Econ 551: Lecture Note 7
Asymptotic Efficiency of Maximum Likelihood
The Hajek Representation Theorem

Professor John Rust

1. Background: For simplicity, we will assume a "cross sectional framework" where we have *IID* observations $\{\tilde{x}_1, \ldots, \tilde{x}_N\}$ from a "true density" $f(x|\theta^*)$, where θ^* is an interior point of a compact parameter space in R^k and $f(x|\theta)$ satisfies standard regularity conditions given, e.g. in White (1982) (θ^* is the unique maximizer of $E\{\log f(\tilde{x}|\theta)\}$ and $\log f(x|\theta)$ is twice continuously differentiable in θ and the expectation of the absolute value of the hessian of $\log f(x|\theta)$ is bounded by an integrable function of x). It would be straightforward to generalize the results presented here to time-series or heterogeneous *INID* observations, but at the cost of extra assumptions and extra notational complexity. The key maintained assumption needed for the results below is the *hypothesis of correct specification:* i.e. that $f(x|\theta^*)$ is the true data generating process for $\{\tilde{x}_1, \ldots, \tilde{x}_N\}$.

2. Motivation: In previous lectures I derived the Cramer-Rao inequality and showed that the inverse of the information matrix equals the *Cramer-Rao lower bound.* While this bound holds for all N (where N is the number of observations), the Cramer-Rao inequality can generally only be used to prove efficiency of *unbiased estimators* by showing their variance equals the Cramer-Rao lower bound. However as we noted in Econ 551, most estimators (including most maximum likelihood estimators) are biased in finite samples. Although the Cramer-Rao inequality holds for biased as well as unbiased estimators, in general it is impossible to evaluate the lower bound when estimators are biased since it involves a formula for the bias term which is generally unknown. The Cramer-Rao lower bound becomes more useful in large samples for *consistent estimators* since consistency implies the estimators are *asymptotically unbiased.* In particular we showed that maximum likelihood estimators are consistent and asymptotically normal under weak regularity conditions. In addition, if the parametric model is correctly specified the asymptotic covariance matrix equals the inverse of the information matrix. This suggests that maximum likelihood estimators are *asymptotically efficient.*

3. The problem of "superefficiency" In Econ 551 we discussed the "superefficiency" counterexamples by Stein and Hodges which showed that one can construct estimators that do better than maximum likelihood (in the sense of having a smaller variance than the ML estimator) if the true parameter θ^* is equal to certain isolated points of the parameter space. A lot of effort and ingenuity has gone into finding ways to rule out these superefficient counterexamples. Statisticians realized that the superefficient estimators were *irregular* in the sense that if the true parameter were arbitrarily close but not equal to a point of superefficiency, the superefficient estimator would do worse than the maximum likelihood estimator. In other words the asymptotic distribution of superefficient estimators can be adversely affected by small perturbations in the true parameter θ^*. Thus, we can rule out the superefficient counterexamples by restricting our attention to *regular estimators,* i.e. those for which the asymptotic distribution is invariant to small perturbations in the value of the true parameter θ^*. This is formalized in the following

Definition: An estimator $\tilde{T}_N = T_N(\tilde{x}_1, \ldots, \tilde{x}_N)$ is a *regular estimator* of a $k \times 1$ parameter vector θ^* if the following conditions hold:

1. $\sqrt{N}(\tilde{T}_N - \theta^*) \underset{d}{\Longrightarrow} \tilde{Y}_{\theta^*}$, where $\tilde{Y}_{\theta^*}$ is some random vector depending on θ^*.

2. Consider a sequence of local alternatives to θ^* of the form $\theta_N = \theta^* + \delta/\sqrt{N}$. Suppose observations $\{\tilde{x}_1, \ldots, \tilde{x}_N\}$ are *triangular array,* i.e. *IID* draws from $f(x|\theta_N)$. Then we have

$$\sqrt{N}(\tilde{T}_N - \theta_N) \underset{d}{\Longrightarrow} \tilde{Y}_{\theta^*} \tag{1}$$

independent of δ and this convergence is uniform for $\|\delta\| < C$ for any constant $C > 0$.

Comment: It is not hard to show that all of the "counterexamples" to the efficiency of maximum likelihood estimation including the Stein estimator and the superefficient estimator of Hodges are "irregular" in the sense of failing to satisfy condition in equation (1): the asymptotic distribution of these estimators depends on the vector δ, i.e. the asymptotic distribution depends on the "direction of approach" of the sequence of local alternatives $\{\theta_N\}$ as they converge to θ^*. By choosing "poor directions of approach" we can show that there are sequences of local alternatives for which the superefficient estimators do worse than maximum likelihood. Of course for our definition to make sense we need to show that the maximum likelihood estimator is regular.

