Statistics 300b: Theory of Statistics
Winter 2017

Lecture 5 – January 24

Lecturer: John Duchi
Scribe: Amanda Zhang

Warning: these notes may contain factual errors

Reading: VDV 4

The method of moments determines estimators by comparing sample and theoretical moments. Let $X_1, \cdots, X_n$ be a sample from a distribution $P_\theta$ that depends on a parameter $\theta$, ranging over some set $\Theta$. Given $f : \mathcal{X} \to \mathbb{R}^d$ with $P_{\theta_0}\|f\|_2^2 < \infty$. By central limit theorem,

$$\sqrt{n}(P_n f - P_{\theta_0} f) \sim N \left( 0, \text{Cov}(f) \right). \tag{1}$$

Let $e : \Theta \to \mathbb{R}^d$ be the vector-valued expectation $e(\theta) = P_{\theta} f$. If $e$ is “nice” in that $e^{-1}(P_{\theta_0} f) = \theta_0$. Then by delta method,

$$\sqrt{n} \left( e^{-1}(P_n f) - e^{-1}(P_{\theta_0} f) \right) = \sqrt{n} \left( e^{-1}(P_n f) - \theta_0 \right) \sim (e(P_{\theta_0} f)')^{-1} N \left( 0, \text{Cov}(f) \right).$$

**Theorem 1. Inverse function theorem** Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable in a neighborhood of $\theta \in \mathbb{R}^d$, where $F'(\theta)$ is invertible, that is, $\det(F'(\theta)) \neq 0$. Then in a neighborhood of $t = F(\theta)$, we have

$$(F^{-1})'(t) = \frac{\partial}{\partial t} F'(t) = [F'(F^{-1}(t))]^{-1} \tag{2}$$

and $(F^{-1})'$ is continuous.

**Theorem 2.** Suppose that $e(\theta) = P_{\theta} f$ is one-to-one on an open set $\Theta \subset \mathbb{R}^d$ and continuously differentiable at $\theta_0$ with nonsingular derivative $e'_{\theta_0}$. Moreover, assume that $P_{\theta_0}\|f\|_2^2 < \infty$. Then moment estimators $\hat{\theta}_n$ exist with probability tending to one and satisfy

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \sim N \left( 0, (e(\theta_0)')^{-1} P_{\theta_0} f f^T ((e(\theta_0)')^{-1})^T \right) \tag{3}$$

**Proof** Continuously differentiability at $\theta_0$ presumes differentiability in a neighborhood and the continuity of $\theta \mapsto e'_{\theta}$ and non-singularity of $e'_{\theta_0}$ imply non-singularity in a neighborhood. Therefore, by the inverse function theorem, there exist open neighborhoods $U$ of $\theta_0$ and $V$ of $P_{\theta_0} f$ such that $e : U \to V$ is a differentiable bijection with a differentiable inverse $e^{-1} : V \to U$. Moment estimators $\hat{\theta}_n = e^{-1}(P_n f)$ exist as soon as $P_n f \in V$, which happens with probability tending to 1 by the law of large numbers. We know, by central limit theorem, that

$$\sqrt{n}(P_n f - P_{\theta_0} f) \sim N \left( 0, \text{Cov}(f) \right). \tag{4}$$

Apply Delta Method, we get:

$$\sqrt{n} \left( e^{-1}(P_n f) - e^{-1}(P_{\theta_0} f) \right) = \sqrt{n}(\hat{\theta}_n - \theta_0) \sim N \left( 0, (e(\theta_0)')^{-1} P_{\theta_0} f f^T ((e(\theta_0)')^{-1})^T \right). \tag{5}$$

\qed


Example 1. Let $X_i$ be i.i.d Bernoulli\{±1\} random variables. Then

$$P_{\theta}(X = x) = \frac{e^{\theta x}}{1 + e^{\theta x}} = \frac{1}{1 + e^{-\theta x}}.$$  

$$e(\theta) = \mathbb{E}_{\theta}[X] = \frac{1}{1 + e^{-\theta}} - \frac{1}{1 + e^{\theta}} = \frac{e^{\theta} - 1}{e^{\theta} + 1}.$$  

Then $e^{-1}(t) = \log \frac{1 + t}{1 - t}$ and $e'(\theta) = \frac{e^\theta}{e^{\theta} + 1} - \frac{e^{-\theta}}{e^{-\theta} + 1} = \frac{e^\theta - 1}{(1 + e^\theta)^2} = P_0(1 - P_0).$ In particular, $(e'(\theta))^{-1} = \frac{1}{P_0(1 - P_0)}.$ The covariance $\text{Cov}_{\theta}(x)$ is $4P_0(1 - P_0).$ Applying Theorem 2, we get

$$\sqrt{n} \left( e^{-1}(\bar{X}_n) - \theta \right) \sim N(0, \frac{4}{P_0(1 - P_0)}).$$

