10.1 引言

一. PCA的主要功能

在信息损失最小的前提下，对高维空间进行**降维**处理。

数据类型： 样本点×变量（定量变量）

$$
\begin{array}{c} \\ e_1 \\ e_2 \\ \mathrm{M} \\ e_n \end{array}
\begin{array}{c}
\begin{array}{cccc} x_1 & x_2 & \mathrm{L} & x_p \end{array} \\
\left[\begin{array}{cccc}
x_{11} & x_{12} & \mathrm{L} & x_{1p} \\
x_{21} & x_{22} & \mathrm{L} & x_{2p} \\
\mathrm{M} & \mathrm{M} & \mathrm{L} & \mathrm{M} \\
x_{n1} & x_{n2} & \mathrm{L} & x_{np}
\end{array}\right]_{n\times p}
\end{array}
\xrightarrow{\text{PCA}}
\begin{array}{c} \\ e_1 \\ e_2 \\ \mathrm{M} \\ e_n \end{array}
\begin{array}{c}
\begin{array}{ccc} y_1 & \mathrm{L} & y_m \end{array} \\
\left[\begin{array}{ccc}
y_{11} & \mathrm{L} & y_{1m} \\
y_{21} & \mathrm{L} & y_{2m} \\
\mathrm{M} & \mathrm{ML} & \mathrm{M} \\
y_{n1} & \mathrm{L} & y_{nm}
\end{array}\right]_{n\times m}
\end{array}
\qquad m \ll p
$$

10.3 数据的原则化处理

（一）“中心化”处理—平移变换

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_j \ , \ i = 1,2,\mathrm{L}\ ,n, \quad j = 1,2,\mathrm{L}\ ,p$$

性质：不变化样本点集合中点与点的相互位置；

$$\begin{bmatrix} x_{11} & \mathrm{L} & x_{1p} \\ \mathrm{M} & & \mathrm{M} \\ x_{n1} & \mathrm{L} & x_{np} \end{bmatrix}_{n\times p} \xrightarrow{\text{中心化}} \begin{bmatrix} \tilde{x}_{11} & \mathrm{L} & \tilde{x}_{1p} \\ \mathrm{M} & & \mathrm{M} \\ \tilde{x}_{n1} & \mathrm{L} & \tilde{x}_{np} \end{bmatrix}_{n\times p}$$

$$\begin{matrix} \bar{x}_1 & \mathrm{L} & \bar{x}_p \end{matrix}$$

（二） 原则化处理：中心化——压缩

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

性质：

- $g^* = 0$ (均值为0)
- $s_j^* = 1, \quad j = 1,2,\ldots,p$ （方差等于1）
- $$r_{jk}^* = \frac{s_{jk}^*}{s_j^* s_k^*} = s_{jk}^*$$

对于原则化数据表：

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_i}$$

（1）变量方差均等于1

（2）有关系数矩阵 ＝ 协方差矩阵

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & \mathrm{L} & r_{1p} \\ r_{21} & 1 & \mathrm{L} & r_{2p} \\ \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ r_{p1} & r_{p2} & \mathrm{L} & 1 \end{bmatrix} = \begin{bmatrix} s_1^2 & s_{12} & \mathrm{L} & s_{1p} \\ s_{21} & s_2^2 & \mathrm{L} & s_{2p} \\ \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ s_{p1} & s_{p2} & \mathrm{L} & s_p^2 \end{bmatrix} = \boldsymbol{V}$$

10.4 PCA的算法

一. PCA对数据系统做“最佳简化”的含意

PCA可在确保信息损失的前提下，经线性变换和舍弃一小部分信息，以少数线性无关的新综合变量取代原始采用的多维有关变量。

$$x_1, x_2, \cdots, x_p \xrightarrow{PCA:\text{平移}\oplus\text{旋转}} y_1, y_2, \cdots, y_m \quad (m < p)$$

