Estimation

Sofia Olhede

EPFL

October 6, 2020

1. Estimation

2. Maximum Likelihood Estimation

Proof of Rao–Blackwell Theorem

- To commence the proof we note that since T is assumed to be sufficient for θ.
- To refer back to Rao-Blackwellising we note that $\mathbb{E}\{\hat{\theta}|T=t\}=h(t)$ is independent of θ.
- Thus $\widehat{\theta}^{*}=\mathbb{E}\{\widehat{\theta}|T\}$ is a well-defined statistic (it is calculated from $\boldsymbol{Y}$ but is not a function of θ).
- We note that

$$\mathbb{E}\{\widehat{\theta}^{*}\}=\mathbb{E}\{\mathbb{E}\{\widehat{\theta}|T\}\}=\mathbb{E}\,\widehat{\theta}=\theta.$$

- By the law of total variance we have due to the non-negativity of variances

$$\mathbb{V}\mathrm{ar}\{\widehat{\theta}\}=\mathbb{V}\mathrm{ar}\{\mathbb{E}\{\widehat{\theta}|T\}\}+\mathbb{E}\,\mathbb{V}\mathrm{ar}\{\widehat{\theta}|T\}\geq\mathbb{V}\mathrm{ar}\{\mathbb{E}\{\widehat{\theta}|T\}\}=\mathbb{V}\mathrm{ar}\{\widehat{\theta}^{*}\}.$$

Proof of Rao–Blackwell Theorem

- We can directly note that

$$\mathbb{V}\text{ar}\{\widehat{\theta}|T\} = \mathbb{E}\{(\widehat{\theta} - \mathbb{E}\{\hat{\theta}|T\})^2|T\} = \mathbb{E}\{\left(\widehat{\theta} - \widehat{\theta}^*\right)^2|T\}.$$

- Thus it follows that $E\,\mathbb{V}\text{ar}\{\widehat{\theta}|T\} = \mathbb{E}\{\left(\widehat{\theta} - \widehat{\theta}^*\right)^2\} > 0$ unless if $\Pr(\widehat{\theta} = \widehat{\theta}^*) = 1$.
- Now assume that $\widehat{\theta}$ is an unbiased estimator of θ and that T and S are θ-sufficient.
- We can also ask ourselves how do $\mathbb{V}\text{ar}\{\mathbb{E}\left(\widehat{\theta}|T\right)\}$ and $\mathbb{V}\text{ar}\{\mathbb{E}\left(\widehat{\theta}|S\right)\}$ relative?
- Our intuition would suggest that whichever of T and S carries the least irrelevant information should do better.
- Formally, if $T = h(S)$ then $\widehat{\theta}^*_T$ should dominate $\widehat{\theta}^*_S$.

Variance preference for sufficient statistics

- Proposition: for $\widehat{\theta}$ an unbiased estimator of θ and T and S sufficient statistics, define

$$\widehat{\theta}^*_T = \mathbb{E}\{\widehat{\theta} \,|\, T\}, \quad \& \quad \widehat{\theta}^*_S = \mathbb{E}\{\widehat{\theta} \,|\, S\}.$$

Then

$$T = h(S) \quad \Longrightarrow \mathbb{V}\mathrm{ar}\{\widehat{\theta}^*_T\} \leq \mathbb{V}\mathrm{ar}\{\widehat{\theta}^*_S\}.$$

- We can deduce from this result that the best way of Rao–Blackwellising is conditioning on the minimally sufficient statistic.

Proof

- We start from the tower property of conditional expectation. If $Y = f(X)$ then we see

$$\mathbb{E}\{Z|Y\} = \mathbb{E}\{\mathbb{E}\{Z|X\}|Y\}.$$

- Since $T = f(S)$ we can deduce that

$$\begin{aligned}\widehat{\theta}^*_T &= \mathbb{E}\{\widehat{\theta}|T\}\\ &= \mathbb{E}\{\mathbb{E}\{\widehat{\theta}|S\}|T\}\\ &= \mathbb{E}\{\widehat{\theta}^*_S|T\}. \qquad (1)\end{aligned}$$

  We can now conclude the desired result using the Rao–Blackwell theorem.

