第十章 典型相关分析

- §10.1 引言
- §10.2 总体典型相关
- §10.3 样本典型相关
- §10.4 典型相关系数的显著性检验

§10.1 引言

- 典型相关分析（canonical correlation analysis）是研究两组变量之间相关关系的一种统计分析方法，它能够有效地揭示两组变量之间的相互线性依赖关系。
- 典型相关分析是由霍特林（Hotelling,1935,1936）首先提出的。

§10.2 总体典型相关

- 一、典型相关的定义及导出
- 二、典型相关变量的性质
- 三、从相关矩阵出发计算典型相关

一、典型相关的定义及导出

- 设$\boldsymbol{x}=(x_1,x_2,\cdots,x_p)'$和$\boldsymbol{y}=(y_1,y_2,\cdots,y_q)'$是两组随机变量，且$V(\boldsymbol{x})=\boldsymbol{\Sigma}_{11}(>0)$，$V(\boldsymbol{y})=\boldsymbol{\Sigma}_{22}(>0)$，$\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{\Sigma}_{12}$，即有

$$V\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}$$

  其中$\boldsymbol{\Sigma}_{21}=\boldsymbol{\Sigma}_{12}'$。

- 我们研究$u=\boldsymbol{a}'\boldsymbol{x}$与$v=\boldsymbol{b}'\boldsymbol{y}$之间的相关关系，其中

$$\boldsymbol{a}=(a_1,a_2,\cdots,a_p)',\quad \boldsymbol{b}=(b_1,b_2,\cdots,b_q)'$$

  现来计算一下u与v的相关系数。

$$\mathrm{Cov}(u,v)=\mathrm{Cov}(\boldsymbol{a}'\boldsymbol{x},\boldsymbol{b}'\boldsymbol{y})=\boldsymbol{a}'\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{b}=\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}$$

$$V(u)=V(\boldsymbol{a}'\boldsymbol{x})=\boldsymbol{a}'V(\boldsymbol{x})\boldsymbol{a}=\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}$$

$$V(v)=V(\boldsymbol{b}'\boldsymbol{y})=\boldsymbol{b}'V(\boldsymbol{y})\boldsymbol{b}=\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}$$

所以，u与v的相关系数

$$\rho(u,v)=\frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}}}$$

由于对任意非零常数k_1和k_2，有

$$\rho(k_1u,k_2v)=\rho(u,v)$$

因此，为避免不必要的结果重复，我们常常限定u与v均为标准化的变量，即附加约束条件

$$V(u)=1,\quad V(v)=1$$

即

$$\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}=1,\quad \boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}=1$$

在此约束条件下，求$\boldsymbol{a}\in R^p$和$\boldsymbol{b}\in R^q$，使得

$$\rho(u,v)=\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}$$

达到最大。

- 容易证明，$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$和$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$有着相同的非零特征值，且皆为正，其个数为$m=\mathrm{rank}(\boldsymbol{\Sigma}_{12})$。将这些正特征值分别记为 $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_m^2 > 0$。设$\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_m$为 $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ 的相应于$\rho_1^2, \rho_2^2, \cdots, \rho_m^2$的特征向量，且满足标准化条件

$$\boldsymbol{a}_i'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_i=1，\ i=1,2,\cdots,m$$

令$\boldsymbol{b}_i = \dfrac{1}{\rho_i}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a}_i$，则有

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{b}_i = \frac{1}{\rho_i}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\left(\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a}_i\right)$$

$$= \frac{1}{\rho_i}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\left(\rho_i^2\boldsymbol{a}_i\right) = \rho_i^2\boldsymbol{b}_i$$

从而$\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_m$为$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ 的相应于$\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_m^2 > 0$ 的特征向量，并且满足

$$\boldsymbol{b}_i'\boldsymbol{\Sigma}_{22}\boldsymbol{b}_i = \frac{1}{\rho_i^2}\boldsymbol{a}_i'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a}_i$$

$$= \frac{1}{\rho_i^2}\boldsymbol{a}_i'\boldsymbol{\Sigma}_{11}\left(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a}_i\right)$$

$$= \frac{1}{\rho_i^2}\boldsymbol{a}_i'\boldsymbol{\Sigma}_{11}\left(\rho_i^2\boldsymbol{a}_i\right) = 1, \quad i = 1, 2, \cdots, m$$

- 可以证明，当取$\boldsymbol{a}=\boldsymbol{a}_1,\boldsymbol{b}=\boldsymbol{b}_1$时，$\rho(u,v)=\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}$达到最大值$\rho_1$(显然$\rho_1\leq 1$)。我们称

$$u_1=\boldsymbol{a}_1'\boldsymbol{x},\ v_1=\boldsymbol{b}_1'\boldsymbol{y}$$

为第一对典型相关变量，称ρ_1为第一个典型相关系数

- 第一对典型相关变量u_1,v_1提取了原始变量$\boldsymbol{x}$与$\boldsymbol{y}$之间相关的主要部分，如果这一部分还显得不够，可以在剩余相关中再求出第二对典型相关变量$u_2=\boldsymbol{a}'\boldsymbol{x},v_2=\boldsymbol{b}'\boldsymbol{y}$，也就是$\boldsymbol{a},\boldsymbol{b}$应满足标准化条件且应使得第二对典型相关变量不包括第一对典型相关

