Stat210B: Theoretical Statistics **Lecture Date: April 19, 2007**

Lecture 27: Optimal Estimators and Functional Delta Method

Lecturer: Michael I. Jordan *Scribe: Guilherme V. Rocha*

1 Achieving Optimal Estimators

From the last class, we know that the best limiting distribution we can hope for a parameter $\psi(\theta)$ is $\mathcal{N}(0, \dot{\psi}_\theta I_\theta \dot{\psi}_\theta^T)$. The next question to ask is whether such bound can be achieved.

This is the theme of our next result:

Lemma 1. ***(Lemma 8.14 in van der Vaart, 1998)***
Assume that the experiment $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean at θ_0 with non-singular Fisher information matrix I_θ. Let ψ be differentiable at θ_0. Let T_n be an estimator sequence in the experiments $(P_\theta^n : \theta \in \mathbb{R}^k)$ such that:

$$\sqrt{n}\,(T_n - \psi(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\psi}_\theta I_\theta^{-1} \dot{\ell}_\theta(X_i) + o_{P_\theta}(1),$$

then T_n is the best regular estimator for $\psi(\theta)$ at θ. Conversely, every best regular estimator sequence satisfies this expansion.

Proof. Let

$$\Delta_{n,\theta} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_\theta(X_i).$$

We know that $\Delta_{n,\theta}$ converges in distribution to a Δ_θ with a $\mathcal{N}(0, I_\theta)$ distribution.

From Theorem 7.2 in van der Vaart (1998) we know that:

$$\log\left(\frac{dP^n_{\theta+\frac{h}{\sqrt{n}}}}{dP^n_\theta}\right) \quad = \quad h^T\Delta_\theta - \frac{1}{2}h^T I_\theta h + o_{P_\theta}(1).$$

Using Slutsky's theorem, we get:

$$\begin{aligned}
\left(\begin{array}{c} \sqrt{n}\,(T_n - \psi(\theta)) \\ \log\left(\frac{dP^n_{\theta+\frac{h}{\sqrt{n}}}}{dP^n_\theta}\right) \end{array}\right) \quad &\overset{\theta}{\rightsquigarrow} \quad \left(\begin{array}{c} \dot{\psi}_\theta I_\theta^{-1}\Delta_\theta \\ h^T\Delta_\theta - \frac{1}{2}h^T I_\theta h \end{array}\right) \\
&\sim \quad \mathcal{N}\left(\left[\begin{array}{c} 0 \\ -\frac{1}{2}h^T I_\theta h \end{array}\right], \left[\begin{array}{cc} \dot{\psi}_\theta I_\theta^{-1}\dot{\psi}_\theta^T & \dot{\psi}_\theta h \\ \dot{\psi}_\theta h^T & h^T I_\theta h \end{array}\right]\right)
\end{aligned}$$

Using Le Cam's third lemma we can conclude that the sequence $\sqrt{n}(T_n - \psi(\theta))$ under $\theta + \frac{h}{\sqrt{n}}$ converges in distribution to $\mathcal{N}(\dot{\psi}_\theta h, \dot{\psi}_\theta I_\theta^{-1}\dot{\psi}_\theta^T)$. Since ψ is differentiable, we have that $\sqrt{n}\left(\psi(\theta + \frac{h}{\sqrt{n}}) - \psi(\theta)\right) \to \dot{\psi}_\theta h$ as $n \to \infty$. We conclude that, under $\theta + \frac{h}{\sqrt{n}}$, $\sqrt{n}(T_n - \psi(\theta + \frac{h}{\sqrt{n}}))$ does not involve h, that is, T_n is regular.

To prove the converse, let T_n and S_n be a two best regular estimator sequences. Along subsequences, it can be shown that:

$$\left(\begin{bmatrix} \sqrt{n}\left(S_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right) \\ \sqrt{n}\left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right) \end{bmatrix}\right) \overset{\theta+\frac{h}{\sqrt{n}}}{\rightsquigarrow} \left(\begin{bmatrix} S - \dot{\psi}_\theta h \\ T - \dot{\psi}_\theta h \end{bmatrix}\right)$$

for a randomized estimator (S, T) in the limiting experiment. Because S_n and T_n are best regular, S and T are best equivariant-in-law. Thus $S = T = \dot{\psi}_\theta X$ almost surely and, as a result, $\sqrt{n}(S_n - T_n)$ converges in distribution to $S - T = 0$. As a result, every two best regular estimator sequences are asymptotically equivalent. To get the result, apply this conclusion to T_n and:

$$S_n \quad = \quad \psi(\theta) + \frac{1}{\sqrt{n}}\dot{\psi}_\theta I_\theta^{-1}\Delta_{n,\theta}$$

□

Remarks on Theorem 8.14:

- From theorem 5.39, an Maximum Likelihod Estimator $\hat{\theta}_n$ satisfies:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_\theta^{-1}\dot{\ell}_\theta(X_i) + o_{P_\theta}(1)$$

  under regularity conditions. It follows that MLEs are asymptotically efficient. This result can be extended to a transforms of the MLE $\psi(\theta)$ for a differentiable ψ by using the delta method and observing that $\psi(\hat{\theta}_n)$ satisfies the expansion in lemma 8.14.

