多元统计分析

协方差结构分析

- 主成分分析 PCA
  - 总体PCA
    - 概念
      - 主成分 $y_1 = v_1^T x$
        - 正定协方差阵二次型的极值定理
        - 方差最大
      - 第一主成分方向 v_1，也成为载荷
        - $v_{ij} \propto \rho_{x_i, y_j}$
          - $v_j = (v_{1j}, \ldots, v_{ij}, \ldots, v_{pj})^T$ 意为给每个变量加的权重
      - 载荷矩阵 V
      - 主成分变换 $y = V^T x$ 或 $y = V^T(x - \mu)$
        - 方差 $var(y) = V^T \Sigma V = \Lambda$
          - $var(y_k) = \lambda_k$
            - 方差贡献率：它能解释的方差占总方差的比例
        - 各个主成分不相关
  - 样本PCA
    - 概念
      - 数据矩阵 X
      - 中心化 X_c
      - 协方差阵 $S = X_c^T X_c / (n-1)$
      - 谱分解 $SV = V\Lambda, \Lambda = diag(\lambda_1, \ldots, \lambda_p)$
      - 主成分变换 $y_i = V^T(x_i - \bar{x})$
        - 第 j 分量成为第 j 主成分
      - 主成分矩阵
        - $Y = (y_1, \ldots, y_n)^T = X_c V$
        - 第 j 列 为所有样本点的第 j 主成分
          - 样本方差为 λ_j
      - 贡献率大于 80，画图时选 $k = 2, 3$
  - 双标图
    - 载荷矩阵V
      - 列代表主成分
      - 行代表变量
        - $y_i = V^T x_i, x_i = V y_i$
          - $x_{ik} \approx v_{k1} y_1 + v_{k2} y_2$
            - 不好解释
      - 点位置和箭头同向，说明变量取值大

- 因子分析
  - 单因子示例
    - 因子
      - 公共因子
        - 因子载荷L
      - 特殊因子
        - 均值为0，方差为1
    - 参数估计
      - 解方程 $\Sigma = LL^T + \Psi$
        - 精确求解
      - 最小二乘/主因子/SVD
      - 极大似然
  - 正交因子模型
    - 模型
      - $x_{p\times1} = \mu_{p\times1} + L_{p\times m} f_{m\times1} + \varepsilon_{p\times1}$
        - m个因子：$Ef = 0, var(f) = I_m$
      - $E\varepsilon = 0, var(\varepsilon) = \Psi = diag(\psi_1, \ldots, \psi_n)$
      - ε 与 f 独立
      - 矩阵表达 $X = FL^T + E$
    - 参数估计
      - 主因子法
        - PCA提供了一个好的逼近
        - $\hat{x} = (V_m \Lambda_m^{1/2})(\Lambda_m^{-1/2} y_m) = \hat{L}\hat{f}$
      - 极大似然法
        - 求解$\Sigma = LL^T + \Psi$的极大似然估计等价于优化问题：
          $\min_{\Sigma = LL^T + \Psi} (\log|\Sigma| + tr(S\Sigma^{-1}))$
          因为解不唯一，增加约束：$L^T \Psi^{-1} L =$ 对角矩阵
        - 拟合优度
          拟合优度：$W = n\log\left(|\hat{L}\hat{L}^T + \hat{\Psi}| / |S|\right)$.
          $\text{p-value} = P(\chi^2_{df} > W)$, $df = \frac{1}{2}[(p-m)^2 - p - m]$.
