



# 系统工程导论

开课单位：清华大学自动化系

主讲教师：胡坚明 副教授



清华大学  
Tsinghua University

# 模块二： 系统分析

## 主成分分析方法

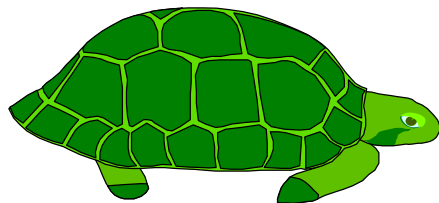
## 内容和重点

### ■本章内容

- 主成分分析基本原理
- 主成分分析计算方法
- 主成分在数据压缩中的作用
- 主成分在线性回归中的作用

## ■引言

**例：**一批龟壳化石的长、宽、高数据，请将全部乌龟分为三类



样本	长	宽	高
$t$	$x_1(t)$	$x_2(t)$	$x_3(t)$
1	93	74	37
2	94	78	35
3	96	80	35
4	101	84	39
5	102	85	38
6	103	81	37
7	104	83	39
8	106	83	39
9	107	82	38
10	112	89	40
11	113	88	40
12	114	86	40

## ■引言

观察：长、宽、高之间近似存在线性关系

若存在一个线性关系

或者 
$$x_1(t) \approx b_0(1) + b_2(1)x_2(t) + b_3(1)x_3(t)$$

或者 
$$x_2(t) \approx b_0(2) + b_1(2)x_1(t) + b_3(2)x_3(t)$$

或者 
$$x_3(t) \approx b_0(3) + b_1(3)x_1(t) + b_2(3)x_2(t)$$

只需要根据两个变量分类！

## ■引言

若存在两个线性关系

或者

$$\begin{aligned}x_1(t) &\approx c_0(1) + c_3(1)x_3(t) \\x_2(t) &\approx c_0(2) + c_3(2)x_3(t)\end{aligned}$$

或者

$$\begin{aligned}x_1(t) &\approx c_0(1) + c_2(1)x_2(t) \\x_3(t) &\approx c_0(3) + c_2(3)x_2(t)\end{aligned}$$

或者

$$\begin{aligned}x_2(t) &\approx c_0(2) + c_1(2)x_1(t) \\x_3(t) &\approx c_0(3) + c_1(3)x_1(t)\end{aligned}$$

只需要根据一个变量分类

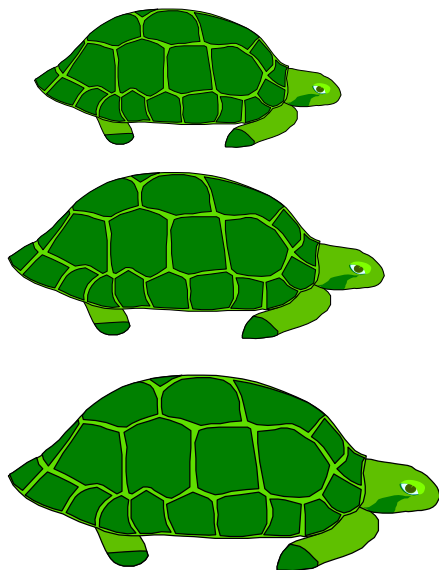
# 主成分分析方法

## ■引言

据长  
分类

据宽  
分类

据高  
分类



样本	长	宽	高
$t$	$x_1(t)$	$x_2(t)$	$x_3(t)$
1	93	74	37
2	94	78	35
3	96	80	35
4	101	84	39
5	102	85	38
6	103	81	37
7	104	83	39
8	106	83	39
9	107	82	38
10	112	89	40
11	113	88	40
12	114	86	40

根据什么变量分类较好?

## ■引言

$$x_1(t) \approx c_0(1) + c_3(1)x_3(t)$$

$$x_2(t) \approx c_0(2) + c_3(2)x_3(t)$$

任取  $y_1(t) = l_0(1) + l_1(1)x_1(t) + l_2(1)x_2(t) + l_3(1)x_3(t)$

只要有关行向量线性无关，就成立

$$x_1(t) \approx p_0(1) + p_1(1)y_1(t)$$

$$x_2(t) \approx p_0(2) + p_1(2)y_1(t)$$

$$x_3(t) \approx p_0(3) + p_1(3)y_1(t)$$



## ■引言

一般性建模问题

$$x = [x_1 \ x_2 \ \cdots \ x_n]^T \longrightarrow \boxed{f = ?} \longrightarrow y$$

如果变量间近似线性相关, 则存在低维向量

$$z = [z_1 \ z_2 \ \cdots \ z_m]^T, m < n \text{ 和 } B \in R^{n \times m} \text{ 使得}$$

$$x \approx Bz, \text{ 于是 } y \approx f(x) \approx f(Bz) = g(z).$$

所以, 一旦知道  $z$  的样本数据, 可考虑低维问题

$$z = [z_1 \ z_2 \ \cdots \ z_m]^T \longrightarrow \boxed{g = ?} \longrightarrow y$$

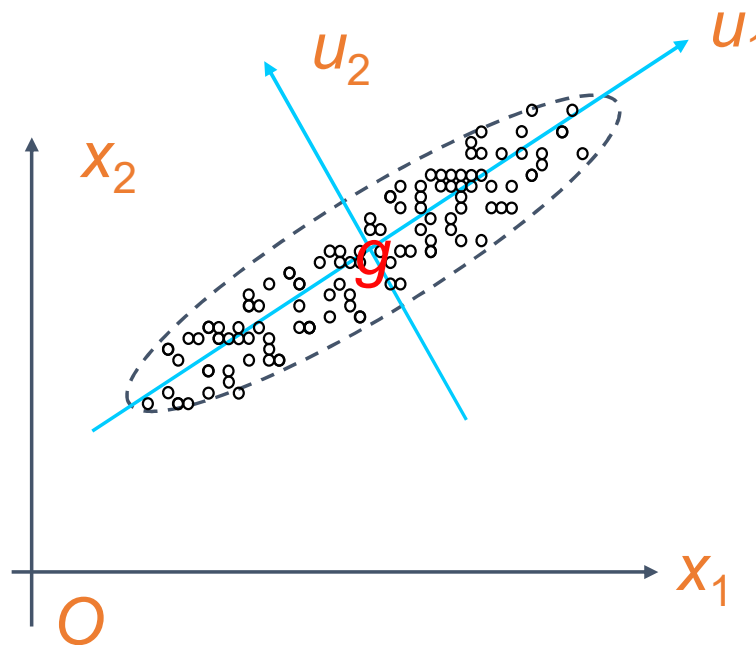
## ■基本原理

- 主成分分析试图从(样本点 $\times$ 变量)的数据表中, 找出最为关键的少数综合变量, 能与原有系统数据保持很高的一致性。
- 实际采用降维方法 (如20维降为2维), **只选择起最主要作用的自变量和因变量来建模。**
- 在数学上, 可以看成将**坐标做平移和旋转变换**, 使得新坐标的原点与样本数据群的重心重合, 第一轴 (称为第一主轴 $u_1$ ) 与数据变异最大的方向对应; 第二轴与数据变异次大的方向对应.....依此类推。经有效舍弃后, 主轴 $u_1 u_2 \dots u_p$ 能十分有效地表示原数据的变异情况。

