第11章 典型相关分析及R使用

武慧
wuh@hit.edu.cn

经济管理学院
哈尔滨工业大学（威海）

1

假设你是校园健康社团的成员，想要调查大学生的熬夜习惯（比如刷手机时长、入睡时间）与日常生活质量（比如课堂注意力、运动积极性、情绪波动）之间的关系，该怎么做？

- 第一组变量（X）：“夜猫子行为”（刷短视频时长、夜间奶茶摄入量、入睡时间）。
- 第二组变量（Y）：“校园战斗力”（早晨迟到次数、下午犯困频率、朋友圈emo文案发帖量）。

单独分析“刷视频时长 vs. 迟到次数”（r=0.6）、“入睡时间 vs. emo发帖量”（r=0.5），但无法揭示变量组的复杂关联。

如果长熬夜组（刷视频3h+奶茶2杯+入睡2AM）的学生，同时具有‘迟到率高+犯困多+emo多’，能否找到两组变量背后的隐藏关联规律？

就像游戏中，英雄的多个技能组合会形成连招——典型相关分析(CCA)的核心正是寻找变量组的“连招效应”！

- 将“夜猫子行为”3个变量合成为1个“熬夜指数”。
- 将“校园战斗力”3个变量合成为1个“日常生活质量指数”。
- 分析这两个指数的相关系数（如0.8），并解释哪些变量对“熬夜指数”贡献最大。

3

本章内容

1. 典型相关分析的基本概念
2. 典型相关分析理论
3. 典型相关分析在R中的实现

4

1. 典型相关分析的基本概念

5

什么是典型相关分析？

- 典型相关分析（Canonical Correlation Analysis，简称CCA）是一种用于研究两个多变量数据集之间的线性关系的统计方法。。
- 具体地，CCA通过寻找每个变量集的线性组合（即典型变量），来最大化两个变量集之间的相关性。
- 通过这两个线性组合，发现两个变量集之间的关系，进而分析它们的相互影响。

6

单选题 1分

典型相关分析的主要目的是

A 研究两个多变量数据集之间关系

B 将数据集降维

C 评估主成分之间的关系

D 确定因变量和自变量的关系

7

单选题 1分

典型相关分析的一个应用场景是:

A 在时间序列数据中寻找周期性变化

B 在两个不同学科之间找出共同规律

C 在经济学中分析消费与收入之间的非线性关系

D 在心理学中分析单一变量与实验条件的因果关系

2. 典型相关分析理论

9

典型相关分析数学原理

- 假设有两个（中心化的）随机向量：

$\mathbf{X} \in \mathbb{R}^{n\times p}$ 表示第一个数据集，有 p 个变量，样本数量为 n。

$\mathbf{Y} \in \mathbb{R}^{n\times q}$ 表示第二个数据集，有 q 个变量，样本数量为 n。

希望找到这两个数据集之间的线性关系。

- 为了实现这一目标，首先构建每个数据集的**典型变量**：

$\mathbf{u} = \mathbf{Xa}$ 为 $\mathbf{X}$ 的线性组合

$\mathbf{v} = \mathbf{Yb}$ 为 $\mathbf{Y}$ 的线性组合

- 然后，**通过调整权重向量$\boldsymbol{a}$和$\boldsymbol{b}$，使得这两个典型变量$\boldsymbol{u}$和$\boldsymbol{v}$之间的相关性最大化。**

10

典型相关分析数学原理

- 目标是最大化：

$$\rho = \frac{\mathrm{cov}(\mathbf{u}, \mathbf{v})}{\sqrt{\mathrm{var}(\mathbf{u}) \cdot \mathrm{var}(\mathbf{v})}}$$

$$\mathbf{u} = \mathbf{Xa}$$

$$\mathbf{v} = \mathbf{Yb}$$

其中，ρ被称为**典型相关系数**。

- 通过选择适当的$\boldsymbol{a}$和$\boldsymbol{b}$，找到最大化ρ的解。

11

典型相关分析数学原理

$$\mathbf{u} = \mathbf{X}\mathbf{a}$$

$$\mathbf{v} = \mathbf{Y}\mathbf{b}$$

$$\mathrm{cov}(\mathbf{u}, \mathbf{v}) = \mathbf{a}^T\mathbf{X}^T\mathbf{Y}\mathbf{b}$$

