Detection & Estimation Theory: Lectures 1 and 2

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning

Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

Technical University of Crete

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- Problem Definition & Basic Assumptions
- Formulation of (Bayesian) Risk
- Minimization of Bayesian Risk: Likelihood Ratio Test (LRT)
- LRT Special Cases: Maximum Aposteriori Probability (MAP) Rule
  - Maximum Likelihood (ML)
- Example

Binary Hypothesis Testing: Problem Definition

- Given a collection of measurements $\mathbf{y} \in \mathcal{Y}$, find *optimal* decision function $\delta(\mathbf{y})$ that splits observation space $\mathcal{Y}$ in two disjoint regions $\mathcal{Y}_0,\ \mathcal{Y}_1$:

$$\delta(\mathbf{y}) = \begin{cases} 0, & \text{under hypothesis } H_0 \text{ (i.e., } \mathbf{y} \in \mathcal{Y}_0) \\ 1, & \text{under hypothesis } H_1 \text{ (i.e., } \mathbf{y} \in \mathcal{Y}_1) \end{cases} \tag{1}$$

  where hypothesis 0 and hypothesis 1 are denoted by H_0, H_1, respectively.

- Vector $\mathbf{y}$ denotes a collection of measurements.
- **Continuous case**: $f_{\mathbf{y}|H_j}\left(\mathbf{y}|H_j\right),\ j \in \{0, 1\}$, i.e., conditional probability density function (pdf) is known and $f_{\mathbf{y}|H_0}\left(\mathbf{y}|H_0\right) \neq f_{\mathbf{y}|H_1}\left(\mathbf{y}|H_1\right)$.
- **Discrete case**: $\Pr\left(\mathbf{y}|H_j\right),\ j \in \{0, 1\}$, i.e., conditional probability mass function (pmf) is known and $\Pr\left(\mathbf{y}|H_0\right) \neq \Pr\left(\mathbf{y}|H_1\right)$.
- Priors $\pi_0 \triangleq \Pr\left(H_0\right) = 1 - \pi_1$, $\pi_1 \triangleq \Pr\left(H_1\right)$ are known.

Binary Hypothesis Testing: Problem Definition

- Given a collection of measurements $\mathbf{y} \in \mathcal{Y}$, find *optimal* decision function $\delta(\mathbf{y})$ that splits observation space $\mathcal{Y}$ in two disjoint regions $\mathcal{Y}_0,\ \mathcal{Y}_1$:

$$\delta(\mathbf{y}) = \begin{cases} 0, & \text{under hypothesis } H_0 \text{ (i.e., } \mathbf{y} \in \mathcal{Y}_0) \\ 1, & \text{under hypothesis } H_1 \text{ (i.e., } \mathbf{y} \in \mathcal{Y}_1) \end{cases} \tag{2}$$

  where hypothesis 0 and hypothesis 1 are denoted by H_0, H_1, respectively.

- We need an *optimality* criterion!

- Bayes comes to help: all uncertainties are quantifiable, all costs and benefits of decision can be measured!

Formulation of (Bayesian) Risk

- Define cost C_{ij} of deciding that H_i holds, when hypothesis H_j is true, $i, j \in \{0, 1\}$.
- Define

$$\Pr\left(\delta(\mathbf{y}) = i | H_j\right) =$$

$$\triangleq \Pr\left(\mathbf{y} \in \mathcal{Y}_i | H_j\right) = \begin{cases} \int_{\mathbf{y}\in\mathcal{Y}_i} f_{\mathbf{y}|H_j}\left(\mathbf{y}|H_j\right) d\mathbf{y}, & \text{(continuous case)} \\ \sum_{\mathbf{y}\in\mathcal{Y}_i} \Pr\left(\mathbf{y}|H_j\right). & \text{(discrete case)} \end{cases} \quad (3)$$

- We are ready to define the *conditional* Bayesian Risk $R\left(\cdot|\cdot\right)$ for decision rule $\delta(\mathbf{y})$ under hypothesis H_j:

$$\begin{aligned} R(\delta(\mathbf{y})|\mathrm{H}_j) &= C_{1j} \Pr(\delta(\mathbf{y}) = 1|H_j) + C_{0j} \Pr(\delta(\mathbf{y}) = 0|H_j) \\ &= \sum_{i=0}^{1} C_{ij} \Pr(\delta(\mathbf{y}) = i|H_j). \end{aligned} \quad (4)$$

Formulation of (Bayesian) Risk

- *Conditional* Bayesian Risk $R(\cdot|\cdot)$ for decision rule $\delta(\mathbf{y})$ under hypothesis H_j:

$$R(\delta(\mathbf{y})|\mathrm{H}_j) = \sum_{i=0}^{1} C_{ij} \, \Pr\left(\delta(\mathbf{y}) = i|H_j\right). \tag{5}$$

- Thus, the average unconditional Bayesian cost of decision rule $\delta(\mathbf{y})$ follows:

$$\begin{aligned}
R(\delta(\mathbf{y})) &= R(\delta(\mathbf{y})|H_0) \ \Pr(H_0) + R(\delta(\mathbf{y})|H_1) \ \Pr(H_1) && (6)\\
&= R(\delta(\mathbf{y})|H_0) \ \pi_0 + R(\delta(\mathbf{y})|H_1) \ \pi_1 && (7)\\
&= \sum_{j=0}^{1} R(\delta(\mathbf{y})|H_j) \ \pi_j && (8)\\
&\overset{Eq. (5)}{=} \sum_{i=0}^{1}\sum_{j=0}^{1} \pi_j \ C_{ij} \ \Pr\left(\delta(\mathbf{y}) = i|H_j\right) && (9)
\end{aligned}$$

Formulation of (Bayesian) Risk

- Exploiting the fact that $\mathcal{Y}_0 \cup \mathcal{Y}_1 = \mathcal{Y}$ and $\mathcal{Y}_0 \cap \mathcal{Y}_1 = \emptyset$:

$$\Pr\left(\delta(\mathbf{y}) = 0 | H_j\right) + \Pr\left(\delta(\mathbf{y}) = 1 | H_j\right) = 1 \Leftrightarrow \tag{10}$$

$$\Pr\left(\mathbf{y} \in \mathcal{Y}_0 | H_j\right) = 1 - \Pr\left(\mathbf{y} \in \mathcal{Y}_1 | H_j\right) \tag{11}$$

- Average Bayesian cost of decision rule $\delta(\mathbf{y})$:

$$R(\delta(\mathbf{y})) = \sum_{i=0}^{1} \sum_{j=0}^{1} \pi_j \; C_{ij} \; \Pr\left(\delta(\mathbf{y}) = i | H_j\right) \tag{12}$$

$$= \sum_{j=0}^{1} \sum_{i=0}^{1} \pi_j \; C_{ij} \; \Pr\left(\mathbf{y} \in \mathcal{Y}_i | H_j\right) \tag{13}$$

$$= \sum_{j=0}^{1} \pi_j \; C_{0j} \; \Pr\left(\mathbf{y} \in \mathcal{Y}_0 | H_0\right) + \pi_j \; C_{1j} \; \Pr\left(\mathbf{y} \in \mathcal{Y}_1 | H_j\right) \tag{14}$$

$$\stackrel{(11)}{=} \sum_{j=0}^{1} \pi_j \; C_{0j} + \sum_{j=0}^{1} \pi_j \; \left(C_{1j} - C_{0j}\right) \; \Pr\left(\mathbf{y} \in \mathcal{Y}_1 | H_j\right) \tag{15}$$

Minimization of Bayesian Risk

- Average Bayesian cost (Bayesian Risk) of decision rule $\delta(\mathbf{y})$:

$$R(\delta(\mathbf{y})) = \sum_{j=0}^{1} \pi_j \; C_{0j} + \sum_{j=0}^{1} \pi_j \; \left(C_{1j} - C_{0j}\right) \; \Pr\left(\mathbf{y} \in \mathcal{Y}_1 | H_j\right) \quad (16)$$

- Notice that the first sum is independent of the measurement data $\mathbf{y}$. The *optimal* decision rule should perform the following minimization:

$$\min_{\delta(\mathbf{y})} \; R(\delta(\mathbf{y})) \quad (17)$$

- Two cases: $\mathbf{y}$ continuous or discrete (solution will be found the same!).

Minimization of Bayesian Risk

- Continuous case: $\Pr\left(\mathbf{y} \in \mathcal{Y}_1 | H_j\right) = \int_{\mathbf{y} \in \mathcal{Y}_1} f_{\mathbf{y}|H_j}\left(\mathbf{y}|H_j\right) d\mathbf{y}$,
- Bayesian Risk of decision rule $\delta(\mathbf{y})$:

$$R(\delta(\mathbf{y})) = \sum_{j=0}^{1} \pi_j \; C_{0j} + \sum_{j=0}^{1} \pi_j \; \left(C_{1j} - C_{0j}\right) \; \int_{\mathbf{y} \in \mathcal{Y}_1} f_{\mathbf{y}|H_j}\left(\mathbf{y}|H_j\right) d\mathbf{y}$$

$$= \sum_{j=0}^{1} \pi_j \; C_{0j} + \int_{\mathbf{y} \in \mathcal{Y}_1} \; \sum_{j=0}^{1} \pi_j \; \left(C_{1j} - C_{0j}\right) f_{\mathbf{y}|H_j}\left(\mathbf{y}|H_j\right) d\mathbf{y} \quad (18)$$

- Remember that $\delta(\mathbf{y})$ controls what data $\mathbf{y}$ is allocated to $\mathcal{Y}_1$ and what data to $\mathcal{Y}_0$. From Eq. (18), $R(\delta(\mathbf{y}))$ is minimized when then integrand of (18) is minimized (i.e., negative or zero):

$$\delta_B(\mathbf{y}) = \arg\min_{\delta(\mathbf{y})} \; R(\delta(\mathbf{y})) \Leftrightarrow$$

$$\text{select } \mathcal{Y}_1 : \left\{ \mathbf{y} \in \mathcal{Y} : \sum_{j=0}^{1} \pi_j \; \left(C_{1j} - C_{0j}\right) f_{\mathbf{y}|H_j}\left(\mathbf{y}|H_j\right) \leq 0 \right\} \quad (19)$$

Minimization of Bayesian Risk

- Discete case: $\Pr(\mathbf{y} \in \mathcal{Y}_1 | H_j) = \sum_{\mathbf{y} \in \mathcal{Y}_1} \Pr(\mathbf{y} | H_j)$,
- Bayesian Risk of decision rule $\delta(\mathbf{y})$:

$$R(\delta(\mathbf{y})) = \sum_{j=0}^{1} \pi_j \; C_{0j} + \sum_{j=0}^{1} \pi_j \;\; (C_{1j} - C_{0j}) \;\; \sum_{\mathbf{y} \in \mathcal{Y}_1} \Pr(\mathbf{y} | H_j)$$

$$= \sum_{j=0}^{1} \pi_j \; C_{0j} + \sum_{\mathbf{y} \in \mathcal{Y}_1} \;\; \sum_{j=0}^{1} \pi_j \;\; (C_{1j} - C_{0j}) \Pr(\mathbf{y} | H_j) \, d\mathbf{y} \tag{20}$$

- Similarly to the continuous case, $R(\delta(\mathbf{y}))$ is minimized when then integrand of (20) is minimized (i.e., negative or zero):

$$\delta_B(\mathbf{y}) = \arg\min_{\delta(\mathbf{y})} \; R(\delta(\mathbf{y})) \Leftrightarrow$$

$$\text{select } \mathcal{Y}_1 : \left\{ \mathbf{y} \in \mathcal{Y} : \sum_{j=0}^{1} \pi_j \;\; (C_{1j} - C_{0j}) \Pr(\mathbf{y} | H_j) \leq 0 \right\} \tag{21}$$

Minimization of Bayesian Risk

- Define likelihood ratio (LR) for continuous or discrete case:

$$L(\mathbf{y}) = \frac{f_{\mathbf{y}|H_1}(\mathbf{y}|H_1)}{f_{\mathbf{y}|H_0}(\mathbf{y}|H_0)} \text{(continuous case)}, L(\mathbf{y}) = \frac{\Pr(\mathbf{y}|H_1)}{\Pr(\mathbf{y}|H_0)} \text{(discrete case)}.$$

- We can safely assume that the LR is finite positive for all cases of interest (see below):
  - LR numerator and denominator both positive: LRT finite positive.
  - LR numerator and denominator both zero: this is impossible, since in that case $\mathbf{y} \notin \mathcal{Y}$ (and we have also assumed that $\Pr(\mathbf{y}|H_1) \neq \Pr(\mathbf{y}|H_0)$).
  - either numerator or denominator (only one of the two) is zero; if numerator is zero then that particular $\mathbf{y}$ cannot occur under H_1; similarly, if denominator is zero then that $\mathbf{y}$ cannot occur under H_0.

Minimization of Bayesian Risk: Likelihood Ratio Test

- We further assume that $C_{01} > C_{11}$, i.e., the cost of wrong decision is strictly higher than the cost of correct decision.
- Continuous case:

$$\text{select } \mathcal{Y}_1 : \left\{ \mathbf{y} \in \mathcal{Y} : \sum_{j=0}^{1} \pi_j \; (C_{1j} - C_{0j}) \, f_{\mathbf{y}|H_j}(\mathbf{y}|H_j) \le 0 \right\} \Leftrightarrow$$

$$\Leftrightarrow \sum_{j=0}^{1} \pi_j \; (C_{1j} - C_{0j}) \, f_{\mathbf{y}|H_j}(\mathbf{y}|H_j) \overset{H_1}{\le} 0 \tag{22}$$

$$\Leftrightarrow \pi_0 \; (C_{10} - C_{00}) \, f_{\mathbf{y}|H_0}(\mathbf{y}|H_0) \overset{H_1}{\le} -\pi_1 \; (C_{11} - C_{01}) \, f_{\mathbf{y}|H_1}(\mathbf{y}|H_1)$$

$$\overset{(C_{01}-C_{11})>0}{\Leftrightarrow} \frac{f_{\mathbf{y}|H_1}(\mathbf{y}|H_1)}{f_{\mathbf{y}|H_0}(\mathbf{y}|H_0)} \overset{H_1}{\ge} \frac{C_{10} - C_{00}}{C_{01} - C_{11}} \; \frac{\pi_0}{\pi_1} \triangleq \tau \tag{23}$$

$$\Leftrightarrow L(\mathbf{y}) \overset{H_1}{\ge} \tau \tag{24}$$

Minimization of Bayesian Risk: Likelihood Ratio Test

- Continuous case:

$$L(\mathbf{y}) \triangleq \frac{f_{\mathbf{y}|H_1}(\mathbf{y}|H_1)}{f_{\mathbf{y}|H_0}(\mathbf{y}|H_0)} \overset{H_1}{\geq} \frac{C_{10}-C_{00}}{C_{01}-C_{11}} \frac{\pi_0}{\pi_1} \triangleq \tau \tag{25}$$

$$\Leftrightarrow L(\mathbf{y}) \overset{H_1}{\geq} \tau \tag{26}$$

  Notice that the values of $\mathbf{y}$ where the integrand goes to zero (or equivalently the (LR) ratio is equal to τ) do not matter; some can be allocated to $\mathcal{Y}_1$ and some (or none) to $\mathcal{Y}_0$).

- Discrete case - with similar reasoning, minimization in Eq. (21) offers the following LR test:

$$L(\mathbf{y}) \triangleq \frac{\Pr(\mathbf{y}|H_1)}{\Pr(\mathbf{y}|H_0)} \overset{H_1}{\geq} \frac{C_{10}-C_{00}}{C_{01}-C_{11}} \frac{\pi_0}{\pi_1} \triangleq \tau \tag{27}$$

$$\Leftrightarrow L(\mathbf{y}) \overset{H_1}{\geq} \tau \tag{28}$$

Minimization of Bayesian Risk: Likelihood Ratio Test

- Thus, optimum Bayesian decision rule $\delta_B(\mathbf{y})$, i.e., rule that minimizes Bayes risk, can be written as follows:

$$\delta_B(\mathbf{y}) = \begin{cases} 1, & \text{if } L(\mathbf{y}) \geq \tau, \\ 0, & \text{if } L(\mathbf{y}) < \tau, \end{cases} \tag{29}$$

or more compactly,

$$L(\mathbf{y}) \overset{H_1}{\geq} \tau. \tag{30}$$

Likelihood Ratio Test (LRT) & Symmetric Costs

- Set symmetric costs, i.e., 1 for (any) erroneous detection and 0 for (any) correct decision:

$$C_{ij} = 1 - \delta_{ij} = \begin{cases} 0, & i = j, \\ 1, & i \neq j, \end{cases} \tag{31}$$

where δ_{ij} denotes the Kronecker delta. For such costs, Bayesian Risk is equivalent to probability of error! From Eq. (4):

$$\begin{aligned}
R(\delta(\mathbf{y})|\mathrm{H}_j) &= C_{1j} \Pr(\delta(\mathbf{y}) = 1|H_j) + C_{0j} \Pr(\delta(\mathbf{y}) = 0|H_j) \Rightarrow \\
R(\delta(\mathbf{y})|\mathrm{H}_0) &= C_{10} \Pr(\delta(\mathbf{y}) = 1|H_0) + C_{00} \Pr(\delta(\mathbf{y}) = 0|H_0) \\
&= \Pr(\delta(\mathbf{y}) = 1|H_0) \equiv \Pr(\mathrm{error}|H_0). && (32) \\
R(\delta(\mathbf{y})|\mathrm{H}_1) &= C_{11} \Pr(\delta(\mathbf{y}) = 1|H_1) + C_{01} \Pr(\delta(\mathbf{y}) = 0|H_1) \\
&= \Pr(\delta(\mathbf{y}) = 0|H_1) \equiv \Pr(\mathrm{error}|H_1) \Rightarrow && (33) \\
R(\delta(\mathbf{y})) &= R(\delta(\mathbf{y})|\mathrm{H}_0)\ \pi_0 + R(\delta(\mathbf{y})|\mathrm{H}_1)\ \pi_1 && (34) \\
&= \Pr(\mathrm{error}|H_0)\ \pi_0 + \Pr(\mathrm{error}|H_1)\ \pi_1 \equiv \Pr(\mathrm{error}).
\end{aligned}$$

- Set symmetric costs $C_{ij} = 1 - \delta_{ij}$, as before. For such costs, Bayesian Risk is equivalent to probability of error!
- Continuous case:[1]

$$L(\mathbf{y}) \triangleq \frac{f_{\mathbf{y}|H_1}(\mathbf{y}|H_1)}{f_{\mathbf{y}|H_0}(\mathbf{y}|H_0)} \overset{H_1}{\geq} \tau \triangleq \frac{C_{10} - C_{00}}{C_{01} - C_{11}} \frac{\pi_0}{\pi_1} = \frac{(1-0)}{(1-0)} \frac{\pi_0}{\pi_1} = \frac{\pi_0}{\pi_1}$$

$$\Leftrightarrow f_{\mathbf{y}|H_1}(\mathbf{y}|H_1)\,\pi_1 \overset{H_1}{\geq} f_{\mathbf{y}|H_0}(\mathbf{y}|H_0)\,\pi_0 \tag{35}$$

$$\Leftrightarrow \frac{f_{\mathbf{y}|H_1}(\mathbf{y}|H_1)\,\pi_1}{f_{\mathbf{y}}(\mathbf{y})} \overset{H_1}{\geq} \frac{f_{\mathbf{y}|H_0}(\mathbf{y}|H_0)\,\pi_0}{f_{\mathbf{y}}(\mathbf{y})} \tag{36}$$

$$\overset{\text{Bayes}(*)}{\Leftrightarrow} \Pr(H_1|\mathbf{y}) \overset{H_1}{\geq} \Pr(H_0|\mathbf{y}) \quad \text{(MAP Rule)} \tag{37}$$

- Discrete case: same rule as above!

[1] (*) holds because the Bayes property holds for continuous distributions as well.

- Set symmetric costs $C_{ij} = 1 - \delta_{ij}$, as before and equal priors $\pi_0 = \pi_1$ (special MAP case):
- Continuous case:

$$\frac{f_{\mathbf{y}|H_1}(\mathbf{y}|H_1)\,\pi_1}{f_{\mathbf{y}}(\mathbf{y})} \overset{H_1}{\geq} \frac{f_{\mathbf{y}|H_0}(\mathbf{y}|H_0)\,\pi_0}{f_{\mathbf{y}}(\mathbf{y})} \tag{38}$$

$$\Rightarrow f_{\mathbf{y}|H_1}(\mathbf{y}|H_1) \overset{H_1}{\geq} f_{\mathbf{y}|H_0}(\mathbf{y}|H_0) \quad \text{(ML Rule)} \tag{39}$$

- Discrete case - same derivation as above:

$$\Pr(\mathbf{y}|H_1) \overset{H_1}{\geq} \Pr(\mathbf{y}|H_0) \quad \text{(ML Rule)} \tag{40}$$

- MAP and ML minimize Bayesian risk and probability of error.

Simple example

- Assume $y_k = m_0 + v_k$ under hypothesis H_0 and $y_k = m_1 + v_k$ under hypothesis H_1, where $m_1 > m_0$ and variables $\{v_k\}$, $k \in \{1, 2, \ldots, M\}$ are derived from white Gaussian noise (WGN), i.e. $v_i \perp v_j, i \neq j$ (statistically independent) and $v_k \sim \mathcal{N}(0, \sigma^2)$. Find optimal decision rule that detects which hypothesis holds.
- Solution: Affine transformation of Gaussian is also Gaussian:

$$y_k \sim \begin{cases} \mathcal{N}(m_0, \sigma^2), & \text{under} \quad H_0 \\ \mathcal{N}(m_1, \sigma^2), & \text{under} \quad H_1 \end{cases} \tag{41}$$

Since $\{v_k\}$ are independent, observations $\{y_k\}$ are independent and the product of their conditional pdfs offers the conditional probability density of each hypothesis and their ratio:

$$f_{\mathbf{y}|H_j}(\mathbf{y} = [y_1 \;\; y_2 \;\; \ldots y_M]|\mathrm{H}_j) = \prod_{k=1}^{M} f_{y_k|H_j}(y_k|\mathrm{H}_j) = \prod_{k=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_k - m_j)^2}{2\sigma^2}\right]$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{M}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{M}(y_k - m_j)^2\right].$$

$$L(\mathbf{y}) \triangleq \frac{f_{\mathbf{y}|H_1}(\mathbf{y}|\mathrm{H}_1)}{f_{\mathbf{y}|H_0}(\mathbf{y}|\mathrm{H}_0)} = \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{M}(y_k - m_1)^2 + \frac{1}{2\sigma^2}\sum_{k=1}^{M}(y_k - m_0)^2\right].$$

$$
\begin{aligned}
L(\mathbf{y}) &= \exp\Big[-\frac{1}{2\sigma^2}\sum_{k=1}^{M}(y_k - m_1)^2 + \frac{1}{2\sigma^2}\sum_{k=1}^{M}(y_k - m_0)^2\Big] && (42)\\
&= \exp\Big[-\frac{1}{2\sigma^2}\sum_{k=1}^{M}(y_k^2 - 2\,m_1\,y_k + m_1^2) + \frac{1}{2\sigma^2}\sum_{k=1}^{M}(y_k^2 - 2\,m_0\,y_k + m_0^2)\Big] && (43)\\
&= \exp\Big[+\frac{m_1}{\sigma^2}\sum_{k=1}^{M} y_k - \frac{1}{2\sigma^2}\sum_{k=1}^{M} m_1^2 - \frac{m_0}{\sigma^2}\sum_{k=1}^{M} y_k + \frac{1}{2\sigma^2}\sum_{k=1}^{M} m_0^2\Big] && (44)\\
&= \exp\Big[+\frac{m_1 - m_0}{\sigma^2}\sum_{k=1}^{M} y_k - \frac{1}{2\sigma^2}\sum_{k=1}^{M} m_1^2 + \frac{1}{2\sigma^2}\sum_{k=1}^{M} m_0^2\Big] && (45)\\
&= \exp\Big[+\frac{m_1 - m_0}{\sigma^2}\sum_{k=1}^{M} y_k - \frac{1}{2\sigma^2}\,M\,m_1^2 + \frac{1}{2\sigma^2}\,M\,m_0^2\Big] && (46)\\
&= \exp\Big[-\frac{M\,(m_1^2 - m_0^2)}{2\sigma^2} + \frac{m_1 - m_0}{\sigma^2}\sum_{k=1}^{M} y_k\Big] && (47)
\end{aligned}
$$

$$L(\mathbf{y}) = \exp\Big[-\frac{M\,(m_1^2 - m_0^2)}{2\sigma^2} + \frac{m_1 - m_0}{\sigma^2}\sum_{k=1}^{M} y_k\Big] \tag{48}$$

$$L(\mathbf{y}) \overset{H_1}{\geq} \tau \Leftrightarrow \ln\left(L(\mathbf{y})\right) \overset{H_1}{\geq} \ln(\tau) \Leftrightarrow \tag{49}$$

$$-\frac{M\,(m_1^2 - m_0^2)}{2\sigma^2} + \frac{m_1 - m_0}{\sigma^2}\sum_{k=1}^{M} y_k \overset{H_1}{\geq} \ln(\tau) \Leftrightarrow \tag{50}$$

$$+\frac{m_1 - m_0}{\sigma^2}\sum_{k=1}^{M} y_k \overset{H_1}{\geq} +\frac{M\,(m_1^2 - m_0^2)}{2\sigma^2} + \ln(\tau) \overset{m_1 > m_0}{\Leftrightarrow} \tag{51}$$

$$+\frac{1}{M}\sum_{k=1}^{M} y_k \overset{H_1}{\geq} +\frac{(m_1 + m_0)}{2} + \frac{\sigma^2}{M(m_1 - m_0)}\ln(\tau), \tag{52}$$

where we used the fact that $(m_1 - m_0) > 0$.

- The left-hand side of the above inequality, i.e., the term $\frac{1}{M}\sum_{k=1}^{M} y_k$, is called the sufficient statistic.
- Observe that the sufficient statistic is the sample mean for $M \to +\infty$, under each hypothesis.

References

Bernard C. Levy, Principles of Signal Detection and Parameter Estimation, Springer 2008.

Thank you!

Detection & Estimation Theory: Lecture 3

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning

Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- One more simple (binary hypothesis testing) example
- Common distributions for Sufficient Statistics
- Gaussian vectors (or jointly Gaussian random variables)

Another simple example

- Assume $y_k = v_k$, where $v_k \sim \mathcal{N}(0, \sigma_0^2)$ under hypothesis H_0 and $y_k = v_k$, where $v_k \sim \mathcal{N}(0, \sigma_1^2)$ under hypothesis H_1, with $\sigma_1^2 > \sigma_0^2$ and variables $\{v_k\}$, $k \in \{1, 2, \dots, M\}$ are derived from white Gaussian noise (WGN), i.e. $v_i \perp v_j, i \neq j$ (statistically independent) and v_k is Gaussian. Find optimal decision rule that detects which hypothesis holds.
- Notice that in this example, the variance of measurements changes per hypothesis (and not the mean, as in the previous example).
- Solution: Affine transformation of Gaussian is also Gaussian:

$$y_k \sim \begin{cases} \mathcal{N}(0, \sigma_0^2), & \text{under} \quad H_0 \\ \\ \mathcal{N}(0, \sigma_1^2), & \text{under} \quad H_1 \end{cases} \tag{1}$$

Since $\{v_k\}$ are independent, observations $\{y_k\}$ are independent and the product of their conditional pdfs offers the conditional probability density of each hypothesis and their ratio, as follows (with $j \in \{0, 1\}$):

$$f_{\mathbf{y}|H_j}(\mathbf{y} = [y_1 \ y_2 \ \dots y_M]|\mathrm{H}_j) = \prod_{k=1}^{M} f_{y_k|H_j}(y_k|\mathrm{H}_j) = \prod_{k=1}^{M} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{y_k^2}{2\sigma_j^2}\right]$$

$$= \frac{1}{(2\pi\sigma_j^2)^{\frac{M}{2}}} \exp\Big[-\frac{1}{2\sigma_j^2} \sum_{k=1}^{M} y_k^2 \Big].$$

$$L(\mathbf{y}) \triangleq \frac{f_{\mathbf{y}|H_1}(\mathbf{y}|\mathrm{H}_1)}{f_{\mathbf{y}|H_0}(\mathbf{y}|\mathrm{H}_0)} = \frac{\sigma_0^M}{\sigma_1^M} \exp\Big[\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2} \right) \sum_{k=1}^{M} y_k^2 \Big].$$

$$L(\mathbf{y}) = \frac{\sigma_0^M}{\sigma_1^M} \exp\Big[\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2} \right) \sum_{k=1}^{M} y_k^2 \Big]. \tag{2}$$

$$L(\mathbf{y}) \overset{H_1}{\geq} \tau \Leftrightarrow \ln\left(L(\mathbf{y})\right) \overset{H_1}{\geq} \ln(\tau) \Leftrightarrow \tag{3}$$

$$-M \ln\left(\frac{\sigma_1}{\sigma_0}\right) + \frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2} \sum_{k=1}^{M} y_k^2 \overset{H_1}{\geq} \ln(\tau) \Leftrightarrow \tag{4}$$

$$\frac{1}{M} \; \frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2} \sum_{k=1}^{M} y_k^2 \overset{H_1}{\geq} \frac{1}{M} \; \ln(\tau) + \ln\left(\frac{\sigma_1}{\sigma_0}\right) \overset{\sigma_1^2 > \sigma_0^2}{\Leftrightarrow} \tag{5}$$

$$\frac{1}{M} \sum_{k=1}^{M} y_k^2 \overset{H_1}{\geq} \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2} \left(\frac{1}{M} \; \ln(\tau) + \ln\left(\frac{\sigma_1}{\sigma_0}\right)\right), \tag{6}$$

where we used the fact in Eq. (6) that $(\sigma_1^2 - \sigma_0^2) > 0$.

- In this case, the sufficient statistic is the term $\frac{1}{M} \sum_{k=1}^{M} y_k^2$.
- Observe that the sufficient statistic is the sample variance (since $\mathbb{E}[y_k] = 0$) for $M \to +\infty$, under each hypothesis (and not the sample mean, as in the previous example).

$$S \triangleq \frac{1}{M} \sum_{k=1}^{M} y_k^2 \overset{H_1}{\geq} \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2} \left(\frac{1}{M} \ln(\tau) + \ln\left(\frac{\sigma_1}{\sigma_0}\right)\right). \tag{7}$$

Additional remarks:

1. Under H_j, for $\lim_{M\to+\infty} S = \sigma_j^2,\ j \in \{0,1\}$.
2. Using $1 - \frac{1}{x} \leq \ln(x) \leq x - 1$, it can be easily shown that $\sigma_0^2 \leq \frac{\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2} \ln\left(\frac{\sigma_1}{\sigma_0}\right)^2 \leq \sigma_1^2$.
3. For $M \to +\infty$, $\frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2} \left(\frac{1}{M} \ln(\tau) + \ln\left(\frac{\sigma_1}{\sigma_0}\right)\right) \to \frac{\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2} \ln\left(\frac{\sigma_1}{\sigma_0}\right)^2$.
4. Under hypothesis H_j, $\frac{M\ S}{\sigma_j^2} = \sum_{k=1}^{M} \left(\frac{y_k}{\sigma_j}\right)^2$ = sum of independent squared zero-mean Gaussians of unit variance: Under hypothesis H_j, $\frac{M\ S}{\sigma_j^2}$ corresponds to Chi-squared distribution with M degrees of freedom [will explain it subsequently].

Useful distributions

- $z = \sum_{i=1}^{M} z_i^2, z_i \sim \mathcal{N}(0,1)$ and $\{z_i\}$ independent, identically distributed (i.i.d.):
  - z distributed according to the Chi-squared distribution with M degrees of freedom and pdf as follows:

$$f_z(z) = \frac{1}{\Gamma(M/2)\, 2^{M/2}}\, z^{\left(\frac{M}{2}-1\right)}\, e^{-z/2}\, u(z), \tag{8}$$

    with $u(z)$ the step function (i.e., $u(z) = 1$ for $z \geq 0$ and zero otherwise) and $\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt$ Euler's gamma function, defined everywhere apart from non-positive integers (and $\Gamma(n) = (n-1)!$ for any positive integer n).
  - $\mathbb{E}[z] = M$, $\sigma_z^2 \stackrel{\triangle}{=}$ variance of $z = \mathrm{var}(z) = 2M$.
  - Special case - $M = 2$: exponential distribution, with p.d.f as follows (since $\Gamma(M = 2/2) = \Gamma(1) = 0! = 1$):

$$f_z(z) = \frac{1}{2} e^{-z/2}\, u(z). \tag{9}$$

Useful distributions

- In general, the pdf of a random variable according to the exponential distribution with parameter $\lambda > 0$ is given by:

$$f_z(z) = \lambda\, e^{-\lambda z}\, u(z), \tag{10}$$

with $\mathbb{E}[z] = 1/\lambda$ and $\sigma_z^2 \stackrel{\triangle}{=} \mathrm{var}(z) = 1/\lambda^2$.

