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

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

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

Definition

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

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

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 $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is a simple function of the PMF's p_θ and $p_{\theta'}$:

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

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 $A \subset E$.

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

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

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

Properties of Total variation

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

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

Exercises

Compute:

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

$E = \{0, 1\}$

$= \frac{1}{2}\Big[\underbrace{|.5 - .9|}_{0.4} + |.5 - .1|\Big] = \frac{0.8}{2} = 0.4$

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 $\quad \Big|\overbrace{\mathbb{P}(X \in \{0, 1\})}^{=1} - \overbrace{\mathbb{P}(X + a \in \{0, 1\})}^{=0}\Big| = 1$

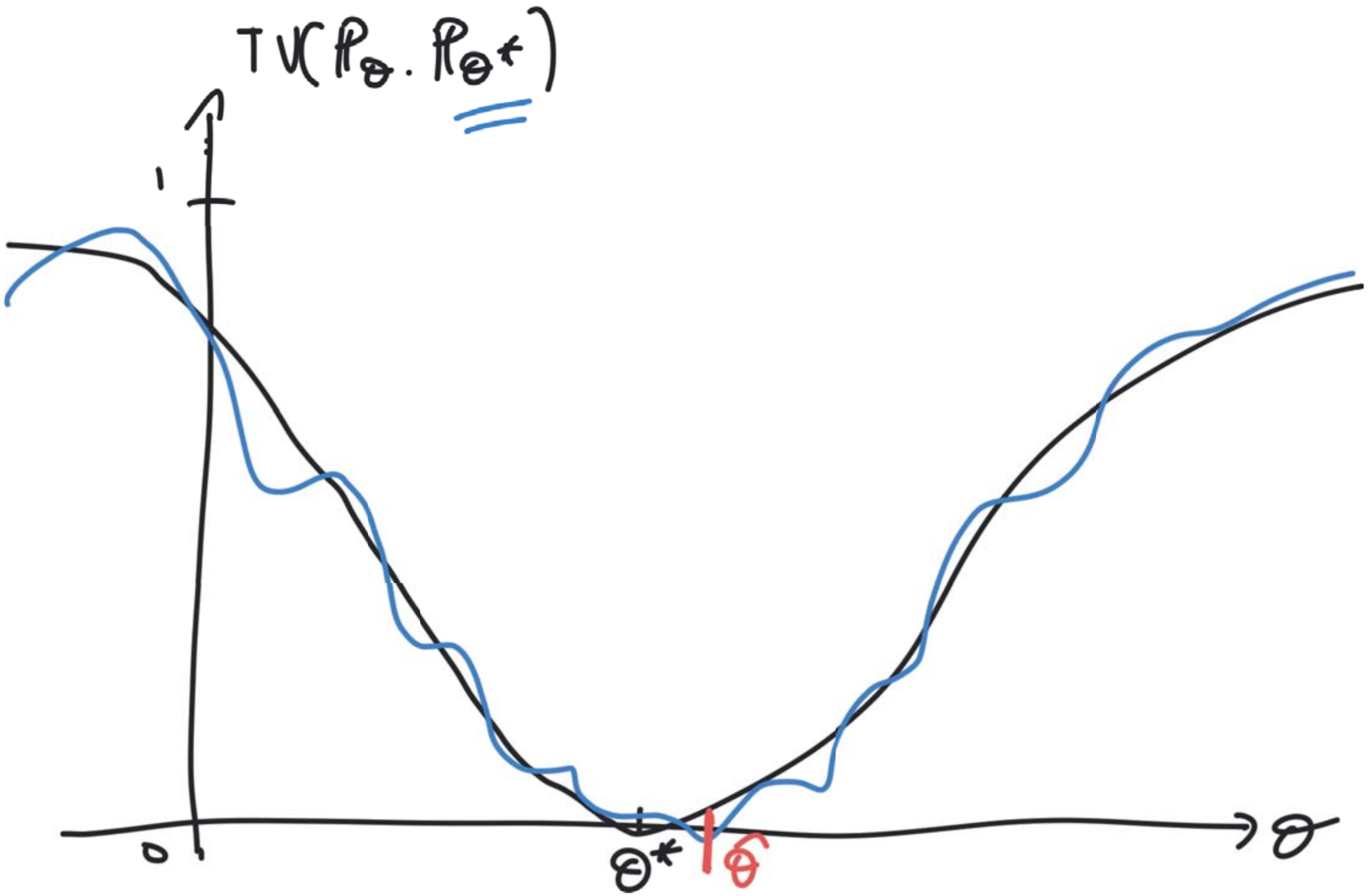
for any $a \in (0, 1)$, where $X \sim \text{Ber}(0.5)$

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

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

An estimation strategy

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



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

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-Leibler*[1] *(KL) divergence* between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is defined by

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

[1] KL-divergence is also know as "relative entropy"

Properties of KL-divergence

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

Not a distance.

