典型相关分析

TUTU

典型相关分析的基本概念

♣ 基本概念

研究两组变量之间相关关系的一种多元统计方法。它能够揭示两组变量之间的内在联系。

♣ 目的

识别并量化两组变量之间的联系，将两组变量相关关系的分析，转化为一组变量的线性组合与另一组变量线性组合之间的相关关系分析。

♣ 应用

被应用于心理学、市场营销等领域。如用于研究个人性格与职业兴趣的关系，市场促销活动与消费者响应之间的关系等问题的分析研究。

典型相关分析与相关分析的异同点

♣ 典型相关分析与相关分析的异同点

- 典型相关分析：典型相关分析是对协方差矩阵的一种理解，是利用综合变量对之间的相关关系来反映两组指标之间的整体相关性的多元统计分析方法。
- 相关分析：相关分析是研究两个或两个以上处于同等地位的随机变量间的相关关系的统计分析方法。
- 联系：典型相关分析与相关分析都是分析变量之间相关性的分析方法，都是线性分析的范畴。
- 区别：简单相关系数描述两组变量的相关关系的缺点：只是孤立考虑单个 X 与单个 Y 间的相关，没有考虑 X、Y 变量组内部各变量间的相关。两组间有许多简单相关系数，使问题显得复杂，难以从整体描述。典型相关是简单相关、多重相关的推广。典型相关是研究两组变量之间相关性的一种统计分析方法。也是一种降维技术。

典型相关分析的基本思想

♣ 基本思想

首先在每组变量中找出变量的线性组合，使其具有最大相关性，然后再在每组变量中找出第二对线性组合，使其分别与第一对线性组合不相关，而第二对本身具有最大的相关性，如此继续下去，直到两组变量之间的相关性被提取完毕为止。有了这样线性组合的最大相关，则讨论两组变量之间的相关，就转化为只研究这些线性组合的最大相关，从而减少研究变量的个数。

典型相关分析的步骤

♣ 步骤

1. 首先分别在每组变量中找出第一对线性组合，使其具有最大相关性。

$$\begin{cases} u_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\ v_1 = b_{11}y_1 + b_{21}y_2 + \cdots + b_{q1}x_q \end{cases}$$

2. 然后再在每组变量中找出第二对线性组合，使其分别与本组内的第一线性组合不相关，第二对本身具有次大的相关性。

$$\begin{cases} u_2 = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\ v_2 = b_{12}y_1 + b_{22}y_2 + \cdots + b_{q2}x_q \end{cases}$$

3. u_2 和 v_2 与 u_1 和 v_1 相互独立，但 u_2 和 v_2 相关。如此继续下去，直至进行到 r 步，两组变量的相关性被提取完为止。$r \leq \min(p, q)$，可以得到 r 组变量。

典型相关的数学描述

♣ 考虑 $\boldsymbol{Z} = (x_1, x_2, \cdots, x_p, y_1, y_2, \cdots, y_q)$，其协方差矩阵为 $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$，记 $\boldsymbol{u}_1 = \boldsymbol{a}_1'\boldsymbol{X}, \boldsymbol{v}_1 = \boldsymbol{b}_1'\boldsymbol{Y}$，则

- $\mathrm{Var}(\boldsymbol{u}_1) = \boldsymbol{a}_1'\mathrm{Var}(\boldsymbol{X})\boldsymbol{a}_1 = \boldsymbol{a}_1'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_1 = 1$
- $\mathrm{Var}(\boldsymbol{v}_1) = \boldsymbol{b}_1'\mathrm{Var}(\boldsymbol{Y})\boldsymbol{b}_1 = \boldsymbol{b}_1'\boldsymbol{\Sigma}_{22}\boldsymbol{b}_1 = 1$
- $\rho_{\boldsymbol{u}_1,\boldsymbol{v}_1} = \boldsymbol{a}_1'\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{b}_1 = \boldsymbol{a}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{b}_1$
- 所以，典型相关分析就是求 $\boldsymbol{a}_1$ 和 $\boldsymbol{b}_1$，使$\rho_{u,v}$ 达到最大。

