第十章、典型相关分析

January 4, 2019

典型相关分析是研究两组多变量之间相关关系的一种统计方法，也是一种降维技术。

最早由Hotelling (1935, 1936) 提出, Cooley and Lohnes (1971), Kshirsagar (1972)和Mardia, Kent, and Bibby (1979) 推动了它的应用.

典型相关分析研究什么样的问题？

例如

- 工厂管理人员需要了解原料的主要质量指标$X_1, X_2, \ldots, X_p$与产品的主要质量指标$Y_1, Y_2, \ldots, Y_q$之间的相关性, 以便提高产品质量;
- 医生要根据一组化验指标确定与一些疾病之间的关系;
- 主教练排兵布阵要考虑自己的队员与对手之间的相生相克以便制定更好的对策;
- ……等等.

案例一：邮电业与国民经济的典型相关分析

问题背景：

- 随着电子信息技术的发展，以电话和快件等为代表的邮电业在近6、70年来得到蓬勃发展，让人们能深刻感受到科技的发展正广泛地影响着人们的生活和工作习惯。
- 邮电业的发展为人们提供了快捷、高效的生活和工作方式。越来越多的人选择网上或电话购物，以快件的形式发送或接收货物。这种消费模式节省了消费者的时间和精力，也极大地降低了商家的成本。降低了整个社会的运行成本，提高了社会的运行效率。

研究问题： 分析我国当前经济与邮电业之间发生相关，研究应该怎样发展邮电业？应该优先发展邮电业哪一行业？

指标选取： 选取邮电业和国民经济各四个主要指标，

- 衡量邮电业指标：X_1函件（亿件），X_2快递（万件），X_3移动电话年末用户(万户)，X_4固定电话年末用户(万户).
- 衡量我国经济指标（单位都是万亿）：Y_1第一产业，Y_2工业，Y_3建筑业，Y_4第三产业.

数据资料：

年份	x_1	x_2	x_3	x_4	y_1	y_2	y_3	y_4
1995	79.55	5582.7	382.9	4070.6	12135.8	24950.8	3728.8	19978.5
1996	78.68	7096.6	685.3	5494.7	14015.4	29447.6	4387.4	23326.2
1997	68.55	6878.9	1323.3	7031.0	14441.9	32921.4	4621.6	26988.1
1998	65.51	7331.8	2386.3	8742.1	14817.6	34018.4	4985.8	30580.5
1999	60.52	9091.3	4329.6	10871.6	14770.0	35861.5	5172.1	33873.4
2000	77.71	11031.4	8453.3	14482.9	14944.7	40033.6	5522.3	38714.0
2001	86.93	12652.7	14522.2	18036.8	15781.3	43580.6	5931.7	44361.6
2002	108.01	14036.2	20600.5	21422.2	16537.0	47431.3	6465.5	49898.9
2003	103.84	17237.8	26996.3	26274.7	17381.7	54945.5	7490.8	58004.7
2004	82.81	19771.9	33482.4	31175.6	21412.7	65210.0	8694.3	64561.3
2005	73.51	22880.3	39340.6	35044.5	22420.0	77230.8	10133.8	73432.9
2006	71.31	26988.0	46105.8	38778.6	24040.0	91310.9	11851.1	84721.4
2007	69.50	120189.6	54730.6	38563.7	28095.0	107387.2	14014.1	100053.5

研究方法： 通过典型相关分析来来找出邮电业和国民经济之间相互影响的内在规律，根据这个规律，给决策者提供一个当前如何发展邮电业的参考。

案例二：典型相关分析在土地利用结构研究中的应用

问题背景：

- 20 世纪90 年代以来, 土地利用与土地覆被变化成为全球环境变化研究的重点领域。
- 目前国际上有关土地利用与土地覆被变化的研究项目多侧重于土地覆被的分类、动态监测和环境影响评价, 对土地结构及其变化的动力机制的研究虽然也开展了一些工作, 但多难以深入。
- 科学家们均已认识到人类活动是影响土地利用变化的主要驱动力, 但在建立分析模型时受到各种限制, 尚不能成功地将社会经济因子的驱动力贡献加以定量分析和模拟。

研究问题： 并以环渤海地区为例, 探讨它在土地利用结构及其变化分析中的具体应用。

指标选取： 选取土地利用类型6个指标和一自然社会经济因子58个变量，

- 自然社会经济因子指标：X_1地形, X_2地貌, X_3气候, 同时期社会经济统计指标.
- 土地利用类型指标：Y_1耕地, Y_2林地, Y_3草地, Y_4水域, Y_5城建矿居用地, Y_6未利用地.

数据资料：

- 土地利用类型数据为1995 年遥感影像解译数据;
- 自然社会经济因子由1995年统计年鉴取得, 样本单元为行政县.

研究方法： 通过典型相关分析来来找出土地利用类型和自然社会经济之间相互影响的内在规律.

