MLE for ND Gaussian PDF

假設我們有一組在高維空間（維度為 d）的點 $x_i, i=1\cdots n$，若這些點的分佈近似橢球狀，則我們可用高斯密度函數 $g(x_i \mu,\Sigma)$ 來描述產生這些點的機率密度函數：

$$g(x;\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

其中 μ 代表此密度函數的中心點，Σ 則代表此密度函數的共變異矩陣（Covariance Matrix），這些參數決定了此密度函數的特性，如函數形狀的中心點、寬窄及走向等。

我們的目標同前，是要求得 PDF 的最佳參數 $[\mu,\Sigma]$ 以描述所觀察到的資料點。在上述 d 維高斯密度函數的假設下，當 $x=x_i$ 時，其機率密度為 $g(x_i;\mu,\Sigma)$，若我們假設 $x_i, i=1\sim n$ 之間為互相獨立的事件，則發生 $X=\{x_1,x_2\cdots x_n\}$ 的機率密度為

$$p(X;\mu,\Sigma)=\prod_{i=1}^{n}g(x_i;\mu,\Sigma)$$

由於 X 是已經發生之事件，因此我們希望找出 $[\mu,\Sigma]$ 值，使得 $p(X;\mu,\Sigma)$ 能有最大值，此即是 MLE 的原則。（注意：μ 是長度為 d 的未知向量，而 Σ 則是 $d\times d$ 的未知方陣。）

欲求得 $p(X;\mu,\Sigma)$ 的最大值，我們通常將之轉化為求下列 $J(\mu,\Sigma)$ 的最大值：

$$\begin{aligned}J(\mu,\Sigma)&=\ln p(X;\mu,\Sigma)\\&=\ln\left[\prod_{i=1}^{n}g(x_i;\mu,\Sigma)\right]\\&=\sum_{i=1}^{n}\ln g(x_i;\mu,\Sigma)\\&=\sum_{i=1}^{n}\left[-\frac{d}{2}\ln(2\pi)-\frac{1}{2}\ln|\Sigma|-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]\\&=-\frac{nd}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma|-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]\end{aligned}$$

欲求最佳的 μ 值，直接求 $J(\mu,\Sigma)$ 對 μ 的梯度：

$$\begin{aligned}\nabla_\mu J(\mu,\Sigma)&=-\frac{1}{2}\sum_{i=1}^{n}\left[-2\Sigma^{-1}(x_i-\mu)\right]\\&=\Sigma^{-1}\left(\sum_{i=1}^{n}x_i-n\mu\right)\end{aligned}$$

令上式等於零，我們就可以得到

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

欲求最佳的$\sum$值，就不是那麼容易，因為$\sum$是一個$d \times d$的方陣，在此我們僅列出其結果：

$$\hat{\sum} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)\left(x_i - \hat{\mu}\right)^T$$

（對上式推導有興趣的同學，可以參考高等多變分析的相關教科書。）