# 一、极大似然估计

## 多元正态似然

假定 $p\times1$ 向量 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n$ 是一个来自均值向量为$\boldsymbol{\mu}$ 和协方差矩阵为$\boldsymbol{\Sigma}$ 的多元正态总体的随机样本. 由于 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n$ 是相互独立的，且每个为 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 分布，所有观测结果的联合密度函数是边缘正态密度之积：

$$\left\{\begin{matrix}\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\\ \text{的联合密度}\end{matrix}\right\}=\prod_{j=1}^{n}\left\{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}e^{-(\boldsymbol{x}_j-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu})/2}\right\}$$
$$=\frac{1}{(2\pi)^{np/2}}\frac{1}{|\boldsymbol{\Sigma}|^{n/2}}e^{-\sum_{j=1}^{n}(\boldsymbol{x}_j-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu})/2} \tag{4-11}$$

**引理1** 设 $A$ 为 $k\times k$ 维对称矩阵，且 $x\in\mathbb{R}^k$ 则有：

(1) $x^TAx = tr(x^TAx) = tr(Ax^Tx) = tr(Axx^T)$

(2) $tr(A) = \sum_{i=1}^{k}\lambda_i$ $\quad\lambda_i$ 为特征值

**Proof:** (1) $x^TAx\in\mathbb{R}$ 为标量，故

$$x^TAx = tr(x^TAx)$$

加之 tr 迹运算的可交换，证毕 □

(2) $A$ 为对称矩，谱分解为

$$A = P\Lambda P^T$$

有 $tr(A) = tr(P\Lambda P^T) = tr(\Lambda PP^T) = tr(\Lambda)$

$= \sum\lambda_i$ 证毕 □

**引理2** 已知一个 $p\times p$ 的正定矩阵 $B$ 和一个标量 $b>0$，对所有正定矩阵 $\Sigma\in\mathbb{R}^{p\times q}$ 有

$$\frac{1}{|\Sigma|^b}\cdot \exp\left\{-\frac{1}{2}tr(\Sigma^{-1}B)\right\} \le \frac{1}{|B|^b}\cdot(2b)^{pb}\cdot e^{-bp}$$

仅当 $\Sigma = \frac{1}{2b}B$ 取等。

Proof：因为 $B$ 正定 $\Rightarrow B = B^{\frac{1}{2}}B^{\frac{1}{2}}$

有 $tr(\Sigma^{-1}B) = tr(\Sigma^{-1}B^{\frac{1}{2}}B^{\frac{1}{2}}) = tr(B^{\frac{1}{2}}\Sigma^{-1}B^{\frac{1}{2}})$

考虑 $y^T(B^{\frac{1}{2}}\Sigma^{-1}B^{\frac{1}{2}})y = (B^{\frac{1}{2}}y)^T\Sigma^{-1}(B^{\frac{1}{2}}y) > 0$ 当 $y \ne 0$

$(\because B^{\frac{1}{2}}$ 正定 $\Rightarrow B^{\frac{1}{2}}y \ne 0$ 当 $y \ne 0)$

设 $\eta_i$ 为 $B^{\frac{1}{2}}\Sigma^{-1}B^{\frac{1}{2}}$ 的特征值，于是 $\forall \eta_i > 0$

由引理1：$tr(B^{\frac{1}{2}}\Sigma^{-1}B^{\frac{1}{2}}) = \sum \eta_i$

而由行列式：$|B^{\frac{1}{2}}\Sigma^{-1}B^{\frac{1}{2}}| = \prod \eta_i = |B||\Sigma|^{-1}$

有 $\frac{1}{|\Sigma|} = \frac{\prod \eta_i}{|B|}$ 代入原式：

$$\frac{(\prod \eta_i)^b}{|B|^b} e^{-\frac{1}{2}\sum \eta_i} \le \frac{1}{|B|^b}(2b)^{pb}\cdot e^{-bp}$$

(=)

$$\prod \eta_i^b e^{-\frac{1}{2}\eta_i} \le (2b)^{pb} e^{-bp}$$

记 $f(x) = x^b e^{-\frac{x}{2}}$ 求导知 $f(x) \le f(2b)$ $\quad (x^b e^{-\frac{x}{2}})' = bx^{b-1}e^{-\frac{x}{2}} - \frac{1}{2}x^b e^{-\frac{x}{2}}$

有 左式 $= \prod f(\eta_i) \le \prod f(2b) = (2b)^{pb}e^{-bp}$ $\square$ $\quad = e^{-\frac{x}{2}}x^{b-1}(b-\frac{x}{2})$

