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

Handout 4: Estimation by the method of Maximum Likelihood

Lecturer & author: Georgios Karagiannis georgios.karagiannis@durham.ac.uk
References: [3, 1]

1 Uniformly strong law of large numbers (USLLN)

Note 1. Assume $X, X_1, ..., X_n \overset{\text{IID}}{\sim} f(\cdot|\theta)$, where $f(\cdot|\theta)$ is a distribution labeled by a parameter $\theta \in \Theta$. Let $U(x, \theta)$ be a measurable function of x for all θ, and let $\mu(\theta) = \mathrm{E}_f(U(X, \theta))$ continuous on θ. Assume that it is $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$. If I want to show that

$$\frac{1}{n}\sum_{i=1}^{n} U(X_i, \hat{\theta}_n) \xrightarrow{\text{a.s.}} \mu(\theta_0), \quad \text{as } n \to \infty \tag{1}$$

I could possibly use the trigonic inequality as

$$|\frac{1}{n}\sum_{i=1}^{n} U(X_i, \hat{\theta}_n) - \mu(\theta_0)| \leq |\frac{1}{n}\sum_{i=1}^{n} U(X_i, \hat{\theta}_n) - \mu(\hat{\theta}_n)| + |\mu(\hat{\theta}_n) - \mu(\theta_0)| \tag{2}$$

$$\leq \sup_{\forall\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^{n} U(X_i, \theta) - \mu(\theta)| + |\mu(\hat{\theta}_n) - \mu(\theta_0)| \tag{3}$$

In (2) it is $|\mu(\hat{\theta}_n) - \mu(\theta_0)| \xrightarrow{\text{a.s.}} 0$ by Slutsky theorem, however we cannot say that the first term reduces to zero by using the SLLN. Creating an even upper boundary in (3), we can use the USLLN (see below) giving conditions such that

$$\sup_{\forall\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^{n} U(X_i, \theta) - \mu(\theta)| \xrightarrow{\text{a.s.}} 0 \tag{4}$$

Theorem 2. *If Θ is finite and SLLN implies $\frac{1}{n}\sum_{i=1}^{n} U(X_i, \theta) \xrightarrow{a.s.} \mu(\theta)$ at each $\theta \in \Theta$, then*

$$\sup_{\forall\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^{n} U(X_i, \theta) - \mu(\theta)| \xrightarrow{a.s.} 0$$

Proof. If $A_n(\theta) = |\frac{1}{n}\sum_{i=1}^{n} U(X_i, \theta) - \mu(\theta)| < \epsilon$ for $\epsilon > 0$, then

$$\mathrm{P}(\sup_{\forall\theta} A_n(\theta), \text{as } n \to \infty) = \mathrm{P}(\cap_{\forall\theta} A_n(\theta), \text{as } n \to \infty) = \prod_{\forall\theta} \underbrace{\mathrm{P}(A_n(\theta), \text{as } n \to \infty)}_{=1, \text{ if SLLN holds}} = 1$$

□

Theorem 3. *If*

1. Θ *is compact*
2. $U(x,\theta)$ *is continuous in* θ *for all* x
3. $|U(x,\theta)| \leq K(x)$*, for some function* $K(\cdot)$ *such that* $E_f(K(X)) < \infty$*.*

then

$$\sup_{\forall\theta\in\Theta} |\frac{1}{n}\sum_{i=1}^{n} U(x_i,\theta) - \mu(\theta)| \xrightarrow{a.s.} 0$$

Proof. It is omitted □

Note 4. Theorem 5 below can be used as a tool to show that the maximizer/minimizer $\hat{\theta}_n$ of $\bar{U}_n(\theta)$ converges to the unique maximizer/minimizer θ_0 of $\mu(\theta)$, if $\bar{U}_n(\theta)$ converges to $\mu(\theta)$, an $n \to \infty$. [1]

Theorem 5. *Assume* $\Theta \subset \mathbb{R}^d$ *is compact, and* $\bar{U}_n(\theta)$ *and* $\mu(\theta)$ *are continuous on* θ*. Let* $\bar{U}_n(\theta) \xrightarrow{a.s.} \mu(\theta)$*, let* θ_0 *be the unique maximizer of* $\mu(\theta) < \infty$*, i.e.* $\theta_0 = \arg\max_{\forall\theta} \mu(\theta)$*, and let* $\hat{\theta}_n$ *be a maximizer of* $\bar{U}_n(\theta)$*, i.e.* $\hat{\theta}_n = \arg\max_{\forall\theta} \bar{U}_n(\theta)$*. If the USLLN in (4) holds then* $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$*.*

Proof. Given in Appendix A. □

2 Set-up and definitions

Notation 6. Let $X, X_1, X_2, ..., X_n$ be a sequence of IID random samples (aka random variables) such that $X_i \sim f(\cdot|\theta)$.

Definition 7. Likelihood function is denoted as

$$L_n(\theta) = L_n(\theta|X_1, ..., X_n) = \prod_{i=1}^{n} f(X_i|\theta)$$

Definition 8. Log Likelihood Function is denoted as

$$\ell_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log(f(X_i|\theta))$$

Definition 9. Maximum likelihood estimator is the statistic $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ such that

$$\hat{\theta}_n \quad = \arg\sup_{\forall\theta\in\Theta} L_n(\theta) \quad = \arg\sup_{\forall\theta\in\Theta} \ell_n(\theta)$$

Definition 10. Likelihood equations are denoted as

$$0 = \dot{\ell}_n(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\ell_n(\theta) \tag{5}$$

if derivative exists

[1] As we will see later, the following result is is used to prove that MLE estimators are consistent.

