目录

起玄🄵摄提格尽强圉协洽

1 Matlab 基础入门

1.1 基本小常识

1. 每一行的语句后面加上分号; 后表示不显示运行结果
2. 注释多行注释快捷键 Ctrl+R 取消注释 Ctrl+T
3. 初始化工作区和命令窗口 clear;clc
4. 输出函数-disp()
5. 行向量 a=[1,2,3] 列向量 b=[1;2;3] 分号可用于分隔行
6. 输入函数-input()
7. M(i,:)-第 i 行所有元素; M(:,j)-第 j 列所有元素;
   M([a,b],:)-只取第 a 行和第 b 行 (共 2 行元素);
   M(a:b,:)-第 a 行到第 b 行 (共 b-a 行元素);
   M(a:i:b,:)-从第 a 行开始, 每次递增 i 个单位, 到第 b 行结束;

1.2 常用函数分类

1.2.1 连接字符串函数

- strcat(str1,str2, $\cdots$,strn) == [str1,str2, $\cdots$,strn]

1.2.2 数字转换字符串函数

- num2str(): 将' 数字' 转换为' 字符串'

1.2.3 sum() 函数

- 对矩阵求和: sum(M,1)-矩阵列求和;
  sum(M,2)-矩阵行求和; sum(M(:))-矩阵所有元素求和;

1.2.4 size() 函数

- $[row, column] = size(M)$;

  r=size(M;1)-only retuen row; c=size(M;2)-only return column;

1.2.5 repmat() 函数

- B=repmat(A,m,n): 将 B 矩阵以 A 矩阵为元素按照 m×n 块复制

1.2.6 find() 函数

- 返回向量或者矩阵中不为 0 的元素的位置的索引.

  dim A=1: ind=find(A,n)-返回前 n 个不为 0 的元素的位置;

  dim A>1: ind=find(A)-先将 A 矩阵按照列存储, 再找出所有不为 0 的元素的位置;

  dim A>1: [r,c]=find(A)-给出非零元素的坐标.

1.3 矩阵运算

1.3.1 矩阵自身的运算

- A÷B == A×inv(B) inv(B)=B^{-1};
- 两个形状相同的矩阵对应元素之间的乘除法需要使用 $\cdot*$ 和 $\cdot\div$;
- 每个元素同时乘方时用 $\cdot\wedge$;

1.3.2 矩阵与常数大小的判断运算

- 共有三种运算符: > < ==
- 返回的是 logic Matrix
- 判断语句格式:if else if else end(无: 和;)

1.4 特征值与特征向量的求解

- Matlab 中求解特征值与特征向量:

  1.E=eig(A): 求出矩阵 A 的全部特征值, 构成向量 E.

  2.[V,D]=eig(A): 求出矩阵 A 的特征值, 构成对角阵 D; 并且求出与特征值对应的特征向量构成的矩阵 V.

2 层次分析法

2.1 权重表格

- **判断矩阵**:

  1. a(i,j) 表示与 j 指标相比,i 坐标的重要程度
  2. i=j means i 与 j 同等重要, 主对角元素为 1
  3. if a(i,j)>0 and a(i,j) × a(j,i)=1 ,then A(i,j) is called 正互反矩阵

- **一致矩阵**:

  充要条件:

  1. a(i,j)>0
  2. a(i,i)=1 (i$\in [1,n]$)
  3. $[a_{i1}, a_{i2}, \cdots, a_{in}] = k_i[a_{11}, a_{12}, \cdots, a_{1n}]$

  lemma:

  1. A 为 n 阶方阵,rank(A)=1, 则 A 有一个特征值 tr(A), 其余特征值为 0
  2. 一致矩阵有一个特征值为 n, 其余特征值为 0

- **一致性检验**:

  步骤

  1. 计算一致性指标 **CI**
     CI $= \frac{\lambda_{\max}-n}{n-1}$
  2. 查找对应的平均随机一致性指标 **RI**
     在表格中根据 n 查找到对应的 **RI**
     (RI 通过随机方法构造样本矩阵, 定义得到)
  3. 计算一致性比例 **CR**
     CR $= \frac{\mathbf{CI}}{\mathbf{RI}}$
  4. if **CR** < 0.1, 则认为矩阵的一致性可以接受
     否则需要对判断矩阵进行修正.

