Topics in statistics III/IV (MATH3361/4071) Michaelmas term, 2020-2021

Handout 5: Improving sub-efficient estimators

Lecturer & author: Georgios Karagiannis georgios.karagiannis@durham.ac.uk
References: [1]

Note 1. Often it is difficult or computationally intractable to calculate the MLE $\hat{\theta}_n$, because it requires to find the roots of the likelihood equation (5 in H/O 4). Hence, we can resort to alternative estimators which may not be asymptotically efficient; E.g. by Method of moments (Section 1). Once we produce a computationally convenient but sub-efficient, it is often desirable to use a procedure able to improve this estimator; E.g. by two step estimators (Section 2).

Summary 2. To address this issue, one can follow the procedure:

1. Calculate a sub-efficient estimator $\tilde{\theta}_n := \tilde{\theta}_n(X_{1:n})$ for θ (by using an alternative estimator method). Ideally, $\tilde{\theta}_n$ should be tractable (or easier to compute), asymptotically Normal, consistent (although in practice this is violated), but not necessarily asymptotically efficient (like the MLE $\hat{\theta}_n$).
   - E.g., in Section 1, we introduce the moments estimators.
2. Use a recursive procedure, by using the sub-efficient estimator $\tilde{\theta}_n$ as a first guess, with purpose to derive an improved estimator in terms of asymptotic variance.
   - E.g., in Section 2, we introduce the Newton, and the Fisher scoring algorithms.

1 Method of moments (MoM)

Note 3. The method of moments is an alternative pretty simple method producing estimators needed in (1).

Notation 4. Let $X, X_1, X_2, ..., X_n$ be a sequence of IID random samples generated from a distribution f_θ labeled by a d-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^d$, and admitting PDF $f(\cdot|\theta)$.

Definition 5. (Method of moments) The (method of) moments estimator $\tilde{\theta}_n := \tilde{\theta}_n(X_{1:n})$ for the parameter θ is produced by solving the equations

$$\bar{g}_j = \mu_j(\theta), \quad \forall j = 1, ..., d$$

where $\bar{g}_j = \frac{1}{n}\sum_{i=1}^n g_j(X_i)$, and $\mu_j(\theta) = \mathrm{E}_f(g_j(X)|\theta)$, for given functions $g_1(\cdot), g_2(\cdot), ..., g_d(\cdot)$.

Note 6. A popular choice for the functions $g_j(\cdot)$ is

$$g_j(x) = x^j, \quad \text{for } j = 1, ..., d.$$

Notation 7. To make the notation compact, we define $g(x) = (g_1(x), ..., g_d(x))^T$, $\bar{g} = (\bar{g}_1, \bar{g}_2, ..., \bar{g}_d)^T$, and $\mu(\theta) = \mathrm{E}_f(g(X)|\theta) = (\mathrm{E}_f(g_1(X)|\theta), \mathrm{E}_f(g_2(X)|\theta), ..., \mathrm{E}_f(g_d(X)|\theta))$.

Proposition 8. *Under mild conditions (given below), moment estimators can be consistent, and asymptotically Normal, however they are not necessarily asymptotically efficient.*

- *The moment estimator is uniquely defined as*

$$\tilde{\theta}_n = \mu^{-1}(\bar{g})$$

if $\mu(\cdot)$ is continuous and 1-1 (bijective).

- *Let θ_0 be the true value of parameter θ. By the CLT, it is*

$$\sqrt{n}(\bar{g} - \mu(\theta_0)) \xrightarrow{D} N(0, \Sigma(\theta_0))$$

where $\mu(\theta_0) = E_f(g(X)|\theta_0)$, and $\Sigma(\theta_0) = var_f(g(X)|\theta_0)$.

