多元统计与矩阵分析

张锋　*8125345@qq.com*

中国地质大学, 计算机学院, 武汉

第7章
对应分析

内容

(1) 基本思想

(2) 原理

(3) 实例

(4) 优缺点

对应分析

- 因子分析只是分析原始变量的因子结构，找出决定原始变量的公共因子，从而使问题的分析简化和清晰。亦称为R型因子分析。
- 若对于样品进行因子分析，称为Q型因子分析。
- 我们对数据同时进行R和Q型因子分析，将他们统一起来，则是对应分析。
- 对应分析可以把变量和样品的载荷反映在相同的公因子轴，把众多的样品和变量同时作到一张图上展示。

内容

(1) 基本思想

(2) 原理

(3) 实例

(4) 优缺点

对应分析的原理

R型因子分析和Q型因子分析是反映一个整体的不同侧面，R型因子分析是从列来讨论（对变量），Q型因子分析是从行来讨论（对样品），因此他们之间存在内在的联系。

设原始数据矩阵为：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p}$$

由于因子分析都是基于协方差矩阵或相关系数矩阵完成的，所以必须从变量和样品的协方差矩阵入手来进行分析。

$$\boldsymbol{X}^{*}=\begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-x_2 & \cdots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-x_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix}_{n\times p}$$

$$\boldsymbol{X}^{*\prime}\cdot\boldsymbol{X}^{*}$$

$$=\begin{pmatrix} x_{11}-\bar{x}_1 & x_{21}-\bar{x}_1 & \cdots & x_{n1}-\bar{x}_1 \\ x_{12}-\bar{x}_2 & x_{22}-x_2 & \cdots & x_{n2}-x_2 \\ \vdots & \vdots & & \vdots \\ x_{1p}-\bar{x}_p & x_{2p}-\bar{x}_p & \cdots & x_{np}-\bar{x}_p \end{pmatrix}\times\begin{pmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-x_2 & \cdots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-x_2 & \cdots & x_{np}-\bar{x}_p \end{pmatrix}$$

变量的积叉矩阵

$$\boldsymbol{\Sigma}_R = (\boldsymbol{X}^*)'\boldsymbol{X}^* \qquad (p \times p)$$

样品的积叉矩阵

$$\boldsymbol{\Sigma}_Q = \boldsymbol{X}^*(\boldsymbol{X}^*)' \qquad (n \times n)$$

显然，变量和样品的积叉矩阵的阶数不同，一般来说，他们的非零特征根也不一样，那么能否将观测值做变换。

$$\boldsymbol{X} \rightarrow \boldsymbol{Z}$$

$\boldsymbol{Z}'\boldsymbol{Z}$和$\boldsymbol{Z}\boldsymbol{Z}'$具有相同的特征根。

规格化矩阵

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p}$$

$x_{i.}$为行和 $\sum_{j=1}^{p} x_{ij}$，

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{matrix} x_{1.} \\ x_{2.} \\ \vdots \\ x_{n.} \end{matrix}$$

$$x_{.1} \quad x_{.2} \quad \cdots \quad x_{.p} \quad x_{..}$$

$x_{.j}$为列和 $\sum_{i=1}^{n} x_{ij}$

$x_{..}$为总和 $\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}$

$$p_{ij} = x_{ij}/x_{..}$$

$$\boldsymbol{X} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1p} \\ p_{21} & p_{22} & \cdots & p_{2p} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{np} \end{bmatrix}_{n\times p}$$

我们可以把 p_{ij} 解释成概率，因为所有的元素之和为1。

行和：$p_{i.} = \sum_{j=1}^{p} p_{ij}$　　列和：$p_{.j} = \sum_{i=1}^{n} p_{ij}$

$$\because \quad \frac{p_{ij}}{p_{i.}} = \frac{x_{ij}/x_{..}}{\sum_{j=1}^{p} p_{ij}} = \frac{x_{ij}/x_{..}}{\sum_{j=1}^{p} x_{ij}/x_{..}} = \frac{x_{ij}}{x_{i.}}$$

$$\therefore \left[\frac{p_{i1}}{p_{i.}} \quad \frac{p_{i2}}{p_{i.}} \quad \cdots \quad \frac{p_{ip}}{p_{i.}}\right] = \left[\frac{x_{i1}}{x_{i.}} \quad \frac{x_{i2}}{x_{i.}} \quad \cdots \quad \frac{x_{ip}}{x_{i.}}\right] i = 1,2,3,\cdots,n$$

称为行形象。

例如：考察某一文章中各种词汇出现的次数，词汇分为如下种类：

n=名词， v=动词， a=形容词， av=副词，

l=冠词， o=其它

x_{ij} 表示在第i 篇文章中属于j 种词汇的次数。

	n	v	a	av	l	o	
i'	80	30	25	15	60	55	$x_{i'.}$ =265
i''	160	60	50	30	120	110	$x_{i''.}$=530

它们的行形象：

i'	0.30	0.11	0.09	0.06	0.23	0.21
i''	0.30	0.11	0.09	0.06	0.23	0.21

这两行的形象相同，由此可以断定这两文的用词手法相同。

$$N(\boldsymbol{R}) = \begin{bmatrix} \frac{p_{11}}{p_{1.}} & \frac{p_{12}}{p_{1.}} & \cdots & \frac{p_{1p}}{p_{1.}} \\ \frac{p_{21}}{p_{2.}} & \frac{p_{22}}{p_{2.}} & \cdots & \frac{p_{2p}}{p_{2.}} \\ \vdots & \vdots & & \vdots \\ \frac{p_{n1}}{p_{n.}} & \frac{p_{n2}}{p_{n.}} & \cdots & \frac{p_{np}}{p_{n.}} \end{bmatrix}$$

行形象点集

第j个列变量的期望：

$$E(\frac{p_{ij}}{p_{i.}}) = \sum_{i=1}^{n} \frac{p_{ij}}{p_{i.}} \times p_{i.} = p_{.j}, j = 1,2,\cdots,p$$

因为原始变量的数量等级可能不同，所以为了尽量减少各变量尺度差异，将行形象中的(各列元素）都除以其期望的平方根。得矩阵D(***R***)

$$D(\boldsymbol{R})=\begin{bmatrix} \dfrac{p_{11}}{p_{1.}\sqrt{p_{.1}}} & \dfrac{p_{12}}{p_{1.}\sqrt{p_{.2}}} & \cdots & \dfrac{p_{1p}}{p_{1.}\sqrt{p_{.p}}} \\ \dfrac{p_{21}}{p_{2.}\sqrt{p_{.1}}} & \dfrac{p_{22}}{p_{2.}\sqrt{p_{.2}}} & \cdots & \dfrac{p_{2p}}{p_{2.}\sqrt{p_{.p}}} \\ \vdots & \vdots & & \vdots \\ \dfrac{p_{n1}}{p_{n.}\sqrt{p_{.1}}} & \dfrac{p_{n2}}{p_{n.}\sqrt{p_{.2}}} & \cdots & \dfrac{p_{np}}{p_{n.}\sqrt{p_{.p}}} \end{bmatrix}$$

第j个列变量的期望：

$$E\left(\frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}}\right)=\sum_{i=1}^{n}\frac{p_{ij}}{p_{i.}\sqrt{p_{.j}}}\times p_{i.}=\frac{1}{\sqrt{p_{.j}}}p_{.j}=\sqrt{p_{.j}},j=1,2,\cdots,p$$

第k个变量与第j个变量的协方差：

$$s_{kj} = \sum_{a=1}^{n} \left[\frac{p_{ak}}{p_{a.}\sqrt{p_{.k}}} - \sqrt{p_{.k}} \right] \left[\frac{p_{aj}}{p_{a.}\sqrt{p_{.j}}} - \sqrt{p_{.j}} \right] \times p_{a.}$$

$$= \sum_{a=1}^{n} \left[\frac{p_{ak}}{\sqrt{p_{a.}}\sqrt{p_{.k}}} - \sqrt{p_{.k}}\sqrt{p_{a.}} \right] \left[\frac{p_{aj}}{\sqrt{p_{a.}}\sqrt{p_{.j}}} - \sqrt{p_{.j}}\sqrt{p_{a.}} \right]$$

$$= \sum_{a=1}^{n} \left[\frac{p_{ak} - p_{a.}p_{.k}}{\sqrt{p_{a.}}\sqrt{p_{.k}}} \right] \left[\frac{p_{aj} - p_{a.}p_{.j}}{\sqrt{p_{a.}}\sqrt{p_{.j}}} \right]$$

$$= \sum_{a=1}^{n} z_{ak} z_{aj}$$

$$z_{ak}=\frac{p_{ak}-p_{a.}p_{.k}}{\sqrt{p_{a.}p_{.k}}}=\frac{x_{ak}-x_{a.}x_{.k}}{\sqrt{x_{a.}x_{.k}}}$$

令$\boldsymbol{Z}$为z_{ij}所组成的矩阵$\boldsymbol{Z}=(z_{ij})_{n\times p}$，则$\boldsymbol{A}=\boldsymbol{Z}'\boldsymbol{Z}$

称$\left[\frac{p_{1j}}{p_{.j}}\quad\frac{p_{2j}}{p_{.j}}\quad\cdots\quad\frac{p_{nj}}{p_{.j}}\right]=\left[\frac{x_{1j}}{x_{.j}}\quad\frac{x_{2j}}{x_{.j}}\quad\cdots\quad\frac{x_{nj}}{x_{.j}}\right]j=1,2,3,\cdots,p$

