

Multivariate Statistical Analysis

多元统计分析

2026年4月14日

已学知识点 (Recap)

第 5 章 多元正态分布的理论

- ▶ $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的概率密度函数为 $f(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$
 $\mathbb{E}(X) = \boldsymbol{\mu}$
 $\text{Cov}(X) = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$
- ▶ 设 $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 矩阵 $\mathcal{A}_{p \times p}$, 向量 $\mathbf{c} \in \mathbb{R}^p$, 则 $Y = \mathcal{A}X + \mathbf{c} \sim N_p(\mathcal{A}\boldsymbol{\mu} + \mathbf{c}, \mathcal{A}\boldsymbol{\Sigma}\mathcal{A}^T)$.
- ▶ 设 $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 作 Mahalanobis 变换: $Y = \boldsymbol{\Sigma}^{-1/2} (X - \boldsymbol{\mu})$, 则 $Y \sim N_p(\mathbf{0}, \mathcal{I}_p)$, 且
 $Y^T Y = (X - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (X - \boldsymbol{\mu}) \sim \chi_p^2$.

已学知识点 (Recap)

第 5 章 多元正态分布的理论

- 设 $X \sim N_p(\mu, \Sigma)$, 分块如下

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

$r \times 1$ (pointing to X_1 and μ_1) $r \times r$ (pointing to Σ_{11})

$$\Sigma_{22 \cdot 1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$$

则 $X_1 \sim N_r(\mu_1, \Sigma_{11})$, $X_{2 \cdot 1} = X_2 - \Sigma_{21} \Sigma_{11}^{-1} X_1 \sim N_{p-r}(\mu_{2 \cdot 1}, \Sigma_{22 \cdot 1})$,

且 X_1 与 $X_{2 \cdot 1}$ 相互独立.

$$\mu_{2 \cdot 1} = \mu_2 - \Sigma_{21} \Sigma_{11}^{-1} \mu_1$$

- X_1 与 X_2 相互独立的充分必要条件是 $\Sigma_{12} = \mathbf{0}$.
- 对于给定的矩阵 \mathcal{A} 与 \mathcal{B} , $\mathcal{A}X$ 与 $\mathcal{B}X$ 相互独立的充分必要条件是 $\mathcal{A} \Sigma \mathcal{B}^T = \mathbf{0}$.
- 给定矩阵 $\mathcal{A}_{q \times p}$, 向量 $c \in \mathbb{R}^q$, 且 $q \leq p$, 则 $Y = \mathcal{A}X + c \sim N_q(\mathcal{A}\mu + c, \mathcal{A} \Sigma \mathcal{A}^T)$.
- 给定 $X_1 = x_1$ 时 X_2 的条件分布 $(X_2 | X_1 = x_1) \sim N_{p-r}(\mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (x_1 - \mu_1), \Sigma_{22 \cdot 1})$.

已学知识点 (Recap)

第 5 章 多元正态分布的理论

- ▶ 设 $X_1 \sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, $(X_2 | X_1 = \mathbf{x}_1) \sim N_{p-r}(\mathcal{A}\mathbf{x}_1 + \mathbf{b}, \boldsymbol{\Omega})$, 其中 $\mathcal{A}_{(p-r) \times r}$, $\mathbf{b} \in \mathbb{R}^{p-r}$,

$\boldsymbol{\Omega}$ 不依赖于 \mathbf{x}_1 , 则 $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{11}\mathcal{A}^T \\ \mathcal{A}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Omega} + \boldsymbol{\Sigma}_{11}\mathcal{A}^T \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \mathcal{A}\boldsymbol{\mu}_1 + \mathbf{b} \end{pmatrix}$$

- ▶ **随机矩阵的分布**: 随机矩阵 $\mathcal{X} = (X_{ij})_{m \times n}$ 的分布定义为其所有元素构成的随机向量

$X = (X_{11}, \dots, X_{1n}, X_{21}, \dots, X_{2n}, \dots, X_{m1}, \dots, X_{mn})$ 的联合分布.

已学知识点 (Recap)

第 5 章 多元正态分布的理论

▶ 非中心 Wishart 分布 $W_p(\Sigma, n, \boldsymbol{\tau})$:

$$\boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{ip} \end{pmatrix}, \quad i = 1, 2, \dots, n$$

○ 假设 $\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_i, \Sigma), \quad i = 1, 2, \dots, n.$

○ 记 $\mathcal{M}_{n \times p} = \begin{pmatrix} \boldsymbol{\mu}_1^T \\ \boldsymbol{\mu}_2^T \\ \vdots \\ \boldsymbol{\mu}_n^T \end{pmatrix} = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1p} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n1} & \mu_{n2} & \cdots & \mu_{np} \end{pmatrix},$

非中心参数

$$\boldsymbol{\tau}_{p \times p} = \mathcal{M}^T \mathcal{M}$$

$$= (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_n) \begin{pmatrix} \boldsymbol{\mu}_1^T \\ \boldsymbol{\mu}_2^T \\ \vdots \\ \boldsymbol{\mu}_n^T \end{pmatrix}$$

$$= \sum_{i=1}^n \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T$$

○ 则称随机矩阵 $\mathcal{A}_{p \times p} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$ 服从非中心 Wishart 分布 $W_p(\Sigma, n, \boldsymbol{\tau})$.

▶ 当 $\boldsymbol{\mu}_i = \mathbf{0}, i = 1, 2, \dots, n$ 时 $\boldsymbol{\tau} = \mathbf{0}$, 此时称 \mathcal{A} 服从中心 Wishart 分布 $\mathcal{A} \sim W_p(\Sigma, n)$.

已学知识点 (Recap)

第 5 章 多元正态分布的理论

随机矩阵 \mathcal{A} 的线性变换

▶ 若 $\mathcal{A} \sim W_p(\Sigma, n)$, 矩阵 $\mathcal{B}_{p \times q}$, 则 $\mathcal{B}^T \mathcal{A} \mathcal{B} \sim W_q(\mathcal{B}^T \Sigma \mathcal{B}, n)$.