- *By Delta method, it is*

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) = \sqrt{n}(\mu^{-1}(\bar{g}) - \mu^{-1}(\mu(\theta_0))) \xrightarrow{D} N(0, \dot{\mu}_{\theta_0}^{-1}\Sigma(\theta_0)(\dot{\mu}_{\theta_0}^{-1})^T) \tag{1.1}$$

where $\Sigma_{g,\theta_0} = var_f(g(X)|\theta_0)$, and $\dot{\mu}_{\theta_0}^{-1}$ is the inverse of the derivative $\dot{\mu}(\theta_0) = \frac{d}{d\theta}\mu(\theta)|_{\theta=\theta_0}$, if $\mu(\cdot)$ is continuously differentiable at $\theta = \theta_0$

Note 9. In (1.1) it is possible that $\dot{\mu}_{\theta_0}^{-1}\Sigma(\theta_0)(\dot{\mu}_{\theta_0}^{-1})^T \neq \mathcal{I}(\theta_0)^{-1}$, hence MoM is sub-efficient (not always asymptotic efficient).

2 Improving sub-efficient estimators

Note 10. The following algorithms are recursive procedures which use sub-efficient estimators (e.g., the moment estimator) as initial guesses, with purpose to produce an improved one.

Notation 11. Let $X, X_1, X_2, ..., X_n$ be a sequence of IID random samples from a distribution f_θ labeled by a d-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^d$, and admitting PDF $f(\cdot|\theta)$.

Notation 12. Let, $\tilde{\theta}_n$ be a sub-efficient estimator of parameter θ. By using one of the following (optimization) procedures, one can use $\tilde{\theta}_n$ as a seed, and iterate hoping they will produce better estimators by moving towards higher likelihood areas.

Definition 13. Newton's algorithm

Set $\breve{\theta}_n^{(0)} = \tilde{\theta}_n$, as an initial guess. For $t = 1, ...,$ iterate

$$\breve{\theta}_n^{(t)} = \breve{\theta}_n^{(t-1)} - \ddot{\ell}_n(\breve{\theta}_n^{(t-1)})^{-1}\dot{\ell}_n(\breve{\theta}_n^{(t-1)}) \tag{2.1}$$

$$\overset{\text{or}}{=} \breve{\theta}_n^{(t-1)} + \mathcal{J}_n(\breve{\theta}_n^{(t-1)})^{-1}\dot{\ell}_n(\breve{\theta}_n^{(t-1)}) \tag{2.2}$$

because $\mathcal{J}_n(\cdot) = -\ddot{\ell}_n(\cdot)$.

Definition 14. Fisher's scoring algorithm

Set $\breve{\theta}_n^{(0)} = \tilde{\theta}_n$, as an initial guess. For $t = 1, ...,$ iterate

$$\breve{\theta}_n^{(t)} = \breve{\theta}_n^{(t-1)} + \frac{1}{n}\mathcal{I}(\breve{\theta}_n^{(t-1)})^{-1}\dot{\ell}_n(\breve{\theta}_n^{(t-1)}) \tag{2.3}$$

Remark 15. The connection between the two algorithms is that

$$\frac{1}{n}\mathcal{J}_n(\theta) \xrightarrow{a.s.} \mathcal{I}(\theta)$$

In fact, Fisher proposed to replace the observed $\mathcal{J}_n$ (in Newton's algorithm) with the expected $\mathcal{I}$; because he observed that Newton's algorithm (2.1) was getting trapped into undesirable values of θ when the sample size n was small.

2.1 One step estimators

Definition 16. One step estimators are the estimators (2.1) and (2.3) using only one iteration $t = 1$, for a given sub-efficient estimator used as an initial guess.

$$\text{Newton alg.} \qquad \breve{\theta}_n = \tilde{\theta}_n - \ddot{\ell}_n(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n)$$

$$\text{Fisher scoring alg.} \qquad \breve{\theta}_n = \tilde{\theta}_n + \frac{1}{n}\mathcal{I}(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n)$$

Note 17. The following theorem proves that under mild conditions, only one iterative step in (2.1) or (2.3) is required to match the asymptotic efficiency of solutions to the likelihood equations, which, from Cramer Theorem 19 in Handout 4, have been shown to have asymptotic variance equal to the Cramer-Rao information bound.

