Factor analysis

因子分析

问题引入

之前我们考虑的数据集x(i)的个数m都远大于特征数n，这样进行EM算法求解就没有问题，如果是n>>m，也就是特征太多或者样本数不足的情况下，我们再用高斯模型对数据进行建模，用最大似然估计求出均值和方差，会得到

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

均值没有什么问题，但是方差矩阵却是奇异矩阵，这意味着逆矩阵不存在，并且 $1/|\Sigma|^{1/2} = 1/0$，但是这些量都是计算多元高斯分布所需要的。

- 此处我不太明白为什么方差矩阵是奇异矩阵，有知道的朋友还望麻烦告知我。github地址： https://github.com/zhoucz97/CS229/issues

更一般地，除非m>>n，否则最大似然估计得到的均值和方差都会不尽人意，但是我们又想用现有数据来进行建模求解，那么怎么办呢？可以对Σ加以限制。

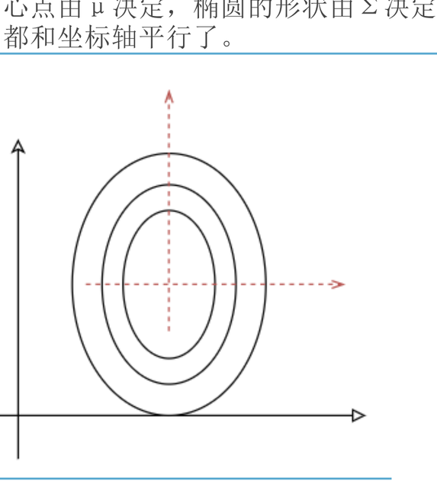
限制Σ

如果我们没有足够的数据来拟合一个完整的协方差矩阵，那么我们可以对矩阵Σ做一些约束。

比如，我们选择拟合一个对角矩阵Σ，一个协方差矩阵的最大似然估计可以由对角矩阵满足

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

- 回想我们之前讨论过的二维多元高斯分布的几何特性，在平面上的投影是个椭圆，中心点由μ决定，椭圆的形状由Σ决定。Σ如果变成对角阵，就意味着椭圆的两个轴都和坐标轴平行了。



- 对角矩阵也意味着各特征间相互独立，丢失了不同维度之间的相关性信息
- Σjj 就是对数据中第 j 个坐标位置的方差值的经验估计（empirical estimate）

有时候我们还要进一步约束，对角矩阵的所有对角元素都必须相同，即 $\Sigma = \sigma^2 I$

σ^2是我们控制的参数，对它的最大似然估计为

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

- 这样一来，上述图片的椭圆就变成圆了

当我们要估计出完整的Σ时，我们需要 m>=n+1 才能保证在最大似然估计下得出的Σ是非 奇异的。然而在上面的任何一种假设限定条件下，只要 m>=2 都可以估计出限定的Σ。

但是这样的缺点就是特征间的相关性信息没有了，我们用因子分析模型来解决这个问题

高斯模型的边缘分布和条件分布

在讲解因子分析（factor analysis）之前，我们要先说一下一个联合多元高斯分布（joint multivariate Gaussian distribution）下的随机变量（random variables）的条件（conditional）和边缘（marginal）分布（distributions）

做如下假设

假设我们有一个值为向量的随机变量 $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$，并且 $x_1 \in \mathbb{R}^r,\ x_2 \in \mathbb{R}^s,\ \text{and } x \in \mathbb{R}^{r+s}$

再假设 $x \sim \mathcal{N}(\mu, \Sigma)$，并且 $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},\quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$，$\mu_1 \in \mathbb{R}^r,\ \mu_2 \in \mathbb{R}^s,\ \Sigma_{11} \in \mathbb{R}^{r\times r},\ \Sigma_{12} \in \mathbb{R}^{r\times s}$

由于协方差矩阵是对称的，所以 $\Sigma_{12} = \Sigma_{21}^T$

基于以上假设，x1,x2是联合多元高斯分布(jointly multivariate Gaussian)

求x1的边缘分布

我们可以看出，$\mathrm{E}[x_1] = \mu_1$，$\mathrm{Cov}(x_1) = \mathrm{E}[(x_1-\mu_1)(x_1-\mu_1)] = \Sigma_{11}$

