MS = MAXIMUMUM LIKELIHOOD ESTIMATION

Let $(\Omega, \mathcal{F}, \{\mathbb{P}_\theta\}_{\theta\in\Theta})$ be a statistical model. Assume that n observations $X_1, \dots, X_n$ are drawn from $\mathbb{P}_{\theta^*}$, for some $\theta^* \in \Theta$. The parameter θ^* is referred to as the true parameter, and the aim is to estimate it based solely on the observations $X_1, \dots, X_n$.

Given our family of suspects $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$, the goal is thus to select which member of the family is the most likely to have given rise to $X_1, \dots, X_n$. The approach discussed in this chapter selects the distribution in $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ that matches the most the empirical distribution of $X_1, \dots, X_n$. The discrepancy between the two distributions is measured by the Kullbach-Leibler divergence, that we introduce in the next section.

I. MAXIMUM LIKELIHOOD ESTIMATION

I.1. Kullbach-Leibler divergence.

The Kullbach-Leibler (KL) divergence between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is defined by:

$$KL(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'}) = \sum_{x\in\Omega} p_\theta(x) \log\left\{\frac{p_\theta(x)}{p_{\theta'}(x)}\right\} \quad \text{if } \Omega \text{ is at most countable,}$$

with $p_\theta(x) = \mathbb{P}_\theta(X = x)$, $X(\omega) = \omega$

and

$$KL(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'}) = \int_\Omega f_\theta(x) \log\left\{\frac{f_\theta(x)}{f_{\theta'}(x)}\right\} dx \quad \text{if } \Omega \text{ is continuous}$$

Assuming that the Xs are absolutely continuous, with density f_θ: $\mathbb{P}_\theta(X \in B) = \int_B f_\theta(x)\,dx$, $X(\omega) = \omega$

↳ Note that $KL(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'}) = \mathbb{E}_\theta\left\{\log\frac{f_\theta(x)}{f_{\theta'}(x)}\right\}$,

and similarly in the discrete case.

x Properties of the KL divergence:

(i) KL divergence is *not* symmetric. In general, we have that $KL(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'}) \neq KL(\mathbb{P}_{\theta'} \,\|\, \mathbb{P}_\theta)$.

⇒ Be careful in which order you write $\mathbb{P}_\theta$ & $\mathbb{P}_{\theta'}$.

Consequence: $KL(.\,\|\,.)$ is *not* a metric.

(ii) $KL(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'}) \geq 0$, and equals zero if and only if $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$

Consequence: $KL(.\,\|\,.)$ can be used as a measure of discrepancy between two probability measures.

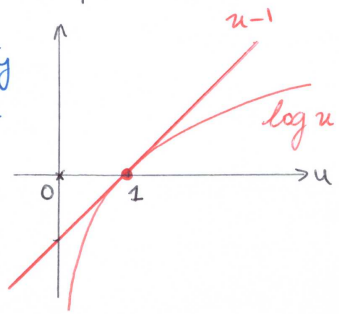
Positivity of the KL divergence follows from GIBB'S INEQUALITY: for any two densities f_θ and $f_{\theta'}$,

$$\int f_\theta(x)\log\{f_\theta(x)\}dx \geq \int f_\theta(x)\log\{f_{\theta'}(x)\}dx,$$

with equality iff $f_\theta(x) = f_{\theta'}(x)$ almost everywhere.

The inequality holds in the discrete case as well,

$$\sum_x p_\theta(x) \log p_\theta(x) \geq \sum_x p_\theta(x) \log p_{\theta'}(x).$$



Proof: we prove Gibb's inequality in the discrete case, the AC case being treated similarly.

Note that $\forall x>0$, $\log x \leq x-1$, with equality for $x=1$.

Thus

$$\sum_x p_\theta(x) \log \frac{p_{\theta'}(x)}{p_\theta(x)} \leq \sum_x p_\theta(x) \left\{ \frac{p_{\theta'}(x)}{p_\theta(x)} - 1 \right\}$$

$$= \sum_x p_{\theta'}(x) - \sum_x p_\theta(x) = 0.$$

Equality occurs at the value $p_{\theta'}(x)/p_\theta(x) = 1$, that is when the two distributions coincide.

Note: Alternatively, Gibb's inequality can be proved using Jensen's inequality. ∎

x Example: KL divergence between two Gaussian densities with common variance: $f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$

$$KL(\mathbb{P}_{\mu,\sigma^2} \| \mathbb{P}_{\mu',\sigma^2}) = \int f_{\mu,\sigma^2}(x) \log \left\{ \frac{f_{\mu,\sigma^2}(x)}{f_{\mu',\sigma^2}(x)} \right\} dx$$

$$= \frac{1}{2\sigma^2} \int \left\{ (x-\mu')^2 - (x-\mu)^2 \right\} f_{\mu,\sigma^2}(x)\, dx$$

$$= \frac{1}{2\sigma^2} \left\{ \mathbb{E}_{\mu,\sigma^2}(X-\mu')^2 - \mathbb{E}_{\mu,\sigma^2}(X-\mu)^2 \right\}$$

$$= \frac{1}{2\sigma^2} \mathbb{E}_{\mu,\sigma^2}\left(-2X(\mu'-\mu) + (\mu')^2 - \mu^2 \right)$$

$$= \frac{1}{2\sigma^2} \left\{ -2\left(\mathbb{E}_{\mu,\sigma^2} X\right)(\mu'-\mu) + (\mu')^2 - \mu^2 \right\}$$