2.2 计算权重

while 判断矩阵的一致性可接受

2.2.1 方法一: 算术平均法求权重

1. 将判断矩阵按照列归一化 (每个元素除以其所在列的和)
2. 将归一化的各列相加 (按行求和)
3. 将相加后得到的向量除以 n 即可得到权重向量

- 数学描述如下所示:

  判断矩阵 $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$

  算术平均法求得的权重向量 $\omega_i = \frac{1}{n} \sum_{j=1}^{n} \frac{a_{ij}}{\sum_{k=1}^{n} a_{kj}} (i = 1, 2, \cdots, \mathrm{n})$

2.2.2 方法二: 几何平均法求权重

1. 将 A 的元素按照行相乘得到一个新的向量
2. 将向量的每个分量开 n 次方
3. 对该列向量进行归一化即可得到权重向量

- 数学描述如下所示:

  判断矩阵 $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$

  几何平均法求得的权重向量 $\omega_i = \frac{(\prod_{j=1}^{n} a_{ij})^{\frac{1}{n}}}{\sum_{k=1}^{n} (\prod_{j=1}^{n} a_{kj})^{\frac{1}{n}}} (i = 1, 2, \cdots, n)$

2.2.3 方法三: 特征值法求权重

前提: 矩阵的一致性可以接受, 仿照一致矩阵权重求法

1. 求出矩阵 A 的最大特征值以及其对应的特征向量
2. 对求出的特征向量进行归一化即可得到权重

- 数学描述如下所示:

判断矩阵 $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$

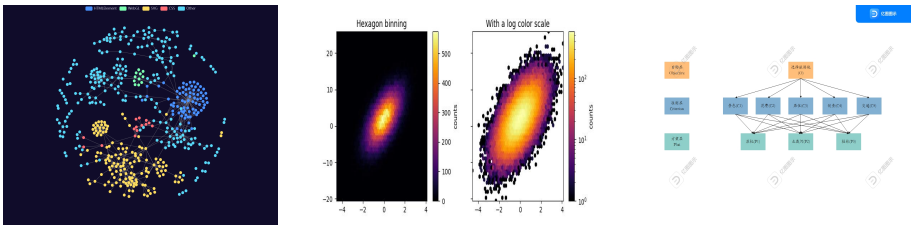
几何平均法求得的权重向量 $\omega_i = \frac{(\prod_{j=1}^{n} a_{ij})^{\frac{1}{n}}}{\sum_{k=1}^{n} (\prod_{j=1}^{n} a_{kj})^{\frac{1}{n}}} (i = 1, 2, \cdots, n)$

注释: 以往论文利用层次分析法解决实际问题时, 都是采用其中某一种方法求解权重, 而不同的计算方法可能会导致结果有所偏差。为了保证结果的鲁棒性, 本文采用三种方法分别求出了权重, 再根据得到的权重矩阵计算各个方案的的得分, 并且进行排序和综合分析, 这样有效避免了采用单一方法所产生的偏差, 得到的结论更全面、有效。[2]

2.3 层次结构图

2.3.1 画图软件指南

1. Echars
2. Matplotlib
3. 亿图示示



(a) Echarts (b) Matplotlib (c) 收费软件

图 1: 层次结构图

3 TOPSIS 算法

3.1 引入 TOPSIS 算法的 reason

1. 充分利用原始数据的信息，并能更精确反映出各个状况/方案之间的优劣与差距
2. 比较对象一般远大于 2 个
3. 比较指标有多个方面
4. 有很多指标不存在理论上的最大值和最小值

3.1.1 计算评分的公式

- $\frac{x-\min}{\max-\min}$

3.1.2 统一指标类型

- 指标正向化:

  将所有的指标 (极大型 or 极小型指标) 全部转化为极大型指标

3.2 具体数学公式

3.2.1 有 n 个评价对象,m 个评价指标, 计算得分情况

1. 标准化矩阵 Z:

$$Z = \begin{pmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{pmatrix}$$

2. 定义最大值 Z^+:

$$Z^+ = (Z_1^+, Z_2^+, \cdots, Z_m^+) =$$

$$(\max\{z_{11}, z_{21}, \cdots, z_{n1}\}, \max\{z_{12}, z_{22}, \cdots, z_{n2}\}, \cdots, \max\{z_{1m}, z_{2m}, \cdots, z_{nm}\})$$

