FDU 统计机器学习 5. 无监督学习

本文参考以下教材:

- 机器学习 (周志华) 第 9 章
- 统计学习方法 (第二版, 李航) 第 14 章

欢迎批评指正!

期末考试:

- 选择题 30 道 (45 分)
- 四道大题 (若干小问, 55 分)

5.1 无监督学习

(无监督学习概论参考统计学习方法 (第二版, 李航) 第 13 章)

5.2 聚类

5.2.1 基本概念

聚类是针对给定的样本，依据其特征的相似度，将其归并到若干个类的数据分析问题.
聚类属于无监督学习，因为它只是根据样本的相似度将其归类，而类别事先并不知道.
我们假设每个样本只能属于一个类，这称为**硬聚类** (hard clustering) 方法.

我们可以将样本集合看作是向量空间中的点集，并以该空间的距离表示样本之间的相似度.
记样本集合 X 为:

$$X = [x^{(1)}, \dots, x^{(n)}] \in \mathbb{R}^{m \times n}$$

其中每个列向量代表一个样本，对应 $\mathbb{R}^m$ 中的向量.
常用的距离有:

- **① Minkowski 距离** $d_p(x, y) := (\sum_{k=1}^m |x_k - y_k|^p)^{\frac{1}{p}}$ $(p \geq 1)$
  常用的 Minkowski 距离有:
  - **Manhattan 距离:** $d_1(x, y) := \sum_{k=1}^m |x_k - y_k|$
  - **Euclid 距离:** $d_2(x, y) := (\sum_{i=1}^m (x_k - y_k)^2)^{\frac{1}{2}}$ (它是 $\mathbb{R}^m$ 的默认度量)
  - **Chebyshev 距离:** $d_\infty(x, y) := \max\{|x_k - y_k| : 1 \leq i \leq m\}$
- **② Mahalanobis 距离:**
  给定协方差矩阵 $\Sigma \in \mathbb{R}^{m \times m}$
  我们定义样本 $x^{(i)}$ 和 $x^{(j)}$ 之间的 Mahalanobis 距离为:

  $$d_\Sigma(x^{(i)}, x^{(j)}) := \sqrt{(x^{(i)} - x^{(j)})^{\mathrm{T}} \Sigma^{-1} (x^{(i)} - x^{(j)})}$$

  特殊地，当 $\Sigma = I_m$ 时，Mahalanobis 距离就是 Euclid 距离.

样本相似度还有其他度量方式:

- **③ 相关系数:**
  样本 $x^{(i)}$ 和 $x^{(j)}$ 之间的相关系数定义为:

$$
\begin{aligned}
r(x^{(i)}, x^{(j)}) &:= \frac{(x^{(i)} - \bar{x}^{(i)}1_m)^{\mathrm{T}}(x^{(j)} - \bar{x}^{(j)}1_m)}{\|x^{(i)} - \bar{x}^{(i)}1_m\|_2\|x^{(j)} - \bar{x}^{(j)}1_m\|_2} \\
&= \frac{(x^{(i)} - \frac{1}{m}1_m1_m^{\mathrm{T}}x^{(i)})^{\mathrm{T}}(x^{(j)} - \frac{1}{m}1_m1_m^{\mathrm{T}}x^{(j)})}{\|x^{(i)} - \frac{1}{m}1_m1_m^{\mathrm{T}}x^{(i)}\|_2\|x^{(j)} - \frac{1}{m}1_m1_m^{\mathrm{T}}x^{(j)}\|_2} \\
&= \frac{(x^{(i)})^{\mathrm{T}}(I_m - \frac{1}{m}1_m1_m^{\mathrm{T}})^2x^{(j)}}{\sqrt{(x^{(i)})^{\mathrm{T}}(I_m - \frac{1}{m}1_m1_m^{\mathrm{T}})^2x^{(i)}}\sqrt{(x^{(j)})^{\mathrm{T}}(I_m - \frac{1}{m}1_m1_m^{\mathrm{T}})^2x^{(j)}}} \\
&= \frac{(x^{(i)})^{\mathrm{T}}(I_m - \frac{1}{m}1_m1_m^{\mathrm{T}})x^{(j)}}{\sqrt{(x^{(i)})^{\mathrm{T}}(I_m - \frac{1}{m}1_m1_m^{\mathrm{T}})x^{(i)}}\sqrt{(x^{(j)})^{\mathrm{T}}(I_m - \frac{1}{m}1_m1_m^{\mathrm{T}})x^{(j)}}}
\end{aligned}
$$

相关系数 $r(x^{(i)}, x^{(j)})$ 的绝对值越接近于 1，相似度就越高.

- **④ 夹角余弦:**
  样本 $x^{(i)}$ 和 $x^{(j)}$ 之间的夹角余弦定义为:

$$
\cos\left(x^{(i)}, x^{(j)}\right) := \frac{(x^{(i)})^{\mathrm{T}}x^{(j)}}{\|x^{(i)}\|_2\|x^{(j)}\|_2}
$$

  夹角余弦 $\cos\left(x^{(i)}, x^{(j)}\right)$ 越接近于 1，相似度就越高.

5.2.2 类内特征

考虑类别 c 的样本集合 D_c，样本个数为 n_c
常用的类内特征:

- **① 类均值 (类中心):**

$$
\mu(D_c) := \frac{1}{n_c}\sum_{i=1}^{n_c} x_i
$$

- **② 样本间平均距离:**

$$
\mathrm{avg}(D_c) := \frac{2}{|D_c|(|D_c| - 1)}\sum_{1 \le i < j \le |D_c|} d(x^{(i)}, x^{(j)})
$$

- **③ 类直径 (diameter):** (即样本间最大距离)

$$
L(D_c) := \max_{x, y \in D_c} d(x, y)
$$

- **④ 样本散布矩阵 (scatter matrix) & 样本协方差矩阵 (covariance matrix):**

$$
\begin{aligned}
\mathrm{Sca}(D_c) &:= \sum_{i=1}^{n_c}(x^{(i)} - \mu(D_c))(x^{(i)} - \mu(D_c))^{\mathrm{T}} \\
\hline
\mathrm{Cov}(D_c) &:= \frac{1}{n_c - 1}\mathrm{Sca}(D_c) \\
&= \frac{1}{n_c - 1}\sum_{i=1}^{n_c}(x^{(i)} - \mu(D_c))(x^{(i)} - \mu(D_c))^{\mathrm{T}}
\end{aligned}
$$

5.2.3 类间差距

考虑类别 c_1, c_2 的样本集合 D_{c_1}, D_{c_2}，样本个数分别为 n_{c_1}, n_{c_2}
常用的刻画类间差距的量:

- **① 最短距离:** (又称单连接)

$$d_{\min}(D_{c_1}, D_{c_2}) := \min\{d(x, z) : x \in D_{c_1}, z \in D_{c_2}\}$$

- **② 最长距离:** (又称完全连接)

$$d_{\max}(D_{c_1}, D_{c_2}) := \max\{d(x, z) : x \in D_{c_1}, z \in D_{c_2}\}$$

- **③ 中心距离:** (类中心之间的距离)

$$d_{\text{center}}(D_{c_1}, D_{c_2}) := d(\mu(D_{c_1}), \mu(D_{c_2}))$$
$$\text{where } \begin{cases} \mu(D_{c_1}) := \frac{1}{n_{c_1}} \sum_{i=1}^{n_{c_1}} x_i \\ \mu(D_{c_2}) := \frac{1}{n_{c_2}} \sum_{i=1}^{n_{c_2}} x_i \end{cases}$$

- **④ 平均距离:**

$$\bar{d}(D_{c_1}, D_{c_2}) := \frac{1}{n_{c_1} n_{c_2}} \sum_{x \in D_{c_1}} \sum_{y \in D_{c_2}} d(x, z)$$

5.2.4 层次聚类

层次聚类假设类别之间存在层次结构，将样本聚到层次化的类中.
它分为两种:

- 聚合聚类 (agglomerative clustering)
- 分裂聚类 (divisive clustering)

假设距离为 Euclid 距离，并用最短距离衡量类间差距，
以类间差距最小化为合并准则，以所有样本聚为一类为停止条件.
(聚合聚类算法，统计学习方法, 算法 14.1)

- ① 基于 n 个样本 $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ 构造 n 个类，每一类只包含一个样本.
- ② 计算类间距离
- ③ 合并类间距离最小的两个类，构建一个新类
  若类的总数为 1，则终止迭代; 否则移步 ②