▶ 设 $\mathcal{A} \sim W_p(\Sigma, m)$, $\mathbf{a} \in \mathbb{R}^p$ 满足 $\mathbf{a}^T \Sigma \mathbf{a} \neq 0$, 则 $\frac{\mathbf{a}^T \mathcal{A} \mathbf{a}}{\mathbf{a}^T \Sigma \mathbf{a}}$ 服从 χ_m^2 分布.

▶ **Cochran 定理**: 设 $\mathcal{X}_{n \times p}$ 是来自正态分布总体 $N_p(\mathbf{0}, \Sigma)$ 的一个数据矩阵, 设 $\mathcal{C}_{n \times n}$ 是一个对称矩阵.

① $\mathcal{X}^T \mathcal{C} \mathcal{X}$ 服从加权 Wishart 随机变量和的分布, 即

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

$$\mathcal{X}^T \mathcal{C} \mathcal{X} = \sum_{i=1}^n \lambda_i W_p(\Sigma, 1),$$

其中 λ_i ($i = 1, 2, \dots, n$) 是 \mathcal{C} 的特征值.

已学知识点 (Recap)

第 5 章 多元正态分布的理论

随机矩阵 \mathcal{A} 的线性变换

▶ 若 $\mathcal{A} \sim W_p(\Sigma, n)$, 矩阵 $\mathcal{B}_{p \times q}$, 则 $\mathcal{B}^T \mathcal{A} \mathcal{B} \sim W_q(\mathcal{B}^T \Sigma \mathcal{B}, n)$.

▶ 设 $\mathcal{A} \sim W_p(\Sigma, m)$, $\mathbf{a} \in \mathbb{R}^p$ 满足 $\mathbf{a}^T \Sigma \mathbf{a} \neq 0$, 则 $\frac{\mathbf{a}^T \mathcal{A} \mathbf{a}}{\mathbf{a}^T \Sigma \mathbf{a}}$ 服从 χ_m^2 分布.

▶ **Cochran 定理**: 设 $\mathcal{X}_{n \times p}$ 是来自正态分布总体 $N_p(\mathbf{0}, \Sigma)$ 的一个数据矩阵, 设 $\mathcal{C}_{n \times n}$ 是一个对称矩阵.

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

② $\mathcal{X}^T \mathcal{C} \mathcal{X}$ 服从 Wishart 分布的充分必要条件是 $\mathcal{C}^2 = \mathcal{C}$.

此时有 $\mathcal{X}^T \mathcal{C} \mathcal{X} \sim W_p(\Sigma, r)$, 其中 $r = \text{rank}(\mathcal{C}) = \text{tr}(\mathcal{C})$.

③ $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \mathcal{X}^T \mathbf{1}_n \sim N_p\left(\mathbf{0}, \frac{1}{n} \Sigma\right)$.

④ $n \mathcal{S} = \mathcal{X}^T \mathcal{H} \mathcal{X}$ 服从 Wishart 分布 $W_p(\Sigma, n - 1)$.

$$\mathcal{H} = \mathcal{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$$

⑤ $\bar{\mathbf{x}}$ 与 \mathcal{S} 相互独立.

已学知识点 (Recap)

第 5 章 多元正态分布的理论

► Wishart 分布的性质:

① 如果 $\mathcal{A} \sim W_p(\Sigma, n)$, 则 $\mathbb{E}(\mathcal{A}) = n\Sigma$.

② 如果 $\mathcal{A}_i \sim W_p(\Sigma, n_i)$, $i = 1, 2, \dots, k$, 则 $\mathcal{A} = \sum_{i=1}^k \mathcal{A}_i \sim W_p(\Sigma, n)$, 其中 $n = \sum_{i=1}^k n_i$.

③ \mathcal{A} 为正定矩阵时, $W_p(\Sigma, n-1)$ 的概率密度函数为

$$f_{\Sigma, n-1}(\mathcal{A}) = \frac{|\mathcal{A}|^{\frac{1}{2}(n-p-2)} e^{-\frac{1}{2}\text{tr}(\mathcal{A}\Sigma^{-1})}}{2^{\frac{1}{2}p(n-1)} \pi^{\frac{1}{4}p(p-1)} |\Sigma|^{\frac{1}{2}(n-1)} \prod_{i=1}^p \Gamma\left(\frac{n-i}{2}\right)}$$

已学知识点 (Recap)

第 5 章 多元正态分布的理论

▶ Hotelling T^2 分布: 若 $Y \sim N_p(\mathbf{0}, \mathcal{I})$, $\mathcal{A} \sim W_p(\mathcal{I}, n)$, 且 Y 与 \mathcal{A} 相互独立, 定义 $n Y^T \mathcal{A}^{-1} Y \sim T_{p, n}^2$.

▶ 如果 $X \sim N_p(\boldsymbol{\mu}, \Sigma)$ 与 $\mathcal{A} \sim W_p(\Sigma, n)$ 相互独立, 则 $n(X - \boldsymbol{\mu})^T \mathcal{A}^{-1}(X - \boldsymbol{\mu}) \sim T_{p, n}^2$.

▶ 若 $\bar{\mathbf{x}}$ 是取自正态总体 $N_p(\boldsymbol{\mu}, \Sigma)$ 的一个样本的样本均值向量, \mathcal{S} 是样本协方差矩阵, 则

$$(n-1)(\bar{\mathbf{x}} - \boldsymbol{\mu})^T \mathcal{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) = n(\bar{\mathbf{x}} - \boldsymbol{\mu})^T \mathcal{S}_u^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim T_{p, n-1}^2. \quad \mathcal{S}_u = \frac{n}{n-1} \mathcal{S}$$

▶ Hotelling T^2 与 F 分布的关系为 $T_{p, n}^2 = \frac{np}{n-p+1} F_{p, n-p+1}$.