## ■几何意义

平移 + 旋转，使得样本点在第一主轴的方差最大

$$\begin{cases} u_1 = x_1 \cos \theta + x_2 \sin \theta - c_1 \\ u_2 = -x_1 \sin \theta + x_2 \cos \theta - c_2 \end{cases}$$



在研究数据分类问题时：

- ☐ A 数据集中更有利于划分
- ☒ B 数据分散更有利于划分
- ☒ C 数据分散程度用数据方差来描述
- ☐ D 数据分散程度越大，方差越小
- ☒ E 数据分散程度越大，方差越大

提交

## ■基本准则

分类变量的分散程度越大越有利

变量的样本均值

$$e(y_1) = \frac{1}{12} \sum_{t=1}^{12} y_1(t)$$

变量的样本方差

$$\delta^2(y_1) = \frac{1}{12-1} \sum_{t=1}^{12} (y_1(t) - e(y_1))^2$$

变量的分散程度可用其样本方差表示

## ■确定分类变量的基本方式

$$\max \sum_{t=1}^{12} (y_1(t) - e(y_1))^2$$

$$y_1(t) = l_0(1) + l_1(1)x_1(t) + l_2(1)x_2(t) + l_3(1)x_3(t)$$

⇓

$$e(y_1) = l_0(1) + l_1(1)e(x_1) + l_2(1)e(x_2) + l_3(1)e(x_3)$$

⇓

$$\begin{aligned} y_1(t) - e(y_1) &= l_1(1)(x_1(t) - e(x_1)) \\ &\quad + l_2(1)(x_2(t) - e(x_2)) + l_3(1)(x_3(t) - e(x_3)) \end{aligned}$$

## ■辅助措施

对组合变量施加规范化约束

**措施1：**对原始变量的尺度规范化

$$y_1(t) - e(y_1) = \tilde{l}_1(1) \frac{x_1(t) - e(x_1)}{\sqrt{\delta^2(x_1)}} \\ + \tilde{l}_2(1) \frac{x_2(t) - e(x_2)}{\sqrt{\delta^2(x_2)}} + \tilde{l}_3(1) \frac{x_3(t) - e(x_3)}{\sqrt{\delta^2(x_3)}}$$

**措施2：**对组合参数的尺度规范化

$$\left(\tilde{l}_1(1)\right)^2 + \left(\tilde{l}_2(1)\right)^2 + \left(\tilde{l}_3(1)\right)^2 = 1$$

在主成分分析的辅助措施中，我们对\_\_\_\_\_进行尺度规范化。

- ☒ A 原始变量
- ☐ B 样本方差
- ☒ C 组合参数
- ☐ D 相关参数

提交



## ■第一主成分

最终的优化模型

$$\begin{aligned} \max \quad & \sum_{t=1}^{12} (l_1(1)\tilde{x}_1(t) + l_2(1)\tilde{x}_2(t) + l_3(1)\tilde{x}_3(t))^2 \\ \text{s.t.} \quad & (l_1(1))^2 + (l_2(1))^2 + (l_3(1))^2 = 1 \end{aligned}$$

其中  $\tilde{x}_i(t) = \frac{x_i(t) - e(x_i)}{\sqrt{\delta^2(x_i)}}$

该问题的最优解

$$\hat{y}_1(t) = \hat{l}_1(1)\tilde{x}_1(t) + \hat{l}_2(1)\tilde{x}_2(t) + \hat{l}_3(1)\tilde{x}_3(t)$$

就是这组样本数据的第一主成分

因为由线性关系

$$x_1(t) \approx c_0(1) + c_3(1)x_3(t)$$

$$x_2(t) \approx c_0(2) + c_3(2)x_3(t)$$

$$\hat{y}_1(t) = \hat{l}_1(1)\tilde{x}_1(t) + \hat{l}_2(1)\tilde{x}_2(t) + \hat{l}_3(1)\tilde{x}_3(t)$$

仍可得到

$$x_1(t) \approx p_0(1) + p_1(1)\hat{y}_1(t)$$

$$x_2(t) \approx p_0(2) + p_1(2)\hat{y}_1(t)$$

$$x_3(t) \approx p_0(3) + p_1(3)\hat{y}_1(t)$$

在当前情况下用第一主成分分类最有利！

若仅存在一个线性关系，例如

$$x_1(t) \approx b_0(1) + b_2(1)x_2(t) + b_3(1)x_3(t)$$

任取

$$y_1(t) = l_0(1) + l_1(1)x_1(t) + l_2(1)x_2(t) + l_3(1)x_3(t)$$

$$y_2(t) = l_0(2) + l_1(2)x_1(t) + l_2(2)x_2(t) + l_3(2)x_3(t)$$

只要有关行向量线性无关，就成立

$$x_1(t) \approx p_0(1) + p_1(1)y_1(t) + p_2(1)y_2(t)$$

$$x_2(t) \approx p_0(2) + p_1(2)y_1(t) + p_2(2)y_2(t)$$

$$x_3(t) \approx p_0(3) + p_1(3)y_1(t) + p_2(3)y_2(t)$$

类似于前面的讨论，可求解

$$\max \sum_{t=1}^{12} \left( (y_1(t))^2 + (y_2(t))^2 \right)$$

$$s.t. \ y_1(t) = l_1(1)\tilde{x}_1(t) + l_2(1)\tilde{x}_2(t) + l_3(1)\tilde{x}_3(t)$$

$$y_2(t) = l_1(2)\tilde{x}_1(t) + l_2(2)\tilde{x}_2(t) + l_3(2)\tilde{x}_3(t)$$

$$(l_1(1))^2 + (l_2(1))^2 + (l_3(1))^2 = 1$$

$$(l_1(2))^2 + (l_2(2))^2 + (l_3(2))^2 = 1$$

$$\text{措施3: } l_1(1)l_1(2) + l_2(1)l_2(2) + l_3(1)l_3(2) = 0$$

保证两个组合向量线性无关！

## ■最终模型

$$\begin{aligned} \max \quad & \sum_{t=1}^{12} \left( (y_1(t))^2 + (y_2(t))^2 \right) \\ \text{s.t.} \quad & y_1(t) = l_1(1)\tilde{x}_1(t) + l_2(1)\tilde{x}_2(t) + l_3(1)\tilde{x}_3(t) \\ & y_2(t) = l_1(2)\tilde{x}_1(t) + l_2(2)\tilde{x}_2(t) + l_3(2)\tilde{x}_3(t) \\ & (l_1(1))^2 + (l_2(1))^2 + (l_3(1))^2 = 1 \\ & (l_1(2))^2 + (l_2(2))^2 + (l_3(2))^2 = 1 \\ & l_1(1)l_1(2) + l_2(1)l_2(2) + l_3(1)l_3(2) = 0 \end{aligned}$$

