

# Stat 710: Mathematical Statistics

## Lecture 14

Jun Shao

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

## Scoring and RLE

The method of estimating  $\theta$  by solving  $s_n(\gamma) = 0$  over  $\gamma \in \Theta$  is called *scoring* and the function  $s_n(\gamma)$  is called the *score* function.

RLE's are not necessarily MLE's.

We may use the techniques discussed in §4.4 to check whether an RLE is an MLE.

However, according to Theorem 4.17, when a sequence of RLE's is consistent, then it is asymptotically efficient.

We may not need to search for MLE's, if asymptotic efficiency is the only criterion to select estimators.

Typically a sequence of MLE's is consistent (and asymptotically efficient), although there are examples in which an RLE sequence is consistent but an MLE sequence is not.

## Example 4.39

Suppose that  $X_i$  has a distribution in a natural exponential family, i.e., the p.d.f. of  $X_i$  is

$$f_{\eta}(x_i) = \exp\{\eta^{\tau} T(x_i) - \zeta(\eta)\} h(x_i).$$

Since  $\partial^2 \log f_{\eta}(x_i) / \partial \eta \partial \eta^{\tau} = -\partial^2 \zeta(\eta) / \partial \eta \partial \eta^{\tau}$ , condition

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

is satisfied.

From Proposition 3.2, other conditions in Theorem 4.16 are also satisfied.

For i.i.d.  $X_i$ 's,

$$s_n(\eta) = \sum_{i=1}^n \left[ T(X_i) - \frac{\partial \zeta(\eta)}{\partial \eta} \right].$$

## Example 4.39 (continued)

If  $\hat{\theta}_n = n^{-1} \sum_{i=1}^n T(X_i) \in \Theta$ , the range of  $\theta = g(\eta) = \partial \zeta(\eta) / \partial \eta$ , then  $\hat{\theta}_n$  is a unique RLE of  $\theta$ , which is also a unique MLE of  $\theta$  since  $\partial^2 \zeta(\eta) / \partial \eta \partial \eta^\tau = \text{Var}(T(X_i))$  is positive definite.

Also,  $\eta = g^{-1}(\theta)$  exists and a unique RLE (MLE) of  $\eta$  is  $\hat{\eta}_n = g^{-1}(\hat{\theta}_n)$ . However,  $\hat{\theta}_n$  may not be in  $\Theta$  and the previous argument fails (e.g., Example 4.29).

What Theorem 4.17 tells us in this case is that as  $n \rightarrow \infty$ ,  $P(\hat{\theta}_n \in \Theta) \rightarrow 1$  and, therefore,  $\hat{\theta}_n$  (or  $\hat{\eta}_n$ ) is the unique asymptotically efficient RLE (MLE) of  $\theta$  (or  $\eta$ ) in the limiting sense.

In an example like this we may directly show that  $P(\hat{\theta}_n \in \Theta) \rightarrow 1$ , using the fact that  $\hat{\theta}_n \rightarrow_{a.s.} E[T(X_1)] = g(\eta)$  (the SLLN).

The next theorem provides a similar result for the MLE or RLE in the GLM (§4.4.2).

Its proof is similar to the proof of Theorem 4.17.

## Example 4.39 (continued)

If  $\hat{\theta}_n = n^{-1} \sum_{i=1}^n T(X_i) \in \Theta$ , the range of  $\theta = g(\eta) = \partial \zeta(\eta) / \partial \eta$ , then  $\hat{\theta}_n$  is a unique RLE of  $\theta$ , which is also a unique MLE of  $\theta$  since  $\partial^2 \zeta(\eta) / \partial \eta \partial \eta^\tau = \text{Var}(T(X_i))$  is positive definite.

Also,  $\eta = g^{-1}(\theta)$  exists and a unique RLE (MLE) of  $\eta$  is  $\hat{\eta}_n = g^{-1}(\hat{\theta}_n)$ . However,  $\hat{\theta}_n$  may not be in  $\Theta$  and the previous argument fails (e.g., Example 4.29).

What Theorem 4.17 tells us in this case is that as  $n \rightarrow \infty$ ,  $P(\hat{\theta}_n \in \Theta) \rightarrow 1$  and, therefore,  $\hat{\theta}_n$  (or  $\hat{\eta}_n$ ) is the unique asymptotically efficient RLE (MLE) of  $\theta$  (or  $\eta$ ) in the limiting sense.

In an example like this we may directly show that  $P(\hat{\theta}_n \in \Theta) \rightarrow 1$ , using the fact that  $\hat{\theta}_n \rightarrow_{a.s.} E[T(X_1)] = g(\eta)$  (the SLLN).

