Statistical Signal Processing

Lecture 4

chapter 1: parameter estimation
deterministic parameters

- some optimality properties
- Maximum Likelihood estimation
- Fischer Information Matrix
- Cramer-Rao lower bound on the MSE

Deterministic Parameter Estimation

Two points of view:

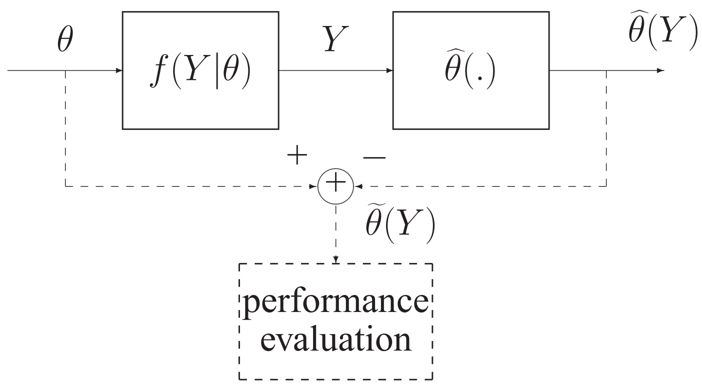
- the parameters θ are unknown deterministic quantities
- the parameters θ are stochastic, but their prior distribution $f(\theta)$ is unknown

The only stochastic description available is the conditional density $f(Y|\theta)$ describing the stochastic relation between the unknown parameters θ and the observed measurements Y.

- since θ is not necessarily a random vector but just a set of parameters on which the distribution of Y depends, we often find the notations

  $$f(Y|\theta) \;=\; f(Y;\theta) \;=\; f_\theta(Y)$$

  but we shall continue to use $f(Y|\theta)$
- expectation now means $E \;=\; E_{Y|\theta}$



Deterministic Parameter Estimation (2)

- an estimator $\widehat{\theta}(Y)$ of θ is again a function of Y (a statistic), with estimation error $\widetilde{\theta} = \theta - \widehat{\theta}(Y)$

- to evaluate the quality of an estimator, we shall again introduce the *risk* function MSE as the average value of the SE *cost* function

$$\text{MSE} = \mathcal{R}(\widehat{\theta}(.)|\theta) \;=\; E_{Y|\theta}\, \|\widetilde{\theta}\|^2 \;=\; \int f(Y|\theta)\, \|\theta - \widehat{\theta}(Y)\|^2 dY$$

  the MSE is a function of θ in general

- minimization of the risk function leads to $\widehat{\theta} = \theta$ (and $\mathcal{R} = 0$): not an acceptable strategy since the resulting $\widehat{\theta}$ depends on the unknown θ

- ideally, would like $\widehat{\theta}(.)$ such that $\mathcal{R}(\widehat{\theta}(.)|\theta)$ is minimized $\forall\theta \in \Theta$: impossible! Consider $\widehat{\theta}(Y) = \theta_0 \in \Theta$: ignores the data Y but $\mathcal{R}(\widehat{\theta}(.)|\theta_0) = 0$

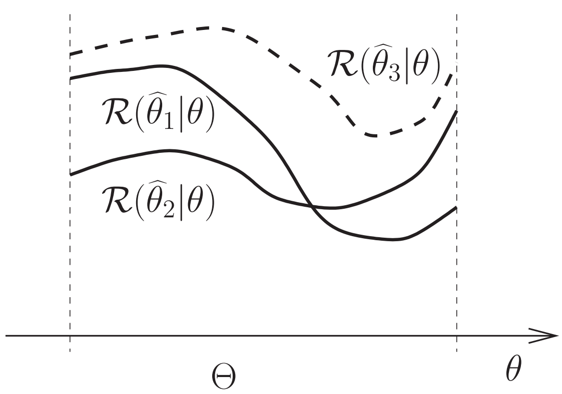
- we shall still evaluate the performance via the MSE, but in the deterministic case, we shall not be able to derive estimators by minimizing the MSE.