- This is equivalent to $z = z_1^2 + z_2^2$, with z_1, z_2 independent and identically distributed according to $\mathcal{N}\left(0, \sigma^2\right)$ and $\mathbb{E}[z] = 1/\lambda = 2\sigma^2$.
  - For the special case of $y = \sqrt{z_1^2 + z_2^2}$, the Rayleigh pdf occurs:

$$f_y(y) = \frac{y}{\sigma^2}\, \exp\left(-\frac{y^2}{2\sigma^2}\right) u(y), \tag{11}$$

with $\mathbb{E}[y] = \sigma\sqrt{\pi/2}$ and $\sigma_y^2 = \mathrm{var}(y) = \frac{4-\pi}{2}\sigma^2$.

Useful distributions

- Set $z = \sum_{i=1}^{2M} z_i^2, z_i \sim \mathcal{N}(0,1)$ and $\{z_i\}$ i.i.d; as explained, z is distributed according to the Chi-squared distribution with $2M$ degrees of freedom.
- What is the distribution of $x = \theta\, z$ $(\theta > 0)$?
  - x can be viewed as the sum of M i.i.d. random variables distributed each according to the exponential distribution with parameter $\lambda = 1/(2\theta)$.
  - for any differentiable and invertible function $x = g(z)$, we do know that the new pdf can be found as follows:

$$f_x(x) = \frac{f_z(z)}{|g'(z)|}\Big|_{z=g^{-1}(x)} \tag{12}$$

and thus,

$$f_x(x) = \frac{f_z(z)}{\theta}\Big|_{z=x/\theta} = \frac{1}{(2\theta)^M\,\Gamma(M)}\, x^{(M-1)} \exp\left(-\frac{x}{2\theta}\right) u(x),$$

which corresponds to the Gamma distribution $(\Gamma(M, 2\theta))$, with parameters $M,\ 2\theta$ and $\mathbb{E}[x] = 2M\theta$, $\mathrm{var}(x) = 4M\theta^2$.

Useful distributions

- Set $z = \sum_{i=1}^{2M} z_i^2, z_i \sim \mathcal{N}(0,1)$ and $\{z_i\}$ i.i.d; as explained, z is distributed according to the Chi-squared distribution with $2M$ degrees of freedom.
- The pdf of $x = \theta\, z$ $(\theta > 0)$ is the pdf of the sum of M i.i.d. exponentials:

$$f_x(x) = \frac{1}{(2\theta)^M\,\Gamma(M)}\, x^{(M-1)} \exp\left(-\frac{x}{2\theta}\right) u(x),$$

  which corresponds, as shown, to the Gamma distribution $\Gamma(M, 2\theta)$, with parameters $M,\ 2\theta$ and $\mathbb{E}[x] = 2M\theta$, $\mathrm{var}(x) = 4M\theta^2$.
- Other distributions of the exponential family to remember:
  - (discrete) Poisson: $\Pr(n) = \frac{\lambda}{n!}\, e^{\lambda}$, $n \in \mathbb{N}$, $\mathbb{E}[n] = \lambda = \mathrm{var}(n)$.
  - (continuous) Laplace: $f_x(x;\mu,\beta) = \frac{1}{2\beta} \exp\left(-\frac{|x-\mu|}{\beta}\right)$, $\mathbb{E}[x] = \mu, \mathrm{var}(x) = 2\beta^2$.

Gaussian vectors (or jointly Gaussian random variables)

- Let $\mathbf{x} = [x_1 \ x_2 \ \ldots \ x_M]^{\mathrm{T}}$, where $x_1, \ x_2, \ \ldots, \ x_M$ are real. Vector $\mathbf{x}$ is Gaussian, or equivalently, $x_1 \ x_2 \ \ldots \ x_M$ are jointly Gaussian, if and only if the pdf of $\mathbf{x}$ (or the joint pdf of $x_1 \ x_2 \ \ldots \ x_M$) is given as follows:

  Covariance form:

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^M \ |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right\} \tag{13}$$

  denoted as $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$, with mean $\mathbf{m} \stackrel{\triangle}{=} \mathbb{E}[\mathbf{x}]$, covariance matrix $\boldsymbol{\Sigma} \stackrel{\triangle}{=} \mathbb{E}\left[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\right]$ and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

Gaussian vectors (or jointly Gaussian random variables)

- Let $\mathbf{x} = [x_1 \ x_2 \ \ldots \ x_M]^{\mathrm{T}}$, where $x_1, \ x_2, \ \ldots, \ x_M$ are real. Vector $\mathbf{x}$ is Gaussian, or equivalently, $x_1 \ x_2 \ \ldots \ x_M$ are jointly Gaussian, if and only if the pdf of $\mathbf{x}$ (or the joint pdf of $x_1 \ x_2 \ \ldots \ x_M$) is given as follows:

  Information form:
  $$f_{\mathbf{x}}(\mathbf{x}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{J}\mathbf{x} + \mathbf{h}^{\mathrm{T}}\mathbf{x}\right\}, \tag{14}$$

  denoted as $\mathbf{x} \sim \mathcal{N}^{-1}(\mathbf{h}, \mathbf{J})$, with potential vector $\mathbf{h} = \mathbf{J}\,\mathbf{m}$ and information (or precision) matrix $\mathbf{J} = \mathbf{\Sigma}^{-1}$.

Gaussian vector properties

- Let $\mathbf{x}$ (real) Gaussian vector. The following hold:
  - Moment generating function (MGF) $M_{\mathbf{x}}(j\mathbf{u})$:

$$M_{\mathbf{x}}(j\mathbf{u}) \triangleq \mathbb{E}\left[e^{j\mathbf{u}^{\mathrm{T}}\mathbf{x}}\right] = \exp\left\{j\mathbf{u}^{\mathrm{T}}\mathbf{m} - \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{u}\right\}. \tag{15}$$

  - All linear combinations of elements of $\mathbf{x}$ are scalar Gaussian random variables: $y = \mathbf{a}^{\mathrm{T}}\mathbf{x}$ is Gaussian for all deterministic $\mathbf{a}$.
  - There exists deterministic matrix $\mathbf{A}$, deterministic vector vector $\mathbf{b}$ and random vector $\mathbf{v}$ of i.i.d. $\mathcal{N}(0,1)$ entries, such that $\mathbf{x} = \mathbf{A}\mathbf{v} + \mathbf{b}$.
  - Affine transformation is also Gaussian, i.e., for any deterministic matrix $\mathbf{A}$ and deterministic vector $\mathbf{b}$, random vector $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ is Gaussian, according to $\mathcal{N}\left(\mathbf{A}\mathbf{m} + \mathbf{b}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^{\mathrm{T}}\right)$. [Simple proof from MGF]

Gaussian vector properties

- Let $\mathbf{x} = [\mathbf{y}^{\mathrm{T}}\ \mathbf{z}^{\mathrm{T}}]^{\mathrm{T}}$ (real) Gaussian vector, where $\mathbf{y}$, $\mathbf{z}$ are (real) vectors of appropriate dimensions. The following properties hold:
  - $\mathbf{y}$ is Gaussian.
  - $\mathbf{z}$ is Gaussian.
  - $\mathbf{y}$ given $\mathbf{z}$ is Gaussian.
  - $\mathbf{z}$ given $\mathbf{y}$ is Gaussian.
  - $\mathbb{E}[\mathbf{y}|\mathbf{z}]$ = affine transformation of $\mathbf{z} \Rightarrow$ Gaussian.
  - $\mathbb{E}[\mathbf{y}\ \mathbf{z}^{\mathrm{T}}] = \mathbb{E}[\mathbf{y}]\ \mathbb{E}[\mathbf{z}]^{\mathrm{T}} \Rightarrow \mathbf{y} \perp \mathbf{z}$, i.e., jointly Gaussian and uncorrelated results to independent!
- However, even if $\mathbf{y}$ is Gaussian and $\mathbf{z}$ is Gaussian, $\mathbf{x} = [\mathbf{y}^{\mathrm{T}}\ \mathbf{z}^{\mathrm{T}}]^{\mathrm{T}}$ may not be Gaussian. In other words, $\mathbf{y}$ and $\mathbf{z}$ may not be necessarily *jointly* Gaussian!

- Counterexample: let $x,\ y$ jointly Gaussian, zero mean, scalar random variables with joint pdf as follows:

$$p_{x,y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\ \exp\left\{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)\right\}, \tag{16}$$

which corresponds to the Gaussian vector $[x\ \ y]^{\mathrm{T}}$, with $\mathbf{m} = [0\ \ 0]^{\mathrm{T}}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$.
- Clearly $p_y(y) = \int_{-\infty}^{+\infty} p_{x,y}(x,y)dx = 2\int_0^{+\infty} p_{x,y}(x,y)dx$ corresponding to $\mathcal{N}(0,\sigma_y^2)$.
- Set the following non-Gaussian pdf:

$$\hat{p}_{x,y}(x,y) = \begin{cases} \frac{1}{\pi\sigma_x\sigma_y}\ \exp\left\{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)\right\}, & \text{if } x\ y > 0, \\ 0, & \text{if } x\ y < 0. \end{cases}$$

In this case, for $y > 0$, $\int_{-\infty}^{+\infty} \hat{p}_{x,y}(x,y)dx = \int_0^{+\infty} \hat{p}_{x,y}(x,y)dx = 2\ \int_0^{+\infty} \hat{p}_{x,y}(x,y)dx = p_y(y)$, i.e., Gaussian.

References

[1] Bernard C. Levy, Principles of Signal Detection and Parameter Estimation, Springer 2008.
[2] Class instructor notes.

Thank you!

Detection & Estimation Theory: Lecture 4

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- Probabilities that fully characterise a test: P_D vs P_F
- Neyman-Pearson Test
  - Derivation

Probability of Detection vs Probability of False Alarm

- There are two probability metrics for test $\delta(\mathbf{y})$ that fully characterise a binary hypothesis testing problem (as well as any binary classifier):
  - Probability of detection $\mathrm{P_D}$:

$$\mathrm{P_D}(\delta) \triangleq \int_{\mathcal{Y}_1} f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1)d\mathbf{y}. \tag{1}$$

    Complementary to the above, Probabilty of a miss is defined as $\mathrm{P_M}(\delta) \triangleq 1 - \mathrm{P_D}(\delta) = \int_{\mathcal{Y}_0} f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1)d\mathbf{y}$.
  - Probability of false alarm $\mathrm{P_F}$:

$$\mathrm{P_F}(\delta) \triangleq \int_{\mathcal{Y}_1} f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0)d\mathbf{y}. \tag{2}$$

- Notice that both are calculated over $\mathcal{Y}_1$, i.e., for space of measurements where decision is $\delta(\mathbf{y}) = 1$.
- Think of a radar system: false alarm is when the radar reports an airplane is coming, when it is not.

Probability of Detection vs Probability of False Alarm

- Remember the Bayes conditional risk:

  $$R(\delta(\mathbf{y})|\mathrm{H}_j) = C_{1j}\ \Pr(\delta(\mathbf{y}) = 1|\mathrm{H}_j) + C_{0j}\ \Pr(\delta(\mathbf{y}) = 0|\mathrm{H}_j)$$

  Thus,

  $$\begin{aligned} R(\delta(\mathbf{y})|\mathrm{H}_0) &= C_{10}\ \Pr(\delta(\mathbf{y}) = 1|\mathrm{H}_0) + C_{00}\ \Pr(\delta(\mathbf{y}) = 0|\mathrm{H}_0) \\ &= C_{10}\ \mathrm{P_F} + C_{00}\ (1 - \mathrm{P_F}). \end{aligned} \tag{3}$$

  and similarly,

  $$\begin{aligned} R(\delta(\mathbf{y})|\mathrm{H}_1) &= C_{11}\ \Pr(\delta(\mathbf{y}) = 1|\mathrm{H}_1) + C_{01}\ \Pr(\delta(\mathbf{y}) = 0|\mathrm{H}_1) \\ &= C_{11}\ \mathrm{P_D} + C_{01}\ (1 - \mathrm{P_D}). \end{aligned} \tag{4}$$

- Thus, the (unconditional) Bayes risk of test δ is fully characterised by the pair $(\mathrm{P_F}, \mathrm{P_D})$ of specific test δ:

  $$\begin{aligned} R(\delta) &= R(\delta(\mathbf{y})|\mathrm{H}_0)\,\pi_0 + R(\delta(\mathbf{y})|\mathrm{H}_1)\,\pi_1 \\ &= C_{00}\,\pi_0 + C_{01}\,\pi_1 + \pi_0\,(C_{10} - C_{00})\ \mathrm{P_F} + \pi_1\,(C_{11} - C_{01})\,\mathrm{P_D} \end{aligned} \tag{5}$$

Probability of Detection vs Probability of False Alarm

- The (unconditional) Bayes risk of test δ is fully characterised by the pair $(\mathrm{P_F}, \mathrm{P_D})$ of specific test δ:

$$
\begin{aligned}
R(\delta) &= R(\delta(\mathbf{y})|\mathrm{H}_0)\,\pi_0 + R(\delta(\mathbf{y})|\mathrm{H}_1)\,\pi_1 \\
&= C_{00}\,\pi_0 + C_{01}\,\pi_1 + \;\pi_0\,(C_{10} - C_{00})\,\mathrm{P_F} + \pi_1\,(C_{11} - C_{01})\,\mathrm{P_D}
\end{aligned} \tag{6}
$$

- Ideally, we would like to have a test (or a binary classifier) with $(\mathrm{P_F}, \mathrm{P_D}) = (0, 1)$. However, this is not feasible!
- Next lecture will offer the feasible pairs $(\mathrm{P_F}, \mathrm{P_D})$, as well as properties of the boundary between feasible and non-feasible pairs for all tests!
- Boundary between feasible and non-feasiible tests: receiver operating characteristic (ROC) [next lecture].

Neyman-Pearson Test

- Problem definition: among the tests that bound false alarm probability, find the test that maximises probability of detection. The formulation is given below:

$$\delta_{\mathrm{NP}} = \arg\max_{\delta \in D_\alpha} \mathrm{P_D}(\delta) \tag{7}$$

$$D_\alpha = \{\text{all tests } \delta : \mathrm{P_F}(\delta) \leq \alpha\}$$

- The problem could be also formulated with bounded probability of detection and minimised probability of false alarm.

Neyman-Pearson Test Derivation

- Constrained optimization problem: we need to set Lagrangian and KKT condition(s).
  - Lagrangian $L(\cdot,\cdot)$ for Lagrange multiplier $\lambda \geq 0$:

$$\begin{aligned} L(\delta,\lambda) &= \mathrm{P_D}(\delta) + \lambda\,(\alpha - \mathrm{P_F}(\delta)) \\ &= \lambda\,\alpha + \int_{\mathcal{Y}_1} \left[f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1) - \lambda\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) \right] d\mathbf{y} \end{aligned} \tag{8}$$

  - KKT condition ($\lambda \geq 0$):

$$\lambda\,(\alpha - \mathrm{P_F}(\delta)) = 0 \tag{9}$$

- Optimal test maximizes Lagrangian in Eq. (10) AND also satisfies KKT condition in Eq. (9).

Neyman-Pearson Test Derivation

- Maximization of Lagrangian $L(\delta,\lambda)(\lambda \geq 0)$:

$$L(\delta,\lambda) = \lambda\,\alpha + \int_{\mathcal{Y}_1} \left[f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1) - \lambda\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) \right] d\mathbf{y} \quad (10)$$

  - if $\mathbf{y} \in \mathcal{Y}_1$ then $\left[f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1) - \lambda\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) \right] > 0$, otherwise the Lagrangian is not maximized.
  - Equivalently, if $\mathbf{y} \in \mathcal{Y}_0$, then $\left[f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1) - \lambda\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) \right] < 0$ and that particular $\mathbf{y}$ is not taken into account in Eq. (10).
  - for any $\mathbf{y}$ with $\left[f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1) - \lambda\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) \right] = 0$, what decision should we adopt?
- Thus, maximization of the Lagrangian results to testing the sign of $\left[f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1) - \lambda\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) \right]$.

Neyman-Pearson Test Derivation

- Maximization of the Lagrangian results to testing the sign of $\left[f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1) - \lambda\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0)\right]$. Thus, optimal test for given $\lambda \geq 0$ follows:

$$\delta(\mathbf{y}) = \begin{cases} 1, & f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1) - \lambda\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) > 0 \\ 0, & f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1) - \lambda\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) < 0 \\ 0 \text{ or } 1 \text{ (TBD)}\ , & f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1) - \lambda\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) = 0. \end{cases} \tag{11}$$

- Setting the likelihood ratio $L(\mathbf{y}) \stackrel{\triangle}{=} f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1)/f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0)$, Eq. (11) is equivalent to:

$$\delta(\mathbf{y}) = \begin{cases} 1, & L(\mathbf{y}) > \lambda, \\ 0, & L(\mathbf{y}) < \lambda, \\ 0 \text{ or } 1 \text{ (TBD)}\ , & L(\mathbf{y}) = \lambda. \end{cases} \tag{12}$$

Neyman-Pearson Test Derivation

- Maximization of the Lagrangian results to the following optimal test for given $\lambda \geq 0$:

$$\delta(\mathbf{y}) = \begin{cases} 1, & L(\mathbf{y}) > \lambda, \\ 0, & L(\mathbf{y}) < \lambda, \\ 0 \text{ or } 1 \text{ (TBD)}, & L(\mathbf{y}) = \lambda. \end{cases} \tag{13}$$

- We need to find out λ and decision for $L(\mathbf{y}) = \lambda$.
- We define the following conditional cumulative distribution function (cdf):

$$F_{\mathrm{L}}\left(l|\mathrm{H}_0\right) \triangleq \Pr\left(L(\mathbf{y}) \leq l|\mathrm{H}_0\right). \tag{14}$$

  - As any cdf, the above should be:
    1. right-continuous,
    2. non-decreasing with increasing l,
    3. 0 for $l \to -\infty$ and
    4. 1 for $l \to +\infty$.

Neyman-Pearson Test Derivation

- Any cdf is "right-continuous": check figure below!

$$F_{\mathrm{L}}\left(l|\mathrm{H}_0\right) \triangleq \Pr\left(L(\mathbf{y}) \leq l|\mathrm{H}_0\right). \qquad (15)$$

- As any cdf, the above should be:
  1. right-continuous,
  2. non-decreasing with increasing l,
  3. 0 for $l \to -\infty$ and
  4. 1 for $l \to +\infty$.

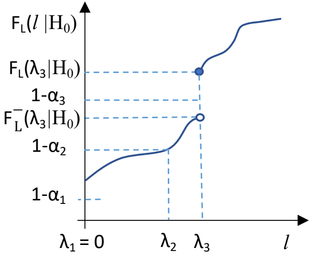
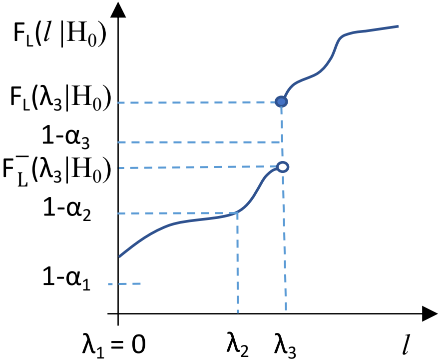


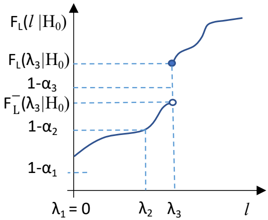
Figure 1: Example cdf with right-continuity.

Neyman-Pearson Test Derivation



- Three cases of likelihood ratio threshold λ occur, depending on value of $1 - \alpha$ vs $F_{\mathrm{L}}\left(l = 0|\mathrm{H}_0\right)$.
- Reminder: α is the upper bound of $\mathrm{P_F}$.

Neyman-Pearson Test Derivation

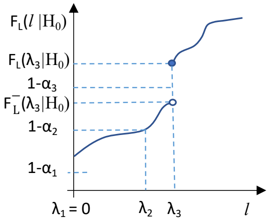


- Case I: $1 - \alpha < F_L\,(l = 0|H_0) \overset{\triangle}{=} f_0 \Leftrightarrow 1 - f_0 < \alpha$:
  - set $\lambda = 0$ and $\delta(\mathbf{y}) = 0$ for $L(\mathbf{y}) = \lambda = 0$, i.e.,

$$\delta(\mathbf{y}) = \begin{cases} 1, & L(\mathbf{y}) > \lambda = 0, \\ 0, & L(\mathbf{y}) \leq \lambda = 0. \end{cases} \tag{16}$$

  - KTT $\lambda\,(\alpha - P_F) = 0$ is satisfied for $\lambda = 0$.
  - $P_F = 1 - \Pr\,(\delta(\mathbf{y}) = 0|H_0) = 1 - \Pr\,(L(\mathbf{y}) \leq 0|H_0) = 1 - f_0 < \alpha \Rightarrow$ probability of false alarm is bounded.

Neyman-Pearson Test Derivation

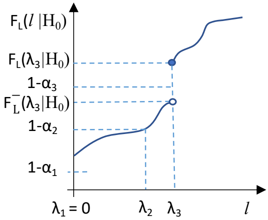


- Case II: $1-\alpha \geq F_{\mathrm{L}}(l=0|\mathrm{H}_0) \stackrel{\triangle}{=} f_0$ and λ^* is in the range of $F_{\mathrm{L}}(l|\mathrm{H}_0)$, i.e., there is λ^* such that $F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0) = 1-\alpha$.
  - set $\lambda = \lambda^*$ and $\delta(\mathbf{y}) = 0$ for $L(\mathbf{y}) = \lambda^*$, i.e.,

$$\delta(\mathbf{y}) = \begin{cases} 1, & L(\mathbf{y}) > \lambda^*, \\ 0, & L(\mathbf{y}) \leq \lambda^*. \end{cases} \tag{17}$$

  - $\mathrm{P_F} = 1 - \Pr(\delta(\mathbf{y}) = 0|\mathrm{H}_0) = 1 - \Pr(L(\mathbf{y}) \leq \lambda^*|\mathrm{H}_0) = 1-(1-\alpha) = \alpha$.
  - KTT $\lambda(\alpha - \mathrm{P_F}) = 0$ is satisfied for $\lambda = \lambda^*$.

Neyman-Pearson Test Derivation

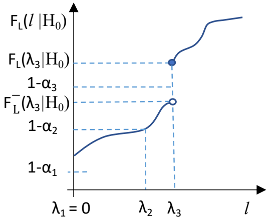


- Case III: $1-\alpha \geq F_{\mathrm{L}}\left(l=0|\mathrm{H}_0\right) \triangleq f_0$ and λ^* is NOT in the range of $F_{\mathrm{L}}\left(l|\mathrm{H}_0\right)$, i.e., $F_{\mathrm{L}}^-\left(\lambda^*|\mathrm{H}_0\right) < 1-\alpha < F_{\mathrm{L}}\left(\lambda^*|\mathrm{H}_0\right)$.
  - set $\lambda = \lambda^*$ and $\delta(\mathbf{y}) = 0$ for $L(\mathbf{y}) = \lambda$, i.e.,

$$\delta(\mathbf{y}) \triangleq \delta_{L,\lambda^*}(\mathbf{y}) = \begin{cases} 1, & L(\mathbf{y}) > \lambda^*, \\ 0, & L(\mathbf{y}) \leq \lambda^*. \end{cases} \tag{18}$$

  - $\mathrm{P_F} = 1 - \Pr\left(\delta(\mathbf{y}) = 0|\mathrm{H}_0\right) = 1 - \Pr\left(L(\mathbf{y}) \leq \lambda^*|\mathrm{H}_0\right) = 1 - F_{\mathrm{L}}\left(\lambda^*|\mathrm{H}_0\right) < \alpha$.
  - KTT $\lambda\left(\alpha - \mathrm{P_F}\right) = 0$ is NOT satisfied!

Neyman-Pearson Test Derivation

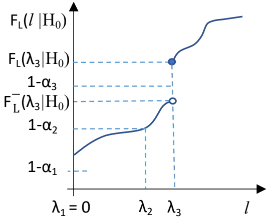


- Case III: $1-\alpha \geq F_L\left(l=0|\mathrm{H}_0\right) \triangleq f_0$ and λ^* is NOT in the range of $F_L\left(l|\mathrm{H}_0\right)$, i.e., $F_L^-\left(\lambda^*|\mathrm{H}_0\right) < 1-\alpha < F_L\left(\lambda^*|\mathrm{H}_0\right)$.
  - set $\lambda = \lambda^*$ and $\delta(\mathbf{y}) = 1$ for $L(\mathbf{y}) = \lambda$, i.e.,

$$\delta(\mathbf{y}) \triangleq \delta_{U,\lambda^*}(\mathbf{y}) = \begin{cases} 1, & L(\mathbf{y}) \geq \lambda^*, \\ 0, & L(\mathbf{y}) < \lambda^*. \end{cases} \tag{19}$$

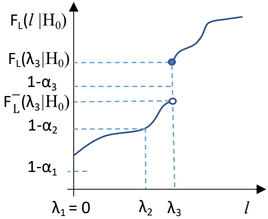
  - $\mathrm{P_F} = 1 - \Pr\left(\delta(\mathbf{y}) = 0|\mathrm{H}_0\right) = 1 - \Pr\left(L(\mathbf{y}) < \lambda^*|\mathrm{H}_0\right) = 1 - F_L^-\left(\lambda^*|\mathrm{H}_0\right) > \alpha$.
  - KTT $\lambda\left(\alpha - \mathrm{P_F}\right) = 0$ is NOT satisfied!

Neyman-Pearson Test Derivation



- Case III: $1-\alpha \geq F_{\mathrm{L}}\left(l=0|\mathrm{H}_0\right) \stackrel{\triangle}{=} f_0$ and λ^* is NOT in the range of $F_{\mathrm{L}}\left(l|\mathrm{H}_0\right)$, i.e., $F_{\mathrm{L}}^{-}\left(\lambda^*|\mathrm{H}_0\right) < 1-\alpha < F_{\mathrm{L}}\left(\lambda^*|\mathrm{H}_0\right)$. So far:
  - Test $\delta_{L,\lambda^*}(\mathbf{y})$ with $\mathrm{P}_{\mathrm{F}}\left(\delta_{L,\lambda^*}\right) < \alpha$.
  - Test $\delta_{U,\lambda^*}(\mathbf{y})$ with $\mathrm{P}_{\mathrm{F}}\left(\delta_{U,\lambda^*}\right) > \alpha$.
  - KTT $\lambda\left(\alpha-\mathrm{P}_{\mathrm{F}}\right)=0$ requires exactly $\mathrm{P}_{\mathrm{F}}=\alpha$.
- Solution?

Neyman-Pearson Test Derivation

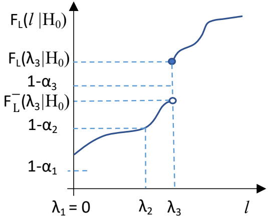


- Case III: $1 - \alpha \geq F_{\mathrm{L}}\left(l = 0|\mathrm{H}_0\right) \stackrel{\triangle}{=} f_0$ and λ^* is NOT in the range of $F_{\mathrm{L}}\left(l|\mathrm{H}_0\right)$, i.e., $F_{\mathrm{L}}^{-}\left(\lambda^*|\mathrm{H}_0\right) < 1 - \alpha < F_{\mathrm{L}}\left(\lambda^*|\mathrm{H}_0\right)$.
  - Solution: set $\lambda = \lambda^*$ and randomize decision for $L(\mathbf{y}) = \lambda$:

$$\delta(\mathbf{y}) = \begin{cases} \delta_{U,\lambda^*}(\mathbf{y}), & \text{with probability } \rho, \\ \delta_{L,\lambda^*}(\mathbf{y}), & \text{with probability } 1-\rho, \end{cases} \tag{20}$$

  - Set $0 < \rho < 1$ such that $\mathrm{P_F} \equiv \alpha$ (and KKT is thus satisfied).

Neyman-Pearson Test Derivation



- Case III: $1-\alpha \geq F_{\mathrm{L}}(l=0|\mathrm{H}_0) \triangleq f_0$ and λ^* is NOT in the range of $F_{\mathrm{L}}(l|\mathrm{H}_0)$, i.e., $F_{\mathrm{L}}^-(\lambda^*|\mathrm{H}_0) < 1-\alpha < F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0)$.
  - Solution: set $\lambda = \lambda^*$ and randomize decision for $L(\mathbf{y}) = \lambda$:

$$\delta(\mathbf{y}) = \begin{cases} \delta_{U,\lambda^*}(\mathbf{y}), & \text{with probability } \rho, \\ \delta_{L,\lambda^*}(\mathbf{y}), & \text{with probability } 1-\rho, \end{cases} \tag{21}$$

$$\rho = \frac{F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0) - (1-\alpha)}{F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0) - F_{\mathrm{L}}^-(\lambda^*|\mathrm{H}_0)}, \; 0 < \rho < 1. \tag{22}$$

  - Such $0 < \rho < 1$ guarantees $\mathrm{P_F} \equiv \alpha$.

Neyman-Pearson Test Derivation

- Case III: $1-\alpha \geq F_{\mathrm{L}}(l=0|\mathrm{H}_0) \stackrel{\triangle}{=} f_0$ and λ^* is NOT in the range of $F_{\mathrm{L}}(l|\mathrm{H}_0)$, i.e., $F_{\mathrm{L}}^{-}(\lambda^*|\mathrm{H}_0) < 1-\alpha < F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0)$.
  - Solution: set $\lambda = \lambda^*$ and randomize decision for $L(\mathbf{y}) = \lambda$:

$$\delta(\mathbf{y}) = \begin{cases} \delta_{U,\lambda^*}(\mathbf{y}), & \text{with probability } \rho, \\ \delta_{L,\lambda^*}(\mathbf{y}), & \text{with probability } 1-\rho, \end{cases} \tag{23}$$

$$\rho = \frac{F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0) - (1-\alpha)}{F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0) - F_{\mathrm{L}}^{-}(\lambda^*|\mathrm{H}_0)},\ 0 < \rho < 1. \tag{24}$$

  - Such ρ guarantees $\mathrm{P_F} \equiv \alpha$.
    Proof:

$$\begin{aligned} \mathrm{P_F} &= \rho\,\mathrm{P_F}(\delta_{U,\lambda^*}) + (1-\rho)\,\mathrm{P_F}(\delta_{L,\lambda^*}) && (25) \\ &= \rho\left(1 - F_{\mathrm{L}}^{-}(\lambda^*|\mathrm{H}_0)\right) + (1-\rho)\,(1 - F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0)) && (26) \\ &= 1 - F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0) + \rho\left(F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0) - F_{\mathrm{L}}^{-}(\lambda^*|\mathrm{H}_0)\right) && (27) \\ &= 1 - F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0) + F_{\mathrm{L}}(\lambda^*|\mathrm{H}_0) - (1-\alpha) && (28) \\ &= \alpha && (29) \end{aligned}$$

Neyman-Pearson Test

- Neyman-Pearson-optimal detector is likelihood ratio test (LRT)!
- As already mentioned, we could have minimized $\mathrm{P_F}$ subject to bounded $\mathrm{P_D}$.
- Next lecture: feasible points $(\mathrm{P_F}, \mathrm{P_D})$ for any test!

References

[1] Bernard C. Levy, Principles of Signal Detection and Parameter Estimation, Springer 2008.
[2] Class instructor notes.

Thank you!

Detection & Estimation Theory: Lectures 5 & 6

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- Receiver Operating Characteristic (ROC)
- ROC Properties
- Remarks
- Examples

ROC definition

- Remember the two probability metrics that fully characterise a test $\delta(\mathbf{y})$ (as well as any binary classifier):
  - Prob. of detection $\mathrm{P_D}$ and prob. of false alarm $\mathrm{P_F}$:

$$\mathrm{P_D}(\delta) \triangleq \int_{\mathcal{Y}_1} f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1)d\mathbf{y}, \mathrm{P_F}(\delta) \triangleq \int_{\mathcal{Y}_1} f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0)d\mathbf{y}.$$

- The Bayes risk of test δ is fully characterised by the pair $(\mathrm{P_F}, \mathrm{P_D})$ of specific test δ (previous lecture).
- Which pairs $(\mathrm{P_F}, \mathrm{P_D})$ are feasible?
- Boundary between feasible and non-feasiible tests = receiver operating characteristic (ROC).
- ROC properties?

