Stat210B: Theoretical Statistics **Lecture Date: January 16, 2007**

Lecture 1

Lecturer: Michael I. Jordan *Scribe: Karl Rohe*

Reading: Chapter two of van der Vaart's book *Asymptotic Statistics*.

1 Convergence

There are four types of convergence that we will discuss.

Definition 1. Weak convergence, also known as convergence in distribution or law, is denoted

$$X_n \xrightarrow{d} X$$

A sequence of random variables X_n converges in law to random variable X if $P(X_n \leq x) \to P(X \leq x)$ for all x at which $P(X \leq x)$ is continuous.

Definition 2. X_n is said to **converge in probability** to X if for all $\epsilon > 0$, $P(\ d(X_n, X) > \epsilon) \to 0$. This is denoted $X_n \xrightarrow{P} X$.

Definition 3. X_n is said to **converge in $\mathbf{r^{th}}$ mean** to X if $E(\ d(X_n, X)^r) \to 0$. This is denoted $X_n \xrightarrow{r} X$.

Definition 4. X_n is said to **converge almost surely** to X if $P(\ \lim_n d(X_n, X) = 0) = 1$. This is denoted $X_n \xrightarrow{\text{a.s.}} X$.

Theorem 5.

- *A.s. convergence implies convergence in probability.*
- *Convergence in r^{th} mean also implies convergence in probability.*
- *Convergence in probability implies convergence in law.*
- *$X_n \xrightarrow{d} c$ implies $X_n \xrightarrow{P} c$. Where c is a constant.*

Theorem 6. ***The Continuous Mapping Theorem***

Let g be continuous on a set C where $P(X \in C) = 1$. Then,

1. $X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$
2. $X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X)$
3. $X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$

Example 7. Let $X_n \xrightarrow{d} X$, where $X \sim N(0, 1)$. Define the function $g(x) = x^2$. The CMT says $g(X_n) \xrightarrow{d} g(X)$. But, $X^2 \sim \chi^2_1$. So, $g(X_n) \xrightarrow{d} \chi^2_1$.

Example 8. Let $X_n = \frac{1}{n}$ and $g(x) = \mathbf{1}_{x>0}$. Then $X_n \xrightarrow{d} 0$ and $g(X_n) \xrightarrow{d} 1$. But, $g(0) \neq 1$.

Theorem 9. ***Slutsky's Theorems***

1. $X_n \xrightarrow{d} X$ *and* $X_n - Y_n \xrightarrow{P} 0$ *together imply* $Y_n \xrightarrow{d} X$.
2. $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{P} c$ *together imply*
$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$$
3. $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{P} c$ *together imply* $X_n + Y_n \xrightarrow{d} X + c$.
4. $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{P} c$ *together imply* $X_n Y_n \xrightarrow{d} Xc$.
5. $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{P} c$ *together imply* $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$ *when* $c \neq 0$.

Example 10. Let X_n be iid with mean μ and variance σ^2. From the Weak Law of Large Numbers we know the sample mean $\bar{X}_n \xrightarrow{P} \mu$. Similarly, $\frac{1}{n}\sum_i X_i^2 \xrightarrow{P} E(X^2)$. By Slutsky's Theorem we know $S_n^2 = \frac{1}{n}\sum_i X_i^2 - \bar{X}^2 \xrightarrow{d} \sigma^2$. Together with the CMT, this implies $S_n \xrightarrow{P} \sigma$. From the CLT $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} N(0,1)$. Together these facts imply

$$t = \sqrt{n-1}\frac{\bar{X}_n - \mu}{S_n} = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}\frac{\sigma}{S_n}\sqrt{\frac{n-1}{n}} \xrightarrow{d} N(0,1)$$

Where this last equality is due to Slutsky. So, the t-statistic is asymptotically normal.

Definition 11. $X_n = o_p(R_n)$, pronounced "X_n is little oh-pee-R_n," means $X_n = Y_n R_n$, where $Y_n \xrightarrow{P} 0$.

Definition 12. $X_n = O_p(R_n)$, pronounced "X_n is big oh-pee-R_n," means $X_n = Y_n R_n$, where $Y_n = O_p(1)$. $O_p(1)$ denotes a sequence Z_n which for any $\epsilon > 0$ there exists an M such that $P(|Z_n| > M) < \epsilon$.

Lemma 13. *Let* $R : \mathbb{R}^k \to \mathbb{R}$ *and* $R(0) = 0$. *Let* $X_n = o_p(1)$. *Then, as* $h \to 0$, *for all* $p > 0$

1. $R(h) = o(\|h\|^p)$ *implies* $R(X_n) = o_p(\|X_n\|^p)$.
2. $R(h) = O(\|h\|^p)$ *implies* $R(X_n) = O_p(\|X_n\|^p)$.

To prove this, apply the CMT to $\frac{R(h)}{\|h\|^p}$.

- Any random variable is **tight**. I.e. for all $\epsilon > 0$, there exists and M such that $P(\|X\| > M) < \epsilon$.
- $\{X_\alpha : \alpha \in A\}$ is called **Uniformly Tight** (UT) if for all $\epsilon > 0$, there exists and M such that $\sup_\alpha P(\|X_\alpha\| > M) < \epsilon$.

Theorem 14. ***Prohorov's theorem (cf. Heine-Borel)***

1. *If* $X_n \xrightarrow{d} X$, *then* X_n *is UT.*
2. *If* $\{X_n\}$ *is UT, then there exits a subsequence* $\{X_{nj}\}$ *with* $X_{nj} \xrightarrow{d} X$ *as* $j \to \infty$ *for some* X.

As we move on in the course we will wish to describe weak convergence for things other than random variables. At this point, the our previous definition will not make sense. We can then use this following theorem as a definition.

Theorem 15. ***Portmanteau***

$$X_n \xrightarrow{d} X \iff Ef(X_n) \to Ef(X) \text{ for all bounded continuous } f.$$

In this theorem, "bounded and continuous" can be replaced with

- "continuous and vanishes outside of compacta"
- "bounded and measurable, such that $P(X \in C(g)) = 1$" where $C(g)$ is the set of g's continuity points.
- "bounded Lipshitz"
- "$f(X) = e^{itX}$." This is the next theorem.

Theorem 16. ***Continuity theorem***

$$X_n \xrightarrow{d} X \iff E\exp(it^T X_n) \to E\exp(it^T X)$$

Example 17. To demonstrate why f must be bounded, observe what happens if $g(x) = x$ and

$$X_n = \begin{cases} n & \text{w.p. } 1/n \\ 0 & \text{otherwise} \end{cases}$$

$X_n \xrightarrow{d} 0$, $Eg(X_n) = 1 \neq Eg(0) = 0$.

Example 18. To demonstrate why f must be continuous, observe what happens if $X_n = 1/n$ and

$$g(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Theorem 19. ***(Scheffè)*** *For random variables, $X_n \geq 0$, if $X_n \xrightarrow{a.s.} X$ and $EX_n \to EX < \infty$, then $E|X_n - X| \to 0$. For densities, if $f_n(x) \to g(x)$ for all x, then $\int |f_n(x) - g(x)|dx \to 0$.*

Stat210B: Theoretical Statistics **Lecture Date: January 18, 2007**

Lecture 2

Lecturer: Michael I. Jordan *Scribe: Ariel Kleiner*

Lemma 1 (Fatou). *If* $X_n \xrightarrow{a.s.} X$ *and* $X_n \geq Y$ *with* $E[|Y|] < \infty$*, then*

$$\liminf_{n \longrightarrow \infty} E[X_n] \geq E[X].$$

Theorem 2 (Monotone Convergence Theorem). *If* $0 \leq X_1 \leq X_2 \cdots$ *and* $X_n \xrightarrow{a.s.} X$*, then*

$$E[X_n] \longrightarrow E[X].$$

Note that the Monotone Convergence Theorem can be proven from Fatou's Lemma.

Theorem 3 (Dominated Convergence Theorem). *If* $X_n \xrightarrow{a.s.} X$ *and* $|X_n| \leq Y, E[|Y|] < \infty$*, then*

$$E[X_n] \longrightarrow E[X].$$

Theorem 4 (Weak Law of Large Numbers). *If* $X_i \overset{i.i.d.}{\sim} X$ *and* $E[|X|] < \infty$*, then*

$$\bar{X}_n \xrightarrow{P} E[X],$$

where $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Theorem 5 (Strong Law of Large Numbers). *If* $X_i \overset{i.i.d.}{\sim} X$ *and* $E[|X|] < \infty$*, then*

$$\bar{X}_n \xrightarrow{a.s.} E[X].$$

Definition 6 (Empirical Distribution Function). Given n i.i.d. data points $X_i \overset{\text{i.i.d.}}{\sim} F$, the empirical distribution function is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{[X_i, \infty)}(x).$$

Note that $F_n(x) \xrightarrow{\text{a.s.}} F(x)$, for each x.

Theorem 7 (Glivenko-Cantelli). *Given* n *i.i.d. data points* $X_i \overset{i.i.d.}{\sim} F$*,*

$$P\{\sup_x |F_n(x) - F(x)| \longrightarrow 0\} = 1$$

That is, the random variable $\sup_x |F_n(x) - F(x)|$ converges to 0, almost surely.

Theorem 8 (Central Limit Thorem). *Given* n *i.i.d. random variables* X_i *from some distribution with mean* μ *and covariance* Σ *(which are assumed to exist),*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma).$$

The following theorem is a generalization of the Central Limit Theorem. It applies to non-i.i.d. (i.e., independent but not identically distributed) random variables as might be arranged in a triangular array as follows, where the random variables within each row are independent:

$$\begin{array}{lll} Y_{11} & & \\ Y_{21} & Y_{22} & \\ Y_{31} & Y_{32} & Y_{33} \\ \vdots & & \end{array}$$

Theorem 9 (Lindeberg-Feller). *For each n, let $Y_{n1}, Y_{n2}, \ldots, Y_{nk_n}$ be independent random variables with finite variance such that $\sum_{i=1}^{k_n} Var(Y_{ni}) \to \Sigma$ and*

$$\sum_{i=1}^{k_n} E\left[\|Y_{ni}\|^2 \mathbb{1}\{\|Y_{ni}\| > \varepsilon\}\right] \overset{n\to\infty}{\longrightarrow} 0, \qquad \forall \varepsilon > 0.$$

Then,

$$\sum_{i=1}^{k_n} (Y_{ni} - E[Y_{ni}]) \xrightarrow{d} N(0, \Sigma).$$

We now consider an example illustrating application of the Lindeberg-Feller theorem.

Example 10 (Permutation Tests). Consider $2n$ paired experimental units in which we observe the results of n treatment experiments X_{nj} and n control experiments W_{nj}. Let $Z_{nj} = X_{nj} - W_{nj}$. We would like to determine whether or not the treatment has had any effect. That is, are the Z_{nj} significantly non-zero? To test this, we condition on $|Z_{nj}|$. This conditioning effectively causes us to discard information regarding the magnitude of Z_{nj} and leaves us to consider only signs. Thus, under the null hypothesis H_0, there are 2^n possible outcomes, all equally probable. We now consider the test statistic

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_{ni}$$

and show that, under H_0,

$$\frac{\sqrt{n}\bar{Z}_n}{\sigma_n} \xrightarrow{d} N(0,1),$$

where $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} Z_{ni}^2$, and we assume that

$$\max_j \frac{Z_{nj}^2}{\sum_i Z_{ni}^2} \longrightarrow 0.$$

Proof. Let

$$Y_{nj} = \frac{Z_{nj}}{\left(\sum_i Z_{ni}^2\right)^{1/2}}.$$

Note that, under H_0, $E[Y_{nj}] = 0$ because H_0 states that X_j and Y_j are identically distributed. Additionally,

we have $\sum_j \text{Var}(Y_{nj}) = 1$. Now observe that, $\forall \varepsilon > 0$,

$$\begin{aligned}
\sum_j E\left[|Y_{nj}|^2 \mathbb{1}\{|Y_{nj}| > \varepsilon\}\right] &= \sum_j \frac{Z_{nj}^2}{\sum_i Z_{ni}^2} \mathbb{1}\left\{\frac{Z_{nj}^2}{\sum_i Z_{ni}^2} > \varepsilon^2\right\} \\
&\leq \left(\sum_j \frac{Z_{nj}^2}{\sum_i Z_{ni}^2}\right) \mathbb{1}\left\{\max_j \frac{Z_{nj}^2}{\sum_i Z_{ni}^2} > \varepsilon^2\right\} \\
&= \mathbb{1}\left\{\max_j \frac{Z_{nj}^2}{\sum_i Z_{ni}^2} > \varepsilon^2\right\} \\
&\to 0
\end{aligned}$$

where the equality in the first line follows from the definition of Y_{nj} and the fact that we are conditioning on the magnitudes of the Z_{nj}, thus rendering Z_{nj}^2 deterministic. The desired result now follows from application of the Lindeberg-Feller theorem. □

We now move on to Chapter 3 in van der Vaart.

Theorem 11 (Delta Method, van der Vaart Theorem 3.1). *Let $\phi : D_\phi \subseteq \mathbb{R}^k \to \mathbb{R}^m$, differentiable at θ. Additionally, let T_n be random variables whose ranges lie in D_ϕ, and let $r_n \to \infty$. Then, given that $r_n(T_n - \theta) \xrightarrow{d} T$,*

(i) $r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \phi'_\theta(T)$

(ii) $r_n(\phi(T_n) - \phi(\theta)) - \phi'_\theta(r_n(T_n - \theta)) \xrightarrow{P} 0$

Proof. Given that $r_n(T_n - \theta) \xrightarrow{d} T$, it follows from Prohorov's Theorem that $r_n(T_n - \theta)$ is uniformly tight (UT). Differentiability implies that

$$\phi(\theta + h) - \phi(\theta) - \phi'_\theta(h) = o(\|h\|)$$

(from the definition of the derivative). Now consider $h = T_n - \theta$ and note that $T_n - \theta \xrightarrow{P} 0$ by UT and $r_n \to \infty$. By Lemma 2.12 in van der Vaart, it follows that

$$\phi(T_n) - \phi(\theta) - \phi'_\theta(T_n - \theta) = o_P(\|T_n - \theta\|).$$

Multiplying through by r_n, we have

$$r_n(\phi(T_n) - \phi(\theta) - \phi'_\theta(T_n - \theta)) = o_P(1),$$

thus proving (ii) above. Slutsky now implies that $r_n\phi'_\theta(T_n - \theta)$ and $r_n(\phi(T_n) - \phi(\theta))$ have the same weak limit. As a result, using the fact that ϕ'_θ is a linear operator and the Continuous Mapping Theorem, we have

$$r_n\phi'_\theta(T_n - \theta) = \phi'_\theta(r_n(T_n - \theta)) \xrightarrow{d} \phi'_\theta(T)$$

and so

$$r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \phi'_\theta(T).$$

□

We now jump ahead to U-statistics.

Definition 12 (U-Statistics). For $\{X_i\}$ i.i.d. and a symmetric kernel function $h(X_1, \ldots, X_r)$, a U-statistic is defined as

$$U = \frac{1}{\binom{n}{r}} \sum_{\beta} h(X_{\beta_1}, \ldots, X_{\beta_r})$$

where β ranges over all subesets of size r chosen from $\{1, \ldots, n\}$.

Note that, by definition, U is an unbiased estimator of $\theta = E[h(X_1, \ldots, X_r)]$ (i.e., $E[U] = \theta$).

Example 13. Consider

$$\theta(F) = E[X] = \int x dF(x).$$

Taking $h(x) = x$,

$$U = \frac{1}{n} \sum_{i} X_i.$$

As an exercise, consider

$$\theta(F) = \int (x - \mu)^2 dF(x)$$

and identify h for the corresponding U-statistic, where $\mu = \int x dF(x)$.

Stat210B: Theoretical Statistics **Lecture Date: January 23, 2007**

Lecture 3

Lecturer: Michael I. Jordan *Scribe: Sahand Negahban*

1 U-statistics

U-statistics are a useful tool. The beauty of the U-statistics framework is that by abstracting away some details, can have a general representation of various meaningful quantities. The theory of U-statistics was initially developed by Hoeffding, one of the pioneers of non-parametric statistics.

Definition 1 (U-statistic). Let $X_i \overset{\text{i.i.d.}}{\sim} F$, $h(x_1, x_2, \ldots, x_r)$ be a symmetric kernel function, and $\theta(F) = E[h(X_1, X_2, \ldots, X_r)]$. A U-statistic U_n is defined as

$$U_n = \frac{1}{\binom{n}{r}} \sum_{\beta} h(X_1, X_2, \ldots, X_r) \tag{1}$$

where β ranges over all subsets of size r chosen from $\{1, 2, \ldots, n\}$. $E[U_n] = \theta(F)$ (i.e. U-statistics are unbiased).

