

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

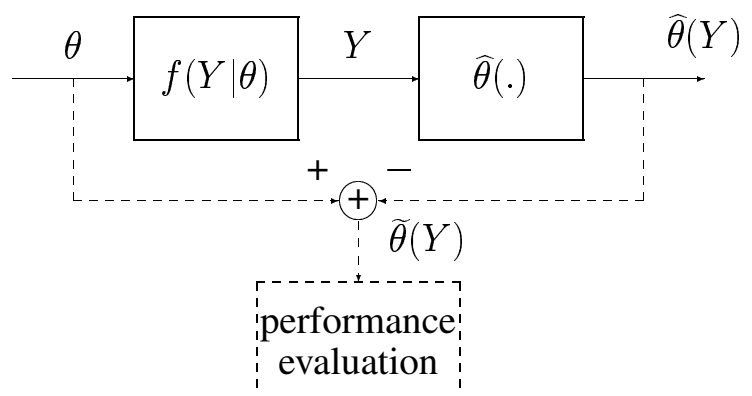
The only stochastic description available is the conditional density $f(Y|\theta)$ describing the stochastic relation between the unknown parameters θ and the observed measurements Y .

- 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 $\hat{\theta}(Y)$ of θ is again a function of Y (a statistic), with estimation error $\tilde{\theta} = \theta - \hat{\theta}(Y)$
- to evaluate the quality of an estimator, we shall again introduce the *risk* function MSE as the average value of the SE *cost* function

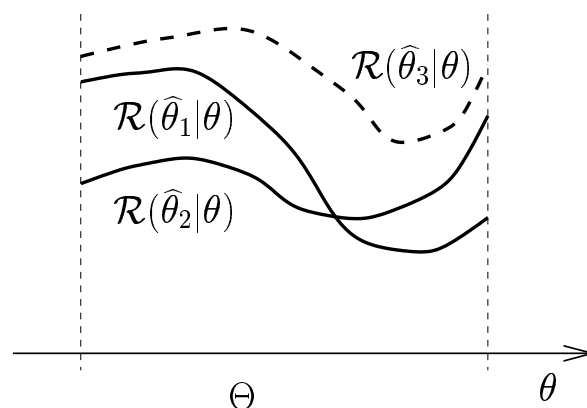
$$\text{MSE} = \mathcal{R}(\hat{\theta}(\cdot)|\theta) = E_{Y|\theta} \|\tilde{\theta}\|^2 = \int f(Y|\theta) \|\theta - \hat{\theta}(Y)\|^2 dY$$

the MSE is a function of θ in general

- minimization of the risk function leads to $\hat{\theta} = \theta$ (and $\mathcal{R} = 0$): not an acceptable strategy since the resulting $\hat{\theta}$ depends on the unknown θ
- ideally, would like $\hat{\theta}(\cdot)$ such that $\mathcal{R}(\hat{\theta}(\cdot)|\theta)$ is minimized $\forall \theta \in \Theta$: impossible! Consider $\hat{\theta}(Y) = \theta_0 \in \Theta$: ignores the data Y but $\mathcal{R}(\hat{\theta}(\cdot)|\theta_0) = 0$
- 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 $\hat{\theta}_1(Y)$ and $\hat{\theta}_2(Y)$, 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

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

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

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

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

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

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

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

$\hat{\theta}$ is *admissible* if no such $\hat{\theta}'$ exists. Example: $\hat{\theta}_3$ in figure above.

Some Optimality Properties (2)

- MSE = $E_{Y|\theta}\|\tilde{\theta}\|^2 = E_{Y|\theta}\tilde{\theta}^T\tilde{\theta} = \text{tr}\{E_{Y|\theta}\tilde{\theta}\tilde{\theta}^T\} = \text{tr}\{R_{\tilde{\theta}\tilde{\theta}}\},$

$R_{\tilde{\theta}\tilde{\theta}} = E_{Y|\theta}\tilde{\theta}\tilde{\theta}^T$ = estimation error correlation matrix

$$\begin{aligned} R_{\tilde{\theta}\tilde{\theta}} &= E_{Y|\theta}(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T = E_{Y|\theta}[\hat{\theta}(-E_{Y|\theta}\tilde{\theta} + E_{Y|\theta}\hat{\theta}) - \theta][\hat{\theta}(-E_{Y|\theta}\tilde{\theta} + E_{Y|\theta}\hat{\theta}) - \theta]^T \\ &= E_{Y|\theta}(\hat{\theta} - E_{Y|\theta}\hat{\theta})(\hat{\theta} - E_{Y|\theta}\hat{\theta})^T + (E_{Y|\theta}\hat{\theta} - \theta)(E_{Y|\theta}\hat{\theta} - \theta)^T \\ &= C_{\hat{\theta}\hat{\theta}} + b_{\hat{\theta}}(\theta)b_{\hat{\theta}}^T(\theta) = C_{\hat{\theta}\hat{\theta}} + (E_{Y|\theta}\tilde{\theta})(E_{Y|\theta}\tilde{\theta})^T \end{aligned}$$