$$\mathrm{var}(\mathbf{u}) = \mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a}$$

$$\mathrm{var}(\mathbf{v}) = \mathbf{b}^T\mathbf{Y}^T\mathbf{Y}\mathbf{b}$$

$$\rho = \frac{\mathrm{cov}(\mathbf{u}, \mathbf{v})}{\sqrt{\mathrm{var}(\mathbf{u}) \cdot \mathrm{var}(\mathbf{v})}}$$

$$\rho = \frac{\mathbf{a}^T\mathbf{X}^T\mathbf{Y}\mathbf{b}}{\sqrt{\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} \cdot \mathbf{b}^T\mathbf{Y}^T\mathbf{Y}\mathbf{b}}}$$

- 为了解的唯一性，通常要求对a和b进行规范化：

$$\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} = 1$$

$$\mathbf{b}^T\mathbf{Y}^T\mathbf{Y}\mathbf{b} = 1$$

12

典型相关分析数学原理

- 为了最大化目标函数ρ并同时满足上述规范化条件，我们可以使用拉格朗日乘子法来求解最优化问题。
- 构造拉格朗日函数：

$$L(\mathbf{a},\mathbf{b},\lambda_1,\lambda_2) = \mathbf{a}^T\mathbf{X}^T\mathbf{Y}\mathbf{b} - \lambda_1(\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} - 1) - \lambda_2(\mathbf{b}^T\mathbf{Y}^T\mathbf{Y}\mathbf{b} - 1)$$

其中，λ_1和λ_2是拉格朗日乘子，表示规范化条件的约束。

- 对$\boldsymbol{a}$和$\boldsymbol{b}$分别求偏导数并令偏导数为零，得到两个方程：

$$\mathbf{X}^T\mathbf{Y}\mathbf{b} = \lambda_1\mathbf{X}^T\mathbf{X}\mathbf{a} \qquad \mathbf{Y}^T\mathbf{X}\mathbf{a} = \lambda_2\mathbf{Y}^T\mathbf{Y}\mathbf{b}$$

13

典型相关分析数学原理

$$\mathbf{X}^T\mathbf{Y}\mathbf{b} = \lambda_1\mathbf{X}^T\mathbf{X}\mathbf{a}$$

$$\mathbf{Y}^T\mathbf{X}\mathbf{a} = \lambda_2\mathbf{Y}^T\mathbf{Y}\mathbf{b}$$

$$\boldsymbol{A}\boldsymbol{a} = \rho^2\boldsymbol{a}$$

$$\boldsymbol{B}\boldsymbol{b} = \rho^2\boldsymbol{b}$$

$$\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{Y})(\boldsymbol{Y}^T\boldsymbol{Y})^{-1}(\boldsymbol{Y}^T\boldsymbol{X})$$

$$\boldsymbol{B} = (\boldsymbol{Y}^T\boldsymbol{Y})^{-1}(\boldsymbol{Y}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{Y})$$

- 通过解这个广义特征值问题，我们可以得到典型相关系数（特征值的平方根），以及每个典型变量的系数（对应的特征向量）。

14

单选题 1分

典型相关分析中，典型变量的线性组合是根据什么原则来构造的?

A 使得两个变量集的协方差最小化

B 使得两个变量集之间的相关性最大化

C 使得每个变量集的方差最大化

D 使得每个变量集的均值为零

15

单选题 1分

在典型相关分析中，两个变量集之间的典型相关系数越大，意味着：

A 两个变量集的线性相关性越强

B 两个变量集没有任何关系

C 其中一个变量集的方差较大

D 两个变量集没有共同的主成分

16

单选题 1分

在典型相关分析中，第一对典型变量的相关系数表示：

A 第一个变量集与第二个变量集之间最强的线性相关性

B 两个变量集的均值差异

C 第一个变量集中的最大方差

D 第二个变量集的最小方差

17

单选题 1分

典型相关分析中，第二对典型变量的相关系数与第一对的相关系数有什么关系?