变量所含的信息，即

$$\rho(u_2,u_1)=\rho(\boldsymbol{a}'\boldsymbol{x},\boldsymbol{a}_1'\boldsymbol{x})=\mathrm{Cov}(\boldsymbol{a}'\boldsymbol{x},\boldsymbol{a}_1'\boldsymbol{x})=\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_1=0$$

$$\rho(v_2,v_1)=\rho(\boldsymbol{b}'\boldsymbol{y},\boldsymbol{b}_1'\boldsymbol{y})=\mathrm{Cov}(\boldsymbol{b}'\boldsymbol{y},\boldsymbol{b}_1'\boldsymbol{y})=\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}_1=0$$

在这些约束条件下使得

$$\rho(u_2,v_2)=\rho(\boldsymbol{a}'\boldsymbol{x},\boldsymbol{b}'\boldsymbol{y})=\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}$$

达到最大。

- 一般地，第$i(1<i\leq m)$对典型相关变量$u_i=\boldsymbol{a}'\boldsymbol{x},v_i=\boldsymbol{b}'\boldsymbol{y}$是指，找出$\boldsymbol{a}\in R^p,\boldsymbol{b}\in R^q$,在约束条件

$$\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}=1，\ \boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}=1$$

$$\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_k=0，\ \boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}_k=0，\ k=1,2,\cdots,i-1$$

下，使得

$$\rho(u_i,v_i)=\rho(\boldsymbol{a}'\boldsymbol{x},\boldsymbol{b}'\boldsymbol{y})=\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}$$

达到最大。当取$\boldsymbol{a}=\boldsymbol{a}_i,\boldsymbol{b}=\boldsymbol{b}_i$时，$\rho(u_i,v_i)$达到最大值$\rho_i$，称它为第$i$个典型相关系数，称$\boldsymbol{a}_i,\boldsymbol{b}_i$为第$i$对典型系数。

二、典型相关变量的性质

- 1.同一组的典型变量互不相关
- 2.不同组的典型变量之间的相关性
- 3.原始变量与典型变量之间的相关系数
- 4.简单相关、复相关和典型相关之间的关系

1.同一组的典型变量互不相关

- 设$\boldsymbol{x},\boldsymbol{y}$的第$i$对典型变量为

$$u_i=\boldsymbol{a}_i'\boldsymbol{x},\quad v_i=\boldsymbol{b}_i'\boldsymbol{y},\qquad i=1,2,\cdots,m$$

则有

$$V(u_i)=\boldsymbol{a}_i'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_i=1,\quad V(v_i)=\boldsymbol{b}_i'\boldsymbol{\Sigma}_{22}\boldsymbol{b}_i=1,\quad i=1,2,\cdots,m$$

$$\rho(u_i,u_j)=\mathrm{Cov}(u_i,u_j)=\boldsymbol{a}_i'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_j=0,\quad 1\le i\ne j\le m$$

$$\rho(v_i,v_j)=\mathrm{Cov}(v_i,v_j)=\boldsymbol{b}_i'\boldsymbol{\Sigma}_{22}\boldsymbol{b}_j=0,\quad 1\le i\ne j\le m$$

2.不同组的典型变量之间的相关性

$$\rho(u_i,v_i)=\rho_i,\quad i=1,2,\cdots,m$$

$$\rho\left(u_i,v_j\right)=\mathrm{Cov}\left(u_i,v_j\right)=\mathrm{Cov}\left(\boldsymbol{a}_i'\boldsymbol{x},\boldsymbol{b}_j'\boldsymbol{y}\right)=\boldsymbol{a}_i'\mathrm{Cov}\left(\boldsymbol{x},\boldsymbol{y}\right)\boldsymbol{b}_j$$

$$=\boldsymbol{\alpha}_i'\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\beta}_j=\rho_j\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_j=0,\qquad 1\le i\ne j\le m$$

记$\boldsymbol{u}=(u_1,u_2,\cdots,u_m)'$，$\boldsymbol{v}=(v_1,v_2,\cdots,v_m)'$，则上述两个性质可用矩阵表示为

$$V(\boldsymbol{u})=\boldsymbol{I}_m,\quad V(\boldsymbol{v})=\boldsymbol{I}_m,\quad \mathrm{Cov}(\boldsymbol{u},\boldsymbol{v})=\boldsymbol{\Lambda}$$

或

$$V\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{pmatrix}=\begin{pmatrix}\boldsymbol{I}_m & \boldsymbol{\Lambda}\\ \boldsymbol{\Lambda} & \boldsymbol{I}_m\end{pmatrix}$$