ROC definition

- Remember the two probability metrics that fully characterise a test $\delta(\mathbf{y})$ (as well as any binary classifier):
  - Prob. of detection $\mathrm{P_D}$ and prob. of false alarm $\mathrm{P_F}$:

$$\mathrm{P_D}(\delta) \triangleq \int_{\mathcal{Y}_1} f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1)d\mathbf{y}, \quad \mathrm{P_F}(\delta) \triangleq \int_{\mathcal{Y}_1} f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0)d\mathbf{y}.$$

- Define likelihood ratio and conditional pdf of likelihood ratio $f_{L|\mathrm{H}_j}(l|\mathrm{H}_j)$:

$$L(\mathbf{y}) \triangleq \frac{f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1)}{f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0)}. \tag{1}$$

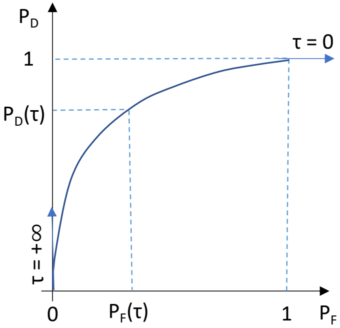
- For likelihood ratio test $L(\mathbf{y}) \overset{H_1}{\geq} \tau$, $\mathrm{P_F}, \mathrm{P_D}$ can be redefined:

$$\mathrm{P_D}(\tau) \triangleq \int_{\tau}^{+\infty} f_{L|\mathrm{H}_1}(l|\mathrm{H}_1)\,dl, \quad \mathrm{P_F}(\tau) \triangleq \int_{\tau}^{+\infty} f_{L|\mathrm{H}_0}(l|\mathrm{H}_0)\,dl.$$

ROC Property 1

- Points $(0, 0)$ and $(1, 1)$ of $(\mathrm{P_F},\ \mathrm{P_D})$ belong to ROC.
  Proof:
  - Always select H_1 (or equivalently, set $\tau = 0$):
  $$\mathrm{P_F}(\tau = 0) = 1,\ \mathrm{P_D}(\tau = 0) = 1. \tag{2}$$
  - Always select H_0 (or equivalently, set $\tau = +\infty$):
  $$\mathrm{P_F}(\tau = +\infty) = 0,\ \mathrm{P_D}(\tau = +\infty) = 0. \tag{3}$$

ROC Property 2



- Slope of ROC at $(\mathrm{P_F}(\tau), \mathrm{P_D}(\tau))$ is equal to τ, i.e., $\frac{d\mathrm{P_D}(\tau)}{d\mathrm{P_F}(\tau)} = \tau$.
  Proof:
  - $\mathrm{P_D}(\delta) = \int_{\tau}^{+\infty} f_{L|\mathrm{H}_1}(l|\mathrm{H}_1)\, d\mathbf{y} = 1 - \int_{-\infty}^{+\tau} f_{L|\mathrm{H}_1}(l|\mathrm{H}_1)\, d\mathbf{y} \Rightarrow$

ROC Property 2

- $\mathrm{P_D}(\tau) = \int_{\tau}^{+\infty} f_{L|\mathrm{H}_1}(l|\mathrm{H}_1)\,dl = 1 - \int_{-\infty}^{+\tau} f_{L|\mathrm{H}_1}(l|\mathrm{H}_1)\,dl \Rightarrow$

$$\frac{d\mathrm{P_D}}{d\tau}(\tau) = -f_{L|\mathrm{H}_1}(\tau|\mathrm{H}_1), \tag{4}$$

$$\text{similarly, } \frac{d\mathrm{P_F}}{d\tau}(\tau) = -f_{L|\mathrm{H}_0}(\tau|\mathrm{H}_0). \tag{5}$$

From Eqs. (4), (5):

$$\frac{d\mathrm{P_D}}{d\mathrm{P_F}}(\tau) = \frac{f_{L|\mathrm{H}_1}(\tau|\mathrm{H}_1)}{f_{L|\mathrm{H}_0}(\tau|\mathrm{H}_0)}. \tag{6}$$

- Recall that:

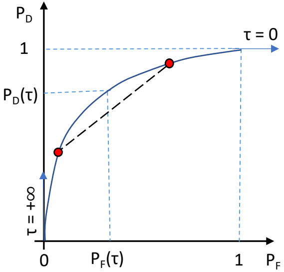
$$\mathrm{P_D}(\tau) = \int_{\tau}^{+\infty} f_{L|\mathrm{H}_1}(l|\mathrm{H}_1)\,dl = \int_{\mathcal{Y}_1} f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1)d\mathbf{y} = \int_{\mathcal{Y}_1} \frac{f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1)}{f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0)} f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0)d\mathbf{y}$$

$$= \int_{\mathcal{Y}_1=\{\mathbf{y}:L(\mathbf{y})\geq\tau\}} L(\mathbf{y})\, f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0)\,d\mathbf{y} = \int_{\tau}^{+\infty} l\cdot f_{L|\mathrm{H}_0}(l|\mathrm{H}_0)\,dl \Rightarrow \tag{7}$$

$$\frac{d\mathrm{P_D}}{d\tau}(\tau) = -\tau\cdot f_{L|\mathrm{H}_0}(\tau|\mathrm{H}_0) \overset{(4)}{\Rightarrow} f_{L|\mathrm{H}_1}(\tau|\mathrm{H}_1) = \tau\, f_{L|\mathrm{H}_0}(\tau|\mathrm{H}_0). \tag{8}$$

From Eqs. (6), (8), the proof is completed.

ROC Property 3



- The domain of feasible points $(\mathrm{P_F}, \mathrm{P_D})$ is convex.
  Proof:
  - We need to show that for any two feasible points $(\mathrm{P_{F1}}, \mathrm{P_{D1}})$, $(\mathrm{P_{F2}}, \mathrm{P_{D2}})$, the line connecting them is also included in the feasible points.
  - Such line is described by $\rho \in [0, 1]$:

$$\mathrm{P_F}(\rho) = \rho\,\mathrm{P_{F1}} + (1-\rho)\,\mathrm{P_{F2}}, \tag{9}$$

$$\mathrm{P_D}(\rho) = \rho\,\mathrm{P_{D1}} + (1-\rho)\,\mathrm{P_{D2}}. \tag{10}$$

ROC Property 3

- Such line is described by $\rho \in [0,1]$:

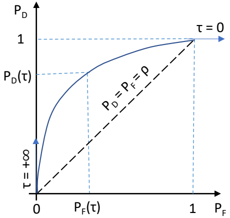
$$\mathrm{P_F}(\rho) = \rho\,\mathrm{P_{F1}} + (1-\rho)\,\mathrm{P_{F2}}, \tag{11}$$
$$\mathrm{P_D}(\rho) = \rho\,\mathrm{P_{D1}} + (1-\rho)\,\mathrm{P_{D2}}. \tag{12}$$

- Why? solve each of the above for ρ and equate: you will find out the line equation connecting the two points.
- Define the following randomized test that selects probabilistic between two tests:

$$\delta(\mathbf{y}) = \begin{cases} \delta_1(\mathbf{y}), & \text{with probability } \rho, \\ \delta_2(\mathbf{y}), & \text{with probability } 1-\rho, \end{cases} \tag{13}$$

  where $\delta_i(\mathbf{y})$ is the feasible test with $\mathrm{P_{F}}_i, \mathrm{P_{D}}_i$, $i \in \{1,2\}$.
- The test above achieves $\mathrm{P_F}(\rho)$ and $\mathrm{P_D}(\rho)$ given in Eqs. (11), (12).

ROC Property 3 Remarks

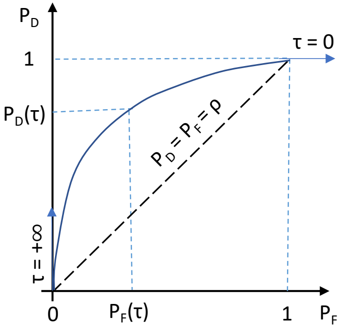


- Define the following randomized test that selects probabilistic between two hypothesis, *independently* of the measurements $\mathbf{y}$:

$$\delta(\mathbf{y}) = \begin{cases} \mathrm{H}_1, & \text{with probability } \rho, \\ \mathrm{H}_0, & \text{with probability } 1-\rho, \end{cases} \tag{14}$$

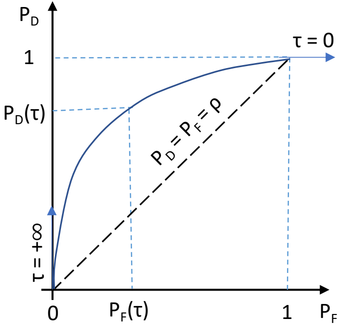
- The test above achieves $\mathrm{P_D}(\rho) = \Pr(\mathcal{Y}_1 | H_1) = \Pr(\mathcal{Y}_1) = \rho$ and $\mathrm{P_F}(\rho) = \Pr(\mathcal{Y}_1 | H_0) = \Pr(\mathcal{Y}_1) = \rho$.
- Thus, line $\mathrm{P_D} = \mathrm{P_F} = \rho$ belongs to the feasible points.

ROC Property 3 Remarks



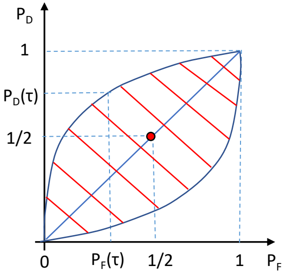
- The domain of feasible points $(\mathrm{P_F}, \mathrm{P_D})$ is convex.
- Line $\mathrm{P_D} = \mathrm{P_F} = \rho$ belongs to the feasible points.
  - ...thus, domain of feasible points are located *below* the ROC curve!
- Domain of feasible points is convex and located below ROC.
  - ...thus, ROC curve is concave!

ROC Property 4



- For tests on the ROC curve, $\mathrm{P_F} \leq \mathrm{P_D}$.
  Proof:
  - ROC curve is concave.
  - $(0, 0)$ and $(1, 1)$ belong to the ROC curve.
    - ...thus, $\mathrm{P_F} \leq \mathrm{P_D}$.

Remarks

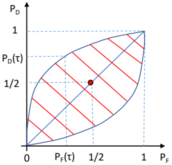


- What about "bad" tests? Are all points with $\mathrm{P_D} \leq \mathrm{P_F}$ feasible? [NO!]
- Assume a "bad" test δ with specific $(\mathrm{P_F}(\delta), \mathrm{P_D}(\delta))$.
- Define the following test $\hat{\delta}$ that flips the decision:

$$\hat{\delta}(\mathbf{y}) = 1 - \delta(\mathbf{y}). \tag{15}$$

- $\hat{\delta}(\mathbf{y})$ achieves $(\widehat{\mathrm{P_F}}, \widehat{\mathrm{P_D}}) = (1 - \mathrm{P_F}(\delta), 1 - \mathrm{P_D}(\delta))$, i.e., region of feasible tests is symmetric around $(1/2, 1/2)$.

Remarks



- Define the following test $\hat{\delta}$ that flips the decision:

$$\hat{\delta}(\mathbf{y}) = 1 - \delta(\mathbf{y}). \tag{16}$$

- $\hat{\delta}(\mathbf{y})$ achieves $(\widehat{\mathrm{P}_\mathrm{F}}, \widehat{\mathrm{P}_\mathrm{D}}) = (1 - \mathrm{P}_\mathrm{F}(\delta), 1 - \mathrm{P}_\mathrm{D}(\delta))$, i.e., region of feasible tests is symmetric around $(1/2, 1/2)$. Proof:

$$\mathrm{P}_\mathrm{F}(\delta) = \int_{\mathcal{Y}_1} f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) d\mathbf{y} \Rightarrow \tag{17}$$

$$\mathrm{P}_\mathrm{F}(\hat{\delta} = 1 - \delta) = \int_{\mathcal{Y}_0} f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0) d\mathbf{y} = 1 - \mathrm{P}_\mathrm{F}(\delta). \tag{18}$$

...and similarly for $\mathrm{P}_\mathrm{D}(\hat{\delta})$.

Examples

- Assume binary hypothesis testing, with $\mathbf{y} \sim \mathcal{N}(\mathbf{m}_j, \mathbf{K}_j)$ under hypothesis H_j, $j \in \{0, 1\}$ and $\mathbf{y} \in \mathbb{R}^{N \times 1}$.

Preliminaries:

1. For any matrix $\mathbf{K}$ with inverse, $\left(\mathbf{K}^{-1}\right)^T = \left(\mathbf{K}^T\right)^{-1}$.
2. Thus, for any *symmetric* matrix $\mathbf{K}$ with inverse, the inverse is also symmetric: $\left(\mathbf{K}^{-1}\right)^T = \mathbf{K}^{-1}$.
3. For any scalar z with $z = \mathbf{a}^{\mathrm{T}}\mathbf{b}$, $z = \mathbf{b}^{\mathrm{T}}\mathbf{a}$, where $\mathbf{a}, \mathbf{b}$ vectors of the same dimension.

$$f_{\mathbf{y}|\mathrm{H}_j}(\mathbf{y}|\mathrm{H}_j) = \frac{1}{\sqrt{(2\pi)^N |\mathbf{K}_j|}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{m}_j)^T \mathbf{K}_j^{-1} (\mathbf{y} - \mathbf{m}_j)\right) \Leftrightarrow \tag{19}$$

$$\frac{f_{\mathbf{y}|\mathrm{H}_1}(\mathbf{y}|\mathrm{H}_1)}{f_{\mathbf{y}|\mathrm{H}_0}(\mathbf{y}|\mathrm{H}_0)} \overset{\mathrm{H}_1}{\geq} \tau \Leftrightarrow \tag{20}$$

$$\frac{1}{2}(\mathbf{y} - \mathbf{m}_0)^T \mathbf{K}_0^{-1} (\mathbf{y} - \mathbf{m}_0) - \frac{1}{2}(\mathbf{y} - \mathbf{m}_1)^T \mathbf{K}_1^{-1} (\mathbf{y} - \mathbf{m}_1) + \frac{1}{2}\ln\left(\frac{|\mathbf{K}_0|}{|\mathbf{K}_1|}\right) \overset{\mathrm{H}_1}{\geq} \ln(\tau) \Leftrightarrow \tag{21}$$

...simple calculations exploiting 2. and 3. above [try them!]...

$$\Leftrightarrow \underbrace{\frac{1}{2}\mathbf{y}^T \left(\mathbf{K}_0^{-1} - \mathbf{K}_1^{-1}\right)\mathbf{y} + \mathbf{y}^T \left(\mathbf{K}_1^{-1}\mathbf{m}_1 - \mathbf{K}_0^{-1}\mathbf{m}_0\right)}_{S(\mathbf{y})} \overset{\mathrm{H}_1}{\geq} \eta, \tag{22}$$

$$\eta = \frac{1}{2}\mathbf{m}_1^T \mathbf{K}_1 \mathbf{m}_1 - \frac{1}{2}\mathbf{m}_0^T \mathbf{K}_0 \mathbf{m}_0 - \frac{1}{2}\ln\left(\frac{|\mathbf{K}_0|}{|\mathbf{K}_1|}\right) + \ln(\tau) \tag{23}$$

Examples

- Assume $\mathbf{K}_0 = \mathbf{K}_1 = \mathbf{K}$. In that case, the test is simplified as follows:

$$\mathbf{y}^T\mathbf{K}^{-1}\underbrace{(\mathbf{m}_1 - \mathbf{m}_0)}_{\Delta\mathbf{m}} = \mathbf{y}^T\mathbf{K}^{-1}\Delta\mathbf{m} \overset{\mathrm{H}_1}{\geq} \eta \Leftrightarrow \tag{24}$$

$$S_s(\mathbf{y}) \triangleq \mathbf{y}^T\mathbf{K}^{-1}\Delta\mathbf{m} - \mathbf{m}_0^T\mathbf{K}^{-1}\Delta\mathbf{m} \overset{\mathrm{H}_1}{\geq} \eta - \mathbf{m}_0^T\mathbf{K}^{-1}\Delta\mathbf{m} \triangleq \eta_s \Leftrightarrow \tag{25}$$

$$S_s(\mathbf{y}) = \Delta\mathbf{m}^T\mathbf{K}^{-1}\mathbf{y} - \Delta\mathbf{m}^T\mathbf{K}^{-1}\mathbf{m}_0 \overset{\mathrm{H}_1}{\geq} \eta - \mathbf{m}_0^T\mathbf{K}^{-1}\Delta\mathbf{m} \triangleq \eta_s, \tag{26}$$

where we have used the fact that $\mathbf{K}^{-1}$ is symmetric; $S_s(\mathbf{y})$ is the shifted sufficient statistic, which is affine transformation of a Gaussian vector, and thus, it is also Gaussian:

$$\mathrm{H}_0 : S_s(\mathbf{y}) \sim \mathcal{N}\left(0, \Delta\mathbf{m}^T\mathbf{K}^{-1}\Delta\mathbf{m}\right) \tag{27}$$

$$\mathrm{H}_1 : S_s(\mathbf{y}) \sim \mathcal{N}\left(\Delta\mathbf{m}^T\mathbf{K}^{-1}\Delta\mathbf{m}, \Delta\mathbf{m}^T\mathbf{K}^{-1}\Delta\mathbf{m}\right) \tag{28}$$

Notice that $\mathbf{K}^{-1}$ (and $\mathbf{K}$) are positive definite, and thus, $\Delta\mathbf{m}^T\mathbf{K}^{-1}\Delta\mathbf{m} > 0$. We set $d^2 \triangleq \Delta\mathbf{m}^T\mathbf{K}^{-1}\Delta\mathbf{m}$. We also need the following definition of the (Gauss) Q-function $Q(x)$ and its properties:

$$Q(x) \triangleq \int_x^{+\infty} \frac{1}{2\pi} e^{-t^2/2} dt = 1 - \int_{-\infty}^{x} \frac{1}{2\pi} e^{-t^2/2} dt, \tag{29}$$

$$Q(-x) = 1 - Q(x), \ \ \frac{dQ(x)}{dx} = -\frac{1}{2\pi} e^{-x^2/2}. \tag{30}$$

$$\mathrm{H}_0 : S_s(\mathbf{y}) \sim \mathcal{N}\left(0, \Delta\mathbf{m}^T\mathbf{K}^{-1}\Delta\mathbf{m}\right) \equiv \mathcal{N}\left(0, d^2\right) \tag{31}$$

$$\mathrm{H}_1 : S_s(\mathbf{y}) \sim \mathcal{N}\left(\Delta\mathbf{m}^T\mathbf{K}^{-1}\Delta\mathbf{m}, \Delta\mathbf{m}^T\mathbf{K}^{-1}\Delta\mathbf{m}\right) \equiv \mathcal{N}\left(d^2, d^2\right) \tag{32}$$

We can now calculate the basic probabilities for the test $S_s(\mathbf{y}) \overset{\mathrm{H}_1}{\geq} \eta_s$:

$$\mathrm{P_D} = \int_{\eta_s}^{+\infty} f_{S_s(\mathbf{y})|\mathrm{H}_1}\left(s|\mathrm{H}_1\right)\, ds = \int_{\eta_s}^{+\infty} \frac{1}{\sqrt{2\pi d^2}} \exp\left(-\frac{\left(s-d^2\right)^2}{2d^2}\right) ds \tag{33}$$

$$\overset{\frac{s-d^2}{d}=t}{=} \int_{\frac{\eta_s-d^2}{d}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = Q\left(\frac{\eta_s-d^2}{d}\right) = 1 - Q\left(d - \frac{\eta_s}{d}\right), \tag{34}$$

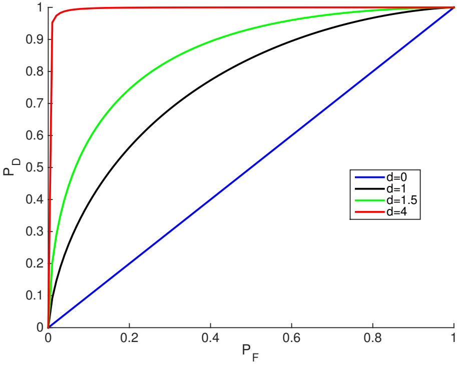
$$\mathrm{P_F} = \int_{\eta_s}^{+\infty} f_{S_s(\mathbf{y})|\mathrm{H}_0}\left(s|\mathrm{H}_0\right)\, ds = \int_{\eta_s}^{+\infty} \frac{1}{\sqrt{2\pi d^2}} \exp\left(-\frac{s^2}{2d^2}\right) ds \tag{35}$$

$$\overset{\frac{s}{d}=t}{=} \int_{\frac{\eta_s}{d}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = Q\left(\frac{\eta_s}{d}\right) \Rightarrow \frac{\eta_s}{d} = Q^{-1}\left(\mathrm{P_F}\right). \tag{36}$$

$$(34), (36) \Rightarrow \mathrm{P_D} = 1 - Q\left(d - Q^{-1}\left(\mathrm{P_F}\right)\right). \tag{37}$$

$$\mathrm{P_D} = 1 - Q\left(d - Q^{-1}\left(\mathrm{P_F}\right)\right). \tag{38}$$

- We need d as large as possible! Why?



Whitening Procedure

- In many cases, it is useful to simplify the observation model with linear transformations. To this end, the eigendecomposition of positive-definite matrix $\mathbf{K}$ is exploited:

$$\mathbf{K}\,\mathbf{P} = \mathbf{P}\boldsymbol{\Lambda} \Leftrightarrow \mathbf{K} \begin{bmatrix} | & | & \cdots & | \\ \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_N \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} | & | & \cdots & | \\ \lambda_1 \mathbf{p}_1 & \lambda_2 \mathbf{p}_2 & \cdots & \lambda_N \mathbf{p}_N \\ | & | & \cdots & | \end{bmatrix} \tag{39}$$

$$= \begin{bmatrix} | & | & \cdots & | \\ \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_N \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix} \Leftrightarrow \mathbf{K} = \mathbf{P}\,\boldsymbol{\Lambda}\,\mathbf{P}^{\mathrm{T}} \tag{40}$$

  with $\mathbf{K}\,\mathbf{p}_i = \lambda_i\,\mathbf{p}_i,\ i \in \{1, 2, \ldots, N\}$, i.e., columns of $\mathbf{P}$ are the eigenvectors of $\mathbf{K}$ and $\{\lambda_i\}$ the corresponding eigenvalues, $\mathbf{P}$ is orthogonal, i.e., $\mathbf{P}\,\mathbf{P}^{\mathrm{T}} = \mathbf{P}^{\mathrm{T}}\,\mathbf{P} = \mathbf{I}_N$ and $\boldsymbol{\Lambda}$ diagonal matrix, with main diagonal the positive eigenvalues of $\mathbf{K}$, i.e., $\boldsymbol{\Lambda} = \mathrm{diag}\,[\lambda_1\,\lambda_2 \ldots \lambda_N]$.

- Set $\boldsymbol{\Lambda}^{-1/2} = \mathrm{diag}\left[\sqrt{\lambda_1}\,\sqrt{\lambda_2}\ldots\sqrt{\lambda_N}\right]$ and multiply $\mathbf{y}$ by $\boldsymbol{\Lambda}^{-1/2}\,\mathbf{P}^{\mathrm{T}}$:

Whitening Procedure

- Multiply $\mathbf{y}$ by $\boldsymbol{\Lambda}^{-1/2}\,\mathbf{P}^{\mathrm{T}}$:

$$\underbrace{\boldsymbol{\Lambda}^{-1/2}\,\mathbf{P}^{\mathrm{T}}\,\mathbf{y}}_{\mathbf{y}_w} = \underbrace{\boldsymbol{\Lambda}^{-1/2}\,\mathbf{P}^{\mathrm{T}}\,\mathbf{m}_j}_{\boldsymbol{\mu}_j} + \underbrace{\boldsymbol{\Lambda}^{-1/2}\,\mathbf{P}^{\mathrm{T}}\,\mathbf{v}}_{\mathbf{v}_w}, \tag{41}$$

$$\Leftrightarrow \mathbf{y}_w = \boldsymbol{\mu}_j + \mathbf{v}_w, \tag{42}$$

with

$$\mathbb{E}[\mathbf{v}_w] = \boldsymbol{\Lambda}^{-1/2}\,\mathbf{P}^{\mathrm{T}}\,\mathbb{E}[\mathbf{v}] = \mathbf{0} \tag{43}$$

$$\mathbb{E}[\mathbf{v}_w\,\mathbf{v}_w^{\mathrm{T}}] = \boldsymbol{\Lambda}^{-1/2}\,\mathbf{P}^{\mathrm{T}}\,\mathbf{P}\,\boldsymbol{\Lambda}\,\mathbf{P}^{\mathrm{T}}\,\mathbf{P}\,\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}\,\boldsymbol{\Lambda}^{-1/2} = \mathbf{I}_N. \tag{44}$$

- Thus, $\mathbf{v}_w \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ and the detection problem is simplified in Eq. (42). Notice that

$$\delta\boldsymbol{\mu} \stackrel{\triangle}{=} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 = \boldsymbol{\Lambda}^{-1/2}\,\mathbf{P}^{\mathrm{T}}\,(\mathbf{m}_1 - \mathbf{m}_0) = \boldsymbol{\Lambda}^{-1/2}\,\mathbf{P}^{\mathrm{T}}\,\delta\mathbf{m} \tag{45}$$

$$\begin{aligned} d^2 &\stackrel{\triangle}{=} \delta\mathbf{m}^{\mathrm{T}}\mathbf{K}^{-1}\delta\mathbf{m} = \delta\mathbf{m}^{\mathrm{T}}\mathbf{P}\,\boldsymbol{\Lambda}^{-1}\,\mathbf{P}^{\mathrm{T}}\delta\mathbf{m} = \delta\mathbf{m}^{\mathrm{T}}\mathbf{P}\,\boldsymbol{\Lambda}^{-1/2}\,\boldsymbol{\Lambda}^{-1/2}\,\mathbf{P}^{\mathrm{T}}\delta\mathbf{m} \equiv \delta\boldsymbol{\mu}^{\mathrm{T}}\,\delta\boldsymbol{\mu} \\ &= \|\delta\boldsymbol{\mu}\|_2^2. \end{aligned} \tag{46}$$

Examples

- Another example:

$$\mathbf{y} = \boldsymbol{\mu}_j + \mathbf{v}, \tag{47}$$

with $\mathbf{v} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_N\right)$, $j \in \{0, 1\}$ and $\pi_0 = \pi_1 = 1/2$.
It can be easily shown that:

1. Minimum probability of error detection rule is the minimum distance rule:
$$\|\mathbf{y} - \boldsymbol{\mu}_0\|_2 \overset{H_1}{\geq} \|\mathbf{y} - \boldsymbol{\mu}_1\|_2$$
2. The probability of error of the above rule is given by:
$$\Pr(e) = Q\left(\frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2}{2\sigma}\right).$$

Proof: simply write the ML rule and exploit the fact that affine transformation of Gaussian vectors is also Gaussian.

References

[1] Bernard C. Levy, Principles of Signal Detection and Parameter Estimation, Springer 2008.
[2] Instructor notes.

Thank you!

Detection & Estimation Theory: Lectures 7 & 8

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- M-ary Hypothesis Testing
- Error Probability Bounds
- Example

M-ary Hypothesis Testing

- Assume M hypotheses $\{\mathrm{H}_j\}$, $j \in \{0,\ 1, \ldots, M-1\}$, with priors $\Pr(\mathrm{H}_j) \triangleq \pi_j$.
  1. Observe $\mathbf{y}$ and find out decision rule $\delta(\mathbf{y}) = j$, i.e., decide H_j for that specific $\mathbf{y}$.
  2. Equivalently, find out $\mathcal{Y}_i = \{\mathbf{y} :\ \delta(\mathbf{y}) = i\}$, with $\mathcal{Y} = \cup_{i=0}^{M-1} \mathcal{Y}_i$ and $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset \, \forall i \neq j$.
- We will revert to Bayesian formulation. Remember Bayes Risk $R(\delta)$ and conditional Bayes Risk $R(\delta|\mathrm{H}_j)$:

$$R(\delta) = \sum_{j=0}^{M-1} R(\delta|\mathrm{H}_j)\pi_j \tag{1}$$

$$R(\delta|\mathrm{H}_j) = \sum_{i=0}^{M-1} C_{ij} \Pr(\delta(\mathbf{y}) = i|\mathrm{H}_j) = \sum_{i=0}^{M-1} C_{ij} \Pr(\mathcal{Y}_i|\mathrm{H}_j)$$

$$\Leftrightarrow R(\delta) = \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} C_{ij} \Pr(\mathcal{Y}_i|\mathrm{H}_j)\, \pi_j, \tag{2}$$

where C_{ij} is the cost of deciding i when H_j holds.

M-ary Hypothesis Testing

- Bayesian formulation:

$$R(\delta) = \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} C_{ij} \, \Pr\left(\mathcal{Y}_i | \mathrm{H}_j\right) \pi_j \tag{3}$$

$$= \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} C_{ij} \int_{\mathcal{Y}_i} f_{\mathbf{y}|\mathrm{H}_j}\left(\mathbf{y}|\mathrm{H}_j\right) \pi_j \, d\mathbf{y} \tag{4}$$

$$= \sum_{i=0}^{M-1} \int_{\mathcal{Y}_i} \sum_{j=0}^{M-1} C_{ij} \, f_{\mathbf{y}|\mathrm{H}_j}\left(\mathbf{y}|\mathrm{H}_j\right) \pi_j \, d\mathbf{y} \tag{5}$$

$$= \sum_{i=0}^{M-1} \int_{\mathcal{Y}_i} \sum_{j=0}^{M-1} C_{ij} \, \Pr\left(\mathrm{H}_j|\mathbf{y}\right) f_{\mathbf{y}}\left(\mathbf{y}\right) d\mathbf{y} \tag{6}$$

$$= \sum_{i=0}^{M-1} \int_{\mathcal{Y}_i} f_{\mathbf{y}}\left(\mathbf{y}\right) \underbrace{\sum_{j=0}^{M-1} C_{ij} \, \Pr\left(\mathrm{H}_j|\mathbf{y}\right)}_{C_i(\mathbf{y})} d\mathbf{y} \tag{7}$$

M-ary Hypothesis Testing

- Bayesian formulation:

$$R(\delta) = \sum_{i=0}^{M-1} \int_{\mathcal{Y}_i} f_{\mathbf{y}}(\mathbf{y}) \underbrace{\sum_{j=0}^{M-1} C_{ij} \Pr(\mathrm{H}_j | \mathbf{y})}_{C_i(\mathbf{y})} \, d\mathbf{y} \tag{8}$$

$$= \sum_{i=0}^{M-1} \int_{\mathcal{Y}_i} C_i(\mathbf{y}) \, f_{\mathbf{y}}(\mathbf{y}) \, d\mathbf{y}. \tag{9}$$

- From Eq. (9),

$$\delta_B(\mathbf{y}) = \arg \min_{i \in \{0,1,\ldots,M-1\}} C_i(\mathbf{y}) \tag{10}$$

i.e., we select H_k if $C_k(\mathbf{y}) \leq C_i(\mathbf{y})$, $\forall i \in \{0, 1, \ldots, M-1\}$.

min Pr(e) rule: MAP rule

- Set symmetric costs:

$$C_{ij} = 1 - \delta_{ij} = \begin{cases} 0, & i = j, \\ 1, & i \neq j. \end{cases} \tag{11}$$

In that case, min of prob. of error is equivalent to risk minimization (as in the binary case):

$$R(\delta|\mathrm{H}_j) = \sum_{i=0}^{M-1} C_{ij} \Pr\left(\mathcal{Y}_i|\mathrm{H}_j\right) = \sum_{i \neq j} \Pr\left(\mathcal{Y}_i|\mathrm{H}_j\right) \tag{12}$$

$$\Rightarrow R(\delta|\mathrm{H}_j) = 1 - \Pr\left(\mathcal{Y}_j|\mathrm{H}_j\right) \equiv \Pr\left(e|\mathrm{H}_j\right) \tag{13}$$

$$R(\delta) = \sum_{j=0}^{M-1} R(\delta|\mathrm{H}_j)\pi_j = \sum_{j=0}^{M-1} \Pr\left(e|\mathrm{H}_j\right)\pi_j \equiv \Pr(e) \tag{14}$$

$$C_i(\mathbf{y}) = \sum_{j=0}^{M-1} C_{ij} \Pr\left(\mathrm{H}_j|\mathbf{y}\right) = \sum_{j \neq i} \Pr\left(\mathrm{H}_j|\mathbf{y}\right) = 1 - \Pr\left(\mathrm{H}_i|\mathbf{y}\right) \tag{15}$$

min $\Pr(e)$ rule: MAP rule

- Set symmetric costs:

$$C_{ij} = 1 - \delta_{ij} = \begin{cases} 0, & i = j, \\ 1, & i \neq j. \end{cases} \tag{16}$$

- Min prob. of error rule:

$$C_i(\mathbf{y}) = \sum_{j=0}^{M-1} C_{ij} \Pr\left(\mathrm{H}_j|\mathbf{y}\right) = \sum_{j \neq i} \Pr\left(\mathrm{H}_j|\mathbf{y}\right) = 1 - \Pr\left(\mathrm{H}_i|\mathbf{y}\right)$$

$$\delta_{MAP}(\mathbf{y}) = \arg \min_{i \in \{0,1,\ldots,M-1\}} \left\{1 - \Pr\left(\mathrm{H}_i|\mathbf{y}\right)\right\} \tag{17}$$

$$= \arg \max_{i \in \{0,1,\ldots,M-1\}} \Pr\left(\mathrm{H}_i|\mathbf{y}\right), \text{ (MAP rule)} \tag{18}$$

$$= \arg \max_{i \in \{0,1,\ldots,M-1\}} \frac{f_{\mathbf{y}|\mathrm{H}_i}\left(\mathbf{y}|\mathrm{H}_i\right) \pi_i}{f_{\mathbf{y}}(\mathbf{y})} \tag{19}$$

$$= \arg \max_{i \in \{0,1,\ldots,M-1\}} f_{\mathbf{y}|\mathrm{H}_i}\left(\mathbf{y}|\mathrm{H}_i\right) \pi_i \tag{20}$$

min Pr(e) rule and equiprobable hypotheses: ML rule

- Set symmetric costs and equiprobable hypotheses ($\pi_j = 1/M$):

$$C_{ij} = 1 - \delta_{ij} = \begin{cases} 0, & i = j, \\ 1, & i \neq j. \end{cases} \tag{21}$$

- Min prob. of error rule:

$$C_i(\mathbf{y}) = \sum_{j=0}^{M-1} C_{ij} \, \Pr\left(\mathrm{H}_j | \mathbf{y}\right) = \sum_{j \neq i} \Pr\left(\mathrm{H}_j | \mathbf{y}\right) = 1 - \Pr\left(\mathrm{H}_i | \mathbf{y}\right)$$

$$\delta_{ML}(\mathbf{y}) = \arg \max_{i \in \{0,1,\ldots,M-1\}} f_{\mathbf{y}|\mathrm{H}_i}\left(\mathbf{y}|\mathrm{H}_i\right) \cdot (\pi_i = 1/M) \tag{22}$$

$$= \arg \max_{i \in \{0,1,\ldots,M-1\}} f_{\mathbf{y}|\mathrm{H}_i}\left(\mathbf{y}|\mathrm{H}_i\right) \text{ (ML rule).} \tag{23}$$

- ...generalisation of the binary hypothesis testing case!