♣ 借助 Lagrange 乘数法，记 $\boldsymbol{M}_1 = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$，$\boldsymbol{M}_2 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$，将典型相关分析转化为求 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 特征根和特征向量的问题，特征向量构成典型变量的系数，特征根为典型变量相关系数的平方。

典型相关的数学描述

♣ 第一对典型变量提取了原始变量 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 之间相关的主要部分，如果这部分还不能足以解释原始变量，可以在剩余的相关中再求出第二对典型变量和他们的典型相关系数。

令 $\boldsymbol{u}_2 = \boldsymbol{a}_2'\boldsymbol{X}, \boldsymbol{v}_2 = \boldsymbol{b}_2'\boldsymbol{Y}$，则在约束条件：

- $\mathrm{Var}(\boldsymbol{u}_2) = \boldsymbol{a}_2'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_2 = 1$
- $\mathrm{Var}(\boldsymbol{v}_2) = \boldsymbol{b}_2'\boldsymbol{\Sigma}_{22}\boldsymbol{b}_2 = 1$
- $\mathrm{Cov}(\boldsymbol{u}_1, \boldsymbol{u}_2) = \boldsymbol{a}_1'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_2 = 0$
- $\mathrm{Cov}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \boldsymbol{b}_1'\boldsymbol{\Sigma}_{22}\boldsymbol{b}_2 = 0$

求使 $\mathrm{Cov}(\boldsymbol{u}_2, \boldsymbol{v}_2) = \boldsymbol{a}_2'\boldsymbol{\Sigma}_{12}\boldsymbol{b}_2$ 达到最大的 $\boldsymbol{a}_2$ 和 $\boldsymbol{b}_2$。

典型变量的性质

♣ 典型变量的性质

- 同一组变量的典型变量之间互不相关：$\mathrm{Cov}(\boldsymbol{u}_k, \boldsymbol{u}_l) = \boldsymbol{a}_k' \boldsymbol{\Sigma}_{11} \boldsymbol{a}_l = 0$，$\mathrm{Cov}(\boldsymbol{v}_k, \boldsymbol{v}_l) = \boldsymbol{b}_k' \boldsymbol{\Sigma}_{22} \boldsymbol{b}_l = 0$
- 不同组变量的典型变量之间的相关性：$\mathrm{Cov}(\boldsymbol{u}_i, \boldsymbol{v}_j) = \begin{cases} \lambda_i, & i = j \\ 0, & i \neq j \end{cases}$
- 各组原始变量被典型变量所解释的方差 (典型冗余分析)：

  $\boldsymbol{X}$ 组原始变量被 $\boldsymbol{u}_i$ 解释的方差比例：$m_{\boldsymbol{u}_i} = \sum_{k=1}^{p} \rho^2_{\boldsymbol{u}_i, \boldsymbol{x}_k} / p$

  $\boldsymbol{X}$ 组原始变量被 $\boldsymbol{v}_i$ 解释的方差比例：$m_{\boldsymbol{v}_i} = \sum_{k=1}^{p} \rho^2_{\boldsymbol{v}_i, \boldsymbol{x}_k} / p$

  $\boldsymbol{Y}$ 组原始变量被 $\boldsymbol{u}_i$ 解释的方差比例：$n_{\boldsymbol{u}_i} = \sum_{k=1}^{q} \rho^2_{\boldsymbol{u}_i, \boldsymbol{y}_k} / q$

  $\boldsymbol{Y}$ 组原始变量被 $\boldsymbol{v}_i$ 解释的方差比例：$n_{\boldsymbol{v}_i} = \sum_{k=1}^{q} \rho^2_{\boldsymbol{v}_i, \boldsymbol{y}_k} / q$

典型冗余分析的内容与作用

♣ 典型冗余分析的内容与作用

- 内容：冗余分析是通过原始变量与典型变量之间的相关性。分析引起原始变量变异的原因。以原始变量为因变量，以典型变量为自变量，建立线性回归模型，则相应的判定系数 R^2 等于因变量与典型变量间的相关系数的平方，它描述了由于因变量与典型变量的线性关系引起的因变量变异在因变量的总变异中比例。
- 作用：分析每组变量提取出的典型变量所能解释的该组样本总方差的比例，从而定量测度典型变量所包含的原始信息量。

