

# Stat 710: Mathematical Statistics

## Lecture 13

Jun Shao

Department of Statistics  
University of Wisconsin  
Madison, WI 53706, USA

# Lecture 13: Asymptotically efficient estimation

## Asymptotic comparison

Let  $\{\hat{\theta}_n\}$  be a sequence of estimators of  $\theta$  based on a sequence of samples  $\{X = (X_1, \dots, X_n) : n = 1, 2, \dots\}$ .

Suppose that as  $n \rightarrow \infty$ ,  $\hat{\theta}_n$  is asymptotically normal (AN) in the sense that

$$[V_n(\theta)]^{-1/2}(\hat{\theta}_n - \theta) \rightarrow_d N_k(0, I_k),$$

where, for each  $n$ ,  $V_n(\theta)$  is a  $k \times k$  positive definite matrix depending on  $\theta$ .

If  $\theta$  is one-dimensional ( $k = 1$ ), then  $V_n(\theta)$  is the asymptotic variance as well as the amse of  $\hat{\theta}_n$  (§2.5.2).

When  $k > 1$ ,  $V_n(\theta)$  is called the *asymptotic covariance matrix* of  $\hat{\theta}_n$  and can be used as a measure of asymptotic performance of estimators.

If  $\hat{\theta}_{1n}$  is AN with asymptotic covariance matrix  $V_{1n}(\theta)$ ,  $j = 1, 2$ , and  $V_{1n}(\theta) \leq V_{2n}(\theta)$  (in the sense that  $V_{2n}(\theta) - V_{1n}(\theta)$  is nonnegative definite) for all  $\theta \in \Theta$ , then  $\hat{\theta}_{1n}$  is said to be asymptotically more efficient than  $\hat{\theta}_{2n}$ .

## Remarks

- Some sequences of estimators are not comparable under this criterion.
- Since the asymptotic covariance matrices are unique only in the limiting sense, we have to make our comparison based on their limits.
- When  $X_i$ 's are i.i.d.,  $V_n(\theta)$  is usually of the form  $n^{-\delta} V(\theta)$  for some  $\delta > 0$  ( $= 1$  in the majority of cases) and a positive definite matrix  $V(\theta)$  that does not depend on  $n$ .

## Information inequality

If  $\hat{\theta}_n$  is AN, it is asymptotically unbiased.

If  $V_n(\theta) = \text{Var}(\hat{\theta}_n)$ , then, under some regularity conditions, it follows from Theorem 3.3 that we have the following information inequality

$$V_n(\theta) \geq [I_n(\theta)]^{-1},$$

where, for every  $n$ ,  $I_n(\theta)$  is the Fisher information matrix for  $X$  of size  $n$ . The information inequality may lead to an optimal estimator

## Remarks

- Some sequences of estimators are not comparable under this criterion.
- Since the asymptotic covariance matrices are unique only in the limiting sense, we have to make our comparison based on their limits.
- When  $X_i$ 's are i.i.d.,  $V_n(\theta)$  is usually of the form  $n^{-\delta} V(\theta)$  for some  $\delta > 0$  ( $= 1$  in the majority of cases) and a positive definite matrix  $V(\theta)$  that does not depend on  $n$ .

## Information inequality

If  $\hat{\theta}_n$  is AN, it is asymptotically unbiased.

If  $V_n(\theta) = \text{Var}(\hat{\theta}_n)$ , then, under some regularity conditions, it follows from Theorem 3.3 that we have the following information inequality

$$V_n(\theta) \geq [I_n(\theta)]^{-1},$$

where, for every  $n$ ,  $I_n(\theta)$  is the Fisher information matrix for  $X$  of size  $n$ . The information inequality may lead to an optimal estimator

Unfortunately, when  $V_n(\theta)$  is an asymptotic covariance matrix, the information inequality may not hold (even in the limiting sense), even if the regularity conditions in Theorem 3.3 are satisfied.

### Example 4.38 (Hodges)

Let  $X_1, \dots, X_n$  be i.i.d. from  $N(\theta, 1)$ ,  $\theta \in \mathcal{R}$ .

Then  $I_n(\theta) = n$ .