▶ 对 $X \sim N_p(\boldsymbol{\mu}, \Sigma)$ 的一个线性变换 $Y = \mathcal{C}X$, 其中 $\mathcal{C}_{q \times p}$ 满足 $q \leq p$. 如果 $\bar{\mathbf{x}}$ 与 \mathcal{S}_X 分别为样本均值向量与样本协方差矩阵, 则有

$$\bar{\mathbf{y}} = \mathcal{C} \bar{\mathbf{x}} \sim N_q\left(\mathcal{C} \boldsymbol{\mu}, \frac{1}{n} \mathcal{C} \Sigma \mathcal{C}^T\right)$$

$$n \mathcal{S}_Y = n \mathcal{C} \mathcal{S}_X \mathcal{C}^T \sim W_q(\mathcal{C} \Sigma \mathcal{C}^T, n-1)$$

$$(n-1)(\mathcal{C} \bar{\mathbf{x}} - \mathcal{C} \boldsymbol{\mu})^T (\mathcal{C} \mathcal{S}_X \mathcal{C}^T)^{-1} (\mathcal{C} \bar{\mathbf{x}} - \mathcal{C} \boldsymbol{\mu}) \sim T_{q, n-1}^2$$

Chapter 6 Theory of Estimation

估计理论

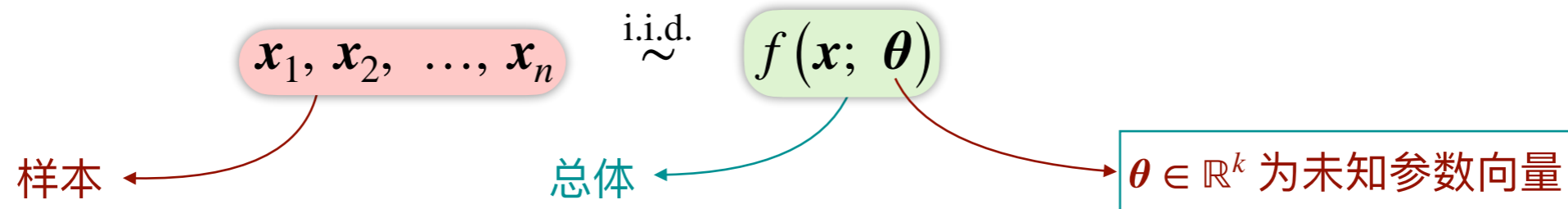
概述

估计理论

似然函数

Cramer-Rao 下界

似然函数



- **目的**: 估计未知参数 $\boldsymbol{\theta} \in \mathbb{R}^k$.

数据矩阵: $\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$

- **似然函数**: 样本 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ 的联合密度 $L(\mathcal{X}; \boldsymbol{\theta})$ 作为 $\boldsymbol{\theta}$ 的一个函数

$$L(\mathcal{X}; \boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{x}_i; \boldsymbol{\theta})$$

似然函数

- 极大似然估计 (MLE):

$$\hat{\theta} = \arg \max_{\theta} L(\mathcal{X}; \theta)$$

- ▶ 一般情形下, 通过使对数似然函数

$$\ell(\mathcal{X}; \theta) = \log L(\mathcal{X}; \theta)$$

达到最大更容易求解, 这是因为对数函数单调且一对一.

$$\Rightarrow \hat{\theta} = \arg \max_{\theta} L(\mathcal{X}; \theta) = \arg \max_{\theta} \ell(\mathcal{X}; \theta)$$

似然函数

- 例: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \mathcal{F})$.

pdf $\Rightarrow f(\mathbf{x}; \boldsymbol{\theta}) = (2\pi)^{-\frac{p}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\theta})^T (\mathbf{x} - \boldsymbol{\theta}) \right\}$

似然函数 $\Rightarrow L(\mathcal{X}; \boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{x}_i; \boldsymbol{\theta})$

$$= (2\pi)^{-\frac{np}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \boldsymbol{\theta})^T (\mathbf{x}_i - \boldsymbol{\theta}) \right] \right\}$$

对数似然函数 $\Rightarrow \ell(\mathcal{X}; \boldsymbol{\theta}) = -\frac{np}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \boldsymbol{\theta})^T (\mathbf{x}_i - \boldsymbol{\theta}) \right]$

$$\begin{aligned}
 (\mathbf{x}_i - \boldsymbol{\theta})^T (\mathbf{x}_i - \boldsymbol{\theta}) &= (\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\theta})^T (\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\theta}) \\
 &= \left[(\mathbf{x}_i - \bar{\mathbf{x}})^T + (\bar{\mathbf{x}} - \boldsymbol{\theta})^T \right] \left[(\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\theta}) \right] \\
 &= (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\bar{\mathbf{x}} - \boldsymbol{\theta}) + 2 (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\mathbf{x}_i - \bar{\mathbf{x}})
 \end{aligned}$$

似然函数

- 例: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \mathcal{F})$.

$$\begin{aligned}
 \ell(\mathcal{X}; \boldsymbol{\theta}) &= -\frac{np}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\bar{\mathbf{x}} - \boldsymbol{\theta}) + 2 (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right] \\
 &= -\frac{np}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right] - \frac{n}{2} (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\bar{\mathbf{x}} - \boldsymbol{\theta}) - \sum_{i=1}^n \left[(\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right]
 \end{aligned}$$

$\rightarrow = 0$

对数似然函数 $\Rightarrow \ell(\mathcal{X}; \boldsymbol{\theta}) = -\frac{np}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \boldsymbol{\theta})^T (\mathbf{x}_i - \boldsymbol{\theta}) \right]$

$$\begin{aligned}
 (\mathbf{x}_i - \boldsymbol{\theta})^T (\mathbf{x}_i - \boldsymbol{\theta}) &= (\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\theta})^T (\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\theta}) \\
 &= \left[(\mathbf{x}_i - \bar{\mathbf{x}})^T + (\bar{\mathbf{x}} - \boldsymbol{\theta})^T \right] \left[(\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\theta}) \right] \\
 &= (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\bar{\mathbf{x}} - \boldsymbol{\theta}) + 2 (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\mathbf{x}_i - \bar{\mathbf{x}})
 \end{aligned}$$

似然函数

- 例: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \mathcal{F})$.

$$\begin{aligned}
 \ell(\mathcal{X}; \boldsymbol{\theta}) &= -\frac{np}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\bar{\mathbf{x}} - \boldsymbol{\theta}) + 2 (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right] \\
 &= -\frac{np}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right] - \frac{n}{2} (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\bar{\mathbf{x}} - \boldsymbol{\theta}) - \sum_{i=1}^n \left[(\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right] \\
 &= -\frac{np}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right] - \frac{n}{2} (\bar{\mathbf{x}} - \boldsymbol{\theta})^T (\bar{\mathbf{x}} - \boldsymbol{\theta})
 \end{aligned}$$

