Lecture #7

- Let us now consider parametric models. We consider a space of functions parametrized by $\Theta \subset \mathbb{R}^D$: $\mathcal{F} = \{ f(x;\theta) \mid \theta \in \Theta \}$. We can think of $f(x;\theta)$ two ways:

  (1.) For fixed θ, $x \mapsto f(x;\theta)$ could be the density/cdf associated to a fixed parameter set.

  (2.) For fixed x, $\theta \mapsto f(x;\theta)$ illustrates how the family changes with respect to the parameter.

- A central question for us is, given data, how to determine a "good" choice of parameter? We'll focus on two high level methods: (1.) Method of moments (MM)
  (2.) Maximum likelihood estimation (MLE)

- Let us recall the definition of moments of a r.v.

Defn. Let X be a r.v. (1.) For $\rho \geq 1$, the ρ^{th} moment of X is $\mathbb{E}(X^\rho)$.

(2.) For $\rho \geq 1$, the ρ^{th} sample moment of X is

$$\frac{1}{n} \sum_{i=1}^{n} X_i^\rho, \quad \text{where } \{X_i\}_{i=1}^n \overset{iid}{\sim} X.$$

- The idea of MM is to choose parameters so that the <u>exact</u> moment matches the <u>sample</u> moment.

- That is, let $X(\theta)$ be the r.v. associated to parameter θ. Then we can compute the (exact) ρ^{th} moment $\alpha_\rho(\theta) = \mathbb{E}(X(\theta)^\rho)$.

- Given data, we can compute sample moments $\hat{\alpha}_\rho = \frac{1}{n}\sum_{i=1}^{n} X_i^\rho$, where $\{X_i\}_{i=1}^n \overset{iid}{\sim} X$.

- The MM idea is to put these in correspondence via a system of equations: choose $\hat{\theta} \in \Theta \subseteq \mathbb{R}^D$ s.t. $\alpha_\rho(\hat{\theta}) = \hat{\alpha}_\rho$, $\rho = 1, \dots, D$.

Remark: We need to do this for $\rho = 1, \dots, D$ in order to have D equations in the D unknowns of θ. This can get a little tedious if D is large!

ex: Consider a 1 parameter family of exponentials: $f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{else} \end{cases}$

Given data $\{X_i\}_{i=1}^n \sim X$, our task is to estimate a good choice of λ. Since there is only one parameter (λ), MM says "choose λ^* so that f_{λ^*} has $\mathbb{E}$ matching the empirical average observed from the data".

So, choose λ^* s.t. $\mathbb{E}(\bar{X}(\lambda)) = \int_0^\infty x \cdot \lambda e^{-\lambda x} dx = \frac{1}{n}\sum_{i=1}^n x_i$.

Through IBP, we see $\int_0^\infty x \cdot \lambda \exp(-\lambda x) dx = \frac{1}{\lambda}$, so we simply set

$\lambda^* = \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^{-1}$ Easy enough!

ex: Let us consider the 2-parameter family of normal distributions, i.e.

$\left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[x-\mu]^2}{2\sigma^2}\right)\right\}_{\substack{\mu\in\mathbb{R} \\ \sigma^2\in\mathbb{R}_+}}$. Given data $\{x_i\}_{i=1}^n \sim X$ (unknown $\bar{X}$), our goal is to estimate "good" parameters σ^2 and μ. The MM suggests choosing $(\hat{\mu}, \hat{\sigma}^2)$ s.t. that $\mathbb{E}(\bar{X}(\hat{\mu}, \hat{\sigma}^2)) = \frac{1}{n}\sum_{i=1}^n x_i$

$$\mathbb{E}\left(\bar{X}(\hat{\mu}, \hat{\sigma}^2)^2\right) = \frac{1}{n}\sum_{i=1}^n x_i^2,$$

$X(\mu, \sigma^2)$ the Gaussian r.v. with parameters (μ, σ^2).

We know $\mathbb{E}(\bar{X}(\hat{\mu}, \sigma^2)) = \hat{\mu}$, so we know the MM choice is

$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$. Natural enough!

- For the second moment, we recall $\mathrm{Var}(\bar{X}) = \mathbb{E}(\bar{X}^2) - \mathbb{E}(\bar{X})^2$ for any $\bar{X}$,

and hence $\mathbb{E}\left(\bar{X}(\hat{\mu}, \hat{\sigma}^2)\right) = \mathrm{Var}\left(\bar{X}(\hat{\mu}, \hat{\sigma}^2)\right) + \mathbb{E}\left(\bar{X}(\hat{\mu}, \hat{\sigma}^2)\right)^2$

$$= \hat{\sigma}^2 + \hat{\mu}^2.$$

- So, MM requires us to solve the system $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\hat{\mu} + \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Again, natural enough.

- The MM has some theoretical guarantees we can prove:

Theorem (Properties of MM): Let $\hat{\theta}_n$ denote the MM estimator. Under appropriate conditions on the family $\mathcal{F} = \{f(x;\theta) \mid \theta \in \Theta\}$,

(1.) $\hat{\theta}_n$ exists with probability 1 (almost surely).

(2.) $\hat{\theta}_n \xrightarrow{P} \theta$.

(3.) $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \rightsquigarrow N(0, \Sigma)$, $\Sigma = g\,\mathbb{E}\left(YY^T\right)g^T$, $Y = (X, X^2, \ldots, X^p)$

$$g = \left(\frac{\partial \alpha_1^{-1}(\theta)}{\partial \theta}, \ldots, \frac{\partial \alpha_p^{-1}(\theta)}{\partial \theta}\right)$$

- Suffice it to say, the MM is almost surely consistent and asymptotically normal.