简介

张伟平

zwp@ustc.edu.cn

Office: 东区管理科研楼 1006

Phone: 63600565

课件 http://staff.ustc.edu.cn/~zwp/

论坛 http://fisher.stat.ustc.edu.cn

1.1 简介

- 多重观测数据: 许多观测或者设计研究中, 每个试验单元的多个指标被同时观测收集

$$Y_i = (Y_{i1}, \ldots, Y_{ip}), i = 1, \ldots, n$$

- 多元分析是**一类**用于分析多重观测数据的方法.
- 基本想法是利用多重观测之间的潜在相关性来提升推断效率
- 一些多元技术基于特定的概率模型, 特别是多元正态分布, 其他不依赖于特定分布的方法称为" 模型自由的" (model-free)

多元方法的应用

- **维数缩减** 通过考虑大量测量变量的少部分组合来降低维数, 同时不损失重要的信息. 用途: 多元数据可视化, 发现重要特征(变量)
  - 消费者价格指数 (CPI) 通过组合一大类商品价格来得到
  - 体脂肪健康指数 (BMI) 通过测量并组合身高和体重观测值来得到
  - MDS 通过研究对象之间某种亲近关系为依据（如距离、相似系数等），将研究对象（样品或变量）在低维空间中给出标度或位置, 以便全面而又直观地再现原始各研究对象之间的关系，同时在此基础上也可按对象点之间距离的远近实现对样品的分类.

- **聚类** 识别观测单元中“相似”的单元
  - 电子商务通过分组聚类出具有相似浏览行为的客户，并分析客户的共同特征，可以更好的帮助电子商务的用户了解自己的客户，向客户提供更合适的服务。
  - 聚类分析被用来在网上进行文档归类来修复信息
- **分类** 使用特定的指标集将观测单元分为事先指定的类
  - 美国国税局使用退税信息 (收入, 扣缴税款, 捐款, 年龄等) 将纳税人分为两组: 需要审查和不需要审查
  - 通过检测铅合金中元素 (铜, 银, 锡, 锑) 的含量, 公安机构可以判断一些子弹是否来自同一批次
- **相关性分析** 变量之间的关联性是什么?
  - 搜索引擎与使用它的人之间的桥梁就是网站的相关性, 用户通过搜索引擎检索跟网站相关的内容找到该网站, 而搜

索引擎通常使用相关性规则, 来展示搜索结果. 一个有极高相关性的匹配是对那个搜索请求排名第一的候选结果.

- **预测** 若变量之间是有关联的, 则可以通过给定的信息来预测另一些变量
  - 利用高中成绩变量与大学成绩变量之间的联系，构造用于预测在大学里会成功与否的指标
  - 基于用户移动通信记录数据, 对用户流失进行预测.
- **假设检验** 可否发现两组或多组响应变量之间的差异?
  - 测量一些与污染有关的变量, 以研究一个城市地区的污染程度是在一周中大致保持不变, 还是在工作日和周末之间会有明显的不同.
  - 利用观测数据来研究职业结构的差异, 以决定支持两个对立的社会理论中的哪一个

数据的组织

- n 个观测, p 个变量:

	变量 1	变量 2	$\cdots$	变量 p
Item 1 Obs 1	x_{11}	x_{12}	$\cdots$	x_{1p}
Item 2 Obs 2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Item n Obs n	x_{n1}	x_{n2}	$\cdots$	x_{np}

x_{ik} : 第 i 个个体的第 k 个变量值. 常用矩阵表示

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad or \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

- 测量的类型
  - Nominal: 类别变量, 顺序没有意义. 例如性别, 颜色
  - Ordinal: 类别变量, 顺序有意义. 例如等级
  - Interval: 数值的差有意义, 没有固定的 0 位置. 例如温度, 海拔, 纪年
  - Ratio: 数值变量, 比值是有意义的, 具有固定的 0 值位置. 例如高度, 年龄, 体重等.