This is is called a divergence

Asymmetry is the key to our ability to estimate it!

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

Maximum likelihood estimation

Estimating the KL

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

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

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

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

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

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

(blue curve)

Maximum likelihood

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

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

This is the **maximum likelihood principle**.

Likelihood, Discrete case (1)

Let $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \ldots, X_n$. 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:

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

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

Likelihood for the Bernoulli model

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

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

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

Likelihood for the Poisson model

Example 2 (Poisson model):

If $X_1, \ldots, X_n \overset{iid}{\sim} \mathsf{Poiss}(\lambda)$ for some $\lambda > 0$:

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

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

$$\mathbb{P}_\lambda(X_i = x_i) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

$$\Rightarrow \underset{\substack{\| \\ \prod_{i=1}^n \mathbb{P}_\lambda(X_i = x_i)}}{L(x_1 \ldots x_n; \lambda)} = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\lambda}$$

Likelihood, Continuous case

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

Definition

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

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

Likelihood for the Gaussian model

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

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

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

Exercises

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

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

b) What is Θ? $(0, \infty)$

c) Find the likelihood of the model.

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

Exercise

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

a) What is E?

$[0, \infty)$

b) What is Θ?

$[0, \infty)$

c) Find the likelihood of the model.

$$L(x_1, \ldots, x_n; b) = \frac{1}{b^n} \mathbb{1}\left(\max_i X_i \leq b\right)$$

Maximum likelihood estimator

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

Definition

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

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

provided it exists.

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

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

Interlude: maximizing/minimizing functions

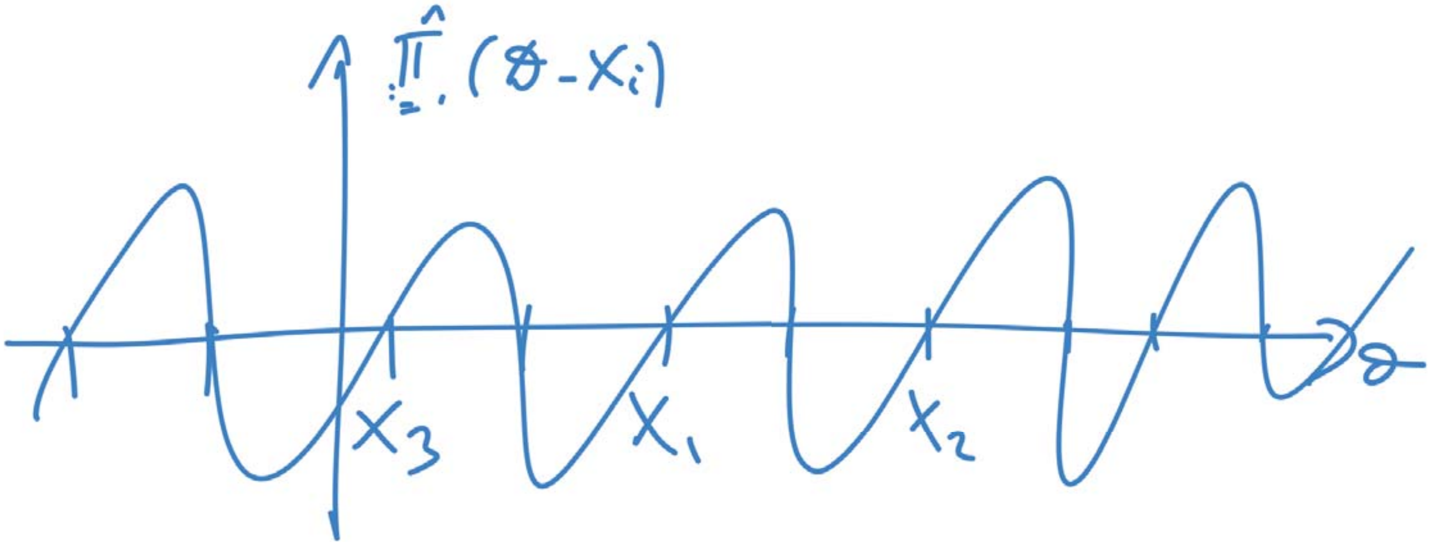
OPTIMIZATION

Note that

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

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

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

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: $h : \Theta \subset \mathbb{R}^d \to \mathbb{R}$, $d \geq 2$, 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:

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

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

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

$x \neq 0$

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

$$\boxed{h'(\theta) = 0\,,}$$

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 $h : \Theta \subset \mathbb{R}^d \to \mathbb{R}$ 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

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

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

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

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

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

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

Poi: $L(x_1,\dots,x_n;\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \cdot e^{-\lambda n}$

$h(\lambda) = \lg L(x_1,\dots,x_n;\lambda) = \sum_{i=1}^n x_i \cdot \lg \lambda - n\lambda - \lg(\prod_{i=1}^n x_i!)$

$h'(\lambda) = \frac{\sum_{i=1}^n x_i}{\lambda} - n$; $h''(\lambda) = -\frac{\sum_{i=1}^n x_i}{\lambda^2} \le 0.$

h is concave, $h'(\hat\lambda) = 0 \Rightarrow \hat\lambda = \frac{1}{n} \cdot \sum_{i=1}^n x_i = \bar X$

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

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

$$\hat S_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar X_n)^2$$

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

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

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

h is concave, $\nabla h(\hat\mu, \hat\sigma^2) = 0 \Leftrightarrow \begin{matrix} \hat\mu = \bar X \\ \hat\sigma^2 = \hat S \end{matrix}$

Consistency of maximum likelihood estimator

Under mild regularity conditions, we have

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

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

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

Moreover, the minimizer of the right-hand side is θ^* if the parameter is identifiable.

Technical conditions allow to transfer this convergence to the minimizers.

Covariance

$$\hat{\theta} = \begin{pmatrix} \overline{X_n} \\ \widehat{S}_n \end{pmatrix}$$

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

- $\text{Cov}(X, \cancel{Y}) =$ 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 $(\alpha, \beta) \in \mathbb{R}^2 \setminus \{(0,0)\}$.

Rademacher

indep

Take $X \sim \mathcal{N}(0, 1)$, $B \sim \text{Ber}(1/2)$, $R = 2B - 1 \sim \text{Rad}(1/2)$. Then

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

$\text{cov}(X,Y) = E[X \cdot Y] - E[X]E[Y]$
$= E[X \cdot R \cdot X]$
$= E[X^2 \cdot R] = E[R] \cdot E[X^2]$
$= 0$

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

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

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

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

The term on the ith row and jth column is

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

In particular, on the diagonal, we have $\Sigma_{ii} = \mathrm{Cov}\left(X^{(i)}, X^{(i)}\right) = \mathrm{Var}\left(X^{(i)}\right)$

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

$$\mathrm{Cov}(AX + B) = \mathrm{Cov}(AX) = A\,\mathrm{Cov}(X)\,A^T = A \Sigma A^T$$

The multivariate Gaussian distribution

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

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

More generally, a Gaussian vector[3] $X \in \mathbb{R}^d$, 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:

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

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

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

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

Equivalently

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

Multivariate Delta method

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

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

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

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

k functions take multiple value(d).

$$\nabla f = \begin{pmatrix} | & | & & | \\ \nabla f_1 & \nabla f_2 & \dots & \nabla f_k \\ | & | & & | \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_k}{\partial x_1} \\ \vdots & \dots & \vdots \\ \frac{\partial f_1}{\partial x_d} & \dots & \frac{\partial f_k}{\partial x_d} \end{pmatrix}.$$

$$\sqrt{n}\,(g(T_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k\Big(0, \underset{k\times k}{\overset{k\times d}{\nabla g(\theta)^T}\ \overset{d\times d}{\Sigma}\ \overset{d\times k}{\nabla g(\theta)}}\Big),$$

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

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 from x_1 to x_d

Fisher Information

Definition: Fisher information

Define the log-likelihood for one observation as:

pdf/pmf $f_\theta(X)$

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

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

d×d cov($\nabla \ell$(θ))

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

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

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

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

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

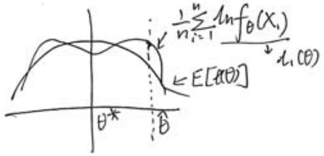
$$\ell''(p) = -\frac{X}{p^2} - \frac{1-X}{(1-p)^2} \qquad -\mathbb{E}[\ell''(p)] = \frac{1}{p(1-p)}$$

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 $\mathbb{P}_\theta$ 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.



