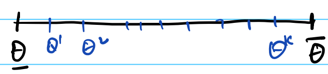
Nov 13, 2023.

Conditions for regularity :-

(i) Θ is compact
(ii) $R(\cdot)$ is continuous
(iii) $\hat{R}(\cdot)$ converges to $R(\cdot)$ uniformly in probability
(iv) $R(\cdot)$ attains a unique global minimum at θ_0

Approach to establish uniform continuity :-



leibzitz constant

$$|f(x) - f(y)| \leq L\,|x-y|$$

If leibstitz continous $\Rightarrow$

- discretize the sample space

For each pt, $\hat{R}(\theta^*) \xrightarrow{P} R(\theta^*)$

$$P\left(\max_{k=1,\dots,K} |\hat{R}(\theta_k) - R(\theta_k)| \geq \varepsilon\right) \leq \sum_{k=1}^{K} P\left(|\hat{R}(\theta_k) - R(\theta_k)| \geq \varepsilon\right)$$

In the context of MLE, we don't need to go into this length to establish uniform continuity.

Instead,

$$E\left[\sup_{\theta \in \Theta} |\log f(Y, \theta)|\right] < \infty$$

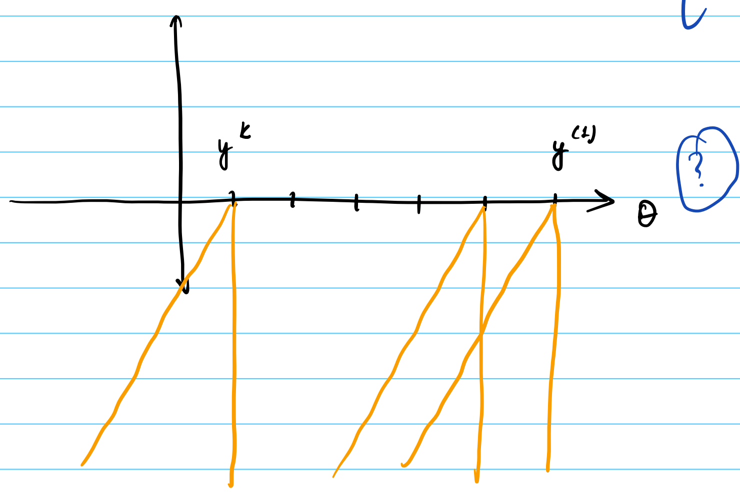
(we can replace cond 3 with this.)

Example : $f(y,\theta) = \begin{cases} e^{-(y-\theta)} & , y \geq \theta \\ 0 & , y < \theta \end{cases}$

log likelihood :

$$\log f(y,\theta) = \begin{cases} \theta - y & , y \geq \theta \\ -\infty & , y < \theta \end{cases}$$

$$\hat{l}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i, \theta) = \begin{cases} -\infty & , \text{if } \theta > y^{(n)} \\ \frac{1}{n}\sum_{i=1}^{n}(\theta - y_i), & \text{if } \theta \leq y^{(n)} \end{cases}$$



$\hat{\theta}_{MLE} = \min_{i=1,\dots,n} y_i$

Convergence to e^x.
(like in the Pset.)

1^{st} order statistic (when arranged decending order) → max.

smallest
~~n^{th}~~ order statistic → min.

Properties of MLE :-

(If another f^n approaches the properties of MLE, that's very useful too!)

* **Score of the MLE:** Gradient of the log likelihood f^n

d×1 vector $\leftarrow$ $s(\theta, y) = \nabla_\theta \log f(y, \theta)$ (log of your dist^n density)

Suppose that θ_0 is the unique global maximum of $L(\theta)$

what does this mean?

$$\theta \Rightarrow f(y, \theta) \Rightarrow L_\theta(\tilde{\theta}) = E_\theta[\log f(Y, \tilde{\theta})]$$

θ is the parameter that induces the dist^n of θ

$\tilde{\theta}$ is the argument of log likelihood f^n

$$\Rightarrow \nabla_{\tilde{\theta}} L_\theta(\tilde{\theta})\Big|_{\tilde{\theta}=\theta} = 0$$