For a fixed constant  $t$ , define

$$\hat{\theta}_n = \begin{cases} \bar{X} & |\bar{X}| \geq n^{-1/4} \\ t\bar{X} & |\bar{X}| < n^{-1/4}, \end{cases}$$

By Proposition 3.2, all conditions in Theorem 3.3 are satisfied.

It can be shown (exercise) that  $\hat{\theta}_n$  is AN with  $V_n(\theta) = V(\theta)/n$ , where  $V(\theta) = 1$  if  $\theta \neq 0$  and  $V(\theta) = t^2$  if  $\theta = 0$ .

If  $t^2 < 1$ , the information inequality does not hold when  $\theta = 0$ .

However, the following result, due to Le Cam (1953), shows that, for i.i.d.  $X_i$ 's, the information inequality holds except for  $\theta$  in a set of Lebesgue measure 0.

Unfortunately, when  $V_n(\theta)$  is an asymptotic covariance matrix, the information inequality may not hold (even in the limiting sense), even if the regularity conditions in Theorem 3.3 are satisfied.

### Example 4.38 (Hodges)

Let  $X_1, \dots, X_n$  be i.i.d. from  $N(\theta, 1)$ ,  $\theta \in \mathcal{R}$ .

Then  $I_n(\theta) = n$ .

For a fixed constant  $t$ , define

$$\hat{\theta}_n = \begin{cases} \bar{X} & |\bar{X}| \geq n^{-1/4} \\ t\bar{X} & |\bar{X}| < n^{-1/4}, \end{cases}$$

By Proposition 3.2, all conditions in Theorem 3.3 are satisfied.

It can be shown (exercise) that  $\hat{\theta}_n$  is AN with  $V_n(\theta) = V(\theta)/n$ , where  $V(\theta) = 1$  if  $\theta \neq 0$  and  $V(\theta) = t^2$  if  $\theta = 0$ .

If  $t^2 < 1$ , the information inequality does not hold when  $\theta = 0$ .

However, the following result, due to Le Cam (1953), shows that, for i.i.d.  $X_i$ 's, the information inequality holds except for  $\theta$  in a set of Lebesgue measure 0.

Unfortunately, when  $V_n(\theta)$  is an asymptotic covariance matrix, the information inequality may not hold (even in the limiting sense), even if the regularity conditions in Theorem 3.3 are satisfied.

### Example 4.38 (Hodges)

Let  $X_1, \dots, X_n$  be i.i.d. from  $N(\theta, 1)$ ,  $\theta \in \mathcal{R}$ .

Then  $I_n(\theta) = n$ .

For a fixed constant  $t$ , define

$$\hat{\theta}_n = \begin{cases} \bar{X} & |\bar{X}| \geq n^{-1/4} \\ t\bar{X} & |\bar{X}| < n^{-1/4}, \end{cases}$$

By Proposition 3.2, all conditions in Theorem 3.3 are satisfied.

It can be shown (exercise) that  $\hat{\theta}_n$  is AN with  $V_n(\theta) = V(\theta)/n$ , where  $V(\theta) = 1$  if  $\theta \neq 0$  and  $V(\theta) = t^2$  if  $\theta = 0$ .

If  $t^2 < 1$ , the information inequality does not hold when  $\theta = 0$ .

However, the following result, due to Le Cam (1953), shows that, for i.i.d.  $X_i$ 's, the information inequality holds except for  $\theta$  in a set of Lebesgue measure 0.

## Theorem 4.16

Let  $X_1, \dots, X_n$  be i.i.d. from a p.d.f.  $f_\theta$  w.r.t. a  $\sigma$ -finite measure  $\nu$  on  $(\mathcal{R}, \mathcal{B})$ , where  $\theta \in \Theta$  and  $\Theta$  is an open set in  $\mathcal{R}^k$ .