$l_1(\theta) = \lg f_\theta(X_1)$

$\frac{\partial}{\partial\theta}\sum_{i=1}^{n} l_i(\hat{\theta}) = 0$, $\hat{\theta}$ is maximizer

$E[l'(\theta^*)] = 0$, θ^* is maximum

$0 = \sum_{i=1}^{n} l_i'(\hat{\theta}) \cong \sum_{i=1}^{n}[l_i'(\theta^*) + (\hat{\theta} - \theta^*) \cdot l''(\theta^*)]$ 在θ^*处的展开

$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(l_i'(\theta^*) - E[l'(\theta^*)]) + (\hat{\theta} - \theta^*) \cdot l''(\theta^*)]$

$0 \cong N(0, \underbrace{Var(l'(\theta^*))}_{I(\theta^*)}) + \sqrt{n}(\hat{\theta} - \theta^*) \cdot \underbrace{\frac{1}{\sqrt{n}\cdot\sqrt{n}}\sum_{i=1}^{n} l''(\theta^*)}_{LLN \xrightarrow[n\to\infty]{P} -I(\theta^*)}$

$\Rightarrow \sqrt{n}(\hat{\theta} - \theta^*) \sim N(0, \frac{\cancel{I(\theta^*)}}{I(\theta^*)^{\cancel{2}}})$

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

- consistent ▶ $\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$ w.r.t. $\mathbb{P}_{\theta^*}$;
- ▶ $\sqrt{n}\left(\hat{\theta}_n^{MLE} - \theta^*\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d\left(0, I(\theta^*)^{-1}\right)$ w.r.t. $\mathbb{P}_{\theta^*}$.

The method of moments

Moments

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

- 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 $\hat{m}_k = \overline{X_n^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, $1 \leq k \leq d$.
- From LLN,

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

More compactly, we say that the whole vector converges:

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

Moments estimator

Let

$$\begin{array}{rcccl} M & : & \Theta & \to & \mathbb{R}^d \\ & & \theta & \mapsto & \underbrace{M(\theta)} = (m_1(\theta), \dots, m_d(\theta)) \,. \end{array}$$

maps moments from 1 to d

Assume M is one to one: unique solution

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

Definition

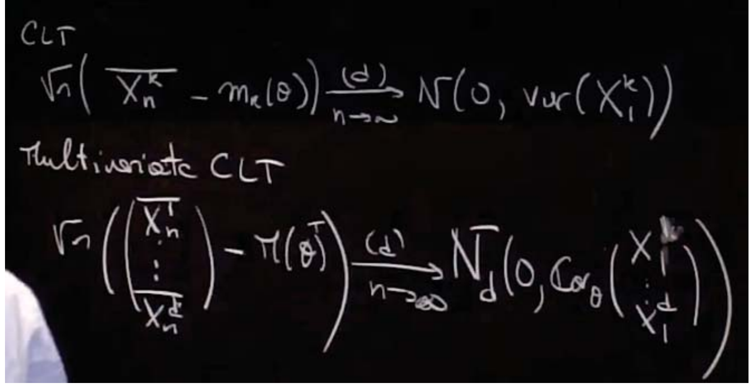
Moments estimator of θ:

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

not depend on θ

provided it exists.