其中$\boldsymbol{\Lambda}=\mathrm{diag}(\rho_1,\rho_2,\cdots,\rho_m)$。

3.原始变量与典型变量之间的相关系数

- 记

$$\boldsymbol{A}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_m)=(a_{ij})_{p\times m}$$
$$\boldsymbol{B}=(\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_m)=(b_{ij})_{q\times m}$$

$$\boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}=\begin{pmatrix}\sigma_{11} & \cdots & \sigma_{1p} & \sigma_{1,p+1} & \cdots & \sigma_{1,p+q}\\ \vdots & & \vdots & \vdots & & \vdots\\ \sigma_{p1} & \cdots & \sigma_{pp} & \sigma_{p,p+1} & \cdots & \sigma_{p,p+q}\\ \sigma_{p+1,1} & \cdots & \sigma_{p+1,p} & \sigma_{p+1,p+1} & \cdots & \sigma_{p+1,p+q}\\ \vdots & & \vdots & \vdots & & \vdots\\ \sigma_{p+q,1} & \cdots & \sigma_{p+q,p} & \sigma_{p+q,p+1} & \cdots & \sigma_{p+q,p+q}\end{pmatrix}$$

则

$$\mathrm{Cov}(\boldsymbol{x},\boldsymbol{u})=\mathrm{Cov}(\boldsymbol{x},\boldsymbol{A}'\boldsymbol{x})=\boldsymbol{\Sigma}_{11}\boldsymbol{A}$$
$$\mathrm{Cov}(\boldsymbol{x},\boldsymbol{v})=\mathrm{Cov}(\boldsymbol{x},\boldsymbol{B}'\boldsymbol{y})=\boldsymbol{\Sigma}_{12}\boldsymbol{B}$$
$$\mathrm{Cov}(\boldsymbol{y},\boldsymbol{u})=\mathrm{Cov}(\boldsymbol{y},\boldsymbol{A}'\boldsymbol{x})=\boldsymbol{\Sigma}_{21}\boldsymbol{A}$$
$$\mathrm{Cov}(\boldsymbol{y},\boldsymbol{v})=\mathrm{Cov}(\boldsymbol{y},\boldsymbol{B}'\boldsymbol{y})=\boldsymbol{\Sigma}_{22}\boldsymbol{B}$$

上述四个等式也可表达为

$$\mathrm{Cov}\left(x_i,u_j\right)=\left(\sigma_{i1},\cdots,\sigma_{ip}\right)\begin{pmatrix}a_{1j}\\ \vdots\\ a_{pj}\end{pmatrix}=\sum_{k=1}^{p}\sigma_{ik}a_{kj}$$

$$\mathrm{Cov}\left(x_i,v_j\right)=\left(\sigma_{i,p+1},\cdots,\sigma_{i,p+q}\right)\begin{pmatrix}b_{1j}\\ \vdots\\ b_{qj}\end{pmatrix}=\sum_{k=1}^{q}\sigma_{i,p+k}b_{kj}$$

$$\mathrm{Cov}(y_i,u_j)=(\sigma_{p+i,1},\cdots,\sigma_{p+i,p})\begin{pmatrix}a_{1j}\\ \vdots\\ a_{pj}\end{pmatrix}=\sum_{k=1}^{p}\sigma_{p+i,k}a_{kj}$$

$$\mathrm{Cov}(y_i,v_j)=(\sigma_{p+i,p+1},\cdots,\sigma_{p+i,p+q})\begin{pmatrix}b_{1j}\\ \vdots\\ b_{qj}\end{pmatrix}=\sum_{k=1}^{q}\sigma_{p+i,p+k}b_{kj}$$

$$i=1,2,\cdots,q,\quad j=1,2,\cdots,m$$

所以

$$\rho(x_i,u_j)=\sum_{k=1}^{p}\sigma_{ik}a_{kj}\Big/\sqrt{\sigma_{ii}},\quad \rho(x_i,v_j)=\sum_{k=1}^{q}\sigma_{i,p+k}b_{kj}\Big/\sqrt{\sigma_{ii}}$$

$$\rho(y_i,u_j)=\sum_{k=1}^{p}\sigma_{p+i,k}a_{kj}\Big/\sqrt{\sigma_{p+i,p+i}},\quad \rho(y_i,v_j)=\sum_{k=1}^{q}\sigma_{p+i,p+k}b_{kj}\Big/\sqrt{\sigma_{p+i,p+i}}$$

$$i=1,2,\cdots,p,\quad j=1,2,\cdots,m$$

4.简单相关、复相关和典型相关之间的关系

- 当$p=q=1$时，$\boldsymbol{x}$与$\boldsymbol{y}$之间的（惟一）典型相关就是它们之间的简单相关；当$p=1$或$q=1$时，$\boldsymbol{x}$与$\boldsymbol{y}$之间的（惟一）典型相关就是它们之间的复相关。可见，复相关是典型相关的一个特例，而简单相关是复相关的一个特例。
- 第一个典型相关系数至少同$\boldsymbol{x}$（或$\boldsymbol{y}$）的任一分量与$\boldsymbol{y}$（或$\boldsymbol{x}$）的复相关系数一样大，即使所有这些复相关系数都较小，第一个典型相关系数仍可能很大；同样，从复相关的定义也可以看出，当$p=1$（或$q=1$）时，x（或y）与$\boldsymbol{y}$（或$\boldsymbol{x}$）之间的复相关系数也不会小于

  x（或y）与$\boldsymbol{y}$（或$\boldsymbol{x}$）的任一分量之间的相关系数，即使所有这些相关系数都较小，复相关系数仍可能很大。

三、从相关矩阵出发计算典型相关

- 有时，x和y的各分量的单位不全相同，我们希望在对各分量作标准化变换之后再作典型相关分析。
- 记$\mu_1=E(x)$，$\mu_2=E(y)$，$D_1=\mathrm{diag}\left(\sqrt{\sigma_{11}},\cdots,\sqrt{\sigma_{pp}}\right)$

