信陽師範学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

第12章 MATLAB多元统计分析

讲授人：牛言涛

日期：2020年5月5日

目录

CONTENTS

A 主成分分析

B 因子分析

C 判别分析

D 聚类分析

E 典型相关分析

F 对应分析

1. 对应分析简介

- Q型分析：样本之间的关系（聚类算法等）；R型分析：变量之间的关系（主成分分析、因子分析等）。有时候我们不仅要弄清样本之间和变量之间的关系，还要弄清样本与变量之间的关系，而对应分析就是这样一种分析方法。（变量就是指特征）。
- Q型和R型因子分析分别反映了数据的不同方面，它们之间必然有内在的联系，对应分析通过巧妙的数学转换，将Q型和R型因子分析有机地结合起来。
- 对应分析(Correspondence Analysis)也称关联分析、R-Q型因子分析，是R型因子分析和Q型因子分析的结合，是近年新发展起来的一种多元统计分析技术，通过分析由定性变量构成的交互汇总表来揭示变量间的联系。可以揭示同一变量的各个类别之间的差异，以及不同变量各个类别之间的对应关系。
- 对应分析主要应用于产品定位、品牌研究、市场细分、竞争分析、广告研究等领域，因为它是一种图形化的数据分析方法，它能够将几组看似没有联系的数据，通过视觉上可以接受的定位图展现出来。

2. 对应分析原理分析

- 定义：研究样本和变量之间的关系。
- 作用：对应分析是分析两组或多组因素之间关系的有效方法，在离散情况下，建立因素间的列联表来对数据进行分析。
- 应用条件：在对数据作对应分析之前，需要先了解因素间是否独立。如果因素之间相互独立，则没有必要进行对应分析。

二维列联表分析		因素B				
		B_1	B_2	$\cdots$	B_c	
因素A	A_1	k_{11}	k_{12}	$\cdots$	k_{1c}	$k_{1\cdot}$
	A_2	k_{21}	k_{22}	$\cdots$	k_{2c}	$k_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	k_{r1}	k_{r2}	$\cdots$	k_{rc}	$k_{r\cdot}$
		$k_{\cdot 1}$	$k_{\cdot 2}$	$\cdots$	$k_{\cdot c}$	$K = k_{\cdot\cdot}$

2. 对应分析原理分析

对应分析的步骤：

将级联表中的数据以矩阵形式表示出来（数据矩阵X），将数据矩阵转化为频率矩阵P，将频率矩阵再进一步转化为过渡矩阵Z。

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \Rightarrow P=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1p} \\ p_{21} & p_{22} & \cdots & p_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{np} \end{bmatrix} \Rightarrow Z=\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

其中 $p_{ij}=\dfrac{x_{ij}}{sum(X)}$，$z_{ij}=\dfrac{p_{ij}-p_{i.}p_{.j}}{\sqrt{p_{i.}p_{.j}}}=\dfrac{x_{ij}-x_{i.}x_{.j}/x_{..}}{\sqrt{x_{i.}x_{.j}}}$

变量的协方差矩阵（一般每行表示一个样本，每列代表一种变量或特征）：$A=Z'Z$；样本的协方差矩阵$B=ZZ'$。

也可从过渡矩阵Z出发，按照奇异值分解进行求解。

2. 对应分析原理分析

A和B有如下定理：设A的非零特征值是λ_i，特征向量是u_i则：

（1）A和B的所有特征值相等。

（2）B的非零特征值λ_i对应的特征向量是$v_i = zu_i$。

计算出A和B的特征值和特征向量之和就要求A和B的因子载荷（对应于R型分析和Q型分析）

$$F = \begin{bmatrix} \sqrt{\lambda_1}u_{11} & \sqrt{\lambda_2}u_{12} & \cdots & \sqrt{\lambda_m}u_{1m} \\ \sqrt{\lambda_1}u_{21} & \sqrt{\lambda_2}u_{22} & \cdots & \sqrt{\lambda_m}u_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\lambda_1}u_{p1} & \sqrt{\lambda_2}u_{p2} & \cdots & \sqrt{\lambda_m}u_{pm} \end{bmatrix}, \quad G = \begin{bmatrix} \sqrt{\lambda_1}v_{11} & \sqrt{\lambda_2}v_{12} & \cdots & \sqrt{\lambda_m}v_{1m} \\ \sqrt{\lambda_1}v_{21} & \sqrt{\lambda_2}v_{22} & \cdots & \sqrt{\lambda_m}v_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\lambda_1}v_{n1} & \sqrt{\lambda_2}v_{n2} & \cdots & \sqrt{\lambda_m}v_{nm} \end{bmatrix}$$

其中F是A的因子载荷，G是B的因子载荷。m为非零特征值数量，p为原始数据的列数，n是行数。在选择A和B的因子载荷之前先要进行矩阵稀疏，思想和主成分分析一样，将特征值从大到小排列，看前面k个特征值的重要性（k个特征值之和占总的特征值之和的比重）。