($\mathbb{E}_{\mu,\sigma^2} X = \mu$)

$$= \frac{(\mu-\mu')^2}{2\sigma^2}$$ ∎

x Example: Similarly, we compute the KL divergence between two multivariate Gaussian distributions $f_{\mu,\Sigma}(\underline{x})$ and $f_{\mu',\Sigma'}(\underline{x})$ as:

$$KL(\mathbb{P}_{\mu,\Sigma} \| \mathbb{P}_{\mu,\Sigma'}) = \int f_{\mu,\Sigma}(\underline{x}) \log \left\{ \frac{f_{\mu,\Sigma}(\underline{x})}{f_{\mu',\Sigma'}(\underline{x})} \right\} d\underline{x}$$

($\underline{x} \in \mathbb{R}^d$)

$$= \frac{1}{2} \log\left\{\frac{|\Sigma'|}{|\Sigma|}\right\} + \mathbb{E}_{\mu,\Sigma}\left\{ -\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu) + \frac{1}{2}(X-\mu')^t (\Sigma')^{-1}(X-\mu') \right\}$$

$|\Sigma|$ = determinant of Σ.

$$= \frac{1}{2}\log\left\{\frac{|\Sigma'|}{|\Sigma|}\right\} - \frac{1}{2}\mathbb{E}_{\mu,\Sigma}\left\{ \mathrm{Tr}\left(\Sigma^{-1}(X-\mu)(X-\mu)^t\right)\right\} + \frac{1}{2}\mathbb{E}_{\mu,\Sigma}\left\{ (X-\mu')^t(\Sigma')^{-1}(X-\mu')\right\}$$

Use: $\mathbb{E}(X^t A X) = \mu^t A \mu + \mathrm{Tr}(A\Sigma)$, for X with mean μ & covariance matrix Σ.

$$= \frac{1}{2}\log\left\{\frac{|\Sigma'|}{|\Sigma|}\right\} - \frac{1}{2}\mathrm{Tr}\left\{\Sigma^{-1}\mathbb{E}_{\mu,\Sigma}(X-\mu)(X-\mu)^t\right\} + \frac{1}{2}\left[(\mu-\mu')^t(\Sigma')^{-1}(\mu-\mu') + \mathrm{Tr}\left((\Sigma')^{-1}\Sigma\right)\right]$$

Use: $\mathbb{E}_{\mu,\Sigma}(XX^t - 2X\mu^t + \mu\mu^t)$
$= \Sigma + \mu\mu^t - 2\mu\mu^t + \mu\mu^t = \Sigma$.

$$= \frac{1}{2}\log\left\{\frac{|\Sigma'|}{|\Sigma|}\right\} - \frac{1}{2}\underbrace{\mathrm{Tr}\left(\Sigma^{-1}\Sigma\right)}_{=d} + \frac{1}{2}\left[(\mu-\mu')^t(\Sigma')^{-1}(\mu-\mu') + \mathrm{Tr}\left\{(\Sigma')^{-1}\Sigma\right\}\right].$$

$$\Rightarrow \boxed{KL(\mathbb{P}_{\mu,\Sigma} \,\|\, \mathbb{P}_{\mu',\Sigma'}) = \frac{1}{2}\left(\log\left\{\frac{|\Sigma'|}{|\Sigma|}\right\} - d + \mathrm{Tr}\left\{(\Sigma')^{-1}\Sigma\right\} + (\mu-\mu')^t(\Sigma')^{-1}(\mu-\mu')\right).}$$

I.2 The Maximum likelihood principle.

- Recall: true model is $\mathbb{P}_{\theta^*}$ for some $\theta^* \in \Theta$. Consider $X \sim \mathbb{P}_{\theta^*}$.

- Idea: find the value of θ which minimizes the KL divergence between $\mathbb{P}_{\theta^*}$ and $\mathbb{P}_\theta$:

$$KL(\mathbb{P}_{\theta^*} \,\|\, \mathbb{P}_\theta) = \underbrace{\mathbb{E}_{\theta^*}\left\{\log f_{\theta^*}(X)\right\}}_{\text{Assuming that } X \text{ is AC with density } f_{\theta^*} \text{ (the discrete case is treated similarly).}} - \underbrace{\mathbb{E}_{\theta^*}\left\{\log f_\theta(X)\right\}}_{\text{Minimization is performed with respect to } \theta. \text{ The first term } \mathbb{E}_{\theta^*}\{\log f_{\theta^*}(X)\} \text{ is independent of } \theta.}$$

$\Rightarrow$ Minimization of the KL divergence is equivalent to the...

$$\boxed{\text{Maximization of } \mathbb{E}_{\theta^*}\left\{\log f_\theta(X)\right\}.}$$

in practice, the expectation $\mathbb{E}_{\theta^*}\{\dots\}$ is estimated from a random sample $X_1, \dots, X_n$, where $X_i \sim \mathbb{P}_{\theta^*}$, iid.