Definition 11. Fisher information is defined as

$$\mathcal{I}(\theta) \quad = \mathrm{E}_\theta(\Psi(X,\theta)\Psi(X,\theta)^T) \quad = \mathrm{E}_\theta((\nabla_\theta \log f(X|\theta))^T(\nabla_\theta \log f(X|\theta)))$$

where

$$\Psi(X,\theta) = (\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X|\theta))^T = \left(\nabla^2_\theta \log f(X|\theta)\right)^\top ;$$
$$\dot{\Psi}(X,\theta) = (\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X|\theta)) = \nabla^2_\theta \log f(X|\theta))$$

Proposition 12. *Extensions of the 1D results from SC2:*

$$E_\theta \Psi(X,\theta) = 0 \tag{6}$$

and

$$\mathcal{I}(\theta) \overset{calc.}{=} Var_\theta(\Psi(X,\theta)) \tag{7}$$
$$\overset{calc.}{=} -E_\theta(\dot{\Psi}(X,\theta)) \tag{8}$$

Proof. It is given as an exercise, however for the solution see `https://en.wikipedia.org/wiki/Fisher_information`. □

Definition 13. Observed information at θ^* is denoted as

$$\mathcal{J}_n(\theta^*) = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell_n(\theta)|_{\theta=\theta_*} = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\sum_{i=1}^n \log(f(X_i|\theta))|_{\theta=\theta_*}$$

Proposition 14. *It is straightforward that for n samples $\{X_1,...,X_n\}$, it is*

$$\mathcal{I}(\theta) = \frac{1}{n}E_\theta(\mathcal{J}_n(\theta))$$

meaning that the Fisher information $\mathcal{I}(\theta)$ is the expected information in $n=1$ sample.

Proposition 15. *By SLLN, as $n \to \infty$, it is* [2]

$$\frac{1}{n}\mathcal{J}_n(\theta) \xrightarrow{a.s.} \mathcal{I}(\theta) \tag{9}$$

Proposition 16. *Let $X, X_1, ..., X_n$ be a sequence of IID random samples from a distribution admitting a density $f(\cdot|\theta)$, $\theta \in \Theta$, then*

$$\frac{1}{\sqrt{n}}\dot{\ell}(\theta) \xrightarrow{D} N(0, \mathcal{I}(\theta)) \tag{10}$$

Proof. It is $\dot{\ell}(\theta) = n\frac{1}{n}\sum_{i=1}^n \Psi(X_i,\theta)$, where $\Psi(X_i,\theta)$ are IID with $\mathrm{E}(\Psi(X_i,\theta)) = 0$, and $\mathrm{Var}(\Psi(X_i,\theta)) = \mathcal{I}(\theta)$. Then by CLT

$$\frac{1}{\sqrt{n}}\dot{\ell}(\theta) = \sqrt{n}\frac{1}{n}\sum_{i=1}^n \Psi(X_i,\theta) \xrightarrow{D} \mathrm{N}(0, \mathcal{I}(\theta))$$

□

[2] It can be extended to cases there $X_1, ..., X_n$ are independent but not IID; see[4]. –Not examinable.

Example 17. If $X \sim \mathrm{N}(\mu, \sigma^2)$ and $\theta = (\mu, \sigma)$, compute $\Psi(X, \theta)$, $\dot{\Psi}(X, \theta)$, and $\mathcal{I}(\theta)$.

Solution.

$$\log(f(x|(\mu,\sigma))) = -\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}$$

$$\frac{\partial}{\partial\mu}\log(f(x|\theta)) = \frac{(x-\mu)}{\sigma^2}; \qquad \frac{\partial}{\partial\sigma}\log(f(x|\theta)) = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}$$

$$\frac{\partial^2}{\partial\mu^2}\log(f(x|\theta)) = \frac{\partial}{\partial\mu}\frac{(x-\mu)}{\sigma^2} = -\frac{1}{\sigma^2}; \qquad \frac{\partial^2}{\partial\sigma^2}\log(f(x|\theta)) = \frac{1}{\sigma^2} - 3\frac{(x-\mu)^2}{\sigma^4}$$

$$\frac{\partial^2}{\partial\mu\partial\sigma}\log f(x|\theta) = -2\frac{(x-\mu)}{\sigma^3}$$

So

$$\Psi(X,\theta) = \begin{bmatrix} \frac{(X-\mu)}{\sigma^2} \\ -\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3} \end{bmatrix} \qquad \dot{\Psi}(X,\theta) = \begin{bmatrix} -\frac{1}{\sigma^2} & -2\frac{(X-\mu)}{\sigma^3} \\ -2\frac{(X-\mu)}{\sigma^3} & \frac{1}{\sigma^2} - 3\frac{(X-\mu)^2}{\sigma^4} \end{bmatrix}$$

$$\mathcal{I}(\theta) = -\mathrm{E}(\dot{\Psi}(X,\theta)) \qquad = -\mathrm{E}\begin{bmatrix} -\frac{1}{\sigma^2} & -2\frac{(X-\mu)}{\sigma^3} \\ -2\frac{(X-\mu)}{\sigma^3} & \frac{1}{\sigma^2} - 3\frac{(X-\mu)^2}{\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & 2\frac{1}{\sigma^2} \end{bmatrix}$$

Note 18. The defined KL divergence can be used as a measure of 'distance' (it is not a distance) between two functions. In stats, it measures the ability of the likelihood ratio to distinguish between models f_0 and f_1.