Example 2 (Sample Variance). Let $\theta(F) = \sigma^2 = \int (X - \mu)^2 dF$ where $\mu = \int x dF(x)$.

$$\begin{aligned}
\theta(F) &\overset{(a)}{=} \int \left(x_1 - \int x_2 dF(x_2) \right)^2 dF(x_1) \\
&= \int x_1^2 dF(x_1) - 2 \int x_1 dF(x_1) \int x_2 dF(x_2) + \left(\int x_2 dF(x_2) \right)^2 \\
&= \int x_1^2 dF(x_1) - \left(\int x_2 dF(x_2) \right)^2 \\
&= \frac{1}{2} \int x_1^2 dF(x_1) + \frac{1}{2} \int x_2^2 dF(x_2) - \int x_1 x_2 dF(x_1) dF(x_2) \\
&= \frac{1}{2} \int (x_1 - x_2)^2 dF(x_1) dF(x_2) \\
&\Rightarrow h(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2.
\end{aligned}$$

Where (a) follows by expanding μ to $\int x_2 dF(x_2)$. Thus, the U-statistic for the variance is

$$\begin{aligned}
U_n &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \frac{1}{2}(X_i - X_j)^2 \\
&= \frac{1}{n(n-1)} \frac{1}{2} \sum_i \sum_j (X_i - X_j)^2
\end{aligned}$$

where the last equality follows because taking the sum over all indices results in double counting the $(X_i - X_j)^2$ terms. Continuing the simplification shows that

$$\begin{aligned} U_n &= \frac{1}{2n(n-1)} \sum_i \sum_j \left[(X_i - \bar{X}) - (X_j - \bar{X})\right]^2 \\ &= \frac{1}{2n(n-1)} \sum_i \sum_j (X_i - \bar{X})^2 + (X_j - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_i (X_i - \bar{X})^2 \\ &= s_n^2. \end{aligned}$$

Thus, s_n^2 is the U-statistic for the variance of a set of samples. Unbiasedness of this statistic follows immediately from the unbiasedness of U-statistics.

1.1 Novel U-statistics

Example 3 (Gini's mean difference).

$$\theta(F) = \int |x_1 - x_2| dF(x_1) dF(x_2) \tag{2}$$

and the corresponding U-statistic is

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} |X_i - X_j|. \tag{3}$$

Example 4 (Quantile statistic.).

$$\theta(F) = \int_{-\infty}^{t} dF(x). \tag{4}$$

$$U_n = \frac{1}{n} \sum 1_{X_i \le t} = F_n(t) \tag{5}$$

where

$$h(x) = 1_{x \le t}. \tag{6}$$

Example 5 (Signed rank statistic). The following statistic can be used in testing whether the location of the samples is 0.

$$\theta(F) = P(X_1 + X_2 > 0) \tag{7}$$

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} 1_{X_1 + X_2 > 0} \tag{8}$$

where

$$h(x_1, x_2) = 1_{x_1 + x_2 > 0} \tag{9}$$

Definition 6 (Two-sample U-statistics). Given $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ define

$$U_n = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_\alpha \sum_\beta h(\overbrace{\underbrace{X_{\alpha_1}, \ldots, X_{\alpha_r}}_{\text{symmetric}}, \underbrace{Y_{\beta_1}, \ldots, Y_{\beta_s}}_{\text{symmetric}}}^{\text{not symmetric}}) \tag{10}$$

where $h(\cdot, \cdot)$ is symmetric in $x_1, \ldots, x_r$ and $y_1, \ldots, y_s$, but not across both sets of inputs.

Example 7 (Mann-Whitney statistic). This statistic is "used to test for a difference in location between the two samples" (van der Vaart, 1998).

$$U_n = \frac{1}{n_1 n_2} \sum_i \sum_j 1_{X_i \leq Y_j} \tag{11}$$

1.2 Variance of U-statistics

The analysis was first done by Hoeffding.
Assume $E[h] < \infty$ and $X_i \overset{\text{i.i.d}}{\sim} F$. Define $h_c(x_1, \ldots, x_c)$ for $c < r$ as

$$h_c(x_1, \ldots, x_c) \quad = E[h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_r)] \tag{12}$$

Remark 8. The following facts follow from the above definition:

1. $h_0 = \theta(F)$
2. $E[h_c(X_1, \ldots, X_c)] = E[h(X_1, \ldots, X_c, X_{c+1}, \ldots, X_r)] = \theta(F)$.

Let $\widehat{h}_c = h_c - E[h_c] = h_c - \theta(F)$, which follows from remark 8. Thus, $E[\widehat{h}_c] = 0$. Define ζ_c as

$$\zeta_c = \text{Var}(h_c(X_1, \ldots, X_c)) = E[\widehat{h}_c^2(X_1, \ldots, X_c)]. \tag{13}$$

Let $B = \{\beta_1, \ldots, \beta_r\}$ and $B' = \{\beta'_1, \ldots, \beta'_r\}$ be two subsets of $\{1, \ldots, n\}$. Let c be the number of integers in common between each of the sets. Let $S = B \cap B'$, $S_1 = S \setminus B$, and $S_2 = S \setminus B'$, which implies that $|S| = c$, and $|S_1| = |S_2| = r - c$. For some subset $A = \{\alpha_1, \ldots, \alpha_r\}$ of $\{1, \ldots, n\}$ let $X_A = \{X_{\alpha_1}, \ldots, X_{\alpha_r}\}$. Thus,

$$\begin{aligned}
E[\widehat{h}(X_{\beta_1}, \ldots, X_{\beta_r})\widehat{h}(X_{\beta'_1}, \ldots, X_{\beta'_r})] &= E[\widehat{h}(X_B)\widehat{h}(X_{B'})] && (14)\\
&= E[\widehat{h}(X_S, X_{S_1})\widehat{h}(X_S, X_{S_2})] && (15)\\
&= E[E[\widehat{h}(X_S, X_{S_1})\widehat{h}(X_S, X_{S_2})|X_S]] && (16)\\
&= E[\widehat{h}_c^2(X_S)] && (17)\\
&= \zeta_c \ , && (18)
\end{aligned}$$

where the second equality follows because h is a symmetric kernel function and the third and fourth equalities follow from iterated expectations, the fact that each X_i is i.i.d., and the definition of h_c.
Remark 9. The number of distinct choices for two sets having c elements in common is

$$\binom{n}{r}\binom{r}{c}\binom{n-r}{r-c} \tag{19}$$

From the definitions

$$U_n - \theta(F) = \frac{1}{\binom{n}{r}} \sum_\beta \widehat{h}(X_{\beta_1}, \ldots, X_{\beta_r}). \tag{20}$$

Thus,

$$\begin{aligned}
\text{Var}(U_n) &= \binom{n}{r}^{-2} \sum_\beta \sum_{\beta'} E[\widehat{h}(X_{\beta_1}, \ldots, X_{\beta_r})\widehat{h}(X_{\beta'_1}, \ldots, X_{\beta'_r})] && (21a)\\
&= \binom{n}{r}^{-2} \sum_{c=0}^{r} \binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}\zeta_c \quad \text{and } \zeta_c = 0 && (21b)\\
&= \sum_{c=1}^{r} \frac{r!^2}{c!(r-c)!^2} \frac{(n-r)(n-r-1)\cdots(n-2r+c+1)}{n(n-1)\cdots(n-r+1)}\zeta_c, && (21c)
\end{aligned}$$

where the term in the summation corresponding to some c is $O(\frac{1}{n^c})$. Thus,

$$\text{Var}(U_n) = O(\frac{1}{n}) + O(\frac{1}{n^2}) + \cdots + O(\frac{1}{n^r}). \tag{22}$$

Example 10 (Sampling variance of the variance). Let $\theta(F) = \sigma^2$. Thus by example 2

$$h(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2 \text{ and } r = 2. \tag{23}$$

Therefore,

$$\widehat{h}(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2 - \sigma^2,$$

$$h_1(X_1) = \frac{1}{2}(X_1^2 - 2X_1\mu + \sigma^2 + \mu^2),$$

so

$$\widehat{h}_1(X_1) = \frac{1}{2}((X_1 - \mu)^2 - \sigma^2)$$

$$\begin{aligned} E[h^2(X_1, X_2)] &= \frac{1}{4}E[((X_1 - \mu) - (X_2 - \mu))^4] \\ &= \frac{1}{4}\sum_{j=0}^{4}\binom{4}{j}E[(X_1 - \mu)^j]E[(X_2 - \mu)^{4-j}] \\ &= \frac{1}{4}(2\mu_4 + 6\sigma^4), \end{aligned}$$

where the final equality follows because $E[(X - \mu)^4] = \mu_4$, $E[(X - \mu)^2] = \sigma^2$, and $E[(X - \mu)] = 0$. Thus, the following equalities follow:

$$\begin{aligned} \zeta_2 &= E[h^2] - \sigma^4 = \frac{\mu_4}{2} + \frac{\sigma^4}{2} \\ \zeta_1 &= E[\widehat{h}_1^2] = \frac{1}{4}\text{Var}((X_1 - \mu)^2) = \frac{1}{4}(\mu_4 - \sigma^4). \end{aligned}$$

Applying equation 21 yields:

$$\begin{aligned} \text{Var}(s_n^2) &= \binom{n}{2}^{-1}(2(n-2)\zeta_1 + \zeta_2) \\ &= \frac{2}{n(n-1)}\left[2(n-1)\zeta_1 - 2\zeta_1 + \zeta_2\right] \\ &= \frac{4\zeta_1}{n} - \frac{4\zeta_1}{n(n-1)} + \frac{2\zeta_2}{n(n-1)} \\ &= \frac{\mu_4 - \sigma^4}{n} + \frac{2\sigma^4}{n(n-1)} = \frac{\mu_4 - \sigma^4}{n} + O(n^{-2}). \end{aligned}$$

The variance is asymptotically the same as what was found in homework 1. However, using the above method also gives the exact value of all higher order terms.

The variance of U-statistics is known, however the question of whether or not U-statistics are asymptotically normal has yet to be answered. *Hájek projections* will help prove that U-statistics do indeed asymptotically go to Gaussians.

References

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

Stat210B: Theoretical Statistics **Lecture Date: January 25, 2007**

Lecture 4

Lecturer: Michael I. Jordan *Scribe: Mike Higgins*

1 Recap

Define the following:

$$h_c(x_1, \ldots, x_c) = E(h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_r))$$

$$\zeta_c = Var(h_c(X_1, \ldots, X_c))$$

Now consider a U-Statistic:

$$U_n = \frac{1}{\binom{n}{r}} \sum_{\beta} h(X_{\beta_1}, \ldots, X_{\beta_r})$$

where $E(h) = \theta$ and

$$\text{Var}(U_n) = \binom{n}{r}^{-2} \sum_{c=0}^{r} \binom{n}{r}\binom{r}{c}\binom{n-r}{r-c}\zeta_c$$

Note that

$$\text{Var}(U_n) = \frac{r^2\zeta_1}{n} + O(n^{-2})$$

1.1 Rao-Blackwellization

Note that we can write $U_n = E(h(X_1, \ldots, X_r)|X_{(1)}, \ldots, X_{(r)})$. Thus, we have the following inequality:

$$\begin{aligned} E(U_n^2) &= E(Eh(X_1, \ldots, X_r)|X_{(1)}, \ldots, X_{(r)})^2 \\ &\leq E(Eh^2(X_1, \ldots, X_r)|X_{(1)}, \ldots, X_{(r)}) \\ &= h^2 \end{aligned}$$

2 Projections

Define $\mathcal{L}_2(P)$ as the set of functions that are finite when squared, and let T and $\{S : S \in \mathcal{S}\}$ belong to $\mathcal{L}_2(P)$.

Definition 1. $\hat{S} \in \mathcal{S}$ is a **projection** of T on $\mathcal{S}$ if and only if $E((T - \hat{S})S) = 0$ for all $S \in \mathcal{S}$

Corollary 2 (From van der Vaart Chapter 11). $E(T^2) = E(T - \hat{S})^2 + E(\hat{S}^2)$

Now consider a sequence of statistics T_n and spaces $\mathcal{S}_n$ (that contain constant real variables) with projections $\hat{S}_n$.

Theorem 3. If $\frac{\text{Var}(T_n)}{\text{Var}(\hat{S}_n)} \to 1$ then

$$\frac{T_n - E(T_n)}{stdev(T_n)} - \frac{\hat{S}_n - E(\hat{S}_n)}{stdev(\hat{S}_n)} \xrightarrow{P} 0$$

Proof: Let $A_n = \frac{T_n - E(T_n)}{stdev(T_n)} - \frac{\hat{S}_n - E(\hat{S}_n)}{stdev(\hat{S}_n)}$. Note that $E(A_n) = 0$ and

$$\text{Var}(A_n) = 2 - 2\left(\frac{\text{Cov}(T_n, \hat{S}_n)}{stdev(T_n) stdev(\hat{S}_n)}\right)$$

Since $(T_n - \hat{S}_n) \perp \hat{S}_n$ ($(T_n - \hat{S}_n)$ is orthogonal to $\hat{S}_n$), we have:

$$\begin{aligned} E(T_n \hat{S}_n) = E(\hat{S}_n^{\,2}) &\Rightarrow \\ \text{Cov}(T_n, \hat{S}_n) = \text{Var}(\hat{S}_n) &\Rightarrow \\ A_n \overset{r=2}{\to} 0 &\Rightarrow \\ A_n \xrightarrow{P} 0 & \end{aligned}$$

2.1 Conditional Expectations are Projections

$\mathcal{S} \equiv$ linear space of all measurable functions $g(Y)$ of Y. Define $E(X|Y)$ as a measurable function of Y that satisfies $E(X - E(X|Y))g(Y) = 0$. As a consequence, we have the following:

- Setting $g \equiv 1$, then $E(X - E(X|Y)) = 0 \Rightarrow E(X) = E(E(X|Y))$
- $E(f(Y)X|Y) = f(Y)E(X|Y)$ because
  $E\left[f(Y)X - f(Y)E(X|Y)\right] g(Y) = E(X - E(X|Y))f(Y)g(Y) = 0$
- $E(E(X|Y, Z)|Y) = E(X|Y)$

2.2 Hájek Projections

Let $X_1, X_2, \ldots, X_n$ be independent, $\mathcal{S} = \{\sum_{i=1}^n g_i(x_i) : g_i \in \mathcal{L}_2(P)\}$. $\mathcal{S}$ is a Hilbert space.

Lemma 3 (11.10 in van der Vaart). *Let T have a finite 2nd moment. Then*

$$\hat{S} = \sum_{i=1}^n E(T|X_i) - (n-1)E(T)$$

Proof:

$$E(E(T|X_i)|X_j) = \begin{cases} E[E(T|X_i)] = E(T) & \text{if } i \neq j \\ E(T|X_i) & \text{if } i = j \end{cases}$$

$$E(\hat{S}|X_j) = \sum_{i \neq j} E(T) - (n-1)E(T) + E(T|X_j) = E(T|X_j)$$

Thus we have that

$$E[(T - \hat{S})g(X_j)] = E[(E(T - \hat{S})|X_j)g(X_j)] = 0.$$

And we conclude $(T - \hat{S}) \perp \mathcal{S}$.

3 Asymptotic Normality of U-Statistics

Assume $E(h^2) < \infty$. Take Hájek projection of $(U_n - \theta)$ onto $\{\sum_{i=1}^n g_i(x_i) : g_i \in \mathcal{L}_2(P)\}$. Define $\hat{U}_n = \widehat{U_n - \theta} = \sum_{i=1}^n E((U-\theta)|X_i)$. We have that

$$E(h(X_{\beta_1}, \ldots, X_{\beta_r}) - \theta | X_i = x) = \begin{cases} h_1(x) & \text{if } i \in \beta \\ 0 & \text{otherwise} \end{cases}$$

Where $h_1(x) = E(h(x_1, X_2, \ldots, X_r) - \theta)$. Now

$$E(U_n - \theta | X_i) = \frac{1}{\binom{n}{r}} \sum_{\beta} E(h(x_{\beta_1}, \ldots, x_{\beta_r} | X_i) - \theta) = \frac{\binom{n-1}{r-1}}{\binom{n}{r}} = \frac{r}{n} h_1(x_i) \Rightarrow$$

$$\hat{U_n} = \frac{r}{n} \sum_{i=1}^n h_1(x_i)$$

Note that $E\hat{U_n} = 0$ and

$$\text{Var}(\hat{U_n}) = \frac{r^2}{n^2} [n[\text{Var}(h(X_1))]] = \frac{r^2}{n} \zeta_1$$

And so we have $\frac{\text{Var}(U_n)}{\text{Var}(\hat{U_n})} \to 1$. By our previous theorem we have that

$$\frac{U_n - \theta}{(\frac{r^2}{n}\zeta_1 + O(n^{-2}))^{\frac{1}{2}}} - \frac{\hat{U_n}}{(\frac{r^2}{n}\zeta_1)^{\frac{1}{2}}} \xrightarrow{P} 0$$

By Slutsky we have

$$\sqrt{n}(U_n - \theta - \hat{U_n}) \xrightarrow{P} 0$$

By CLT we have

$$\sqrt{n}\hat{U_n} \xrightarrow{d} N(0, r^2\zeta_1)$$

And by Slutsky again we have

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, r^2\zeta_1)$$

Stat210B: Theoretical Statistics **Lecture Date: February 01, 2007**

Lecture 4

Lecturer: Michael I. Jordan *Scribe: Sriram Sankararaman*

Empirical Process theory allows us to prove uniform convergence laws of various kinds. One of the ways to start Empirical Process theory is from the Glivenko-Cantelli theorem. Recall the Glivenko-Cantelli theorem.

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\xi_i \leq t\}} \quad (1)$$

$$F(t) = \mathbb{P}\{\xi \leq t\} \quad (2)$$

We would like to show that $\sup_t |F_n(t) - F(t)| \xrightarrow{P} 0$. The proof makes use of the compactness of the class of indicator functions on the real line to break this class into bins and bound the oscillations in each bin. This leads to the question of whether the same idea can be generalized to other function classes.