Theorem 18. *Let $\tilde{\theta}_n$ be a strongly consistent sequence such that*

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{D} N(0, \Sigma(\theta_0))$$

for some $\Sigma(\theta_0) > 0$, where θ_0 is the true value of the parameter. Assume assumptions of Theorem 19 are satisfied. Then the one-step estimators

$$\textit{Newton alg.} \qquad \breve{\theta}_n = \tilde{\theta}_n - \ddot{\ell}_n(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n) \tag{2.4}$$

$$\textit{Fisher scoring alg.} \qquad \breve{\theta}_n = \tilde{\theta}_n + \frac{1}{n}\mathcal{I}(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n) \tag{2.5}$$

are asymptotically equivalent to the MLE, and hence asymptotically efficient. [1]

[1] This theorem can be seen as an exercise. If it was given as an exercise, it would be given with several sub-questions leading to the final result; e.g.

1. Expand $\dot{\ell}_n(\tilde{\theta}_n)$ around $\hat{\theta}_n$ by Mean value theorem
2. Show that $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) = \left(I - \left(\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n)\right)^{-1} \int_0^1 \frac{1}{n}\ddot{\ell}_n(\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n))\mathrm{d}u\right) \times \left(\sqrt{n}(\tilde{\theta}_n - \theta_0) - \sqrt{n}(\hat{\theta}_n - \theta_0)\right)$
3. Show that $\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) \xrightarrow{\text{as}} -\mathcal{I}(\theta_0)$
4. Show that $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) \xrightarrow{\text{as}} 0$
5. Show that $\breve{\theta}_n$ asymptotically equivalent to the MLE $\hat{\theta}_n$
6. Show that $\breve{\theta}_n$ is asymptotically efficient.

Proof. I 'll show the prove for the statement for (2.4), and you can do the same for (2.5), for practice.

Let $\hat{\theta}_n$ be the MLE (or precisely the consistent likelihood equation root of Theorem 19). To prove the statement of the Cramer Theorem for (2.4), I need to show that

$$\breve{\theta}_n - \hat{\theta}_n \xrightarrow{\mathrm{P}} 0$$

(see Definition 23), which, according to Slutsky Theorem 14 and Cramer Theorem 19, implies that

$$\sqrt{n}(\breve{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta_0)^{-1})$$

Then, from (2.4)

$$\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) = \sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n - \ddot{\ell}_n(\tilde{\theta}_n)^{-1}\dot{\ell}_n(\tilde{\theta}_n))$$

Then by using the Mean value theorem (1st order Taylor) to expand $\dot{\ell}_n(\tilde{\theta}_n)$ around $\hat{\theta}_n$, I get

$$\dot{\ell}_n(\tilde{\theta}_n) = \dot{\ell}_n(\hat{\theta}_n) + \int_0^1 \ddot{\ell}_n(\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n))\mathrm{d}u\,(\tilde{\theta}_n - \hat{\theta}_n)$$

So

$$\begin{aligned}
\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) =& \sqrt{n}\left[(\tilde{\theta}_n - \hat{\theta}_n) - \ddot{\ell}_n(\tilde{\theta}_n)^{-1}\left(\dot{\ell}_n(\hat{\theta}_n) + \int_0^1 \ddot{\ell}_n(\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n))\mathrm{d}u\,(\tilde{\theta}_n - \hat{\theta}_n)\right)\right] \\
=& \left(I - \ddot{\ell}_n\tilde{\theta}_n)^{-1}\int_0^1 \ddot{\ell}_n(\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n))\mathrm{d}u\right)\,\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) \\
=& \left(I - \left(\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n)\right)^{-1}\int_0^1 \frac{1}{n}\ddot{\ell}_n(\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n))\mathrm{d}u\right) \times \left(\sqrt{n}(\tilde{\theta}_n - \theta_0) - \sqrt{n}(\hat{\theta}_n - \theta_0)\right) \\
=& \underbrace{\left(I - \underbrace{\left(\underbrace{\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n)}_{\xrightarrow{as}-\mathcal{I}(\theta_0)}\right)^{-1}}_{\xrightarrow{as}-\mathcal{I}(\theta_0)^{-1}} \int_0^1 \underbrace{\frac{1}{n}\ddot{\ell}_n(\underbrace{\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n)}_{\xrightarrow{as}\theta_0})}_{\xrightarrow{as}\int_0^1 \mathrm{E}_f(\ddot{\ell}_n(\theta_0))\mathrm{d}u=-\mathcal{I}(\theta_0)}\,\mathrm{d}u\right)}_{\xrightarrow{as}0} \times \underbrace{\left(\underbrace{\sqrt{n}(\tilde{\theta}_n - \theta_0)}_{\xrightarrow{D}\mathrm{N}(0,\mathcal{I}(\theta_0)^{-1})} - \underbrace{\sqrt{n}(\hat{\theta}_n - \theta_0)}_{\xrightarrow{D}\mathrm{N}(0,\mathcal{I}(\theta_0)^{-1})}\right)}_{\xrightarrow{D}\mathrm{N}(0,2\mathcal{I}(\theta_0)^{-1})\ \text{does not become infinity}\ a.s.} \\
& \xrightarrow{as}0 \\
=& \xrightarrow{as} 0
\end{aligned}$$