证明第二个式子

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

划红线第二处为x1的方差的定义，与划红线第一处相对应，证毕

即联合多元高斯分布的边缘分布也是高斯分布，所以 $x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$

求x1|x2的条件分布

$$x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$$

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2),$$
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

- 吴恩达老师课上只说用概率密度公式进行计算，还说特别麻烦，没有给出过程，我也不知道咋算出来的....
- 如果想了解 具体的推导过程，可以参见 Chuong B. Do 写的《Gaussian processes》。

因子分析模型

1 首先构建一个(x,z)的联合分布，z是隐藏随机变量并且 $z \in \mathbb{R}^k$

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

模型参数

- vector $\mu \in \mathbb{R}^n$
- 矩阵 $\Lambda \in \mathbb{R}^{n\times k}$
- diagonal matrix $\Psi \in \mathbb{R}^{n\times n}$
- 通常k<n

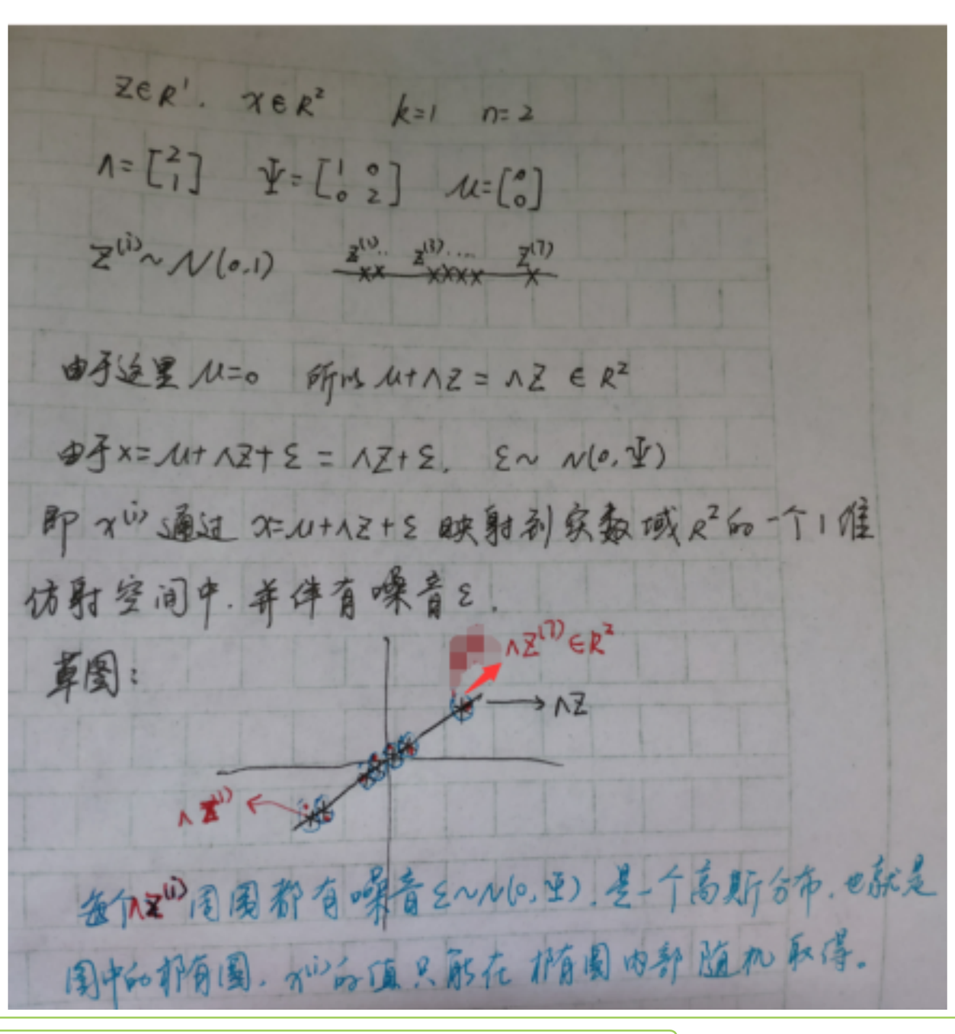
每个数据点 x(i) 都是通过在一个 k 维度的多元高斯分布 z(i) 中取样获得的。然后，通过计算 μ+Λz(i)，就可以映射到实数域 Rn 中的一个 k 维仿射空间（k-dimensional affine space），在 μ + Λz(i) 上加上协方差 Ψ 作为噪音，就得到了 x(i)。

