高斯分布（正态分布）

一、极大似然估计

Data: $X=(x_1, x_2, \cdots, x_N)^T=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}_{N\times p}$，$x_i \in \mathbb{R}^p$，$x_i \overset{iid}{\sim} N(\mu, \Sigma)$

iid 即独立同分布

$p=1$：

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$p>1$：

$$p(x)=\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

令参数 $\theta=(\mu, \Sigma)$

则极大似然估计 MLE：$\theta_{MLE}=\arg\max_{\theta} P(X|\theta)$

为简化计算，令 $p=1$，则 $\theta=(\mu, \sigma^2)$

推导：

$\because x_i$ 独立同分布，则

$$\begin{aligned}\log P(X|\theta) &= \log\prod_{i=1}^{N} p(x_i|\theta)=\sum_{i=1}^{N}\log p(x_i|\theta) \\ &=\sum_{i=1}^{N}\log\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &=\sum_{i=1}^{N}\left[\log\frac{1}{\sqrt{2\pi}}+\log\frac{1}{\sigma}-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\end{aligned}$$

$$\begin{aligned}\mu_{MLE} &= \arg\max_{\mu}\log P(X|\theta) \\ &=\arg\max\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2} \\ &=\arg\min\sum_{i=1}^{N}(x_i-\mu)^2\end{aligned}$$

对 μ 求导：

$$\frac{\partial}{\partial\mu}\sum_{i=1}^{N}(x_i-\mu)^2=\sum_{i=1}^{N}2(x_i-\mu)\cdot(-1)=0$$

$$\sum_{i=1}^{N}(x_i-\mu)=0$$

$$\sum_{i=1}^{N}x_i-\underbrace{\sum_{i=1}^{N}\mu}_{N\cdot\mu}=0$$

$$\mu_{MLE}=\frac{1}{N}\sum_{i=1}^{N}x_i$$ 即样本的均值。

$$\sigma^2_{MLE} = \arg\max_{\sigma} P(X|\theta) = \arg\max \underbrace{\left(-\log\sigma - \frac{1}{2\sigma^2}(x_i-\mu)^2\right)}_{L}$$

$$\frac{\partial L}{\partial \sigma} = \sum_{i=1}^{N} \frac{1}{\sigma} - \sum (x_i-\mu)^2 \cdot (-2)\sigma^{-3} = 0$$

$$\sum_{i=1}^{N}\left[-\frac{1}{\sigma} + (x_i-\mu)^2\sigma^{-3}\right] = 0$$

$$\sum_{i=1}^{N}\left[-\sigma^2 + (x_i-\mu)^2\right] = 0$$

$$-\sum_{i=1}^{N}\sigma^2 + \sum_{i=1}^{N}(x_i-\mu)^2 = 0$$

$$\sum_{i=1}^{N}\sigma^2 = \sum_{i=1}^{N}(x_i-\mu)^2$$

$$\boxed{\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{MLE})^2}$$

σ^2_{MLE}是有偏估计，μ_{MLE}是无偏估计

σ^2_{MLE}：$E[\sigma^2_{MLE}] = \frac{N-1}{N}\sigma^2$，真正的无偏应为 $\hat{\sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_{MLE})^2$

μ_{MLE}：$E[\mu_{MLE}] = \frac{1}{N}\sum_{i=1}^{N}E[x_i] = \frac{1}{N}\sum_{i=1}^{N}\mu = \frac{1}{N}\cdot N\cdot\mu = \mu$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{MLE})^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x_i\mu_{MLE} + \mu_{MLE}^2) = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \underbrace{\frac{1}{N}\sum_{i=1}^{N}2x_i\mu_{MLE}}_{2\cdot\mu_{MLE}\cdot\mu_{MLE}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\mu_{MLE}^2}_{\mu_{MLE}^2}$$

$$= \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \mu_{MLE}^2$$

$$E[\sigma^2_{MLE}] = E\left[\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \mu^2_{MLE}\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \mu^2 - (\mu^2_{MLE} - \mu^2)\right]$$

$$= E\left(\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \mu^2\right) - E(\mu^2_{MLE} - \mu^2)$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N}(x_i^2-\mu^2)\right] - \left[E(\mu^2_{MLE}) - E(\mu^2)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}E(x_i^2-\mu^2) - \left[E(\mu^2_{MLE}) - \mu^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(E(x_i^2) - E^2(x_i)\right) - \left[E(\mu^2_{MLE}) - E^2(\mu_{MLE})\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sigma^2 - D(\mu_{MLE})$$

$$D(\mu_{MLE}) = D\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right) = \frac{1}{N^2}\sum_{i=1}^{N}D(x_i) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 = \frac{1}{N}\sigma^2$$

$$\therefore E[\sigma^2_{MLE}] = \sigma^2 - \frac{1}{N}\sigma^2 = \frac{N-1}{N}\sigma^2$$