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

Some Optimality Properties (3)

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

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

(inf \approx min, sup \approx max).

A minimax estimator minimizes the maximum risk over Θ .

A minimax $\hat{\theta}$ is difficult to obtain in general.

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

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

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

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

- 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

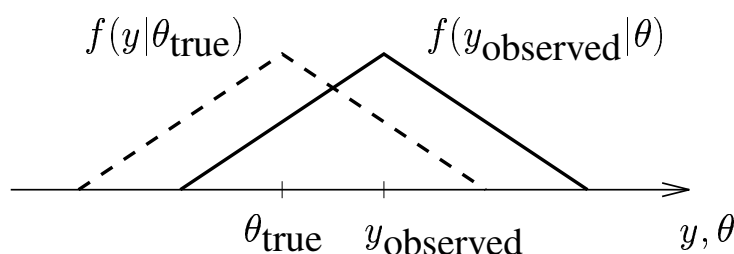
- the maximum likelihood (ML) estimation philosophy is to choose that value of the parameters that renders the observations most likely:

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

example:

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

$$\hat{\theta}_{ML}(y) = y$$



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

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

- 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 $\hat{\theta}_{ML}$ 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 $\sigma_{\theta_i}^2 \rightarrow \infty$ (information $\rightarrow 0$). Indeed

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

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

($\text{MSE}_{MAP} = \int_{\Theta} \text{MSE}_{ML}(\theta) f(\theta) d\theta$, 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. $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$

- Q: $\hat{\theta}_1 = \hat{\mu}_{ML}$, $\hat{\theta}_2 = \hat{\sigma}_{ML}^2$

- A:

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

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

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

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

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

ML Estimation: Example 1 (2)

bias calculations

- $E[\hat{\mu}_{ML}|\mu, \sigma^2] = E[\bar{y}|\mu, \sigma^2] = \frac{1}{n} \sum_{i=1}^n E[y_i|\mu, \sigma^2] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$ unbiased!

- note: with $\bar{y} = \frac{1}{n} \mathbf{1}^T Y$, we get

$$\begin{aligned} n \hat{\sigma}_{ML}^2 &= \sum_{i=1}^n (y_i - \bar{y})^2 = \left\| \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} \right\|^2 = \|Y - \bar{y} \mathbf{1}\|^2 = (Y - \bar{y} \mathbf{1})^T (Y - \bar{y} \mathbf{1}) \\ &= (Y - \mu \mathbf{1} + \mu \mathbf{1} - \bar{y} \mathbf{1})^T (Y - \mu \mathbf{1} + \mu \mathbf{1} - \bar{y} \mathbf{1}) = (Y - \mu \mathbf{1} - (\bar{y} - \mu) \mathbf{1})^T (\cdots) = (Y - \mu \mathbf{1})^T (Y - \mu \mathbf{1}) \\ &\quad + (\bar{y} - \mu)^2 \underbrace{\mathbf{1}^T \mathbf{1}}_{=n} - 2(\bar{y} - \mu) \underbrace{\mathbf{1}^T (Y - \mu \mathbf{1})}_{=n(\bar{y} - \mu)} = \underbrace{(Y - \mu \mathbf{1})^T (Y - \mu \mathbf{1})}_{\sum_{i=1}^n (y_i - \mu)^2} - \frac{1}{n} (Y - \mu \mathbf{1})^T \mathbf{1} \mathbf{1}^T (Y - \mu \mathbf{1}) \end{aligned}$$