只有该项与 $\boldsymbol{\theta}$ 有关

- ▶ 显然, $l(\mathcal{X}; \boldsymbol{\theta})$ 在 $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ 取得极大值.
- ▶ 密度函数为上述 $f(\mathbf{x}; \boldsymbol{\theta})$ 的分布中, 参数 $\boldsymbol{\theta} = \boldsymbol{\mu}$ 的极大似然估计量 (MLE) 为

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$$

似然函数

- 例: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$\text{pdf} \implies f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

- ▶ $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 这里将 $\boldsymbol{\Sigma}$ 看作是一个向量. 因为 $\boldsymbol{\Sigma}$ 对称, 未知参数 $\boldsymbol{\theta}$ 实际上只有

$$\left\{ p + \frac{1}{2}p(p+1) \right\} \text{ 维.}$$

- ▶ 似然函数为

$$L(\mathcal{X}; \boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{x}_i; \boldsymbol{\theta}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \right\}$$

- ▶ 从而得对数似然函数

$$\ell(\mathcal{X}; \boldsymbol{\theta}) = \log L(\mathcal{X}; \boldsymbol{\theta}) = -\frac{n}{2} \log |2\pi\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right]$$

似然函数

- 例: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$\begin{aligned}
 (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) &= (\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu}) \\
 &= \left[(\mathbf{x}_i - \bar{\mathbf{x}})^T + (\bar{\mathbf{x}} - \boldsymbol{\mu})^T \right] \boldsymbol{\Sigma}^{-1} \left[(\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\mu}) \right] \\
 &= (\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) + 2 (\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})
 \end{aligned}$$

$$\begin{aligned}
 \sum_{i=1}^n \left[(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] &= \sum_{i=1}^n \left[(\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \right] + \\
 &\quad n (\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) + 2 \sum_{i=1}^n \left[(\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \right]
 \end{aligned}$$

▶ 从而得对数似然函数

$$\ell(\mathcal{X}; \boldsymbol{\theta}) = \log L(\mathcal{X}; \boldsymbol{\theta}) = -\frac{n}{2} \log |2\pi\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right]$$

= 0

似然函数

- 例: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$\begin{aligned}
 & \mathcal{A}_{n \times p}, \mathcal{B}_{p \times q}, \mathcal{C}_{q \times n} \implies \\
 & \text{tr}(\mathcal{A}\mathcal{B}\mathcal{C}) = \text{tr}(\mathcal{B}\mathcal{C}\mathcal{A}) = \text{tr}(\mathcal{C}\mathcal{A}\mathcal{B})
 \end{aligned}$$

$$\begin{aligned}
 \left[(\mathbf{x}_i - \bar{\mathbf{x}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \right]_{1 \times 1} &= \text{tr} \left[(\mathbf{x}_i - \bar{\mathbf{x}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \right] \\
 &= \text{tr} \left[\boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^\top \right]
 \end{aligned}$$

$$\begin{aligned}
 \sum_{i=1}^n \left[(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] &= \sum_{i=1}^n \left[(\mathbf{x}_i - \bar{\mathbf{x}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \right] + n (\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \\
 &= \sum_{i=1}^n \text{tr} \left[\boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^\top \right] + n (\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})
 \end{aligned}$$

▶ 从而得对数似然函数

$$\ell(\mathcal{X}; \boldsymbol{\theta}) = \log L(\mathcal{X}; \boldsymbol{\theta}) = -\frac{n}{2} \log |2\pi\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right]$$

似然函数

- 例: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$\begin{aligned}
 &= \text{tr} \{ \boldsymbol{\Sigma}^{-1} n \mathcal{S} \} + n (\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \\
 &= \text{tr} \left\{ \boldsymbol{\Sigma}^{-1} \sum_{i=1}^n [(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T] \right\} + n (\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})
 \end{aligned}$$

$$\begin{aligned}
 \sum_{i=1}^n [(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})] &= \sum_{i=1}^n [(\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})] + n (\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \\
 &= \sum_{i=1}^n \text{tr} [\boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T] + n (\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})
 \end{aligned}$$

- 从而得对数似然函数

$$\ell(\mathcal{X}; \boldsymbol{\theta}) = \log L(\mathcal{X}; \boldsymbol{\theta}) = -\frac{n}{2} \log |2\pi\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^n [(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})]$$

似然函数

- 例: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- ▶ 可求得极大似然估计量为

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}, \quad \hat{\boldsymbol{\Sigma}} = \mathcal{S}$$

- ▶ 注意: 协方差矩阵的无偏估计量 $\mathcal{S}_u = \frac{n}{n-1} \mathcal{S}$ 不是 $\boldsymbol{\Sigma}$ 的极大似然估计量!

- ▶ 从而得对数似然函数

$$\ell(\mathcal{X}; \boldsymbol{\theta}) = -\frac{n}{2} \log |2\pi\boldsymbol{\Sigma}| - \frac{n}{2} \text{tr}(\boldsymbol{\Sigma}^{-1}\mathcal{S}) - \frac{n}{2} (\bar{\mathbf{x}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})$$

似然函数

- 例：考虑线性回归模型

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$\mathbf{x}_i \in \mathbb{R}^p$ ← (red arrow from \mathbf{x}_i to \mathbb{R}^p)
 → $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ (teal arrow from ε_i to $N(0, \sigma^2)$)

- ▶ 参数： $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma)$ 是一个 $(p + 1)$ 维的参数向量.