下面考虑取等条件：

当 $\Sigma = \frac{1}{2b}B$ 时：$B^{\frac{1}{2}}\Sigma^{-1}B^{\frac{1}{2}} = (2b)I_p \Rightarrow \eta_i = 2b$ 代入等号成立。

定理 极大似然估计 设 $X_1, X_2, \cdots, X_n$ 是来自 $N(\mu, \Sigma)$ 的随机样本，$X_i \in \mathbb{R}^p$ 则：

$$\hat{\mu} = \bar{X} \qquad \hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{X})(x_j-\bar{X})^T =: \frac{n-1}{n}S$$

是 $\mu$ 与 $\Sigma$ 的最大似然估计。

Proof: $L(\mu,\Sigma) = (2\pi)^{-\frac{np}{2}} \cdot |\Sigma|^{-\frac{n}{2}} \cdot \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)\right\}$

其中 $(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) = tr\left((x_j-\mu)^T\Sigma^{-1}(x_j-\mu)\right)$

$= tr\left(\Sigma^{-1}(x_j-\mu)(x_j-\mu)^T\right)$

故 $\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu) = \sum_{j=1}^{n} tr\left(\Sigma^{-1}(x_j-\mu)(x_j-\mu)^T\right)$

$= tr\left\{\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)^T\right)\right\}$

其中 $\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)^T = \sum_{j=1}^{n}(x_j-\bar{x}+\bar{x}-\mu)(x_j-\bar{x}+\bar{x}-\mu)^T$

$= \sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^T + \sum_{j=1}^{n}(\bar{x}-\mu)(\bar{x}-\mu)^T + 0$

故 $L(\mu,\Sigma) = (2\pi)^{-\frac{np}{2}}|\Sigma|^{-\frac{n}{2}} \cdot \exp\left\{-\frac{1}{2}\, tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^T + \sum_{j=1}^{n}(\bar{x}-\mu)(\bar{x}-\mu)^T\right)\right]\right\}$

而 $tr\left(\Sigma^{-1}\left(\sum_{j=1}^{n}(\bar{x}-\mu)(\bar{x}-\mu)^T\right)\right) = \sum_{j=1}^{n} tr\left(\Sigma^{-1}(\bar{x}-\mu)(\bar{x}-\mu)^T\right) = n\, tr\left((\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)\right)$

$= n\cdot(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu) \in \mathbb{R}$ 为标量

只有 $(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)$ 与 $\mu$ 有关，当 $\hat{\mu}=\bar{x}$ 时，$L(\mu,\Sigma)$ 变大

故 $L(\bar{x},\Sigma) = (2\pi)^{-np/2}|\Sigma|^{-\frac{n}{2}} \cdot \exp\left\{-\frac{1}{2}\, tr\left(\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^T\right)\right) + 0\right\}$

由引理 2：$b = \frac{n}{2}$ $\quad B = \sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^T$

取等条件：$\hat{\Sigma} = \frac{1}{2b}B = \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^T$

故 $\max L(\mu,\Sigma) = L(\hat{\mu},\hat{\Sigma})$：

$$\hat{\mu} = \bar{X} \qquad \hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{X})(x_j-\bar{X})^T =: \frac{n-1}{n}S \qquad \square$$

另证：

$$L(\mu,\Sigma) = (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)\right\}$$

$$\log L(\mu,\Sigma) = -\frac{p}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)^T\Sigma^{-1}(x_j-\mu)$$

当 $\hat{\mu} = \bar{x}$ 有

$$\log L(\hat{\mu},\Sigma) = -\frac{p}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\underbrace{\sum_{j=1}^{n}(x_j-\bar{x})^T\Sigma^{-1}(x_j-\bar{x})}_{tr(\Sigma^{-1}\cdot n\hat{\Sigma})}$$

$$\frac{\partial \det(X)}{\partial X} = \det(X)\cdot(X^{-1})^T \qquad \frac{\partial x^T B x}{\partial x} = (B+B^T)x$$

$$\frac{\partial\, tr(X^{-1})}{\partial X} = -(X^{-1}X^{-1})^T \qquad \frac{\partial\, tr(AX^{-1}B)}{\partial X} = -(X^{-1}BAX^{-1})^T$$

$$\frac{\partial L(\hat{\mu},\Sigma)}{\partial \Sigma} = -\frac{n}{2}\cdot|\Sigma|\cdot(\Sigma^{-1})^T\cdot\frac{1}{|\Sigma|} + \frac{1}{2}(\Sigma^{-1}n\hat{\Sigma}\Sigma^{-1})^T$$

$$= -\frac{n}{2}(\Sigma^{-1} - \Sigma^{-1}\hat{\Sigma}\Sigma^{-1})$$

当 $\hat{\Sigma} = \Sigma$ 时 上式为 0

## 二、极大似然估计量的性质

### (一). 不变性

极大似然估计量具有不变性，设 $\hat{\theta}$ 是 $\theta$ 的 MLE

考虑函数 $h(\theta)$，则 $h(\theta)$ 的 MLE 也为 $\hat{h}_{MLE}(\theta) = h(\hat{\theta})$