Suppose that for every  $x$  in the range of  $X_1$ ,  $f_\theta(x)$  is twice continuously differentiable in  $\theta$  and satisfies

$$\frac{\partial}{\partial \theta} \int \psi_\theta(x) d\nu = \int \frac{\partial}{\partial \theta} \psi_\theta(x) d\nu$$

for  $\psi_\theta(x) = f_\theta(x)$  and  $\psi_\theta(x) = \partial f_\theta(x) / \partial \theta$ ; the Fisher information matrix

$$I_1(\theta) = E \left\{ \frac{\partial}{\partial \theta} \log f_\theta(X_1) \left[ \frac{\partial}{\partial \theta} \log f_\theta(X_1) \right]^\tau \right\}$$

is positive definite; and for any given  $\theta \in \Theta$ , there exists a positive number  $c_\theta$  and a positive function  $h_\theta$  such that  $E[h_\theta(X_1)] < \infty$  and

$$\sup_{\gamma: \|\gamma - \theta\| < c_\theta} \left\| \frac{\partial^2 \log f_\gamma(x)}{\partial \gamma \partial \gamma^\tau} \right\| \leq h_\theta(x)$$

for all  $x$  in the range of  $X_1$ , where  $\|A\| = \sqrt{\text{tr}(A^\tau A)}$  for any matrix  $A$ .

## Theorem 4.16 (continued)

If  $\hat{\theta}_n$  is an estimator of  $\theta$  (based on  $X_1, \dots, X_n$ ) and is AN with  $V_n(\theta) = V(\theta)/n$ , then there is a  $\Theta_0 \subset \Theta$  with Lebesgue measure 0 such that the information inequality holds if  $\theta \notin \Theta_0$ .

Proof: see the textbook

Points at which the information inequality does not hold are called points of superefficiency.

Motivated by the fact that the set of superefficiency points is of Lebesgue measure 0 under regularity conditions, we have the following definition.

### Definition 4.4 (Asymptotic efficiency)

Assume that the Fisher information matrix  $I_n(\theta)$  is well defined and positive definite for every  $n$ .

A sequence of estimators  $\{\hat{\theta}_n\}$  that is AN is said to be *asymptotically efficient* or *asymptotically optimal* if and only if  $V_n(\theta) = [I_n(\theta)]^{-1}$ .

## Theorem 4.16 (continued)

If  $\hat{\theta}_n$  is an estimator of  $\theta$  (based on  $X_1, \dots, X_n$ ) and is AN with  $V_n(\theta) = V(\theta)/n$ , then there is a  $\Theta_0 \subset \Theta$  with Lebesgue measure 0 such that the information inequality holds if  $\theta \notin \Theta_0$ .

**Proof:** see the textbook

Points at which the information inequality does not hold are called points of superefficiency.

Motivated by the fact that the set of superefficiency points is of Lebesgue measure 0 under regularity conditions, we have the following definition.

### Definition 4.4 (Asymptotic efficiency)

Assume that the Fisher information matrix  $I_n(\theta)$  is well defined and positive definite for every  $n$ .

A sequence of estimators  $\{\hat{\theta}_n\}$  that is AN is said to be *asymptotically efficient* or *asymptotically optimal* if and only if  $V_n(\theta) = [I_n(\theta)]^{-1}$ .

## Theorem 4.16 (continued)

If  $\hat{\theta}_n$  is an estimator of  $\theta$  (based on  $X_1, \dots, X_n$ ) and is AN with  $V_n(\theta) = V(\theta)/n$ , then there is a  $\Theta_0 \subset \Theta$  with Lebesgue measure 0 such that the information inequality holds if  $\theta \notin \Theta_0$ .

**Proof:** see the textbook

Points at which the information inequality does not hold are called points of superefficiency.

Motivated by the fact that the set of superefficiency points is of Lebesgue measure 0 under regularity conditions, we have the following definition.

### Definition 4.4 (Asymptotic efficiency)

Assume that the Fisher information matrix  $I_n(\theta)$  is well defined and positive definite for every  $n$ .

A sequence of estimators  $\{\hat{\theta}_n\}$  that is AN is said to be *asymptotically efficient* or *asymptotically optimal* if and only if  $V_n(\theta) = [I_n(\theta)]^{-1}$ .

## Theorem 4.16 (continued)

If  $\hat{\theta}_n$  is an estimator of  $\theta$  (based on  $X_1, \dots, X_n$ ) and is AN with  $V_n(\theta) = V(\theta)/n$ , then there is a  $\Theta_0 \subset \Theta$  with Lebesgue measure 0 such that the information inequality holds if  $\theta \notin \Theta_0$ .