- ▶ 数据：

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

- ▶ 似然函数：

$$L(\mathbf{y}, \mathcal{X}; \boldsymbol{\theta}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi} \sigma} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 \right\}$$

似然函数

- 例：考虑线性回归模型

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$\mathbf{x}_i \in \mathbb{R}^p$ ← (red arrow from \mathbf{x}_i to \mathbb{R}^p)
 → (blue arrow from ε_i to $N(0, \sigma^2)$)
 $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

- ▶ 对数似然函数：

$$\begin{aligned}
 \ell(y, \mathcal{X}; \boldsymbol{\theta}) &= \log \left\{ \prod_{i=1}^n \frac{1}{\sqrt{2\pi} \sigma} \exp \left[-\frac{1}{2\sigma^2} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 \right] \right\} \\
 &= -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 \\
 &= -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} (\mathbf{y} - \mathcal{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathcal{X}\boldsymbol{\beta}) \\
 &= -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} (\mathbf{y}^T \mathbf{y} + \boldsymbol{\beta}^T \mathcal{X}^T \mathcal{X} \boldsymbol{\beta} - 2\boldsymbol{\beta}^T \mathcal{X}^T \mathbf{y})
 \end{aligned}$$

似然函数

- 例：考虑线性回归模型

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$\mathbf{x}_i \in \mathbb{R}^p$ (indicated by a red arrow pointing to \mathbf{x}_i)
 $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ (indicated by a blue arrow pointing to ε_i)

- ▶ 关于参数求导得：

$$\begin{cases} \frac{\partial \ell}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2} (2\mathcal{X}^T \boldsymbol{\beta} - 2\mathcal{X}^T \mathbf{y}) \\ \frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} [(\mathbf{y} - \mathcal{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathcal{X}\boldsymbol{\beta})] \end{cases}$$

与最小二乘估计量相同

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{0} \implies \mathcal{X}^T \boldsymbol{\beta} - \mathcal{X}^T \mathbf{y} = \mathbf{0} \implies \hat{\boldsymbol{\beta}} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathbf{y}$$

$$\frac{\partial \ell}{\partial \sigma} = 0 \implies \sigma^2 = \frac{1}{n} (\mathbf{y} - \mathcal{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathcal{X}\boldsymbol{\beta})$$

$$\implies \hat{\sigma}^2 = \frac{1}{n} (\mathbf{y} - \mathcal{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathcal{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n} \|\mathbf{y} - \mathcal{X}\hat{\boldsymbol{\beta}}\|^2$$

残差平方和

似然函数

- 例：考虑线性回归模型

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$\mathbf{x}_i \in \mathbb{R}^p$ ← (red arrow) (green arrow) → $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} \\ \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2p} \\ \vdots \\ \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \iff \mathbf{y} = \mathcal{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

→ $\sim N_n(\mathbf{0}, \sigma^2 \mathcal{I}_n)$

- ▶ 当 \mathbf{x}_i 的值给定时，我们有

$$\mathbf{E}(\mathbf{y}) = \mathcal{X}\boldsymbol{\beta} \implies \mathbf{E}(\hat{\boldsymbol{\beta}}) = \mathbf{E}\left[(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathbf{y}\right] = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathbf{E}(\mathbf{y}) = \boldsymbol{\beta}$$

$$\begin{aligned}
 \text{Var}(\mathbf{y}) = \sigma^2 \mathcal{I}_n \implies \text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}\left[(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathbf{y}\right] \\
 &= (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \text{Var}(\mathbf{y}) \left[(\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T\right]^T \\
 &= (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \sigma^2 \mathcal{I}_n \mathcal{X} (\mathcal{X}^T \mathcal{X})^{-1} = \sigma^2 (\mathcal{X}^T \mathcal{X})^{-1}
 \end{aligned}$$

Cramer-Rao 下界

- 在估计理论中，一个重要的问题是估计量 $\hat{\theta}$ 是否具有某些特定的性质.

- ▶ 无偏性: $E(\hat{\theta}) = \theta$

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} \text{ 无偏估计量}$$

$$\implies \hat{\boldsymbol{\Sigma}} = \mathcal{S} \text{ 有偏估计量}$$

- ▶ 有效性: 我们聚焦无偏性, 寻找其中方差可能达到最小的一个估计量.
- ▶ 一致性: 估计量 $\hat{\theta}$ 收敛于未知参数 θ .
- ▶ Cramer-Rao 下界给出了任意一个无偏估计量的方差可能取到的最小值.

Cramer-Rao 下界

- 评分函数 (score function) 及其性质:

- ▶ 评分函数 $s(\mathcal{X}; \theta)$ 是对数似然函数关于 $\theta \in \mathbb{R}^k$ 的导数

$$s(\mathcal{X}; \theta) = \frac{\partial}{\partial \theta} \ell(\mathcal{X}; \theta) = \frac{1}{L(\mathcal{X}; \theta)} \cdot \frac{\partial}{\partial \theta} L(\mathcal{X}; \theta)$$

- ▶ Fisher 信息矩阵: 评分函数的协方差矩阵

$$\mathcal{F}_n = \text{Var} \left[s(\mathcal{X}; \theta) \right]$$

定理 6.1 设 $s = s(\mathcal{X}; \theta)$ 是评分函数, 如果 $\hat{\theta} = t = t(\mathcal{X}; \theta)$ 是 \mathcal{X} 与 θ 的任一函数, 则在正则条件下有

$$\text{E}(st^T) = \frac{\partial}{\partial \theta} \text{E}(t^T) - \text{E} \left(\frac{\partial t^T}{\partial \theta} \right).$$

证明留作习题

Cramer-Rao 下界

- 评分函数 (score function) 及其性质:

推论 6.1 设 $s(\mathcal{X}; \theta)$ 是评分函数, $\hat{\theta} = t = t(\mathcal{X})$ 是 θ 的任意一个无偏估计量, 即

$\mathbb{E}(t) = \theta$, 则

$$\mathbb{E}(st^T) = \text{Cov}(s, t) = \mathcal{F}_k.$$

证明也留作习题.