1 Empirical Process Theory

Denote $\sup_t |\cdot|$ by $||\cdot||$. To bound the difference $||F_n(t) - F(t)||$, we compare two independent copies of the empirical quantity - $F_n(t)$ and $F'_n(t)$. A symmetrization lemma is used to bound the former in terms of the latter.

1.1 First Symmetrization

Lemma 1. *(Pollard, 1984, Section II.8, p. 14) Let $Z(t)$ and $Z'(t)$ be independent stochastic processes. Suppose that $\exists \alpha, \beta > 0$ such that $\mathbb{P}\{|Z'(t)| \leq \alpha\} \geq \beta, \quad \forall t$. Then*

$$\mathbb{P}\{\sup_t |Z(t)| > \epsilon\} \leq \beta^{-1}\mathbb{P}\{\sup_t |Z(t) - Z'(t)| > \epsilon - \alpha\} \quad (3)$$

An application of Lemma 1 can be seen by setting $Z(t) = F_n(t) - F(t)$ and $Z'(t) = F'_n(t) - F(t)$.

Proof. Suppose that the event $\{\sup_t |Z(t)| > \epsilon\}$ occurs. Choose $\tau \ni |Z(\tau)| > \epsilon$. Note that τ is a random variable. By definition of τ,

$$\mathbb{P}\{\sup_t |Z(t)| > \epsilon\} \leq \mathbb{P}\{|Z(\tau)| > \epsilon\} \quad (4)$$

From the independence of Z and Z', we have

$$\mathbb{P}\{|Z'(t)| < \alpha | Z\} \geq \beta \quad (5)$$

Suppose that both $\{|Z(\tau)| > \epsilon\}$ and $\{|Z'(\tau)| \leq \alpha\}$ occur. Then we have

$$\{|Z(\tau) - Z'(\tau)| \geq \epsilon - \alpha\} \quad (6)$$

Also

$$\begin{aligned}\beta\{|Z(\tau)| > \epsilon\} &\leq \mathbb{P}\{|Z'(\tau)| \leq \alpha, |Z(\tau)| > \epsilon | Z\} \\ &\leq \mathbb{P}\{|Z'(\tau)| > \alpha, |Z(\tau)| > \epsilon\} \qquad (7)\end{aligned}$$

Here $\{|Z(\tau)| > \epsilon\}$ is an indicator function on the event $\{|Z(\tau)| > \epsilon\}$. The inequality 7 uses the independence of Z and Z'. From Equation 6

$$\begin{aligned}\beta\{|Z(\tau)| > \epsilon\} &\leq \mathbb{P}\{|Z'(\tau) - Z(\tau)| \geq \epsilon - \alpha\} \\ &\leq \mathbb{P}\{\sup_t |Z(t) - Z'(t)| \geq \epsilon - \alpha\} \qquad (8)\end{aligned}$$

The proof follows from Equations 8 and 4. □

1.1.1 Example

$$U_n(\omega, t) = n^{-\frac{1}{2}} \sum_{i=1}^{n} (\{\xi_i(\omega) \leq t\} - t)$$

where $\xi_i \overset{iid}{\sim} Unif(0, 1)$.

For fixed value of t,

$$U_n \sim \frac{Bin(n, t)}{n^{\frac{1}{2}}} - \frac{t}{n^{\frac{1}{2}}}$$

$$\begin{aligned}\mathbb{P}\{|F_n(t) - F(t)| > \frac{\epsilon}{2}\} &\leq \frac{4}{\epsilon^2} E(F_n(t) - F(t))^2 \\ &= \frac{4}{\epsilon^2} E(\frac{1}{n} \sum_i \{\xi \leq t\} - F(t))^2 \\ &= \frac{4}{n\epsilon^2} E(\{\xi \leq t\} - F(t))^2 \\ &= \frac{4F(t)(1 - F(t))}{n\epsilon^2} \\ &\leq \frac{1}{n\epsilon^2} \\ &= \frac{1}{2} \quad \text{for} \quad n \geq \frac{2}{\epsilon^2}\end{aligned}$$

1.2 Second Symmetrization

The second symmetrization lemma allows us to replace the difference $F_n - F_n'$ with a single empirical quantity consisting of n observations. We can further bound the latter so that the bound is independent of the data ξ.

Define Rademacher variables $\{\sigma_i\} \overset{iid}{\in} \{-1, +1\}$. For any choice of $\{\sigma_i\}$, the distribution of $(\{\xi_i \leq t\} - \{\xi_i' \leq t\})$ is equal to the distribution of $\sigma_i(\{\xi_i \leq t\} - \{\xi_i' \leq t\})$. We change notation here so that $P_n = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\xi_i \leq t\}}$. P_n' is defined similarly.

Lemma 2. *Pollard (1984, II.8,p. 15)* $\mathbb{P}\{||P_n - P'_n|| > \frac{\epsilon}{2}\} \leq 2\mathbb{P}\{\underset{t}{sup}|\frac{1}{n}\sum_i \sigma_i\{\xi_i \leq t\}| \geq \frac{\epsilon}{4}\}$

Proof.

$$\begin{aligned}
&\mathbb{P}\{||P_n - P'_n|| > \frac{\epsilon}{2}\} \\
&\qquad = \mathbb{P}\{\frac{1}{n}\underset{t}{sup}|\sum_i \sigma_i(\{\xi_i \leq t\} - \{\xi'_i \leq t\})| \geq \frac{\epsilon}{2}\} \\
&\qquad \leq \mathbb{P}\{\underset{t}{sup}|\frac{1}{n}\sum_i \sigma_i\{\xi_i \leq t\}| \geq \frac{\epsilon}{4}\} + \mathbb{P}\{\underset{t}{sup}|\frac{1}{n}\sum_i \sigma_i\{\xi'_i \leq t\}| \geq \frac{\epsilon}{4}\} \qquad (9) \\
&\qquad = 2\mathbb{P}\{\underset{t}{sup}|\frac{1}{n}\sum_i \sigma_i\{\xi_i \leq t\}| \geq \frac{\epsilon}{4}\}
\end{aligned}$$

□

Inequality 9 was derived using the equivalence of the two random quantities and the triangle inequality.

1.3 Hoeffding bound for independent RVs

We state here the Hoeffding bound which we use to bound the quantity $\frac{1}{n}\sum_i \sigma_i\{\xi_i \leq t\}$. Consider n independent RVs $\{Y_i\}$s so that $EY_i = 0$ and $a_i \leq Y_i \leq b_i$.

Theorem 3 (Hoeffding Bound). $\mathbb{P}\{\sum_{i=1}^n Y_i > \eta\} \leq 2e^{-\frac{2\eta^2}{\sum_i (b_i - a_i)^2}}$

The proof proceeds by considering the random variable $e^{s\sum_i Y_i}$ where s is a free parameter. Using Markov's inequality,

$$\begin{aligned}
\mathbb{P}\{e^{s\sum_i Y_i} > e^{s\eta}\} &\leq \frac{Ee^{s\sum_i Y_i}}{e^{s\eta}} \\
&\leq \frac{\prod_i Ee^{sY_i}}{e^{s\eta}}
\end{aligned}$$

Minimizing s gives the necessary bound.

References

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

Stat210B: Theoretical Statistics **Lecture Date: February 6th, 2007**

Uniformly Strong Law of Large Numbers

Lecturer: Michael I. Jordan *Scribe: Jian Ding*

In this lecture, we try to generalize the Glivenko-Cantelli theorem.

Let $\xi_1, \xi_2, \cdots, \xi_n \sim P$ and are i.i.d. sequences. We define $Pf := \mathbb{E}(f(X))$, in which $X \sim P$. We also define $P_n f$ with respect to the empirical measure but puts mass $\frac{1}{n}$ at $\{\xi_1, \cdots, \xi_n\}$. Notice that by definition $P_n f = \frac{1}{n}\sum_{i=1}^n f(\xi_i)$.

We point out that $P_n f - Pf$ is an object of interest; and $\sup_{f\in\mathcal{F}} |P_n f - Pf|$ is of even more interest. For example, let $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}\}$, then $P_n f - Pf$ becomes $F_n(t) - F(t)$ and $\sup_{f\in\mathcal{F}} |\cdot|$ becomes $\sup_t |F_n(t) - F(t)|$. In general, we are interested in statistics defined on a family of stochastic processes with index set $\mathcal{F}$.

Uniform Law of Large Numbers

Define $\|P_n - P\| := \sup_{f\in\mathcal{F}} |P_n f - Pf|$. Recalling the discussion in last lecture, we get

$$\begin{aligned}\mathbb{P}\{\|P_n - P\| > \epsilon\} &\leq 2\mathbb{P}\{\|P_n - P_n'\| > \frac{\epsilon}{2}\} \\ &\leq 4\mathbb{P}\{\|P_n^0\| > \frac{\epsilon}{4}\}\end{aligned}$$

where P_n^0 is a signed measure putting mass $\frac{1}{n}\sigma_i$ at $\{\xi_1, \cdots, \xi_n\}$. Again, σ_i independently pick value uniformly on $\{1, -1\}$.

Specialize $\mathcal{F}$ to indicators

Let $I_j = (-\infty, t_j]$ where $\{t_j\}$ lie between the points ξ_i, i.e., $t_0 < \xi_1 < t_1 < \xi_2 < t_2 < \cdots$. Consider

$$\begin{aligned}&\mathbb{P}\{\|P_n^0\| > \frac{\epsilon}{4}|\xi\} \\ &\qquad = \mathbb{P}\{\bigcup_{j=0}^n \{|P_n^0 I_j| > \frac{\epsilon}{4}\}|\xi\} \\ &\qquad \leq \sum_{j=0}^n \mathbb{P}\{|P_n^0 I_j| > \frac{\epsilon}{4}|\xi\} \\ &\qquad \leq (n+1)\max_j \mathbb{P}\{|P_n^0 I_j| > \frac{\epsilon}{4}|\xi\}.\end{aligned}$$

Recall Hoeffding's inequality. Let Y_i be independent, $\mathbb{E}(Y_i) = 0$, $a_i \leq Y_i \leq b_i$. Then, $\mathbb{P}\{|Y_1 + Y_2 + \cdots + Y_n| > \eta\} \leq \exp\{-\frac{2\eta^2}{\sum_i (b_i - a_i)}\}$. We apply this to $\sigma_i\{\xi_i \leq t\}$, and conclude

$$\begin{aligned}\mathbb{P}\{|P_n^0\{(-\infty, t]\}| > \frac{\epsilon}{4}|\xi\} &\leq 2\exp(-\frac{2(n\epsilon/4)^2}{4n}) \\ &\leq 2\exp(-\frac{n\epsilon^2}{32}),\end{aligned}$$

notice that this is independent of ξ, so $\mathbb{P}\{\|P_n - P\| > \epsilon\} \leq 8(n+1)\exp(-\frac{n\epsilon^2}{32})$, i.e., we get Uniform Law of Large Numbers in probability and also almost surely (by Borel-Cantelli).

The conclusion, namely, Glivenko-Cantelli theorem is not new. However, this method can be generalized to richer class of functions immediately.

VC Classes

Consider a collection $\mathcal{C}$ of subsets of some set $\mathcal{X}$, and consider points $\xi_1, \cdots, \xi_n$ from $\mathcal{X}$. Define $\Delta_n^{\mathcal{C}} := \#\{C \bigcap \{\xi_1, \cdots, \xi_n\} : C \in \mathcal{C}\}$; $m(n) := \max_{\xi_1,\cdots,\xi_n} \Delta_n^{\mathcal{C}}(\xi_1, \cdots, \xi_n)$; $V^{\mathcal{C}} := \min\{n : m(n) < 2^n\}$.

Examples

1, $\mathcal{X} = \mathbb{R}, \mathcal{C} = \{(-\infty, t]\}$. Then, $V^{\mathcal{C}} = 2$.

2, $\mathcal{X} = \mathbb{R}, \mathcal{C} = \{(s, t] : s < t\}$. Then, $V^{\mathcal{C}} = 3$.

3, $\mathcal{X} = \mathbb{R}^d, \mathcal{C} = \{(-\infty, t] : t \in \mathbb{R}^d\}$. Then, $V^{\mathcal{C}} = d + 1$.

4, Rectangles in $\mathbb{R}^d$. $V^{\mathcal{C}} = 2d + 1$.

Sauer's Lemma

Lemma 1.

$$m(n) \leq \sum_{j=0}^{V^{\mathcal{C}}} \binom{n}{j} \leq (\frac{ne}{V^{\mathcal{C}} - 1})^{V^{\mathcal{C}}-1}.$$

Proof. We prove the second part.

$$\begin{aligned}
\sum_{j=0}^{S} \binom{n}{j} &= 2^n \sum_{j=0}^{S} \binom{n}{j} (\frac{1}{2})^n \\
&= 2^n \mathbb{P}(Y \leq S), \quad Y \sim Bin(n, \frac{1}{2}) \\
&\leq 2^n \mathbb{E}(\theta)^{Y-S}, \quad 0 \leq \theta \leq 1 \\
&= 2^n \theta^{-S} (\frac{1}{2} + \frac{\theta}{2})^n, \quad \text{take } \theta = \frac{S}{n} \\
&= (\frac{n}{S})^S (1 + \frac{S}{n})^n \\
&\leq (\frac{n}{S})^S e^S.
\end{aligned}$$

□

This suggests

$$\begin{aligned}
\mathbb{P}\{\|P_n^0\| > \frac{\epsilon}{4}|\xi\} &= \mathbb{P}\{\bigcup_{i=0}^{m(n)} |P_n^0 \widetilde{f}_i| > \frac{\epsilon}{4}|\xi\} \\
&\quad (\widetilde{f}_i \text{ are indicators of subsets that achieve } m(n)). \\
&\leq \sum_{i=1}^{m(n)} \mathbb{P}\{|P_n^0 \widetilde{f}_i| > \frac{\epsilon}{4}|\xi\} \\
&\leq m_n \max(\cdot).
\end{aligned}$$

Then if $\mathcal{F}$ is a VC class, (i.e., $V^{\mathcal{C}} < \infty$), then

$$\mathbb{P}\{\|P_n - P\| > \epsilon\} \leq \text{ (Poly in } n)(\exp(-Cn)).$$

Stat210B: Theoretical Statistics **Lecture Date: February 8, 2007**

Lecture 7

Lecturer: Michael I. Jordan *Scribe: Kurt Miller*

1 Properties of VC-classes

1.1 VC preservation

Let $\mathcal{C}$ and $\mathcal{D}$ be VC-classes (i.e. classes with finite VC-dimension). Then so are

- $\{C^{\complement} : C \in \mathcal{C}\}$
- $\{C \cup D : C \in \mathcal{C}, D \in \mathcal{D}\}$
- $\{C \cap D : C \in \mathcal{C}, D \in \mathcal{D}\}$
- $\phi(C)$ where ϕ is 1-1
- $\{C \times D : C \in \mathcal{C}, D \in \mathcal{D}\}$

1.2 Half spaces

Let $\mathcal{G}$ be a finite-dimensional vector space of functions. Let $\mathcal{C} = \{g \geq 0 : g \in \mathcal{G}\}$ or more formally $\mathcal{C} = \{\{\omega : g(\omega) \geq 0\} : g \in \mathcal{G}\}$. Then $V^{\mathcal{C}} \leq \dim \mathcal{G} + 1$.

1.3 Subgraphs

Definition 1. A *subgraph* of $f : \mathcal{X} \to \mathcal{R}$ is the subset $\mathcal{X} \times \mathcal{R}$ given by $\{(x, t) : t \leq f(x)\}$.

A collection $\mathcal{F}$ is a *VC-subgraph class* if the collection of subgraphs is a VC-class.

2 Covering Number

We now begin to explore a more powerful method of defining complexity than VC-dimension.

2.1 Definitions

Definition 2 (Covering Number). (Pollard, 1984, p. 25) Let Q be a probability measure on S and $\mathcal{F}$ be a class of functions in $\mathcal{L}^1(Q)$, i.e. $\forall f \in \mathcal{F}$, $\int |f| dQ < \infty$. For each $\varepsilon > 0$ define the $\mathcal{L}_1$ covering number $N_1(\varepsilon, Q, \mathcal{F})$ as the smallest value of m for which there exist functions $g_1, \ldots, g_m$ (not necessarily in $\mathcal{F}$) such that $\min_j Q|f - g_j| \leq \varepsilon$ for each f in $\mathcal{F}$. For definiteness set $N_1(\varepsilon, Q, \mathcal{F}) = \infty$ if no such m exists.

Note that the set $\{g_j\}$ that achieves this minimum is not necessarily unique.

Definition 3 (Metric Entropy). Define $H_1(\varepsilon, Q, \mathcal{F}) = \log N_1(\varepsilon, Q, \mathcal{F})$ as the $\mathcal{L}_1$ metric entropy of $\mathcal{F}$.

More generally, $H_p(\varepsilon, Q, \mathcal{F})$ uses the $\mathcal{L}_p(Q)$ norm. Write this as $\|g\|_{p,Q} = (\int |g|^p dQ)^{1/p}$.

Definition 4 (Totally bounded). A class is called *totally bounded* if $\forall \varepsilon,\ H_p(\varepsilon, Q, \mathcal{F}) < \infty$

Another kind of entropy:

Definition 5 (Entropy with bracketing). Let $N_{p,B}(\varepsilon, Q, \mathcal{F})$ be the smallest value of m for which there exist pairs of functions $\{(g_j^L, g_j^U)\}_{j=1}^m$ such that $\forall j$, $\|g_j^U - g_j^L\|_{p,Q} < \varepsilon$ and $\forall f \in \mathcal{F}, \exists j(f)$ s.t. $g_{j(f)}^L \le f \le g_{j(f)}^U$. Then we define the *entropy with bracketing* as $H_{p,Q}(\varepsilon, Q, \mathcal{F}) = \log N_{p,Q}(\varepsilon, Q, \mathcal{F})$.

