Standford CS229 2022Fall，第13讲：因子分析

因子分析

当我们拥有来自多个高斯分布混合的数据 $x^{(i)} \in \mathbb{R}^d$ 时，可以应用 EM 算法来拟合混合模型。在这种情况下，我们通常设想的是数据量足够大，能够清晰地辨别出数据中的多高斯结构。例如，当我们的训练集大小 n 显著大于数据维度 d 时，就是这种情况。

现在，考虑一个 $d \gg n$ 的场景。在这样的问题中，即使使用单个高斯分布来建模数据也可能很困难，更不用说高斯混合模型了。具体来说，由于 n 个数据点仅张成 $\mathbb{R}^d$ 的一个低维子空间，如果我们把数据建模为高斯分布，并使用通常的最大似然估计器来估计均值和协方差：

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

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

我们会发现矩阵 Σ 是奇异的。这意味着 Σ^{-1} 不存在，且 $|\Sigma|^{1/2} = 1/0$。但这两个项都是计算多元高斯分布密度所必需的。另一种表述这种困难的方式是，参数的最大似然估计会导致一个高斯分布，其所有概率都集中在由数据张成的仿射空间上[1]，这对应于一个奇异的协方差矩阵。

[1]这是满足 $x = \sum_{i=1}^{n} \alpha_i x^{(i)}$ 的点集，其中 α_i 满足 $\sum_{i=1}^{n} \alpha_i = 1$。

更一般地说，除非 n 比 d 大出相当的数量，否则均值和协方差的最大似然估计可能会很差。
尽管如此，我们仍然希望能够在数据上拟合一个合理的高斯模型，并可能捕捉到数据中一些有趣的协方差结构。我们该如何做到这一点呢？

在下一节中，我们首先回顾两种允许我们在少量数据下拟合 Σ 的限制，但它们都不能为我们的问题提供令人满意的解决方案。接着，我们将讨论稍后需要的一些高斯分布的性质，特别是如何找到高斯分布的边缘分布和条件分布。最后，我们介绍因子分析模型及其 EM 算法。

1. 对 Σ 的限制

如果我们没有足够的数据来拟合完整的协方差矩阵，我们可以对考虑的矩阵 Σ 的空间施加一些限制。例如，我们可以选择拟合一个对角协方差矩阵 Σ。在这种设置下，读者可以很容易地验证，协方差矩阵的最大似然估计是由对角矩阵 Σ 给出的，该矩阵满足：

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

因此，Σ_{jj} 只是数据第 j 个坐标的方差的经验估计值。

回想一下，高斯密度的等高线是椭圆。一个对角 Σ 对应于一个高斯分布，其椭圆的主轴与坐标轴对齐。

有时，我们可能会进一步限制协方差矩阵，不仅要求它是对角的，而且其对角线元素必须全部相等。
在这种设置下，我们有 $\Sigma = \sigma^2 I$，其中 σ^2 是我们控制的参数。
σ^2 的最大似然估计可以求得为：

$$\sigma^2 = \frac{1}{nd}\sum_{j=1}^{d}\sum_{i=1}^{n}(x_j^{(i)}-\mu_j)^2.$$

这个模型对应于使用密度等高线为圆形（在二维中）或球体/超球体（在更高维度中）的高斯分布。

如果我们正在为数据拟合一个完整、无约束的协方差矩阵 Σ，则必须满足 $n \geq d+1$，才能使 Σ 的最大似然估计非奇异。在上述两种限制下，当 $n \geq 2$ 时，我们就可以获得非奇异的 Σ。

然而，将 Σ 限制为对角矩阵也意味着将数据的不同坐标 x_i, x_j 建模为不相关且独立的。通常，我们希望能够捕捉数据中一些有趣的关联结构。如果我们使用上述任一限制，我们将无法做到这一点。在本讲义中，我们将描述因子分析模型，它使用的参数比对角 Σ 更多，能够捕捉数据中的一些相关性，同时又无需拟合完整的协方差矩阵。