▶ 可以证明

$$\mathbb{E}\left[s(\mathcal{X}; \theta)\right] = \mathbf{0} \implies \mathbb{E}(ss^T) = \text{Var}(s) = \mathcal{F}_n$$

$s = t$

$$\mathcal{F}_n = \mathbb{E}(ss^T) = -\mathbb{E}\left(\frac{\partial}{\partial \theta} s^T\right)$$

$$= -\mathbb{E}\left[\frac{\partial^2}{\partial \theta \partial \theta^T} \ell(\mathcal{X}; \theta)\right]$$

定理 6.1 设 $s = s(\mathcal{X}; \theta)$ 是评分函数, 如果 $\hat{\theta} = t = t(\mathcal{X}; \theta)$ 是 \mathcal{X} 与 θ 的任一函数, 则在正则条件下有

$$\mathbb{E}(st^T) = \frac{\partial}{\partial \theta} \mathbb{E}(t^T) - \mathbb{E}\left(\frac{\partial t^T}{\partial \theta}\right).$$

$$s(\mathcal{X}; \theta) = \frac{\partial}{\partial \theta} \ell(\mathcal{X}; \theta)$$

Cramer-Rao 下界

- 例：考虑 $x_1, x_2, \dots, x_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\theta}, \mathcal{F})$.

$$\begin{aligned}
 s(\mathcal{X}; \boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathcal{X}; \boldsymbol{\theta}) \\
 &= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\theta}} \left[\sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\theta})^T (\mathbf{x}_i - \boldsymbol{\theta}) \right] \\
 &= n(\bar{\mathbf{x}} - \boldsymbol{\theta})
 \end{aligned}$$

$$f(\mathbf{x}; \boldsymbol{\theta}) = (2\pi)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\theta})^T (\mathbf{x} - \boldsymbol{\theta}) \right\}$$

$$L(\mathcal{X}; \boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{x}_i; \boldsymbol{\theta})$$

$$= (2\pi)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\theta})^T (\mathbf{x}_i - \boldsymbol{\theta}) \right\}$$

$$\ell(\mathcal{X}; \boldsymbol{\theta}) = \log L(\mathcal{X}; \boldsymbol{\theta})$$

$$= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\theta})^T (\mathbf{x}_i - \boldsymbol{\theta})$$

$$\Rightarrow \mathcal{F}_n = \text{Var} [s(\mathcal{X}; \boldsymbol{\theta})] = \text{Var} [n(\bar{\mathbf{x}} - \boldsymbol{\theta})] = n\mathcal{F}_p$$

Cramer-Rao 下界

定理 6.2 (Cramer-Rao) 设 $\hat{\theta} = t = t(\mathcal{X})$ 是 θ 的任意一个无偏估计量, 则在正则条件

下有

$$\text{Var}(t) \geq \mathcal{F}_n^{-1}$$

其中

$$\mathcal{F}_n = \mathbb{E} \left\{ s(\mathcal{X}; \theta) \left[s(\mathcal{X}; \theta) \right]^T \right\} = \text{Var} \left[s(\mathcal{X}; \theta) \right]$$

是 Fisher 信息矩阵.

Cramer-Rao 下界

证明：令

$$Y = \mathbf{a}^T \mathbf{t}, \quad Z = \mathbf{c}^T \mathbf{s}$$

$\mathbf{a}, \mathbf{c} \in \mathbb{R}^p$ 评分函数

$$\text{Cov}(\mathbf{s}, \mathbf{t}) = \mathcal{F}$$

$$\Rightarrow \text{Cov}(Y, Z) = \text{Cov}(\mathbf{a}^T \mathbf{t}, \mathbf{c}^T \mathbf{s}) = \mathbf{a}^T \text{Cov}(\mathbf{t}, \mathbf{s}) \mathbf{c} = \mathbf{a}^T \mathbf{c}$$

$$\text{Var}(Z) = \text{Var}(\mathbf{c}^T \mathbf{s}) = \mathbf{c}^T \text{Var}(\mathbf{s}) \mathbf{c} = \mathbf{c}^T \mathcal{F}_n \mathbf{c}$$

$$\Rightarrow \rho_{Y, Z}^2 = \frac{[\text{Cov}(Y, Z)]^2}{\text{Var}(Y) \text{Var}(Z)} = \frac{(\mathbf{a}^T \mathbf{c})^2}{\mathbf{a}^T \text{Var}(\mathbf{t}) \mathbf{a} \cdot \mathbf{c}^T \mathcal{F}_n \mathbf{c}} \leq 1$$

对任意 $\mathbf{c} \neq \mathbf{0}$ 成立

定理 6.2 (Cramer-Rao) 设 $\hat{\theta} = t = t(\mathcal{X})$ 是 θ 的任意一个无偏估计量，则在正则条件下有

下有

$$\text{Var}(t) \geq \mathcal{F}_n^{-1}$$

其中

$$\mathcal{F}_n = \text{E} \left\{ s(\mathcal{X}; \theta) [s(\mathcal{X}; \theta)]^T \right\} = \text{Var} [s(\mathcal{X}; \theta)]$$

是 Fisher 信息矩阵。

推论 6.1 设 $s(\mathcal{X}; \theta)$ 是评分函数， $\hat{\theta} = t = t(\mathcal{X})$ 是 θ 的任意一个无偏估计量，即 $\text{E}(t) = \theta$ ，则

$\text{E}(t) = \theta$ ，则

$$\text{E}(st^T) = \text{Cov}(s, t) = \mathcal{F}_k.$$

Cramer-Rao 下界

证明：令

$$Y = \mathbf{a}^T \mathbf{t}, \quad Z = \mathbf{c}^T \mathbf{s}$$

$\mathbf{a}, \mathbf{c} \in \mathbb{R}^p$ 评分函数

定理 6.2 (Cramer-Rao) 设 $\hat{\theta} = t = t(x)$ 是 θ 的任意一个无偏估计量，则在正则条件下有

下有

$$\text{Var}(t) \geq \mathcal{F}_n^{-1}$$

其中

$$\mathcal{F}_n = \text{E} \left\{ \mathbf{s}(x; \theta) [\mathbf{s}(x; \theta)]^T \right\} = \text{Var} [\mathbf{s}(x; \theta)]$$