Deterministic Parameter Estimation (3)

- given two estimators $\widehat{\theta}_1(Y)$ and $\widehat{\theta}_2(Y)$, one is usually not uniformly better than the other one (see figure)

- a uniformly minimum risk estimator does not exist in general

- consider some other desirable properties



Some Optimality Properties

- estimator *bias* : average deviation from the true parameter

$$b_{\hat\theta}(\theta) \;=\; -E_{Y|\theta}\tilde\theta \;=\; E_{Y|\theta}\left(\hat\theta(Y)-\theta\right) \;=\; E_{Y|\theta}\hat\theta(Y) \;-\; \theta$$

  unbiased estimator: $b_{\hat\theta}(\theta) = 0,\ \forall\theta\in\Theta$
  Unbiasedness is a weak property: estimator can be correct on the average, but with large deviations. Good estimators exist that are biased.

- Example: mean of Gaussian i.i.d. variables

$$\text{i.i.d.}\quad y_i \sim \mathcal{N}(\theta,1)\ ,\quad i=1,\ldots,n$$

  Consider $\hat\theta(Y) = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the sample mean.

$$E_{Y|\theta}\hat\theta = E_{Y|\theta}\overline{y} = E_{Y|\theta}\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} E_{Y|\theta}y_i = \frac{1}{n}\sum_{i=1}^{n}\theta = \frac{n\,\theta}{n} = \theta \text{ : unbiased!}$$

- $\hat\theta(.)$ is *inadmissible* if another estimator $\hat\theta'(.)$ has uniformly lower risk:

$$\forall\theta\in\Theta : \mathcal{R}(\hat\theta'|\theta) \le \mathcal{R}(\hat\theta|\theta)\ ,\quad \exists\theta_0\in\Theta : \mathcal{R}(\hat\theta'|\theta_0) < \mathcal{R}(\hat\theta|\theta_0)$$

  $\hat\theta$ is *admissible* if no such $\hat\theta'$ exists. Example: $\hat\theta_3$ in figure above.

Some Optimality Properties (2)

- MSE = $E_{Y|\theta}\|\tilde\theta\|^2 = E_{Y|\theta}\,\tilde\theta^T\tilde\theta = \text{tr}\left\{E_{Y|\theta}\,\tilde\theta\tilde\theta^T\right\} = \text{tr}\left\{R_{\tilde\theta\tilde\theta}\right\}$,
  $R_{\tilde\theta\tilde\theta} = E_{Y|\theta}\,\tilde\theta\,\tilde\theta^T$ = estimation error correlation matrix

$$R_{\tilde\theta\tilde\theta} \;=\; E_{Y|\theta}(\hat\theta-\theta)(\hat\theta-\theta)^T \;=\; E_{Y|\theta}[\underline{\hat\theta\ (-E_{Y|\theta}\hat\theta} + \underline{E_{Y|\theta}\hat\theta)-\theta}][\underline{\hat\theta\ (-E_{Y|\theta}\hat\theta} + \underline{E_{Y|\theta}\hat\theta)-\theta}]^T$$

$$= \; E_{Y|\theta}(\hat\theta-E_{Y|\theta}\hat\theta)(\hat\theta-E_{Y|\theta}\hat\theta)^T + (E_{Y|\theta}\hat\theta-\theta)(E_{Y|\theta}\hat\theta-\theta)^T$$

$$= \; C_{\hat\theta\hat\theta} + b_{\hat\theta}(\theta)b^T_{\hat\theta}(\theta) \;=\; C_{\tilde\theta\tilde\theta} + (E_{Y|\theta}\,\tilde\theta)\,(E_{Y|\theta}\,\tilde\theta)^T$$

where we used: $\quad C_{\hat\theta\hat\theta} \;=\; C_{\tilde\theta\tilde\theta}$

Some Optimality Properties (3)

- $\widehat{\theta}(Y)$ is said to be *minimax* if it satisfies

$$\sup_{\theta\in\Theta} \mathcal{R}(\widehat{\theta}|\theta) \;=\; \inf_{\widehat{\theta}'} \sup_{\theta\in\Theta} \mathcal{R}(\widehat{\theta}'|\theta)$$

  (inf $\approx$ min, sup $\approx$ max).
  A minimax estimator minimizes the maximum risk over Θ.
  A minimax $\widehat{\theta}$ is difficult to obtain in general.