Definition 19. Kullback-Leibler (KL) divergence is defined as

$$\mathrm{KL}(f_0, f_1) = \mathrm{E}_0(\log\frac{f_0(X)}{f_1(X)}) = \int \log\frac{f_0(X)}{f_1(X)} f_0(X)\mathrm{d}X$$

where $\mathrm{KL}(f_0, f_1) = \infty$ if $f_0(x) = \infty$ and $f_1(x) = 1$, and $\mathrm{KL}(f_0, f_1) = 0$ if $f_0(x) = 0$ and $f_1(x) = \infty$.

Lemma 20. *(Shannon-Kolmogorov Information Inequality) Let f_0 and f_1 (like $f_0(\cdot) = f(\cdot|\theta_0)$ and $f_1(\cdot) = f(\cdot|\theta_1)$) be PDFs of corresponding distributions with respect to x. Then*

$$KL(f_0, f_1) = E_0 \log\frac{f_0(X)}{f_1(X)} = \int \log\frac{f_0(X)}{f_1(X)} f_0(X)\,dX \geq 0$$

with the equality iff $f_0(x) = f_1(x)$ a.s.

Proof. Given as an Exercise **??** in the Exercise sheet. □

Example 21. Let distributions f_0 & f_1 with densities $f(\cdot|\theta_0)$ & $f(\cdot|\theta_0)$. Let $X_1, ...X_n$ is a sequence of IID random samples from $\mathrm{d}f_0$. Let $\ell_n(\theta_0)$ & $\ell_n(\theta_1)$ be log-likelihoods based on densities $f(\cdot|\theta_0)$ & $f(\cdot|\theta_0)$ and given a realisation of samples $X_1, ...X_n$. Then

$$\mathrm{E}_{f_0}(\ell_n(\theta_0)) \geq \mathrm{E}_{f_0}(\ell_n(\theta_1))$$

where $\mathrm{E}_{f_0}(\spadesuit) = \int \spadesuit f(X|\theta_0)\mathrm{d}X$ denotes expectation w.r.t $f(X|\theta_0)\mathrm{d}X$.

Proof. From Lemma 20, it is $\mathrm{E}_{f_0}(\ell_n(\theta_0)) \geq \mathrm{E}_{f_0}(\ell_n(\theta_1)) \Leftrightarrow \mathrm{KL}(f_0, f_1) \geq 0$. □

3 On the consistency & asymptotic distribution of the "MLE"

Note 22. The theorem below provides a tool that allows to find the asymptotic distribution of MLE, under certain conditions.

Theorem 23. *(Cramer Theorem) Let $X, X_1, ..., X_n$ be a sequence of IID random samples from distribution with density $f(\cdot|\theta)$, $\theta \in \Theta$, and let θ_0 denote the true value of the parameter. If:*

C.1 *Θ is an open subset of $\mathbb{R}^d$*

C.2 *$\ddot{\Psi}(x, \theta)$ exists, and is continuous for all x, and the derivative sign can pass under the integral sign.*

C.3 *Each component of $\dot{\Psi}(x, \theta)$ is bounded in absolute value by a function $K(x)$ where $E_{\theta_0}(K(X)) < \infty$*

C.4 *$\mathcal{I}(\theta) = -E_{\theta_0}(\dot{\Psi}(X, \theta))$ is positive definite*

C.5 *(Identifiability ass.) $f(x|\theta) = f(x|\theta)$ implies $\theta = \theta_0$ a.s.*

then there exists sequence $\hat{\theta}_n$ of roots of the likelihood equations, where

1. *it is strongly consistent*
$$\hat{\theta}_n \xrightarrow{a.s.} \theta_0 \tag{11}$$
2. *has asymptotic distribution such as*
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, \mathcal{I}(\theta_0)^{-1}) \tag{12}$$

Proof. Proof is given in Appendix B (It is a good exercise for practice). □

Remark 24. Theorem 23 says that given a set of regularity conditions C.1-C.5, there exists a sequence of the likelihood equation roots (aka $\hat{\theta}_n$ such that $\dot{\ell}_n(\theta)|_{\theta=\hat{\theta}_n} = 0$) which is strongly consistent and asymptotically Normal with Fisher information as covariance matrix. It does not explicitly refers to MLE. MLE may not be the consistent root, or it may not be a likelihood equation root at all.

Remark 25. If the root of likelihood equation (aka $\hat{\theta}_n$ such that $\dot{\ell}_n(\theta)|_{\theta=\hat{\theta}_n} = 0$) is unique and the conditions of Theorem 23 are satisfied, then $\hat{\theta}_n$ refers to the MLE (aka $\hat{\theta}_n = \arg\sup_{\forall\theta} \ell_n(\theta)$). Otherwise MLE may not be the consistent root, or it may not be a likelihood equation at all.

Remark 26. Theorem 23 combined with Delta methods imply that continuous functions of MLE are asymptotically Normal with mean and covariance that can be computed.

Remark 27. The conclusions of Proposition 16, and Cramer's Theorem 23 remain valid even when the samples are independent but not identically distributed. For instance, Theorem 44 is a restatement of Cramer's Theorem 23.

