## CS/CNS/EE/IDS 165: Foundations of Machine Learning and Statistical Inference

# UMVU, Intro to Estimation, and Cramer-Rao Bound

http://tensorlab.cms.caltech.edu/users/anima/cms165-2020.html

Anima Anandkumar
Computing and Mathematical Sciences
California Institute of Technology, Pasadena CA 91125
anima@caltech.edu

# Outline

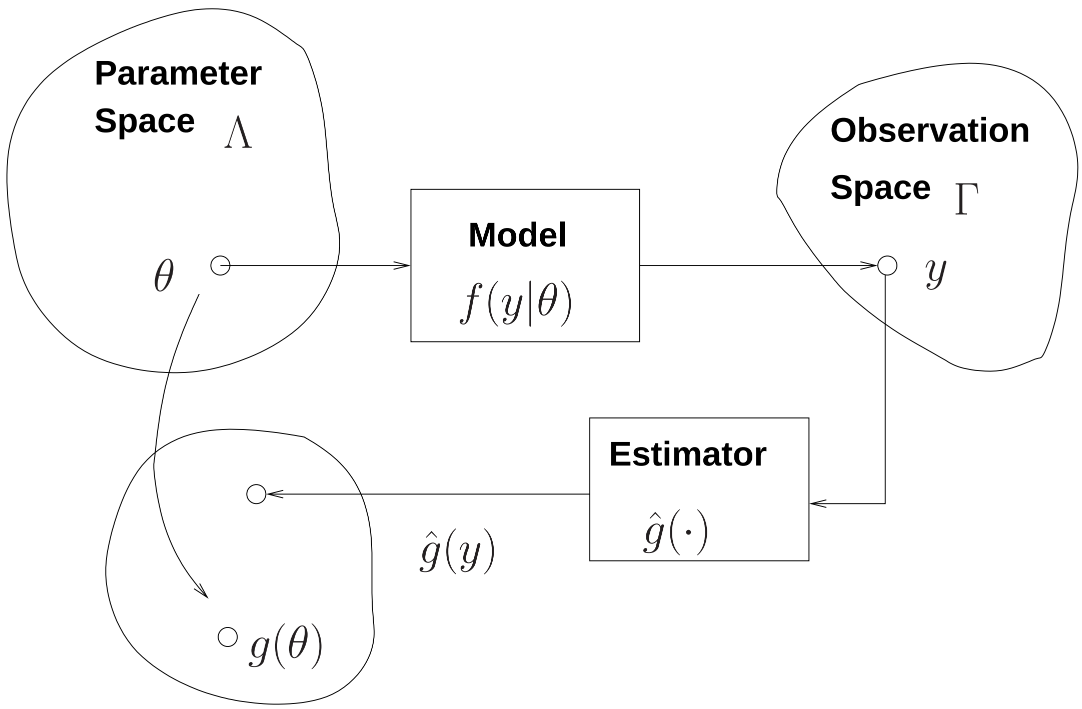
## Main Topics

- Point estimation.
- Mean square error and bias.
- Uniformly minimum variance unbiased (UMVU) estimator.
- Rao-Balckwell and Lehmann-Scheffé Theorems.

## References:


1. H.V. Poor, An Introduction to Signal Detection and Estimation, 2nd Ed., Springer-Verlag, 1994, Chapter IV-C.
2. P.J. Bickel and K.A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, Prentice Hall, Englewood Cliffs, NJ, 1977.
3. E.L. Lehmann, Theory of Point Estimation, Chapman & Hall, New York, 1991.

# Point Estimation




**The Problem** Given the observation $Y = y$ drawn from $f(y|\theta)$ with unknown deterministic parameter $\theta \in \Lambda$, estimate $g(\theta)$ with some "optimal" estimator $\hat{g}(y)$.

