Stats 300b: Theory of Statistics **Winter 2019**

Lecture 4 – January 17

Lecturer: John Duchi *Scribe: Esha Maiti, Doug Callahan*

Warning: *these notes may contain factual errors*

Reading: Van der Vaart Ch. 4

Outline: Moment methods

- inverse function theorem, definition
- applications in exponential family models
- asymptotic normality in exponential family models
- efficiency of estimators in particular asymptotic relative efficiency

1 Moment methods and the inverse function theorem

Say we have a function $f : X \to \mathbb{R}^d$, and $P||f||^2 = E_P[||f(X)||^2] = \int ||f(x)||^2 dP(x) < \infty$.

Define $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$, the sample mean of $f(X)$.

Then, by the central limit theorem, $\sqrt{n}(P_n f - Pf) \xrightarrow{d} \mathcal{N}(0, \Sigma)$, where $\Sigma = \text{Cov}(f(X))$.

Let us suppose we have a family $\{P_\theta\}_{\theta\in\Theta}$ indexed by parameter θ. We have expectation mapping $e(\theta) := E_{P_\theta}[f(X)] = P_\theta f$. Since $f : X \to \mathbb{R}^d$, we have $e : \theta \to \mathbb{R}^d$.

Suppose e^{-1} exists; we might expect $e^{-1}(P_n f) \approx e^{-1}(P_\theta f) = \theta$. Furthermore, if it were differentiable (i.e., $(e^{-1})'(t) = \frac{\partial}{\partial t}(e^{-1})'(t)$ exists at $t = P_n f$), then we could immediately use the delta method to get asymptotic normality, parameter estimates, etc.:

$$\sqrt{n}(e^{-1}(P_n f) - e^{-1}(P_\theta f)) = \sqrt{n}(e^{-1}(P_n f) - \theta) \xrightarrow{d} \mathcal{N}(0, [\nabla e^{-1}(P_\theta f)]^T \Sigma [\nabla e^{-1}(P_\theta f)])$$

We understand whether or not the inverse of a function is differentiable from the inverse function theorem.

Lemma 1 (Inverse function theorem)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable in a neighborhood of a point $\theta \in \mathbb{R}^d$, where $F'(\theta) \in \mathbb{R}^{d\times d}$ is invertible. Then, in a neighborhood of $t = F(\theta)$, we have $(F^{-1})'(t) = \frac{\partial}{\partial t}F^{-1}(t) = \frac{1}{F'(F^{-1}(t))}$, and this derivative is continuous.*

Proof Let $t = F(\theta)$. Then, $\theta = F^{-1}(t)$. Let δ be a small change in t, and Δ be the corresponding small change in θ. Then:

$$\Delta \approx \frac{\partial \theta}{\partial t}\delta = \frac{\partial}{\partial t}F^{-1}(t)\delta = (F^{-1})'(t)\delta \qquad (*)$$

And by Taylor series expansion:

$$F(\theta + \Delta) = F(\theta) + F'(\theta)\Delta + O(||\Delta||^2)$$

Then, for small Δ, we have $t + \delta = F(\theta + \Delta) = F(\theta) + F'(\theta)\Delta$, i.e.,
$\delta \approx F'(\theta)\Delta$, i.e., $\Delta \approx (F'(\theta))^{-1}\delta$. From (*), we then have $(F^{-1})'(t) = (F'(\theta))^{-1} = (F'(F^{-1}(t)))^{-1}$.
$\square$

Theorem 2. *Let $e(\theta) = P_\theta f$ be one-to-one on some open set $\Theta \subset \mathbb{R}^d$, and continuously differentiable near θ_0, where $e'(\theta_0) \in \mathbb{R}^{d\times d}$ is non-singular.*

If $P_{\theta_0}||f||^2 < \infty$, then:

1. *$P_n f \in dom(e^{-1})$ eventually*
2. *$\hat{\theta}_n = e^{-1}(P_n f)$ satisfies $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, [(e'(\theta_0))^{-1}]^T \operatorname{Cov}_{\theta_0}(f)(e'(\theta_0))^{-1})$, where $\operatorname{Cov}_{\theta_0}(f) = P_{\theta_0}((f - P_{\theta_0} f)(f - P_{\theta_0} f)^T)$.*