根据数据的测量水平选择合适的统计方法

数据缺失问题

- 测量很多变量时, 常遇到其中一些变量观测值缺失的情况
- 一种做法: **完全数据分析** —删除观测变量中具有缺失值的个体, 使用没有缺失值的个体观测进行推断
- 问题: 可能会导致许多观测被删除, 从而大大减少了样本量
- 问题: 可以导致有偏估计, 除非缺失数据为 **MCAR**(missing completely at random) 的 (缺失是完全随机的, 与观测到的变量和感兴趣的参数等独立)
- 更佳的解决方法: 使用多重插值 (multiple imputation), 对缺失值进行补缺.

描述性统计

- 均值向量: p 维随机变量 $\mathbf{x} = (X_1, \ldots, X_p)^T$ 的 (总体) 均值向量 μ 为

$$\mu = (\mu_1, \ldots, \mu_p)^T$$

其中 $\mu_j = EX_j$.

- 若样本为 $\mathbf{x}_1, \ldots, \mathbf{x}_n$, 则 μ 的矩估计为样本均值 $\bar{\mathbf{x}}$:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = (\bar{x}_1, \ldots, \bar{x}_p)^T$$

其中 $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$.

- $\mathbf{x}$ 的协方差矩阵 Σ:

$$Cov(\mathbf{x}) = \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

- Σ 的对角元为方差: $\sigma_{jj} = \sigma_j^2 = Var(X_j)$, Σ 的非对角元为协方差: $\sigma_{ij} = cov(X_i, X_j) = \sigma_{ji}$

- Σ 可以由样本协方差矩阵 $\mathbf{S}$ 来估计:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

- $\mathbf{S}$ 的对角元为样本方差, 非对角元为两个变量的样本协方差

$\bar{\mathbf{x}}$ 和 $\mathbf{S}$ 的几何解释

由于 p 元变量的 n 个观测阵为

$$\mathbf{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

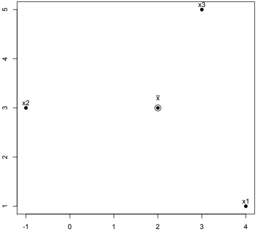
因此可以在 n 维或者 p 维空间里表示这些样本点:

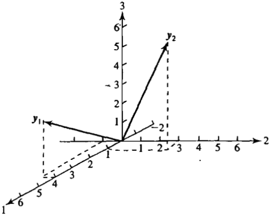
- 视 $\mathbf{X}$ 为 p 维空间里的 n 个点: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$

例如当

$$\mathbf{X} = \begin{pmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{pmatrix}, \qquad \text{有} \quad \bar{\mathbf{x}} = \begin{pmatrix} \frac{4-1+3}{3} \\ \frac{1+3+5}{3} \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

从而可以图示为 (左图:$p=2$ 空间里的 $n=3$ 个点; 右图: $n=3$ 维空间里的 $p=2$ 个向量)



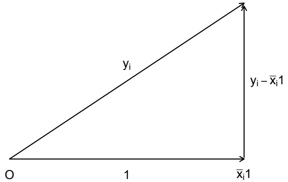


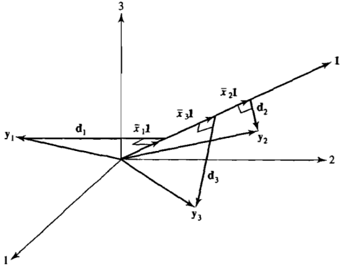
- 视 $\mathbf{X}$ 为 n 维空间里的 p 个向量:$\mathbf{X}=[\mathbf{y}_1,\ldots,\mathbf{y}_p]$, 如上右图所示.

对样本均值, 注意到 $\mathbf{y}_i$ 在单位向量 $\frac{1}{\sqrt{n}}\mathbf{1}$ 上的投影为

$$\mathbf{y}_i'\left(\frac{1}{\sqrt{n}}\mathbf{1}\right)\frac{1}{\sqrt{n}}\mathbf{1}=\bar{x}_i\mathbf{1}$$

从而 $\bar{x}_i = \mathbf{y}_i'\mathbf{1}/n$.