Assuming we can swap differentiables with integrals (requires last class) :-

$$\nabla_{\tilde{\theta}} L_\theta(\tilde{\theta})\Big|_{\tilde{\theta}=\theta} = 0 = E_\theta\left[\nabla_{\tilde{\theta}} \log f(Y, \tilde{\theta})\right]\Big|_{\tilde{\theta}=\theta}$$

$$= E_\theta[s(\theta, Y)]$$

random vector $\Leftarrow$ $E_\theta[s(\theta, Y)] = 0$ $\longrightarrow$ system of equations

$$E_\theta[s(\theta, Y)\, s(\theta, Y)'] = I_\theta \longrightarrow \text{Covariance Matrix}$$

d×d matrix.

$\boxed{\text{Information Matrix}}$

(Since $E_\theta[s(\theta, Y)] = 0$, we don't need to recenter like in the usual cases.

Now,

$$E_\theta[s(\theta, Y)] = 0 \quad \forall\, \theta \in H$$

Now this can be treated as a curve & derivatives can be taken anywhere on the curve

For dimension k:

kth component

$$E_\theta[s^k(\theta, Y)] = 0 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} s^k(\theta, y)\, f(y, \theta)\, \underbrace{dy^1 \cdots dy^J}$$

depending on the dimensions of y.

By def;

$$s^k(\theta, y) = \frac{\partial \log f(y, \theta)}{\partial \theta^k}$$

$$\Rightarrow 0 = \int_{-\infty}^{+\infty} \nabla_\theta s^k(\theta, y)\, f(y, \theta)\, dy + \int_{-\infty}^{+\infty} s^k(\theta, y)\, \nabla_\theta f(y, \theta)\, dy$$

$$s(\theta, y) = \nabla_\theta \log f(y, \theta) = \frac{1}{f(y, \theta)} \nabla_\theta f(y, \theta)$$

$$0 = \int_{-\infty}^{+\infty} \nabla_\theta s^k(\theta, y)\, f(y, \theta)\, dy + \int_{-\infty}^{+\infty} s^k(\theta, y)\, \overbrace{\nabla_\theta f(y, \theta) \cdot \frac{1}{f(y, \theta)}}^{\text{score}} \cdot f(y, \theta)\, dy$$

$$0 = \underbrace{\int_{-\infty}^{+\infty} \nabla_\theta s^k(\theta, y)\, f(y, \theta)\, dy}_{\cdots\cdots} + \int_{-\infty}^{+\infty} s^k(\theta, y)\, s(\theta, y)\, f(y, \theta)\, dy$$

$$\cdots \cdots \cdots + E[s^k(\theta, Y)\, s(\theta, Y)]$$

$$\nabla_\theta s^k(\theta, y) = \begin{pmatrix} \dfrac{\partial^2 \log f(y,\theta)}{\partial\theta^1 \partial\theta^k} \\ \vdots \\ \dfrac{\partial^2 \log f(y,\theta)}{\partial\theta^n \partial\theta^k} \end{pmatrix}$$

$$L(\theta) = E[\log f(Y, \theta)]$$

Create similar vector for each dimension

$$\begin{pmatrix} \dfrac{\partial^2 L(\theta)}{\partial\theta^1 \partial\theta^k} \\ \vdots \\ \dfrac{\partial^2 L(\theta)}{\partial\theta^n \partial\theta^k} \end{pmatrix}$$

$$0 = \begin{pmatrix} \dfrac{\partial^2 L(\theta)}{\partial\theta^1 \partial\theta^k} \\ \vdots \\ \dfrac{\partial^2 L(\theta)}{\partial\theta^n \partial\theta^k} \end{pmatrix} + E[s^k(\theta, Y) s(\theta, Y)]$$