Proof: see the textbook

Points at which the information inequality does not hold are called points of superefficiency.

Motivated by the fact that the set of superefficiency points is of Lebesgue measure 0 under regularity conditions, we have the following definition.

## Definition 4.4 (Asymptotic efficiency)

Assume that the Fisher information matrix  $I_n(\theta)$  is well defined and positive definite for every  $n$ .

A sequence of estimators  $\{\hat{\theta}_n\}$  that is AN is said to be *asymptotically efficient* or *asymptotically optimal* if and only if  $V_n(\theta) = [I_n(\theta)]^{-1}$ .

## Estimating a function of $\theta$

Suppose that we are interested in estimating  $\vartheta = g(\theta)$ , where  $g$  is a differentiable function from  $\Theta$  to  $\mathcal{R}^p$ ,  $1 \leq p \leq k$ .

If  $\hat{\theta}_n$  is AN, then, by Theorem 1.12(i),  $\hat{\vartheta}_n = g(\hat{\theta}_n)$  is asymptotically distributed as  $N_p(\vartheta, [\nabla g(\theta)]^\tau V_n(\theta) \nabla g(\theta))$ .

Thus, the information inequality becomes

$$[\nabla g(\theta)]^\tau V_n(\theta) \nabla g(\theta) \geq [\tilde{I}_n(\vartheta)]^{-1},$$

where  $\tilde{I}_n(\vartheta)$  is the Fisher information matrix about  $\vartheta$  contained in  $X$ .

If  $p = k$  and  $g$  is one-to-one, then

$$[\tilde{I}_n(\vartheta)]^{-1} = [\nabla g(\theta)]^\tau [I_n(\theta)]^{-1} \nabla g(\theta)$$

and, therefore,  $\hat{\vartheta}_n$  is asymptotically efficient if and only if  $\hat{\theta}_n$  is asymptotically efficient.

For this reason, in the case of  $p < k$ ,  $\hat{\vartheta}_n$  is considered to be asymptotically efficient if and only if  $\hat{\theta}_n$  is asymptotically efficient, and we can focus on the estimation of  $\theta$  only.

## Asymptotic efficiency of MLE's and RLE's in the i.i.d. case

Under some regularity conditions, a root of the likelihood equation (RLE), which is a candidate for an MLE, is asymptotically efficient.

### Theorem 4.17

Assume the conditions of Theorem 4.16.

(i) There is a sequence of estimators  $\{\hat{\theta}_n\}$  such that

$$P(s_n(\hat{\theta}_n) = 0) \rightarrow 1 \quad \text{and} \quad \hat{\theta}_n \rightarrow_p \theta,$$

where  $s_n(\gamma) = \partial \log \ell(\gamma) / \partial \gamma$ .

(ii) Any consistent sequence  $\tilde{\theta}_n$  of RLE's is asymptotically efficient.

### Remarks

- Part (i) is asymptotic existence and consistency.
- If the RLE is unique, then it is consistent and asymptotically efficient, whether or not it is MLE.
- If there are more than one sequences of RLE, the theorem does not tell which one is consistent and asymptotically efficient.
- An MLE sequence is often consistent, but this needs to be verified.

## Asymptotic efficiency of MLE's and RLE's in the i.i.d. case

Under some regularity conditions, a root of the likelihood equation (RLE), which is a candidate for an MLE, is asymptotically efficient.

### Theorem 4.17

Assume the conditions of Theorem 4.16.

(i) There is a sequence of estimators  $\{\hat{\theta}_n\}$  such that

$$P(s_n(\hat{\theta}_n) = 0) \rightarrow 1 \quad \text{and} \quad \hat{\theta}_n \rightarrow_p \theta,$$

where  $s_n(\gamma) = \partial \log \ell(\gamma) / \partial \gamma$ .

(ii) Any consistent sequence  $\tilde{\theta}_n$  of RLE's is asymptotically efficient.