The next theorem provides a similar result for the MLE or RLE in the GLM (§4.4.2).

Its proof is similar to the proof of Theorem 4.17.

## Theorem 4.18

Consider the GLM with  $\phi_i = \phi/t_i$  and  $t_i$ 's in a fixed interval  $(t_0, t_\infty)$ ,  $0 < t_0 \leq t_\infty < \infty$ .

Assume that the range of the unknown parameter  $\beta$  is an open subset of  $\mathcal{R}^p$ ; at the true value of  $\beta$ ,  $0 < \inf_i \varphi(\beta^\tau Z_i) \leq \sup_i \varphi(\beta^\tau Z_i) < \infty$ , where  $\varphi(t) = [\psi'(t)]^2 \zeta''(\psi(t))$ ; as  $n \rightarrow \infty$ ,  $\max_{i \leq n} Z_i^\tau (Z^\tau Z)^{-1} Z_i \rightarrow 0$  and  $\lambda_- [Z^\tau Z] \rightarrow \infty$ , where  $Z$  is the  $n \times p$  matrix whose  $i$ th row is the vector  $Z_i$  and  $\lambda_- [A]$  is the smallest eigenvalue of  $A$ .

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

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

where  $s_n(\beta) = \partial \log \ell(\beta, \phi) / \partial \beta$  is the score function.

(ii) Let  $I_n(\beta) = \text{Var}(s_n(\beta))$ . Then

$$[I_n(\beta)]^{1/2} (\hat{\beta}_n - \beta) \rightarrow_d N_p(0, I_p).$$

(iii) If  $\phi$  is known or the p.d.f. indexed by  $\theta = (\beta, \phi)$  satisfies the conditions for  $f_\theta$  in Theorem 4.16, then  $\hat{\beta}_n$  is asymptotically efficient.

## One-Step MLE

Assume the conditions in Theorem 4.16.

Let  $s_n(\gamma)$  be the score function.

Let  $\hat{\theta}_n^{(0)}$  be an estimator of  $\theta$  that may not be asymptotically efficient.

The estimator

$$\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} - [\nabla s_n(\hat{\theta}_n^{(0)})]^{-1} s_n(\hat{\theta}_n^{(0)})$$

is the first iteration in computing an MLE (or RLE) using the Newton-Raphson iteration method with  $\hat{\theta}_n^{(0)}$  as the initial value and, therefore, is called the *one-step* MLE.

Without any further iteration,  $\hat{\theta}_n^{(1)}$  is asymptotically efficient under some conditions.

### Theorem 4.19

Assume that the conditions in Theorem 4.16 hold and that  $\hat{\theta}_n^{(0)}$  is  $\sqrt{n}$ -consistent for  $\theta$  (Definition 2.10).

- (i) The one-step MLE  $\hat{\theta}_n^{(1)}$  is asymptotically efficient.
- (ii) The one-step MLE obtained by replacing  $\nabla s_n(\gamma)$  with its expected value,  $-I_n(\gamma)$  (the Fisher-scoring method), is asymptotically efficient.

## One-Step MLE

Assume the conditions in Theorem 4.16.

Let  $s_n(\gamma)$  be the score function.

Let  $\hat{\theta}_n^{(0)}$  be an estimator of  $\theta$  that may not be asymptotically efficient.

The estimator

$$\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} - [\nabla s_n(\hat{\theta}_n^{(0)})]^{-1} s_n(\hat{\theta}_n^{(0)})$$

is the first iteration in computing an MLE (or RLE) using the Newton-Raphson iteration method with  $\hat{\theta}_n^{(0)}$  as the initial value and, therefore, is called the *one-step* MLE.

Without any further iteration,  $\hat{\theta}_n^{(1)}$  is asymptotically efficient under some conditions.

## Theorem 4.19

Assume that the conditions in Theorem 4.16 hold and that  $\hat{\theta}_n^{(0)}$  is  $\sqrt{n}$ -consistent for  $\theta$  (Definition 2.10).

- (i) The one-step MLE  $\hat{\theta}_n^{(1)}$  is asymptotically efficient.
- (ii) The one-step MLE obtained by replacing  $\nabla s_n(\gamma)$  with its expected value,  $-I_n(\gamma)$  (the Fisher-scoring method), is asymptotically efficient.