Example 28. (Which will be used as a Proposition later on) Show that given that the assumptions [C.1-C.5] of Theorem 23 are satisfied, and that $\mathcal{I}(\theta)$ and $\mathcal{J}_n(\theta)$ are continuous on θ, then

$$\sqrt{n}\mathcal{I}(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, I) \tag{13}$$

$$\sqrt{n}\mathcal{I}(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, I) \tag{14}$$

$$\mathcal{J}_n(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathrm{N}(0, I) \tag{15}$$

where $A^{1/2}$ denotes the lower triangular matrix of the Cholesky decomposition of A; i.e., $A = A^{1/2}(A^{1/2})^T$.

Solution.

- Eq 13 results from Cramer Theorem, and the properties of covariance matrix.
- Eq. 14 results by using Cramer Theorem and Slutsly theorems. Precisely, because $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$, Slutsky implies $\mathcal{I}(\hat{\theta}_n) \xrightarrow{a.s.} \mathcal{I}(\theta_0)$ which implies $\mathcal{I}(\hat{\theta}_n)^{-1/2}\mathcal{I}(\theta_0)^{-1/2} \xrightarrow{a.s.} I$. Therefore, by Slutsky
$$\underbrace{\mathcal{I}(\hat{\theta}_n)^{-1/2}\mathcal{I}(\theta_0)^{-1/2}\sqrt{n}\mathcal{I}(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0)}_{=\sqrt{n}\mathcal{I}(\hat{\theta}_n)^{1/2}(\hat{\theta}_n-\theta_0)} \xrightarrow{D} \underbrace{I \times \mathrm{N}(0, I)}_{=\mathrm{N}(0,I)}$$
- Eq. 15 results by using the USLLN and Slutsly theorems. So I just need to show that
$$\frac{1}{n}\mathcal{J}_n(\hat{\theta}_n) \xrightarrow{a.s.} \mathcal{I}(\theta_0)$$
Set $U(x, \theta) = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(f(x|\theta))$, and $\mathcal{I}(\theta) = \mathrm{E}(U(x, \theta))$. Then
$$|\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(f(x_i|\hat{\theta}_n))\right)}_{U(x_i,\hat{\theta}_n)} - \mathcal{I}(\theta_0)| \leq |\frac{1}{n}\sum_{i=1}^{n}U(x_i, \hat{\theta}_n) - \mathcal{I}(\hat{\theta}_n)| + |\mathcal{I}(\hat{\theta}_n) - \mathcal{I}(\theta_0)| \tag{16}$$
$$\leq \sup_{|\hat{\theta}_n - \theta_0| \leq \delta} |\frac{1}{n}\sum_{i=1}^{n}U(x_i, \theta) - \mathcal{I}(\theta)| + |\mathcal{I}(\hat{\theta}_n) - \mathcal{I}(\theta_0)| \tag{17}$$
The first term converges to zero because the assumptions of the USLLN are satisfied. The second term converges to zero because $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$ and hence $\mathcal{I}(\hat{\theta}_n) \xrightarrow{a.s.} \mathcal{I}(\theta_0)$ by using Slutsky theorem.
So by Slutsy $(\frac{1}{n}\mathcal{J}_n(\hat{\theta}_n))^{1/2}\mathcal{I}(\theta_0)^{-1/2} \xrightarrow{a.s.} I$, and by Slutsky again
$$\underbrace{(\frac{1}{n}\mathcal{J}_n(\hat{\theta}_n))^{1/2}\mathcal{I}(\theta_0)^{-1/2}I(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0)}_{=\mathcal{J}_n(\hat{\theta}_n)^{1/2}(\hat{\theta}_n-\theta_0)} \xrightarrow{D} \underbrace{I \times \mathrm{N}(0, I)}_{=\mathrm{N}(0,I)}$$

4 The information inequality

Notation 29. Let $X_1, X_2, ...$ be IID random variables (unseen observations) with a distribution f_θ labeled by a parameter $\theta \in \Theta$. Let $\hat{\theta}_n := \hat{\theta}_n(X_{1:n})$ be a sequence of estimators of θ.

Definition 30. For two covariance matrices Σ_1 and Σ_2 we say $\Sigma_1 \geq \Sigma_2$ if $\Sigma_1 - \Sigma_2$ is positive semi-definite[3].

Note 31. The theorem below provides a lower bound (Cramer-Rao lower bound) of the variance of an estimator $\hat{\theta}_n$ of a multivariate parameter θ.

Theorem 32. *(Information inequality theorem) Let $X_1, ..., X_n \in \mathbb{R}^d$ be a sequence of IID random samples from a distribution $f_\theta(\cdot)$ labeled by an parameter $\theta \in \Theta \subset \mathbb{R}^r$ and admitting PDF $f(\cdot|\theta)$. Consider an estimator $\hat{\theta}_n := \hat{\theta}_n(X_{1:n}) \in \Theta \subset \mathbb{R}^r$ such that $g_n(\theta) = E_{f_\theta}(\hat{\theta}_n)$ exists on Θ.*

Then

$$var_{f_\theta}(\hat{\theta}_n) \geq \frac{1}{n}\dot{g}_n(\theta)\mathcal{I}(\theta)^{-1}\dot{g}_n(\theta)^T \tag{18}$$

where $\mathcal{I}(\theta)$ is the Fisher's information matrix.

- *we assumed that, $\frac{d}{d\theta}f(x|\theta)$ exists ; $\frac{d}{d\theta}$ can pass under the integral sign in $\int f(X|\theta)dX$ and $\int \hat{\theta}_n(X)f(X|\theta)dX$.*

Proof. It is given in the Appendix D. □

Definition 33. Cramer-Rao lower bound (CRLB) is called the quantity $\frac{1}{n}\dot{g}_n(\theta)\mathcal{I}(\theta)^{-1}\dot{g}_n(\theta)^T$ in (18).