称：$L = L(v_1, v_2, \cdots, v_m)$ 为“主超平面”；

称：$L_2 = L(v_1, v_2)$ 为“主平面”；

二. PCA算法中的几种要素

输入—输出：

$$\begin{array}{c} \\ e_1 \\ e_2 \\ \mathrm{M} \\ e_n \end{array}\begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \mathrm{L} & x_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \mathrm{L} & x_{1p} \\ x_{21} & x_{22} & \mathrm{L} & x_{2p} \\ \mathrm{M} & \mathrm{M} & \mathrm{L} & \mathrm{M} \\ x_{n1} & x_{n2} & \mathrm{L} & x_{np} \end{bmatrix}_{n\times p} \end{array} \xrightarrow{\text{平移⊕旋转}} \begin{array}{c} \\ e_1 \\ e_2 \\ \mathrm{M} \\ e_n \end{array}\begin{array}{c} \begin{array}{ccc} y_1 & \mathrm{L} & y_m \end{array} \\ \begin{bmatrix} y_{11} & \mathrm{L} & y_{1m} \\ y_{21} & \mathrm{L} & y_{2m} \\ \mathrm{M} & \mathrm{ML} & \mathrm{M} \\ y_{n1} & \mathrm{L} & y_{nm} \end{bmatrix}_{n\times m} \end{array} \quad m < p$$

（1）平移变换：把原点移到重心：

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_j \ , \ i = 1,2,\mathrm{L} \ ,n,$$

$$i = 1.2.\mathrm{L} \ . \ p$$

（2）旋转变换，得到“主轴”： $u_1, u_2, \ldots, u_p \in R^p$

其中，u_1相应数据变异最大的方向，u_2与u_1 垂直，相应于数据变异第二大方向，…

所以$u_1, \ldots, u_p$是原则正交的，即：$u_j' u_k = \begin{cases} 1 & j \neq k \\ 0 & j = k \end{cases}$

（3）求样本点e_i 在 u_h轴上的投影坐标

$$y_h(i) = e_i' \cdot u_h$$

全部样本点在u_h 上的投影构成“第h主成份 y_h”：

$$y_h = (y_h(1), y_h(2), \mathrm{L}\ , y_h(n))' \in R^n$$

在主成份中，$\mathrm{Var}(y_1) \to \max$

而 $y_2 \perp y_1$, 且$\mathrm{Var}(y_2)$ 是次大的……

（4）在 $\boldsymbol{u}_h$主轴上，$\boldsymbol{e}_i$的投影坐标是$y_h(i)$

$$y_h(i)=\boldsymbol{e}_i'\cdot\boldsymbol{u}_h,\quad i=1,2,\mathrm{L}\ ,n$$

第h主成份为：

$$\boldsymbol{y}_h=(y_h(1),y_h(2),\mathrm{L}\ ,y_h(n))'$$

$$=\begin{pmatrix}\boldsymbol{e}_1'\cdot\boldsymbol{u}_h\\ \mathrm{M}\\ \boldsymbol{e}_n'\cdot\boldsymbol{u}_h\end{pmatrix}=\begin{pmatrix}\boldsymbol{e}_1'\\ \mathrm{M}\\ \boldsymbol{e}_n'\end{pmatrix}\cdot\boldsymbol{u}_h=\boldsymbol{X}\boldsymbol{u}_h$$

$$=(\boldsymbol{x}_1,\mathrm{L}\ ,\boldsymbol{x}_p)\begin{pmatrix}u_h(1)\\ \mathrm{M}\\ u_h(p)\end{pmatrix}=\sum_{j=1}^{p}u_h(j)\boldsymbol{x}_j$$

y_h是原变量x_1，…，x_P的线性组合，组合系数为$u_h(1),\ldots,u_h(p)$

PAC算法推导：

不妨设变量 $\mathbf{X}_1, \mathbf{X}_2, \mathrm{L}\ , \mathbf{X}_p$ 都是中心化的，$\mathbf{Y}_h \in L\left(\mathbf{X}_1, \mathbf{X}_2, \mathrm{L}\ , \mathbf{X}_p\right)$