高维情形：$x \in \mathbb{R}^p$，$x=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$ $\mu=\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$ $\Sigma=\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & & \sigma_{pp} \end{pmatrix}_{p\times p}$

通常 Σ 是半正定的，即 $\Sigma_{ij}=\Sigma_{ji}$

此处假设 Σ 为正定

$$X\sim N(\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{p}{2}}\cdot|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

$(x-\mu)^T\Sigma^{-1}(x-\mu)$：马氏距离，（$x$ 和 μ 之间的）

Σ 是协方差矩阵，如果 Σ 为 I（单位阵），则马氏距离=欧氏距离

$\Sigma=U\Lambda U^T$，$UU^T=U^TU=I$，$\Lambda=\mathrm{diag}(\lambda_i)$ → 对角阵

$U=(u_1,u_2,\cdots u_p)_{p\times p}$.

$$=(u_1,u_2,\cdots u_p)\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}\begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{pmatrix}$$

$$=(u_1\lambda_1,u_2\lambda_2,\cdots u_p\lambda_p)\begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{pmatrix}$$

$$=\sum_{i=1}^{p}u_i\lambda_i u_i^T \quad \because U\text{正交，则 } U^T=U^{-1}$$

则 $\Sigma^{-1}=(U\Lambda U^T)^{-1}=(U^T)^{-1}\Lambda^{-1}U^{-1}=U\Lambda^{-1}U^T$，$\Lambda^{-1}=\mathrm{diag}(\frac{1}{\lambda_i})$，$i=1,\cdots,p$.

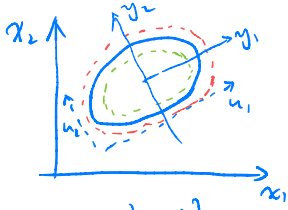
$$=\sum_{i=1}^{p}u_i\frac{1}{\lambda_i}u_i^T$$

则 $(x-\mu)^T\Sigma^{-1}(x-\mu)=(x-\mu)^T\sum_{i=1}^{p}u_i\frac{1}{\lambda_i}u_i^T(x-\mu)=\sum_{i=1}^{p}(x-\mu)^Tu_i\frac{1}{\lambda_i}u_i^T(x-\mu)$

令 $y_i=(x-\mu)^Tu_i$，则 $=\sum_{i=1}^{p}y_i\frac{1}{\lambda_i}y_i^T=\boxed{\sum_{i=1}^{p}\frac{y_i^2}{\lambda_i}}$

如果 $p=2$，那么 $(x-\mu)^T\Sigma^{-1}(x-\mu)=\frac{y_1^2}{\lambda_1}+\frac{y_2^2}{\lambda_2}$，令该值为 1.

则有 $\frac{y_1^2}{\lambda_1}+\frac{y_2^2}{\lambda_2}=1$，其为椭圆曲线.



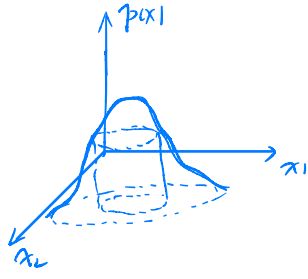
$y_i=(x-\mu)^Tu_i$ 表示，将 x 往中心移，$(x-\mu)$ 这个向量在 u_i 上的投影，即将 $(x-\mu)$ 在 u_i 上的投影作为 y_i

我们发现当 $\frac{y_1^2}{\lambda_1}+\frac{y_2^2}{\lambda_2}$ 取不同值时，设为 r，对于某个固定的 r，都能得到不同的椭圆曲线，是一个等高线.