Remark 34. Notice that the CRLB is attained if and only if $\hat{\theta}_n$ and $\Psi(X_{1:n}, \theta)$ are linearly related.

Example 35. If the estimator $\hat{\theta}_n$ has bias $b_n(\theta) = \mathrm{E}_{f_\theta}\hat{\theta}_n - \theta$ then

$$\mathrm{var}_{f_\theta}(\hat{\theta}_n) \geq \frac{1}{n}(I + \dot{b}_n)\mathcal{I}(\theta)^{-1}(I + \dot{b}_n)^T \tag{19}$$

Solution. It is $b_n(\theta) = \underbrace{\mathrm{E}_{f_\theta}\hat{\theta}_n}_{g} - \theta \implies \dot{b}_n(\theta) = \dot{g}_n(\theta) - I$. So replacing the terms in (18), I get (19).

Proposition 36. *If $\hat{\theta}_n$ is an unbiased estimator of θ, Theorem 32 implies that*

$$var_{f_\theta}(\hat{\theta}_n) \geq \frac{1}{n}\mathcal{I}(\theta)^{-1} \tag{20}$$

Definition 37. Best unbiased estimator (BUE) of θ is called the estimator $\hat{\theta}_n$ which has the lowest variance compared to other unbiased estimators. (I.e., its variance equal to the lower bound in (20))

Example 38. (Cont. of Example 17) If $X_1, ..., X_n \overset{IID}{\sim} \mathrm{N}(\mu, \sigma^2)$ then, the Cramer-Rao lower bound of the variance of an unbiased estimator $\hat{\theta}_n$ of $\theta = (\mu, \sigma)$ is

$$\mathrm{var}_{\mathrm{N}(\mu,\sigma^2)}(\hat{\theta}_n)\frac{1}{n}\mathcal{I}(\theta)^{-1} = \frac{1}{n}\begin{bmatrix}\frac{1}{\sigma^2} & 0\\ 0 & 2\frac{1}{\sigma^2}\end{bmatrix}^{-1} = \frac{1}{n}\begin{bmatrix}\sigma^2 & 0\\ 0 & 2\sigma^2\end{bmatrix}$$

[3]Matrix A is called positive semidefinite iff for any z , $zAz^T \geq 0$. It is symbolized as $A \geq 0$

Exercise sheet

Exercise #25

5 Asymptotic efficiency

Note 39. Let $X, X_1, X_2, ...$ be a sequence of IID random variables with a distribution f_θ labeled by a parameter $\theta \in \Theta$. Let $\hat{\theta}_n := \hat{\theta}_n(X_{1:n})$ be a sequence of estimators of θ.

Definition 40. Estimator $\hat{\theta}_n$ is called asymptotically efficient estimator of θ if

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

and $\Sigma(\theta) = \mathcal{I}(\theta)^{-1}$ for all $\theta \in \Theta$, whatever the true value of the parameter θ is. Also, $\hat{\theta}_n$ is called asymptotically sub-efficient estimator of θ if $\Sigma(\theta) \neq \mathcal{I}(\theta)^{-1}$ for some $\theta \in \Theta$.

Note 41. There are counter-examples of asymptotically super-efficient estimators... where $\Sigma(\theta) \geq \mathcal{I}(\theta)$. This is because the Information inequality theorem refers to the exact Variance, and not the asymptotic one.

Remark 42. We observe that MLE estimators $\hat{\theta}_n$ are asymptotically efficient under the assumptions C.1-C.5 of Theorem 23.

Exercise sheet

Exercise #28

References

[1] Tom M Apostol. *Mathematical analysis; 2nd ed.* Addison-Wesley Series in Mathematics. Addison-Wesley, Reading, MA, 1974. URL `https://cds.cern.ch/record/105425`.

[2] R. A. Bradley and J. J. Gart. The asymptotic properties of ml estimators when sampling from associated populations. *Biometrika*, 49(1/2):205–214, 1962.

[3] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

[4] T. A. Severini. *Likelihood methods in statistics*. Oxford University Press, 2000.

Appendix

A Proof of Theorem 5.

Not examinable

Proof. For every $\epsilon > 0$, let $S_\epsilon = \{\theta \in \Theta : |\theta - \theta_0| \geq \epsilon\}$. Because S_ϵ is compact, $\mu(\theta)$ is continuous on S, and θ_0 is the unique maximizer, it is

$$\underbrace{\sup_{\theta \in S_\epsilon} \mu(\theta)}_{=c_\epsilon} < \mu(\theta_0).$$

Consider the event,

$$A_{n,\epsilon} = \{X_{1:n} : \sup_{\theta \in \Theta} |\bar{U}_n(\theta) - \mu(\theta)| < \delta_\epsilon\}. \tag{21}$$

Let $c_\epsilon = \sup_{\theta \in S} \mu(\theta)$, and pick a $\delta_\epsilon > 0$ such that

$$0 < \delta_\epsilon < \frac{\mu(\theta_0) - c_\epsilon}{2} \quad \Longrightarrow \quad c_\epsilon + \delta_\epsilon < \mu(\theta_0) - \delta_\epsilon. \tag{22}$$