3. 对应分析检验

- χ^2卡方检验：理论计算出来的频率与实际收集到的数据统计出来的频率之间总是存在一些偏差，把每一个指定值的偏差以平方的形式加起来，如果这个值比较小，则说明分布拟合得较好，如果这个值很大，则说明实际收集到的数据与目标分布并不相同，需要去寻找其它恰当的分布。计算公式如下：

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(f_i - F_i\right)^2}{F_i}$$，其中f_i是实际观察到的量，F_i是运用目标分布计算出的量。

- 对数据进行对应分析之前需要先对其进行卡方检验，检验数据之间是否独立，如果数据直接独立的话就没有必要进行对应分析。其卡方检验公式如下：

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left[k_{ij} - \hat{E}\left(k_{ij}\right)\right]^2}{\hat{E}\left(k_{ij}\right)} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left[k_{ij} - k_{i.}k_{.j}/k\right]^2}{k_{i.}k_{.j}/k} \sim \chi^2\left((n-1)(p-1)\right)$$

其中k_{ij}是矩阵中的每个元素，k为所有元素总和，$k_{i.}, k_{.j}$分别表示第i行元素之和及第j列元素之和。

4. 对应分析MATLAB实现一

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

CorrespA_RQ.m

```
function carq = CorrespA_RQ(data,label)
% CA_demo函数用来进行对应分析
% data是样本数据，label对应变量名称和行名称
% carq是输出结构体变量，其中包括对应分析过程的结果和结论

    %% 变量的处理，计算频率矩阵P和过渡矩阵Z
    T = sum(sum(data)); %计算总和
    P = data/T;    %计算对应矩阵P
    r = sum(P,2); %计算边缘分布，行和
    c = sum(P);  %计算边缘分布，列和
    Z = (P - r*c)./sqrt((r*c)); %计算标准化数据B，即过渡矩阵

    %% 求过渡矩阵Z的奇异值分解
    [u,D,v] = svd(Z,'econ');  %对标准化后的数据阵Z作奇异值分解
    G = u*sqrt(D);
    F = v*sqrt(D);

    %% 求解Z'*Z主惯量和累积贡献率
    lamda = diag(D).^2;    %计算Z'*Z的特征值，即计算主惯量
    con_rate = lamda/sum(lamda);  %计算贡献率

    %% 卡方检验
    ksi2square = T*(lamda);  %计算卡方统计量的分解
    T_ksi2square = sum(ksi2square); %计算总卡方统计量
    def = (size(data,1) - 1)*(size(data,2) - 1); %自由度
    pChisq = 1-chi2cdf(T_ksi2square,def); %求卡方检验概率值

    %% 对应分析可视化，只绘制前两个维度，并标记每一维贡献率
    num = size(G,1);  %样本点的个数b
    rang = minmax(G(:,[1,2])');  %坐标的取值范围
    delta = (rang(:,2)-rang(:,1))/(5*num); %画图的标注位置调整量
    ch = cellstr(label(1,2:size(label,2))); %对应列变量名称
    yb = cellstr(label(2:size(label,1),1))'; %对应行样本名称
    h1 = plot(G(:,1),G(:,2),'b*','LineWidth',1.3);  %画行点散布图
    hold on
    text(G(:,1)-delta(1),G(:,2)-3*delta(2),yb) %对行点进行标注
    h2 = plot(F(:,1),F(:,2),'rH','LineWidth',1.3); %画列点散布图
    text(F(:,1)+delta(1),F(:,2),ch) %对列点进行标注
    h = refline(0,0); h.Color = 'k'; h.LineStyle = ':'; %添加水平辅助线
    h1m = minmax(h1.YData);h2m = minmax(h2.YData);
    mind = min(h1m(1),h2m(1));
    maxd = max(h1m(2),h2m(2));
    % 添加垂直辅助线
    plot(zeros(1,10),linspace(mind-5*delta(2),maxd+5*delta(1),10),'k:');
    xtext = strcat('dimension1(',num2str(con_rate(1)*100),'%)');
    ytext = strcat('dimension2(',num2str(con_rate(2)*100),'%)');
    xlabel(xtext), ylabel(ytext)
    title('Correspondence analysis chart');

    %% 输出变量的组合
    carq.lamda = lamda;
    carq.con_rate = con_rate;
    carq.F = F;
    carq.G = G;
    carq.Total_ksi2square = T_ksi2square;
    carq.pChisq = pChisq;
end
```