Finally, using $\|g\|_\infty \triangleq \sup_{x \in \mathcal{X}} |g(x)|$, let $N_\infty(\varepsilon, \mathcal{F})$ be the smallest m such that there exists a set $\{g_j\}_{j=1}^m$ such that $\sup_{f \in \mathcal{F}} \min_{j=1,\ldots,m} \|f - g_j\|_\infty < \varepsilon$. Then $H_\infty(\varepsilon, \mathcal{F}) = \log N_\infty(\varepsilon, \mathcal{F})$.

2.2 Relationship of the various entropies

Using the definitions above, we have that

1. $H_1(\varepsilon, Q, \mathcal{F}) \le H_{p,B}(\varepsilon, Q, \mathcal{F}),\ \forall \varepsilon > 0$
2. $H_{p,B}(\varepsilon, Q, \mathcal{F}) \le H_\infty(\varepsilon/2, \mathcal{F}),\ \forall \varepsilon > 0$

Can these quantities be computed for normal classes of functions? Yes, but you would generally look them up in a big book. We'll look at how to compute one of these quantities here.

2.3 Examples

Example 6. Let $\mathcal{F} = \{f : [0,1] \to [0,1],\ |f'| \le 1\}$ (i.e. functions from $[0,1]$ to $[0,1]$ with first derivatives bounded by 1). Then $H_\infty(\varepsilon, \mathcal{F}) \le A\frac{1}{\varepsilon}$ where A is a constant that we will compute.

Proof. Let $0 = a_0 < a_1 < \cdots < a_m = 1$ where $a_k = k\varepsilon$ and $k = 0, \ldots, m$. Let $B_1 = [a_0, a_1]$ and $B_k = (a_{k-1}, a_k]$. For each $f \in \mathcal{F}$, define

$$\tilde{f} = \sum_{k=1}^{m} \varepsilon \left\lfloor \frac{f(a_k)}{\varepsilon} \right\rfloor 1_{B_k}$$

$\tilde{f}$ takes on values in εk where k is an integer. We also have $\|\tilde{f} - f\|_\infty \le 2\varepsilon$, because $|\tilde{f}(a_{k-1}) - f(a_{k-1})| \le \varepsilon$ by construction and $|f(a_k)^- - f(a_{k-1})| \le \varepsilon$ since f' is bounded by 1.

We now count the number of possible $\tilde{f}$ obtained by this construction. At a_0, there are $\lfloor 1/\varepsilon \rfloor + 1$ choices for $\tilde{f}(a_0)$ since $\tilde{f}$ only takes on values of εk in $[0,1]$. Furthermore, combining previous results gives us

$$\begin{aligned} |\tilde{f}(a_k) - \tilde{f}(a_{k-1})| &\le |\tilde{f}(a_k) - f(a_k)| + |f(a_k) - f(a_{k-1})| + |f(a_{k-1}) - \tilde{f}(a_{k-1})| \\ &\le 3\varepsilon. \end{aligned}$$

Therefore, having chosen $\tilde{f}(a_{k-1})$, $\tilde{f}$ can take on at most 7 distinct values at a_k. Therefore

$$N_\infty(2\varepsilon, \mathcal{F}) \leq \left(\left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1 \right) 7^{\lfloor 1/\varepsilon \rfloor}$$

which gives us that

$$H_\infty(2\varepsilon, \mathcal{F}) \leq \frac{1}{\varepsilon} \log 7 + \log(\lfloor 1/\varepsilon \rfloor + 1)$$

so our constant can be chosen as any constant that $> \log 7$. □

A seminal paper in this field is by Birman and Solomjak in 1967. They present other examples of metric entropy calculations, including:

Example 7. Let $\mathcal{F} = \{f : [0,1] \to [0,1] : \int (f^{(m)}(x))^2 dx \leq 1\}$. Then $H_\infty(\varepsilon, \mathcal{F}) \leq A\varepsilon^{-1/m}$.

Example 8. Let $\mathcal{F} = \{f : \mathcal{R} \to (0,1) : f \text{ is increasing}\}$. Then $H_{p,B}(\varepsilon, Q, \mathcal{F}) \leq A\frac{1}{\varepsilon}$.

Example 9. Let $\mathcal{F} = \{f : \mathcal{R} \to [0,1] : \int |df| \leq 1\}$, the class of bounded variation. Then $H_{p,B}(\varepsilon, Q, \mathcal{F}) \leq A\frac{1}{\varepsilon}$.

Lemma 10 (Ball covering lemma). *A ball $B_d(R)$ in $\mathcal{R}^d$ of radius R can be covered by*

$$\left(\frac{4R + \varepsilon}{\varepsilon} \right)^d$$

balls of radius ε.

Proof. Let $\{c_j\}_{j=1}^m$ be a packing of size ε (Euclidean norm). This implies that balls of radius ε with centers at $\{c_j\}$ cover $B_d(R)$ (otherwise we could add more points c_j to the packing). Let B_j be the ball of radius $\varepsilon/4$ centered at c_j. We must have that $B_i \cap B_j$ is empty for $i \neq j$. Therefore $\{B_j\}$ are disjoint and

$$\cup_j B_j \subset B_d(R + \varepsilon/4).$$

A ball of radius ρ has volume $C_d\rho^d$ where C_d is a constant that depends on the dimension d. Therefore, the volume of the union $\cup_j B_j$ is $MC_d(\varepsilon/4)^d$ and since it is a subset of $B_d(R + \varepsilon/4)$, we have

$$MC_d \left(\frac{\varepsilon}{4} \right)^d \leq C_d \left(R + \frac{\varepsilon}{4} \right)^d.$$

With a simple manipulation of this equation, we get that

$$M \leq \left(\frac{4R + \varepsilon}{\varepsilon} \right)^d$$

□

References

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

Stat210B: Theoretical Statistics **Lecture Date: February 13, 2007**

P-Glivenko-Cantelli

Lecturer: Michael I. Jordan *Scribe: Christopher Hundt*

1 P-Glivenko-Cantelli

Definition 1 (P-Glivenko-Cantelli)**.** A class $\mathcal{F}$ is P-Glivenko-Cantelli if

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| \xrightarrow{\text{a.s.}} 0.$$

Definition 2 (envelope)**.** An envelope for a class $\mathcal{F}$ of functions is a function F such that $PF < \infty$ and, for all $f \in \mathcal{F}$, $|f| \leq F$.

Theorem 3. *(Pollard, 1984, Theorem 24) Let $\mathcal{F}$ be a permissible[1] class of functions with envelope F. If $\frac{1}{n}H_1(\varepsilon, P_n, \mathcal{F}) \xrightarrow{P} 0$ for all $\varepsilon > 0$ then $\|P_n - P\| \triangleq \sup_{f\in\mathcal{F}} |P_n f - Pf| \xrightarrow{a.s.} 0$.*

Remark 4. The condition that $\frac{1}{n}H_1(\varepsilon, P_n, \mathcal{F}) \xrightarrow{P} 0$ is natural in the sense that we want to make sure that the covering number does not grow exponentially fast. See Pollard (1984) for more discussion of this theorem and its conditions.

Proof. In lectures 5 and 6, we proved Glivenko-Cantelli for a special class of functions, namely indicators. This proof extends it to more general classes of functions. The proof will be similar, but some changes will need to be made.

As before, we will prove convergence in probability. A reverse-martingale argument can be used to extend the proof to show convergence almost surely.

Since $PF < \infty$, for any $\varepsilon > 0$ there exists a K such that $PF\{F > K\} < \varepsilon$. It follows that

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| \leq \sup_{f\in\mathcal{F}} |P_n f\{F \leq K\} - Pf\{F \leq K\}| + \sup_{f\in\mathcal{F}} |P_n f\{F > K\}| + \sup_{f\in\mathcal{F}} |Pf\{F > K\}|. \quad (1)$$

Furthermore, since F is an envelope,

$$\sup_{f\in\mathcal{F}} |P_n f\{F > K\}| + \sup_{f\in\mathcal{F}} |Pf\{F > K\}| \leq P_n F\{F > K\} + PF\{F > K\} \xrightarrow{\text{a.s.}} 2PF\{F > K\} < 2\varepsilon.$$

Since this is true for all ε, inequality (1) means that

$$\sup_{f\in\mathcal{F}} |P_n f\{F \leq K\} - Pf\{F \leq K\}| \xrightarrow{P} 0 \implies \sup_{f\in\mathcal{F}} |P_n f - Pf| \xrightarrow{P} 0.$$

This tells us that we can proceed under the assumption that $|f| \leq K$ for all $f \in \mathcal{F}$.

[1] Permissibility is a concept from measure theory that is not important for this class; see Pollard (1984, Appendix C, Definition 1) for details.

Now lecture 5 used two symmetrization arguments to establish bounds that helped in proving Glivenko-Cantelli for indicator functions. Both these bounds apply in this more general case, and the proofs are similar, so we will repeat only the conclusion from lecture 6, which is

$$P\{\|P_n - P\| > \varepsilon\} \le 4P\{\|P_n^0\| > \tfrac{\varepsilon}{4}\} \text{ for } n \ge \frac{2}{\varepsilon^2},$$

where P_n^0 is the signed measure putting mass $\pm\frac{1}{n}$ at each of the observed data points $\xi = \{\xi_1, \ldots, \xi_n\}$. We will now continue, working conditionally with ξ.

Given any ξ, choose $g_1, \ldots, g_M$, where $M = N_1(\frac{\varepsilon}{8}, P_n, \mathcal{F})$ such that $\min_j P_n|f - g_j| < \frac{\varepsilon}{8}$ for all $f \in \mathcal{F}$. Denote f^* as the g_j that achieves the minimal P_n-norm distance from f. Now

$$\begin{aligned}
P\{\|P_n^0\| > \tfrac{\varepsilon}{4}|\xi\} &\le P\{\sup_{f\in\mathcal{F}}(|P_n^0 f^*| + |P_n^0(f - f^*)|) > \tfrac{\varepsilon}{4}|\xi\} \\
&\le P\{\sup_{f\in\mathcal{F}}(|P_n^0 f^*| + P_n|f - f^*|) > \tfrac{\varepsilon}{4}|\xi\} \\
&\le P\{\max_j |P_n^0 g_j| > \tfrac{\varepsilon}{8}|\xi\} && \text{since } P_n|f - f^*| < \tfrac{\varepsilon}{8} \\
&= P\{\bigcup_{j=1}^M |P_n^0 g_j| > \tfrac{\varepsilon}{8}|\xi\} \\
&\le \sum_{j=1}^M P\{|P_n^0 g_j| > \tfrac{\varepsilon}{8}|\xi\} \\
&\le N_1(\tfrac{\varepsilon}{8}, P_n, \mathcal{F}) \max_j P\{|P_n^0 g_j| > \tfrac{\varepsilon}{8}|\xi\} \\
&\le N_1(\tfrac{\varepsilon}{8}, P_n, \mathcal{F}) \max_j 2\exp\left(-2\frac{\left(\frac{n\varepsilon}{8}\right)^2}{\sum_{i=1}^n (2g_j(\xi_i))^2}\right) && \text{by Hoeffding} \\
&\le 2N_1(\tfrac{\varepsilon}{8}, P_n, \mathcal{F}) \exp\left(-\frac{n\varepsilon^2}{128K^2}\right) && \text{since } |g_j| \le K
\end{aligned}$$

Note that this bound does not depend on the data!

To complete the proof we must integrate over ξ: for the event $\{\log N_1(\frac{\varepsilon}{8}, P_n, \mathcal{F}) \le \frac{n\varepsilon^2}{256K^2}\}$ we can use the bound just obtained, replacing $N_1(\frac{\varepsilon}{8}, P_n, \mathcal{F})$ with the upper bound $e^{n\varepsilon^2/256K^2}$. Otherwise, we will use 1 as a bound. That is,

$$\begin{aligned}
P\{\|P_n^0\| > \tfrac{\varepsilon}{4}\} &\le P\{\log N_1(\tfrac{\varepsilon}{8}, P_n, \mathcal{F}) \le \tfrac{n\varepsilon^2}{256K^2}\} 2\exp\left(-\frac{n\varepsilon^2}{256K^2}\right) + P\{\log N_1(\tfrac{\varepsilon}{8}, P_n, \mathcal{F}) > \tfrac{n\varepsilon^2}{256K^2}\} \\
&\le \underbrace{2\exp\left(-\frac{n\varepsilon^2}{256K^2}\right)}_{\to 0} + \underbrace{P\{\log N_1(\tfrac{\varepsilon}{8}, P_n, \mathcal{F}) > \tfrac{n\varepsilon^2}{256K^2}\}}_{\xrightarrow{P} 0}.
\end{aligned}$$

□

Example 5 (A non-GC class)**.** Suppose $\mathcal{F} = \{1_A : A \subset \mathbf{R}\}$, $P = U(0,1)$, and $\mathcal{X} = (0,1)$. Consider $A = \{x_1, \ldots, x_n\}$. Then $P1_A \equiv 0$, but $P_n 1_A = 1$ for some subsets.

2 Glivenko-Cantelli and VC dimension

Lemma 6 (Approximation Lemma)**.** *(Pollard, 1984, Lemma 25) Let $\mathcal{F}$ be a class of functions with envelope F and let $\mathcal{Q}$ be a probability measure such that $\mathcal{Q}F < \infty$. Suppose graphs of $\mathcal{F}$ have finite VC dimension $\mathcal{V}$. Then*

$$N_1(\varepsilon \mathcal{Q}F, \mathcal{Q}, \mathcal{F}) \leq A\mathcal{V}(16e)^{\mathcal{V}}\varepsilon^{-(\mathcal{V}-1)}.$$

Remark 7. The exponential dependence of N_1 on $\mathcal{V}$ shown in this lemma gives an intuition for the use of the word "dimension" in *VC dimension*.

Remark 8. This lemma implies that $H_1 \leq C + (\mathcal{V} - 1)\log\frac{1}{\varepsilon}$.

Remark 9. See van der Vaart (1998, Lemma 19.15) for a tighter result.

References

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

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

Stat210B: Theoretical Statistics **Lecture Date: February 15, 2007**

Applications of ULLNs: Consistency of M-estimators

Lecturer: Michael I. Jordan *Scribe: Blaine Nelson*

1 M and Z-estimators (van der Vaart, 1998, Section 5.1, p. 41–54)

Definition 1 (**M-estimator).** An estimator $\hat{\theta}_n$ defined as a maximizer of the expression:

$$M_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i) \tag{1}$$

for some function $m_\theta(\cdot)$. If there is a unique solution, the estimator can be expressed simply as

$$\hat{\theta}_n = \text{argmax}_{\theta\in\Theta} M_n(\theta) \ .$$

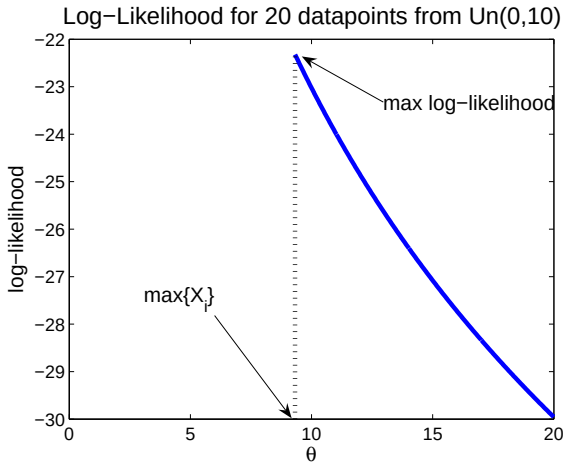
Definition 2 (**Z-estimator (estimating equations)).** An estimator $\hat{\theta}_n$ that can be expressed as the *root* of the expression:

$$\Phi_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \phi_\theta(X_i)$$

for some function $\phi_\theta(\cdot)$; that is, a solution to

$$\Phi_n\left(\hat{\theta}_n\right) = 0$$

M-estimators first were introduced in the context of robust estimation by Peter J. Huber as a generalization of the *maximum likelihood estimator* (MLE): $m_\theta(x) = \log p_\theta(x)$. In the literature, they are often confused with Z-esimators because of the relationship between optimization and differentiation. In fact under certain conditions, they are equivalent via the relationship $\phi_\theta(x) = \nabla_\theta[m_\theta(x)]$. If m_θ is everywhere differentiable w.r.t. θ then the M-estimator is a Z-estimator. A simple example where this fails is the estimation of the parameter θ for the distribution $\text{Un}(0,\theta)$. In this model, the log-likelihood is discontinuous in θ but the MLE is well defined as $\hat{\theta}_n = \max\{X_i\}_{i=1}^n$, which occurs at this discontinuity as show in the following figure:



As is clear, the log-likelihood is $-\infty$ before the MLE and decreasing after it. Hence, the maximum of the log-likelihood occurs at this point of discontinuity even though the derivative is not 0 there (it is not defined).

2 Consistency of M-estimators (van der Vaart, 1998, Section 5.2, p. 44–51)

Definition 3 (Consistency). An estimator is *consistent* if $\hat{\theta}_n \xrightarrow{P} \theta_0$ (alternatively, $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$) for any $\theta_0 \in \Theta$, where θ_0 is the true parameter being estimated.

