏导，并令其为 0，有：

$$\frac{\partial s}{\partial w_1} = E_0^T F_0 c_1 - 2\lambda_1 w_1 = 0,$$

$$\frac{\partial s}{\partial c_1} = F_0^T E_0 w_1 - 2\lambda_2 c_1 = 0,$$

$$\frac{\partial s}{\partial \lambda_1} = -(w_1^T w_1 - 1) = 0$$

$$\frac{\partial s}{\partial \lambda_2} = -(c_1^T c_1 - 1) = 0$$

偏最小二乘法的算法推导

由上述式子可以推出 $2\lambda_1 = 2\lambda_2 = w_1^T E_0^T F_0 c_1 = \langle E_0 w_1, F_0 c_1 \rangle$ 。

记 $\theta_1 = 2\lambda_1$

θ_1 是优化问题的目标函数值，有 $E_0^T F_0 F_0^T E_0 w_1 = \theta_1^2 w_1$ 。
同理可得， $F_0^T E_0 E_0^T F_0 c_1 = \theta_1^2 c_1$ 。

w_1 是矩阵 $E_0^T F_0 F_0^T E_0$ 的**特征向量**，对应的特征值为 θ_1^2 。 θ_1 是目标函数值，它要求取最大值，所以 w_1 是对应于矩阵 $E_0^T F_0 F_0^T E_0$ **最大特征值的单位特征向量**。而另一方面， c_1 是对应于矩阵 $F_0^T E_0 E_0^T F_0$ 最大特征值 θ_1^2 的单位特征向量。

偏最小二乘法的算法推导

然后,分别求 E_0 和 F_0 对 t_1 , u_1 的3个回归方程:

$$\begin{aligned}E_0 &= t_1 p_1^T + E_1 \\F_0 &= u_1 q_1^T + F_1^* \\F_0 &= t_1 r_1^T + F_1\end{aligned}$$

回归系数向量

OLS求解

PCA中的负载系数

$$\begin{aligned}p_1 &= \frac{E_0^T t_1}{\|t_1\|^2} \\q_1 &= \frac{F_0^T u_1}{\|u_1\|^2} \\r_1 &= \frac{F_0^T t_1}{\|t_1\|^2}\end{aligned}$$

E_1 、 F_1^* 、 F_1 分别是3个回归方程的残差矩阵。

偏最小二乘法的算法推导

Step2

用残差矩阵 E_1 和 F_1 取代 E_0 和 F_0 ，求第2个轴 w_2 和 c_2 以及第2个成分 t_2 和 u_2 ，有

$$t_2 = E_1 w_2$$

$$u_2 = F_1 c_2$$

$$\theta_2 = \langle t_2, u_2 \rangle = w_2^T E_1^T F_1 c_2$$

w_2 是对应于矩阵 $E_1^T F_1 F_1^T E_1$ 最大特征值 θ_2^2 的特征向量，
 c_2 是对应于矩阵 $F_1^T E_1 F_1 E_1^T$ 最大特征值的特征向量。

计算回归系数：

$$p_2 = \frac{E_1^T t_2}{\|t_2\|^2}, \quad r_2 = \frac{F_1^T t_2}{\|t_2\|^2}$$

回归方程：


$$E_1 = t_2 p_2^T + E_2$$
$$F_1 = t_2 r_2^T + E_2$$



偏最小二乘法的算法推导

如此计算下去，如果X的秩是A，则会有

$$\begin{aligned}E_0 &= t_1 p_1^T + t_2 p_2^T + \cdots + t_A p_A^T \\F_0 &= t_1 r_1^T + t_2 r_2^T + \cdots + t_A r_A^T\end{aligned}$$



t_1, t_2, \cdots, t_A 均可以表示成 $E_{01}, E_{02}, \cdots, E_{0p}$ 的线性组合

上面的式子还可以还原成 $y_k^* = F_{0k}$ 关于 $x_j^* = E_{0j}$ 的回归方程形式，即

$$y_k^* = a_{k1} x_1^* + a_{k2} x_2^* + \cdots + a_{kp} x_p^* + F_{AK}$$

其中， F_{AK} 是残差矩阵 F_A 的第 k 列。



隐变量个数确定

- 在许多情形下，偏最小二乘回归方程并不需要选用全部的成分进行回归建模，而是可以像在主成分分析时一样，采用截尾的方式选择前 m 个成分($m < A$, $A = \text{秩}(X)$)，仅用这 m 个成分就可以得到一个预测性能较好的模型。
- 事实上，如果后续的成分已经不能为解释提供更有意义的信息时，采用过多的成分只会破坏对统计趋势的认识，引导错误的预测结论。
- 在偏最小二乘回归建模中，究竟应该选取多少个成分为宜，这可通过考察增加一个新的成分后，能否对模型的预测功能有明显的改进来考虑。