最优解就是第一和第二主成分

## ■一般情况

给定一组样本数据：

$$x_i(t), t = 1, 2, \dots, N, i = 1, 2, \dots, n$$

首先求出其规格化的数据：

$$\tilde{x}_i(t) = \frac{x_i(t) - e(x_i)}{\sqrt{\delta^2(x_i)}} \Rightarrow \begin{aligned} e(\tilde{x}_i) &= 0 \\ \delta^2(\tilde{x}_i) &= 1 \end{aligned}$$

## ■一般情况

确定m个主成分的优化模型为

$$\max \sum_{t=1}^N \sum_{k=1}^m (y_k(t))^2$$

$$s.t. \ y_k(t) = \sum_{i=1}^n l_i(k) \tilde{x}_i(t), \quad k = 1, 2, \dots, m \leq n$$

$$\sum_{i=1}^n (l_i(k))^2 = 1, \quad k = 1, 2, \dots, m$$

$$\sum_{i=1}^n l_i(k) l_i(j) = 0, \quad \forall k \neq j$$

## ■一般情况

符号约定:

$$\tilde{x}(t) = \begin{bmatrix} \tilde{x}_1(t) \\ \tilde{x}_2(t) \\ \vdots \\ \tilde{x}_n(t) \end{bmatrix} \quad y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_m(t) \end{bmatrix}$$

$$l(k) = \begin{bmatrix} l_1(k) \\ l_2(k) \\ \vdots \\ l_n(k) \end{bmatrix} \quad \begin{aligned} \tilde{X} &= [\tilde{x}(1) \quad \tilde{x}(2) \quad \cdots \quad \tilde{x}(N)] \\ Y &= [y(1) \quad y(2) \quad \cdots \quad y(N)] \\ L &= [l(1) \quad l(2) \quad \cdots \quad l(m)] \end{aligned}$$



# 主成分分析方法——基本原理



## ■一般情况

$$\begin{aligned} \max \quad & \sum_{t=1}^N \sum_{k=1}^m (y_k(t))^2 \\ \text{s.t.} \quad & y_k(t) = \sum_{i=1}^n l_i(k) \tilde{x}_i(t) \\ & \sum_{i=1}^n (l_i(k))^2 = 1 \\ & \sum_{i=1}^n l_i(k) l_i(j) = 0 \end{aligned} \quad \Rightarrow \quad \begin{aligned} \max \quad & \sum_{k=1}^m \sum_{t=1}^N (y_k(t))^2 \\ \text{s.t.} \quad & y_k(t) = l^T(k) \tilde{x}(t) \\ & L^T L = I_m \end{aligned}$$

## ■一般情况

$$\Rightarrow \max \sum_{k=1}^m \sum_{t=1}^N l^T(k) \tilde{x}(t) \tilde{x}^T(t) l(k)$$
$$s.t. L^T L = I_m$$

因为  $\sum_{t=1}^N \tilde{x}(t) \tilde{x}^T(t) = \tilde{X} \tilde{X}^T$

$$\Rightarrow \max \sum_{k=1}^m l^T(k) \tilde{X} \tilde{X}^T l(k)$$
$$s.t. L^T L = I_m$$

## ■结论

用  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$

表示  $\tilde{X}\tilde{X}^T$  的顺序递减的特征根,

$q(1), q(2), \cdots, q(n)$  是它们对应的

规范化的特征向量, 则所求主成分为

$$\hat{y}_k(t) = q^T(k) \tilde{x}(t), k = 1, 2, \cdots, m$$

## ■有关性质

- 主成分样本均值

$$e(\hat{y}_k) = \sum_{i=1}^n q_i(k) e(\tilde{x}_i) = 0, \forall k$$

- 主成分样本方差

$$\begin{aligned} (N-1)\delta^2(\hat{y}_k) &= \sum_{t=1}^N q^T(k) \tilde{x}(t) \tilde{x}^T(t) q(k) \\ &= q^T(k) \tilde{X} \tilde{X}^T q(k) = \lambda_k q^T(k) q(k) = \lambda_k \end{aligned}$$

$$\Rightarrow \delta^2(\hat{y}_k) = \lambda_k / (N-1), \forall k$$

## ■有关性质

- 主成分的样本方差之和

$$\begin{aligned}(N-1) \sum_{k=1}^n \delta^2(\hat{y}_k) &= \sum_{k=1}^n \sum_{t=1}^N \tilde{x}^T(t) q(k) q^T(k) \tilde{x}(t) \\&= \sum_{t=1}^N \tilde{x}^T(t) \left( \sum_{k=1}^n q(k) q^T(k) \right) \tilde{x}(t) \quad \boxed{Q Q^T = I_n} \\ \sum_{t=1}^N \tilde{x}^T(t) \tilde{x}(t) &= (N-1) \sum_{i=1}^n \delta^2(\tilde{x}_i) = (N-1)n \\ \Rightarrow \sum_{k=1}^n \delta^2(\hat{y}_k) &= n\end{aligned}$$

注意此处k的取值范围是1到n，而不是1到m，怎样理解？

## ■有关性质

- 样本相关矩阵特征根

定义样本相关矩阵

$$\text{因为 } \frac{1}{N-1} \tilde{X} \tilde{X}^T q(k) = \frac{1}{N-1} \lambda_k q(k)$$

所以样本相关矩阵的特征根  $\bar{\lambda}_k(R) = \frac{1}{N-1} \lambda_k$

$$\Rightarrow \sigma^2(\hat{y}_k) = \bar{\lambda}_k(R), \forall k$$

## ■计算方法

- 分类变量的个数选择准则

设定方差阈值

$$0 < \gamma < 1, \quad e.g. \quad \gamma = 0.85$$

选择最小的  $\hat{m}$  作为  $m$ , 满足

$$\frac{\sum_{k=1}^{\hat{m}} \delta^2(\hat{y}_k)}{\sum_{k=1}^n \delta^2(\hat{y}_k)} = \frac{1}{n} \sum_{k=1}^{\hat{m}} \delta^2(\hat{y}_k) \geq \gamma$$

取前  $m$  个主成分为分类变量

## ■计算方法

乌龟数例的计算结果

$$\delta^2(\hat{y}_1) = 2.67$$

$$\delta^2(\hat{y}_2) = 0.22$$

$$\delta^2(\hat{y}_3) = 0.11$$


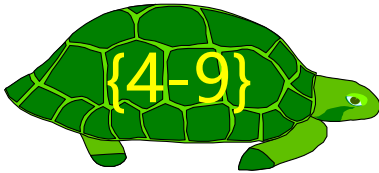
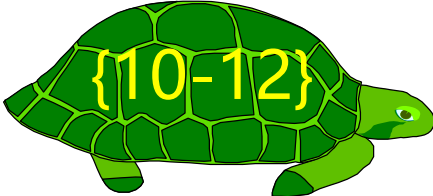
$$\frac{1}{3}\delta^2(\hat{y}_1) = 0.89$$