上述算法的复杂度为 $O(n^3 d)$ (其中 d 为特征空间的维度)

例 14.1 给定 5 个样本的集合，样本之间的欧氏距离由如下矩阵 D 表示：

$$D=[d_{ij}]_{5\times 5}=\begin{bmatrix}0&7&2&9&3\\7&0&5&4&6\\2&5&0&8&1\\9&4&8&0&5\\3&6&1&5&0\end{bmatrix}$$

其中 d_{ij} 表示第 i 个样本与第 j 个样本之间的欧氏距离。显然 D 为对称矩阵。应用聚合层次聚类法对这 5 个样本进行聚类。

解 (1) 首先用 5 个样本构建 5 个类，$G_i=\{x_i\}$，$i=1,2,\cdots,5$，这样，样本之间的距离也就变成类之间的距离，所以 5 个类之间的距离矩阵亦为 D。

(2) 由矩阵 D 可以看出，$D_{35}=D_{53}=1$ 为最小，所以把 G_3 和 G_5 合并为一个新类，记作 $G_6=\{x_3,x_5\}$。

(3) 计算 G_6 与 G_1，G_2，G_4 之间的最短距离，有

$$D_{61}=2,\quad D_{62}=5,\quad D_{64}=5$$

又注意到其余两类之间的距离是

$$D_{12}=7,\quad D_{14}=9,\quad D_{24}=4$$

显然，$D_{61}=2$ 最小，所以将 G_1 与 G_6 合并成一个新类，记作 $G_7=\{x_1,x_3,x_5\}$。

(4) 计算 G_7 与 G_2，G_4 之间的最短距离，

$$D_{72}=5,\quad D_{74}=5$$

又注意到

$$D_{24}=4$$

显然，其中 $D_{24}=4$ 最小，所以将 G_2 与 G_4 合并成一新类，记作 $G_8=\{x_2,x_4\}$。

(5) 将 G_7 与 G_8 合并成一个新类，记作 $G_9=\{x_1,x_2,x_3,x_4,x_5\}$，即将全部样本聚成 1 类，聚类终止。 ■

上述层次聚类过程可以用下面的层次聚类图表示。

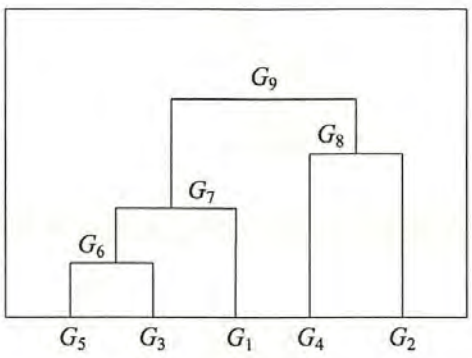


图 14.2 层次聚类图

5.2.5 K-均值聚类

给定样本 $x^{(1)}, \dots, x^{(n)}$
记 $\pi = [\pi(1), \dots, \pi(K)]$ 为 $\{1, \dots, n\}$ 的一个 K 类划分
在划分 π 下，我们记第 k 类样本构成的集合为 $D_k := \{x^{(i)} : i \in \pi(k)\}\ (k = 1, \dots, K)$
记类中心为 $\mu(D_k) := \frac{1}{|D_k|} \sum_{x \in D_k} x\ (\forall\ k = 1, \dots, K)$
则我们可以定义损失函数为样本与其所属类别的类中心的 Euclid 距离平方之和:

$$\mathrm{Loss}(\pi) := \sum_{k=1}^{K} \sum_{i \in \pi(k)} \|x^{(i)} - \mu(D_K)\|^2$$

K-均值聚类的目标就是求解 $\{1, \dots, n\}$ 的最优 K 类划分 $\pi_\star$ 以最小化损失函数:

$$\begin{aligned} \pi_\star &:= \arg\min_{\pi} \mathrm{Loss}(\pi) \\ &= \arg\min_{\pi} \sum_{k=1}^{K} \sum_{i \in \pi(k)} \|x^{(i)} - \mu(D_K)\|^2 \end{aligned}$$

但上述最优化问题是一个组合优化问题
其中 $\{1, \dots, n\}$ 的 K 类划分 π 的可能取法有 $\frac{1}{K!} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} K^n$ 种 (关于 n 是指数级的)
因此 K-均值聚类的最优化问题是一个 NP-hard 问题.
在实际应用中，我们只能使用迭代法求解.

(K-均值聚类算法, 统计学习方法, 算法 14.2)
给定样本 $x^{(1)}, \dots, x^{(n)}$

- ① 初始化: 规定一组初始均值 $\mu_1, \dots, \mu_K$ (可以是随机选取 K 个样本点)
- ② 将每个样本分配给最接近的均值对应的聚类集合:
  遍历 $i = 1, \dots, n$
  样本 $x^{(i)}$ 隶属的类别序号为 $\arg\min_{k \in \{1, \dots, K\}} \|x^{(i)} - \mu_k\|^2$，分配给对应的聚类集合.
  最终得到聚类集合 $D_1, \dots, D_K$
- ③ 更新聚类均值:

$$\mu_k = \mathrm{average}(D_k) = \frac{1}{|D_k|} \sum_{x \in D_k} x\ (\forall\ k = 1, \dots, K)$$

- ④ 收敛检查:
  计算当前步骤和前几步中均值的残差的 Euclid 范数.
  若低于某个预设定的阈值，则停止迭代; 否则返回步骤 ②

K-均值聚类算法仅能保证收敛到某个局部最优解，而不能保证收敛到全局最优解.
选取不同的初始聚类中心，会得到不同的聚类结果，实际使用时可能需要多次实验取最优结果.

K-均值聚类算法的类别数 K 需要预先指定，而实际应用中最优的类别数 K 是未知的.
不同的 K 对应的聚类结果的质量可以用类别的平均直径来衡量.
当类别数 K 超过某个值而平均直径改变很小时，这个值就是最优的类别数.
我们可以基于二分法来快速选取近似最优的类别数 K

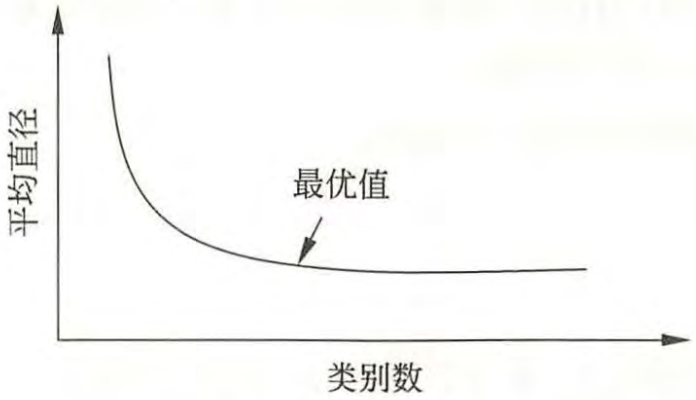


图 14.3　类别数与平均直径的关系

K-均值聚类算法的稳健性很差，离群点 (outliers) 可能对聚类中心有显著影响.
我们可以将其改进为 K-中心点 (medoids) 聚类算法，使用中心点来代替均值:
(K-中心点聚类算法)
给定样本 $x^{(1)}, \ldots, x^{(n)}$

- ① 初始化: 随机选取 K 个样本点作为中心点 $m_1, \ldots, m_K$
- ② 将每个样本分配给最接近的中心点对应的聚类集合:
  遍历 $i = 1, \ldots, n$
  样本 $x^{(i)}$ 隶属的类别序号为 $\arg\min_{k \in \{1, \ldots, K\}} \|x^{(i)} - m_k\|^2$，分配给对应的聚类集合.
  最终得到聚类集合 $D_1, \ldots, D_K$
- ③ 更新聚类中心点，即在每个聚类 D_k 中选择一个样本点作为新的中心点:

$$m_k = \text{medoid}(D_k) = \arg\min_{x \in D_k} \left\{ \sum_{y \in D_k} \|y - x\|^2 \right\} \quad (k = 1, \ldots, K)$$

- ④ 收敛检查:
  若新的中心点和旧的中心点完全相同，则停止迭代; 否则返回步骤 ②

上述算法的计算复杂度要高于 K-均值算法，但稳健性更高.

