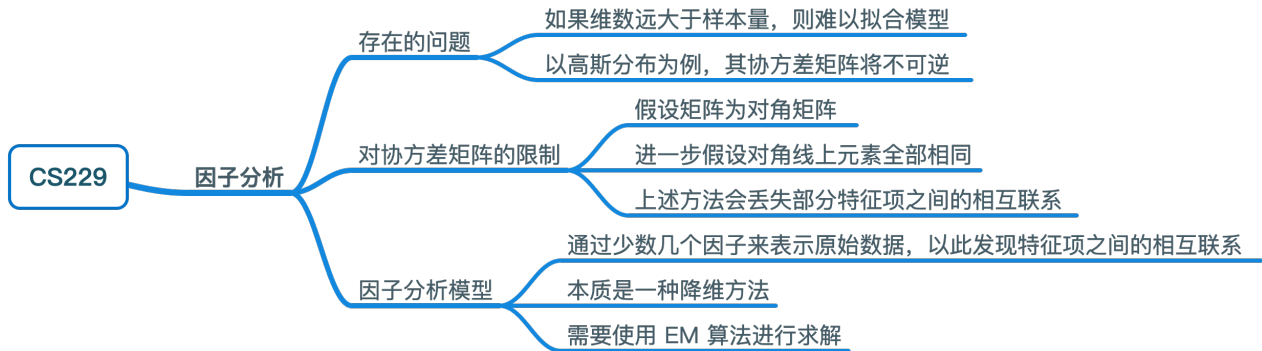


第十章 因子分析



存在的问题

- 在之前的推导中，我们通常假定拥有足够的数据来拟合模型，即 $m \gg n$ （样本量远大于维数）
- 但是如果维数远大于样本量，则难以对模型进行拟合
 - 以简单的高斯分布为例，通过极大似然法可以得到：

$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$
$$\Sigma = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

- 可以发现 Σ 是奇异矩阵，即特征值为 0，不满秩的矩阵
 - 这意味着我们无法对 Σ 求逆，且 $1/|\Sigma|^{1/2} = 0$
 - 因此我们无法写出该分布的概率密度函数，也就无法对其建模
- 可以将其理解为线性方程组求解，未知数的个数比方程数目多，因而无法完全求出所有未知数
 - 原文使用仿射空间进行解释，并不是很懂(⊙⊙⊙)
- 我们可以通过一些方法解决这个问题，在接下来的几节中：
 - 我们首先会对协方差矩阵添加两种可能的限制，来帮助求解
 - 但这些方法并不能完美解决问题
 - 之后我们会介绍高斯分布的某些性质，并提出因子分析模型及其 EM 求解

对协方差矩阵的限制

- 对协方差矩阵的限制可以分为两种
- 第一种限制是假设矩阵为对角矩阵
 - 基于该假设，最大似然估计的结果为：

$$\Sigma_{jj} = \frac{1}{m} \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

- 对二维高斯分布来说，其概率密度在平面上的投影轮廓为椭圆
- 当协方差矩阵为对角矩阵时，椭圆的轴与坐标轴平行

- 第二种限制是进一步假设对角线上的元素全部相同

- 此时 $\Sigma = \sigma^2 I$ ，其中最大似然估计表明：

$$\sigma^2 = \frac{1}{mn} \sum_{j=1}^n \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

- 此时投影轮廓为圆（高维情况下为球面或超球面）

- 在没有限制的情况下，我们需要 $m > n + 1$ 来保证 Σ 的最大似然估计不是奇异矩阵

- 在上述两个限制中的任意一个下，我们只需要 $m \geq 2$ 来保证非奇异

- 但是在上述限制下，我们会丢失部分特征项之间的相互联系

- 因子分析模型能够解决上述问题

高斯分布的边缘和条件分布

- 在介绍因子分析前，我们先介绍联合多元高斯分布的[边缘分布](#)和条件分布

- 假定我们有一个由两个变量组合而成的随机变量：

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- 其中 $x_1 \in \mathbb{R}^r$ ， $x_2 \in \mathbb{R}^s$ ，因此 $x \in \mathbb{R}^{r+s}$

- 假定 $x \sim \mathcal{N}(\mu, \Sigma)$ ，其中

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

- 这里 $\mu_1 \in \mathbb{R}^r$ ， $\mu_2 \in \mathbb{R}^s$ ， $\Sigma_{11} \in \mathbb{R}^{r \times r}$ ， $\Sigma_{12} \in \mathbb{R}^{r \times s}$ ，以此类推