Proof By the inverse function theorem, there exists a neighborhood ν (or an open set) around $e(\theta_0)$ s.t. (e^{-1}) exists on ν and is continuous.
As $P_n f \xrightarrow{a.s.} P_{\theta_0} f$, $P_n f \in \nu$ eventually, so $e^{-1}(P_n f)$ exists.
Now, apply the delta method:

$$\sqrt{n}(P_n f - P_{\theta_0} f) \xrightarrow{d} \mathcal{N}(0, \operatorname{Cov}(f)) \Longrightarrow$$

$$\sqrt{n}(e^{-1}(P_n f) - \theta_0) \xrightarrow{d} (e^{-1})'(e(\theta_0))Z = (e'(\theta_0))^{-1}Z, \text{ where } Z \sim \mathcal{N}(0, I_{d\times d})$$

$\square$

To get the mathematical form of the statement above, we note that:

$$\sqrt{n}(P_n f - P_{\theta_0} f) \xrightarrow{d} \mathcal{N}(0, \operatorname*{Cov}_{\theta_0} f) \Longrightarrow$$

$$\sqrt{n}(e^{-1}(P_n f) - \theta_0) \xrightarrow{d} \mathcal{N}(0, (e^{-1})'(e(\theta_0)) \operatorname*{Cov}_{\theta_0}(f)[(e^{-1})'(e(\theta_0))]^\intercal)$$

i.e.,

$$\sqrt{n}(e^{-1}(P_n f) - \theta_0) \xrightarrow{d} \mathcal{N}(0, e'(\theta_0)^{-1} \operatorname*{Cov}_{\theta_0}(f)(e'(\theta_0)^{-1})^\intercal)$$

where Lemma 1 was used. This gives a number of powerful moment-matching estimators.

Example 1. Estimate the mean of a Bernoulli distribution on $\{\pm 1\}$.

$P_\theta(x) = \frac{e^{\theta x}}{1+e^{\theta x}} = \frac{1}{1+e^{-\theta x}}$. Then,
$e(\theta) = E_\theta(x) = \frac{1}{1+e^{-\theta}} - \frac{1}{1+e^{\theta}} = \frac{e^\theta - 1}{e^\theta + 1}$.
$t = e(\theta) \Leftrightarrow \theta = \log(\frac{1+t}{1-t})$
Let $p_\theta = P(x = 1) = \frac{e^\theta}{1+e^\theta}$. Then,
$e'(\theta) = \frac{(e^\theta+1)e^\theta - (e^\theta-1)e^\theta}{(e^\theta+1)^2} = \frac{2e^\theta}{(1+e^\theta)^2} = 2p_\theta(1 - p_\theta)$.

Thus, $e'(\theta)^{-1} = \frac{1}{2p_\theta(1-p_\theta)}$.
Also, $E_\theta(x^2) = 1$. Thus, $\mathrm{Cov}_\theta(x) = 1 - e(\theta)^2 = 1 - \frac{(e^\theta-1)^2}{(e^\theta+1)^2} = \frac{4e^\theta}{(e^\theta+1)^2} = 4p_\theta(1-p_\theta)$.
So, if $\hat{\theta}_n = \log \frac{1+\bar{x}_n}{1-\bar{x}_n} = \mathrm{argmax}_\theta \sum_{i=1}^n \log p_\theta(x_i)$, then $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \frac{1}{p_\theta(1-p_\theta)})$.
Thus, we have the asymptotic distribution. As expected, it is more difficult to find estimators when $e(\theta)$ gets close to 1 or -1.

2 Exponential family models

Definition 2.1 (Exponential family). *A family $\{P_\theta\}_{\theta\in\Theta}$ is a regular exponential family with respect to a base measure μ if there exists a function $T : x \to \mathbb{R}^d$ (sufficient statistic) and density $p_\theta(x) = e^{\theta^\intercal T(x) - A(\theta)}$, where $A(\theta) = \log \int e^{\theta^\intercal T(x)} d\mu(x)$ and is called the log-partition or cumulant generating function.*

Standard results:

(1) $A(\theta)$ is convex in θ, and ∞-differentiable on its domain, $\{\theta : A(\theta) < \infty\}$.