- Lemma 8.14 suggests that Rao score functions leading to tests constructed from the scores are asymptotically efficient.

2 Functional Delta Method

The functional delta method aims at extending the delta method to a nonparametric context. The high level idea is to interpret a statistic as a functional ϕ mapping from the space of probability distributions D to the real line $\mathbb{R}$ and use a notion of derivative of this functional to obtain the asymptotic distribution of $\phi(\hat{\mathbb{F}}_n)$.

We have proven before that:

$$\begin{array}{rc} \sup_x \|\hat{\mathbb{F}}_n(x) - F(x)\| \xrightarrow{p} 0, & \text{(Glivenko-Cantelli Theorem)} \\ \sqrt{n}\left(\hat{\mathbb{F}}_n - F\right) \overset{F}{\rightsquigarrow} \mathbb{G}_F, & \text{(Donsker Theorem)} \end{array}$$

where $\hat{\mathbb{F}}_n$ is the empirical distribution function based in n samples and $\mathbb{G}_F$ is the Brownian Bridge.

Our goal now is to find conditions on the functionals ϕ so we can extend the above modes of convergence to $\phi(\hat{\mathbb{F}}_n)$. As we will see in detail below, consistency of the sequence $\phi(\hat{\mathbb{F}}_n)$ follows easily from assuming ϕ to be continuous with respect to the supremum norm: this a natural extension of the continuous mapping theorem. The generalization of the delta method to functionals is more involved as different notions of differentiability of functionals exist.

Before we jump into the continuous mapping theorem and the functional delta method, a few examples of statistical functionals are in order.

2.1 Examples of statistical functionals

- The mean: $\mu(F) = \int X dF(X)$;
- The variance: $\text{Var}(F) = \int (X - \mu(F))^2 dF(X)$;
- Higher order moments: $\mu_k(F) = \int (X - \mu(F))^k dF(X)$;
- The Kolmogorov-Smirnov statistics: $K(F) = \sup_x \|F(x) - F_0(x)\|$, where F_0 is a fixed hypothesized distribution;
- The Crámer-von Mises statistics: $C(F) = \int (F(x) - F_0(x))^2 \, dF_0(x)$, where F_0 is a fixed hypothesized distribution;
- V-statistics: $\phi(F) = \mathbb{E}_F (T(X_1, X_2, \ldots, X_p))$ where $X_1, X_2, \ldots, X_p$ are independent copies of F-distributed random variables;
- Quantile functional: $\phi(F) = F^{-1}(p) \triangleq \inf_x\{x : F(x) \geq p\}$;

2.2 Consistency of statistical functionals

One possible assumption to ensure that $\phi(\hat{\mathbb{F}}_n)$ converges to $\phi(F)$ is continuous with respect to the supremum norm. Formally, this is defined as:

Definition 2. Continuity of a functional
Let D be the space of distributions and $\phi : D \to \mathbb{R}$. We say ϕ is continuous (with respect to the supremum norm) at F if:

$$\sup_x \|F_n(x) - F(x)\| \to 0 \Rightarrow \phi(F_n) \to \phi(F).$$

The next result is an extension of the continuous mapping theorem to functionals and can be used to establish the consistency of statistical functionals.

Theorem 3. ***Continuous Mapping Theorem for Statistical Functionals*** *Let $\phi : D \to \mathbb{R}$ be a continuous functional at F. It follows that:*

$$\textit{If } \|F_n - F\|_\infty \xrightarrow{p} 0, \ \textit{then } \phi(F_n) - \phi(F) \xrightarrow{p} 0.$$

Proof. From continuity of ϕ (with respect to the sup norm), we have that for every $\varepsilon > 0$, there exists $\delta > 0$ such that:

$$\|F_n - F\|_\infty \leq \delta \Rightarrow \|\phi(F_n) - \phi(F)\| \leq \varepsilon.$$

Hence,

$$0 \leq \mathbb{P}\left(\|\phi(F_n) - \phi(F)\| > \varepsilon\right) \leq \mathbb{P}\left(\|F_n - F\|_\infty > \delta(\varepsilon)\right) \to 0,$$

where the last convergence is due to $\|F_n - F\|_\infty \xrightarrow{p} 0$ by hypothesis. □

2.2.1 Examples of continuous statistical functionals

The following two functionals are continuous with respect to the supremum norm:

- $\phi(F) = F(a)$: The distribution function evaluated at a point.