  - 因子分析
    - 方差分解
      - $trace(\Sigma) = tr(LL^T) + tr(\Psi) = h_1^2 + \ldots + h_p^2 + tr(\Psi)$
        - 变量 x_i 的共性方差：$h_i^2 = l_{i1}^2 + \ldots + l_{im}^2 = \|l_i\|^2$
        - $\psi_i = \sigma_{ii} - \|l_i\|^2 = \sigma_{ii} - h_i^2$
      - $L = cov(x, f)$
        - 行代表某变量共性方差
        - 列代表某因子载荷
        - 单个因子的方差贡献
      - $\Sigma = LL^T + \Psi$
        - 因子方差贡献率
          - 第 i 个变量中公共因子所能解释的方差的比例h_i^2/σ_{ii}
          - 公共因子能解释总方差的比例 $\frac{tr(LL^T)}{tr(\Sigma)} = \frac{h_1^2 + \ldots + h_p^2}{\sigma_{11} + \ldots + \sigma_{pp}}$
          - 单个因子的方差贡献 $SS_j = \|L_{(j)}\|^2$
            - L 的列平方和
            - 贡献率 $SS_j / tr(\Sigma)$
              - 第 j 个因子所能解释的总方差

- 典则相关分析 CCA
  - 矩阵的极分解
    - $A = UDV^T$
    - $A = \Theta R_1 = R_2 \Theta$
    - $R_1 = (A^T A)^{1/2}, R_2 = (AA^T)^{1/2}, \Theta = UV^T$
  - 普氏分析
    - 最优正交变换
      - $\max_{Q\text{正交}} tr(X^T Y Q^T)$
      - $X^T Y = UDV^T$
      - $Q = UV^T$
  - CCA
    - 一个实数度量两个向量的相关性
    - 概念
      - 典则相关系数
        - $\rho(a, b) = corr(a^T x, b^T y) \triangleq corr(\xi, \eta)$
      - 典则变量
        - 第一对典则变量
          - $\xi_1 = a_1^T x, \eta_1 = b_1^T y$, 使得 $\rho(a, b)$ 达到最大
        - 第二对典则变量
          - $\xi_2 = a_2^T x, \eta_2 = b_2^T y$, 分别与 ξ_1, η_1 不相关
          - 以此类推
    - 结果
      - 命题3: 满本Result10.1: 假设$\Sigma = cov\begin{pmatrix} \mathbf{y}_{\cdot m} \\ \mathbf{x}_{\cdot m} \end{pmatrix} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}, r = rank(\Sigma_{xy})$,
        假设$\Sigma_{yy}^{-1/2}\Sigma_{xy}\Sigma_{xx}^{-1/2}$的奇异值分解为$\Sigma_{yx} = UD\Psi^T$,
        其中$D = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$为$\Phi = \Sigma_{yy}^{-1}\Sigma_{yx}$或$\Psi = \Sigma_{xx}^{-1}\Sigma_{xy}$
        的只回非零特征根，$U = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$, $V = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$, $U^T U = V^T V = I_r$,
        则$\mathbf{x}$的所有典则变量$\mathbf{x}_{m\times1} = (\xi_1, \ldots, \xi_r)^T = \Psi^T \Sigma_{xx}^{-1/2} \mathbf{x}$, $\mathbf{y}$的所有典则变量
        $\mathbf{y}_{m\times1} = (\eta_1, \ldots, \eta_r)^T = U^T \Sigma_{yy}^{-1/2} \mathbf{y}$, 且 $var(\mathbf{x}_{m\times1}) = var(\mathbf{y}_{m\times1}) = I_r, cov(\mathbf{x}_{m\times1}, \mathbf{y}_{m\times1}) = D$.