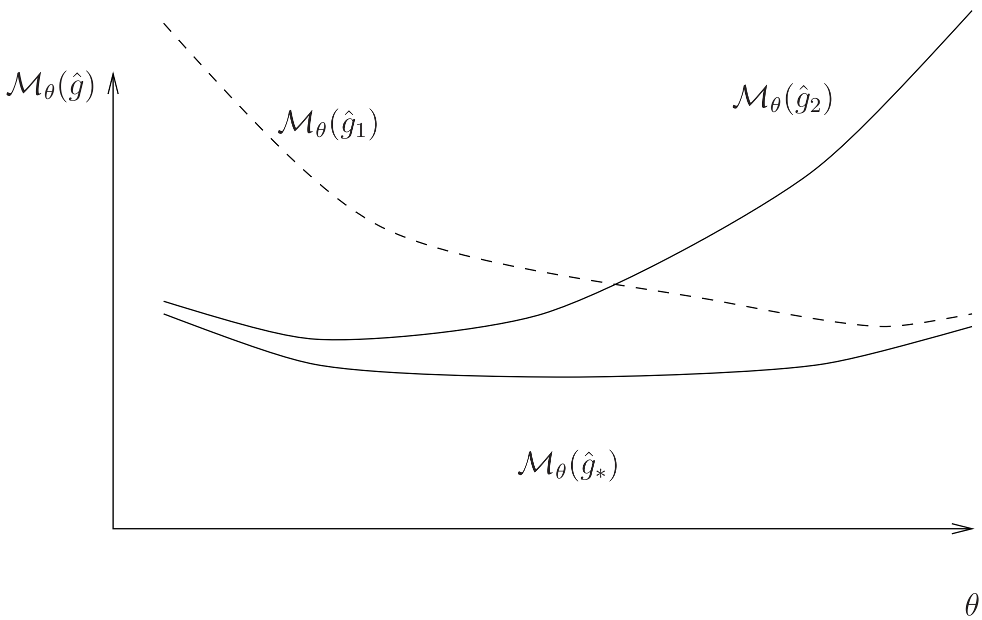
**The optimality criterion:** Minimize the mean square error

$$\mathcal{M}_\theta(\hat{g}) \triangleq \mathbb{E}(||\hat{g}(Y) - g(\theta)||^2)$$

**Remark:** Note that the MSE is in general a function of $\theta$.

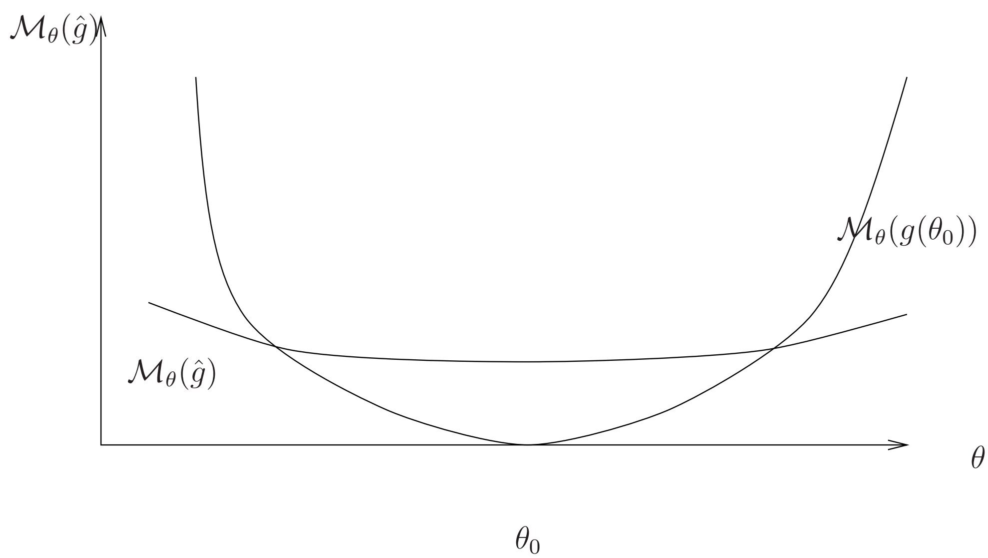
# Does the Best Estimator Exist?

## The Uniformly Best Estimator




$$\mathcal{M}_\theta(\hat{g}_*) \leq \mathcal{M}_\theta(\hat{g}), \quad \forall \theta, \hat{g}$$

The Uniformly Best Estimator does not exist!



# MSE and Bias

## MSE, Covariance and Bias

Let $\hat{\theta}(Y)$ be an estimator of $\theta \in \mathcal{R}^k$. Then

$$\begin{aligned}\mathcal{M}(\hat{\theta}) &\triangleq \mathbb{E}(||\hat{\theta} - \theta||^2)\\ &= \mathbb{E}(||\hat{\theta} - \mathbb{E}(\hat{\theta})||^2) + ||\underbrace{\mathbb{E}(\hat{\theta}) - \theta}_{B(\theta)}||^2\\ &= \text{tr}\{\text{Cov}(\hat{\theta})\} + ||B(\theta)||^2.\end{aligned}$$

where the bias of an estimator is defined by

$$B(\theta) \triangleq \mathbb{E}(\hat{\theta} - \theta)$$

**Remarks** Bias introduces systematic errors to MSE. If $B(\boldsymbol{\theta})$ is known, then removing bias reduces MSE.