- We can now judge the quality of any given estimator. For certain cases we even know how good performance we can hope for.

ML Estimation

- How do we then come up with (good) estimators?
- We do want a method that works in general to come up with estimators.
- We even want methods that give good estimators. We shall do so using the method of maximum likelihood.
- This estimator was proposed by Ron Fisher in 1921. A more detailed exposition was provided in "On the mathematical foundations of theoretical statistics." Philos. Trans. Roy. Soc. London Ser. A 222, 309–368.
- How do we motivate the principle of maximum likelihood? Fisher: "*On the bases that the purpose of the statistical reduction of data is to obtain statistics which shall contain as much as possible, ideally the whole, of the relevant information contained in the sample . . .* ". Let us see how that is done in practice.

ML Estimation II

- Statistics is about "inverse probability" (more to follow).
- Likelihood from a probability perspective: Given a parameter $\theta \in \Theta$ then for any $(y_1, \ldots, y_n)^T \in \mathcal{Y}^n$ then we can evaluate

  $$(y_1, \ldots, y_n) \mapsto \Pr_\theta(Y_1 = y_1, \ldots, Y_n = y_n),$$

  namely how the probability varies as a function of the sample.
- Likelihood from a statistics perspective: Given a sample $(y_1, \ldots, y_n)^T \in \mathcal{Y}^n$ then for any $\theta \in \Theta$ we can calculate

  $$\theta \mapsto \Pr_\theta(Y_1 = y_1, \ldots, Y_n = y_n).$$

- How the probability varies as a function of θ or the model.
- The maximum likelihood principle is to select a value of θ that makes the observations most likely.

ML Estimation III

- Definition: Let $(Y_1, \dots, Y_n)$ be a sample of random variables with joint density/frequency $f(y_1, \dots, y_n; \theta)$ where $\theta \in \mathbb{R}^p$. The likelihood of θ is defined as
$$L(\theta) = f(Y_1, \dots, Y_n; \theta).$$
- If $(Y_1, \dots, Y_n)^T$ has i.i.d. entries, each with density/frequency $f(y_j; \theta)$ then
$$L(\theta) = \prod_{j=1}^{n} f(y_j; \theta).$$
We get the following estimation method.
- Definition: (Maximum Likelihood Estimation). In the same context, a maximum likelihood estimator (MLE) of $\widehat{\theta}$; is an estimator such that
$$L(\theta) \leq L(\widehat{\theta}), \quad \forall \theta \in \Theta.$$

ML Estimation IV

- When there exists a unique maximum, we speak of the MLE $\widehat{\theta} = \arg_{\theta\in\Theta}\max L(\theta)$.
- The likelihood is a random function.
- It is the joint density/frequency of the sample, but viewed as a function of θ.
- It is NOT the probability of θ.
- $L(\theta)$ is the answer to the question how does the joint density / probability of the sample vary as we vary θ.
- In the discrete case it is exactly "the probability of observing our sample" as a function of θ.
- If a sufficient statistic T exists for θ then Fisher–Neyman factorisation implies

$$L(\theta) = g(T(\mathrm{Y});\theta)h(\mathrm{Y}) \propto g(T(\mathrm{Y});\theta).$$

ML Estimation V

- Any MLE depends on data only through a sufficient statistic.
- Since the sufficient statistic was arbitrary, if a minimally sufficient statistic exists, the MLE will have used an estimator that has achieved the maximal sufficient reduction of the data.
- MLE's are also equivariant. Sometimes this is also known as invariance. If $g : \Theta \mapsto \Theta'$ is a bijection, and if $\widehat{\theta}$ is the MLE of θ, then $g(\widehat{\theta})$ is the MLE of $g(\theta)$.
- When the likelihood is differentiable in θ, its maximum must solve

$$\nabla_\theta L(\theta)|_{\theta=\widehat{\theta}} = 0.$$

- We must verify it is a maximum (don't want the minimum likelihood solution!)

