Theoretical Statistics. Lecture 1.

Peter Bartlett

1. Organizational issues.
2. Overview.
3. Stochastic convergence.

Organizational Issues

- Lectures: Tue/Thu 11am–12:30pm, 332 Evans.
- Peter Bartlett. bartlett@stat. Office hours: Tue 1-2pm, Wed 1:30-2:30pm (Evans 399).
- GSI: Siqi Wu. siqi@stat. Office hours: Mon 3:30-4:30pm, Tue 3:30-4:30pm (Evans 307).
- http://www.stat.berkeley.edu/∼bartlett/courses/210b-spring2013/
  Check it for announcements, homework assignments, ...
- Texts:
  Asymptotic Statistics, Aad van der Vaart. Cambridge. 1998.
  Convergence of Stochastic Processes, David Pollard. Springer. 1984.
  Available on-line at
  `http://www.stat.yale.edu/∼pollard/1984book/`

Organizational Issues

- **Assessment:**

  Homework Assignments (60%): posted on the website.

  Final Exam (40%): scheduled for Thursday, 5/16/13, 8-11am.

- Required background:

  Stat 210A, and either Stat 205A or Stat 204.

Asymptotics: Why?

Example: We have a sample of size n from a density p_θ. Some estimator gives $\hat{\theta}_n$.

- Consistent? i.e., $\hat{\theta}_n \to \theta$? Stochastic convergence.
- Rate? Is it optimal? Often no finite sample optimality results. Asymptotically optimal?
- Variance of estimate? Optimal? Asymptotically?
- Distribution of estimate? Confidence region. Asymptotically?

Asymptotics: Approximate confidence regions

Example: We have a sample of size n from a density p_θ. Maximum likelihood estimator gives $\hat{\theta}_n$.

Under mild conditions, $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$ is asymptotically $N\left(0, I_\theta^{-1}\right)$. Thus $\sqrt{n} I_\theta^{1/2}(\hat{\theta}_n - \theta) \sim N(0, I)$, and $n(\hat{\theta}_n - \theta)^T I_\theta(\hat{\theta}_n - \theta) \sim \chi^2(k)$.

So we have an approximate $1 - \alpha$ confidence region for θ:

$$\left\{\theta : (\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n}(\theta - \hat{\theta}_n) \leq \frac{\chi^2_{k,\alpha}}{n}\right\}.$$

Overview of the Course

1. Tools for consistency, rates, asymptotic distributions:
   - Stochastic convergence.
   - Concentration inequalities.
   - Projections.
   - U-statistics.
   - Delta method.
2. Tools for richer settings (eg: function space vs $\mathbb{R}^k$)
   - Uniform laws of large numbers.
   - Empirical process theory.
   - Metric entropy.
   - Functional delta method.

3. Tools for asymptotics of likelihood ratios:
   - Contiguity.
   - Local asymptotic normality.
4. Asymptotic optimality:
   - Efficiency of estimators.
   - Efficiency of tests.
5. Applications:
   - Nonparametric regression.
   - Nonparametric density estimation.
   - M-estimators.
   - Bootstrap estimators.

Convergence in Distribution

$X_1, X_2, \ldots, X$ are random vectors,

Definition: X_n **converges in distribution** (or **weakly converges**) to X (written $X_n \rightsquigarrow X$) means that their distribution functions satisfy $F_n(x) \rightarrow F(x)$ at all continuity points of F.

Review: Other Types of Convergence

d is a distance on $\mathbb{R}^k$ (for which the Borel σ-algebra is the usual one).

Definition: X_n **converges almost surely** to X (written $X_n \overset{as}{\to} X$) means that $d(X_n, X) \to 0$ a.s.

Definition: X_n **converges in probability** to X (written $X_n \overset{P}{\to} X$) means that, for all $\epsilon > 0$,

$$P\left(d(X_n, X) > \epsilon\right) \to 0.$$

Review: Other Types of Convergence

Theorem:

$$X_n \stackrel{as}{\to} X \Longrightarrow X_n \stackrel{P}{\to} X \Longrightarrow X_n \rightsquigarrow X,$$

$$X_n \stackrel{P}{\to} c \Longleftrightarrow X_n \rightsquigarrow c.$$

NB: For $X_n \stackrel{as}{\to} X$ and $X_n \stackrel{P}{\to} X$, X_n and X must be functions on the sample space of the same probability space. But not convergence in distribution.

Convergence in Distribution: Equivalent Definitions

Theorem: [Portmanteau] The following are equivalent:

1. $P(X_n \le x) \to P(X \le x)$ for all continuity points x of $P(X \le \cdot)$.
2. $\mathbf{E} f(X_n) \to \mathbf{E} f(X)$ for all bounded, continuous f.
3. $\mathbf{E} f(X_n) \to \mathbf{E} f(X)$ for all bounded, Lipschitz f.
4. $\mathbf{E} e^{it^T X_n} \to \mathbf{E} e^{it^T X}$ for all $t \in \mathbb{R}^k$. (Lévy's Continuity Theorem)
5. for all $t \in \mathbb{R}^k$, $t^T X_n \rightsquigarrow t^T X$. (Cramér-Wold Device)
6. $\liminf \mathbf{E} f(X_n) \ge \mathbf{E} f(X)$ for all nonnegative, continuous f.
7. $\liminf P(X_n \in U) \ge P(X \in U)$ for all open U.
8. $\limsup P(X_n \in F) \le P(X \in F)$ for all closed F.
9. $P(X_n \in B) \to P(X \in B)$ for all continuity sets B (i.e., $P(X \in \partial B) = 0$).