Example

- Under hypohesis H_i, with $i \in \{0, 1, \dots, M-1\}$, $\pi_i = 1/M$ and $\mathbf{y} = [y_1 \ y_2 \ \dots \ y_N]^{\mathrm{T}}$:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{m}_i, \mathbf{K}) \,. \tag{24}$$

  What is the minimum probability of error detection rule?

- Minimum probability of error detection rule for equiprobable hypotheses is the ML rule:

$$\delta_{ML}(\mathbf{y}) = \arg \max_{i \in \{0,1,\dots,M-1\}} f_{\mathbf{y}|\mathrm{H}_i}(\mathbf{y}|\mathrm{H}_i) \tag{25}$$

$$= \arg \max_{i \in \{0,1,\dots,M-1\}} \ln \left[f_{\mathbf{y}|\mathrm{H}_i}(\mathbf{y}|\mathrm{H}_i) \right] \tag{26}$$

$$= \arg \max_{i \in \{0,1,\dots,M-1\}} \ln \left[-\frac{1}{2} \left((\mathbf{y} - \mathbf{m}_i)^{\mathrm{T}} \mathbf{K}^{-1} (\mathbf{y} - \mathbf{m}_i) \right) - \frac{N}{2} \ln(2\pi) - \frac{\ln(|\mathbf{K}|)}{2} \right] \tag{27}$$

$$= \arg \min_{i \in \{0,1,\dots,M-1\}} \ln \left[\left((\mathbf{y} - \mathbf{m}_i)^{\mathrm{T}} \mathbf{K}^{-1} (\mathbf{y} - \mathbf{m}_i) \right) \right] \tag{28}$$

$$= \arg \min_{i \in \{0,1,\dots,M-1\}} \left((\mathbf{y} - \mathbf{m}_i)^{\mathrm{T}} \mathbf{K}^{-1} (\mathbf{y} - \mathbf{m}_i) \right) \tag{29}$$

- Note that through whitening $\mathbf{z} = \mathbf{\Lambda}^{-1/2} \mathbf{P}^{\mathrm{T}} \mathbf{y}$, i.e., using $\mathbf{K} = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^{\mathrm{T}}$, under hypothesis H_i:

$$\mathbf{z} \sim \mathcal{N}\left(\mathbf{\Lambda}^{-1/2} \mathbf{P}^{\mathrm{T}} \mathbf{m}_i, \mathbf{I}_N \right) . \tag{30}$$

  ...analytical proof in the previous lectures.

Error Probability Bounds

- Set symmetric costs:

$$C_{ij} = 1 - \delta_{ij} = \begin{cases} 0, & i = j, \\ 1, & i \neq j. \end{cases} \tag{31}$$

As we have already seen:

$$R(\delta|\mathrm{H}_j) = \sum_{i \neq j} \Pr\left(\mathcal{Y}_i|\mathrm{H}_j\right) = 1 - \Pr\left(\mathcal{Y}_j|\mathrm{H}_j\right) \equiv \Pr\left(e|\mathrm{H}_j\right) \triangleq \Pr\left(\mathcal{Y}_j^c|\mathrm{H}_j\right) \tag{32}$$

$$R(\delta) = \sum_{j=0}^{M-1} R(\delta|\mathrm{H}_j)\pi_j = \sum_{j=0}^{M-1} \Pr\left(e|\mathrm{H}_j\right)\pi_j \equiv \Pr(e) \tag{33}$$

- So, $\mathcal{Y}_j^c$ is the region where hypothesis H_j is NOT selected. Define formally the following:

$$\mathcal{Y}_j^c = \bigcup_{k \neq j} \mathcal{E}_{kj}, \tag{34}$$

$$\mathcal{E}_{kj} = \left\{\mathbf{y} : \Pr\left(\mathrm{H}_k|\mathbf{y}\right) > \Pr\left(\mathrm{H}_j|\mathbf{y}\right)\right\} \tag{35}$$

$$= \left\{\mathbf{y} : \frac{f_{\mathbf{y}|\mathrm{H}_k}\left(\mathbf{y}|\mathrm{H}_k\right)}{f_{\mathbf{y}|\mathrm{H}_j}\left(\mathbf{y}|\mathrm{H}_j\right)} > \frac{\pi_j}{\pi_k}\right\}, \tag{36}$$

i.e., $\mathcal{E}_{kj}$ is the region of $\{\mathbf{y}\}$'s where hypothesis H_k is preferred over H_j.

Error Probability Bounds

- The areas $\{\mathcal{E}_{kj}\}$ for given j usually overlap.
- It is often possible to find a subset of such areas, such that:

$$\mathcal{Y}_j^c = \bigcup_{k \in N(j)} \mathcal{E}_{kj}, \text{ with } N(j) \subset \{0,\ 1,\ \ldots,\ M-1\} \,/\, j \quad (37)$$

  i.e., $N(j)$ does not include element j.

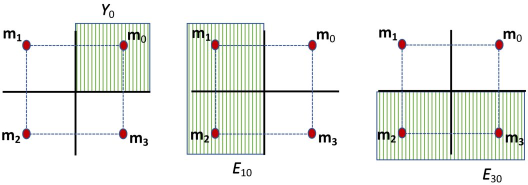
- Thus,

$$\Pr(e|\mathrm{H}_j) \triangleq \Pr\left(\mathcal{Y}_j^c|\mathrm{H}_j\right) \leq \underbrace{\sum_{k \in N(j)} \Pr(\mathcal{E}_{kj}|\mathrm{H}_j)}_{\text{Improved union bound}} \quad (38)$$

$$\max_{k \neq j} \Pr(\mathcal{E}_{kj}|\mathrm{H}_j) \leq \Pr\left(\mathcal{Y}_j^c|\mathrm{H}_j\right) \quad (39)$$

- The above bounds are usually simple to calculate!

Example



- Find error probability for optimal detection for the following:

$$\mathrm{H}_j : \ \mathbf{y} = \boldsymbol{m}_j + \mathbf{v}, \tag{40}$$

with $\mathbf{v} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_2\right)$, $j \in \{0, 1, 2, 3\}$ and $\pi_j = 1/4$.

- MAP is simplified to ML, simplified to minimum distance. In addition:

$$\mathcal{Y}_0^c = \mathcal{E}_{10} \cup \mathcal{E}_{30} \tag{41}$$

$$d \triangleq \|\boldsymbol{m}_1 - \boldsymbol{m}_0\|_2 = \|\boldsymbol{m}_2 - \boldsymbol{m}_1\|_2 = \|\boldsymbol{m}_3 - \boldsymbol{m}_2\|_2 = \|\boldsymbol{m}_3 - \boldsymbol{m}_0\|_2 \tag{42}$$

$$\Pr(e|\mathrm{H}_0) = \Pr\left(\mathcal{Y}_0^c|\mathrm{H}_0\right) = \Pr(\mathcal{E}_{10} \cup \mathcal{E}_{30}|\mathrm{H}_0) \leq \underbrace{\Pr(\mathcal{E}_{10}|\mathrm{H}_0)}_{Q\left(\frac{d}{2\sigma}\right)} + \underbrace{\Pr(\mathcal{E}_{30}|\mathrm{H}_0)}_{Q\left(\frac{d}{2\sigma}\right)} \tag{43}$$

$$\Leftrightarrow \Pr(e|\mathrm{H}_0) \leq 2Q\left(\frac{d}{2\sigma}\right), \tag{44}$$

where the latter is due to minimum distance binary error detection in white Gaussian noise (reminder in next slide).

Reminder

- Reminder:

$$\mathrm{H}_j: \ \mathbf{y} = \boldsymbol{\mu}_j + \mathbf{v}, \tag{45}$$

with $\mathbf{v} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_N\right)$, $j \in \{0, 1\}$ and $\pi_0 = \pi_1 = 1/2$.

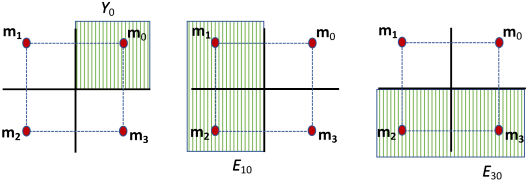
1. Minimum probability of error detection rule is the minimum distance rule:

$$\|\mathbf{y} - \boldsymbol{\mu}_0\|_2 \overset{H_1}{\geq} \|\mathbf{y} - \boldsymbol{\mu}_1\|_2$$

2. The probability of error of the above rule is given by:

$$\Pr(e) = Q\left(\frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2}{2\sigma}\right).$$

Example



- Error analysis:

$$\mathcal{Y}_0^c = \mathcal{E}_{10} \cup \mathcal{E}_{30} \tag{46}$$

$$d \triangleq \|\boldsymbol{m}_1 - \boldsymbol{m}_0\|_2 = \|\boldsymbol{m}_2 - \boldsymbol{m}_1\|_2 = \|\boldsymbol{m}_3 - \boldsymbol{m}_2\|_2 = \|\boldsymbol{m}_3 - \boldsymbol{m}_0\|_2 \tag{47}$$

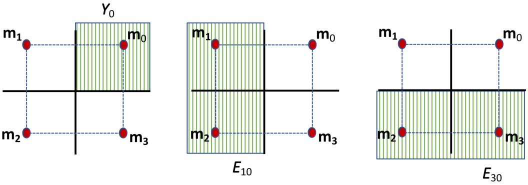
$$\Pr(e|\mathrm{H}_0) = \Pr\left(\mathcal{Y}_0^c|\mathrm{H}_0\right) = \Pr(\mathcal{E}_{10} \cup \mathcal{E}_{30}|\mathrm{H}_0) \le \underbrace{\Pr(\mathcal{E}_{10}|\mathrm{H}_0)}_{Q\left(\frac{d}{2\sigma}\right)} + \underbrace{\Pr(\mathcal{E}_{30}|\mathrm{H}_0)}_{Q\left(\frac{d}{2\sigma}\right)} = 2Q\left(\frac{d}{2\sigma}\right)$$

$$Q\left(\frac{d}{2\sigma}\right) = \Pr(\mathcal{E}_{10}|\mathrm{H}_0) = \Pr(\mathcal{E}_{30}|\mathrm{H}_0) \le \Pr(e|\mathrm{H}_0) \tag{48}$$

$$\Pr(e|\mathrm{H}_0) = \Pr(e)\,, \text{ due to symmetry} \tag{49}$$

$$\Rightarrow Q\left(\frac{d}{2\sigma}\right) \le \Pr(e) \le 2Q\left(\frac{d}{2\sigma}\right), \tag{50}$$

Example



- Error analysis with improved union bound:

$$d \triangleq \|\boldsymbol{m}_1 - \boldsymbol{m}_0\|_2 = \|\boldsymbol{m}_2 - \boldsymbol{m}_1\|_2 = \|\boldsymbol{m}_3 - \boldsymbol{m}_2\|_2 = \|\boldsymbol{m}_3 - \boldsymbol{m}_0\|_2 \tag{51}$$

$$Q\left(\frac{d}{2\sigma}\right) \leq \Pr(e) \leq 2Q\left(\frac{d}{2\sigma}\right) \tag{52}$$

- Exact error analysis:

$$\Pr(e) = 1 - \left(1 - Q\left(\frac{d}{2\sigma}\right)\right)^2 \tag{53}$$

$$= 2Q\left(\frac{d}{2\sigma}\right) - \left[Q\left(\frac{d}{2\sigma}\right)\right]^2 \tag{54}$$

Thus, upper error probability bound is tight!

References

[1] Bernard C. Levy, Principles of Signal Detection and Parameter Estimation, Springer 2008.
[2] Instructor notes.

Thank you!

Detection & Estimation Theory: Lectures 9-10

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning

Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- Bayesian Estimation: Problem Definition
    - Details on Problem Formulation
- Optimum Bayesian Estimator
    - Bayesian MSE Estimator
        MSE Performance Evaluation
    - Bayesian MAE Estimator
    - Bayesian MAP Estimator

Parameter Estimation Theory: Bayesian Formulation

- Formulation for Bayesian estimation is similar to Bayesian detection!
  1. Observe $\mathbf{y} \in \mathbb{R}^n$ for estimation of parameter vector $\mathbf{x} \in \mathbb{R}^m$.
  2. Detection: find out decision rule $\delta(\mathbf{y}) = j$, where j is discrete.
  3. Estimation: find out estimate $\hat{\mathbf{x}}(\mathbf{y}) \in \mathbb{R}^m$.
- Formulation for Bayesian estimation requires the following:

  1. Observation model: $f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$, i.e., conditional p.d.f. from measurements!
  2. Prior density: $f_{\mathbf{x}}(\mathbf{x})$, i.e., prior p.d.f. density of the unknown parameter. Notice that in the Bayesian formulation the unknown parameter is assumed random!
  3. Cost function: $C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x}) \equiv C(\hat{\mathbf{x}}, \mathbf{x})$, i.e., the cost of estimating $\mathbf{x}$ as $\hat{\mathbf{x}}(\mathbf{y})$.
- Derivations in Bayesian estimation proceed alongside similar lines to Bayesian detection!
  ...some details on the problem formulation follow...

Details on problem formulation

- Observation model $f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ is defined explicitly or indirectly through a measurement model.
- One example: $\mathbf{y} = \mathbf{h}(\mathbf{x}) + \mathbf{v}$, with $f_{\mathbf{v}}(\mathbf{v})$ known and $\mathbf{h}(\mathbf{x})$ is a deterministic vector function of $\mathbf{x}$; assume $n = m$.
  $\Rightarrow f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \overset{n=m}{=} f_{\mathbf{v}}(\mathbf{y} - \mathbf{h}(\mathbf{x}))$.

Proof:

$$\mathbf{J}_{n\times m}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_m} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_m} \end{bmatrix} = \begin{bmatrix} \nabla y_1 \\ \nabla y_2 \\ \vdots \\ \nabla y_n \end{bmatrix}, \tag{1}$$

$$\text{where } \nabla f = [\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_m}], \tag{2}$$

$$\mathbf{y} = \mathbf{h}(\mathbf{x}) + \mathbf{v} \Rightarrow$$

$$f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \overset{n=m}{=} \frac{f_{\mathbf{v}}(\mathbf{v})}{\det(\mathbf{J}(\mathbf{x},\mathbf{y}))}\Big|_{\mathbf{v}=\mathbf{y}-\mathbf{h}(\mathbf{x})} = \frac{f_{\mathbf{v}}(\mathbf{y}-\mathbf{h}(\mathbf{x}))}{\det(\mathbf{I}_n)} = f_{\mathbf{v}}(\mathbf{y}-\mathbf{h}(\mathbf{x})) \tag{3}$$

Details on problem formulation

- Prior density $f_{\mathbf{x}}(\mathbf{x})$ is known.
- ...unfortunately, prior density biases the estimator towards more probable values of $\mathbf{x}$, i.e., values of $\mathbf{x}$, where $f_{\mathbf{x}}(\mathbf{x})$ is larger.
- ...remember that we don't know anything about $f_{\mathbf{x}}(\mathbf{x})$ (i.e., it is random), apart from possible values.

Details on problem formulation

- Cost function: $C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x}) \equiv C(\hat{\mathbf{x}}, \mathbf{x})$, i.e., the cost of estimating $\mathbf{x}$ as $\hat{\mathbf{x}}(\mathbf{y})$.

$$C(\hat{\mathbf{x}}, \mathbf{x}) = C(\hat{\mathbf{x}} - \mathbf{x}) \equiv L(\mathbf{e}) \tag{4}$$

- Loss function $L(\mathbf{e})$ is a not decreasing function of error $\mathbf{e} \stackrel{\triangle}{=} \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}$.
- 3 different versions of loss function are typically used:
  1. $L_{\text{MSE}}(\mathbf{e}) = ||\mathbf{e}||_2^2$
  2. $L_{\text{MAE}}(\mathbf{e}) = ||\mathbf{e}||_1$
  3. $L_\epsilon(\mathbf{e})$: notch function.

Details on problem formulation

- 3 different versions of loss function are typically used:
  1. Euclidean norm or 2-norm squared:

     $$L_{\mathrm{MSE}}(\mathbf{e}) = ||\mathbf{e}||_2^2 = \mathbf{e}^{\mathrm{T}}\,\mathbf{e} = \sum_{i=1}^{m} e_i^2. \tag{5}$$

     Due to the square, it penalises more larger errors; improbable instances of $\mathbf{x}$ matter a lot; sensitive to modelling errors.
  2. Sum norm or 1-norm:

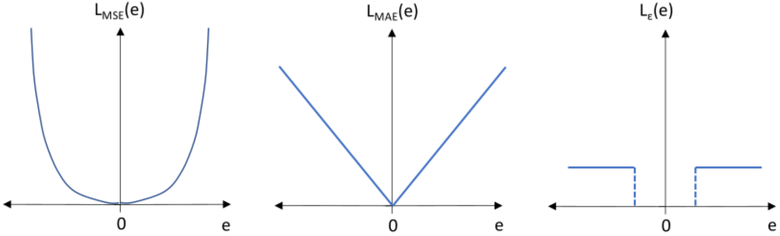
     $$L_{\mathrm{MAE}}(\mathbf{e}) = ||\mathbf{e}||_1 = \sum_{i=1}^{m} |e_i|. \tag{6}$$

     It weights equally the magnitude of all errors.
  3. using the infinity norm:

     $$L_{\epsilon}(\mathbf{e}) = \begin{cases} 0, & \text{if } ||\mathbf{e}||_\infty < \epsilon \\ 1, & \text{otherwise,} \end{cases} \tag{7}$$

     where infinity norm is given by $||\mathbf{e}||_\infty = \max_i |e_i|$; notch function cares about small errors and not about errors above ϵ.

Details on problem formulation: error (cost) functions



- The first two are convex, while the last one is non-convex (for scalar error).

Optimum Bayesian Estimator

- Cost is a function of random vectors, since both $\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x}$ are random.
- Bayesian objective: find $\hat{\mathbf{x}}(\mathbf{y}) = [x_1(\mathbf{y}) \;\; x_2(\mathbf{y}) \;\; \ldots \;\; x_m(\mathbf{y})]^{\mathrm{T}}$ by minimising the expected cost:

$$\min \mathbb{E}\left[C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x})\right], \text{ where}$$

$$\mathbb{E}\left[C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x})\right] = \int_{\mathbf{x}} \int_{\mathbf{y}} C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x}) \; f_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}) \; d\mathbf{x} \; d\mathbf{y} \tag{8}$$

$$= \int_{\mathbf{y}} \left[\int_{\mathbf{x}} C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x}) \; f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \; d\mathbf{x} \right] \; f_{\mathbf{y}}(\mathbf{y}) \; d\mathbf{y}. \tag{9}$$

- Notice that posterior p.d.f. $f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ and measurements p.d.f. $f_{\mathbf{y}}(\mathbf{y})$ can be (at least in principle) known:

$$f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \frac{f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \; f_{\mathbf{x}}(\mathbf{x})}{f_{\mathbf{y}}(\mathbf{y})}, \; f_{\mathbf{y}}(\mathbf{y}) = \int_{\mathbf{x}} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \; f_{\mathbf{x}}(\mathbf{x}) \; d\mathbf{x} \tag{10}$$

Optimum Bayesian Estimator

$$\min \mathbb{E}\left[C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x})\right], \text{ where}$$

$$\mathbb{E}\left[C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x})\right] = \int_{\mathbf{y}} \left[\int_{\mathbf{x}} C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x})\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x}\right]\ f_{\mathbf{y}}(\mathbf{y})\ d\mathbf{y}. \tag{11}$$

- Notice that since measurements p.d.f. $f_{\mathbf{y}}(\mathbf{y})$ is non-negative for each given $\mathbf{y}$, the term between brackets above is minimised for each given $\mathbf{y}$ according to the following:

$$\begin{aligned}
\hat{\mathbf{x}}(\mathbf{y}) &= \arg\min_{\mathbf{x}} \int_{\mathbf{x}} C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x})\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} && (12)\\
&= \arg\min_{\mathbf{x}} \frac{1}{f_{\mathbf{y}}(\mathbf{y})} \int_{\mathbf{x}} C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x})\ f_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y})\ d\mathbf{x} && (13)\\
&= \arg\min_{\mathbf{x}} \int_{\mathbf{x}} C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x})\ f_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y})\ d\mathbf{x} && (14)\\
&= \arg\min_{\mathbf{x}} \int_{\mathbf{x}} C(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{x})\ f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\ f_{\mathbf{x}}(\mathbf{x})\ d\mathbf{x}. && (15)
\end{aligned}$$

Optimum Bayesian Estimator: MSE case

- For minimum square error (MSE) loss function:

$$C(\hat{\mathbf{x}}(\mathbf{y}),\mathbf{x}) = (\underbrace{\hat{\mathbf{x}}-\mathbf{x}}_{\mathbf{e}})^{\mathrm{T}}(\hat{\mathbf{x}}-\mathbf{x}) = L_{\mathrm{MSE}}(\mathbf{e}) = ||\mathbf{e}||_2^2 \tag{16}$$

$$= \hat{\mathbf{x}}^{\mathrm{T}}\hat{\mathbf{x}} - \hat{\mathbf{x}}^{\mathrm{T}}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\hat{\mathbf{x}} + \mathbf{x}^{\mathrm{T}}\mathbf{x} \tag{17}$$

- Since $\mathbf{a}^{\mathrm{T}}\mathbf{b} = \mathbf{b}^{\mathrm{T}}\mathbf{a}$, $\frac{\partial \mathbf{b}^{\mathrm{T}}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{b}$ and $\frac{\partial \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \left(\mathbf{A}+\mathbf{A}^{\mathrm{T}}\right)\mathbf{x}$,

$$\frac{\partial C(\hat{\mathbf{x}}(\mathbf{y}),\mathbf{x})}{\partial \hat{\mathbf{x}}} = 2\left(\mathbf{I}+\mathbf{I}^{\mathrm{T}}\right)\hat{\mathbf{x}} - \mathbf{x} - \mathbf{x} = 2(\hat{\mathbf{x}}-\mathbf{x}) \tag{18}$$

- Denoting $\nabla_{\hat{\mathbf{x}}} = \left[\frac{\partial}{\partial \hat{x}_1} \quad \frac{\partial}{\partial \hat{x}_2} \quad \ldots \quad \frac{\partial}{\partial \hat{x}_m}\right]^{\mathrm{T}} = \frac{\partial}{\partial \hat{\mathbf{x}}}$, and the following (scalar) function of $\hat{\mathbf{x}}$ (from Eq. (12)),

$$J(\hat{\mathbf{x}}|\mathbf{y}) = \int_{\mathbf{x}} C(\hat{\mathbf{x}}(\mathbf{y}),\mathbf{x})\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} \tag{19}$$

$$\Rightarrow \nabla_{\hat{\mathbf{x}}} J(\hat{\mathbf{x}}|\mathbf{y}) = \int_{\mathbf{x}} \nabla_{\hat{\mathbf{x}}} C(\hat{\mathbf{x}}(\mathbf{y}),\mathbf{x})\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} = \mathbf{0} \tag{20}$$

$$= \int_{\mathbf{x}} 2(\hat{\mathbf{x}}-\mathbf{x})\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} = \mathbf{0} \tag{21}$$

Optimum Bayesian Estimator: MSE case

- ...continued from previous page...

$$\int_{\mathbf{x}} 2(\hat{\mathbf{x}} - \mathbf{x})\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} = \mathbf{0} \tag{22}$$

$$\Leftrightarrow \int_{\mathbf{x}} \hat{\mathbf{x}}\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} = \int_{\mathbf{x}} \mathbf{x}\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} \tag{23}$$

$$\Leftrightarrow \hat{\mathbf{x}} \int_{\mathbf{x}} \ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} = \int_{\mathbf{x}} \mathbf{x}\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} \stackrel{\triangle}{=} \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right] \tag{24}$$

$$\Leftrightarrow \hat{\mathbf{x}}(\mathbf{y})_{\mathrm{MSE}} = \int_{\mathbf{x}} \mathbf{x}\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} \stackrel{\triangle}{=} \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]. \tag{25}$$

- Bayesian MSE estimator is the conditional (on $\mathbf{y}$) mean!

MSE Performance Evaluation

- First, conditional mean square error is calculated:

$$
\begin{aligned}
J(\hat{\mathbf{x}}(\mathbf{y})|\mathbf{y}) &\overset{\hat{\mathbf{x}}(\mathbf{y})\equiv\hat{\mathbf{x}}}{=} \int_{\mathbf{x}} (\hat{\mathbf{x}}-\mathbf{x})^{\mathrm{T}}(\hat{\mathbf{x}}-\mathbf{x})\ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} && (26)\\
&\overset{\triangle}{=} \mathbb{E}_{\mathbf{x}|Y=\mathbf{y}}\left[(\hat{\mathbf{x}}-\mathbf{x})^{\mathrm{T}}(\hat{\mathbf{x}}-\mathbf{x})|Y=\mathbf{y}\right] && (27)\\
&\overset{(*)}{=} \mathbb{E}_{\mathbf{x}|Y=\mathbf{y}}\left[\mathrm{Trace}\left\{(\hat{\mathbf{x}}-\mathbf{x})^{\mathrm{T}}(\hat{\mathbf{x}}-\mathbf{x})\right\}|Y=\mathbf{y}\right] && (28)\\
&\overset{(**)}{=} \mathrm{Trace}\left\{\mathbb{E}_{\mathbf{x}|Y=\mathbf{y}}\left[(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^{\mathrm{T}}|Y=\mathbf{y}\right]\right\} && (29)\\
&= \mathrm{Trace}\left\{\mathbf{K}_{\mathbf{x}|Y=\mathbf{y}}\right\}, && (30)
\end{aligned}
$$

where $\mathrm{Trace}(\mathbf{A}\,\mathbf{B}) = \mathrm{Trace}(\mathbf{B}\,\mathbf{A})$ property was used in $(*)$ and $\mathbb{E}\left[\mathrm{Trace}\{\cdot\}\right] = \mathrm{Trace}\{\mathbb{E}[\cdot]\}$ property in $(**)$, and

$$
\begin{aligned}
\mathbf{K}_{\mathbf{x}|Y=\mathbf{y}} \equiv \mathbf{K}_{\mathbf{x}|Y=\mathbf{y}}(\mathbf{y}) &\overset{\triangle}{=} \mathbb{E}_{\mathbf{x}|Y=\mathbf{y}}\left[(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^{\mathrm{T}}|Y=\mathbf{y}\right] && (31)\\
&= \int_{\mathbf{x}} (\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]-\mathbf{x})(\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]-\mathbf{x})^{\mathrm{T}} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x} && (32)\\
&= \int_{\mathbf{x}} (\mathbf{x}-\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right])(\mathbf{x}-\mathbb{E}\left[\mathbf{x}|\mathbf{y}\right])^{\mathrm{T}} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\ d\mathbf{x}. && (33)
\end{aligned}
$$

MSE Performance Evaluation

- Then, (unconditional) minimum mean square error (MMSE) is calculated:

$$\begin{aligned}
\text{MMSE} &\triangleq \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x})^{\mathrm{T}}(\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x})\right] &(34)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\text{Trace}\left\{(\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x})^{\mathrm{T}}(\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x})\right\}\right] &(35)\\
&= \text{Trace}\left\{\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x})(\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x})^{\mathrm{T}}\right]\right\} &(36)\\
&= \text{Trace}\left\{\mathbf{K}_{\mathrm{E}}\right\} = \text{Trace}\left\{\mathbf{K}_{\mathrm{X/Y}}\right\}, &(37)
\end{aligned}$$

where $\text{Trace}(\mathbf{A}\,\mathbf{B}) = \text{Trace}(\mathbf{B}\,\mathbf{A})$ and $\mathbb{E}\left[\text{Trace}\{\cdot\}\right] = \text{Trace}\{\mathbb{E}[\cdot]\}$ properties were again used and $\mathbf{K}_{\mathrm{E}}$ follows:

$$\begin{aligned}
\mathbf{K}_{\mathrm{E}} &\triangleq \int_{\mathbf{y}}\int_{\mathbf{x}} (\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{y}])(\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{y}])^{\mathrm{T}} f_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})\; d\mathbf{x}\; d\mathbf{y} &(38)\\
&= \int_{\mathbf{y}}\int_{\mathbf{x}} (\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{y}])(\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{y}])^{\mathrm{T}} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) f_{\mathbf{y}}(\mathbf{y})\; d\mathbf{x}\; d\mathbf{y}\\
&= \int_{\mathbf{y}}\left[\int_{\mathbf{x}} (\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{y}])(\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{y}])^{\mathrm{T}} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\; d\mathbf{x}\right] f_{\mathbf{y}}(\mathbf{y})\; d\mathbf{y}\\
&\overset{Eq.\ (33)}{=} \int_{\mathbf{y}} \mathbf{K}_{\mathbf{x}|Y=\mathbf{y}}\; f_{\mathbf{y}}(\mathbf{y})\; d\mathbf{y}. &(39)
\end{aligned}$$

Optimum Bayesian Estimator: MAE case

- For minimum absolute error (MAE) loss function:

$$C(\hat{\mathbf{x}}(\mathbf{y}),\mathbf{x}) = ||\hat{\mathbf{x}}-\mathbf{x}||_1 = |\hat{x}_1 - x_1| + |\hat{x}_2 - x_2| + \ldots + |\hat{x}_m - x_m|$$

$$J(\hat{\mathbf{x}}|\mathbf{y}) = \int_{\mathbf{x}} ||\hat{\mathbf{x}}-\mathbf{x}||_1 \; f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \; d\mathbf{x} \tag{40}$$

- Denoting $\mathrm{sign}(z) = +1$ if $z \geq 0$ and $\mathrm{sign}(z) = -1$, otherwise, the following are calculated:

$$\frac{\partial J(\hat{\mathbf{x}}|\mathbf{y})}{\partial \hat{x}_i} = \int_{\mathbf{x}} \mathrm{sign}(\hat{x}_i - x_i) \; f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \; d\mathbf{x} \tag{41}$$

$$= \int_{x_i} \mathrm{sign}(\hat{x}_i - x_i) \times$$

$$\left[\int_{x_1} \int_{x_2} \ldots \int_{x_{i-1}} \int_{x_{i+1}} \ldots \int_{x_m} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) dx_1 \; dx_2 \ldots dx_{i-1} dx_{i+1} \ldots dx_m \right] dx_i$$

$$= \int_{x_i} \mathrm{sign}(\hat{x}_i - x_i) \; f_{x_i|\mathbf{y}}(x_i|\mathbf{y}) \; dx_i . \tag{42}$$

Optimum Bayesian Estimator: MAE case

- Continuing from previous slide,

$$\frac{\partial J(\hat{\mathbf{x}}|\mathbf{y})}{\partial \hat{x}_i} = 0 \Leftrightarrow \int_{x_i} \operatorname{sign}(\hat{x}_i - x_i)\ f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\ dx_i = 0 \Leftrightarrow \tag{43}$$

$$\int_{-\infty}^{\hat{x}_i} \operatorname{sign}(\hat{x}_i - x_i)\ f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\ dx_i + \int_{\hat{x}_i}^{+\infty} \operatorname{sign}(\hat{x}_i - x_i)\ f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\ dx_i = 0 \Leftrightarrow \tag{44}$$

$$\int_{-\infty}^{\hat{x}_i} f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\ dx_i - \int_{\hat{x}_i}^{+\infty} f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\ dx_i = 0 \Leftrightarrow \tag{45}$$

$$I_1 \triangleq \int_{-\infty}^{\hat{x}_i} f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\ dx_i = \int_{\hat{x}_i}^{+\infty} f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\ dx_i \triangleq I_2 \tag{46}$$

- Since $I_1 + I_2 = 1$ and from above $I_1 = I_2$, the Bayesian MAE estimate is the median of the posterior density:

$$\int_{-\infty}^{\hat{x}_i^{\mathrm{MAE}}} f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\ dx_i = \int_{\hat{x}_i^{\mathrm{MAE}}}^{+\infty} f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\ dx_i = 1/2. \tag{47}$$

Optimum Bayesian Estimator: MAE case

- Thus, the $i-$th entry of $\hat{\mathbf{x}}_{\mathrm{MAE}}(\mathbf{y})$ is the median of the posterior density $f_{x_i|\mathbf{y}}(x_i|\mathbf{y})$:

$$\int_{-\infty}^{\hat{x}_i^{\mathrm{MAE}}} f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\; dx_i = \int_{\hat{x}_i^{\mathrm{MAE}}}^{+\infty} f_{x_i|\mathbf{y}}(x_i|\mathbf{y})\; dx_i = 1/2. \qquad (48)$$

Optimum Bayesian Estimator: Notch Cost Function

- For notch cost function $L_\epsilon(\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x}) = \begin{cases} 0, & \text{if } ||\mathbf{e}||_\infty < \epsilon \\ 1, & \text{otherwise,} \end{cases}$, where $||\mathbf{e}||_\infty = \max_i |e_i|$, assuming $\mathbf{e} = [e_1 \quad e_2 \ldots e_m]^{\mathrm{T}}$. Thus,

$$J(\hat{\mathbf{x}}|\mathbf{y}) = \int_{\mathbf{x}} L_\epsilon(\hat{\mathbf{x}} - \mathbf{x}) \ f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \ d\mathbf{x} = \int_{||\hat{\mathbf{x}}-\mathbf{x}||_\infty \geq \epsilon} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \ d\mathbf{x}$$
$$= 1 - \int_{||\hat{\mathbf{x}}-\mathbf{x}||_\infty < \epsilon} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \ d\mathbf{x} \tag{49}$$

- For $\epsilon \to 0^+$, minimization of $J(\cdot)$ above is equivalent to maximizing the following:

$$\int_{||\hat{\mathbf{x}}-\mathbf{x}||_\infty < \epsilon} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \ d\mathbf{x} \approx (2\epsilon)^m \ f_{\mathbf{x}|\mathbf{y}}(\hat{\mathbf{x}}|\mathbf{y}) \tag{50}$$

- In other words, for small ϵ,

$$\hat{\mathbf{x}}(\mathbf{y}) \equiv \hat{\mathbf{x}}_{\mathrm{MAP}}(\mathbf{y}) = \arg\max_{\mathbf{x}} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}). \tag{51}$$

Optimum Bayesian Estimator: Notch Cost Function

- For notch cost function $L_\epsilon(\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x})$ and small ϵ,

$$\hat{\mathbf{x}}(\mathbf{y}) \equiv \hat{\mathbf{x}}_{\mathrm{MAP}}(\mathbf{y}) = \arg\max_{\mathbf{x}} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \tag{52}$$

- ...the Bayesian estimate becomes the mode (i.e., maximum) of the posterior density (MAP estimate).
- It makes sense if the posterior density $f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ has a single dominant peak, or multiple peaks of the same size.
- Example: for jointly Gaussian $\mathbf{x}, \mathbf{y}$, conditional mean, median and mode coincide, i.e.,

$$\hat{\mathbf{x}}_{\mathrm{MSE}}(\mathbf{y}) = \hat{\mathbf{x}}_{\mathrm{MAE}}(\mathbf{y}) = \hat{\mathbf{x}}_{\mathrm{MAP}}(\mathbf{y}) = \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right].$$

References

[1] Bernard C. Levy, Principles of Signal Detection and Parameter Estimation, Springer 2008.
[2] Instructor notes.

