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.

1. Maximum likelihood estimation (MLE): a generic approach with very good properties
2. Method of moments: a (fairly) generic and easy approach that extends the setup where $\theta = \mathbb{E}[X]$
3. M-estimators: a generalization of MLE, flexible, and close to machine learning

Distance measures

probability distributions

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

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

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 $A \subset E$.

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

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

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

An estimation strategy

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

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

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

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

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

Properties of KL-divergence

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

Not a distance.

This is is called a *divergence*.

Asymmetry is the key to our ability to estimate it!

Maximum likelihood estimation

Estimating the KL

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

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

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

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

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

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

Maximum likelihood

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

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

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:

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

Likelihood for the Bernoulli model

Example 1 (Bernoulli trials): If $X_1, \ldots, X_n \overset{iid}{\sim} \mathsf{Ber}(p)$ 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 $X_1, \ldots, X_n \stackrel{iid}{\sim}$ Poiss(λ) for some $\lambda > 0$:

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

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

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. $X_1, \dots, X_n$. Assume that all the $\mathbb{P}_\theta$ have density f_θ.

Definition

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

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

Likelihood for the Gaussian model

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

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

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

Likelihood for the Uniform model

Example (Uniform model): If $X_1, \ldots, X_n \stackrel{iid}{\sim} \text{Unif}(0, \theta)$, for some $\theta > 0$:

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

$$L(x_1, \ldots, x_n; \theta) = \frac{1}{\theta^n} \mathbb{1}\{x_{(n)} \le \theta\}$$

where $x_{(n)} = \max_i x_i$

Likelihood for the Mixture of two Gaussians model

Example 1 (Mixture of Gaussians model): If $X_1, \ldots, X_n$ are i.i.d from a mixture of two Gaussians, with means $\mu_1, \mu_2 \in \mathbb{R}$, variances, $\sigma_1^2, \sigma_2^2 > 0$ and $\pi \in (0,1)$

- $E = \mathbb{R}$;
- $\Theta = \mathbb{R} \times \mathbb{R} \times (0,1)$
- $\forall (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad \forall (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty),$

$$L(x_1, \ldots, x_n; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi) =$$

$$\frac{1}{(\sqrt{2\pi})^n} \prod_{i=1}^n \left\{ \frac{\pi}{\sigma_1} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right) + \frac{1-\pi}{\sigma_2} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right) \right\}.$$

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:

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

provided it exists.

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

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

Interlude: maximizing/minimizing functions

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:

- $\Theta = \mathbb{R}$, $h(\theta) = -\theta^2$,
- $\Theta = (0, \infty)$, $h(\theta) = \sqrt{\theta}$,
- $\Theta = (0, \infty)$, $h(\theta) = \log\theta$,
- $\Theta = [0, \pi]$, $h(\theta) = \sin(\theta)$
- $\Theta = \mathbb{R}$, $h(\theta) = 2\theta - 3$

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.

Examples of maximum likelihood estimators

- Bernoulli trials: $\hat{p}_n^{\text{MLE}} = \bar{X}_n$.
- Poisson model: $\hat{\lambda}_n^{\text{MLE}} = \bar{X}_n$.
- Gaussian model: $\left(\hat{\mu}_n, \hat{\sigma}_n^2\right)^{\text{MLE}} = \left(\bar{X}_n, \hat{S}_n\right)$.
- Uniform model: $\hat{\theta}^{\text{MLE}} = X_{(n)} = \max_i X_i$
- Mixture of Gaussians: no closed form. Need to use an optimization algorithm, for example EM.

The EM algorithm

To maximize the (log-) likelihood in mixtures of Gaussians, we often use the popular Expectation-Maximization (EM) algorithm.