是 Fisher 信息矩阵。

▶ 因为

$$\frac{(\mathbf{a}^T \mathbf{c})^2}{\mathbf{c}^T \mathcal{F}_n \mathbf{c}} = \frac{(\mathbf{a}^T \mathbf{c})^T (\mathbf{a}^T \mathbf{c})}{\mathbf{c}^T \mathcal{F}_n \mathbf{c}} = \frac{\mathbf{c}^T \mathbf{a} \mathbf{a}^T \mathbf{c}}{\mathbf{c}^T \mathcal{F}_n \mathbf{c}}$$

$$\Rightarrow \rho_{Y, Z}^2 = \frac{[\text{Cov}(Y, Z)]^2}{\text{Var}(Y) \text{Var}(Z)} = \frac{(\mathbf{a}^T \mathbf{c})^2}{\mathbf{a}^T \text{Var}(t) \mathbf{a} \cdot \mathbf{c}^T \mathcal{F}_n \mathbf{c}} \leq 1$$

Cramer-Rao 下界

证明：令

$$Y = \mathbf{a}^T \mathbf{t}, \quad Z = \mathbf{c}^T \mathbf{s}$$

$\mathbf{a}, \mathbf{c} \in \mathbb{R}^p$ ← → 评分函数

定理 6.2 (Cramer-Rao) 设 $\hat{\theta} = t = t(x)$ 是 θ 的任意一个无偏估计量, 则在正则条件

下有

$$\text{Var}(t) \geq \mathcal{F}_n^{-1}$$

其中

$$\mathcal{F}_n = \text{E} \left\{ \mathbf{s}(x; \theta) [\mathbf{s}(x; \theta)]^T \right\} = \text{Var} [\mathbf{s}(x; \theta)]$$

是 Fisher 信息矩阵.

▶ 因为

$$\frac{(\mathbf{a}^T \mathbf{c})^2}{\mathbf{c}^T \mathcal{F}_n \mathbf{c}} = \frac{(\mathbf{a}^T \mathbf{c})^T (\mathbf{a}^T \mathbf{c})}{\mathbf{c}^T \mathcal{F}_n \mathbf{c}} = \frac{\mathbf{c}^T \mathbf{a} \mathbf{a}^T \mathbf{c}}{\mathbf{c}^T \mathcal{F}_n \mathbf{c}}$$

定理 2.5 若 \mathcal{A} 与 \mathcal{B} 为对称矩阵且 $\mathcal{B} > 0$, 则 $\frac{\mathbf{x}^T \mathcal{A} \mathbf{x}}{\mathbf{x}^T \mathcal{B} \mathbf{x}}$ 的最大值等于 $\mathcal{B}^{-1} \mathcal{A}$ 的最

大特征值. 更一般地, 我们有

$$\max_x \frac{\mathbf{x}^T \mathcal{A} \mathbf{x}}{\mathbf{x}^T \mathcal{B} \mathbf{x}} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p = \min_x \frac{\mathbf{x}^T \mathcal{A} \mathbf{x}}{\mathbf{x}^T \mathcal{B} \mathbf{x}}$$

其中 $\lambda_1, \lambda_2, \dots, \lambda_p$ 是 $\mathcal{B}^{-1} \mathcal{A}$ 的特征值. 使得 $\frac{\mathbf{x}^T \mathcal{A} \mathbf{x}}{\mathbf{x}^T \mathcal{B} \mathbf{x}}$ 达到最大(最小)的向

量是 $\mathcal{B}^{-1} \mathcal{A}$ 的最大(最小)特征值对应的 $\mathcal{B}^{-1} \mathcal{A}$ 的特征向量. 如果

$\mathbf{x}^T \mathcal{B} \mathbf{x} = 1$, 我们则有

$$\max_x \mathbf{x}^T \mathcal{A} \mathbf{x} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p = \min_x \mathbf{x}^T \mathcal{A} \mathbf{x}$$

Cramer-Rao 下界

证明：令

$$Y = a^T t, \quad Z = c^T s$$

$a, c \in \mathbb{R}^p$ ← → 评分函数

定理 6.2 (Cramer-Rao) 设 $\hat{\theta} = t = t(x)$ 是 θ 的任意一个无偏估计量，则在正则条件

下有

$$\text{Var}(t) \geq \mathcal{F}_n^{-1}$$

其中

$$\mathcal{F}_n = E \left\{ s(x; \theta) [s(x; \theta)]^T \right\} = \text{Var} [s(x; \theta)]$$

是 Fisher 信息矩阵。

▶ 因为

$$\frac{(a^T c)^2}{c^T \mathcal{F}_n c} = \frac{(a^T c)^T (a^T c)}{c^T \mathcal{F}_n c} = \frac{c^T a a^T c}{c^T \mathcal{F}_n c}$$

$$\max_c \left(\frac{c^T a a^T c}{c^T \mathcal{F}_n c} \right) = \max_{c^T \mathcal{F}_n c = 1} (c^T a a^T c)$$

$$= \max_{c^T \mathcal{F}_n c = 1} (a^T c c^T a)$$

$$= a^T \mathcal{F}_n^{-1} a$$

$$c c^T \mathcal{F}_n c c^T = c c^T \implies \begin{cases} c c^T \mathcal{F}_n &= \mathcal{I} \\ \mathcal{F}_n c c^T &= \mathcal{I} \end{cases}$$

$$\implies c c^T = \mathcal{F}_n^{-1}$$

定理 2.5 若 \mathcal{A} 与 \mathcal{B} 为对称矩阵且 $\mathcal{B} > 0$ ，则 $\frac{x^T \mathcal{A} x}{x^T \mathcal{B} x}$ 的最大值等于 $\mathcal{B}^{-1} \mathcal{A}$ 的最

大特征值。更一般地，我们有

$$\max_x \frac{x^T \mathcal{A} x}{x^T \mathcal{B} x} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p = \min_x \frac{x^T \mathcal{A} x}{x^T \mathcal{B} x}$$

其中 $\lambda_1, \lambda_2, \dots, \lambda_p$ 是 $\mathcal{B}^{-1} \mathcal{A}$ 的特征值。使得 $\frac{x^T \mathcal{A} x}{x^T \mathcal{B} x}$ 达到最大(最小)的向

量是 $\mathcal{B}^{-1} \mathcal{A}$ 的最大(最小)特征值对应的 $\mathcal{B}^{-1} \mathcal{A}$ 的特征向量。如果 $x^T \mathcal{B} x = 1$ ，我们则有

$$\max_x x^T \mathcal{A} x = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p = \min_x x^T \mathcal{A} x$$