为列形象。

$$N(Q)=\begin{bmatrix} \frac{p_{11}}{p_{.1}} & \frac{p_{12}}{p_{.2}} & \cdots & \frac{p_{1p}}{p_{.p}} \\ \frac{p_{21}}{p_{.1}} & \frac{p_{22}}{p_{.2}} & \cdots & \frac{p_{2p}}{p_{.p}} \\ \vdots & \vdots & & \vdots \\ \frac{p_{n1}}{p_{.1}} & \frac{p_{n2}}{p_{.2}} & \cdots & \frac{p_{np}}{p_{.p}} \end{bmatrix}$$

第i个行变量的期望：

$$E(\frac{p_{ij}}{p_{.j}})=\sum_{j=1}^{p}\frac{p_{ij}}{p_{.j}}.p_{.j}=p_{i.}$$

因为原始变量的数量等级可能不同，所以为了尽量减少各变量尺度差异，将列形象中的各行元素均除以其期望的平方根。得矩阵D(Q)

$$D(Q)=\begin{bmatrix} \dfrac{p_{11}}{p_{.1}\sqrt{p_{1.}}} & \dfrac{p_{12}}{p_{.2}\sqrt{p_{1.}}} & \cdots & \dfrac{p_{1p}}{p_{.p}\sqrt{p_{1.}}} \\ \dfrac{p_{21}}{p_{.1}\sqrt{p_{2.}}} & \dfrac{p_{22}}{p_{.2}\sqrt{p_{2.}}} & \cdots & \dfrac{p_{2p}}{p_{.p}\sqrt{p_{2.}}} \\ \vdots & \vdots & & \vdots \\ \dfrac{p_{n1}}{p_{.1}\sqrt{p_{n.}}} & \dfrac{p_{n2}}{p_{.2}\sqrt{p_{n.}}} & \cdots & \dfrac{p_{np}}{p_{.p}\sqrt{p_{n.}}} \end{bmatrix}$$

第i个行变量的期望：

$$E\left(\frac{p_{ij}}{p_{.j}\sqrt{p_{i.}}}\right)=\sum_{j=1}^{p}\frac{p_{ij}}{p_{.j}\sqrt{p_{i.}}}\cdot p_{.j}=\sqrt{p_{i.}}$$

第k个样品与第 l 个样品的协方差：

$$b_{kl}=\sum_{i=1}^{p}\left[\frac{p_{ki}}{p_{.i}\sqrt{p_{k.}}}-\sqrt{p_{k.}}\right]\left[\frac{p_{li}}{p_{.i}\sqrt{p_{l.}}}-\sqrt{p_{l.}}\right]\times p_{.i}$$

$$=\sum_{i=1}^{p}\left[\frac{p_{ki}}{\sqrt{p_{.i}}\sqrt{p_{k.}}}-\sqrt{p_{.i}}\sqrt{p_{k.}}\right]\left[\frac{p_{li}}{\sqrt{p_{.i}}\sqrt{p_{l.}}}-\sqrt{p_{.i}}\sqrt{p_{l.}}\right]$$

$$=\sum_{i=1}^{p}\left[\frac{p_{ki}-p_{k.}p_{.i}}{\sqrt{p_{.i}}\sqrt{p_{k.}}}\right]\left[\frac{p_{li}-p_{i.}p_{l.}}{\sqrt{p_{l.}}\sqrt{p_{.i}}}\right]$$

$$=\sum_{i=1}^{p}z_{ki}z_{li}$$

令$\boldsymbol{Z}$为z_{ij}所组成的矩阵$\boldsymbol{Z}=(z_{ij})_{n\times p}$，则$\boldsymbol{B}=\boldsymbol{ZZ'}$

将原矩阵变换成矩阵$\boldsymbol{Z}$，很容易求出$\boldsymbol{A}$和$\boldsymbol{B}$存在着的简单对应关系。由特征根和特征向量的性质，A和B有相同的非零特征根。

设 λ_k是$\boldsymbol{A}=\boldsymbol{Z'Z}$的非零特征根，则 $\boldsymbol{Z'Zu_k}=\lambda_k\boldsymbol{u_k}$

在上式的两边都左乘$\boldsymbol{Z}$，则

$$\boldsymbol{ZZ'(Zu_k)}=\lambda_k\boldsymbol{(Zu_k)}$$

可见 λ_k 也是$\boldsymbol{ZZ'}$的特征根，相应的特征向量是 $\boldsymbol{Zu_k}$

因此将原始数据矩阵$\boldsymbol{X}$变换成矩阵$\boldsymbol{Z}$，则变量和样品的协差阵分别可表示为$\boldsymbol{A}=\boldsymbol{Z}'\boldsymbol{Z}$和$\boldsymbol{B}=\boldsymbol{Z}\boldsymbol{Z}'$，$\boldsymbol{A}$和$\boldsymbol{B}$具有相同的非零特征值，相应的特征向量有很密切的关系。

这样就可以用相同的因子轴去同时表示变量和样品，把变量和样品同时反映在具有相同坐标轴的因子平面上。

行列独立性检验

检验假设：

H0：行列变量相互独立

H1：行列变量相互不独立

检验统计量：$\chi^2 = T\sum_{i}^{n}\sum_{j=1}^{p}\frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$　　注意$T = x_{..} = \sum_{i=1}^{n}\sum_{j=1}^{p}x_{ij}$

当T充分大，χ^2近似服从自由度为(n-1)(p-1)的卡方分布，如果

$$\chi^2 > \chi_\alpha^2\,(n-1)(p-1)$$

则拒绝H0，否则接受。

行列独立性检验

令$\boldsymbol{Z}$为z_{ij}所组成的矩阵，$\boldsymbol{A}=\boldsymbol{Z}'\boldsymbol{Z}$ 则

$$\boldsymbol{Z}=(z_{ij})_{n\times p}$$

$$z_{ij}=\frac{p_{ij}-p_{i.}p_{.j}}{\sqrt{p_{i.}p_{.j}}}$$

总惯量：$\frac{\chi^2}{T}=\sum_{i}^{n}\sum_{j=1}^{p}\frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$

记$k=rank(\boldsymbol{Z})\leq \min(n-1,p-1)$，则

$$\boldsymbol{Z}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}'=\sum_{i=1}^{k}\boldsymbol{\lambda_i}\boldsymbol{u_i}\boldsymbol{v_i}'$$

$$\boldsymbol{U}=(\boldsymbol{u_1},\boldsymbol{u_2},\dots,\boldsymbol{u_k}),\ \boldsymbol{V}=(\boldsymbol{v_1},\boldsymbol{v_2},\dots,\boldsymbol{v_k}),\ \boldsymbol{\Lambda}=\boldsymbol{diag}(\boldsymbol{\lambda_1},\boldsymbol{\lambda_2},\dots,\boldsymbol{\lambda_k})$$

$\boldsymbol{u_1},\boldsymbol{u_2},\dots,\boldsymbol{u_k}$n维正交单位向量，$\boldsymbol{v_1},\boldsymbol{v_2},\dots,\boldsymbol{v_k}$p维正交单位向量。

$\boldsymbol{U}'\boldsymbol{U}=\boldsymbol{V}'\boldsymbol{V}=\boldsymbol{I}$，$\boldsymbol{\lambda_1},\boldsymbol{\lambda_2},\dots,\boldsymbol{\lambda_k}$是$\boldsymbol{Z}$的$\boldsymbol{k}$个奇异值。

$\boldsymbol{\lambda_1^2}$，$\boldsymbol{\lambda_2^2}$，…，$\boldsymbol{\lambda_k^2}$是$\boldsymbol{Z}\boldsymbol{Z}'$的正特征值。

$$\sum_{i}^{n}\sum_{j=1}^{p}\frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}}=tr(\boldsymbol{Z}\boldsymbol{Z}')=\sum_{i=1}^{k}\boldsymbol{\lambda_i^2}$$

内容

(1) 基本思想

(2) 原理

(3) 实例

(4) 优缺点

对应图

设$\lambda_1 \geq \lambda_2 \cdots \geq \lambda_i (0<i<\min(n,p))$为矩阵**A**和**B**的非零特征根，取前两个，其相应的特征向量为

$$\mathbf{u_1} = [u_{11} \quad u_{21} \quad \cdots \quad u_{p1}]'$$

$$\mathbf{u_2} = [u_{12} \quad u_{22} \quad \cdots \quad u_{p2}]'$$

$$\mathbf{v_1} = [v_{11} \quad v_{21} \quad \cdots \quad v_{n1}]'$$

$$\mathbf{v_2} = [v_{12} \quad v_{22} \quad \cdots \quad v_{n2}]'$$

因子载荷矩阵的含义是原始变量与公共因子之间的相关系数，所以如果我们构造一个平面直角坐标系，将第一公共因子的载荷与第二个公共因子的载荷看成平面上的点，在坐标系中绘制散点图，则构成对应图。

对应分析的步骤

1、获取对应分析数据

首先要规定研究的目的，然后选择对应分析中所需数据，应该包括的背景资料。

2、建立列联表

3、对应分析

4、对应图并解释结果的意义。

对应分析图解读

1、总体观察

2、观察邻近区域

3、向量分析——偏好排序

从中心向任意点连线--向量，例如从中心向“工资性收入”作向量，然后让所有的地区往这条向量及延长线上作垂线，垂点越靠近向量正向的表示工资性收入比重越高。

4-向量的夹角——余弦定理

从向量夹角的角度看不同地区或不同收入来源之间的相似情况，从余弦定理的角度看相似性！

5、从距离中的位置看：越靠近中心，越没有特征，越远离中心，说明特征越明显。各个类别之间的距离表示相对密切关系

案例1

例 7.3 交叉列联表(表7-6)总结了260个消费者对于四种不同软件的性能评价，假设B软件是公司自己的产品，试分析消费者对B软件的评价如何，与其他竞争对手的产品形象有何不同。