典型变量的性质

♣ 典型变量的性质

- 原始变量与典型变量之间的相关系数 (典型载荷分析):

  $\boldsymbol{X}$ 典型变量系数矩阵 $\boldsymbol{A} = (a_{ij})_{p\times r}$,

  $\boldsymbol{Y}$ 典型变量系数矩阵 $\boldsymbol{B} = (b_{ij})_{q\times r}$:

  - $\mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{u}_j) = \sum_{k=1}^{p} a_{kj}\sigma_{\boldsymbol{x}_i,\boldsymbol{x}_k}$, $\rho(\boldsymbol{x}_i, \boldsymbol{u}_j) = \sum_{k=1}^{p} a_{kj}\sigma_{\boldsymbol{x}_i,\boldsymbol{x}_k}/\sqrt{\sigma_{\boldsymbol{x}_i,\boldsymbol{x}_i}}$
  - $\mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{v}_j) = \sum_{k=1}^{q} b_{kj}\sigma_{\boldsymbol{x}_i,\boldsymbol{y}_k}$, $\rho(\boldsymbol{x}_i, \boldsymbol{v}_j) = \sum_{k=1}^{q} b_{kj}\sigma_{\boldsymbol{x}_i,\boldsymbol{y}_k}/\sqrt{\sigma_{\boldsymbol{x}_i,\boldsymbol{x}_i}}$
  - $\mathrm{Cov}(\boldsymbol{y}_i, \boldsymbol{u}_j) = \sum_{k=1}^{p} a_{kj}\sigma_{\boldsymbol{y}_i,\boldsymbol{x}_k}$, $\rho(\boldsymbol{y}_i, \boldsymbol{u}_j) = \sum_{k=1}^{p} a_{kj}\sigma_{\boldsymbol{y}_i,\boldsymbol{x}_k}/\sqrt{\sigma_{\boldsymbol{y}_i,\boldsymbol{y}_i}}$
  - $\mathrm{Cov}(\boldsymbol{y}_i, \boldsymbol{v}_j) = \sum_{k=1}^{q} b_{kj}\sigma_{\boldsymbol{y}_i,\boldsymbol{y}_k}$, $\rho(\boldsymbol{y}_i, \boldsymbol{v}_j) = \sum_{k=1}^{q} b_{kj}\sigma_{\boldsymbol{y}_i,\boldsymbol{y}_k}/\sqrt{\sigma_{\boldsymbol{y}_i,\boldsymbol{y}_i}}$

♣ 在实际中，可以使用样本的协方差或相关系数矩阵进行分析。

典型相关系数的显著性检验

♣ Bartlett 检验

- $H_0 : \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_r = 0$，$H_1 : \lambda_{k+1} \neq 0$
- 似然比统计量：$\Lambda_k = \prod_{i=k+1}^{r} (1 - \hat{\lambda}_i^2)$
- $Q_m = -m_k \ln \Lambda_k \to \chi^2(f_k)$，其中 $f_k = (p-k)(q-k)$，
  $m_k = (n-k-1) - \frac{1}{2}(p+q+1)$
- 从 $k = 0$ 时开始，若拒绝原假设，则认为 $\lambda_1 > 0$，继续 $k = 1$，直至 $k = j$ 时，不拒绝原假设，则 $\lambda_j = \lambda_{j+1} = \cdots = \lambda_r = 0$，提取 $j-1$ 对典型变量进行分析

从相关矩阵出发计算典型相关

♣ 从相关矩阵出发计算典型相关

- 为消除量纲影响，对数据先做标准化变换，然后再做典型相关分析。
- 显然，经标准化变换之后的协差阵就是相关系数矩阵，因而，也即通常应从相关矩阵出发进行典型相关分析。

典型相关分析的 SAS 代码

♣ SAS 代码:

```
/*with后是第二组变量，var后是第一组变量*/
/*vdep以var为因变量，with为自变量，进行多元回归分析*/
/*vreg以with为因变量，var为自变量，进行多元回归分析*/
proc cancorr data=yourdata out=out1 outstat=outstat1 all
    vdep / vreg;
with y1-y3;
var x1-x3;
run;
```