Uniformly minimum risk estimators may be found if we restrict the class of estimators.

- $\widehat{\theta}$ is a *uniformly minimum variance unbiased estimator* (UMVUE) if it is unbiased and if for any other unbiased estimator $\widehat{\theta}'$: $\quad R_{\tilde{\theta}\tilde{\theta}} \;\leq\; R_{\tilde{\theta}'\tilde{\theta}'}\,,\;\; \forall\theta \in \Theta$, or

$$E_{Y|\theta}(\widehat{\theta}(Y)-\theta)(\widehat{\theta}(Y)-\theta)^T \;\leq\; E_{Y|\theta}(\widehat{\theta}'(Y)-\theta)(\widehat{\theta}'(Y)-\theta)^T$$

  note: variance = tr {covariance matrix}, $\;\text{MSE}_{\widehat{\theta}} \;=\; \text{tr}\,\{R_{\tilde{\theta}\tilde{\theta}}\}$

- UMVUE are highly desirable but they may not exist or be difficult to compute. They can be computed if a *complete sufficient statistic* can be found.

Maximum Likelihood Estimation

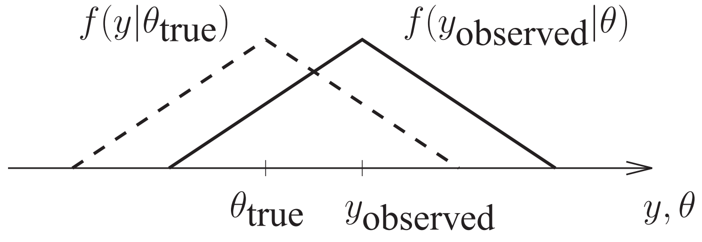
- the maximum likelihood (ML) estimation philosophy is to choose that value of the parameters that renders the observations most likely:

$$\widehat{\theta}_{ML}(Y) \;=\; \arg\max_{\theta\in\Theta}\; f(Y|\theta)$$

example:

- $y = \theta + v$, $f_{\mathrm{v}}(v) = \begin{cases} 1-|v| & , \; |v| \leq 1 \\ 0 & , \; |v| > 1 \end{cases} \qquad f(y|\theta) = f_{\mathrm{v}}(y-\theta)$

$$\widehat{\theta}_{ML}(y) \;=\; y$$



ML Estimation: Remarks

- $f(Y|\theta)$ is called the *likelihood function*. In order to emphasize the dependence on θ and the fact that the observation Y is fixed, it is often denoted as

$$l(\theta;Y) = f(Y|\theta) \qquad\qquad L(\theta;Y) = \ln f(Y|\theta)$$

- since the logarithmic function is strictly monotone, the maximum point of $f(Y|\theta)$ corresponds with the maximum point of $\ln f(Y|\theta)$, called the *log likelihood function*

- Often $f(Y|\theta)$ satisfies certain regularity conditions so that $\widehat{\theta}_{ML}$ is a solution of

$$\frac{\partial}{\partial\theta}\ln f(Y|\theta) = 0 .$$

  The conditions for a maximum (rather than another form of extremum) need to be verified of course.

- The ML estimator is given by the *global* maximum of $f(Y|\theta)$. If there are several local maxima, all of them need to be examined and compared to find the global maximum.

ML Estimation: Remarks (2)

- Even if $f(Y|\theta)$ satisfies regularity conditions, the maximum may occur at the boundary of the parameter space Θ (which may not necessarily be $(-\infty,\infty)$ for every θ_i). In that case, the maximum is not a local extremum.