于是, 偏差向量

$$\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1}$$

表示了第 i 个变量偏离样本均值第 i 个分量的程度. 上右图表示了三个变量的 n 个观测点中, 每个变量偏离平均值的程度.

- 第 i 个偏差的长度平方: $L_{d_i}^2 = \mathbf{d}_i'\mathbf{d}_i = \sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2$
- 任何两个偏差 $\mathbf{d}_i$ 和 $\mathbf{d}_k$ 有

$$\mathbf{d}_i'\mathbf{d}_k = \sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) = L_{d_i}L_{d_k}cos(\theta_{ik})$$

- 从而夹角的余弦值为样本相关系数:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = cos(\theta_{ik})$$

定理 1. 若样本 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 为简单随机样本, 则

$$E\mathbf{S} = \Sigma.$$

证明. 由 $E\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} E\mathbf{x}_i = \mu$, 以及

$$\begin{aligned}
\mathbf{S} &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mu + \mu - \bar{\mathbf{x}})(\mathbf{x}_i - \mu + \mu - \bar{\mathbf{x}})' \\
&= \frac{1}{n-1}[\sum_{i=1}^{n}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)' - n(\bar{\mathbf{x}} - \mu)(\bar{\mathbf{x}} - \mu)']
\end{aligned}$$

注意到 $E(\bar{\mathbf{x}} - \mu)(\bar{\mathbf{x}} - \mu)' = Cov(\bar{\mathbf{x}}) = Cov(\mathbf{x}) = \Sigma$, 从而

$$E\mathbf{S} = \frac{1}{n-1}[n\Sigma - \Sigma] = \Sigma.$$

$\square$

- 利用矩阵运算, 有

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1} \quad \mathbf{S} = \frac{1}{n-1}\mathbf{X}'(I - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}$$

- **广义样本方差**: $|\mathbf{S}|$, 用以描述样本观测值变差的程度
- 在 n 维空间下, p 个偏差所围成的矩形区域体积与广义样本方差有关系: $|\mathbf{S}| = (n-1)^{-p}(\text{体积})^2$
- 在 p 维空间下, 记 $\bar{\mathbf{x}}$ 表示样本平均值, 有

$$\{\mathbf{x} : (\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) \le c^2\}\text{的体积} = k_p|\mathbf{S}|^{1/2}c^p$$

- **总样本方差**: $tr(\mathbf{S}) = s_{11} + s_{22} + \cdots + s_{pp}$
- 协方差受量纲影响, 相关系数则更方便解释.

定理 2. *广义样本方差为零, 当且仅当偏差矩阵的列向量之间线性相关.*

证明. 若 $\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$ 的列向量之间线性相关, 则存在非零常向量 $\mathbf{a}$ 使得

$$(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = 0$$

从而 $\mathbf{Sa}=0$, 由于 $\mathbf{a}\neq 0$, 所以 $|\mathbf{S}|=0$.

另一方面, 如果 $|\mathbf{S}|=0$, 则存在 $\mathbf{S}$ 的列的某个线性组合 $\mathbf{Sa}$, 使得 $\mathbf{Sa}=0$, 从而有 $(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')\mathbf{a}=0$, 用 $\mathbf{a}'$ 左乘即有

$$\mathbf{a}'(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')\mathbf{a}=0$$

因此 $(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')\mathbf{a}=0$. □

定理 3. *1. 如果 $n\leq p$, 则 $|\mathbf{S}|=0$.*

2. 设 $\mathbf{x}_1,\ldots,\mathbf{x}_n$ 为独立的 p 维样本, $\mathbf{S}$ 为样本协方差矩阵, 则

(a) 若对任意非零常向量 $\mathbf{a}$, 线性组合 $\mathbf{a}'\mathbf{x}_j$ 有正的方差 $(j=1,\ldots,n)$, 则在 $p<n$ 时,$\mathbf{S}$ 以概率 1 满秩且 $|\mathbf{S}|>0$.