多元统计分析

多元正态分布

- 多元正态 $X \sim N_p(\mu, \Sigma)$
  - 二次型求期望公式
    - $E(x^T C x) = \mu^T C \mu + tr(\Sigma C)$
  - 密度
    - $\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$
  - 矩母函数
    - $E\exp(t^T x) = \exp\left(t^T \mu + \frac{1}{2} t^T \Sigma t\right), \forall t \in R^p$
  - 条件和边际分布
    - 经典打洞
- 球对称分布
  - 球对称
    - 正交旋转不变
      - $Ox \overset{d}{=} x$
    - 密度只依赖于 $\|x\|$
  - χ, χ^2 分布
    - χ^2分布
      - 密度
        - $p_{\chi^2}(z) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}, z > 0$
    - χ_n 分布
      - 密度
        - $p_r(r) = \frac{1}{2^{n/2-1}\Gamma(n/2)} r^{n-1} e^{-r^2/2}, r > 0$
      - 分析数字特征
  - 单位球面均匀分布：$U(S^{n-1})$
    - 单位球面上唯一的正交不变分布
    - 定义
      - 概率和面积占比相同
      - $x = \|x\| \frac{x}{\|x\|} \triangleq ru$
        - $r \sim \chi_n$
        - $u \sim U(S^{n-1})$
        - 两者独立
    - 特例：n维标准正态
      - 是多元标准正态的刻画（等价定义）
        - 利用定理中 $p(x) = g(\|x\|)$ 以及$g(r)$和$p(r)$的关系
          - 球坐标换元
      - 球面均匀分布只依赖"密度只与r有关"，与正态形式无关
    - 边际分布
      - $z_i \sim \Gamma(\alpha_i), T = \sum_{i=1}^{n} z_i, x_i = \frac{z_i}{T}$
      - Dirichlet分布
        - $x = (x_1, \ldots, x_{n-1})$与 T独立，服从 $Dirichlet(\alpha_1, \ldots, \alpha_{n-1}; \alpha_n)$分布
      - 边际分布定理
        - 均匀球面分布
        - 定理3: 第一讲定理1: 假设$\mathbf{u} = (u_1, \ldots, u_n)^T \sim U(S^{n-1})$，则对任何$1 \leq k \leq n-1$, $\mathbf{u}_k = (u_1, \ldots, u_k)^T$的概率密度
          $p(\mathbf{u}_k) = \frac{\Gamma(n/2)}{\pi^{k/2}\Gamma((n-k)/2)}(1 - \|\mathbf{u}_k\|^2)^{(n-k)/2-1}, 0 < \|\mathbf{u}_k\| < 1.$
  - 低维例子计算
    - 证明独立性

多元正态总体的统计推断

- 正态总体的均值检验
  - 均值估计 $\bar{x}$；协方差估计 $\hat{S} = (n-1)S/n = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}_i)^T$
    - 无偏性
      - $E(\bar{x}) = \mu, var(\bar{x}) = \frac{\Sigma}{n}$；$E(S) = \Sigma$
    - 极大似然估计
      - $\bar{x}_i)^T$
  - 均值检验
    - 单正态：Hotelling T^2
      - 直接利用正态总体的 Hotelling T^2检验 $\mu = \mu_0$
      - 应用：成对比较问题
    - 两正态均值是否相同：Hotelling T^2
      - 定理2:假设两个多元n维总体（方差相同但均值可以不同）：
        $\mathbf{x}_{1\cdot n} \ldots \mathbf{x}_{n_1} \ iid \sim N_p(\boldsymbol{\mu}_1, \Sigma)$, 样本均值和方差为$\bar{\mathbf{x}}_{\cdot1}, S_1$; $\bar{x}_{\cdot1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, S_1 = \sum_{i=1}^{n_1}(x_i - \bar{x}_{\cdot1})(x_i - \bar{x}_{\cdot1})^T/(n_1 - 1)$
        $\mathbf{x}_{2\cdot n} \ldots \mathbf{x}_{n_2} \ iid \sim N_p(\boldsymbol{\mu}_2, \Sigma)$, 样本均值和方差为$\bar{\mathbf{x}}_{\cdot2}, S_2$; $\bar{x}_{\cdot2} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i, S_2 = \sum_{i=1}^{n_2}(x_i - \bar{x}_{\cdot2})(x_i - \bar{x}_{\cdot2})^T/(n_2 - 1)$
        令$S_{pooled} = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)S_1 + (n_2 - 1)S_2)$, 总样本量$n = n_1 + n_2$,则
