

18.650 – Fundamentals of Statistics

3. Methods for estimation

Goals

In the kiss example, the estimator was **intuitively** the right thing to do: $\hat{p} = \bar{X}_n$.

In view of LLN, since $p = \mathbb{E}[X]$, we have \bar{X}_n so $\hat{p} \approx p$ for n large enough.

If the parameter is $\theta \neq \mathbb{E}[X]$? How do we perform?

1. Maximum likelihood estimation: a generic approach with very good properties
2. Method of moments: a (fairly) generic and easy approach
3. M-estimators: a flexible approach, close to machine learning

Total variation distance

Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with a sample of i.i.d. r.v. X_1, \dots, X_n . Assume that there exists $\theta^* \in \Theta$ such that $X_1 \sim \mathbb{P}_{\theta^*}$: θ^* is the **true** parameter.

Statistician's goal: given X_1, \dots, X_n , find an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ such that $\mathbb{P}_{\hat{\theta}}$ is close to \mathbb{P}_{θ^*} for the true parameter θ^* .

This means: $| \mathbb{P}_{\hat{\theta}}(A) - \mathbb{P}_{\theta^*}(A) |$ is **small** for all $A \subset E$.

Definition

The *total variation distance* between two probability measures \mathbb{P}_θ and $\mathbb{P}_{\theta'}$ is defined by

$$\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \max_{A \subset E} | \mathbb{P}_\theta(A) - \mathbb{P}_{\theta'}(A) |.$$

Total variation distance between discrete measures

Assume that E is discrete (i.e., finite or countable). This includes Bernoulli, Binomial, Poisson, . . .

Therefore X has a PMF (probability mass function):
 $\mathbb{P}_\theta(X = x) = p_\theta(x)$ for all $x \in E$,

$$p_\theta(x) \geq 0, \quad \sum_{x \in E} p_\theta(x) = 1$$

The total variation distance between \mathbb{P}_θ and $\mathbb{P}_{\theta'}$ is a simple function of the PMF's p_θ and $p_{\theta'}$:

$$\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \sum_{x \in E} |p_\theta(x) - p_{\theta'}(x)|.$$

注意support

Total variation distance between continuous measures

Assume that E is continuous. This includes Gaussian, Exponential,

...

Assume that X has a density $\mathbb{P}_\theta(X \in A) = \int_A f_\theta(x) dx$ for all $A \subset E$.

$$f_\theta(x) \geq 0, \quad \int_E f_\theta(x) dx = 1.$$

The total variation distance between \mathbb{P}_θ and $\mathbb{P}_{\theta'}$ is a simple function of the densities f_θ and $f_{\theta'}$:

$$\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \int |f_\theta(x) - f_{\theta'}(x)| dx.$$

注意support

Properties of Total variation

- $\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \text{TV}(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ (symmetric)
- $\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0, \quad \text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq 1$ (positive)
- If $\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$ then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ (definite)
- $\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq \text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + \text{TV}(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$ (triangle inequality)

These imply that the total variation is a *distance* between probability distributions.

Exercises

Compute:

a) $\text{TV}(\text{Ber}(0.5), \text{Ber}(0.1)) = \frac{1}{2} [|P_{0.5}(0) - P_{0.1}(0)| + |P_{0.5}(1) - P_{0.1}(1)|]$

$E = 0.15$

$$= \frac{1}{2} [|\underbrace{0.5 - 0.1}_{0.4}| + |0.5 - 0.1|] = \frac{0.8}{2} = 0.4$$

b) $\text{TV}(\text{Ber}(0.5), \text{Ber}(0.9)) = 0.4$

c) $\text{TV}(\text{Exp}(1), \text{Unif}[0, 1]) = \frac{1}{e}$

d) $\text{TV}(X, X + a) = 1$

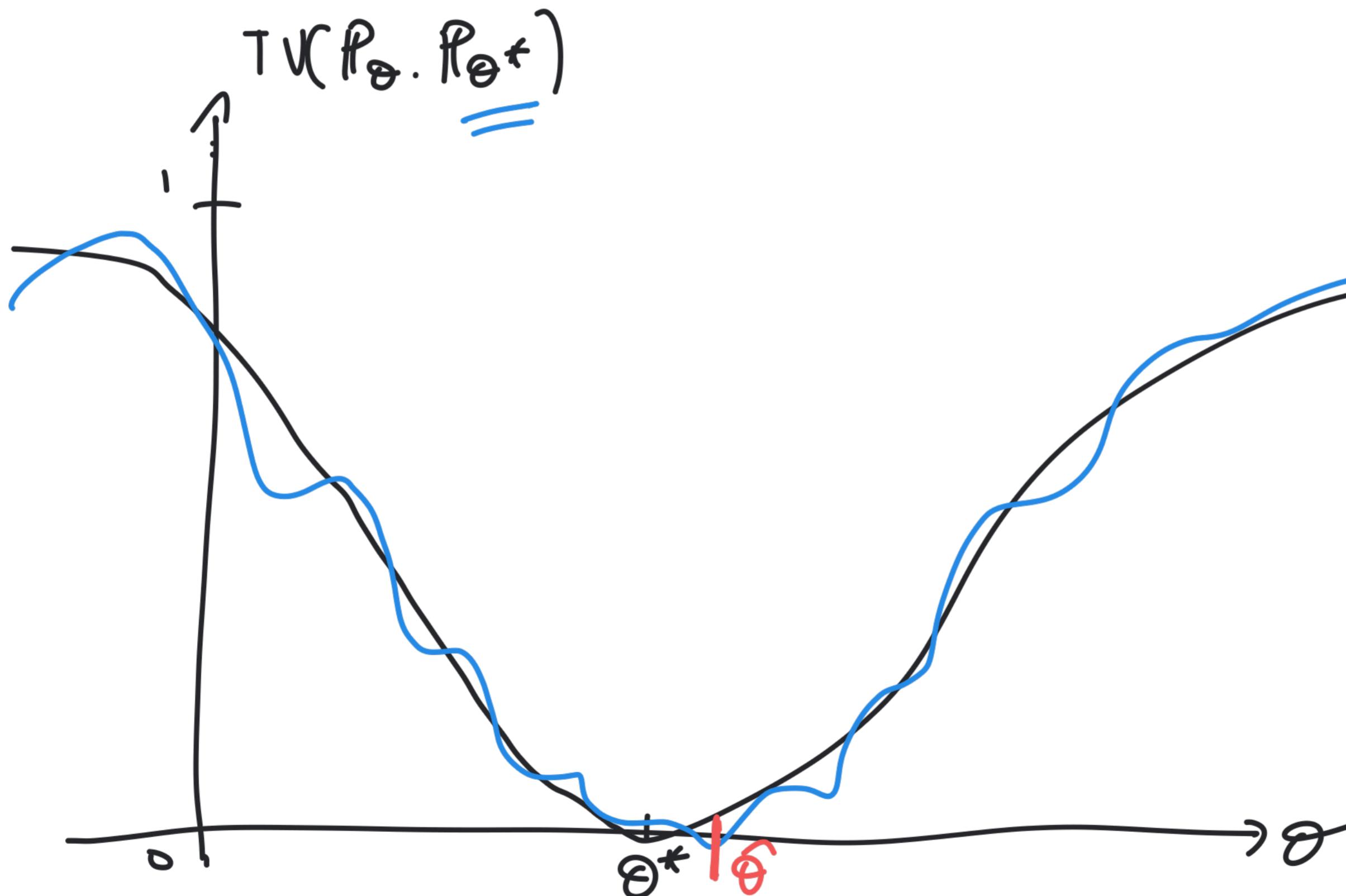
for any $a \in (0, 1)$, where $X \sim \text{Ber}(0.5)$

e) $\text{TV}(2\sqrt{n}(\bar{X}_n - 1/2), Z) = 1$

where $X_i \stackrel{i.i.d.}{\sim} \text{Ber}(0.5)$ and $Z \sim \mathcal{N}(0, 1)$

An estimation strategy

Build an estimator $\widehat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$ for all $\theta \in \Theta$. Then find $\hat{\theta}$ that minimizes the function $\theta \mapsto \widehat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$.



problem: Unclear how to build $\widehat{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$!