Thank you!

Detection & Estimation Theory: Lectures 11-12

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- Examples of Bayesian Estimation
- Properties of Bayesian MSE Estimator
- MMSE Estimation in Linear Gaussian Systems

Bayesian Estimation Example 1

- The following (exponential) p.d.f.s are given: $f_{y|x}(y|x) = xe^{-xy}u(y), f(x) = ae^{-ax}u(x)$
- Need to compute $f_y(y)$ to compute $f_{x|y}(x|y)$:

$$
\begin{aligned}
f(x,y) = f(y|x)f(x) = axe^{-(a+y)x}u(x)u(y) \Rightarrow \\
f(y) = \int_0^{+\infty} axe^{-(a+y)x}u(y)dx = \frac{a}{(y+a)^2}u(y) \\
\Rightarrow f(x|y) = \frac{f_{x,y}(x,y)}{f(y)} = (y+a)^2 xe^{-(y+a)x}u(x)
\end{aligned}
$$

- MSE estimator:

$$
\hat{x}_{\text{MSE}}(y) = \int_{-\infty}^{+\infty} xf(x|y)dx = \cdots = \frac{2}{y+a}
$$

Bayesian Estimation Example 1

- MAE estimator:

$$\frac{1}{2} = \int_{-\infty}^{\hat{x}} f(x|y)dx = (y+a)^2 \int_0^{\hat{x}} xe^{-(y+a)x}dx = \cdots =$$

$$= \left[1 + (a+y)\hat{x}\right] e^{-(a+y)\hat{x}} \overset{c=(a+y)\hat{x}}{=} (1+c)e^{-c} \Rightarrow$$

$$e^c = 2(1+c) \Rightarrow c = \ln\left[2\left(1+c\right)\right] \Rightarrow c \approx 1.68 \Rightarrow$$

$$\hat{x}_{\text{MAE}}(y) = \frac{c}{a+y} = \frac{1.68}{y+a}$$

- MAP estimator:

$$\frac{\partial f_{x|y}(x|y)}{\partial x} = (y+a)^2 e^{-(y+a)x} + (y+a)^2 x e^{-(y+a)x}(-(y+a)) = 0 \Rightarrow$$

$$(y+a)^2 e^{-(y+a)x}\left(1-(y+a)x\right) = 0 \Rightarrow$$

$$\hat{x}_{\text{MAP}} = \frac{1}{y+a}$$

Bayesian Estimation Example 2

- $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$ jointly Gaussian $\Leftrightarrow \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$

$$\mathbf{m} \triangleq \begin{bmatrix} \mathbb{E}[\mathbf{x}] \\ \mathbb{E}[\mathbf{y}] \end{bmatrix} = \begin{bmatrix} \mathbf{m}_x \\ \mathbf{m}_y \end{bmatrix}, \mathbf{K} \triangleq \mathbb{E}\left\{ \begin{bmatrix} \mathbf{x} - \mathbf{m}_x \\ \mathbf{y} - \mathbf{m}_y \end{bmatrix} \begin{bmatrix} (\mathbf{x} - \mathbf{m}_x)^{\mathrm{T}} & (\mathbf{y} - \mathbf{m}_y)^{\mathrm{T}} \end{bmatrix} \right\}$$
$$= \begin{bmatrix} \mathbf{K}_X & \mathbf{K}_{XY} \\ \mathbf{K}_{YX} & \mathbf{K}_Y \end{bmatrix}, \mathbf{K}_{XY} = \mathbf{K}_{YX}^{\mathrm{T}}$$

- $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$ jointly Gaussian $\Rightarrow$

$$f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{m}_{X|Y}, \mathbf{K}_{X|Y}\right)$$

$$\begin{aligned} \mathbf{m}_{X|Y} &= \mathbf{m}_x + \mathbf{K}_{XY}\mathbf{K}_Y^{-1}(\mathbf{y} - \mathbf{m}_y) \\ \mathbf{K}_{X|Y} &= \mathbf{K}_X - \mathbf{K}_{XY}\mathbf{K}_Y^{-1}\mathbf{K}_{YX} \\ \hat{\mathbf{x}}_{\mathrm{MSE}}(\mathbf{y}) &= \mathbb{E}[\mathbf{x}|\mathbf{y}] \equiv \mathbf{m}_{X|Y} = \mathbf{m}_x + \mathbf{K}_{XY}\mathbf{K}_Y^{-1}(\mathbf{y} - \mathbf{m}_y) \\ \mathrm{MMSE} &\triangleq \mathbb{E}\left[||\mathbf{x} - \hat{\mathbf{x}}_{\mathrm{MSE}}||_2^2\right] = \mathrm{Trace}\left(\mathbf{K}_{X|Y}\right) \end{aligned}$$

Bayesian Estimation Example 2

- The three estimates coincide (conditional mean, median, maximum):

$$\hat{\mathbf{x}}_{\mathrm{MSE}}(\mathbf{y}) = \hat{\mathbf{x}}_{\mathrm{MAE}}(\mathbf{y}) = \hat{\mathbf{x}}_{\mathrm{MAP}}(\mathbf{y}) = \mathbf{m}_x + \mathbf{K}_{XY}\mathbf{K}_Y^{-1}(\mathbf{y} - \mathbf{m}_y).$$

Properties of MSE Estimator

1. **Orthogonality Property**:

$$\mathbb{E}\Big[(\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{y}])\ g(\mathbf{y}) \Big] = \mathbf{0} \tag{1}$$

- The above states that the expected value of the inner product of the error vector with any function of the measurements is always zero.
- This property is just an expression of the fact that the conditional mean $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ extracts all the information in $\mathbf{y}$ that can be used to reduce the MSE.
- Notice that $g(\mathbf{y})$ can be scalar or (row) vector.

Proof.

$$\textit{1st method: } \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{y}])\ \ g(\mathbf{y})\right] = \mathbb{E}\left[\mathbf{x}\ g(\mathbf{y})\right] - \mathbb{E}\left[\mathbb{E}[\mathbf{x}|\mathbf{y}]g(\mathbf{y})\right]$$
$$= \mathbb{E}\left[\mathbf{x}\ g(\mathbf{y})\right] - \mathop{\mathbb{E}}_{Y}\left[\mathop{\mathbb{E}}_{X|Y}\left[\mathbf{x}\ g(\mathbf{y})|\mathbf{y}\right]\right]$$
$$\overset{(*)}{=} \mathbb{E}\left[\mathbf{x}\ g(\mathbf{y})\right] - \mathbb{E}\left[\mathbf{x}\ g(\mathbf{y})\right] = \mathbf{0}$$

where at step $(*)$ the law of iterated expectation was used:

$$E[h(x,y)] = \mathop{E}_{Y}\mathop{E}_{X|Y}\left[h(x,y)|y\right] \Leftrightarrow$$
$$\int\int h(x,y)f(x,y)\ dx\ dy = \int_y\int_x h(x,y)\ f(x|y)\ dx\ f(y)\ dy$$

$$\textit{2nd method: } \mathop{\mathbb{E}}_{X,Y}[\mathbf{x}\ g(\mathbf{y})] = \mathop{\mathbb{E}}_{Y}\left\{\mathop{\mathbb{E}}_{X|Y}\left[\mathbf{x}\ g(\mathbf{y})|\ \mathbf{y}\right]\right\} = \mathop{\mathbb{E}}_{Y}\left[\mathop{\mathbb{E}}_{X|Y}\left[\mathbf{x}|\mathbf{y}\right]\ \ g(\mathbf{y})\right]$$

$\square$

Properties of MSE Estimator

2. **Uniqueness Property**:

   $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ is the **unique** vector function $\in \mathbb{R}^m$ that adheres to the orthogonality property.

Proof.

Suppose that $\mathbf{h}(\mathbf{y})$ is another function with $\mathbb{E}\left[(\mathbf{x}-\mathbf{h}(\mathbf{y}))\,g(\mathbf{y})\right]=\mathbf{0}$ for all functions $g(\cdot)$. Then, it follows:

$$
\begin{aligned}
\mathbb{E}\left[||\mathbb{E}[\mathbf{x}|\mathbf{y}]-\mathbf{h}(\mathbf{y})||_2^2\right] &= \mathbb{E}\left[\big(\underbrace{\mathbb{E}(\mathbf{x}|\mathbf{y})-\mathbf{h}(\mathbf{y})}_{g(\mathbf{y})}\big)^{\mathrm{T}}\big(\mathbb{E}[\mathbf{x}|\mathbf{y}]-\mathbf{x}+\mathbf{x}-\mathbf{h}(\mathbf{y})\big)\right] \\
&= \mathbb{E}\left[g^{\mathrm{T}}(\mathbf{y})\,(\mathbb{E}[\mathbf{x}|\mathbf{y}]-\mathbf{x})+g^{\mathrm{T}}(\mathbf{y})\,(\mathbf{x}-\mathbf{h}(\mathbf{y}))\right] \\
&= \mathbf{0}+\mathbf{0}\ \text{(due to orthogonality principle)} \Rightarrow \\
\mathbb{E}[\mathbf{x}|\mathbf{y}] &= \mathbf{h}(\mathbf{y}),
\end{aligned}
$$

since at the last step, the expected value of a non-negative random variable is zero only when the variable is always zero. □

Properties of MSE Estimator

3. **Variance reduction**:

   if $\mathbf{K}_X = \mathbb{E}[(\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)^{\mathrm{T}}]$

   and

   $$\mathbf{K}_E = \mathbb{E}\left[\left(\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{y}]\right)\left(\mathbf{x} - \mathbb{E}(\mathbf{x}|\mathbf{y})\right)^{\mathrm{T}}\right] \equiv \mathbf{K}_{X|Y}$$

   then

   - $\mathbf{K}_E \leq \mathbf{K}_X$ i.e., $\mathbf{K}_X - \mathbf{K}_E$ is positive semi-definite,
   - $\mathbf{K}_E = \mathbf{K}_x$ if and only if (iff) $\mathbf{m}_x = \mathbb{E}[\mathbf{x}|\mathbf{y}]$ i.e., knowledge of the observation $\mathbf{y}$ does not improve the estimate of $\mathbf{x}$.

Properties of MSE Estimator

Proof.

$\mathbf{x} - \mathbf{m}_x = \mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{y}] + \mathbb{E}[\mathbf{x}|\mathbf{y}] - \mathbf{m}_x$, therefore

$$\begin{aligned}
\mathbf{K}_X = \mathbf{K}_E &+ \mathbb{E}\left[\big(\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{y}]\big)\big(\mathbb{E}[\mathbf{x}|\mathbf{y}] - \mathbf{m}_x\big)^{\mathrm{T}}\right] \\
&+ \mathbb{E}\left[\big(\mathbb{E}[\mathbf{x}|\mathbf{y}] - \mathbf{m}_x\big)\big(\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{y}]^{\mathrm{T}}\big)\right] \\
&+ \mathbb{E}\Big[\big(\underbrace{\mathbb{E}[\mathbf{x}|\mathbf{y}] - \mathbf{m}_x}_{\Delta(\mathbf{y})}\big)\big(\mathbb{E}[\mathbf{x}|\mathbf{y} - \mathbf{m}_x]\big)^{\mathrm{T}}\Big] \Leftrightarrow
\end{aligned}$$

$$\begin{aligned}
\mathbf{K}_X = \mathbf{K}_E &+ \mathbb{E}\Big[\big(\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{y}]\big)\Delta(\mathbf{y})^{\mathrm{T}}\Big] + \mathbb{E}\Big[\Delta(\mathbf{y})\big(\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{y}]\big)^{\mathrm{T}}\Big] \\
&+ \underbrace{\mathbb{E}\left[\Delta(\mathbf{y})\Delta(\mathbf{y})^{\mathrm{T}}\right]}_{\mathbf{K}_\Delta} = \mathbf{K}_E + \mathbf{K}_\Delta \Leftrightarrow
\end{aligned}$$

$\mathbf{K}_X - \mathbf{K}_E = \mathbf{K}_\Delta = \mathbb{E}[\Delta(\mathbf{y})\Delta(\mathbf{y})^{\mathrm{T}}] \geq 0$ since

$\mathbf{z}^{\mathrm{T}}\mathbb{E}\left[\Delta(\mathbf{y})\Delta(\mathbf{y})^{\mathrm{T}}\right]\mathbf{z} = \mathbb{E}\Big[\underbrace{\mathbf{z}^{\mathrm{T}}\Delta\mathbf{y}}_{\mathbf{z}_0^{\mathrm{T}}}\Delta\mathbf{y}^{\mathrm{T}}\mathbf{z}\Big] = \mathbb{E}\left[||\mathbf{z}_0||_2^2\right] \geq 0$ □

Properties of MSE Estimator

...proof continued.

- $\mathbf{K}_\Delta = \mathbf{0}$ (all elements zero) $\Rightarrow$ Trace$(\mathbf{K}_\Delta) = 0$ and the following holds:

$$\begin{aligned}\mathrm{Trace}(\mathbf{K}_\Delta) &= \mathrm{Trace}\left[\mathbb{E}[\Delta\,\Delta^{\mathrm{T}}]\right] = \mathbb{E}[\mathrm{Trace}(\Delta\,\Delta^{\mathrm{T}})]\\ &= \mathbb{E}[\Delta^{\mathrm{T}}\,\Delta] = \mathbb{E}[||\Delta||_2^2] = 0 \Rightarrow\\ \Delta \equiv \Delta(\mathbf{y}) = \mathbf{0} &\Rightarrow \mathbb{E}[\mathbf{x}|\mathbf{y}] = \mathbf{m}_x\end{aligned}$$

- For the other direction, i.e., $\Delta(\mathbf{y}) = \mathbf{0} \Rightarrow \mathbf{K}_\Delta = \mathbf{0}$ the proof is trivial.

$\square$

MMSE Estimation in Linear Gaussian Systems

- Let $\mathbf{x} \in \mathbb{R}^{D_x}$ to be estimated and $\mathbf{y} \in \mathbb{R}^{D_y}$ with

  $$p(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x), p(\mathbf{y}|\mathbf{x}) \sim \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_y)$$

  where $\mathbf{A}$ (deterministic) $D_y \times D_x$ real matrix, $\mathbf{b} \in \mathbb{R}^{D_y}$ and $\boldsymbol{\mu}_y = \mathbb{E}[\boldsymbol{y}] = \boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b}$.

  Then, $p(\mathbf{x}|\mathbf{y}) \sim \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y})$ with
  $\boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x\right]$
  and
  $\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{A}\right)^{-1}$

- Thus,

  $$\hat{\mathbf{x}}_{\mathrm{MMSE}}(\mathbf{y}) \equiv \mathbb{E}[\mathbf{x}|\mathbf{y}] \equiv \boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x\right].$$

MMSE Estimation in Linear Gaussian Systems

Proof (1/4):

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \Rightarrow$$

$$\log p(\mathbf{x},\mathbf{y}) = \log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}) =$$

$$= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_x)^{\mathrm{T}}\boldsymbol{\Sigma}_x^{-1}(\mathbf{x}-\boldsymbol{\mu}_x) - \frac{1}{2}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b})^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}) +$$

$$+ \text{constant terms}$$

$$= -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_x^{-1}\mathbf{x} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{y} - \frac{1}{2}(\mathbf{A}\mathbf{x})^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}(\mathbf{A}\mathbf{x}) + \mathbf{y}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}(\mathbf{A}\mathbf{x}) +$$

$$+ \text{ linear terms } + \text{ constant terms}$$

$$= -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_x^{-1}\mathbf{x} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{A}\mathbf{x} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{y} + +\mathbf{y}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}(\mathbf{A}\mathbf{x}) +$$

$$+ \text{ linear terms } + \text{ constant terms}$$

Continue proof (2/4):

$$= -\frac{1}{2}\begin{bmatrix}\mathbf{x}^{\mathrm{T}} & \mathbf{y}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1} \\ -\boldsymbol{\Sigma}_y^{-1}\mathbf{A} & \boldsymbol{\Sigma}_y^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{x} \\ \mathbf{y}\end{bmatrix} + \text{ constant terms}$$

$$= -\frac{1}{2}\begin{bmatrix}\mathbf{x}^{\mathrm{T}} & \mathbf{y}^{\mathrm{T}}\end{bmatrix}\boldsymbol{\Sigma}^{-1}\begin{bmatrix}\mathbf{x} \\ \mathbf{y}\end{bmatrix} \Rightarrow$$

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix}\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1} \\ -\boldsymbol{\Sigma}_y^{-1}\mathbf{A} & \boldsymbol{\Sigma}_y^{-1}\end{bmatrix} \triangleq \boldsymbol{\Lambda} = \begin{bmatrix}\boldsymbol{\Lambda}_{xx} & \boldsymbol{\Lambda}_{xy} \\ \boldsymbol{\Lambda}_{yx} & \boldsymbol{\Lambda}_{yy}\end{bmatrix} \tag{2}$$

Useful (for the proof) Theorem

Continue proof (3/4): The following theorem will be utilized; its proof will be given in the problem sets and can be found in various textbooks, e.g., Chapter 4 in *"Machine Learning, a Probabilistic Perspective" by Kevin Murphy*):

- Assume $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$ Gaussian vector, i.e. $\mathbf{x}_1, \mathbf{x}_2$ jointly Gaussians, with $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \boldsymbol{\Sigma} \triangleq \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\right] = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \boldsymbol{\Lambda} \triangleq \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}$ $(\mathbf{A^{**}})$

Then:

$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}),\ p(\mathbf{x}_1) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}),\ p(\mathbf{x}_2) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}),$$

$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) = \boldsymbol{\Lambda}_{11}^{-1}[\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)] \\ &= \boldsymbol{\Sigma}_{1|2}[\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)] \qquad (3) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1} \qquad (4) \end{aligned}$$

MMSE Estimation in Linear Gaussian Systems

Continue proof (4/4).

Thus, from ($\mathbf{A}$**) and Eq. (4),

$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Lambda}_{xx}^{-1} = (\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{A})^{-1}$$

From from ($\mathbf{A}$**), Eq. (3), and $\boldsymbol{\mu}_y = \mathbf{A}\boldsymbol{\mu}_x + \boldsymbol{b}$,

$$\begin{aligned}
\boldsymbol{\mu}_{x|y} &= \boldsymbol{\Sigma}_{x|y}\big[\boldsymbol{\Sigma}_{x|y}^{-1}\boldsymbol{\mu}_x - \boldsymbol{\Lambda}_{xy}(\boldsymbol{y} - \boldsymbol{\mu}_y)\big] \\
&= \boldsymbol{\Sigma}_{x|y}\big[(\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{A})\boldsymbol{\mu}_x + \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \mathbf{A}\boldsymbol{\mu}_x - \boldsymbol{b})\big] \\
&= \boldsymbol{\Sigma}_{x|y}\big[\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x + \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{A}\boldsymbol{\mu}_x + \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{b}) - \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{A}\boldsymbol{\mu}_x\big] \\
&= \boldsymbol{\Sigma}_{x|y}\big[\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x + \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{b})\big]
\end{aligned}$$

□

- Important Remark: for the above $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$, the following can be also shown:

$$p(\boldsymbol{y}) = \mathcal{N}(\mathbf{A}\boldsymbol{\mu}_x + \boldsymbol{b}, \boldsymbol{\Sigma}_y + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)$$

MMSE in Linear Gaussian Systems: Example

$f(\boldsymbol{y}_i|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}, \boldsymbol{\Sigma}_y), f(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$

- We observe $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N$, which are i.i.d.
- We would like to estimate $\boldsymbol{x}$ based on $\boldsymbol{y}_0$, where

$$\boldsymbol{y}_0 = \frac{1}{N} \Sigma_{i=1}^{N} \boldsymbol{y}_i \tag{5}$$

  - Notice that $\boldsymbol{y}_0 \sim \mathcal{N}(\boldsymbol{x}, \frac{1}{N}\boldsymbol{\Sigma}_y)$. This can be easily shown by the fact that affine transformation of a Gaussian vector is again a Gaussian vector[1] and the fact that:

$$\boldsymbol{y}_0 = \frac{1}{N} \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}}_{\boldsymbol{B}} \underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}}_{\boldsymbol{y}}$$

[1] if $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$ then $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \sim \mathcal{N}(\mathbf{A}\mathbf{m} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}})$

MMSE in Linear Gaussian Systems: Example

- Thus, $f(\boldsymbol{y}_0|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}, \frac{1}{N}\boldsymbol{\Sigma}_y)$, i.e., $\mathbf{A}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{x} \Rightarrow \mathbf{A} = \mathbf{I}, \boldsymbol{b} = \mathbf{0}$ and $f(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.
- With direct application of the above theorem, $f(\boldsymbol{x}|\boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left[\mathbf{I}^{\mathrm{T}}\left(\frac{1}{N}\boldsymbol{\Sigma}_y\right)^{-1}(\boldsymbol{y}_0 - \mathbf{0}) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right]$$

and

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Sigma}_0^{-1} + \left(\frac{1}{N}\boldsymbol{\Sigma}_y\right)^{-1}\right)^{-1} = \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}_y^{-1}\right)^{-1}$$

Therefore,

$$\hat{\boldsymbol{x}}(\boldsymbol{y})_{\mathrm{MSE}} \equiv \boldsymbol{\mu} = \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}_y^{-1}\right)^{-1}\left(N\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}_0 + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right).$$

Thank you!

Detection & Estimation Theory: Lecture 13

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

Technical University of Crete

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- Bayesian Linear Estimators
  - Derivation
- Remarks
  - Proof on Remark 4
- Linear MSE Estimator Example

Bayesian Linear Estimators: Problem Formulation

- Assume $\mathbf{m}_x = \mathbb{E}[\mathbf{x}]$, $\mathbf{m}_y = \mathbb{E}[\mathbf{y}]$ and the joint covariance matrix $\mathbf{K} = \mathbb{E}\left[\begin{bmatrix}\mathbf{x}-\mathbf{m}_x \\ \mathbf{y}-\mathbf{m}_y\end{bmatrix}\begin{bmatrix}(\mathbf{x}-\mathbf{m}_x)^{\mathrm{T}} & (\mathbf{y}-\mathbf{m}_y)^{\mathrm{T}}\end{bmatrix}\right] = \begin{bmatrix}\mathbf{K}_X & \mathbf{K}_{XY} \\ \mathbf{K}_{YX} & \mathbf{K}_Y\end{bmatrix}$ are known
- $\mathbf{K} > 0$ ($\mathbf{K}^{-1}$ exists); otherwise a non-trivial linear combination in vector $\mathbf{y}-\mathbf{m}_y$ exists, so we could replace observation $\mathbf{y}$ by a vector of smaller dimension!
- $f(\mathbf{y}|\mathbf{x})$ and $f(\mathbf{x})$ are UNKNOWN!
- $\mathbf{y} \in \mathbb{R}^n, \mathbf{x} \in \mathbb{R}^m$
- we are looking for $\hat{\mathbf{x}}_L(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{b}$ that minimizes $\mathbb{E}\left[\|\mathbf{e}\|_2^2\right]$ (MSE) with the following:
  1. $\mathbf{A}$ a $m \times n$ matrix
  2. $\mathbf{b} \in \mathbb{R}^m$
  3. $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}_L(\mathbf{y}) = \mathbf{x} - A\mathbf{y} - \mathbf{b}$
  4. $\mathbb{E}[\mathbf{e}] \equiv \mathbf{m}_e = \mathbf{m}_x - \mathbf{A}\mathbf{m}_y - \mathbf{b}$

Bayesian Linear Estimators

- Notice that

$$\begin{aligned}\hat{\mathbf{x}}(\mathbf{y}) = \mathbb{E}[\mathbf{x}|\mathbf{y}] &\Rightarrow \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{y}])\cdot \mathbf{g}(\mathbf{y})] = 0 \text{ (orthogonality property)}\\ &\Rightarrow \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{y}])] = 0 \qquad (g(\mathbf{y}) = 1)\\ &\Rightarrow \mathbb{E}[\mathbf{e}] = 0\end{aligned}$$

- For $\hat{\mathbf{x}}_L(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{b} \Rightarrow \mathbb{E}[\mathbf{e}] = \mathbf{m}_x - \mathbf{A}\mathbf{m}_y - \mathbf{b} \neq 0$
- Thus,

$$\mathbf{K}_E \triangleq \mathbb{E}\left[(\mathbf{e}-\mathbf{m}_e)(\mathbf{e}-\mathbf{m}_e)^{\mathrm{T}}\right] \tag{1}$$

$$\mathbf{e}-\mathbf{m}_e = (\mathbf{x}-\mathbf{A}\mathbf{y}-\mathbf{b}) - (\mathbf{m}_x - \mathbf{A}\mathbf{m}_y - \mathbf{b}) \tag{2}$$

$$= \begin{bmatrix}\mathbf{I}_m & -\mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{x}-\mathbf{m}_x\\ \mathbf{y}-\mathbf{m}_y\end{bmatrix} \tag{3}$$

- From Eq. (1) and Eq. (3):

$$\mathbf{K}_E = \begin{bmatrix}\mathbf{I}_m & -\mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{K}_X & \mathbf{K}_{XY}\\ \mathbf{K}_{YX} & \mathbf{K}_Y\end{bmatrix}\begin{bmatrix}\mathbf{I}_m\\ -\mathbf{A}^{\mathrm{T}}\end{bmatrix}$$

Bayesian Linear Estimators

- MMSE:

$$\mathbb{E}\left[\|\mathbf{e}\|_2^2\right] = \mathbb{E}\left[\mathbf{e}^{\mathrm{T}}\mathbf{e}\right] = \|\mathbf{e}\|_2^2 - \|\mathbf{m}_e\|_2^2 + \|\mathbf{m}_e\|_2^2 \tag{4}$$

$$= \mathbb{E}\left[(\mathbf{e}-\mathbf{m}_e)^{\mathrm{T}}(\mathbf{e}-\mathbf{m}_e)\right] + \|\mathbf{m}_e\|_2^2 \tag{5}$$

$$= \mathrm{Trace}\left(\mathbf{K}_E\right) + \|\mathbf{m}_e\|_2^2 \tag{6}$$

- $\mathbf{K}_E$ depends on $\mathbf{A}$, $\mathbf{m}_e$ depends on $\mathbf{A}$ and $\mathbf{b}$
- Set $\mathbf{m}_e = 0 \Rightarrow \mathbf{m}_x - \mathbf{A}\mathbf{m}_y - \mathbf{b} = 0 \Rightarrow$

$$\mathbf{b} = \mathbf{m}_x - \mathbf{A}\mathbf{m}_y \tag{7}$$

- Now we need to find $\mathbf{A}$. We work as follows:

$$\begin{aligned}
\mathbf{K}_E &= \begin{bmatrix}\mathbf{I}_m & -\mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{K}_X & \mathbf{K}_{XY}\\ \mathbf{K}_{YX} & \mathbf{K}_{\mathbf{Y}}\end{bmatrix}\begin{bmatrix}\mathbf{I}_m\\ -\mathbf{A}^{\mathrm{T}}\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{K}_X - \mathbf{A}\mathbf{K}_{YX} & \mathbf{K}_{XY} - \mathbf{A}\mathbf{K}_Y\end{bmatrix}\begin{bmatrix}\mathbf{I}_m\\ -\mathbf{A}^{\mathrm{T}}\end{bmatrix}\\
&= \mathbf{K}_X - \mathbf{A}\mathbf{K}_{YX} - \mathbf{K}_{XY}\mathbf{A}^{\mathrm{T}} + \mathbf{A}\mathbf{K}_Y\mathbf{A}^{\mathrm{T}}\\
&= \mathbf{K}_{XY}\mathbf{K}_Y^{-1}\mathbf{K}_{YX} - \mathbf{K}_{XY}\mathbf{A}^{\mathrm{T}} - \mathbf{A}\mathbf{K}_{YX} + \mathbf{A}\mathbf{K}_Y\mathbf{A}^{\mathrm{T}} + \mathbf{S}\\
&= \left(\mathbf{K}_{XY} - \mathbf{A}\mathbf{K}_Y\right)\left(\left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1}\right)^{\mathrm{T}} - \mathbf{A}^{\mathrm{T}}\right) + \mathbf{S}\\
&= \left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1} - \mathbf{A}\right)\mathbf{K}_Y\left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1} - \mathbf{A}\right)^{\mathrm{T}} + \mathbf{S}
\end{aligned}$$

Bayesian Linear Estimators

- Schur complement $\mathbf{S} \triangleq \mathbf{K}_X - \mathbf{K}_{XY}\mathbf{K}_Y^{-1}\mathbf{K}_{YX}$
  - Schur complement of $\mathbf{K}_\mathbf{Y}$ in $\mathbf{K}$ plays a role in evaluating the determinant and the inverse of block matrices.
  - Schur complement is constant and known and does not depend on $\mathbf{A}$

  Thus, $\mathrm{Trace}\left(\mathbf{K}_E\right) =$ $\mathrm{Trace}\left(\left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1} - \mathbf{A}\right)\mathbf{K}_Y\left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1} - \mathbf{A}\right)^{\mathrm{T}}\right) + \mathrm{Trace}\left(\mathbf{S}\right)$, since $\mathrm{Trace}\left(\mathbf{A}+\mathbf{B}\right) = \mathrm{Trace}\left(\mathbf{A}\right) + \mathrm{Trace}\left(\mathbf{B}\right)$.
- $\mathrm{Trace}\left(\left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1} - \mathbf{A}\right)\mathbf{K}_Y\left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1} - \mathbf{A}\right)^{\mathrm{T}}\right) \geq 0$, since $\left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1} - \mathbf{A}\right)\mathbf{K}_Y\left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1} - \mathbf{A}\right)^{\mathrm{T}}$ is positive semi-definite. Thus, $\mathrm{Trace}\left(\left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1} - \mathbf{A}\right)\mathbf{K}_Y\left(\mathbf{K}_{XY}\mathbf{K}_Y^{-1} - \mathbf{A}\right)^{\mathrm{T}}\right) = 0 \Leftrightarrow$ $\mathbf{K}_{XY}\mathbf{K}_Y^{-1} = \mathbf{A}$.
- Thus, $\mathrm{Tr}\left(\mathbf{K}_\mathbf{E}\right)$ is minimized iff

$$\mathbf{K}_{XY}\mathbf{K}_Y^{-1} = \mathbf{A} \tag{8}$$

- From Eq. (7) and Eq. (8), $\hat{\mathbf{x}}_L\left(\mathbf{y}\right) = \mathbf{m}_x + \mathbf{K}_{XY}\mathbf{K}_Y^{-1}\left(\mathbf{y} - \mathbf{m}_y\right)$.

