www.mast.queensu.ca/∼blevit/

Background – 1. Cramer-Rao inequality

Observation vector, data:

$$\mathbf{X} = (X_1, ..., X_n) \quad \in \quad \mathcal{X} \quad \subset \quad \mathbb{R}^n$$

Here data can be real- or vector valued, components of the data can be independent or dependent, identically distributed or not; it can even be a random process. Denote the density of the data by

$$f(\mathbf{x}|\theta), \quad \mathbf{x} \in \mathcal{X}, \qquad \theta \in \Theta = (a, b) \subset \mathbb{R}^1$$

Assume that

$$f(\mathbf{x}|\theta) > 0, \quad \mathbf{x} \in \mathcal{X}, \quad \int_{\mathcal{X}} f(\mathbf{x}|\theta)\, d\mathbf{x} = 1.$$

$f(\mathbf{x}|\cdot)$ piecewise continuosly differentiable

Note on terminology: outside Statistics, a better known knotion of *Information* is due to C. Shannon (about 1950). However, The Fisher information introduced earlier (1910-20) is more useful in Statistics.

Score function:

$$l(\mathbf{x}|\theta) := \frac{\partial}{\partial\theta} \log f(\mathbf{x}|\theta) = \frac{\frac{\partial}{\partial\theta} f(\mathbf{x}|\theta)}{f(\mathbf{x}|\theta)}$$

Expected scores are zero:

$$\mathbf{E}_{\mathbf{X}|\theta} \left(\frac{\partial}{\partial\theta} \log f(\mathbf{X}|\theta) \right) = 0$$

Indeed,

$$\frac{d}{d\theta} \int_{\mathcal{X}} f(\mathbf{x}|\theta)\, d\mathbf{x} = \frac{d}{d\theta} 1 = 0$$

$$0 = \frac{d}{d\theta} \int_{\mathcal{X}} f(\mathbf{x}|\theta)\, d\mathbf{x} = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} f(\mathbf{x}|\theta)\, d\mathbf{x} = \int_{\mathcal{X}} l(\mathbf{x}|\theta) f(\mathbf{x}|\theta)\, d\mathbf{x} = \mathbf{E}_{\mathbf{X}|\theta} l(\mathbf{X}|\theta)$$

The Fisher information $I_{\mathbf{X}}(\theta)$, contained in the data $\mathbf{X}$ about the parameter θ, for a given θ:

$$I_{\mathbf{X}}(\theta) = \mathbf{Var}_{\mathbf{X}|\theta} \left(\frac{\partial}{\partial\theta} \log f(\mathbf{X}|\theta) \right) = \mathbf{E}_{\mathbf{X}|\theta} \left(\frac{\partial}{\partial\theta} \log f(\mathbf{X}|\theta) \right)^2 =$$

$$\int \left(\frac{\partial}{\partial\theta} \log f(\mathbf{x}|\theta) \right)^2 f(\mathbf{x}|\theta)\, d\mathbf{x} = \int \frac{\left(\frac{\partial}{\partial\theta} f(\mathbf{x}|\theta) \right)^2}{f(\mathbf{x}|\theta)}\, d\mathbf{x}.$$

The Cramer-Rao inequality: For any unbiased estimator $\bar{\psi} = \bar{\psi}(\mathbf{X})$ of $\psi(\theta)$, $\mathbf{E}_{\mathbf{X}|\theta} \left(\bar{\psi}(\mathbf{X}) \right) = \psi(\theta)$,

$$\mathbf{E}_{\mathbf{X}|\theta} (\bar{\theta} - \theta)^2 \geq \frac{\left(\psi'(\theta) \right)^2}{I_{\mathbf{X}}(\theta)}.$$

Example 1. $\mathbf{X} = (X_1, ..., X_n)$ *i.i.d.*, $X_1 \sim f(x|\theta)$. The Fisher Information contained in any X_i individually:

$$I(\theta) = \mathbf{Var}_{X_i|\theta}\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right).$$