- It is a *heuristic*. In particular, it can fail to find the MLE.
- Some very recent guarantees have been proved but require structural assumptions and/or good initialization.
- In practice, the algorithm is started from different random initializations and the solution with largest log-likelihood is kept in the end.
- The EM algorithm was introduced in 1977 and is still hugely popular

TITLE	CITED BY	YEAR
Maximum Likelihood from Incomplete Data Via the *EM* Algorithm AP Dempster, NM Laird, DB Rubin Journal of the Royal Statistical Society: Series B (Methodological) 39 (1), 1-22	61636	1977

Likelihood

To illustrate EM, assume that $\pi = 1/2$, and $\sigma_1^2 = \sigma_2^2 = 1$. Recall that the PDF is

$$f(x) = \frac{1}{2\sqrt{2\pi}} \left\{ e^{-\frac{(x-\mu_1)^2}{2}} + e^{-\frac{(x-\mu_2)^2}{2}} \right\}.$$

So log-likelihood is:

$$\ell(x_1, \ldots, x_m; \mu_1, \mu_2) = \sum_{i=1}^{n} \log \left[e^{-\frac{(x_i-\mu_1)^2}{2}} + e^{-\frac{(x_i-\mu_2)^2}{2}}) \right] - n \log(2\sqrt{2\pi})$$

Not easily tractable.

Complete observations

We also have the *sampling* description:

$$X = Z X^{(1)} + (1 - Z) X^{(2)}$$

Z is a *latent* variable with pmf $p(z) = \begin{cases} 1/2 & \text{if } z = 0 \\ 1/2 & \text{if } z = 1 \end{cases}$

What if we observed both (Z, X)? Their joint density is

$$\begin{aligned} f(x, z) &= p(z) \cdot f(x|z) \\ &= \frac{1}{2} \cdot \left(z \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2}} + (1 - z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2}} \right) \\ &= \frac{1}{2\sqrt{2\pi}} e^{-z\frac{(x-\mu_1)^2}{2}} e^{-(1-z)\frac{(x-\mu_1)^2}{2}} \\ &= \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{z(x-\mu_1)^2 + (1-z)(x-\mu_2)^2}{2} \right) \end{aligned}$$

Complete likelihood

The complete likelihood becomes

$$L^{\mathsf{comp}}\left((x_1, z_1), \ldots, (x_n, z_n); \mu_1, \mu_2\right) = \prod_{i=1}^{n} \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{z_i(x_i-\mu_1)^2 + (1-z_i)(x_i-\mu_2)^2}{2}\right)$$

and the corresponding complete log-likelihood is

$$\ell^{\mathsf{comp}}(\mu_1, \mu_2) = -n\log(2\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}[Z_i(X_i-\mu_1)^2 + (1-Z_i)(X_i-\mu_2)^2]$$

Easy to maximize:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} Z_i X_i}{\sum_{i=1}^{n} Z_i}, \qquad \hat{\mu}_2 = \frac{\sum_{i=1}^{n}(1-Z_i)X_i}{\sum_{i=1}^{n}(1-Z_i)}$$

but requires knowledge of the Z_is which we don't have...

The E step

Idea: replace unknown Z_i by its (conditional) Expectation:

- First attempt: $Z_i \approx \mathbb{E}[Z_i] = 1/2$. This is **too rough**!
- Second attempt: $Z_i \approx \mathbb{E}[Z_i|X_i]$. This is much better!

$$\begin{aligned}
\mathbb{E}[Z_i|X_i] &= \mathbb{P}[Z_i = 1|X_i] \\
&= \frac{f(X_i|Z_i = 1)\mathbb{P}[Z_i = 1]}{f(X_i|Z_i = 1)\mathbb{P}[Z_i = 1] + f(X_i|Z_i = 0)\mathbb{P}[Z_i = 0]} \\
&\hphantom{=} \hfill \text{(Bayes formula)} \\
&= \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - \mu_1)^2}{2}} \cdot \frac{1}{2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - \mu_1)^2}{2}} \frac{1}{2} + \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - \mu_2)^2}{2}} \cdot \frac{1}{2}} \\
&= \frac{e^{-\frac{(X_i - \mu_1)^2}{2}}}{e^{-\frac{(X_i - \mu_1)^2}{2}} + e^{-\frac{(X_i - \mu_2)^2}{2}}} =: w_i \in (0, 1)
\end{aligned}$$

Note that w_i depends on μ_1, μ_2.

The M step

If we replace Z_i by $\mathbb{E}[Z_i|X_i] = w_i$ in the complete log-likelihood, we get

$$\tilde{\ell}(\mu_1, \mu_2) = -n \log(2\sqrt{2\pi}) - \frac{1}{2} \sum_{i=1}^{n} [w_i \ (X_i - \mu_1)^2 + (1 - w_i) \ (X_i - \mu_2)^2]$$

Which is easy to maximize. It yields

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}, \qquad \hat{\mu}_2 = \frac{\sum_{i=1}^n (1 - w_i) X_i}{\sum_{i=1}^n (1 - w_i)}$$

The EM algorithm

Input data: $X_1, \ldots, X_n$.