3. 定义最小值 Z^-:

$$Z^- = (Z_1^-, Z_2^-, \cdots, Z_m^-) =$$

$$(\min\{z_{11}, z_{21}, \cdots, z_{n1}\}, \min\{z_{12}, z_{22}, \cdots, z_{n2}\}, \cdots, \min\{z_{1m}, z_{2m}, \cdots, z_{nm}\})$$

4. 定义第 i 个评价对象与最大值的距离 D_i^+:

   $D_i^+ = \sqrt{\sum_{j=1}^{m} \omega_j (Z_j^+ - z_{ij})^2}$ ω_j 为指标权重

5. 定义第 i 个评价对象与最小值的距离 D_i^-:

   $D_i^- = \sqrt{\sum_{j=1}^{m} \omega_j (Z_j^- - z_{ij})^2}$ ω_j 为指标权重

6. 得出第 i 个评价对象未归一化的得分:

   $S_i = \frac{D_i^-}{D_i^- + D_i^+}$

3.3 指标转化公式

3.3.1 极小型指标转化为极大型指标

- 极小型指标转化为极大型指标: $\max -x$

3.3.2 中间型指标转化为极大型指标

中间型指标: 指标值既不要太大, 也不要太小, 取特定值最好 (例如 PH 值检验)

- $\{x_i\}$ 为一组中间型指标序列, 且最佳值为 x_{best}, 则正向化公式如下
- $M = \max\{|x_i - x_{best}|\}, \tilde{x}_i = 1 - \frac{|x_i - x_{best}|}{M}$

3.3.3 区间型指标转化为极大型指标

区间型指标: 指标落在某个区间最好 (例如人体的体温范围)

1. $\{x_i\}$ 为一组中间型指标序列, 且最佳区间为 [a,b], 那么正向化公式如下:

2. $M = \max\{a - \min\{x_i\}, \max\{x_i\} - b\}$,

$$\tilde{x}_i = \begin{cases} 1 - \frac{a-x}{M} & \text{if } x < a \\ 1 & \text{if } a \le x \le b \\ 1 + \frac{x-b}{M} & \text{if } x > b \end{cases}$$

3.4 Matlab 相关代码

3.4.1 unfinished

- 详情参考 youtube.[清风数学建模].02-01TOPSIS 模型部分 [3]

3.5 基于熵权法对模型的修正

3.5.1 依据的原理

- 指标变异程度越小, 所反应的信息量也越少, 其对应的权重也越低
- 层次分析法的确定依赖专家, 主观性较强

3.5.2 熵权法的计算步骤-未完全理解

- x-可能发生的某种情况,p(x)-该情况发生的概率 $(0 \leq p(x) \leq 1)$
- 定义 $I(x) = -\ln(p(x))$
- 定义事件 **X** 的信息熵为:$H(X) = \sum_{i=1}^{n} [p(x_i)I(x_i)] = -\sum_{i=1}^{n} [p(x_i)\ln(p(x_i))]$
- while $p(x_i) = \frac{1}{n}$ 时,$H(X)_{\max} = \ln n$

1. 判断输入矩阵中是否存在负数,if 存在则需要重新标准化到非负区间
2. 计算第 j 项指标下第 i 个样本所占的比重, 并将其看作相对熵计算中用到的概率
3. 计算每个指标的信息熵, 并且计算信息效用值, 并且得到每个指标的熵权

4 插值算法

4.1 Lagrange 插值法

- 拉格朗日插值多项式 $P(x)$: $P(x) = \sum_{i=0}^{n} y_i \ell_i(x)$
- 拉格朗日基函数 $\ell_i(x)$: $\ell_i(x) = \prod_{\substack{0 \le j \le n \\ j \ne i}} \frac{x - x_j}{x_i - x_j}$
- $\ell_i(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}$

4.1.1 局限性

1. 插值多项式次数高, 精度未必提升
2. 多项式 dd 次数越高, 误差可能越大

4.2 分段线性插值

- 分段二次插值多项式 $P_i(x) = a_i(x - x_i)^2 + b_i(x - x_i) + c_i$
- 插值条件 $P_i(x_i) = y_i \quad P_i(x_{i+1}) = y_{i+1}$
- 连续可导条件（如果适用）$P_i'(x_{i+1}) = P_{i+1}'(x_{i+1})$