Then

$$I_{\mathbf{X}}(\theta) = \mathbf{Var}_{\mathbf{X}|\theta}\left(\frac{\partial}{\partial\theta}\log f(X_1|\theta)\cdots f_(X_n|\theta)\right) = \mathbf{Var}_{\mathbf{X}|\theta}\left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right) =$$

$$\sum_{i=1}^{n}\mathbf{Var}_{X_i|\theta}\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right) = nI(\theta).$$

Example 2. $\mathbf{X} = (X_1, ..., X_n) \quad i.i.d., \quad X_1 \quad \sim \quad \mathcal{N}(\theta, \sigma^2)$

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}$$

$$\log f(x_i|\theta) = \log\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_i-\theta)^2}{2\sigma^2}, \qquad \frac{\partial}{\partial\theta}\log f(X_i|\theta) = \frac{X_i-\theta}{\sigma^2}$$

$$I(\theta) = \mathbf{Var}_{X_i|\theta}\left(\frac{X_i-\theta}{\sigma^2}\right) = \frac{1}{\sigma^2}$$

$$I_{\mathbf{X}}(\theta) = \frac{n}{\sigma^2}$$

Thus for any unbiased estimator $\tilde{\theta}$ of θ,

$$\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta}-\theta)^2 \geq \frac{1}{I_{\mathbf{X}}(\theta)} = \frac{\sigma^2}{n}.$$

The MLE is the best unbiased estimator:

$$\hat{\theta} = \frac{X_1 + \cdots + X_n}{n}, \qquad \mathbf{E}_{\mathbf{X}|\theta}\hat{\theta} = \theta$$

$$\mathbf{E}_{\mathbf{X}|\theta}(\hat{\theta}-\theta)^2 = \frac{\sigma^2}{n} \equiv \frac{1}{I_{\mathbf{X}}(\theta)}.$$

An unbiased estimator treats all $\theta \in \mathbb{R}$ fily, or equally. But suppose next that $\theta \in \Theta = (a, b) \neq \mathbb{R}$. Then an estimator $\tilde{\theta}$ does not need to be unbiased. We can consider the **maximal risk:**

$$r(\tilde{\theta}, \Theta) = \sup_{a<\theta<b}\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta}-\theta)^2$$

and the minimax risk:

$$r(\Theta) = \inf_{\tilde{\theta}}\sup_{a<\theta<b}\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta}-\theta)^2$$

Minimax estimators are only known in case $b - a \leq 0.1$.

Elements of the Decision Theory

Let $\lambda(\theta) > 0, \quad \theta \in (a, b)$, be a *weight* function:

$$\int_a^b \lambda(\theta)\, d\theta = 1.$$

Then the *average, or Bayes, risk of an etimator* $\tilde{\theta}$ is

$$R(\tilde{\theta},\lambda) := \int_a^b \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta}-\theta)^2\lambda(\theta)\,d\theta \le \sup_{a<\theta<b}\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta}-\theta)^2\int_a^b\lambda(\theta)\,d\theta = \sup_{a<\theta<b}\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta}-\theta)^2 = r(\tilde{\theta},\Theta),$$

i.e. the maximal risk does not exceed the Bayes risk of an estimator. Also

$$r(\tilde{\theta},\Theta) \ge \inf_{\tilde{\theta}} R(\tilde{\theta},\lambda) = \mathcal{R}(\lambda).$$

and therefore,

$$r(\Theta) = \inf_{\tilde{\theta}} r(\tilde{\theta},\Theta) \ge \mathcal{R}(\lambda).$$