1. Initialize $\hat{\mu}_1, \hat{\mu}_2$ (e.g. independent $\mathcal{N}(0,1)$)

2. Repeat until convergence:
   - Compute weights (E-step):

$$w_i \leftarrow \frac{e^{-\frac{(X_i-\mu_1)^2}{2}}}{e^{-\frac{(X_i-\mu_1)^2}{2}} + e^{-\frac{(X_i-\mu_2)^2}{2}}}, \quad i = 1, \ldots, n$$

   - Update centers (M-step):

$$\hat{\mu}_1 \leftarrow \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}, \qquad \hat{\mu}_2 \leftarrow \frac{\sum_{i=1}^n (1-w_i) X_i}{\sum_{i=1}^n (1-w_i)}$$

Consistency of maximum likelihood estimator

Under mild regularity conditions, we have

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

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

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

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.

Fisher Information

Definition: Fisher information

Define the log-likelihood for one observation as:

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

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

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

Equivalence of the two definitions

We write it in the case of a continuous r.v. with pdf f_θ. It all starts with the $\star$ identity (we will write $\overset{\star}{=}$ when we use it):

$$\int f_\theta(x)dx = 1 \;\Rightarrow\; \frac{d}{d\theta}\int f_\theta(x)dx = \boxed{\int \frac{d}{d\theta} f_\theta(x)dx = 0} \qquad (\star)$$

We now compute $\text{var}\big[\ell'(\theta)\big]$ and $-\mathbb{E}\big[\ell''(\theta)\big]$ and check that they are indeed equal. First we compute derivatives:

$$\ell'(\theta) = \frac{d}{d\theta}\log f_\theta(x) = \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)}\,, \qquad \ell''(\theta) = \frac{\frac{d^2}{d\theta^2} f_\theta(x)}{f_\theta(x)} - \Big(\frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)}\Big)^2.$$

The first identity gives

$$\text{var}[\ell'(\theta)] = \mathbb{E}[(\ell'(\theta))^2] - (\mathbb{E}[\ell'(\theta)])^2 = \int \Big(\frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)}\Big)^2 f_\theta(x)dx - \Big(\int \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x)dx\Big)^2$$

$$= \int \frac{(\frac{d}{d\theta} f_\theta(x))^2}{f_\theta(x)}dx - \Big(\int \frac{d}{d\theta} f_\theta(x)dx\Big)^2 \overset{\star}{=} \int \frac{(\frac{d}{d\theta} f_\theta(x))^2}{f_\theta(x)}dx$$

Moreover, the second identity gives

$$\mathbb{E}[\ell''(\theta)] = \int \frac{\frac{d^2}{d\theta^2} f_\theta(x)}{f_\theta(x)} f_\theta(x)dx - \int \Big(\frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)}\Big)^2 f_\theta(x)dx$$

$$= \frac{d}{d\theta}\int \frac{d}{d\theta} f_\theta(x)dx - \int \frac{\big(\frac{d}{d\theta} f_\theta(x)\big)^2}{f_\theta(x)}dx \overset{\star}{=} -\int \frac{(\frac{d}{d\theta} f_\theta(x))^2}{f_\theta(x)}dx = -\text{var}[\ell'(\theta)]\,.$$

Fisher information of the Bernoulli experiment

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

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

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

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

Asymptotic normality of the MLE

Theorem

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

1. The parameter is identifiable.
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. $I(\theta) \neq 0$ in a neighborhood of θ^*;
5. A few more technical conditions.

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

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

An idea of the proof

We can use a technique resembling what we used for the Δ-method. How? We need to write the MLE as the function of an average. Write $\ell_i(\theta) := \log f_\theta(X_i)$ and by CLT, we have

$$\sqrt{n}\Big\{\frac{1}{n}\sum_{i=1}^{n}\ell_i'(\theta) - \mathbb{E}[\ell'(\theta)]\Big\} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathsf{var}[\ell'(\theta)])$$

Note first that, $\mathbb{E}[\ell'(\theta)] \overset{\star}{=} 0$ and $\mathsf{var}[\ell'(\theta)] = I(\theta)$. Moreover, to make the MLE appear, recall that since it maximizes the log likelihood so that $\displaystyle\sum_{i=1}^{n}\ell_i'(\hat{\theta}^{\mathrm{MLE}}) = 0$.