- The ML estimator can be seen as a limiting case of the MAP estimator when the prior distribution $f(\theta)$ becomes uninformative (uniform distribution). For those components θ_i of θ for which the support is unbounded, this means that $\sigma^2_{\theta_i} \to \infty$ (information $\to 0$). Indeed

$$\begin{aligned}\widehat{\theta}_{MAP}(Y) &= \arg\max_{\theta\in\Theta} f(\theta|Y) = \arg\max_{\theta\in\Theta} \frac{f(Y|\theta)f(\theta)}{f(Y)} \\ &= \arg\max_{\theta\in\Theta} f(Y|\theta)f(\theta) \overset{f(\theta)=c^t}{=} \arg\max_{\theta\in\Theta} f(Y|\theta) = \widehat{\theta}_{ML}(Y)\end{aligned}$$

  But in the deterministic case, θ is fixed, whereas in the Bayesian case θ is random, hence e.g. the MSE is different for both formulations
  (MSE$_{MAP} = \int_\Theta$ MSE$_{ML}(\theta)\, f(\theta)\, d\theta$, averaged with prior distribution for θ).

ML Estimation: Example 1

- Given: $y_i = \mu + \sigma v_i,\ v_i \sim \mathcal{N}(0,1)$ i.i.d. or $y_i \sim \mathcal{N}(\mu,\sigma^2)$ i.i.d. $\qquad \theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$

  $Y = \mu\,\mathbf{1} + \sigma\,V\ ,\quad V \sim \mathcal{N}(0, I_n)$
- Q: $\widehat{\theta}_1 = \widehat{\mu}_{ML}, \widehat{\theta}_2 = \widehat{\sigma^2}_{ML}$
- A:

$$f(Y|\mu,\sigma^2) = \prod_{i=1}^{n} f(y_i|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{\exp[-\frac{(y_i-\mu)^2}{2\sigma^2}]}{\sqrt{2\pi\sigma^2}} = (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2]$$

$$L(\theta;Y) = \ln l(\theta;Y) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

$$\begin{cases} \frac{\partial}{\partial\mu}L(\theta;Y) = 0 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) & (1) \\ \frac{\partial}{\partial\sigma^2}L(\theta;Y) = 0 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2 & (2) \end{cases}$$

$$\begin{cases} (1) \Rightarrow\ \widehat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y} & \text{sample mean} \\ (2) \Rightarrow\ \widehat{\sigma^2}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \overline{(y-\overline{y})^2} & \text{sample variance} \end{cases}$$

ML Estimation: Example 1 (2)

bias calculations

- $E[\widehat{\mu}_{ML}|\mu,\sigma^2] = E[\overline{y}|\mu,\sigma^2] = \frac{1}{n}\sum_{i=1}^{n} E[y_i|\mu,\sigma^2] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$ unbiased!
- note: with $\overline{y} = \frac{1}{n}\mathbf{1}^T Y$, we get

$$n\,\widehat{\sigma^2}_{ML} = \sum_{i=1}^{n}(y_i-\overline{y})^2 = \left\| \begin{bmatrix} y_1-\overline{y} \\ \vdots \\ y_n-\overline{y} \end{bmatrix} \right\|^2 = \|Y-\overline{y}\mathbf{1}\|^2 = (Y-\overline{y}\mathbf{1})^T(Y-\overline{y}\mathbf{1})$$

$$= (Y-\mu\mathbf{1}+\mu\mathbf{1}-\overline{y}\mathbf{1})^T(Y-\mu\mathbf{1}+\mu\mathbf{1}-\overline{y}\mathbf{1}) = (Y-\mu\mathbf{1}-(\overline{y}-\mu)\mathbf{1})^T(\cdots) = (Y-\mu\mathbf{1})^T(Y-\mu\mathbf{1})$$