### Remarks

- Part (i) is asymptotic existence and consistency.
- If the RLE is unique, then it is consistent and asymptotically efficient, whether or not it is MLE.
- If there are more than one sequences of RLE, the theorem does not tell which one is consistent and asymptotically efficient.
- An MLE sequence is often consistent, but this needs to be verified.

## Asymptotic efficiency of MLE's and RLE's in the i.i.d. case

Under some regularity conditions, a root of the likelihood equation (RLE), which is a candidate for an MLE, is asymptotically efficient.

### Theorem 4.17

Assume the conditions of Theorem 4.16.

(i) There is a sequence of estimators  $\{\hat{\theta}_n\}$  such that

$$P(s_n(\hat{\theta}_n) = 0) \rightarrow 1 \quad \text{and} \quad \hat{\theta}_n \rightarrow_p \theta,$$

where  $s_n(\gamma) = \partial \log \ell(\gamma) / \partial \gamma$ .

(ii) Any consistent sequence  $\tilde{\theta}_n$  of RLE's is asymptotically efficient.

### Remarks

- Part (i) is asymptotic existence and consistency.
- If the RLE is unique, then it is consistent and asymptotically efficient, whether or not it is MLE.
- If there are more than one sequences of RLE, the theorem does not tell which one is consistent and asymptotically efficient.
- An MLE sequence is often consistent, but this needs to be verified.

## Proof of Theorem 4.17

(i) Let  $B_n(c) = \{\gamma : \|[I_n(\theta)]^{1/2}(\gamma - \theta)\| \leq c\}$  for  $c > 0$ .

Since  $\Theta$  is open, for each  $c > 0$ ,  $B_n(c) \subset \Theta$  for sufficiently large  $n$ .

Since  $B_n(c)$  shrinks to  $\{\theta\}$  as  $n \rightarrow \infty$ , the asymptotic existence of  $\hat{\theta}_n$  is implied by the fact that for any  $\varepsilon > 0$ , there exists  $n_0 > 1$  such that

$$P(\log \ell(\gamma) - \log \ell(\theta) < 0 \quad \text{for all } \gamma \in \partial B_n(c)) \geq 1 - \varepsilon, \quad n \geq n_0, \quad (1)$$

where  $c = 4\sqrt{k/\varepsilon}$  and  $\partial B_n(c)$  is the boundary of  $B_n(c)$ .

For a proof of the measurability of  $\hat{\theta}_n$ , see Serfling (1980, p147).

For  $\gamma \in \partial B_n(c)$ , the Taylor expansion gives

$$\begin{aligned} \log \ell(\gamma) - \log \ell(\theta) &= c\lambda^\tau [I_n(\theta)]^{-1/2} s_n(\theta) \\ &\quad + (c^2/2)\lambda^\tau [I_n(\theta)]^{-1/2} \nabla s_n(\gamma^*) [I_n(\theta)]^{-1/2} \lambda, \end{aligned} \quad (2)$$

where  $\lambda = [I_n(\theta)]^{1/2}(\gamma - \theta)/c$  satisfying  $\|\lambda\| = 1$ ,  $\nabla s_n(\gamma) = \partial s_n(\gamma)/\partial \gamma$ , and  $\gamma^*$  lies between  $\gamma$  and  $\theta$ .

## Proof of Theorem 4.17 (continued)

Note that

$$\begin{aligned} E \frac{\|\nabla s_n(\gamma^*) - \nabla s_n(\theta)\|}{n} &\leq E \max_{\gamma \in B_n(c)} \frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n} \\ &\leq E \max_{\gamma \in B_n(c)} \left\| \frac{\partial^2 \log f_\gamma(X_1)}{\partial \gamma \partial \gamma^\tau} - \frac{\partial^2 \log f_\theta(X_1)}{\partial \theta \partial \theta^\tau} \right\| \\ &\rightarrow 0, \end{aligned} \tag{3}$$

which follows from (a)  $\partial^2 \log f_\gamma(x) / \partial \gamma \partial \gamma^\tau$  is continuous in a neighborhood of  $\theta$  for any fixed  $x$ ; (b)  $B_n(c)$  shrinks to  $\{\theta\}$ ; and (c) for sufficiently large  $n$ ,

$$\max_{\gamma \in B_n(c)} \left\| \frac{\partial^2 \log f_\gamma(X_1)}{\partial \gamma \partial \gamma^\tau} - \frac{\partial^2 \log f_\theta(X_1)}{\partial \theta \partial \theta^\tau} \right\| \leq 2h_\theta(X_1)$$

under the regularity condition.