## Unbiased Estimator

An estimator $\hat{g}(y)$ of $g(\theta)$ is unbiased if

$$\mathbb{E}_Y\{\hat{g}(Y)\} = g(\theta), \quad \forall \theta \in \Lambda$$

### Some Notations:

For $\theta = (\theta_1, \cdots, \theta_n)^\intercal$ and $n \times n$ matrix $\mathbf{C} = [C_{ij}]$,

- $||\theta||^2 \triangleq \sum_i |\theta_i|^2$.
- $\text{tr}\{\mathbf{C}\} \triangleq \sum_i C_{ii}$

# Examples of Unbiased Estimator

Let $X_1, \cdots, X_N$ be i.i.d. Gaussian with mean $\mu$ and variance $\sigma^2$.

- an unbiased estimator for $\mu$ is

$$\hat{\mu} = \frac{X_1 + \cdots + X_N}{N}, \qquad \mathbb{E}\{\hat{\mu}\} = \mu$$

- an unbiased estimator for $\sigma^2$ with known $\mu$ is

$$\widehat{\sigma^2} = \frac{(X_1 - \mu)^2 + \cdots + (X_N - \mu)^2}{N}, \qquad \mathbb{E}\{\widehat{\sigma^2}\} = \sigma^2$$

- a biased estimator for $\sigma^2$ with unknown $\mu$ is

$$\widehat{\sigma^2} = \frac{(X_1 - \hat{\mu})^2 + \cdots + (X_N - \hat{\mu})^2}{N}, \qquad \mathbb{E}\{\hat{\sigma}^2\} = \frac{N-1}{N}\sigma^2$$

- an unbiased variance estimator with unknown $\mu$:

$$\widehat{\sigma^2} = \frac{(X_1 - \hat{\mu})^2 + \cdots + (X_N - \hat{\mu})^2}{N-1}$$

# Existence of Unbiased Estimator

---

**A Counter Example:** Let $X$ be distributed according to the binomial distribution $\mathcal{B}(\theta, n)$ and $g(\theta) = \frac{1}{\theta}$. Is there an unbiased estimator?

If $\hat{g}(X)$ is unbiased, then

$$\mathbb{E}_X(\hat{g}) = \sum_{k=0}^{n} \hat{g}(k) \binom{n}{k} \theta^k (1-\theta)^{n-k} = \frac{1}{\theta}.$$

Therefore, no unbiased estimator exists for $\frac{1}{\theta}$. However, there exists an unbiased estimator for $\theta$ as

$$\mathbb{E}(\hat{g}(X)) = \mathbb{E}(\frac{X}{n}) = \theta.$$

**Remarks** An unbiased estimator may be desirable, but

- it may not exist;
- it may not be invariant under transformations;
- biased estimator may be satisfactory;
- the best estimator among the class of unbiased estimator may have larger MSE than those of biased estimators.

# UMVU

**UMVU** An estimator $\hat{g}$ of $\mathbf{g}(\boldsymbol{\theta})$ is uniformly minimum variance unbiased (UMVU) if

- $\mathbb{E}(\hat{g}(Y)) = g(\theta)$ for all $\theta$;
- $\mathcal{M}_\theta(\hat{g}) \leq \mathcal{M}_\theta(\hat{g}')$ for any unbiased $\hat{g}'$.

## In Search of UMVU

- Improve the estimator by the use of sufficient statistics.
- Check if the estimator is already UMVU by the use of Cramér-Rao bound.

## Caution:

- UMVU may not exist.
- UMVU may be uniformly worse than some biased estimator.

# The Rao-Blackwell Theorem

**Theorem** (Rao-Blackwell)
Suppose that $T(Y)$ is sufficient for $\theta$ and that $\hat{g}$ is an estimator for $g(\theta)$ with $\mathbb{E}(|\hat{g}(Y)|_1) < \infty$ for all $\theta$. Let

$$\hat{g}_*(y) \stackrel{\Delta}{=} \mathbb{E}(\hat{g}(Y)|T(Y) = T(y)).$$

Then for all $\boldsymbol{\theta}$

$$\mathbb{E}(||\hat{g}_*(Y) - g(\theta)||^2) \leq \mathbb{E}(||\hat{g}(Y) - g(\theta)||^2).$$

If components of $\hat{g}$ have finite variances, then the strict inequality holds unless $\hat{g}_*(Y) \stackrel{\text{a.s.}}{=} \hat{g}(Y)$.

## Remarks

- Conditioning on any sufficient statistic always reduces MSE.
- Rao-Blackwell does not imply optimality.
- Why do we require $T$ be sufficient?

