Lecture 8

- An alternative to MM, and one that is in some ways more principled, is the notion of maximum likelihood estimation (MLE).

- The idea here is to choose parameter $\hat{\theta} \in \Theta \subset \mathbb{R}^D$ that "fits" the data best, in the sense of being the "likeliest" choice of parameter.

- To make this precise, we need to focus on the case that $\mathcal{F} = \{f(x;\theta) \mid \theta \in \Theta\}$ is a family of probability density functions. Let $\{X_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{F}$ be an iid sample from X, unknown. We do not assume the density of X is in $\mathcal{F}$, though it's possible.

Def: Let $\{f(x;\theta) \mid \theta \in \Theta\}_{\Theta \subset \mathbb{R}^D}$ be a parametric family of densities. Let data $\{x_i\}_{i=1}^n \subset \mathbb{R}$ be observed. The likelihood function associated to $\mathcal{F}$ and the data $\{x_i\}_{i=1}^n$ is $\mathcal{L}_n : \Theta \to \mathbb{R}_{\geq 0}$

$$\theta \mapsto \prod_{i=1}^{n} f(x_i;\theta).$$

The idea is that even if $f(x_i;\theta) \neq \mathbb{P}(x_i \text{ occurs under parameter } \theta)$, it is true that $f(x_i;\theta)$ large suggests x_i is likely under parameter θ.

- So, $\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ is large if the observed data $\{x_i\}_{i=1}^n$ is likely under θ.

- Notice we need a parametric family *and* data to define a likelihood function

Remark: While for all θ, $\int_{\mathbb{R}} f(x;\theta)\,dx = 1$, it need not be the case (and generically is not) that $\int_{\Theta} f(x;\theta)\,d\theta = 1$ for any x.

The data variable (x) and the parameter variable (θ) behave very differently.

- Given the likelihood function $\mathcal{L}_n$, the ideal MLE is simple: make $\mathcal{L}_n$ maximal:

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}}\ \mathcal{L}_n(\theta), \quad \text{i.e. } \mathcal{L}_n(\hat{\theta}_n) = \max_{\theta\in\Theta} \mathcal{L}_n(\theta).$$

[redacted]

- So, given data and a model, all we have to do is do calculus on $\mathcal{L}_n$! Easier said than done.

- Because taking $\frac{\partial}{\partial \theta}$ of a product is tedious, it is preferrable often to consider not $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(x_i;\theta)$, but $\log(\mathcal{L}_n(\theta)) = \sum_{i=1}^n \log(f(x_i;\theta))$.

This log-likelihood is often easier to handle. Since $y \mapsto \log(y)$ is increasing, the optima don't change.

ex: Let $\mathcal{F} = \{f_\lambda \mid \lambda > 0\}$ be the collection of exponential densities,

$$f_\lambda(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$$

assumed $x_i > 0 \ \forall i$

Then given data $\{x_i\}_{i=1}^n \sim \underline{X}$, we can compute the likelihood and log-likelihood:

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^n f(x_i;\lambda)$$

$$= \lambda^n \prod_{i=1}^n \exp(-\lambda x_i)$$

$$= \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right)$$

$$\log(\mathcal{L}_n(\lambda)) = \log\left(\lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right)\right)$$

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

it's typically to see some kind of "normalizing constant" appearing

To find the MLE estimator $\hat{\lambda}$, we can solve

$$\frac{d}{d\lambda}\left[\log\left(\mathcal{L}_n(\lambda)\right)\right] = 0 \text{, i.e. } \frac{d}{d\lambda}\left[n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i\right] = 0$$

$$\Leftrightarrow \quad \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\Leftrightarrow \quad \lambda = \frac{n}{\sum_{i=1}^{n} x_i}, \text{ same as MM.}$$

- In the previous example, the log-likelihood was differentiable, so we were able to use calculus. When it is not differentiable, we need to think harder!

ex: Let $\mathcal{F} = \{f_\theta \mid \theta \in \mathbb{R}_+\}$, where $f_\theta = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta) \\ 0, & \text{else} \end{cases}$

be the family of uniform densities on $[0, \theta)$. Then given data $\{x_i\}_{i=1}^{n} \overset{iid}{\sim} X$,

assumed to be positive, what is $\hat{\theta}_{MLE}$? Writing out the likelihood gives

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\theta} \cdot \begin{cases} 1, & x_i \le \theta \\ 0, & x_i > \theta \end{cases}$$

$$= \frac{1}{\theta^n} \cdot \prod_{i=1}^{n} \begin{cases} 1, & x_i \le \theta \\ 0, & x_i > \theta \end{cases}$$

Notice that $\prod_{i=1}^{n} \begin{cases} 1, & x_i \le \theta \\ 0, & x_i > \theta \end{cases} = \begin{cases} 1, & x_i \le \theta \text{ for all } i \\ 0, & \text{else.} \end{cases}$

So, $\mathcal{L}_n(\theta) = \underbrace{\frac{1}{\theta^n}}_{\text{(I)}} \cdot \underbrace{\begin{cases} 1, & \max_i x_i \le \theta \\ 0, & \text{else} \end{cases}}_{\text{(II)}}$

Notice the tension between (I) and (II): to keep (I) large, we want θ small. But (II) switches from 0 to 1 only once $\theta \ge \max_i x_i$. So,

$$\hat{\theta}_{MLE} = \max_i x_i.$$