表 7-6　消费者对四种软件的性能评价

软件性能 / 软件名称	易学	操作简单	运行速度快	可视化	算法丰富	界面友好	扩展能力强
A 软件	140	120	130	100	140	110	130
B 软件	160	150	180	180	160	160	160
C 软件	170	180	110	115	120	140	100
D 软件	100	150	200	150	180	120	170

案例1

例 7.3 交叉列联表(表7-6)总结了260个消费者对于四种不同软件的性能评价，假设B软件是公司自己的产品，试分析消费者对B软件的评价如何，与其他竞争对手的产品形象有何不同。

表 7-6　消费者对四种软件的性能评价

软件性能 / 软件名称	易学	操作简单	运行速度快	可视化	算法丰富	界面友好	扩展能力强
A 软件	140	120	130	100	140	110	130
B 软件	160	150	180	180	160	160	160
C 软件	170	180	110	115	120	140	100
D 软件	100	150	200	150	180	120	170

案例1

表 7-6 汇集了消费者对四种不同软件性能评价的所有信息，可以初步了解消费者对不同软件的形象认知情况，但是很难对品牌和性能之间的关联有一个整体的认识，无法整体把握消费者对于不同软件的形象认知情况。为此，我们进行对应分析。

将表 7-6 中的数据除以总和4 025，得到对应矩阵，见表 7-7。

表 7-7　消费者对四种软件性能评价的对应矩阵

软件性能 / 软件名称	易学	操作简单	运行速度快	可视化	算法丰富	界面友好	扩展能力强
A 软件	0.034 8	0.029 8	0.032 3	0.024 8	0.034 8	0.027 3	0.032 3
B 软件	0.039 8	0.037 3	0.044 7	0.044 7	0.039 8	0.039 8	0.039 8
C 软件	0.042 2	0.044 7	0.027 3	0.028 6	0.029 8	0.034 8	0.024 8
D 软件	0.024 8	0.037 3	0.049 7	0.037 3	0.044 7	0.029 8	0.042 2

案例1

表 7-7　消费者对四种软件性能评价的对应矩阵

软件性能 / 软件名称	易学	操作简单	运行速度快	可视化	算法丰富	界面友好	扩展能力强
A 软件	0.034 8	0.029 8	0.032 3	0.024 8	0.034 8	0.027 3	0.032 3
B 软件	0.039 8	0.037 3	0.044 7	0.044 7	0.039 8	0.039 8	0.039 8
C 软件	0.042 2	0.044 7	0.027 3	0.028 6	0.029 8	0.034 8	0.024 8
D 软件	0.024 8	0.037 3	0.049 7	0.037 3	0.044 7	0.029 8	0.042 2

可计算得行轮廓的矩阵为

$$N(\boldsymbol{R})=\begin{bmatrix} 0.160\ 9 & 0.137\ 9 & 0.149\ 4 & 0.114\ 9 & 0.160\ 9 & 0.126\ 4 & 0.149\ 4 \\ 0.139\ 1 & 0.130\ 4 & 0.156\ 5 & 0.156\ 5 & 0.139\ 1 & 0.139\ 1 & 0.139\ 1 \\ 0.181\ 8 & 0.192\ 5 & 0.117\ 6 & 0.123\ 0 & 0.128\ 3 & 0.149\ 7 & 0.107\ 0 \\ 0.093\ 5 & 0.140\ 2 & 0.186\ 9 & 0.140\ 2 & 0.168\ 2 & 0.112\ 1 & 0.158\ 9 \end{bmatrix}$$

可计算得列轮廓的矩阵为

$$N(\boldsymbol{Q})=\begin{bmatrix} 0.245\ 6 & 0.200\ 0 & 0.209\ 7 & 0.183\ 5 & 0.233\ 3 & 0.207\ 5 & 0.232\ 1 \\ 0.280\ 7 & 0.250\ 0 & 0.290\ 3 & 0.330\ 3 & 0.266\ 7 & 0.301\ 9 & 0.285\ 7 \\ 0.298\ 2 & 0.300\ 0 & 0.177\ 4 & 0.211\ 0 & 0.200\ 0 & 0.264\ 2 & 0.178\ 6 \\ 0.175\ 4 & 0.250\ 0 & 0.322\ 6 & 0.275\ 2 & 0.300\ 0 & 0.226\ 4 & 0.303\ 6 \end{bmatrix}$$

案例1

经计算，奇异值、主惯量以及贡献率等的计算结果见表 7-8，总惯量的79.58% 可由第一维来解释，前两维解释了90.90% 的总惯量，即解释了列联表数据90.90% 的变差。

表 7-8　奇异值、主惯量以及贡献率

项目	奇异值	主惯量	卡方	贡献率	累积贡献率
1	0.135 88	0.018 46	74.316 6	79.58	79.58
2	0.051 25	0.002 63	10.571 4	11.32	90.90
3	0.045 95	0.002 11	8.498 7	9.10	100.00
		0.023 20	93.386 7		

案例1

行变量和列变量前两维的坐标矩阵分别如表 7-9 和 7-10 所示。

表 7-9　行坐标

软件名称	Dim1	Dim2
A 软件	0.013 2	0.063 7
B 软件	−0.021 5	−0.075 0
C 软件	0.211 5	0.007 6
D 软件	−0.172 4	0.022 2

表 7-10　列坐标

软件性能	Dim1	Dim2
易学	0.221 0	0.014 6
操作简单	0.129 6	0.035 4
运行速度快	−0.158 8	0.001 6
可视化	−0.055 3	−0.104 9
算法丰富	−0.088 9	0.059 2
界面友好	0.096 2	−0.046 7