4. 对应分析MATLAB实现二

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

CorrespA_RQ.m

```
function carq = CorrespA_RQ(data,label)
% CA_demo函数用来进行对应分析
% data是样本数据，label对应变量名称和行名称
% carq是输出结构体变量，其中包括对应分析过程的结果和结论

    %% 变量的处理，计算频率矩阵P和过渡矩阵Z
    T = sum(sum(data)); %计算总和
    P = data/T;   %计算对应矩阵P
    r = sum(P,2); %计算边缘分布，行和
    c = sum(P);  %计算边缘分布，列和
    Z = (P - r*c)./sqrt((r*c)); %计算标准化数据B，即过渡矩阵

    %% 求过渡矩阵Z的奇异值分解，计算因子载荷F、G
    [u,D,v] = svd(Z,'econ');  %对标准化后的数据阵Z作奇异值分解
    beta = diag(r.^(-1/2))*u;  %求加权特征向量
    G = beta*D;   %求行轮廓坐标G
    alpha = diag(c.^(-1/2))*v;   %求加权特征向量
    F = alpha*D;   %求列轮廓坐标F

    %% 求解Z'*Z主惯量和累积贡献率
    lamda = diag(D).^2;   %计算Z'*Z的特征值，即计算主惯量
    con_rate = lamda/sum(lamda);  %计算贡献率

    %% 卡方检验
    ksi2square = T*(lamda);  %计算卡方统计量的分解
    T_ksi2square = sum(ksi2square); %计算总卡方统计量
    def = (size(data,1) - 1)*(size(data,2) - 1); %自由度
    pChisq = 1-chi2cdf(T_ksi2square,def); %求卡方检验概率值

    %% 对应分析可视化，只绘制前两个维度，并标记每一维贡献率
    num = size(G,1);  %样本点的个数b
    rang = minmax(G(:,[1,2])');  %坐标的取值范围
    delta = (rang(:,2)-rang(:,1))/(5*num); %画图的标注位置调整量
    ch = cellstr(label(1,2:size(label,2))); %对应列变量名称
    yb = cellstr(label(2:size(label,1),1))'; %对应行样本名称
    h1 = plot(G(:,1),G(:,2),'b*','LineWidth',1.3);  %画行点散布图
    hold on
    text(G(:,1)-delta(1),G(:,2)-3*delta(2),yb) %对行点进行标注
    h2 = plot(F(:,1),F(:,2),'rH','LineWidth',1.3); %画列点散布图
    text(F(:,1)+delta(1),F(:,2),ch) %对列点进行标注
    h = refline(0,0); h.Color = 'k'; h.LineStyle = ':'; %添加水平辅助线
    h1m = minmax(h1.YData);h2m = minmax(h2.YData);
    mind = min(h1m(1),h2m(1));
    maxd = max(h1m(2),h2m(2));
    % 添加垂直辅助线
    plot(zeros(1,10),linspace(mind-5*delta(2),maxd+5*delta(1),10),'k:');
    xtext = strcat('dimension1(',num2str(con_rate(1)*100),'%)');
    ytext = strcat('dimension2(',num2str(con_rate(2)*100),'%)');
    xlabel(xtext), ylabel(ytext); title('Correspondence analysis chart');
    %根据行坐标第一维进行分类
    rowclass = yb(G(:,1)>0);  %提出第一类样本点
    %根据列坐标第一维进行分类
    colclass = ch(F(:,1)>0);  %提出第一类变量
    %% 输出变量的组合
    carq.lamda = lamda;
    carq.con_rate = con_rate;
    carq.F = F; carq.G = G;
    carq.Total_ksi2square = T_ksi2square;
    carq.pChisq = pChisq;
    carq.FirstSample = rowclass;
    carq.FirstVar = colclass;
end
```

5. 案例分析——产品命名

例1：起名为“波澜”恰当吗?

中美纯水有限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。

后来中美纯水有限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。其部分数据如下表：

	玉泉	雪源	春溪	期望	波澜	天山绿	中美纯	雪浪花
雪糕	50	442	27	21	14	50	30	258
纯水	508	110	272	51	83	88	605	79
碳酸饮料	55	68	93	36	71	47	37	77
果汁饮料	109	95	149	41	36	125	44	65
…	…	…	…	…	…	…	…	…