Lemma: *The maximum likelihood estimator $\hat{\theta}_N$ is a regular estimator of θ^*.*

Proof: (sketch) Expanding the first order condition for $\hat{\theta}_N$ about the true parameter vector $\theta_N = \theta^* + \delta/\sqrt{N}$ and solving for $\sqrt{N}(\hat{\theta}_N - \theta_N)$ we get:

$$\sqrt{N}(\hat{\theta}_N - \theta_N) = -H_N(\overline{\theta}_N)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial \log}{\partial \theta} f(x_i|\theta^* + \delta/\sqrt{N}) \tag{2}$$

where H_N is the average of the hessian of the log-likelihood terms and $\overline{\theta}_N$ is a point on the line segment between θ_N and $\hat{\theta}_N$. Applying the Lindeberg-Levy Central Limit Theorem we can show that

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial \log}{\partial \theta} f(x_i|\theta^* + \delta/\sqrt{N}) \underset{d}{\Longrightarrow} N(0, I(\theta^*)) \tag{3}$$

and using the fact that the average hessian of the log-likelihood converges uniformly to $H(\theta) = E\{\partial^2 \log f(x|\theta)/\partial\theta\partial\theta'\}$ and $\overline{\theta}_N$ converges with probability 1 to θ^*, we have that

$$\sqrt{N}(\hat{\theta}_N - \theta_N) \underset{d}{\Longrightarrow} N(0, I^{-1}(\theta^*)) \tag{4}$$

independent of δ, i.e. the ML estimator $\hat{\theta}_N$ is a regular estimator of θ^*.

4. Log-Likelihood Ratios and Local Asymptotic Normality. Hajek, LeCam and others realized that the asymptotic properties of maximum likelihood estimators were a result of a property they termed *local asymptotic normality:* i.e. the log-likelihood ratio converges to particular a normal random variable. To state this property formally, it it helpful to consider the following sequence of local alternatives $\{\theta_N\}$ to θ^* formed as follows:

$$\theta_N = \theta^* + I^{-1/2}(\theta^*)\delta/\sqrt{N} \tag{5}$$

where $I(\theta^*)$ is the information matrix.

Definition: The parametric model $f(x|\theta)$ is said to have the *local asymptotic normality* property at θ^* iff for any $\delta \in R^k$ we have:

$$\mathrm{LLR}_N(\delta) = \log\left[\frac{L_N(\theta_N)}{L_N(\theta^*)}\right] \underset{d}{\Rightarrow} \mathrm{LLR}_\infty(\delta) = \langle \delta, \tilde{Z}\rangle - \langle \delta, \delta\rangle/2 \tag{6}$$

where $\langle \delta, \delta\rangle = \delta'\delta = \|\delta\|^2$,

$$L_N(\theta) = \prod_{i=1}^{N} f(x_i|\theta), \tag{7}$$

is the *likelihood* for N observations, $\{\theta_N\}$ is the sequence of local alternatives given in equation (5), and $\tilde{Z} \sim N(0, I)$ where I is the $k \times k$ identity matrix.

Lemma: *If $f(x|\theta)$ satisfies standard regularity conditions (e.g. White 1982), then the parametric model $f(x|\theta)$ has the LAN property.*

Proof: Expand $\log L_N(\theta_N)$ in a second-order Taylor series about θ^* we get:

$$\log L_N(\theta^* + I^{-1/2}(\theta^*)\delta/\sqrt{N}) = \log L_N(\theta^*) + \frac{\partial \log L_N}{\partial \theta}(\theta^*) I^{-1/2}(\theta^*)\delta/\sqrt{N} + \tag{8}$$

$$\frac{1}{2N}\delta' I^{-1/2}(\theta^*)\frac{\partial^2 \log L_N}{\partial\theta\partial\theta'}(\overline{\theta}_N) I^{-1/2}(\theta^*)\delta \tag{9}$$

where $\overline{\theta}_N$ is on the line segment between θ^* and θ_N. Using the Central Limit Theorem it is easy to show that the second term on the right hand side of equation (8) converges in distribution to $\tilde{Z}'\delta$ where $\tilde{Z} \sim N(0, I)$ and I is the $k \times k$ identity matrix. Using the uniform law of large numbers, it is easy to show the third term on the right hand side of equation (8) converges with probability 1 to $\delta'\delta/2 = \|\delta\|^2/2$.

Hajek Representation Theorem: *Suppose the parametric model $f(x|\theta)$ has the LAN property at θ^*. Then if $\tilde{T}_N$ is any regular estimator of θ^* we have:*

$$\sqrt{N} I^{1/2}(\theta^*)(\tilde{T}_N - \theta^*) \underset{d}{\Rightarrow} \tilde{Z} \oplus \tilde{X} \tag{10}$$

where $\tilde{Z} \sim N(0, I)$ (I is the $k \times k$ identity matrix) and $\oplus$ denotes convolution, i.e. the random vectors $\tilde{Z}$ and $\tilde{Y}$ are independently distributed.