K-均值算法只能处理形如超球体的聚类，而处理不了非凸的聚类.
我们可以将其改进为**谱聚类** (spectral clustering)
给定样本 $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$

- ① 使用 Gauss 径向基距离计算邻接矩阵 (adjacency matrix):

$$W := [w_{ij}]_{i,j=1}^n = \left[\exp\left(-\frac{\|x^{(i)} - x^{(j)}\|}{2\sigma^2} \right) \right]_{i,j=1}^n$$

- ② 计算度矩阵 (degree matrix):

$$D := \text{diag}\{d_1, \ldots, d_n\}$$
$$\text{where } d_i := \sum_{j=1}^n w_{ij} \ (i = 1, \ldots, n)$$

- ③ 计算 Laplace 矩阵:

$$L := D - W$$

or

$$L := D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

- ④ 计算 L 的谱分解 $L = Q\Lambda Q^{\mathrm{T}}$
  其中 $Q \in \mathbb{R}^{n\times n}$ 为正交阵，而 $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$ 为对角元为实数的对角阵 (对角元升序排列)
  取前 $d_0 \ll d$ 小的特征值 $\lambda_1, \ldots, \lambda_{d_0}$ 及其特征向量 $q_1, \ldots, q_{d_0}$ (特征值越小，表示图的切割越容易)
  定义 $\tilde{Q} := [q_1, \ldots, q_{d_0}] \in \mathbb{R}^{n\times d_0}$
- ⑤ 将 $\tilde{Q}$ 的 n 个 d_0 维行向量视为样本点 $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ 的低维嵌入.
  使用 K-均值算法对它们进行聚类，并将聚类结果应用到样本点 $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ 上.

5.2.6 区别

层次聚类和 K-均值聚类的区别:

- ① 类别数
  K-均值算法要求用户在开始时就指定类别数，错误的 K 值可能会导致聚类结果不理想.
  层次聚类的最大优势之一是不需要预先指定类别数，
  算法会生成一个包含所有数据点的树状图 (Dendrogram)，它显示了各个数据点如何逐步合并成不同的簇.
  用户可以根据树状图的结构自由选择合适的聚类数目.
- ② 稳定性
  K-均值算法对初始中心的选择非常敏感，不同的初始化可能导致不同的结果.
  此外，如果数据中存在噪声或离群点，K-均值也可能产生不稳定的聚类结果.
  层次聚类不依赖于随机初始化，因此它通常较为稳定.
  无论数据点的顺序如何，层次聚类的结果在算法运行的过程中都会逐步合并数据点.
  层次聚类对于离群点的稳健性较好，因为离群点不会影响整体的合并过程，直到最终合并阶段才可能被视为一类.
- ③ 计算复杂度
  K-均值算法的时间复杂度为 $O(nKd)$，层次聚类的时间复杂度为 $O(n^2 \log(n)d)$
  因此在处理大规模数据时 K-均值算法更便宜.

5.2.7 性能评价

(1) 外部指标 (与某个参考模型进行比较)
给定 n 个样本 $x^{(1)}, \ldots, x^{(n)}$
设我们的聚类模型的类别为 $c_1, \ldots, c_K$，样本标签为 $y^{(1)}, \ldots, y^{(n)}$，
而参考模型的类别为 $c_1^\star, \ldots, c_S^\star$，样本标签为 $y_\star^{(1)}, \ldots, y_\star^{(n)}$

- ① 进行样本两类配对，定义:

$$a := \#\{(x^{(i)}, x^{(j)}) : 1 \le i < j \le n \text{ such that } \begin{cases} y^{(i)} = y^{(j)} \\ y_\star^{(i)} = y_\star^{(j)} \end{cases}\}$$

$$b := \#\{(x^{(i)}, x^{(j)}) : 1 \le i < j \le n \text{ such that } \begin{cases} y^{(i)} = y^{(j)} \\ y_\star^{(i)} \ne y_\star^{(j)} \end{cases}\}$$

$$c := \#\{(x^{(i)}, x^{(j)}) : 1 \le i < j \le n \text{ such that } \begin{cases} y^{(i)} \ne y^{(j)} \\ y_\star^{(i)} = y_\star^{(j)} \end{cases}\}$$

$$d := \#\{(x^{(i)}, x^{(j)}) : 1 \le i < j \le n \text{ such that } \begin{cases} y^{(i)} \ne y^{(j)} \\ y_\star^{(i)} \ne y_\star^{(j)} \end{cases}\}$$

其中 a, d 代表我们的聚类模型与参考模型的相同性，而 b, c 代表我们的聚类模型与参考模型的差异性

- ② 计算以下评价指标:

$$\text{Jaccard Coefficient} := \frac{a}{a+b+c}$$
$$\text{Fowlkes-Mallows Index} := \sqrt{\frac{a}{a+b} \times \frac{a}{a+c}}$$
$$\text{Rand Index} := \frac{2(a+d)}{n(n-1)}$$

上述指标越大，我们的聚类模型和参考模型的性能就越接近.

---

(2) 内部指标

- **① 类内特征:**
类均值 (类中心):

$$\mu(D_c) := \frac{1}{n_c}\sum_{i=1}^{n_c} x_i$$

样本间平均距离:

$$\text{avg}(D_c) := \frac{2}{|D_c|(|D_c|-1)} \sum_{1\le i<j\le |D_c|} d(x^{(i)}, x^{(j)})$$

类直径 (diameter): (即样本间最大距离)

$$L(D_c) := \max_{x,y\in D_c} d(x,y)$$

我们希望样本间平均距离和类直径 (样本间最大距离) 尽可能小.

- **② 类间差距:**
最短距离: (又称单连接)

$$d_{\min}(D_{c_1}, D_{c_2}) := \min\{d(x,z) : x \in D_{c_1}, z \in D_{c_2}\}$$

中心距离: (类中心之间的距离)

$$d_{\text{center}}(D_{c_1}, D_{c_2}) := d(\mu(D_{c_1}), \mu(D_{c_2}))$$
$$\text{where } \begin{cases} \mu(D_{c_1}) := \frac{1}{n_{c_1}}\sum_{i=1}^{n_{c_1}} x_i \\ \mu(D_{c_2}) := \frac{1}{n_{c_2}}\sum_{i=1}^{n_{c_2}} x_i \end{cases}$$

我们希望类别间最短距离和中心距离尽可能大.

- **③ 评价指标:**

$$\text{Davies-Bouldin\_Index} := \frac{1}{K}\sum_{k=1}^{K} \max_{j\neq k}\left(\frac{\text{avg}(D_k)+\text{avg}(D_j)}{d_{\text{center}}(D_k, D_j)}\right)$$
$$\text{Down\_Index} := \min_{1\le k\le K}\min_{j\neq k}\left(\frac{d_{\min}(D_k, D_j)}{\max_{1\le p\le K} L(D_p)}\right)$$

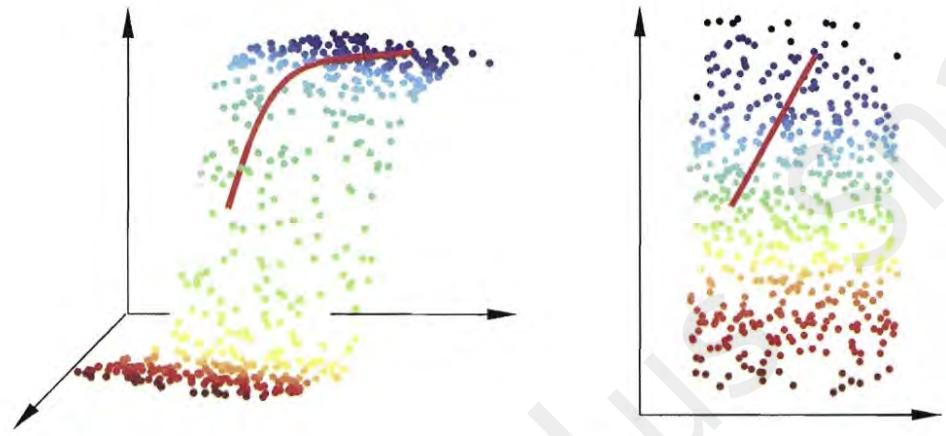
我们希望 DB 指数尽可能小 (分子分母分别代表类内差距和类间差距),
而 Down 指数尽可能大 (分子分母分别代表类间差距和类内差距)

5.3 主成分分析

5.3.1 低维嵌入

维数灾难 (curse of dimensionality)
在高维情形下出现的数据样本稀疏、距离计算困难等问题.
它是所有机器学习方法共同面临的严重障碍.