Based on the event $A_{n,\epsilon}$ in 21, if I pick a δ_ϵ as in (22), it is

$$\sup_{\theta \in S_\epsilon} \bar{U}_n(\theta) < \sup_{\theta \in S_\epsilon} \mu(\theta) + \delta_\epsilon = c_\epsilon + \delta_\epsilon < \mu(\theta_0) - \delta_\epsilon \overset{(21)}{\leq} \bar{U}_n(\theta_0).$$

So if $\hat{\theta}_n \in S$ (aka $\hat{\theta}_n$ away from θ_0), then θ_0 would yield a strictly larger value of $\bar{U}_n(\cdot)$, which is a contradiction to $\hat{\theta}_n$ being the maximizer of $\bar{U}_n$. So it has to be $\hat{\theta}_n \notin S$. Then, because

$$\{x_{1:n} : |\hat{\theta}_n - \theta_0| < \epsilon\} \supset A_{n,\epsilon}. \tag{23}$$

holds for any $\epsilon > 0$, it is

$$\begin{aligned} \Pr(\{|\hat{\theta}_n - \theta_0| < \epsilon\},\, n \to \infty) &\leq \Pr(A_{n,\epsilon},\, n \to \infty) \\ &= \Pr(\sup_{\forall \theta \in \Theta} |\bar{U}_n(\theta) - \mu(\theta)|,\, n \to \infty) \overset{\text{USLLN}}{=} 1 \end{aligned} \tag{24}$$

Hence, it is$\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$. □

B Proof of Cramer Theorem 23

To prove that MLE is consistent we use Theorem 5 and (25). Essentially this is because we actually need to show that th

Proof. of Cramer Theorem 23

1. Let $X, X_1, ...X_n$ be a sequence of IID random samples from sampling distribution $f(\cdot|\theta_0)$ with density $f(x|\theta_0)$, and θ_0 is the real value of the unknown parameter $\theta \in \Theta$.

   MLE (ML equations roots) $\hat{\theta}_n$ is the maximizer of the function

$$\frac{1}{n}(\ell_n(\theta) - \ell_n(\theta_0)) = \frac{1}{n}\log\frac{L_n(\theta)}{L_n(\theta_0)} = \frac{1}{n}\sum_{i=1}^{n}\overbrace{\log\frac{f(X_i|\theta)}{f(X_i|\theta_0)}}^{=U(X_i,\theta)}$$

   The real value θ_0 is the maximizer of

$$\mathrm{E}_{\theta_0}\log\frac{f(X|\theta)}{f(X|\theta_0)} = -\mathrm{KL}(f_0, f_1) \ : \begin{cases} < 0 & \text{, if } f(X|\theta) \neq f(X|\theta_0),\ a.s. \\ = 0 & \text{, if } f(X|\theta) = f(X|\theta_0),\ a.s. \end{cases}$$

   because $-\mathrm{KL}(f_0, f_1) < 0$ if $f(X|\theta) \neq f(X|\theta_0)$ a.s., and $-\mathrm{KL}(f_0, f_1) = 0$ if $f(X|\theta) = f(X|\theta_0)$ a.s. .

   Moreover from the SLLN, it is

$$\frac{1}{n}\log\frac{L_n(\theta)}{L_n(\theta_0)} = \frac{1}{n}\sum_{i=1}^{n}\overbrace{\log\frac{f(X_i|\theta)}{f(X_i|\theta_0)}}^{U(X_i,\theta)=} \xrightarrow{a.s.} \mathrm{E}_{\theta_0}\overbrace{\log\frac{f(X|\theta)}{f(X|\theta_0)}}^{\overbrace{\phantom{\log\frac{f(X|\theta)}{f(X|\theta_0)}}}^{\mu(\theta)=}\ U(X,\theta)=} \tag{25}$$

   where $f_\theta(\cdot) = f(\cdot|\theta)$, $f_{\theta_0}(\cdot) = f(\cdot|\theta_0)$, and $U(X_i, \theta)$ & $\mu(\theta)$ show the correspondence to (4).

   Then from Theorems 2 and 5 we have $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$.

2. Regarding the asymptotic distribution: I expand $\dot{\ell}(\theta)$ around θ_0 by Taylor expansion (MVT...) as

$$\dot{\ell}_n(\theta) = \dot{\ell}_n(\theta_0) + \int_0^1 \sum_{i=1}^{n} \dot{\Psi}(X_i, \theta_0 + u(\theta - \theta_0))\mathrm{d}u\ (\theta - \theta_0) \tag{26}$$

   Let $\hat{\theta}_n$ is any consistent ML equation root such that $\dot{\ell}_n(\hat{\theta}_n) = 0$. Then by setting $\theta = \hat{\theta}_n$, dividing by $\sqrt{n}$, and rearranging a bit the terms in (26), I get

$$C_n\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}}\dot{\ell}_n(\theta_0) \xrightarrow{D} Z \sim \mathrm{N}(0, \mathcal{I}(\theta_0)^{-1}) \tag{27}$$

   where

$$C_n = -\int_0^1 \sum_{i=1}^{n} \dot{\Psi}(X_i, \theta_0 + u(\hat{\theta}_n - \theta_0))\mathrm{d}u$$

   To derive (12) from (27), I need to show that $C_n \xrightarrow{a.s.} \mathcal{I}(\theta_0)$.