  To establish the continuity of this function, notice that for a sequence of distributions such that $\|F_n - F\|_\infty \to 0$:

  $$0 \leq \|F_n(a) - F(a)\| \leq \sup_x \|F_n(x) - F(x)\| \to 0$$

- $\phi(F) = \int (F(x) - F_0(x))^2 dF_0(x)$: The Crámer-von Mises functional.

  Again, take a sequence of distributions such that $\|F_n - F\|_\infty \to 0$. We have:

  $$\begin{aligned} 0 \leq \|\phi(F_n) - \phi(F)\| &= \|\int \left[F_n^2 - F^2 + 2F_0(F - F_n)\right] dF_0\| \leq \int |F_n - F| \underbrace{|2F_0 - F - F_n|}_{\leq 2} dF_0 \\ &\leq 2 \sup_x |F_n(x) - F(x)| \int dF_0 \to 0 \end{aligned}$$

2.3 Limiting distribution of statistical functionals

Recall our goal to determine the limiting distribution of $\phi(\hat{\mathbb{F}}_n)$. Heuristically, we might hope to derive it if we can find a linear ϕ'_F resulting in an approximation of the sort:

$$\begin{aligned} \phi(\hat{\mathbb{F}}_n) - \phi(F) &= \phi'_F(\hat{\mathbb{F}}_n - F) + \text{ some residual} \\ &= \sqrt{n}\phi'_F(\hat{\mathbb{G}}_n) + \text{ some residual} \end{aligned}$$

As before, we must keep track of the behavior of the residual term as n grows. The fact that ϕ operates on the infinite dimensional space of distribution functions will require us to be more careful and resort to some concepts of functional analysis. Namely, we will be looking at the notions of Gateaux and Hadamard derivatives.

We start by considering Gateaux derivatives. To see how they can be used, notice that from linearity of ϕ'_F, we have that:

$$\phi'_F(\hat{\mathbb{G}}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi'_F(\delta_{X_i} - F),$$

where δ_{X_i} is the distribution function concentrating all mass at X_i. For each of the terms in the sum, we can consider a "directional" derivative of ϕ in the direction $\delta_{X_i} - F$. That is the Gateaux derivative which in this case is defined as:

$$\phi'_F(\delta_{X_i} - F) = \left.\frac{d}{dt}\left[\phi\left((1-t)F + t\delta_{X_i}\right)\right]\right|_{t=0}.$$

The expression for $\phi'_F(\delta_{X_i} - F)$ above is known in the robust statistics literature as the influence function:

$$IF_{\phi,F}(x) = \left.\frac{d}{dt}\left[\phi\left((1-t)F + t\delta_x\right)\right]\right|_{t=0}.$$

To some extent, it measures how much the statistic $\phi(F)$ is affected by adding a new observation at x to the sample. A related concept is the gross-error sensitivity defined as:

$$\gamma^* = \sup_x IF_{\phi,F}(x).$$

For robustness, γ^* must be bounded.

Going back to the approximation of $\phi(\hat{\mathbb{F}}_n) - \phi(F)$, we now write:

$$\phi(\hat{\mathbb{F}}_n) - \phi(F) = \frac{1}{n}\sum_{i=1}^{n} IF_{\phi,F}(X_i) + R_n.$$

We now have:

$$\mathbb{E}_F(\phi'_F\left(\delta_{X_i} - F\right)) = \phi'_F\left(\int (\delta_x - P)dP(x)\right) = 0,$$

If we assume:

$$\mathrm{Var}_F(\phi'_F\left(\delta_{X_i} - F\right)) = \int \left(IF_{\phi,F}(x)\right)^2 dF(x) < \infty.$$

and the residual term R_n can be controlled somehow (more on this in later classes), a central limit theorem will hold for the (random variable) $\phi'_F(\delta_{X_i} - F)$ and we can expect that:

$$\sqrt{n}\left(\phi(\hat{\mathbb{F}}_n) - \phi(F)\right) \overset{F}{\rightsquigarrow} \mathcal{N}(0, \lambda^2).$$

As is the case in multivariate calculus, the existence of Gateaux (directional) derivative does not ensure that the residual of the approximation is well behaved (differentiability). In the next classes, we will study the residual term more closely.

Coming up next

In the next classes, we will:

- make this heuristic of the functional delta method more precise;
- look more closely at the notion of Hadamard derivative;
- use Hadamard derivative to determine conditions that ensure the functional delta method works

References

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.