Kullback-Leibler (KL) divergence

There are **many** distances between probability measures to replace total variation. Let us choose one that is more convenient.

Definition

The *Kullback-Leibler¹ (KL) divergence* between two probability measures \mathbb{P}_θ and $\mathbb{P}_{\theta'}$ is defined by

$$\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \begin{cases} \sum_{x \in E} p_\theta(x) \log \left(\frac{p_\theta(x)}{p_{\theta'}(x)} \right) & \text{if } E \text{ is discrete} \\ \int_E f_\theta(x) \log \left(\frac{f_\theta(x)}{f_{\theta'}(x)} \right) dx & \text{if } E \text{ is continuous} \end{cases}$$

¹KL-divergence is also known as “relative entropy”

Properties of KL-divergence

- ▶ $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \neq \text{KL}(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ in general
- ▶ $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0$
- ▶ If $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$ then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ (definite) ✓
- ▶ $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \not\leq \text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + \text{KL}(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$ in general

Not a distance.

This is called a *divergence*

Asymmetry is the key to our ability to estimate it!

θ^* unique minimizer of $\theta \mapsto \text{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$

Maximum likelihood estimation

Estimating the KL

$$\text{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \mathbb{E}_{\theta^*} \left[\log \left(\frac{p_{\theta^*}(X)}{p_\theta(X)} \right) \right]$$

$$= \mathbb{E}_{\theta^*} [\log p_{\theta^*}(X)] - \mathbb{E}_{\theta^*} [\log p_\theta(X)]$$

So the function $\theta \mapsto \text{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ is of the form:

“constant” – $\mathbb{E}_{\theta^*} [\log p_\theta(X)]$

Can be estimated: $\mathbb{E}_{\theta^*}[h(X)] \rightsquigarrow \frac{1}{n} \sum_{i=1}^n h(X_i)$ (by LLN)

$$\widehat{\text{KL}}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \text{“constant”} - \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i)$$

(blue circ)

Maximum likelihood

$$\widehat{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \text{"constant"} - \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i)$$

$$\begin{aligned}\min_{\theta \in \Theta} \widehat{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) &\Leftrightarrow \min_{\theta \in \Theta} -\frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i) \\ &\Leftrightarrow \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log \hat{p}_\theta(X_i) \\ &\Leftrightarrow \max_{\theta \in \Theta} \log \left[\prod_{i=1}^n \hat{p}_\theta(X_i) \right] \\ &\Leftrightarrow \max_{\theta \in \Theta} \prod_{i=1}^n p_\theta(X_i)\end{aligned}$$

This is the **maximum likelihood principle**.

Likelihood, Discrete case (1)

Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with a sample of i.i.d. r.v. X_1, \dots, X_n . Assume that E is discrete (i.e., finite or countable).

Definition

The *likelihood* of the model is the map L_n (or just L) defined as:

$$\begin{aligned} L_n : E^n \times \Theta &\rightarrow \mathbb{R} \\ (x_1, \dots, x_n, \theta) &\mapsto \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n]. \end{aligned}$$

$= \prod_{i=1}^n \mathbb{P}_\theta[X_i = x_i]$

Likelihood for the Bernoulli model

Example 1 (Bernoulli trials): If $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Ber}(p)$ for some $p \in (0, 1)$:

- ▶ $E = \{0, 1\}$; ✓
- ▶ $\Theta = (0, 1)$; ✓
- ▶ $\forall (x_1, \dots, x_n) \in \{0, 1\}^n, \quad \forall p \in (0, 1),$

$$\begin{aligned} L(x_1, \dots, x_n, p) &= \prod_{i=1}^n \mathbb{P}_p[X_i = x_i] \\ &= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}. \end{aligned}$$

Likelihood for the Poisson model

Example 2 (Poisson model):

If $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Poiss}(\lambda)$ for some $\lambda > 0$:

- ▶ $E = \mathbb{N}$;
- ▶ $\Theta = (0, \infty)$;
- ▶ $\forall (x_1, \dots, x_n) \in \mathbb{N}^n, \quad \forall \lambda > 0,$

$$L(x_1, \dots, x_n, \lambda) = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \dots x_n!}.$$

$$\begin{aligned} P(X_i = x_i) &= \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \\ \Rightarrow L(x_1, \dots, x_n, \lambda) &= \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\lambda} \\ &\stackrel{!}{=} \prod_{i=1}^n P_\lambda(X_i = x_i) \end{aligned}$$

Likelihood, Continuous case

Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with a sample of i.i.d. r.v. X_1, \dots, X_n . Assume that all the \mathbb{P}_θ have density f_θ .

Definition

The *likelihood* of the model is the map L defined as:

$$\begin{aligned} L &: E^n \times \Theta && \rightarrow \mathbb{R} \\ &(x_1, \dots, x_n, \theta) && \mapsto \prod_{i=1}^n f_\theta(x_i). \end{aligned}$$

Likelihood for the Gaussian model

Example 1 (Gaussian model): If $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some $\mu \in \mathbb{R}, \sigma^2 > 0$:

- ▶ $E = \mathbb{R}$;
- ▶ $\Theta = \mathbb{R} \times (0, \infty)$
- ▶ $\forall (x_1, \dots, x_n) \in \mathbb{R}^n, \quad \forall (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty),$

$$L(x_1, \dots, x_n; \mu, \sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right).$$

Exercises

Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with $X_1, \dots, X_n \sim \text{Exp}(\lambda), \lambda > 0$.

a) What is E ? $(0, \infty)$

b) What is Θ ? $(0, \infty)$

c) Find the likelihood of the model.

$$L(x_1, \dots, x_n; \lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \mathbb{1}(\min_i x_i > 0)$$

Exercise

Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with $X_1, \dots, X_n \sim \text{Unif}[0, b]$ for some $b > 0$.

a) What is E ?

$$[0, \infty)$$

b) What is Θ ?

$$[0, \infty)$$

c) Find the likelihood of the model.

$$L(x_1, \dots, x_n; b) = \frac{1}{b^n} \prod_{i=1}^n \mathbb{1}(x_i \leq b)$$

Maximum likelihood estimator

Let X_1, \dots, X_n be an i.i.d. sample associated with a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ and let L be the corresponding likelihood.

Definition

The *maximum likelihood estimator* of θ is defined as:

$$\hat{\theta}_n^{MLE} = \operatorname{argmax}_{\theta \in \Theta} L(X_1, \dots, X_n, \theta),$$

provided it exists.