        $T^2 = \frac{1}{1/n_1 + 1/n_2}(\bar{\mathbf{x}}_{\cdot1} - \bar{\mathbf{x}}_{\cdot2} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^T S_{pooled}^{-1}(\bar{\mathbf{x}}_{\cdot1} - \bar{\mathbf{x}}_{\cdot2} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) \sim \frac{(n-2)p}{n-p-1} F_{p,n-p-1}$
  - 轮廓分析
    - 检验$C\mu_1 = C\mu_2$
  - 多正态（MANOVA）：Wilks检验（近似）
    - 检验原假设 $\mu_1 = \ldots = \mu_g$
    - 组间差异 B, 组内差异 W
    - 如果组间差异较大，否定原假设，认为各组之间存在差异
    - Wilks检验：U小时否定原假设
      - 令Wilks检验统计量(Wilks' Lambda)
        $\Lambda^* = \frac{\det(W)}{\det(W+B)}$
        则 H_0成立时 $T_1 = -n\log\Lambda^* \xrightarrow{d} \chi^2_{(k-1)p}$,
        或Bartlett修正：$T_2 = -(n-1-\frac{p+g}{2})\log\Lambda^* \xrightarrow{d} \chi^2_{(k-1)p}$,
        当T_1或$T_2 \geq \chi^2_{p(k-1)}(\alpha)$ 或时，在α水平下否定原假设。
      - $U = (B+W)^{-1/2} W (B+W)^{-1/2}$
        - 检验统计量：$\Lambda^* = |U| = |W|/|B+W| \sim \Lambda_p(m_1, m_2) \overset{d}{=} \Pi_{i=1}^{m_1} beta(\frac{m_1 - p + i}{2}, \frac{p}{2})$
        - 从似然比出发得到的近似逼近，叫 Wilks 检验
- $Wishart$分布
  - 定义
    - 定义(Wishart分布) 设$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \ iid \sim N_p(\mathbf{0}, \Sigma)$, 记$Z = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)^T$,
      则随机矩阵$W = Z^T Z = \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T$的分布称为自由度为$m$，参数为$\Sigma$的Wishart分布, 记为$W \sim W_p(m, \Sigma)$.
  - 标准 $Wishart$分布
    - 协方差为 I_p
  - 性质 $W \sim W_p(m, \Sigma)$
    - 线性变换
      - $AWA^T \sim W_q(m, A\Sigma A^T), A_{q\times p}$
      - 标准化：$\Sigma^{-1/2} W \Sigma^{-1/2} \sim W_p(m, I_p) = W_p(m)$
      - $t^T W t / t^T \Sigma t \sim \chi^2_m$
    - $Wishart$矩阵分块
      - $\sim W_p(m, \Sigma)$, 定义$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}$, $\begin{array}{l} \Sigma_{11} \ \Sigma_{12} \\ \Sigma_{21} \ \Sigma_{22} \end{array}$
        令$k = 1$, 则
        1. Σ_{11}: $W_{11} \sim W_{p-1}(m, \Sigma_{11})$,
        2. 记$W_{22\cdot1} = W_{22} - W_{21}W_{11}^{-1}W_{12} \sim W_p(m - 1, \Sigma_{22\cdot1})$
        3. $[W_{12} \sim N(W_{11}\Sigma_{11}^{-1}\Sigma_{12}, W_{22\cdot1}\sigma^{-1})$, $\sigma^{-1} \sim \chi^2_{m-1}]$
        $vec(W_{12}^T)|W_{22} \sim N(vec(\Sigma_{12}\Sigma_{11}^{-1}W_{22}^{-1}), I)$
      - 令$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}$, $\begin{array}{l} \Sigma_{11} \ \Sigma_{12} \\ \Sigma_{21} \ \Sigma_{22} \end{array}$, $\begin{array}{l} \Sigma^{11} \ \Sigma^{12} \\ \Sigma^{21} \ \Sigma^{22} \end{array}$
        $\begin{array}{l} \Sigma_{11} \ \Sigma_{12} \\ \Sigma_{21} \ \Sigma_{22} \end{array} = \begin{pmatrix} \sum_{i=1}^{n} X_i^2 & \sum_{i=1}^{n} X_i X_j \\ \sum_{i=1}^{n} X_i X_j & \sum_{i=1}^{n} X_j^2 \end{pmatrix} \sim W_2(m, \Sigma)$, 则
        $X_i \sim N_1(\mu_i, \sigma_i)$, $w_{ij} = \sum_{k=1}^{n} X_i^k - \sigma_{ij} \chi_i^2$;
        $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} (X_i)^2$, $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} \chi_i^2 + n_{11}\mu_1^2 \sim \chi^2_{n,\lambda}$, $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} \chi_i^2 + n_{22}\mu_2^2 \sim \chi^2_{n,\lambda}$