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

$$\begin{aligned} E[\hat{\sigma}_{ML}^2|\mu, \sigma^2] &= \frac{1}{n} E_{Y|\mu, \sigma^2} \sum_{i=1}^n (y_i - \mu)^2 - \frac{1}{n^2} \text{tr}\{\mathbf{1} \mathbf{1}^T \overbrace{E_{Y|\mu, \sigma^2} (Y - \mu \mathbf{1})(Y - \mu \mathbf{1})^T}^{E_V V V^T}\} \\ &= \sigma^2 - \frac{1}{n^2} \text{tr}\{\mathbf{1} \mathbf{1}^T \sigma^2 I_n\} = \sigma^2 - \frac{1}{n^2} \sigma^2 \underbrace{\mathbf{1}^T I_n \mathbf{1}}_{=n} = (1 - \frac{1}{n}) \sigma^2 = \frac{n-1}{n} \sigma^2 \neq \sigma^2 \end{aligned}$$

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

however, can show: $\text{Var}\{\hat{\sigma}_{ub}^2\} \geq \text{Var}\{\hat{\sigma}_{ML}^2\}$ (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: $\hat{\theta}_{ML}$

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

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

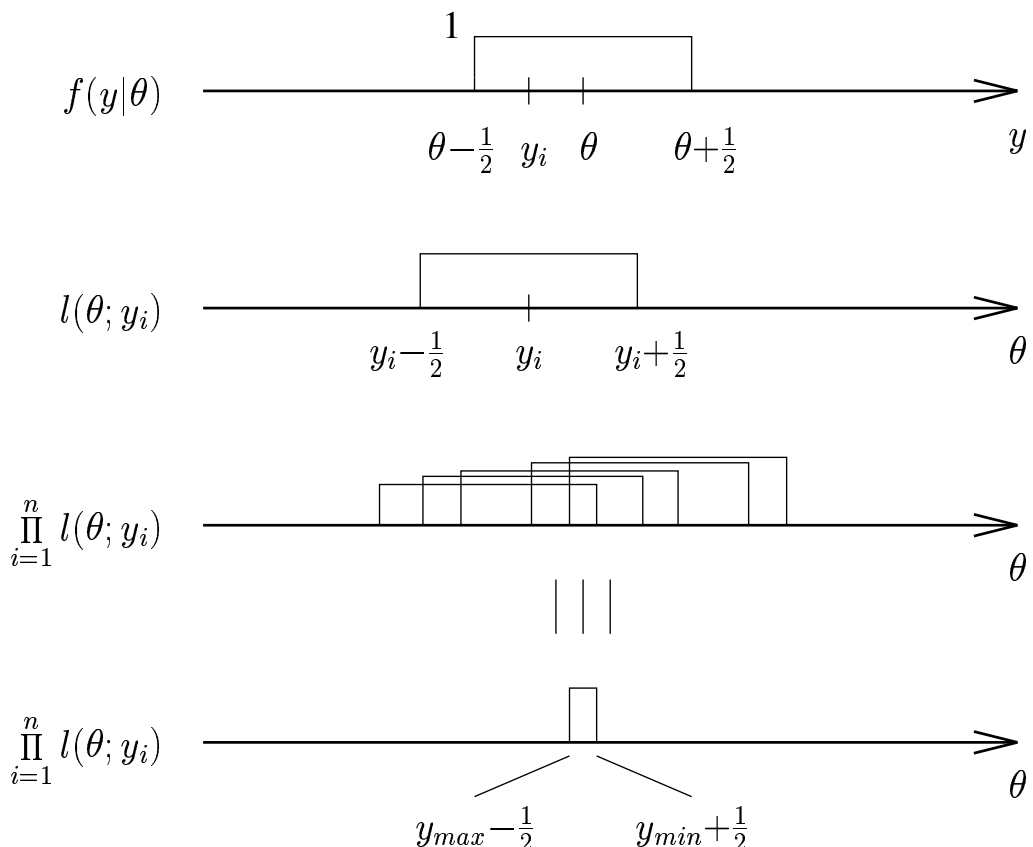
hence

$$\begin{aligned} f(Y|\theta) &= \prod_{i=1}^n f(y_i|\theta) = \prod_{i=1}^n I_{[\theta - \frac{1}{2}, \theta + \frac{1}{2}]}(y_i) = \prod_{i=1}^n I_{[y_i - \frac{1}{2}, y_i + \frac{1}{2}]}(\theta) \\ &= I_{\bigcap_{i=1}^n [y_i - \frac{1}{2}, y_i + \frac{1}{2}]}(\theta) = I_{[y_{\max} - \frac{1}{2}, y_{\min} + \frac{1}{2}]}(\theta) \end{aligned}$$