Proof:

$$\begin{aligned}
\mathbb{E}(||\hat{g}_*(Y) - g(\theta)||^2) &= \mathbb{E}(||\hat{g}_*(Y) - \mathbb{E}(\hat{g}_*(Y))||^2) + ||\mathbb{E}(\hat{g}_*(Y)) - g(\theta)||^2 \\
\mathbb{E}(||\hat{g}(Y) - g(\theta)||^2) &= \mathbb{E}(||\hat{g}(Y) - \mathbb{E}(\hat{g}(Y))||^2) + ||\mathbb{E}(\hat{g}(Y)) - g(\theta)||^2
\end{aligned}$$

But $\mathbb{E}(\hat{g}_*(Y)) = \mathbb{E}(\hat{g}(Y))$, and it is always true that

$$\text{Cov}(\mathbb{E}(\hat{g}(Y)|T(Y)) \leq \text{Cov}(\hat{g}(Y))$$

with equality iff $\hat{g}(y) = \mathbb{E}(\hat{g}|T(Y) = T(y)) \stackrel{\text{a.s.}}{=} \hat{g}_*(y)$

**Note:** For two symmetrical (Hermitian) matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \geq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite, *i.e.*, for any vector $\mathbf{x}$, $\mathbf{x}^{\mathsf{T}}(\mathbf{A} - \mathbf{B})\mathbf{x} \geq 0$.

# Example

Let $Y_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, 1), i = 1, \cdots, N$. To estimate $\mu$,

- consider the simple estimator $\hat{\mu}(Y) = Y_1$
- $T(Y) = \sum Y_i$ is a sufficient statistic
- improve $\hat{\mu}$ by

$$\hat{\mu}_*(y) = \mathbb{E}(Y_1 | T(Y) = \sum_i y_i)$$

- Recall that If

$$x = \begin{bmatrix} y \\ z \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yz} \\ \boldsymbol{\Sigma}_{zy} & \boldsymbol{\Sigma}_{zz} \end{bmatrix}), \quad (1)$$

then $f(y|z)$ is the Gaussian density with

$$\mathbb{E}(y|z) = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}(\mathbf{z} - \boldsymbol{\mu}_z) \quad (2)$$

$$\text{Cov}(y, y^T|\mathbf{z}) = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yz}\Sigma_{zz}^{-1}\boldsymbol{\Sigma}_{zy} \quad (3)$$

- Since

$$\begin{pmatrix} \hat{\mu}(Y) \\ T(Y) \end{pmatrix} \sim \mathcal{N}(\begin{pmatrix} \mu \\ N\mu \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & N \end{pmatrix})$$

The conditional density of $\hat{\mu}$ is also Gaussian with

$$\hat{\mu}|T \sim \mathcal{N}(\frac{t}{N}, \frac{N-1}{N})$$

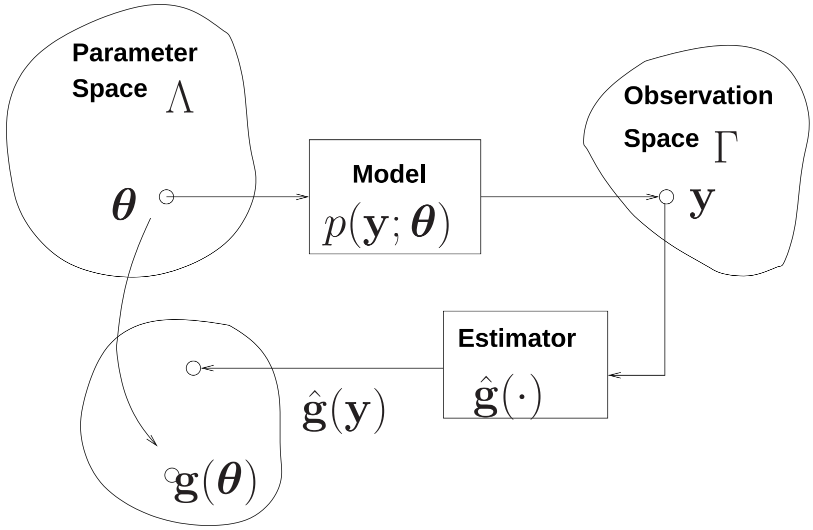
- by the Rao-Blackwell Theorem,

$$\hat{\mu}_* = E(y_1|T = t) = \frac{1}{N}\sum_i y_i$$

has a lower MSE.

- But we still don't know if $\hat{\mu}_*$ is UMVU.

# The Estimation Problem




Given random observation

$$\mathbf{Y} \sim p(\mathbf{y};\boldsymbol{\theta}), \quad \theta \in \Lambda,$$

estimate $g(\boldsymbol{\theta})$

- **Estimator:** $\hat{\mathrm{g}}(\cdot)$, a function of random vector $\mathbf{Y}$.
- **Estimate:** $\hat{\mathrm{g}}(\mathbf{y})$. A realization of the estimator corresponding to the observation $\mathbf{y}$.