缓解维数灾难的一个重要途径是**降维** (dimension reduction)
即通过某种数学变换将原始高维属性空间转变为一个低维子空间,
在这个子空间中样本密度大幅提高，距离计算也变得更为容易.
为什么能进行降维?
这是因为在很多时候，人们观测或收集到的数据样本虽是高维的,
但与学习任务密切相关的也许仅是某个低维分布，即高维空间中的一个低维嵌入 (embedding)



(a) 三维空间中观察到的样本点　　(b) 二维空间中的曲面

图 10.2 低维嵌入示意图

主成分分析 (principal component analysis, PCA)
利用正交变换把由线性相关变量表示的观测数据转换为少数几个由线性无关变量 (称为主成分) 表示的数据.
即利用一个超平面对样本进行恰当表达 (简单地将样本投影到该超平面上):

- ① 最近重构性: 样本点到超平面的距离都足够近
- ② 最大可分性: 样本点在超平面上的投影尽可能分开

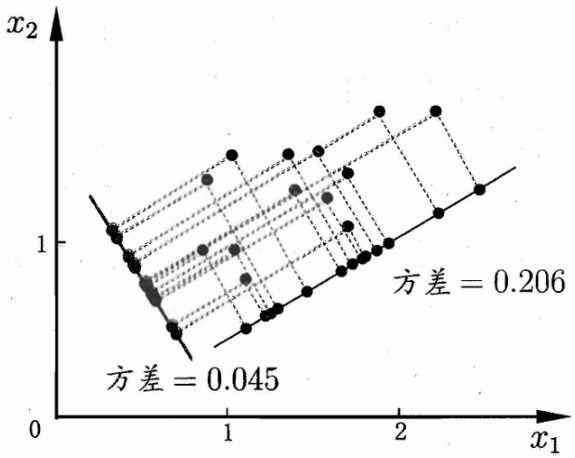


图 10.4 使所有样本的投影尽可能分开(如图中红线所示), 则需最大化投影点的方差

5.3.2 奇异值分解的视角

几何最小二乘问题 (主成分分析):
给定数据点 $y_1, \dots, y_n \in \mathbb{C}^d$ 和维度 $k \ll q := \max\{n, d\}$
要找最佳的 k 维仿射空间 $\{Ax + b : x \in \mathbb{C}^d\}$ 使得 $y_1, \dots, y_n$ 的投影方差最大化，即使得垂直部分最小化.

任意给定 k 维仿射空间 $V := \{Ax + b : x \in \mathbb{C}^k\}$
其中 $A \in \mathbb{C}^{d\times k}, b \in \mathbb{C}^d$ 且 $\operatorname{rank}(A) = \dim(\operatorname{span}(A)) = k$
记 y_i 在 $V := \{Ax + b : x \in \mathbb{C}^k\}$ 中的投影为 $\tilde{y}_i$
则依向量 b 平移后的点 $y_i - b$ 和 $\tilde{y}_i - b$ 具有如下关系:

$$\begin{gathered}\tilde{y}_i - b = P(y_i - b) = AA^\dagger(y_i - b) \\ \Leftrightarrow \\ \tilde{y}_i = AA^\dagger(y_i - b) + b\end{gathered}$$

其中 $P := AA^\dagger$ 是 $\mathbb{C}^d$ 到 $\operatorname{span}(A)$ 的正交投影算子
记 $P_\perp := I_d - AA^\dagger$
于是垂直分量的 l_2 范数和为:

$$\begin{aligned}\sum_{i=1}^n \|\tilde{y}_i - y_i\|_2^2 &= \sum_{i=1}^n \|AA^\dagger(y_i - b) + b - y_i\|_2^2 \\ &= \sum_{i=1}^n \|(I - AA^\dagger)(y_i - b)\|_2^2 \\ &= \sum_{i=1}^n \|P_\perp(y_i - b)\|_2^2\end{aligned}$$

我们的目标是最小化垂直分量的 l_2 范数和:

$$\min_{A\in\mathbb{C}^{d\times k}, b\in\mathbb{C}^d:\operatorname{rank}(A)\le k} \sum_{i=1}^n \|P_\perp(y_i - b)\|_2^2 \quad (\text{where } P_\perp := I_d - AA^\dagger)$$

目标函数对 b 求梯度可得:

$$\begin{aligned}\nabla_b \left\{\sum_{i=1}^n \|P_\perp(y_i - b)\|_2^2\right\} &= -2\sum_{i=1}^n P_\perp(y_i - b) \\ &= -2P_\perp\left(\sum_{i=1}^n y_i - nb\right)\end{aligned}$$

令 $\nabla_b \left\{\sum_{i=1}^n \|P_\perp(y_i - b)\|_2^2\right\} = -2P_\perp\left(\sum_{i=1}^n y_i - nb\right) = 0_d$ 可得 $b_\star = \frac{1}{n}\sum_{i=1}^n y_i$
这表明无论 $A \in \mathbb{C}^{d\times k}$ 如何选取，数据点 $y_1, \dots, y_n$ 的重心总是截距向量 b 的一个最优解.
(但并非唯一的最优解，它处于 $\operatorname{span}(A_\star)$ 中的分量是可以松动的，其中 $A_\star$ 是 A 的一个最优解)
将 $b = b_\star = \frac{1}{n}\sum_{i=1}^n y_i$ 代入原问题，就将其等价转换为:

$$\min_{A\in\mathbb{C}^{d\times k}:\operatorname{rank}(A)\le k} \sum_{i=1}^n \|P_\perp(y_i - b_\star)\|^2 \quad \left(\text{where } \begin{cases}P_\perp := I_d - AA^\dagger \\ b_\star = \frac{1}{n}\sum_{i=1}^n y_i\end{cases}\right)$$

记 $Z := [y_1 - b_\star, \dots, y_n - b_\star] \in \mathbb{C}^{d\times n}$
则目标函数可化简为:

$$\begin{aligned}\sum_{i=1}^{n}\|P_\perp(y_i-b_\star)\|^2 &= \|P_\perp Z\|_F^2\\ &= \|(I_d-AA^\dagger)Z\|_F^2\\ &= \|Z-AA^\dagger Z\|_F^2\end{aligned}$$

因此优化问题变为:

$$\min_{A\in\mathbb{C}^{d\times k}:\mathrm{rank}(A)\le k}\|Z-AA^\dagger Z\|_F^2$$

设 $Z\in\mathbb{C}^{d\times n}$ 的奇异值分解为 $Z:=\sum_{i=1}^{q}u_i\sigma_i v_i^H=U\Sigma V^H$
其中 $q:=\max\{n,d\}$，奇异值按非增次序排列: $\sigma_1\ge\cdots\ge\sigma_q$，对角阵 $\Sigma:=\mathrm{diag}\{\sigma_1,\ldots,\sigma_q\}$
而 $U:=[u_1,\ldots,u_q]\in\mathbb{C}^{d\times q}$ 和 $V:=[v_1,\ldots,v_q]\in\mathbb{C}^{d\times q}$ 的列向量组标准正交.
假设 $\sigma_k>\sigma_{k+1}$，则根据 **Eckart-Young 定理**可知上述优化问题的最优解 $A_\star$ 必然满足:

$$A_\star A_\star^\dagger Z=\sum_{i=1}^{k}u_i\sigma_i v_i^H$$

因此 $A_\star$ 只需保证 $A_\star A_\star^\dagger=\sum_{i=1}^{k}u_k u_k^H=U_kU_k^H$ 即可:
(其中 $U_k:=[u_1,\ldots,u_k]\in\mathbb{C}^{d\times k}$)

$$\begin{aligned}(A_\star A_\star^\dagger)Z &= \left(\sum_{i=1}^{k}u_k u_k^H\right)\left(\sum_{i=1}^{q}u_i\sigma_i v_i^H\right)\\ &= \sum_{i=1}^{k}u_k\left(\sum_{i=1}^{q}(u_k^Hu_i)\sigma_i v_i^H\right)\quad\left(\text{note that } u_k^Hu_i=\delta_{k,i}=\begin{cases}1 & \text{if } i=k\\ 0 & \text{otherwise}\end{cases}\right)\\ &= \sum_{i=1}^{k}u_k\sigma_k v_k^H\end{aligned}$$

$A_\star$ 的最简单取法便是 $A_\star=U_k=[u_1,\ldots,u_k]\in\mathbb{C}^{d\times k}$

5.3.3 谱分解的视角

设 $x=[x_1,\ldots,x_d]^{\mathrm{T}}$ 是 d 维随机变量，均值向量和协方差矩阵为:

$$\mu:=\mathrm{E}[x]=[\mu_1,\ldots,\mu_d]^{\mathrm{T}}$$
$$\Sigma:=\mathrm{Cov}(x)=\mathrm{E}[(x-\mu)(x-\mu)^{\mathrm{T}}]$$

> 实际问题中，为消除量纲的影响，我们通常对 x 的各个分量进行标准化:
>
> $$z_i=\frac{x_i-\mathrm{E}[x_i]}{\sqrt{\mathrm{Var}(x_i)}}\quad(i=1,\ldots,d)$$
>
> 此时的协方差矩阵 $\mathrm{Cov}(z)$ 就是原变量 x 的相关矩阵.