Remarks

1. $\mathbb{E}\left[\mathbf{e}\right] = \mathbf{m}_e = \mathbf{0}$
2. $\hat{\mathbf{x}}_L\left(\mathbf{y}\right) = \mathbf{m}_x + \mathbf{K}_{XY}\mathbf{K}_Y^{-1}\left(\mathbf{y} - \mathbf{m}_y\right) = \hat{\mathbf{x}}_{\mathrm{MSE}}\left(\mathbf{y}\right)$
   for $\mathbf{x}, \mathbf{y}$ jointly Gaussians

   $$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}_x \\ \mathbf{m}_y \end{bmatrix}, \begin{bmatrix} \mathbf{K}_X & \mathbf{K}_{XY} \\ \mathbf{K}_{YX} & \mathbf{K}_Y \end{bmatrix} \right)$$

   Linear estimate and MSE estimate coincide for the Gaussian case, i.e., all MSE estimates will necessarily be linear.
3. As promised, linear-least-square estimate $\hat{\mathbf{x}}_L\left(\mathbf{y}\right)$ requires knowledge of first and second moments and not knowledge of $f_{\mathbf{y}|\mathbf{x}}\left(\mathbf{y}|\mathbf{x}\right), f_{\mathbf{x}}\left(\mathbf{x}\right)$.
4. Orthogonality property holds only for linear functions of $\mathbf{y}$:

   $$\mathbb{E}\left[\left(\mathbf{x} - \hat{\mathbf{x}}_L\left(\mathbf{y}\right)\right) \cdot \mathbf{g}\left(\mathbf{y}\right)\right] = \mathbf{0}$$

   for all linear functions $\mathbf{g}\left(\mathbf{y}\right) = g_0 + \mathbf{y}^{\mathrm{T}}\mathbf{g}_1$ where g_0 a real scalar and $\mathbf{g}_1$ a vector in $\mathbb{R}^n$.

Proof.

- Estimation error: $\mathbf{x} - \hat{\mathbf{x}}_L(\mathbf{y}) = \begin{bmatrix} \mathbf{I}_m & -\mathbf{K}_{XY}\mathbf{K}_Y^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mathbf{m}_x \\ \mathbf{y} - \mathbf{m}_y \end{bmatrix}$
- $\mathbf{g}(\mathbf{y}) = g_0 + \mathbf{m}_y^{\mathrm{T}}\mathbf{g}_1 + (\mathbf{y} - \mathbf{m}_y)^{\mathrm{T}}\mathbf{g}_1$ where $g_0 + \mathbf{m}_y^{\mathrm{T}}\mathbf{g}_1$ is a constant term.
- Therefore,

$$
\begin{aligned}
\mathbb{E}\left[(\mathbf{x} - \hat{\mathbf{x}}_L(\mathbf{y}))\,\mathbf{g}(\mathbf{y})\right] &= \mathbb{E}\left[(\mathbf{x} - \hat{\mathbf{x}}_L(\mathbf{y}))\,(\mathbf{y} - \mathbf{m}_y)^{\mathrm{T}}\,\mathbf{g}_1\right] \\
&= \begin{bmatrix} \mathbf{I}_m & -\mathbf{K}_{XY}\mathbf{K}_Y^{-1} \end{bmatrix} \mathbb{E}\left[\begin{bmatrix} \mathbf{x} - \mathbf{m}_x \\ \mathbf{y} - \mathbf{m}_y \end{bmatrix} \begin{bmatrix} \mathbf{y} & -\mathbf{m}_y \end{bmatrix}^{\mathrm{T}}\right] \mathbf{g}_1 \\
&= \underbrace{\begin{bmatrix} \mathbf{I}_m & -\mathbf{K}_{XY}\mathbf{K}_Y^{-1} \end{bmatrix}}_{m\times(m+n)} \underbrace{\begin{bmatrix} \mathbf{K}_{XY} \\ \mathbf{K}_Y \end{bmatrix}}_{(m+n)\times n} \underbrace{\mathbf{g}_1}_{n\times 1} \\
&= \underbrace{(\mathbf{K}_{XY} - \mathbf{K}_{XY})}_{m\times n} \underbrace{\mathbf{g}_1}_{n\times 1} \\
&= \mathbf{0}
\end{aligned}
$$

$\square$

Remarks (Cont'd)

Proof.

5. $\hat{\mathbf{x}}_L(\mathbf{y})$ is the unique linear estimator that adheres to the orthogonality property.
   - Suppose that $\mathbf{h}(\mathbf{y})$ is another linear estimator with the property $\mathbb{E}\left[(\mathbf{x} - \mathbf{h}(\mathbf{y})) \cdot \mathbf{g}(\mathbf{y})\right] = \mathbf{0}$.
   - Thus
   $\mathbb{E}\left[\|\hat{\mathbf{x}}_L(\mathbf{y}) - \mathbf{h}(\mathbf{y})\|_2^2\right] = \mathbb{E}\left[\mathbf{g}^{\mathrm{T}}(\mathbf{y})(\hat{\mathbf{x}}_L(\mathbf{y}) - \mathbf{x} + \mathbf{x} - \mathbf{h}(\mathbf{y}))\right]$
   $= \mathbb{E}\left[\mathbf{g}^{\mathrm{T}}(\mathbf{y})(\hat{\mathbf{x}}_L(\mathbf{y}) - \mathbf{x})\right] + \mathbb{E}\left[\mathbf{g}^{\mathrm{T}}(\mathbf{y})(\mathbf{x} - \mathbf{h}(\mathbf{y}))\right] = \mathbf{0} + \mathbf{0}$
   $= \mathbf{0}$,
   since $\mathbf{g}(\mathbf{y}) = \hat{\mathbf{x}}_L(\mathbf{y}) - \mathbf{h}(\mathbf{y})$ is linear (because $\hat{\mathbf{x}}_L(\mathbf{y}), \mathbf{h}(\mathbf{y})$ are linear in $\mathbf{y}$).

□

Example

- Assume $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}$
  1. $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$
  2. $\mathbf{v}$ uncorrelated with $\mathbf{x}$
  3. $\mathbb{E}\left[\mathbf{v}\right] = 0$ and $\mathbb{E}\left[\mathbf{v}\mathbf{v}^{\mathrm{T}}\right] = \mathbf{R} > 0$, i.e., $\mathbf{R}$ is positive-definite
  4. $\mathbf{H}$ (known) constant matrix
- We need to find the $\hat{\mathbf{x}}_L\left(\mathbf{y}\right)$:
  - $\mathbf{m}_y = \mathbf{H}\mathbf{m}_x$
  - $$\begin{aligned}\mathbf{K}_{YX} &= \mathbb{E}\left[\left(\mathbf{y} - \mathbf{m}_y\right)\left(\mathbf{x} - \mathbf{m}_x\right)^{\mathrm{T}}\right] \\ &= \mathbb{E}\left[\left(\mathbf{H}\left(\mathbf{x} - \mathbf{m}_x\right) + \mathbf{v}\right)\left(\mathbf{x} - \mathbf{m}_x\right)^{\mathrm{T}}\right] \\ &= \mathbf{H}\mathbf{K}_X\end{aligned}$$
    since $\mathbb{E}\left[\mathbf{v}\right] = 0$ and $\mathbf{v}, \mathbf{x}$ are uncorrelated
  - $$\begin{aligned}\mathbf{K}_Y &= \mathbb{E}\left[\left(\mathbf{y} - \mathbf{m}_y\right)\left(\mathbf{y} - \mathbf{m}_y\right)^{\mathrm{T}}\right] \\ &= \mathbb{E}\left[\left(\mathbf{H}\left(\mathbf{x} - \mathbf{m}_x\right) + \mathbf{v}\right)\left(\mathbf{H}\left(\mathbf{x} - \mathbf{m}_x\right) + \mathbf{v}\right)^{\mathrm{T}}\right] \\ &= \mathbf{H}\mathbf{K}_X\mathbf{H}^{\mathrm{T}} + \mathbf{R}\end{aligned}$$

Example

- Therefore,
  - $\hat{\mathbf{x}}_L(\mathbf{y}) = \mathbf{m}_x + \mathbf{K}_X\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{K}_X\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}(\mathbf{y} - \mathbf{H}\mathbf{m}_x)$
  - $\mathbf{K}_L = \mathbf{K}_X - \mathbf{K}_X\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{K}_X\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}\mathbf{H}\mathbf{K}_X \equiv \mathbf{K}_E$
- We can show that

$$\mathbf{K}_L^{-1} = \mathbf{K}_X^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} \qquad \text{(A)}$$

i.e. $\mathbf{K}_L = \left(\mathbf{K}_X^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-1}$
("Sherman-Morrison-Woodbury" identity)
Proof:

  - $(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}$
  - $(\mathbf{K}_L)^{-1} = \mathbf{K}_X - \mathbf{K}_X\mathbf{H}^{\mathrm{T}}\left(\mathbf{R} + \mathbf{H}\mathbf{K}_X\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\mathbf{K}_X$,
    where $\mathbf{A} = \mathbf{K}_X^{-1}$, $\mathbf{B} = \mathbf{H}^{\mathrm{T}}$, $\mathbf{C} = \mathbf{R}^{-1}$, $\mathbf{D} = \mathbf{H}$

Example

- Denote $\mathbf{G} \overset{\triangle}{=} \mathbf{K}_X \mathbf{H}^{\mathrm{T}} \left(\mathbf{H}\mathbf{K}_X\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1} \qquad (\alpha)$
- It is true that $\mathbf{G} = \mathbf{K}_L \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \qquad (\beta)$
- Proof:
  1. $\mathbf{K}_L^{-1} \cdot \mathbf{G} \cdot \left(\mathbf{H}\mathbf{K}_X\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right) \overset{(\beta)}{=} \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} \left(\mathbf{H}\mathbf{K}_X\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right) \qquad (6)$
  2. $\mathbf{K}_L^{-1} \cdot \mathbf{G} \cdot \left(\mathbf{H}\mathbf{K}_X\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right) \overset{(\alpha)}{=} \left(\mathbf{K}_X^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right) \mathbf{K}_X \mathbf{H}^{\mathrm{T}} \ (7)$
     which is true
  3. (6) = (7) after simple manipulations...
- Also

$$\begin{aligned}
\mathbf{K}_L^{-1} \hat{\mathbf{x}}_L(\mathbf{y}) &= \left(\mathbf{K}_X^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)\left(\mathbf{m}_x + \mathbf{G} \cdot (\mathbf{y} - \mathbf{H}\mathbf{m}_x)\right) \\
&= \mathbf{K}_X^{-1}\mathbf{m}_x + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{m}_x + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{H}\mathbf{m}_x) \\
&= \mathbf{K}_X^{-1}\mathbf{m}_x + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y} \qquad (\mathrm{B})
\end{aligned}$$

- (A), (B) are used in the derivation of Kalman Filter.

Example

- Since
  1. $\mathbf{G} \equiv \mathbf{K}_L \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1}$
  2. $\hat{\mathbf{x}}_L(\mathbf{y}) = \mathbf{m}_x + \underbrace{\mathbf{K}_X \mathbf{H}^{\mathrm{T}} \left(\mathbf{H}\mathbf{K}_X \mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}}_{\mathbf{G}} (\mathbf{y} - \mathbf{H}\mathbf{m}_x)$
  3. $\mathbf{K}_L \equiv \left(\mathbf{K}_X^{-1} + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}\right)^{-1}$
- Then $\hat{\mathbf{x}}_L(\mathbf{y}) = \mathbf{m}_x + \left(\mathbf{K}_X^{-1} + \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}\mathbf{m}_x)$
- Notice that the above expression is regularly used in various textbooks.

Thank you!

Detection & Estimation Theory: Lectures 14-16

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- Estimation of Non-random Parameters
  - ML Estimation Examples
- Performance: Cramer-Rao Bound
- Cramer-Rao Bound Derivation
- Existence of Efficient Estimator

Estimation of Non-random Parameters

- Alternative view: $f_{\mathbf{x}}(\mathbf{x})$ not available, $\mathbf{x} \in \mathbb{R}^m$ is viewed as unknown and non-random!
  1. Likelihood function $f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ of measurements vector $\mathbf{y}$ when the parameter vector is $\mathbf{x}$.
  2. One simple solution: maximum likelihood (ML) estimate!

$$\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y}) = \arg\max_{\mathbf{x}\in\mathbb{R}^m} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \tag{1}$$

$$= \arg\max_{\mathbf{x}\in\mathbb{R}^m} \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \tag{2}$$

...the latter (logarithmic) is convenient for p.d.f. in the exponential family (Poisson, Exponential, Gaussian):

$$f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = \exp\left(\mathbf{x}^{\mathrm{T}}\mathbf{s}(\mathbf{y}) - \mathbf{t}(\mathbf{x})\right)$$

- Bias: $\mathbf{b} \triangleq \mathbb{E}\left[\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\right] = \mathbf{x} - \mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y})\right]$
- Bias of ML estimate may not be zero.

Estimation of Non-random Parameters: Bias

- Bias: $\mathbf{b} \triangleq \mathbb{E}\left[\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\right] = \mathbf{x} - \mathbb{E}\left[\hat{\mathbf{x}}(\mathbf{y})\right]$
- ...is the expected value of the error.
- ...is a weak metric, since it does not ensure that for a single measurement vector $\mathbf{y}$ the estimate will offer the true parameter vector $\mathbf{x}$.

Estimation of Non-random Parameters: Examples

- Assume $\mathbf{y} \sim \mathcal{N}\left(A\mathbf{s}, \sigma^2 \mathbf{I}_N\right)$, where $\mathbf{y} \in \mathbb{R}^N$ and $\mathbf{s}, \sigma^2$ known.
- $\hat{A}_{\mathrm{ML}}(\mathbf{y})$?

$$\ln\left[f_{\mathbf{y}|A}(\mathbf{y}|A)\right] = -\frac{1}{2\sigma^2}||\mathbf{y} - A\mathbf{s}||_2^2 + \ln\left[\frac{1}{\sqrt{(2\pi\sigma^2)^N}}\right] \tag{3}$$

$$\Rightarrow \arg\max_{A\in\mathbb{R}} \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = \arg\min_{A}||\mathbf{y} - A\mathbf{s}||_2^2 \tag{4}$$

$$= \arg\min_{A}\Big(\mathbf{y} - A\mathbf{s})^{\mathrm{T}}(\mathbf{y} - A\mathbf{s}\Big) \tag{5}$$

$$= \arg\min_{A}\left(||\mathbf{s}||_2^2 A^2 - 2\mathbf{s}^{\mathrm{T}}\mathbf{y}A + ||\mathbf{y}||_2^2\right) \tag{6}$$

$$= \frac{\mathbf{s}^{\mathrm{T}}\mathbf{y}}{||\mathbf{s}||_2^2} \tag{7}$$

$$\Rightarrow \hat{A}_{\mathrm{ML}}(\mathbf{y}) = \frac{\mathbf{s}^{\mathrm{T}}\mathbf{y}}{||\mathbf{s}||_2^2}. \tag{8}$$

Estimation of Non-random Parameters: Examples

- Assume $\mathbf{y} = [y_1 \;\; y_2 \dots y_N]^{\mathrm{T}}$ of N independent and identically distributed (i.i.d.) r.v.'s $\{y_k\}$, with $y_k \sim \mathcal{N}(m, v)$.
- Estimation problems:
  1. Estimate m with v known.
  2. Estimate v with m known.
  3. Estimate m, v.
- Check bias of estimate(s).

Estimation of Non-random Parameters: Examples

- Assume $\mathbf{y} = [y_1 \;\; y_2 \ldots y_N]^{\mathrm{T}}$ of N independent and identically distributed (i.i.d.) r.v.'s $\{y_k\}$, with $y_k \sim \mathcal{N}(m, v)$.
- Estimation problem:
  1. Estimate m with v known. Check bias of estimate.

Solution: ...this is the previous example, with $A \equiv m$, $\mathbf{s} = [1 \;\; 1 \ldots 1]^{\mathrm{T}}$ and $\sigma^2 = v$, Thus,

$$\hat{m}_{\mathrm{ML}}(\mathbf{y}) = \frac{\mathbf{s}^{\mathrm{T}}\mathbf{y}}{N} = \frac{\sum_{k=1}^{N} y_k}{N} \tag{9}$$

$$\mathbb{E}\left[\hat{m}_{\mathrm{ML}}(\mathbf{y})\right] = \frac{Nm}{N} = m \text{ (unbiased estimate)} \tag{10}$$

Estimation of Non-random Parameters: Examples

- Estimation problem:
  2. Estimate v with m known. Check bias of estimate.

Solution: $\ln\left[f_{\mathbf{y}|v}(\mathbf{y}|v)\right] = -\frac{N}{2}\ln(2\pi v) - \frac{1}{2v}\sum_{k=1}^{N}(y_k - m)^2$

$$\frac{d}{dv}\ln\left[f_{\mathbf{y}|v}(\mathbf{y}|v)\right] = 0 \Rightarrow \frac{1}{2v}\left[-N + \frac{1}{v}\sum_{k=1}^{N}(y_k - m)^2\right] = 0 \quad (11)$$

$$\Rightarrow \hat{v}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N}(y_k - m)^2 \quad (12)$$

- Need to make sure that $\frac{d^2}{dv^2}\ln\left[f_{\mathbf{y}|v}(\mathbf{y}|v)\right] < 0$ at $v = \hat{v}_{\mathrm{ML}}$:

$$\frac{d^2}{dv^2}\ln\left[f_{\mathbf{y}|v}(\mathbf{y}|v)\right] = \ldots = \frac{1}{v^2}\left(\frac{N}{2} - \frac{1}{v}\sum_{k=1}^{N}(y_k - m)^2\right) \quad (13)$$

$$\overset{v=\hat{v}_{\mathrm{ML}}}{=} -\frac{1}{\hat{v}_{\mathrm{ML}}^2}\frac{N}{2} < 0 \quad (14)$$

Estimation of Non-random Parameters: Examples

- Assume $\mathbf{y} = [y_1 \;\; y_2 \dots y_N]^{\mathrm{T}}$ of N independent and identically distributed (i.i.d.) r.v.'s $\{y_k\}$, with $y_k \sim \mathcal{N}(m, v)$.
- Estimation problem:

  2 Estimate v with m known. Check bias of estimate.

$$\hat{v}_{\mathrm{ML}} = \frac{1}{N} \sum_{k=1}^{N} (y_k - m)^2 \tag{15}$$

$$\mathbb{E}\left[\hat{v}_{\mathrm{ML}}\right] = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left[(y_k - m)^2\right] = \frac{Nv}{N} = v \text{ (unbiased estimate).} \tag{16}$$

Estimation of Non-random Parameters: Examples

- Estimation problem:
  3. Estimate m, v. Check bias of estimates.

$$\frac{\partial}{\partial m} \ln \left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v) \right] = 0 \Rightarrow \frac{1}{v} \sum_{k=1}^{N} (y_k - m) = 0 \tag{17}$$

$$\Rightarrow \hat{m}_{\mathrm{ML}} = \frac{1}{N} \sum_{k=1}^{N} y_k \tag{18}$$

$$\frac{\partial}{\partial v} \ln \left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v) \right] = 0 \Rightarrow \frac{1}{2v} \left[-N + \frac{1}{v} \sum_{k=1}^{N} (y_k - m)^2 \right] = 0 \tag{19}$$

$$\Rightarrow \hat{v}_{\mathrm{ML}} = \frac{1}{N} \sum_{k=1}^{N} (y_k - \hat{m}_{\mathrm{ML}})^2 \tag{20}$$

- How do we know that the above maximise the log-likelihood?

Estimation of Non-random Parameters: Examples

- Estimation problem:
  3. Estimate m, v. Check bias of estimates.

$$\hat{m}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N} y_k \tag{21}$$

$$\hat{v}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N} (y_k - \hat{m}_{\mathrm{ML}})^2 \tag{22}$$

$$g(m,v) \triangleq \ln\left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v)\right] \tag{23}$$

$$\mathbf{H}_g(m,v) = \begin{bmatrix} \frac{\partial^2 g(m,v)}{\partial m^2} & \frac{\partial^2 g(m,v)}{\partial m\,\partial v} \\ \frac{\partial^2 g(m,v)}{\partial v\,\partial m} & \frac{\partial^2 g(m,v)}{\partial v^2} \end{bmatrix} \tag{24}$$

- Need to check that the Hessian matrix $\mathbf{H}_g$ on $g(m,v)$ for $m = \hat{m}_{\mathrm{ML}}$ and $v = \hat{v}_{\mathrm{ML}}$ is non-negative definite (left as an exercise for the reader).

Estimation of Non-random Parameters: Examples

- Estimation problem:

    3 Estimate m, v. Check bias of estimates.

$$\mathbb{E}\left[\hat{v}_{\mathrm{ML}}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{m}_{\mathrm{ML}})^2\right] \tag{25}$$

$$= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} y_k^2\right] + \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\hat{m}_{\mathrm{ML}}^2\right] - \mathbb{E}\left[\frac{2}{N}\sum_{k=1}^{N}(\hat{m}_{\mathrm{ML}}\, y_k)\right] \tag{26}$$

$$= \ldots = \frac{N-1}{N} v \neq v \text{ (biased estimate)} \tag{27}$$

- That is why numerical packages utilise the following, non-ML, unbiased variance estimate:

$$\frac{1}{N-1}\sum_{k=1}^{N}(y_k - \hat{m}_{\mathrm{ML}})^2$$

Performance of Non-random Parameter Estimation

- Apart from bias, we need the mean square error (MSE) and the corresponding error matrix:

$$\mathbf{C}_{\mathrm{E}} = \mathbb{E}\left[\left(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\right)\left(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\right)^{\mathrm{T}}\right] \tag{28}$$

$$\mathrm{MSE} = \mathrm{Trace}\left(\mathbf{C}_{\mathrm{E}}\right) = \mathbb{E}\left[||\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})||_2^2\right] \tag{29}$$

- Denote the gradient (vector) $\nabla_x = [\frac{\partial}{\partial x_1} \quad \frac{\partial}{\partial x_2} \dots \frac{\partial}{\partial x_m}]^{\mathrm{T}}$ and the Hessian (matrix) $\nabla_x \nabla_x^{\mathrm{T}}$.

Non-random Parameter Estimation:Cramer-Rao Bound

- $m \times m$ Fisher Information matrix $\mathbf{J}(\mathbf{x})$ characterises the information in $\mathbf{y} \in \mathbb{R}^n$ about the parameter vector $\mathbf{x} \in \mathbb{R}^n$:

$$\mathbf{J}(\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{y}} \left[\left[\nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \right] \left[\nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \right]^{\mathrm{T}} \right] \tag{30}$$

- It turns out that

$$\mathbf{J}(\mathbf{x}) = -\mathbb{E}_{\mathbf{y}} \left[\underbrace{\nabla_x \nabla_x^{\mathrm{T}} \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})}_{\text{Hessian of the log-likelihood}} \right] \tag{31}$$

Theorem

For any ***unbiased*** *estimator* $\hat{\mathbf{x}}(\mathbf{y})$*, the MSE is bounded from the Cramer-Rao bound, which stems from the diagonal elements of the inverse Fisher Information matrix:*

$$\mathbb{E}\left[||\mathbf{x}_i - \hat{\mathbf{x}}(\mathbf{y})_i||_2^2 \right] \geq \left[\mathbf{J}^{-1}(\mathbf{x}) \right]_{ii}, \tag{32}$$

where $\mathbf{a}_i, \mathbf{A}_{ii}$ *denotes the* $i-th$ *element and* $i-th$ *diagonal element of vector* $\mathbf{a}$ *and matrix* $\mathbf{A}$*, respectively.*

- Assume $\mathbf{y} = [y_1 \;\; y_2 \ldots y_N]^{\mathrm{T}}$, with $\{y_k\}$ i.i.d. and $y_k \sim \mathcal{N}(m, v)$.

$$\ln\left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v)\right] = -\frac{N}{2}\ln(2\pi v) - \frac{1}{2v}\sum_{k=1}^{N}(y_k - m)^2$$

$$\frac{\partial}{\partial m}\ln\left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v)\right] = +\frac{1}{v}\sum_{k=1}^{N}(y_k - m) \tag{33}$$

$$\frac{\partial^2}{\partial m^2}\ln\left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v)\right] = -\frac{N}{v} \tag{34}$$

$$\frac{\partial}{\partial v}\ln\left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v)\right] = -\frac{N}{2v} + \frac{1}{2v^2}\sum_{k=1}^{N}(y_k - m)^2 \tag{35}$$

$$\frac{\partial^2}{\partial v^2}\ln\left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v)\right] = +\frac{N}{2v^2} - \frac{1}{v^3}\sum_{k=1}^{N}(y_k - m)^2 \tag{36}$$

$$\frac{\partial^2}{\partial m\,\partial v}\ln\left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v)\right] = -\frac{1}{v^2}\sum_{k=1}^{N}(y_k - m) \tag{37}$$

Cramer-Rao Bound Example

- ...will use the Hessian version of $\mathbf{J}$. Thus,

$$-\mathbb{E}_{\mathbf{y}}\left[\frac{\partial^2}{\partial m^2}\ln\left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v)\right]\right] = -\mathbb{E}_{\mathbf{y}}\left[-\frac{N}{v}\right] = \frac{N}{v} \tag{38}$$

$$\begin{aligned}-\mathbb{E}_{\mathbf{y}}\left[\frac{\partial^2}{\partial v^2}\ln\left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v)\right]\right] &= -\mathbb{E}_{\mathbf{y}}\left[+\frac{N}{2v^2} - \frac{1}{v^3}\sum_{k=1}^{N}(y_k - m)^2\right] = \\ &= +\frac{N}{2v^2}\end{aligned} \tag{39}$$

$$-\mathbb{E}_{\mathbf{y}}\left[\frac{\partial^2}{\partial m\,\partial v}\ln\left[f_{\mathbf{y}|m,v}(\mathbf{y}|m,v)\right]\right] = -\mathbb{E}_{\mathbf{y}}\left[-\frac{1}{v^2}\sum_{k=1}^{N}(y_k - m)\right] = 0 \tag{40}$$

Cramer-Rao Bound Example

- Now, Fisher Information matrix and its inverse can be calculated:

$$\mathbf{J}(m,v) = \begin{bmatrix} \frac{N}{v} & 0 \\ 0 & \frac{N}{2v^2} \end{bmatrix} \Leftrightarrow \mathbf{J}^{-1}(m,v) = \begin{bmatrix} \frac{v}{N} & 0 \\ 0 & \frac{2v^2}{N} \end{bmatrix} \tag{41}$$

- Therefore, for any unbiased estimate of m, v,

$$\mathbb{E}\left[(m - \hat{m}(\mathbf{y})^2)\right] \geq \frac{v}{N} \tag{42}$$

$$\mathbb{E}\left[(v - \hat{v}(\mathbf{y})^2)\right] \geq \frac{2v^2}{N} \tag{43}$$

Schur Complement Properties

- Before offering the derivation, we first list some basic properties. For any symmetric matrix $\mathbf{M}$ of the following form:[1]

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^{\mathrm{T}} & \mathbf{C} \end{bmatrix} \tag{44}$$

- If $\mathbf{C}$ is invertible then:
  1. $\mathbf{M} > \mathbf{0}$ iff $\mathbf{C} > \mathbf{0}$ and $\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^{\mathrm{T}} > \mathbf{0}$
  2. For $\mathbf{C} > \mathbf{0}$: $\mathbf{M} \geq \mathbf{0} \Leftrightarrow \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^{\mathrm{T}} \geq \mathbf{0}$
  3. Schur complement $\mathbf{M}|\mathbf{C} \overset{\triangle}{=} \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^{\mathrm{T}} \Rightarrow \det(\mathbf{M}) = \det(\mathbf{M}|\mathbf{C})\det(\mathbf{C})$
- If $\mathbf{A}$ is invertible then:
  1. $\mathbf{M} > \mathbf{0}$ iff $\mathbf{A} > \mathbf{0}$ and $\mathbf{C} - \mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B} > \mathbf{0}$
  2. For $\mathbf{A} > \mathbf{0}$: $\mathbf{M} \geq \mathbf{0} \Leftrightarrow \mathbf{C} - \mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B} \geq \mathbf{0}$
  3. Schur complement $\mathbf{M}|\mathbf{A} \overset{\triangle}{=} \mathbf{C} - \mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B} \Rightarrow \det(\mathbf{M}) = \det(\mathbf{M}|\mathbf{A})\det(\mathbf{A})$

[1] matrix inequality in the positive semi-definite sense: $\mathbf{A} \geq \mathbf{0} \Leftrightarrow \mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} \geq 0, \forall \mathbf{z}$

Cramer-Rao Bound Derivation

- Proof: First, we show the two equivalent forms of the Fisher $m \times m$ matrix $\mathbf{J}(\mathbf{x})$:

$$\int f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})d\mathbf{y} = 1 \tag{45}$$

$$\nabla_x \ln\left[f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] = \frac{1}{f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})}\nabla_x f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \Rightarrow \tag{46}$$

$$\nabla_x^{\mathrm{T}} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = \nabla_x^{\mathrm{T}} \ln\left[f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \tag{47}$$

$$\overset{(45)}{\Rightarrow} \int \nabla_x^{\mathrm{T}} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})d\mathbf{y} \overset{(47)}{=} \int \nabla_x^{\mathrm{T}} \ln\left[f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})d\mathbf{y} = \mathbf{0}$$

$$\overset{\nabla_x}{\Rightarrow} \int \nabla_x \nabla_x^{\mathrm{T}} \ln\left[f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})d\mathbf{y} +$$

$$+ \int \nabla_x \ln\left[f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] \nabla_x^{\mathrm{T}} \ln\left[f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})d\mathbf{y} = \mathbf{O} \tag{48}$$

$$\Rightarrow \mathbf{J}_{ij}(\mathbf{x}) = \mathbb{E}_{\mathbf{y}}\left[\frac{\partial}{\partial x_i}\ln\left[f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] \frac{\partial}{\partial x_j}\ln\left[f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right]\right]$$

$$= -\mathbb{E}_{\mathbf{y}}\left[\frac{\partial^2}{\partial x_i \partial x_j}\ln\left[f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right]\right] \tag{49}$$

Cramer-Rao Bound Derivation

- Next, we define the $2m \times 1$ vector $\mathbf{z}$, corresponding positive semi-definite matrix $\mathbf{C}_z$ and bias $\mathbf{b}(\mathbf{x})$:

$$\mathbf{z} = \begin{bmatrix} \underbrace{\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})}_{\mathbf{e}} - \mathbf{b}(\mathbf{x}) \\ \\ \nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \end{bmatrix}, \mathbf{C}_z = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} = \mathbb{E}\left[\mathbf{z}\mathbf{z}^{\mathrm{T}}\right] \geq 0 \tag{50}$$

- The following hold:
  1. $\mathbf{C}_{11} = \mathbb{E}\left[(\mathbf{e} - \mathbf{x})(\mathbf{e} - \mathbf{x})^{\mathrm{T}}\right] = \mathbf{C}_{\mathrm{E}} - \mathbf{b}(\mathbf{x})\mathbf{b}^{\mathrm{T}}(\mathbf{x})$.
  2. $\mathbf{C}_{22} = \mathbf{J}(\mathbf{x})$.
  3. $\mathbf{C}_{12} = \mathbf{C}_{21}^{\mathrm{T}} = \mathbb{E}\left[(\mathbf{x} - \hat{\mathbf{x}} - \mathbf{b}(\mathbf{x}))\nabla_x^{\mathrm{T}} \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] = -\mathbf{I}_m + \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x})$.

- The proof for 3. above follows.