2. 高斯分布的边缘分布和条件分布

在描述因子分析之前，我们先偏离主题，谈谈如何找到具有联合多元高斯分布的随机变量的条件分布和边缘分布。

假设我们有一个向量值随机变量

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

其中 $x_1 \in \mathbb{R}^r$, $x_2 \in \mathbb{R}^s$, 且 $x \in \mathbb{R}^{r+s}$。假设 $x \sim N(\mu, \Sigma)$，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，$\mu_1 \in \mathbb{R}^r$, $\mu_2 \in \mathbb{R}^s$, $\Sigma_{11} \in \mathbb{R}^{r\times r}$, $\Sigma_{12} \in \mathbb{R}^{r\times s}$，等等。注意，由于协方差矩阵是对称的，$\Sigma_{12} = \Sigma_{21}^T$。

根据我们的假设，x_1 和 x_2 是联合多元高斯分布的。

x_1 的边缘分布是什么？不难看出 $E[x_1] = \mu_1$，且 $\mathrm{Cov}(x_1) = E[(x_1-\mu_1)(x_1-\mu_1)] = \Sigma_{11}$。要看到后者成立，请注意，根据 x_1 和 x_2 的联合协方差定义，我们有

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

比较第二行和最后一行矩阵的左上子块即可得到结果。

由于高斯分布的边缘分布本身也是高斯分布，因此我们有 x_1 的边缘分布为 $x_1 \sim N(\mu_1, \Sigma_{11})$。

另外，我们可以问，给定 x_2 时，x_1 的条件分布是什么？通过参考多元高斯分布的定义，可以证明 $x_1|x_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$，其中

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

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{2}$$

当我们下一节处理因子分析模型时，这些用于寻找高斯分布的条件分布和边缘分布的公式将非常有用。

3. 因子分析模型

在因子分析模型中，我们假设 (x, z) 的联合分布如下，其中 $z \in \mathbb{R}^k$ 是一个潜在的随机变量：

$$z \sim N(0, I)$$

$$x|z \sim N(\mu + \Lambda z, \Psi).$$

这里，模型的参数是向量 $\mu \in \mathbb{R}^d$，矩阵 $\Lambda \in \mathbb{R}^{d\times k}$，以及对角矩阵 $\Psi \in \mathbb{R}^{d\times d}$。$k$ 的值通常选择小于 d。

因此，我们想象每个数据点 $x^{(i)}$ 是通过采样一个 k 维多元高斯分布 $z^{(i)}$ 生成的。然后，通过计算 $\mu + \Lambda z^{(i)}$ 将其映射到 $\mathbb{R}^d$ 的一个 d 维仿射空间。最后，通过向 $\mu + \Lambda z^{(i)}$ 添加协方差为 Ψ 的噪声来生成 $x^{(i)}$。

等价地（请自行验证），我们也可以根据以下方式定义因子分析模型：

$$z \sim N(0, I)$$
$$\epsilon \sim N(0, \Psi)$$
$$x = \mu + \Lambda z + \epsilon$$

其中 ϵ 和 z 是独立的。

让我们精确地计算一下我们的模型所定义的分布。我们的随机变量 z 和 x 具有联合高斯分布 $\begin{bmatrix} z \\ x \end{bmatrix} \sim N(\mu_{zx}, \Sigma)$。

我们现在来求 μ_{zx} 和 Σ。

我们知道 $E[z] = 0$，因为 $z \sim N(0, I)$。此外，我们有

$$E[x] = E[\mu + \Lambda z + \epsilon] = \mu + \Lambda E[z] + E[\epsilon] = \mu.$$

将这些结合起来，我们得到 $\mu_{zx} = \begin{bmatrix} 0 \\ \mu \end{bmatrix}$。

接下来，为了找到 Σ，我们需要计算 $\Sigma_{zz} = E[(z - E[z])(z - E[z])^T]$（$\Sigma$ 的左上块），$\Sigma_{zx} = E[(z - E[z])(x - E[x])^T]$（右上块），以及 $\Sigma_{xx} = E[(x - E[x])(x - E[x])^T]$（右下块）。