ML Estimation VI

- If the likelihood is twice differentiable in θ, we can verify this by checking
$$-\nabla_\theta^2 L(\theta)\big|_{\theta=\widehat{\theta}} \succ 0.$$
- The negative of the Hessian is positive definite. In one dimension, his reduces to the standard second derivative criterion.
- Solving $\nabla_\theta L(\theta)\big|_{\theta=\widehat{\theta}} = 0$ when Y_i is independent corresponds to solving an equation after using the Leibnitz rule.
- Instead of maximising $L(\theta)$ we take an increasing function of $L(\theta)$ and maximise that.
- We chose to use $\log(x)$ as that function which is strictly increasing.
- When the Y_i are independent then $\ell(\theta) = \log L(\theta)$ is a sum rather than a product
$$\ell(\theta) = \log\left(\prod_{j=1}^{n} f(y_j;\theta)\right) = \sum_{j=1}^{n} \log(f(y_j;\theta)).$$

ML Estimation VI

- Of course, under twice differentiability, verification of a maximum can be checked again by whether or not

$$\nabla_\theta L(\theta)\big|_{\theta=\widehat{\theta}} = 0, \quad -\nabla_\theta^2 L(\theta)\big|_{\theta=\widehat{\theta}} \succ 0.$$

- Example (MLE for Bernoulli trials). Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{Bern}(p)$. The likelihood is

$$L(p) = \prod_{j=1}^{n} f(Y_j; p) = \prod_{j=1}^{n} p^{Y_j}(1-p)^{1-Y_j} = p^{\sum_j Y_j}(1-p)^{n-\sum_j Y_j}.$$

- This gives a loglikelihood (or a loglihood) of

$$\ell(p) = \log(p^{\sum_j Y_j}(1-p)^{n-\sum_j Y_j}) = \sum_j Y_j \log(p) + (n - \sum_j Y_j) \log(1-p).$$

ML Estimation VII

- This is twice differentiable. We can therefore calculate

$$\frac{d}{dp}\ell(p) = \sum_j Y_j \frac{1}{p} - (n - \sum_j Y_j)\frac{1}{1-p}. \qquad (2)$$

  Setting this equal to 0 and solving yields $\widehat{p} = \frac{1}{n}\sum_j Y_j = \overline{Y}$.

- We now calculate the second derivative which yields

$$\frac{d^2}{dp^2}\ell(p) = -\sum_j Y_j \frac{1}{p^2} - (n - \sum_j Y_j)\frac{1}{(1-p)^2} < 0. \qquad (3)$$

- Thus it follows that

$$\widehat{p} = \overline{Y},$$

  is a unique MLE for the Bernoulli success probability.

ML Estimation VIII

- Example (MLE for Exponential trials). Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{Expl}(\lambda)$.
- The likelihood for this data takes the form

$$L(\lambda) = \prod_{j=1}^{n} f(Y_j; \lambda) = \prod_{j=1}^{n} \lambda e^{-\lambda Y_j} = \lambda^n e^{-\lambda \sum_j Y_j}.$$

- This gives a loglikelihood (or a loglihood) of

$$\ell(\lambda) = \log(L(\lambda)) = n \log \lambda - \lambda n \bar{Y}.$$

- Differentiating we find

$$\frac{d}{d\lambda} \ell(\lambda) = n\lambda^{-1} - n\bar{Y}.$$

Thus we take $\widehat{\lambda} = \bar{Y}^{-1}$.

ML Estimation IX

- Differentiating yet one more time we get

$$\frac{d^2}{d\lambda^2}\ell(\lambda) = -n\lambda^{-2} < 0.$$

- Thus we determine that

$$\widehat{\lambda} = \bar{Y}^{-1},$$

is a unique MLE.