## Proof

Since  $\widehat{\theta}_n^{(0)}$  is  $\sqrt{n}$ -consistent, we can focus on the event  $\widehat{\theta}_n^{(0)} \in A_\varepsilon = \{\gamma: \|\gamma - \theta\| \leq \varepsilon\}$  for a sufficiently small  $\varepsilon$  such that  $A_\varepsilon \subset \Theta$ . From the mean-value theorem,

$$s_n(\widehat{\theta}_n^{(0)}) = s_n(\theta) + \left[ \int_0^1 \nabla s_n(\theta + t(\widehat{\theta}_n^{(0)} - \theta)) dt \right] (\widehat{\theta}_n^{(0)} - \theta).$$

Substituting this into the formula for  $\widehat{\theta}_n^{(1)}$ , we obtain that

$$\widehat{\theta}_n^{(1)} - \theta = -[\nabla s_n(\widehat{\theta}_n^{(0)})]^{-1} s_n(\theta) + [I_k - G_n(\widehat{\theta}_n^{(0)})](\widehat{\theta}_n^{(0)} - \theta),$$

where

$$G_n(\widehat{\theta}_n^{(0)}) = [\nabla s_n(\widehat{\theta}_n^{(0)})]^{-1} \int_0^1 \nabla s_n(\theta + t(\widehat{\theta}_n^{(0)} - \theta)) dt.$$

From the proof of Theorem 4.17,

$$\| [I_n(\theta)]^{1/2} [\nabla s_n(\widehat{\theta}_n^{(0)})]^{-1} [I_n(\theta)]^{1/2} + I_k \| \rightarrow_p 0.$$

## Proof (continued)

Using an argument similar to those in the proof of Theorem 4.17, we can show that

$$\|G_n(\hat{\theta}_n^{(0)}) - I_k\| \rightarrow_p 0.$$

These results and the fact that  $\sqrt{n}(\hat{\theta}_n^{(0)} - \theta) = O_p(1)$  imply

$$\sqrt{n}(\hat{\theta}_n^{(1)} - \theta) = \sqrt{n}[I_n(\theta)]^{-1} s_n(\theta) + o_p(1).$$

This proves (i).

The proof for (ii) is similar.

## Example 4.40

Let  $X_1, \dots, X_n$  be i.i.d. from the Weibull distribution  $W(\theta, 1)$ , where  $\theta > 0$  is unknown.

Note that

$$s_n(\theta) = \frac{n}{\theta} + \sum_{i=1}^n \log X_i - \sum_{i=1}^n X_i^\theta \log X_i$$

## Proof (continued)

Using an argument similar to those in the proof of Theorem 4.17, we can show that

$$\|G_n(\hat{\theta}_n^{(0)}) - I_k\| \rightarrow_p 0.$$

These results and the fact that  $\sqrt{n}(\hat{\theta}_n^{(0)} - \theta) = O_p(1)$  imply

$$\sqrt{n}(\hat{\theta}_n^{(1)} - \theta) = \sqrt{n}[I_n(\theta)]^{-1} s_n(\theta) + o_p(1).$$

This proves (i).

The proof for (ii) is similar.

## Example 4.40

Let  $X_1, \dots, X_n$  be i.i.d. from the Weibull distribution  $W(\theta, 1)$ , where  $\theta > 0$  is unknown.

Note that

$$s_n(\theta) = \frac{n}{\theta} + \sum_{i=1}^n \log X_i - \sum_{i=1}^n X_i^\theta \log X_i$$

## Example 40 (continued)

Then

$$\nabla s_n(\theta) = -\frac{n}{\theta^2} - \sum_{i=1}^n X_i^\theta (\log X_i)^2.$$

Hence, the one-step MLE of  $\theta$  is

$$\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} \left[ 1 + \frac{n + \hat{\theta}_n^{(0)} (\sum_{i=1}^n \log X_i - \sum_{i=1}^n X_i^{\hat{\theta}_n^{(0)}} \log X_i)}{n + (\hat{\theta}_n^{(0)})^2 \sum_{i=1}^n X_i^{\hat{\theta}_n^{(0)}} (\log X_i)^2} \right].$$

Usually one can use a moment estimator (§3.5.2) as the initial estimator  $\hat{\theta}_n^{(0)}$ .

In this example, a moment estimator of  $\theta$  is the solution of  $\bar{X} = \Gamma(\theta^{-1} + 1)$ .

Results similar to that in Theorem 4.19 can be obtained in the GLM.