- 因为协方差矩阵 Σ 是对称的，所以 $\Sigma_{12} = \Sigma_{21}^T$

- 基于上述定义，我们可以求出 x_1 的[边缘分布](#)：

$$E[x_1] = \mu_1$$

$$\text{Cov}(x_1) = E[(x_1 - \mu_1)(x_1 - \mu_1)^T] = \Sigma_{11}$$

- 关于协方差公式的证明如下：

$$\begin{aligned} \text{Cov}(x) &= \Sigma \\ &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \\ &= E[(x - \mu)(x - \mu)^T] \\ &= E \left[\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \right] \\ &= E \begin{bmatrix} (x_1 - \mu_1)(x_1 - \mu_1)^T & (x_1 - \mu_1)(x_2 - \mu_2)^T \\ (x_2 - \mu_2)(x_1 - \mu_1)^T & (x_2 - \mu_2)(x_2 - \mu_2)^T \end{bmatrix} \end{aligned}$$

- 因为边缘分布本身也是高斯分布，所以有：

$$x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$$

- 类似地，我们可以推导出条件分布 $x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$ ，其中：

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \quad (1)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \quad (2)$$

- 推导过程省略

因子分析模型

模型的提出

- 因子分析模型是指通过少数几个潜在变量（因子）来表示原始数据，以此发现特征项之间的相互联系，探求原始数据的基本结构

- 其本质是一种降维方法

- 在因子模型中，我们提出如下的联合分布 (x, z) :

$$\begin{aligned} z &\sim \mathcal{N}(0, I) \\ x|z &\sim \mathcal{N}(\mu + \Lambda z, \Psi) \end{aligned}$$

- 其中 $z \in \mathbb{R}^k$ 是潜在随机变量

- 模型的参数包括:

- 向量 $\mu \in \mathbb{R}^n$
- 矩阵 $\Lambda \in \mathbb{R}^{n \times k}$
- 对角矩阵 $\Psi \in \mathbb{R}^{n \times n}$

- k 应该小于 n

- 因此我们可以想象 $x^{(i)}$ 是通过 k 维多元高斯分布 $z^{(i)}$ 进行采样生成的

- 首先通过计算 $\mu + \Lambda z^{(i)}$ 将其映射至 k 维仿射空间 \mathbb{R}^n

- 然后通过加上协方差噪声 Ψ 生成 $x^{(i)}$

- 我们也可以将上述模型表示为如下形式:

$$\begin{aligned} z &\sim \mathcal{N}(0, I) \\ \epsilon &\sim \mathcal{N}(0, \Psi) \\ x &= \mu + \Lambda z + \epsilon \end{aligned}$$

- 其中 ϵ 和 z 是独立的

模型的求解

- 我们的随机变量 z 和 x 有如下的联合高斯分布:

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{zx}, \Sigma)$$

- 下面将分别求解 μ_{zx} 和 Σ

- 根据 $z \sim \mathcal{N}(0, I)$ ，我们有 $E[z] = 0$ ，而

$$\begin{aligned} E[x] &= E[\mu + \Lambda z + \epsilon] \\ &= \mu + \Lambda E[z] + E[\epsilon] \\ &= \mu \end{aligned}$$

- 将两者结合在一起，得到:

$$\mu_{zx} = \begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}$$

- 为了求解 Σ ，我们需要计算 Σ_{zz} （矩阵左上角）、 Σ_{zx} （矩阵右上角）和 Σ_{xx} （矩阵右下角）

- 矩阵左下角与右上角对称，计算其一即可

- 由于 $z \sim \mathcal{N}(0, I)$ ，根据之前提出的边缘分布性质，有 $\Sigma_{zz} = \text{Cov}(z) = I$

- Σ_{zx} 的求解如下:

$$\begin{aligned}
\Sigma_{zx} &= \mathbb{E}[(z - \mathbb{E}[z])(x - \mathbb{E}[x])^T] \\
&= \mathbb{E}[z(\mu + \Lambda z + \epsilon - \mu)^T] \\
&= \mathbb{E}[zz^T]\Lambda^T + \mathbb{E}[z\epsilon^T] \\
&= \Lambda^T
\end{aligned}$$