Cramer-Rao 下界

证明：令

$$Y = \mathbf{a}^T \mathbf{t}, \quad Z = \mathbf{c}^T \mathbf{s}$$

$\mathbf{a}, \mathbf{c} \in \mathbb{R}^p$ 评分函数

定理 6.2 (Cramer-Rao) 设 $\hat{\theta} = \mathbf{t} = \mathbf{t}(\mathcal{X})$ 是 θ 的任意一个无偏估计量，则在正则条件

下有

$$\text{Var}(\mathbf{t}) \geq \mathcal{F}_n^{-1}$$

其中

$$\mathcal{F}_n = \text{E} \left\{ \mathbf{s}(\mathcal{X}; \theta) [\mathbf{s}(\mathcal{X}; \theta)]^T \right\} = \text{Var} [\mathbf{s}(\mathcal{X}; \theta)]$$

是 Fisher 信息矩阵。

▶ 所以

$$\frac{\mathbf{a}^T \mathcal{F}_n^{-1} \mathbf{a}}{\mathbf{a}^T \text{Var}(\mathbf{t}) \mathbf{a}} \leq 1, \quad \forall \mathbf{a} \in \mathbb{R}^p, \quad \mathbf{a} \neq \mathbf{0}$$

已证得 $\Rightarrow \rho_{Y,Z}^2 = \frac{[\text{Cov}(Y, Z)]^2}{\text{Var}(Y) \text{Var}(Z)} = \frac{(\mathbf{a}^T \mathbf{c})^2}{\mathbf{a}^T \text{Var}(\mathbf{t}) \mathbf{a} \cdot \mathbf{c}^T \mathcal{F}_n \mathbf{c}} \leq 1$

$\Rightarrow \max_c \frac{(\mathbf{a}^T \mathbf{c})^2}{\mathbf{c}^T \mathcal{F}_n \mathbf{c}} = \mathbf{a}^T \mathcal{F}_n^{-1} \mathbf{a}$

Cramer-Rao 下界

证明：令

$$Y = \mathbf{a}^T \mathbf{t}, \quad Z = \mathbf{c}^T \mathbf{s}$$

$\mathbf{a}, \mathbf{c} \in \mathbb{R}^p$ 评分函数

定理 6.2 (Cramer-Rao) 设 $\hat{\theta} = \mathbf{t} = \mathbf{t}(\mathcal{X})$ 是 θ 的任意一个无偏估计量，则在正则条件

下有

$$\text{Var}(\mathbf{t}) \geq \mathcal{F}_n^{-1}$$

其中

$$\mathcal{F}_n = \text{E} \left\{ \mathbf{s}(\mathcal{X}; \theta) [\mathbf{s}(\mathcal{X}; \theta)]^T \right\} = \text{Var} [\mathbf{s}(\mathcal{X}; \theta)]$$

是 Fisher 信息矩阵。

▶ 所以

$$\frac{\mathbf{a}^T \mathcal{F}_n^{-1} \mathbf{a}}{\mathbf{a}^T \text{Var}(\mathbf{t}) \mathbf{a}} \leq 1, \quad \forall \mathbf{a} \in \mathbb{R}^p, \quad \mathbf{a} \neq \mathbf{0}$$

$$\implies \mathbf{a}^T \mathcal{F}_n^{-1} \mathbf{a} \leq \mathbf{a}^T \text{Var}(\mathbf{t}) \mathbf{a}, \quad \forall \mathbf{a} \in \mathbb{R}^p, \quad \mathbf{a} \neq \mathbf{0}$$

$$\implies \mathbf{a}^T [\text{Var}(\mathbf{t}) - \mathcal{F}_n^{-1}] \mathbf{a} \geq 0, \quad \forall \mathbf{a} \in \mathbb{R}^p, \quad \mathbf{a} \neq \mathbf{0}$$

$$\implies \text{Var}(\mathbf{t}) - \mathcal{F}_n^{-1} \geq 0$$

$$\implies \text{Var}(\mathbf{t}) \geq \mathcal{F}_n^{-1}$$

Cramer-Rao 下界

定理 6.3 假设 $\{x_i\}_{i=1}^n$ 是独立同分布的样本. 若 $\hat{\theta}$ 是 $\theta \in \mathbb{R}^k$ 的极大似然估计, 即

$\hat{\theta} = \arg \max_{\theta} L(\mathcal{X}; \theta)$, 则在一定的正则条件下, 当 $n \rightarrow \infty$ 时有:

$$\sqrt{n} (\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N_k(\mathbf{0}, \mathcal{F}_1^{-1})$$

其中 \mathcal{F}_1 表示样本容量 $n = 1$ 时的 Fisher 信息.

- 由定理 6.3 的结果我们看到, 在正则性条件下
 - ▶ 极大似然估计 (MLE) 是渐进无偏的
 - ▶ 极大似然估计 (MLE) 是渐进有效的 (最小方差)
 - ▶ 极大似然估计 (MLE) 是渐进正态分布的
 - ▶ 极大似然估计 (MLE) 是 θ 的一致估计量

Cramer-Rao 下界

- 由多元正态分布的性质，我们有

$$n \left(\hat{\theta} - \theta \right)^T \mathcal{F}_1 \left(\hat{\theta} - \theta \right) \xrightarrow{\mathcal{L}} \chi_p^2, \quad n \rightarrow \infty$$

如果 $X \sim N_p(\mu, \Sigma)$ ，则 Mahalanobis 变换

$$Y = \Sigma^{-1/2} (X - \mu) \sim N_p(\mathbf{0}, \mathcal{I}_p)$$

且

$$Y^T Y = (X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi_p^2$$

- 如果 $\widehat{\mathcal{F}}_1$ 是 \mathcal{F}_1 的一个一致估计量，即 $\widehat{\mathcal{F}}_1 = \mathcal{F}_1(\hat{\theta})$ ，则等价地有

$$n \left(\hat{\theta} - \theta \right)^T \widehat{\mathcal{F}}_1 \left(\hat{\theta} - \theta \right) \xrightarrow{\mathcal{L}} \chi_p^2, \quad n \rightarrow \infty$$

- 上式在关于 θ 的假设检验问题和置信域的构造当中很有用，详见第 7 章.