Convergence in Distribution: Equivalent Definitions

Example: [Why do we need continuity?]

Consider $f(x) = 1[x > 0]$, $X_n = 1/n$. Then $X_n \to 0$, $f(x) \to 1$, but $f(0) = 0$.

[Why do we need boundedness?]

Consider $f(x) = x$,

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

Then $X_n \rightsquigarrow 0$, $\mathbf{E}f(X_n) \to 1$, but $f(0) = 0$.

Relating Convergence Properties

Theorem:

$$X_n \rightsquigarrow X \text{ and } d(X_n, Y_n) \xrightarrow{P} 0 \Longrightarrow Y_n \rightsquigarrow X,$$

$$X_n \rightsquigarrow X \text{ and } Y_n \rightsquigarrow c \Longrightarrow (X_n, Y_n) \rightsquigarrow (X, c),$$

$$X_n \xrightarrow{P} X \text{ and } Y_n \xrightarrow{P} Y \Longrightarrow (X_n, Y_n) \xrightarrow{P} (X, Y).$$

Relating Convergence Properties

Example: NB: **NOT** $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y \Longrightarrow (X_n, Y_n) \rightsquigarrow (X, Y)$.
(joint convergence versus marginal convergence in distribution)
Consider X, Y independent $N(0,1)$, $X_n \sim N(0,1)$, $Y_n = -X_n$. Then $X_n \rightsquigarrow X$, $Y_n \rightsquigarrow Y$, but $(X_n, Y_n) \rightsquigarrow (X, -X)$, which has a very different distribution from that of (X, Y).

Relating Convergence Properties: Continuous Mapping

Suppose $f : \mathbb{R}^k \to \mathbb{R}^m$ is "almost surely continuous" (i.e., for some S with $P(X \in S)$=1, f is continuous on S).

Theorem: [Continuous mapping]

$$X_n \rightsquigarrow X \Longrightarrow f(X_n) \rightsquigarrow f(X).$$

$$X_n \stackrel{P}{\to} X \Longrightarrow f(X_n) \stackrel{P}{\to} f(X).$$

$$X_n \stackrel{as}{\to} X \Longrightarrow f(X_n) \stackrel{as}{\to} f(X).$$

Relating Convergence Properties: Continuous Mapping

Example: For $X_1, \ldots, X_n$ i.i.d. mean μ, variance σ^2, we have

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \rightsquigarrow N(0,1).$$

So

$$\frac{n}{\sigma^2}(\bar{X}_n - \mu)^2 \rightsquigarrow (N(0,1))^2 = \chi_1^2.$$

Example: We also have $\bar{X}_n - \mu \rightsquigarrow 0$ hence $(\bar{X}_n - \mu)^2 \rightsquigarrow 0$. Consider $f(x) = 1[x > 0]$. Then $f((\bar{X}_n - \mu)^2) \rightsquigarrow 1 \neq f(0)$.
(The problem is that f is not continuous at 0, and $P_X(0) > 0$, for X satisfying $(\bar{X}_n - \mu)^2 \rightsquigarrow X$.)

Relating Convergence Properties: Slutsky's Lemma

Theorem: $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$ imply

$$X_n + Y_n \rightsquigarrow X + c,$$

$$Y_n X_n \rightsquigarrow cX,$$

$$Y_n^{-1} X_n \rightsquigarrow c^{-1} X.$$

(Why does $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ not imply $X_n + Y_n \rightsquigarrow X + Y$?)

Relating Convergence Properties: Examples

Theorem: For i.i.d. Y_t with $\mathbf{E}Y_1 = \mu$, $\mathbf{E}Y_1^2 = \sigma^2 < \infty$,

$$\sqrt{n}\frac{\bar{Y}_n - \mu}{S_n} \rightsquigarrow N(0,1),$$

where

$$\bar{Y}_n = n^{-1}\sum_{i=1}^{n} Y_i,$$

$$S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2.$$

Proof:

$$S_n^2 = \underbrace{\frac{n}{n-1}}_{\xrightarrow{P} 1} \left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}_{\xrightarrow{P} \mathbf{E}Y_1^2} - \left(\underbrace{\bar{Y}_n}_{\xrightarrow{P} \mathbf{E}Y_1} \right)^2 \right)$$

(weak law of large numbers)

$$\xrightarrow{P} \mathbf{E}Y_1^2 - (\mathbf{E}Y_1)^2$$

(continuous mapping theorem, Slutsky's Lemma)

$$= \sigma^2.$$

Also

$$\underbrace{\sqrt{n}\left(\bar{Y}_n - \mu\right)}_{\rightsquigarrow N(0,\sigma^2)} \underbrace{\frac{1}{S_n}}_{\xrightarrow{P} 1/\sigma}$$

(central limit theorem)

$$\rightsquigarrow N(0,1)$$

(continuous mapping theorem, Slutsky's Lemma)

Showing Convergence in Distribution

Recall that the **characteristic function** demonstrates weak convergence: $X_n \rightsquigarrow X \iff \mathbf{E}e^{it^TX_n} \to \mathbf{E}e^{it^TX}$ for all $t \in \mathbb{R}^k$.

