18.650
Statistics for Applications

Chapter 3: Maximum Likelihood Estimation

Total variation distance (1)

Let $\big(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$ 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: $\big|\mathbb{P}_{\hat\theta}(A) - \mathbb{P}_{\theta^*}(A)\big|$ 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

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

Total variation distance (2)

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 $p_{\theta'}$:

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

Total variation distance (3)

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 $f_{\theta'}$:

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

Total variation distance (4)

Properties of Total variation:

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

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

Total variation distance (5)

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 $\theta \mapsto \widehat{\mathsf{TV}}(\mathbb{P}_\theta, \mathbb{P}_{\theta^*})$.

Total variation distance (5)

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 (1)

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

Definition

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

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

Kullback-Leibler (KL) divergence (2)

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)
- $\mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \nleq \mathsf{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + \mathsf{KL}(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$ in general

Not a distance.

This is is called a *divergence*.

Asymmetry is the key to our ability to estimate it!

Kullback-Leibler (KL) divergence (3)

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

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

So the function $\theta \mapsto \mathsf{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$ is of the form:
"constant" $- \mathbb{E}_{\theta^*}\big[\log p_\theta(X)\big]$

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

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

Kullback-Leibler (KL) divergence (4)

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

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

This is the **maximum likelihood principle**.

Interlude: maximizing/minimizing functions (1)

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

Interlude: maximizing/minimizing functions (2)

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:

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

Interlude: maximizing/minimizing functions (3)

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:

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

h is concave $\quad \Leftrightarrow \quad x^\top \nabla^2 h(\theta) x \leq 0 \quad \forall x \in \mathbb{R}^d,\ \theta \in \Theta$.

h is strictly concave $\quad \Leftrightarrow \quad x^\top \nabla^2 h(\theta) x < 0 \quad \forall x \in \mathbb{R}^d,\ \theta \in \Theta$.

Examples:

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

Interlude: maximizing/minimizing functions (4)

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

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:

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

Likelihood, Discrete case (2)

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, Discrete case (3)

Example 2 (Poisson model):

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

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

$$\begin{aligned} L(x_1, \ldots, x_n, p) &= \prod_{i=1}^{n} \mathbb{P}_\lambda[X_i = x_i] \\ &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda_i^x}{x_i!} \\ &= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \ldots x_n!}. \end{aligned}$$

Likelihood, Continuous 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 all the $\mathbb{P}_\theta$ have density f_θ.

Definition

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

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

Likelihood, Continuous case (2)

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

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

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

Maximum likelihood estimator (1)

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

Definition

The *likelihood estimator* of θ is defined as:

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

provided it exists.

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

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

Maximum likelihood estimator (2)

Examples

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

Maximum likelihood estimator (3)

Definition: Fisher information

Define the log-likelihood for one observation as:

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

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

$$I(\theta) = \mathbb{E}\left[\nabla \ell(\theta) \nabla \ell(\theta)^\top\right] - \mathbb{E}\left[\nabla \ell(\theta)\right] \mathbb{E}\left[\nabla \ell(\theta)\right]^\top = -\mathbb{E}\left[\nabla^2 \ell(\theta)\right].$$

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

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

Maximum likelihood estimator (4)

Theorem

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

1. The model is identified.
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.

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

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