偏最小二乘法应用广泛的原因

- PLS方法与多元线性回归、主成分回归等分析方法比较，求得的模型的残差平方和(SSE)差别不大，但预报残差平方和较小(PRESS)，因而具有较高的预报稳定性。
- PLS方法比较适合用于处理变量多而样本数又少的问题，是一种高效地抽提信息的方法。

SSE: *Sum of Squares for Error*

PRESS: *predicted residual sum of squares*



PLS的性质

- 得分向量之间相互正交 $t_i' t_j = 0$
- 权重系数之间互相正交 $w_i' w_j = 0$
- 当 $l > h$, $t_h' E_l = 0$
- 当 $l = h$, $p_h' w_l = 1$; 当 $l > h$, $p_h' w_l = 0$
- 当 $l > h$, $w_h' E_l' = 0$



PLS

- $T = XR$
- $R = W^*(P'W)^{-1}$



一种观点

- 由对X和对Y的两个主成分分析步骤和一个回归步骤组成

这样互相独立求出的 u 与 t ，与PCR法没有本质区别，为了使由Y得出的 u 能与X得出的 t 之间有良好的线性关系，可以让Y分解为U时引入有关T的信息，或从X分解出T时引入U的信息，将上述两个独立的分解过程合而为一，得到PLS解。

PLS vs PCA

	PCA	PLS
分析对象	X	X 和 Y
目标函数	$\max < Xw_1, Xw_1 >$ $s.t \quad w_1^r w_1 = 1$	$\max < Xw_1, Yc_1 >$ $s.t \begin{cases} w_1^r w_1 = 1 \\ c_1^r c_1 = 1 \end{cases}$
解	$X^r X w_1 = \lambda w_1$	$X^r Y Y^r X w_1 = \lambda w_1$
潜变量	$t = Xw$	$t = Xw$ $u = Yc$
关系公式	$X = TW^r + E$	$X = TP^r + E$ $Y = TQ^r + F$



OLS与PLS的比较

	OLS	PLS
样本量	远大于自变量个数	可以少于自变量个数
自变量	线性无关	可以存在多重共线性
因变量个数	单个	多个

- OLS: 直接实施对X和Y的回归
- PLS: 对X提取m个成分, 通过实时对X的回归, 表达成关于原变量Y的回归方程



小结

- 偏最小二乘法不仅将响应矩阵 Y 进行分解提取主因子，还将自变量矩阵 X 进行分解提取主因子，因而只有更强的提供信息的能力。
- PLS方法在建立模型时，又不同于简单的将 X 和 Y 分别作主成分分解。而是在将 X 分解为 T 的时候引入有关 U 的信息，在将 Y 分解为 U 的时候引入 T 的相关信息，这样做可以保证 T 和 U 之间具有良好的线性关系。



小结

- 偏最小二乘回归是一种多因变量对多自变量的回归建模方法。特别当各变量集合内部存在较高度度的相关性时，用偏最小二乘回归进行回归建模分析，比对逐个因变量做多元回归更加有效，其结论更加可靠，整体性更强。
- 在建模的同时实现了数据结构的简化,可以在二维平面上对多维数据的特性进行观察,图形功能强大。
- 因此,许多统计分析专家称PLS为第二代回归分析方法。

偏最小二乘法应用之一

软测量工具箱(*Soft Sensor Analytics Toolbox*)