Remark (log-likelihood estimator): In practice, we use the fact that

$$\hat{\theta}_n^{MLE} = \operatorname{argmax}_{\theta \in \Theta} \log L(X_1, \dots, X_n, \theta).$$

Interlude: maximizing/minimizing functions

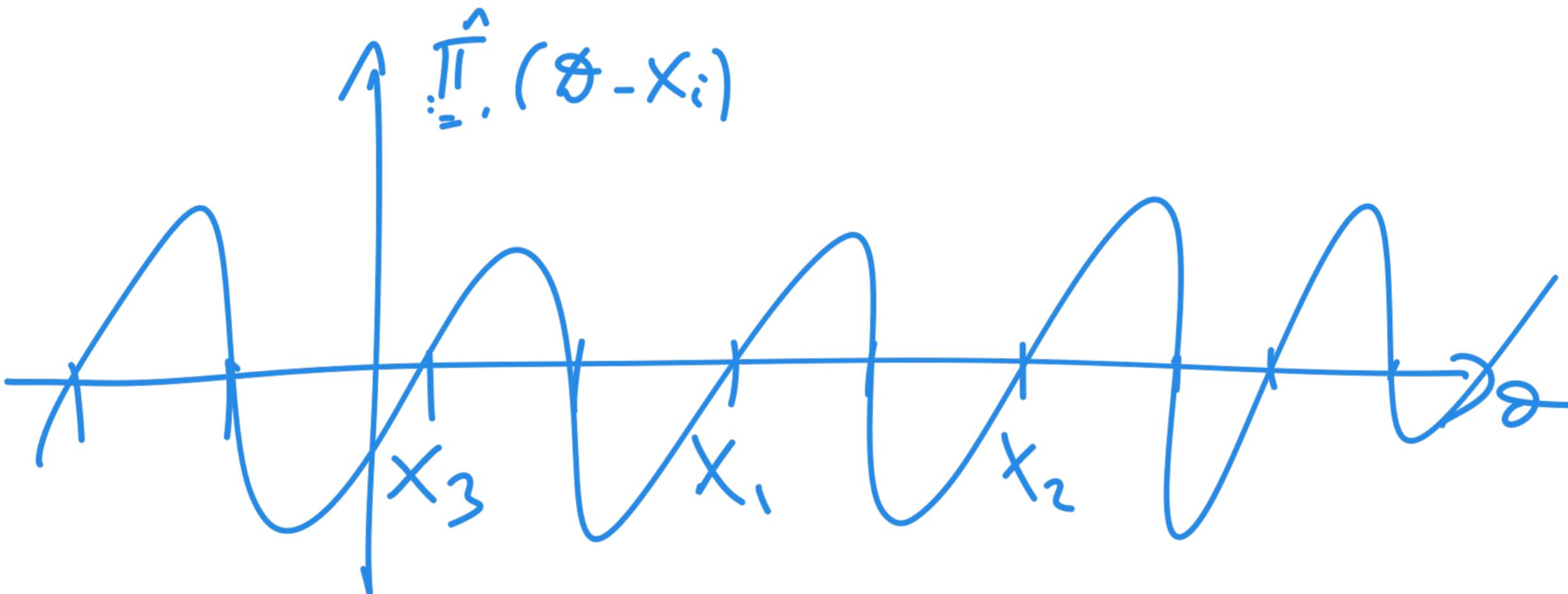
optimization

Note that

$$\min_{\theta \in \Theta} -h(\theta) \Leftrightarrow \max_{\theta \in \Theta} h(\theta)$$

In this class, we focus on **maximization**.

Maximization of arbitrary functions can be difficult:



Example: $\theta \mapsto \prod_{i=1}^n (\theta - X_i)$

Concave and convex functions

Definition

A function twice differentiable function $h : \Theta \subset \mathbb{R} \rightarrow \mathbb{R}$ is said to be **concave** if its second derivative satisfies

$$h''(\theta) \leq 0, \quad \forall \theta \in \Theta$$

It is said to be **strictly concave** if the inequality is strict: $h''(\theta) < 0$

Moreover, h is said to be (strictly) **convex** if $-h$ is (strictly) concave, i.e. $h''(\theta) \geq 0$ ($h''(\theta) > 0$).

Examples:

- $\Theta = \mathbb{R}, h(\theta) = -\theta^2, h'(\theta) = -2\theta, h''(\theta) = -2 < 0$ (s. concave)
- $\Theta = (0, \infty), h(\theta) = \sqrt{\theta}, h'(\theta) = \frac{1}{2\sqrt{\theta}}, h''(\theta) = -\frac{1}{4\theta^{3/2}} < 0$ (s. concave)
- $\Theta = (0, \infty), h(\theta) = \log \theta, h'(\theta) = \frac{1}{\theta}, h''(\theta) = -\frac{1}{\theta^2} < 0$ (s. concave)
- $\Theta = [0, \pi], h(\theta) = \sin(\theta), h'(\theta) = \cos(\theta), h''(\theta) = -\sin(\theta) \leq 0$ (concave)
- $\Theta = \mathbb{R}, h(\theta) = 2\theta - 3, h'(\theta) = 2, h''(\theta) = 0 \begin{cases} \leq 0 \\ \geq 0 \end{cases}$ Both

Multivariate concave functions

More generally for a *multivariate* function: $h : \Theta \subset \mathbb{R}_{\neq}^d \rightarrow \mathbb{R}$, $d \geq 2$, define the

- *gradient vector*: $\nabla h(\theta) = \begin{pmatrix} \frac{\partial h}{\partial \theta_1}(\theta) \\ \vdots \\ \frac{\partial h}{\partial \theta_d}(\theta) \end{pmatrix} \in \mathbb{R}^d$

► *Hessian matrix*:

$$\mathbf{H}h(\theta) = \begin{pmatrix} \frac{\partial^2 h}{\partial \theta_1 \partial \theta_1}(\theta) & \cdots & \frac{\partial^2 h}{\partial \theta_1 \partial \theta_d}(\theta) \\ & \ddots & \\ \frac{\partial^2 h}{\partial \theta_d \partial \theta_1}(\theta) & \cdots & \frac{\partial^2 h}{\partial \theta_d \partial \theta_d}(\theta) \end{pmatrix} \in \mathbb{R}^{d \times d}$$

h is concave $\Leftrightarrow x^\top \mathbf{H}h(\theta)x \leq 0 \quad \forall x \in \mathbb{R}^d, \theta \in \Theta.$

h is strictly concave $\Leftrightarrow x^\top \mathbf{H}h(\theta)x < 0 \quad \forall x \in \mathbb{R}^d, \theta \in \Theta.$
 $x \neq 0$

Examples:

- $\Theta = \mathbb{R}^2$, $h(\theta) = -\theta_1^2 - 2\theta_2^2$ or $h(\theta) = -(\theta_1 - \theta_2)^2$
- $\Theta = (0, \infty)$, $h(\theta) = \log(\theta_1 + \theta_2)$,

Optimality conditions

Strictly concave functions are easy to maximize: if they have a maximum, then it is **unique**. It is the unique solution to

$$h'(\theta) = 0,$$

or, in the multivariate case

$$\nabla h(\theta) = 0 \in \mathbb{R}^d.$$

There are many algorithms to find it numerically: this is the theory of “convex optimization”. In this class, often a **closed form formula** for the maximum.