Thus, **the minimax risk is at least as large as the Bayes risk. The Wald principle: typically**

$$r(\Theta) = \sup_{\lambda} R(\lambda).$$

In the Bayesian setting:
$\lambda(\theta)$ is viewed as the *prior density* of θ,
$f_{\mathbf{X},\theta}(\mathbf{x},\theta) = f(\mathbf{x}|\theta)\lambda(\theta)$ is viewed as the *joint density* of $\mathbf{X}$ and θ,

$$R(\tilde{\theta},\lambda) = \mathbf{E}_\theta\left(\mathbf{E}_{\mathbf{X}|\boldsymbol{\theta}}(\tilde{\theta}(\mathbf{X})-\theta)^2\right) = \mathbf{E}_{\mathbf{X},\boldsymbol{\theta}}(\tilde{\theta}(\mathbf{X})-\theta)^2,$$

according to the *Law of Total Expectation.* Since here we are dealing simultaneously with the conditional and average (Bayes) risk, and since above we have defined the conditional Fisher information, given θ, it suggests itself naturally to define the **Bayes, or average, Fisher information contained in X, about parameter** θ:

$$I_{\mathbf{X}} = \mathbf{E}_{\mathbf{X},\theta}\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2.$$

By the Law of Total Expectation,

$$I_{\mathbf{X}} = \mathbf{E}_\theta\left\{\mathbf{E}_{\mathbf{X}|\theta}\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right\} = \mathbf{E}_\theta\left(I_{\mathbf{X}}(\theta)\right) = \int_a^b I_{\mathbf{X}}(\theta)\lambda(\theta)\,d\theta.$$

Knowing the prior distribution of θ also provides some information about θ. Again the expected scores are zero:

$$\mathbf{E}_\theta\left(\frac{d}{d\theta}\log\lambda(\theta)\right) = 0.$$

We can measure this information by the **Fisher information corresponding to the prior density:**

$$I_\lambda = \mathbf{E}_\theta\left(\frac{d}{d\theta}\log\lambda(\theta)\right)^2.$$

Now we can define the **Total Fisher information** jointly contained in the data, $\mathbf{X}$, and in the prior distribution of θ:

$$I = \mathbf{E}_{\mathbf{X}\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta}\log\left(f(\mathbf{X}|\boldsymbol{\theta})\lambda(\boldsymbol{\theta})\right)\right)^2.$$

Fisher information is additive :

$$I = I_{\mathbf{X}} + I_\lambda = \int I_{\mathbf{X}}(\theta)\lambda(\theta)\,dt + I_\lambda.$$

Indeed,

$$I = \mathbf{E}_{\mathbf{X}\theta}\left(\frac{\partial}{\partial\theta}\log\left(f(\mathbf{X}|\theta)\lambda(\theta)\right)\right)^2 = \mathbf{E}_{\mathbf{X}\theta}\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X}|\theta) + \frac{\partial}{\partial\theta}\log\lambda(\theta)\right)^2 =$$

$$\mathbf{E}_{\boldsymbol{\theta}\mathbf{X}}\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X}|\theta)\right)^2 + \mathbf{E}_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta}\log\lambda(\theta)\right)^2 + 2\mathbf{E}_{\mathbf{X}\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\frac{\partial}{\partial\theta}\log\lambda(\theta)\right) =$$

$$I_{\mathbf{X}} + I_\lambda + 2\mathbf{E}_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta}\log\lambda(\theta)\mathbf{E}_{\mathbf{X}|\theta}\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = I_{\mathbf{X}} + I_\lambda.$$

Corollary.

$$\mathbf{X} = (X_1, ..., X_n) \quad \sim \quad i.i.d. \quad f(x|\theta)$$

Consider the amount of information contained in every single observation individually:

$$I(\theta) = \mathbf{E}_{X_i|\theta}\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2.$$

Then the total Fisher information

$$I = I_{\mathbf{X}} + I_\lambda = n\int I(\theta)\lambda(\theta)\,d\theta + I_\lambda.$$

In particular, in the case

$$\mathbf{X} = (X_1, ..., X_n) \quad \sim \quad i.i.d. \quad \mathcal{N}(\theta, \sigma^2),$$

we have

$$I = \frac{n}{\sigma^2} + I_\lambda.$$

van Trees inequality: a Bayesian Cramer-Rao lower bound; see [2], [1]. Assume that $\lambda(\theta)$ is piece-wise continuously differentiable and $\lambda(a) = \lambda(b) = 0$. Then for **any** estimator $\tilde{\psi} = \tilde{\psi}(\mathbf{X})$ of $\psi(\theta)$ (cf. the Cramer-Rao lower bound)

$$R(\tilde{\psi}, \lambda) \geq \frac{\left(\mathbf{E}_\theta\psi'(\theta)\right)^2}{I}.$$