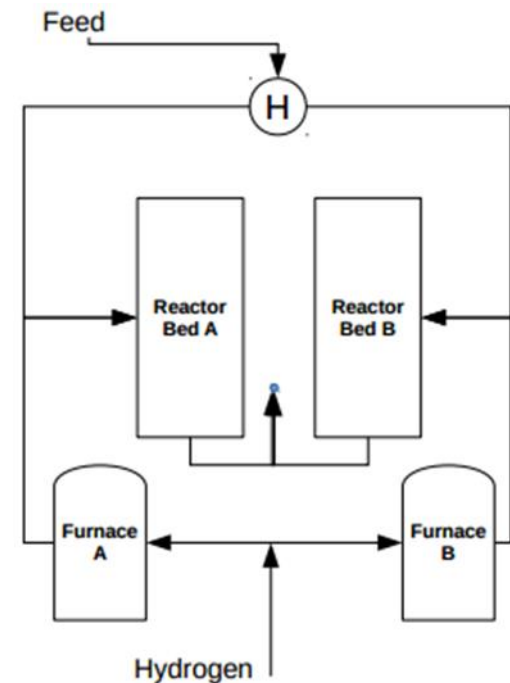


偏最小二乘法应用之一

软测量工具箱(*Soft Sensor Analytics Toolbox*)

利用软测量工具箱中的偏最小二乘回归来对从某一化学反应过程中获取的数据（已进行标准化处理）进行分析，建立软传感器来准确估计相关常规变量，如温度，压力，流量等，与产品（**AL201.EL**）的关系。数据储存在 Data1.xlsx。

Synbol	Description	Avaliability
T151	Feed Temperature	Real-time
F151	Feed flowrate A	Real-time
DIC151	Feed density	Real-time
FI151	Feed flowrate B	Real-time
TI1511	Reactor A Temperature 1	Real-time
TI1512	Reactor A Temperature 2	Real-time
TI1513	Reactor A Temperature 3	Real-time
TI1514	Reactor A Temperature 4	Real-time
TI1515	Reactor B Temperature 1	Real-time
TI1516	Reactor B Temperature 2	Real-time
TI1517	Reactor B Temperature 3	Real-time
TI1518	Reactor B Temperature 4	Real-time
AIC101	Feed chemical content	2 hours
AI201.EL	Product chemical content	6 hours



偏最小二乘法应用之一

软测量工具箱(Soft Sensor Analytics Toolbox)