设协方差矩阵 Σ 的谱分解为 $\Sigma=U\Lambda U^{\mathrm{T}}$
其中 $U=[u_1,\ldots,u_d]\in\mathbb{R}^{d\times d}$ 为正交矩阵，$\Lambda=\mathrm{diag}\{\lambda_1,\ldots,\lambda_d\}$ (其中 $\lambda_1\ge\cdots\ge\lambda_d$)
记 $y=U^{\mathrm{T}}x=[y_1,\ldots,y_d]^{\mathrm{T}}$，我们分量 $y_1,\ldots,y_d$ 为第 $1,\ldots,d$ 个**总体主成分**.
其性质如下:

- ① 协方差矩阵为对角阵:

$$
\begin{aligned}
\mathrm{Cov}(y) &= \mathrm{E}[(y-U^{\mathrm{T}}\mu)(y-U^{\mathrm{T}}\mu)^{\mathrm{T}}] \quad (\text{note that } \mathrm{E}[y]=\mathrm{E}[U^{\mathrm{T}}x]=U^{\mathrm{T}}\mu)\\
&= \mathrm{E}[(U^{\mathrm{T}}x-U^{\mathrm{T}}\mu)(U^{\mathrm{T}}x-U^{\mathrm{T}}\mu)^{\mathrm{T}}]\\
&= U^{\mathrm{T}}\mathrm{E}[(x-\mu)(x-\mu)^{\mathrm{T}}]U\\
&= U^{\mathrm{T}}\cdot\mathrm{Cov}(x)\cdot U\\
&= U^{\mathrm{T}}\Sigma U\\
&= \Lambda
\end{aligned}
$$

- ② 方差之和不变:

$$
\begin{aligned}
\sum_{i=1}^{d}\mathrm{Var}(y_i) &= \mathrm{tr}\left(\mathrm{Cov}(y)\right)\\
&= \mathrm{tr}\left(\Lambda\right)=\sum_{i=1}^{d}\lambda_i\\
&= \mathrm{tr}\left(\Lambda UU^{\mathrm{T}}\right)\\
&= \mathrm{tr}\left(U\Lambda U^{\mathrm{T}}\right)\\
&= \mathrm{tr}\left(\Sigma\right)\\
&= \mathrm{tr}\left(\mathrm{Cov}(x)\right)\\
&= \sum_{i=1}^{d}\mathrm{Var}(x_i)
\end{aligned}
$$

- ③ 因子负荷量 (factor loading)
  即第 j 个总体主成分 y_j 和变量 x_i 的相关系数为 $\rho(x_i,y_j)$

$$
\begin{aligned}
\rho(x_i,y_j) &= \frac{\mathrm{Cov}(x_i,y_j)}{\sqrt{\mathrm{Var}(x_i)}\sqrt{\mathrm{Var}(y_j)}}\\
&= \frac{\mathrm{Cov}(e_i^{\mathrm{T}}x,u_j^{\mathrm{T}}x)}{\sqrt{\sigma_{ii}}\sqrt{\lambda_j}} \quad \left(\text{note that } \begin{cases}\mathrm{Cov}(x)=\Sigma=[\sigma_{ij}]_{i,j=1}^{d}\\ \mathrm{Cov}(y)=\Lambda=\mathrm{diag}\{\lambda_1,\ldots,\lambda_d\}\end{cases}\right)\\
&= \frac{e_i^{\mathrm{T}}\Sigma u_j}{\sqrt{\sigma_{ii}}\sqrt{\lambda_j}} \quad (\text{note that } \Sigma u_j=u_j\lambda_j)\\
&= \frac{e_i^{\mathrm{T}}u_j\lambda_j}{\sqrt{\lambda_j}\sqrt{\sigma_{ii}}}\\
&= \frac{\sqrt{\lambda_j}u_{ij}}{\sqrt{\sigma_{ii}}} \quad (\text{note that } U=[u_1,\ldots,u_j]=[u_{ij}]_{i,j=1}^{d})
\end{aligned}
$$

- ④ 第 j 个总体主成分 y_j 和所有变量 $x_1,\ldots,x_d$ 的相关系数满足:

$$
\begin{aligned}
\sum_{i=1}^{d}\sigma_{ii}\rho^2(x_i,y_j) &= \sum_{i=1}^{d}\sigma_{ii}\left(\frac{\sqrt{\lambda_j}u_{ij}}{\sqrt{\sigma_{ii}}}\right)^2\\
&= \lambda_j\sum_{i=1}^{d}u_{ij}^2\\
&= \lambda_j\|u_j\|^2
\end{aligned}
$$

- ⑤ 第 i 个变量 x_i 和所有总体主成分 $y_1,\ldots,y_d$ 的相关系数满足:

$$
\begin{aligned}
\rho^2(x_i,(y_1,\dots,y_d)) &= \sum_{j=1}^{d}\rho^2(x_i,y_j) \quad (\text{note that } y_1,\dots,y_d \text{ are uncorrelated}) \\
&= \sum_{j=1}^{d}\left(\frac{\sqrt{\lambda_j}u_{ij}}{\sqrt{\sigma_{ii}}}\right)^2 \\
&= \frac{1}{\sigma_{ii}}\sum_{j=1}^{d}\lambda_j u_{ij}^2 \\
&= \frac{1}{\sigma_{ii}}(e_i^{\mathrm{T}}U)\Lambda(e_i^{\mathrm{T}}U)^{\mathrm{T}} \quad \left(\text{note that } \begin{cases} U=[u_1,\dots,u_d]=[u_{ij}]_{i,j=1}^{d} \\ e_i^{\mathrm{T}}U=[u_{i1},\dots,u_{id}] \\ \Lambda=\mathrm{diag}\{\lambda_1,\dots,\lambda_d\} \end{cases}\right) \\
&= \frac{1}{\sigma_{ii}}e_i^{\mathrm{T}}(U\Lambda U^{\mathrm{T}})e_i \\
&= \frac{1}{\sigma_{ii}}e_i^{\mathrm{T}}\Sigma e_i \quad (\text{note that } \Sigma=[\sigma_{ij}]_{i,j=1}^{d}) \\
&= \frac{1}{\sigma_{ii}}\sigma_{ii} \\
&= 1
\end{aligned}
$$

我们定义第 i 个总体主成分 y_i 的方差贡献率为:

$$
\eta(y_i) := \frac{\mathrm{Var}(y_i)}{\sum_{j=1}^{d}\mathrm{Var}(y_j)} = \frac{\lambda_i}{\sum_{j=1}^{d}\lambda_j} \quad (i=1,\dots,d)
$$

我们通常取 k 使得累积方差贡献率 $\sum_{i=1}^{k}\eta(y_i)=\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{d}\lambda_i}$ 达到 $70\%\sim80\%$ 以上.