Corollary. Under the above assumptions,

$$\sup_{a<\theta<b}\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\psi} - \psi(\theta))^2 \geq \frac{\left(\mathbf{E}_\theta\psi'(\theta)\right)^2}{I}.$$

Proof. Let us take a look at the following integral

$$J := \int\int(\tilde{\psi}(\mathbf{x}) - \psi(\theta))\frac{\partial}{\partial\theta}\left(f(\mathbf{x}|\theta)\lambda(\theta)\right)\,d\theta d\mathbf{x} =$$

$$\int\left((\tilde{\psi}(\mathbf{x}) - \psi(\theta))f(\mathbf{x}|\theta)\lambda(\theta)\Big|_a^b - \int\frac{\partial}{\partial\theta}(\tilde{\psi}(\mathbf{x}) - \psi(\theta))f(\mathbf{x}|\theta)\lambda(\theta)\,d\theta\right)d\mathbf{x} =$$

$$\int\left(\int\psi'(\theta)f(\mathbf{x}|\theta)\pi(\theta)\,d\theta\right)d\mathbf{x} = \int\psi'(\theta)\lambda(\theta)\,d\theta = \mathbf{E}_\theta\psi'(\theta).$$

On the other hand,

$$\mathbf{E}_\theta \psi'(\theta) = J = \int\int (\tilde{\psi}(\mathbf{x}) - \psi(\theta)) \frac{\frac{\partial}{\partial\theta}\left(f(\mathbf{x}|\theta)\lambda(\theta)\right)}{f(\mathbf{x}|\theta)\lambda(\theta)} f(\mathbf{x}|\theta)\lambda(\theta)\, d\theta d\mathbf{x} =$$

$$\mathbf{E}_{\mathbf{X}\boldsymbol{\theta}}\left[(\tilde{\psi}(\mathbf{X}) - \psi(\boldsymbol{\theta}))\left(\frac{\partial}{\partial\theta}\log\left(f(\mathbf{X}|\boldsymbol{\theta})\lambda(\boldsymbol{\theta})\right)\right)\right].$$

By the Cauchy-Schwarz inequality:

$$\left(\mathbf{E}_\theta \psi'(\theta)\right)^2 \le \mathbf{E}_{\mathbf{X}\boldsymbol{\theta}}(\tilde{\psi}(\mathbf{X}) - \psi(\boldsymbol{\theta}))^2 \mathbf{E}_{\mathbf{X}\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta}\log\left(f(\mathbf{X}|\boldsymbol{\theta})\lambda(\boldsymbol{\theta})\right)\right)^2 = I\mathbf{E}_{\boldsymbol{\theta}\mathbf{X}}(\tilde{\psi}(\mathbf{X}) - \psi(\boldsymbol{\theta}))^2 = IR(\tilde{\psi}, \lambda).$$

□

Corollary 1.

$$\mathbf{X} = (X_1, ..., X_n) \quad \sim \quad i.i.d. \quad f(x|\theta), \quad \theta \in (a, b)$$

Then, for any piece-wise continuously differentiable prior density $\lambda(\theta) > 0,\ \theta \in (a, b),\ \lambda(a) = \lambda(b) = 0$, and any estimator $\tilde{\psi}$ of $\psi(\theta)$,

$$\max_{a<\theta<b} \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\psi} - \psi(\theta))^2 \ge \frac{\left(\int_a^b \psi'(\theta)\lambda(\theta)\, d\theta\right)^2}{n\int_a^b I(\theta)\lambda(\theta)\, d\theta + I_\lambda}.$$

In particular, if

$$\mathbf{X} = (X_1, ..., X_n) \quad \sim \quad i.i.d. \quad \mathcal{N}(\theta, \sigma^2), \quad \theta \in (a, b)$$

Then

$$\max_{a<\theta<b} \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \ge \frac{1}{\frac{n}{\sigma^2} + I_\lambda}.$$

Note. Of the two terms appearing in the denominator of the van Trees inequality, the first one is essentially determined by the Fisher information $I(\theta)$, whereas the prior knowledge that $\theta \in (a, b)$ is mainly captured by the second term in the denominator. The following makes this dependence more explicit.