1 Exponential Family

Given a measure $\mu$, we define an exponential family of probability distributions as those distributions whose density (relative to $\mu$) have the following general form:

$$p(x|\theta) = h(x) \exp[\theta^T T(x) - A(\theta)],$$  

for a parameter vector $\theta$, often referred to as the canonical parameter, and for given functions $T$ and $h$. The statistic $T(X)$ is referred to as a sufficient statistic; the function $A(\theta)$ is known as the cumulant function. Integrating (6) with respect to the measure $\mu$, we have

$$A(\theta) = \log \int h(x) \exp[\theta^T T(x)] d\mu(x).$$  

The set of parameters $\theta$ for which the integral in (7) is finite is referred to as the natural parameter space:

$$\mathcal{N} = \{ \theta : \int h(x) \exp\{\theta^T T(x)\} d\mu(x) < \infty \}.$$  

We will restrict ourselves to exponential families for which the natural parameter space is a nonempty open set. Such families are referred to as regular.

Proposition 3. $A(\theta)$ is convex and infinitely differentiable.

As a consequence, we can calculate expectation and variance by differentiating $A$ with respect to $\theta$:

$$\frac{\partial A}{\partial \theta^T} = \frac{\int T(x) \exp\{\theta^T T(x)\} h(x) d\mu(x)}{\int \exp\{\theta^T T(x)\} h(x) d\mu(x)} = \int T(x) \exp\{\theta^T T(x) - A(\theta)\} h(x) d\mu(x) = \mathbb{E}[T(X)].$$

$$\frac{\partial^2 A}{\partial \theta \partial \theta^T} = \int T(x)(T(x) - \mathbb{E}[T(X)])^T \exp\{\theta^T T(x) - A(\theta)\} h(x) d\mu(x) = \mathbb{E}[T(X)T(X)^T] - \mathbb{E}[T(X)]\mathbb{E}[T(X)]^T = \text{Var}[T(X)].$$
In general, higher-order moments of sufficient statistic can be obtained by taking higher-order derivatives of $A$.

With the above techniques, it’s not hard to obtain maximum likelihood estimates of the mean parameter in exponential family distributions. Consider an i.i.d. data set, $S = \{X_1, \cdots, X_n\}$. The log likelihood is:

$$\ell(\theta|S) = \log \left( \prod_{i=1}^{n} h(X_i) \right) + \theta^{T} \left( \sum_{i=1}^{n} T(X_i) \right) - n A(\theta).$$

(9)

Taking the graduate with respect to $\theta$ yields:

$$\nabla_{\theta} \ell = \sum_{i=1}^{n} T(X_i) - n \nabla_{\theta} A(\theta),$$

(10)

and setting it to zero gives:

$$\nabla_{\theta} A(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} T(X_i).$$

(11)

Finally, defining $\mu = \mathbb{E}[T(X)]$, we obtain

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^{n} T(X_i)$$

(12)

as the general formula for maximum likelihood estimation of the mean parameter in the exponential family.

**Theorem 4.** Let $\Theta \subset \mathbb{R}^d$ be open. Let the (exponential) family of densities $p_{\theta}$ be given by (6) and be of full rank, meaning $\text{Cov}_{\theta}(T) > 0$. Then the likelihood equation $\nabla_{\theta} A(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} T(X_i)$ has a unique solution $\hat{\theta}_n$ with probability tending to 1 and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow_{P_{\theta_0}} N \left( 0, \nabla^2 A(\theta_0)^{-1} \right)$$

(13)

**Proof** By central limit theorem, we know

$$\sqrt{n}(P_n T - P_{\theta_0} T) \rightsquigarrow N \left( 0, \text{Cov}(T) \right).$$

Define $e(\theta) = P_{\theta} T$ as before. Then $e(\theta) = P_n T = \nabla A(\theta)$ and $(e(\theta_0))^{-1} = (\nabla^2 A(\theta_0))^{-1}$. Since $\text{Cov}_{\theta_0}(T) = \nabla^2 A(\theta_0)^{-1}$, apply Theorem 2 and (13) follows. \qed

**Remark:** in exponential family, Fisher information

$$I(\theta) = \mathbb{E}_{\theta}[\nabla \ell_{\theta} \nabla^{T} \ell_{\theta}] = \text{Cov}(T) = \nabla^2 A(\theta).$$