    - 性质 $W \sim W_p(m, \Sigma)$
      - 一维、倒数
        - 引理1：假设$W \sim W_p(m), m \geq p$, 对任何常数向量$\mathbf{t} \in R^p$, $\|\mathbf{t}\| = 1$, 有 $\mathbf{t}^T W \mathbf{t} \sim \chi^2_m$, $\frac{1}{\mathbf{t}^T W^{-1} \mathbf{t}} \sim \chi^2_{m-p+1}$
      - 含$Wishart$矩阵的二次型
        - F分布的导出
          - 定理5：假设$W \sim W_p(m, \Sigma), m \geq p, \boldsymbol{x} \sim N(\mathbf{0}, \Sigma), \Sigma > 0$, 且$W$与$\boldsymbol{x}$独立，则 $\left(\frac{m-p+1}{p}\right) \times \boldsymbol{x}^T W^{-1} \boldsymbol{x} \sim F_{p,m-p+1}$
        - 样本均值和样本方差的分布（正态总体）
          - 定理6：设$\mathbf{x}_1, \ldots, \mathbf{x}_n \ iid \sim N_p(\boldsymbol{\mu}, \Sigma)$, 假设$n > p$, 则
            (1) $\bar{\mathbf{x}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\Sigma\right)$, $(n-1)S \sim W_p(n-1, \Sigma)$; $\bar{\mathbf{x}}$与S独立.
            (3) $T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu})' S^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$.
      - 应用：统计推断
        - 置信椭球；Hotelling T检验
  - 样本均值 $\bar{x} = \sum_{i=1}^{n} x_i / n$；样本方差 $S = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}_i)^T/(n-1)$
    - 统计量
    - 无偏性
    - 极大似然估计
  - 多正态（MANOVA）：Wilks检验（近似）
  - 轮廓分析
    - 检验$C\mu_1 = C\mu_2$
- 样本数字特征
  - $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in R^{n\times p}$
  - 样本均值：$\bar{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i / n = X^T \mathbf{1} / n$,
  - 样本方差：$S = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T / (n-1) = X^T P_c X / (n-1)$.

聚类、分类

- 距离和相似度
  - 距离
    - 欧式距离矩阵
      - 例子
      - $D = (d_{ij})_{n\times n}$
      - 存在 $x_1, \ldots, x_n \in \mathbb{R}^p$
        - $d_{ij} = \|x_i - x_j\| = \sqrt{(x_i - x_j)^T (x_i - x_j)}$
      - 等价条件
        - $\tilde{D} = (-\frac{1}{2}d_{ij}^2)$, $J_n = P_{1\cdot} = I_n - 1_n^T 1_n / n$
        - 距离矩阵欧式等价于 $J_n \tilde{D} J_n$是欧式内积矩阵
          - 左乘是列中心化，右乘是行中心化
          - $J_n A_{n\times p} J_p$
            - anova 双向中心化
  - 相似度
    - 定义
      - 距离的减函数
        - $d_{max} - d$
      - 内积
        - $s(x, y) = x^T y = (\|x\|^2 + \|y\|^2 - \|x - y\|^2)/2$
    - 欧式相似性矩阵
      - 存在 $x_1, \ldots, x_n \in \mathbb{R}^p$
        - $S = (s_{ij})_{n\times n} = (x_i^T x_j)_{n\times n} = XX^T$
          - $X_{n\times p} = (x_1, \ldots, x_n)^T$
      - 等价于半正定 $S \geq 0$
  - 欧式距离矩阵和中心化内积唯一互相确定
    - $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$
    - $s_{ij} = -\frac{1}{2}(d_{ij}^2 - \bar{d}_{i\cdot}^2 - \bar{d}_{j\cdot}^2 + \bar{d}_{\cdot\cdot}^2)$
- 多维标度法 MDS
  - 数据的低维表示
  - 目标
    - 极小化 $Stress$ 函数
      $Stress = \sum_{i,j}(s_{ij} - \mathbf{x}_i^T \mathbf{x}_j)^2 = \|S - XX^T\|_F^2, rank(X) = k$,
      或 $Stress = \sum_{i,j}(d_{ij} - \|\mathbf{x}_i - \mathbf{x}_j\|)^2$.