It is

$$|C_n - \mathcal{I}(\theta_0)| \leq |C_n + \mathcal{I}(\theta_0)| \leq \int_0^1 |\sum_{i=1}^n \dot{\Psi}(X_i, \theta_0 + u(\hat{\theta}_n - \theta_0)) + \mathcal{I}(\theta_0)| \mathrm{d}u$$
$$\leq \int_0^1 \underbrace{\sup_{\theta \in \{\theta : |\hat{\theta}_n - \theta_0| \leq \delta\}} |\sum_{i=1}^n \dot{\Psi}(X_i, \theta_0 + u(\hat{\theta}_n - \theta_0)) - \mathrm{E}_{\theta_0}(\dot{\Psi}(X, \theta))|}_{=(ii)}$$
$$+ \underbrace{|\mathrm{E}_{\theta_0}(\dot{\Psi}(x, \theta)) + \mathcal{I}(\theta_0)|}_{=(i)} \mathrm{d}u$$

- About (ii)... $\dot{\Psi}(X, \theta)$ is continuous at θ_0 from [C.3] so by definition, $(\forall \epsilon' > 0)(\exists \delta > 0)$ where

$$|\mathrm{E}_{\theta_0}(\dot{\Psi}(X, \theta)) + \mathcal{I}(\theta_0)| < \epsilon \qquad \text{whenever} \quad |\hat{\theta}_n - \theta_0| < \delta \tag{28}$$

- About (i)... I can restrict the neighborhood below the sup to be smaller than that of (28). Also the conditions of Theorem 3, are satisfied from C.2 and C.3. So $(\forall \epsilon'' > 0)(\exists N > 0)(\forall n > N)$

$$\sup_{\theta \in \{\theta : |\hat{\theta}_n - \theta_0| \leq \delta\}} |\sum_{i=1}^n \dot{\Psi}(X_i, \theta_0 + u(\hat{\theta}_n - \theta_0)) - \mathrm{E}_{\theta_0}(\dot{\Psi}(X, \theta))| < \epsilon''$$

So I get that $(\forall \epsilon > 0)(\exists N > 0)(\forall n > N)$ $|C_n - \mathcal{I}(\theta_0)| < \epsilon = \max(\epsilon', \epsilon'')$, and hence $C_n \xrightarrow{D} \mathcal{I}(\theta_0)$.

By using Slutsky theorems,

$$\cancel{C_n^{-1} C_n} \cancel{\xrightarrow{I}} \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{I}(\theta_0) Z \sim \mathrm{N}(0, \mathcal{I}(\theta_0) \cancel{\mathcal{I}(\theta_0)^{-1} \mathcal{I}(\theta_0)}) \cancel{\xrightarrow{I}}$$

□

C Extension of consistency and asymptotic Normality to non-IID cases

Not examinable

The Theorem below is a re-statement of Cramer's Theorem 23, for independent but non-identically distributed samples.

Note 43. While Cramer's Theorem 23 can be used for instance for to find the asymptotic Normality of the MLE in contingency tables under Multinational Sampling, it is not suitable to do the same for the same application but under product of Multinationals. The modification below can.

Theorem 44. *For $i = 1, ..., k$, let $X_{i,1}, ..., X_{i,m_i}$, be a sequence of m_i IID random samples drawn from sampling distribution $df_i(X_{i,1}, ..., X_{i,m_i}|\theta)$ labeled by a d-dimensional parameter $\theta \in \Theta$; namely*

$$X_{1,1}, ..., X_{1,m_1} \overset{IID}{\sim} df_1(\cdot|\theta) \quad ; \cdots ; \quad X_{i,1}, ..., X_{i,m_i} \overset{IID}{\sim} df_i(\cdot|\theta) \quad ; \cdots ; \quad X_{k,1}, ..., X_{k,m_k} \overset{IID}{\sim} df_k(\cdot|\theta)$$

Let the Cramer theorem conditions [C.1-C.5] hold for each $f_i(\cdot|\theta)$. Let the true value of θ be θ_0. Let $n = \sum_{i=1}^{k} m_i$.
Then there exists sequence $\hat{\theta}_n$ of roots of the likelihood equations, where

- *it is strongly consistent*
$$\hat{\theta}_n \xrightarrow{as} \theta_0$$
- *it has asymptotic distribution such as*
$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, \mathcal{I}(\theta_0)^{-1}) \tag{29}$$
where $\mathcal{I}(\theta) = \sum_{i=1}^{k} \varpi_i \mathcal{I}^{(i)}(\theta)$ is the expected information matrix, $\mathcal{I}^{(i)}(\theta)$ is the Fisher information matrix for $f_i(\cdot|\theta)$, $\varpi_i = m_i / \sum_{i=1}^{k} m_i$ for $i = 1, ..., k$.

Proof. [2] The proof is omitted and can be found in [2]; we offer the rational of the proof. The log-likelihood is
$$\ell_n(\theta) = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \log(f_i(x_{i,j}|\theta)) \tag{30}$$
and the roots of the likelihood equations are such as $0 = \ell_n(\theta)|_{\theta=\hat{\theta}_n}$. Since $\ell_n(\theta)$, and hence its Taylor expansion, is a sum of k independent terms of the kind considered in the proof of Theorem 23, then (29) could be true. □

D Proof of information inequality theorem 32

A complete proof of information inequality theorem Theorem 32 is given below.