**Example 2 (Linear Regression).** Let $(x, y) \in \mathbb{R}^d \times R$ be i.i.d samples with density

$$p_{\theta}(y|x) = \exp \left( -\frac{1}{2} (X^{T} \theta - y)^2 \right),$$

where $Y|X = x$ follows $N(\theta^{T} x, 1)$. Then $L_n(\theta) = \sum_{i=1}^{N} \log P_{\theta}(y_i|x_i) = -\frac{1}{2} ||x_\theta - y_\theta||^2_2$ and $\hat{\theta}_n = \arg\max_{\theta} ||X_{\theta} - Y||^2_2 = (X^{T} X)^{-1} X^{T} Y$. Furthermore, $\ell_{\theta}(Y|X = x) = -\frac{1}{2} (x^{T} \theta - y)^2 \Rightarrow \nabla \ell_{\theta}(y|X = x) = x(x^{T} \theta - y) \Rightarrow \nabla^{2} \ell_{\theta} = xx^{T}$. Thus, $I(\theta) = \mathbb{E}[XX^{T}]$. Apply Theorem 4, we obtain

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, \mathbb{E}[XX^{T}]^{-1}).$$
Definition 1.1. **efficient** We say an estimator \( \hat{\theta}_n \) is efficient for a parameter \( \theta \) in model \( \{P_\theta\} \) if
\[
\sqrt{n}(\hat{\theta}_n - \theta) \sim_{P_\theta} N(0, I_\theta^{-1}).
\]

Definition 1.2. **asymptotic relative efficiency (ARE)** Let \( \hat{\theta}_n \) and \( T_n \) be estimators of parameter \( \theta \in \mathbb{R} \). Assume that
\[
\sqrt{n}(\hat{\theta}_n - \theta) \sim N(0, \sigma^2(\theta)).
\]
Let \( m(n) \to \infty \) such that
\[
\sqrt{n}(T_{m(n)} - \theta) \sim N(0, \sigma^2(\theta)).
\]
The asymptotic relative efficiency of \( \hat{\theta}_n \) with respect to \( T_n \) is
\[
\lim_{n \to \infty} \inf \frac{m(n)}{n}.
\]

The intuition here is if \( ARE = c \gg 1 \), then \( T_n \) requires sample size \( C_n \gg n \) to get estimate of quality as \( \hat{\theta}_n \). We can also see the interpretation through confidence interval: if ARE of \( \hat{\theta}_n \) vs \( T_n \) is \( c \), then the asymptotic \( 1 - \alpha \) confidence interval for \( \theta \) take \( Z_{1-\alpha/2} \) such that
\[
Pr \left( \left| Z \right| \geq Z_{1-\alpha/2} \right) = \alpha,
\]
where \( \alpha \sim N(0,1) \). The confidence intervals of \( \hat{\theta}_n \) and \( T_n \) are:
\[
C_{\hat{\theta}_n} : \left( \hat{\theta}_n - Z_{1-\alpha/2} \sqrt{\frac{\alpha^2}{n}}, \hat{\theta}_n + Z_{1-\alpha/2} \sqrt{\frac{\alpha^2}{n}} \right);
\]
\[
C_{T_n} : \left( T_n - Z_{1-\alpha/2} \sqrt{\frac{m(n) \sigma^2}{n}}, T_n + Z_{1-\alpha/2} \sqrt{\frac{m(n) \sigma^2}{n}} \right).
\]
Then we have
\[
\lim_{n \to \infty} P_\theta(\theta \in C_{\hat{\theta}_n}) = \lim_{n \to \infty} P_\theta(\theta \in C_{T_n}) = 1 - \alpha.
\]
Furthermore,
\[
\frac{\text{length} C(C_{T_n})}{\text{length} C(\hat{\theta}_n)} = \sqrt{\text{ARE of } \hat{\theta}_n \text{ with respect to } T_n} = \sqrt{\frac{m(n)}{n}}.
\]

**Proposition 5.** Suppose \( \hat{\theta}_n \) and \( T_n \) are estimators of \( \theta \) such that
\[
\sqrt{n}(\hat{\theta}_n - \theta) \sim N(0, \sigma^2(\theta));
\]
\[
\sqrt{n}(T_n - \theta) \sim N(0, \tau^2(\theta)).
\]
Then the ARE of \( \hat{\theta}_n \) with respect to \( T_n \) is \( \frac{\tau^2(\theta)}{\sigma^2(\theta)} \). (In higher dimensions, it is roughly \( \text{Tr}(\tau^2(\theta)/\sigma^2(\theta)^{-1}) \).

**Proof** Let \( m(n) = \lceil \frac{\sqrt{n} \cdot \sigma^2}{\tau^2} \cdot n \rceil \). Then
\[
\sqrt{n}(T_{m(n)} - \theta) = \sqrt{\frac{n}{m(n)}} \sqrt{m(n) (T_{m(n)} - \theta)} \sim N(0, \sigma^2(\theta)).
\]
Thus, ARE is \( \frac{m(n)}{n} = \frac{\tau^2}{\sigma^2} \). \( \square \)

If \( \tau^2 > \sigma^2 \), we prefer \( \hat{\theta}_n \) over \( T_n \), because \( T_n \) requires \( \frac{\tau^2}{\sigma^2} \) times the sample size \( \hat{\theta}_n \) does for the similar quality.