为什么要进行典型相关分析？

假设有两组变量$(X_1, \cdots X_p)$与$(Y_1, \cdots Y_q)$, 我们的研究目标是要找出这两组变量的相关关系, 如何给出两组变量之间的相关性定量描述？

(1)当$p = q = 1$时,用**相关系数**度量两个随机变量的相关关系

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

(2)当$p \geqslant 1, q = 1$(或$q \geqslant 1, p = 1$)时,用全相关系数度量一个随机变量与一组随机变量的相关关系. 假设p 维随机向量$X = (X_1, X_2, \cdots, X_p)'$, 并设$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_{p+1}(\mu, \Sigma)$, 其中,

$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \sigma_{YY} \end{bmatrix}$, 则称

$$R = \sqrt{\frac{\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}}{\sigma_{YY}}},$$

为Y与$X_1, X_2, \cdots, X_p$的全相关系数.

(3)当$p, q > 1$时, 如何分析两组向量之间的相关性呢?

典型相关分析的主要思想

利用主成分分析思想,可以把多个变量与多个变量之间的相关化为两个新的综合变量之间的相关.

即求$\alpha = (\alpha_1, \cdots, \alpha_p)'$和$\beta = (\beta_1, \cdots, \beta_q)'$,使得新的综合变量

$$\begin{aligned} V &= \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_p X_p, \\ W &= \beta_1 Y_1 + \beta_2 Y_2 + \cdots + \beta_q Y_q, \end{aligned}$$

之间有最大可能的相关,基于这个思想就产生了典型相关分析(Canonical correlation analysis).

10.1 总体典型相关

定义10.1.1 设$X=(X_1,\ldots,X_p)'$ 和$Y=(Y_1,\ldots,Y_q)'$为两组随机变量, $p+q$维随机向量$(X',Y')'$ 的均值向量为0, 协方差阵为$\Sigma>0$,(不妨设$p\le q$), 如果存在$a_1=(a_{11},\ldots,a_{p1})'$ 和$b_1=(b_{11},\ldots,b_{q1})'$, 使得

$$\rho(a_1'X,b_1'y)=\max_{\mathrm{Var}(\alpha'X)=1,\mathrm{Var}(\beta'Y)=1}\rho(\alpha'X,\beta'Y)$$

则称$a_1'X$ 和$b_1'Y$ 是X,Y 的第一典典型相关变量, 它们之间的相关系数为第一典型相关系数.

如果存在$a_k = (a_{1k}, \ldots, a_{pk})'$ 和$b_k = (b_{1k}, \ldots, b_{qk})'$, 使得

1. $a_k'X, b_k'Y$和前面$k-1$对典型相关变量都不相关;
2. $\mathrm{Var}(\alpha'X) = 1, \mathrm{Var}(\beta'Y) = 1$;
3. $a_k'X, b_k'Y$的相关系数最大;

则称$a_k'X, b_k'Y$是X, Y的**第k对典型相关变量**,它们之间的相关系数称为**第k个典型相关系数**$(k = 2, \cdots, p)$.

假设：随机向量$Z=(X',Y')'$, $\mathsf{Var}[X]=\Sigma_{11}$, $\mathsf{Var}[Y]=\Sigma_{22}$, $\mathsf{Cov}(X,Y)=\Sigma_{12}$. 已知

$$(1)\qquad \mathsf{E}[Z]=0.$$

$$(2)\qquad \mathsf{Var}[Z]=\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}>0.$$

1.第一对典型相关变量的求法

令$V=\alpha'X, W=\beta'Y$,则V,W的相关系数

$$\rho(V,W)=\frac{\alpha'\Sigma_{12}\beta}{\sqrt{\alpha'\Sigma_{11}\alpha}\sqrt{\beta'\Sigma_{22}\beta}}$$

求第一对典型相关变量就等价于求α和β,使得在条件$\mathrm{Var}(\alpha'X)=1$和$\mathrm{Var}(\beta'Y)=1$下,

$$\rho(\alpha'X,\beta'Y)=\alpha'\Sigma_{12}\beta$$

达到最大.

用拉格朗日乘子法,令

$$\varphi(\alpha,\beta)=\alpha'\Sigma_{12}\beta-\frac{\lambda_1}{2}(\alpha'\Sigma_{11}\alpha-1)-\frac{\lambda_2}{2}(\beta'\Sigma_{22}\beta-1)$$

对上式的α,β求偏导,并令其为零,得

$$\begin{cases}\dfrac{\partial\varphi}{\partial\alpha}=\Sigma_{12}\beta-\lambda_1\Sigma_{11}\alpha=0,\\[2ex]\dfrac{\partial\varphi}{\partial\beta}=\Sigma_{21}\alpha-\lambda_2\Sigma_{22}\beta=0,\end{cases}$$

再分别用α',β'左乘上面的方程,得

$$\lambda_1=\lambda_2=\alpha'\Sigma_{12}\beta=\rho(V,W)\overset{def}{=}\lambda,$$

则上述方程组等价于

$$\begin{cases}-\lambda\Sigma_{11}\alpha+\Sigma_{12}\beta=0,\\\Sigma_{21}\alpha-\lambda\Sigma_{22}\beta=0.\end{cases}\tag{1}$$

由方程组(1)不难得到

$$(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \lambda^2 I_p)\alpha = 0.$$

然后由p阶特征方程求解λ和α.类似地,可通过求解q阶特征方程

$$(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \lambda^2 I_q)\beta = 0.$$

来得到λ和β. 故等价于求解下面的方程组

$$\begin{cases} |\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \lambda^2 I_p| = 0, \\ |\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \lambda^2 I_q| = 0. \end{cases}$$

记:

$$\begin{aligned} M_1 &= \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \\ M_2 &= \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \end{aligned}$$

- 记：$A=\Sigma_{11}^{-1/2}$, $B=\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{22}^{-1/2}\Sigma_{21}$.
  则$M_1=AB$.
- AB 与$BA=\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}$有相同特征根.
- 记$T=\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$, $BA=TT'$, 即TT'与BA 有相同特征根.
- $M_2=T'T$, 与TT' 有相同特征根.
- **结论：**
  1. M_1 与TT' 有相同的特征根;
  2. 类似, M_2 与$T'T$ 有相同的特征根;
  3. M_1 与M_2 有相同特征根, 且非零特征根的个数至多p个.