$$+(\overline{y}-\mu)^2\underbrace{\mathbf{1}^T\mathbf{1}}_{=n} - 2(\overline{y}-\mu)\underbrace{\mathbf{1}^T(Y-\mu\mathbf{1})}_{=n(\overline{y}-\mu)} = \underbrace{(Y-\mu\mathbf{1})^T(Y-\mu\mathbf{1})}_{\sum_{i=1}^{n}(y_i-\mu)^2} - \frac{1}{n}(Y-\mu\mathbf{1})^T\mathbf{1}\mathbf{1}^T(Y-\mu\mathbf{1})$$

hence $\widehat{\sigma^2}_{ML}$ is biased:

$$E[\widehat{\sigma^2}_{ML}|\mu,\sigma^2] = \frac{1}{n}E_{Y|\mu,\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 - \frac{1}{n^2}\text{tr}\{\mathbf{1}\mathbf{1}^T\ \overbrace{E_{Y|\mu,\sigma^2}(Y-\mu\mathbf{1})(Y-\mu\mathbf{1})^T}^{=\sigma^2 E_V VV^T}\}$$

$$= \sigma^2 - \frac{1}{n^2}\text{tr}\{\mathbf{1}\mathbf{1}^T\sigma^2 I_n\} = \sigma^2 - \frac{1}{n^2}\sigma^2\underbrace{\mathbf{1}^T I_n\mathbf{1}}_{=n} = (1-\frac{1}{n})\sigma^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

- unbiased variance estimate: $\widehat{\sigma^2}_{ub} = \frac{n}{n-1}\widehat{\sigma^2}_{ML} = \frac{1}{n-1}\sum_{i-1}^{n}(y_i-\overline{y})^2$

  however, can show: $Var\{\widehat{\sigma^2}_{ub}\} \geq Var\{\widehat{\sigma^2}_{ML}\}$ (and similarly for MSE).

ML Estimation: Example 2

- given: $y_i \sim \mathcal{U}[\theta-\frac{1}{2},\theta+\frac{1}{2}]$ i.i.d. $f(y_i|\theta)=\begin{cases}1 & , \; y_i\in[\theta-\frac{1}{2},\theta+\frac{1}{2}]\\ 0 & , \; \text{elsewhere}\end{cases}$

- Q: $\widehat{\theta}_{ML}$

- A: use the indicator function $I_A(x)=\begin{cases}1 & , \; x\in A\\ 0 & , \; x\notin A\end{cases}$

$$f(y_i|\theta)=I_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(y_i)=1 \text{ if } \theta-\frac{1}{2}\le y_i\le\theta+\frac{1}{2}\Leftrightarrow y_i-\frac{1}{2}\le\theta\le y_i+\frac{1}{2}$$