取第一个主成分为分类变量

$$\hat{y}_1(t) = 0.5900\tilde{x}_1(t) + 0.5731\tilde{x}_2(t) + 0.5687\tilde{x}_3(t)$$



# 主成分分析方法——有关性质

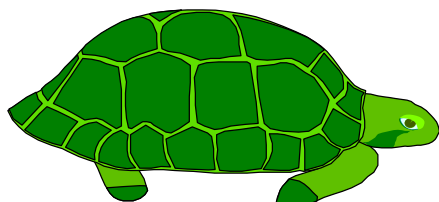
样本	长	宽	高	主成分	分类结果
$t$	$x_1(t)$	$x_2(t)$	$x_3(t)$	$\hat{y}_1(t)$	
1	93	74	37	-2.4310	
2	94	78	35	-2.4410	
3	96	80	35	-2.0023	
4	101	84	39	0.2349	
5	102	85	38	0.1351	
6	103	81	37	-0.6467	
7	104	83	39	0.3475	
8	106	83	39	0.5134	
9	107	82	38	0.1407	
10	112	89	40	2.1487	
11	113	88	40	2.0953	
12	114	86	40	1.9055	

# 主成分分析方法——数据压缩

例：一批龟壳

化石的长

宽高数据



样本	长	宽	高
$t$	$x_1(t)$	$x_2(t)$	$x_3(t)$
1	93	74	37
2	94	78	35
3	96	80	35
4	101	84	39
5	102	85	38
6	103	81	37
7	104	83	39
8	106	83	39
9	107	82	38
10	112	89	40
11	113	88	40
12	114	86	40

如前所述，若存在一个线性关系

或者 
$$x_1(t) \approx b_0(1) + b_2(1)x_2(t) + b_3(1)x_3(t)$$

或者 
$$x_2(t) \approx b_0(2) + b_1(2)x_1(t) + b_3(2)x_3(t)$$

或者 
$$x_3(t) \approx b_0(3) + b_1(3)x_1(t) + b_2(3)x_2(t)$$

总之，可以用两个变量的样本数据近似恢复三个变量的样本数据。

若采用规格化的数据，前面三式可化为

或者  $\tilde{x}_1(t) \approx \tilde{b}_2(1)\tilde{x}_2(t) + \tilde{b}_3(1)\tilde{x}_3(t)$

或者  $\tilde{x}_2(t) \approx \tilde{b}_1(2)\tilde{x}_1(t) + \tilde{b}_3(2)\tilde{x}_3(t)$

或者  $\tilde{x}_3(t) \approx \tilde{b}_1(3)\tilde{x}_1(t) + \tilde{b}_2(3)\tilde{x}_2(t)$

可以用两个变量的规格化数据近似  
恢复三个变量的规格化数据。

更好的做法是极小化逼近误差

$$\sum_{t=1}^{12} \sum_{i=1}^3 \left( \tilde{x}_i(t) - (l_i(1)y_1(t) + l_i(2)y_2(t)) \right)^2$$

确定存储什么数据，其中采用规格化的样本数据是为了平衡不同变量的逼近误差。

有了规格化的样本数据，只要再记住原变量的样本均值和方差，即可恢复原数据。

同样，若存在两个线性关系

或者  $x_1(t) \approx c_0(1) + c_3(1)x_3(t)$   
 $x_2(t) \approx c_0(2) + c_3(2)x_3(t)$

或者  $x_1(t) \approx c_0(1) + c_2(1)x_2(t)$   
 $x_3(t) \approx c_0(3) + c_2(3)x_2(t)$

或者  $x_2(t) \approx c_0(2) + c_1(2)x_1(t)$   
 $x_3(t) \approx c_0(3) + c_1(3)x_1(t)$

总之，可以用一个变量的样本数据近似恢复三个变量的样本数据。

## ■一般情况

给定一组样本数据：

$$x_i(t), t = 1, 2, \dots, N, i = 1, 2, \dots, n$$

首先求出其规格化的数据：

$$\tilde{x}_i(t) = \frac{x_i(t) - e(x_i)}{\sqrt{\delta^2(x_i)}}$$

# 主成分分析方法——数据压缩

## ■一般情况

然后求解优化问题

$$\min \sum_{t=1}^N \sum_{i=1}^n \left( \tilde{x}_i(t) - \sum_{k=1}^m l_i(k) y_k(t) \right)^2$$



## ■一般情况

然后求解优化问题

$$\min \sum_{t=1}^N \sum_{i=1}^n \left( \tilde{x}_i(t) - \sum_{k=1}^m l_i(k) y_k(t) \right)^2$$

可将其写成

$$\min \sum_{t=1}^N (\tilde{x}(t) - Ly(t))^T (\tilde{x}(t) - Ly(t))$$

由于  $L$  和  $y(t)$  均为变量，为使解比较确定，应对它们加一定的限制。

## ■性质

对于数据压缩问题

$$\min \sum_{t=1}^N (\tilde{x}(t) - Ly(t))^T (\tilde{x}(t) - Ly(t))$$

若列向量  
线性相关

$$\begin{aligned} & \Downarrow \\ & \tilde{x}(t) - \sum_{k=1}^m l(k) y_k(t) \\ & \Downarrow \\ & \tilde{x}(t) - \sum_{k=1}^{m-1} \tilde{l}(k) \tilde{y}_k(t) \end{aligned}$$

## ■性质

对于数据压缩问题

$$\min \sum_{t=1}^N (\tilde{x}(t) - Ly(t))^T (\tilde{x}(t) - Ly(t))$$

有如下两个性质：

- ① 假定  $L$  列满秩, 不影响优化效果
- ② 假定  $L$  满足下式也不影响优化效果

$$L^T L = I_m$$

这组样本数据的前  $m$  个主成分就是该问题的一个最优解。

因为当  $L$  列满秩时, 存在可逆矩阵  $G$   
和满足  $P^T P = I_m$  的  $P$ , 使得  $L = PG$

于是 
$$\min \sum_{t=1}^N (\tilde{x}(t) - Ly(t))^T (\tilde{x}(t) - Ly(t))$$

$$\min \sum_{t=1}^N (\tilde{x}(t) - PGy(t))^T (\tilde{x}(t) - PGy(t))$$

数据压缩问题成为

$$\min \sum_{t=1}^N (\tilde{x}(t) - \tilde{L}\tilde{y}(t))^T (\tilde{x}(t) - \tilde{L}\tilde{y}(t))$$

$$s.t. \tilde{L}^T \tilde{L} = I_m$$

$$P = \tilde{L}, Gy(t) = \tilde{y}(t)$$

下面说明，这组样本数据的前  $m$  个主成分就是该问题的一个最优解。

为简化符号，考虑下述优化问题

$$\begin{aligned} \min \sum_{t=1}^N (x(t) - Ly(t))^T (x(t) - Ly(t)) \\ s.t. L^T L = I_m \end{aligned}$$