The goal is now to maximize $\frac{1}{n}\sum_{i=1}^n \log f_\theta(X_i)$

$\frac{1}{n}\log\left(\prod_{i=1}^n f_\theta(X_i)\right)$

Empirical mean; converges a.s. to $\mathbb{E}_{\theta^*}\{\log f_\theta(X_i)\}$, SLLN.

$\Leftrightarrow \boxed{\max_\theta \log \prod_{i=1}^n f_\theta(X_i)}$

The Maximum likelihood principle.

Note that $\prod_{i=1}^n f_\theta(X_i) = f_\theta(\underline{X}) =$ joint density of $X_1, \dots, X_n$.

The function $\ell(\underline{X}; \theta) = \log f_\theta(\underline{X})$ is known as the log-likelihood function.

In the discrete case, replace f_θ by the probability masses.

Summary: The maximum likelihood principle returns the value of θ maximizing $\begin{cases} \log \prod_{i=1}^n f_\theta(x_i) & \text{in the AC case} \\ \log \prod_{i=1}^n p_\theta(x_i) & \text{in the discrete case,} \end{cases}$

having observed $X_1 = x_1, \dots, X_n = x_n$.

We write $\hat{\theta}_{ML} \in \underset{\theta\in\Theta}{\mathrm{argmax}}\ \ell(\underline{X}; \theta)$,

where

$$\ell(\underline{X}; \theta) = \begin{cases} \sum_{i=1}^n f_\theta(X_i) & \text{in the AC case} \\ \sum_{i=1}^n p_\theta(X_i) & \text{in the discrete case.} \end{cases}$$

$\Rightarrow \hat{\theta}_{ML}$ is a Random Variable. We need to investigate its properties.

Remark: Denoting $\hat{\mathbb{P}}_n$ the empirical distribution placing a mass $1/n$ on each observation $X_1, \dots, X_n$, we see that the maximum likelihood estimator $\hat{\theta}_{ML}$ corresponds to the value of $\theta \in \Theta$ minimizing the KL divergence between $\hat{\mathbb{P}}_n$ and $\mathbb{P}_\theta$:

$$\forall \theta \in \Theta, \quad KL\left(\hat{\mathbb{P}}_n \,\|\, \mathbb{P}_{\hat{\theta}_{ML}}\right) \leq KL\left(\hat{\mathbb{P}}_n \,\|\, \mathbb{P}_\theta\right).$$

since we replaced $\mathbb{P}_{\theta^*}$ (true distribution) with $\hat{\mathbb{P}}_n$ on the top of page 6, when estimating $\mathbb{E}_{\theta^*}\{\dots\}$ using the empirical mean.

I.3. Alternatives to KL divergence.

- The KL divergence is a tool to quantify how far away two probability measures are from each other. However, there are many alternatives, and we introduce a selected few in this section.

- One potential drawback to the KL divergence is its assymetry. Therefore, one often work with the **HELLINGER METRIC**

$$H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \begin{cases} \left[\int \left(\sqrt{f_\theta(x)} - \sqrt{f_{\theta'}(x)}\right)^2 dx\right]^{1/2} & \text{if AC cont.} \\ \left[\sum_x \left(\sqrt{p_\theta(x)} - \sqrt{p_{\theta'}(x)}\right)^2\right]^{1/2} & \text{if discrete.} \end{cases}$$

It clearly satisfies $H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = H(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$, and in fact satisfies all the properties of a metric.

Moreover, $\boxed{H(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq \sqrt{KL(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'})}}$

↳ an so a small KL divergence between two distributions implies that the two distributions are close as measured by the Hellinger metric.

proof $= H^2(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \int \left(\sqrt{f_\theta(x)} - \sqrt{f_{\theta'}(x)}\right)^2 dx$

$$= 2 - 2\int \sqrt{f_\theta(x) f_{\theta'}(x)}\, dx$$

$$(*) \quad = 2\left(1 - \int \sqrt{\frac{f_{\theta'}(x)}{f_\theta(x)}}\, f_\theta(x)\, dx\right)$$

$$= 2\left(1 - \mathbb{E}_\theta\left\{\sqrt{\frac{f_{\theta'}(x)}{f_\theta(x)}}\right\}\right)$$

$1 - x \leq -\log x$ ↳

$$\leq -2 \log\left(\mathbb{E}_\theta\left\{\sqrt{\frac{f_{\theta'}(x)}{f_\theta(x)}}\right\}\right)$$

Jensen: $-\log \mathbb{E}X \leq -\mathbb{E}\log X$ ↳

$$\leq -2\, \mathbb{E}_\theta\left\{\log \sqrt{\frac{f_{\theta'}(x)}{f_\theta(x)}}\right\}$$

$$= \mathbb{E}_\theta\left\{\log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right)\right\}$$

$$= KL(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'}).$$

We can repeat the previous steps, replacing equality $(*)$ with $2\left(1 - \int \sqrt{\frac{f_\theta(x)}{f_{\theta'}(x)}}\, f_{\theta'}(x)\, dx\right)$, so that we obtain as well that $H^2(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq KL(\mathbb{P}_{\theta'} \,\|\, \mathbb{P}_\theta)$. ■

x Example: Consider again two univariate Gaussian densities with different means, and common variance.

$$H(\mathbb{P}_{\mu,\sigma^2}, \mathbb{P}_{\mu',\sigma^2}) = 2\left(1 - \int \sqrt{f_{\mu,\sigma^2}(x)\, f_{\mu',\sigma^2}(x)}\, dx\right)$$