求第主成份 $\mathbf{Y}_h = \sum_{j=1}^{p} u_{hj} \mathbf{X}_j$

经过旋转变换得到的Y，是X的线性组合

$$\mathrm{D}\left(\mathbf{Y}_h\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_{ih} - 0\right)^2 = \frac{1}{n}\sum_{i=1}^{n} y_{ih}^2 = \frac{1}{n}\left\|\mathbf{Y}_h\right\|^2 = \frac{1}{n}\left\langle \mathbf{Y}_h, \mathbf{Y}_h \right\rangle$$

$$= \frac{1}{n}\left\langle \left(u_{h1}\mathbf{X}_1 + u_{h2}\mathbf{X}_2 + \mathrm{L}\ + u_{hp}\mathbf{X}_p\right), \left(u_{h1}\mathbf{X}_1 + u_{h2}\mathbf{X}_2 + \mathrm{L}\ + u_{hp}\mathbf{X}_p\right) \right\rangle$$

$$= \frac{1}{n}\left(u_{h1}, u_{h2}, \mathrm{L}\ , u_{hp}\right) \begin{pmatrix} \left\langle \mathbf{X}_1, \mathbf{X}_1 \right\rangle & \left\langle \mathbf{X}_1 \cdot \mathbf{X}_2 \right\rangle & \mathrm{L} & \left\langle \mathbf{X}_1 \cdot \mathbf{X}_p \right\rangle \\ \left\langle \mathbf{X}_2 \cdot \mathbf{X}_1 \right\rangle & \left\langle \mathbf{X}_2, \mathbf{X}_2 \right\rangle & \mathrm{L} & \left\langle \mathbf{X}_2 \cdot \mathbf{X}_p \right\rangle \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \left\langle \mathbf{X}_p \cdot \mathbf{X}_1 \right\rangle & \left\langle \mathbf{X}_p \cdot \mathbf{X}_2 \right\rangle & \mathrm{L} & \left\langle \mathbf{X}_p, \mathbf{X}_p \right\rangle \end{pmatrix} \begin{pmatrix} u_{h1} \\ u_{h2} \\ \mathrm{M} \\ u_{hp} \end{pmatrix}$$

$$= \mathbf{u}_h' \mathbf{V} \mathbf{u}_h$$

$$\max \quad \sum_{h=1}^{m} \mathbf{u}_h' \mathbf{V} \mathbf{u}_h$$

$$\text{s.t.} \quad \begin{cases} \|\mathbf{u}_h\| = 1 \\ \mathbf{u}_h' \mathbf{u}_l = 0 \qquad h \neq l \\ \mathbf{u}_1' \mathbf{V} \mathbf{u}_1 \geq \mathbf{u}_2' \mathbf{V} \mathbf{u}_2 \geq \mathrm{L} \geq \mathbf{u}_m' \mathbf{V} \mathbf{u}_m \\ k = 1, 2, \mathrm{L}\ , p, l \neq k \end{cases}$$

所以 $\mathbf{u}_1, \mathbf{u}_2, \mathrm{L}\ \mathbf{u}_p$ 是矩阵 $\mathbf{V}$ 的特征向量，

相应的特征值是 $\lambda_1 \geq \lambda_2 \geq \mathrm{L} \geq \lambda_p$

三. PCA的计算措施（一般情况下）

（1）数据的原则化

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_{j}}{s_{i}}$$

为以便起见，仍记 $(x_{ij}^{*})=(x_{ij})=\boldsymbol{X}_{n\times p}$。

（2）计算原则化数据表 $\boldsymbol{X}_{n\times p}$ 的协方差矩阵 $\boldsymbol{V}$。

（3）求 $\boldsymbol{V}$ 的前 m 个特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m>0$，

以及相应的特征向量： $\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_m$ (主轴)

它们是原则正交的： $\mathbf{u}_h'\mathbf{u}_l=\begin{cases}1 & h=l\\ 0 & h\neq l\end{cases}$