软测量工具箱的主界面

选择数据预处理的方式

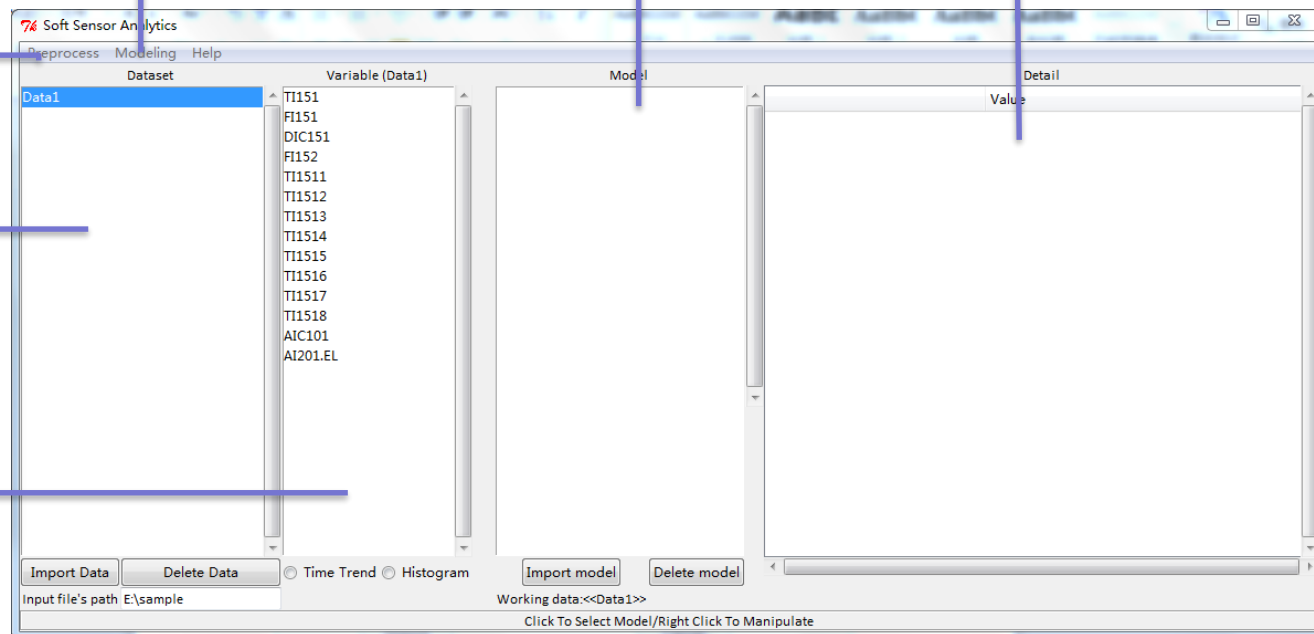
显示所建立的模型

选择建立模型的方式

显示模型的具体信息

显示导入的数据集以及经过预处理后的数据集

显示数据集中包含的变量





偏最小二乘法应用之一

软测量工具箱(*Soft Sensor Analytics Toolbox*)

- 导入需要分析的数据 *Data1.xlsx*
- 对数据进行预处理

(1) 剔除缺失值:

在数据测量过程中，通常会出现一些缺失值，因此要将这些缺失值剔除。

(2) 下采样处理:

在实验室中获取的数据变化缓慢，因此需要进行下采样处理，即对于产品（**AL201.EL**）的样值序列每隔几个样值取样一次。

(3) 分离数据:

将经过上述处理数据样本分为两个部分，一部分作为训练集，用来建立**PLS**模型，另一部分作为检验集，用来检验所建模型的预测效果。

偏最小二乘法应用之一

软测量工具箱(Soft Sensor Analytics Toolbox)

(4) 相关性分析:



由图可知，每个变量之间高度相关，所以选择偏最小二乘法 (PLS) 来建立回归模型。偏最小二乘法通常用于处理具有共线性和高维度的数据。

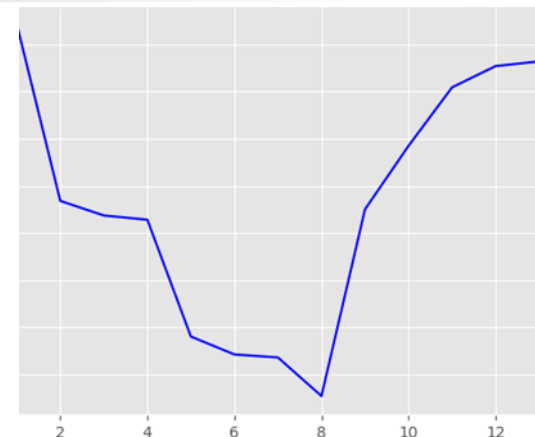
偏最小二乘法应用之一

软测量工具箱 (Soft Sensor Analytics Toolbox)

■ 建立模型

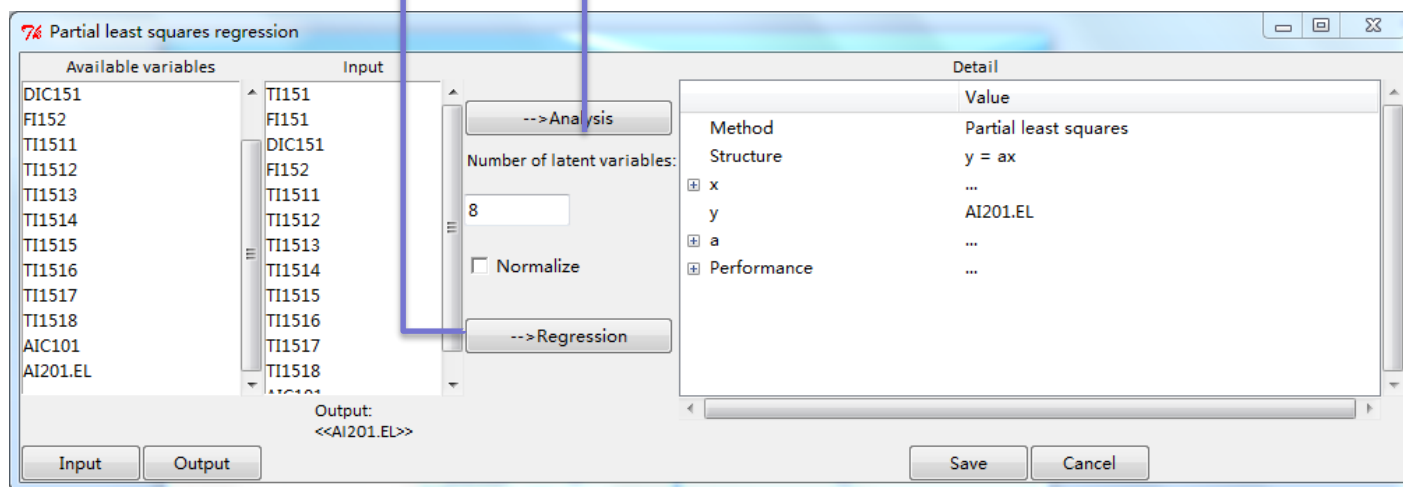
(1) 确定偏最小二乘法特征向量个数:

对所有潜在变量进行10折交叉验证, 从而确定偏最小二乘法中的特征向量个数。由右图可知, 特征向量个数的最佳选择是“8”。



回归分析

确定偏最小二乘法特征向量个数

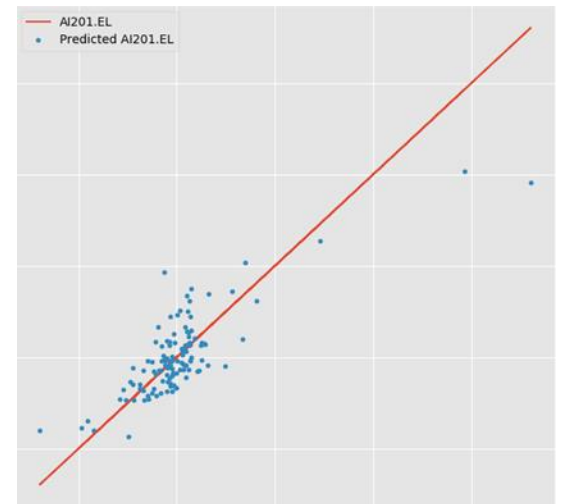
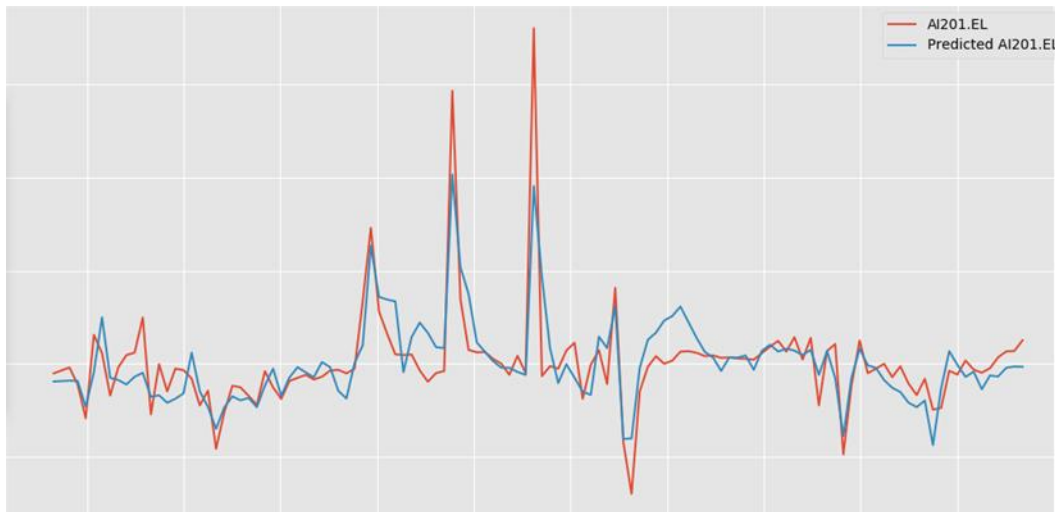


偏最小二乘法应用之一

软测量工具箱(*Soft Sensor Analytics Toolbox*)

(2) 回归分析:

利用偏最小二乘法建立回归模型并生成散点图和时间趋势图。



偏最小二乘法应用之二

水文相关分析(Related Analysis of Hydrology)