ML Estimation X

- Example (MLE for Gaussian trials). Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$.
- This process has two parameters, $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$.
- The likelihood is then

$$L(\mu, \sigma^2) = \prod_{j=1}^{n} f(Y_j; \mu, \sigma^2) = \prod_{j=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(Y_j - \mu)^2\} \right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (Y_j - \mu)^2\}.$$

- Taking logarithms we get

$$\ell(\mu, \sigma^2) = -(n/2)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{n} (Y_j - \mu)^2.$$

We can now seek to maximise this quantity.

ML Estimation XI

- Taking partial derivatives we get

$$\frac{\partial}{\partial\mu}\ell(\mu,\sigma^2) = 2\frac{1}{2\sigma^2}\sum_{j=1}^{n}(Y_j-\mu)$$

$$\frac{\partial}{\partial\sigma^2}\ell(\mu,\sigma^2) = -(n/2)\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{j=1}^{n}(Y_j-\mu)^2.$$

- This leads to a set of equations that we can solve for $\widehat{\mu}$ and $\widehat{\sigma^2}$:

$$\begin{pmatrix}\widehat{\mu} & \widehat{\sigma^2}\end{pmatrix} = \begin{pmatrix}\bar{Y} & \frac{1}{n}\sum_{j=1}^{n}(Y_j-\bar{Y})^2\end{pmatrix}.$$

- Now we need to verify that this corresponds to a maximum, by calculating second and mixed partial derivatives.

ML Estimation XII

- Note that

$$\frac{\partial^2}{\partial \mu^2}\ell(\mu,\sigma^2) = -\frac{n}{\sigma^2}, \quad \frac{\partial^2}{\partial (\sigma^2)^2}\ell(\mu,\sigma^2) = \frac{(n/2)}{\sigma^4} - \frac{2}{2\sigma^6}\sum_{j=1}^{n}(Y_j-\mu)^2$$

$$\frac{\partial^2}{\partial \mu \partial \sigma}\ell(\mu,\sigma^2) = -2\frac{1}{\sigma^4}\sum_{j=1}^{n}(Y_j-\mu).$$

We now evaluate the derivatives and get

$$\frac{\partial^2}{\partial \mu^2}\ell(\widehat{\mu},\widehat{\sigma^2}) = -\frac{n}{\widehat{\sigma^2}}, \quad \frac{\partial^2}{\partial (\sigma^2)^2}\ell(\widehat{\mu},\widehat{\sigma^2}) = \frac{(n/2)}{\widehat{\sigma^4}} - \frac{2n}{2\widehat{\sigma^4}} = -\frac{(n/2)}{\widehat{\sigma^4}}$$

$$\frac{\partial^2}{\partial \mu \partial \sigma}\ell(\widehat{\mu},\widehat{\sigma^2}) = 0.$$

Thus the Hessian is diagonal. We can also note that the negative of the matrix is positive definite and diagonal- we have a maximum.

ML Estimation XIII

- Example (MLE for Poisson observations). Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{Poisson}(\mu)$.
- This process has one parameter, $\mu \in \mathbb{R}^+$.
- The likelihood is then

$$L(\mu) = \prod_{j=1}^{n} f(Y_j; \mu) = \prod_{j=1}^{n} \left\{ \frac{e^{-\mu} \mu^{Y_j}}{Y_j!} \right\} = e^{-n\mu} \mu^{\sum_{j=1}^{n} Y_j} \frac{1}{\prod_j Y_j!}.$$