案例1

将各行变量和列变量的值置于同一坐标系中，构成对应分析图，如图 7-2 所示。

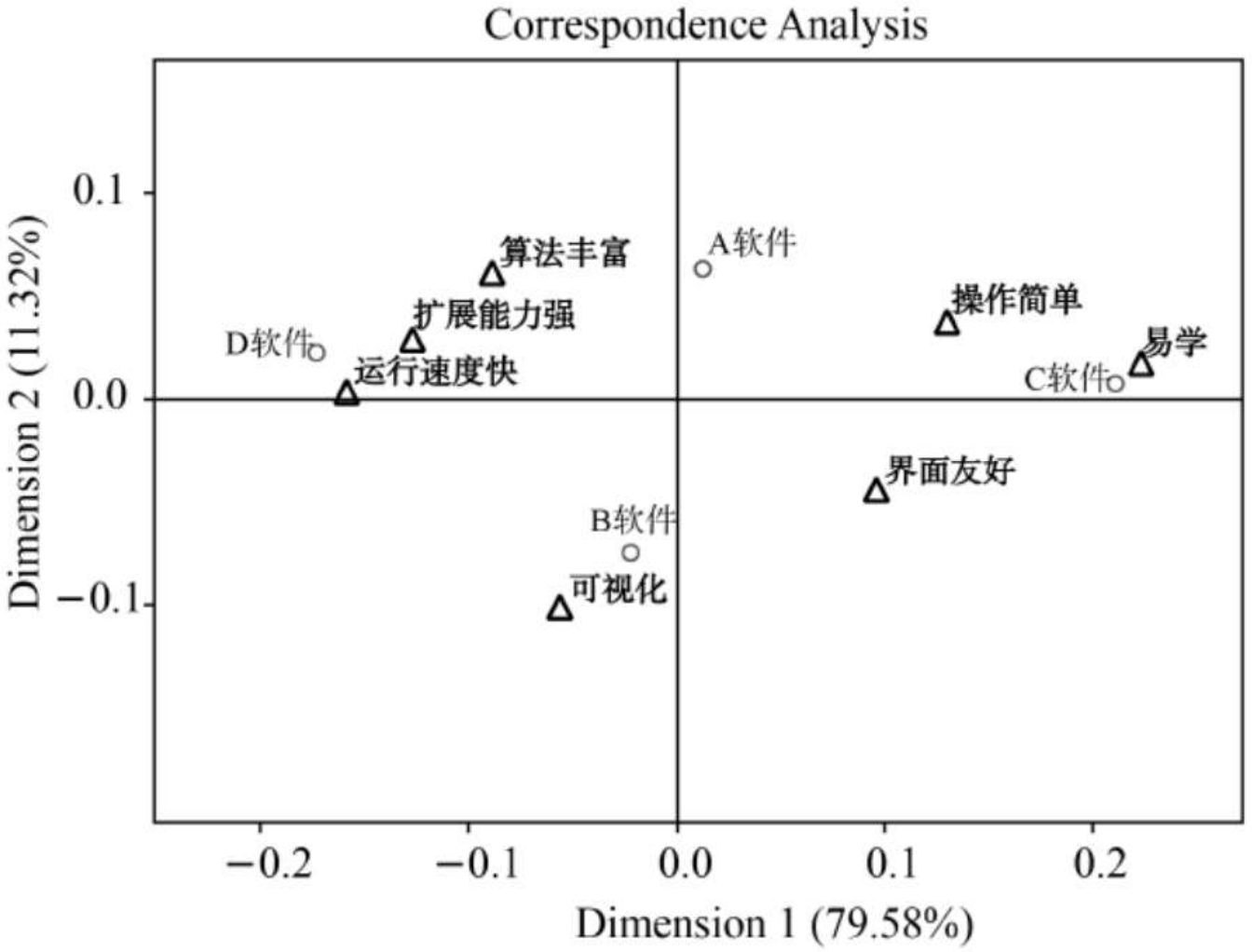


图 7-2　消费者对四种软件性能评价的对应分析图

案例1

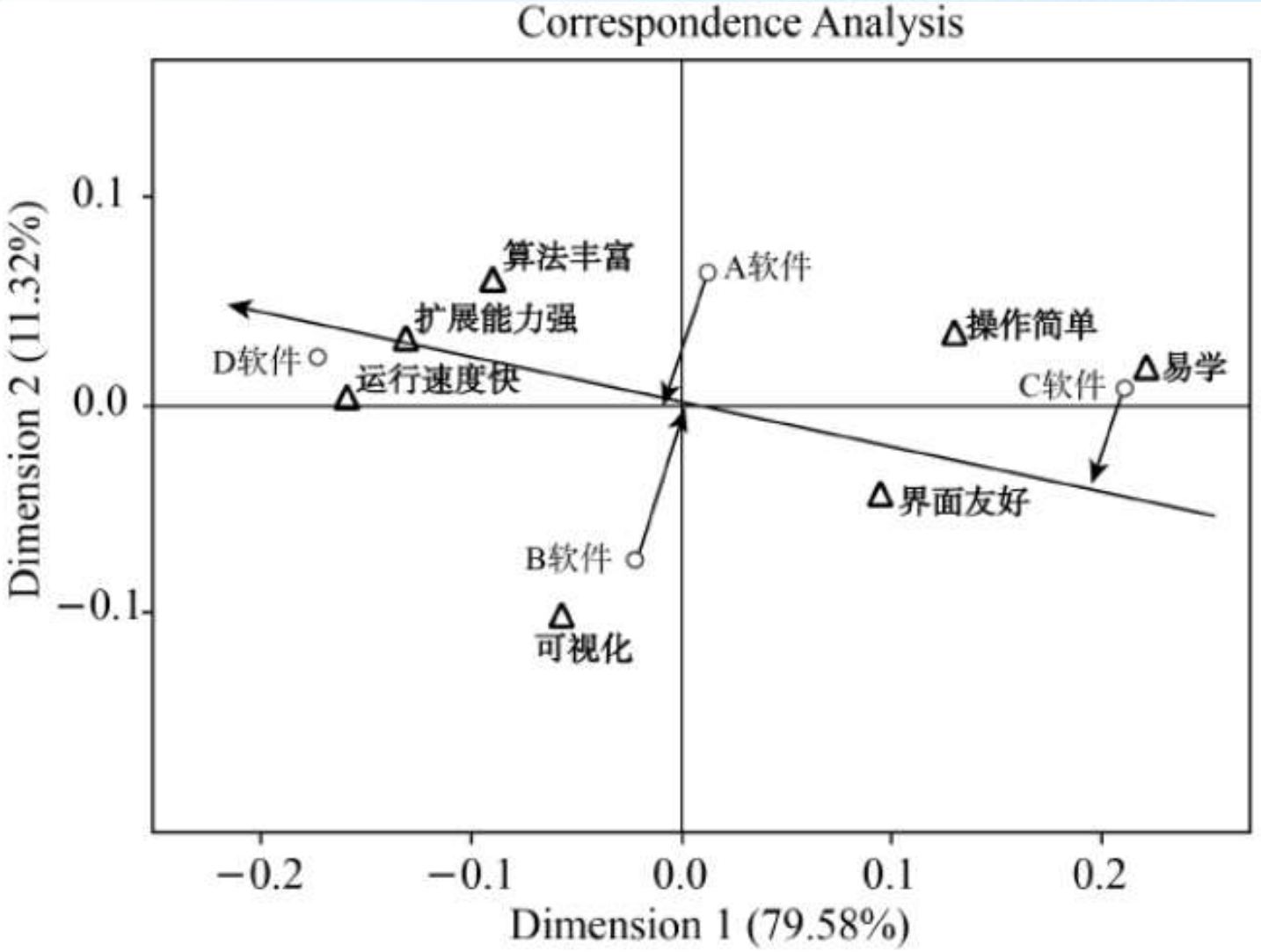


图 7-3　消费者对四种软件性能评价的对应分析图

案例2

表 7-11　学生体质及体测成绩

体质状况 / 体测成绩	偏瘦	正常	超重	肥胖
不及格	92	527	225	121
及格	212	4 671	451	63
良好	3	777	11	0
优秀	0	47	0	0

案例2

【输出 7-1】

Row Profiles				
	偏瘦	正常	超重	肥胖
不及格	0.09534	0.54611	0.23316	0.12539
及格	0.03928	0.86548	0.08356	0.01167
良好	0.00379	0.98230	0.01391	0.00000
优秀	0.00000	1.00000	0.00000	0.00000

输出 7-1 是行轮廓，是列联表中每一行数值除以行和得到的结果。

【输出 7-2】

Column Profiles				
	偏瘦	正常	超重	肥胖
不及格	0.299674	0.087512	0.327511	0.657609
及格	0.690554	0.775656	0.656477	0.342391
良好	0.009772	0.129027	0.016012	0.000000
优秀	0.000000	0.007805	0.000000	0.000000

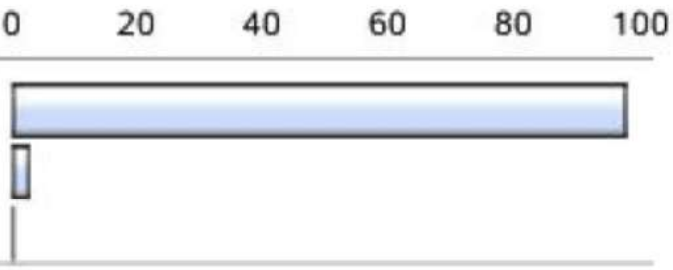
输出 7-2 是列轮廓，是列联表中每一列数值除以列和得到的结果。

案例2

【输出 7-3】

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.35244	0.12421	894.345	97.59	97.59
0.05540	0.00307	22.099	2.41	100.00
0.00110	0.00000	0.009	0.00	100.00
	0.12729	916.452	100.00	



Degrees of Freedom = 9

输出7-3是各维汇总表，其中Singular是奇异值，Principal inertia是主惯量，Percent是惯量的百分比，最后一列数据是惯量占比的累计值。从中可以看出，第一维和第二维的惯量比例占总惯量的100%，因此前两维解释了列联表数据100%的变异。

案例2

【输出 7-4】

Row Coordinates		
	Dim1	Dim2
不及格	0.8680	-0.0348
及格	-0.0975	0.0281
良好	-0.3698	-0.1389
优秀	-0.4068	-0.1732

输出 7-4 是 R 型因子分析中的公因子载荷，表示“样品”投影到公共因子 Dim1 和 Dim2 的坐标值(行坐标)。

在以 Dim1 为横坐标、Dim2 为纵坐标的直角坐标系内，每个成绩等级和每种体质状况就是一个点，例如不及格的坐标(0.868 0，－0.034 8) 在第四象限。

案例2

【输出 7-5】

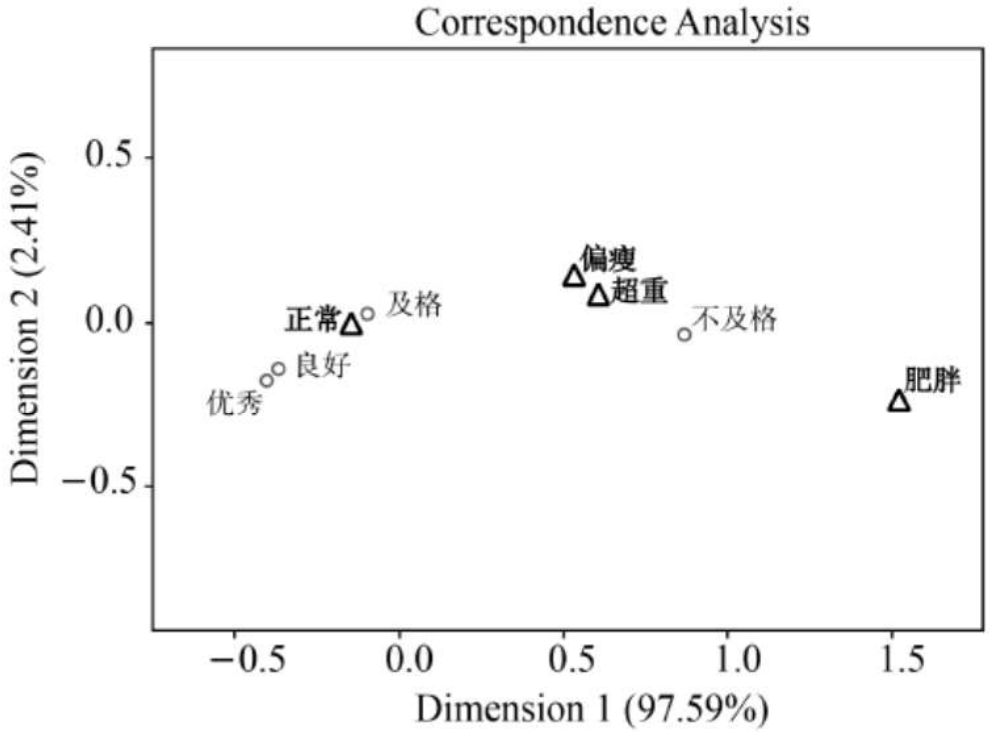
Column Coordinates		
	Dim1	Dim2
偏瘦	0.5369	0.1373
正常	-0.1434	-0.0096
超重	0.6083	0.0869
肥胖	1.5250	-0.2396