- 设TT' 的p个特征根依次为$\lambda_1^2 \geq \cdots \geq \lambda_p^2 > 0$; 相应的特征向量为$l_1, \cdots, l_p$. 满足

$$\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}l_k = \lambda_k^2 l_k.$$

- 相应地, $T'T$只能有p个非零特征根$\lambda_1^2 \geq \cdots \geq \lambda_p^2 > 0$, 另有$q-p$个0特征根.

结论：求典型相关变量及典型相关系数的求解问题等价于求解TT' 的最大特征值及相应的特征向量.

- 对M_1,选择最大的特征根λ_1^2, $a_1 = \Sigma_{11}^{-1/2} l_1$为特征向量, 满足

$$\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11}^{-1/2} l_1) = \lambda_1^2(\Sigma_{11}^{-1/2} l_1).$$

  其中: $a_1'\Sigma_{11}a_1 = 1$.

- 对M_2, 选择最大的特征根λ_1^2, $b_1 = \lambda_1^{-1}\Sigma_{22}^{-1}\Sigma_{21}a_1$为特征向量, 满足

$$\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}b_1 = \lambda_1^2 b_1.$$

  其中: $b_1'\Sigma_{22}b_1 = 1$.

2.典型相关变量的一般求法

定理10.1.1 设$Z=(X',Y')'$. 已知

$$\mathrm{E}(Z)=0,\quad \mathrm{D}(Z)=\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}>0$$

记: $T=\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$, 并设$p$阶方阵$TT'$的非零特征值依次为$\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_p^2>0(\lambda_i>0,i=1,\cdots,p)$; 而$l_1,l_2,\cdots,l_p$为其相应的单位正交特征向量.令

$$a_k=\Sigma_{11}^{-1/2}l_k,\quad b_k=\lambda_k^{-1}\Sigma_{22}^{-1}\Sigma_{21}a_k\qquad(k=1,2,\cdots,p).$$

则$V_k=a_k'X,W_k=b_k'Y$为X,Y的第k对典型相关变量,λ_k为第k个典型相关系数.

- 以上定理10.1.1中,我们假定$\Sigma > 0$,一般情况下协方差阵非负定,从而$\Sigma_{11}^{-1}, \Sigma_{22}^{-1}$不一定存在.
- 但注意到方程组(1)总有非零解,因此可以用**广义逆矩阵**来求解.

定义10.1.2 给定一个矩阵A如果有矩阵D满足

$$ADA = A, \quad DAD = D, \quad (AD)' = AD, \quad (DA)' = DA$$

则称D是A的加号逆,记作A^+(A^+是存在唯一的).

定理10.1.2 设$Z=(X',Y')'$. 已知

$$\mathrm{E}(Z)=0,\quad \mathrm{D}(Z)=\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\geqslant 0.$$

- 记$T=(\Sigma_{11}^{1/2})^{+}\Sigma_{12}(\Sigma_{22}^{1/2})^{+}$.
- m=rank$(TT')\leqslant$min(p,q).
- 设p阶方阵TT'的非零特征值依次为$\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_m^2>0$.
- $l_1,l_2,\cdots,l_m$为其相应的单位正交特征向量.
- 令：$a_k=(\Sigma_{11}^{1/2})^{+}l_k, b_k=\lambda_k^{-1}(\Sigma_{22}^{1/2})^{+}\Sigma_{21}a_k$.
  $k=1,2,\cdots,m$.

则$V_k=a_k'X, W_k=b_k'Y$为X,Y的第k对典型相关变量,λ_k为第k个典型相关系数.

三、典型变量的性质

性质1 设$V_k = a_k'X, W_k = b_k'Y$为X, Y 的第k 对典型相关变量$(k = 1, \cdots, p)$; 令$V = (V_1, \cdots, V_p)', W = (W_1, \cdots, W_p)'$. 则

$$D\begin{bmatrix} V \\ W \end{bmatrix} = \begin{bmatrix} I_p & \Lambda \\ \Lambda & I_p \end{bmatrix}$$

其中$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots \lambda_p)$.

说明:

1. $V_i \quad (i = 1, \cdots, p)$互不相关, $\mathsf{Var}(V_i) = 1, \ \ i = 1, \cdots, p$;
2. $W_j \quad (j = 1, \cdots, p)$互不相关, $\mathsf{Var}(W_i) = 1, \ \ i = 1, \cdots, p$;
3. 且V_i与$W_j (i \neq j)$互不相关; 而$\rho(V_i, W_i) = \lambda_i \quad (i = 1, \cdots, p)$.

性质2 原始变量与典型变量之间的相关性.

