

# Math Basis 01

Chen Gong

18 October 2019

本节的主要目的是从频率派的角度使用极大似然估计，通过观察到数据，是观察到的数据出现的概率最大化，来对高斯分布的参数进行估计。并且分析了高斯分布的参数， $\mu$ ， $\sigma^2$  的无偏性和有偏性。其中， $\mu$  是关于参数的无偏估计，而  $\sigma$  是有偏估计。

数据矩阵为：（这样可以保证每一行为一个数据点）

$$X = (x_1, x_2, \dots, x_N)^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{32} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix}_{N \times P} \quad (1)$$

在数据矩阵的基础上，有  $x_i \in \mathbb{R}$ ， $x_i \sim \mathcal{N}(\mu, \Sigma)$ ，那么参数为  $\theta = \mathcal{N}(\mu, \Sigma)$ 。

## 1 求解目的

首先对于单变量的高斯分布  $\mathcal{N}(\mu, \sigma^2)$ ，概率密度函数为：

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

然而对于多变量的高斯分布  $\mathcal{N}(\mu, \Sigma)$ ，概率密度函数为：

$$p(X) = \frac{1}{\sqrt{2\pi}^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu) \right\} \quad (3)$$

我们希望通过观察到的数据来计算参数  $\theta$  的值，那么我们使用极大似然估计的优化目标为  $\theta_{MLE} = \operatorname{argmax}_{\theta} p(x|\theta)$ 。于是我们可以转化为  $\theta_{MLE} = \operatorname{argmax}_{\theta} \log p(x|\theta)$ 。那么，计算公式可以化简为：

$$\log p(x|\theta) = \log \prod_{i=1}^N p(x_i|\theta) = \sum_{i=1}^N \log p(x_i|\theta) \quad (4)$$

$$= \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi}} + \log \frac{1}{\sigma} - \frac{(x - \mu)^2}{2\sigma^2} \quad (5)$$

## 2 极大似然法求解参数 $\mu$ 和 $\sigma^2$

在求解  $\mu_{MLE}$  时，计算目标为  $\frac{\partial \log p(x|\theta)}{\partial \mu}$ ，推导公式如下：

$$\frac{\partial \log p(x|\theta)}{\partial \mu} = \sum_{i=1}^N -\frac{(x_i - \mu)}{\sigma^2} = 0 \quad (6)$$

$$\sum_{i=1}^N x_i = \sum_{i=1}^N \mu \quad (7)$$

$$\mu_{MLE} = \frac{1}{N} \sum_{i=1}^N x_i \quad (8)$$

在求解  $\sigma_{MLE}^2$  时，计算目标为  $\frac{\partial \log p(x|\theta)}{\partial \sigma}$ ，推导公式如下：

$$\frac{\partial \log p(x|\theta)}{\partial \sigma} = \sum_{i=1}^N -\frac{1}{\sigma} - \frac{1}{2}(x_i - \mu)^2(-2)\sigma^{-3} = 0 \quad (9)$$

$$\sum_{i=1}^N \sigma^2 = \sum_{i=1}^N (x_i - \mu)^2 \quad (10)$$

$$\sigma_{MLE}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (11)$$

实际上这里的  $\mu$  是  $\mu_{MLE}$ ，所以，

$$\sigma_{MLE}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_{MLE})^2 \quad (12)$$

## 3 验证 $\mu_{MLE}$ 和 $\sigma_{MLE}^2$ 的无偏性

首先需要明确什么是无偏估计，所谓无偏估计也就是， $\mathbb{E}(\hat{x}) = x$ 。那么利用这个性质我们就可以很方便的判断一个估计是否为无偏估计。

### 3.1 验证 $\mu_{MLE}$ 的无偏性

$$\begin{aligned} \mathbb{E}[\mu_{MLE}] &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N x_i\right] \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i] \\ &= \frac{1}{N} N\mu = \mu \end{aligned} \quad (13)$$

根据上述的推导，我们可以得出  $\mu_{MLE}$  是无偏估计。

### 3.2 验证 $\sigma_{MLE}^2$ 的无偏性

$$\begin{aligned}
\mathbb{E}[\sigma_{MLE}^2] &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N (x_i - \mu_{MLE})^2\right] \\
&= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N (x_i^2 - 2\mu_{MLE}x_i + \mu_{MLE}^2)\right] \\
&= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N (x_i^2 - \mu_{MLE}^2)\right] \\
&= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N (x_i^2 - \mu^2) - (\mu_{MLE}^2 - \mu^2)\right] \\
&= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N (x_i^2 - \mu^2)\right] - \mathbb{E}[(\mu_{MLE}^2 - \mu^2)] \\
&= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N (x_i^2 - (\frac{1}{N} \sum_{i=1}^N x_i)^2)\right] - \mathbb{E}[(\mu_{MLE}^2 - \mu^2)] \\
&= \frac{1}{N} \sum_{i=1}^N (\mathbb{E}[x_i^2] - \mathbb{E}[x]^2) - \mathbb{E}[(\mu_{MLE}^2 - \mu^2)] \\
&= \sigma^2 - (\mathbb{E}[\mu_{MLE}^2] - \mathbb{E}[\mu^2]) \\
&= \sigma^2 - (\mathbb{E}[\mu_{MLE}^2] - \mathbb{E}[\mathbb{E}[\mu_{MLE}]^2]) \\
&= \sigma^2 - (\mathbb{E}[\mu_{MLE}^2] - \mathbb{E}[\mu_{MLE}]^2) \\
&= \sigma^2 - \text{Var}[\mu_{MLE}] \\
&= \sigma^2 - \text{Var}\left[\frac{1}{N} \sum_{i=1}^N x_i\right] \\
&= \sigma^2 - \frac{1}{N^2} \text{Var}\left[\sum_{i=1}^N x_i\right] \\
&= \sigma^2 - \frac{1}{N^2} \sum_{i=1}^N \text{Var}[x_i] \\
&= \sigma^2 - \frac{1}{N^2} N\sigma^2 \\
&= \frac{N-1}{N} \sigma^2
\end{aligned} \tag{14}$$

有上述推导我们可以得出， $\sigma_{MLE}^2$  为有偏估计量，而且和真实值比较偏小。为什么会造成这个结果呢？主要原因是出在  $\mu_{MLE}$  上，因为我们在求  $\sigma_{MLE}^2$  时使用的是  $\mu_{MLE}$  而不是  $\mu$ 。而  $\mu_{MLE}$  是拟合数据得到的，所以波动的角度讲，肯定会比使用真实的  $\mu$  算出来要小。所以在高斯分布中，利用极大似然估计得到的  $\sigma_{MLE}^2$  的值，是比真实值偏小的有偏估计。