### (二). 充分统计量

若 $T(X)$ 是参数 $\theta$ 的充分统计量，则任何关于 $\theta$ 的推断都应由 $T(X)$ 完全给出（在样本 $X$ 上）。也即若 $X$ 与 $Y$ 为 2 个样本，且 $T(X)$ 与 $T(Y)$ 是 $\theta$ 的统计量，则有 $T(X) = T(Y)$

例：设 $X_1, X_2, \cdots, X_n$ 是来自 $N(\mu, \Sigma)$ 的随机样本，则 $\bar{X}$，$S$ 为充分统计量。

### (三). $\bar{X}$ 与 $S$ 的抽样分布

① 一元情形：

$X_1, X_2, \cdots, X_n \ iid \sim N(\mu, \sigma^2)$ 记：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

则：(1) $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

(2) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

(3) $\bar{X} \perp S^2$

(4) $\frac{\sqrt{n}\cdot(\bar{X} - \mu)}{S} \sim t(n-1)$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2}\Big/(n-1)} \sim t(n-1)$$

② 多元情形：

设 $X_1, X_2, \cdots X_n$ 为 n 个来自 $N_p(\mu, \Sigma)$ 的随机向量，$X_i \in \mathbb{R}^p$

记：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T$$

则有：

(1) $\bar{X} \sim N_p(\mu, \frac{1}{n}\Sigma)$

(2) $(n-1)S \sim W_{n-1}(0, \Sigma)$

(3) $\bar{X} \perp S^2$

$Cov(X_i-\bar{X})$

$= \Sigma + \frac{1}{n}\Sigma - Cov(X_i, \bar{X})$

$= \Sigma + \frac{1}{n}\Sigma - \frac{1}{n}\Sigma - 0$

$n(\bar{X}-\mu)^T S^{-1}(\bar{X}-\mu) \overset{a}{\sim} \chi^2(p)$

$X_i - \bar{X} \sim N_p(0, \Sigma)$

## (四) Wishart 分布

设 $X_1, X_2, \cdots, X_n$ 为 $N_p(\mu, \Sigma)$ 的样本 $X_i \in \mathbb{R}^p$ 则

$$W = \sum_{i=1}^{n} X_i X_i^T$$

服从自由度为 n 的 Wishart 分布，记作 $W \sim W_n(\Delta, \Sigma)$

其中：

$$\Delta = \mu\mu^T$$

当 $\mu = 0$ 时，称为中心 Wishart 分布，记作 $W_n(0, \Sigma)$

性质：

① 若 $A_j \sim W_{n_j}(\Delta_j, \Sigma)$ $j=1,2,\cdots,m$ 且 $A_j$ 相互独立，则有

$$\sum_{j=1}^{m} A_j \sim W_n(\Delta, \Sigma)$$ 独立可加

其中 $n = \sum_{j=1}^{m} n_j$ $\Delta = \sum_{j=1}^{m} \Delta_j$

② 若 $W \sim W_n(\Delta, \Sigma)$ 则有： $W = \sum X_i X_i^T$

$$CWC^T \sim W_n(C\Delta C^T, C\Sigma C^T)$$

$CWC^T = \sum (CX_i)(CX_i)^T$

## 三、$\overline{X}$ 与 $S$ 的大样本特性

$\mu\mu^T \to (C\mu)(C\mu)^T$

### (一) 大数定律

$Cov(CX_i) = C\Sigma C^-$

设 $Y_1, Y_2, \cdots, Y_n$ 是来自 $E(Y_i) = \mu$ 总体的独立观测，则有

$$\overline{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)$$

满足： $P(|\overline{Y} - \mu| < \varepsilon) \longrightarrow 1 \quad \forall \varepsilon$ 在 $n \to \infty$，即为

依概率收敛： $\overline{Y} \xrightarrow{P} \mu$

多元情形：

$$\overline{X} \xrightarrow{P} \mu \in \mathbb{R}^p$$

$$S \xrightarrow{P} \Sigma \in \mathbb{R}^{p \times p}$$

具体而言： $X_i \xrightarrow{P} \mu_i \qquad S_{ij} \xrightarrow{P} \sigma_{ij}$

### (二) 中心极限定理

设 $X_1, X_2, \cdots, X_n$ 来自均值为 $\mu$，协方差阵为 $\Sigma$ 的独立观测，则有

$$\sqrt{n}(X - \mu) \overset{a}{\sim} N_p(0, \Sigma)$$

要求 $n \to \infty$ 且 $n >> p$ （ $\overset{a}{\sim}$ 表示依分布收敛）

特别地，对 $S$ 有 Chi-Square 近似

$$n(\overline{X} - \mu)^T S^{-1} (\overline{X} - \mu) \overset{a}{\sim} \chi^2(p)$$