记$A=(a_1,a_2,\cdots,a_p)$为$p\times p$矩阵, $B=(b_1,b_2,\cdots,b_p)$为$q\times p$矩阵. 则典型随机向量

$$\begin{aligned} V &= (V_1,\cdots,V_p)'=(a_1'X,\cdots,a_p'X)'=A'X; \\ W &= (W_1,\cdots,W_p)'=(b_1'Y,\cdots,b_p'Y)'=B'Y. \end{aligned}$$

则

1. $\mathrm{Cov}(X,V)=\Sigma_{11}A$; ($X$原始变量与$X$的典型变量)
2. $\mathrm{Cov}(X,W)=\Sigma_{12}B$; ($X$ 原始变量与Y的典型变量)
3. $\mathrm{Cov}(Y,V)=\Sigma_{21}A$; ($Y$ 原始变量与X的典型变量)
4. $\mathrm{Cov}(Y,W)=\Sigma_{22}B$. ($Y$ 原始变量与Y的典型变量)

说明：

1. 假定原始变量为标准化变量, 则上面求出的原始变量和典型变量之间的协方差阵即为相关系数矩阵.

2. 四个相关系数矩阵中各列（或各行）相关系数的平方和将用于典型冗余分析.

性质3 线性变换不变性

设X和Y分别为p维和q维随机向量,令

$$X^* = C'X + d, Y^* = G'Y + h,$$

其中C为$p \times p$非退化矩阵,d为p维向量,G为$q \times q$非退化矩阵,h为q维向量.则

1. X^*和Y^*的典型相关变量为$V_i^* = (a_i^*)'X^*$和$W_i = (b_i^*)'Y^*$,

$$(a_i)^* = C^{-1}a_i, b_i^* = G^{-1}b_i \quad (i = 1, \cdots, p),$$

其中：a_i, b_i是X和Y的第i对典型相关变量的系数;

2. X^*和Y^*的典型相关变量的典型相关系数为

$$\rho[(a_i^*)'X^*, (b_i^*)'Y^*] = \rho(a_i'X, b_i'Y),$$

即线性变换不改变相关性.

Example

已知p维随机变量X和q维随机向量Y的协方差阵分别为Σ_{11}和Σ_{22}. 试从$Z=(X',Y')'$ 的相关阵R 出发求典型相关变量和典型相关系数.

- $X^*=CX$, $Y^*=GY$, 其中$C=(diag(\Sigma_{11})^{-1/2}$, $G=(diag(\Sigma_{22})^{-1/2}$. 则有$\mathsf{Var}[Z^*]=\mathsf{Var}[(X^*)',(Y^*)')']=R$.
- $T^*=R_{11}^{-1/2}R_{12}R_{22}^{-1/2}$, $T^*(T^*)'$的特征根为$(\lambda_k^*)^2$, $k=1,\cdots,p$.
- $V_k^*=a_k^*X^*$, $W_k^*=b_k^*Y^*$ 为$X^*=CX$, $Y^*=GY$的典型变量.
- 由性质3, $a_k^*=C^{-1}a_k$, $b_k^*=G^{-1}b_k$.
- 故得$a_k=Ca_k^*$, $b_k=Gb_k^*$.

Example

已知标准化变量$X=(X_1,X_2)'$和$Y=(Y_1,Y_2)'$的相关阵分别为

$$R=\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix},$$

其中，

$$R_{11}=\begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}, R_{22}=\begin{bmatrix} 1 & \nu \\ \nu & 1 \end{bmatrix}, R_{12}=R_{21}=\begin{bmatrix} \beta & \beta \\ \beta & \beta \end{bmatrix},$$

其中: $0<\beta<1$. 试求X,Y 的典型相关变量和典型相关系数.

- $M^* = R_{11}^{-1}R_{12}R_{22}^{-1}R_{21} = \frac{2\beta^2}{(1+\alpha)(1+\nu)}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.
- M^* 的特征值为

$$\lambda_1^2 = \frac{4\beta^2}{(1+\alpha)(1+\nu)},\ \lambda_2^2 = 0.$$

- M_1^* 对应于λ_1^2 的特征向量为$(1/\sqrt{2}, 1/\sqrt{2})$.

- 满足$a'R_{11}a = 1$的向量,

$$a = \frac{1}{\sqrt{2(1+\alpha)}}\begin{bmatrix}1\\1\end{bmatrix}, b = \frac{1}{\sqrt{2(1+\nu)}}\begin{bmatrix}1\\1\end{bmatrix}.$$

- 第一对典型相关变量

$$\begin{aligned} V_1 &= a'X = \frac{1}{\sqrt{2(1+\alpha)}}(X_1 + X_2), \\ W_1 &= b'X = \frac{1}{\sqrt{2(1+\nu)}}(Y_1 + Y_2), \end{aligned}$$

- 第一对典型相关变量的相关系数

$$\rho_1 = \frac{2\beta}{\sqrt{(1+\alpha)(1+\nu)}}.$$

10.2 样本典型相关

设总体$Z=(X_1,\cdots,X_p,Y_1,\cdots,Y_q)'$, 在实际问题中, 总体的均值$\mathrm{E}(Z)=\mu$和协方差阵$\mathrm{D}(Z)=\Sigma$是未知的,因而无法求得总体典型相关变量和典型相关系数. 首先需要根据观测到的样本数据阵对Σ进行估计.