下面说明，这组样本数据的前  $m$  个主成分就是该问题的一个最优解。

为简化符号，考虑下述优化问题

$$\min \sum_{t=1}^N (x(t) - Ly(t))^T (x(t) - Ly(t))$$

$$s.t. L^T L = I_m$$

先求各  $y(t)$  的最优解

$$\frac{\partial \left( \sum_{t=1}^N (x(t) - Ly(t))^T (x(t) - Ly(t)) \right)}{\partial (y(t))}$$

$$= -2L^T (x(t) - Ly(t))$$

$$\Rightarrow L^T x(t) = L^T L \hat{y}(t) = \hat{y}(t)$$

系统工程导论

将  $y(t)$  的最优解代入目标函数, 可得

$$\begin{aligned} & \sum_{t=1}^N (x(t) - L \hat{y}(t))^T (x(t) - L \hat{y}(t)) \\ &= \sum_{t=1}^N (x(t) - L L^T x(t))^T (x(t) - L L^T x(t)) \\ &= \sum_{t=1}^N x(t)^T (I_n - L L^T)^2 x(t) \end{aligned}$$

注意:  $L \in R^{n \times m}$ , 若  $m < n$   $LL^T = I_n \neq I_m$

因为  $L^T L = I_m$

$$\begin{aligned} \Rightarrow & (I - LL^T)(I - LL^T) \\ &= I - LL^T - LL^T + LL^T LL^T \\ &= I - LL^T \end{aligned}$$

$$\begin{aligned} \Rightarrow & \sum_{t=1}^N x^T(t)(I_n - LL^T)^2 x(t) \\ &= \sum_{t=1}^N x^T(t)x(t) - \sum_{t=1}^N x^T(t)LL^T x(t) \end{aligned}$$



原问题等价于

$$\min \sum_{t=1}^N x^T(t) x(t) - \sum_{t=1}^N x^T(t) L L^T x(t)$$
$$s.t. \quad L^T L = I_m$$

$$\Rightarrow \max \sum_{t=1}^N x^T(t) L L^T x(t)$$
$$s.t. \quad L^T L = I_m$$

由于  $L = [l(1) \quad l(2) \quad \cdots \quad l(m)]$  所以  $L^T x(t) = \begin{bmatrix} l^T(1)x(t) \\ l^T(2)x(t) \\ \vdots \\ l^T(m)x(t) \end{bmatrix}$

$$\Rightarrow x^T(t) L L^T x(t) = \sum_{k=1}^m l^T(k) x(t) x^T(t) l(k)$$

$$\begin{aligned} \Rightarrow \sum_{t=1}^N x^T(t) L L^T x(t) &= \sum_{t=1}^N \sum_{k=1}^m l^T(k) x(t) x^T(t) l(k) \\ &= \sum_{k=1}^m l^T(k) \left( \sum_{t=1}^N x(t) x^T(t) \right) l(k) = \sum_{k=1}^m l^T(k) X X^T l(k) \end{aligned}$$

最终可知，求数据压缩问题等价于求解

$$\max \sum_{k=1}^m l^T(k) \tilde{X} \tilde{X}^T l(k)$$

$$s.t. \quad L^T L = I_m$$

注意下标  
应该是m

并且，最优的压缩变量是  $\hat{y}_k(t) = q^T(k) \tilde{x}(t), k=1, 2, \dots, m$

它就是前 m 个主成分。

## ■相对逼近误差

$$\begin{aligned}
 & \frac{\sum_{t=1}^N \tilde{x}^T(t) \tilde{x}(t) - \sum_{t=1}^N \tilde{x}^T(t) \hat{L} \hat{L}^T \tilde{x}(t)}{\sum_{t=1}^N \tilde{x}^T(t) \tilde{x}(t)} = 1 - \frac{\sum_{t=1}^N \hat{y}^T(t) \hat{y}(t)}{\sum_{t=1}^N \tilde{x}^T(t) \tilde{x}(t)} = 1 - \frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^n \lambda_i} = 1 - \frac{\sum_{i=1}^m \lambda_i / (N-1)}{\sum_{i=1}^n \lambda_i / (N-1)} \\
 & = 1 - \frac{\sum_{k=1}^m \delta^2(\hat{y}_k)}{\sum_{k=1}^n \delta^2(\hat{y}_k)} = \frac{\sum_{k=1}^n \delta^2(\hat{y}_k) - \sum_{k=1}^m \delta^2(\hat{y}_k)}{\sum_{k=1}^n \delta^2(\hat{y}_k)} = \frac{\sum_{k=m+1}^n \delta^2(\hat{y}_k)}{\sum_{k=1}^n \delta^2(\hat{y}_k)} = \frac{\sum_{i=m+1}^n \lambda_i}{\sum_{i=1}^n \lambda_i} = \frac{\sum_{k=m+1}^n \delta^2(\hat{y}_k)}{n}
 \end{aligned}$$

对于乌龟数例，若用第一个主成分压缩原数据，只需存储：

$$\hat{y}_1(t) = 0.5900\tilde{x}_1(t) + 0.5731\tilde{x}_2(t) + 0.5687\tilde{x}_3(t)$$

$$t = 1, 2, \dots, 12$$

$$q(1) = \begin{bmatrix} 0.5900 \\ 0.5731 \\ 0.5687 \end{bmatrix}$$

12个投影结果（第一主成分），1个主成分投影方向（3维），  
6个均值和方差，共21个数据。  
压缩率=21/36=58%

# 主成分分析方法——数据压缩



因为  $\frac{1}{3}\sigma^2(\hat{y}_1) = 0.89$

令  $\hat{x}(t) = \begin{bmatrix} 0.5900 \\ 0.5731 \\ 0.5687 \end{bmatrix} \hat{y}_1(t)$

相对逼近误差为

$$\frac{\sum_{t=1}^{12} (\tilde{x}(t) - \hat{x}(t))^T (\tilde{x}(t) - \hat{x}(t))}{\sum_{t=1}^{12} \tilde{x}^T(t) \tilde{x}(t)} = 11\%$$

$$\begin{aligned} \sigma^2(\hat{y}_1) &= 2.67 \\ \sigma^2(\hat{y}_2) &= 0.22 \\ \sigma^2(\hat{y}_3) &= 0.11 \end{aligned}$$

$$\frac{\sum_{k=m+1}^n \sigma^2(\hat{y}_k)}{n}$$

## ■注意

- 能够利用主成分有效压缩数据，是因为数据本身具有可压缩性。
- 换句话说，就是样本相关矩阵的特征根相差很大，其本质是变量间近似线性相关！

## 例题

对某组10维规格化向量的数据压缩问题，用第一至第四个主成分进行压缩和用第一至第五个主成分进行压缩的误差相同；用第一、第三个主成分进行压缩和用第二、第四个主成分进行压缩的误差相同；用第一个主成分进行压缩相对误差是0.6。请求出每个主成分的样本方差。

## 解：

由第一条条件，我们有：

$$\lambda_5 + \cdots + \lambda_{10} = \lambda_6 + \cdots + \lambda_{10} \Rightarrow \lambda_i = 0, i \geq 5$$