hence

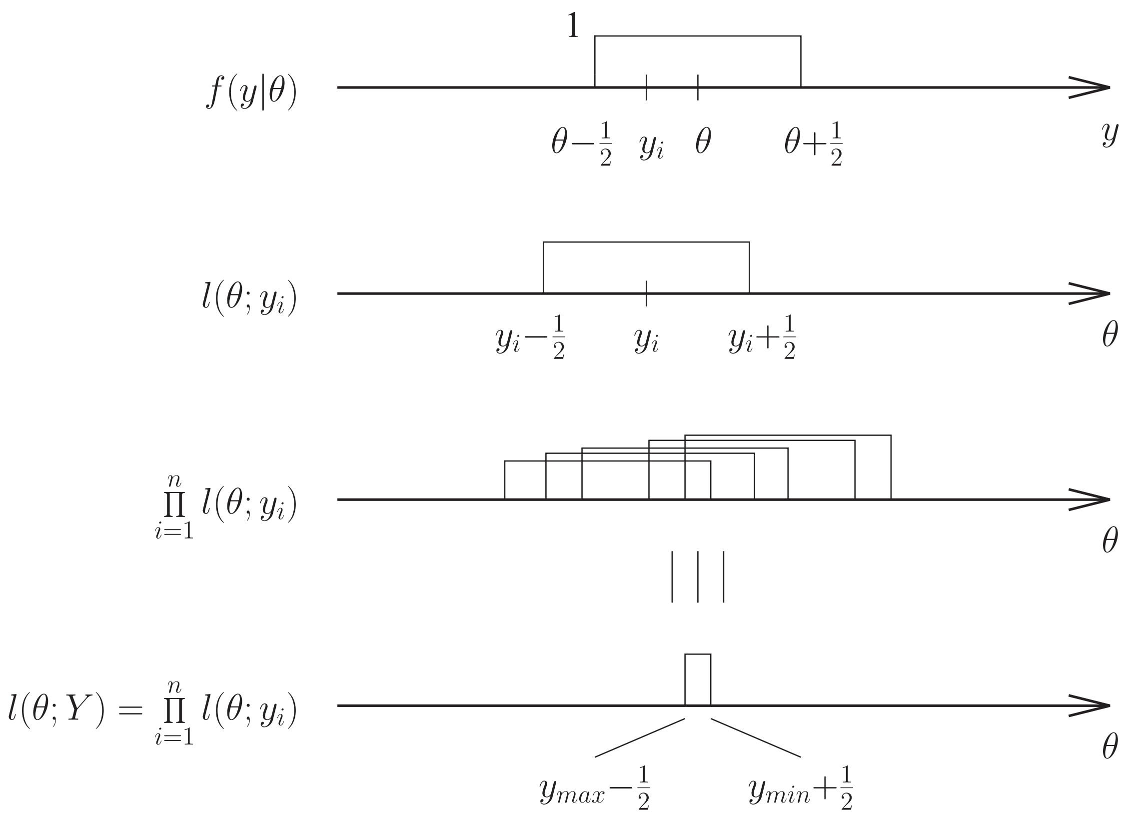
$$f(Y|\theta)=\prod_{i=1}^{n} f(y_i|\theta)=\prod_{i=1}^{n} I_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(y_i)=\prod_{i=1}^{n} I_{[y_i-\frac{1}{2},y_i+\frac{1}{2}]}(\theta)$$

$$=I_{\bigcap_{i=1}^{n}[y_i-\frac{1}{2},y_i+\frac{1}{2}]}(\theta)=I_{[y_{max}-\frac{1}{2},y_{min}+\frac{1}{2}]}(\theta)$$

hence $\widehat{\theta}\in[y_{max}-\frac{1}{2},y_{min}+\frac{1}{2}]$ a whole interval!

- choose $\widehat{\theta}_{ML}=\dfrac{y_{min}+y_{max}}{2}$

ML Estimation: Example 2 (2)



Fisher Information Matrix

- The information matrix for deterministic parameters is defined as

$$J(\theta) \;=\; R_{\frac{\partial L}{\partial \theta},\frac{\partial L}{\partial \theta}} \;=\; E_{Y|\theta}\left(\frac{\partial\,\ln\,f(Y|\theta)}{\partial\theta}\right)\left(\frac{\partial\,\ln\,f(Y|\theta)}{\partial\theta}\right)^T \;=\; -E_{Y|\theta}\,\frac{\partial}{\partial\theta}\left(\frac{\partial\,\ln\,f(Y|\theta)}{\partial\theta}\right)^T$$

  It can again be shown to satisfy all the properties we specified for an information matrix. The second equality can be shown as before. Note that $J(\theta)$ now depends on the true parameter value θ.

- unbiased estimators: $b_{\widehat{\theta}}(\theta) \;=\; E_{Y|\theta}\widehat{\theta}(Y) \;-\; \theta \;=\; 0 \;\;, \;\forall\theta \in \Theta$

- **Lemma 0.1 (Unit Cross Correlation)** *For any unbiased estimator* $\widehat{\theta}(Y)$

$$E_{Y|\theta}\,\frac{\partial\,\ln\,f(Y|\theta)}{\partial\theta}\,\left(\widehat{\theta}-\theta\right)^T \;=\; I\;.$$

  In words, the cross correlation matrix between $\frac{\partial\,\ln\,f(Y|\theta)}{\partial\theta}$ and the estimation error of any unbiased estimator is the identity matrix.

