Asymptotic comparison

Let $\{\widehat{\theta}_n\}$ be a sequence of estimators of θ based on a sequence of samples $\{X = (X_1, ..., X_n) : n = 1, 2, ...\}$.
Suppose that as $n \to \infty$, $\widehat{\theta}_n$ is asymptotically normal (AN) in the sense that

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

where, for each n, $V_n(\theta)$ is a $k \times k$ positive definite matrix depending on θ.
If θ is one-dimensional ($k = 1$), then $V_n(\theta)$ is the asymptotic variance as well as the amse of $\widehat{\theta}_n$ (§2.5.2).
When $k > 1$, $V_n(\theta)$ is called the *asymptotic covariance matrix* of $\widehat{\theta}_n$ and can be used as a measure of asymptotic performance of estimators.
If $\widehat{\theta}_{jn}$ is AN with asymptotic covariance matrix $V_{jn}(\theta)$, $j = 1, 2$, and $V_{1n}(\theta) \le V_{2n}(\theta)$ (in the sense that $V_{2n}(\theta) - V_{1n}(\theta)$ is nonnegative definite) for all $\theta \in \Theta$, then $\widehat{\theta}_{1n}$ is said to be asymptotically more efficient than $\widehat{\theta}_{2n}$.

Remarks

- Some 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 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 $\widehat{\theta}_n$ is AN, it is asymptotically unbiased.
If $V_n(\theta) = \mathrm{Var}(\widehat{\theta}_n)$, 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 $X_1,...,X_n$ be i.i.d. from $N(\theta,1)$, $\theta \in \mathscr{R}$, and $\bar{X}$ be the sample mean.
The Fisher information is $I_n(\theta) = n$.
By Proposition 3.2, all conditions in Theorem 3.3 are satisfied.
By the CLT,

$$\sqrt{n}(\bar{X} - \theta) \to_d N(0,1)$$

In fact, $\bar{X}$ achieves the information lower bound, $\mathrm{Var}(\bar{X}) = n^{-1}$.

For a fixed constant t, define

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

Consider first $\theta \neq 0$.
By the SLLN, $\bar{X} \to_{a.s} \theta \neq 0$, hence,

$$P(|\bar{X}| < n^{-1/4}) \to 0$$

This means that the asymptotic distribution of $\widehat{\theta}_n - \theta$ is the same as that of $\bar{X} - \theta$, i.e.,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \to_d N(0,1), \quad \theta \neq 0$$

Consider now $\theta = 0$.
By the CLT, $\sqrt{n}\bar{X} \to_d N(0,1)$ and hence

$$\begin{aligned} P(|\bar{X}| < n^{-1/4}) &= P(\sqrt{n}|\bar{X}| < n^{1/4}) \\ &= \Phi(n^{1/4}) - \Phi(-n^{1/4}) + o(1) \\ &\to 1 \end{aligned}$$

where Φ is the c.d.f. of $N(0,1)$.
It shows that the asymptotic distribution of $\widehat{\theta}_n$ is the same as that of $t\bar{X}$, i.e.,

$$\sqrt{n}\widehat{\theta}_n \to_d N(0, t^2) \quad \theta = 0$$

If $t^2 < 1$, $\widehat{\theta}_n$ is asymptotically more efficient than $\bar{X}$ when $\theta = 0$.

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

Example 4.38 shows that $\theta = 0$ is a single sppereffiency point.

However, the following result, due to Le Cam (1953), shows that, for i.i.d. X_i's, the set of superefficiency points is of Lebesgue measure 0, under regularity conditions.

Theorem 4.16

Let $X_1,...,X_n$ be i.i.d. from a p.d.f. f_θ w.r.t. a σ-finite measure ν on $(\mathscr{R},\mathscr{B})$, where $\theta \in \Theta$ and Θ is an open set in $\mathscr{R}^k$.
Suppose that for every x in the range of X_1, $f_\theta(x)$ is twice continuously differentiable in θ 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 $= \partial f_\theta(x)/\partial\theta$; 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_θ and a positive function h_θ such that $E[h_\theta(X_1)] < \infty$ and

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

for all x in the range of X_1, where $\|A\| = \sqrt{\mathrm{tr}(A^\tau A)}$ for any matrix A.
If $\widehat{\theta}_n$ is an estimator of θ (based on $X_1,...,X_n$) 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 $\theta \notin \Theta_0$.