（4）在 $\boldsymbol{u}_h$主轴上，$\boldsymbol{e}_i$的投影坐标是$y_h(i)$

$$y_h(i)=\boldsymbol{e}_i'\cdot\boldsymbol{u}_h,\quad i=1,2,\mathrm{L}\ ,n$$

第h主成份为：

$$\boldsymbol{y}_h=(y_h(1),y_h(2),\mathrm{L}\ ,y_h(n))'$$

$$=\begin{pmatrix}\boldsymbol{e}_1'\cdot\boldsymbol{u}_h\\ \mathrm{M}\\ \boldsymbol{e}_n'\cdot\boldsymbol{u}_h\end{pmatrix}=\begin{pmatrix}\boldsymbol{e}_1'\\ \mathrm{M}\\ \boldsymbol{e}_n'\end{pmatrix}\cdot\boldsymbol{u}_h=\boldsymbol{X}\boldsymbol{u}_h$$

$$=(\boldsymbol{x}_1,\mathrm{L}\ ,\boldsymbol{x}_p)\begin{pmatrix}u_h(1)\\ \mathrm{M}\\ u_h(p)\end{pmatrix}=\sum_{j=1}^{p}u_h(j)\boldsymbol{x}_j$$

y_h是原变量x_1，…，x_P的线性组合，组合系数为$u_h(1),\ldots,u_h(p)$

四、主成份的统计特征

第h主成份 $\boldsymbol{y}_h = (y_h(1), y_h(2), \mathrm{L}\ , y_h(n))' \in \boldsymbol{R}^n$

$$y_h(i) = \boldsymbol{e}_i' \cdot \boldsymbol{u}_h, \quad i = 1,2,\mathrm{L}\ ,n$$

①y_h 的均值为0。

$$\frac{1}{n}\sum_{i=1}^{n} y_h(i) = 0$$

② y_h的方差等于λ_h。

$$Var(y_h) = \frac{1}{n}\sum_{i=1}^{n}(y_h(i) - 0)^2 = \lambda_h$$

③ y_j与y_k的协方差等于0：

$$Cov(\mathbf{y}_h, \mathbf{y}_l) = 0 \ , \quad \forall h \neq l$$

总结：PCA算法的输入与输出

$$\boldsymbol{X}_{n\times p}=\begin{bmatrix} & \mathrm{M} & \\ \mathrm{L} & x_{ij} & \mathrm{L} \\ & \mathrm{M} & \end{bmatrix}_{n\times p}$$

n个样本点，p个变量

①$\lambda_1\geq\lambda_2\geq\ldots\ldots\geq\lambda_m$

$\mathrm{Var}(\boldsymbol{y}_1),\mathrm{Var}(\boldsymbol{y}_2),\ldots,\ \mathrm{Var}(\boldsymbol{y}_m)$

②$\boldsymbol{u}_1,\ \boldsymbol{u}_2,\ \ldots\ldots,\ \boldsymbol{u}_m\in\boldsymbol{R}^P$ (主轴)

③$\boldsymbol{y}_1,\ \boldsymbol{y}_2,\ \ldots\ldots,\ \boldsymbol{y}_m\in\boldsymbol{R}^n$ (主成份)

$$\boldsymbol{Y}_{n\times m}=\begin{bmatrix} & \mathrm{M} & \\ \mathrm{L} & y_{ij} & \mathrm{L} \\ & \mathrm{M} & \end{bmatrix}_{n\times m}$$

n个样本点，m个变量

总结：经过主成份分析，

$$\begin{bmatrix} x_{11} & x_{12} & \mathrm{L} & x_{1p} \\ x_{21} & x_{22} & \mathrm{L} & x_{2p} \\ \mathrm{M} & \mathrm{M} & \mathrm{L} & \mathrm{M} \\ x_{n1} & x_{n2} & \mathrm{L} & x_{np} \end{bmatrix}_{n\times p} \xrightarrow{PCA} \begin{bmatrix} y_{11} & \mathrm{L} & y_{1m} \\ y_{21} & \mathrm{L} & y_{2m} \\ \mathrm{M} & \mathrm{L} & \mathrm{M} \\ y_{n1} & \mathrm{L} & y_{nm} \end{bmatrix}_{n\times m} \qquad m < p$$

均值 $g = (\bar{x}_1, \bar{x}_2, \mathrm{L}\ , \bar{x}_p) \rightarrow g = (0, \mathrm{L}\ , 0)$

方差 $s_1^2, s_2^2, \mathrm{L}\ , s_p^2 \rightarrow \lambda_1 \geq \lambda_2 \geq \mathrm{L} \geq \lambda_m$

$\mathrm{Cov}(x_j, x_k) \neq 0\ \ (j \neq k) \rightarrow \mathrm{Cov}(y_j, y_k) = 0\ \ (j \neq k)$