Suppose $(a, b) = (\vartheta - \delta, \vartheta + \delta)$ where $\vartheta = (a + b)/2$ and $\delta = (b - a)/2$. Let **(rescaling)**

$$\lambda(\theta) = \frac{1}{\delta}\lambda_0\left(\frac{\theta - \vartheta}{\delta}\right)$$

where $\lambda_0(\theta)$ is a density on $(-1, 1)$, with the corresponding Fisher information

$$I_{\lambda_0} = \int_{-1}^{1} \frac{(\lambda_0'(\theta))^2}{\lambda_0(\theta)}\, d\theta < \infty.$$

Then

$$I_\lambda = \int_{\vartheta-\delta}^{\vartheta+\delta} \frac{(\lambda'(\theta))^2}{\lambda(\theta)}\, d\theta = \int_{\vartheta-\delta}^{\vartheta+\delta} \frac{\left(\frac{1}{\delta^2}\lambda_0'\left(\frac{\theta-\vartheta}{\delta}\right)\right)^2}{\frac{1}{\delta}\lambda_0\left(\frac{\theta-\vartheta}{\delta}\right)}\, d\theta = \frac{1}{\delta^2}\int_{-1}^{1} \frac{(\lambda_0'(\theta))^2}{\lambda_0(\theta)}\, d\theta = \delta^{-2} I_{\lambda_0}.$$

Corollary 2.

$$\mathbf{X} = (X_1, ..., X_n) \quad \sim \quad i.i.d. \quad f(x|\theta), \quad \theta \in (a, b)$$

Then, for any piece-wise continuously differentiable prior density $\lambda_0(\theta) > 0,\ \theta \in (-1, 1),\ \lambda_0(\pm 1) = 0$, and any estimator $\tilde{\psi}$ of $\psi(\theta)$,

$$\max_{a<\theta<b} \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\psi} - \psi(\theta))^2 \ge \frac{\left(\int_a^b \psi'(\theta)\lambda(\theta)\, d\theta\right)^2}{n\int_a^b I(\theta)\lambda(\theta)\, d\theta + \delta^{-2} I_{\lambda_0}}.$$

In particular, if

$$\mathbf{X} = (X_1, ..., X_n) \quad \sim \quad i.i.d. \quad \mathcal{N}(\theta, \sigma^2), \quad \theta \in (a, b)$$

Then

$$\max_{a<\theta<b} \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \geq \frac{1}{\frac{n}{\sigma^2} + \frac{I_{\lambda_0}}{\delta^2}}.$$

One can further improve the above lower bound by minimizing the Fisher information I_{λ_0}. Thus consider the following minimization problem:

$$I_{\lambda_0} \to \min, \quad \int_{-1}^{1} \lambda_0(\theta)\, d\theta = 1.$$

Denote

$$\lambda_0(\theta) = \omega^2(\theta), \quad \omega(\theta) > 0, \quad \theta \in (-1, 1), \quad \omega(\pm 1) = 0.$$

Then

$$I_{\lambda_0} = \int_{-1}^{1} \frac{(\lambda_0'(\theta))^2}{\lambda_0(\theta)}\, d\theta = \int_{-1}^{1} \frac{(2\omega'(\theta)\omega(\theta))^2}{\omega^2(\theta)}\, d\theta = 4\int_{-1}^{1} (\omega'(\theta))^2\, d\theta \quad \to \quad \min$$

