ECON 481-3
LECTURE 9: CONVOLUTION THEOREMS

Ivan A. Canay
Northwestern University

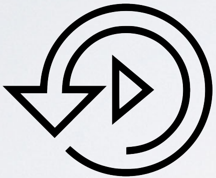
Past & Future

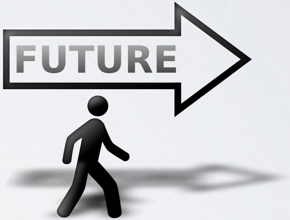
So Far

- Local Asymptotic Normality
- Differentiability in Quadratic Mean
- Limit Distribution under Contig. Alt.
- Symmetric Location Model

TODAY

- Hodges' Estimator
- Supper-Efficiency
- Convolution Theorems
- Anderson's Lemma





Convolution Theorems

- Consider the following generic version of an estimation problem.
- **Data**: $X_i, i = 1, \ldots, n$ i.i.d. with distribution $P \in \mathbf{P} = \{P_\theta : \theta \in \Theta\}$.
- **Estimator**: we wish to estimate $\psi(\theta)$ using the data and that we have an estimator $T_n = T_n(X_1, \ldots, X_n)$ such that for each $\theta \in \Theta$,

  $$\sqrt{n}\,(T_n - \psi(\theta)) \xrightarrow{d} L_\theta$$

  under P_θ - for short we may write "under θ" today.
- **Question**: What is the "best" possible limit distribution for such an estimator?
- It is natural to measure "best" in terms of concentration, and we can measure concentration with a loss function.

Bowl-shaped loss function

- **Loss function**: simply any function $\ell(x)$ that takes values in $[0, \infty)$.

- A loss function is said to be "bowl-shaped" if the sublevel sets

$$\{x : \ell(x) \leqslant c\}$$

  are convex and symmetric about the origin.

- A common bowl-shaped loss function on $\mathbf{R}$ is mean-squared error loss: $\ell(x) = x^2$.

- For a given loss function $\ell(x)$, a limit distribution will be considered "good" if

$$\int \ell(x) dL_\theta \quad \text{is small} .$$

- **Example**: If the estimator T_n is asymptotically normal,

$$L_\theta = N(\mu(\theta), \sigma^2(\theta)) \ ,$$

  then to minimize the mean-squared error loss it is optimal to have $\mu(\theta) = 0$ and $\sigma^2(\theta)$ as small as possible. But we do not want to restrict attention to asymptotically normal estimators.

Hodges' Estimator and Superefficiency

- Consider $\mathbf{P} = \{P_\theta = N(\theta, 1) : \theta \in \mathbf{R}\}$ and $\psi(\theta) = \theta$.

- A natural **estimator** of θ is the sample mean: $T_n = \bar{X}_n$.

- This estimator has many finite-sample optimality properties (it's minimax for every bowlshaped loss function, it's minimum variance unbiased, etc.)

- We might reasonably expect it to be optimal asymptotically as well.

- A second estimator of θ, S_n, can be defined as follows:

$$S_n = \begin{cases} T_n & \text{if } |T_n| \geqslant n^{-1/4} \\ 0 & \text{if } |T_n| < n^{-1/4} \end{cases}.$$

  In words, $S_n = T_n$ when T_n is "far" from zero and $S_n = 0$ when T_n is "close" to zero.

- **Immediate**: $\sqrt{n}\,(T_n - \theta) \sim N(0, 1)$. But how does S_n behave asymptotically?

Asymptotic behavior of S_n

$$S_n = T_n \, I\{\, |T_n| \geqslant n^{-1/4} \,\}$$

First consider the case where $\theta \neq 0$.

Asymptotic behavior of S_n

$$S_n = T_n \, I\{ \, |T_n| \geqslant n^{-1/4} \, \}$$

Next consider the case where $\theta = 0$.