$D_2=\mathrm{diag}\left(\sqrt{\sigma_{p+1,p+1}},\cdots,\sqrt{\sigma_{p+q,p+q}}\right)$，$R=\begin{pmatrix}R_{11}&R_{12}\\R_{21}&R_{22}\end{pmatrix}$为$\begin{pmatrix}x\\y\end{pmatrix}$的相关矩阵。对$x$和$y$的各分量作标准化变换，即令

$$x^*=D_1^{-1}(x-\mu_1),\quad y^*=D_2^{-1}(y-\mu_2)$$

现在来求x^*和y^*的典型相关变量，$i=1,2,\cdots,m$。

$$V\left(\boldsymbol{x}^*\right)=\boldsymbol{D}_1^{-1}V\left(\boldsymbol{x}\right)\boldsymbol{D}_1^{-1}=\boldsymbol{D}_1^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{D}_1^{-1}=\boldsymbol{R}_{11}$$

$$V\left(\boldsymbol{y}^*\right)=\boldsymbol{D}_2^{-1}V\left(\boldsymbol{y}\right)\boldsymbol{D}_2^{-1}=\boldsymbol{D}_2^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{D}_2^{-1}=\boldsymbol{R}_{22}$$

$$\mathrm{Cov}\left(\boldsymbol{x}^*,\boldsymbol{y}^*\right)=\boldsymbol{D}_1^{-1}\mathrm{Cov}\left(\boldsymbol{x},\boldsymbol{y}\right)\boldsymbol{D}_2^{-1}=\boldsymbol{D}_1^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{D}_2^{-1}=\boldsymbol{R}_{12}$$

$$\mathrm{Cov}\left(\boldsymbol{y}^*,\boldsymbol{x}^*\right)=\boldsymbol{D}_2^{-1}\mathrm{Cov}\left(\boldsymbol{y},\boldsymbol{x}\right)\boldsymbol{D}_1^{-1}=\boldsymbol{D}_2^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{D}_1^{-1}=\boldsymbol{R}_{21}$$

于是

$$\begin{aligned}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}&=\left(\boldsymbol{D}_1^{-1}\boldsymbol{\Sigma}_{11}\boldsymbol{D}_1^{-1}\right)^{-1}\boldsymbol{D}_1^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{D}_2^{-1}\left(\boldsymbol{D}_2^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{D}_2^{-1}\right)^{-1}\boldsymbol{D}_2^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{D}_1^{-1}\\&=\boldsymbol{D}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{D}_1^{-1}\end{aligned}$$

因为

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a}_i=\rho_i^2\boldsymbol{a}_i$$

$$\boldsymbol{D}_1\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{D}_1^{-1}\left(\boldsymbol{D}_1\boldsymbol{a}_i\right)=\rho_i^2\left(\boldsymbol{D}_1\boldsymbol{a}_i\right)$$

所以

$$\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{a}_i^* = \rho_i^2\boldsymbol{a}_i^*$$

式中$\boldsymbol{a}_i^* = \boldsymbol{D}_1\boldsymbol{a}_i$，有$\boldsymbol{a}_i^{*\prime}\boldsymbol{R}_{11}\boldsymbol{a}_i^* = \boldsymbol{a}_i'\boldsymbol{D}_1\boldsymbol{R}_{11}\boldsymbol{D}_1\boldsymbol{a}_i = \boldsymbol{a}_i'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_i = 1$。同理

$$\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{b}_i^* = \rho_i^2\boldsymbol{b}_i^*$$

式中$\boldsymbol{b}_i^* = \boldsymbol{D}_2\boldsymbol{b}_i$，有$\boldsymbol{b}_i^{*\prime}\boldsymbol{R}_{22}\boldsymbol{b}_i^* = \boldsymbol{b}_i'\boldsymbol{D}_2\boldsymbol{R}_{22}\boldsymbol{D}_2\boldsymbol{b}_i = \boldsymbol{b}_i'\boldsymbol{\Sigma}_{22}\boldsymbol{b}_i = 1$。由此可见，$\boldsymbol{a}_i^*, \boldsymbol{b}_i^*$为$\boldsymbol{x}^*$和$\boldsymbol{y}^*$的第$i$对典型系数，其第$i$个典型相关系数仍为$\rho_i$，在标准化变换下具有不变性，这一点与主成分分析有所不同。

- $\boldsymbol{x}^*$和$\boldsymbol{y}^*$的第i对典型变量$u_i^* = \boldsymbol{a}_i^{*\prime}\boldsymbol{x}^*, v_i^* = \boldsymbol{b}_i^{*\prime}\boldsymbol{y}^*$具有零均值，且与$\boldsymbol{x}$和$\boldsymbol{y}$的第$i$对典型变量$u_i=\boldsymbol{a}_i'\boldsymbol{x}$，$v_i=\boldsymbol{b}_i'\boldsymbol{y}$只相差一个常数。
- 例10.2.1　设$\boldsymbol{x},\boldsymbol{y}$有如下相关矩阵：

$$\boldsymbol{R}_{11} = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}, \quad \boldsymbol{R}_{22} = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}, \quad \boldsymbol{R}_{12} = \begin{pmatrix} \beta & \beta \\ \beta & \beta \end{pmatrix} = \beta\boldsymbol{1}\boldsymbol{1}'$$