我们定义前 k 个总体主成分 $y_1,\dots,y_k$ 对变量 x_i 的贡献率为:

$$
\begin{aligned}
\rho^2(x_i,(y_1,\dots,y_k)) &= \sum_{j=1}^{k}\rho^2(x_i,y_j) \\
&= \sum_{j=1}^{k}\left(\frac{\sqrt{\lambda_j}u_{ij}}{\sqrt{\sigma_{ii}}}\right)^2 \\
&= \frac{1}{\sigma_{ii}}\sum_{j=1}^{k}\lambda_j u_{ij}^2
\end{aligned}
$$

5.3.4 算法

设对随机变量 $x\in\mathbb{R}^d$ 进行 n 次独立观测得到样本矩阵 $X:=[x_1,\dots,x_n]^{\mathrm{T}}\in\mathbb{R}^{n\times d}$

定义样本均值向量、样本协方差矩阵和样本相关系数矩阵为:

$$
\begin{aligned}
\hat{\mu} &:= \frac{1}{n} X^{\mathrm{T}} 1_n \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i \in \mathbb{R}^d
\end{aligned}
$$

$$
\begin{aligned}
\hat{\Sigma} &:= \frac{1}{n-1} X^{\mathrm{T}} (I_n - \frac{1}{n} 1_n 1_n^{\mathrm{T}}) X \\
&= \frac{1}{n-1} (X^{\mathrm{T}} - \frac{1}{n} X^{\mathrm{T}} 1_n 1_n^{\mathrm{T}})(X^{\mathrm{T}} - \frac{1}{n} X^{\mathrm{T}} 1_n 1_n^{\mathrm{T}})^{\mathrm{T}} \\
&= \frac{1}{n-1} (X^{\mathrm{T}} - \hat{\mu} 1_n^{\mathrm{T}})(X^{\mathrm{T}} - \hat{\mu} 1_n^{\mathrm{T}})^{\mathrm{T}} \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^{\mathrm{T}} \\
&= [\hat{\sigma}_{ij}]_{i,j=1}^{d} \in \mathbb{R}^{d \times d}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\mathrm{P}} &:= (\mathrm{diag}(\hat{\Sigma}))^{-\frac{1}{2}} \hat{\Sigma} (\mathrm{diag}(\hat{\Sigma}))^{-\frac{1}{2}} \\
&= \left[\hat{\rho}_{ij} := \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii} \hat{\sigma}_{jj}}} \right]_{i,j=1}^{d} \in \mathbb{R}^{d \times d}
\end{aligned}
$$

若做变换 $\tilde{X} := (X - 1_n \hat{\mu}^{\mathrm{T}})(\mathrm{diag}(\hat{\Sigma}))^{-\frac{1}{2}} \in \mathbb{R}^{n \times d}$ 将观测标准化，
则我们有 $\hat{\mathrm{P}} = \frac{1}{n-1} \tilde{X}^{\mathrm{T}} \tilde{X}$ 成立.

记 $q = \min\{n, d\}$
设 $\tilde{X} \in \mathbb{R}^{n \times d}$ 的奇异值分解为 $\tilde{X} = U \Sigma V^{\mathrm{T}} = \sum_{i=1}^{q} u_i \sigma_i v_i^{\mathrm{T}}$
其中 $U \in \mathbb{R}^{n \times q}$ 和 $V \in \mathbb{R}^{d \times q}$ 列标准正交，对角阵 $\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_q\}$ 满足 $\sigma_1 \geq \cdots \geq \sigma_q$

给定维度 $k \ll d$，记 $\begin{cases} U_k := [u_1, \ldots, u_k] \in \mathbb{R}^{n \times k} \\ V_k := [v_1, \ldots, v_k] \in \mathbb{R}^{d \times k} \\ \Sigma_k := \mathrm{diag}\{\sigma_1, \ldots, \sigma_k\} \end{cases}$
定义样本主成分为:

$$
\begin{aligned}
Z_k &:= \tilde{X} V_k \\
&= \left(\sum_{i=1}^{q} u_i \sigma_i v_i^{\mathrm{T}} \right) [v_1, \ldots, v_k] \\
&= [u_1 \sigma_1, \ldots, u_k \sigma_k] \\
&= U_k \Sigma_k \in \mathbb{R}^{n \times k}
\end{aligned}
$$

记 $\tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_d] \in \mathbb{R}^{n \times d}$
我们希望寻找 $B = [b_1, \ldots, b_d] \in \mathbb{R}^{k \times d}$ 使得 $\|\tilde{X} - Z_k B\|_F$ 达到最小.
即要求解:

$$
\min_{B \in \mathbb{R}^{k \times d}} \|\tilde{X} - Z_k B\|_F^2 := \sum_{i=1}^{d} \|\tilde{x}_i - Z_k b_i\|_2^2
$$

对于任意 $i = 1, \ldots, d$，考虑求解:

$$
\min_{b_i \in \mathbb{R}^k} \|\tilde{x}_i - Z_k b_i\|_2^2
$$

对应的最小二乘解为 $b_i := (Z_k^{\mathrm{T}} Z_k)^{-1} Z_k^{\mathrm{T}} \tilde{x}_i \ (i = 1, \ldots, d)$
因此我们有:

$$
\begin{aligned}
B &= [b_1, \dots, b_d] \\
&= [(Z_k^{\mathrm{T}} Z_k)^{-1} Z_k^{\mathrm{T}} \tilde{x}_1, \dots, (Z_k^{\mathrm{T}} Z_k)^{-1} Z_k^{\mathrm{T}} \tilde{x}_d] \\
&= (Z_k^{\mathrm{T}} Z_k)^{-1} Z_k^{\mathrm{T}} [\tilde{x}_1, \dots, \tilde{x}_d] \\
&= (Z_k^{\mathrm{T}} Z_k)^{-1} Z_k^{\mathrm{T}} \tilde{X} \\
&= [(U_k \Sigma_k)^{\mathrm{T}} (U_k \Sigma_k)]^{-1} (U_k \Sigma_k)^{\mathrm{T}} U \Sigma V^{\mathrm{T}} \\
&= (\Sigma_k)^{-2} \Sigma_k U_k^{\mathrm{T}} \left(\sum_{i=1}^{q} u_i \sigma_i v_i^{\mathrm{T}} \right) \\
&= \Sigma_k^{-1} \begin{bmatrix} u_1^{\mathrm{T}} \\ \vdots \\ u_k^{\mathrm{T}} \end{bmatrix} \left(\sum_{i=1}^{q} u_i \sigma_i v_i^{\mathrm{T}} \right) \\
&= \begin{bmatrix} \sigma_1^{-1} & & \\ & \ddots & \\ & & \sigma_k^{-1} \end{bmatrix} \begin{bmatrix} \sigma_1 v_1^{\mathrm{T}} \\ \vdots \\ \sigma_k v_k^{\mathrm{T}} \end{bmatrix} \\
&= \begin{bmatrix} v_1^{\mathrm{T}} \\ \vdots \\ v_k^{\mathrm{T}} \end{bmatrix} \\
&= V_k^{\mathrm{T}}
\end{aligned}
$$

这样我们就找到了 $B \in \mathbb{R}^{k \times d}$ 的一个解 $B = V_k^{\mathrm{T}}$
它使得 $\|\tilde{X} - Z_k B\|_F$ 最小化 (其中 $Z_k = U_k \Sigma_k \in \mathbb{R}^{n \times k}$ 是样本主成分)

(主成分分析, 统计学习方法, 算法 16.1)
给定 n 个样本观测 $x_1, \dots, x_n \in \mathbb{R}^d$ 和降维空间的维数 $k \ll d$

- ① 构造并标准化样本矩阵:

$$
\begin{aligned}
X &:= [x_1, \dots, x_n]^{\mathrm{T}} \in \mathbb{R}^{n \times d} \\
\mu &:= \frac{1}{n} X^{\mathrm{T}} 1_n \in \mathbb{R}^d \\
\Sigma &:= \frac{1}{n-1} (X - 1_n \mu^{\mathrm{T}})^{\mathrm{T}} (X - 1_n \mu^{\mathrm{T}}) \in \mathbb{R}^{d \times d} \\
\tilde{X} &:= (X - 1_n \mu^{\mathrm{T}}) (\mathrm{diag}(\hat{\Sigma}))^{-\frac{1}{2}} \in \mathbb{R}^{n \times d}
\end{aligned}
$$

- ② 对标准化的样本矩阵 $\tilde{X}$ 做 k 阶截断的奇异值分解:

$$
\sum_{i=1}^{k} u_i \sigma_i v_i^{\mathrm{T}} = U_k \Sigma_k V_k^{\mathrm{T}} \text{ where } \begin{cases} U_k := [u_1, \dots, u_k] \in \mathbb{R}^{n \times k} \\ V_k := [v_1, \dots, v_k] \in \mathbb{R}^{d \times k} \\ \Sigma_k := \mathrm{diag}\{\sigma_1, \dots, \sigma_k\} \end{cases}
$$

  并得到样本主成分矩阵 $Z_k := U_k \Sigma_k \in \mathbb{R}^{n \times k}$，完成数据降维.