Super-efficiency

- For $\theta \neq 0$: $\sqrt{n}(S_n - \theta) \xrightarrow{d} N(0,1)$ under P_θ.
- For $\theta = 0$: $a_n(S_n - \theta) \xrightarrow{d} 0$ under any sequence a_n, including $\sqrt{n}$.
- The estimator is said to be **superefficient** at $\theta = 0$.
- Let L_θ denote the limit distribution of T_n and L'_θ denote the limit distribution of S_n.
- It follows from the above discussion that for $\theta \neq 0$

$$\int x^2 dL_\theta = \int x^2 dL'_\theta$$

  and for $\theta = 0$,

$$\int x^2 dL'_\theta = 0 < 1 = \int x^2 dL_\theta \ .$$

- **Thus**: S_n appears to be a better estimator of θ than T_n.

Appearances can be deceiving

- Reasoning again reflects the poor use of asymptotics. Our hope is that

$$\int x^2 dL'_\theta$$

is a reasonable approximation to the finite-sample expected loss

$$E_\theta\left[\left(\sqrt{n}\,(S_n - \theta)\right)^2\right] .$$

- **Finite-samples**: for θ "far" from zero, we might expect $S_n = T_n$, so L'_θ may be a reasonable approximation to the distribution of $\sqrt{n}\,(S_n - \theta)$; for "close" to zero, on the other hand, S_n will frequently differ from T_n, so the distribution of $\sqrt{n}\,(S_n - \theta)$ may be quite different from L'_θ.

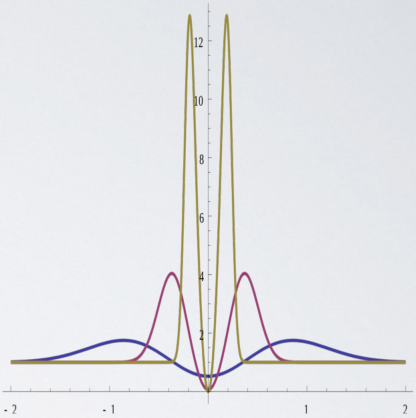


FIGURE: Risk of S_n

QUESTIONS?

Going for a better approximation

- Consider $\theta_n = \frac{h}{n^{1/4}}$ where $0 < h < 1$.

- We are redefining $T_n = \bar{X}_{n,n}$, where $X_{i,n}, i = 1, \ldots, n$ are i.i.d. with distribution $P_{\theta_n} = N(\theta_n, 1)$).

- **Finite Sample distribution**: As before,

$$\sqrt{n}\,(T_n - \theta_n) \sim N(0,1) \;\text{ under } P_{\theta_n} \;.$$

- **Question**: how does S_n behave under θ_n? Star by noticing that

$$\begin{aligned} P_{\theta_n}\left\{|T_n| < n^{-1/4}\right\} &= P_{\theta_n}\left\{-n^{-1/4} < T_n < n^{-1/4}\right\} \\ &= P_{\theta_n}\left\{\sqrt{n}(-n^{-1/4} - \theta_n) < Z_n < \sqrt{n}(n^{-1/4} - \theta_n)\right\} \\ &= P_{\theta_n}\left\{-n^{1/4}(1+h) < Z_n < n^{1/4}(1-h)\right\} \to 1\,. \end{aligned}$$

- Earlier this probability tended to 0 under $\theta \neq 0$, but now under $\theta_n = \frac{h}{n^{1/4}}$, this probability tends to 1.

Lesson from the local approximation

- **Result**: under θ_n we have $S_n = 0$ with probability approaching 1. Hence, under θ_n,

$$\sqrt{n}(S_n - \theta_n) = -n^{1/4}h$$

  with probability approaching 1, and $-n^{1/4}h \to -\infty$.