## Example 40 (continued)

Then

$$\nabla s_n(\theta) = -\frac{n}{\theta^2} - \sum_{i=1}^n X_i^\theta (\log X_i)^2.$$

Hence, the one-step MLE of  $\theta$  is

$$\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} \left[ 1 + \frac{n + \hat{\theta}_n^{(0)} (\sum_{i=1}^n \log X_i - \sum_{i=1}^n X_i^{\hat{\theta}_n^{(0)}} \log X_i)}{n + (\hat{\theta}_n^{(0)})^2 \sum_{i=1}^n X_i^{\hat{\theta}_n^{(0)}} (\log X_i)^2} \right].$$

Usually one can use a moment estimator (§3.5.2) as the initial estimator  $\hat{\theta}_n^{(0)}$ .

In this example, a moment estimator of  $\theta$  is the solution of  $\bar{X} = \Gamma(\theta^{-1} + 1)$ .

Results similar to that in Theorem 4.19 can be obtained in the GLM.

## Bayes estimators

Bayes estimators are often asymptotically efficient

It can be checked if explicit forms of Bayes estimators are available.

The following is a general result.

### Theorem 4.20

Assume the conditions of Theorem 4.16.

Let  $\pi(\gamma)$  be a prior p.d.f. (which may be improper) w.r.t. the Lebesgue measure on  $\Theta$  and  $p_n(\gamma)$  be the posterior p.d.f., given  $X_1, \dots, X_n$ ,  $n = 1, 2, \dots$

Assume that there exists an  $n_0$  such that  $p_{n_0}(\gamma)$  is continuous and positive for all  $\gamma \in \Theta$ ,  $\int p_{n_0}(\gamma) d\gamma = 1$  and  $\int \|\gamma\| p_{n_0}(\gamma) d\gamma < \infty$ .

Suppose further that, for any  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that

$$\lim_{n \rightarrow \infty} P \left( \sup_{\|\gamma - \theta\| \geq \varepsilon} \frac{\log \ell(\gamma) - \log \ell(\theta)}{n} > -\delta \right) = 0$$

and

## Bayes estimators

Bayes estimators are often asymptotically efficient  
It can be checked if explicit forms of Bayes estimators are available.  
The following is a general result.

### Theorem 4.20

Assume the conditions of Theorem 4.16.

Let  $\pi(\gamma)$  be a prior p.d.f. (which may be improper) w.r.t. the Lebesgue measure on  $\Theta$  and  $p_n(\gamma)$  be the posterior p.d.f., given  $X_1, \dots, X_n$ ,  $n = 1, 2, \dots$

Assume that there exists an  $n_0$  such that  $p_{n_0}(\gamma)$  is continuous and positive for all  $\gamma \in \Theta$ ,  $\int p_{n_0}(\gamma) d\gamma = 1$  and  $\int \|\gamma\| p_{n_0}(\gamma) d\gamma < \infty$ .

Suppose further that, for any  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that

$$\lim_{n \rightarrow \infty} P \left( \sup_{\|\gamma - \theta\| \geq \varepsilon} \frac{\log \ell(\gamma) - \log \ell(\theta)}{n} > -\delta \right) = 0$$

and

## Theorem 4.20 (continued)

$$\lim_{n \rightarrow \infty} P \left( \sup_{\|\gamma - \theta\| \leq \delta} \frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n} \geq \varepsilon \right) = 0,$$

where  $\ell(\gamma)$  is the likelihood function and  $s_n(\gamma)$  is the score function.

(i) Let  $p_n^*(\gamma)$  be the posterior p.d.f. of  $\sqrt{n}(\gamma - T_n)$ , where  $T_n = \theta + [I_n(\theta)]^{-1} s_n(\theta)$  and  $\theta$  is the true parameter value, and let  $\psi(\gamma)$  be the p.d.f. of  $N_k(0, [I_1(\theta)]^{-1})$ .

Then

$$\int (1 + \|\gamma\|) |p_n^*(\gamma) - \psi(\gamma)| d\gamma \rightarrow_p 0.$$

(ii) The Bayes estimator of  $\theta$  under the squared error loss is asymptotically efficient.

Proof: omitted

## Conclusions from Theorem 4.20

- Result (i) shows that the posterior p.d.f. is approximately normal with mean  $\theta + [I_n(\theta)]^{-1} s_n(\theta)$  and covariance matrix  $[I_n(\theta)]^{-1}$ . This result is useful in Bayesian computation; see Berger (1985, §4.9.3).
- Result (i) shows that the posterior distribution and its first-order moments converge to the degenerate distribution at  $\theta$  and its first-order moments, which implies the consistency and asymptotic unbiasedness of Bayes estimators such as the posterior means.
- The Bayes estimator under the squared error loss is asymptotically efficient, which provides an additional support for the early suggestion that the Bayesian approach is a useful method for generating estimators.
- The results hold regardless of the prior being used, indicating that the effect of the prior declines as  $n$  increases.