这里$|\alpha| < 1, |\gamma| < 1$，可以保证$\boldsymbol{R}_{11}^{-1}, \boldsymbol{R}_{22}^{-1}$存在。

$$\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}=\frac{1}{1-\alpha^2}\begin{pmatrix}1 & -\alpha\\ -\alpha & 1\end{pmatrix}\beta\mathbf{1}\mathbf{1}'\bullet\frac{1}{1-\gamma^2}\begin{pmatrix}1 & -\gamma\\ -\gamma & 1\end{pmatrix}\beta\mathbf{1}\mathbf{1}'$$

$$=\frac{\beta^2}{(1-\alpha^2)(1-\gamma^2)}\begin{pmatrix}1-\alpha\\ 1-\alpha\end{pmatrix}(1-\gamma,1-\gamma)\mathbf{1}\mathbf{1}'$$

$$=\frac{\beta^2}{(1+\alpha)(1+\gamma)}\mathbf{1}\mathbf{1}'\mathbf{1}\mathbf{1}'=\frac{2\beta^2}{(1+\alpha)(1+\gamma)}\mathbf{1}\mathbf{1}'$$

由于$\mathbf{1}\mathbf{1}'$有惟一的非零特征值$\mathbf{1}'\mathbf{1}=2$，故$\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}$有惟一非零特征值

$$\rho_1^2=\frac{4\beta^2}{(1+\alpha)(1+\gamma)}$$

- 在约束条件$\boldsymbol{a}_1^{*\prime}\boldsymbol{R}_{11}\boldsymbol{a}_1^{*}=1$ 下，相应于特征值ρ_1^2的特征向量为 $\boldsymbol{a}_1^{*}=\left[2(1+\alpha)\right]^{-1/2}\mathbf{1}$ 。同理，在约束条件$\boldsymbol{b}_1^{*\prime}\boldsymbol{R}_{22}\boldsymbol{b}_1^{*}=1$ 下，

$\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}$相应于特征值$\rho_1^2$的特征向量为$\boldsymbol{b}_1^* = \left[2(1+\gamma)\right]^{-1/2}\mathbf{1}$。所以，第一对典型相关变量为

$$\boldsymbol{a}_1^{*\prime}\boldsymbol{x}^* = \left[2(1+\alpha)\right]^{-1/2}\mathbf{1}'\boldsymbol{x}^*, \boldsymbol{b}_1^{*\prime}\boldsymbol{y}^* = \left[2(1+\gamma)\right]^{-1/2}\mathbf{1}'\boldsymbol{y}^*$$

其中$\boldsymbol{x}^*$和$\boldsymbol{y}^*$分别是对$\boldsymbol{x}$和$\boldsymbol{y}$各分量标准化后的向量。第一个典型相关系数为$\rho_1 = 2|\beta|/\left[(1+\alpha)(1+\gamma)\right]^{1/2}$。由于$|\alpha|<1$，$|\gamma|<1$，故$\rho_1>|\beta|$，表明第一个典型相关系数大于两组原始变量之间的相关系数。

§10.3 样本典型相关

- 设数据矩阵为

$$(\boldsymbol{X} \quad \boldsymbol{Y}) = \begin{pmatrix} \boldsymbol{x}_1' & \boldsymbol{y}_1' \\ \vdots & \vdots \\ \boldsymbol{x}_n' & \boldsymbol{y}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} & y_{11} & \cdots & y_{1q} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} & y_{n1} & \cdots & y_{nq} \end{pmatrix}$$

则样本协方差矩阵为

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix}$$

$\boldsymbol{S}$可用来作为$\boldsymbol{\Sigma}$的估计。当$n>p+q$时，$\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}$和$\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}$可分别作为$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$和$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$的估计；它们的非零特征值$r_1^2 \geq r_2^2 \geq \cdots \geq r_m^2$可用来估计$\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_m^2$；

- 相应的特征向量 $\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_m$作为$a_1,a_2,\cdots,a_m$的估计， $\hat{b}_1,\hat{b}_2,\cdots,\hat{b}_m$作为$b_1,b_2,\cdots,b_m$的估计。$r_j^2$的正平方根$r_j$称为第$j$个样本典型相关系数，$\hat{a}_j'x$和$\hat{b}_j y$称为第$j$对样本典型相关变量，$j=1,2,\cdots,m$。将样本$(x_i,y_i)$，$i=1,2,\cdots,n$经中心化后代入$m$对典型变量，即令

$$u_{ij}=\hat{a}_j'(x_i-\bar{x}),\quad v_{ij}=\hat{b}_j'(y_i-\bar{y}),\quad i=1,2,\cdots,n,\ j=1,2,\cdots,m$$