输出 7-5 是 Q 型因子分析中的公因子载荷，表示变量投影到公共因子 Dim1 和 Dim2 的坐标值(列坐标)。

在以 Dim1 为横坐标、Dim2 为纵坐标的直角坐标系内，每种体质状况就是一个点，例如偏瘦(0.536 9，0.137 3) 在第一象限。

案例2

【输出 7-6】



输出 7-6 是对应分析结果图，根据对应分析的思路，可以通过观察邻近区域进行关联性分析。

案例3

表 7-12　四种居民收入来源指标数据

单位：元

地区	工资性收入	经营净收入	财产净收入	转移净收入
北京	35 216.6	1 408.3	9 305.9	11 299.0
天津	23 165.0	3 262.2	3 504.9	7 090.3
河北	13 003.5	3 210.8	1 467.3	3 802.6
山西	11 957.1	2 624.1	1 227.9	4 610.9
内蒙古	13 899.7	6 363.8	1 287.6	4 661.2
辽宁	14 596.2	4 881.9	1 342.7	7 014.6
吉林	10 631.3	4 712.7	898.7	5 125.6
黑龙江	10 318.8	4 499.3	993.0	5 394.8
上海	34 365.4	1 532.6	9 030.1	14 059.9
江苏	20 399.2	4 994.2	3 238.6	6 392.1
浙江	24 137.3	7 123.4	4 741.6	6 043.4
安徽	11 920.9	4 878.9	1 227.7	3 835.8

案例3

【输出 7-7】

Row Profiles				
	工资性收入	经营净收入	财产净收入	转移净收入
北京	0.615354	0.024608	0.162606	0.197432
天津	0.625702	0.088114	0.094670	0.191514
河北	0.605259	0.149449	0.068297	0.176995
山西	0.585558	0.128506	0.060132	0.225803
内蒙古	0.530274	0.242779	0.049122	0.177825
辽宁	0.524375	0.175385	0.048237	0.252003
吉林	0.497527	0.220546	0.042058	0.239869
黑龙江	0.486600	0.212172	0.046827	0.254401
上海	0.582583	0.025982	0.153084	0.238352
江苏	0.582433	0.142593	0.092468	0.182506
浙江	0.574073	0.169420	0.112773	0.143734
安徽	0.545247	0.223155	0.056153	0.175445
福建	0.578417	0.186374	0.096017	0.139192
江西	0.569782	0.170706	0.063409	0.196102
山东	0.576766	0.218812	0.068002	0.136420
河南	0.501145	0.226797	0.061344	0.210714
湖北	0.497980	0.217084	0.063219	0.221718
湖南	0.512340	0.194060	0.070416	0.223165

案例3

【输出 7-8】

Column Profiles				
	工资性收入	经营净收入	财产净收入	转移净收入
北京	0.078043	0.010758	0.140973	0.072678
天津	0.051336	0.024921	0.053095	0.045607
河北	0.028817	0.024528	0.022228	0.024459
山西	0.026498	0.020046	0.018601	0.029658
内蒙古	0.030803	0.048615	0.019506	0.029982
辽宁	0.032347	0.037294	0.020340	0.045120
吉林	0.023560	0.036002	0.013614	0.032969
黑龙江	0.022867	0.034371	0.015043	0.034701
上海	0.076157	0.011708	0.136795	0.090437
江苏	0.045207	0.038152	0.049061	0.041116
浙江	0.053491	0.054418	0.071829	0.038873
安徽	0.026418	0.037271	0.018598	0.024673
福建	0.038516	0.042781	0.043706	0.026902
江西	0.027819	0.028731	0.021163	0.027790
山东	0.034421	0.045015	0.027742	0.023631
河南	0.022400	0.034946	0.018744	0.027338
湖北	0.026218	0.039398	0.022752	0.033881
湖南	0.026231	0.034251	0.024644	0.033163
广东	0.051087	0.033772	0.054566	0.012398
广西	0.021760	0.038304	0.017747	0.025084
海南	0.029632	0.032740	0.020684	0.022712
重庆	0.027931	0.030685	0.023109	0.038638
四川	0.022191	0.032572	0.020646	0.031773
贵州	0.019153	0.029251	0.013684	0.021336

案例3

【输出 7-9】

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.21306	0.04540	36481.6	72.51	72.51
0.11371	0.01293	10390.0	20.65	93.17
0.06541	0.00428	3438.0	6.83	100.00
	0.06260	50309.6	100.00	

Degrees of Freedom = 90

案例3

【输出 7-10】

Row Coordinates		
	Dim1	Dim2
北京	-0.4467	0.0021
天津	-0.1914	-0.0026
河北	-0.0120	-0.0373
山西	-0.0335	0.0875
内蒙古	0.2324	-0.0373
辽宁	0.0997	0.1522
吉林	0.2096	0.1203
黑龙江	0.1864	0.1561
上海	-0.4172	0.1061
江苏	-0.0661	-0.0281
浙江	-0.0487	-0.1308
安徽	0.1754	-0.0434
福建	0.0164	-0.1397
江西	0.0494	0.0100
山东	0.1358	-0.1429
河南	0.1844	0.0432
湖北	0.1620	0.0709
湖南	0.0981	0.0745
广东	-0.1457	-0.3404
广西	0.2418	0.0052
海南	0.0878	-0.0892
重庆	0.0535	0.1415
四川	0.1393	0.1165

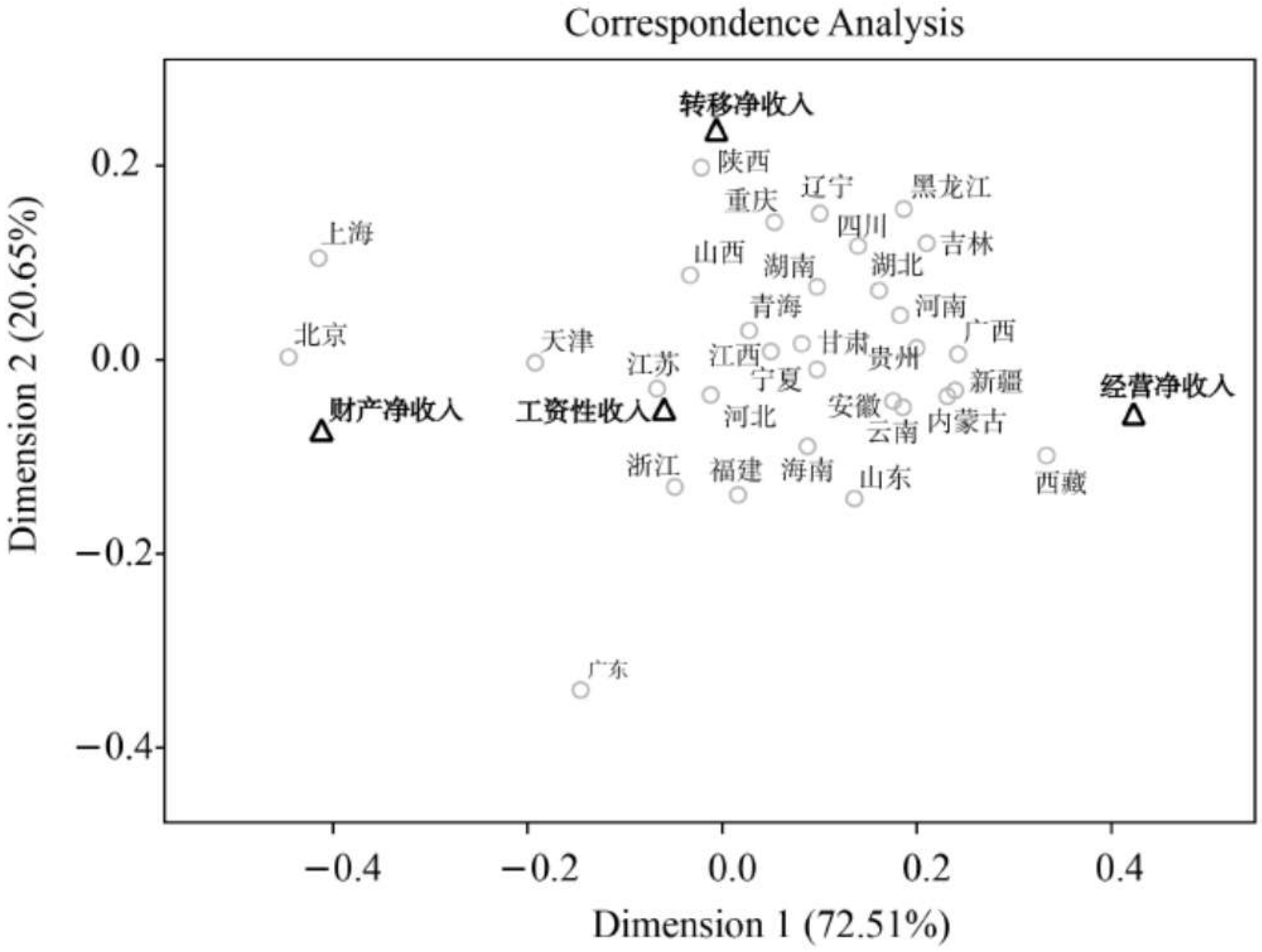
案例3

【输出 7-11】

Column Coordinates		
	Dim1	Dim2
工资性收入	-0.0609	-0.0517
经营净收入	0.4242	-0.0591
财产净收入	-0.4127	-0.0755
转移净收入	-0.0053	0.2318

案例3

【输出 7-12】


Correspondence Analysis
转移净收入
陕西
辽宁
黑龙江
重庆
四川
吉林
上海
山西
湖南
湖北
青海
河南
北京
天津
甘肃
广西
江苏
江西
贵州
宁夏
新疆
经营净收入
财产净收入
工资性收入
安徽
河北
云南
内蒙古
浙江
福建
海南
山东
西藏
广东
0.2
0.0
−0.2
−0.4
Dimension 2 (20.65%)
−0.4
−0.2
0.0
0.2
0.4
Dimension 1 (72.51%)