5.4 因子分析

5.4.1 正交因子分析

正交因子分析 (Orthogonal Factor Analysis) 是通过寻找较少的潜在变量 (即公共因子 F)
来解释原始变量之间的相关性的一种降维方法.

设 $X = [X_1, \dots, X_p]^{\mathrm{T}} \in \mathbb{R}^p$，假设:

$$
\begin{aligned}
X_1 - \mu_1 &= a_{11}F_1 + \cdots + a_{1q}F_q + \varepsilon_1 \\
&\cdots \\
X_p - \mu_p &= a_{p1}F_1 + \cdots + a_{pq}F_q + \varepsilon_p \\
&\Leftrightarrow \\
X - \mu &= AF + \varepsilon
\end{aligned}
$$

其中 $q \ll p$ (降维)，$\mu \in \mathbb{R}^p$ 是均值向量
$A \in \mathbb{R}^{p\times q}$ 称为**载荷矩阵** (loading matrix)，$F \in \mathbb{R}^q$ 称为**公共因子**，$\varepsilon \in \mathbb{R}^p$ 称为**特殊因子**.
公共因子不可观测.
为识别它们，做如下假定:

- ① $\mathrm{E}[F] = 0_q, \mathrm{Cov}[F] = I_q$ (不同因子之间不存在相关性)
- ② $\mathrm{E}[\varepsilon] = 0_p, \mathrm{Cov}[\varepsilon] = \Phi = \mathrm{diag}\{\phi_1^2, \ldots, \phi_p^2\}$ (不同特殊因子之间不存在相关性)
- ③ $\mathrm{Cov}(F, \varepsilon) = [\mathrm{Cov}(F_i, \varepsilon_j)] = 0_{q\times p}$ (因子与特殊因子之间不存在相关性)

根据上述假设我们有:

- $\mathrm{Var}(X_i) = \mathrm{Cov}(X_i, X_i) = \|a_i\|_2^2 + \phi_i^2$
  其中 $a_i \in \mathbb{R}^q$ 是载荷矩阵 $A \in \mathbb{R}^{p\times q}$ 的第 i 个行向量
  我们称 $\|a_i\|_2^2$ 是 X_i 的**共性方差**，即 X_i 的方差中由公共因子解释的部分.
  而 ϕ_i^2 称为 X_i 的**特殊方差**，即 X_i 的方差中由特殊因子解释的部分.
- 记 $\Sigma := \mathrm{Cov}(X) = \mathrm{E}[(X-\mu)(X-\mu)^{\mathrm{T}}]$，则我们有:

  $$
  \begin{aligned}
  \Sigma &= \mathrm{Cov}(X) \\
  &= [\mathrm{Cov}(X_i, X_j)]_{i,j=1}^p \\
  &= [a_i^{\mathrm{T}} a_j + \phi_{ii}\delta_{i,j}]_{i,j=1}^p \\
  &= AA^{\mathrm{T}} + \Phi
  \end{aligned}
  $$

  其中 $\delta_{i,j} = \mathbb{1}\{i = j\}$ 为 Kronecker 记号.
  这说明特殊因子仅在方差中有贡献，对协方差没有贡献.

我们通过载荷矩阵 A 的系数 a_{ij} 来解释公共因子 F_i.

- 可以通过对绝对值较大的载荷系数来解释 (载荷系数的正负本身没有意义)
  例如学生成绩预测中的理科因子 (与理科强相关) 和文科因子 (与文科强相关) 等等
- 载荷系数的正负对比有意义

5.4.2 模型估计

根据模型假设我们有:

$$
\Sigma = \mathrm{Cov}(X) = AA^{\mathrm{T}} + \Phi
$$

其中 $\Sigma \in \mathbb{R}^{p\times p}, A \in \mathbb{R}^{p\times q}$，而 $\Phi = \mathrm{diag}\{\phi_1, \ldots, \phi_p\}$
若直接对 Σ 进行估计，则参数量为 $\frac{1}{2}p(p+1)$
而通过对 A 和 Φ 进行估计，则参数量减少为 $pq + p = p(q+1)$ (其中 $q \ll p$)
这也体现了降维的思想.

载荷矩阵 $A \in \mathbb{R}^{p\times q}$ 有无穷多个解.
给定载荷矩阵 $A \in \mathbb{R}^{p\times q}$，则对于任意正交矩阵 $Q \in \mathbb{R}^{q\times q}$，$AQ$ 都能作为载荷矩阵.
因此可以先得到载荷矩阵 A 的一个初始估计，再经过作用正交变换来得到更好的解释.

下面考虑如何得到 A 的一个初始估计:

- **① 主成分法**
  设对随机变量 $x \in \mathbb{R}^p$ 进行 n 次独立观测得到样本矩阵 $X := [x_1, \dots, x_n]^{\mathrm{T}} \in \mathbb{R}^{n \times d}$
  定义样本均值向量和样本协方差矩阵为:

$$\begin{aligned}
\hat{\mu} &:= \frac{1}{n} X^{\mathrm{T}} 1_n \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i \in \mathbb{R}^d \\
\hline
\hat{\Sigma} &:= \frac{1}{n-1} X^{\mathrm{T}} (I_n - \frac{1}{n} 1_n 1_n^{\mathrm{T}}) X \\
&= \frac{1}{n-1} (X^{\mathrm{T}} - \frac{1}{n} X^{\mathrm{T}} 1_n 1_n^{\mathrm{T}})(X^{\mathrm{T}} - \frac{1}{n} X^{\mathrm{T}} 1_n 1_n^{\mathrm{T}})^{\mathrm{T}} \\
&= \frac{1}{n-1} (X^{\mathrm{T}} - \hat{\mu} 1_n^{\mathrm{T}})(X^{\mathrm{T}} - \hat{\mu} 1_n^{\mathrm{T}})^{\mathrm{T}} \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^{\mathrm{T}} \\
&= [\hat{\sigma}_{ij}]_{i,j=1}^{d} \in \mathbb{R}^{d \times d}
\end{aligned}$$

  设 $\hat{\Sigma} = [\sigma_{ij}]_{i,j=1}^{p} \in \mathbb{R}^{p \times p}$ 的谱分解为:

$$\hat{\Sigma} = U \Lambda U^{\mathrm{T}} = \sum_{i=1}^{p} \lambda_i u_i u_i^{\mathrm{T}}$$

  其中 $U = [u_1, \dots, u_p] \in \mathbb{R}^{p \times p}$ 为正交矩阵，$\Lambda = \mathrm{diag}\{\lambda_1, \dots, \lambda_p\}$
  (特征值按非增次序排列, 与惯用的约定方式相反)
  我们记 $\hat{A} := [\sqrt{\lambda_1} u_1, \dots, \sqrt{\lambda_q} u_q] \in \mathbb{R}^{p \times q}$ (截断前 q 阶奇异值)
  此时 $\hat{A}\hat{A}^{\mathrm{T}} = \sum_{i=1}^{q} \lambda_i u_i u_i^{\mathrm{T}}$ 是 $\hat{\Sigma} = U \Lambda U^{\mathrm{T}} = \sum_{i=1}^{p} \lambda_i u_i u_i^{\mathrm{T}}$ 的秩不超过 q 的最佳逼近.

  记 $\hat{A}$ 的行向量为 $\hat{a}_1, \dots, \hat{a}_p \in \mathbb{R}^q$，则特殊因子的方差的估计为:

$$\hat{\phi}_i^2 := \hat{\sigma}_{ii} - \|\hat{a}_i\|_2^2 \quad (i = 1, \dots, p)$$

- **② 最大似然估计**
  假定 F 和 ε 服从多元正态分布:

$$\begin{aligned}
F &= [F_1, \dots, F_q]^{\mathrm{T}} \sim N(0_q, I_q) \\
\varepsilon &= [\varepsilon_1, \dots, \varepsilon_p]^{\mathrm{T}} \sim N(0_p, \Phi) \text{ where } \Phi := \mathrm{diag}\{\phi_1^2, \dots, \phi_p^2\}
\end{aligned}$$

  于是 $X = AF + \varepsilon$ 也服从多元正态分布:

$$X = [X_1, \dots, X_p]^{\mathrm{T}} \sim N(\mu, \Sigma) \quad (\text{where } \Sigma = AA^{\mathrm{T}} + \Phi)$$

  给定独立样本 $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^p$，其对数似然函数为:

$$\mathcal{L}(\mu, A, \Phi) = \log \left\{ \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}} (\det(AA^{\mathrm{T}} + \Phi))^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (X^{(i)} - \mu)^{\mathrm{T}} (AA^{\mathrm{T}} + \Phi)^{-1} (X^{(i)} - \mu) \right) \right\}$$

  由于载荷矩阵的不唯一性，故我们可限定 $A^{\mathrm{T}} \Phi^{-1} A$ 为对角阵来得到唯一解.
  可以使用 EM 算法简化求解极大似然解的过程.