Remember the definition of the factorial sign $n! = \Gamma(n+1) = n(n-1)(n-2)\ldots 1$. Note that $0! = 1$, $1! = 1$, $2! = 2$, $3! = 6$ etc. The loglihood is

$$\ell(\mu) = -n\mu + \sum_{j=1}^{n} Y_j \log(\mu) - \log \prod_j Y_j!$$

$$\frac{\partial}{\partial \mu} \ell(\mu) = -n + \sum_{j=1}^{n} Y_j / \mu.$$

ML Estimation XIV

- This is zero if $\mu = \bar{Y}$. We check the second derivative: $\frac{\partial^2}{\partial \mu^2}\ell(\mu) = -\sum_{j=1}^n Y_j/\mu^2$.
- Example (MLE for Uniform Distribution – a non–differentiable case). This example is useful as it shows that differentiation does not always work.
- The uniform distribution takes the form of $f(Y_j;\theta) = \theta^{-1}\mathrm{I}(Y_j \in (0,\theta))$ here. The likelihood takes the form of

$$L(\theta) = \prod_j \theta^{-1}\mathrm{I}(Y_j \in (0,\theta)) = \theta^{-n}\mathrm{I}(\max_j Y_j \in (0,\theta))\mathrm{I}(\min_j Y_j \in (0,\theta)).$$

We cannot differentiate this sensibly in theta as $\frac{\partial}{\partial\theta}L(\theta)$ has a discontinuity. We see that θ^{-n} has derivative $-n\theta^{-n+1}$, which is negative and so is decreasing in θ. As $\theta > Y_i$ for all i, that means we must take $\widehat{\theta} = \max_j Y_j$.

ML Estimation XV

- The next property we shall cover is the equi–variance or invariance of MLEs. If $g(\theta)$ is a bijection, recall that if we are attempting to estimate $\tau = g(\theta)$ then if we form the likelihood

$$L(\theta) = \prod_{j=1}^{n} f(Y_j; \theta),$$

- We can use the chain rule to determine

$$\frac{d}{d\tau} L(\tau) = \frac{d}{d\theta} L(\theta) \frac{d\theta}{d\tau},$$

and as $g(\theta)$ is a bijection, we know that $\frac{d\theta}{d\tau} \neq 0$. Thus we can easily map between zeros.

ML Estimation XVI

- Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$. Suppose we are interested in estimating $\tau = \Pr\{Y_1 \leq y\}$. Assume that $y \in \mathbb{R}$ and fix $\sigma \in \mathbb{R}^+$.
- We note that

$$\tau(\mu) = \Pr\{Y_1 \leq y\} = \Pr\{\frac{Y_1 - \mu}{\sigma} \leq \frac{y - \mu}{\sigma}\} = \Phi\left\{\frac{y - \mu}{\sigma}\right\}.$$

- We note that

$$\frac{d\tau}{d\mu} = \phi\left\{\frac{y - \mu}{\sigma}\right\}(-\frac{1}{\sigma}) < 0 \ \forall \ \mu.$$

- This implies that we have a bijective map, and so $\widehat{\tau}$ is $\tau(\widehat{\mu})$.

ML Estimation–Exponential Family

- Assume that $Y_1, \ldots, Y_n$ are all drawn from the distribution f which is specified by

$$f(y) = \exp\{\phi T(y) - \gamma(\phi) + S(y)\}, \quad y \in \mathcal{Y}.$$

- We set ϕ as the natural parameter.
- Assume also that $\phi = \eta(\theta)$, where $\theta \in \Theta$ is the usual parameter.
- Assume that $\eta : \Theta \to \Phi$ is a differentiable bijection. Thus $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$. $d = \gamma \circ \eta$.
- Thus the density can be rewritten:

$$f(y) = \exp\{\phi T(y) - \gamma(\phi) + S(y)\} = \exp\{\eta(\theta) T(y) - d(\theta) + S(y)\}.$$

- Equivariance implies that if $\widehat{\theta}$ is an MLE for θ then so is $\eta(\widehat{\theta})$ for $\phi = \eta(\theta)$. If $\widehat{\phi}$ is an MLE of ϕ then $\eta^{-1}(\widehat{\phi})$ is an MLE of $\eta^{-1}(\phi)$.