Statistical analysis



the map function can be anything i like(map x to cos(6*pi*x))

Delta method; g = inverse(x)

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

Method of moments (5)

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

- Let $g_1, \ldots, g_d : E \to \mathbb{R}$ be given functions, chosen by the practitioner.

  e.g. $g_k(x) = \cos(2\pi k \dot{x})$

  Less functions. Wavelets are these kind of function which contain a lot of information

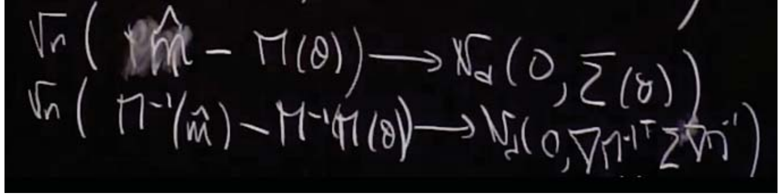
- Previously, $g_k(x) = x^k$, $\quad x \in E = \mathbb{R}$, for all $k = 1, \ldots, d$.

- Define $m_k(\theta) = \mathbb{E}_\theta[g_k(X)]$, for all $k = 1, \ldots, d$.

- Let $\Sigma(\theta) = \text{Cov}_\theta(g_1(X_1), g_2(X_1), \ldots, g_d(X_1))$ be the covariance matrix of the random vector $(g_1(X_1), g_2(X_1), \ldots, g_d(X_1))$, 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



Applying the multivariate CLT and Delta method yields:

Theorem

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

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

MLE vs. Moment estimator

- Comparison of the quadratic risks: In general, the MLE is more accurate.

  bias +variance

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

- MLE still gives good results if model is misspecified

  hard to find maximum

- Computational issues: Sometimes, the MLE is intractable but MM is easier (polynomial equations)

M-estimation

M-estimators

Idea:

- Let $X_1, \ldots, X_n$ 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 $\rho : E \times \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the set of all possible values for the unknown μ^*, such that:

expectation of sth

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

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

Examples (1)

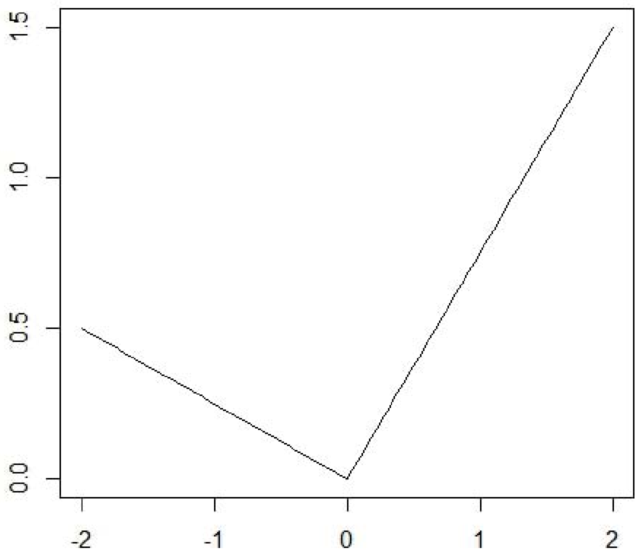
- If $E = \mathcal{M} = \mathbb{R}$ and $\rho(x, \mu) = (x - \mu)^2$, for all $x \in \mathbb{R}, \mu \in \mathbb{R}$: $\mu^* = \mathbb{E}[X]$
- If $E = \mathcal{M} = \mathbb{R}^d$ and $\rho(x, \mu) = \|x - \mu\|_2^2$, for all $x \in \mathbb{R}^d, \mu \in \mathbb{R}^d$: $\mu^* = \mathbb{E}[X] \in \mathbb{R}^d$
- 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 $\rho(x, \overset{\mu}{\cancel{\theta}}) = -\log L_1(x, \overset{\mu}{\cancel{\theta}})$, 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 $\mathbb{E}$ with $\frac{1}{n}\sum_{i=1}^{n}$

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

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

empirical: average of sample

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

Statistical analysis

Hessian

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

$$J(\theta) = K(\theta) = I(\theta)$$

(Fisher information)

Asymptotic normality

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

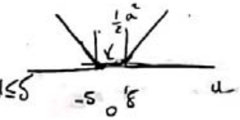
1. μ^* is the only minimizer of the function $\mathcal{Q}$;
2. $J(\mu)$ is invertible for all $\mu \in \mathcal{M}$;
3. A few more technical conditions.