Exercises

a) Which one of the following functions are concave on $\Theta = \mathbb{R}^2$?

1. $h(\theta) = -(\theta_1 - \theta_2)^2 - \theta_1\theta_2$
2. $h(\theta) = -(\theta_1 - \theta_2)^2 + \theta_1\theta_2$
3. $h(\theta) = (\theta_1 - \theta_2)^2 - \theta_1\theta_2$
4. Both 1. and 2.
5. All of the above
6. None of the above

b) Let $h : \Theta \subset \mathbb{R}^d \rightarrow \mathbb{R}$ be a function whose hessian matrix $\mathbf{H}h(\theta)$ has a positive diagonal entry for some $\theta \in \Theta$. Can h be concave? Why or why not?

Examples of maximum likelihood estimators

$$\text{Ber: } L(x_1, \dots, x_n; p) = \prod_{i=1}^n P_p[X_i=x_i] = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}$$

$$h(p) := \underbrace{\lg L(x_1, \dots, x_n; p)}_{S_n} + (n - \sum_{i=1}^n x_i) \cdot \underbrace{\lg(1-p)}_{n-S_n}$$

$$h'(p) = \frac{1}{p} S_n - \frac{1}{1-p}(n-S_n); h''(p) = -\frac{1}{p^2} \cdot S_n - \frac{1}{(1-p)^2}(n-S_n) \leq 0$$

$$h \text{ is concave, } h'(\hat{p})=0 \Leftrightarrow \frac{1}{\hat{p}} S_n - \frac{1}{1-\hat{p}}(n-S_n)=0$$

$$\Rightarrow \hat{p} = \frac{S_n}{n} = \bar{x}$$

► Bernoulli trials: $\hat{p}_n^{MLE} = \bar{X}_n$.

► Poisson model: $\hat{\lambda}_n^{MLE} = \bar{X}_n$.

► Gaussian model: $(\hat{\mu}_n, \hat{\sigma}_n^2) = (\bar{X}_n, \hat{S}_n)$.

$$\hat{S}_n = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

$$\text{Gaussian: } L(x_1, \dots, x_n; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

$$h(\mu, \sigma^2) = \lg L(x_1, \dots, x_n; \mu, \sigma^2) = -n \cdot \lg(2\pi\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}$$

$$\nabla h(\mu, \sigma^2) = \begin{cases} \frac{\partial}{\partial \mu} h(\mu, \sigma^2) = \frac{1}{\sigma^2} \cdot \sum_{i=1}^n (x_i - \mu) \\ \frac{\partial}{\partial \sigma^2} h(\mu, \sigma^2) = \frac{n}{2\sigma^2} \lg \sigma^2 - n \lg(\sqrt{2\pi}) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \end{cases}$$

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

$$h \text{ is concave, } \nabla h(\hat{\mu}, \hat{\sigma}^2) = 0 \Leftrightarrow \begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \hat{S}_n \end{cases}$$

Consistency of maximum likelihood estimator

Under mild regularity conditions, we have

$$\hat{\theta}_n^{MLE} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \theta^*$$



This is because for all $\theta \in \Theta$

$$\frac{1}{n} \log L(X_1, \dots, X_n, \theta) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \text{"constant"} - KL(P_{\theta^*}, P_\theta)$$

Moreover, the minimizer of the right-hand side is $\hat{\theta}^*$ if the parameter is identifiable

Technical conditions allow to transfer this convergence to the minimizers.

Covariance

$$\hat{\theta} = \begin{pmatrix} \bar{X}_n \\ \hat{S}_n \end{pmatrix}$$

How about asymptotic normality?

In general, when $\theta \subset \mathbb{R}^d, d \geq 2$, its coordinates are not necessarily independent.

The **covariance** between two random variables X and Y is

$$\begin{aligned}\text{Cov}(X, Y) &:= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[X \cdot Y] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[X \cdot (Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[(X - \mathbb{E}[X])Y]\end{aligned}$$

Properties

- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ ✓
- If X and Y are independent, then $\text{Cov}(X, Y) = 0$



In general, the **converse is not true** except if $(X, Y)^\top$ is a **Gaussian vector**, i.e., $\alpha X + \beta Y$ is Gaussian for all $(\alpha, \beta) \in \mathbb{R}^2 \setminus \{(0, 0)\}$.
Take $X \xrightarrow{\text{indep.}} \mathcal{N}(0, 1)$, $B \sim \text{Ber}(1/2)$, $R = 2B - 1 \sim \text{Rad}(1/2)$. Then

Rademacher

1/2的概率是1, 1/2的概率是-1

$$Y = R \cdot X \sim \mathcal{N}(0, 1)$$

But taking $\alpha = \beta = 1$, we get

$$\begin{aligned}\text{Cov}(X, Y) &= E[X \cdot Y] - E[X]E[Y] \\ &= E[X \cdot R \cdot X] \\ &= E[X^2 \cdot R] = E[R] \cdot E[X] \\ &= 0\end{aligned}$$

$$X + Y = \begin{cases} 2X & \text{with prob. } 1/2 \\ 0 & \text{with prob. } 1/2 \end{cases} \quad \text{Conditionally on } X$$

Actually $\text{Cov}(X, Y) = 0$ but they are not independent: $|X| = |Y|$



Covariance matrix

The covariance matrix of a random vector $X = (X^{(1)}, \dots, X^{(d)})^\top \in \mathbb{R}^d$ is given by

$$\Sigma = \mathbf{Cov}(X) = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^\top]$$

This is a matrix of size $d \times d$

The term on the i th row and j th column is

$$\Sigma_{ij} = \mathbb{E}[(X^{(i)} - \mathbb{E}(X^{(i)}))(X^{(j)} - \mathbb{E}(X^{(j)}))] = \text{Cov}(X^{(i)}, X^{(j)})$$

In particular, on the diagonal, we have

$$\Sigma_{ii} = \text{Cov}(X^{(i)}, X^{(i)}) = \text{Var}(X^{(i)})$$

Recall that for $X \in \mathbb{R}$, $\text{Var}(aX + b) = a^2 \text{Var}(X)$. Actually, if $X \in \mathbb{R}^d$ and A, B are matrices:

$$\text{Cov}(AX + B) = \text{Cov}(AX) = A \text{Cov}(X) A^\top = A \Sigma A^\top$$

The multivariate Gaussian distribution

If $(X, Y)^\top$ is a Gaussian vector then its pdf depends on 5 parameters:

$$\mathbb{E}[X], \text{Var}(X), \mathbb{E}[Y], \text{Var}(Y) \quad \text{and} \quad \text{Cov}(X, Y)$$

More generally, a Gaussian vector³ $X \in \mathbb{R}^d$, is completely determined by its expected value and $\mathbb{E}[X] = \mu \in \mathbb{R}^d$ covariance matrix Σ . We write

$$X \sim \mathcal{N}_d(\mu, \Sigma).$$

It has pdf over \mathbb{R}^d given by:

$$f(x) = f(x^1, \dots, x^d) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

³As before, this means that $\alpha^\top X$ is Gaussian for any $\alpha \in \mathbb{R}^d, \alpha \neq 0$.

The multivariate CLT

The CLT may be generalized to averages or random vectors (also vectors of averages).