由第二条条件，我们有：

$$\lambda_2 + \lambda_4 = \lambda_1 + \lambda_3 \Rightarrow \lambda_2 = \lambda_1, \lambda_4 = \lambda_3$$

## 例题

对某组10维规格化向量的数据压缩问题，用第一至第四个主成分进行压缩和用第一至第五个主成分进行压缩的误差相同；用第一、第三个主成分进行压缩和用第二、第四个主成分进行压缩的误差相同；用第一个主成分进行压缩相对误差是0.6。请求出每个主成分的样本方差。

## 解：

由第三条件，我们有：

$$(\lambda_2 + \lambda_3 + \lambda_4) / (\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4) = 0.6$$

$$(\lambda_1 + 2\lambda_3) / (2\lambda_1 + 2\lambda_3) = 0.6$$

则：  $\lambda_1 = 4\lambda_3$

$$\sum \delta^2(\hat{y}_k) = 10$$

$$\delta^2(y_k) = \bar{\lambda}_k = \frac{\lambda_k}{N-1}$$

$$\sum_{k=1}^n \delta^2(\hat{y}_k) = n$$

$$\delta^2(\hat{y}_1) = \delta^2(\hat{y}_2) = 4; \quad \delta^2(\hat{y}_3) = \delta^2(\hat{y}_4) = 1; \text{其余为} 0$$



## ■主成分压缩数据所需要存储的数据个数

假设样本数据是  $n$  维向量，共有  $N$  个

我们提取出了  $m$  个主成分，则共需要存储数据

$$m \times N + m \times n + n + n$$

在  $m$  个主成分方向上的投影

$m$  个主成分向量

样本方差

样本均值

## ■主成分数据压缩计算过程

1. 对样本数据进行归一化  $\frac{x_i(t) - e(x_i)}{\sqrt{\delta^2(x_i)}}$
2. 计算归一化样本数据协方差矩阵  $XX^T$
3. 计算  $XX^T$  的特征值和对应的特征向量
4. 对特征值进行排序，选取前  $m$  个特征值所对应的特征向量  $q(1), q(2), \dots, q(m)$  作为主成分方向
5. 计算各样本数据在主成分方向上的投影  
$$y(t) = [q(1), q(2), \dots, q(m)]^T x(t)$$
6. 计算**压缩率**:  $\eta = \frac{m \times N + m \times n + n + n}{n \times N}$

## ■基于PCA的海量数据压缩实例

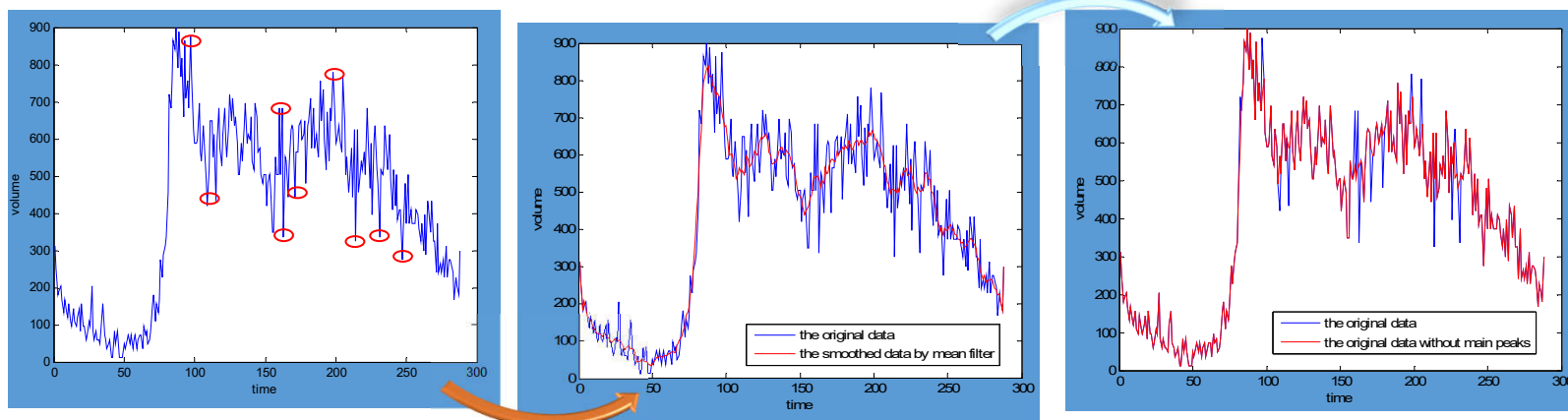
- ① 交通数据预处理
- ② 压缩与恢复性能评价指标
- ③ 主要研究结果
- ④ 其他结果分析
- ⑤ 软件展示

# 主成分分析方法——数据压缩

## ① 交通数据预处理

- 交通流中的“尖峰”
  - 这些非线性的尖峰严重影响了PCA对数据的压缩恢复效果。
- 利用均值滤波器提取“尖峰”
  - 尖峰时刻的交通流量在压缩前单独存储。
  - 用均值滤波器平滑后的流量值代替尖峰值。

$$MF(i) = \frac{1}{(2K+1)} \sum_{m=-K}^K x(i+m)$$



# 主成分分析方法——数据压缩

## ②压缩与恢复性能评价指标

CR

压缩比

$$CR = \frac{\text{原始数据所占用的字节数}}{\text{压缩存储数据所占用的字节数}}$$



APRE

均方根百分比误差

$$APRE(i, j) = \frac{\|f_i^j - \hat{f}_i^j\|}{\|f_i^j\|} \quad \begin{array}{l} f_i^j \text{ 原始数据向量} \\ \hat{f}_i^j \text{ 恢复数据向量} \end{array}$$



R

相关系数

$$R(i, j) = \frac{(f_i^j - \overline{f_i^j})^T \cdot (\hat{f}_i^j - \overline{\hat{f}_i^j})}{\|f_i^j - \overline{f_i^j}\| \cdot \|\hat{f}_i^j - \overline{\hat{f}_i^j}\|} \quad \begin{array}{l} \overline{f_i^j} \text{ 原始数据均值向量} \\ \overline{\hat{f}_i^j} \text{ 恢复数据均值向量} \end{array}$$