Proof. Let $\Psi(X_{1:n}, \theta) = (\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X_{1:n}|\theta))^T = (\frac{\mathrm{d}}{\mathrm{d}\theta} \log \prod_{i=1}^{n} f(X|\theta))^T = \sum_{i=1}^{n} \Psi(X_i, \theta)$, where $\Psi(X_i, \theta) = (\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X_i|\theta))^T$. It is

$$\begin{aligned}
\mathrm{E}_{f_\theta} \Psi(X_{1:n}, \theta) =& 0 \quad \text{(you have proved it before)} \\
\dot{g}_n(\theta) =& \frac{\mathrm{d}}{\mathrm{d}\theta} \int \hat{\theta}_n f(X_{1:n}|\theta) \mathrm{d}X = \int \hat{\theta}_n(x_{1:n}) \frac{\frac{\mathrm{d}}{\mathrm{d}\theta} f(X_{1:n}|\theta)}{f(X_{1:n}|\theta)} f(X|\theta) \mathrm{d}X \\
=& \int \hat{\theta}_n \frac{\mathrm{d}}{\mathrm{d}\theta} \log f(X_{1:n}|\theta) f(X_{1:n}|\theta) \mathrm{d}X = \mathrm{E}_{f_\theta}(\hat{\theta}_n \Psi(X_{1:n}, \theta) - \cancelto{0}{\mathrm{E}_\theta \Psi(X_{1:n}, \theta)}) \\
=& \mathrm{cov}_{f_\theta}(\hat{\theta}_n, \Psi(X_{1:n}, \theta))
\end{aligned} \tag{31}$$

For any $a, \gamma \in \mathbb{R}^r$, $\xi = a^T \hat{\theta}_n$ and $\zeta = \gamma^T \Psi(X_{1:n}, \theta)$. It is

$$\begin{aligned}
\mathrm{cov}_{f_\theta}(\xi, \zeta) =& a^T \mathrm{cov}_{f_\theta}(\hat{\theta}_n, \Psi(X_{1:n}, \theta)) \gamma = a^T \dot{g}_n(\theta) \gamma \\
\mathrm{var}_{f_\theta}(\zeta) =& \gamma^T \mathrm{var}_{f_\theta}(\Psi(X_{1:n}, \theta)) \gamma = \gamma^T \left(n \mathcal{I}(\theta) \right) \gamma \\
\mathrm{var}_{f_\theta}(\xi) =& a^T \mathrm{var}_{f_\theta}(\hat{\theta}_n) a
\end{aligned}$$

So

$$1 \geq (\mathrm{corr}_{f_\theta}(\xi,\zeta))^2 = \frac{a^T \dot{g}_n(\theta)\gamma}{a^T \mathrm{var}_{f_\theta}(\hat{\theta}_n) a \, \gamma^T \left(n\mathcal{I}(\theta)\right)\gamma}$$

the right maximizes[4] with respect to γ at $\gamma^* = \frac{\mathcal{I}(\theta)^{-1}\dot{g}_n(\theta)^T a}{(a^T\mathcal{I}(\theta)a)^{1/2}}$ for all a and gives

$$1 \geq \frac{a^T \dot{g}_n(\theta)\left(n\mathcal{I}(\theta)\right)^{-1}\dot{g}_n(\theta)^T a}{a^T \mathrm{var}_{f_\theta}(\hat{\theta}_n) a}$$

namely

$$a^T\left(\mathrm{var}_{f_\theta}(\hat{\theta}_n) - \frac{1}{n}\dot{g}_n(\theta)\mathcal{I}(\theta)^{-1}\dot{g}_n(\theta)^T\right)a \geq 0$$

which completes the proof. □

Note 45. An alternative but incomplete proof of information inequality theorem Theorem 32 is given below.

Proof. Let $\Psi(x,\theta) = (\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\theta))^T$.

It is

$$\begin{aligned}
\mathrm{E}_{f_\theta}\Psi(X,\theta) =& 0 \quad \text{(you have proved it before)} \\
\dot{g}_n(\theta) =& \frac{\mathrm{d}}{\mathrm{d}\theta}\int \hat{\theta}_n(x) f(x|\theta)\mathrm{d}x = \int \hat{\theta}_n(x)\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}f(x|\theta)}{f(x|\theta)} f(x|\theta)\mathrm{d}x \\
=& \int \hat{\theta}_n(x)\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x|\theta) f(x|\theta)\mathrm{d}x = \mathrm{E}_{f_\theta}(\hat{\theta}_n(x)(\Psi(x,\theta) - \cancelto{0}{\mathrm{E}_\theta\Psi(X,\theta)})) \\
=& \mathrm{cov}_{f_\theta}(\hat{\theta}_n(x), \Psi(x,\theta))
\end{aligned} \tag{32}$$

So

$$\begin{aligned}
0 \leq& \mathrm{var}_{f_\theta}(\hat{\theta}_n - \dot{g}(\theta)\mathcal{I}(\theta)^{-1}\Psi(x,\theta)) \\
=& \mathrm{var}_{f_\theta}(\hat{\theta}_n) - 2\frac{1}{n}\dot{g}(\theta)\mathcal{I}(\theta)^{-1}\dot{g}(\theta)^T + \frac{1}{n}\dot{g}(\theta)\mathcal{I}(\theta)^{-1}\mathcal{I}(\theta)\mathcal{I}(\theta)^{-1}\dot{g}(\theta)^T \\
=& \mathrm{var}_{f_\theta}(\hat{\theta}_n) - \frac{1}{n}\dot{g}(\theta)\mathcal{I}(\theta)^{-1}\dot{g}(\theta)^T
\end{aligned}$$

and the proof is done □

[4] Fact: from Linear algebra: It is $\max_{\forall x}\frac{(a^T x)^2}{x^T B x} = a^T B^{-1} a$ where the maximum is attained at $x^* = \frac{B^{-1}a}{(x^T B^{-1}x)^{1/2}}$