Therefore, we can write we can write

$$\sqrt{n}\Big\{\frac{1}{n}\sum_{i=1}^{n}(\ell_i'(\hat{\theta}^{\mathrm{MLE}}) - \ell_i'(\theta)\Big\} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, I(\theta))$$

Now we start being more informal. Using a first order Taylor expansion (which is justified because the MLE is consistent), we have that

$$\frac{1}{\hat{\theta}^{\mathrm{MLE}} - \theta}\Big\{\frac{1}{n}\sum_{i=1}^{n}(\ell_i'(\hat{\theta}^{\mathrm{MLE}}) - \ell_i'(\theta)\Big\} \approx \Big\{\frac{1}{n}\sum_{i=1}^{n}\ell_i''(\theta)\Big\} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[\ell_i''(\theta)] = -I(\theta) \quad \text{(LLN)}$$

The two above displays together with Slutsky yield

$$-I(\theta)\sqrt{n}(\hat{\theta}^{\mathrm{MLE}} - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, I(\theta))$$

Dividing both sides by $-I(\theta)$ yields an asymptotic variance of $I(\theta)/I(\theta)^2 = 1/I(\theta)$.

M-estimation

MLE Strategy

Observe $X_1, \dots, X_n \sim \mathbb{P}_{\theta^*}$, i.i.d, θ^* unknown.

1. Ideal loss function: $\theta \mapsto \mathsf{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ minimized at $\theta = \theta^*$
2. Observe that $\mathsf{KL}(\mathbb{P}, \mathbb{P}_\theta) = -\mathbb{E}\log[p_\theta(X)]$ (plus additive constant)
3. Estimate by $-\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\log[p_\theta(X_i)]$ (-log-likelihood)
4. $\hat{\theta} := \operatorname{argmin}\left\{-\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\log[p_\theta(X_i)]\right\}$

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:

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

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

Examples (1)

- If $E = \mathcal{M} = \mathbb{R}$ and $\rho(x, \mu) = (x - \mu)^2$, for all $x \in \mathbb{R}, \mu \in \mathbb{R}$: $\mu^* =$
- If $E = \mathcal{M} = \mathbb{R}^d$ and $\rho(x, \mu) = \|x - \mu\|_2^2$, for all $x \in \mathbb{R}^d, \mu \in \mathbb{R}^d$: $\mu^* =$
- 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 $\mathbb{P}$.

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

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

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

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

The method of moments

Moments

Let X be a random variable with distribution $\mathbb{P}_\theta$ (write $\mathbb{E}_\theta$ for its expectation).

Definition

For $k = 1, 2, \ldots$, the **moment** of order k of X is given by

$$m_k = m_k(\theta) = \mathbb{E}_\theta[X^k]$$

Example 1: $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\begin{aligned} m_1 &= \mathbb{E}[X] = \mu \\ m_2 &= \mathbb{E}[X^2] \\ &= \mathrm{var}[X] + (\mathbb{E}[X])^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

Example 2: $X \sim \mathrm{Ber}(p)$

$$\begin{aligned} m_1 &= \mathbb{E}[X] &= p \\ m_k &= \mathbb{E}[X^k] &= p \end{aligned}$$

Moment generating function

For many distributions[2] $\mathbb{P}$, all the moments of $X \sim \mathbb{P}$ are contained in a *single* function called Moment Generating Function, or simply MGF:

$$M_X(t) = \mathbb{E}[e^{tX}] \qquad , t \in \mathbb{R}\,.$$

The moments are given by successive[3] derivatives of $M_X(\cdot)$ at $t = 0$:

$$\begin{aligned}
M_X^{(1)}(t) &= \mathbb{E}\left[\tfrac{d}{dt}e^{tX}\right] &&= \mathbb{E}\left[Xe^{tX}\right] = \mathbb{E}[X] = m_1 \quad \text{for } t = 0 \\
M_X^{(2)}(t) &= \mathbb{E}\left[\tfrac{d^2}{dt^2}e^{tX}\right] &&= \mathbb{E}\left[X^2e^{tX}\right] = \mathbb{E}[X^2] = m_2 \quad \text{for } t = 0 \\
&\vdots \\
M_X^{(k)}(t) &= \mathbb{E}\left[\tfrac{d^k}{dt^k}e^{tX}\right] &&= \mathbb{E}\left[X^ke^{tX}\right] = \mathbb{E}[X^k] = m_k \quad \text{for } t = 0
\end{aligned}$$