  - 算法
    - $cMDS$
      - 1. 计算$\tilde{D} = \left(-\frac{1}{2}d_{ij}^2\right)$, 将$\tilde{D}$双向中心化：$S = J\tilde{D}J$
      - 2. 谱分解$S = U\Lambda U^T$.
      - 3. 对$k < rank(S)$, 取k个最大特征根组成对角阵$\Lambda_k = diag(\lambda_1, \ldots, \lambda_k)$,
        对应的特征向量按列排列成U_k, 记$X_k = U_k \Lambda_k^{1/2}$,
        X_k的n个行向量即是R^k中最优的n个向量。
    - 对XX^T谱分解
    - $isoMDS$
      - Kruskal - Shepard non - metric scaling：
        对于给定的距离矩阵$D = (d_{ij})$，求$\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^k$和单调增函数$f$, 极小化
        $Stress(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n; f) = \frac{\sum_{i\neq j}(f(d_{ij}) - \|\mathbf{x}_i - \mathbf{x}_j\|)^2}{\sum_{i\neq j}\|\mathbf{x}_i - \mathbf{x}_j\|^2}$.
    - $non - metricMDS$
      - 保序
- 图像压缩
- 应用
  - 网页搜索排序
    - 概念
      - 邻接矩阵
      - 链出多 Hub 大，链入多 Authority 大
        - Hub, Authority
    - $HITS$ 算法
      - 思想
        - 权威性网页由重要Hub链入，重要Hub网页链出到权威网页
      - 解法
        - 权威性 v, Hub u
        - SVD
        - 幂次迭代
          网页重要性计算的$HITS$迭代算法（*power method*）：
          初始$\mathbf{v}^{(0)} = (1, 1, \ldots, 1)^T / n$,
          $\mathbf{u}^{(k)} = A \cdot \mathbf{v}^{(k)}$, $\mathbf{u}^{(k)} \leftarrow \mathbf{u}^{(k)} / \|\mathbf{u}^{(k)}\|$
          $\mathbf{v}^{(k+1)} = A^T \mathbf{u}^{(k)}$, $\mathbf{v}^{(k+1)} \leftarrow \mathbf{v}^{(k+1)} / \|\mathbf{v}^{(k+1)}\|$
    - $Pagerank$ 算法
      - 不区分重要性
      - 思想：重要的网页应该链接重要的网页, 而且被对外链接较少的重要网页链接。
      - 得分 $x_i = \sum_{k: k\to i} \frac{x_k}{d_k}$
      - 计算
        PageRank算法：
        邻接矩阵为$A, D = diag(A(\mathbf{1}_n)), H = D^{-1}A$,
        记x_i为网页i的重要性得分(*score*), *PageRank* 假设模型：$\mathbf{x} = H^T \mathbf{x}$
        迭代求解$\mathbf{x}$: $\mathbf{x}^{(k+1)} = H^T \mathbf{x}^{(k)}, k = 0, 1, 2, \ldots$.
- 聚类分析
  - 层次聚类
    - 集合距离
      - 单连结 Single：链状结构（长链），非线性分类
      - 完全连结 Complete：紧密
      - 平均连结 average
      - Centroid, median linkage：逆序，树倒长
      - 距离最小的两个cluster合并成一个更大的cluster
  - K-means
    - 划分给定时，求各类中心；中心给定时，划分数据点
      - 聚类方法
    - 多次聚类避免局部极小
  - K-medoid
    - 只用距离，无原始数据
    - 递归迭代算法
      K - medoid clustering算法：
      1. 给定$\{1, 2, \ldots, n\}$的一个划分$C_1, \ldots, C_K$, 对每一个类$1 \leq k \leq K$, 求其中心c_k，即在类内与其它个体总距离最小的点$m_k \in C_k$：
         $m_k = \arg\min_{m \in C_k} \sum_{j \in C_k} d_{ij}$
      2. 给定当前的中心$\{m_k, k = 1, \ldots, K\}$, 对将每个个体$i$划分到离它最近的中心所属的类中，即求
         $k_i = \arg\min_{k} d_{m_k i}, 1 \leq k_i \leq K$.
         并将样本点i划归到第k_i类，由此得到更新的划分$C_1, \ldots, C_K$.