Hence $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) \xrightarrow{as} 0$, implying $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) \xrightarrow{P} 0$.

Analytically,

- It is

$$\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) \xrightarrow{as} -\mathcal{I}(\theta_0)$$

by using the the USLLN trick (1). Namely

$$|\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) - (-\mathcal{I}(\theta_0))| = |\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) \pm \mathcal{I}(\tilde{\theta}_n) + \mathcal{I}(\theta_0))|$$
$$\leq \underbrace{\sup_{\theta \in \{\theta : |\tilde{\theta}_n - \theta_0| \leq \delta\}} |\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}_n) - (-\mathcal{I}(\tilde{\theta}_n))|}_{=(i)} + \underbrace{|\mathcal{I}(\tilde{\theta}_n) - \mathcal{I}(\theta_0)|}_{=(ii)}$$

Here, term (i) converges to zero from the USLLN because the assumptions of Theorem 3 (Handout 4) are satisfied by the conditions of Cramer Theorem 19 (Handout 4). Also (ii) converges to zero from Stutsly theorem.

- It is
$$\hat{\theta}_n + u(\tilde{\theta}_n - \hat{\theta}_n) \xrightarrow{as} \theta_0$$
because $\tilde{\theta}_n \xrightarrow{as} \theta_0$ and $\hat{\theta}_n \xrightarrow{as} \theta_0$ as strongly consistent, and by using Slutsky theorem. From the SLLN, I get
$$\frac{1}{n}\ddot{\ell}_n(\theta_0) \xrightarrow{as} -\mathcal{I}(\theta_0)$$
because
$$\frac{1}{n}\ddot{\ell}_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(X_i|\theta_0) \xrightarrow{as} \mathrm{E}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(X|\theta_0) = -\mathcal{I}(\theta_0)$$
Consequently,
$$\int_0^1 \frac{1}{n}\ddot{\ell}_n(\theta_0)\mathrm{d}u \xrightarrow{as} \int_0^1 \mathrm{E}_f(\dot{\Psi}(X, \theta_0))\mathrm{d}u = \int_0^1 1\mathrm{d}u \times \mathrm{E}_f(\dot{\Psi}(X, \theta_0)) = -\mathcal{I}(\theta_0)$$

- It is
$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta_0)^{-1})$$
and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta_0)^{-1})$ by assumption and by Cramer Theorem 19 (Handout 4). So
$$\sqrt{n}(\tilde{\theta}_n - \theta_0) + \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, \Sigma(\theta_0) + \mathcal{I}(\theta_0)^{-1})$$
does not became infinity as it is bounded in probability.

- Since $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) \xrightarrow{as} 0$ then $\sqrt{n}(\breve{\theta}_n - \hat{\theta}_n) \xrightarrow{P} 0$ which implies that $\breve{\theta}_n$ and $\hat{\theta}_n$ are asymptotic equivalent.

- Because $\breve{\theta}_n$ and $\hat{\theta}_n$ are asymptotic equivalent, they asymptotically follow the same distribution, so
$$\sqrt{n}(\breve{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta_0)^{-1})$$
as well, which implies that $\breve{\theta}_n$ is asymptotic efficient.

□

Example 19. Consider random sample $X_1, ..., X_n \overset{IID}{\sim} f(a,b)$, $a>0$, $b>0$ with PDF

$$f(x|a,b) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-x\frac{1}{b}} 1(x>0)$$

Assume that b is known, and a is unknown.