Again, using $1-x \leq -\log x$, $x>0$ ↳

$$\leq -2 \log \left\{ \int \sqrt{f_{\mu,\sigma^2}(x) f_{\mu',\sigma^2}(x)}\, dx \right\}$$

↑ This quantity inside the log is known as the AFFINITY between $\mathbb{P}_{\mu,\sigma^2}$ and $\mathbb{P}_{\mu',\sigma^2}$.

$$= -2 \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(-\frac{(x-\mu)^2}{4\sigma^2} - \frac{(x-\mu')^2}{4\sigma^2} \right) dx \right\}$$

$$= -2 \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(-\frac{(\mu-\mu')^2}{8\sigma^2} \right) \exp\left(-\frac{\left(x - \frac{\mu+\mu'}{2}\right)^2}{2\sigma^2} \right) dx \right\}$$

$$= -2 \log \exp\left\{ -\frac{(\mu-\mu')^2}{8\sigma^2} \right\}$$

$$= \frac{(\mu-\mu')^2}{4\sigma^2}$$

■

- The TOTAL VARIATION DISTANCE is also a common choice. It is defined as

$$TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \sup_{A \in \mathcal{F}} |\mathbb{P}_\theta(A) - \mathbb{P}_{\theta'}(A)|.$$

↖ One can show that it is a distance / metric.

In the discrete case, the TV distance between $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is a simple function of the probability mass functions $p_\theta(x)$ and $p_\theta(x')$:

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

And in the AC case, one can show that

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

Moreover, it is known that $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq \sqrt{\frac{1}{2} KL(\mathbb{P}_\theta \| \mathbb{P}_{\theta'})}$.

- Finally, note that the KL divergence can be made symmetrical, by considering for instance

$$KL(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}) + KL(\mathbb{P}_{\theta'} \| \mathbb{P}_\theta),$$

or

$$\min \left\{ KL(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}),\ KL(\mathbb{P}_{\theta'} \| \mathbb{P}_\theta) \right\}.$$

Alternatively, the JENSEN-SHANNON DIVERGENCE (JS) is sometimes used, and is a symmetrized & smoothed version of the KL divergence. It is defined as

$$JS(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}) = \frac{1}{2} KL(\mathbb{P}_\theta \| \mathbb{Q}_{\theta,\theta'}) + \frac{1}{2} KL(\mathbb{P}_{\theta'} \| \mathbb{Q}_{\theta,\theta'}),$$

where

$$\mathbb{Q}_{\theta,\theta'} := \frac{1}{2}(\mathbb{P}_\theta + \mathbb{P}_{\theta'}).$$

- Other alternatives: Levy-Prokhorov metric (metrizes weak convergence), Wasserstein-Kantorovich metric, ...

II. PROPERTIES OF THE MLE

II.1. Introductive examples.

x Example: Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ iid. Put $\theta = (\mu, \sigma^2)$

The log-likelihood is

$$\ell(\underline{X}; \theta) = \log \left\{ \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \right\} \right\}$$

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

Setting the partial derivatives to zero yields:

$$\begin{cases} \dfrac{\partial \ell}{\partial \mu} = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i - \mu) = \frac{n}{\sigma^2}(\bar{X} - \mu) = 0 \\ \dfrac{\partial \ell}{\partial \sigma} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases}$$

The first equation gives $\hat{\mu}_{ML} = \bar{X}$, that we can substitute into the second equation, to obtain

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Observations:

- $\hat{\mu}_{ML} = \bar{X} \xrightarrow{a.s} \mu$ as $n \to +\infty$ (SLLN) consistency. Moreover, $\hat{\mu}_{ML}$ is unbiased for μ.
- $\hat{\sigma}^2_{ML} \xrightarrow{a.s} \sigma^2$ as $n \to +\infty$ (consistency). However, $\hat{\sigma}^2_{ML}$ is a biased estimator of σ^2 (we need to renormalize by $(n-1)$ to obtain an unbiased estimator). ■

x Example: $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$ iid. The log-lik is

$$\ell(\underline{X}, \lambda) = \log\left\{ \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right\}$$

$$= -n\lambda - \sum_{i=1}^{n}\log(x_i!) + \log\lambda \sum_{i=1}^{n} x_i.$$

$$\frac{\partial \ell(\underline{X}; \lambda)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \hat{\lambda}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Observations: $\hat{\lambda}_{ML}$ is unbiased for λ, consistent, and efficient (it has the smallest possible variance amongst the class of unbiased estimators) (see page 21/22 in MS = PARAMETRIC INFERENCE). ■

II.2. General properties of the MLE.

The Maximum likelihood principle returns the value $\hat{\theta}_{ML}$ of $\theta \in \Theta$ maximizing $\log \prod_{i=1}^{n} f_\theta(X_i) = \sum_{i=1}^{n} \ell(X_i; \theta)$ (see page 6)

As $n \to +\infty$, expect the MLE $\hat{\theta}_{ML}$ to be close to the value of θ maximizing $\mathbb{E}_{\theta^*}\{\ell(X_1; \theta)\} = \mathbb{E}_{\theta^*}\{\log f_\theta(X_1)\}$, since $\frac{1}{n}\sum_{i=1}^{n} \ell(X_i; \theta) \xrightarrow{a.s.} \mathbb{E}_{\theta^*}\{\ell(X_1; \theta)\}$ as $n \to +\infty$.