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

- choose $\hat{\theta}_{ML} = \frac{y_{\min} + y_{\max}}{2}$

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: $b_{\hat{\theta}}(\theta) = E_{Y|\theta} \hat{\theta}(Y) - \theta = 0$, $\forall \theta \in \Theta$
- **Lemma 0.1 (Unit Cross Correlation)** For any unbiased estimator $\hat{\theta}(Y)$

$$E_{Y|\theta} \frac{\partial \ln f(Y|\theta)}{\partial \theta} (\hat{\theta} - \theta)^T = I.$$

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 $\hat{\theta}(Y)$ of θ is unbiased, then the covariance matrix of the parameter estimation errors $\tilde{\theta}$ is bounded below by the inverse of the information matrix:

$$C_{\tilde{\theta}\tilde{\theta}} = R_{\tilde{\theta}\tilde{\theta}} = E_{Y|\theta} (\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \geq J^{-1}(\theta)$$

with equality iff

$$\hat{\theta}(Y) - \theta = J^{-1}(\theta) \frac{\partial \ln f(Y|\theta)}{\partial \theta} \quad a.e. (\theta)$$

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

Remarks:

- when equality holds, we can integrate to get

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

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 $\hat{\theta}(Y)$ is a sufficient statistic.

Cramer-Rao Bound: Remarks

- the CRB $J^{-1}(\theta)$ only depends on $f(Y|\theta)$, not on $\hat{\theta}(Y)$
- the (deterministic) CRB has two uses:
 - (i) evaluate unbiased estimators: $\hat{\theta}$ with $b_{\hat{\theta}}(\theta) \equiv 0$: if $C_{\hat{\theta}\hat{\theta}} - J^{-1}(\theta)$ small enough, then $\hat{\theta}$ good enough
 - (ii) find UMVUE: $\min_{\hat{\theta}: b_{\hat{\theta}} \equiv 0} C_{\hat{\theta}\hat{\theta}} \geq J^{-1}(\theta)$.

If $\hat{\theta}$ is efficient ($\forall \theta \in \Theta$), $C_{\hat{\theta}\hat{\theta}} = J^{-1}(\theta)$, then $\hat{\theta}$ is UMVUE!

- **Theorem** Suppose $\hat{\theta}_{ML}$ is obtained by $\frac{\partial}{\partial \theta} f(Y|\theta)|_{\theta=\hat{\theta}_{ML}} = 0$. Then if an efficient estimator exists, it is $\hat{\theta}_{ML}$.

Proof: $\hat{\theta}_{eff}$ satisfies

$$\frac{\partial \ln f(Y|\theta)}{\partial \theta} = \underbrace{J(\theta)}_{>0} [\hat{\theta}_{eff} - \theta]$$

For $\theta = \hat{\theta}_{ML}$, LHS = 0, hence RHS = 0 : $\hat{\theta}_{eff} = \hat{\theta}_{ML}$.

- 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, $\theta = \mu$
- $f(Y|\mu) = \prod_{i=1}^n f(y_i|\mu) = (2\pi\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2]$
- $\frac{\partial \ln f(Y|\mu)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu)$, $\frac{\partial^2 \ln f(Y|\mu)}{\partial \mu^2} = -\frac{n}{\sigma^2}$
- $J = -E_{Y|\mu} \frac{\partial^2 \ln f(Y|\mu)}{\partial \mu^2} = \frac{n}{\sigma^2}$, $C_{\hat{\mu}\hat{\mu}} = E_{Y|\mu} (\hat{\mu} - \mu)^2 \geq J^{-1} = \frac{\sigma^2}{n}$
- $\hat{\mu}_{ML} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, $E_{Y|\mu} \hat{\mu}_{ML} = \mu$: unbiased
- $C_{\hat{\mu}\hat{\mu}} = E_{Y|\mu} (\hat{\mu} - \mu)^2 = E_{Y|\mu} \left(\frac{1}{n} \sum_{i=1}^n (y_i - \mu) \right)^2$
 $= \frac{1}{n^2} \left(\sum_{i=1}^n \underbrace{E(y_i - \mu)^2}_{=\sigma^2} + \sum_{i \neq j} \underbrace{E(y_i - \mu)(y_j - \mu)}_{=(E y_i - \mu)(E y_j - \mu) = 0} \right) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n} = J^{-1}$
- efficient: $\frac{\partial \ln f(Y|\mu)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu) = \frac{n}{\sigma^2} (\bar{y} - \mu) = J (\hat{\mu}_{ML} - \mu)$

The Deterministic Linear Model

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