ML Estimation–Exponential Family

- Consistency of MLE in $\theta \in \mathbb{R}$. Let $Y_1, \ldots, Y_n \sim f(y;\theta_0)$ where $f \in C^1$ wrt to θ. Assume that for all n there is an unique MLE $\widehat{\theta}_n$. We will show $\widehat{\theta}_n \xrightarrow{p} \theta$.
- Define

$$\Xi_n(u) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial u}\log\left(\frac{f(Y_i;u)}{f(Y_i;\theta)}\right)\right], \quad \Xi(u) = \mathbb{E}\left[\frac{\partial}{\partial u}\log\left(\frac{f(Y_i;u)}{f(Y_i;\theta)}\right)\right],$$

- so that $\Xi_n(\widehat{\theta}_n) = 0$ by uniqueness of MLEs;
- $\Xi(\theta) = 0$ uniquely, assuming regularity allowing interchange of $\mathbb{E}()$ and $\frac{\partial}{\partial u}$.
- The law of large numbers implies that $\Xi_n(u) \xrightarrow{p} \Xi(u)$.
- This implies $\widehat{\theta}_n$ converges to θ. Consistency is equivalent to this statement.

ML Estimation–Exponential Family

- Consistency of MLE in $\mathbb{R}^k$ for exponential families.
- Consider $Y_1, \dots, Y_n \sim f(y;\phi)$ from a k exponential family:

$$f(y) = \exp\{\sum_{j=1}^{k} \phi_j T_j(y) - \gamma(\phi) + S(y)\}.$$

- The likelihood and log-likelihoods are given by

$$L(\boldsymbol{\phi}) = \exp\{\boldsymbol{\phi}^T \boldsymbol{\tau} - n\gamma(\boldsymbol{\phi})\}, \quad \ell(\boldsymbol{\phi}) = \boldsymbol{\phi}^T \boldsymbol{\tau} - n\gamma(\boldsymbol{\phi}).$$

  where $\boldsymbol{\tau} = (\tau_1, \dots, \tau_k)^T$ and $\tau_j(\boldsymbol{y}) = \sum_i T_j(y_i)$.
- If it exists the MLE $\widehat{\boldsymbol{\phi}}$ must satisfy

$$\nabla_{\phi} \ell(\widehat{\boldsymbol{\phi}}_n) = 0 \Rightarrow \nabla_{\phi} \gamma(\widehat{\boldsymbol{\phi}}_n) = n^{-1} \boldsymbol{\tau}.$$

  Furthermore, existence of the MLE guarantees uniqueness by strict concavity.

ML Estimation–Exponential Family

- Provided it exists, the MLE of the natural parameter in a k-parameter natural exponential family with open parameter space Φ is consistent.
- Assuming we can get consistency, we can focus on understanding the sampling distribution of the MLE.
- For simplicity, assume $X_1, \ldots, X_n$ are iid with density/frequency) $f(x;\theta)$ for $\theta \in \Theta$. Write
- Let $\ell(x_i;\theta) = \log f(x_i;\theta)$.
- Let $\ell'(x_i;\theta)$, $\ell''(x_i;\theta)$ and $\ell'''(x_i;\theta)$ denote the partial derivatives wrt θ.
- We need some regularity conditions:
- A1: Θ is an open subset of $\mathbb{R}$.
- A2: The support of f, $\mathrm{supp}\, f$ is independent of θ.
- A3: f is thrice continuously differentiable w.r.t. θ for all the support of f.

ML Estimation–Exponential Family

- A4: $\mathbb{E}(\ell'(X_i;\theta)) = 0$ for all θ and $\mathbb{V}\mathrm{ar}\{\ell'(X_i;\theta)\} = \mathcal{I}_1(\theta)$.
- A5: $-\,\mathbb{E}(\ell''(X_i;\theta)) = \mathscr{I}_1(\theta)$.
- A6: $\exists M(x) > 0$ and $\delta > 0$ such that $\mathbb{E}\,M(x) < \infty$ and

$$|\theta - \theta_0| < \delta \Rightarrow |\ell'''(x;\theta)| < M(x).$$