4.3 Newton 插值法

- 牛顿插值多项式 $P(x) = \sum_{i=0}^{n} a_i \prod_{j=0}^{i-1}(x - x_j)$
- 一阶差商 $f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}$
- 二阶差商 $f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i}$
- 一般 k 阶差商 $f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}$

4.4 Hermite 插值法

- Hermite 插值多项式 $H(x) = \sum_{i=0}^{n} \left[(1 - 2(x - x_i)h_i'(x_i)) \left(\frac{(x-x_i)^2}{h_i(x_i)^2} \right) y_i + (x - x_i) \left(\frac{(x-x_i)^2}{h_i(x_i)^2} \right) y_i' \right]$
- 基函数 $h_i(x) = \prod_{\substack{0 \le j \le n \\ j \ne i}} (x - x_j)$
- 基函数在插值点的导数 $h_i'(x_i) = \frac{d}{dx} \left(\prod_{\substack{0 \le j \le n \\ j \ne i}} (x - x_j) \right) \Bigg|_{x = x_i}$

4.5 Matlab 分段三次 Hermite 插值法

4.5.1 unfinished

5　拟合算法

6 相关系数

6.1 Pearson 相关系数

1. **总体均值** $(E(X))$: 也称为数学期望或期望值，是随机变量 X 取值的加权平均值，权重是相应取值的概率。在实际中，总体均值反映了随机变量的中心趋势。公式表示:

$$E(X) = \sum_i x_i \cdot P(X = x_i) = \frac{\sum_{i=1}^{n} x_i}{n}$$

其中，x_i 是随机变量的取值，$P(X = x_i)$ 是随机变量取值为 x_i 的概率。

2. **总体协方差** $Cov(X, Y)$: 用于衡量两个随机变量 X 和 Y 之间的线性关系。如果协方差为正，说明两个变量趋向于同向变化；如果协方差为负，说明它们趋向于反向变化。协方差为零表示两个变量是线性不相关的。公式表示:

$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \frac{\sum_{i=1} (x_i - \mu_X)(y_i - \mu_Y)}{n}$$

其中，$\mu_X = E(X)$ 和 $\mu_Y = E(Y)$ 分别是 X 和 Y 的均值。

3. **协方差的量纲与标准化**: 协方差的大小与变量的单位或量纲有关，因此不同量纲的协方差不适合直接比较。为了消除单位的影响，我们可以通过标准化将协方差转化为相关系数。

标准差:

$$\sigma_X = \sqrt{E[(X - \mu_X)^2]} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - E(X))^2}{n}}$$

σ_X 是随机变量 X 的标准差，反映了随机变量值的离散程度。类似地，σ_Y 是随机变量 Y 的标准差。

Pearson 相关系数:

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} \frac{(x_i - E(X))}{\sigma_X} \frac{(y_i - E(Y))}{\sigma_Y}}{n}$$

通过标准化协方差，我们得到两个变量的相关系数 ρ_{XY}，它取值在 -1 和 1 之间，反映了变量之间的线性关系强度。由此可得 Pearson 相关系数公式。

4. **总体与样本关系**: 通过样本统计量估计总体统计量，主要包括样本均值 (平均水平) 和样本标准差（偏离程度)。

5. **总体 Pearson 相关系数**: 对于随机变量 X 和 Y，总体 Pearson 相关系数定义为:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} \frac{(x_i - E(X))}{\sigma_X} \frac{(y_i - E(Y))}{\sigma_Y}}{n}$$

其中，$\mathrm{Cov}(X,Y)$ 表示 X 和 Y 的协方差，σ_X 和 σ_Y 分别是 X 和 Y 的标准差。

6. **样本 Pearson 相关系数**: 样本的 Pearson 相关系数为:

$$r_{XY} = \frac{Cov(X,Y)}{S_X S_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

公式 = 表达式 1（理论公式）= 表达式 2（Google or chatGPT 给的）= 表达式 3(建模课程教的)

其中，$\bar{x}$ 和 $\bar{y}$ 分别为 x 和 y 的样本均值，n 是样本大小,S_X 和 S_Y 是样本的标准差

7. 使用 Pearson 相关系数前应先确定两个变量是否线性相关（即先画散点图，再算相关系数)。

8. **相关系数的范围**: Pearson 相关系数的取值范围为 $[-1,1]$，其中:

   - $r = 1$ 表示完全正线性相关;
   - $r = -1$ 表示完全负线性相关;
   - $r = 0$ 表示无线性相关。

6.1.1 Pearson 相关系数的注意事项

- 非线性相关性也可能导致高的线性相关系数，需谨慎解释。
- 离群点对相关系数的影响很大，可能会显著改变相关系数的值。
- 高相关系数并不一定表示两者具有因果关系。
- 相关系数为 0 只能说明没有线性相关性，但可能存在其他更复杂的关系（如非线性关系）。

6.2 假设检验

1. **确定原假设和备择假设**：

$$H_0 : \rho = 0 \quad \text{(无相关性)}$$

$$H_1 : \rho \neq 0 \quad \text{(有相关性)}$$

2. **构造检验统计量**：在 H_0 成立的条件下，构造 t 检验统计量：

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

   其中，r 是样本的 Pearson 相关系数，n 是样本大小。

3. **确定临界值**：在给定的显著性水平 α 下，t 统计量服从自由度为 $n-2$ 的 t 分布。

4. **计算 P 值**：通过计算得到的 t 值与临界值进行比较，或者计算 P 值判断是否拒绝原假设。

6.2.1 P 值

- **P 值解释**：P 值表示在原假设为真时，观测到或更极端数据的概率。双侧检验的 P 值是单侧检验的两倍。
- 若 P 值小于显著性水平 α，则拒绝原假设；否则无法拒绝原假设。

6.3 Pearson 相关系数的假设检验条件

- 实验数据通常是成对的来自正态分布的总体，因为 Pearson 相关系数后通常会使用 t 检验进行检验，而 t 检验基于正态分布假设。
- 实验数据之间的差距不能太大，因为 Pearson 相关系数对异常值敏感。
- 每组样本之间应是独立抽样的，这是构造 t 统计量的前提。

6.4 正态分布检验（检验可以使用 python 完成）

- **JB 检验**（大样本 $n > 30$）：找出 H_0：服从正态分布和 H_1：不服从正态分布，计算 JB 统计量的 P 值后，与 0.05 比较，判断是否拒绝原假设。
- **Shapiro-Wilk 检验**（小样本 $n : 3 \sim 50$）：找出 H_0：服从正态分布和 H_1：不服从正态分布，计算 S-W 统计量的 P 值后，与 0.05 比较，判断是否拒绝原假设。
- **Q-Q 图**：通过 Q-Q 图判断样本数据是否近似于正态分布，只需观察图上点是否接近直线。

6.5 Spearman 相关系数

- **定义**：Spearman 相关系数用于衡量两个变量之间的单调关系，不要求数据服从正态分布。其公式为：

  $$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

  其中，d_i 是每对数据排名差的平方，n 是样本大小。
- 适用于非正态分布、非线性关系或定序数据的情况。
- 在小样本情况下，通过查表获得 Spearman 等级相关的临界值。
- 在大样本情况下，计算检验值并求出 P 值与 0.05 比较，判断是否拒绝原假设。

6.5.1 Spearman 与 Pearson 相关系数的选择

- 当数据连续且服从正态分布并且线性相关时，Pearson 相关系数更合适。
- 当上述任一条件不满足时，选择 Spearman 相关系数。
- Spearman 相关系数可用于定序数据，而 Pearson 相关系数不可。
- Pearson 相关系数衡量线性相关性，而 Spearman 相关系数衡量单调相关性。

  （单调关系，变量倾向于朝着相同的方向相对方向移动，但不一定以恒定的速率移动）

  （线性关系，变量沿着相同的方向以恒定的速率移动）

7 典型相关分析

7.1 定义

1. 在每组变量（每一组中有若干变量）中找出变量的线性组合，使得两组的线性组合之间具有最大的相关系数
2. 选取和最初挑选的这对线性组合不相关的线性组合，使其配对，并且选择相关系数最大的一对
3. 重复上述步骤，直至两组变量之间的相关性提取完毕
4. 选出的线性组合配对称为典型变量，它们的相关系数被称为典型相关系数。典型相关系数度量了这两组变量之间的强度。

7.2 典型相关分析的思路

假设两组变量分别为：

$$X^{(1)}=\left(X_1^{(1)},X_2^{(1)},\ldots,X_p^{(1)}\right),\quad X^{(2)}=\left(X_1^{(2)},X_2^{(2)},\ldots,X_q^{(2)}\right)$$