Let $X_1, \dots, X_n \in \mathbb{R}^d$ be independent copies of a random vector X such that $\mathbb{E}[X] = \mu$, $\text{Cov}(X) = \Sigma$,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}_d(0, \Sigma)$$

Equivalently

$$\sqrt{n} \Sigma^{-\frac{1}{2}} (\bar{X}_n - \mu) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}_d(0, I_d)$$

$$I_d = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{bmatrix}$$

Multivariate Delta method

Let $(T_n)_{n \geq 1}$ sequence of random vectors in \mathbb{R}^d such that

$$\sqrt{n}(T_n - \theta) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}_d(0, \Sigma),$$

for some $\theta \in \mathbb{R}^d$ and some covariance $\Sigma \in \mathbb{R}^{d \times d}$.

Let $g : \mathbb{R}^d \rightarrow \mathbb{R}^k$ ($k \geq 1$) be continuously differentiable at θ .

Then, k functions take multiple value(d).

$$\begin{aligned} \nabla f &= \begin{pmatrix} \nabla f_1 & \nabla f_2 & \dots & \nabla f_k \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_k}{\partial x_1} \\ \vdots & \dots & \vdots \\ \frac{\partial f_1}{\partial x_d} & \dots & \frac{\partial f_k}{\partial x_d} \end{pmatrix}. \end{aligned} \quad \sqrt{n}(g(T_n) - g(\theta)) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}_k(0, \nabla g(\theta)^T \Sigma \nabla g(\theta)),$$

$k \times d$ $d \times d$ $d \times k$

$$\sqrt{n}(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\vec{\theta})) \xrightarrow[n \rightarrow \infty]{(d)} \nabla \mathbf{g}(\vec{\theta})^T \mathbf{T} \sim \mathcal{N}\left(\mu, \nabla \mathbf{g}(\vec{\theta})^T \Sigma_{\mathbf{X}} \nabla \mathbf{g}(\vec{\theta})\right) \quad (\mathbf{T} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{X}})).$$

$k \times k$

where $\nabla g(\theta) = \frac{\partial g}{\partial \theta}(\theta) = \left(\frac{\partial g_j}{\partial \theta_i} \right)_{\substack{1 \leq i \leq d \\ 1 \leq j \leq k}} \in \mathbb{R}^{d \times k}$.

rows are gradients of function from g_1 to g_k
columns are function g_j take partial derivative with respect to x_1 to x_d

Fisher Information

Definition: Fisher information

Define the log-likelihood for **one observation** as:

$$\ell(\theta) = \log L_1(X, \theta), \quad \theta \in \Theta \subset \mathbb{R}^d$$

Assume that ℓ is a.s. twice differentiable. Under some regularity conditions, the *Fisher information* of the statistical model is defined as:

$$I(\theta) = \mathbb{E}\left[\nabla\ell(\theta)\nabla\ell(\theta)^T\right] - \mathbb{E}\left[\nabla\ell(\theta)\right]\mathbb{E}\left[\nabla\ell(\theta)\right]^T = -\mathbb{E}\left[\mathbf{H}\ell(\theta)\right].$$

If $\Theta \subset \mathbb{R}$, we get:

$$I(\theta) = \text{var}[\ell'(\theta)] = -\mathbb{E}[\ell''(\theta)]$$

Fisher information of the Bernoulli experiment

Let $X \sim \text{Ber}(p)$.

$$\ell(p) = X \log p + (1-X) \log (1-p)$$

X is r.v., nothing random in p
Both of X and p are unknown

$$\ell'(p) = \frac{X}{p} - \frac{1-X}{1-p}$$

$$\text{var}[\ell'(p)] = \frac{1}{p(1-p)}$$

$$\ell''(p) = -\frac{X}{p^2} - \frac{1-X}{(1-p)^2}$$

$$-\mathbb{E}[\ell''(p)] = \frac{1}{p(1-p)}$$

Asymptotic normality of the MLE

Theorem

Let $\theta^* \in \Theta$ (the *true parameter*). Assume the following:

1. The parameter is identifiable. ✓

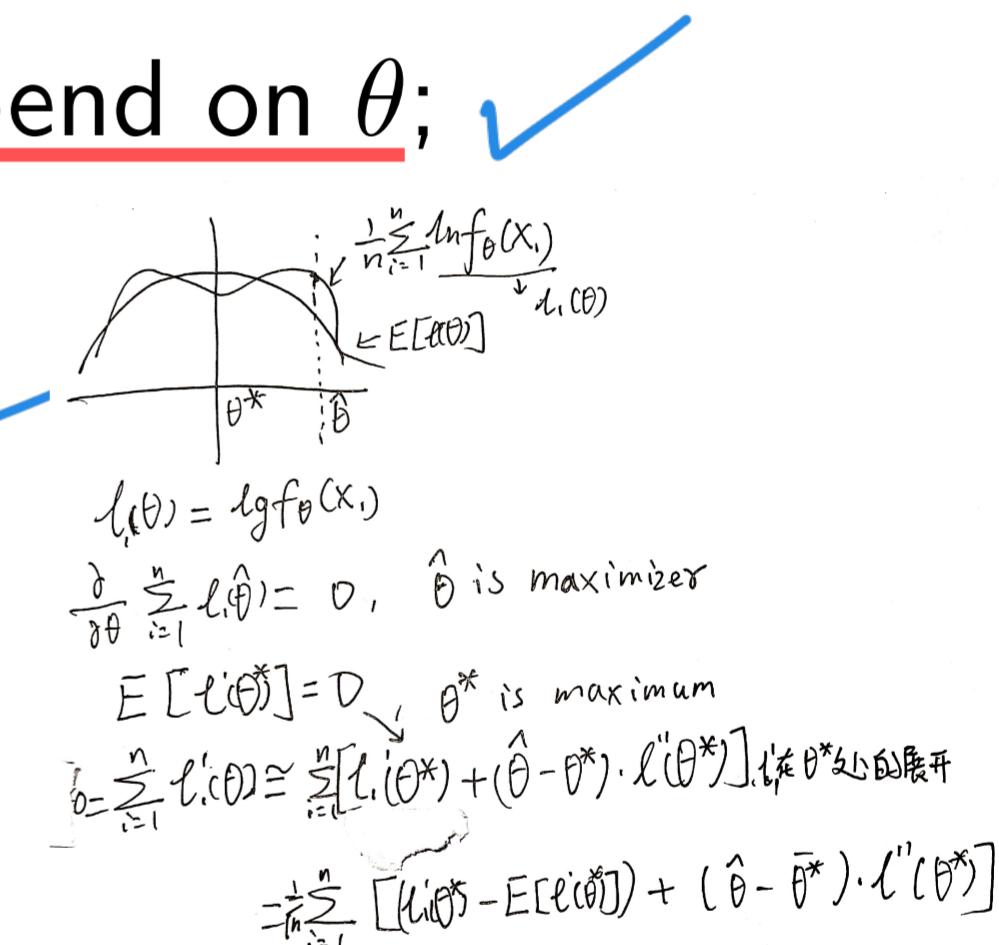
KL to 0 , theta to thetaMLE

2. For all $\theta \in \Theta$, the support of \mathbb{P}_θ does not depend on θ ; ✓

3. θ^* is not on the boundary of Θ ; ✓

4. $I(\theta)$ is invertible in a neighborhood of θ^* ; ✓

5. A few more technical conditions. ↗



Then, $\hat{\theta}_n^{MLE}$ satisfies:

consistent

► $\hat{\theta}_n^{MLE} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \theta^*$ w.r.t. \mathbb{P}_{θ^*} ;

► $\sqrt{n} (\hat{\theta}_n^{MLE} - \theta^*) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}_d (0, I(\theta^*)^{-1})$ w.r.t. \mathbb{P}_{θ^*} .

$$\begin{aligned} 0 &\stackrel{D}{=} N(0, \frac{\text{Var}(l'(0^*))}{I(0^*)}) + l'(0^*) \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n l''(\theta^*) \\ &\Rightarrow J_n(\hat{\theta} - \theta^*) \sim N(0, \frac{I(\hat{\theta})}{I(\theta^*)}) \end{aligned}$$

The method of moments

Moments

Let X_1, \dots, X_n be an i.i.d. sample associated with a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$.