Proof: (sketch) Since T_N is assumed to be a regular estimator of θ^*, then its asymptotic distribution does not depend on the direction of approach of a sequence of local alternatives $\{\theta_N\}$

$$\sqrt{N} I^{1/2}(\theta^*)(\tilde{T}_N - \theta_N) = \sqrt{N} I^{1/2}(\theta^*)\left[\tilde{T}_N - \theta^* + I^{-1/2}(\theta^*)\delta/\sqrt{N}\right] \underset{d}{\Rightarrow} \tilde{Y}, \tag{11}$$

where $\tilde{Y}$ is some $k \times 1$ random vector that is independent of the direction of approach of θ_N to θ^* as indexed by δ. Let $\phi_{\tilde{Y}}(t)$ denote the characteristic function of $\tilde{Y}$, i.e.

$$\phi_{\tilde{Y}}(t) = E\left\{\exp\left[i\langle t, \tilde{Y}\rangle\right]\right\}, \tag{12}$$

where $i = \sqrt{-1}$, and $\langle t, Y\rangle$ denotes the inner product of vectors t and Y in R^k. from the density $f(x|\theta_N)$. Since convergence in distribution implies pointwise convergence of characteristic functions we have:

$$\lim_{N\to\infty} E_{\theta_N}\left\{\exp\left[i\langle t, \sqrt{N} I^{1/2}(\theta^*)(\tilde{T}_N - \theta^*) - \delta\rangle\right]\right\} = \phi_{\tilde{Y}}(t), \tag{13}$$

for each $t \in R^k$, where E_{θ_N} denotes the expectation of with respect to the underlying random variables $\{\tilde{x}_1, \ldots, \tilde{x}_N\}$ which are *IID* draws from $f(x|\theta_N)$. Now we can write

$$\phi_{\tilde{Y}}(t) = \exp\{-\langle t, t\rangle/2\} \times \exp\{\langle t, t\rangle/2\}\phi_{\tilde{Y}}(t). \tag{14}$$

Notice that $\exp\{-\langle t, t\rangle/2\}$ is the CF of a $N(0, I)$ random vector. If we can show that $\exp\{\langle t, t\rangle/2\}\phi_{\tilde{Y}}(t)$ is also a CF, then by the Inversion theorem for characteristic functions it follows that $\exp\{\langle t, t\rangle/2\}\phi_{\tilde{Y}}(t)$ is the CF for some random vector $\tilde{X}$ and $\tilde{Y} = \tilde{Z} \oplus \tilde{X}$, so the Hajek Representation Theorem (the result in equation (10)) will follow as a special case when $\delta = 0$. To show that $\exp\{\langle t, t\rangle/2\}\phi_{\tilde{Y}}(t)$ is a CF we appeal to the

Lévy-Cramer Continuity Theorem: *If a sequence of random vectors $\{\tilde{W}_n\}$ with corresponding characteristic functions $\phi_N(t)$ converge pointwise to some function $\mu(t)$ and if $\mu(t)$ is continuous at $t = 0$, then $\mu(t)$ is the characteristic function of some random vector $\tilde{W}$ and $\tilde{W}_N \underset{d}{\Rightarrow} \tilde{W}$.*

Since $\phi_{\tilde{Y}}(t)$ is a CF, *Bochner's Theorem* (theorem providing necessary and sufficient conditions for a function to be a CF for some random vector) guarantees that it is continuous at $t = 0$, so the product $\exp\{\langle t, t\rangle/2\}\phi_{\tilde{Y}}(t)$ is also continuous at $t = 0$. So to complete the proof of the Hajek Representation Theorem we need to show there is a sequence of characteristic functions converging pointwise to $\exp\{\langle t, t\rangle/2\}\phi_{\tilde{Y}}(t)$. Let $\tilde{\mathcal{F}}_N \equiv \sqrt{N} I^{1/2}(\theta^*)(\tilde{T}_N - \theta^*)$. Then we can write

$$\begin{aligned}
& E_{\theta_N}\left\{\exp\left[i\langle t, \sqrt{N} I^{1/2}(\theta^*)(\tilde{T}_N - \theta^*) - \delta\rangle\right]\right\} \\
= & \int_{x_1} \cdots \int_{x_N} \exp\left\{i\langle t, \tilde{\mathcal{F}}_N - \delta\rangle\right\} \prod_{i=1}^{N} f(x_i|\theta_N) dx_i \\
= & \int_{x_1} \cdots \int_{x_N} \exp\left\{i\langle t, \tilde{\mathcal{F}}_N - \delta\rangle\right\} L_N(\theta_N) dx_1 \cdots dx_N \\
= & \int_{x_1} \cdots \int_{x_N} \exp\left\{i\langle t, \tilde{\mathcal{F}}_N - \delta\rangle\right\} \frac{L_N(\theta_N)}{L_N(\theta^*)} L_N(\theta^*) dx_1 \cdots dx_N.
\end{aligned} \tag{15}$$