- Denote by L the limiting distribution of T_n under θ_n and by L' the limiting distribution of S_n under θ_n (in this case L' is degenerate at $-\infty$). It follows that

$$\int x^2 dL' = \infty > 1 = \int x^2 dL \, .$$

- **Lesson**: S_n "buys" its better asymptotic performance at 0 at the expense of worse behavior for points "close" to zero. The definition of "close" changes with n, so this feature is not borne out by a pointwise asymptotic comparison for every $\theta \in \Theta$.

- This example is quite famous and is due to Hodges: S_n is often referred to as **Hodges' estimator**.

QUESTIONS?

Efficiency of Maximum likelihood

- **Background**: Theorems that in some way show that a normal distribution with mean zero and covariance matrix equal to the inverse of the Fisher information is a "best possible" limit distribution have a long history, starting with Fisher in the 1920s and with important contributions by Cramér, Rao, Stein, Rubin, Chernoff and others.
- "The" theorem referred to is not true, at least not without a number of qualifications.
- The above example illustrates this and shows that it is impossible to give a non-trivial definition of "best" to the limit distributions L_θ.
- In fact, it is not even enough to consider L_θ under every $\theta \in \Theta$. For some fixed $\theta' \in \Theta$, we could always construct an estimator whose limit distribution was equal to L_θ for $\theta \neq \theta'$, but "better" at $\theta = \theta'$ by using the trick due to Hodges.
- Hájek and Le Cam contributed to this issue, and eventually gave a complete explanation.
- Under certain conditions, the "best" limit distributions are in fact the limit distributions of maximum likelihood estimators, but to make this idea precise is a bit tricky (convolution theorems)

Regular Estimators

Definition

T_n is called a sequence of **locally regular estimators** of $\psi(\theta)$ at the point θ_0 if, for every h

$$a_n\Big(T_n - \psi(\theta_0 + h/a_n)\Big) \xrightarrow{d} L_{\theta_0} \text{ under } P_{\theta_0+h/a_n}$$

as $a_n \to \infty$ (typically, $a_n = \sqrt{n}$), where the limit distribution might depend on θ_0 but not on h.

- A regular estimator sequence attains its limit distribution in a “locally uniform” manner.
- **Intuition**: a small change in the parameter should not change the distribution of the estimator too much; a disappearing small change should not change the (limit) distribution at all.

Definition

A model $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ is called **differentiable in quadratic mean** at θ if there exists a measurable function $\dot{\ell}_\theta$ such that, as $h \to 0$,

$$\int \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h' \dot{\ell}_\theta \sqrt{p_\theta} \right]^2 d\mu = o(\|h\|^2),$$

where p_θ is the density of P_θ w.r.t. some measure μ.

- Typically, $\dot{\ell}_\theta = \partial(\log p_\theta)/\partial\theta = \frac{\dot{p}_\theta}{p_\theta}$
- QMD is the condition that gives us LAN
- Theorems on local optimality of tests and estimators use a condition like QMD or require LAN directly.

Convolution Theorems

Hájek's convolution theorem shows that the limiting distribution of any regular estimator T_n can be written as a convolution of $N(0, \cdot)$ and "noise".

Theorem (Hájek Convolution Theorem)

Suppose that (1) $\mathbf{P}$ is differentiable in quadratic mean at each θ with non-singular Fisher information matrix

$$I_\theta = E_\theta[\dot{\ell}_\theta \dot{\ell}_\theta'] \ ,$$

and that (2) ψ is differentiable at every θ. (3) Let T_n be an at θ regular estimator sequence with limit distribution L_θ.

Then, there exist distributions M_θ such that

$$L_\theta = N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta') * M_\theta \ .$$

In particular, if L_θ has covariance matrix Σ_θ, then the matrix $\Sigma_\theta - \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta'$ is nonnegative-definite.

The notation $*$ denotes the "convolution" operation between two distributions and should be interpreted as follows: If $X \sim F$ and $Y \sim G$ and $X \perp Y$, then $X + Y \sim F * G$.