Motivated by Theorem 4.16, 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 $\{\widehat{\theta}_n\}$ 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 θ

Suppose that we are interested in estimating $\vartheta = g(\theta)$, where g is a differentiable function from Θ to $\mathscr{R}^p$, $1 \leq p \leq k$.
If $\widehat{\theta}_n$ is AN, then, by Theorem 1.12(i), $\widehat{\vartheta}_n = g(\widehat{\theta}_n)$ 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 ϑ 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, $\widehat{\vartheta}_n$ is asymptotically efficient if and only if $\widehat{\theta}_n$ is asymptotically efficient.
For this reason, in the case of $p < k$, $\widehat{\vartheta}_n$ is considered to be asymptotically efficient if and only if $\widehat{\theta}_n$ is asymptotically efficient, and we can focus on the estimation of θ 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) Asymptotic existence and consistency.
There is a sequence of estimators $\{\widehat{\theta}_n\}$ such that

$$P\big(s_n(\widehat{\theta}_n) = 0\big) \to 1 \qquad \text{and} \qquad \widehat{\theta}_n \to_p \theta,$$

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

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

Remarks

- 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 $c > 0$ and $\partial B_n(c)$ be the boundary of $B_n(c)$.

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

If $\log \ell(\gamma) - \log \ell(\theta) < 0$ for all $\gamma \in \partial B_n(c)$, then $\log \ell(\gamma)$ has a local maximum point $\widehat{\theta}_n$ inside $B_n(c)$ and $\widehat{\theta}_n$ must satisfy $s_n(\widehat{\theta}_n) = 0$.

This means

$$\left\{\text{there exists } \widehat{\theta}_n \text{ such that } s_n(\widehat{\theta}_n) = 0 \text{ and } \widehat{\theta}_n \in B_n(c)\right\}$$

$$\supset \left\{\log \ell(\gamma) - \log \ell(\theta) < 0 \quad \text{for all } \gamma \in \partial B_n(c)\right\}$$

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

Since $I_n(\theta) = nI_1(\theta) \to 0$ as $n \to \infty$, $B_n(c)$ shrinks to $\{\theta\}$ as $n \to \infty$.

Hence, the asymptotic existence and consistency of $\widehat{\theta}_n$ is implied by

$$\lim_{n\to\infty} P\Big\{ \log \ell(\gamma) - \log \ell(\theta) < 0 \quad \text{for all } \gamma \in \partial B_n(c) \Big\} = 0$$

To prove this, we use the definition of limit.

For any $\varepsilon > 0$, we want to show that there exists $n_0 > 1$ such that

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

where we choose $c = 4\sqrt{k/\varepsilon}$.

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

$$\log \ell(\gamma) - \log \ell(\theta) = (\gamma - \theta)^\tau s_n(\theta) + \frac{1}{2}(\gamma - \theta)^\tau \nabla s_n(\gamma^*)(\gamma - \theta)$$

where

$$\nabla s_n(\gamma) = \partial s_n(\gamma)/\partial \gamma$$

and γ^* lies between γ and θ.

Let $\lambda = [I_n(\theta)]^{1/2}(\gamma - \theta)/c$.
Then $\|\lambda\| = 1$ and for $\gamma \in \partial B_n(c)$,

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

Note that

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

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

(c) under the regularity condition, for sufficiently large n,

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

By the SLLN (Theorem 1.13) and Proposition 3.1,

$$n^{-1}\nabla s_n(\theta) \to_{a.s.} -I_1(\theta) \quad (\text{i.e., } \|n^{-1}\nabla s_n(\theta) + I_1(\theta)\| \to_{a.s.} 0).$$

These results, together with (2), imply that

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

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

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

This completes the proof of (i).

Proof of Theorem 4.17 (ii)

Let $A_\varepsilon = \{\gamma : \|\gamma - \theta\| \le \varepsilon\}$ for $\varepsilon > 0$.

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

If $\{\tilde\theta_n\}$ is a sequence of consistent RLE's, then for any $\varepsilon > 0$,

$$P(s_n(\tilde\theta_n) = 0 \text{ and } \tilde\theta_n \in A_\varepsilon) \to 1$$

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

$$-s_n(\theta) = \left[\int_0^1 \nabla s_n\big(\theta + t(\tilde\theta_n - \theta)\big)\,dt\right](\tilde\theta_n - \theta).$$

Note that

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

Using the argument in proving (3) and the fact that $P(\tilde\theta_n \in A_\varepsilon) \to 1$ for arbitrary $\varepsilon > 0$, we obtain that

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

Since $n^{-1}\nabla s_n(\theta) \to_{a.s.} -I_1(\theta)$ 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

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

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

Scoring and RLE

The method of estimating θ 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.
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 consistenct, although there are examples in which an RLE sequence is consistent but not an MLE.

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 Θ and $p_n(\gamma)$ be the posterior p.d.f., given $X_1,...,X_n$, $n = 1,2,....$
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

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

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

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 θ is the true parameter value, and let $\psi(\gamma)$ be the p.d.f. of $N_k(0, [I_1(\theta)]^{-1})$.
Then
$$\int (1 + \|\gamma\|) \left| p_n^*(\gamma) - \psi(\gamma) \right| d\gamma \to_p 0.$$

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

Conclusions from Theorem 4.20

- The posterior p.d.f. is approximately normal with mean $\theta + [I_n(\theta)]^{-1} s_n(\theta)$ and covariance matrix $[I_n(\theta)]^{-1}$.
- The Bayes estimator under the squared error loss is consistent and 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 $n \to \infty$.