## Notations

- We use $\hat{\boldsymbol{\theta}}$ to denote an estimate/estimator of $\boldsymbol{\theta}$, $\hat{\mathrm{g}}$ of $\mathrm{g}(\boldsymbol{\theta})$.

# Examples

## Sinusoid in noise

$$
\begin{aligned}
Y_k &= \cos(\theta k) + N_k, \quad N_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad k = 1, \cdots, N \\
\hat{\theta} &= \arg\max_{\theta} |\sum_k y_k e^{-j\theta k}|^2
\end{aligned}
$$

## Uniform distribution with unknown interval

$$
\begin{aligned}
Y_k &\overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, \theta), \quad k = 1, \cdots, N \\
\hat{\theta} &= \max\{y_k\}
\end{aligned}
$$

## The Gaussian Model

$$
\begin{aligned}
Y_k &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad k = 1, \cdots, N, \qquad \boldsymbol{\theta} = [\mu, \sigma^2] \\
\hat{\mu} &= \frac{1}{N}\sum_{k=1}^{N} y_k, \quad \hat{\sigma^2} = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{\mu})^2
\end{aligned}
$$

## Gaussian Signal in Gaussian Noise

$$
\begin{aligned}
Y_k &= \Theta + N_k, \quad k = 1, \cdots, N, \\
\Theta &\sim \mathcal{N}(0, \sigma_\theta^2), N_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_n^2), \\
\hat{\theta} &= \frac{1}{N}\sum_{k=1}^{N} y_k, \qquad \hat{\theta}_1 = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_n^2/N}(\frac{1}{N}\sum_{k=1}^{N} y_k)
\end{aligned}
$$

# Issues and Approaches

## Issues

- What do we mean by optimal?
- How to find an optimal estimator?
- Is an estimator good on the average?
- Are there limits on the performance?
- Does the estimator utilize data efficiently?
- Does the performance improve when the sample size increases?

## Approaches

- The Bayesian estimation for random parameters.
  - Minimum mean square error estimator (MMSE).
  - Maximum a posteriori estimator.
  - Minimax estimator.
- Point Estimation for deterministic parameters.
  - Uniform minimum variance unbiased estimator (UMVU).
  - Maximum likelihood estimator.
  - Moment estimator.

# References

1. H. V. Poor, An Introduction to Signal Detection and Estimation, 2nd Ed., Springer Verlag, 1994, Chapter 4.
2. S. M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice Hall, 1993.
3. L. L. Scharf, Statistical Signal Processing: Detection, Estimation and Time Series Analysis, Addison-Wesley, 1991, Chapter 3, 5-9.
4. H.L. Van Trees, Detection, Estimation, and Modulation Theory, vol. I. Wiley, New York, 1968, Chap. 2.
5. E.L. Lehmann, Theory of Point Estimation, Wiley, 1986.

# Outline

## Topics

- Fisher information matrix and CRB.
- CRB for functions of parameters.
- CRB for Gaussian models.
- Chapman-Robbins, Bhattachayya bounds.
- CRB for random parameters.
- CRB for complex models.

## References:


1. H.V. Poor, An Introduction to Signal Detection and Estimation, 2nd Ed., Springer-Verlag, 1994, Chapter IV-C.
2. S. M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice Hall, 1993.
3. P.J. Bickel and K.A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, Prentice Hall, Englewood Cliffs, NJ, 1977.
4. E.L. Lehmann, Theory of Point Estimation, Chapman & Hall, New York, 1991.

# Motivations

## To Find UMVU:

1. Find the complete sufficient $\mathbf{T} = \mathbf{t}(\mathbf{Y})$.
2. Two ways:
   (a) Find an unbiased estimator $\hat{\mathbf{g}}(\mathbf{T})$.
   (b) Find any unbiased estimator $\hat{\mathbf{g}}(\mathbf{Y})$ and $\hat{\mathbf{g}}_*(\mathbf{T}) = \mathbb{E}(\hat{\mathbf{g}}(\mathbf{Y})|\mathbf{T})$

## Difficulties:

1. Complete sufficient statistics may be difficult to find.
2. $\hat{\mathbf{g}}_*(\mathbf{T}) = \mathbb{E}(\hat{\mathbf{g}}(\mathbf{Y})|\mathbf{T})$ may be hard to compute.
3. It is difficult to know, without finding UMVU, whether certain performance can be achieved.