## ③主要研究结果

### ➤主成分分析

- 第一主成分贡献率达80.97%。
- 前25个主成分贡献率和达92.88%。

### ➤压缩和恢复



### ➤结果

- CR(压缩比)为6.2。
- 平均APRE为13%。
- 平均相关系数为0.9524。

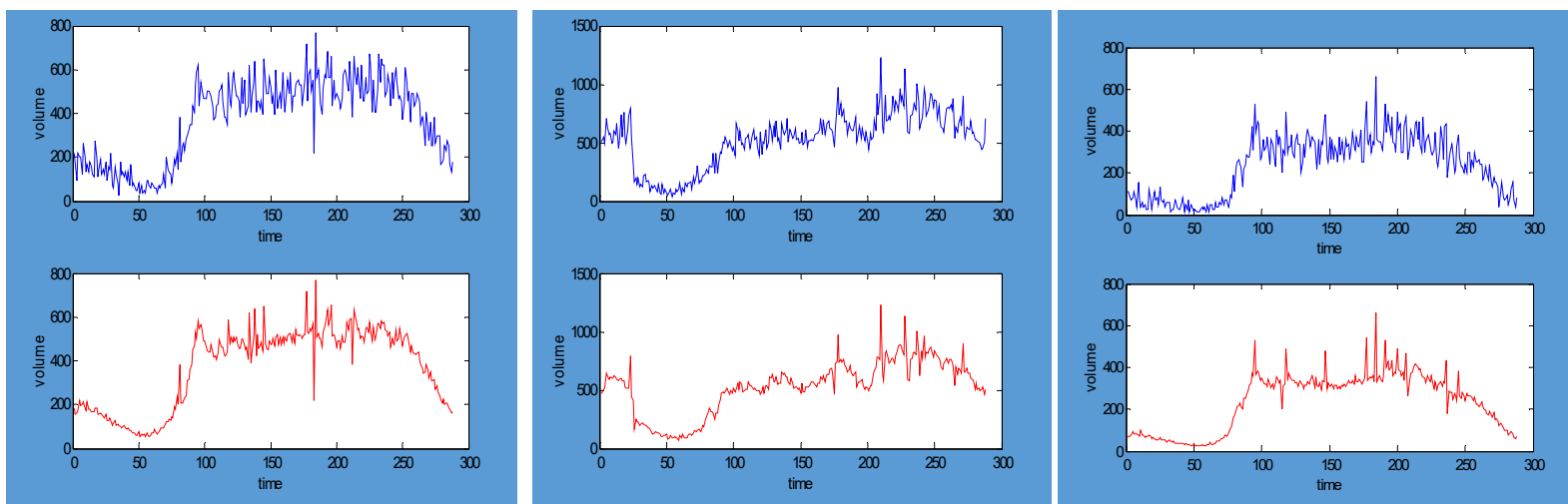
# 主成分分析方法——数据压缩

## ③主要实验结果

对于具有不同特征的交通流均可以较好的恢复。

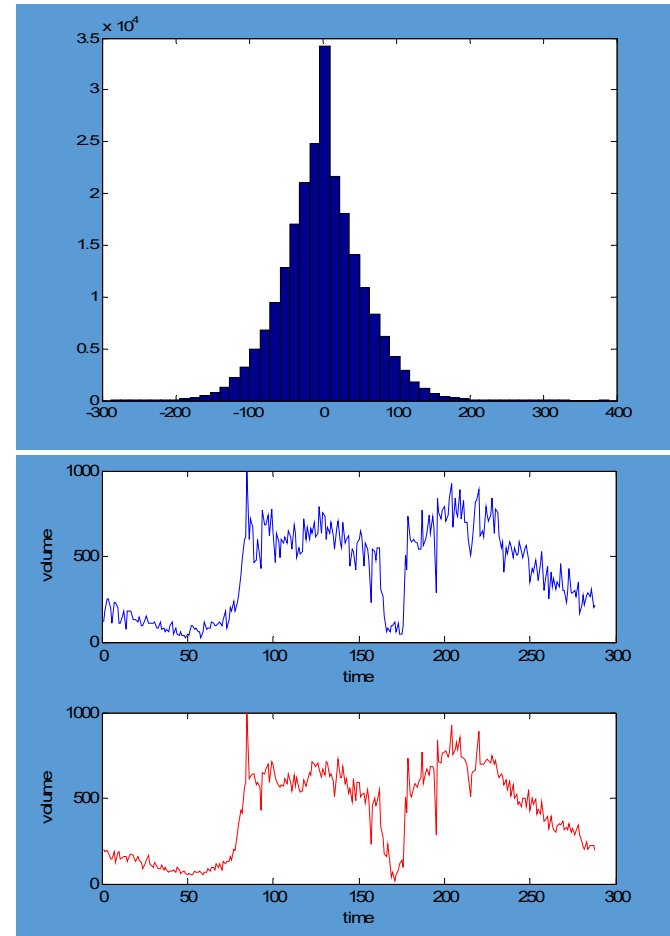
- 不同的早、晚高峰时间
- 同一时段不同的流量值
- 特殊的“尖峰”

Fig 原始数据与恢复数据



## ④其他结果分析

- 恢复误差分布
  - 具有正态分布的特征
  - 绝大多数恢复数据点与原始数据点基本相同。
- 对异常数据具有鲁棒性
  - “尖峰”点
  - 交通流异常（如在白天出现堵车现象时的交通流）





# 主成分分析方法——数据压缩

Table 各路口压缩和恢复性能参数

路口编号	1	2	3	4	5	6	7
总数据点个数	9216	18432	13824	9216	23040	27648	9216
提前存储的尖峰点个数	386	762	583	393	924	1111	353
提前存储数据点占总数据点百分比	4.188%	4.134%	4.217%	4.264%	4.010%	4.018%	3.830%
APRE (误差)	12.14%	10.75%	9.259%	9.100%	10.44%	9.871%	9.592%
R (相关度)	0.952	0.964	0.977	0.978	0.964	0.968	0.969

8	9	10	11	12	13	14	15	16	17
9216	9216	13824	27648	9216	9216	9216	13824	4608	13824
376	379	525	1068	380	384	341	538	170	584
4.080%	4.112%	3.798%	3.864%	4.123%	4.167%	3.700%	3.892%	3.689%	4.225%
9.703%	8.915%	8.328%	9.235%	9.029%	8.191%	6.995%	11.17%	14.36%	9.178%
0.958	0.965	0.981	0.977	0.980	0.984	0.987	0.968	0.949	0.978

# 主成分分析方法——数据压缩

基于主成分分析(PCA)的数据压缩

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)

后退 搜索 文件夹 Folder Sync

地址(C:) 转到

文件和文件夹

- 重命名这个
- 移动这个
- 复制这个
- 将这个文件移到Web
- 以电子方式发布此文件
- 删除这个

其它位置

- mat
- 我的文档
- 共享文档
- 我的电脑
- 网上邻居

详细信息

Original Data  
MATLAB MAT-file  
修改日期: 2008年9月5日, 17:30  
大小: 253 KB

文件 数据预处理

交通流量数据

交通记录时间

- 0点0分
- 0点5分
- 0点10分
- 0点15分
- 0点20分
- 0点25分

预处理操作

记录时间 20060810

显示数据

- ☒ 处理前数据
- ☐ 处理后数据

数据压缩

打开原始文件

基本PCA

请输入主成分数

选项序号

压缩

选择类型

- ☒ 压缩全部
- ☐ 按时间序列
- ☐ 按日期
- ☐ 按地点

改善PCA

分离矩阵阈值

第一矩阵主成分数目

第二矩阵贡献率阈值

选择类型

- ☒ 无补偿
- ☐ 有补偿

压缩

交通记录地点

- 月坛北街东口1
- 月坛北街东口2
- 月坛北街东口3
- 月坛北街东口4
- 儿童医院1
- 儿童医院2
- 儿童医院3
- 儿童医院4
- 儿童医院8
- 月坛南桥1
- 月坛南桥2
- 月坛南桥3
- 月坛南桥4
- 月坛南桥7
- 月坛北桥1
- 月坛北桥2
- 月坛北桥3
- 月坛北桥4
- 月坛北桥5
- 展览馆汽车站1
- 展览馆汽车站3
- 审计署1
- 审计署3
- 东直门北街西口2
- 东直门北街西口1
- 东直门北街西口3