Cramer-Rao Bound Derivation

- Next, we show $\mathbf{C}_{12} = \mathbf{C}_{21}^{\mathrm{T}} = -\mathbf{I}_m + \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x})$.

$$\mathbf{C}_{12} = \mathbf{C}_{21}^{\mathrm{T}} = \mathbb{E}\left[(\mathbf{x} - \hat{\mathbf{x}} - \mathbf{b}(\mathbf{x}))\nabla_x^{\mathrm{T}} \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] \tag{51}$$

$$= \int_{\mathbf{y}} (\mathbf{x} - \hat{\mathbf{x}} - \mathbf{b}(\mathbf{x}))\underbrace{\nabla_x^{\mathrm{T}} \ln\left[f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})}_{\nabla_x^{\mathrm{T}} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})} d\mathbf{y} \tag{52}$$

$$\mathbb{E}\left[(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{b}(\mathbf{x})\right] = \mathbf{0} \tag{53}$$

$$\Rightarrow \int_{\mathbf{y}} (\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{b}(\mathbf{x}) f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) d\mathbf{y} = 0 \tag{54}$$

$$\overset{\nabla_x^{\mathrm{T}}}{\Rightarrow} \underbrace{\int_{\mathbf{y}} (\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{b}(\mathbf{x}) \nabla_x^{\mathrm{T}} f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) d\mathbf{y}}_{\mathbf{C}_{12}} +$$

$$+ \int_{\mathbf{y}} (\mathbf{I}_m - \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x})) f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) d\mathbf{y} = \mathbf{O} \tag{55}$$

$$\Rightarrow \mathbf{C}_{12} = -\mathbf{I}_m + \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x}) \tag{56}$$

Cramer-Rao Bound Derivation

- In summary:
  1. $\mathbf{C}_{11} = \mathbb{E}\left[(\mathbf{e}-\mathbf{x})(\mathbf{e}-\mathbf{x})^{\mathrm{T}}\right] = \mathbf{C}_{\mathrm{E}} - \mathbf{b}(\mathbf{x})\mathbf{b}^{\mathrm{T}}(\mathbf{x})$.
  2. $\mathbf{C}_{22} = \mathbf{J}(\mathbf{x})$.
  3. $\mathbf{C}_{12} = -\mathbf{I}_m + \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x})$.

$$\mathbf{z} = \begin{bmatrix} \underbrace{\mathbf{x}-\hat{\mathbf{x}}(\mathbf{y})}_{\mathbf{e}} - \mathbf{b}(\mathbf{x}) \\ \\ \nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \end{bmatrix}, \mathbf{C}_z = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} = \mathbb{E}\left[\mathbf{z}\mathbf{z}^{\mathrm{T}}\right] \geq 0 \tag{57}$$

- We assume that positive semi-definite $\mathbf{J}(\mathbf{x})$ is invertible, i.e., it is positive definite. We also known that $\mathbf{C}_z$ is positive semi-definite.
- Thus, the Schur complement is also positive semi-definite:

$$\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}^{\mathrm{T}} \geq \mathbf{0} \tag{58}$$

Cramer-Rao Bound Derivation

- In summary:
  1. $\mathbf{C}_{11} = \mathbb{E}\left[(\mathbf{e}-\mathbf{x})(\mathbf{e}-\mathbf{x})^{\mathrm{T}}\right] = \mathbf{C}_{\mathrm{E}} - \mathbf{b}(\mathbf{x})\mathbf{b}^{\mathrm{T}}(\mathbf{x})$.
  2. $\mathbf{C}_{22} = \mathbf{J}(\mathbf{x})$.
  3. $\mathbf{C}_{12} = -\mathbf{I}_m + \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x})$.

$$\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}^{\mathrm{T}} \geq \mathbf{0} \Rightarrow \tag{59}$$

$$\mathbf{C}_{\mathrm{E}} - \mathbf{b}(\mathbf{x})\mathbf{b}^{\mathrm{T}}(\mathbf{x}) - \left(\mathbf{I}_m - \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x})\right)\mathbf{J}^{-1}(\mathbf{x})\left(\mathbf{I}_m - \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x})\right)^{\mathrm{T}} \geq \mathbf{0} \tag{60}$$

- For unbiased estimator, i.e., $\mathbf{b}(\mathbf{x}) = \mathbf{0}$, the above is simplified to:

$$\mathbf{C}_{\mathrm{E}} - \mathbf{J}^{-1}(\mathbf{x}) \geq \mathbf{0}$$

which completes the proof, if we consider that the diagonal elements of a positive semi-definite matrix are non-negative. ■

Existence of Efficient Estimator - Some Properties

- From Cramer-Rao proof, the following positive semi-definite matrix was utilized:

$$\mathbf{C}_z = \begin{bmatrix} \mathbf{C}_{11} = \mathbf{C}_{\mathrm{E}} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} = \mathbf{J}(\mathbf{x}) \end{bmatrix} = \mathbb{E}\left[\mathbf{z}\mathbf{z}^{\mathrm{T}}\right] \geq 0 \Rightarrow \tag{61}$$

$$\mathbf{C}_z^{-1} = \begin{bmatrix} \mathbf{S}^{-1} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}, \mathbf{z} = \begin{bmatrix} \mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{b} \\ \nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \end{bmatrix}, \tag{62}$$

- The following properties hold:
  1. $\mathbf{S} = \mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}$.
  2. $\mathbf{C}_z > 0 \Rightarrow \mathbf{S} > 0$.
  3. $\mathbf{C}_z \geq 0 \Rightarrow \mathbf{S} \geq 0$.
  4. $\mathbf{S} = \mathbf{0} \Rightarrow 2m \times 2m$ matrix $\mathbf{C}_z$ is of rank m.[2]

[2] $\mathbf{A} \geq 0$ in the positive-semi definite sense, i.e., $\mathbf{A} \geq 0 \Leftrightarrow \mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} \geq 0$

Efficient Estimator

- From property 3 in the previous slide, it follows for any biased estimator, with bias $\mathbf{b}(\mathbf{x})$:

$$\mathbf{S} = \mathbf{C}_{\mathrm{E}} - \mathbf{b}(\mathbf{x})\mathbf{b}(\mathbf{x})^{\mathrm{T}} - \left(\mathbf{I}_m - \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x})\right)\mathbf{J}^{-1}\left(\mathbf{I}_m - \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x})\right)^{\mathrm{T}} \geq 0 \tag{63}$$

- For an unbiased estimator (i.e., $\mathbf{b}(\mathbf{x}) = \mathbf{0}$), the above leads to the Cramer-Rao bound:

$$\mathbf{S} \geq \mathbf{0} \Leftrightarrow \mathbf{C}_{\mathrm{E}} \geq \mathbf{J}^{-1}. \tag{64}$$

- Efficient estimator is the unbiased estimator for which $\mathbf{C}_{\mathrm{E}} \equiv \mathbf{J}^{-1}$.
- Thus, for an efficient estimator it holds that $\mathbf{S} = \mathbf{0}$; from property 4 in the previous slide, matrix $\mathbf{C}_z$ is of rank m for an efficient estimator.

Efficient Estimator

- For an efficient estimator it holds that $\mathbf{S} = \mathbf{0}$; from property 4 in the previous slide, matrix $\mathbf{C}_z$ is of rank m for an efficient estimator.
- ...the above means that m rows can be written as a linear combination of the other m rows; having in mind the definition of $\mathbf{C}_z$ and the definition of $\mathbf{z}$, the above can be stated as follows:

$$\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}) = \mathbf{M}\,\nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \Rightarrow \tag{65}$$

$$(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}))\nabla_x^{\mathrm{T}} \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = \mathbf{M}\,\nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\nabla_x^{\mathrm{T}} \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \Rightarrow$$

$$\mathbb{E}_{\mathbf{y}}\left[(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y}))\nabla_x^{\mathrm{T}} \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right] = \mathbf{M}\,\mathbb{E}_{\mathbf{y}}\left[\nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\nabla_x^{\mathrm{T}} \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\right]$$

$$\Rightarrow \mathbf{C}_{12} = \mathbf{M}\,\mathbf{J}$$

$$\Rightarrow \mathbf{M} = \mathbf{C}_{12}\,\mathbf{J}^{-1} = \left(-\mathbf{I}_m + \nabla_x^{\mathrm{T}}\mathbf{b}(\mathbf{x})\right)\mathbf{J}^{-1} = -\mathbf{J}^{-1} \tag{66}$$

$$\Rightarrow \hat{\mathbf{x}}(\mathbf{y}) = \mathbf{x} - \mathbf{M}\,\nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \tag{67}$$

$$\Rightarrow \hat{\mathbf{x}}(\mathbf{y}) = \mathbf{x} + \mathbf{J}^{-1}\,\nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \tag{68}$$

Efficient Estimator Existence

- Thus, an efficient estimator has the following form:

$$\Rightarrow \hat{\mathbf{x}}(\mathbf{y}) = \mathbf{x} + \mathbf{J}^{-1} \nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) \tag{69}$$

- The left-hand side (LHS) is independent of $\mathbf{x}$; thus, an efficient estimator exists iff the right-hand side (RHS) of the above equation is independent of $\mathbf{x}$.
- Notice that $\nabla_x \ln f_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = \mathbf{0} \Rightarrow \hat{\mathbf{x}}(\mathbf{y}) = \mathbf{x}$.
- Remarks follow:
  1. If an efficient exists, it must be a stationary point of the likelihood function; if there is only one such point, it must be the ML estimator.
  2. If the likelihood function has a single maximum and the estimator is efficient, it must be the ML estimator.
  3. ...the above does not mean that all ML estimators are efficient... they may not be!
- ...will see an example at next lecture.

References

[1] Bernard C. Levy, Principles of Signal Detection and Parameter Estimation, Springer 2008.
[2] Instructor notes.

Thank you!

Detection & Estimation Theory: Lectures 17-18

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

Technical University of Crete

“Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete” (MIS 5150766), under the call for proposals “Support of the Internationalization of Higher Education - Technical University of Crete” (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme “Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- ML Estimates and 1-1 Functions
- Phase Estimation Example
  - Geometric interpretation
- Sufficient Statistic
- UMVU Estimator
- Complete Sufficient Statistic
- The RBLS Theorem (& Proof)

ML Estimates and 1-1 Functions

- Suppose $\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y})$ and $\mathbf{z} = \mathbf{g}(\mathbf{x})$ with $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{z})$, i.e., $\mathbf{g}(\mathbf{x}) = \mathbf{z}$ is a "1-1" mapping.
- Thus, $f_{\mathbf{y}}(\mathbf{y}|\mathbf{z}) = f_{\mathbf{y}}(\mathbf{y}|\underbrace{\mathbf{g}^{-1}(\mathbf{z})}_{\mathbf{x}})$

- Then if $\hat{\mathbf{x}}_{\mathrm{ML}} = \mathbf{g}^{-1}(\hat{\mathbf{z}}) \Rightarrow \hat{\mathbf{z}}_{\mathrm{ML}} = \mathbf{g}(\hat{\mathbf{x}}_{\mathrm{ML}})$.
- However, the transformation does not preserve unbiasedness or efficiency.

Phase estimation example

Now let's examine a phase estimation example:

$$\mathbf{y} = \begin{bmatrix} y_c \\ y_s \end{bmatrix} = A \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} + \mathbf{v},\ \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_2),$$

where σ and A are known and θ is unknown.

Thus, $\mathbf{y} \sim N\left(\overbrace{A \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}}^{\boldsymbol{\mu}}, \sigma^2 \mathbf{I}_2\right)$.

So,

$$\begin{aligned}
f_{\mathbf{y}}(\mathbf{y}|\theta) &= \frac{1}{\sqrt{(2\pi)^2 \sigma^4}} e^{-\frac{1}{2\sigma^2}\|\mathbf{y}-\boldsymbol{\mu}\|^2} \\
\Rightarrow \ln f_{\mathbf{y}}(\mathbf{y}|\theta) &= -\frac{1}{2} \ln (2\pi)^2 \sigma^4 - \frac{1}{2\sigma^2}\left[(y_c - A\cos\theta)^2 + (y_s - A\sin\theta)^2\right] \\
\Rightarrow L(\mathbf{y}) &\stackrel{\triangle}{=} \ln f_{\mathbf{y}}(\mathbf{y}|\theta) \\
&= -\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(y_c^2 + y_s^2 + A^2) + \frac{1}{\sigma^2}(A y_c \cos\theta + A y_s \sin\theta)
\end{aligned}$$

Phase estimation example (cont.)

$$\frac{\partial L}{\partial \theta} = \frac{A}{\sigma^2}(-y_c \sin\theta + y_s \cos\theta) = \emptyset$$

$$\Rightarrow y_c \sin\theta = y_s \cos\theta \Rightarrow \tan\theta = \frac{y_s}{y_c}$$

$$\Rightarrow \hat{\theta}_{\mathrm{ML}} = \tan^{-1}\frac{y_s}{y_c}$$

$$\begin{aligned}\frac{\partial^2 L}{\partial \theta^2} &= \frac{A}{\sigma^2}(-y_c \cos\theta - y_s \sin\theta) \\ &= -\frac{A}{\sigma^2}(y_c \cos\theta + y_s \sin\theta)\end{aligned}$$

$$\begin{aligned}d(\theta) = -\mathbb{E}\left[\frac{\partial^2 L}{\partial \theta^2}\right] &= \frac{A}{\sigma^2}\mathbb{E}[y_c \cos\theta + y_s \sin\theta] \\ &= \frac{A}{\sigma^2}\mathbb{E}[A\cos^2\theta + A\sin^2\theta] = \frac{A^2}{\sigma^2}\end{aligned}$$

Phase estimation example (cont.)

$$\mathbb{E}[(\theta - \hat{\theta}(\mathbf{y})^2] \geq \mathrm{J}^{-1}(\theta) = \frac{\sigma^2}{A^2} \simeq \frac{1}{\mathrm{SNR}}$$

- $\mathbb{E}[\hat{\theta}_{\mathrm{ML}}(\mathbf{y})] = ?$
- Is $\hat{\theta}_{\mathrm{ML}}(\mathbf{y})$ efficient?

$$\text{Set} \qquad \left.\begin{aligned} y_c &= r\cos\phi \\ y_s &= r\sin\phi \end{aligned}\right\} \hat{\theta}_{\mathrm{ML}} = \tan^{-1}\frac{y_s}{y_c} = \phi$$

$$\left.\begin{aligned} r > 0, \quad y_c^2 + y_s^2 &= r^2 \\ \tan^{-1}\frac{y_s}{y_c} &= \phi \end{aligned}\right\} \qquad \begin{bmatrix} r \\ \phi \end{bmatrix} \leftarrow \begin{bmatrix} y_c \\ y_s \end{bmatrix}$$

Phase estimation example (cont.)

$$\underset{\begin{bmatrix} y_c \\ y_s \end{bmatrix} \to \begin{bmatrix} r \\ \phi \end{bmatrix}}{\text{Jacobian}} = \begin{bmatrix} \nabla_{\mathbf{y}}^{\intercal} r \\ \nabla_{\mathbf{y}}^{\intercal} \phi \end{bmatrix} = \begin{bmatrix} \frac{\partial r}{\partial y_c} & \frac{\partial r}{\partial y_s} \\ \frac{\partial \phi}{\partial y_c} & \frac{\partial \phi}{\partial y_s} \end{bmatrix} = \begin{bmatrix} \frac{\cancel{2} y_c}{\cancel{2}\sqrt{y_c^2+y_s^2}} & \frac{\cancel{2} y_s}{\cancel{2}\sqrt{y_c^2+y_s^2}} \\ \frac{-y_s \frac{1}{y_c^2}}{1+(\frac{y_s}{y_c})^2} & \frac{\frac{1}{y_c}}{1+(\frac{y_s}{y_c})^2} \end{bmatrix}$$

So,

$$|\det(\text{Jacobian})| = \left| \frac{1}{\sqrt{y_c^2+y_s^2} \cdot \left(1+(\frac{y_s}{y_c})^2\right)} + \frac{\frac{y_s^2}{y_c^2}}{\left(1+\left(\frac{y_s}{y_c}\right)^2\right) \cdot \sqrt{y_c^2+y_s^2}} \right| = \frac{1}{\sqrt{y_c^2+y_s^2}}$$

Phase estimation example (cont.)

$$\text{Hence,} \quad f_{r,\phi}(r,\phi|\theta) = \left.\frac{f_{\mathbf{y}}(\mathbf{y})}{|\text{Jacobian}|}\right|_{\substack{y_c \,=\, r\cos\phi \\ y_s \,=\, r\sin\phi}}$$

$$= \frac{\frac{1}{\sqrt{(2\pi)^2(\sigma^2)^2}}\, e^{-\frac{1}{2\sigma^2}\|\mathbf{y}-\boldsymbol{\mu}\|^2}}{\frac{1}{\sqrt{y_c^2+y_s^2}}}$$

$$= \frac{\sqrt{y_c^2+y_s^2}}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}[y_c^2+y_s^2+A^2-2Ay_c\cos\theta-2Ay_s\sin\theta]}$$

$$= \frac{r}{2\pi\sigma^2}\, \underbrace{e^{-\frac{1}{2\sigma^2}[r^2+A^2-2Ar\cos(\phi-\theta)]}}_{\text{even function of } \phi-\theta}$$

$$\text{Thus,} \quad \mathbb{E}[\hat{\theta}_{\text{ML}}] = \mathbb{E}[\phi] = \theta + \cancelto{\emptyset}{\mathbb{E}[\phi-\theta]} = \theta$$

Therefore $\hat{\theta}_{\text{ML}}$ is unbiased.

Phase estimation example (cont.)

Remember that

$$\hat{\mathbf{x}}(\mathbf{y}) = \mathbf{x} + \mathrm{J}^{-1}(\mathbf{x})\nabla_{\mathbf{x}} \ln\left(f_{\mathbf{y}}(\mathbf{y}|\mathbf{x})\right)$$

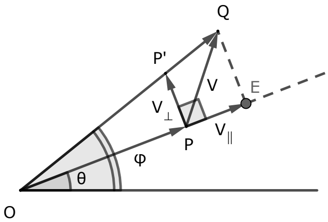
unbiased efficient estimator exists if and only if the RHS does not depend on $\mathbf{x}$.

$$\hat{\mathbf{x}}(\mathbf{y}) = \theta + \frac{\sigma^2}{A^2}\frac{A}{\sigma^2}(-y_c \sin\theta + y_s \cos\theta) = \theta + \frac{r}{A}\sin\left(\phi - \theta\right)$$

As long as $A \simeq r \rightarrow \phi \simeq \theta$ and since $\hat{\theta}_{\mathrm{ML}}(\mathbf{y}) = \phi \simeq \theta$ then,
$\theta + \frac{r}{A}(\phi - \theta) = \phi = \hat{\theta}_{\mathrm{ML}}(\mathbf{y})$
Thus,

$$\mathbb{E}[(\phi - \theta)^2] = \mathbb{E}[(\hat{\theta}_{\mathrm{ML}} - \theta)^2] \simeq \mathrm{J}^{-1} = \frac{\sigma^2}{A^2} = \frac{1}{\mathrm{SNR}}$$

Geometric interpretation



- $|\overrightarrow{\mathrm{OP}}| = A, \ \overrightarrow{\mathrm{OQ}} = \mathbf{y}$
- $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_2)$
- $\mathbf{V} = \mathbf{V}_\parallel + \mathbf{V}_\perp$ (orthogonal vector subspace)
- $|\mathbf{V}_\perp|$ is $\mathcal{N}(0, \sigma^2)$

$$\sin(\phi - \theta) \simeq \phi - \theta \simeq \frac{|\mathbf{V}_\perp|}{A} \sim \mathcal{N}(0, \frac{\sigma^2}{A^2})$$

$$\Rightarrow \mathbb{E}[\phi - \theta] = \emptyset \quad \text{and} \quad \mathbb{E}[(\phi - \theta)^2] = \frac{\sigma^2}{A^2}$$

Thus, efficient estimator for high SNR.

Sufficient Statistic

- $\mathbf{y}$ is a sufficient statistic of $\mathbf{x}$ if $f_{\mathbf{y}|\mathbf{s}}$ is independent of $\mathbf{x}$ i.e., all information about $\mathbf{x}$ has been "squeezed" in $f_{\mathbf{s}}(\mathbf{s}|\mathbf{x})$ and there is no leftover information about $\mathbf{x}$ that could be extracted from $f_{\mathbf{y}|\mathbf{s}}$, which means that the latter is independent of $\mathbf{x}$.
- In practice, sufficient statistic $\mathbf{s}(\mathbf{y})$ can be directly found if $f_{\mathbf{y}}(\mathbf{y}|\mathbf{x})$ belongs to the exponential class of densities:

$$f_{\mathbf{y}}(\mathbf{y}|\mathbf{x}) = u(\mathbf{y}) \cdot \exp[\mathbf{x}^{\intercal}\mathbf{s}(\mathbf{y}) - t(\mathbf{t})],$$

  which includes discrete Poisson, Exponential and Gaussian distributions as special cases.

Example 1

- iid $\{y_k\}$'s, $y_k \sim \mathcal{N}(m,u)$, $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_{\mathrm{N}} \end{bmatrix}^\mathsf{T}$

$$f_{\mathbf{y}}(m,u) = \prod_{k=1}^{\mathrm{N}} \frac{1}{\sqrt{2\pi u}} \cdot e^{\frac{1}{2u}}(y_k - m)^2$$

$$= \left(\frac{1}{\sqrt{2\pi u}}\right)^{\mathrm{N}} \cdot e^{-\frac{1}{2u}\sum_{k=1}^{\mathrm{N}}(y_k - m)^2}$$

$$= \frac{1}{(2\pi u)^{\frac{\mathrm{N}}{2}}} \cdot e^{-\frac{1}{2u}\left[\sum_{k=1}^{\mathrm{N}} y_k^2 + \mathrm{N}m^2 - 2m\sum_{k=1}^{\mathrm{N}} y_k\right]}$$

$$= \frac{1}{(2\pi u)^{\frac{\mathrm{N}}{2}}} \cdot e^{-\frac{\mathrm{N}m^2}{2u}} \cdot e^{-\frac{1}{2u}\sum_{k=1}^{\mathrm{N}} y_k^2 + \frac{m}{u}\sum_{k=1}^{\mathrm{N}} y_k}$$

Example 1 (cont.)

$$f_{\mathbf{y}}(m,u) = \frac{1}{(2\pi u)^{\frac{\mathrm{N}}{2}}} \cdot e^{-\frac{\mathrm{N}m^2}{2u}} \cdot e^{-\frac{1}{2u}\sum_{k=1}^{\mathrm{N}} y_k^2 + \frac{m}{u}\sum_{k=1}^{\mathrm{N}} y_k}$$

$$= \frac{1}{(2\pi u)^{\frac{\mathrm{N}}{2}}} \cdot e^{-\frac{\mathrm{N}m^2}{2u}} \cdot \exp\left(\underbrace{\begin{bmatrix} \mathrm{N}\frac{m}{u} & -\frac{N}{2u} \end{bmatrix}}_{\mathbf{x}^\intercal} \underbrace{\begin{bmatrix} \frac{1}{\mathrm{N}}\sum_{k=1}^{\mathrm{N}} y_k \\ \frac{1}{\mathrm{N}}\sum_{k=1}^{\mathrm{N}} y_k^2 \end{bmatrix}}_{\mathbf{s}(\mathbf{y})=\begin{bmatrix} s_1 \\ s_2 \end{bmatrix}} \right)$$

i.e. $\mathbf{s}(\mathbf{y})$ is a sufficient statistic for estimating parameter $\mathbf{x}$.

Example 1 (cont.)

Sufficient statistic requires the definition of the unknown parameter

- if u = unknown $\rightarrow$ sufficient statistic is $s_1(\mathbf{y})$
- if m = unknown $\rightarrow$ sufficient statistic is $s_2(\mathbf{y})$

Notice that we have shown that:

$$\hat{m}_{\mathrm{ML}} = \frac{\sum y_k}{\mathrm{N}} = s_1(\mathbf{y})$$

$$\begin{aligned}
\hat{u}_{\mathrm{ML}} = \frac{\sum (y_k - \hat{m}_{\mathrm{ML}})^2}{\mathrm{N}} &= \frac{\sum y_k^2}{\mathrm{N}} + \hat{m}_{\mathrm{ML}}^2 - \frac{2\hat{m}_{\mathrm{ML}}}{\mathrm{N}} \sum y_k \\
&= \frac{\sum y_k^2}{\mathrm{N}} - \hat{m}_{\mathrm{ML}}^2 = s_2(\mathbf{y}) - s_1^2(\mathbf{y})
\end{aligned}$$

Example 2

- iid $\{y_k\}$'s, $y_k \sim$ exponential with known parameter $1/\theta$,
  $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_\mathrm{N} \end{bmatrix}^\mathsf{T}$

$$f_{\mathbf{y}|\theta} = \left(\frac{1}{\theta}\right)^\mathrm{N} \cdot e^{-\frac{1}{\theta}\sum y_k} \prod_{k=1}^{\mathrm{N}} u(y_k) \Rightarrow \left| \begin{array}{l} x = \dfrac{1}{\theta} \\ \\ s(\mathbf{y}) = \sum y_k \end{array} \right.$$

$$L(\mathbf{y}|\theta) = \ln\left[f(\mathbf{y}|\theta)\right] = -\mathrm{N}\ln\theta - \frac{s(\mathbf{y})}{\theta},\ y_k \geq \emptyset$$

$$\frac{\partial}{\partial\theta}L(\mathbf{y}|\theta) = -\frac{\mathrm{N}}{\theta} + \frac{s(\mathbf{y})}{\theta^2} = 0 \Rightarrow \hat{\theta}_\mathrm{ML}(\mathbf{y}) = \frac{s(\mathbf{y})}{\mathrm{N}}$$

$$\frac{\partial^2}{\partial\theta^2}L(\mathbf{y}|\theta) = \frac{\mathrm{N}}{\theta^2} - \frac{2s(\mathbf{y})}{\theta^3} = \frac{1}{\theta^2}\left(\mathrm{N} - \frac{2s(\mathbf{y})}{\theta}\right) = \frac{\mathrm{N}}{\theta^2}\left(1 - \frac{2\hat{\theta}}{\theta}\right)$$

Example 2 (cont.)

$$\mathbb{E}\left[\hat{\theta}_{\mathrm{ML}}(\mathbf{y})\right] = \frac{1}{\mathrm{N}}\sum \mathbb{E}[y_k] = \frac{\cancel{\mathrm{N}}\theta}{\cancel{\mathrm{N}}} = \theta, \text{ (unbiased estimate)}$$

$$\mathrm{J}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}L(\mathbf{y}|\theta)\right] = -\frac{\mathrm{N}}{\theta^2} + \frac{2}{\theta^3}\mathbb{E}[s(\mathbf{y})] = -\frac{\mathrm{N}}{\theta^2} + \frac{2}{\theta^3}\mathrm{N}\theta = \frac{\mathrm{N}}{\theta^2}$$

Thus $\mathbb{E}[(\theta - \hat{\theta}_{\mathrm{ML}})^2] \geq \mathrm{J}^{-1}(\theta) = \frac{\theta^2}{\mathrm{N}}$

$$\begin{aligned}
\mathbb{E}\left[(\theta - \hat{\theta}_{\mathrm{ML}})^2\right] &= \theta^2 + \mathbb{E}\left[\hat{\theta}^2_{\mathrm{ML}}\right] - 2\theta\mathbb{E}\left[\hat{\theta}_{\mathrm{ML}}\right] \\
&= \mathbb{E}\left[\hat{\theta}^2_{\mathrm{ML}}\right] - \theta^2 = \frac{1}{\mathrm{N}^2}\mathbb{E}\left[(\textstyle\sum y_i)^2\right] - \theta^2 \\
&= \frac{1}{\mathrm{N}^2}\left[\mathrm{N}\cdot\mathbb{E}\left[y_i^2\right] + 2\mathbb{E}^2\left[y_i\right]\cdot\binom{\mathrm{N}}{2}\right] - \theta^2
\end{aligned}$$

Example 2 (cont.)

Since $\mathbb{E}[y_i] = \theta = 1/\lambda$
and $\mathbb{E}[y_i^2] - \mathbb{E}^2[y_i] = 1/\lambda^2 = \theta^2 \Rightarrow \mathbb{E}[y_i^2] = 2\theta^2$:

$$\begin{aligned}\mathbb{E}\left[(\theta - \hat{\theta}_{\mathrm{ML}})^2\right] &= \frac{1}{\mathrm{N}^2}\left[\mathrm{N} \cdot 2\theta + 2\theta \cdot \frac{\mathrm{N}(\mathrm{N}-1)}{2}\right] - \theta^2 \\ &= \frac{1}{\mathrm{N}}\left(2\theta^2 + (\mathrm{N}-1)\theta^2\right) - \theta^2 = \frac{\theta^2}{\theta} \equiv \mathrm{J}^{-1}(\theta)\end{aligned}$$

Thus, in this case, the ML unbiased estimate is <u>efficient</u> and $s(\mathbf{y})$ for the parameter θ is $s(\mathbf{y}) = \sum y_k$.

Discussion on unbiased estimates

Set $\mathrm{J}(\hat{\mathbf{x}}, \mathbf{x}) = \mathbb{E}\left[\|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\|_2^2\right] = \mathrm{MSE}$
"Uniform minimum variance <u>unbiased</u> estimate $\hat{\mathbf{x}}_{\mathrm{UMVUE}}(\mathbf{y})$"

$\mathrm{J}(\hat{\mathbf{x}}_{\mathrm{UMVUE}}, \mathbf{x}) \leq \mathrm{J}(\hat{\mathbf{x}}, \mathbf{x})$ for all other unbiased $\hat{\mathbf{x}}$ estimate.

- How do we find it?
  - search for ML estimate, see if it is unbiased and see if it is efficient.
  - if that approach fails, look for <u>complete</u>, <u>sufficient statistic</u>, as well as an unbiased estimator $\check{\mathbf{x}}(\mathbf{y})$.
- Apply RBLS theorem: if $\mathbf{s}(\mathbf{y})$ is <u>complete</u> sufficient statistic, then the estimate $\hat{\mathbf{x}}(\mathbf{s})$ (stemming from RBLS) is a UMVUE of $\mathbf{x}$.

Complete Sufficient Statistic

- What is a <u>complete</u> sufficient statistic?

- Let $\mathbf{s}(\mathbf{y})$ is a sufficient statistic for parameter $\mathbf{x}$.
- $\mathbf{s}$ is complete if $\mathbb{E}\left[\mathbf{h}(\mathbf{s})\right] = \mathbf{0} \Leftrightarrow \mathbf{h}(\mathbf{s}) = \mathbf{0} \Leftrightarrow$ there is at <u>most</u> one unbiased

  estimator of $\mathbf{x}$ depending on $\mathbf{s}$ only.

Note: if $\mathbf{h}(\mathbf{s}) = \mathbf{0} \Rightarrow \mathbb{E}\left[\mathbf{h}(\mathbf{s})\right] = \mathbf{0}$ is trivial.
Obviously, $\mathbb{E}\left[\mathbf{h}(\mathbf{s})\right] = \mathbf{0} \Rightarrow \mathbf{h}(\mathbf{s}) = \mathbf{0}$ is non-trivial

Complete Sufficient Statistic

- How do we find sufficient statistics which are complete? In general it is hard.

- However, for $f_{\mathbf{y}}(\mathbf{y}|\mathbf{x}) = u(\mathbf{y}) \cdot \exp\left[\mathbf{x}^\intercal \mathbf{s}(\mathbf{y}) - t(\mathbf{x})\right]$, $\mathbf{s}(\mathbf{y})$ is complete sufficient statistic!
- ...the above includes Poisson, Exponential and Gaussian.

Rao-Blackwell-Lehmann-Sheffe (RBLS) Theorem

The Rao-Blackwell-Lehmann-Sheffe Theorem states that for an unbiased estimate $\check{\mathbf{x}}(\mathbf{y})$ of $\mathbf{x}$ and a sufficient statistic $\mathbf{s}(\mathbf{y})$, the estimate can be improved:

$$\text{If } \mathbb{E}[\check{\mathbf{x}}(\mathbf{y})] = \mathbf{x}, \tag{1}$$

$$\text{then } \hat{\mathbf{x}}(\mathbf{s}) = \mathbb{E}[\check{\mathbf{x}}(\mathbf{y})|\mathbf{s}] \text{ is unbiased}$$

$$\text{with } \hat{\mathbf{K}}(\mathbf{x}) \leq \check{\mathbf{K}}(\mathbf{x}) \text{ i.e., their differ. is positive semi-definite} \tag{2}$$

$$\text{and } \check{\mathbf{K}}(\mathbf{x}) = \mathbb{E}[(\mathbf{x} - \check{\mathbf{x}})(\mathbf{x} - \check{\mathbf{x}})^\mathsf{T}]$$

$$\hat{\mathbf{K}}(\mathbf{x}) = \mathbb{E}[(\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^\mathsf{T}]$$

- If $\mathbf{s}$ is complete then $\hat{\mathbf{x}}(\mathbf{s})$ is a uniformly minimum-variance unbiased estimator (UMVUE).