10.5 PCA的辅助分析技术

一．怎样选用精度合适的主超平面

1. *m*维主超平面的精度测量

主成份分析前，$X_{n\times p}$数据中的全部变异信息：

$$\sum_{j=1}^{p}\mathrm{Var}(x_j)=\sum_{j=1}^{p}s_j^2=p$$

原则化

主成份分析后保存的数据变差：

$\mathrm{Var}(y_1)=\lambda_1,\mathrm{Var}(y_2)=\lambda_2,\ \ldots,\ \mathrm{Var}(y_m)=\lambda_m$

$$\sum_{h=1}^{m}\mathrm{Var}(y_h)=\sum_{h=1}^{m}\lambda_h$$

形象地看：$[\mathbf{x}_1, \mathbf{x}_2, \mathrm{L}\ , \mathbf{x}_p] \xrightarrow{PCA} [\mathbf{y}_1, \mathbf{y}_2, \mathrm{L}\ , \mathbf{y}_m]$

方差：$s_1^2,\ s_2^2, \mathrm{L}\ , s_p^2 \qquad \lambda_1, \lambda_2, \mathrm{L}\ , \lambda_m$

注意：$\sum_{h=1}^{p} \lambda_h = \sum_{h=1}^{p} \mathrm{Var}(y_h) = \sum_{j=1}^{p} s_j^2$

所以，定义“**合计贡献率**”：

$$Q_m = \frac{\sum_{h=1}^{m} \lambda_h}{\sum_{j=1}^{p} s_j^2} = \frac{1}{p} \sum_{h=1}^{m} \lambda_h$$

原则化

2.、怎样选用合适精度的$u_1,\ldots,u_m$ 。

根据合计贡献率能够拟定所要选用的成份的个数。

例.管理期刊评价

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Loadings	
	Total	% of Variance	Cumulative %	Total	% of Variance
1	1.532	38.293	38.293	1.532	38.293
2	1.026	25.644	63.937	1.026	25.644
3	.886	22.150	86.087		
4	.557	13.913	100.000		

（2）若希望Q_m在80%左右，应选用 3 个主成份。

某些科技问题的合计贡献率要求在90%以上。但对复杂的社会科学、行为科学或经济学中的数据，能到达60%也能够考虑。

二. 主成份的命名

主成份y_1，…，y_m 是原变量x_1，…,x_p 的线性组合。
原变量x_1，…,x_p 都有明确的物理含意。

问题： y_1，…，y_m的物理含意是什么？

1. 作用：指出影响系统构造的主要原因和主要特征。

例 ①：分析各阶层人员生活状态

发展中国家：y_1——食品， y_2——穿着

发达国家： y_1——住宅， y_2——旅游

以此能够划分不同社会阶层的生活档次。

（在这个方向，人们的生活水平差距最大）

例②：中国城市经济分析：

1984： y_1—综合水平， y_2——工农业投入国家。

1988： y_1—综合水平，y_2——外贸，科技。

中国改革开放以来，因为开放程度不同，使中国各地域经济水平差距逐渐拉大。所以，加大开放力度，发展高科技产业是城市发展的主要工作方面。

2. 措施：专业知识 + 数学手段

数学手段： 研究 y_h与 x_1， …,x_p 的有关关系。

对于原则化数据能够证明：

$$r(y, x_j) = \sqrt{\lambda_i} \cdot u_i(j)$$

所以：

$$r(y_1, x_1) = \sqrt{\lambda_1}u_1(1)$$

$$r(y_1, x_2) = \sqrt{\lambda_1}u_1(2)$$

L L

$$r(y_1, x_p) = \sqrt{\lambda_1}u_1(p)$$

第一种主轴：

$$u_1 = (u_1(1), u_1(2), \mathrm{L}\ , u_1(p))'$$

由此可见，仅差一种常量倍$\sqrt{\lambda_1}$：

$(u_1(1)\ u_1(2)\ \mathrm{L}\ \ u_1(p))$ 是 y_1与 x_1，…,x_p 的有关系数。

所以，能够经过观察 $u_1 = (u_1(1), u_1(2), \mathrm{L}\ , u_1(p))'$ 来拟定y_1的含意。

例.管理期刊分类评估

$$r(\mathbf{y}_1, \mathbf{x}_j) = \sqrt{\lambda_1} u_1(j)$$

$$r(\mathbf{y}_2, \mathbf{x}_i) = \sqrt{\lambda_2} u_2(j)$$

Component Matrix[a]