组号	H_{\max}/m	$Q_{\max}/(\text{m}^3 \cdot \text{s}^{-1})$	$H_{\text{起涨}}/\text{m}$	$P_{\text{区}}/\text{mm}$	$P_{\text{下}}/\text{mm}$
1	28.8	5485	27.19	204	134
2	26.67	1725	26.46	66	61
3	27.53	2115	26.41	192	213
4	27.45	2308	23.94	84	71
5	26.61	1688	25.74	41	19
6	28.45	4310	26.92	249	198
7	27.32	2115	26.74	75	70
8	28.34	3690	27.86	117	121
9	28.04	3630	25.23	123	143
10	27.85	2910	26.90	80	71
11	27.95	2940	26.12	125	127
12	27.90	3330	26.44	66	69
13	27.00	2250	26.63	32	18
14	26.78	1750	26.36	36	39
15	27.64	2406	26.06	99	102
16	28.16	3358	26.00	169	126
17	26.46	1644	20.27	119	44
18	27.06	2174	23.92	11	12
19	26.85	1706	25.96	150	113
20	27.59	2532	25.78	91	94
21	28.74	4440	28.11	186	276
22	27.93	2387	24.24	197	171
23	27.32	2010	25.06	77	74

利用偏最小二乘法对淮河王家坝水文站最高洪水水位 H_{\max} 与各因子进行相关分析。该站水位流量关系呈绳套形而极为复杂，最高洪水水位受洪峰流量、起涨水位、站以上区间平均雨量和站以下几个雨量站的平均雨量的综合影响。现以洪峰流量 Q_{\max} 、起涨水位 $H_{\text{起涨}}$ 、站以上区间平均雨量 $P_{\text{区}}$ 和站以下几个雨量站的平均降雨量 $P_{\text{下}}$ 作为自变量与最高洪水水位 H_{\max} 建立模型。基本数据资料共23组，如表格所示。

偏最小二乘法应用之二

水文相关分析(Related Analysis of Hydrology)

■ 计算变量相关系数

	Q_{max}	H_{qz}	P_q	P_x	H_{max}
Q_{max}	1.000	0.431	0.668	0.581	0.928
H_{qz}		1.000	0.170	0.346	0.520
P_q			1.000	0.901	0.691
P_x				1.000	0.715
H_{max}					1.000

由上表可知，自变量之间存在明显的相互关系，用一般最小二乘法建立模型。会导致结构不合理、成果不稳定，因此不用一般最小二乘回归方法，而采用偏最小二乘回归方法。

偏最小二乘法应用之二

水文相关分析(Related Analysis of Hydrology)

■ 建立模型

用前18组作拟合，后5组作检验，按照偏最小二乘回归的算法，思路和公式建立模型，得到的回归方程为：

$$H'_{max} = 0.0005Q_{max} + 0.0491H_{qz} + 0.0005P_q + 0.0019P_x + 24.68 \quad (1)$$

为了进行对比分析，同样用前18组作拟合，后5组作检验，按照最小二乘回归的算法，思路和公式建立模型，得到的回归方程为：

$$H''_{max} = 0.0006Q_{max} + 0.0108H_{qz} - 0.0041P_q + 0.0065P_x + 24.68 \quad (2)$$

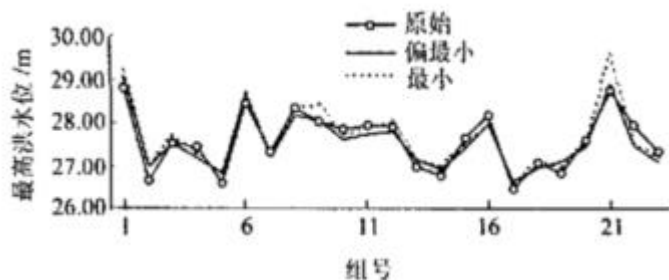
由式(1)和式(2)可知，较之最小二乘回归模型偏最小二乘回归模型的结构符合水文现象的变化特征。因为最高洪水位 H_{max} 与洪峰流量 Q_{max} 、起涨水位 H_{qz} 、站以上区间平均雨量 P_q 和站以下几个雨量站的平均降雨量 P_x 都是正相关的，而最小二乘回归方法给出的站以上区间平均雨量 P_q 的回归系数 -0.0041 根本不符合物理意义。

偏最小二乘法应用之二

水文相关分析(Related Analysis of Hydrology)