This is a constrained minimization problem:

$$\int_{-1}^{1} (\omega'(\theta))^2\, d\theta \quad \to \quad \min, \qquad \int_{-1}^{1} \omega^2(\theta)\, d\theta = 1.$$

Euler-Lagrange equation:

$$\omega''(\theta) + c\omega(\theta) = 0, \quad \omega(\pm 1) = 0$$

where c is the Lagrange multiplier. The only positive solution of this equation is

$$\omega(\theta) = \cos\left(\frac{\pi\theta}{2}\right)$$

Thus the density λ_0 minimizing I_{λ_0} **(the least favorable density)** is

$$\lambda_0(\theta) = \cos^2\left(\frac{\pi\theta}{2}\right), \quad |\theta| \leq 1$$

and

$$I_{\lambda_0} = \pi^2.$$

Corollary 3.

$$\mathbf{X} = (X_1, ..., X_n) \quad \sim \quad i.i.d. \quad f(x|\theta), \quad \theta \in (a, b)$$

Then, assuming $\lambda(\theta)$ is the least favorable density on $(a, b) = (\vartheta - \delta, \vartheta + \delta)$, for any estimator $\tilde{\psi}$ of $\psi(\theta)$,

$$\max_{a<\theta<b} \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\psi} - \psi(\theta))^2 \geq \frac{\left(\int_a^b \psi'(\theta)\lambda(\theta)\, d\theta\right)^2}{n\int_a^b I(\theta)\lambda(\theta)\, d\theta + \pi^2\delta^{-2}}.$$

In particular, if

$$\mathbf{X} = (X_1, ..., X_n) \quad \sim \quad i.i.d. \quad \mathcal{N}(\theta, \sigma^2), \quad \theta \in (a, b)$$

Then for any estimator $\tilde{\theta}$ of θ

$$\max_{a<\theta<b} \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \geq \frac{1}{\frac{n}{\sigma^2} + \frac{\pi^2}{\delta^2}}.$$

Applications 1. Asymptotic lower bound. Let

$$\mathbf{X} = (X_1, ..., X_n) \quad \sim \quad i.i.d. \quad f(x|\theta)$$

Let $I(\theta)$ be the Fisher information contained in any X_i individually. Then, for any prior density λ on $[a, b]$, and any estimator $\tilde{\theta}$,

$$\sup_{\theta\in(a,b)} n\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \geq \frac{n}{n\int_a^b I(\theta)\lambda(\theta)\,d\theta + I_\lambda} = \frac{1}{\int_a^b I(\theta)\lambda(\theta)\,d\theta + n^{-1}I_\lambda}.$$

Thus

$$\lim_{n\to\infty} \sup_{\theta\in(a,b)} n\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \geq \frac{1}{\int_a^b I(\theta)\lambda(\theta)\,d\theta}.$$

Also, for any $(\vartheta - \delta, \vartheta + \delta) \subset (a, b)$

$$\lim_{n\to\infty} \sup_{\theta\in(a,b)} n\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \geq \lim_{n\to\infty} \sup_{\theta\in(\vartheta-\delta,\vartheta+\delta)} n\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \geq \frac{1}{\int_{\vartheta-\delta}^{\vartheta+\delta} I(\theta)\lambda(\theta)\,d\theta}.$$

Therefore, assuming that $I(\theta)$ is a continuous function, and letting $\delta \to 0$, we get, for any $a < \vartheta < b$,

$$\lim_{n\to\infty} \sup_{\theta\in[a,b]} n\mathbf{E}(\tilde{\theta} - \theta)^2 \geq \lim_{\delta\to 0} \frac{1}{\sup_{\theta\in[\vartheta-\delta,\vartheta+\delta]} I(\theta)} = \frac{1}{I(\vartheta)}.$$

Finally, for any (a, b), and any sequence of estimators $\tilde{\theta}_n$,

$$\lim_{n\to\infty} \sup_{\theta\in(a,b)} n\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \geq \sup_{\theta\in(a,b)} \frac{1}{I(\theta)}.$$