5.4.3 因子旋转

得到因子载荷矩阵 $A \in \mathbb{R}^{p\times q}$ 的初始估计 $\hat{A}$ 后，可以进行因子旋转提高其可解释性.
因子旋转的目标为: **("稀疏" 编码)**

- ① 对于任意因子 F_i $(i = 1, \ldots, q)$ 而言，
  变量 $X = [X_1, \ldots, X_p]^{\mathrm{T}}$ 只有少数分量在该因子上载荷矩阵的系数的绝对值较大，
  而其余分量在该因子上的载荷矩阵的系数接近于 0
  (换言之，$A \in \mathbb{R}^{p\times q}$ 的每一列都是 "稀疏的"，即绝对值较大的元素比较少)
- ② 对于变量 $X = [X_1, \ldots, X_p]^{\mathrm{T}}$ 的任意分量 X_i 而言，
  它只在少数因子上的载荷矩阵的系数的绝对值较大，在其余因子上的载荷矩阵的系数接近于 0
  (换言之，$A \in \mathbb{R}^{p\times q}$ 的每一行都是 "稀疏的"，即绝对值较大的元素比较少)
- ③ 任何两个因子对应的载荷呈现不同的模式，
  因而在解释时这两个因子具有不同的含义.
  (换言之，$A \in \mathbb{R}^{p\times q}$ 的不同列有着不同的 pattern)

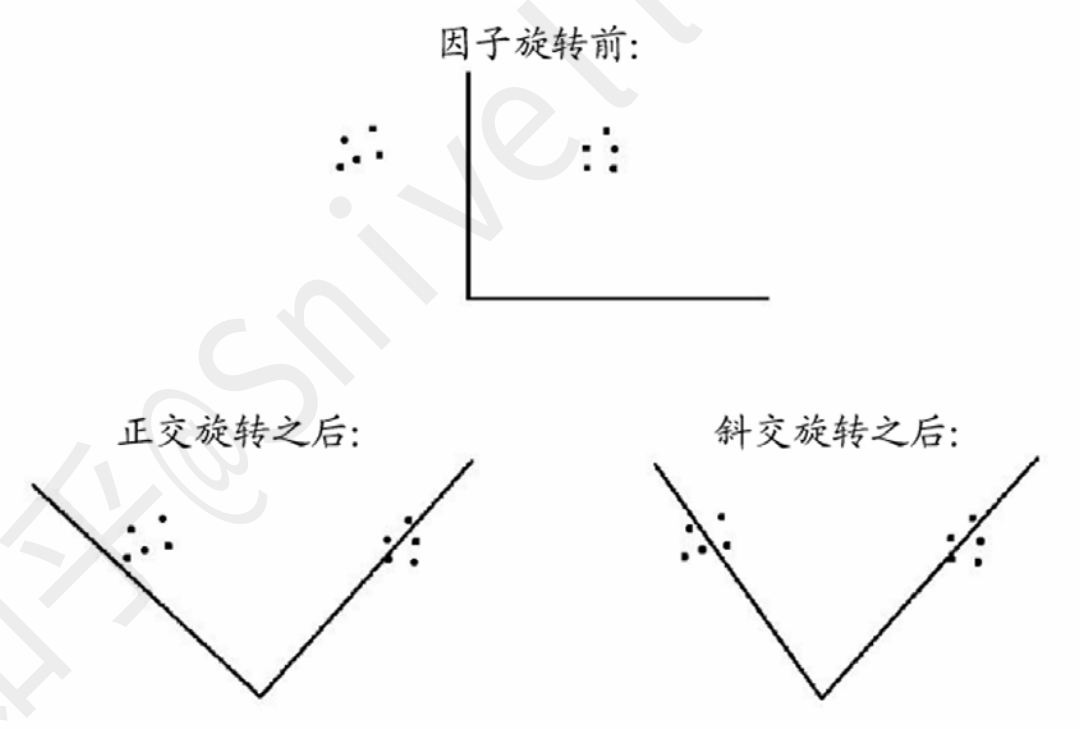
因子旋转的类型:

- ① 正交旋转: 采用正交矩阵对因子载荷矩阵进行旋转，能够保持因子之间的正交性.

$$\begin{aligned} X - \mu &= AF + \varepsilon \\ &= (AQ)(Q^H F) + \varepsilon \end{aligned}$$

  其中 $\mathrm{Cov}(Q^H F) = Q^H \mathrm{Cov}(F) Q = Q^H I_q Q = I_q$ (说明因子之间仍是不相关的，即正交的)

- ② 斜交旋转: 采用非正交矩阵对因子载荷矩阵进行旋转，
  可以更好地简化载荷矩阵，提高因子的可解释性，但旋转后的因子之间存在相关性.



应用最广泛的因子旋转方法:
最大方差旋转 (varimax rotation)
它是一种正交旋转，目的是使载荷平方方差最大化:
(少量元素绝对值大，而大多数元素绝对值小，因此我们最大化元素的绝对值 (或平方) 的方差)

$$\mathrm{avg}(A \odot A) := \frac{1}{pq} \sum_{i,j=1}^{p} a_{ij}^2$$

$$\max_{Q\in\mathbb{R}^{q\times q}: Q^{\mathrm{T}}Q = I_q} \sum_{i,j=1}^{q} [(AQ)_{[i,j]}^2 - \mathrm{avg}((AQ)\odot(AQ))]^2$$

5.4.4 探索性因子分析

因子数目 q 的选择:

- **① Kaiser 准则:**
  共性方差占总方差比例大于平均解释比: **(存疑: 行向量?)**

$$\frac{\|a_j\|_2^2}{\sum_{i=1}^p \sigma_{ii}^2} = \frac{\sum_{i=1}^p a_{ij}^2}{\sum_{i=1}^p \sigma_{ii}^2} > \frac{1}{p} \quad (\forall\, j = 1, \dots, q)$$

  其中 $A = [a_1, \dots, a_q] \in \mathbb{R}^{p\times q}$，而 $\Sigma = \mathrm{Cov}(X) = AA^{\mathrm{T}} + \Phi$

- **② 崖底碎石图 (scree plot)**
  选择斜率变化点 (朱老师称之为拐点，我并不认同) 的前一点

- **③ 假设检验:**
  若载荷矩阵由最大似然估计得到，则可以使用假设检验.

因子得分: 对公共因子 $F = [F_1, \dots, F_q]^{\mathrm{T}}$ 的估计值.
可以通过最小二乘法得到.
假设样本矩阵为 $X = [x^{(1)}, \dots, x^{(n)}] \in \mathbb{R}^{p\times n}$
公共因子矩阵 $F = [f^{(1)}, \dots, f^{(n)}] \in \mathbb{R}^{q\times n}$
特殊因子矩阵 $\mathrm{E} = [\varepsilon^{(1)}, \dots, \varepsilon^{(n)}] \in \mathbb{R}^{p\times n}$
因此因子模型 $x = Af + \varepsilon$ 的样本形式为:

$$X = AF + E$$

设载荷矩阵 $A \in \mathbb{R}^{p\times q}$ 已知，考虑以下优化问题:

$$\min_{F\in\mathbb{R}^{q\times n}} \|X - AF\|_F^2 := \sum_{i=1}^n \|x^{(i)} - Af^{(i)}\|_2^2$$

取最小二乘解 $\hat{F} := (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}X$，作为因子得分.

因子分析和主成分分析的区别:

- 因子分析有 "模型" $(x = Af + \varepsilon)$，因子不可观测，存在识别性问题
- 主成分分析解释输入变量的总变异，因子分析解释公共变异性
- 因子模型中有因子旋转的问题

其他降维方法:

- 核主成分分析 (Kernalized PCA)
- 流形学习 (Manifold Learning)

The End