      3. 迭代1-2
  - 混合高斯分布模型
    - 迭代算法：一步更新点的类标记概率k，一步更新类的参数μ_k, Σ
      - M - step（Maximization - 极大似然）：估计类的中心c_i
        当知道每个样本所属类别时(即已知G_i $i = 1, \ldots, n$), 求各类中心c_i：
        $\hat{\boldsymbol{\mu}}_k = \sum_{i=1}^{n} \mathbf{x}_i / n_k = \sum_{i:G_i=k} \mathbf{x}_i / n_k$（第k类内样本的平均值）
        以及$\hat{\Sigma}_k = \sum_{i:G_i=k}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T / n_k, \hat{p}_k = \sum_{i=1}^{n} I_{\{G_i=k\}} / n, k = 1, \ldots, K$.
      - E - step（Expectation - 条件期望）：预测G_i
        给定$\boldsymbol{\theta}_k, k = 1, \ldots, K$, 将每个数据点划分到概率最大的类中：
        若$f_k(\mathbf{x}_i) > f_l(\mathbf{x}_i), \forall l \neq k$, 则$G_i$预测为$\hat{G}_i = k$,
        即$\hat{G}_i = \arg\max_k P(G_i = k \mid \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k)$.
- 谱聚类
  - 基于相似度矩阵的特征向量
  - 图、加权图
    - 拉普拉斯矩阵：$L = D - W$
      - 二次型 $x^T L x = \frac{1}{2}\sum_{i,j} w_{i,j}(x_i - x_j)^2$
      - 0 特征根重数等于互不连通的全连通子图个数
  - 聚类
    - mincut
      - 两类
        拓展的mincut问题：假设加权图是全连通的，L的特征根$0 = \lambda_1 < \lambda_2 \leq \lambda_3, \ldots$
        $\min_{\mathbf{x}\in R^n} \mathbf{x}^T L \mathbf{x} = \min_{\mathbf{x}\in R^n} \frac{1}{2}\sum_{ij} w_{ij}(x_i - x_j)^2$, s.t. $\mathbf{x}^T \mathbf{x} = 1, \mathbf{x}^T \mathbf{1} = 0$,
        由引理2及二次型极值定理（第9讲定理1），最优解为最小非0特征根λ_2的特征向量。对该特征向量的分量进行聚类，或更简单地，通过排序或考察它们的符号即可将节点分为两类。
      - k类
        假设加权图是全连通的，L的特征根$0 = \lambda_1 < \lambda_2 \leq \lambda_3, \ldots$, 考虑二次优化
        $\min_{V\in R^{n\times k}} \operatorname{tr}(V^T L V)$, s.t. $V^T V = I_{k\times k}, V^T \mathbf{1} = 0$,
        由引理2及二次型极值定理（第9讲定理1），最优解为$V = (\mathbf{v}_1, \ldots, \mathbf{v}_{k-1})$, $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$分别是$L$的最小非0特征根$\lambda_2, \ldots, \lambda_{k-1}$的特征向量。
        对$V = (\mathbf{v}_1, \ldots, \mathbf{v}_{k-1})$的各行进行通常的聚类，将节点聚集为k类。
    - 谱聚类
      - 假设$W_{n\times n}$是n个体/物体的相似度矩阵，令$w_{ii} = 0$,（如果是0-1矩阵，可认为是普通图的邻接矩阵，否则认为是加权图的权重矩阵）。
      - 计算度数矩阵$D = diag(\mathbf{d})$, 其中$\mathbf{d} = (d_1, \ldots, d_n)^T = W\mathbf{1}$. 拉普拉斯$L = D - W$
      - 计算L的谱分解，取其前k个最小特征根的对应的特征向量
        $V = (\mathbf{v}_1, \ldots, \mathbf{v}_k)$
        包括0特征根(可能重复)及其它最小非0的特征根的特征向量.
      - 记P各行为$\mathbf{y}_1, \ldots, \mathbf{y}_n \in R^k$, 即$V = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$
      - 对$\mathbf{y}_1, \ldots, \mathbf{y}_n \in R^k$进行$K$-means聚类，将$n$个节点聚成$K$类。