已知总体Z的n次观测数据为:

$$Z_{(t)}=\begin{bmatrix} X_{(t)} \\ Y_{(t)} \end{bmatrix}_{(p+q)\times 1} \qquad (t=1,2,\cdots,n)$$

样本均值

$$\bar{Z} = \frac{1}{n}\sum_{t=1}^{n} Z_{(t)}$$

样本协方差阵

$$S = \frac{1}{n-1}\sum_{t=1}^{n}(Z_{(t)} - \bar{Z})(Z_{(t)} - \bar{Z})'$$

若假定$Z \sim N_{p+q}(\mu, \Sigma)$,则

$$\hat{\Sigma} = \frac{1}{n}\sum_{t=1}^{n}(Z_{(t)} - \bar{Z})(Z_{(t)} - \bar{Z})'$$

为协方差阵Σ的最大似然估计

设$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$,$\Sigma_{11}$为$p$阶矩阵.将$S$相应剖分为

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

显然,$S_{ij}(i,j=1,2)$是Σ_{ij}的无偏估计量.

下面从样本协方差阵出发,讨论两组变量间的关系.

一、样本典型相关变量和典型相关系数

1.从样本协方差阵S出发导出样本典型变量和样本典型相关系数(不妨设$S>0$)

- 令$\hat{T}=S_{11}^{-1/2}S_{12}S_{22}^{-1/2}$,
- 设$\hat{T}\hat{T}'$的特征值依次为$\hat{\lambda}_1^2\geqslant\hat{\lambda}_2^2\geqslant\cdots\geqslant\hat{\lambda}_p^2>0$,
  $(\hat{\lambda}_i>0.i=1,\cdots,p)$.
- $\hat{l}_k(k=1,\cdots,p)$为$\hat{T}\hat{T}'$的特征根$\hat{\lambda}_k^2$所对应的正交特征向量.
- 令

$$\begin{cases}\hat{a}_k=S_{11}^{-1/2}\hat{l}_k\\ \hat{b}_k=\hat{\lambda}_k^{-1}S_{22}^{-1}S_{21}\hat{a}_k,\end{cases}$$

- $\hat{V}_k=\hat{a}_k'X,\hat{W}_k=\hat{b}_k'Y$为$X,Y$的第$k$对样本典型相关变量;而$\hat{\lambda}_k$为$X,Y$的第$k$个样本典型相关系数.

2.从样本相关阵R出发导出样本典型变量和样本典型相关系数

- 设样本相关阵$R=(r_{ij})$,其中$r_{ij}=s_{ij}/\sqrt{s_{ii}s_{jj}}$,$s_{ij}$为样本协方差阵$S$的元素.
- 把R相应剖分为

$$R=\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

- 记:
$$D_1 = \begin{bmatrix} \sqrt{s_{11}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{s_{pp}} \end{bmatrix},$$
$$D_2 = \begin{bmatrix} \sqrt{s_{p+1,P+1}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{s_{p+q,p+q}} \end{bmatrix}.$$

- 记$\tilde{T} = R_{11}^{-1/2} R_{12} R_{22}^{-1/2}$,
- $\tilde{T}\tilde{T}'$的特征值依次为$\hat{\lambda}_1^2 \geqslant \hat{\lambda}_2^2 \geqslant \cdots \geqslant \hat{\lambda}_p^2 > 0$.
- X, Y的第k对样本典型相关变量为

$$\begin{cases} \hat{V}_k = (D_1^{-1} R_{11}^{-1/2} \hat{l}_k)' X = \hat{a}_k' X, \\ \hat{W}_k = (D_2^{-1} \hat{\lambda}_k^{-1} R_{22}^{-1} R_{21} R_{11}^{-1/2} \hat{l}_k)' Y = \hat{b}_k' Y, \end{cases}$$

- $\hat{\lambda}_k$为X, Y的第k个样本典型相关系数.

二、典型相关系数的显著性检验

1.检验$H_0 : \Sigma_{12} = 0$

设总体$Z \sim N_{p+q}(\mu, \Sigma)$,用似然比检验法可导出检验$H_0$的似然比统计量

$$\Lambda = \frac{|S|}{|S_{11}||S_{22}|} \tag{1}$$

其中$p+q$阶矩阵S是Σ的最大似然估计,S_{11}, S_{22}分别是Σ_{11}, Σ_{22}的最大似然估计.

协方差阵分解

利用矩阵的行列式及其分块行列式的关系,可得

$$\begin{aligned} |S| &= |S_{22}| \cdot |S_{11} - S_{12}S_{22}^{-1}S_{21}| \\ &= |S_{22}| \cdot |S_{11}| \cdot |I_p - S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}| \end{aligned}$$

故对假设的检验可采用下面的统计量

$$\Lambda = |I_p - S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}| = \prod_{i=1}^{p}(1 - \hat{\lambda}_i^2)$$

统计量的近似分布

当$n \to \infty$时,在H_0成立时有

$$P\{-m\ln\Lambda \leqslant C\} \approx .P\{V \leqslant C\}$$

其中

- $m = n - 1 - \frac{1}{2}(p + q + 1), f = pq,$
- $V = -m\ln\Lambda$ 近似服从$\chi^2(f)$的统计量.