Now here is the heuristic part of the proof (for a fully rigorous proof that justifies this step see Ibragimov, I.A. and R.Z. Khas'minskii (1981) *Statistical Estimation – Asymptotic Theory* Springer Verlag). By the LAN condition in equation (6) we have for large N:

$$\frac{L_N(\theta_N)}{L_N(\theta^*)} \simeq \exp\left\{\langle \delta, \tilde{Z}_N\rangle - \langle \delta, \delta\rangle/2\right\}. \tag{16}$$

where $\tilde{Z}_N \underset{d}{\Rightarrow} \tilde{Z} \sim N(0, I)$. So substituting equation (16) into equation (15) we get:

$$\begin{aligned}
& E_{\theta_N}\left\{\exp\left[i\langle t, \sqrt{N} I^{1/2}(\theta^*)(\tilde{T}_N - \theta^*) - \delta\rangle\right]\right\} \\
= & \int_{x_1} \cdots \int_{x_N} \exp\left\{i\langle t, \tilde{\mathcal{F}}_N - \delta\rangle\right\} \frac{L_N(\theta_N)}{L_N(\theta^*)} L_N(\theta^*) dx_1 \cdots dx_N \\
= & \int_{x_1} \cdots \int_{x_N} \exp\left\{i\langle t, \tilde{\mathcal{F}}_N - \delta\rangle + \langle \delta, \tilde{Z}_N\rangle - \langle \delta, \delta\rangle/2\right\} L_N(\theta^*) dx_1 \cdots dx_N.
\end{aligned} \tag{17}$$

This implies that

$$\lim_{N\to\infty} \int_{x_1} \cdots \int_{x_N} \exp\left\{i\langle t, \tilde{\mathcal{F}}_N\rangle + \langle \delta, \tilde{Z}_N\rangle\right\} L_N(\theta^*) dx_1 \cdots dx_N = \phi_{\tilde{Y}}(t) \exp\left\{i\langle t, \delta\rangle + \langle \delta, \delta\rangle/2\right\}. \tag{18}$$

Now let $\delta = -it$ (this can be justified by a bit of complex analysis — Mantel's Theorem). Then we have:

$$\lim_{N\to\infty} \int_{x_1} \cdots \int_{x_N} \exp\left\{ i\langle t, \tilde{\mathcal{F}}_N - \tilde{Z}_N \rangle \right\} L_N(\theta^*) dx_1 \cdots dx_N = \phi_{\tilde{Y}}(t) \exp\left\{ \langle t, t \rangle / 2 \right\}. \tag{19}$$

This equation states that the characteristic function of the random vector $\tilde{\mathcal{F}}_N - \tilde{Z}_N$ converges pointwise to the function $\phi_{\tilde{Y}}(t)\exp\{\langle t,t\rangle/2\}$. It follows from the Lévy-Cramer Continuity Theorem that this function is a characteristic function for some random vector $\tilde{X}$. In summary we have:

$$\begin{aligned} &\tilde{\mathcal{F}}_N \equiv \sqrt{N} I^{1/2}(\theta^*)(\tilde{T}_N - \theta^*) \underset{d}{\Rightarrow} \tilde{Y} \quad \text{with CF} \quad \phi_{\tilde{Y}}(t) \\ &(\tilde{\mathcal{F}}_N - \tilde{Z}_N) \underset{d}{\Rightarrow} \tilde{X} \quad \text{with CF} \quad \phi_{\tilde{X}}(t) = \phi_{\tilde{Y}}(t) \exp\{\langle t,t\rangle/2\} \\ &\tilde{Z}_N \underset{d}{\Rightarrow} \tilde{Z} \quad \text{with CF} \quad \phi_{\tilde{Z}}(t) = \exp\{-\langle t,t\rangle/2\}. \end{aligned} \tag{20}$$

It follows that

$$\phi_{\tilde{Y}}(t) = \phi_{\tilde{Z}}(t)\phi_{\tilde{X}}(t) \tag{21}$$

so by the Lévy-Cramer continuity theorem and the inversion theorem for characteristic functions we have:

$$\tilde{\mathcal{F}}_N \equiv \sqrt{N} I^{1/2}(\theta^*)(\tilde{T}_N - \theta^*) \underset{d}{\Rightarrow} \tilde{Y} = \tilde{Z} \oplus \tilde{X}. \tag{22}$$