Moreover, Gibb's inequality informs us that the value of θ maximizing $\mathbb{E}_{\theta^*}\{\log f_\theta(X_1)\} = \int f_{\theta^*}(x) \log f_\theta(x)\, dx$ is precisely θ^*; so that when n is large, one expects $\hat{\theta}_{ML}$ to be close (in some sense) to the true parameter θ^*.

× **Example**: Let $X_1, \dots, X_n \sim \mathcal{P}(\lambda^*)$ iid

Then $\ell(X_1; \theta) = -\theta + X_1 \log\theta - \log(X_1!)$,

so that

$$\mathbb{E}_{\lambda^*}\{\ell(X_1;\theta)\} = \mathbb{E}_{\lambda^*}\{-\theta + X_1\log\theta - \log(X_1!)\}$$
$$= -\theta + \lambda^*\log\theta - \mathbb{E}_{\lambda^*}\{\log(X_1!)\}.$$

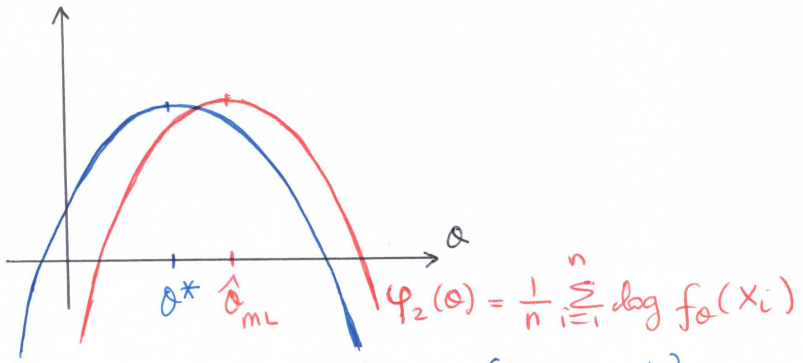
The value of θ maximizing $\theta \mapsto \mathbb{E}_{\lambda^*}\{\ell(X_1;\theta)\}$ satisfies

$$\frac{\partial\, \mathbb{E}_{\lambda^*}\{\ell(X_1;\theta)\}}{\partial\theta} = -1 + \frac{\lambda^*}{\theta} = 0$$

$\Rightarrow$ The maximum occurs at $\theta = \lambda^*$ = the true parameter ■

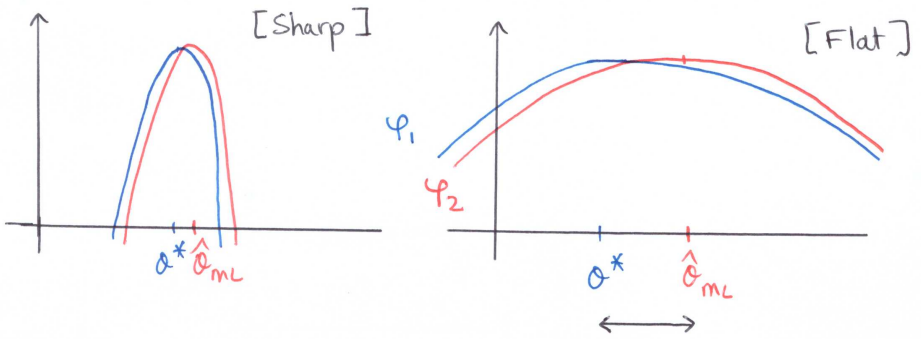
• **What we know so far**: expect $\hat\theta_{ML}$ to be close to the true parameter θ^* = maximum point of the function $\varphi_1(\theta) = \mathbb{E}_{\theta^*}\ell(X_1,\theta)$

↳ Plot on a same graph $\varphi_1(\theta)$ and $\varphi_2(\theta) = \frac{1}{n}\ell(\underline{X};\theta)$, where $\ell(\underline{X};\theta) = \log\prod_{i=1}^n f_\theta(X_i)$.



• **Observation**: the sharper the peak of $\varphi_1(\theta)$, the closer should be the maxima of φ_1 and φ_2.

Compare:



↳ In both cases, φ_1 and φ_2 are close together $\forall\theta\in\Theta$. However, because of the flat peak of φ_1 in the second case, the maxima are far from each other.

$\Rightarrow$ Estimation of θ^* is related to the curvature of $\varphi_1(\theta)$ at its maximum.

The curvature is defined in terms of the second order derivative, but is not equal to it.

Ex: the function $x \mapsto x^2$ has a constant second order derivative equal to 2, but its curvature is not constant: the graph of $x \mapsto x^2$ is "flatter" as we move away from the origin.

However, at the point where the derivative vanishes (i.e. at a local maximum or minimum for example), the curvature is given by $|f''(.)|$ at that point.