Cramer-Rao Bound

- **Theorem (CRB for Deterministic Parameters)** *If the estimator* $\widehat{\theta}(Y)$ *of* θ *is unbiased, then the covariance matrix of the parameter estimation errors* $\widetilde{\theta}$ *is bounded below by the inverse of the information matrix:*

$$C_{\widetilde{\theta}\widetilde{\theta}} = R_{\widetilde{\theta}\widetilde{\theta}} \;=\; E_{Y|\theta}\,(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \geq J^{-1}(\theta)$$

  with equality iff

$$\widehat{\theta}(Y) - \theta \;=\; J^{-1}(\theta)\frac{\partial\,\ln\,f(Y|\theta)}{\partial\,\theta} \qquad a.e.\;(\theta)$$

  An estimator that achieves the lower bound $(\forall\theta \in \Theta)$ is called *efficient*.

Remarks:

- when equality holds, we can integrate to get

$$f(Y|\theta) \;=\; h(Y)\;\exp[c_1^T(\theta)\widehat{\theta}(Y) - c_0(\theta)]$$

  where $\frac{\partial c_1^T(\theta)}{\partial\theta} \;=\; J(\theta)$ and $\frac{\partial c_0(\theta)}{\partial\theta} \;=\; J(\theta)\theta$. Hence $\{f(Y|\theta),\,\theta\,\in\,\Theta\}$ forms an exponential family and $\widehat{\theta}(Y)$ is a sufficient statistic.

Cramer-Rao Bound: Remarks

- the CRB $J^{-1}(\theta)$ only depends on $f(Y|\theta)$, not on $\widehat{\theta}(Y)$
- the (deterministic) CRB has two uses:
  (i) evaluate unbiased estimators: $\widehat{\theta}$ with $b_{\widehat{\theta}}(\theta) \equiv 0$: if $C_{\widetilde{\theta}\widetilde{\theta}} - J^{-1}(\theta)$ small enough, then $\widehat{\theta}$ good enough
  (ii) find UMVUE: $\min\limits_{\widehat{\theta}:b_{\widehat{\theta}}\equiv 0} C_{\widetilde{\theta}\widetilde{\theta}} \geq J^{-1}(\theta)$.
  If $\widehat{\theta}$ is efficient ($\forall\theta \in \Theta$), $C_{\widetilde{\theta}\widetilde{\theta}} = J^{-1}(\theta)$, then $\widehat{\theta}$ is UMVUE!
- **Theorem** *Suppose* $\widehat{\theta}_{ML}$ *is obtained by* $\frac{\partial}{\partial\theta} f(Y|\theta)\big|_{\theta=\widehat{\theta}_{ML}} = 0$. *Then if an efficient estimator exists, it is* $\widehat{\theta}_{ML}$.
  Proof: $\widehat{\theta}_{eff}$ satisfies
  $$\frac{\partial \ln f(Y|\theta)}{\partial \theta} = \underbrace{J(\theta)}_{>0}[\widehat{\theta}_{eff} - \theta]$$
  For $\theta = \widehat{\theta}_{ML}$, LHS = 0, hence RHS = 0 : $\widehat{\theta}_{eff} = \widehat{\theta}_{ML}$.
- If $J(\theta)$ is singular $\Rightarrow$ (local) *unidentifiability*. E.g. linear model with $n < m$.