■ 检验模型

淮河王家坝水文站洪水水位计算成果表



结合图表可知，从总体上看，偏最小二乘回归方法建立的模型，其预测精度优于一般最小二乘回归方法建立的模型。

组号	H_{max}/m	$Q_{\text{max}}/(\text{m}^3 \cdot \text{s}^{-1})$	H_{max}/m	P_{H}/mm	P_{Q}/mm	M_{max}/m	误差 1/m	M_{max}/m	误差 1/m
1	28.8	5485	27.19	204	134	29.11	0.31	29.18	0.38
2	26.67	1725	26.46	66	61	26.99	0.32	27.10	0.43
3	27.53	2115	26.41	192	213	27.53	0.00	27.71	0.18
4	27.45	2308	25.94	84	71	27.18	-0.27	27.32	-0.13
5	26.61	1688	25.74	41	19	26.84	0.23	26.81	0.20
6	28.45	4310	26.92	249	198	28.65	0.20	28.70	0.25
7	27.32	2115	26.74	75	70	27.22	-0.10	27.27	-0.05
8	28.34	3690	27.86	117	121	28.18	-0.16	28.38	0.04
9	28.04	3630	25.23	123	143	28.96	-0.03	28.44	0.40
10	27.85	2910	26.90	80	71	27.63	-0.22	27.73	-0.12
11	27.95	2940	26.12	125	127	27.73	-0.22	27.92	-0.03
12	27.90	3330	26.44	66	69	27.80	-0.10	28.02	0.12
13	27.00	2250	26.63	32	18	27.16	0.16	27.18	0.18
14	26.78	1750	26.36	36	39	26.94	0.16	27.00	0.22
15	27.64	2406	26.06	99	102	27.40	-0.24	27.54	-0.10
16	28.16	3358	26.00	169	128	27.96	-0.20	27.98	-0.02
17	26.46	1644	20.27	119	44	26.64	0.18	26.56	0.10
18	27.06	2174	27.92	11	12	26.97	-0.09	27.16	0.19
19	26.85	1706	25.96	150	113	27.10	0.25	26.98	0.13
20	27.59	2532	25.78	91	94	27.43	-0.16	27.40	0.03
21	28.74	4440	28.11	186	276	28.89	0.15	29.56	0.67
22	27.93	2387	24.24	197	171	27.49	-0.44	27.56	-0.37
23	27.32	3010	25.06	77	74	27.09	-0.23	27.30	-0.12
拟合阶段平均误差							0.17		0.17
检验阶段平均误差							0.19		0.20

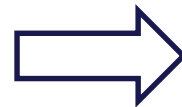
偏最小二乘法可以做什么？

几个案例——生物医学



偏最小二乘法在基因微阵列数据分析中的应用

随着生物信息学的发展，高通量技术的出现，在21世纪初PLS引入基因微阵列数据的分析。目前PLS在微阵列数据分析中主要有四个方面的应用。



差异表达基因的选取



差异表达基因的分类
(判别)



差异表达基因的生存
分析



基因调控网络的构建

偏最小二乘法可以做什么？

几个案例——工业技术

偏最小二乘法在电厂监测数据预测中的应用



数据挖掘技术在火电厂中的应用已成为国内外的研究热点，但是在建立数据仓库和数据挖掘时，往往会遇到由于**仪表故障而导致的测量数据的缺失和失真**，这给数据挖掘过程带来了许多困难，甚至会造成灾难性后果。

为了解决上述问题，提出了利用现场实测数据**建立电厂实时监测数据预测模型**。基于电厂监测参数间普遍存在相关性的特点，提出了利用偏最小二乘回归方法建模，能有效地克服自变量间的多重相关性问题，计算结果更为可靠。

回归家族





线性回归复习

- 最小二乘
- 岭回归
- **Lasso**回归，弹性网回归（同时实现变量选择+回归）
- **PCR**
- **PLS**
- 关键变量选择； 关键潜变量选择；



非线性回归拓展

- 回归模型的因变量是自变量的一次以上函数形式，回归规律在图形上表现为形态各异的各种曲线。
- ◆ **可线性化处理的：**通过变量变换，将非线性回归化为线性回归，然后用线性回归方法处理。
- ◆ **不可线性化的：**映射到高维核空间做线性回归（KPLS等）；函数曲线法（逻辑回归、神经网络等）