A 第二对典型变量的相关系数一定小于第一对

B 第二对典型变量的相关系数大于第一对

C 第二对典型变量的相关系数与第一对相等

D 第二对典型变量的相关系数无任何固定关系

18

3. 典型相关分析在R中的实现

19

案例

- 分析环境污染与居民健康状况之间的关系。环境污染可能会影响居民的健康，因此了解两者之间的相关性对政策制定和公共卫生工作具有重要意义。

变量集1：环境污染数据（X）

- ➢空气污染指数（PM2.5）
- ➢水污染指数（COD）
- ➢土壤污染指数（Heavy Metals）

变量集2：健康数据（Y）

- ➢居民寿命期望（Life Expectancy）
- ➢患病率（Disease Incidence Rate）
- ➢心理健康评分（Mental Health Index）

20

```
> head(X)
   PM25   COD Heavy_Metals
1 78.59 45.46        30.07
2 80.89 47.94        30.47
3 91.57 70.69        33.74
4 86.53 64.05        30.58
5 69.46 50.76        31.24
6 77.76 45.36        29.09
```

```
> head(Y)
  Life_Expectancy Disease_Rate Mental_Health
1           77.83         0.14         46.83
2           76.58         0.15         53.56
3           82.97         0.22         55.07
4           78.07         0.14         55.41
5           71.90         0.12         45.79
6           74.26         0.08         48.82
```

执行典型相关分析

```
# 执行典型相关分析
cca_result <- cancor(X, Y)
```

```
> # 查看典型相关系数
> cca_result$cor
[1] 0.92646056 0.13758841 0.05408389
```

- 典型相关系数的值介于 0 和 1 之间。值越接近 1，表示两个变量集之间的相关性越强，值越接近 0，表示相关性越弱。
- 例如，第一对典型变量之间的典型相关系数为 0.9265，这表示环境污染和健康状况之间有较强的关联。第二对和第三对典型变量的相关性较弱，分别为 0.1376 和 0.0541。

22

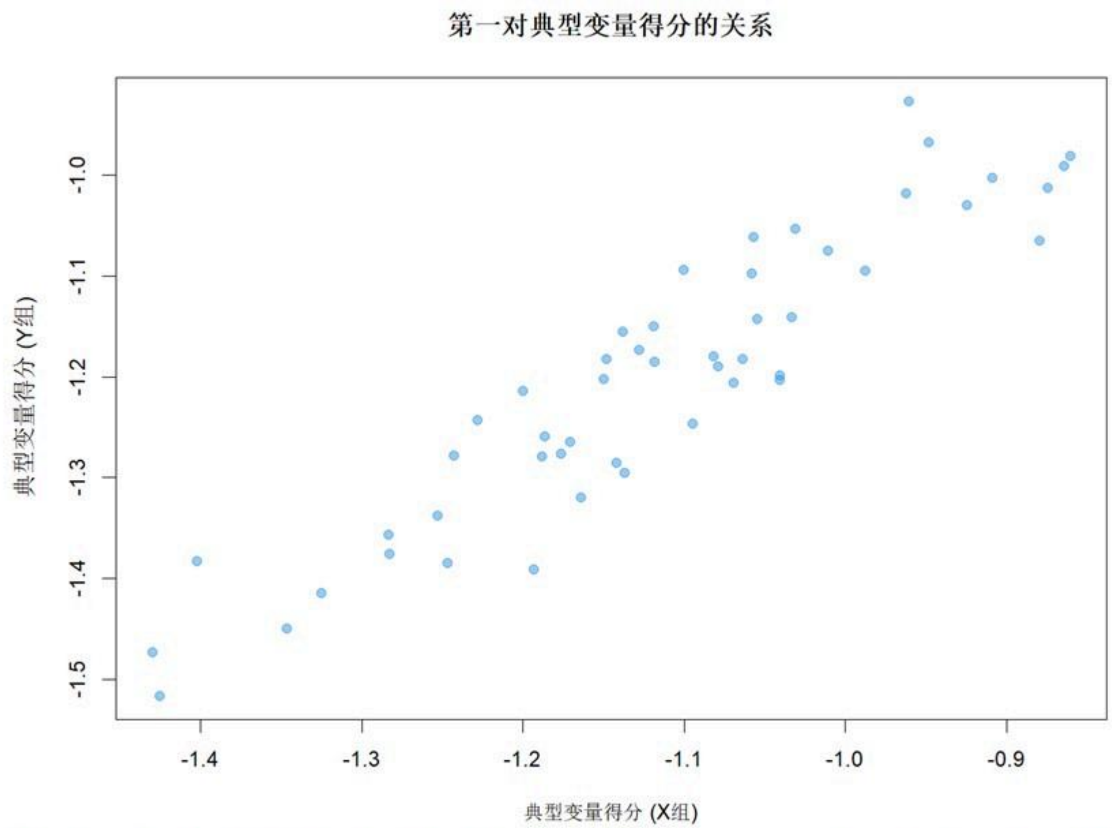
```
> # 查看X集的典型变量系数
> cca_result$xcoef
                    [,1]         [,2]        [,3]
PM25         -0.004007183  0.009868911  0.01410954
COD          -0.002851339  0.015211333 -0.02450344
Heavy_Metals -0.020291868 -0.062962811  0.01350960

> # 查看Y集的典型变量系数
> cca_result$ycoef
                      [,1]        [,2]         [,3]
Life_Expectancy -0.008475010  0.03900780  0.046492026
Disease_Rate    -1.203941079  0.45769021 -5.344812674
Mental_Health   -0.006714409 -0.02132995  0.008068132
```