(b) 若对所有 j, $\mathbf{a}'\mathbf{x}_j$ 以概率 1 为 (相同) 常数, 则 $|\mathbf{S}|=0$.

- 相关矩阵 ρ:

$$Corr(\mathbf{X}) = \rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

其中 $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$.

- 样本相关系数阵则为

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$$

其中 $\mathbf{D} = diag\{\mathbf{S}\}$

- 类似于广义样本方差, 可以通过 $|\mathbf{R}|$ 定义标准化变量的广义样本方差.

距离度量

- 在一些多元方法中, 两个观测之间的距离是非常重要的
- 欧氏距离是最简单的距离:

$$d_{ij} = [\sum_{k=1}^{p}(x_{ik} - x_{jk})^2]^{1/2} = \|\mathbf{x}_i - \mathbf{x}_j\|$$

- 欧式距离中各变量 (坐标) 同等重要
- 标准化距离: 当 p 个变量的散布不同时, 直观上在计算距离时候, 应该给予那些测量精度较高的维数以更大的权重, 权重和标准差的逆成比例, 从而计算距离前需要对数据进行标准化:

$$d(\mathbf{x}, \mathbf{y}) = [\sum_{j=1}^{p}\left(\frac{x_j - y_j}{s_j}\right)^2]^{1/2} = \sqrt{\|(\mathbf{x} - \mathbf{y})D^{-1}(\mathbf{x} - \mathbf{y})}$$

其中 $D = diag\{s_{jj}\}$

- 这种距离定义忽略了变量之间的相关性, 假定了变量之间是相互独立的

- 统计距离: 当各坐标不独立, 且散布不同时, 则可以对前述距离进行推广. 对两个随机变量的观测 $\mathbf{x}$ 和 $\mathbf{y}$, 定义

  $$d(\mathbf{x}, \mathbf{y}) = [(\mathbf{x} - \mathbf{y})^T A^{-1} (\mathbf{x} - \mathbf{y})]^{1/2}$$

  其中 $A = Cov(\mathbf{x} - \mathbf{y})$. (Mahalnobis distance)

- 一般的距离度量可以定义为

  $$d(\mathbf{x}, \mathbf{y}) = [(\mathbf{x} - \mathbf{y})^T A (\mathbf{x} - \mathbf{y})]^{1/2}$$

  其中 A 为对称正定矩阵.

距离度量的性质

- $\mathbf{x}$ 和 $\mathbf{y}$ 之间的任何距离度量 $d(\mathbf{x},\mathbf{y})$ 如果满足下述性质, 都是有效的:
  - $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$
  - $d(\mathbf{x},\mathbf{y}) > 0$, 若 $\mathbf{x} \neq \mathbf{y}$
  - $d(\mathbf{x},\mathbf{y}) = 0$, 若 $\mathbf{x} = \mathbf{y}$
  - $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$ $\forall\mathbf{z}$(三角不等式)

常用距离

- Euclidean: $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$
- Maximum: $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_\infty$
- Manhattan: $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1$
- Canberra: $d(\mathbf{x}, \mathbf{y}) = \sum \frac{|x_i - y_i|}{|x_i + y_i|}$
- binary: $d(\mathbf{x}, \mathbf{y}) = \frac{\#\{0,1\}}{\#\{0,1\} + \{1,1\}}$, $\mathbf{x}, \mathbf{y}$ 的元素为 0 或 1.
- Minkowski: $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p$

多元数据分析的目标

- 有些多元数据分析是**探索性的**(exploratory), 研究者的目的仅仅是搜寻数据中的模式 (patterns) 和确切的性状.
- 探索性方法多使用描述性统计方法, 数据的缩减以及图形.
- 当研究者以检验某个特定假设为目的时, 这时的多元分析方法称为是**验证性的**(confirmatory).
- 因此, 在验证性多元分析中, 可以使用显著性检验方法
- 许多验证性多元分析方法假定了一些特定的条件以保证结论是有效的