操作已完成

# 主成分分析方法——线性回归



龟壳化石数据

样本 $t$	长 $x_1(t)$	宽 $x_2(t)$	高 $x_3(t)$	重量 $z(t)$
1	93	74	37	97
2	94	78	35	101
3	96	80	35	106
4	101	84	39	112
5	102	85	38	114
6	103	81	37	107
7	104	83	39	112
8	106	83	39	111
9	107	82	38	107
10	112	89	40	118
11	113	88	40	116
12	114	86	40	115

规格化的重量和长宽高之间存在线性关系

$$\tilde{z}(t) \approx a_1 \tilde{x}_1(t) + a_2 \tilde{x}_2(t) + a_3 \tilde{x}_3(t)$$

参数估计: 
$$\min \sum_{t=1}^{12} \left( \tilde{z}(t) - \sum_{i=1}^3 a_i \tilde{x}_i(t) \right)^2$$

$$\hat{a}_1 = -0.19, \quad \hat{a}_2 = 0.97, \quad \hat{a}_3 = 0.23$$

相对拟合误差:

$$e(\hat{z}, \tilde{z}) = \frac{\sum_{t=1}^{12} (\hat{z}(t) - \tilde{z}(t))^2}{\sum_{t=1}^{12} (\tilde{z}(t))^2} = 2.4\%$$

# 主成分分析方法——线性回归



	样本	长	宽	高	重量
另一组样本	$t$	$x_1(t)$	$x_2(t)$	$x_3(t)$	$z(t)$
	13	116	90	43	124
	14	117	90	41	121
	15	117	91	41	119
	16	119	93	41	120
	17	120	89	40	121
	18	120	93	44	126
	19	121	95	42	126
	20	125	93	45	129
	21	127	96	45	131
	22	128	95	45	131
	23	131	95	46	127
	24	135	106	47	138

# 主成分分析方法——线性回归



预报精度:  $\hat{z}(t) = \hat{a}_1 \tilde{x}_1(t) + \hat{a}_2 \tilde{x}_2(t) + \hat{a}_3 \tilde{x}_3(t)$   
 $t = 13, 14, \dots, 24$

$$e(\hat{z}, \tilde{z}) = \frac{\sum_{t=13}^{24} (\hat{z}(t) - \tilde{z}(t))^2}{\sum_{t=13}^{24} (\tilde{z}(t))^2} = 20.9\%$$

利用全部样本回归  $\min \sum_{t=1}^{24} \left( \tilde{z}(t) - \sum_{i=1}^3 a_i \tilde{x}_i(t) \right)^2$

参数估计:

$$\hat{a}_1 = -0.04, \quad \hat{a}_2 = 0.67, \quad \hat{a}_3 = 0.38 \quad e(\hat{z}, \tilde{z}) = 8.9\%$$

原因分析

$$\min \sum_{t=t_1}^{t_2} \left( \tilde{z}(t) - \sum_{i=1}^3 a_i \tilde{x}_i(t) \right)^2$$
$$\Downarrow$$
$$\min (\tilde{Z} - a^T \tilde{X})(\tilde{Z} - a^T \tilde{X})^T$$

求偏导并令偏导等于0，可求得

$$\hat{a} = (\tilde{X}\tilde{X}^T)^{-1} \tilde{X}\tilde{Z}^T$$

# 主成分分析方法——线性回归



假定最好的参数值是  $a^*$  , 误差  $\eta = \tilde{Z} - (a^*)^T \tilde{X}$

参数估计误差为  $a^* - \hat{a} = a^* - (\tilde{X}\tilde{X}^T)^{-1} \tilde{X}\tilde{Z}^T$

$$= a^* - (\tilde{X}\tilde{X}^T)^{-1} \tilde{X}(\tilde{X}^T a^* + \eta)$$

$$= -(\tilde{X}\tilde{X}^T)^{-1} \tilde{X}\eta$$

因为  $\tilde{X}\tilde{X}^T = Q\Lambda Q^T \Rightarrow (\tilde{X}\tilde{X}^T)^{-1} = Q\Lambda^{-1}Q^T$

$$\Lambda^{-1} = \text{diag}(\lambda_1^{-1} \quad \lambda_2^{-1} \quad \lambda_3^{-1}) \Rightarrow a^* - \hat{a} = Q\Lambda^{-1}Q^T \tilde{X}\eta$$

$\lambda_2^{-1}$  和  $\lambda_3^{-1}$  很大, 可能使参数估计误差很大!



由主成分分析知

$$\tilde{x}_1(t) \approx \hat{l}_1(1)\hat{y}_1(t)$$

$$\tilde{x}_2(t) \approx \hat{l}_2(1)\hat{y}_1(t)$$

$$\tilde{x}_3(t) \approx \hat{l}_3(1)\hat{y}_1(t)$$

$$\tilde{z}(t) \approx a_1\tilde{x}_1(t) + a_2\tilde{x}_2(t) + a_3\tilde{x}_3(t)$$

$\Rightarrow$

$$\tilde{z}(t) \approx b\hat{y}_1(t)$$

$$\min \sum_{t=1}^{12} (\tilde{z}(t) - b\hat{y}_1(t))^2 \Rightarrow \hat{b} = 0.5732$$

$$\hat{z}(t) = \hat{b} \left( \hat{l}_1(1)\tilde{x}_1(t) + \hat{l}_2(1)\tilde{x}_2(t) + \hat{l}_3(1)\tilde{x}_3(t) \right) \quad t = 1, 2, \dots, 24$$

拟合误差12.2%      预报误差12.4%

## ■一般情况

$$z(t) \approx a^T x(t)$$

$$x(t) \approx \hat{L} \hat{y}(t) \quad \Rightarrow \quad z(t) \approx b^T \hat{y}(t)$$

$$\hat{y}(t) = \hat{L}^T x(t)$$

没有病态问题



$$\hat{b} = (\hat{Y} \hat{Y}^T)^{-1} \hat{Y} Z^T = \text{diag}(\lambda_1^{-1} \lambda_2^{-1} \cdots \lambda_m^{-1}) \hat{Y} Z^T$$

$\Rightarrow$

$$\hat{z}(t) = \hat{b}^T \hat{y}(t) = \hat{b}^T \hat{L}^T x(t)$$

# 主成分分析方法

- 主要内容回顾

- 主成分分析基本原理
- 基于主成分分析的数据压缩
- 基于主成分分析的回归分析