5. 案例分析——产品命名

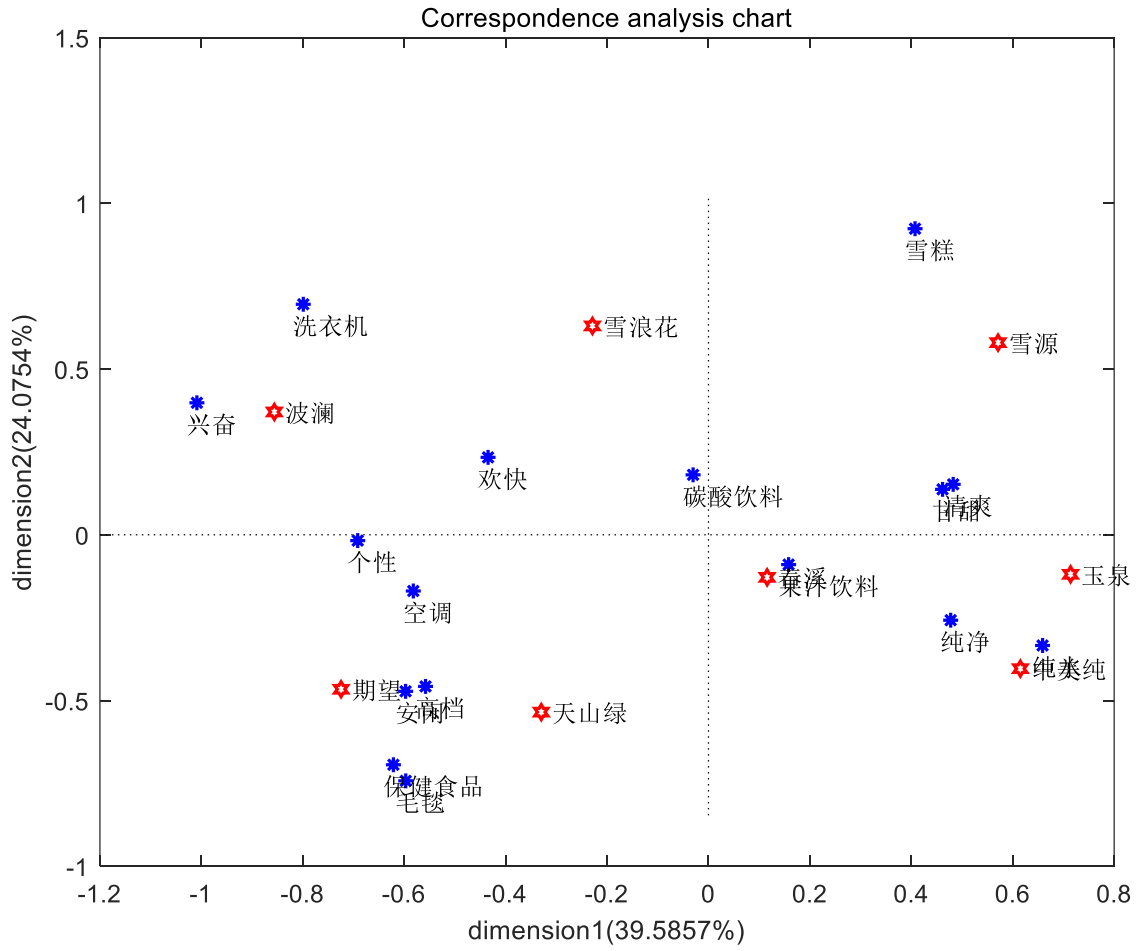
信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

	玉泉	雪源	春溪	期望	波澜	天山绿	中美纯	雪浪花
雪糕	50	442	27	21	14	50	30	258
纯水	508	110	272	51	83	88	605	79
碳酸饮料	55	68	93	36	71	47	37	77
果汁饮料	109	95	149	41	36	125	44	65
保健食品	34	29	45	302	37	135	42	18
空调	11	28	112	146	113	39	28	31
洗衣机	30	12	54	64	365	42	8	316
毛毯	2	4	17	36	29	272	9	35
清爽	368	322	167	53	57	129	149	170
甘甜	217	237	142	41	34	95	119	116
欢快	19	25	185	105	123	44	22	193
纯净	142	140	128	47	38	123	330	68
安闲	16	16	106	166	81	164	21	36
个性	2	14	9	72	94	41	37	42
兴奋	4	11	10	78	248	35	17	81
高档	3	5	19	107	63	126	63	49

案例求解与分析

```
>> [data,label] = xlsread('watername.xlsx');
>> carq = CorrespA_RQ(data,label)
```

- 通过检验，说明变量之间是相关的。
- 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。
- 中美纯水公司的产品是“纯水”，如果想要使该名称给人们一种“纯净”的感觉，那么“中美纯”将是最好的商品名称。
- 如果想要使该名称给人们一种“清爽、甘甜”的感觉，那么“玉泉”将是最好的商品名称。



5. 案例分析——地区与收入关系

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

例2：按现行统计报表制度，农村居民可支配收入主要由四部分构成，即工资性收入、经营净收入、财产净收入、转移净收入。选取全国31个省、直辖市、自治区农村居民人均可支配收入的数据。试进行对应分析，揭示全国农村居民人均可支配收入的特征以及各省、直辖市、自治区与各收入类型间的关系。