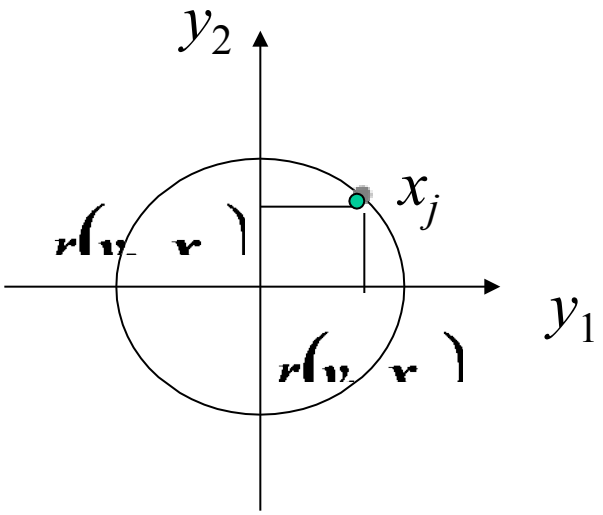
	Component	
	1	2
BYCISHU	.784	-3.5E-02
ZAIWENL	.102	.948
YZQIKAN	.802	-.249
NSFC	.513	.254

Extraction Method: Principal Component Ana

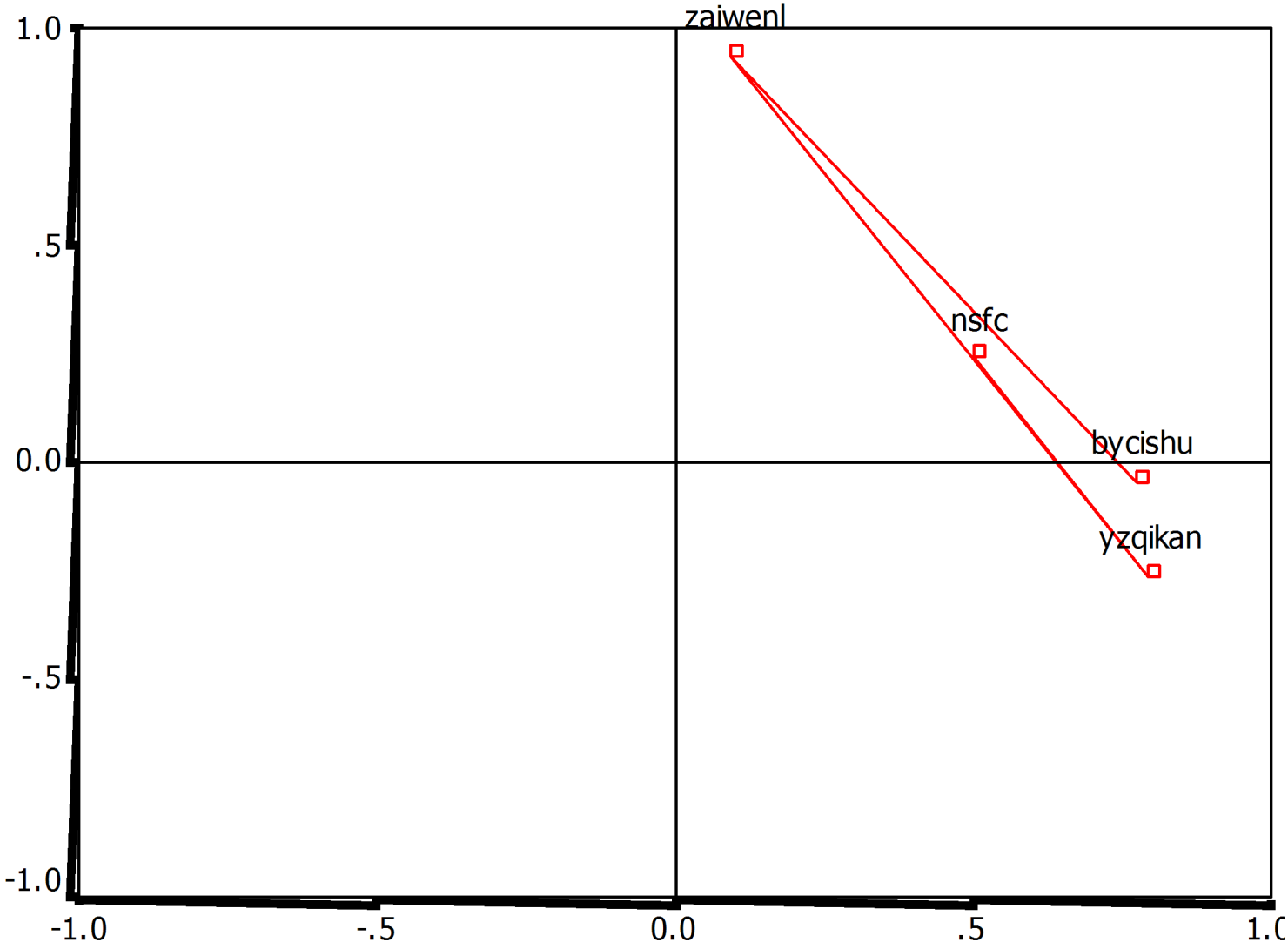
（2）有关圆图（Component Plot）

若 $m=2$

$$\begin{cases} \text{横坐标} r(y_1, x_j) \\ \text{纵坐标} r(y_2, x_j) \end{cases}$$



Component Plot



三. 判断“特异点”（e_k）

“特异点”：

在PCA中，若有e_k远离数据分布的平均水平，能够用“点对主成份方差的贡献”来测量。

如：

$$\mathrm{Var}(y_1)=\lambda_1=\frac{1}{n}\sum_{i=1}^{n}y_1^2(i)$$

则定义“e_i 对$\mathrm{Var}(y_1)$的贡献”为：