Properties of the MLE $\hat\theta_{ML}$ should be related to the magnitude of $\varphi_1''(\theta)$ evaluated at $\theta = \theta^*$. ⇐

Provided we can exchange $\mathbb{E}_{\theta^*}\{\dots\}$ and derivatives (in other words, under technical conditions), the second order partial derivative of φ_1 is given by

$$\varphi_1''(\theta) = \mathbb{E}_{\theta^*}\{\ell''(X;\theta)\} = -I_1(\theta)$$

↑ Fisher information for one observation (see p.20 in chapter MS: PARAMETRIC ESTIMATION)

⇒ Evaluated at $\theta=\theta^*$ = the true parameter, our intuition tells us that the higher $I_1(\theta^*)$, the better the properties of $\hat{\theta}_{ML}$ (hence the name Fisher "information") (since the higher the curvature at the maximum). And indeed, we have the following result:

Theorem: Let. $X_1,\dots,X_n$ iid $\sim \mathbb{P}_{\theta^*}$ for some $\theta^*\in\Theta$.

- $\mathcal{P} = \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ a family of suspects
- $\forall\theta\in\Theta$, the support of $\mathbb{P}_\theta$ does not depend on θ, and θ^* is not on the boundary of Θ
- $I_1(\theta) = -\mathbb{E}_{\theta^*}\{\ell''(X;\theta)\}$ is invertible in a neighborhood of θ^*, where

$$\ell(X;\theta) = \begin{cases} \log f_\theta(X) & \text{in the AC case} \\ \log p_\theta(X) & \text{in the discrete case} \end{cases}$$

- Some additional technical conditions.

Then $\hat{\theta}_{ML}$ satisfies:

(i) $\hat{\theta}_{ML} \xrightarrow{\mathbb{P}} \theta^*$ as $n\to+\infty$, with respect to $\mathbb{P}_{\theta^*}$. (consistency).

(ii) The MLE is asymptotically normal:

$$n^{1/2}(\hat{\theta}_{ML}-\theta^*) \xrightarrow{d} \mathcal{N}(0,\ I_1^{-1}(\theta^*))$$

(iii) Convergence of moments:

↘ $\mathbb{E}_{\theta^*}\hat{\theta}_{ML} = \theta^* + o(n^{-1/2})$ (asymp. unbiased)

↘ $\mathbb{E}_{\theta^*}(\hat{\theta}_{ML}-\theta^*)^2 = \dfrac{1+o(1)}{n\,I_1(\theta^*)}$

The higher $I_1(\theta^*)$, the smaller the MSE. The theorem agrees with our intuition. Moreover, recall the Cramer-Rao lower bound for unbiased estimators: $1/I_n(\theta)$ (page 21 in MS: PARAMETRIC ESTIMATION)

⇒ $\hat{\theta}_{ML}$ is asymptotically efficient.

proof: We admit consistency. We outline the proof for the asymptotic normal distribution, omitting the technical details.

- Taylor expanding $\ell'(X;\theta^*)$ around the MLE yields:

$$\ell'(X;\theta^*) = \underbrace{\ell'(X;\hat{\theta}_{ML})}_{=0} + (\theta^*-\hat{\theta}_{ML})\,\ell''(X;\tilde{\theta}),$$

where $\tilde{\theta}$ is a point between θ^* and $\hat{\theta}_{ML}$. Thus:

$$\sum_{i=1}^n \ell'(X_i;\theta^*) = -(\hat{\theta}_{ML}-\theta^*)\sum_{i=1}^n \ell''(X_i;\tilde{\theta})$$

$\times n^{-1/2}$ ↓

$$\underbrace{n^{-1/2}\sum_{i=1}^n \ell'(X_i;\theta^*)}_{\text{(I)}} = \underbrace{n^{1/2}(\hat{\theta}_{ML}-\theta^*)}_{\text{quantity of interest}}(-1)\ \underbrace{n^{-1}\sum_{i=1}^n \ell''(X_i;\tilde{\theta})}_{\text{(II)}}.$$

↘ (I) $= n^{1/2} \sum_{i=1}^{n} \ell'(X_i ; \theta^*) \xrightarrow{d} \mathcal{N}(0, I_1(\theta^*))$,

since $\mathbb{E}_{\theta^*}\{\ell'(X_1 ; \theta^*)\} = 0$

& $\mathbb{E}_{\theta^*}\{(\ell'(X_1 ; \theta^*))^2\} = I_1(\theta^*)$

+ Central limit Theorem.

↘ (II) Consistency of the MLE : $\hat{\theta}_{ML} \xrightarrow{P} \theta^*$.

Thus, $\tilde{\theta} \xrightarrow{P} \theta^*$, as $n \to +\infty$.

The WLLN + Slutsky theorem given

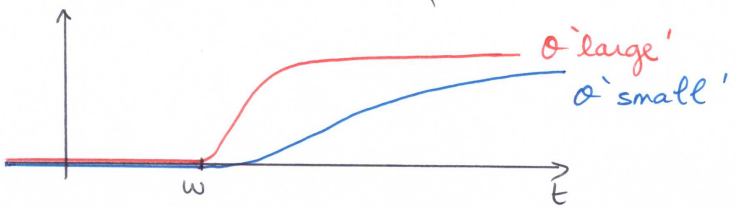
$$n^{-1} \sum_{i=1}^{n} \ell''(X_i ; \tilde{\theta}) \xrightarrow{P} \mathbb{E}_{\theta^*}\{\ell''(X_1 ; \theta^*)\} = -I_1(\theta^*)$$

Putting pieces together, we get $n^{1/2}(\hat{\theta}_{ML} - \theta^*) \xrightarrow{d} \mathcal{N}(0, I_1^{-1}(\theta^*))$ as required. ∎

II.3. Illustration.