案例3

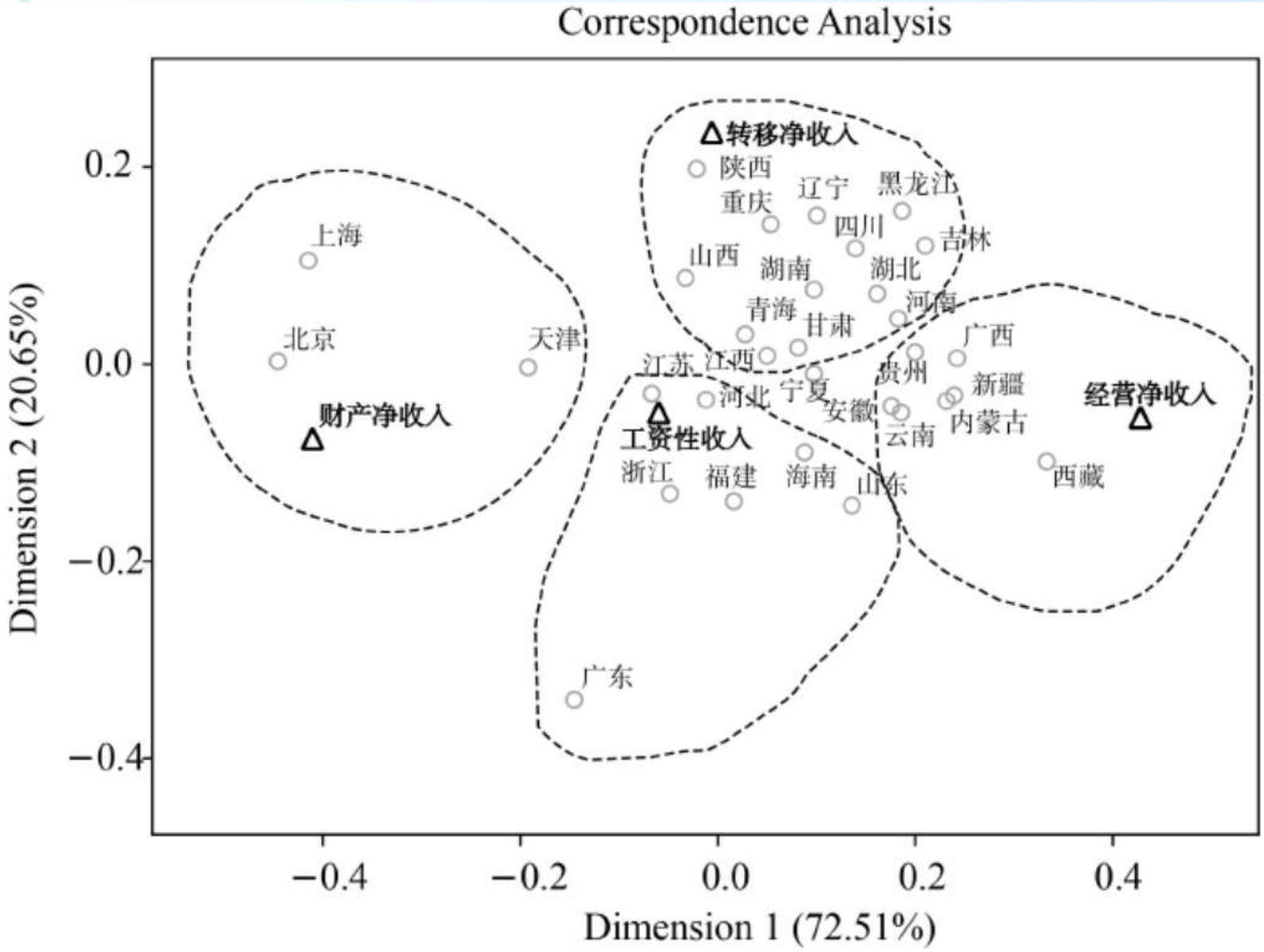


图 7-5　临近区域关联性

案例3

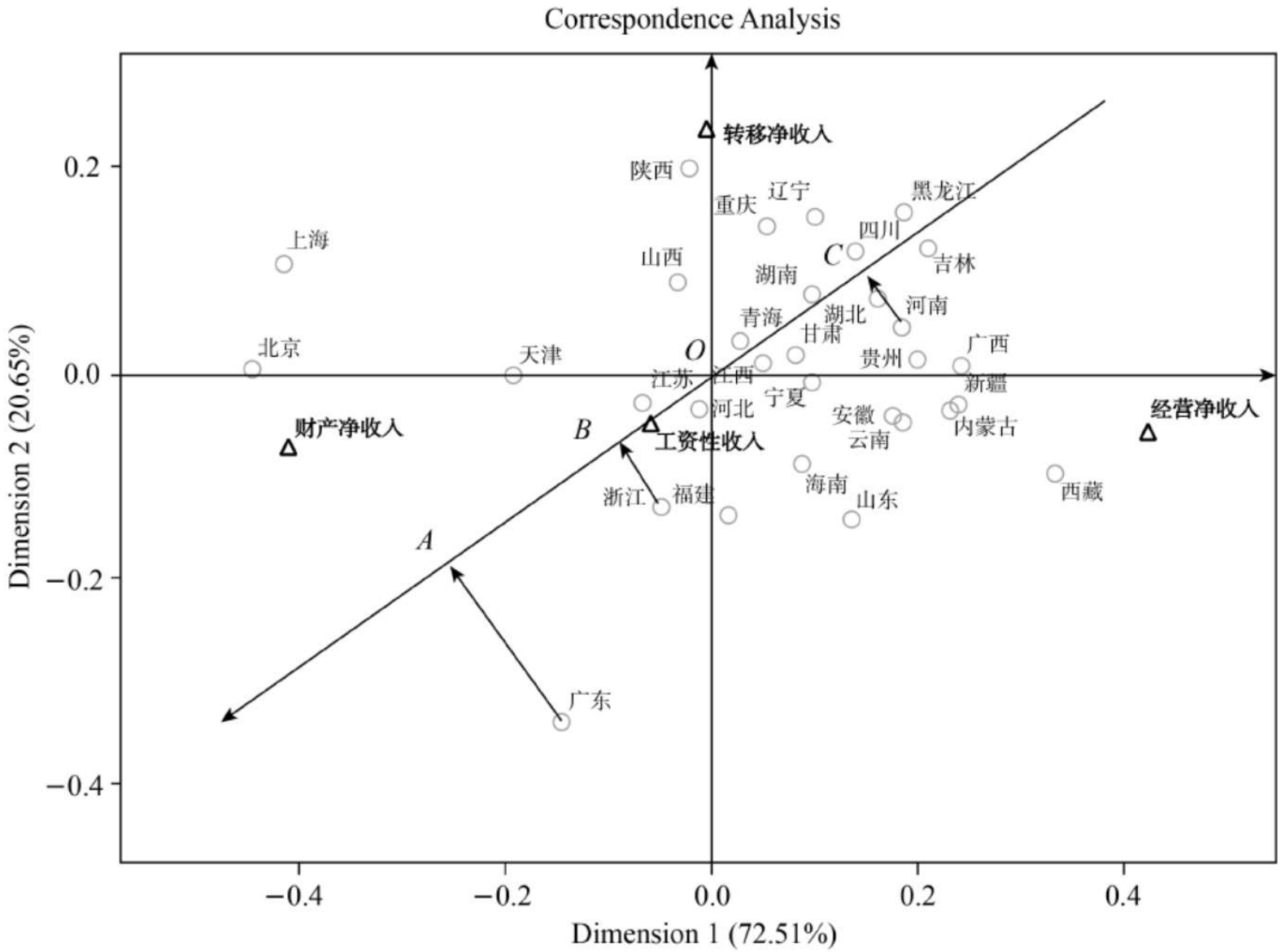


图 7-6　向量分析图

案例3

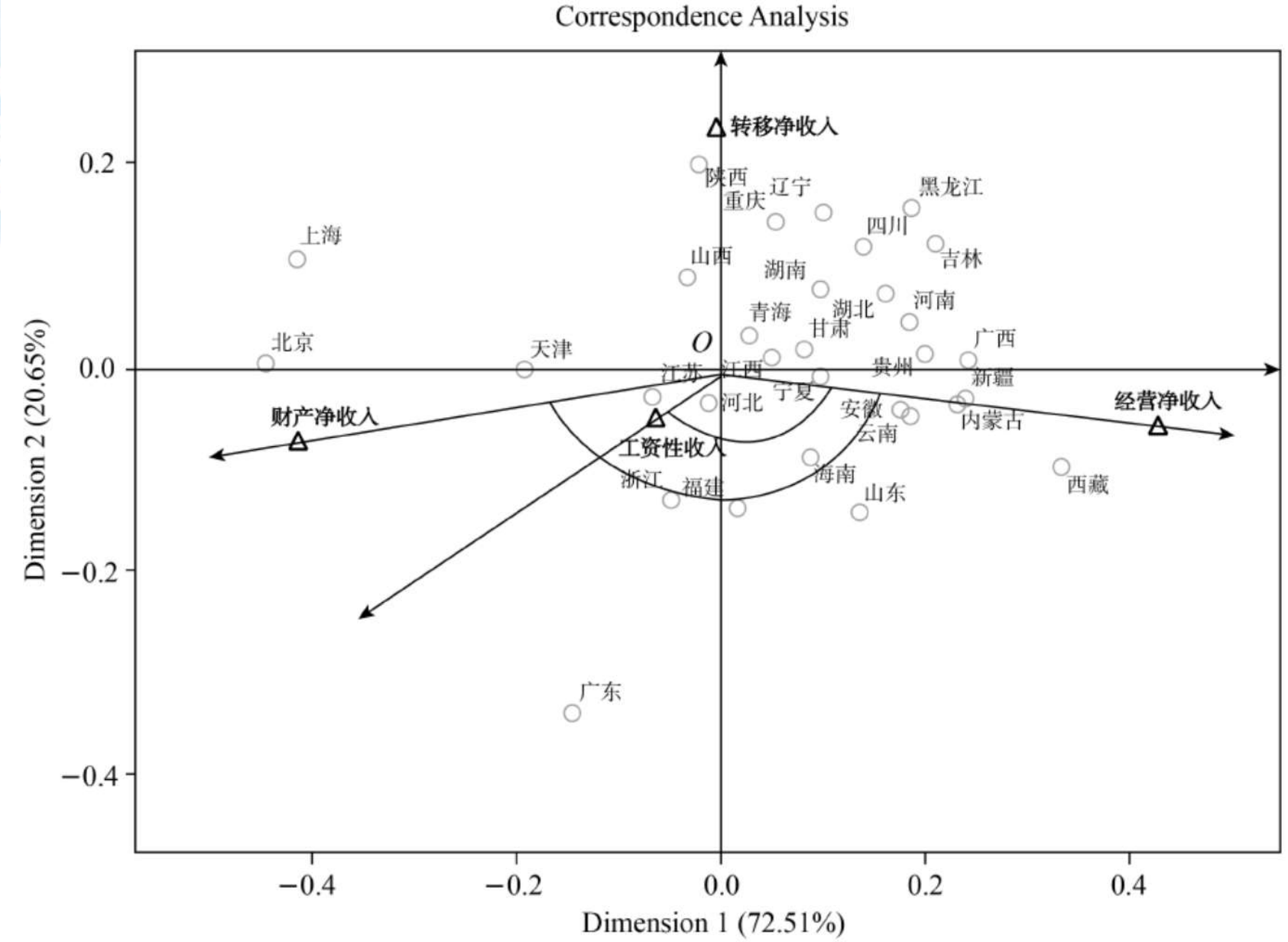


图 7-7　不同收入来源夹角图

案例3

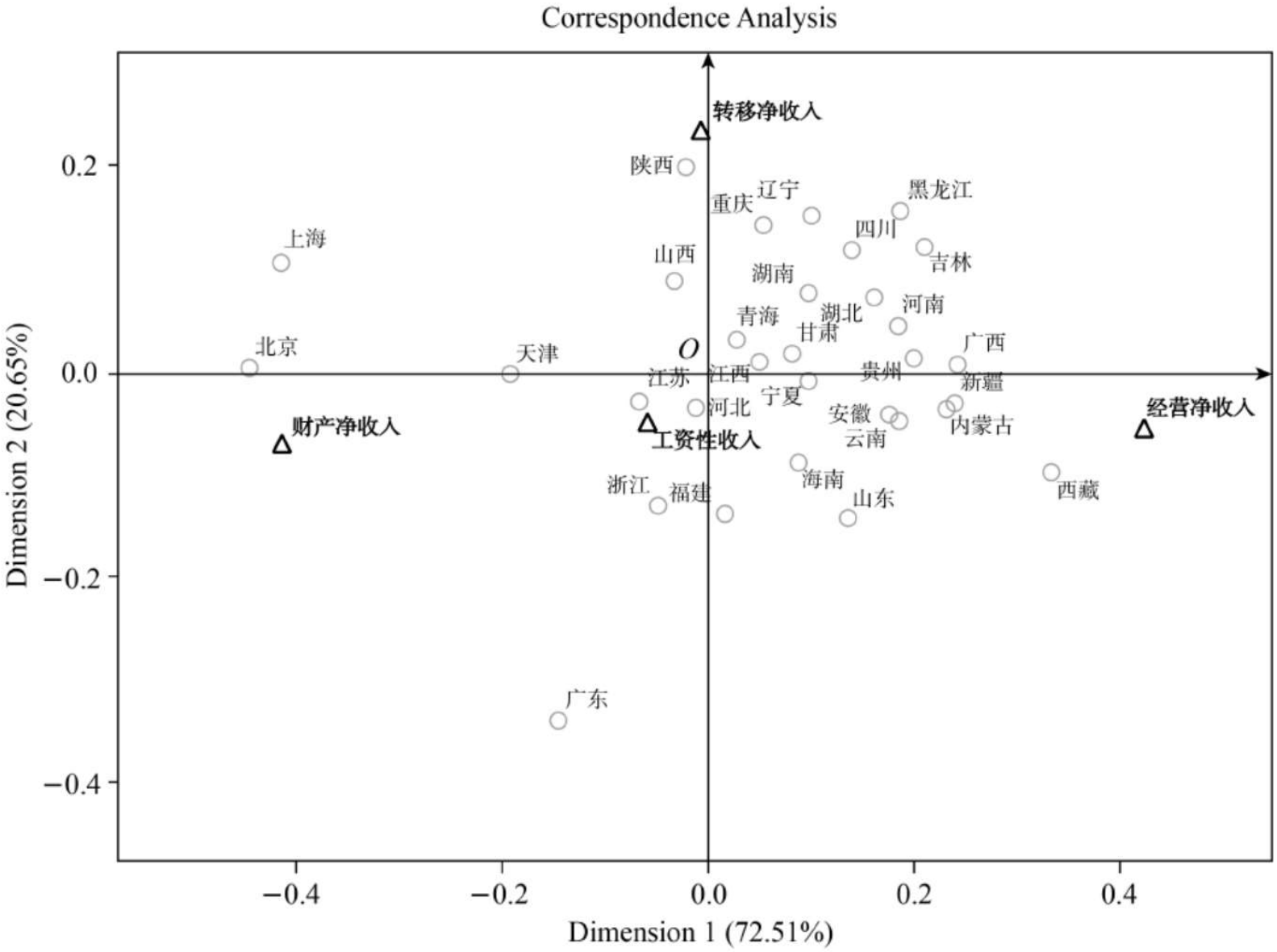


图 7-8　距原点距离图

案例4

起名为“波澜”恰当吗

娃哈哈公司欲为其新推出的一种纯水产品起一个合适的名字，为此专门委托了当地的策划咨询公司，取了一个名字“波澜”。一个好的名字至少应该满足两个条件：

1）会使消费者联想到正确的产品“纯水”；

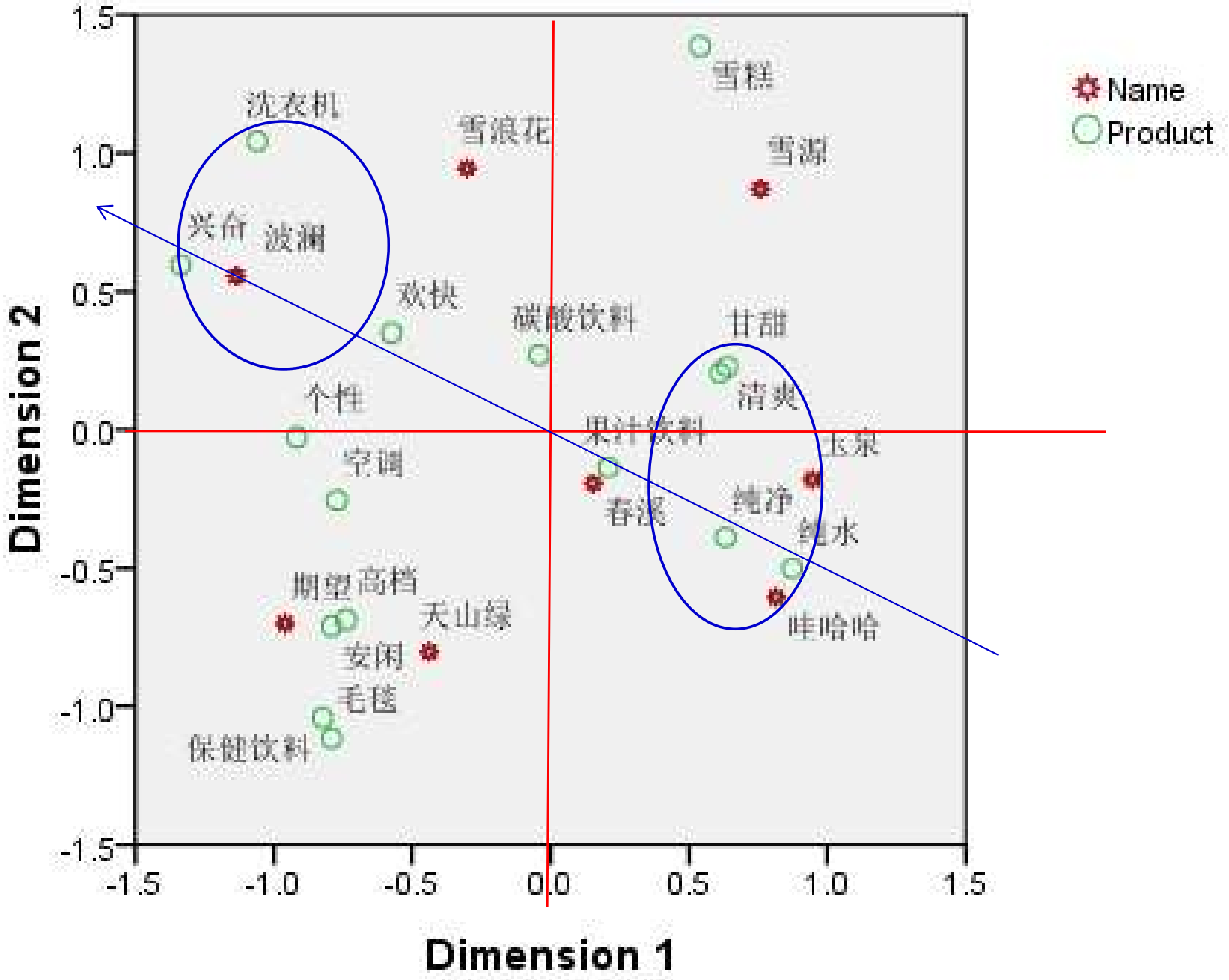
2）会使消费者产生与正确产品密切相关的联想，如“纯净”、“清爽”等。

娃哈哈集团决定对“波澜”这一名称方案进行品牌测试。采用调查问卷的形式进行消费者调查，以便最终确定品牌名称。

- 哇哈哈公司委托调查统计研究所，进行了一次全面的市场研究，在调查中还包括简单的名称测试。调查的代码和含义如下：

代码	含义	代码	含义	代码	含义
Name1	玉泉	Product1	雪糕	Feel1	清爽
Name2	雪源	Product2	纯水	Feel2	甘甜
Name3	春溪	Product3	碳酸饮料	Feel3	欢快
Name4	期望	Product4	果汁饮料	Feel4	纯净
Name5	波澜	Product5	保健食品	Feel5	安闲
Name6	天山绿	Product6	空调	Feel6	个性
Name7	哇哈哈	Product7	洗衣机	Feel7	兴奋
Name8	雪浪花	Product8	毛毯	Feel8	高档