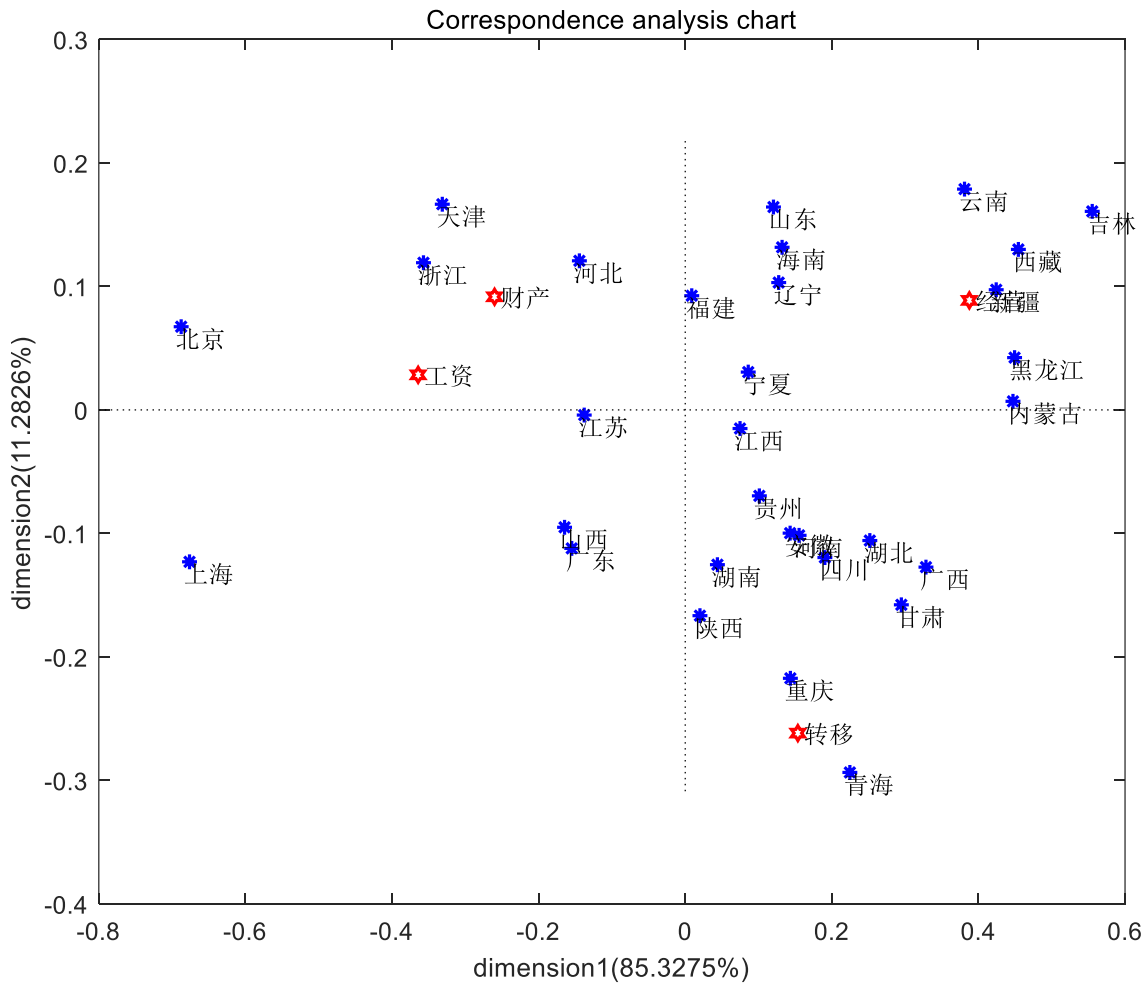
province	工资	经营	财产	转移	province	工资	经营	财产	转移
北京	16637.5	2061.9	1350.1	2260	湖北	4023	5534	158.6	3009.3
天津	12048.1	5309.4	893.7	1824.4	湖南	4946.2	4138.6	143.1	2702.5
河北	6263.2	3970	257.5	1428.6	广东	7255.3	3883.6	365.8	3007.5
山西	5204.4	2729.9	149	1999.1	广西	2848.1	4759.2	149.2	2603
内蒙古	2448.9	6215.7	452.6	2491.7	海南	4764.9	5315.7	139.1	1623.1
辽宁	5071.2	5635.5	257.6	1916.4	重庆	3965.6	4150.1	295.8	3137.3
吉林	2363.1	7558.9	231.8	1969.1	四川	3737.6	4525.2	268.5	2671.8
黑龙江	2430.5	6425.9	572.7	2402.6	贵州	3211	3115.8	67.1	1696.3
上海	18947.9	1387.9	859.6	4325	云南	2553.9	5043.7	152.2	1270.1
江苏	8731.7	5283.1	606	2984.8	西藏	2204.9	5237.9	148.7	1502.3
浙江	14204.3	5621.9	661.8	2378.1	陕西	3916	3057.9	159	2263.6
安徽	4291.4	4596.1	186.7	2646.2	甘肃	2125	3261.4	128.4	1942
福建	6785.2	5821.5	255.7	2136.9	青海	2464.3	3197	325.2	2677.8
江西	4954.7	4692.3	204.4	2286.4	宁夏	3906.1	3937.5	291.8	1716.3
山东	5569.1	6266.6	358.7	1759.7	新疆	2527.1	5642	222.8	1791.3
河南	4228	4643.2	168	2657.6					

案例求解与分析

```
>> [data,label] = xlsread('dirs_shouru.xlsx');
>> carq = CorrespA_RQ(data,label)
```

```
carq =
  包含以下字段的 struct:

              lamda: [4×1 double]
           con_rate: [4×1 double]
     cumsum_con_rate: [4×1 double]
                  F: [4×4 double]
                  G: [31×4 double]
    Total_ksi2square: 5.5148e+04
             pChisq: 0
         FirstSample: {1×23 cell}
            FirstVar: {'经营'  '转移'}
```