- ▶ Assume that $E \subseteq \mathbb{R}$ and $\Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$.
- ▶ *Population moments*: Let $m_k(\theta) = \mathbb{E}_\theta[X_1^k]$, $1 \leq k \leq d$.
- ▶ *Empirical moments*: Let $\hat{m}_k = \overline{\bar{X}_n^k} = \frac{1}{n} \sum_{i=1}^n \hat{X}_i^k$, $1 \leq k \leq d$.

MM的估计是这样算的：

find relationships between my parameter and my moments.

找出某一moment和参数的关系，再用样本均数替代期望

对于正态r.v.，一次moment给了我mu参数，二次moment给了我一些关于sigma的信息。

- ▶ From LLN,

$$\hat{m}_k \xrightarrow[n \rightarrow \infty]{\mathbb{P}/a.s} m_k(\theta)$$

More compactly, we say that the whole vector converges:

$$(\hat{m}_1, \dots, \hat{m}_d) \xrightarrow[n \rightarrow \infty]{\mathbb{P}/a.s} (m_1(\theta), \dots, m_d(\theta))$$

Moments estimator

Let

$$\begin{aligned} M &: \Theta \rightarrow \mathbb{R}^d \\ \theta &\mapsto \underbrace{M(\theta)}_{\text{maps moments from 1 to } d} = (m_1(\theta), \dots, m_d(\theta)). \end{aligned}$$

maps moments from 1 to d

Assume M is one to one: 每一级moment给了一些关于参数的信息
假设有唯一解，直接把参数解出来

$$\theta = M^{-1}(m_1(\theta), \dots, m_d(\theta)).$$

Definition

Moments estimator of θ :

not depend on θ
but depend on data

$$\hat{\theta}_n^{MM} = M^{-1}(\hat{m}_1, \dots, \hat{m}_d),$$

provided it exists.

delta method
 $g = M^{-1}$

Statistical analysis

- Recall $M(\theta) = (m_1(\theta), \dots, m_d(\theta))$;
- Let $\hat{M} = (\hat{m}_1, \dots, \hat{m}_d)$. ✓
- Let $\Sigma(\theta) = \text{Cov}_{\theta}(X_1, X_1^2, \dots, X_1^d)$ ✓ be the covariance matrix of the random vector $(X_1, X_1^2, \dots, X_1^d)$, which we assume to exist.
- Assume M^{-1} is continuously differentiable at $M(\theta)$.

Delta method; $g = \text{inverse}(x)$

the map function can be anything (like map x to $\cos(6\pi x)$)

The image shows handwritten mathematical notes on a blackboard. At the top, it says "CLT" and then shows the formula $\sqrt{n} \left(\bar{X}_n^k - m_k(\theta) \right) \xrightarrow[n \rightarrow \infty]{(d)} N(0, \text{Var}(X_1^k))$. Below this, it says "Multivariate CLT" and shows the formula $\sqrt{n} \left(\begin{pmatrix} \bar{X}_n^1 \\ \vdots \\ \bar{X}_n^d \end{pmatrix} - M(\theta) \right) \xrightarrow[n \rightarrow \infty]{(d)} N_d(0, \Sigma_{\theta} \begin{pmatrix} X_1^1 \\ \vdots \\ X_1^d \end{pmatrix})$.

Method of moments (5)

Remark: The method of moments can be extended to more general moments, even when $E \not\subset \mathbb{R}$.

- ▶ Let $g_1, \dots, g_d : E \rightarrow \mathbb{R}$ be given functions, chosen by the practitioner.
e.g. $g_k(x) = \cos(2\pi k x)$
- ▶ Previously, $\underline{g_k(x)} = \underline{x^k}$, $x \in E = \mathbb{R}$, for all $k = 1, \dots, d$.
We want use less functions. Wavelets are these kind of function which contain a lot of information
- ▶ Define $m_k(\theta) = \mathbb{E}_\theta[g_k(X)]$, for all $k = 1, \dots, d$.
- ▶ Let $\Sigma(\theta) = \text{Cov}_\theta(g_1(X_1), g_2(X_1), \dots, g_d(X_1))$ be the covariance matrix of the random vector $(g_1(X_1), g_2(X_1), \dots, g_d(X_1))$, which we assume to exist.
- ▶ Assume M is one to one and M^{-1} is continuously differentiable at $M(\theta)$.
get back theta

Generalized method of moments

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_0) &\xrightarrow{d} \mathcal{N}(0, \Sigma(\theta_0)) \\ \sqrt{n}(\nabla M^{-1}(\hat{\theta}) - \nabla M^{-1}(\theta_0)) &\xrightarrow{d} \mathcal{N}(0, \nabla M^{-1}(\theta_0)^\top \Sigma(\theta_0)^{-1} \nabla M')\end{aligned}$$

Applying the multivariate CLT and Delta method yields:

Theorem

$$\sqrt{n}(\hat{\theta}_n^{MM} - \theta) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \Gamma(\theta)) \quad (\text{w.r.t. } \mathbb{P}_\theta),$$

$$\text{where } \Gamma(\theta) = \left[\frac{\partial M^{-1}}{\partial \theta}(M(\theta)) \right]^\top \Sigma(\theta) \left[\frac{\partial M^{-1}}{\partial \theta}(M(\theta)) \right].$$

MLE vs. Moment estimator

- ▶ Comparison of the quadratic risks: In general, the MLE is more accurate.
bias² + variance

Cramer Rao Bound: No unbiased estimator, no matter how you cook it up, none of them can actually have a variance which is smaller than the 1/Fisher information.

- ▶ MLE still gives good results if model is misspecified



- ▶ Computational issues: Sometimes, the MLE is intractable but MM is easier (polynomial equations)
hard to find maximum

M-estimation

M-estimators

Idea:

- ▶ Let X_1, \dots, X_n be i.i.d with some unknown distribution \mathbb{P} in some sample space E ($E \subseteq \mathbb{R}^d$ for some $d \geq 1$).
-
- ▶ No statistical model needs to be assumed (similar to ML).
- ▶ Goal: estimate some parameter μ^* associated with \mathbb{P} , e.g. its mean, variance, median, other quantiles, the true parameter in some statistical model...
- ▶ Find a function $\rho : E \times \mathcal{M} \rightarrow \mathbb{R}$, where \mathcal{M} is the set of all possible values for the unknown μ^* , such that:

通过函数找到data的信息

expectation of sth

$$Q(\mu) := \mathbb{E} [\rho(X_1, \mu)]$$

achieves its minimum at $\mu = \mu^*$.