By the SLLN (Theorem 1.13) and Proposition 3.1,

$n^{-1} \nabla s_n(\theta) \rightarrow_{a.s.} -l_1(\theta)$  (i.e.,  $\|n^{-1} \nabla s_n(\theta) + l_1(\theta)\| \rightarrow_{a.s.} 0$ ).

## Proof of Theorem 4.17 (continued)

These results, together with (2), imply that

$$\log \ell(\gamma) - \log \ell(\theta) = c \lambda^\tau [I_n(\theta)]^{-1/2} s_n(\theta) - [1 + o_p(1)] c^2 / 2. \quad (4)$$

Note that  $\max_\lambda \{ \lambda^\tau [I_n(\theta)]^{-1/2} s_n(\theta) \} = \| [I_n(\theta)]^{-1/2} s_n(\theta) \|$ .

Hence, (1) follows from (4) and

$$\begin{aligned} P(\| [I_n(\theta)]^{-1/2} s_n(\theta) \| < c/4) &\geq 1 - (4/c)^2 E \| [I_n(\theta)]^{-1/2} s_n(\theta) \|^2 \\ &= 1 - k(4/c)^2 = 1 - \varepsilon \end{aligned}$$

This completes the proof of (i).

(ii) Let  $A_\varepsilon = \{ \gamma : \|\gamma - \theta\| \leq \varepsilon \}$  for  $\varepsilon > 0$ .

Since  $\Theta$  is open,  $A_\varepsilon \subset \Theta$  for sufficiently small  $\varepsilon$ .

Let  $\{\tilde{\theta}_n\}$  be a sequence of consistent RLE's, i.e.,

$P(s_n(\tilde{\theta}_n) = 0 \text{ and } \tilde{\theta}_n \in A_\varepsilon) \rightarrow 1$  for any  $\varepsilon > 0$ .

Hence, we can focus on the set on which  $s_n(\tilde{\theta}_n) = 0$  and  $\tilde{\theta}_n \in A_\varepsilon$ .

Using the mean-value theorem for vector-valued functions, we obtain

$$-s_n(\theta) = \left[ \int_0^1 \nabla s_n(\theta + t(\tilde{\theta}_n - \theta)) dt \right] (\tilde{\theta}_n - \theta).$$

Note that

$$\frac{1}{n} \left\| \int_0^1 \nabla s_n(\theta + t(\tilde{\theta}_n - \theta)) dt - \nabla s_n(\theta) \right\| \leq \max_{\gamma \in A_\varepsilon} \frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n}.$$

Using the argument in proving (3) and the fact that  $P(\tilde{\theta}_n \in A_\varepsilon) \rightarrow 1$  for arbitrary  $\varepsilon > 0$ , we obtain that

$$\frac{1}{n} \left\| \int_0^1 \nabla s_n(\theta + t(\tilde{\theta}_n - \theta)) dt - \nabla s_n(\theta) \right\| \rightarrow_p 0.$$

Since  $n^{-1} \nabla s_n(\theta) \rightarrow_{a.s.} -I_1(\theta)$  and  $I_n(\theta) = nI_1(\theta)$ ,

$$-s_n(\theta) = -I_n(\theta)(\tilde{\theta}_n - \theta) + o_p(\|I_n(\theta)(\tilde{\theta}_n - \theta)\|).$$

This and Slutsky's theorem (Theorem 1.11) imply that  $\sqrt{n}(\tilde{\theta}_n - \theta)$  has the same asymptotic distribution as

$$\sqrt{n}[I_n(\theta)]^{-1} s_n(\theta) = n^{-1/2}[I_1(\theta)]^{-1} s_n(\theta) \rightarrow_d N_k(0, [I_1(\theta)]^{-1})$$

by the CLT (Corollary 1.2), since  $\text{Var}(s_n(\theta)) = I_n(\theta)$ .