典型相关分析的结果可以为后续的回归分析提供线索。请问哪种环境污染对健康影响最大？

- 典型变量系数反映了每个原始变量对典型变量的贡献，这些系数有助于我们理解每个变量在典型变量中的相对重要性。

计算和绘制典型变量得分

```
# 计算X组的典型变量得分
T_X <- as.matrix(X) %*% cca_result$xcoef
# 计算Y组的典型变量得分
T_Y <- as.matrix(Y) %*% cca_result$ycoef
# 绘制第一对典型变量的得分散点图
plot(T_X[, 1], T_Y[, 1],
     xlab = "典型变量得分 (X组)",
     ylab = "典型变量得分 (Y组)",
     main = "第一对典型变量得分的关系",
     pch = 19, col = rgb(0.2, 0.6, 0.9, 0.5))
```



- 通过绘制典型变量得分的散点图，可以直观地看到样本在这两个维度上的分布情况，进而可以发现不同样本群体的特征差异。

单选题 1分

在R中，进行典型相关分析时使用的函数是：

A cor()

B lm()

C cancor()

D prcomp()

单选题 1分

使用 cancor() 函数时，输入的两个数据集必须满足的条件是：

A 两个数据集必须包含相同数量的观察值

B 两个数据集必须包含相同数量的变量

C 两个数据集必须独立

D 两个数据集必须有相同的列名

27

单选题 1分

cancor() 输出结果中的 "xcoef" 和 "ycoef" 含义是:

A x 和 y 数据集中的每个变量系数

B x 和 y 数据集的相关性

C x 和 y 数据集的典型变量系数

D x 和 y 数据集的得分

28

单选题 1分

使用 cancor() 进行典型相关分析时，典型变量系数（canonical coefficients）表示：

A 典型变量的标准差

B 每个变量集的均值大小

C 每个变量对典型变量的贡献程度

D 典型相关系数的平方

单选题 1分

典型变量得分（canonical scores）主要反映：

A 每个变量对数据集整体的贡献

B 每个观察值在典型变量上的投影值

C 两个数据集的协方差

D 每个变量集的标准差

30

本章小结

- 典型相关分析的基本概念：定义、应用、与主成分分析的区别与联系。
- 典型相关分析理论：典型相关系数、典型变量的求解方式。
- 典型相关分析在R中的实现：典型相关系数、典型变量系数、典型变量得分的解读。（cancor）

31