Examples (1)

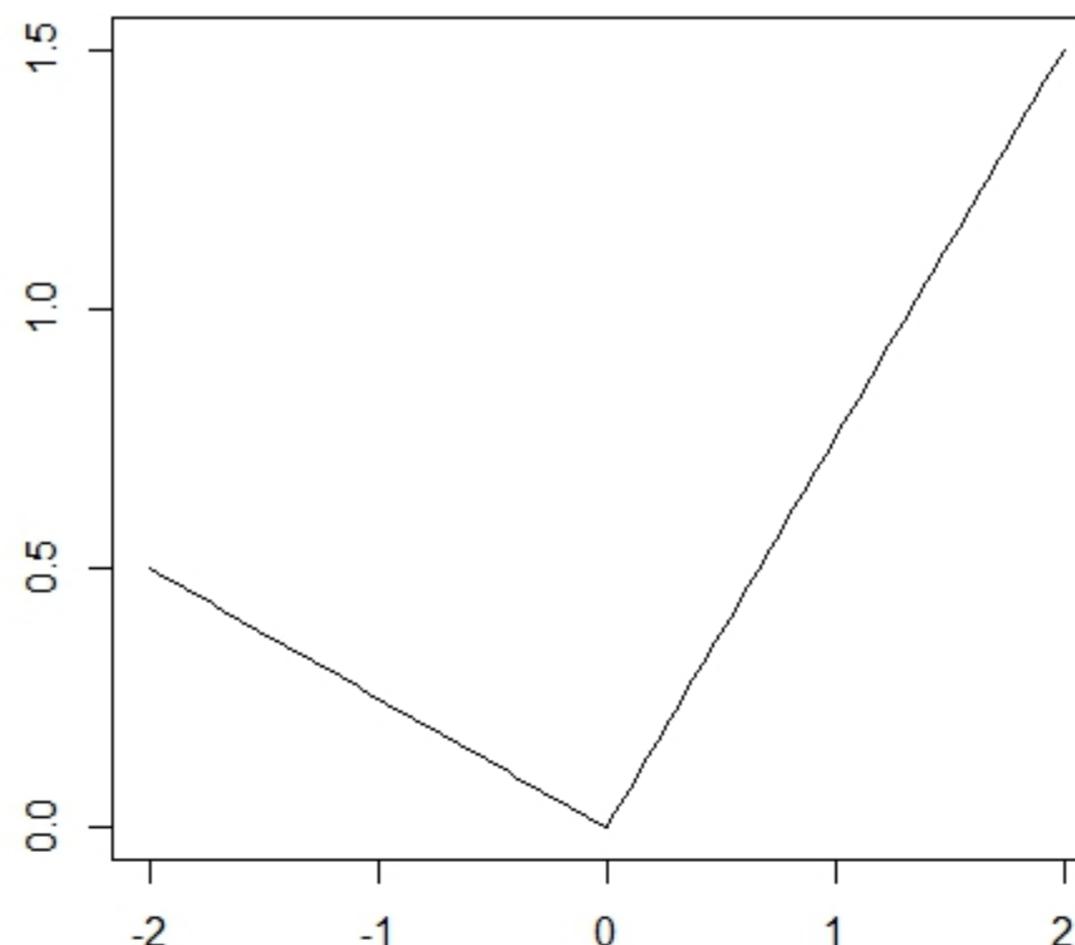
- ▶ If $E = \mathcal{M} = \mathbb{R}$ and $\rho(x, \mu) = (x - \mu)^2$, for all $x \in \mathbb{R}, \mu \in \mathbb{R}$:
 $\mu^* = \mathbb{E}[X]$
- ▶ If $E = \mathcal{M} = \mathbb{R}^d$ and $\rho(x, \mu) = \|x - \mu\|_2^2$, for all
 $x \in \mathbb{R}^d, \mu \in \mathbb{R}^d$: $\mu^* = \mathbb{E}[X] \in \mathbb{R}^d$
- ▶ If $E = \mathcal{M} = \mathbb{R}$ and $\rho(x, \mu) = |x - \mu|$, for all $x \in \mathbb{R}, \mu \in \mathbb{R}$:
 μ^* is a median of \mathbb{P} .

Examples (2)

If $E = \mathcal{M} = \mathbb{R}$, $\alpha \in (0, 1)$ is fixed and $\rho(x, \mu) = C_\alpha(x - \mu)$, for all $x \in \mathbb{R}, \mu \in \mathbb{R}$: μ^* is a α -quantile of \mathbb{P} .

Check function

$$C_\alpha(x) = \begin{cases} -(1 - \alpha)x & \text{if } x < 0 \\ \alpha x & \text{if } x \geq 0. \end{cases}$$



MLE is an M-estimator

Assume that $(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$ is a statistical model associated with the data.

Theorem

Let $\mathcal{M} = \Theta$ and $\rho(x, \theta) = -\log L_1(x, \theta)$, provided the likelihood is positive everywhere. Then,

$$\mu^* = \theta^*,$$

where $\mathbb{P} = \mathbb{P}_{\theta^*}$ (i.e., θ^* is the true value of the parameter).

Definition

replace E with $\frac{1}{n} \sum_{i=1}^n$

- ▶ Define $\hat{\mu}_n$ as a minimizer of:

$$Q_n(\mu) := \frac{1}{n} \sum_{i=1}^n \rho(X_i, \mu).$$

empirical: average of sample

- ▶ Examples: Empirical mean, empirical median, empirical quantiles, MLE, etc.

Statistical analysis

- Let $J(\mu) = + \frac{\partial^2 Q}{\partial \mu \partial \mu^\top}(\mu)$ ($= + \mathbb{E} \left[\frac{\partial^2 \rho}{\partial \mu \partial \mu^\top}(X_1, \mu) \right]$ under some regularity conditions).
- Let $K(\mu) = \text{Cov} \left[\frac{\partial \rho}{\partial \mu}(X_1, \mu) \right]$.
- **Remark:** In the log-likelihood case (write $\mu = \theta$),

$$J(\theta) = K(\theta) = \mathcal{I}(\theta) \quad (\text{Fisher information})$$

Asymptotic normality

Let $\mu^* \in \mathcal{M}$ (the *true* parameter). Assume the following:

1. μ^* is the only minimizer of the function Q ;
2. $J(\mu)$ is invertible for all $\mu \in \mathcal{M}$;
3. A few more technical conditions.

Then, $\hat{\mu}_n$ satisfies:

- $\hat{\mu}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \mu^*$;
- $\sqrt{n} (\hat{\mu}_n - \mu^*) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N} \left(0, J(\mu^*)^{-1} K(\mu^*) J(\mu^*)^{-1} \right)$.