Theorem (Almost Everywhere Convolution Theorem)

Suppose that (1) $\mathbf{P}$ *is differentiable in quadratic mean at each* θ *with norming rate* a_n *and non-singular Fisher information matrix*

$$I_\theta = E_\theta[\dot{\ell}_\theta \dot{\ell}'_\theta] \, ,$$

and that (2) ψ *is differentiable at every* θ*.* (3) *Let* T_n *be* **any** *estimator such that for every* θ

$$a_n(T_n - \psi(\theta)) \xrightarrow{d} L_\theta$$

under θ*.*

Then, there exist distributions M_θ *such that for* **almost every** θ *w.r.t. Lebesgue measure*

$$L_\theta = N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}'_\theta) * M_\theta \, .$$

QUESTIONS?

Comments

- **Remarkable theorem**: yields the assertion of Hájek's convolution theorem at almost every parameter value θ, without having to impose the regularity requirement on the estimator sequence.

- **Indeed**: Le Cam showed that it is roughly true that any estimator sequence T_n is "almost Hájek regular" at almost every parameter θ

- The convolution property implies that the covariance matrix of L_θ, if it exists, must be bounded below by the inverse Fisher information.

- This theorem does not contradict the results of the previous section. In that case:

$$\mathbf{P} = \{N(\theta,1) : \theta \in \mathbf{R}\}, \quad \psi(\theta) = \theta, \quad \text{and} \quad N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta') = N(0,1) .$$

- For every $\theta \neq 0$,

$$\sqrt{n}(S_n - \theta) \xrightarrow{d} N(0,1)$$

under P_θ, so the theorem is satisfied for M_θ the distribution with unit mass at 0.

Anderson's Lemma

$N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta')$ is the limit distribution of the MLE of $\psi(\theta)$. In order to assert that this is in fact the "best" limit distribution for more general loss functions, we need the following lemma.

Lemma (Anderson's Lemma)

For **any** *bowl-shaped loss function ℓ on $\mathbf{R}^k$, every probability distribution M on $\mathbf{R}^k$, and every covariance matrix Σ,*

$$\int \ell(x) dN(0, \Sigma) \leqslant \int \ell(x) d(N(0, \Sigma) * M) .$$

- If "best" is measured by any bowl-shaped loss function, then maximum likelihood estimators are "best" for almost every θ w.r.t. Lebesgue measure.
- **Lesson**: the possibility of improvement over the $N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta')$-limit is restricted on a null set of parameters.
- Improvement is also possible by considering special loss function (e.g., the James-Stein's estimator).
- An important part of convolution theorems is the assumption that the model is QMD. The differentiability of ψ is also key.

MLE in models that are not QMD

Example

Suppose $\mathbf{P} = \{P_\theta = U(0,\theta) : \theta > 0\}$ and $\psi(\theta) = \theta$ (Recall that $\mathbf{P}$ is nowhere QMD so the model does not satisfy the conditions of the previous Theorems). We know that the MLE of θ is

$$X_{(n)} = \max\{X_1, \ldots, X_n\}$$

and that

$$n(\theta - X_{(n)}) \xrightarrow{d} L_\theta, \quad \text{where } L_\theta \text{ has density } \frac{1}{\theta} \exp\{-w/\theta\}\,. \tag{1}$$

Clearly, the estimator is **not** asymptotically normal. Although it converges at rate n, much faster than the usual $\sqrt{n}$ rate, the fact that the limiting distribution lies completely to one side of the true parameter suggests that even better estimators may exists.

Claim: for $\ell(x) = x^2$, MLE is sub-optimal and dominated by $\tilde{\theta} = X_{(n)} + X_{(n)}/n$.

MLE dominated in the Uniform case

$n(\theta - X_{(n)}) \xrightarrow{d} L_\theta$ where L_θ has density $\dfrac{1}{\theta}\exp\{-w/\theta\}$ so if $W \sim L_\theta \Rightarrow E(W) = \theta$

THE END!