$$\frac{\frac{1}{n}y_1^2(i)}{\lambda_1}=CTR_1(i)$$

若CTR_1（BJ）=1/3，说明有1/3的$Var(y_1)$是由BJ提供的

所以，BJ是特异点。

一般地，定义“e_i 对Var(y_h)的贡献”**CTR$_h$(i)：**

$$CTR_h(i)=\frac{y_h^2(i)}{n\cdot\lambda_h}$$

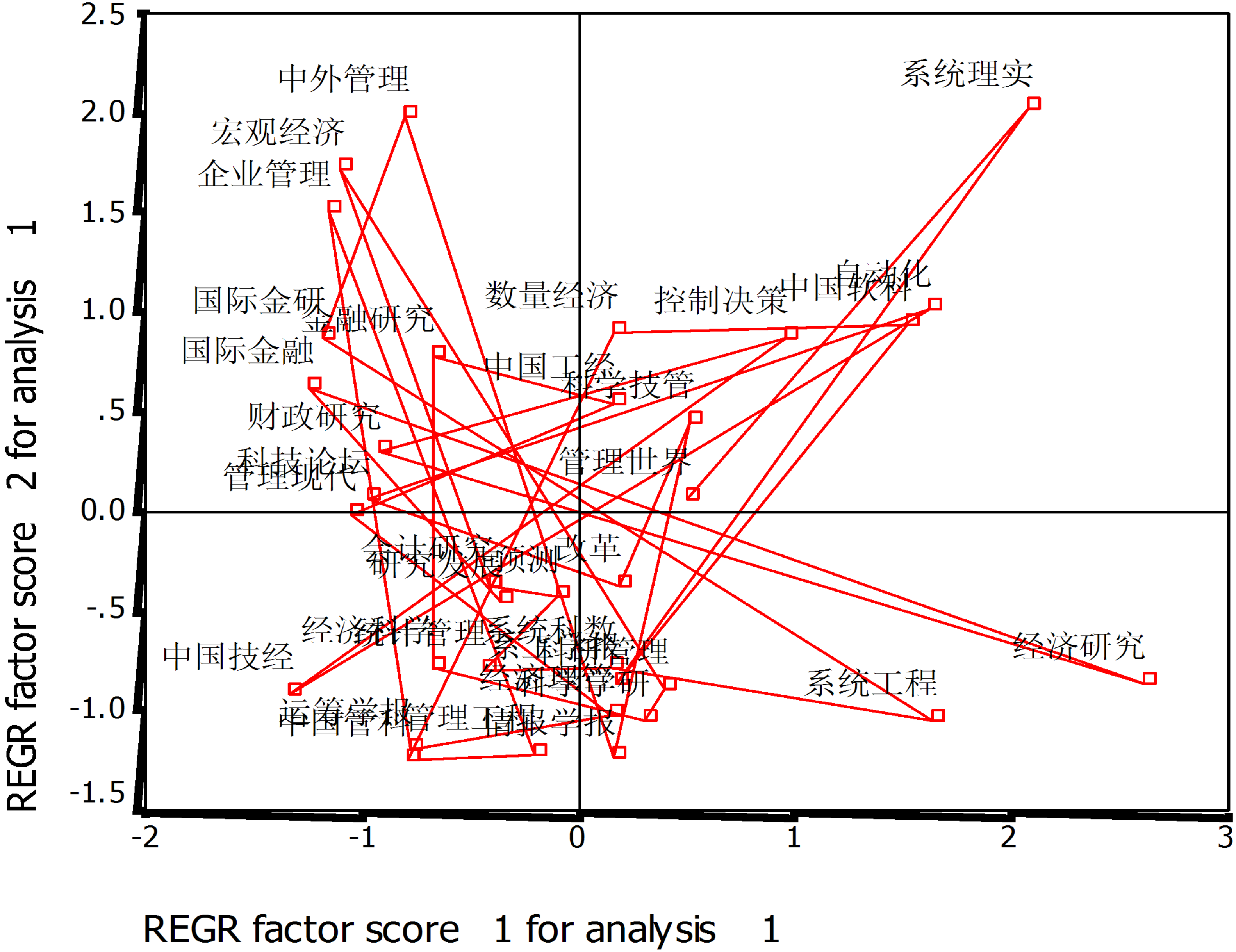
CTR(i)过大解原因：

（1）数据本身的特异性（BJ，SH, GZ, SZ, TJ）

（2）数据统计上的错误

处理措施：除去这些特异点，能够提升分析精度，图示也愈加清楚。

四. 主平面图



中外管理
系统理实
宏观经济
企业管理
国际金研
金融研究
数量经济
控制决策
中国软科
自动化
国际金融
中国工经
科学技管
财政研究
科技论坛
管理现代
管理世界
研究发展
预测
改革
中国技经
经济研究
系统工程
2.5
2.0
1.5
1.0
.5
0.0
-.5
-1.0
-1.5
-2
-1
0
1
2
3
REGR factor score 2 for analysis 1
REGR factor score 1 for analysis 1

10.6 利用主成份分析构造评估函数

PCA将一种高维变量系统有效的降至 1 维

例1：Kendall [英] 评估英国各地域农业生产水平。48个郡，10种农作物：小麦（x_1)、大麦（x_2)、燕麦（x_3)、土豆（x_4)、菜豆（x_5)、马铃薯（x_6)、萝卜（x_7)、饲料甜菜（x_8)、临时牧场干草（x_9)、永久牧场干草（x_{10})。（精度：47.6%）

$$Y_1 = 0.39\,x_1 + 0.37\,x_2 + 0.39\,x_3 + 0.27\,x_4 + 0.22\,x_5$$
$$+ 0.30\,x_6 + 0.32\,x_7 + 0.26\,x_8 + 0.24\,x_9 + 0.34\,x_{10}$$

第一主成份$\boldsymbol{y}_1$与 $\boldsymbol{x}_1$，…,$\boldsymbol{x}_{10}$ 均正有关。所以$\boldsymbol{y}_1$称为——“水平因子”，可用于评估排序。即：某个样本点在$\boldsymbol{y}_1$上取值很大时，它在$\boldsymbol{x}_1$，…,$\boldsymbol{x}_{10}$取值都会很大。