分别在两组变量中选取若干有代表性的综合变量 U_i,V_i，使得每一个综合变量是原变量的线性组合，即

$$U_i=a_1^{(i)}X_1^{(1)}+a_2^{(i)}X_2^{(1)}+\cdots+a_p^{(i)}X_p^{(1)}\triangleq\mathbf{a}^{(i')}\mathbf{X}^{(1)}$$

$$V_i=b_1^{(i)}X_1^{(2)}+b_2^{(i)}X_2^{(2)}+\cdots+b_q^{(i)}X_q^{(2)}\triangleq\mathbf{b}^{(i')}\mathbf{X}^{(2)}$$

注意：综合变量的组数是不确定的，如果第一组就能代表原样本数据大部分的信息，那么一组就足够了。假设第一组反映的信息不够，我们就需要找第二组了。并且为了让第二组的信息更有效，需要保证两组的信息不相关。

不相关：$\mathrm{cov}(U_1,U_2)=\mathrm{cov}(V_1,V_2)=0$

第一组满足的条件：

在 $\mathrm{var}(U_1)=\mathrm{var}(V_1)=1$ 满足的条件下，找到 $a^{(1)}$ 和 $b^{(1)}$ 两组系数，使得 $\rho(U_1,V_1)$ 最大。

（为什么要固定这个条件：因为相关系数与量纲无关：$\rho(U_1,V_1)=\rho(aU_1,bV_1)$）

7.3 多元统计-unlearned

7.3.1 随机向量

假设 X 是一个 n 维随机向量，表示为

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix},$$

其中 X_i $(i = 1, 2, \ldots, n)$ 均为随机变量。随机向量可以用来描述多维数据或多元随机变量的集合。

7.3.2 随机向量的期望

随机向量 X 的期望（或均值）定义为每个分量的期望组成的向量：

$$E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix},$$

其中 $E(X_i)$ 表示 X_i 的数学期望。

7.3.3 协方差矩阵

协方差矩阵 $D(X)$ 是随机向量各分量之间的协方差的集合，定义为：

$$D(X) = \mathrm{Cov}(X, X) = E\left[(X - E(X))(X - E(X))^T\right],$$

它是一个 $n \times n$ 的对称矩阵，矩阵的每个元素 $D_{ij}(X)$ 表示 X_i 和 X_j 之间的协方差，即

$$D_{ij}(X) = \mathrm{Cov}(X_i, X_j) = E\left[(X_i - E(X_i))(X_j - E(X_j))\right].$$

对角线元素 $D_{ii}(X)$ 是 X_i 的方差。

7.3.4 线性变换的协方差

假设 a 是一个 $n \times 1$ 的列向量（常数向量），则随机向量 X 经过线性变换后的方差为：

$$D(a^T X) = E\left[(a^T X - E(a^T X))(a^T X - E(a^T X))^T\right] = a^T E\left[(X - E(X))(X - E(X))^T\right] a = a^T D(X) a.$$

这表明线性变换后的方差可以表示为协方差矩阵 $D(X)$ 与向量 a 的双线性形式。

7.3.5 协方差矩阵的性质

协方差矩阵 $D(X)$ 具有以下重要性质：

- 对称性：$D(X)$ 是对称矩阵，即 $D(X)^T = D(X)$。
- 正定性：对于任意非零向量 a，$a^T D(X) a \geq 0$，即协方差矩阵是正定的。

7.3.6 联合协方差

如果 X 和 Y 均为 n 维随机向量，那么 X 和 Y 的联合协方差矩阵为：

$$\mathrm{Cov}(X, Y) = E\left[(X - E(X))(Y - E(Y))^T\right],$$

其中 $\mathrm{Cov}(X, Y)$ 是一个 $n \times n$ 的矩阵，表示 X 的各分量与 Y 的各分量之间的协方差。

此外，对于线性变换后的随机向量 $a^T X$ 和 $b^T Y$，它们的协方差为：

$$\mathrm{Cov}(a^T X, b^T Y) = E\left[(a^T X - E(a^T X))(b^T Y - E(b^T Y))^T\right] = a^T E\left[(X - E(X))(Y - E(Y))^T\right] b = a^T \mathrm{Cov}$$