- 最后一步的推导使用了 $\mathbb{E}[zz^T] = \text{Cov}(z) + (\mathbb{E}[z])^2 = \text{Cov}(z)$
- 以及 $\mathbb{E}[z\epsilon^T] = \mathbb{E}[z]\mathbb{E}[\epsilon^T] = 0$
 - 因为 z 和 ϵ 相互独立
- 类似地, Σ_{xx} 的求解如下:

$$\begin{aligned}
\Sigma_{xx} &= \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T] \\
&= \mathbb{E}[(\mu + \Lambda z + \epsilon - \mu)(\mu + \Lambda z + \epsilon - \mu)^T] \\
&= \mathbb{E}[\Lambda zz^T \Lambda^T + \epsilon z^T \Lambda^T + \Lambda z \epsilon^T + \epsilon \epsilon^T] \\
&= \Lambda \mathbb{E}[zz^T] \Lambda^T + \mathbb{E}[\epsilon \epsilon^T] \\
&= \Lambda \Lambda^T + \Psi
\end{aligned}$$

- 综上所述, 得:

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda \Lambda^T + \Psi \end{bmatrix}\right) \quad (3)$$

- 因此, x 的边缘分布为 $x \sim \mathcal{N}(\mu, \Lambda \Lambda^T + \Psi)$
 - 给定一个训练集 $\{x^{(i)}; i = 1, \dots, m\}$, 我们可以得出如下的对数似然函数:

$$\begin{aligned}
\ell(\mu, \Lambda, \Psi) &= \sum_{i=1}^m \log p(x^{(i)}; \mu, \Lambda, \Psi) \\
&= \log \prod_{i=1}^m \frac{1}{(2\pi)^{n/2} |\Lambda \Lambda^T + \Psi|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu)^T (\Lambda \Lambda^T + \Psi)^{-1} (x^{(i)} - \mu)\right)
\end{aligned}$$

- 对该函数进行最大化估计是难以求出闭合解的
- 我们将使用 EM 算法来求解该问题

因子分析模型的 EM 求解

E-step

- 在 E-step 中, 我们需要计算 $Q_i(z^{(i)}) = p(z^{(i)} | x^{(i)}; \mu, \Lambda, \Psi)$
- 将 (3) 式代入之前推导出的条件分布公式 (1) 和 (2), 可以得出 $z^{(i)} | x^{(i)}; \mu, \Lambda, \Psi \sim \mathcal{N}(\mu_{z^{(i)} | x^{(i)}}, \Sigma_{z^{(i)} | x^{(i)}})$, 其中:

$$\begin{aligned}
\mu_{z^{(i)} | x^{(i)}} &= \Lambda^T (\Lambda \Lambda^T + \Psi)^{-1} (x^{(i)} - \mu), \\
\Sigma_{z^{(i)} | x^{(i)}} &= I - \Lambda^T (\Lambda \Lambda^T + \Psi)^{-1} \Lambda
\end{aligned}$$

- 因此, 基于上述定义, 我们有:

$$Q_i(z^{(i)}) = \frac{1}{(2\pi)^{k/2} |\Sigma_{z^{(i)} | x^{(i)}}|^{1/2}} \exp\left(-\frac{1}{2}(z^{(i)} - \mu_{z^{(i)} | x^{(i)}})^T (\Sigma_{z^{(i)} | x^{(i)}})^{-1} (z^{(i)} - \mu_{z^{(i)} | x^{(i)}})\right)$$

M-step

- 在 M-step 中, 我们需要最大化:

$$\sum_{i=1}^m \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)}{Q_i(z^{(i)})} dz^{(i)} \quad (4)$$

- 下面介绍关于参数 Λ 的优化方法，其他两个参数的推导省略
- 我们可以将 (4) 式简化为：

$$\sum_{i=1}^m \int_{z^{(i)}} Q_i(z^{(i)}) [\log p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)})] dz^{(i)} \quad (5)$$

$$= \sum_{i=1}^m \mathbb{E}_{z^{(i)} \sim Q_i} [\log p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)})] \quad (6)$$

- 因为 $z \sim \mathcal{N}(0, I)$ ，所以 $\log p(z^{(i)})$ 与 Λ 无关
- $\log Q_i(z^{(i)})$ 中的参数来自上一次迭代，与本次迭代中的参数无关
- 综上所述，我们需要优化的函数为：

$$\begin{aligned} & \sum_{i=1}^m \mathbb{E} [\log p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)] \\ &= \sum_{i=1}^m \mathbb{E} \left[\log \frac{1}{(2\pi)^{n/2} |\Psi|^{1/2}} \exp \left(-\frac{1}{2} (x^{(i)} - \mu - \Lambda z^{(i)})^T \Psi^{-1} (x^{(i)} - \mu - \Lambda z^{(i)}) \right) \right] \\ &= \sum_{i=1}^m \mathbb{E} \left[-\frac{1}{2} \log |\Psi| - \frac{n}{2} \log(2\pi) - \frac{1}{2} (x^{(i)} - \mu - \Lambda z^{(i)})^T \Psi^{-1} (x^{(i)} - \mu - \Lambda z^{(i)}) \right] \end{aligned}$$