M-estimators in robust statistics

Example: Location parameter

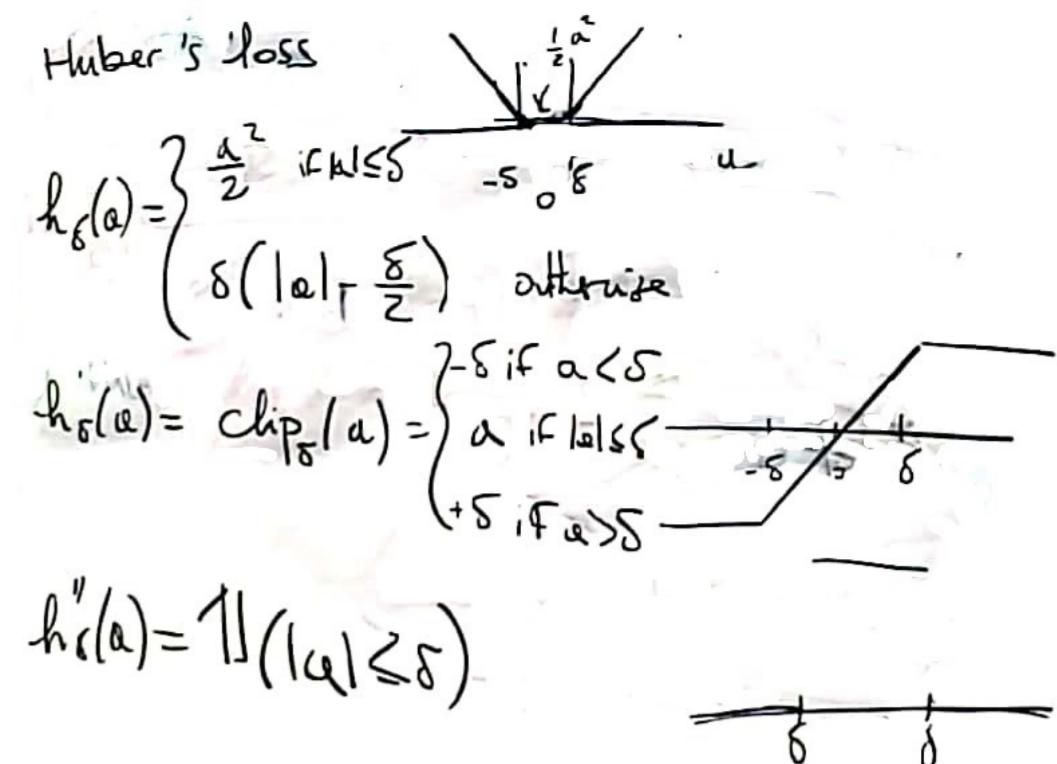
If X_1, \dots, X_n are i.i.d. with density $f(\cdot - m)$, where:

- ▶ f is an unknown, positive, even function (e.g., the Cauchy density);
- ▶ m is a real number of interest, a *location parameter*;

How to estimate m ?

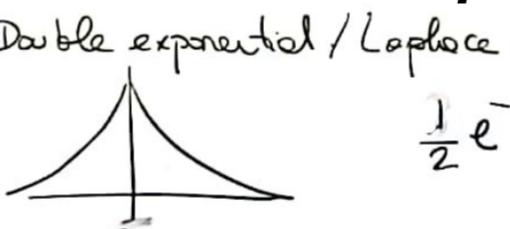
- ▶ M-estimators: empirical mean, empirical median,
- ▶ Compare their risks or asymptotic variances;
- ▶ The empirical median is more *robust*.

median minimizes, $f(\mu) = E[|X - \mu|] \geq E[|X - \text{median}|]$; but the absolute function is not der



$$f(x; x_0, \gamma) = \frac{1}{\pi \gamma \left[1 + \left(\frac{x-x_0}{\gamma} \right)^2 \right]} = \frac{1}{\pi \gamma} \left[\frac{\gamma^2}{(x-x_0)^2 + \gamma^2} \right],$$

(1) 数学期望 $E\zeta$ 不存在, 即 $\int_{-\infty}^{+\infty} |x| f_C(x) dx = \int_{-\infty}^{+\infty} |x| \frac{1}{\pi \gamma^2 + (x-x_0)^2} dx = \infty$
 (2) 方差 $D\zeta$ 不存在。
 (3) 高阶矩均不存在。



$$\frac{1}{2} e^{-|x-m|}$$

$$\log L(x_1, \dots, x_n; m) = \sum_{i=1}^n \log \left(\frac{1}{2} e^{-|x_i-m|} \right)$$

$$\hat{m}_{\text{MLE}} = \arg \min_m \sum_{i=1}^n |x_i - m| \quad (\text{empirical median})$$

$$\begin{aligned} P(x, m) &= h_\delta(x-m) \quad \text{Huber's loss} \\ J(m) &= E[h'_\delta(x-m)] = E[\mathbb{1}(|x-m| \leq \delta)] \\ &= P(|x-m| \leq \delta) = 2 \int_0^\delta f_0(x) dx \\ &= \int_0^\delta e^{-x} dx \quad \text{by symmetry} \\ &= 1 - e^{-\delta} \end{aligned}$$

$$\begin{aligned} K(m) &= \text{Var}(h_\delta(x-m)) = \text{Var}(\text{clip}_\delta(x-m)) \\ E[\text{clip}_\delta(x-m)] &= 0 \quad \text{by symmetry} \\ E[\text{clip}_\delta^2(x-m)] &= 2 \left[\int_0^\delta x^2 f_0(x) dx + \int_\delta^\infty \delta^2 f_0(x) dx \right] \\ &= 2 \left[\int_0^\delta x^2 \frac{e^{-x}}{2} dx + \delta^2 \int_\delta^\infty \frac{e^{-x}}{2} dx \right] \\ \int_0^\delta x^2 \cdot e^{-x} dx &= -x^2 e^{-x} \Big|_0^\delta + 2 \int_0^\delta x \cdot e^{-x} dx \\ &= -\delta^2 e^{-\delta} + 2[-x \cdot e^{-x}]_0^\delta + \int_0^\delta e^{-x} dx \\ &= -\delta^2 e^{-\delta} + 2\delta e^{-\delta} + 2 - 2e^{-\delta} \end{aligned}$$

$$\begin{aligned} K(m) &= -\delta^2 e^{-\delta} - 2\delta e^{-\delta} + 2 - 2e^{-\delta} + 2e^{-\delta} \\ \hat{m}_\delta &= \arg \min_{\mu \in R} \sum_{i=1}^n h_\delta(X_i - \mu) \end{aligned}$$

$$\begin{aligned} \ln(\hat{m}_\delta - m) &\xrightarrow{n \rightarrow \infty} N(0, \frac{K(m)}{J(m)}) = N(0, \frac{2(1 - \delta e^{-\delta} - e^{-\delta})}{(1 - e^{-\delta})^2}) \\ \xrightarrow{\text{asymptotic variance}} 2 &\quad \text{the minimum asymptotic variance} \\ 2 & \quad \text{is for } \delta = 0 \Rightarrow \hat{m}_0 = \hat{m}_{\text{MLE}} \\ \hat{m}_\infty &= \arg \min_m \sum_{i=1}^n (x_i - m)^2 \end{aligned}$$

Recap

- ▶ Three principled methods for estimation: maximum likelihood, Method of moments, M-estimators
- ▶ Maximum likelihood is an example of M -estimation
- ▶ Method of moments inverts the function that maps parameters to moments
- ▶ All methods yield to asymptotic normality under regularity conditions
- ▶ Asymptotic covariance matrix can be computed using multivariate Δ -method
- ▶ For MLE, asymptotic covariance matrix is the inverse Fisher information matrix