In particular, if $X_1, ..., X_n \quad \sim \quad i.i.d. \quad \mathcal{N}(\theta, \sigma^2)$, then for any any sequence of estimators $\tilde{\theta}_n$ and any $a < b$,

$$\lim_{n\to\infty} \sup_{\theta\in(a,b)} n\mathbf{E}(\tilde{\theta}_n - \theta)^2 \geq \sigma^2.$$

For any (a, b), an equality is achieved in this inequality, for the MLE $\hat{\theta}_n$. Hence $\hat{\theta}_n$ is **locally asymptotically minimax**, among all possible estimators (not only unbiased). Thus, we still can write

$$\sigma^2(1 + o(1)) \leq n\mathbf{E}(\tilde{\theta}_n - \theta)^2 \leq \sigma^2$$

if we agree to interprete this in the locally asymptotically minimax (LAM) sense.

This is the best possible and complete characterization of the asymptotic optimality, among all possible estimators. Note, that a valid point-wise lower bound, in the class of all estimators, is impossible because for any θ there exists an estimator which, for this particular θ, has a zero mean squared error. However, this result is, in a certain sense trivial, because it does not reflect the influence of the prior knowledge that $\theta \in (a, b)$ on the optimal estimator. This is because such prior information effects only the second order optimal properties.

Application 2. Second order minimax estimation. Let

$$X_1, ..., X_n \quad \sim \quad \mathcal{N}(\theta, \sigma^2), \quad I(\theta) = \frac{1}{\sigma^2} \quad \theta \in (-\delta, \delta)$$

Then for any estimator $\tilde{\theta}$

$$\sup_{\theta\in(-\delta,\delta)} n\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta}_n - \theta)^2 \geq \frac{n}{n\sigma^{-2} + \pi^2\delta^{-2}} = \frac{\sigma^2}{1 + \frac{\pi^2\sigma^2}{\delta^2 n}} \geq \sigma^2\left(1 - \frac{\pi^2\sigma^2}{n\delta^2}\right).$$

This lower bound is asymptotically exact (achievable) to the order n^{-1}. An estimator achieving this bound is second order asymptotically minimax. Note that letting $\delta \to \infty$ we get

$$\sup_{\theta\in(-\infty,\infty)} n\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta}_n - \theta)^2 \geq \sigma^2.$$

Thus $\hat{\theta}_n = \bar{X}_n$ is minimax in $\mathbb{R}$, for any n. This raises next question whether itis also admissible.

Application 3. Admissibility. *Definition.* $\tilde{\theta}$ is admissible if there is no other estimator $\bar{\theta}$ such that

$$\mathbf{E}_{\mathbf{X}|\theta}(\bar{\theta} - \theta)^2 \leq \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2, \quad \text{for all} \quad \theta \in \Theta,$$

$$\mathbf{E}_{\mathbf{X}|\theta}(\bar{\theta} - \theta)^2 < \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2, \quad \text{for some} \quad \theta \in \Theta.$$

For any $n \geq 1$ the MLE $\hat{\theta}_n = \bar{X}_n$ is admissible. Let us proof it in the case $\sigma^2 = 1,\ n = 1$. The general case can be either treated quite similarly, or directly reduced to it by sufficiency.

$$\mathbf{E}_{\mathbf{X}|\theta}(\hat{\theta} - \theta)^2 = 1$$

Let $\lambda(\theta)$ be a prior density on $(-a, a)$. According to the van Trees inequality, for any estimator $\tilde{\theta}$,

$$\sup_{|\theta|<a} \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \geq \frac{1}{1 + I_\lambda}. \tag{2}$$

Suppose there is an estimator $\tilde{\theta}$ such that for all θ

$$\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \leq \mathbf{E}_{\mathbf{X}|\theta}(\hat{\theta} - \theta)^2 = 1,$$

whereas for some θ_0

$$\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta_0)^2 < \mathbf{E}_{\mathbf{X}|\theta}(\hat{\theta} - \theta)^2 = 1.$$