- 具体的求导过程如下（上式只有最后一项与参数相关）：

$$\begin{aligned} & \nabla_{\Lambda} \sum_{i=1}^m -\mathbb{E} \left[\frac{1}{2} (x^{(i)} - \mu - \Lambda z^{(i)})^T \Psi^{-1} (x^{(i)} - \mu - \Lambda z^{(i)}) \right] \\ &= \sum_{i=1}^m \nabla_{\Lambda} \mathbb{E} \left[-\text{tr} \frac{1}{2} z^{(i)T} \Lambda^T \Psi^{-1} \Lambda z^{(i)} + \text{tr} z^{(i)T} \Lambda^T \Psi^{-1} (x^{(i)} - \mu) \right] \\ &= \sum_{i=1}^m \nabla_{\Lambda} \mathbb{E} \left[-\text{tr} \frac{1}{2} \Lambda^T \Psi^{-1} \Lambda z^{(i)} z^{(i)T} + \text{tr} \Lambda^T \Psi^{-1} (x^{(i)} - \mu) z^{(i)T} \right] \\ &= \sum_{i=1}^m \mathbb{E} \left[-\Psi^{-1} \Lambda z^{(i)} z^{(i)T} + \Psi^{-1} (x^{(i)} - \mu) z^{(i)T} \right] \end{aligned}$$

- 第一步首先将连乘打开，然后由于结果均为实数，所以用迹替换（ $\text{tra} = a$ ）
- 第二步使用迹的性质 $\text{tr} AB = \text{tr} BA$
- 第三步使用了迹的多条性质：

$$\begin{aligned} \nabla_{A^T} f(A) &= (\nabla_A f(A))^T \\ \nabla_{A^T} \text{tr} ABA^T C &= B^T A^T C^T + BA^T C \\ \nabla_A \text{tr} AB &= B^T \end{aligned}$$

- 将求导结果设为 0 并简化，得到：

$$\sum_{i=1}^m \Lambda \mathbb{E}_{z^{(i)} \sim Q_i} [z^{(i)} z^{(i)T}] = \sum_{i=1}^m (x^{(i)} - \mu) \mathbb{E}_{z^{(i)} \sim Q_i} [z^{(i)T}]$$

- 因此，我们可以解得：

$$\Lambda = \left(\sum_{i=1}^m (x^{(i)} - \mu) \mathbb{E}_{z^{(i)} \sim Q_i} [z^{(i)T}] \right) \left(\sum_{i=1}^m \mathbb{E}_{z^{(i)} \sim Q_i} [z^{(i)} z^{(i)T}] \right)^{-1} \quad (7)$$

- 该结果与线性回归中正规方程的解在形式上类似，因为二者都是在寻找两个变量之间的线性关系

- 对 (7) 式中的期望值进行求解，得到：

$$\begin{aligned} \mathbb{E}_{z^{(i)} \sim Q_i} \left[z^{(i)T} \right] &= \mu_{z^{(i)}|x^{(i)}}^T \\ \mathbb{E}_{z^{(i)} \sim Q_i} \left[z^{(i)} z^{(i)T} \right] &= \mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}} \end{aligned}$$

- 第二项的求解在源于公式 $\mathbb{E}(YY^T) = \mathbb{E}(Y)\mathbb{E}(Y^T) + \text{Cov}(Y)$
- 协方差项在求解中容易被忽略，需要注意
- 综上所述，可以得到 Λ 在 M-step 的更新公式为：

$$\Lambda = \left(\sum_{i=1}^m (x^{(i)} - \mu) \mu_{z^{(i)}|x^{(i)}}^T \right) \left(\sum_{i=1}^m \mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}} \right)^{-1} \quad (8)$$

- M-step 的其他两个参数的更新公式如下：

$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

- 该公式与参数无关，因此只需要计算一次即可

$$\Phi = \frac{1}{m} \sum_{i=1}^m x^{(i)} x^{(i)T} - x^{(i)} \mu_{z^{(i)}|x^{(i)}}^T \Lambda^T - \Lambda \mu_{z^{(i)}|x^{(i)}} x^{(i)T} + \Lambda (\mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}}) \Lambda^T$$

- 只取 Ψ 的对角元素组成 Ψ (即令 $\Psi_{ii} = \Phi_{ii}$)