现在，由于 $z \sim N(0, I)$，我们很容易发现 $\Sigma_{zz} = \mathrm{Cov}(z) = I$。另外，

$$E[(z - E[z])(x - E[x])^T] = E[z(\mu + \Lambda z + \epsilon - \mu)^T] = E[zz^T]\Lambda^T + E[z\epsilon^T] = \Lambda^T.$$

在最后一步中，我们使用了 $E[zz^T] = \mathrm{Cov}(z)$（因为 z 的均值为零），以及 $E[z\epsilon^T] = E[z]E[\epsilon^T] = 0$（因为 z 和 ϵ 是独立的，因此它们乘积的期望等于它们期望的乘积）。

类似地，我们可以如下找到 Σ_{xx}：

$$E[(x - E[x])(x - E[x])^T] = E[(\mu + \Lambda z + \epsilon - \mu)(\mu + \Lambda z + \epsilon - \mu)^T] = E[\Lambda z z^T \Lambda^T + \epsilon z^T \Lambda^T + \Lambda z \epsilon^T + \epsilon\epsilon^T] = \Lambda E[zz^T]\Lambda^T + E[\epsilon\epsilon^T] = \Lambda\Lambda^T + \Psi.$$

将所有内容放在一起，我们因此有 $\begin{bmatrix} z \\ x \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}\right)$. \tag{3}

因此，我们也看到 x 的边缘分布由 $x \sim N(\mu, \Lambda\Lambda^T + \Psi)$ 给出。因此，给定一个训练集 $\{x^{(i)}; i = 1, \ldots, n\}$，我们可以写出参数的对数似然：

$$\ell(\mu, \Lambda, \Psi) = \log \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2}|\Lambda\Lambda^T + \Psi|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu)^T(\Lambda\Lambda^T + \Psi)^{-1}(x^{(i)} - \mu)\right).$$

为了进行最大似然估计，我们希望最大化这个量。但是显式地最大化这个公式是很困难的（你自己试试看），并且我们不知道有任何算法可以在闭式解中做到这一点。因此，我们将改用 EM 算法。在下一节中，我们推导因子分析的 EM 算法。

4. 因子分析的 EM 算法

E 步的推导很简单。我们需要计算 $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi)$。通过将方程 (3) 中给出的分布代入用于寻找高斯分布条件分布的公式 (1-2)，我们发现 $z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi \sim N(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}})$，其中

$$\mu_{z^{(i)}|x^{(i)}} = \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(x^{(i)} - \mu),$$

$$\Sigma_{z^{(i)}|x^{(i)}} = I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda.$$

因此，使用这些定义，我们有

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

现在我们来推导 M 步。在这里，我们需要关于参数 μ, Λ, Ψ 最大化

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

我们将只推导关于 Λ 的优化，而将 μ 和 Ψ 的更新推导留给读者作为练习。

我们可以将方程 (4) 简化如下：

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

$$= \sum_{i=1}^{n} E_{z^{(i)}\sim Q_i} [\log p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)})] \tag{6}$$

这里，“$z^{(i)} \sim Q_i$” 下标表示期望是相对于从 Q_i 中抽取的 $z^{(i)}$ 而言的。在后续发展中，如果没有歧义的风险，我们将省略此下标。去掉不依赖于参数的项，我们发现我们需要最大化：

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

让我们关于 Λ 最大化这个表达式。只有上面的最后一项依赖于 Λ。取导数，并利用 $\operatorname{tr} a = a$（对于 $a \in \mathbb{R}$），$\operatorname{tr} AB = \operatorname{tr} BA$，以及 $\nabla_A \operatorname{tr} ABA^T C = CAB + C^T AB^T$ 这些事实，我们得到：

$$\nabla_\Lambda \sum_{i=1}^{n} -E\left[\frac{1}{2}(x^{(i)}-\mu-\Lambda z^{(i)})^T \Psi^{-1}(x^{(i)}-\mu-\Lambda z^{(i)})\right] = \sum_{i=1}^{n} \nabla_\Lambda E\left[-\frac{1}{2}\operatorname{tr} z^{(i)T}\Lambda^T\Psi^{-1}\Lambda z^{(i)} + \operatorname{tr} z^{(i)T}\Lambda^T\Psi^{-1}(x^{(i)}-\mu)\right] = \sum_{i=1}^{n} \nabla_\Lambda E\left[-\frac{1}{2}\operatorname{tr} \Lambda^T\Psi^{-1}\Lambda\right.$$