拒绝域

当样本容量n足够大时,由样本值计算样本典型相关系数$\hat{\lambda}_i^2(i=1,2,\cdots,p)$及

$$\begin{aligned} Q_1 &= -m\ln\prod_{i=1}^{p}(1-\hat{\lambda}_i^2) = -m\sum_{i=1}^{p}\ln(1-\hat{\lambda}_i^2), \\ P_{value} &= P\{V > Q_1\} \quad (V \sim \chi^2(f)). \end{aligned}$$

如果$P_{value} < \alpha$ (例如：$\alpha = 0.05$), 则否定H_0,即两组变量X和Y相关;否则与H_0相容.

2.检验$H_0 : \lambda_k = 0 (k = 2, 3, \cdots, p)$

- **检验统计量** 采用Bartlett提出的大样本χ^2检验,取检验统计量为

$$Q_k = -[n - k - \frac{1}{2}(p + q + 1)] \sum_{i=k}^{p} \ln(1 - \hat{\lambda}_i^2).$$

- **统计量的分布** $Q_k \sim \chi^2(f_k)$, $f_k = (p - k + 1)(q - k + 1)$.
- **统计量的值** 由样本值计算样本典型相关系数$\hat{\lambda}_i^2$ $i = 1, \cdots, p$及$Q_k = c$值.
- **计算**P **值** $P_{value}(k) = P\{Q_k > c\}$.

- **检验结论** 对给定的显著水平α,
  - $P_{value}(k) < \alpha$,则否定H_0,即第k个典型相关变量显著不等于0.
  - $P_{value}(k) \geq \alpha$, 认为$\lambda_k = 0$.

- **检验过程**

  对$H_0^{(k)}$从$k = 2$开始逐个检验,直到某个k_0,使$H_0^{(k_0)}$相容时为止. 这是说明第k_0个及以后的所有典型相关系数均为0.

三、样本典型变量的得分值

- 假设经检验,有$r(r \leqslant p)$个典型相关系数显著不等于0, 则得到r 对典型相关变量$(V_i, W_i)(i = 1, \cdots, r)$.
- 将样品代入典型变量中,令

$$v_{ti} = a_i'(X_{(t)} - \bar{X}), \quad w_{ti} = b_i'(Y_{(t)} - \bar{Y})$$

其中：$i = 1, \cdots, r$; $t = 1, \cdots, n$, 称(v_{ti}, w_{ti})为第t个样品$Z_{(t)}$的第i对**样本典型变量的得分值.**

- 对每个i,可用$(v_{ti}, w_{ti})(t = 1, 2, \cdots, n)$来绘制散点的散布图,散点应近似在一直线上,若有异常点,则应分析原因.

载入CCA模块

- 载入数据$D <- read.table("ex1021.txt", header = F)$
- 数据标准化$D1 <- scale(D, center = TRUE, scale = TRUE)$
- 计算典型相关$re <- cc(D1[, 1:2], D1[, 3:4])$

载入CCA时需要载入的包：

```
载入需要的程辑包：fda
载入需要的程辑包：splines
载入需要的程辑包：zoo
载入程辑包：  'zoo'
载入需要的程辑包：Matrix
载入需要的程辑包：lattice
载入需要的程辑包：RCurl
载入需要的程辑包：bitops
载入程辑包：  'fda'
载入需要的程辑包：fields
载入需要的程辑包：spam
载入需要的程辑包：grid
载入程辑包：  'spam'
载入需要的程辑包：maps
```

例10.2.1

为了了解某矿区下部矿Pt(铂)、Pd(钯) 与Cu(铜)、Ni(镍) 的共生组合规律。我们从其钻孔中取出27 个样品（数据见下表）。试用典型相关分析研究Pt、Pd 与Cu、Ni 的相关关系。

序号	Pt X_1	Pd X_2	Cu Y_1	Ni Y_2
1	0.14	0.3	0.03	0.14
2	0.2	0.5	0.14	0.22
3	0.06	0.11	0.03	0.02
4	0.07	0.11	0.04	0.13
5	0.12	0.22	0.06	0.12
6	0.52	0.87	0.19	0.2
7	0.23	0.47	0.14	0.1
8	1.19	0.38	0.09	0.11
9	0.37	0.66	0.14	0.15
10	0.36	0.6	0.12	0.14
11	0.42	0.77	0.17	0.1
12	0.35	0.85	0.3	0.19
13	0.5	0.87	0.23	0.22
14	0.56	1.15	0.29	0.28

序号	Pt X_1	Pd X_2	Cu Y_1	Ni Y_2
15	0.43	0.9	0.13	0.22
16	0.47	0.97	0.26	0.22
17	0.49	0.79	0.21	0.2
18	0.47	0.77	0.51	0.22
19	0.4	0.88	0.33	0.19
20	0.66	1.3	0.21	0.3
21	0.63	1.3	0.45	0.28
22	0.52	1.43	0.31	0.23
23	0.44	0.87	0.17	0.25
24	0.03	0.07	0.05	0.08
25	0.2	0.28	0.04	0.08
26	0.04	0.1	0.11	0.07
27	0.17	0.28	0.15	0.09

```
$cor
[1] 0.894617586 0.009495891

$xcoef
         [,1]       [,2]
V1  0.02625195 -1.2753660
V2 -1.01608697  0.7712424

$ycoef
         [,1]     [,2]
V3 -0.3231000  1.328168
V4 -0.7513873 -1.141858
```

10.3 典型冗余分析

- 由样本数据阵Z计算样本协方差阵S,
- 由S矩阵求出样本典型变量,
- 计算原始变量与r对典型变量之间的相关系数矩阵(或称典型结构).