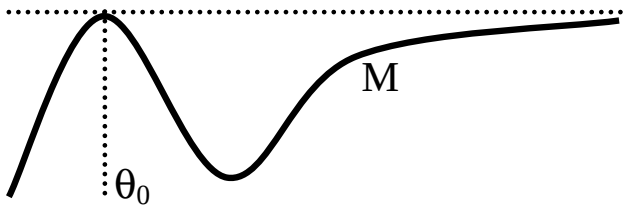
Theorem 4. *(van der Vaart, 1998, Theorem 5.7, p. 45) Let M_n be random functions and M be a fixed function such that $\forall\, \epsilon > 0$:*

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0 \tag{2}$$

$$\sup_{\{\theta \mid d(\theta,\theta_0) \geq \epsilon\}} M(\theta) < M(\theta_0) \tag{3}$$

Then, any sequence $\hat{\theta}_n$ with $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$ converges in probability to θ_0.

Notice, condition (2) is a restriction on the random functions M_n, whereas condition (3) ensures that θ_0 is a *well-separated* maximum of M; i.e., only θ close to θ_0 achieve a value $M(\theta)$ close to the maximum (See figure below):



Finally it is worth noting that sequences $\hat{\theta}_n$ that *nearly maximize* M_n (i.e., $M_n(\hat{\theta}_n) \geq sup_\theta M_n(\theta) - o_p(1)$) meet the above requirement on $\hat{\theta}_n$.

Proof. We are assuming that our $\hat{\theta}_n$ satisfies, $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$. Then, uniform convergence of M_n to M implies

$$\begin{aligned}
\Rightarrow \quad & M_n(\theta_0) \xrightarrow{P} M(\theta_0) \\
\Rightarrow \quad & M_n(\hat{\theta}_n) \geq M(\theta_0) - o_p(1) \\
\Rightarrow \quad & M(\theta_0) \leq M_n(\hat{\theta}_n) + o_p(1) \\
\Rightarrow \quad & M(\theta_0) - M(\hat{\theta}_n) \leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_p(1) \\
& \leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + o_p(1) \\
& \xrightarrow{P} 0 \quad \text{(by condition (2))}
\end{aligned}$$

Now, by condition (3), $\forall\ \epsilon > 0$, $\exists\ \eta$ such that $M(\theta) < M(\theta_0) - \eta$ is satisfied $\forall\ \theta : d(\theta, \theta_0) \geq \epsilon$. Thus $\{d(\hat{\theta}_n, \theta_0) \geq \epsilon\} \subseteq \{M(\hat{\theta}_n) < M(\theta_0) - \eta\}$.

$$\Rightarrow P\left(d(\hat{\theta}_n, \theta_0) \geq \epsilon\right) \leq \underbrace{P\left(M(\hat{\theta}_n) < M(\theta_0) - \eta\right)}_{\xrightarrow{P} 0 \quad \text{(as shown above)}}$$

□

The primary drawback of this approach is that it requires the metric entropy to achieve condition (2).

3 Consistency of the MLE (non-parametric)

We assume that we have n i.i.d. samples from some (unknown) distribution P; i.e., $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$. Further, we assume P has a density $p_0 = \frac{dP}{d\mu}$. For the family of densities, $\mathcal{P}$, we will consider the *maximum likelihood estimator* (MLE) amongst $\mathcal{P}$ as

$$\hat{p}_n = \text{argmax}_{p \in \mathcal{P}} \int \log p dP_n$$

where $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$—the empirical distribution. To further formalize this, we consider the following definitions.

Definition 5 (**Kullback-Leibler (KL)-divergence).** The Kullback-Leibler divergence between two densities is defined as,

$$K(p_0, p) = \int \log \frac{p_0(x)}{p(x)} dP(x) \ .$$

(Recall, $K(p_0, p)$ is always non-negative and is 0 if and only if $p_0(x) = p(x)$ almost everywhere.)

Definition 6 (**Maximum Likelihood Estimator (MLE)).** The maximum-likelihood estimator $\hat{p}_n$ is the minimizer of

$$\int \log \frac{p_0(x)}{\hat{p}_n(x)} dP(x)$$

where P has a density p_0. This implies

$$\int \log \frac{\hat{p}_n}{p_0} dP_n \leq 0 \tag{4}$$

Given these definitions, we now derive a bound on the KL-divergence between the true density p_0 and the MLE $\hat{p}_n$:

$$\begin{aligned}
\Rightarrow \quad & \int \log \frac{p_0(x)}{\hat{p}_n(x)} dP_n(x) \leq 0 \\
\Rightarrow \quad & \int \log \frac{p_0(x)}{\hat{p}_n(x)} dP_n(x) - K\left(p_0, \hat{p}_n\right) + K\left(p_0, \hat{p}_n\right) \leq 0 \\
\Rightarrow \quad & K\left(p_0, \hat{p}_n\right) \leq \left| \int \log \frac{p_0(x)}{\hat{p}_n(x)} dP_n(x) - \int \log \frac{p_0(x)}{\hat{p}_n(x)} dP(x) \right| \\
& \qquad = \left| \int \log \frac{\hat{p}_n(x)}{p_0(x)} d(P_n - P)(x) \right| \ .
\end{aligned}$$

Thus, we need a ULLN for the family of functions: $\mathfrak{F} = \{\log \frac{p}{p_0}\{p_0 > 0\} \mid p \in \mathcal{P}\}$. To this end, we use the following distance measure:

Definition 7 (**Hellinger Distance).**

$$h\left(p_1, p_2\right) = \left(\frac{1}{2} \int \left(p_1^{1/2}(x) - p_2^{1/2}(x) \right)^2 d\mu(x) \right)^{\frac{1}{2}}$$

Unlike the KL-divergence, Hellinger distance is a proper distance metric (non-negative, symmetric, transitive, and 0 if and only if $p_1 = p_2$ almost everywhere). Moreover, Hellinger is appealing as the square-root of a density lies in $\mathcal{L}_2$. Further we have the following:

Lemma 8.

$$h^2(p_1, p_2) \leq \frac{1}{2} K(p_1, p_2)$$

Proof. We use the inequality $\log(x) \leq x - 1$ in the form $\frac{1}{2}\log(v) \leq v^{1/2} - 1$. This gives the following:

$$\begin{aligned}
\Rightarrow \quad & \frac{1}{2} \log \frac{p_2(x)}{p_1(x)} \leq \frac{p_2^{1/2}(x)}{p_1^{1/2}(x)} - 1 \\
\Rightarrow \quad & \frac{-1}{2} K(p_1, p_2) \leq \int_{p_1 > 0} \frac{p_2^{1/2}(x)}{p_1^{1/2}(x)} p_1(x) \mu(dx) - 1 \\
\Rightarrow \quad & \frac{1}{2} K(p_1, p_2) \geq \underbrace{\frac{1}{2}}_{\frac{1}{2}\int_{p_1>0} p_1(x)\mu(dx)} + \underbrace{\frac{1}{2}}_{\frac{1}{2}\int_{p_1>0} p_2(x)\mu(dx)} - \int_{p_1>0} \frac{p_2^{1/2}(x)}{p_1^{1/2}(x)} p_1(x)\mu(dx) \\
\Rightarrow \quad & \frac{1}{2} K(p_1, p_2) \geq \int_{p_1>0} \frac{1}{2} p_1(x) - p_1^{1/2}(x) p_2^{1/2}(x) + \frac{1}{2} p_2(x) \mu(dx) \\
\Rightarrow \quad & \frac{1}{2} K(p_1, p_2) \geq \underbrace{\frac{1}{2} \int \left(p_1^{1/2}(x) - p_2^{1/2}(x) \right)^2 \mu(dx)}_{= h^2(p_1, p_2)}
\end{aligned}$$

□

Unfortunately, though, $\mathfrak{F}$ is hard to work with (p's are not bounded away from 0). Instead we will work with the family

$$\mathfrak{G} \triangleq \{\frac{1}{2} \log \frac{p + p_0}{2p_0}\{p_0 > 0\} \mid p \in \mathcal{P}\}$$

which is bounded below by $\frac{1}{2}\log\frac{1}{2}$.

Lemma 9.

$$h^2\left(\frac{\hat{p}_n + p_0}{2}, p_0\right) \leq \int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n + p_0}{2p_0} d(P_n - P)$$

Proof. Concavity of the logarithm implies

$$\begin{aligned}
\Rightarrow \qquad & \log\frac{\hat{p}_n + p_0}{2} \geq \frac{1}{2}\log\hat{p}_n + \frac{1}{2}\log p_0 \\
\Rightarrow \qquad & \log\frac{\hat{p}_n + p_0}{2} - \log p_0 \geq \frac{1}{2}\log\hat{p}_n - \frac{1}{2}\log p_0 \\
\Rightarrow \qquad & \log\frac{\hat{p}_n + p_0}{2p_0}\{p_0 > 0\} \geq \frac{1}{2}\log\frac{\hat{p}_n}{p_0}\{p_0 > 0\}
\end{aligned}$$

Now, by the definition of the MLE (Eq. (4)):

$$\begin{aligned}
\Rightarrow \qquad & 0 \leq \int_{p_0>0} \frac{1}{4}\log\frac{\hat{p}_n}{p_0} dP_n \\
\Rightarrow \qquad & 0 \leq \int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n + p_0}{2p_0} dP_n \\
& = \int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n + p_0}{2p_0} d(P_n - P) + \underbrace{\int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n + p_0}{2p_0} dP}_{=-\frac{1}{2}K\left(p_0, \frac{\hat{p}_n + p_0}{2}\right)} \\
& \leq \int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n + p_0}{2p_0} d(P_n - P) - h^2\left(\frac{\hat{p}_n + p_0}{2}, p_0\right) \quad \text{(by Lemma 8)} \\
\Rightarrow \qquad & h^2\left(\frac{\hat{p}_n + p_0}{2}, p_0\right) \leq \int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n + p_0}{2p_0} d(P_n - P)
\end{aligned}$$

□

Thus, elements of our family $\mathfrak{G}$ have Hellinger distance 0 that goes to 0. To connect this back to our orginal family $\mathfrak{F}$, we have the following Lemma:

Lemma 10.

$$h^2(p, p_0) \leq 16h^2(\bar{p}, p_0)$$

where $\bar{p} \triangleq \frac{p+p_0}{2}$.

Finally, we arrive at the following Theorem:

Theorem 11. *Let* $\mathfrak{G} = \{\frac{1}{2}\log\frac{\bar{p}}{p_0}\{p_0 > 0\} \mid p \in \mathcal{P}\}$ *and let* $G = \sup_{g\in\mathfrak{G}}|g|$. *Assume that* $\int G dP < \infty$ *and* $\forall\epsilon > 0 \quad \frac{1}{n}H_1(\epsilon, P_n, \mathfrak{G}) \xrightarrow{P} 0$, *then*

$$h(\hat{p}_n, p_0) \xrightarrow{a.s.} 0$$

Example 12 (**Logistic Regression for nonparameteric links).** We are given data pairs: (Y_i, Z_i) and we assume the conditional distribution of Y follows a particular functional form:

$$P(Y = 1|Z = z) = F_{\theta_0}(z)$$

where F_θ is an increasing function of z for every $\theta \in \Theta$ and $\theta_0 \in \Theta$ is the true parameter.

Let μ be (counting measure on $\{0, 1\}) \times Q$ where Q is the distribution of Z. Now, the family of joint densities we obtain is

$$\mathcal{P} = \{p_\theta(y, z) = yF_\theta(z) + (1 - y)(1 - F_\theta(z))\}$$

which has the following properties:

- $\sup_{p\in\mathcal{P}} p \leq 1$.
- $H_B(\epsilon, \mu, \mathcal{P}) \leq A\epsilon^{-1}$ (for increasing functions).

Hence we have

$$h\left(\hat{p}_n, p_0\right) \xrightarrow{P} 0$$

References

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

Stat210B: Theoretical Statistics **Lecture Date: February 20, 2007**

Lecture 10

Lecturer: Michael I. Jordan *Scribe: Alex Shyr*

In Empirical Process theory, the notion of a sequence of stochastic processes converging to another process is important. The scalar analogy of this convergence is the CLT. This lecture is an introduction to Donsker's Theorem, one of the fundamental theorems of Empirical Process theory.

1 Weak Convergence (aka Conv. in Law, Conv. in Distribution)

Given the usual sample space $(\Omega, \mathcal{F}, P)$, random element $X : \Omega \to \mathcal{X}$. Let $\mathcal{A}$ be a σ-field of $\mathcal{X}$.

Define $C(\mathcal{X}, \mathcal{A})$ to be the space of continuous, bounded function class on $\mathcal{X}$, which is measurable on $\mathcal{A}$.

A sequence of probability measures Q_n converges weakly to Q if $Q_n f \to Qf, \forall f \in C(\mathcal{X}, \mathcal{A})$. Note that $\mathcal{A}$ must be smaller than the Borel σ-field $\mathcal{B}(\mathcal{X})$. An alternative field that works is the projection σ-field generated by the coordinate projection maps.

2 Continuous Mapping Theorem (van der Vaart, 1998, Cha.18)

Since weak convergence does not hold for all probability measures, we need conditions on the set $\mathcal{C}$ on which the limiting random element concentrates.

Definition 1. A set $\mathcal{C}$ is *separable* if it has a countable, dense subset.

A point X in $\mathcal{X}$ is *regular* if

$$\forall \text{ neighborhood } V \text{ of } X, \exists \text{ a uniformly continuous } g \text{ with } g(X) = 1 \text{ and } g \leq V.$$

Theorem 2. *Let H be an $\mathcal{A}/\mathcal{A}'$ measurable map from $\mathcal{X}$ into another metric space $\mathcal{X}'$. If H is continuous at each point of some separable, $\mathcal{A}$-measurable set $\mathcal{C}$ of regular points, then*

$$X_n \xrightarrow{\mathcal{L}} X \text{ and } P(X \in \mathcal{C}) = 1 \qquad \Rightarrow \qquad HX_n \xrightarrow{\mathcal{L}} HX$$

Some useful notes:

- a common function space $\mathcal{X}$ is $D[0,1]$, which is the set of all $\mathbf{R}$-valued, *cadlag* functions
- $d(x,y) = \sup_{0 \leq t \leq 1} |x(t) - y(t)|$ defines a metric, and closed balls for d generate the projection σ-field
- every point of $D[0,1]$ is regular, but $D[0,1]$ is not separable ...
- BUT the limit processes we will talk about concentrate on $C[0,1]$, which is separable.

Theorem 3 ("Stochastic equicontinuity" or "Asymptotic tightness"). *Let $X_1, ..., X_n$ be random elements of $D[0,1]$. Suppose that $P(X \in \mathcal{C}) = 1$ for some separable $\mathcal{C}$. Then $X_n \xrightarrow{\mathcal{L}} X$ iff*

(i) *Fidi convergence of X_n to X (ie. $\Pi_S X_n \xrightarrow{\mathcal{L}} \Pi_S X \quad \forall$ finite $S \subseteq [0,1]$)*

(ii) $\forall \epsilon > 0, \delta > 0, \exists$ *a grid* $0 = t_0 < t_1 < ... < t_n = 1$ *s.t.* $\limsup_n P\{\max_i \sup_{J_i} |X_n(t) - X_n(t_i)| > \delta\} < \epsilon$, *where* $J_i = [t_i, t_{i+1})$

3 Donsker's Theorem (for standard empirical process)

The first version of Donsker's theorem deals with the convergence of the empirical process U_n of random variables drawn uniformly from the unit interval, where

$$U_n t = \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} 1_{\{\xi_i \leq t\}} - t), \qquad \text{and } \xi_i \overset{iid}{\sim} U[0,1]$$

Definition 4 (Brownian Bridge). U is a Brownian Bridge iff

(i) $\forall$ finite subset $S \in [0,1]$, $\Pi_S U$ is Gaussian with zero mean,

(ii) covariances $E[U(s)U(t)] = s(1-t), \quad \forall 0 \leq s \leq t \leq 1,$ and

(iii) U only has continuous sample paths.

Theorem 5. $U_n \xrightarrow{\mathcal{L}} U$, *where U is a Brownian Bridge.*

Proof. First check (i) of Theorem 3.

$$\begin{aligned} E[U_n s U_n t] &= \frac{1}{n}\sum_i E[(1_{\{\xi_i \leq t\}} - t)(1_{\{\xi_i \leq s\}} - s)] \\ &= \frac{1}{n}\sum_i \{P(\xi_i \leq s) - tP(\xi_i \leq s) - sP(\xi_i \leq t) + st\} \\ &= s(1-t) \end{aligned}$$

□

References

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

Stat210B: Theoretical Statistics **Lecture Date: February 21, 2008**

Weak Convergence in General Metric Spaces

Lecturer: Michael I. Jordan *Scribe: Yueqing Wang*

1 General Metric (Norm) Space

The objects of interest are functions from a sample space to a general metric space, where each point is a function. Then we can try to use statistical properties, e.g. goodness of fit, to test certain assumptions.

Example 1 (*Cramér-von Mises*)**.** Let P_n be the empirical probability measures of a random sample $X_1, \ldots, X_n$ of real-valued random variables. The *Cramér-von Mises statistic* for testing the (null) hypothesis that the underlying probability measure is a given P is given by

$$\int (P_n f - Pf)^2 \mathrm{d}P,$$

which can be considered as a measure for the distance between P_n and P. If the distribution of this statistic is known, we can test the hypothesis. P can be very complex. But if the class $\mathcal{F}$ of measurable functions is *P-Donsker*, the *Cramér-von Mises statistic* converges to a Brownian Bridge.