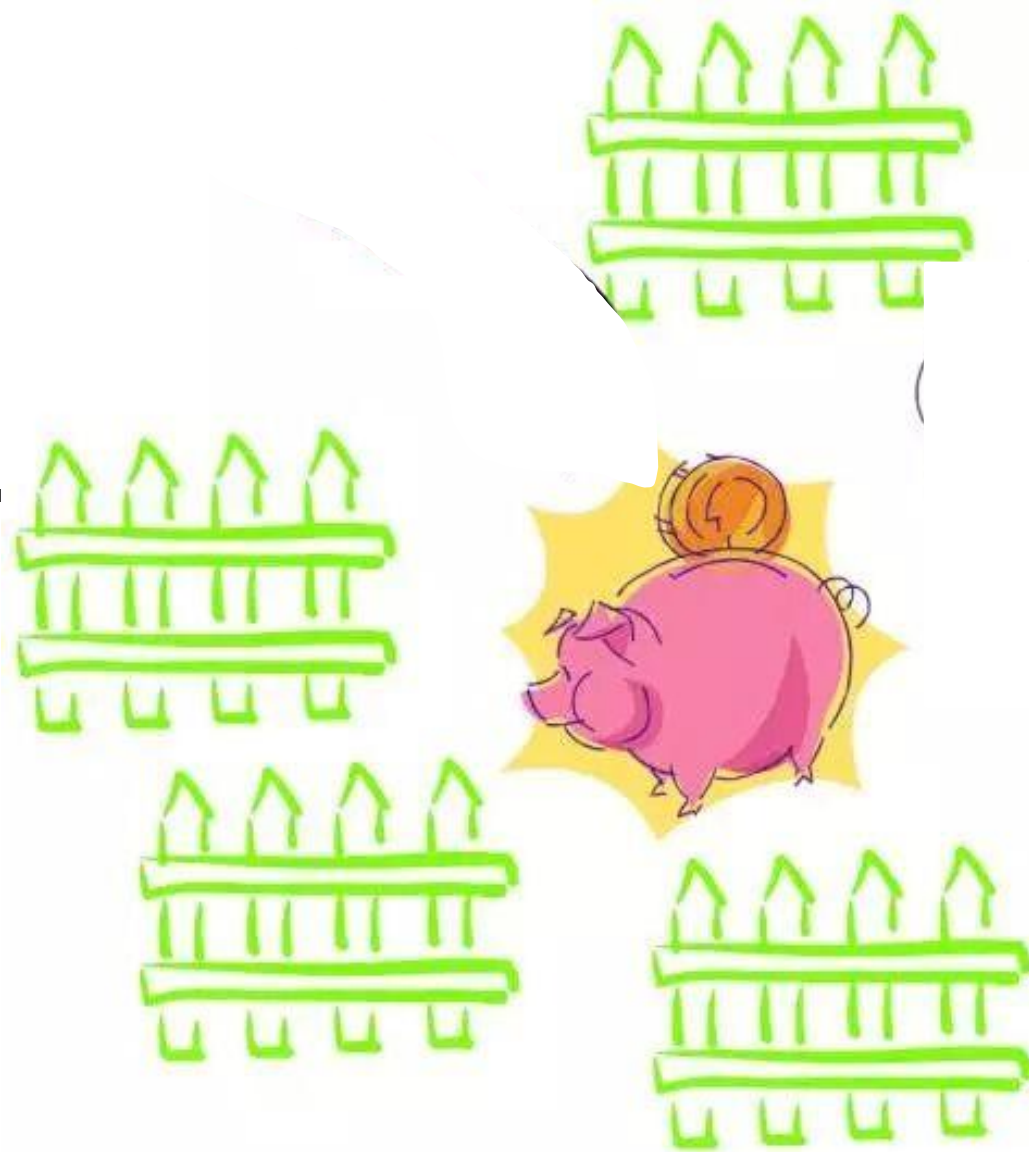
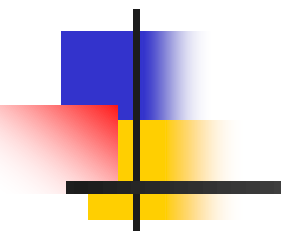
回归分析的用途

- 变量指标未来趋势的**预测**
- 不可直接测量的关键指标的**软测量**
- 关键**影响**因素的识别
- 通过自变量来实现对因变量的**调控**



End

回归分析的耿直



回归分析的耿直与欺骗