假定两组变量都为标准化变量. 记原始变量X(或Y)与典型变量V(或W)的相关阵为

$$\begin{aligned} R(X,V) &= R_{11}A = (R_{11}a_1, \cdots, R_{11}a_r) \\ &\overset{def}{=} \begin{bmatrix} r(X_1,V_1) & \cdots & r(X_1,V_r) \\ \vdots & & \vdots \\ r(X_p,V_1) & \cdots & r(X_p,V_r) \end{bmatrix}_{p\times p} \end{aligned}$$

$$\begin{aligned} R(Y,W) &= R_{22}B = (R_{22}b_1, \cdots, R_{22}b_r) \\ &\overset{def}{=} \begin{bmatrix} r(Y_1,W_1) & \cdots & r(Y_1,W_r) \\ \vdots & & \vdots \\ r(Y_q,W_1) & \cdots & r(Y_q,W_r) \end{bmatrix}_{q\times p} \end{aligned}$$

一、总变差的百分比

设$\text{rank}(\Sigma_{12}) = r \leqslant \min(p, q)$.类似与主成分分析,

- 把V_k看成时由第一组标准化变量X提取的成分,
  $R(X, V) = R_{11}A = [r(X_j, V_k)]_{p\times r}$
- W_k看成时由第二组标准化变量Y提取的成分,
  $R(Y, W) = R_{22}B = [r(Y_j, W_k)]_{q\times r}$.

分别计算第k列平方和除以原变量组变差总和p或q.

记

$$R_d(X;V_k)=\frac{1}{p}\sum_{j=1}^{p}r^2(X_j,V_k),$$

$$R_d(Y;W_k)=\frac{1}{q}\sum_{j=1}^{q}r^2(Y_j,W_k)$$

并称

- $R_d(X;V_k)$为第k个典型变量V_k解释本组变量X总变差的百分比.
- $R_d(Y;W_k)$为第k个典型变量W_k解释本组变量Y总变差的百分比.

记

$$R_d(X;V_1,\cdots,V_m)=\frac{1}{p}\sum_{k=1}^{m}\sum_{j=1}^{p}r^2(X_j,V_k),$$

$$R_d(Y;W_1,\cdots,W_m)=\frac{1}{q}\sum_{k=1}^{m}\sum_{j=1}^{q}r^2(Y_j,W_k),$$

并称

- $R_d(X;V_1,\cdots,V_m$ 为前$m(m\leqslant r)$个典型变量V_k 解释本组变量X **总变差的累计百分比**.
- $R_d(Y;W_1,\cdots,W_m$ 为前$m(m\leqslant r)$个典型变量W_k解释本组变量Y **总变差的累计百分比**.

二、典型冗余分析

意义： 典型冗余分析即分析典型变量解释另一组变量总变差的百分比. 冗余测度体现了两组变量之间的相关程度.

计算方法： 利用典型变量解释本组变差的百分比来计算解释另一组变差百分比的公式: $k = 1, \cdots, r$,

$$R_d(X; W_k) = \frac{1}{p}\sum_{j=1}^{p} r^2(X_j, W_k) = \lambda_k^2 R_d(X; V_k),$$

$$R_d(Y; V_k) = \frac{1}{q}\sum_{j=1}^{q} r^2(Y_j, V_k) = \lambda_k^2 R_d(Y; W_k).$$

冗余解释：

- $R_d(X;W_k)$表示第一组中典型变量解释原变量组的变差被第二组中典型变量重复解释的百分比, 简称为第一组典型变量的**冗余测度;**
- $R_d(Y;V_k)$表示第二组中典型变量解释原变量组的变差被第一组中典型变量重复解释的百分比, 简称为第二组典型变量的**冗余测度;**

冗余测度的大小表示这对典型变量能够对另一组变量解释的程度大小，它将为进一步讨论多对多建模提供一些有用的信息.