The time T of occurrence of cancer in a tissue, assumed random, can be modelled as the time of occurrence of cancer in one individual cell composing the tissue. Given that a tissue contains a large number of cells, asymptotic theory of extreme values indicate that possible distributions for T can be of three different types (Gumbel, '58, Statistics of Extremes). Following the work of (Pike, '66, A Method of Analysis of a Certain Class of Experiments in Carcinogenesis, Biometrics), we consider that the time of

appearance of cancer, in days, is modelled as $F_\theta(t) = \mathbb{P}(T \leq t) = 1 - e^{-\theta(t-w)^k}$ for $t \geq w$, ——— (*) with $k > 0$, $\theta > 0$. The parameter θ has a huge effect on the distribution of T. For k fixed, we have



A large value of θ indicates that T is more likely to be small. The goal is to quantify the effect of a treatment on the time of appearance of cancer; i.e. on the value of θ.

Pike considers a population of mice, on which we are testing the effect of a pre-treatment on a toxic product. The time t (after the start of the experiment) at which the first effects of the toxic product were observed is recorded for each mouse. In this experiment, we assume k and w known, where

$w \approx$ time it takes for the toxic product to be active

$k \approx$ death rate of the mice, function of their age.

Observations $t_1, \dots, t_n$ (n = number of mice) are assumed to be n independent observations of $T_1, \dots, T_n$, each following the distribution (*), with density

$$f_\theta(t) = k\theta(t-w)^{k-1} e^{-\theta(t-w)^k} \mathbb{1}(t > w)$$

The joint distribution of $T_1, \dots, T_n$ is

$$f_\theta(\underline{t}) = \theta^n k^n \left(\prod_{i=1}^n (t_i - \omega)^{k-1} \mathbb{1}(t_i > \omega) \right) e^{-\theta \sum_{i=1}^n (t_i-\omega)^k},$$

and the log-likelihood is

$$\ell(\underline{T}\,; \theta) = n \log \theta - \theta \sum_{i=1}^n (T_i - \omega)^k + \text{something independent of } \theta.$$

$$\frac{\partial \ell(\underline{T}; \theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^n (T_i - \omega)^k,$$

so that the ML estimator of θ is $\boxed{\hat{\theta}_{ML} = \dfrac{n}{\sum_{i=1}^n (T_i - \omega)^k}}$

Fisher information associated with one observation is

$$I_1(\theta) = -\mathbb{E}_\theta\{\ell''(T; \theta)\} = \frac{1}{\theta^2}.$$

$\Rightarrow \hat{\theta}_{ML}$ is asymptotically normally distributed, and

$$n^{1/2}(\hat{\theta}_{ML} - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2) \qquad (**)$$

↑ We can use this result to construct an asymptotic confidence interval for θ: replace the asymptotic variance θ^2 by the consistent estimator $\hat{\theta}_{ML}^2$. Slutsky theorem and (**) ensure that

$$\frac{n^{1/2}(\hat{\theta}_{ML} - \theta)}{\hat{\theta}_{ML}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to +\infty$$

↑ A pivotal statistic.

Select $z_{1-\alpha/2} = \left(1 - \frac{\alpha}{2}\right)$ quantile of the standard normal distribution, so that

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leqslant \frac{n^{1/2}(\hat{\theta}_{ML} - \theta)}{\hat{\theta}_{ML}} \leqslant z_{1-\frac{\alpha}{2}} \right) \approx 1 - \alpha,$$

and we deduce that $\boxed{\left[\hat{\theta}_{ML} - z_{1-\frac{\alpha}{2}} n^{-1/2} \hat{\theta}_{ML},\ \hat{\theta}_{ML} + z_{1-\frac{\alpha}{2}} n^{-1/2} \hat{\theta}_{ML} \right]}$ is an asymptotic CI for θ, with nominal level $(1-\alpha)$

- Numerical application: $\omega = 100$, $k = 3$, $n = 17$

$$\sum_{i=1}^n (t_i - 100)^3 = 33\,175\,533$$

Observations =

143	164	188	188	190	192	206	209	213	216
220	227	230	234	246	265	304			

We obtain $\hat{\theta}_{ML} = 5.124 \cdot 10^{-7}$

$$CI = \left[2.7 \cdot 10^{-7},\ 7.6 \cdot 10^{-7} \right]$$

It turns out that some mice die before the end of the experiment, for reasons that are independent of the injection of the toxic product (such as disease, wounds, ...) These mice are removed from the experiment, and the time s of removal is recorded. For a given mouse, we do not observe the time of appearance t of the first effects directly, but $r = \min(s, t)$. The value s corresponds to the observation of a random variable S with distribution $G_s(s) = \mathbb{P}(S \leqslant s) = \int_{-\infty}^{s} g_s(u)\,du$. The functions G_s and g_s are unknown, and we are not trying

to estimate them. We observe $X := \mathbb{1}(T \leq S)$ as well, where $X = 0$ if the mouse is removed from the experiment before the first effects of the toxic product appear, and $X = 1$ otherwise. We write $p = \mathbb{P}(X = 1)$, so that $X \sim B(p)$.