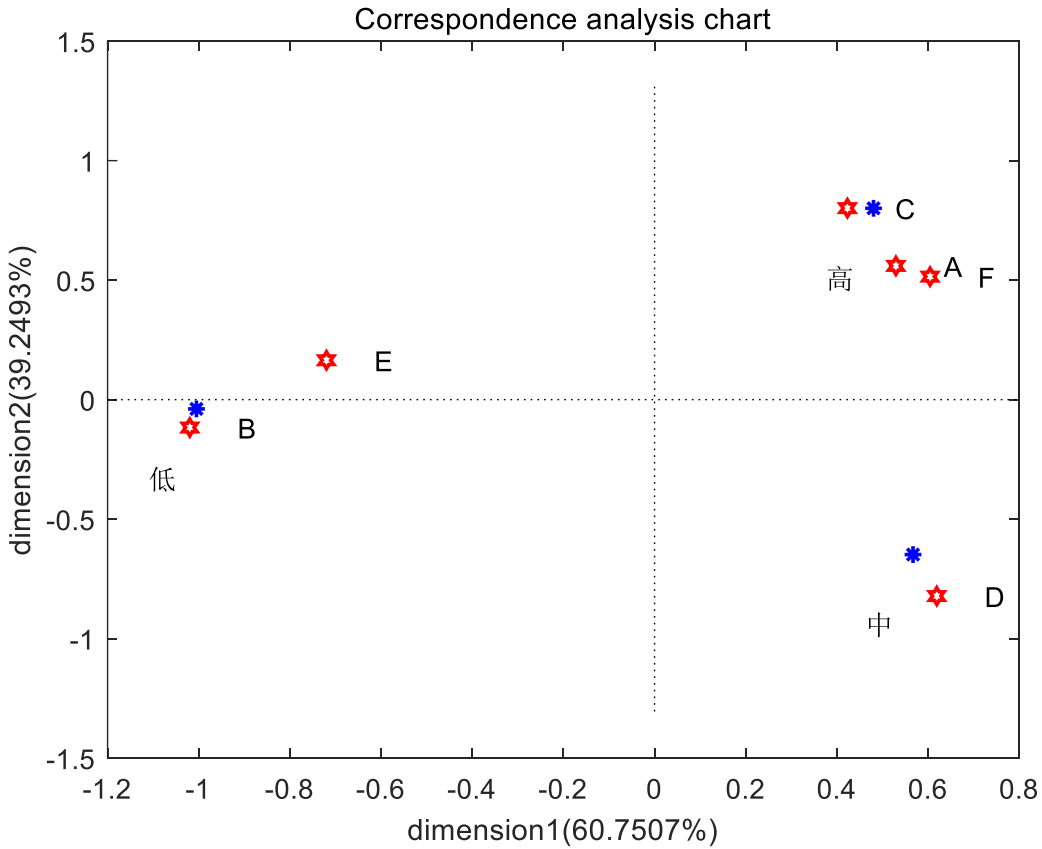
从图中看出，财产和工资收入主要有天津、河北和浙江，北京、江苏等地。经营收入主要是新疆、西藏、云南、黑龙江、吉林等地。转移收入主要是重庆、青海、陕西、甘肃等地。

5. 案例分析——收入与品牌选择

例3：某企业对218名受访人员进行了收入水平和品牌选择关系的调查研究，收入水平分高、中、低3个等级，品牌有A~F 6个。得到表所示的调查数据，对其进行对应分析。

收入水平	品牌					
	A	B	C	D	E	F
低	2	49	4	4	15	1
中	7	7	5	49	2	7
高	16	3	23	5	5	14

```
>> [data,label] = xlsread('pinpaixuanze.xlsx');
>> carq = CorrespA_RQ(data,label)
% 通过卡方检验。
% 从图中可以看出，高水平收入主要青睐于品牌A、C、F；中收入人群主要青睐于D品牌；低收入群体主要选择B和E品牌。
```