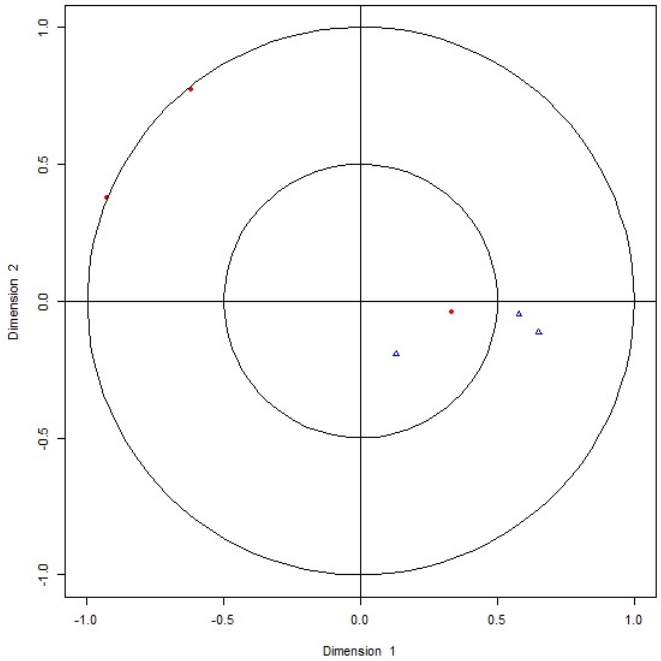
例10.3.1（康复俱乐部20名成员测试数据的典型相关分析）

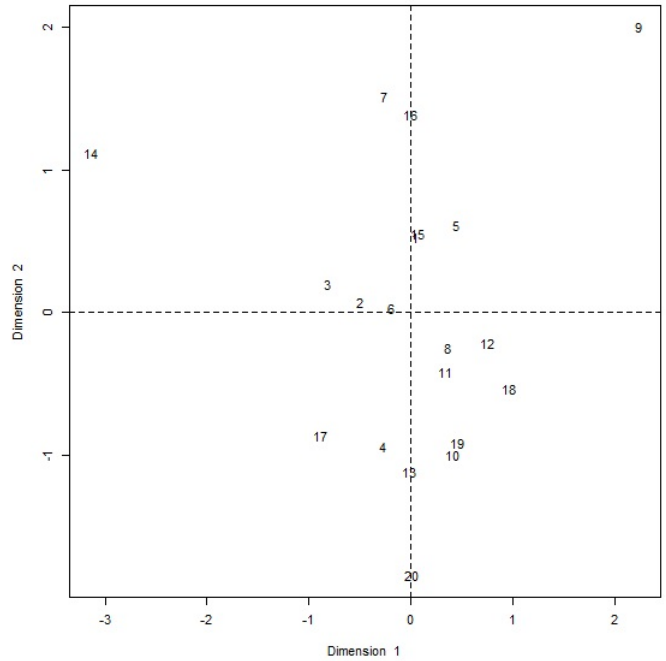
康复俱乐部对20名中年人测量了三个生理指标：WEIGHT（体重）、WAIST（腰围）、PULSE（脉搏），以及三个训练指标：CHINS（单杠）、SITUPS（仰卧起坐）和JUMP（跳高）。

试分析生理指标和训练指标这两组变量间的相关性。

体重	腰围	脉搏	单杠	仰卧起坐	跳高
191	36	50	5	162	60
193	38	58	12	101	101
189	35	46	13	155	58
211	38	56	8	101	38
176	31	74	15	200	40
169	34	50	17	120	38
154	34	64	14	215	105
193	36	46	6	70	31
176	37	54	4	60	25
156	33	54	15	225	73
189	37	52	2	110	60
162	35	62	12	105	37
182	36	56	4	101	42
167	34	60	6	125	40
154	33	56	17	251	250
166	33	52	13	210	115
247	46	50	1	50	50
202	37	62	12	210	120
157	32	52	11	230	80
138	33	68	2	110	43

	WilksL	F	df1	df2	p
1	0.3503905	2.04823353	9	34.22293	0.06353094
2	0.9547227	0.17578229	4	30.00000	0.94912025
3	0.9947336	0.08470926	1	16.00000	0.77475327


Dimension 2
Dimension 1
1.0
0.5
0.0
-0.5
-1.0


Dimension 2
Dimension 1
9
7
16
14
5
15
3
2
6
8
12
11
18
17
4
19
10
13
20

$xcoef

	[,1]	[,2]	[,3]
V1	0.77539761	1.8843672	-0.1909822
V2	-1.57934657	-1.1806411	0.5060195
V3	0.05912012	0.2311068	1.0507838

$ycoef

	[,1]	[,2]	[,3]
V4	0.3494969	0.3755436	-1.2965937
V5	1.0540110	-0.1234905	1.2367934
V6	-0.7164267	-1.0621670	-0.4188073

```
$scores$corr.X.xscores
          [,1]       [,2]        [,3]
V1 -0.6206424  0.7723919 -0.13495886
V2 -0.9254249  0.3776614 -0.03099486
V3  0.3328481 -0.0414842  0.94206752
$scores$corr.Y.xscores
          [,1]       [,2]        [,3]
V4 0.5789047 -0.0475222 -0.04671717
V5 0.6505914 -0.1149232  0.00395139
V6 0.1290401 -0.1922586 -0.01697689
```

```
$scores$corr.X.yscores
          [,1]        [,2]         [,3]
V1 -0.4937881  0.154907853 -0.009794003
V2 -0.7362756  0.075742277 -0.002249306
V3  0.2648166 -0.008319907  0.068366110
$scores$corr.Y.yscores
          [,1]       [,2]        [,3]
V4 0.7276254 -0.2369522 -0.64375064
V5 0.8177285 -0.5730231  0.05444915
V6 0.1621905 -0.9586280 -0.23393722
```

典型冗余分析

（1）计算冗余测度：

```
BX<-cc2$scores$corr.X.xscores（解释组内的比例）
RX<-(diag(t(BX)%*%BX))/2
RX
BXY<-cc2$scores$corr.X.yscores（解释另一组的比例）
RXY<-(diag(t(BXY)%*%BXY))/2
RXY
```

（2）各组典型变量解释组内的比例

[1] 0.4508, 0.2470, 0.3022

（3）各组典型变量解释另一组的比例（即冗余）

[1] 0.2854, 0.0099, 0.0016