则称u_{ij}为第i个样品x_i的第j个样本典型变量得分，称v_{ij}为第i个样品y_i的第j个样本典型变量得分。由约束条件 可得

$$\hat{a}_j'S_{11}\hat{a}_j=1$$

- 同理可得 $\dfrac{1}{n-1}\sum_{i=1}^{n}u_{ij}^2=\dfrac{1}{n-1}\hat{a}_j'\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'\hat{a}_j=\hat{a}_j'S_{11}\hat{a}_j=1,\quad j=1,2,\cdots,m$

$$\frac{1}{n-1}\sum_{i=1}^{n}v_{ij}^2=1,\quad j=1,2,\cdots,m$$

- 对每个j，可画出(u_{ij},v_{ij})，$i=1,2,\cdots,n$的散点图，该图也可用来检查是否有异常值出现。

例10.3.1 某康复俱乐部对20名中年人测量了三个生理指标：体重(x_1)、腰围(x_2)、脉搏(x_3)和三个训练指标：引体向上(y_1)、起坐次数(y_2)、跳跃次数(y_3)。其数据列于表10.3.1。

表10. 3. 1 某康复俱乐部的生理指标和训练指标数据

编 号	x_1	x_2	x_3	y_1	y_2	y_3
1	191	36	50	5	162	60
2	189	37	52	2	110	60
3	193	38	58	12	101	101
4	162	35	62	12	105	37
5	189	35	46	13	155	58
6	182	36	56	4	101	42
7	211	38	56	8	101	38

8	167	34	60	6	125	40
9	176	31	74	15	200	40
10	154	33	56	17	251	250
11	169	34	50	17	120	38
12	166	33	52	13	210	115
13	154	34	64	14	215	105
14	247	46	50	1	50	50
15	193	36	46	6	70	31
16	202	37	62	12	210	120
17	176	37	54	4	60	25
18	157	32	52	11	230	80
19	156	33	54	15	225	73
20	138	33	68	2	110	43

$$\hat{\boldsymbol{R}}_{11}=\begin{pmatrix}1 & & \\ 0.870 & 1 & \\ -0.366 & -0.353 & 1\end{pmatrix},\quad \hat{\boldsymbol{R}}_{22}=\begin{pmatrix}1 & & \\ 0.696 & 1 & \\ 0.496 & 0.669 & 1\end{pmatrix}$$

$$\hat{\boldsymbol{R}}_{12}=\hat{\boldsymbol{R}}'_{21}=\begin{pmatrix}-0.390 & -0.493 & -0.226 \\ -0.552 & -0.646 & -0.192 \\ 0.151 & 0.225 & 0.035\end{pmatrix}$$

$\hat{\boldsymbol{R}}_{11}^{-1}\hat{\boldsymbol{R}}_{12}\hat{\boldsymbol{R}}_{22}^{-1}\hat{\boldsymbol{R}}_{21}$的特征值分别为0.6630、0.0402和0.0053，于是

$$r_1=0.797,\ r_2=0.201,\ r_3=0.073$$

相应的样本典型变量系数为

$$\hat{\boldsymbol{a}}_1^* = \begin{pmatrix} -0.775 \\ 1.579 \\ -0.059 \end{pmatrix}, \quad \hat{\boldsymbol{a}}_2^* = \begin{pmatrix} -1.884 \\ 1.181 \\ -0.231 \end{pmatrix}, \quad \hat{\boldsymbol{a}}_3^* = \begin{pmatrix} -0.191 \\ 0.506 \\ 1.051 \end{pmatrix}$$

$$\hat{\boldsymbol{b}}_1^* = \begin{pmatrix} -0.350 \\ -1.054 \\ 0.716 \end{pmatrix}, \quad \hat{\boldsymbol{b}}_2^* = \begin{pmatrix} -0.376 \\ 0.124 \\ 1.062 \end{pmatrix}, \quad \hat{\boldsymbol{b}}_3^* = \begin{pmatrix} -1.297 \\ 1.237 \\ -0.419 \end{pmatrix}$$

因此，第一对样本典型变量为

$$u_1^* = -0.775x_1^* + 1.579x_2^* - 0.059x_3^*$$

$$v_1^* = -0.350y_1^* - 1.054y_2^* + 0.716y_3^*$$

- 如果需要，第二对样本典型变量为

$$u_2^* = -1.884x_1^* + 1.181x_2^* - 0.231x_3^*$$

$$v_2^* = -0.376y_1^* + 0.124y_2^* + 1.062y_3^*$$

例10.3.2 在研究组织结构对“职业满意度”的影响时，作为其中一部分，邓讷姆(Dunham)调查了职业满意度与职业特性相关的程度。对从一大型零售公司各分公司挑出的$n=784$个行政人员，测量了$p=5$个职业特性变量：用户反馈(x_1)、任务重要性(x_2)、任务多样性(x_3)、任务特性(x_4)及自主权(x_5)和$q=7$个职业满意度量：主管满意度(y_1)、事业前景满意度(y_2)、财政满意度(y_3)、工作强度满意度(y_4)、公司地位满意度(y_5)、工种满意度(y_6)及总体满意度(y_7)。对784个被测者的样本相关矩阵为