1. Find the moment estimator $\tilde{a}$ of a by using the first raw moments
2. Is the moment estimator $\tilde{a}$ consistent and asymptotically Normal?
3. Find the one step estimator by Fisher scoring algorithm.

Hint-1 Digamma function $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$

Hint-2 Trigamma function $\psi_1(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2}\log\Gamma(x)$

Solution.

1. The first raw moment is

$$\mu_1(a) = \mathrm{E}(X) = \int_0^\infty x \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-x/b} \mathrm{d}x = \int_0^\infty \frac{1}{\frac{1}{a}\Gamma(a+1)\frac{1}{b}b^{a+1}} x^{(a+1)-1} e^{-x/b} \mathrm{d}x = ab$$

   and the sample one

$$m_1 = \bar{X}$$

   From the method of moments I get

$$m_1 = \mu_1(\tilde{a}) \implies \tilde{a} = \frac{m_1}{b} = \frac{\bar{X}}{b}$$

2. The moment estimator is consistent based on the SLLN, $\bar{X} \xrightarrow{D} E(X)$, and by Slutsky $m_1 = \frac{1}{b}\bar{X} \xrightarrow{D} \frac{1}{b}E(X) = \mu_1$. The asymptotic distribution is
   - by CLT
$$\sqrt{n}(\bar{X} - \mu_1) \xrightarrow{D} \mathrm{N}(0, ab^2)$$
   - so by Delta method with $g(\bar{X}) = \frac{\bar{X}}{b}$
$$\sqrt{n}(\frac{\bar{X}}{b} - \frac{\mu_1}{b}) = \sqrt{n}(\tilde{a} - a) \xrightarrow{D} \mathrm{N}(0, \cancel{\frac{ab^2}{b^2}}^{=a})$$

3. For the Fisher algorithm, I need to find $\mathcal{I}(a)^{-1}$. It is

$$\begin{aligned}
\log f(x|a) &= -\log\Gamma(a) - a\log(b) - \frac{1}{b}x + (a-1)\log(x) \\
\frac{\mathrm{d}}{\mathrm{d}a}\log f(x|a) &= -\psi(a) - \log(b) + \log(x) \\
\frac{\mathrm{d}^2}{\mathrm{d}a^2}\log f(x|a) &= -\psi_1(a) \qquad (2.6) \\
\mathcal{I}(a) &= \psi_1(a) \\
\mathcal{I}(a)^{-1} &= 1/\psi_1(a)
\end{aligned}$$

$$\ell_n(\theta) = -n \log \Gamma(a) - na \log(b) - \frac{1}{b} \sum_{i=1}^{n} x_i + (a-1) \sum_{i=1}^{n} \log(x_i)$$

$$\dot{\ell}_n(\theta) = -n\psi(a) - n \log(b) + \sum_{i=1}^{n} \log(x_i)$$

$$\ddot{\ell}_n(\theta) = -n\psi_1(a)$$

The Fisher recursion is

$$\breve{\theta}_n = \tilde{\theta}_n + \frac{1}{n} \mathcal{I}(\tilde{\theta}_n)^{-1} \dot{\ell}_n(\tilde{\theta}_n)$$

$$\breve{\theta}_n = \frac{\bar{X}}{b} + \frac{1}{\psi_1(\frac{\bar{X}}{b})} (-\psi(\frac{\bar{X}}{b}) - \log(b) + \frac{1}{n} \sum_{i=1}^{n} \log(X_i))$$

Additionally for the Newton recursion I need

$$\ddot{\ell}_n(\theta) = -n\psi_1(a)$$

The Newton recursion is

$$\begin{aligned} \breve{\theta}_n &= \tilde{\theta}_n - \ddot{\ell}_n(\tilde{\theta}_n)^{-1} \dot{\ell}_n(\tilde{\theta}_n) \\ &= \frac{\bar{X}}{b} + \frac{1}{\psi_1(\frac{\bar{X}}{b})} (-\psi(\frac{\bar{X}}{b}) - \log(b) + \frac{1}{n} \sum_{i=1}^{n} \log(X_i)) \end{aligned}$$

Here, Fisher and Newton recursion lead to the same results, because the 2nd-derivative (2.6) does not depend on the random sample. However, this does not happen always.

Exercise sheet

Exercise #29
Exercise #30

References

[1] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.