5. 案例分析——车辆购买因素与车型关系

例4：汽车细分市场购买因素。得到表所示的调查数据，对其进行对应分析。

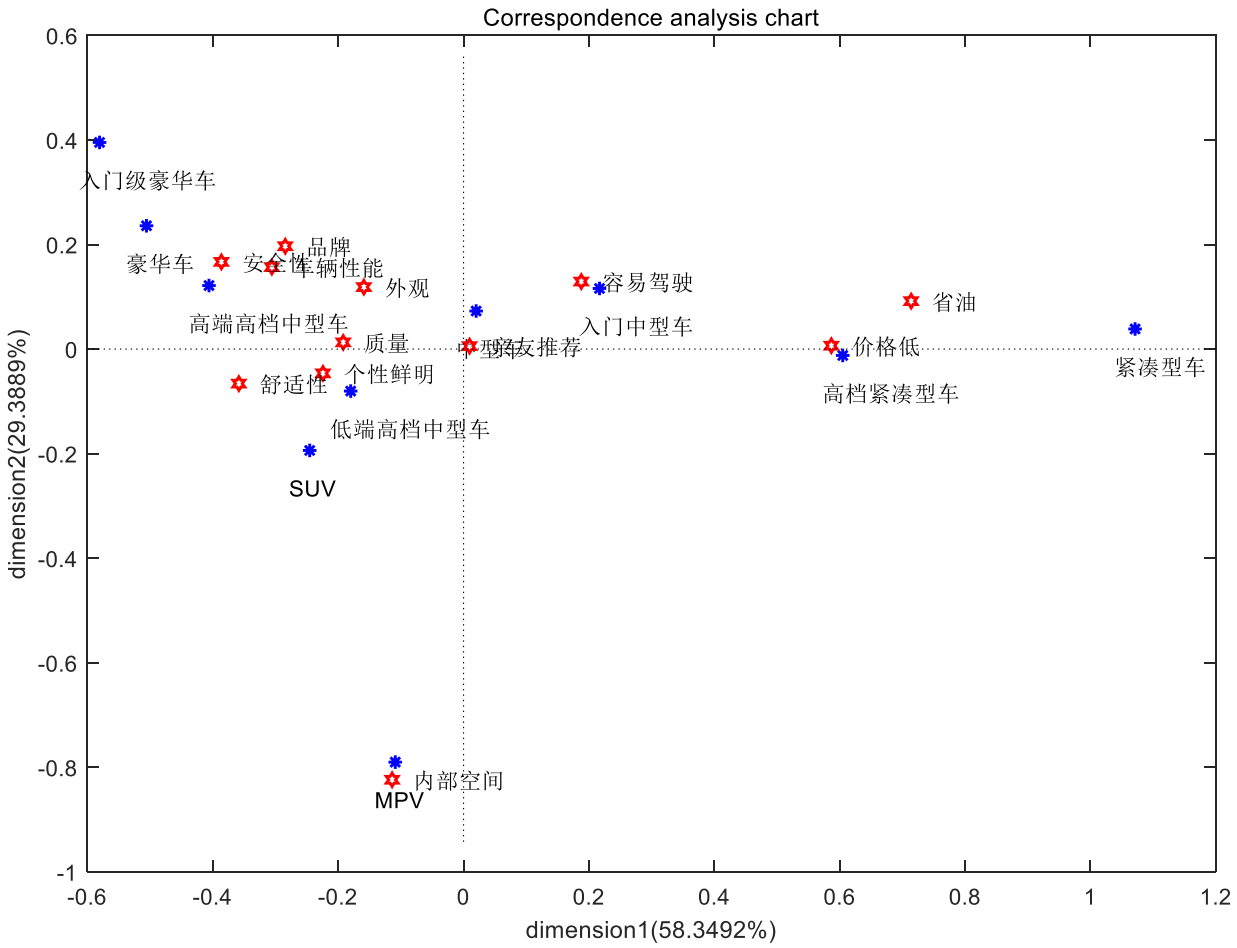
乘用车类型	品牌	亲友推荐	省油	价格低	质量	外观	个性鲜明	安全性	舒适性	容易驾驶	内部空间	车辆性能
紧凑型车	20	26	210	207	23	40	16	5	4	14	32	8
高档紧凑型车	21	25	83	102	21	45	21	12	2	7	28	12
入门中型车	221	136	318	393	227	297	99	118	42	49	107	142
中型车	290	214	317	374	238	512	153	220	61	45	196	202
低端高档中型车	39	45	14	75	53	83	32	58	19	5	56	45
高端高档中型车	209	91	65	82	196	372	123	227	71	20	119	198
入门级豪华车	50	7	1	4	24	26	13	25	6	4	1	16
豪华车	138	23	21	5	80	59	46	56	19	9	23	52
SUV	23	33	19	37	38	85	56	39	15	7	61	37
MPV	76	51	61	97	103	86	61	41	32	10	297	40

李晓娜，史占国．基于对应分析的汽车细分市场购买因素研究[J]．武汉理工大学学报（信息与管理工程版）.2011(33-1), 159-162.