Definition 2 (Uniform Norm)**.** The uniform norm on function spaces is defined as

$$\|Z\| = \sup_{t \in T} |Z(t)|. \tag{1}$$

Example 3. Some commonly used general metric spaces:

- $C[a, b]$. All the continuous functions on $[a, b] \in \mathbf{R}$.
- $D[a, b]$. (*Cadlag* functions). All the functions that have limit from the left and are continuous from the right.
- $\ell^\infty[a, b]$. All bounded functions.

And we have,

$$C[a, b] \subseteq D[a, b] \subseteq \ell^\infty[a, b]$$

Note. $C[a, b]$ is separable, i.e. it has a countable dense subset. $D[a, b]$ isn't separable. Hence, $\ell^\infty[a, b]$ is not separable, neither. Most of the empirical processes are in $D[a.b]$ because of the jumps; most limiting processes are in $C[a, b]$.

2 Weak Convergence

Definition 4 (Random Element)**.** The *Borel σ-field* on a metric space $\mathbb{D}$ is the smallest σ-field that contains the open sets (and then also the closed sets). A function defined relative to (one or two) metric

spaces is called *Borel-measurable* if it is measurable relative to the Borel σ-field(s). A Borel-measurable map $X : \Omega \to \mathbb{D}$ defined on a probability space $(\Omega, \mathcal{U}, P)$ is referred to as a *random element* with values in $\mathbb{D}$.

Definition 5. *Random Elements* (R.E.) X_n converging weakly to the random element X means $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$, for all bounded and continuous function f.

Note. For random elements, Continuous Mapping Theorem still holds. If random elements $X_n \xrightarrow{d} X$ and functions $g_n \to g$ are continuous, it follows that

$$g_n(X_n) \xrightarrow{d} g(X)$$

Definition 6. A random element is *tight* if $\forall \epsilon > 0$, $\exists$ a compact set K such that

$$\mathbb{P}(X \notin K) \leq \epsilon.$$

Definition 7. $X = \{X_t : t \in T\}$ is a collection of random variables, where $X_t : \Omega \to \mathbb{R}$ is defined on $(\Omega, \mathcal{U}, P)$. A *sample path* is defined as $t \to X_t(\omega)$.

Theorem 8 (**Converge Weakly to a Tight Random Element**). *A sequence of maps $X_n : \Omega_n \to l^\infty(T)$ converge weakly to a tight R.E. iff*

(i) (Fidi Convergence) $(X_{n,t_1}, \ldots, X_{n,t_k})$ converges weakly in $\mathbf{R}^k$ for each finite set $(t_1, \ldots, t_k)$.

(ii) (Asymptotic Partition) $\forall \epsilon, \eta > 0$, exists a partition of T into finitely many sets $T_1, \ldots, T_k$ such that

$$\limsup_{n \to \infty} \mathbb{P}(\sup_i \sup_{s,t \in T_i} |X_{n,s} - X_{n,t}| \geqslant \epsilon) \leqslant \eta.$$

3 The Donsker Theorems

Theorem 9 (**Classical Donsker Theorem**). *If $X_1, \ldots$ are i.i.d. random variables with distribution function F, where F is uniform distribution function on the real line and $\{\mathbb{F}_n\}$ are the empirical processes: $\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1} \mathbf{1}_{\{X_i \leq t\}}$. Then for fixed $(t_1, \ldots, t_k)$, it follows that,*

$$\sqrt{n}(\mathbb{F}_n(t_1) - F(t_1), \ldots, \mathbb{F}_n(t_k) - F(t_k)) \xrightarrow{d} (\mathbb{G}_F(t_1), \ldots, \mathbb{G}_F(t_k)),$$

where $\{\mathbb{G}_F(t_i)\}$ are zero-mean Gaussian with covariance $t_i \wedge t_j - t_i t_j$.

Theorem 10 (**Donsker**). *If $X_1, \ldots$ are i.i.d. random variables with distribution function F, then the sequence of empirical processes $\sqrt{n}(\mathbb{F}_n - F)$ converges in distribution in the space $\mathrm{D}[-\infty, \infty]$ to a tight random element $\mathbb{G}_F$ (i.e. a Brownian Bridge), whose marginal distributions are zero-mean normal with covariance function: $\mathbb{E}\mathbb{G}_F(t_i)\mathbb{G}_F(t_j) = F(t_i \wedge t_j - F(t_i)F(t_j))$.*

Denote empirical processes as follows: $\mathbb{G}_n = \sqrt{n}(P_n - P)$ and thus $\mathbb{G}_n f = \sqrt{n}(Pf_n - Pf)$.

Definition 11 (P-Donsker). $\mathcal{F}$ is *$P-$Donsker* if $\mathbb{G}_n$ converges weakly to a tight limit process in $l^\infty(\mathcal{F})$ which is a P-Brownian Bridge $\mathbb{G}_P$ with zero mean and covariance function $\mathbb{E}\mathbb{G}_P f \mathbb{G}_P g = Pfg - PfPg$.

Definition 12. Define the *Bracketing Integral* as,

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))} \mathrm{d}\epsilon$$

Theorem 13. *If $J_{[]}(1, \mathcal{F}, L_2(P)) < \infty$, $\mathcal{F}$ is P-Donsker.*

Example 14. $\mathcal{F} = \{1_{(-\infty,t]} : t \in \mathbf{R}\}$. By calculating the bracketing number, it follows that $\log N_{[]} \to \frac{1}{\epsilon^2}$. Hence there exists limits for $J_{[]}(1, \mathcal{F}, L_2(P))$. By the above theorem we know that this function space is *P-Donsker* and the empirical processes will converge to a Brownian Bridge.

Example 15 (Lipschitz Classes are P-Donsker)**.** Let $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbf{R}^d\}$ be a Lipschitz function class. i.e. given x (fixed), if

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leqslant m(x)\|\theta_1 - \theta_2\|, \forall \theta_1, \theta_2,$$

then,

$$N_{[]}(\epsilon\|m\|_{p,r}, \mathcal{F}, L_r(P)) \leqslant k(\frac{\text{diameter } \Theta}{\epsilon})^d,$$

where k is a certain constant.

Proof. The brackets $(f_\theta - \epsilon m, f_\theta + \epsilon m)$ for θ have size smaller than $2\epsilon\|m\|_{p,r}$. And they cover $\mathcal{F}$ because,

$$f_{\theta_1} - \epsilon m \leq f_{\theta_2} \leq f_{\theta_1} + \epsilon m, \text{ if } \|\theta_1 - \theta_2\| \leq t.$$

Hence, we need at most $(\frac{\text{diam } \Theta}{\epsilon})^d$ cubes of size ϵ to cover Θ, and then use balls to cover the cubes. □

References

References

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

Stat210B: Theoretical Statistics **Lecture Date: February 26, 2008**

The Chaining Lemma

Lecturer: Michael I. Jordan *Scribe: Fabian Wauthier*

Recall from last time the definition of a P-Donsker class.

Definition. A class of functions is called *P-Donsker* if $\mathbb{G}_n$ converges weakly to a tight limit process in $l^\infty(\mathcal{F})$, which is a P-Brownian bridge $\mathbb{G}_p$ with zero mean and covariance function $E(\mathbb{G}_p f \mathbb{G}_p g) = Pfg - PfPg$. Here the empirical process $\mathbb{G}_n$ is defined as $\mathbb{G}_n = \sqrt{n}(P_n - P)$. This means in particular, that for any finite collection of functions, the elements $\mathbb{G}_n f$ converge to a zero mean multivariate Gaussian, with aforementioned covariance function.

Furthermore, recall Theorem 19.5 stated last time.

Theorem 19.5 (Donsker). *Every class $\mathcal{F}$ of measurable functions with $J_{[]}(1, \mathcal{F}, L_2(P)) < \infty$ is P-Donsker. Here we defined $J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2(P))} d\varepsilon$.*

In this lecture we will be concerned mostly with proving the Chaining Lemma, which is instrumental to the proof of this theorem. Before commencing the presentation, we first illustrate some properties of P-Donsker classes.

Combining P-Donsker classes

The definition of P-Donsker classes gives rise to an algebra for combining any two P-Donsker classes. In particular, suppose that $f \in \mathcal{F}$ and $g \in \mathcal{G}$ are both P-Donsker. If $\phi(\cdot, \cdot)$ is a Lipschitz transformation, then $\phi(f, g)$ is P-Donsker. Examples of such Lipschitz transformations include: $f + g$, $f \wedge g$, $f \vee g$, fg if $\mathcal{F}$ and $\mathcal{G}$ are uniformly bounded, and $1/f$ if $\mathcal{F}$ is bounded away from zero.

Chaining Lemma

In this lecture we give a thorough treatment of the core of empirical process theory by proving the Chaining Lemma (lemma 19.34 in van der Vaart). The presentation is based on section 19.6 in van der Vaart (1998). We begin by stating two relevant lemmas. The first one, Bernstein's inequality, represents a tightening of the Hoeffding bound we previously discussed. This strengthening will be required for the following argument.

Lemma 19.32 (Bernstein's inequality). *For one function f and any $x > 0$,*

$$P(|\mathbb{G}_n f| > x) \leq 2 \exp\left\{ -\frac{1}{4} \frac{x^2}{Pf^2 + x||f||_\infty / \sqrt{n}} \right\}. \tag{1}$$

Note that as in Hoeffding, the upper bound is twice the exponential of some function. Here, the Pf^2 term in the exponential accounts for something like the variance, whereas in Hoeffding there was an upper bound on the variance through terms $\sum_i (b_i - a_i)^2$. An additional term has also been introduced to the denominator.

The next lemma will relate Bernstein's inequality to finite collections.

Lemma 19.33. *For any finite class $\mathcal{F}$ of bounded, measurable and square-integrable functions, with $|\mathcal{F}|$ elements, we have*

$$E(||\mathbb{G}_n||_{\mathcal{F}}) \lesssim \max_f \frac{||f||_\infty}{\sqrt{n}} \log(1+|\mathcal{F}|) + \max_f ||f||_{P,2}\sqrt{\log(1+|\mathcal{F}|)}. \tag{2}$$

Here, we have adopted the notation $\lesssim$ to express that the left hand side is less than the right hand side, up to a universal multiplicative constant. The proof idea behind this Lemma lies in breaking the left hand side into two pieces using the triangle inequality, and then applying Bernstein's inequality to both.

We now turn to the Chaining Lemma. The motivation for this lemma lies in the difficulty of carrying out an independent analysis of fluctuations for each element f of an uncountably infinite set of functions $\mathcal{F}$. To get control over the infinite set, we need to tie functions together to a finite number of grid cells. We can introduce suitable structure on $\mathcal{F}$ via a multi-resolution grid. At the coarse top level very few cells partition $\mathcal{F}$; at progressively deeper levels each grid cell is partitioned into a set of smaller cells. By choosing one representative function for each grid cell, the fluctuations between any two functions in $\mathcal{F}$ can be related to fluctuations along edges on the grid tree.

Lemma 19.34 (Chaining Lemma). *Define* $\text{Log } x = 1 \vee \log(x)$ *and* $a(\delta) = \delta/\sqrt{\text{Log } N_{[]}(\delta, \mathcal{F}, L_2(P))}$. *For any class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$ so that, for some common δ^2, $Pf^2 \leq \delta^2, \forall f \in \mathcal{F}$, and F an envelope function,*

$$E(||\mathbb{G}_n||_{\mathcal{F}}) \lesssim J_{[]}(\delta, \mathcal{F}, L_2(P)) + \sqrt{n}PF\left\{F > \sqrt{n}a(\delta)\right\}. \tag{3}$$

Proof. We begin the proof by focussing on the first term on the right hand side. For $|f| \leq g$ by the triangle inequality

$$\begin{aligned} |\mathbb{G}f| &= \sqrt{n}|P_n f - Pf| && (4)\\ &\leq \sqrt{n}(P_n|f| + P|f|) && (5)\\ &\leq \sqrt{n}(P_n g + Pg). && (6) \end{aligned}$$

This implies that for an envelope function F

$$\begin{aligned} E\left(||\mathbb{G}_n f\left\{F > \sqrt{n}a(\delta)\right\}||_{\mathcal{F}}\right) &\leq \sqrt{n}E\left(P_n F\left\{F > \sqrt{n}a(\delta)\right\} + PF\left\{F > \sqrt{n}a(\delta)\right\}\right) && (7)\\ &= 2\sqrt{n}PF\left\{F > \sqrt{n}a(\delta)\right\}. && (8) \end{aligned}$$

This demonstrates the inequality for the second term on the right hand side. We continue the derivation on $||\mathbb{G}_n f\left\{F \leq \sqrt{n}a(\delta)\right\}||$ and show that it is less than or equal to $J_{[]}(\delta, \mathcal{F}, L_2(P))$. Since the set of remaining functions we work with has shrunk, it has smaller bracketing number than $\mathcal{F}$. For notational convenience, continue by assuming that $f \leq \sqrt{n}a(\delta), \forall f \in \mathcal{F}$. At this point we turn to the multi-resolution structure on $\mathcal{F}$ which we previously noted. Choose an integer q_0 such that $4\delta \leq 2^{-q_0} \leq 8\delta$. Also choose a nested sequence of partitions $\mathcal{F}_{q_i}$ of $\mathcal{F}$ indexed by integers $q \geq q_0$; that is, if at level q there are N_q disjoint sets, then $\mathcal{F} = \cup_{i=1}^{N_q}\mathcal{F}_{q_i}$. Choose this nested sequence of partitions and measurable functions $\Delta_{q_i} \leq 2F$, so that

$$\begin{aligned} \sum_{q \geq q_0} 2^{-q}\sqrt{\text{Log } N_q} &\lesssim \int_0^\delta \sqrt{\text{Log } N_{[]}(\varepsilon, \mathcal{F}, L_2(P))}d\varepsilon && (9)\\ \sup_{f,g \in \mathcal{F}_{q_i}} |f - g| \leq \Delta_{q_i}, &\quad P\Delta_{q_i}^2 < 2^{-2q}. && (10) \end{aligned}$$

The functions Δ_{q_i} are the difference between upper and lower brackets and act as envelopes.

We continue by choosing a representative function within each cell of each level. Fix for each level $q > q_0$ and each partition $\mathcal{F}_{q_i}$ one representative f_{q_i} and define, if $f \in \mathcal{F}_{q_i}$

$$\pi_q f = f_{q_i} \text{ (Nearest neighbor function)} \tag{11}$$
$$\Delta_q f = \Delta_{q_i}. \tag{12}$$

Here is where $\mathcal{F}$ is attributed a finite representation. At scale q, $\pi_q f$ and $\Delta_q f$ run over N_q functions as f runs over $\mathcal{F}$. Define

$$a_q = 2^{-q}/\sqrt{\text{Log } N_{q+1}}, \tag{13}$$
$$A_{q-1} f = \mathbb{I}\left\{\Delta_{q_0} f \le \sqrt{n} a_{q_0}, \ldots, \Delta_{q-1} f \le \sqrt{n} a_{q-1}\right\}, \tag{14}$$
$$B_q f = \mathbb{I}\left\{\Delta_{q_0} f \le \sqrt{n} a_{q_0}, \ldots, \Delta_{q-1} f \le \sqrt{n} a_{q-1}, \Delta_q f > \sqrt{n} a_q\right\}. \tag{15}$$

Now decompose the difference between any f and the representative $\pi_{q_0} f$ using the newly defined sets as a telescoping sum,

$$f - \pi_{q_0} f = \sum_{q_0+1}^{\infty} (f - \pi_q f) B_q f + \sum_{q_0+1}^{\infty} (\pi_q f - \pi_{q-1} f) A_{q-1} f. \tag{16}$$

We observe that either all of the $B_q f$ are zero[1] in which case the $A_{q-1} f$ are 1 (we always have small fluctuations). Alternatively, one $B_{q_1} f = 1$ for some $q_1 > q_0$ (and zero for all other q), in which case $A_q f = 1$ for $q < q_1$ and $A_q f = 0$ for $q \ge q_1$. In that last case we have a sequence of small fluctuations, followed by one large fluctuation

$$f - \pi_{q_0} f = (f - \pi_{q_1} f) + \sum_{q_0+1}^{q_1} (\pi_q f - \pi_{q-1} f) A_{q-1} f. \tag{17}$$

By the construction of partitions and our choice of q_0 we have

$$2a(\delta) = \frac{2\delta}{\sqrt{\text{Log } N_{[]}(\delta, \mathcal{F}, L_2(P))}} \tag{18}$$
$$\le \frac{2^{-q_0}}{\sqrt{\text{Log } N_{q_0+1}}} \tag{19}$$
$$= a_{q_0}. \tag{20}$$

This implies that $\Delta_{q_0} f \le a_{q_0}\sqrt{n}$ and therefore $A_{q_0} f = 1$. Furthermore, nesting implies $\Delta_q f B_q f \le \Delta_{q-1} f B_q f \le \sqrt{n} a_{q-1}$. The last inequality holds if $B_q f = 0$ and also if $B_q f = 1$ by definition. It follows that since $B_q f$ is an indicator where $\Delta_q f > \sqrt{n} a_q$ that $\sqrt{n} a_q P(\Delta_q f B_q f) \le P(\Delta_q f B_q f)^2 = P(\Delta_q f)^2 B_q f \le 2^{-2q}$ by the choice of $\Delta_q f$. We now apply the empirical process $\mathbb{G}_n$ to both series on the right of the equation 16 and use the triangular inequality on the supremum over absolute values. Because $|\mathbb{G}_n f| \le \mathbb{G}_n g + 2\sqrt{n} P g$ for $|f| < g$ we get, by applying Lemma 19.33

$$E\left\|\sum_{q_0+1}^{\infty} \mathbb{G}_n (f - \pi_q f) B_q f\right\|_{\mathcal{F}} \le \sum_{q_0+1}^{\infty} E||\mathbb{G}_n \Delta_q f B_q f||_{\mathcal{F}} + \sum_{q_0+1}^{\infty} 2\sqrt{n}||P\Delta_q f B_q f||_{\mathcal{F}} \tag{21}$$
$$\overset{19.33}{\lesssim} \sum_{q_0+1}^{\infty} \left[a_{q-1}\text{Log } N_q + 2^{-q}\sqrt{\text{Log } N_q} + \frac{4}{a_q} 2^{-2q}\right]. \tag{22}$$

We note that the third term arises in part from our earlier observation that $P(\Delta_q f B_q f) \le 2^{-2q}/\sqrt{n} a_q$. However, it was unclear in class where the additional factor of 2 stems from. All three terms in the infinite

[1] There is a typo in van der Vaart (1998) page 287, where the author states that "either all $B_q f$ are *1*".

sum will become essentially like the middle one, which we know from inequality 9 can be bounded by a multiple of $J_{[]}(\delta, \mathcal{F}, L_2(P))$. Thus we have bounded one more term.