$$\hat{\boldsymbol{R}}_{11}=\begin{pmatrix} 1.00 & & & & \\ 0.49 & 1.00 & & & \\ 0.53 & 0.57 & 1.00 & & \\ 0.49 & 0.46 & 0.48 & 1.00 & \\ 0.51 & 0.53 & 0.57 & 0.57 & 1.00 \end{pmatrix}$$

$$\hat{\boldsymbol{R}}_{22}=\begin{pmatrix}1.00 & & & & & & \\ 0.43 & 1.00 & & & & & \\ 0.27 & 0.33 & 1.00 & & & & \\ 0.24 & 0.26 & 0.25 & 1.00 & & & \\ 0.34 & 0.54 & 0.46 & 0.28 & 1.00 & & \\ 0.37 & 0.32 & 0.29 & 0.30 & 0.35 & 1.00 & \\ 0.40 & 0.58 & 0.45 & 0.27 & 0.59 & 0.31 & 1.00\end{pmatrix}$$

$$\hat{\boldsymbol{R}}_{12}=\hat{\boldsymbol{R}}'_{21}=\begin{pmatrix}0.33 & 0.32 & 0.20 & 0.19 & 0.30 & 0.37 & 0.21 \\ 0.30 & 0.21 & 0.16 & 0.08 & 0.27 & 0.35 & 0.20 \\ 0.31 & 0.23 & 0.14 & 0.07 & 0.24 & 0.37 & 0.18 \\ 0.24 & 0.22 & 0.12 & 0.19 & 0.21 & 0.29 & 0.16 \\ 0.38 & 0.32 & 0.17 & 0.23 & 0.32 & 0.36 & 0.27\end{pmatrix}$$

样本典型相关系数和样本典型变量系数列于表10.3.2中。

表10.3.2 典型相关系数和典型变量系数

标准化变量	$\hat{a}_1^*$	$\hat{a}_2^*$	$\hat{a}_3^*$	$\hat{a}_4^*$	$\hat{a}_5^*$
x_1^*	0.42	0.34	−0.86	−0.79	0.03
x_2^*	0.20	−0.67	0.44	−0.27	0.98
x_3^*	0.17	−0.85	−0.26	0.47	−0.91
x_4^*	−0.02	0.36	−0.42	1.04	0.52
x_5^*	0.46	0.73	0.98	−0.17	−0.44
r_j	0.55	0.24	0.12	0.07	0.06
标准化变量	$\hat{b}_1^*$	$\hat{b}_2^*$	$\hat{b}_3^*$	$\hat{b}_4^*$	$\hat{b}_5^*$
y_1^*	0.43	−0.09	0.49	−0.13	−0.48
y_2^*	0.21	0.44	−0.78	−0.34	−0.75
y_3^*	−0.04	−0.09	−0.48	−0.61	0.35
y_4^*	0.02	0.93	−0.01	0.40	0.31
y_5^*	0.29	−0.10	0.28	−0.45	0.70
y_6^*	0.52	−0.55	−0.41	0.69	0.18
y_7^*	−0.11	−0.03	0.93	0.27	−0.01

第一对样本典型变量为

$$u_1^* = 0.42x_1^* + 0.20x_2^* + 0.17x_3^* - 0.02x_4^* + 0.46x_5^*$$

$$v_1^* = 0.43y_1^* + 0.21y_2^* - 0.04y_3^* + 0.02y_4^* + 0.29y_5^* + 0.52y_6^* - 0.11y_7^*$$

- 根据典型系数， 主要代表了用户反馈和自主权这两个变量，三个任务变量显得并不重要；而 主要代表了主管满意度和工种满意度变量，其次代表了事业前景满意度和公司地位满意度变量。我们也可从相关系数的角度来解释典型变量，原始变量与第一对典型变量间的样本相关系数列于表10.3.3中。

表10.3.3 原始变量与典型变量的样本相关系数

原始变量	样本典型变量		原始变量	样本典型变量	
$\boldsymbol{x}$	u_1^*	v_1^*	$\boldsymbol{y}$	u_1^*	v_1^*
x_1：用户反馈	0.83	0.46	y_1：主管满意度	0.42	0.76
x_2：任务重要性	0.73	0.40	y_2：事业前景满意度	0.36	0.64
x_3：任务多样性	0.75	0.42	y_3：财政满意度	0.21	0.39
x_4：任务特性	0.62	0.34	y_4：工作强度满意度	0.21	0.38
x_5：自主权	0.86	0.48	y_5：公司地位满意度	0.36	0.65
			y_6：工种满意度	0.45	0.80
			y_7：总体满意度	0.28	0.50

- 所有五个职业特性变量与第一典型变量u_1^*有大致相同的相关系数，故u_1^*可以解释为职业特性变量，这与基于典型系数的解释不同。v_1^*主要代表了主管满意度、事业前景满意度、公司地位满意度和工种满意度，v_1^*可以解释为职业满意度—公司地位变量，这与基于典型系数的解释基本相一致。第一对典型变量u_1^*与v_1^*的样本相关系数$r_1=0.55$，可见，职业特性与职业满意度之间有一定程度的相关性。