案例求解与分析

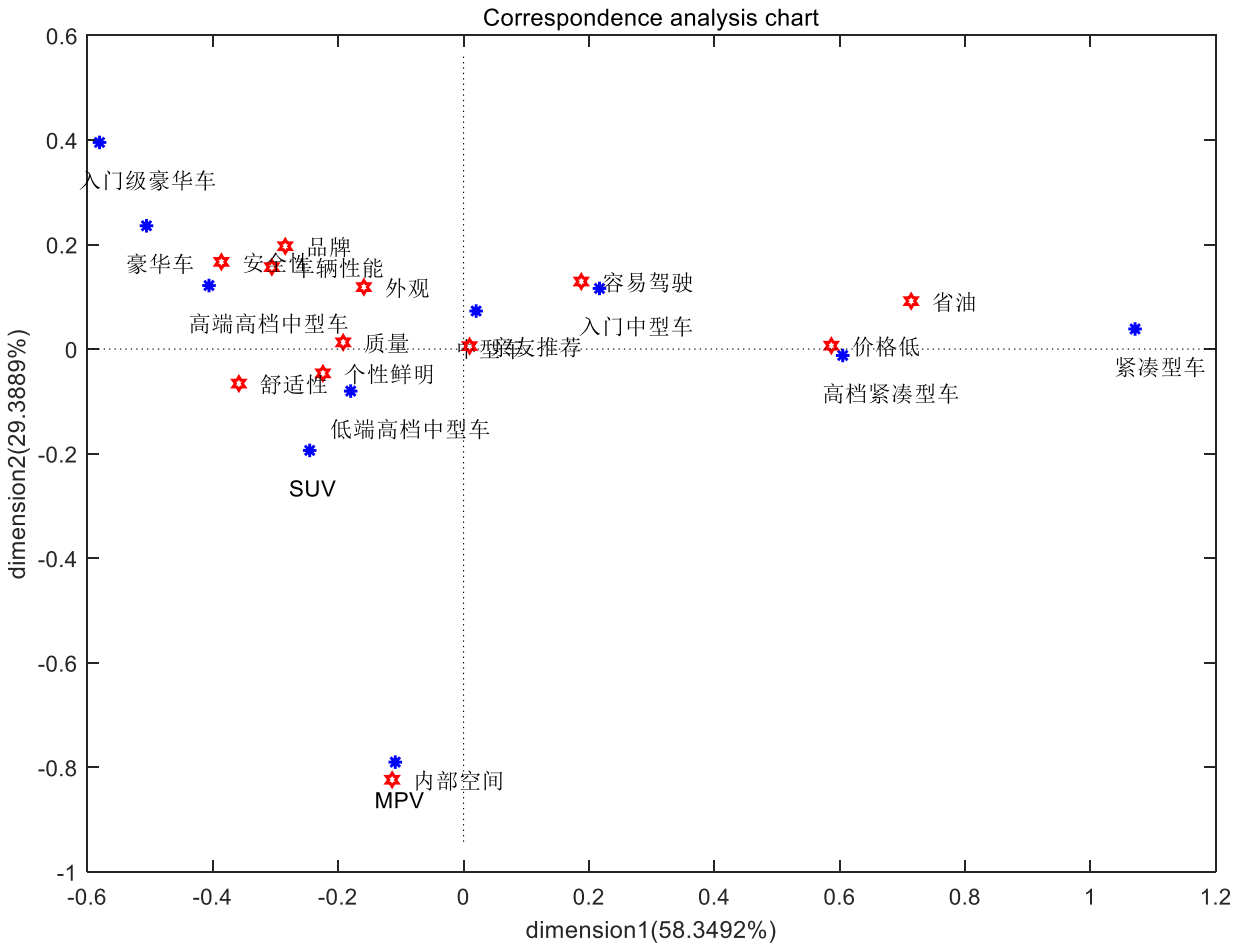
```
>> [data,label] = xlsread('car.xlsx');
>> carq = CorrespA_RQ(data,label)
```

- 从图中看出，车辆类型与购买因素的对应关系，两者之间的距离越近，则表明其关系越密切，车辆类型之间的距离越近，表明它们之间的竞争越激烈。
- 随着轿车类型由小型车向大型车的提升，消费者购买时考虑的因素从价格、省油和亲友推荐等车辆的外在属性，逐步过渡到考虑容易驾驶、外观、质量和舒适等车辆本身的属性，直至对车辆的性能、安全和品牌信誉这些更高属性的追求。
- 外形和质量处在入门级中型车、中型车、低端高档中型车、SUV、高端高档中型车、豪华车和入门级豪华车的中间地带，说明这两个因素被多个车辆类型的消费者看重。



案例求解与分析

- 紧凑型车和高档紧凑型车与价格和省油对应，说明这两种类型的消费者在购买时首先考虑的就是价格因素。
- 入门中型车与中型车位置接近，关系密切，具有共同的消费特征。购买时，对车辆本身要求容易驾驶、外形好看、质量可靠和亲友推荐，以及个性与消费者个人情趣相吻合。
- 低端高档中型车与SUV位置接近，具有共同的消费特征。要求车辆个性、质量和舒适性。
- 高端高档中型车和豪华车位置接近，说明两者由较强的竞争关系。两种车与安全性、性能和品牌紧密对应，另外，还与外观、质量和舒适性有较强的相关性。注意，入门级豪华车与豪华车存在较强的竞争关系。
- MPV与空间对应，还与个性鲜明、舒适性有一定的相关性。




01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
EDUCATION
BUSINESS
PHOTO BOOK

感谢聆听