Since its true for any θ, transposing it!—

$$0 = \left(\frac{\partial^2 L(\theta)}{\partial\theta^1 \partial\theta^k} \cdots\cdots \frac{\partial^2 L(\theta)}{\partial\theta^n \partial\theta^k} \right) + E\Big[\underbrace{s^k(\theta, Y) s(\theta, Y)'}_{\text{stacking it together}}\Big] \quad k = 1, \ldots d$$

Transpose

$$0 = \begin{pmatrix} \frac{\partial^2 L(\theta)}{\partial \theta^1 \partial \theta^1} & \cdots & \frac{\partial^2 L(\theta)}{\partial \theta^1 \partial \theta^d} \end{pmatrix} + E[s^1(\theta, Y)\, s(\theta, Y)']$$

$$\begin{pmatrix} \frac{\partial^2 L(\theta)}{\partial \theta^1 \partial \theta^2} & \cdots & \frac{\partial^2 L(\theta)}{\partial \theta^d \partial \theta^2} \end{pmatrix} + E[s^2(\theta, Y)\, s(\theta, Y)']$$

$$\vdots \qquad \vdots$$

$$\begin{pmatrix} \frac{\partial^2 L}{\partial \theta^1 \partial \theta^d} & & \frac{\partial^2 L(\theta)}{\partial \theta^d \partial \theta^d} \end{pmatrix} + E[s^d(\theta, Y)\, s(\theta, Y)']$$

Writing it in matrix form:

$$E \begin{pmatrix} s^1(\theta, Y)\, s(\theta, Y)' \\ s^2(\theta, Y)\, s(\theta, Y)' \\ \vdots \\ s^d(\theta, Y)\, s(\theta, Y)' \end{pmatrix} = E[s(\theta, Y)\, s(\theta, Y)'] = I_\theta$$

$$H(\theta) = \begin{pmatrix} \frac{\partial^2 L(\theta)}{\partial \theta^1 \partial \theta^1} & \cdots & \frac{\partial^2 L(\theta)}{\partial \theta^1 \partial \theta^d} \\ \vdots & & \\ \frac{\partial^2 L(\theta)}{\partial \theta^d \partial \theta^1} & \cdots & \frac{\partial^2 L(\theta)}{\partial \theta^d \partial \theta^d} \end{pmatrix}$$

look for cases where symmetry of derivatives don't exist.

Hessian

Combining two together;

$$H(\theta) + I_\theta = 0$$

$$\Rightarrow \underbrace{H(\theta) = -I_\theta}_{\text{Information Matrix Equality}}$$

[Singularity of Hessian $\Longrightarrow$ singularity of Information Matrix)

Nov 15, 2023

$$H(\theta) = -I_\theta.$$

singular $\Rightarrow$ Not invertible

Score: $\hat{I}_\theta = \frac{1}{n}\sum_{i=1}^{n} s(\theta, y_i)\, s(\theta, y_i)'$

By design, positive semidefinite matrix.

$$\sqrt{n}\,(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0,\ H^{-1}(\theta_0)\,\Sigma\, H(\theta_0)^{-1}\right)$$

covariance of gradient of the objective fn.

In case of MLE: $\Sigma = I_{\theta_0}$

$H(\theta_0) = -I_{\theta_0}$ $\rightarrow$ how?

from last class

$$H(\theta_0)^{-1}\,\Sigma\, H(\theta_0)^{-1} = I_\theta^{-1}$$

Some stuff on invertability here. If I is singular, how do we have this?

Theorem: suppose that
(Rao-Cramer's)

variance matrix

$$\lim_{n \to \infty} \operatorname{Var}\left(\sqrt{n}\,(\hat{\theta} - \theta_0)\right) = C, \quad 0 < C < +\infty$$

Neither collapsing nor diverging to ∞.

then $\operatorname{Var}\left(\sqrt{n}\,(\hat{\theta} - \theta_0)\right) - I_\theta^{-1}$ is positive semidefinite.

$$\Rightarrow \operatorname{Var}\left(\sqrt{n}\,(\hat{\theta} - \theta_0)\right) \geq I_\theta^{-1}$$

Confusion the & equal!!

Any regular estimator will have variance at most as small as MLE estimator.