对于 $X\sim N(\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

它的概率密度函数 $p(x)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\Delta\right)$，对于 $p(x)$ 来讲，x 有 n 是自变量，μ 与 Σ 是模型的参数。

$0\le p(x)\le 1$，当 $p(x)$ 取某一值时，总会有一个 $\Delta=r_1$ 与之对应。



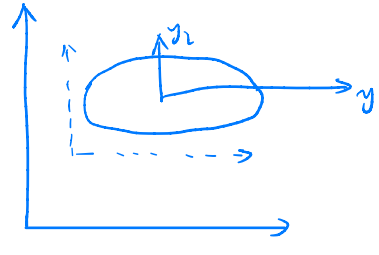
局限性：

① $\Sigma_{p\times p}$：元素共 p^2 个，又因其对称，则共 $\frac{p(p+1)}{2}$ 个参数，为 $O(p^2)$，因此，当面对实际问题时，尤其高维的问题，p 会很大，参数会过多。

因此我们会对方差矩阵做简化。

$\Sigma=\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_p\end{pmatrix}$，只在对角线存在值，其余全为 0。

那么 Σ 就可以不用做特征值分解，其中的 $y_i=(x-\mu)^Tu_i$ 就变为了 $y_i=(x-\mu)^Tv_i$，图象变为↓，其与原来 x_1,x_2 一致。



若 $\lambda_1=\lambda_2=\cdots\lambda_p$，则为圆形，称为各向同性。

③ 高斯分布本身无法表达模型，不是好的表达模型。
可用混合模型解决（多个高斯）

P5 求边缘概率和条件概率

设 $x=\begin{pmatrix}x_1\\x_2\\\vdots\\x_p\end{pmatrix}$ $\mu=\begin{pmatrix}\mu_1\\\mu_2\\\vdots\\\mu_p\end{pmatrix}$ $\Sigma=\begin{pmatrix}\sigma_{11}&\cdots&\sigma_{1p}\\\sigma_{21}&\ddots&\vdots\\\vdots&\ddots&\\\sigma_{p1}&\cdots&\sigma_{pp}\end{pmatrix}$

已知：$x=\begin{pmatrix}x_a\to m\\x_b\to n\end{pmatrix}$ $m+n=p$ $\mu=\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix}$ $\Sigma=\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}$

求 $P(x_a)$，$P(x_b|x_a)$
对称的 $P(x_b)$，$P(x_a|x_b)$

定理：$X\sim N(\mu,\Sigma)$，$y=AX+B$，则 $y\sim N(A\mu+B,A\Sigma A^T)$

$P(x_a)$：$x_a=\underbrace{(I_m,0)}_{A}\underbrace{\begin{pmatrix}x_a\\x_b\end{pmatrix}}_{X}$，$E[x_a]=(I_m,0)\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix}=\mu_a$

$$Var[x_a]=(I_m,0)\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}\begin{pmatrix}I_m\\0\end{pmatrix}=(\Sigma_{aa}\ \Sigma_{ab})\begin{pmatrix}I_m\\0\end{pmatrix}=\Sigma_{aa}$$

所以 $x_a\sim N(\mu_a,\Sigma_{aa})$

$P(x_b|x_a)$：

$$\text{定义}\begin{cases}x_{b\cdot a}=x_b-\Sigma_{ba}\Sigma_{aa}^{-1}x_a\\\mu_{b\cdot a}=\mu_b-\Sigma_{ba}\Sigma_{aa}^{-1}\mu_a\\\Sigma_{bb\cdot a}=\Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\end{cases}$$

$$x_{ba}=(-\Sigma_{ba}\Sigma_{aa}^{-1}\quad I_n)\begin{pmatrix}x_a\\x_b\end{pmatrix}$$

$$\text{则 } E[x_{b\cdot a}]=(-\Sigma_{ba}\Sigma_{aa}^{-1}\quad I_n)\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix}=\mu_b-\Sigma_{ba}\Sigma_{aa}^{-1}\mu_a=\mu_{b\cdot a}.$$

$$Var[X_{b\cdot a}] = \begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1} & I_n \end{pmatrix} \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \begin{pmatrix} -\Sigma_{aa}^{-1}\Sigma_{ba}^{T} \\ I_n \end{pmatrix}$$

$$= \begin{pmatrix} 0 & \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \end{pmatrix} \begin{pmatrix} -\Sigma_{aa}^{-1}\Sigma_{ba}^{T} \\ I \end{pmatrix} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} = \Sigma_{bb\cdot a}$$

$\therefore X_{b\cdot a} \sim N(\mu_{b\cdot a}, \Sigma_{bb\cdot a})$

$\because X_b = X_{b\cdot a} + \underbrace{\Sigma_{ba}\Sigma_{aa}^{-1}X_a}_{B}$

此处 B 视作常数，$X_{b\cdot a}$ 与 X_a 相互独立，故 $X_b|X_a = X_b$.

$\therefore E[X_b|X_a] = \mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}X_a$

$Var[X_b|X_a] = Var[X_{b\cdot a}] = \Sigma_{bb\cdot a}$

故 $X_b|X_a \sim N(\mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}X_a, \Sigma_{bb\cdot a})$

P6. 已知