## An alternative strategy:

- Find a tight lower bound on MSE among all unbiased estimators.
- Check if the lower bound can be achieved.

# Schur Complement

## Block Diagonalization

Consider a block matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where $\mathbf{A}_{ii}$ are square and nonsingular. Matrix $\mathbf{A}$ can be diagonalized by

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Delta}_{11} \end{bmatrix}.$$

where the Schur Complement of $\mathbf{A}_{11}$ is defined as

$$\mathbf{\Delta}_{11} \triangleq \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$$

## Decorrelation

If $\mathbf{A} \geq \mathbf{0}$ is the covariance matrix[†] of a zero mean random vector $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$. The vector $\mathbf{x}$ can be decorrelated via transform

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{X}_1 \end{bmatrix}$$

with covariance $\text{Cov}(\mathbf{Y}) = \text{diag}\{\mathbf{A}_{11}, \mathbf{\Delta}_{11}\}$, and

$$\mathbf{\Delta}_{11} \triangleq \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \geq 0$$

with equality iff

$$\mathbf{X}_2 = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{X}_1 \quad \text{a.s.}$$

[†] By $\mathbf{A} \geq 0$ we mean that matrix $\mathbf{A}$ is positive semidefinite, *i.e.*, , for any column vector $\mathbf{v}$, $\mathbf{v}'\mathbf{A}\mathbf{v} \geq 0$, which implies that all diagonal blocks of $\mathbf{A}$ are also positive semidefinite.

# Score Function and Fisher Information

## Definition

Consider the real vector model $f(\mathbf{y}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathcal{R}^K$. The score function is defined by

$$\mathbf{s}(\mathbf{y};\boldsymbol{\theta}) \triangleq \begin{bmatrix} \frac{\partial}{\partial\theta_1} \ln f(\mathbf{y}|\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial\theta_K} \ln f(\mathbf{y}|\boldsymbol{\theta}) \end{bmatrix}$$

Under regularity conditions, $\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{s}(\mathbf{Y};\boldsymbol{\theta})) = \mathbf{0}$.

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}(\frac{\partial}{\partial\theta_i} \ln f(\mathbf{Y}|\boldsymbol{\theta})) &= \int f(\mathbf{y}|\boldsymbol{\theta}) \frac{\partial}{\partial\theta_i} \ln f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \\ &= \int \frac{\partial}{\partial\theta_i} f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} = \frac{\partial}{\partial\theta_i} \int f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \end{aligned}$$

## Fisher Information Matrix

The covariance matrix of $\mathbf{s}(\mathbf{Y};\boldsymbol{\theta})$ is the Fisher Information Matrix

$$\mathbf{I}(\boldsymbol{\theta}) \triangleq \mathbb{E}(\mathbf{s}(\mathbf{Y};\boldsymbol{\theta})\mathbf{s}'(\mathbf{Y};\boldsymbol{\theta})) \geq \mathbf{0}$$

The $(i,j)$th entry of $\mathbf{I}(\boldsymbol{\theta})$ can also be written as

$$\mathbf{I}_{ij}(\boldsymbol{\theta}) = \mathbb{E}(\frac{\partial}{\partial\theta_i} \ln f(\mathbf{y}|\boldsymbol{\theta}) \frac{\partial}{\partial\theta_j} \ln f(\mathbf{y}|\boldsymbol{\theta})) = -\mathbb{E}\frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln f(\mathbf{y}|\boldsymbol{\theta})$$

where the second equality is based on

$$\mathbb{E}(\frac{1}{f(\mathbf{y}|\boldsymbol{\theta})} \frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\mathbf{y}|\boldsymbol{\theta})) = \mathbf{0}$$

# The Cramér-Rao Lower Bound

**Theorem** (The scalar case.)

Given $\mathbf{Y} \sim f(\mathbf{y}|\theta)$, let $\hat{\theta}$ be a scaler unbiased estimator of $\theta$. Then, under regularity conditions[‡],

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where $I(\theta)$ is the Fisher Information. The equality holds if and only if the scoring function satisfies

$$s(y;\theta) \triangleq \frac{\partial}{\partial\theta} \ln f(y|\theta) = I(\theta)(\hat{\theta}(y) - \theta)$$

Proof:

- For any unbiased estimator $\hat{\theta}$, Consider vector $\mathbf{z} \triangleq \begin{bmatrix} \mathbf{s}(\mathbf{y};\theta) \\ \hat{\theta}(y) - \theta \end{bmatrix}$. We have $\mathbb{E}(\mathbf{z}) = \mathbf{0}$.
- Compute the covariance $\text{Cov}(\mathbf{z}) = \begin{bmatrix} I(\theta) & 1 \\ 1 & \text{Var}(\hat{\theta}) \end{bmatrix}$. The Schur complement of $I(\theta)$ implies

$$\text{Var}(\hat{\theta}) - I^{-1}(\theta) \geq 0$$

with equality holds if and only if

$$\hat{\theta}(y) - \theta = I^{-1}(\theta)s(\mathbf{y};\theta) \quad \text{almost surely}$$