To establish a similar bound for the second part of equation 16, note that there are at most N_q functions $\pi_q f - \pi_{q-1} f$ and at most N_{q-1} indicators $A_{q-1} f$. Nesting implies $|\pi_q f - \pi_{q-1} f| A_{q-1} f \leq \Delta_{q-1} f A_{q-1} f \leq \sqrt{n} a_{q-1}$. The $L_2(P)$ norm of $|\pi_q f - \pi_{q-1} f|$ is upper bounded by $2^{-(q+1)}$. Now using Lemma 19.33 we find that

$$E \left|\left| \sum_{q_0+1}^{\infty} \mathbb{G}_n (\pi_q f - \pi_{q-1} f) A_{q-1} f \right|\right|_{\mathcal{F}} \lesssim \sum_{q_0+1}^{\infty} \left[a_{q-1} \text{Log } N_q + 2^{-q} \sqrt{\text{Log } N_q} \right]. \tag{23}$$

As before, note that the first and second terms on the right are identical and that each can be bounded by a multiple of $J_{[]}(\delta, \mathcal{F}, L_2(P))$.

As the final step in this proof we need to establish a bound for terms $\pi_{q_0} f$. Note that for the envelope function F, we have $|\pi_{q_0} f| \leq F$. Also, recall that since early in the derivation we are only considering the class of functions $f \{F \leq \sqrt{n} a(\delta)\}$ where f ranges over $\mathcal{F}$, so that $F \leq \sqrt{n} a(\delta)$. Moreover, $\sqrt{n} a(\delta) \leq \sqrt{n} a_{q_0}$ by a similar argument as in derivation 18-20. Recall also that one of the preconditions of this lemma is that $Pf^2 < \delta^2, \forall f \in \mathcal{F}$, so that in particular $P(\pi_{q_0} f)^2 \leq \delta^2$. Applying Lemma 19.33 again, we find that

$$E ||\mathbb{G}_n \pi_{q_0} f||_{\mathcal{F}} \lesssim a_{q_0} \text{Log } N_{q_0} + \delta \sqrt{\text{Log } N_{q_0}}. \tag{24}$$

By the choice of q_0 at the onset and inequality 9, both terms can be bounded by a multiple of $J_{[]}(\delta, \mathcal{F}, L_2(P))$.

This concludes the proof of Lemma 19.34. We summarise briefly. The proof was carried out by using an envelope function F to split the function space $\mathcal{F}$ into two sets. In inequality 8 we quickly saw that one set gives rise to one of the terms in the final result. We then defined a multi-resolution tree on the remaining subset of $\mathcal{F}$ so that we could consider fluctuations via suitably defined events $A_{q-1} f$ and $B_q f$. In the following we repeatedly applied Lemma 19.33 to yield inequalities 22, 23, and 24, each of which can be upper bounded by a multiple of $J_{[]}(\delta, \mathcal{F}, L_2(P))$. In the final result, these three parts are represented by one copy of $J_{[]}(\delta, \mathcal{F}, L_2(P))$. □

References

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

Stat210B: Theoretical Statistics **Lecture Date: March 07, 2007**

Asymptotics of Empirical Processes

Lecturer: Michael I. Jordan *Scribe: Wei-Chun Kao*

1

Lemma 1. *(van der Vaart, 1998, lemma 19.24) Let $\mathcal{F}$ be P-donsker, and $\hat{f}_n$ be a random sequence of functions taking values in $\mathcal{F}$ s.t.*

$$\int (f_n(x) - f_0(x))^2 dP(x) \xrightarrow{P} 0$$

for some f_0 in $\mathcal{L}_2(P)$, then we have

$$G_n(\hat{f}_n - f_0) \xrightarrow{P} 0,$$

and

$$G_n \hat{f}_n \xrightarrow{d} G_P f_0$$

with $G_n = \sqrt{n}(P_n - P)$.

Proof. Sketch: Uses uniform continuity of sample paths of G_p with CMT. □

2 An Example: Mean Absolute Deviation

Define mean absolute deviation

$$M_n = \frac{1}{n} \sum_i |X_i - \bar{X}_n|$$

Let F denote the unknown CDF. W.l.o.g, let $Fx = 0$. Let $F_n|X - \theta| = \frac{1}{n}\sum_i |X_i - \theta|$. If $Fx^2 < \infty$, and if $\theta \in \Theta$ for a compact Θ, then $\{|x - \theta|\}$ is F-Donsker (van der Vaart, 1998, example 19.7)

$$F(|x - \bar{X}_n| - |x|)^2 \leq |\bar{X}_n|^2 \xrightarrow{P} 0$$

By Lemma 1, we have

$$G_n|x - \bar{X}_n| - G_n|x| \xrightarrow{P} 0 \tag{1}$$

Consider

$$\sqrt{n}(M_n - F|x|) = \sqrt{n}(F|x - \bar{X}_n| - F|x|) + G_n|x| + o_P(1), \tag{2}$$

assume that $\theta \mapsto F|x - \theta|$ is differentiable at $\theta = 0$, differentiate $F|x - \theta|$ at $\theta = 0$ we have the derivative:

$$2F(0) - 1.$$

Apply Delta Method on $\sqrt{n}(F|x-\bar{X}_n| - F|x|)$, we have

$$\begin{aligned}\sqrt{n}(F|x-\bar{X}_n| - F|x|) &= -(2F(0)-1)\sqrt{n}((x-\bar{X}_n)-x) + o_P(1) &(3)\\ &= (2F(0)-1)\sqrt{n}\bar{X}_n &(4)\\ &= (2F(0)-1)\sqrt{n}(F_n - F)x + o_P(1) &(5)\\ &= (2F(0)-1)G_n x + o_P(1). &(6)\end{aligned}$$

Therefore, we have

$$\begin{aligned}\sqrt{n}(M_n - F|x|) &= ((2F(0)-1)x + |x|) + o_P(1) &(7)\\ &\xrightarrow{d} G_P((2F(0)-1)x + |x|). &(8)\end{aligned}$$

Thus, M_n is AN with mean 0 and variance equals to variance of $(2F(0)-1)X_1 + |X_1|$. We lose $(2F(0)-1)X$ term by not knowing the mean of X. When the mean and median are the same, $2F(0)-1=0$, in which case we don't incur any extra variance by having to estimate the location parameter.

3 AN of Z-estimators

Definition 2. A function $\psi_\theta(x)$ is Lipschitz if $\exists$ a function $\dot{\psi}(x)$ s.t.

$$||\psi_{\theta_1}(x) - \psi_{\theta_2}(x)|| \le \dot{\psi}(x)||\theta_1 - \theta_2||$$

$\forall \theta_1, \theta_2$ in some neighborhood of θ_0 and $P\dot{\psi}^2 \le \infty$.

Theorem 3. *(van der Vaart, 1998, Theorem 5.21) For each θ_0 in an open subset of Euclidean space, let $\psi_\theta(x)$ be Lipschitz. Assume $P||\psi_{\theta_0}||^2 \le \infty$, $P\psi_\theta$ is differentiable at θ_0 with derivative V_{θ_0} (note that it is different from "ψ_θ is differentiable"). Let*

$$P_n\psi_{\hat{\theta}_n} = o_P(n^{-1/2}) (\text{ a "near zero"}). \tag{9}$$

Assume $\hat{\theta}_n \xrightarrow{P} \theta_0$ (consistency). Then we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_i \psi_{\theta_0}(x_i) + o_P(1),$$

and

$$\hat{\theta}_n - \theta_0 \text{ is AN with zero mean and covariance } V_{\theta_0}^{-1}(P\psi_{\theta_0}\psi_{\theta_0}')V_{\theta_0}^{-T}$$

Proof. (van der Vaart, 1998, example 19.7) shows that Lipschitz functions are P-Donsker. Apply Lemma 1 we have

$$G_n\psi_{\hat{\theta}_n} - G_n\psi_{\theta_0} \xrightarrow{P} 0$$

By assumption that $\sqrt{n}P_n\psi_{\hat{\theta}_n} = o_P(1)$, we have

$$\begin{aligned}G_n\psi_{\hat{\theta}_n} &= -\sqrt{n}P\psi_{\hat{\theta}_n} + o_P(1) &(10)\\ &= \sqrt{n}P(\psi_{\theta_0} - \psi_{\hat{\theta}_n}) + o_P(1), &(11)\end{aligned}$$

with $P\psi_{\theta_0} = 0$ by definition. Apply Delta Method, or (van der Vaart, 1998, Lemma 2.12), we have

$$\sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n) + \sqrt{n}\ o_P(||\hat{\theta}_n - \theta_0||) = G_n\psi_{\theta_0} + o_P(1). \tag{12}$$

By invertability of V_{θ_o}, we have

$$\begin{aligned} \sqrt{n}||\hat{\theta}_n - \theta_o|| &\leq ||V_{\theta_0}^{-1}||\sqrt{n}||V_{\theta_0}(\hat{\theta}_n - \theta_0)|| \quad (13) \\ &= O_P(1) + o_P(\sqrt{n}||\hat{\theta}_n - \theta_0||). \quad (14) \end{aligned}$$

Inequality (14) is obtained by plugging (12) into (13) and using triangle inequality. Therefore, we have

$$\hat{\theta}_n \text{ is } \sqrt{n} - \text{consistent}. \quad (15)$$

By (12) and (15), we have

$$\sqrt{n}V_{\theta_0}(\hat{\theta}_n - \theta_0) = -G_n\psi_{\theta_0} + o_P(1). \quad (16)$$

Multiply both side by $V_{\theta_0}^{-1}$ to get the result. □

References

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

Stat210B: Theoretical Statistics **Lecture Date: 3/8/2007**

Lecture 15: Asymptotic Testing

Lecturer: Michael I. Jordan *Scribe: Dapo Omidiran*

Reading: Chapter 7 of van der Vaart (1998).

1 Asymptotic Testing

Setup: We are given:

- A parametric model: $P_\theta, \theta \in \Theta$
- A null hypothesis: $\theta = \theta_0$
- An alternative hypothesis: $\theta = \theta_1$

Our test then consists of computing the log likelihood ratio:

$$\lambda = \log \left(\frac{\prod_i p_{\theta_1}(X_i)}{\prod_i p_{\theta_0}(X_i)} \right),$$

and accepting the alternative hypothesis if λ is sufficiently large.

Example 1. (Normal location model)

Let $P_\theta = N(\theta, \sigma^2)$, with σ^2 known. After some algebra, we see that

$$\lambda = \frac{n}{\sigma^2}[(\theta_1 - \theta_0)\bar{X}_n - \frac{1}{2}(\theta_1^2 - \theta_0^2)].$$

We can study the distribution under each hypothesis.

Under θ_0, we can use the WLLN to conclude:

$$\lambda \xrightarrow{P} -\frac{n}{\sigma^2}\frac{1}{2}(\theta_1 - \theta_0)^2 \to -\infty$$

Notice that this a good thing. Asymptotically, we will never reject the null hypothesis; our test is "consistent". However, this is also somewhat vacuous, as almost any reasonable test will give the same result.

We should instead look at the rates at which our test converges. One approach is to use large deviations (pioneered by Hoeffding in the '60s?) However, we won't go that route. Instead, we will "shrink" θ towards θ_0 as n increases (e.g., $\theta_1 = \theta_0 + \frac{h}{\sqrt{n}}$.)

In some sense, this $\sqrt{n}$ behavior is the right shrinkage factor for "regular" data, such as iid data.

This approach was first developed for testing, but is applicable to estimation as well.

So, let's study shrinking alternatives:

Example 2. (Normal location model revisited)

$\lambda = h\sqrt{n}\frac{\bar{X_n}-\theta_0}{\sigma^2} - \frac{h^2}{2\sigma^2} = h\bar{Z_n} - \frac{h^2}{2\sigma^2}$ (where $Z_n = \sqrt{n}\frac{\bar{X_n}-\theta_0}{\sigma^2} \overset{H_0}{\sim} N(0, \frac{1}{\sigma^2})$)

Note that this is a quadratic in h. Hence:

$$\lambda \sim N(-\frac{h^2}{2\sigma^2}, \frac{h^2}{\sigma^2})$$

The mean is -$\frac{1}{2}$ the variance!

Is this behavior specific to the Normal distribution? Let's check the exponential family:

Example 3. (Exponential Family)

$$\begin{aligned} p_\theta(x) &= \exp\left[\theta T(x) - A(\theta)\right] \\ \lambda &= hn^{-\frac{1}{2}}\sum_i T(X_i) - n[A(\theta_0 + hn^{-\frac{1}{2}}) - A(\theta_0)] \\ &= hn^{-\frac{1}{2}}\sum_i T(X_i) - n[A'(\theta_0)hn^{-\frac{1}{2}} + \frac{1}{2}A''(\theta_0)h^2n^{-1} + o(n^{-1})] \\ &= hZ_n - \frac{1}{2}h^2A''(\theta_0) + o(1) \end{aligned}$$

Where $Z_n = n^{\frac{1}{2}}\sum_i T(X_i) - E_{\theta_0}[T(X_i)]$ (As $A'(\theta_0) = E_{\theta_0}[T(X_i)]$).

Asymptotically, the mean is again -1/2 the variance.

How much further can we go?

The key property is quadratic mean differentiability (QMD), essentially a notion of smoothness relevant for asymptotic statistics.

In particular, we want a smoothness condition. However, we are constrained by the following:

- We want to avoid assuming that derivatives exist for all x (i.e., for each x, a derivative exists at each value of θ)
- We also want to avoid explicit conditions on higher derivatives.

Solution: We will work with square roots of densities. Classical (Frechet) differentiability of $\sqrt{p_\theta}$ (Again, note that x is held fixed, and θ is the variable):

$$\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - h^T\eta_{\theta_0}(x) = o(||h||).$$

To weaken this somewhat stringent condition, we only ask that it hold in the quadratic mean:

Definition 4. QMD

P_θ is QMD at θ_0 if

$$\int\left(\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2}h^T\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}\right)^2 d\mu = o(||h^2||)$$

for some function $\dot{\ell}_{\theta_0}$.

Keep in mind that these (p_{θ_0}, p_θ, etc) are all functions of x. $\dot{\ell}_{\theta_0}$ is *not* the derivative of some ℓ_{θ_0}, but is instead some function.

Why do we define things in this weird way? If classical derivatives *do* exist:

$$\frac{\partial}{\partial\theta}\sqrt{p_\theta} = \frac{1}{2}\sqrt{p_\theta}\frac{\partial}{\partial\theta}\log p_\theta.$$

So we associate $\dot{\ell}_\theta$ is the score function, in this case.

Theorem 5 (Theorem 7.2, van der Vaart (1998) p94)**.** *If Θ is an open subset of $\mathcal{R}^K$ and $P_\theta,\ \theta \in \Theta$ is QMD. Then:*

- $P_\theta \dot{\ell}_\theta = 0$ *(Like score functions),*
- *and* $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ *exists (Fisher information),*
- *and* $\lambda = \dfrac{\prod_i p_{\theta+\frac{h_n}{\sqrt{n}}}(X_i)}{\prod_i p_\theta(X_i)} = \frac{1}{\sqrt{n}}\sum_i h^T\dot{\ell}_\theta(X_i) - \frac{1}{2}h^T I_\theta h + o_{p_\theta}(1)$.

Where $h_n \to h \neq 0$. Note that this implies that:

$$\lambda \xrightarrow{d} N(-\frac{1}{2}h^T I_\theta h, h^T I_\theta h)\ .$$

Proof. (Partial Proof) Let

$$\begin{aligned} p_n &= p_\theta + \frac{h_n}{\sqrt{n}}\ , \\ p &= p_\theta\ , \\ g &= h^T\dot{\ell}_\theta\ . \end{aligned}$$

By the definition of QMD, it follows that:

$$\begin{aligned} & \int (\sqrt{p_n} - \sqrt{p} - \frac{1}{2}g\sqrt{p})^2 d\mu = o(n^{-1})\ , \\ \implies\ & n^{1/2}(\sqrt{p_n} - \sqrt{p}) \xrightarrow{QM} \frac{1}{2}g\sqrt{p}\ , \\ \implies\ & \sqrt{p_n} \xrightarrow{QM} \sqrt{p}\ . \end{aligned}$$

We recall that $\int f_n g_n \to \int fg$ if $f_n \to$ f and $g_n \to g$.