Random sample collected is $(R_1, X_1), \dots, (R_n, X_n)$.
⇒ We need to derive the joint distribution of (R, X).
We have

- $\mathbb{P}(R > r, X = 1) = \mathbb{P}(S \geq T > r)$

$$= \int \mathbb{1}(s \geq t > r)\, g_S(s)\, f_\theta(t)\, ds\, dt$$

$$= \int \mathbb{1}(t > r)\, f_\theta(t) \times \left[\int \mathbb{1}(s \geq t)\, g_S(s)\, ds \right] dt$$

$$= \int_r^{+\infty} f_\theta(t)\,(1 - G_S(t))\, dt .$$

- $\mathbb{P}(R > r, X = 0) = \mathbb{P}(T > S > r)$

$$= \int \mathbb{1}(t > s > r)\, g_S(s)\, f_\theta(t)\, ds\, dt$$

$$= \int \mathbb{1}(s > r)\, g_S(s) \times \left[\int \mathbb{1}(t > s)\, f_\theta(t)\, dt \right] ds$$

$$= \int_r^{+\infty} g_S(s)\,(1 - F_\theta(t))\, dt$$

Differentiating with respect to r, we obtain

$$f(r, x=1) = k\theta(r-\omega)^{k-1} e^{-\theta(r-\omega)^k} \mathbb{1}(r > \omega)\,(1 - G_S(r))$$

$$= \theta\, e^{-\theta(r-\omega)_+^k} \underbrace{k\,(r-\omega)_+^{k-1}\,(1 - G_S(r))}_{\text{independent of } \theta =: c(r,1)}$$

where $(u)_+ = \max(u, 0)$

$$= \theta\, c(r,1) \exp\left(-\theta (r-\omega)_+^k\right)$$

Likewise,

$$f(r, x=0) = g(r)\, e^{-\theta(r-\omega)_+^k} \quad \text{(} c(r,0) := g(r) \text{)}$$

$$= c(r,0) \exp\left(-\theta(r-\omega)_+^k\right)$$

Summarizing, $\boxed{f(r, x) = c(r, x)\, \theta^x\, e^{-\theta(r-\omega)_+^k}}$

The joint density of the random sample $(R_1, X_1) \dots (R_n, X_n)$ is

$$f(\underline{r}, \underline{x}; \theta) = \prod_{i=1}^n f(r_i, x_i)$$

$$= \theta^{\sum x_i}\, e^{-\theta \sum (r_i - \omega)_+^k} \prod_{i=1}^n c(r_i, x_i),$$

and the log-likelihood is

$$\tilde{\ell}(\underline{r}, \underline{x}; \theta) = (\log\theta) \sum x_i - \theta \sum (r_i - \omega)_+^k + \text{something independent of } \theta.$$

$$\frac{\partial \tilde{\ell}(\underline{r}, \underline{x}; \theta)}{\partial \theta} = \frac{1}{\theta} \sum x_i - \sum (r_i - \omega)_+^k$$

$$\Rightarrow \boxed{\tilde{\theta}_{ML} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n (R_i - \omega)_+^k}}$$

Fisher information, associated with one observation, is

$$\tilde{I}_1(\theta) = -\mathbb{E}_\theta\{\tilde{\ell}''(R, X; \theta)\} = \frac{\mathbb{E}_\theta X_1}{\theta^2} = \frac{p}{\theta^2}.$$

⇒ Asymptotic properties of $\tilde{\theta}_{ML}$:

$$n^{1/2}(\tilde{\theta}_{ML} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\theta^2}{p}\right)$$

As before, to construct a CI for θ, based on its limiting distribution, we need a consistent estimator of the asymptotic variance.

Put $N_n := \sum_{i=1}^{n} X_i$ = number of mice on which toxic effects were observed.

Then $\frac{N_n}{n} \xrightarrow{a.s.} p$ & $\tilde{\theta}_{ML} \xrightarrow{P} \theta$ $\Bigg\}$ $\frac{N_n}{n} \frac{1}{\tilde{\theta}_{ML}^2} \xrightarrow{P} \tilde{I}_1(\theta)$ i.e. consistent.

Using Slutsky, we conclude that

$$\frac{n^{1/2}(\tilde{\theta}_{ML} - \theta)}{\tilde{\theta}_{ML}/\sqrt{N_n/n}} = \frac{N_n^{1/2}(\tilde{\theta}_{ML} - \theta)}{\tilde{\theta}_{ML}} \xrightarrow{d} \mathcal{N}(0,1).$$

⇒ $$\left[\tilde{\theta}_{ML} - z_{1-\frac{\alpha}{2}} N_n^{-1/2} \tilde{\theta}_{ML},\ \tilde{\theta}_{ML} + z_{1-\frac{\alpha}{2}} N_n^{-1/2} \tilde{\theta}_{ML}\right]$$

is an asymptotic CI for θ, with nominal level $1-\alpha$.

- Numerical Application: Use data page 20, plus two extra observations 216, 244, corresponding to the time these two mice were removed from the experiment. We obtain $\tilde{\theta}_{ML} = 4.5 \cdot 10^{-7}$, and $[2.4 \cdot 10^{-7},\ 6.6 \cdot 10^{-7}]$.