(2) $\frac{\partial^k}{\partial\theta_1^{\alpha_1} \cdot \ldots \cdot \partial\theta_d^{\alpha_d}} e^{A(\theta)} = \int T_1(x)^{\alpha_1} \cdot \ldots \cdot T_d(x)^{\alpha_d} e^{\theta^\intercal T(x)} d\mu(x)$, $\alpha_i \in \mathbb{N}$ and $\sum_{i=1}^d \alpha_i = k$ (equivalently, $\frac{\partial^k}{\partial\theta_1^{\alpha_1} \cdot \ldots \cdot \partial\theta_d^{\alpha_d}} A(\theta) = E_{p_\theta}(T_1(x)^{\alpha_1} \cdot \ldots \cdot T_d(x)^{\alpha_d})$)

(3) For the gradient, we have $\frac{\partial}{\partial\theta} A(\theta) = \nabla A(\theta) = \frac{1}{\int e^{\theta^\intercal T} d\mu} \int T e^{\theta^\intercal T} d\mu = E_\theta T$.

(4) For the Hessian, we have:

$$\nabla^2 A(\theta) = \nabla\nabla^\intercal A(\theta) = \int T(x)T(x)^\intercal dp_\theta(x) - \left(\int T dp_\theta\right)\left(\int T dp_\theta\right)^\intercal = \underset{\theta}{\mathrm{Cov}}(T(x))$$

Note that in our earlier notation, $e(\theta) = E_\theta[T(x)] = \nabla A(\theta)$, so $e'(\theta) = \nabla^2 A(\theta) = \mathrm{Cov}_\theta(T) \geq 0$.

Aside: Suppose we use maximum likelihood to estimate θ. Let log likelihood $L_n(\theta) := \sum_{i=1}^n \log p_\theta(x_i) = \sum_{i=1}^n \theta^\intercal T(x_i) - nA(\theta)$. As $\theta \longmapsto A(\theta)$ is smooth and convex, our solutions are characterized by $\nabla L_n(\theta) = 0$, i.e.,

$$\nabla L_n(\theta) = n(P_n T - P_\theta T) = n(P_n T - e(\theta)) \Longrightarrow \hat{\theta}_{\mathrm{ML}} = e^{-1}(P_n T) \quad \text{(moment estimator)}$$

3 Asymptotic normality and efficiency

Theorem 3. *Suppose $\{p_\theta\}$ is full rank, i.e., $\nabla^2 A(\theta) > 0$, i.e., $\mathrm{Cov}_\theta(T) > 0$, or the covariance is full rank. Then, the solution to $P_n T = \frac{1}{n}\sum_{i=1}^n T(x_i) = E_\theta(T)$ exists eventually (when $x_i \overset{iid}{\sim} p_{\theta_0}$), and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, (e'(\theta_0))^{-1}\mathrm{Cov}(T)(e'(\theta_0))^{-1}) = \mathcal{N}(0, (\nabla^2 A(\theta_0))^{-1}) = \mathcal{N}(0, I_{\theta_0}^{-1})$, where I_{θ_0} is the Fisher information.*

Proof The above is proven by noticing that $I_\theta = -\nabla^2 A(\theta) = -\mathrm{Cov}(T)$, and applying general moment-method asymptotics.

Notice that we do not require consistency $\hat{\theta}_n \xrightarrow{p} \theta_0$ for these theorems. □

Example 2. Linear regression.

$Y_i|X_i \sim \mathcal{N}({X_i}^\intercal\theta_0, \delta^2)$
$p_\theta(Y_i|X_i) \propto e^{-\frac{1}{2\delta^2}({X_i}^\intercal\theta - Y_i)^2}$
$X = \begin{bmatrix} X_1^\intercal \\ \vdots \\ X_n^\intercal \end{bmatrix} \in \mathbb{R}^{n\times d}$, $L_n(\theta) = \sum_{i=1}^n \log p_\theta(Y_i|X_i) = -\frac{1}{2\delta^2}||X\theta - Y||_2^2$
$\nabla L_n(\theta) = (-X^\intercal X\theta + X^\intercal Y)/\delta^2 = 0 \Longrightarrow \hat{\theta} = (X^\intercal X)^{-1}X^\intercal Y$.
The Fisher information is obtained as $I_\theta = -\nabla^2 L_n(\theta) = (X^\intercal X)/\delta^2$.
Thus, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \delta^2(X^\intercal X)^{-1})$.