2 与之等价，我们定义因子分析模型

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

where ϵ and z are independent

举例



在模型中，x,z服从联合高斯分布 $\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{zx}, \Sigma)$

3 接下来我们要找出 μ_{zx} and Σ

因为 $z \sim \mathcal{N}(0, I)$，所以有 $\mathrm{E}[z] = 0$，所以

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

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

接下来计算Σ

因为 $z \sim \mathcal{N}(0, I)$，所以显然 $\Sigma_{zz} = \mathrm{Cov}(z) = I$

Σzx=Cov(z,x)= $\mathrm{E}[(z-\mathrm{E}[z])(x-\mathrm{E}[x])^T] = \mathrm{E}[z(\mu+\Lambda z+\epsilon-\mu)^T] = \mathrm{E}[zz^T]\Lambda^T + \mathrm{E}[z\epsilon^T] = \Lambda^T$

- E(z)=0, E(x)=μ
- Cov(z)=E((z-Ez)(z-Ez).T)=E(zz.T)=I

同理可得Σxz=Λ

Σxx=Cov(x)=

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

所以得到

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

我们还能发现x的边缘分布为 $\mathcal{N}(\mu, \Lambda\Lambda^T + \Psi)$

因此，给出一组数据集{x1,x2,...xm}，我们能够写出参数的最大似然函数估计

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

然而对上面的式子进行最大化是很难的，所以接下来我们使用EM算法

针对因子分析模型的EM算法

目的，求出 μ, Λ, Ψ 三个参数

E-step

就是求 $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi)$

只需把 $\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}\right)$ 代入 $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$, $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$ 中

我们会发现 $z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi \sim \mathcal{N}(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}})$

并且

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

进而得到

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

M-step

最大化下面这个关于参数 μ, Λ, Ψ 的函数值

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

1 本文中仅仅对 Λ 进行优化

We can simplify Equation (4) as follows:

$$\sum_{i=1}^{m}\int_{z^{(i)}} Q_i(z^{(i)})\left[\log p(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)})\right] dz^{(i)} \quad (5)$$
$$= \sum_{i=1}^{m} \mathrm{E}_{z^{(i)}\sim Q_i}\left[\log p(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)})\right] \quad (6)$$

"$z^{(i)} \sim Q_i$" 表示这个期望是从关于zi的Qi中取得的，后面无特殊情况，我们会将这个标志省略掉

2 因为只优化Λ，所以去掉无关参数后

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

3 对其求偏导

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

- 1式到2式拆开，并使用了公式 $\mathrm{tr}\, a = a \ (\text{for } a \in \mathbb{R})$
- 2式到3式使用公式 $\mathrm{tr}AB = \mathrm{tr}BA$
- 3式到4式使用公式 $\nabla_A \mathrm{tr}ABA^TC = CAB + C^TAB$

4 设为0解得

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

5 整理得到

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

有一个很有意思的地方，上式和用最小二乘线性回归推出的正规方程很类似 "$\theta^T = (y^TX)(X^TX)^{-1}.$"

6 紧接着，为了能完成M步骤的更新，我们要解出（7）式中的期望值

由于 Q_i being Gaussian with mean $\mu_{z^{(i)}|x^{(i)}}$ and covariance $\Sigma_{z^{(i)}|x^{(i)}}$

我们很容易得到

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

- 第二个式子利用了 $\mathrm{E}[YY^T] = \mathrm{E}[Y]\mathrm{E}[Y]^T + \mathrm{Cov}(Y)$

7 最后得到更新规则

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

注意，上式中的 $\Sigma_{z^{(i)}|x^{(i)}}$，这是一个后验分布 $p(z^{(i)}|x^{(i)})$ 的协方差，而zi是隐含随机变量，具有不确定性，这一点要考虑到

同理，给出参数μ 和 Ψ的更新规则

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

- 由公式可以看出，μ在整个过程中只需计算一次，因为它只与数据集有关

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

然后设置 $\Psi_{ii} = \Phi_{ii}$（也就是说，设Ψ为一个仅包含Φ中对角线元素的对角矩阵）