**Generalization** For biased estimator, $\mathbb{E}(\hat{\theta}) = \Phi(\theta)$, then

$$\text{Var}(\hat{\theta}) \geq \frac{[\Phi'(\theta)]^2}{I(\theta)}$$

with equality iff

$$s(y;\theta) = I(\theta)(\hat{\theta}(y) - \Phi(\theta))$$

[‡]The regularity conditions involve (i) The support of $p(\mathbf{x};\theta)$ does not depend on $\theta$. (ii) All derivatives exist. (iii) Switch between $\mathbb{E}\{\cdot\}$ and $\frac{\partial}{\partial\theta}$.

# An Alternative Proof

- Unbiasedness:

$$\mathbb{E}(\hat{\theta}) = \theta \rightarrow \int (\hat{\theta} - \theta) f dy = 0 \rightarrow \int (\hat{\theta} - \theta) \frac{\partial}{\partial \theta} f dy = 1.$$

- Variation of the likelihood function $f(\mathbf{y}|\theta)$ at the true parameter:

$$\frac{\partial f}{\partial \theta} = f \frac{\partial}{\partial \theta} \ln f, \quad \mathbb{E}(\frac{\partial}{\partial \theta} \ln f) = 0$$

- Substitution:

$$\mathbb{E}\{(\hat{\theta} - \theta) \frac{\partial}{\partial \theta} \ln f\} = 1.$$

- Schwarz Inequality: $|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2) E(Y^2)$ with equality iff $Y = cX$.

$$\mathsf{Var}(\theta) \mathbb{E}(\frac{\partial}{\partial \theta} \ln f)^2 \geq 1$$

with equality only when

$$c(\theta)(\hat{\theta} - \theta) = \frac{\partial}{\partial \theta} \ln f$$

- Note:

$$\frac{\partial}{\partial \theta} \int f \frac{\partial}{\partial \theta} \ln f = \mathbb{E}(\frac{\partial}{\partial \theta} \ln f)^2 + \mathbb{E}(\frac{\partial^2}{\partial \theta^2} \ln f) = 0$$

- Finally, to find $c(\theta)$, because $\hat{\theta}$ is unbiased,

$$c(\theta) = -\mathbb{E}(\frac{\partial^2}{\partial \theta^2} \ln f) = I(\theta)$$

# Efficiency

**Definition** An unbiased estimator is efficient if it achieves CRB.

**Theorem**

If there exists an efficient estimator $\hat{\theta}$, then the distribution of the observation must be belong to the exponential family. The efficient estimator can be found by the maximum likelihood (ML) estimator:

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta} \ln f(\mathbf{y}|\theta)$$

Proof: If the CRB is achieved by an unbiased estimator $\hat{\theta}(\mathbf{y})$,

$$\frac{\partial}{\partial\theta} \ln f(\mathbf{y}|\theta) = I(\theta)(\hat{\theta}(\mathbf{y}) - \theta) \quad \text{a.s.}$$

which implies

$$f(\mathbf{y}|\theta) = h(\mathbf{y}) exp\{\hat{\theta} \int_{-\infty}^{\theta} I(u)du - \int_{-\infty}^{\theta} I(u)udu\}$$

and $\hat{\theta}$ is a complete sufficient statistic. To show that $\hat{\theta}$ is the maximum likelihood estimator, we note that

$$\frac{\partial}{\partial\theta} \ln f(\mathbf{y}|\theta)|_{\theta=\hat{\theta}_{\text{ML}}} = I(\theta)(\hat{\theta} - \hat{\theta}_{\text{ML}}) = 0.$$

**Remark** An efficient estimator is UMVU but a UMVU estimator may not be efficient (when CRB is not achievable).

# Example: Estimating Signal Amplitude

**Example:** Sinusoid in Noise:

$$x_n = \alpha cos(\omega_0 n + \phi) + w_n, \quad n = 0, \cdots, N-1,$$

where $w_n \sim \mathcal{N}(0, \sigma^2)$ and i.i.d.. All variables except $\alpha$ are known. In vector form:

$$\mathbf{x} = \mathbf{h}\alpha + \mathbf{w},$$

where

$$\begin{aligned} \mathbf{x} &= [x_0, \cdots, x_{N-1}]^t, \mathbf{w} = [w_0, \cdots, w_{N-1}]^t, \\ \mathbf{h} &= [cos(\phi), \cdots, cos(\omega_0(N-1) + \phi)]^t; \end{aligned} \tag{1}$$