By continuity of the inner product:

$$Pg = \int gpd\mu = \int \frac{1}{2}g\sqrt{p}2\sqrt{p}d\mu = \lim_{n\to\infty}\int \sqrt{n}(\sqrt{p_n} - \sqrt{p})(\sqrt{p_n} + \sqrt{p})d\mu = 0\ .$$

Define:

$$W_{n,i} = 2\left(\frac{\sqrt{p_n(X_i)}}{\sqrt{p(X_i)}} - 1\right)$$

We use the fact that $\log(1+x) = x - \frac{1}{2}x^2 + x^2R(x)$ (where R(x) $\to$ 0 as x$\to$ 0) to conclude that:

$$\log \prod_i p_n(X_i)/p(X_i) = 2\sum_i \log(1 + \frac{1}{2}W_{n,i}) = \sum_i W_{n,i} - \frac{1}{4}\sum_i W_{n,i}^2 + \frac{1}{2}\sum_i W_i^2 R(W_{n,i}) \ .$$

As:

$$E_p\left(\sum_i W_{n,i}\right) = 2n\left(\int \sqrt{p_n}\sqrt{p}d\mu - 1\right) = -n\int (\sqrt{p_n} - \sqrt{p})^2 d\mu \to -\int \frac{1}{4}g^2 p d\mu \ ,$$

where $Pg^2 = \int \frac{1}{4}g^2 p d\mu = h^T(\int \dot{\ell}_\theta \dot{\ell}_\theta^T dP)h = h^T I_\theta h$.

Look at the remainder of the proof in van der Vaart (1998). □

References

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

Stat210B: Theoretical Statistics **Lecture Date: March 13, 2007**

Quadratic Mean Differentiability and Contiguity

Lecturer: Michael I. Jordan *Scribe: Maximilian Kasy*

Lemma 1 (Sufficient conditions for QMD). *Fix $\theta_0 \in \Theta \backslash \partial\Theta$.*

- *assume $p_\theta^{1/2}$ is an absolutely continuous function of θ in some neighbourhood of θ_0 for μ-almost all x.*
- *assume the derivative p'_θ at θ_0 exists for μ-almost all x.*
- *asume that the Fisher information exists and is continuous at θ_0.*

Then p_θ is QMD.

Example 2.
- exponential families
- location families $p_\theta(x) = f(x-\theta)$ where $f^{1/2}$ is absolutely continuous and f' exists almost everywhere. Note: $I = -\mathbb{E}\left(\left(\frac{f'(x-\theta)}{f(x-\theta}\right)^2\right)$ exists!
- e.g. the Cauchy location model $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$
- e.g. the Laplace location model $f(x) = \frac{1}{2}e^{-|x|}$

Example 3. A family that is not QMD: The $Uniform[0,\theta]$ distributions. Proof:

$$\int \left(\sqrt{n}\left(p_{\theta+h/\sqrt{n}}^{1/2}(x) - p_\theta^{1/2}(x)\right)\right)^2 \geq n\left(\int_\theta^{\theta+h/\sqrt{n}} \frac{1}{\theta+h/\sqrt{n}}dx\right) \to \infty \; as \; n \to \infty$$

Reminder of the definition of absolute continuity: $Q << P$ iff $P(A) = 0 \Rightarrow Q(A) = 0, \quad \forall A$

Theorem 4 (Radon Nikodym).

$$Q << P \Rightarrow \exists g : Q(A) = \int_A g \, dP$$

Write: $g =: \frac{dQ}{dP}$
Likelihood ratio=Radon Nikodym derivative=g

Lemma 5 (Lemma 6.2, van der Vaart (1998) p86). *Let P and Q have densities p and q w.r.t. μ. Then we can write:*

$$Q = Q^a + Q^\perp, \quad where \tag{1}$$
$$Q^a(A) = Q(A \cap p > 0) \tag{2}$$
$$Q^\perp(A) = Q(A \cap p = 0) \tag{3}$$

With this notation we have:

1. $Q^a << P, Q^\perp \perp P$
2. $Q^a(A) = \int_A (q/p)dP$ *for all measurable sets* A
3. $Q << P$ *iff* $Q(p=0)=0$ *iff* $\int (q/p)dP = 1$

Corollary 6 (change of measure). *Let* $Q << P$, *let* $P^{X,V}$ *denote the law of the pair* $(X,V) = (X, \frac{dQ}{dP})$ *under* P. *Then*

$$Q(X \in B) = \mathbb{E}_P\left[1_B(X)\frac{dQ}{dP}\right] = \int_{B\times\mathbb{R}} v dP^{X,V}(x,v)$$

Definition 7 (Contiguity). Q_n is contiguous with respect to P_n, in symbols $Q_n \triangleleft P_n$, if $\forall\{A_n\} : P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$

Example 8 (Absolute continuity does not imply contiguity). Let $P_n = N(0,1)$, $Q_n = N(\xi_n, 1)$, $\xi_n \to \infty$, $A_n = \{x : |x - \xi_n| < 1\}$. Then $Q_n << P_n$ but not $Q_n \triangleleft P_n$.

Example 9 (Contiguity does not imply absolute continuity). Let $P_n = Uni[0,1]$, $Q_n = Uni[0,\theta_n]$, $\theta_n \to 1$, $\theta_n > 1$. Then $Q_n \triangleleft P_n$ but not $Q_n << P_n$.

$$\mathbb{E}_{P_n}\left[\frac{dQ_n}{dP_n}\right] = \int \frac{q_n}{p_n} p_n d\mu = \int_{p_n>0} q_n d\mu = Q_n\{x : p_n > 0\} \le 1$$

Hence $\frac{dQ_n}{dP_n}$ is uniformly tight. Prohorov's theorem then implies that for all subsequences of $\frac{dQ_n}{dP_n}$ there exists a further weakly converging subsequence. As the following lemma shows, the limit points determine contiguity.

Lemma 10 (Le Cam's first lemma, Lemma 6.4, van der Vaart (1998) p88). *The following statements are equivalent:*

- $Q_n \triangleleft P_n$
- *If* $\frac{dQ_n}{dP_n} \overset{P_n}{\rightsquigarrow} V$ *along a subsequence, then* $\mathbb{E}V = 1$
- $T_n \overset{P_n}{\to} 0$ *implies* $T_n \overset{Q_n}{\to} 0$

Corollary 11 (Asymptotic log normality, van der Vaart (1998) p89). *Suppose* $\frac{dQ_n}{dP_n} \overset{P_n}{\rightsquigarrow} \exp\left[N(\mu,\sigma^2)\right]$. *Then* $Q_n \triangleleft\triangleright P_n$ *iff* $\mu = -\frac{1}{2}\sigma^2$.

Proof. Idea of proof: Let $Z \sim N(\mu,\sigma^2)$. By Le Cam's first lemma, we need $\mathbb{E}e^Z = 1$ for $Q_n \triangleleft P_n$. But $\mathbb{E}e^Z = \exp(\mu + \frac{1}{2}\sigma^2)$. (Characteristic functions!) □

Example 12. Let $P_n = N(0,1)$, $Q_n = N(\xi_n, 1)$. Then $\frac{dQ_n}{dP_n} = \exp[\xi_n x - \frac{1}{2}\xi_n^2]$. This converges if $\xi_n \to \xi$ with $|\xi| < \infty$ which yields $\exp[\xi x - \frac{1}{2}\xi^2]$ in the limit, hence we get contiguity for $|\xi| < \infty$

Example 13. Let $X_i \overset{i.i.d.}{\sim} N(\xi,1)$, let P_n be the joint distribution for $\xi = 0, i = 1\dots n$ and Q_n for $\xi = \xi_n, i = 1\dots n$. Then $\log\frac{dQ_n}{dP_n} = \xi_n \sum_i X_i - \frac{n\xi_n^2}{2}$, hence $\frac{dQ_n}{dP_n} \sim \exp\{N(-\frac{n\xi_n^2}{2}, n\xi_n^2)\}$. Therefore we need $\xi_n = O(n^{-1/2})$.

Example 14 (QMD families, Theorem 7.2, van der Vaart (1998) p94).

$$\log\frac{dP_{\theta_0+h/\sqrt{n}}}{dP_{\theta_0}} = \frac{1}{\sqrt{n}}\sum_i h^t \dot{l}_{\theta_0}(X_i) - \frac{1}{2}h^t I_{\theta_0} h + o_{P_{\theta_0}}(1)$$

We get mean $= -\frac{1}{2}$ variance in the limit, i.e. for qmd families $P_{\theta_0+h/\sqrt{n}} \triangleleft P_{\theta_0}$.

References

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

Stat210B: Theoretical Statistics **Lecture Date: March 15, 2007**

Change of Measure and Contiguity

Lecturer: Michael I. Jordan *Scribe: Aria Haghighi*

In the last lecture, we discussed *contiguity* of measure as the analogue of absolute continuity for asymptotic statistics. In this lecture, we will use contiguity to establish results change-of-measure results for statistical hypothesis testing. We briefly recall the definition of contiguity here,

Definition 1 (Contiguity)**.** Let Q_n and P_n be sequences of measures. We say that Q_n is contiguous w.r.t to P_n, denoted $Q_n \triangleleft P_n$, if for each sequence of measurable sets A_n,[1] we have that

$$P_n(A_n) \to 0 \Rightarrow Q_n(A_n) \to 0$$

We also showed that $Q_n \triangleleft P_n$ if and only if whenever the Radon-Nikodym derivative, $\frac{dQ_n}{dP_n}$, converges weakly under P_n to a random variable V (i.e.$\frac{dQ_n}{dP_n} \overset{P_n}{\rightsquigarrow} V$), then we have $EV = 1$.[2] We also saw that a distribution being in the Quadratic Mean Derivative (QMD) family implied contiguity for shrinking alternatives in statistical testing. Formally, for QMD families P_θ, we have that $P^n_{\theta_0 + \frac{h}{\sqrt{n}}} \triangleleft P^n_{\theta_0}$ (by Theorem 7.2 in van der Vaart (1998) pg. 94). We now state an important result regarding the joint distribution of test statistics and the likelihood ratio:

Lemma 2 (Theorem 6.6 in van der Vaart (1998) pg. 90)**.** *Let P_n and Q_n be sequences of measures such that $Q_n \triangleleft P_n$. Let X_n be a sequence of test statistic random variables. Suppose that we have,*

$$\left(X_n, \frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightsquigarrow} (X, V)$$

for limiting random variables X and V. Then we have that $L(B) = E\mathbf{1}_B(X)V$ defines a measure. Furthermore, $X_n \overset{Q_n}{\rightsquigarrow} L$.

Proof. By contiguity, we have that $EV = 1$, which in turn implies that L must be a probability measure. Using Portmanteau's lemma and a standard induction over measurable functions gives that $X_n \overset{Q_n}{\rightsquigarrow} L$. □

Typically, we have that (X, V) is bi-variate normal. In this case we have a very appealing result about the asymptotic distribution of the test statistic under Q_n.

Lemma 3 (LeCan's Third Lemma, pg. 90 van der Vaart (1998))**.** *Suppose that*

$$(X_n, \log \frac{dQ_n}{dP_n}) \overset{P_n}{\rightsquigarrow} \mathcal{N}\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right),$$

where τ and σ are scalars.[3] *Then we have that*

$$X_n \overset{Q_n}{\rightsquigarrow} \mathcal{N}(\mu + \tau, \Sigma)$$

[1] Where measurable means with respect to the underlying Borel set of Q_n, which may change with n.

[2] Note that by Prohorov's theorem that $\frac{dQ_n}{dP_n}$ has a convergent subsequence so the theorem isn't vacuous.

[3] Note that we have that the mean of $\log \frac{dQ_n}{dP_n}$ must be $-\frac{1}{2}\sigma^2$.

This lemma shows that under the alternative distribution Q_n, the limiting distribution of the test statistic X_n is also normal but has a mean shifted by $\tau = \lim_{n\to\infty} \mathrm{Cov}(X_n, \log \frac{dQ_n}{dP_n})$.

Proof. Suppose that (X, W) be the limiting distribution on the RHS of the above. By the continuous mapping theorem we have that,

$$(X_n, \frac{dQ_n}{dP_n}) \overset{P_n}{\rightsquigarrow} (X, e^W)$$

Since we have that $W \sim \mathcal{N}(-\frac{1}{2}\sigma^2, \sigma^2)$, we have that $Q_n \triangleleft P_n$. We have by theorem 6.4 then, that X_n converges weakly to L under Q_n, where $L = E\mathbf{1}_B(X)e^W$. We are going to determine the distribution of L via it's characteristic function,[4]

$$\begin{aligned}
\int e^{it^T x} dL(x) &= E\left[e^{it^T X + W}\right] \\
&= E\left[e^{it^T X + i(-i)W}\right] \\
&= \exp\left\{it^T\mu - \frac{1}{2}\sigma^2 - \frac{1}{2}(t^T, -i)\begin{pmatrix}\Sigma & \tau \\ \tau^T & \sigma^2\end{pmatrix}\begin{pmatrix}t \\ -i\end{pmatrix}\right\} \\
&= e^{it^T(\mu+\tau) - \frac{1}{2}t^T\Sigma t} \\
&\Rightarrow L \sim \mathcal{N}(\mu + \tau, \Sigma)
\end{aligned}$$

□

where the last line is obtained by recognizing the form of the RHS of the previous equation as the characteristic function of the normal distribution.

Example 4 (Asymptotically Linear Statistics)**.** Suppose that P_θ is a family of QMD measures. We are interested in the asymptotic behavior of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. We will consider the following setting,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}}\sum_i \psi_{\theta_0}(X_i) + o_P(1)$$

where $\mathrm{Var}_{\theta_0}\psi_{\theta_0}(X) = \tau^2 < \infty$ and $E_{\theta_0}\psi_{\theta_0} = 0$. Furthermore, we assume that under H_0 (i.e. when $\theta = \theta_0$), we have by the CLT that,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \tau^2)$$

Since P_θ is in the QMD family, we have the following expression,

$$\left(\sqrt{n}(\hat{\theta}_n - \theta_0), \frac{dP_{\theta_0 + \frac{h}{\sqrt{n}}}}{dP_{\theta_0}}\right) = \left(\frac{1}{\sqrt{n}}\sum_i \left[\psi_{\theta_0}(X_i), h^T\dot{\ell}_{\theta_0}(X_i)\right] + \left[0, -\frac{1}{2}h^T I_{\theta_0} h\right] + o_P(1)\right)$$

Using the bivariate CLT, we have that the RHS above converges to a normal distribution where the covariance between $\sqrt{n}(\hat{\theta}_n - \theta_0)$ and $\frac{dP_{\theta_0+\frac{h}{\sqrt{n}}}}{dP_{\theta_0}}$ is given by $\tau = \mathrm{Cov}_{\theta_0}(\psi_{\theta_0}(X), h^T\dot{\ell}_{\theta_0}(X))$.

Our next example builds upon the previous one:

Example 5 (T-Statitic for Location Families)**.** Suppose that $f(X - \theta)$ is a density for a QMD location family. We are interested in testing $\theta_0 = 0$. We define the t-statitic as,

$$t_n = \sqrt{n}\frac{\overline{X}_n}{S_n} = \sqrt{n}\frac{\overline{X}_n}{\sigma} + o_{P_{\theta_0}}(1)$$

[4] Which uniquely determines a distribution.

where the second equality uses a delta method argument. This yields that the t-statistic is an asymptotic linear statistic as in example 4. We are interested in the behavior of t_n under the alternative $\theta_h = \frac{h}{\sqrt{n}}$. Recall that, $\dot{\ell}_{\theta_0} = -\frac{f'(x)}{f(x)}$. Using example 4 and the fact that $\psi_{\theta_0}(X_i) = \frac{X_i}{\sigma}$, we have that

$$\begin{aligned}\tau &= -\frac{h}{\sigma}\text{Cov}(X_i, \frac{f'_{\theta_0}(X_i)}{f_{\theta_0}(X_i)}) \\ &= -\frac{h}{\sigma}\int x\frac{f'}{f}df = -\frac{h}{\sigma}\int xf'dx \\ &= \frac{h}{\sigma}, \text{ using integration by part.}\end{aligned}$$

We therefore have that under shrinking alternatives, $t_n \overset{\frac{h}{\sqrt{n}}}{\rightsquigarrow} \mathcal{N}(\frac{h}{\sigma}, 1)$.

Example 6 (Sign Test for Location Families)**.** We suppose again that $f(X-\theta)$ is a density for QMD family of distributions. We also suppose that $f(\cdot)$ is continuous at the origin and that $P_{\theta=0}(X > 0) = \frac{1}{2}$. We define the sign statistic,

$$s_n = \frac{1}{\sqrt{n}}\sum_i (1_{X>0} - \frac{1}{2})$$

We again suppose we are interested in testing whether $\theta_0 = 0$. Under the alternative hypothesis $\theta_h = \frac{h}{\sqrt{n}}$, we have

$$\begin{aligned}\tau &= -h\text{Cov}_{\theta_0}(1_{X>0}, \frac{f'(X)}{f(X)}) \\ &= -h\int_0^\infty f'(X)dx = hf(0)\end{aligned}$$

Under the alternative hypothesis, the asymptotic distribution of s_n is normal with mean $hf(0)$.

References

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