§10.4 典型相关系数的显著性检验

- 一、全部总体典型相关系数均为零的检验
- 二、部分总体典型相关系数为零的检验

一、全部总体典型相关系数均为零的检验

- 设$(\boldsymbol{x}',\boldsymbol{y}')' \sim N_{p+q}(\boldsymbol{\mu},\boldsymbol{\Sigma}),\boldsymbol{\Sigma}>0$。又设$\boldsymbol{S}$为样本协方差矩阵，且$n>p+q$。
- 考虑假设检验问题：

$$H_0: \rho_1=\rho_2=\cdots=\rho_m=0$$

$$H_1: \rho_1,\rho_2,\cdots,\rho_m\text{至少有一个不为零}$$

其中$m=\min\{p,q\}$。若检验接受H_0，则认为讨论两组变量之间的相关性没有意义；若检验拒绝H_0，则认为第一对典型变量是显著的。(10.4.1)式实际上等价于假设检验问题

$$H_0: \boldsymbol{\Sigma}_{12}=\boldsymbol{0},\quad H_1: \boldsymbol{\Sigma}_{12}\neq\boldsymbol{0}$$

H_0成立表明$\boldsymbol{x}$与$\boldsymbol{y}$互不相关。

检验统计量为

$$\Lambda_1 = \prod_{i=1}^{m} \left(1 - r_i^2\right)$$

对于充分大的n，当H_0成立时，统计量

$$Q_1 = -\left[n - \frac{1}{2}(p+q+3)\right] \ln \Lambda_1 \dot{\sim} \chi^2(pq)$$

在给定的α下，若$Q_1 \geq \chi_\alpha^2(pq)$，则拒绝$H_0$，认为典型变量$u_1$与$v_1$之间的相关性是显著的；否则，就认为第一个典型相关系数不显著。

例10.4.1 在例10.3.1中，假设为多元正态数据，欲检验：

$$H_0: \rho_1=\rho_2=\rho_3=0, \quad H_1: \rho_1\neq 0$$

它的似然比统计量为

$$\Lambda_1 = (1-r_1^2)(1-r_2^2)(1-r_3^2)$$

$$= (1-0.6330)(1-0.0402)(1-0.0053) = 0.3504$$

$$Q_1 = -\left[20-\frac{1}{2}(3+3+3)\right]\ln\Lambda_1 = -15.5\times\ln 0.3504 = 16.255$$

查χ^2分布表得，$\chi^2_{0.10}(9)=14.684$，$\chi^2_{0.05}(9)=16.919$，因此在α=0.10的显著性水平下，拒绝原假设H_0，也即认为至少有一个典型相关是显著的。

二、部分总体典型相关系数为零的检验

- 若H_0：$\rho_1=\rho_2=\cdots=\rho_m=0$经检验被拒绝，则应进一步检验假设

$$H_0：\rho_2=\cdots=\rho_m=0$$

$$H_1：\rho_2,\cdots,\rho_m\text{至少有一个不为零}$$

若原假设H_0被接受，则认为只有第一对典型变量是有用的；
若原假设H_0被拒绝，则认为第二对典型变量也是有用的。

- 如此进行下去，直至对某个k，假设H_0：$\rho_{k+1}=\cdots=\rho_m=0$被接受，这时可认为只有前$k$对典型变量是显著的。
- 对于假设检验问题

$$H_0：\rho_{k+1}=\cdots=\rho_m=0$$

$$H_1：\rho_{k+1},\cdots,\rho_m\text{至少有一个不为零}$$

其检验统计量为

$$\Lambda_{k+1}=\prod_{i=k+1}^{m}\left(1-r_i^2\right)$$

对于充分大的n，当H_0为真时，统计量

$$Q_{k+1}=-\left[n-k-\frac{1}{2}(p+q+3)+\sum_{i=1}^{k}r_i^{-2}\right]\ln\Lambda_{k+1}$$

近似服从自由度为$(p-k)(q-k)$的χ^2分布。给定显著性水平α，若$Q_{k+1}\geq\chi_\alpha^2\left[(p-k)(q-k)\right]$，则拒绝原假设$H_0$，认为第$k+1$个典型相关系数$\rho_{k+1}$是显著的，即第$k+1$对典型变量显著相关。

- 以上的一系列检验实际上是一个序贯检验，检验直到对某个k值H_0未被拒绝为止。事实上，检验的总显著性水平已不是α了，且难以确定。还有，检验的结果易受样本容量大小的影响。因此，检验的结果只宜作为确定典型变量个数的重要参考依据，而不宜作为惟一的依据。通常选择尽可能小的k。

· 例10.4.2 在例10.3.1中，欲进一步检验：

$$H_0: \rho_2=\rho_3=0, \quad H_1: \rho_2\neq 0$$

检验统计量为

$$\Lambda_2 = (1-r_2^2)(1-r_3^2) = (1-0.0402)(1-0.0053) = 0.9547$$

$$Q_2 = -\left[20-1-\frac{1}{2}(3+3+3)+r_1^{-2}\right]\ln\Lambda_2$$

$$= -16.08\times \ln 0.9547 = 0.745 < 7.779 = \chi^2_{0.10}(4)$$

故接受原假设H_0，即认为第二个典型相关是不显著的。因此，只有一个典型相关是显著的。