7.3.7 补充知识点

1. ** 随机向量的期望 **：对于一个随机向量 X，其期望是一个向量，表示每个分量的期望值。这些期望值可以看作是随机变量的 “中心位置”。
2. ** 协方差矩阵 **：协方差矩阵不仅描述了单个变量的方差（即对角线元素），还描述了不同变量之间的线性相关性（即非对角线元素）。如果协方差为正，表示两个变量正相关；如果为负，表示负相关；如果为零，表示不相关。

3. ** 线性变换的协方差 **：线性变换会改变随机向量的方差结构。通过矩阵 $a^T D(X)a$，我们可以量化这种变换对方差的影响。

4. ** 协方差矩阵的应用 **：协方差矩阵在数据分析、特征提取（如主成分分析）中起着关键作用，它反映了多维数据的分布结构和变量之间的相互关系。

7.4 没学明白

8 数学规划模型

8.1 线性规划

- 目标函数 $f(x)$ 和约束条件均是决策变量的线性表达式

8.1.1 Matlab 中线性规划的标准型

数学表达式

$$\min C^{\mathrm{T}}X \quad (\text{内积 } C=\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix},\ X=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},\ \text{n 为决策变量个数})\ \mathsf{S.T.}$$

$$\begin{cases} Ax \le b & (\text{不等式约束}) \\ Aeq \cdot x = beq & (\text{等式约束}) \\ lb \le x \le ub & (\text{上下界约束}) \end{cases}$$

8.1.2 Matlab 求解线性规划的命令

1. $[x, fval] = linprog[c, A, b, Aeq, beq, lb, ub, X0]$
2. X0 代表给定 Matlab 迭代求解的初始数值, 可以不用给
3. $[c, A, b, Aeq, beq, lb, ub]$ 的意义和标准型中的意义一致
4. if 不存在不等式或者等式约束, 可以用 [] 代替 [A, b, Aeq, beq]
5. if 某个 x_i 无上下界, 可以 set lb(i)=-inf　ub(i)=+inf
6. retuen 的 x 表示取得最小值时候 x 的取值; fval 表示最优解得到的最小值
7. 不是所有的线性规划都有唯一解, 可能无解或无穷解
8. if 求解的是 max, 对 fval 求负值

8.2 非线性规划

- 目标函数 $f(x)$ 或者约束条件有一个是决策变量的非线性表达式

8.2.1 Matlab 中非线性规划的标准型

数学表达式

$$\min f(x)$$

$$\mathbf{S.T.} \quad \begin{cases} Ax \le b, & Aeq \cdot x = beq \quad \text{（线性约束）} \\ C(x) \le 0, & Ceq(x) = 0 \quad \text{（非线性约束）} \\ lb \le x \le ub \end{cases}$$

8.2.2 Matlab 求解非线性规划的函数

1. $[x, fval] = fmincon(@fun, x0, A, b, Aeq, beq, lb, ub, @nonlfun, option)$
2. 非线性规划中对于初始值 X0 的选取重要, 非线性规划求解的是局部最优解
3. if 需要寻找" 全局最优解",first-给定不同的初始值, 在其中寻找到最优解:
   second-先用蒙特卡洛模拟, 得到蒙特卡洛求解, 再将这个解作为初始值求解最优解
4. "option"选项可以给定特定的求解算法:1.interior-point(内点法) 2.sqp(序列二次规划) 3.active-set(有效集法) 4.trust-region-reflective(信赖域反射算法)
5. 使用上述四种算法, 观测求解的结果是否有改变, 体现出模型的稳健性 [4]
6. "@fun" 表示目标函数, 我们需要编写一个独立的"m 文件" 存储目标函数:
   $function f = fun(x)$　可以借助例题去理解
7. "@nonlfun" 表示非线性部分的约束, 我们同样需要编写一个独立的"m 文件" 存储约束条件:

$function[c, ceq] = nonlfun(x)$
c = [非线性不等式约束 1; 非线性不等式约束 p;]
ceq = [非线性等式约束 1; 非线性等式约束 q;]

8. 写入 matlab 时将 $f = x_1^2 + 3x_2$ 修改为 $f = x(1)^2 + 3 * x(2)$

8.3 整数规划

- 要求变量取整数的数学规划 (线性整数规划/非线性整数规划)-目前针对于线性整数规划

8.3.1 Matlab 整数规划求解

1. 线性规划函数-$[x, fval] = linprog[c, A, b, Aeq, beq, lb, ub, X0]$
2. 线性整数规划求解-$[x, fval] = intlinprog[c, intcon, A, b, Aeq, beq, lb, ub]$
3. intlinprog 不能只能初始值
4. 加入 incon 参数可以指定哪些决策变量是整数　example: 决策变量有 3 个 x_1, x_2, x_3　x_1, x_3 是整数, 则 $intcon = [1, 3]$