	name1	name2	name3	name4	name5	name6	name7	name8
product1	50	442	27	21	14	50	30	258
product2	508	110	272	51	83	88	605	79
product3	55	68	93	36	71	47	37	77
product4	109	95	149	41	36	125	44	65
product5	34	29	45	302	37	135	42	18
product6	11	28	112	146	113	39	28	31
product7	30	12	54	64	365	42	8	316
product8	2	4	17	36	29	272	9	35
feel1	368	322	167	53	57	129	149	170
feel2	217	237	142	41	34	95	119	116
feel3	19	25	185	105	123	44	22	193
feel4	142	140	128	47	38	123	330	68
feel5	16	16	106	166	81	164	21	36
feel6	2	14	9	72	94	41	37	42
feel7	4	11	10	78	248	35	17	81
feel8	3	5	19	107	63	126	63	49


Name
Product
洗衣机
雪糕
雪浪花
雪源
兴奋
波澜
欢快
碳酸饮料
甘甜
个性
清爽
果汁饮料
玉泉
空调
纯净
春溪
纯水
期望
高档
天山绿
哇哈哈
安闲
毛毯
保健饮料
Dimension 1
Dimension 2
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5

由直观图可以看出，“波澜”与“洗衣机”产品相联系，引起的感觉是“兴奋”，因此“波澜”不是合适的纯净水品牌名称。

哇哈哈公司的产品是“纯水”如果想要使该名称给人们一种“纯净”的感觉，那么“哇哈哈” 将是最好的商品名称。如果想要使该名称给人们一种“清爽” 的感觉，那么“玉泉”将是最好的商品名称。。

结论：

不如叫“玉泉”或“娃哈哈”吧！

python例题

本次使用的数据是收入数据，原数据有8993个观测值和14个变量，这里只取没有缺失值的6876个观测值以及2个变量,并且合并了少数 变量的水平.我们取的变量为:
income (收入，水平，取值为: [0,10)， [10,20), [20,30)，[30 , 50), [50,75), 75+,单位:千美元);

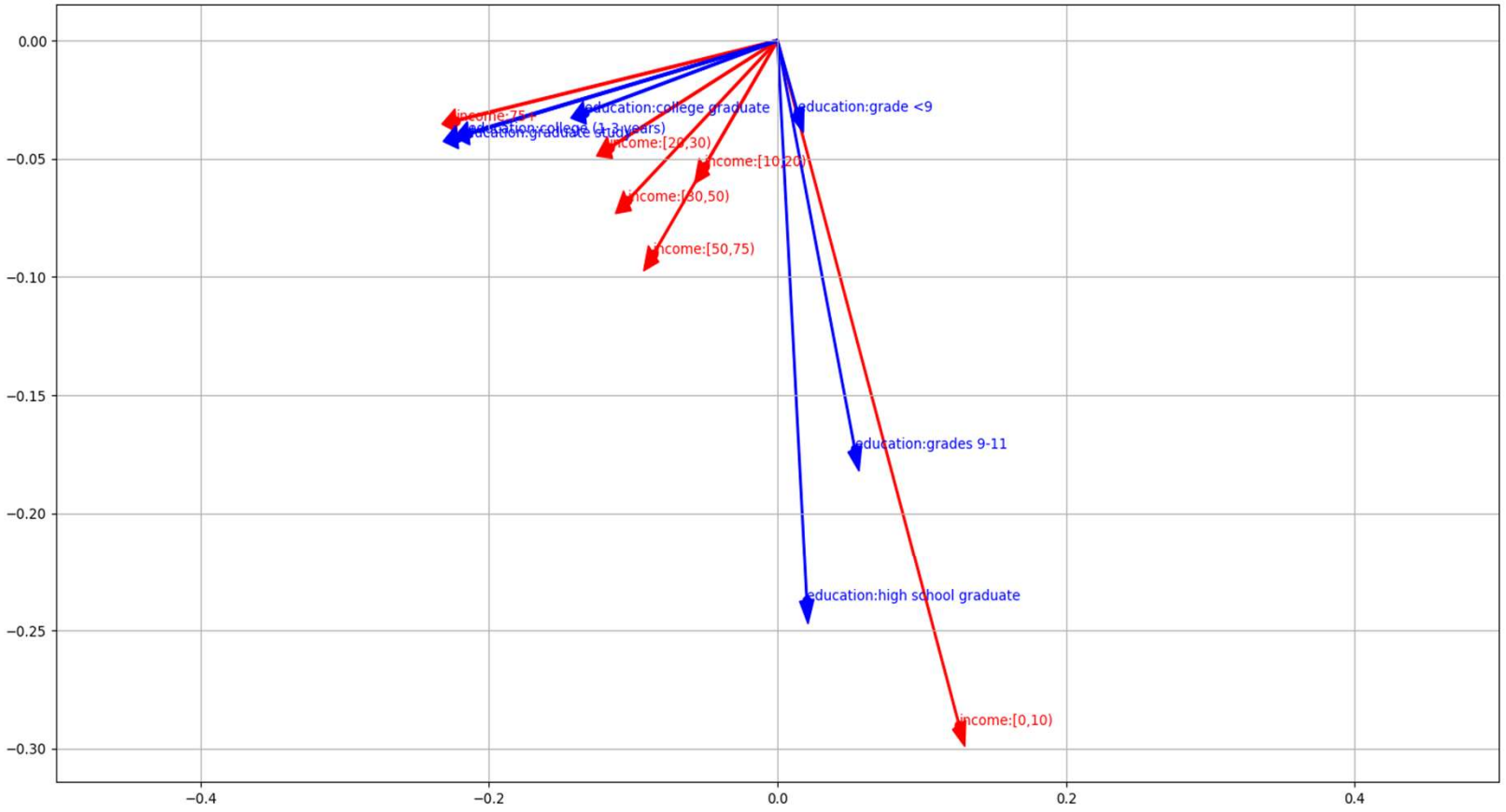
education (教育,水平)取值为:
college graduate(大学毕业)
grades 9-11(高中)
college (1-3 years)(大一到大三)，
grade<=11 (最多11 年级)
graduate study(研究生),
high school graduate (中学毕业);

代码：

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
#输入模块和数据整理，把分类变量变成哑元变量
v=pd.read_csv('IncomeCA.csv',index_col=False)
id =v.columns[np.array([0,1])]
w=v[id]
Z=w
for i in range(2):
    Z=pd.concat([Z,pd.get_dummies(w[id[i]])],axis=1)
#按照前面公式做二元对应分析：
z=np.array(Z[Z.columns[2:]])
P=z[:,:6].T.dot(z[:,6:])/z.shape[0]
r=np.sum(P,axis=1)
c=np.sum(P,axis=0)
Dr=np.diag(np.sqrt(r))
Dc=np.diag(np.sqrt(c))
r.shape=len(r),1
c.shape=len(c),1
S=np.dot(Dr.dot(P-r.dot(c.T)),Dc)
U,s,V=np.linalg.svd(S)
X=Dr.dot(U)
Y=Dc.dot(V.T)
```

```
#画出收入数据的二元对应分析图
T1=id[0]+":"+Z.columns[2:8]
T2=id[1]+":"+Z.columns[8:14]
plt.xlim(-0.5,0.5)
plt.scatter(X[:,0],X[:,1],color='r')
for i in range(6):
    plt.text(X[i,0],X[i,1],T1[i],color='r')
    plt.arrow(0,0,X[i,0],X[i,1],color='r',head_width=0.01,head_length=0.01)
plt.scatter(Y[:,0],Y[:,1],color='b')
for i in range(6):
    plt.text(Y[i, 0], Y[i, 1], T2[i], color='b')
    plt.arrow(0,0,Y[i, 0], Y[i,1],  color='b', head_width=0.01,
head_length=0.01)
plt.grid()
plt.show()
```

结果如图：


0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.30
−0.4
−0.2
0.0
0.2
0.4
education:college graduate
education:grade <9
income:[20,30)
income:[10,20)
income:[30,50)
income:[50,75)
education:grades 9-11
education:high school graduate
income:[0,10)

内容

(1) 基本思想

(2) 对应分析原理

(3) 对应分析实例

(4) 对应分析方法的优缺点

对应分析方法的优缺点

（1）定性变量划分的类别越多，这种方法的优越性越明显

（2）揭示行变量类间与列变量类间的联系

（3）直观地表达变量所含类别间的关系

（4）不能用于相关关系的假设检验

（5）受极端值的影响

（6）维数自定