# 四、评估正态性假定

## (一) Q-Q 图　　Normal

作 Q-Q Plot 步骤：

(1) 原始观测值排序：$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$

定们对应的概率：$P(X \le x_{(j)}) = \frac{j-\frac{1}{2}}{n}$

(2) 标准正态分位数：$q_{(1)} \le q_{(2)} \le \cdots \le q_{(n)}$

(3) 作图：$(q_{(i)}, x_{(i)})$

理应有 $x_{(j)} = \mu + \sigma \cdot q_{(j)}$ 线性关系（若 $X \sim N(\mu, \sigma)$）

## Q-Q Plot 相关系数

也可从 $q_{(j)}$ 与 $x_{(j)}$ 的相关系数评估 $X$ 是否近似

$$r_Q = \frac{\sum_{j=1}^{n}(x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^{n}(x_{(j)} - \bar{x})^2} \cdot \sqrt{\sum_{j=1}^{n}(q_{(j)} - \bar{q})^2}}$$

查表判断：

表 4.2　对正态性的 Q-Q 图相关系数检验的临界点

| 样本容量 $n$ | 显著性水平 $\alpha$ | | |
|---|---|---|---|
| | 0.01 | 0.05 | 0.10 |
| 5 | 0.829 9 | 0.878 8 | 0.903 2 |
| 10 | 0.880 1 | 0.919 8 | 0.935 1 |
| 15 | 0.912 6 | 0.938 9 | 0.950 3 |
| 20 | 0.926 9 | 0.950 8 | 0.960 4 |
| 25 | 0.941 0 | 0.959 1 | 0.966 5 |
| 30 | 0.947 9 | 0.965 2 | 0.971 5 |
| 35 | 0.953 8 | 0.968 2 | 0.974 0 |
| 40 | 0.959 9 | 0.972 6 | 0.977 1 |
| 45 | 0.963 2 | 0.974 9 | 0.979 2 |
| 50 | 0.967 1 | 0.976 8 | 0.980 9 |
| 55 | 0.969 5 | 0.978 7 | 0.982 2 |
| 60 | 0.972 0 | 0.980 1 | 0.983 6 |
| 75 | 0.977 1 | 0.983 8 | 0.986 6 |
| 100 | 0.982 2 | 0.987 3 | 0.989 5 |
| 150 | 0.987 9 | 0.991 3 | 0.992 8 |
| 200 | 0.990 5 | 0.993 1 | 0.994 2 |
| 300 | 0.993 5 | 0.995 3 | 0.996 0 |

当 $r_Q < r^*$ 时，拒绝正态性假定

## (二). 评估二元或多元正态性 $\chi^2$

样本 $X_1, X_2, \cdots, X_n \sim N_p(\mu, \Sigma)$ 则有

$$d_i^2 = (X_i-\mu)^T \Sigma^{-1} (X_i-\mu) \sim \chi^2(p)$$

$$\hat{d}_i^2 = (X_i-\bar{X})^T S^{-1} (X_i-\bar{X}) \overset{a}{\sim} \chi^2(p)$$

类似地画 Q-Q Plot：

(1). 观测值计算后排序：$\hat{d}^2_{(1)} \le \hat{d}^2_{(2)} \le \cdots \le \hat{d}^2_{(n)}$

它们的概率：$P(D \le \hat{d}^2_{(j)}) = \dfrac{j-0.5}{n}$

(2). $\chi^2$ 分布的分位点：$q_{\chi^2(p)}\left(\dfrac{j-0.5}{n}\right)$

(3) 作图 $\left(q_{\chi^2(p)}\left(\dfrac{j-0.5}{n}\right), \hat{d}^2_{(j)}\right)$

当大约有一半的 $d_j^2$ 小于或等于 $q_{\chi^2(p)}(0.5)$；图像近似是一条通过原点且斜率为 1 的直线.

# 五、数据清洗 & 异常值点

## (一). 一元数据

$Q_1$ $Q_3$ 为四分位点，则

$$IQR = Q_3 - Q_1$$

于是在 $(Q_1 - 1.5\,IQR,\ Q_3 + 1.5\,IQR)$ 之外的点，可经验地认为是异常值点.

## (二). 多元数据

对 $\hat{d}_j^2$ 进行检验

六. 变换到近似正态