Since $\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2$ is a continuous function of θ (why?), there exist an $\varepsilon > 0$ such that

$$\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \leq 1 - \varepsilon, \quad \theta \in (\theta_0 - \varepsilon, \theta + \varepsilon),$$

or equivalently,

$$\mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta} - \theta)^2 \leq 1 - \varepsilon \mathbf{1}_{(\theta_0-\varepsilon,\theta_0+\varepsilon)}(\theta), \quad \theta \in \mathbf{R}^1. \tag{3}$$

Consider the following even non-negative continuous function

$$g(\theta) = \begin{cases} 1 & |\theta| \leq a \\ \frac{(|\theta|-a-b)^2}{b^2} & a \leq |\theta| \leq a+b \\ 0 & |\theta| > a+b \end{cases}$$

where a is large enough so that

$$(\theta_0 - \varepsilon, \theta_0 + \varepsilon) \subset (-a, a).$$

Since

$$\int g(\theta)\, d\theta = 2\int_0^\infty g(\theta)\, d\theta =$$

$$2\int_0^a d\theta + 2\int_a^b \frac{(\theta - a - b)^2}{b^2}\, d\theta = 2a + 2\,\frac{(\theta - a - b)^3}{3b^2}\bigg|_a^{a+b} = 2a - 2\frac{(-b)^3}{3b^2} = 2\left(a + \frac{b}{3}\right),$$

we define a prior density as:

$$\lambda(\theta) = \frac{g(\theta)}{2(a+\frac{b}{3})} = \begin{cases} 1 & |\mu| \le a \\ \frac{(|\theta|-a-b)^2}{b^2} & a \le |\mu| \le a+b \\ 0 & |\mu| > a+b \end{cases}$$

We can evaluate directly

$$I_\lambda = \int \frac{(\lambda')^2}{\lambda}\, d\theta = 2\int_0^{a+b} \frac{(\lambda')^2}{\lambda}\, d\theta = 2\int_a^{a+b} \frac{(\lambda')^2}{\lambda}\, d\theta$$

$$\lambda(\theta) = \frac{(\theta-a-b)^2}{2(a+\frac{b}{3})b^2}, \quad \lambda'(\theta) = \frac{\lambda-a-b}{(a+\frac{b}{3})b^2}$$

Since for $a \le \theta \le a+b$,

$$\frac{(\lambda'(\theta))^2}{\lambda(\theta)} = \frac{\frac{(\theta-a-b)^2}{(a+\frac{b}{3})^2 b^4}}{\frac{(\theta-a-b)^2}{2(a+\frac{b}{3})b^2}} = \frac{2}{(a+\frac{b}{3})b^2}$$

we find

$$I_\lambda = \frac{4}{(a+\frac{b}{3})b}$$

Now by (3)

$$\int \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta}-\theta)^2\lambda(\theta)\, d\theta \le 1-\varepsilon\int \mathbf{1}[\theta_0-\varepsilon,\theta_0+\varepsilon]\lambda(\mu)\, d\mu = 1-\varepsilon\int_{\theta_0-\varepsilon}^{\theta_0+\varepsilon} \frac{d\mu}{2\left(a+\frac{b}{3}\right)} = 1-\frac{\varepsilon^2}{(a+\frac{b}{3})}$$

On the other hand from (2)

$$\int \mathbf{E}_{\mathbf{X}|\theta}(\tilde{\theta}-\theta)^2\lambda(\theta)\, d\theta \ge \frac{1}{1+I_\lambda} \ge 1-I_\lambda = 1-\frac{4}{(a+\frac{b}{3})b}$$

These two inequalitites are compatible only if

$$\frac{\varepsilon^2}{(a+\frac{b}{3})} \le \frac{4}{(a+\frac{b}{3})b} \quad \Longrightarrow \quad \varepsilon^2 \le \frac{4}{b}.$$

Choosing b large enough, we get a contradiction!

REFERENCES

[1] R. Gill and B. Levit, Math. Meth. Statist, v. 1(1996)

[2] van Trees, Detection, Estimation and Modulation Theory (1968), vol. I.