令其为零并简化，我们得到：

$$\sum_{i=1}^{n} \Lambda E_{z^{(i)}\sim Q_i}[z^{(i)}z^{(i)T}] = \sum_{i=1}^{n} (x^{(i)}-\mu) E_{z^{(i)}\sim Q_i}[z^{(i)T}].$$

因此，解出 Λ，我们得到

$$\Lambda = \left(\sum_{i=1}^{n} (x^{(i)}-\mu) E_{z^{(i)}\sim Q_i}[z^{(i)T}]\right)\left(\sum_{i=1}^{n} E_{z^{(i)}\sim Q_i}[z^{(i)}z^{(i)T}]\right)^{-1}. \tag{7}$$

有趣的是，注意到这个方程与我们为最小二乘回归推导出的正规方程“$\theta^T = (y^T X)(X^T X)^{-1}$”有着密切的关系。类比在于，这里的 x 是 z 的线性函数（加上噪声）。
鉴于 E 步已经找到了 z 的“猜测”，我们现在尝试估计连接 x 和 z 的未知线性关系 Λ。因此，我们得到类似于正规方程的结果也就不足为奇了。然而，这里有一个重要的区别：我们很快就会看到这个区别。
为了完成我们的 M 步更新，让我们计算方程 (7) 中的期望值。根据我们对 Q_i 的定义，它是一个均值为 $\mu_{z^{(i)}|x^{(i)}}$、协方差为 $\Sigma_{z^{(i)}|x^{(i)}}$ 的高斯分布，我们很容易发现

$$E_{z^{(i)}\sim Q_i}[z^{(i)T}] = \mu_{z^{(i)}|x^{(i)}}^T$$

$$E_{z^{(i)}\sim Q_i}[z^{(i)}z^{(i)T}] = \mu_{z^{(i)}|x^{(i)}}\mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}}.$$

后者来自于这样一个事实：对于随机变量 Y，$\operatorname{Cov}(Y) = E[YY^T] - E[Y]E[Y]^T$，因此 $E[YY^T] = E[Y]E[Y]^T + \operatorname{Cov}(Y)$。将这些代回方程 (7)，我们得到 Λ 的 M 步更新：

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

重要的是要注意到，这个方程的右侧存在 $\Sigma_{z^{(i)}|x^{(i)}}$。这是后验分布 $p(z^{(i)}|x^{(i)})$ 中 $z^{(i)}$ 的协方差，M 步必须考虑到后验分布中关于 $z^{(i)}$ 的这种不确定性。
在推导 EM 时，一个常见的错误是假设在 E 步中，我们只需要计算潜在随机变量 z 的期望 $E[z]$，然后将其代入 M 步的优化中所有出现 z 的地方。虽然这在像高斯混合模型这样简单的问题中有效，但在我们对因子分析的推导中，我们还需要 $E[zz^T]$ 以及 $E[z]$；正如我们所见，$E[zz^T]$ 和 $E[z]E[z]^T$ 相差一个量 $\Sigma_{z|x}$。
因此，M 步的更新必须考虑到后验分布 $p(z^{(i)}|x^{(i)})$ 中 z 的协方差。
最后，我们还可以找到参数 μ 和 Ψ 的 M 步优化。不难证明，第一个由下式给出：

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

由于这不会随着参数的变化而改变（即，与 Λ 的更新不同，右边不依赖于 $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)};\mu,\Lambda,\Psi)$，而后者又依赖于参数），因此可以一次性计算，无需在算法运行过程中进一步更新。类似地，对角矩阵 Ψ 可以通过计算

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

然后令 $\Psi_{ii} = \Phi_{ii}$（即，让 Ψ 成为仅包含 Φ 对角元素的对角矩阵）来找到。