8.3.2 Malab 线性 0-1 规划求解

1. 依旧使用 intlinprog 函数, 不过在 lb 和 ub 上做改变
2. example: 决策变量有 3 个 x_1, x_2, x_3　x_1, x_3 是 0/1 变量,
   则 $intcon = [1, 3]$, $lb = \begin{bmatrix} 0 \\ -\inf \\ 0 \end{bmatrix}$ $ub = \begin{bmatrix} 1 \\ +\inf \\ 1 \end{bmatrix}$

8.4 最大最小化模型

在最不利的条件下，寻求最有利的策略。对每一个 x_i 先求出目标 $f_i(x)$ 的最大值，再求出这些最大值中的最小值。

示例: 急救中心选址，规划急救中心到所有地点最大距离的最小值；
投资规划中最大风险的最低限度等。

8.4.1 最大最小化问题的一般数学模型:

$$\min\{\max[f_1(x), f_2(x), \cdots, f_m(x)]\}$$

约束条件:

$$\text{S.T.} \quad \begin{cases} Ax \leq b \\ Aeq \cdot x = beq \\ C(x) \leq 0 \\ Ceq(x) = 0 \\ VLB \leq x \leq VUB \end{cases}$$

8.5 多目标规划模型

if 一个规划问题中有多个目标, 例如企业在保证利润同时保证产生污染最少, 这种情况下我们可以对多目标函数进行加权组合, 使得问题转换为单目标规划

8.5.1 求解思路

1. 首先, 将多个目标函数统一为最大化 or 最小化问题后进行加权组合
2. if 目标函数的量纲不同, 则需要对其进行标准化之后再进行加权, 标准化的方法是用目标函数除以一个常量, 该常量是该目标函数的某个取值
3. 对目标函数进行加权求和时, 权重需要专家填写, 实际建模过程中可以令权重相同

9 回归分析

回归分析的任务任务就是，通过研究自变量 X 和因变量 Y 的相关关系，尝试去解释 Y 的形成机制，进而达到通过 X 去预测 Y 的目的

回归分析的使命：*1.* 识别重要变量； *2.* 判断相关性方向； *3.* 估计权重（回归系数）

而根据因变量 Y 的类型可以将回归分析分成以下 *5* 大类：

类型	模型	Y 的特点	例子
线性回归	OLS、GLS（最小二乘）	连续数值型变量	GDP、产量、收入
0-1 回归	logistic 回归	二值变量（0-1）	是否违约、是否得病
定序回归	probit 定序回归	定序变量	等级评定（优良差）
计数回归	泊松回归（泊松分布）	计数变量	每分钟车流量
生存回归	Cox 等比例风险回归	生存变量（截断数据）	企业、产品的寿命

表 1: 回归分析的分类

9.1 线性回归

9.1.1 数据分类与对应的建模方法

1. 横截面数据（多元线性回归）：在某一个时间点收集到的不同对象的数据

   （example：某一年 GDP 数据 or 今年的体测数据）

2. 时间序列的数据（AR、MA、ARMA）：对同一个对象在不同时间连续观察所得到的数据

   （example：每年的身高体重数据 or 中国历年来的 GDP 数据）

3. 面板数据（固定效应和随机效应）：横截面数据与时间需略数据综合起来的一种数据资源

   （example：2008-2018 年，我国不同省份的 GDP 数值）

9.1.2 数据的收集

1. 宏观数据:【简道云汇总】、【虫部落数据搜索】、【数据平台】[5]
2. 微观数据:【人大经济论坛】[5]

9.2 0-1 回归

9.3 定序回归

9.4 计数回归

9.5 生存回归

10 引用

参考文献

[1] 姜启源，谢金星，叶俊，数学模型（第四版）[M]. 北京: 高等交易出版社，2011.

[2] [清风数学建模] 层次分析法

[3] [清风数学建模] TOPSIS 算法 02-01.35min 处

[4] [清风数学建模] 规划模型 24-11

[5] [清风数学建模] 回归模型 07-02.8min 处