Cramer-Rao Bound: Example

- i.i.d. $y_i \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 known, $\theta = \mu$
- $f(Y|\mu) = \prod\limits_{i=1}^{n} f(y_i|\mu) = (2\pi\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(y_i-\mu)^2]$
- $\dfrac{\partial \ln f(Y|\mu)}{\partial\mu} = \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}(y_i-\mu)$, $\quad \dfrac{\partial^2 \ln f(Y|\mu)}{\partial\mu^2} = -\dfrac{n}{\sigma^2}$
- $J = -E_{Y|\mu}\dfrac{\partial^2 \ln f(Y|\mu)}{\partial\mu^2} = \dfrac{n}{\sigma^2}$, $\quad C_{\widetilde{\mu}\widetilde{\mu}} = E_{Y|\mu}(\widehat{\mu}-\mu)^2 \geq J^{-1} = \dfrac{\sigma^2}{n}$
- $\widehat{\mu}_{ML} = \overline{y} = \dfrac{1}{n}\sum\limits_{i=1}^{n} y_i$, $\quad E_{Y|\mu}\widehat{\mu}_{ML} = \mu$: unbiased
- $C_{\widetilde{\mu}\widetilde{\mu}} = E_{Y|\mu}(\widehat{\mu}-\mu)^2 = E_{Y|\mu}\left(\dfrac{1}{n}\sum\limits_{i=1}^{n}(y_i-\mu)\right)^2$
  $= \frac{1}{n^2}(\sum\limits_{i=1}^{n}\underbrace{E(y_i-\mu)^2}_{=\sigma^2} + \sum\limits_{i\neq j}\underbrace{E(y_i-\mu)(y_j-\mu)}_{=(Ey_i-\mu)(Ey_j-\mu)=0}) = \dfrac{1}{n^2}n\,\sigma^2 = \dfrac{\sigma^2}{n} = J^{-1}$
- efficient: $\dfrac{\partial \ln f(Y|\mu)}{\partial\mu} = \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}(y_i-\mu) = \dfrac{n}{\sigma^2}(\overline{y}-\mu) = J\ (\widehat{\mu}_{ML}-\mu)$

The Deterministic Linear Model

- $Y = H\theta + V\ ,\ \ V \sim \mathcal{N}(0, C_{VV})$
- $f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta) = f_{\mathbf{V}}(Y - H\theta) = \frac{1}{\sqrt{(2\pi)^n \det C_{VV}}}\, e^{-\frac{1}{2}(Y-H\theta)^T C_{VV}^{-1}(Y-H\theta)}$
- $\frac{\partial\ \ln\ f_{\mathbf{V}}(Y - H\theta)}{\partial\theta} = H^T C_{VV}^{-1}(Y - H\theta) = 0$
  $\Rightarrow\ \ \widehat{\theta}_{ML} = (H^T C_{VV}^{-1} H)^{-1}\, H^T C_{VV}^{-1} Y$
- $\frac{\partial}{\partial\theta}\left(\frac{\partial\ \ln\ f_{\mathbf{V}}(Y - H\theta)}{\partial\theta}\right)^T = -\underbrace{H^T \underbrace{C_{VV}^{-1}}_{>0} H}_{>0} = -J < 0\ \ \Rightarrow$ maximum!
  assuming H full column rank
- $\widetilde{\theta} = \theta - \widehat{\theta} = -(H^T C_{VV}^{-1} H)^{-1}\, H^T C_{VV}^{-1} V\ ,\ \ E_{Y|\theta}\widetilde{\theta} = E_V \widetilde{\theta} = 0 \Rightarrow$ unbiased!
- $C_{\widetilde{\theta}\widetilde{\theta}} = R_{\widetilde{\theta}\widetilde{\theta}} = E_{Y|\theta}\widetilde{\theta}\widetilde{\theta}^T = E_V \widetilde{\theta}\widetilde{\theta}^T = (H^T C_{VV}^{-1} H)^{-1} = J^{-1}$: efficient!
- $\frac{\partial\ \ln\ f_{\mathbf{V}}(Y - H\theta)}{\partial\theta} = H^T C_{VV}^{-1} Y - H^T C_{VV}^{-1} H\theta = J(\widehat{\theta} - \theta)$: efficient