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

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

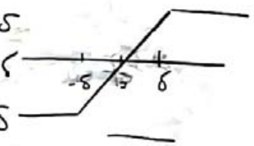
M-estimators in robust statistics

Huber's loss



$$h_\delta(a) = \begin{cases} \frac{a^2}{2} & \text{if } |a| \le \delta \\ \delta(|a| - \frac{\delta}{2}) & \text{otherwise} \end{cases}$$

$$h'_\delta(a) = \mathrm{clip}_\delta(a) = \begin{cases} -\delta & \text{if } a < \delta \\ a & \text{if } |a| \le \delta \\ +\delta & \text{if } a > \delta \end{cases}$$



$$h''_\delta(a) = \mathbb{1}(|a| \le \delta)$$

Example: Location parameter

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

- f is an unknown, positive, even function (e.g., the Cauchy density);

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

(1)数学期望 $E\zeta$ 不存在，即

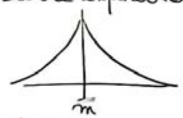
$$\int_{-\infty}^{+\infty} |x| f_\zeta(x) dx = \int_{-\infty}^{+\infty} |x| \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x - x_0)^2} dx = \infty$$

(2)方差 $D\zeta$ 不存在。

(3)高阶矩均不存在。

- m is a real number of interest, a *location parameter*;

Double exponential / Laplace.



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

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

$$\hat{m}^{MLE} = \underset{m}{\mathrm{argmin}} \sum_{i=1}^{n} |X_i - m| \quad \text{(empirical median)}$$

$\rho(X, m) = h_\delta(X - m)$ Huber's loss

$$J(m) = E[h''_\delta(X-m)] = E[\mathbb{1}(|X-m| \le \delta)]$$
$$= P(|X - m| \le \delta) = 2\int_0^\delta f_0(x) dx$$
$$= \int_0^\delta e^{-x} dx \quad \text{by symmetry}$$
$$= 1 - e^{-\delta}$$

$$K(m) = \mathrm{Var}(h'_\delta(X-m)) = \mathrm{Var}(\mathrm{Clip}_\delta(X-m))$$

$E[\mathrm{Clip}_\delta(X-m)] = 0$ by symmetry

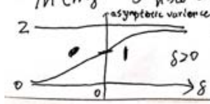
$$E[\mathrm{Clip}_\delta^2(X-m)] = 2\left[\int_0^\delta x^2 f_0(x) dx + \int_\delta^\infty \delta^2 f_0(x) dx\right]$$
$$= \cancel{2}\left[\int_0^\delta x^2 \cdot \frac{e^{-x}}{\cancel{2}} + \delta^2 \int_\delta^\infty \delta^2 \cdot \frac{e^{-x}}{2} dx\right]$$

$$\int_0^\delta x^2 \cdot e^{-x} dx = -x^2 e^{-x}\Big|_0^\delta + 2\int_0^\delta x \cdot e^{-x} dx$$
$$= -\delta^2 e^{-\delta} + 2\left[-x \cdot e^{-x}\Big|_0^\delta + \int_0^\delta e^{-x} dx\right]$$
$$= -\delta^2 \cdot e^{-\delta} - 2\delta e^{-\delta} + 2 - 2 \cdot e^{-\delta}$$

$$K(m) = \cancel{-\delta^2 e^{-\delta}} - 2\delta e^{-\delta} + 2 - 2e^{-\delta} + \cancel{\delta^2 e^{-\delta}}$$

$$\hat{m}_\delta = \underset{\mu \in \mathbb{R}}{\mathrm{argmin}} \sum_{i=1}^{n} h_\delta(X_i - \mu)$$

$$\sqrt{n}(\hat{m}_\delta - m) \xrightarrow[n \to \infty]{(d)} N\left(0, \frac{K(m)}{J^2(m)}\right) = N\left(0, \frac{2(1 - \delta e^{-\delta} - e^{-\delta})}{(1 - e^{-\delta})^2}\right)$$



$\delta > 0$

the minimum asymptotic variance is for $\delta = 0 \Rightarrow \hat{m}_0 = \hat{m}^{MLE}$

$\hat{m}_\infty = \bar{X}_n$ (Expectation)

$$\hat{m}_\infty = \underset{\mu}{\mathrm{argmin}} \sum_{i=1}^{n} (X_i - \mu)^2$$

How to estimate m ?

- M-estimators: empirical mean, empirical median, ...
- Compare their risks or asymptotic variances;

median minimizes, f(mu) = E[|X - mu|] >= E[|X - median|]; but the absolute function is not derivative.

- The empirical median is more *robust*.

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