1. Log-likelihood function. Denote $\mathbf{x} = [x_0, \cdots, x_{N-1}]'$.

$$\ln f(\mathbf{x}|\alpha) = -\frac{||\mathbf{x} - \mathbf{h}\alpha||^2}{2\sigma^2} + const.$$

2. The score function:

$$s(\mathbf{x}; \alpha) = \frac{||\mathbf{h}||^2}{\sigma^2}\left(\frac{\mathbf{x}^t\mathbf{h}}{||\mathbf{h}||^2} - \alpha\right)$$

3. Fisher Information:

$$I(\alpha) = \frac{||\mathbf{h}||^2}{\sigma^2}$$

4. CRLB:

$$\text{Var}(\hat{\alpha}) \geq \frac{\sigma^2}{||\mathbf{h}||^2}$$

   with equality with the least squares estimator

$$\hat{\alpha}_{LS} = \arg\min_{\alpha} ||\mathbf{x} - \alpha\mathbf{h}||^2 = \frac{\mathbf{x}^t\mathbf{h}}{||\mathbf{h}||^2},$$

   The least squares estimator is unbiased and is UMVU.

5. Asymptotic Performance: As $N \to \infty$, $Var(\alpha_{LS}) \to 0$. Consistent.

The estimator $\hat{\alpha}_{LS}$ is (i) UMVU, (ii) efficient, (iii) Gaussian, (iii) and consistent.

# Example: Estimating Signal Phase

**Example:** Sinusoid in Noise:

$$x_n = \alpha cos(\omega_0 n + \phi) + w_n, \quad n = 0, \cdots, N-1,$$

where $w_n \sim \mathcal{N}(0, \sigma^2)$ and i.i.d.. All variables except $\phi$ are known. In vector form:

$$\mathbf{x} = \mathbf{h}\alpha + \mathbf{w},$$

where

$$\begin{aligned} \mathbf{x} &= [x_0, \cdots, x_{N-1}]^t, \mathbf{w} = [w_0, \cdots, w_{N-1}]^t, \\ \mathbf{h} &= [cos(\phi), \cdots, cos(\omega_0(N-1) + \phi)]^t; \end{aligned} \tag{2}$$

1. Log-likelihood function. Denote $\mathbf{x} = [x_0, \cdots, x_{N-1}]'$.

$$\ln f(\mathbf{x}|\phi) = -\frac{||\mathbf{x} - \mathbf{h}\alpha||^2}{2\sigma^2} + const.$$

2. The score function:

$$s(\mathbf{x}; \phi) = -\frac{\alpha}{\sigma^2}(\sum_i x_i sin(i\omega_0 + \phi) - \frac{\alpha}{2}\sum_i sin(2i\omega_0 + 2\phi))$$

3. Fisher Information:

$$\begin{aligned} I(\phi) &= \frac{\alpha^2}{\sigma^2}(\sum_i cos^2(i\omega_0 + \phi) - \sum_i cos(2i\omega_0 + 2\phi)) \\ &= \frac{N\alpha^2}{2\sigma^2} - \frac{\alpha^2}{2\sigma^2}\sum_i cos(2i\omega_0 + 2\phi) \approx \frac{N\alpha^2}{2\sigma^2} \end{aligned}$$

4. CRLB:

$$\text{Var}(\hat{\phi}) \geq \frac{2\sigma^2}{N\alpha^2}$$

but unachievable. (Not the one-parameter exp. family.!)

# Example: UMVU and CRB

**Example:** Let $X$ has the Poisson Distribution with parameter $\theta$:

$$\Pr\{X = k\} = \frac{e^{-\theta}\theta^k}{k!}, \quad k = 0, 1 \cdots \tag{3}$$

To estimate $e^{-\theta}$, consider the estimator

$$T(X) = \begin{cases} 1 & \text{if } X = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

**UMVU** For an estimator $g(x)$ to be unbiased, we have

$$\sum g(k) \frac{e^{-\theta}\theta^k}{k!} = e^{-\theta}, \quad \forall \theta$$

which implies that $g(X) = T(X)$, *i.e.*, there is only one unbiased estimator. Hence $T$ is UMVU.

**CRB**

$$\text{CRB} = \theta e^{-2\theta} \tag{5}$$

$$\text{Var}(T) = e^{-2\theta}(e^{\theta} - 1) \geq \theta e^{-2\theta}. \tag{6}$$

**Remark:**

The UMVU estimator may not achieve CRLB.