---

[2] It may be infinite for some t. For if X has a Cauchy distribution with pdf given by $f(x) = \frac{1}{\pi(1+x^2)}$

[3] For a function $f(t)$, we write $f^{(k)}(t) = \frac{d^k}{dt^k}f(t)$ for its kth derivative.

MGF of a Standard Gaussian

Consider the Standard Gaussian r.v. $Z \sim \mathcal{N}(0,1)$. We compute its MGF:

$$M_Z(t) = \mathbb{E}[e^{tZ}] = \frac{1}{\sqrt{2\pi}} \int e^{tz} e^{-\frac{z^2}{2}} dz$$

To compute it, we use a standard trick when manipulating Gaussians: *completing the square*

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}} \int e^{tz} e^{-\frac{z^2}{2}} dz &= \frac{1}{\sqrt{2\pi}} \int e^{-\frac{(z-t)^2}{2}} e^{\frac{t^2}{2}} dz \\
&= e^{\frac{t^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \int e^{-\frac{(z-t)^2}{2}} dz \\
&= e^{\frac{t^2}{2}} \cdot 1
\end{aligned}$$

Therefore

$$M_Z(t) = e^{\frac{t^2}{2}}$$

Moments of a Standard Gaussian

We have seen that for any r.v X,

$$m_k = M_X^{(k)}(0)\,, \quad k = 1, 2, \ldots$$

If $X = Z \sim \mathcal{N}(0,1)$, compute

$$M_Z^{(k)}(0) = \frac{d^k}{dt^k} e^{\frac{t^2}{2}} \Big|_{t=0}$$

It yields

$$\begin{aligned}
M_Z^{(1)}(t) &= te^{\frac{t^2}{2}} && \Rightarrow m_1 = M_Z^{(1)}(0) = 0 \\
M_Z^{(2)}(t) &= e^{\frac{t^2}{2}} + t^2 e^{\frac{t^2}{2}} && \Rightarrow m_2 = M_Z^{(2)}(0) = 1 \\
M_Z^{(3)}(0) &= 0 \\
M_Z^{(4)}(0) &= 3
\end{aligned}$$

Sample moments

- Statistical model $\big(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$.
- Assume $\theta \subset \mathbb{R}^d$ (d parameters to estimate)
- Moments $m_k(\theta) = \mathbb{E}[X^k], k = 1, 2, \dots$
- Let $X_1, \dots, X_n$ be an i.i.d. observations from this model

The kth **sample moment** is

$$\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k\,.$$

By LLN, we have

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

Methods of moments estimator

Definition

The **methods of moments estimator** $\hat{\theta}_n \in \mathbb{R}^d$ satisfies

$$\begin{aligned} m_1(\hat{\theta}_n) &= \hat{m}_1 \\ m_2(\hat{\theta}_n) &= \hat{m}_2 \\ \vdots \quad & \quad \vdots \\ m_d(\hat{\theta}_n) &= \hat{m}_d \end{aligned}$$

This a system of d equations with d unknowns.

Ex. 1: $X \sim \mathcal{N}(\mu, \sigma^2)(d=2)$

$$\begin{aligned} m_1 &= \mu \\ m_2 &= \sigma^2 + \mu^2 \end{aligned}$$

$$(\hat{\mu}_n, \hat{\sigma}_n^2) = \left(\bar{X}_n, \bar{X}_n^2 - (\bar{X}_n)^2\right)$$

Ex. 2: $X \sim \text{Ber}(p)(d=1)$

$$m_1 = p$$

$$\hat{p}_n = \bar{X}_n$$

Recap

- Three principled methods for estimation: maximum likelihood, M-estimation, and the method of moments.
- Maximum likelihood is an example of M-estimation
- MLE tends to be best: asymptotic variance is smallest, given by inverse Fisher information.