Theorem: [Lévy's Continuity Theorem]
If $\mathbf{E}e^{it^TX_n} \to \phi(t)$ for all t in $\mathbb{R}^k$, and $\phi : \mathbb{R}^k \to \mathbb{C}$ is continuous at 0, then $X_n \rightsquigarrow X$, where $\mathbf{E}e^{it^TX} = \phi(t)$.

Special case: $X_n = Y$. So X, Y have same distribution iff $\phi_X = \phi_Y$.

Showing Convergence in Distribution

Theorem: [Weak law of large numbers]

Suppose $X_1, \ldots, X_n$ are i.i.d. Then $\bar{X}_n \xrightarrow{P} \mu$ iff $\phi'_{X_1}(0) = i\mu$.

Proof:

We'll show that $\phi'_{X_1}(0) = i\mu$ implies $\bar{X}_n \xrightarrow{P} \mu$. Indeed,

$$\begin{aligned}
\mathbf{E}e^{it\bar{X}_n} &= \phi^n(t/n) \\
&= (1 + ti\mu/n + o(1/n))^n \\
&\to \underbrace{e^{it\mu}}_{=\phi_\mu(t)}.
\end{aligned}$$

Lévy's Theorem implies $\bar{X}_n \rightsquigarrow \mu$, hence $\bar{X}_n \xrightarrow{P} \mu$.

Showing Convergence in Distribution

e.g., $X \sim N(\mu, \Sigma)$ has characteristic function

$$\phi_X(t) = \mathbf{E}e^{it^T X} = e^{it^T\mu - t^T\Sigma t/2}.$$

Theorem: [Central limit theorem]
Suppose $X_1, \ldots, X_n$ are i.i.d., $\mathbf{E}X_1 = 0$, $\mathbf{E}X_1^2 = 1$. Then $\sqrt{n}\bar{X}_n \rightsquigarrow N(0, 1)$.

Proof: $\phi_{X_1}(0) = 1$, $\phi'_{X_1}(0) = i\mathbf{E}X_1 = 0$, $\phi''_{X_1}(0) = i^2\mathbf{E}X_1^2 = -1$.

$$\begin{aligned}
\mathbf{E}e^{it\sqrt{n}\bar{X}_n} &= \phi^n(t/\sqrt{n}) \\
&= \left(1 + 0 - t^2\mathbf{E}Y^2/(2n) + o(1/n)\right)^n \\
&\to e^{-t^2/2} \\
&= \phi_{N(0,1)}(t).
\end{aligned}$$

Uniformly tight

Definition:
X is **tight** means that for all $\epsilon > 0$ there is an M for which

$$P(\|X\| > M) < \epsilon.$$

$\{X_n\}$ is **uniformly tight** (or **bounded in probability**) means that for all $\epsilon > 0$ there is an M for which

$$\sup_n P(\|X_n\| > M) < \epsilon.$$

(so there is a compact set that contains each X_n with high probability.)

Notation: Uniformly tight

Theorem: [Prohorov's Theorem]

1. $X_n \rightsquigarrow X$ implies $\{X_n\}$ is uniformly tight.
2. $\{X_n\}$ uniformly tight implies that for some X and some subsequence, $X_{n_j} \rightsquigarrow X$.

Notation for rates: o_P, O_P

Definition:

$$X_n = o_P(1) \iff X_n \xrightarrow{P} 0,$$

$$X_n = o_P(R_n) \iff X_n = Y_n R_n \text{ and } Y_n = o_P(1).$$

$$X_n = O_P(1) \iff X_n \text{ uniformly tight}$$

$$X_n = O_P(R_n) \iff X_n = Y_n R_n \text{ and } Y_n = O_P(1).$$

(i.e., o_P, O_P specify *rates* of growth of a sequence. o_P means strictly slower (sequence Y_n converges in probability to zero). O_P means within some constant (sequence Y_n lies in a ball).

Relations between rates

$$o_P(1) + o_P(1) = o_P(1).$$

$$o_P(1) + O_P(1) = O_P(1).$$

$$o_P(1) O_P(1) = o_P(1).$$

$$(1 + o_P(1))^{-1} = O_P(1).$$

$$o_P(O_P(1)) = o_P(1).$$

$$X_n \xrightarrow{P} 0,\ R(h) = o(\|h\|^p) \Longrightarrow R(X_n) = o_P(\|X_n\|^p).$$

$$X_n \xrightarrow{P} 0,\ R(h) = O(\|h\|^p) \Longrightarrow R(X_n) = O_P(\|X_n\|^p).$$