Proof of RBLS Theorem

Proof of RBLS Theorem:
Let's start from the last one and assume that (1), (2) are true:

- If $\hat{\mathbf{s}}$ is complete, then there is at most one unbiased estimate of $\mathbf{x}$ that depends on $\mathbf{s} : \hat{\mathbf{x}}(\mathbf{s})$.
- Suppose that there is a second $\hat{\mathbf{x}}_2(\mathbf{y})$ that achieves smaller $\hat{\mathbf{K}}_2(\mathbf{x}) < \hat{\mathbf{K}}(\mathbf{x})$.
- If we condition on $\mathbf{s}$, then we must get $\hat{\mathbf{x}}(\mathbf{s})$ with $\hat{\mathbf{K}}(\mathbf{x}) \leq \hat{\mathbf{K}}_2(\mathbf{x})$ which is a contradiction.
- Thus, there is no other estimator that minimizes the mean squared error, meaning that $\hat{\mathbf{x}}(\mathbf{s})$ is UMVUE.

Proof of RBLS Theorem (cont.)

Now lets prove that $\mathop{\mathbb{E}}_{\mathbf{s}}[\hat{\mathbf{x}}(\mathbf{s})] = \mathbf{x}$

$$\mathop{\mathbb{E}}_{\mathbf{s}}[\hat{\mathbf{x}}(\mathbf{s})] = \mathop{\mathbb{E}}_{\mathbf{s}}\left[\mathop{\mathbb{E}}_{\mathbf{y}|\mathbf{s}}[\check{\mathbf{x}}(\mathbf{y})|\mathbf{s}]\right] \overset{\triangle}{=} \mathbb{E}[\check{\mathbf{x}}] = \mathbf{x}$$

law of iterated/repeated expectation

Finally we need to prove that $\hat{\mathbf{K}}_{\mathbf{x}} \leq \check{\mathbf{K}}_{\mathbf{x}}$:

$$\begin{aligned}
\check{\mathbf{K}}_{\mathbf{x}} &= \mathbb{E}[(\mathbf{x} - \check{\mathbf{x}})(\mathbf{x} - \check{\mathbf{x}})^{\mathsf{T}}] \\
&= \mathbb{E}[(\underline{\mathbf{x} - \hat{\mathbf{x}}} + \hat{x} - \check{\mathbf{x}})(\underline{\mathbf{x} - \hat{\mathbf{x}}} + \hat{\mathbf{x}} - \check{\mathbf{x}})^{\mathsf{T}}] \\
&= \hat{\mathbf{K}}_{\mathbf{x}} + \cancelto{\emptyset^*}{\mathbb{E}\left[(\mathbf{x} - \hat{\mathbf{x}})\underbrace{(\hat{\mathbf{x}} - \check{\mathbf{x}})}_{\mathbf{g}(\mathbf{y})}{}^{\mathsf{T}}\right]} + \cancelto{\emptyset^*}{\mathbb{E}\left[\underbrace{(\hat{\mathbf{x}} - \check{\mathbf{x}})}_{\mathbf{g}(\mathbf{y})}(\mathbf{x} - \hat{\mathbf{x}})^{\mathsf{T}}\right]} + \\
&\quad + \mathbb{E}[(\hat{\mathbf{x}} - \check{\mathbf{x}})(\hat{\mathbf{x}} - \check{\mathbf{x}})] \qquad (3)
\end{aligned}$$

---

* $\hat{\mathbf{x}} = \mathbb{E}[\check{\mathbf{x}}(\mathbf{y})|\mathbf{s}]$ is the MSE estimate of $\check{\mathbf{x}}(\mathbf{y})$ given $\mathbf{s}$ thus $\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})$ is orthogonal to any function of $\mathbf{s}$ (or $\mathbf{y}$).

Proof of RBLS Theorem (cont.)

Proof.

Thus from (3) we are left with:

$$\check{\mathbf{K}}_{\mathbf{x}} = \hat{\mathbf{K}}_{\mathbf{x}} + \mathbb{E}[(\hat{\mathbf{x}} - \check{\mathbf{x}})(\hat{\mathbf{x}} - \check{\mathbf{x}})]$$
$$\Rightarrow \check{\mathbf{K}}_{\mathbf{x}} - \hat{\mathbf{K}}_{\mathbf{x}} = \text{covariance matrix}$$
$$\Rightarrow \check{\mathbf{K}}_{\mathbf{x}} - \hat{\mathbf{K}}_{\mathbf{x}} \geq 0 \text{, i.e., } \check{\mathbf{K}}_{\mathbf{x}} - \hat{\mathbf{K}}_{\mathbf{x}} \text{ is positive semi-definite.}$$

□

Example

- iid s $y_k = \begin{bmatrix} y_1 & y_2 & \dots & y_{\mathrm{N}} \end{bmatrix}^{\mathsf{T}}$, $y_k \sim \frac{1}{\theta} \cdot e^{-\frac{y_k}{\theta}}$

$$f(\mathbf{y}|\theta) = \frac{1}{\theta^{\mathrm{N}}} \exp\left(-\frac{s(\mathbf{y})}{\theta}\right) \prod_{k=1}^{\mathrm{N}} u(y_k),\ s(\mathbf{y}) = \sum_{k=1}^{\mathrm{N}} y_k$$

$$\hat{\theta}_{\mathrm{ML}}(\mathbf{y}) = \frac{s(\mathbf{y})}{\mathrm{N}}$$

$$\check{\theta}(\mathbf{y}) = y_1 \text{ since } \mathbb{E}[y_1] = \theta \quad \text{(unbiased estimate)}$$

$$s(\mathbf{y}) = \sum y_k = \underline{\text{complete}} \text{ sufficient statistic}$$

Thus, according to RBLS,

$$\hat{\theta}(s) = \mathop{\mathbb{E}}_{\mathbf{y}}\left[\mathbf{x}(\check{\mathbf{y}})|\mathbf{s}\right] = \mathop{\mathbb{E}}_{\mathbf{y}_1}\left[y_1|\mathbf{s}\right] \quad \text{is an UMVUE}$$

$$= \int y_1 \cdot f_{y_1|s}(y_1|s)\, dy_1$$

Example (cont.)

- Thus, I need to find the $f_{y_1|s}$, $s = \sum_{k=1}^{N} y_k$

$$\text{Set } \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ s \end{bmatrix}}_{\tilde{\mathbf{y}}} = \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 1 & 1 & \dots & 1 & 1 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix}}_{\mathbf{y}}$$

$$\text{Jacobian} = \begin{bmatrix} \nabla_{\mathbf{y}}^{\intercal} y_1 \\ \nabla_{\mathbf{y}}^{\intercal} y_2 \\ \vdots \\ \nabla_{\mathbf{y}}^{\intercal} y_{N-1} \\ \nabla_{\mathbf{y}}^{\intercal} s \end{bmatrix} \overset{*}{=} \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 1 & 1 & \dots & 1 & 1 \end{bmatrix} = \mathbf{A}$$

---

* $\mathbf{A} \rightarrow$ upper diagonal $\Rightarrow \det(\mathbf{A})$ = product of diagonal elements, and $\det(\mathbf{A}) = 1$

Example (cont.)

$$\nabla_{\mathbf{y}} = \begin{bmatrix} \frac{\partial}{\partial y_1} \\ \frac{\partial}{\partial y_2} \\ \vdots \\ \frac{\partial}{\partial y_{\mathrm{N}}} \end{bmatrix} \quad \text{thus} \quad \begin{aligned} f_{\tilde{\mathbf{y}}}(\tilde{\mathbf{y}}) &\triangleq f_{y_1, y_2, \ldots, y_{\mathrm{N}-1}, s}(\tilde{\mathbf{y}}) \\ &= \left.\frac{f_{\mathbf{y}}(\mathbf{y})}{|\det(\text{Jacobian})|}\right|_{\mathbf{y}=\mathbf{A}^{-1}\cdot\tilde{\mathbf{y}}} \end{aligned}$$

$$s = \sum_{k=1}^{\mathrm{N}-1} y_k + y_n \Rightarrow y_n = s - \sum_{k=1}^{\mathrm{N}-1} y_k$$

$$\Rightarrow f_{y_1, y_2, \ldots, y_{\mathrm{N}-1}, s}(\tilde{\mathbf{y}}) = \frac{1}{\theta^{\mathrm{N}}} \cdot \exp\left(-\frac{s}{\theta}\right) \left(\prod_{k=1}^{\mathrm{N}-1} u(y_k)\right) u\left(s - \sum_{k=1}^{\mathrm{N}-1} y_k\right)$$

Example (cont.)

$$\begin{aligned}
f_{y_1,s} &= \int f_{y_1,y_2,\ldots,y_{\mathrm{N}-1},s}(\tilde{\mathbf{y}})\, dy_2 dy_3 \ldots dy_{\mathrm{N}-1} \\
&= \tfrac{1}{\theta^{\mathrm{N}}} \exp\!\left(-\tfrac{s}{\theta}\right) \int \textstyle\prod_{k=1}^{\mathrm{N}-2} u(y_k) \left[\int u(y_{\mathrm{N}-1}) u\!\left(s - \sum_{k=1}^{\mathrm{N}-1} y_k\right) dy_{\mathrm{N}-1}\right] dy_2 dy_3 \ldots dy_{\mathrm{N}-2} \\
&\overset{*}{=} \tfrac{1}{\theta^{\mathrm{N}}} \exp\!\left(-\tfrac{s}{\theta}\right) \int \textstyle\prod_{k=1}^{\mathrm{N}-2} u(y_k) \cdot u\!\left(s - \sum_{k=1}^{\mathrm{N}-2} y_k\right) \cdot \left(s - \sum_{k=1}^{\mathrm{N}-2} y_k\right) dy_2 dy_3 \ldots dy_{\mathrm{N}-2} \\
&= \cdots = \frac{1}{\theta^{\mathrm{N}}} \exp\left(-\frac{s}{\theta}\right) \frac{(s - y_1)^{\mathrm{N}-2}}{(\mathrm{N}-2)!} u(s - y_1) u(y_1)
\end{aligned}$$

$s(y) : \sum_{k=1}^{\mathrm{N}} y_k =$ sum of N identically distributed exponentials with parameters $\frac{1}{\theta}$ eachs $\Rightarrow s$: Gamma distribution with $f(s) = \Gamma(\mathrm{N}, \theta) = s^{\mathrm{N}-1} \frac{\exp(-s/\theta)}{\Gamma(\mathrm{N}) \cdot \theta^{\mathrm{N}}}$, $\mathbb{E}[s] = \mathrm{N} \cdot \theta$, $\mathrm{Var}(s) = \mathrm{N} \cdot \theta^2$

$^*y_{\mathrm{N}-1} \geq 0,\ s - \sum_{k=1}^{\mathrm{N}-1} y_k \geq 0 \Rightarrow s - \sum_{k=1}^{\mathrm{N}-2} y_k \geq y_{\text{N-1}} \geq 0$

Example (cont.)

- Alternatively, we could integrate $f_{y_1,s}$

$$
\begin{aligned}
\int f_{y_1,s}(y_1,s)\,dy_1 &= \frac{1}{\theta^{\mathrm{N}}}e^{-s/\theta}\frac{1}{(\mathrm{N}-2)!}\int_0^s (s-y_1)^{N-24}\,dy_1 = \\
&= \frac{1}{\theta^{\mathrm{N}}}e^{-s/\theta}\frac{1}{(\mathrm{N}-2)!}\left[-\frac{(s-y_1)^{\mathrm{N}-1}}{\mathrm{N}-1}\right]_0^s \\
&= \frac{1}{\theta^{\mathrm{N}}}e^{-s/\theta}\frac{1}{(\mathrm{N}-1)!}s^{\mathrm{N}-1} \\
&= \frac{1}{\theta^{\mathrm{N}}}e^{-s/\theta}\frac{1}{\Gamma(\mathrm{N})}s^{\mathrm{N}-1}
\end{aligned}
$$

Example (cont.)

- Thus, $f_{y_1|s} = \frac{f_{y_1,s}}{f_s} = \dfrac{\cancel{\frac{1}{\theta^{\mathrm{N}}}}\cancel{e^{-s/\theta}}\frac{(s-y_1)^{\mathrm{N}-2}}{(\mathrm{N}-2)!}u(s-y_1)u(y_1)}{\cancel{\frac{1}{\theta^{\mathrm{N}}}}\frac{1}{(\mathrm{N}-1)!}s^{\mathrm{N}-1}\cancel{e^{-s/\theta}}}$

$$= \frac{(\mathrm{N}-1)}{s^{\mathrm{N}-1}}(s-y_1)^{\mathrm{N}-2}u(s-y_1)u(y_1)$$

- and $\mathbb{E}[y_1|s] = \int y_1 \cdot f_{y_1|s}(y_1|s)\,dy_1$

$$= \int_0^s y_1 \frac{(\mathrm{N}-1)}{s^{\mathrm{N}-1}}(s-y_1)^{\text{N-2}}\,dy_1 = \cdots = \frac{s}{\mathrm{N}}$$

Thus, the RBLS procedure provides $\hat{\theta}(s) = \frac{s}{\mathrm{N}} = \frac{\sum y_k}{\mathrm{N}} \equiv \hat{\theta}_{\mathrm{ML}}$, which can be shown to be both unbiased <u>and</u> efficient!

s complete $\Rightarrow$ at <u>most</u> one unbiased estimate which is a function of s only.

- Trick : find one unbiased estimate which is a function of s only, provided that s is also a complete sufficient statistic. That will be the UMVUE!

Below, we offer an example where transformation offers ML estimate, without preserving unbiased property.

$$\text{if } \hat{\mathbf{x}}(\mathbf{y}) = \hat{\mathbf{x}}_{\text{ML}}(\mathbf{y}),\ \mathbf{z} = \mathbf{g}(\mathbf{x}) \quad \text{1-1} \quad \left[\mathbf{x} = \mathbf{g}^{-1}(\mathbf{z})\right]$$

$$\text{then } \hat{\mathbf{z}}_{\text{ML}}(\mathbf{y}) = \mathbf{g}(\mathbf{x}(\hat{\mathbf{y}})) \quad \text{“ML estimate is parameterization independent”}$$

Example (cont.)

$$\hat{\theta}_{\mathrm{ML}} = \frac{s(\mathbf{y})}{\mathrm{N}} = \frac{\sum y_k}{\mathrm{N}}$$

$$\theta = \frac{1}{\lambda} \Rightarrow \lambda = \frac{1}{\theta} \Rightarrow \hat{\lambda}_{\mathrm{ML}}(\mathbf{y}) = \frac{1}{\hat{\theta}_{\mathrm{ML}}(\mathbf{y})} = \frac{\mathrm{N}}{s}$$

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathrm{N}}{s}\right] &= \int \frac{\mathrm{N}}{s} f_s(s)\, ds = \frac{\mathrm{N}}{\Gamma(\mathrm{N})\theta^{\mathrm{N}}} \int_{-\infty}^{+\infty} s^{\mathrm{N}-2} e^{-s/\theta} ds \\
&= \frac{\mathrm{N}}{\mathrm{N}-1}\frac{1}{\theta} \int \frac{1}{\theta^{\mathrm{N}-1}} \frac{1}{\Gamma(\mathrm{N}-1)} s^{\mathrm{N}-2} e^{-s/\theta}\, ds \\
&= \frac{\mathrm{N}}{\mathrm{N}-1}\frac{1}{\theta} = \frac{\mathrm{N}}{\mathrm{N}-1}\lambda, \text{ thus the estimator is biased}
\end{aligned}$$

Example (cont.)

- Set $\hat{\lambda}_0 = \frac{\mathrm{N}-1}{\mathrm{N}}\hat{\lambda}_{\mathrm{ML}} = \frac{\mathrm{N}-1}{s}, \quad \mathbb{E}[\hat{\lambda}_0] = \lambda$ (unbiased)
- $\hat{\lambda}_0(s) =$ in a function of s (which is a complete sufficient statistic) only, thus $\hat{\lambda}_0(s) = \hat{\lambda}_{\mathrm{UMVUE}}$
- It can be easily shown that:

$$\left.\begin{array}{l} \text{CRLB: } \lambda^2/\mathrm{N} \\ \mathbb{E}[(\lambda-\lambda_{\mathrm{UMVUE}})^2] = \lambda^2/(\mathrm{N}-2) \\ \mathbb{E}[(x-\hat{\lambda}_{\mathrm{ML}})^2] = \dfrac{\mathrm{N}+2}{(\mathrm{N}-1)(\mathrm{N}-2)}\lambda^2 \end{array}\right\} \begin{array}{l} \mathbb{E}[(\lambda-\lambda_{\mathrm{ML}})^2] > \mathbb{E}[(\lambda-\lambda_{\mathrm{UMVUE}})^2] > \\ \qquad\qquad > \text{CRLB} \end{array}$$

<u>Important Remark</u>: the conditional mean given the <u>complete</u> sufficient statistic should always give the same estimator.

References

Bernard C. Levy, Principles of Signal Detection and Parameter Estimation, Springer 2008.

Thank you!

Detection & Estimation Theory: Lecture 19

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

Technical University of Crete

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

BLUE (Best Linear Unbiased Estimator)

BLUE: Problem Definition

- Suppose that we want to estimate parameter vector $\boldsymbol{\theta}_{(p\times 1)}$ based on measurements $\mathbf{y}_{(N\times 1)}$ with linear estimator:

$$\hat{\boldsymbol{\theta}} = \mathbf{A}_{(p\times N)} \cdot \mathbf{y} \tag{1}$$

- We require unbiased estimator:

$$\mathbb{E}\left[\hat{\boldsymbol{\theta}}\right] = \mathbf{A} \cdot \mathbb{E}\left[\mathbf{y}\right] = \boldsymbol{\theta} \tag{2}$$

that can be achieved if and only if $\mathbb{E}\left[\mathbf{y}\right] = \mathbf{H}_{(N\times p)} \cdot \boldsymbol{\theta}_{(p\times 1)} \Rightarrow$

$$\mathbf{A}_{(p\times N)} \cdot \mathbf{H}_{(N\times p)} = \mathbf{I}_p \tag{3}$$

  - $\mathrm{rank}(\mathbf{I}_p) = p = \min\left(\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{H})\right) = \min(N, p)$, thus $N \geq p$

- Set $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^{\mathrm{T}} \\ \mathbf{a}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{a}_p^{\mathrm{T}} \end{bmatrix}$ (4) and $\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_p \end{bmatrix}$ (5)

BLUE Derivation

- from (1) and (4) $\Rightarrow$
$$\hat{\theta}_i = \mathbf{a}_i^{\mathrm{T}} \cdot \mathbf{y} \qquad (6)$$
- from (4), (5) and (3) $\Rightarrow \mathbf{a}_i^{\mathrm{T}} \mathbf{h}_j = \delta_{ij}$
- $\mathrm{var}(\hat{\theta}_i) = \mathbb{E}\left[\left(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i]\right)^2\right] = \mathbb{E}\left[\left(\mathbf{a}_i^{\mathrm{T}} \mathbf{y} - \mathbf{a}_i^{\mathrm{T}} \mathbb{E}[\mathbf{y}]\right)^2\right]$

$$= \mathbb{E}\left[\left[\mathbf{a}_i^{\mathrm{T}} (\mathbf{y} - \mathbb{E}[\mathbf{y}])\right]^2\right]$$

$$= \mathbb{E}\left[\mathbf{a}_i^{\mathrm{T}} (\mathbf{y} - \mathbb{E}[\mathbf{y}]) (\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}} \mathbf{a}_i\right] = \mathbf{a}_i^{\mathrm{T}} \mathbf{K}_{\mathbf{y}} \ \mathbf{a}_i$$

$$\Rightarrow \mathrm{var}(\hat{\theta}_i) = \mathbf{a}_i^{\mathrm{T}} \mathbf{K}_{\mathbf{y}} \ \mathbf{a}_i \qquad (7)$$

BLUE Derivation

- Minimize $\mathrm{var}(\hat{\theta}_i) = \mathbf{a}_i^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}\ \mathbf{a}_i$, for $i = 1, 2, \cdots, p$ subject to the constraints $\mathbf{a}_i^{\mathrm{T}} \mathbf{h}_j = \delta_{ij},\ i, j \in \{1, 2, \cdots, p\}$
- We have p constraints for each $\mathbf{a}_i$. Since each $\mathbf{a}_i$ is free to assume any value, independently of the others, we actually have p separate minimization problems linked only by the constraints:

$$J_i = \mathbf{a}_i^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}\ \mathbf{a}_i + \sum_{j=1}^{p} \left(\lambda_j^{(i)} (\mathbf{a}_i^{\mathrm{T}} \mathbf{h}_j - \delta_{ij}) \right), \boldsymbol{\lambda}_i = \left[\lambda_1^{(i)}\ \lambda_2^{(i)} \cdots\ \lambda_p^{(i)} \right]^{\mathrm{T}}$$

$$\frac{\partial J_i}{\partial \mathbf{a}_i} = 2\mathbf{K}_{\mathbf{y}} \mathbf{a}_i + \sum_{j=1}^{p} \lambda_j^{(i)} \mathbf{h}_j = 2\mathbf{K}_{\mathbf{y}} \mathbf{a}_i + \mathbf{H} \boldsymbol{\lambda}_i = \mathbf{0}$$

$$\Rightarrow \frac{\partial J_i}{\partial \mathbf{a}_i} = \mathbf{0} \Rightarrow \mathbf{a}_i = -\frac{1}{2} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \boldsymbol{\lambda}_i \tag{8}$$

BLUE Derivation

- To find $\boldsymbol{\lambda}_i$, we need to exploit the constraints:

$$\mathbf{a}_i^{\mathrm{T}} \cdot \mathbf{h}_j = \mathbf{h}_j^{\mathrm{T}} \mathbf{a}_i = \delta_{ij}, \qquad j = 1, 2, \cdots, p$$

  where $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$.

- from (5) and the above $\Rightarrow \mathbf{H}^{\mathrm{T}} \cdot \mathbf{a}_i = \mathbf{e}_i$, which is a vector with all zeros apart from position i, where it is one:

$$\mathbf{H}^{\mathrm{T}} \cdot \mathbf{a}_i = \mathbf{e}_i \Rightarrow \mathbf{H}^{\mathrm{T}} \cdot \left(-\frac{1}{2} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \boldsymbol{\lambda}_i \right) = \mathbf{e}_i$$

$$\Rightarrow -\frac{1}{2} \boldsymbol{\lambda}_i = \underbrace{\left(\mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \right)^{-1}}_{\text{assuming invertibility of } \mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H}} \tag{9}$$

- from (8) and (9) $\Rightarrow$

$$\mathbf{a}_{i_{opt}} = \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \left(\mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \right)^{-1} \mathbf{e}_i$$

BLUE Derivation

$$\hat{\boldsymbol{\theta}} = \mathbf{A} \cdot \mathbf{y} = \begin{bmatrix} \mathbf{a}_1^{\mathrm{T}} \\ \mathbf{a}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{a}_p^{\mathrm{T}} \end{bmatrix} \cdot \mathbf{y} = \underbrace{\begin{bmatrix} \mathbf{e}_1^{\mathrm{T}} \\ \mathbf{e}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{e}_p^{\mathrm{T}} \end{bmatrix}}_{\mathbf{I}_p} \cdot \left(\mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \right)^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \cdot \mathbf{y}$$

$$\Rightarrow \hat{\boldsymbol{\theta}} = \left(\mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \right)^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \cdot \mathbf{y}$$

Example 1

- $\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, $\mathbf{w}$ has zero mean $\mathbb{E}[\mathbf{w}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] = \mathbf{K}_{\mathbf{y}}$, so $\mathbb{E}[\mathbf{y}] = \mathbf{H}\boldsymbol{\theta}$
- $$\begin{aligned}\hat{\boldsymbol{\theta}} - \mathbb{E}\left[\hat{\boldsymbol{\theta}}\right] &= \left(\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y} - \mathbb{E}\left[\hat{\boldsymbol{\theta}}\right] \\ &= \cancel{\left(\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{H}\boldsymbol{\theta}} - \cancelto{\boldsymbol{\theta}}{\mathbb{E}\left[\hat{\boldsymbol{\theta}}\right]} + \\ &\quad + \left(\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{w} \\ &= \left(\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{w}\end{aligned}$$
- Thus,

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}} - \mathbb{E}\left[\hat{\boldsymbol{\theta}}\right]\right)\left(\hat{\boldsymbol{\theta}} - \mathbb{E}\left[\hat{\boldsymbol{\theta}}\right]\right)^{\mathrm{T}}\right] = \\ &= \left(\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\underbrace{\mathbf{K}_{\mathbf{y}}\mathbf{K}_{\mathbf{y}}^{-1}}_{\mathbf{I}}\mathbf{H}\left(\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{H}\right)^{-1} \\ &= \left(\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{H}\right)^{-1} \equiv \mathbf{C}_{\hat{\boldsymbol{\theta}}}\end{aligned}$$

$$\Rightarrow \mathrm{var}(\hat{\theta}_i) = \left[\left(\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{H}\right)^{-1}\right]_{ii} \tag{10}$$

Example 1

- This could be also seen from

$$\hat{\theta}_i = \mathbf{a}_{i_{opt}}^{\mathrm{T}} \cdot \mathbf{y} = \mathbf{e}_i^{\mathrm{T}} \left(\mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \right)^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \cdot \mathbf{y}$$

and

$$\begin{aligned}
\mathrm{var}\hat{\theta}_i &= \mathbf{a}_{i_{opt}}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}} \; \mathbf{a}_{i_{opt}} \\
&= \mathbf{e}_i^{\mathrm{T}} \left(\mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \right)^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{K}_{\mathbf{y}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \left(\mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \right)^{-1} \mathbf{e}_i \\
&= \mathbf{e}_i^{\mathrm{T}} \left(\mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \right)^{-1} \mathbf{e}_i \\
&\overset{(7)}{=} \left[\left(\mathbf{H}^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \mathbf{H} \right)^{-1} \right]_{ii}
\end{aligned}$$

Remarks

- *Remark 1:* MVUE for linear Gaussian Case $\equiv$ BLUE, if $\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}, \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, then $\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y} \equiv$ BLUE is also UMVUE
- *Remark 2 (Gauss-Markov Theorem):* if the data are of the general linear model form $\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, where $\mathbf{H}$ is a <u>known</u> $N \times p$ matrix, $\boldsymbol{\theta}$ is a $p \times 1$ vector of parameters to be estimated, and $\mathbf{w}$ is a $N \times 1$ noise vector with zero mean and covariance $\mathbf{C}$[1], then BLUE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y} \quad \text{and} \quad \mathbf{C}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{H}\right)^{-1}$$

[1] N should be greater or equal than p ($N \geq p$)

Example 2

- $y[n] = A + w[n],\ n = 0, 1, \cdots, N-1 : w[n]$ white noise with variance σ^2 (not necessarily Gaussian)

$$\mathbb{E}\left[w[n]\right] = 0 \quad \text{and} \quad \mathbb{E}\left[w[n]w[n+m]\right] = \sigma^2 \delta[m]$$

$$\mathbf{y} = \underbrace{\begin{bmatrix} y[0] \\ y[1] \\ \vdots \\ y[N-1] \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}_{\mathbf{H}_{N\times 1}} \cdot A = \underbrace{\begin{bmatrix} w[0] \\ w[1] \\ \vdots \\ w[N-1] \end{bmatrix}}_{\mathbf{w}}$$

$$\mathbb{E}[\mathbf{y}] = \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}_{\mathbf{H}} \cdot A, \quad \mathbb{E}[\mathbf{w}] = 0 \quad \text{and} \quad \mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] = \sigma^2 \mathbf{I}_N$$

Example 2

- Thus,

$$\hat{\boldsymbol{\theta}} = \left(\frac{N}{\sigma^2}\right)^{-1} \cdot \frac{1}{\sigma^2} \sum_{n=0}^{N-1} y[n]$$

$$= \frac{1}{N} \sum_{n=1}^{N-1} y[n] = \hat{A}$$

$$\mathbb{E}\left[A - \hat{A}\right] = \left(\frac{1}{\sigma^2} N\right)^{-1} = \frac{\sigma^2}{N}$$

Example 3

- We had seen that for y_k i.i.d. $\sim \underbrace{\mathcal{N}(\cancel{m}^{0}, u)}_{m\ known} = \mathcal{N}(0, u)$

$$\hat{u}_{\text{ML}} = \frac{1}{N}\sum_{k=1}^{N} y_k^2 \quad \text{and} \quad \text{CRLB} = \frac{2u^2}{N}$$

- Is $\hat{u}_{\text{ML}}$ efficient?
- What is the BLUE estimate of u?

Example 3

- $\mathbb{E}[\hat{u}_{\mathrm{ML}}] = \frac{1}{N} N u = u$ unbiased
- $$\mathbb{E}\left[(\hat{u}_{\mathrm{ML}} - u)^2\right] = \mathbb{E}\left[\hat{u}_{\mathrm{ML}}^2\right] - u^2 = \frac{1}{N^2}\left(\sum_{k=1}^{N} y_k^2\right)^2 - u^2$$
$$= \frac{1}{N^2}\left(N\mathbb{E}[y_k^4] + \binom{N}{2} 2u^2\right) - u^2$$
$$\overset{2}{=} \frac{1}{N^2}\left(N \cdot 3u^2 + \frac{N!}{\cancel{2}!(N-2)!}\cancel{2}u^2\right) - u^2$$
$$= \frac{3u^2}{N} + \frac{N-1}{N}u^2 - u^2 = \frac{2u^2}{N} \equiv \mathrm{CRLB}$$
- Thus, $\hat{u}_{\mathrm{ML}}(\mathbf{y}) = \frac{1}{N}\sum_{k=1}^{N} y_k^2$ is efficient.

---

[2] $y \sim \mathcal{N}(0, u), \sigma_u = \sqrt{u} \Rightarrow$

$$\mathbb{E}[y^n] = \begin{cases} (\sigma_u)^n \cdot 1 \cdot 3 \cdots (n-1), & \text{n even,} \\ \emptyset, & \text{n odd,} \end{cases}$$

Example 3 - BLUE estimate

- Search for BLUE of u:

$$\hat{u}_{\text{BLUE}}(\mathbf{y}) = \mathbf{a}^{\mathrm{T}}\mathbf{y} = \sum_{k=1}^{N} a_k y_k$$

$$\mathbb{E}[\hat{u}_{\text{BLUE}}] = \sum a_k \cancelto{\emptyset}{\mathbb{E}[y_k]} = \emptyset \neq u \quad \text{Biased}$$

- However, we can use $z_k = y_k^2$ (data transformation) and test BLUE on the transformed data:[3]

$$\hat{u}_{\text{BLUE}}(\mathbf{z}) = \sum a_k z_k = \sum_{k=1}^{N} a_k y_k^2 \Rightarrow \tag{11}$$

$$\mathbb{E}[\hat{u}_{\text{BLUE}}] = \underbrace{\sum_{k=1}^{N} a_k u}_{\sum_{k=1}^{N} a_k = 1} = u \tag{12}$$

[3] z_k iid, $\sum a_k = 1 \Rightarrow a_k = \frac{1}{N}$

References

Bernard C. Levy, Principles of Signal Detection and Parameter Estimation, Springer 2008.

Thank you!

Detection & Estimation Theory: Lectures 20-21

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Outline

- Introduction to Composite Hypothesis Testing and UMP/GLRT
- GLRT: Examples and Properties
- Asymptotic Optimality of the GLRT

Composite Hypothesis Testing

- Composite Hypothesis Testing: Problem of both Detection and Estimation!
- Problem definition:
  1. $\mathcal{H}_0 : \mathbf{y} \sim f_{\mathbf{y}}\left(\mathbf{y}|\mathbf{x}, \mathcal{H}_0\right)$, $\mathbf{x} \in \mathcal{X}_0$
     $\mathcal{H}_1 : \mathbf{y} \sim f_{\mathbf{y}}\left(\mathbf{y}|\mathbf{x}, \mathcal{H}_1\right)$, $\mathbf{x} \in \mathcal{X}_1$
  2. $\mathbf{x}$ defines $\mathcal{H}_j$; if $\mathcal{X}_0 \equiv \mathcal{X}_1$ there would be no way to distinguish between $\mathcal{H}_0, \mathcal{H}_1 \Rightarrow$ no way to estimate $\mathbf{x}$
- Example: $y[k] = As[k] + v[k]$, $k \in \{1, \ldots, N\}$
  1. A unknown, v WGN with $v \sim \mathcal{N}\left(0, \sigma^2\right)$, σ^2 unknown
  2. $\mathbf{y} = \begin{bmatrix} y[1] \\ y[2] \\ \vdots \\ y[N] \end{bmatrix} = A \begin{bmatrix} s[1] \\ s[2] \\ \vdots \\ s[N] \end{bmatrix} + \begin{bmatrix} v[1] \\ v[2] \\ \vdots \\ v[N] \end{bmatrix}$
  3. $\mathcal{H}_0$: $A = 0$, σ^2 unknown $\Rightarrow \mathcal{X}_0 = \left\{\left(A, \sigma^2\right) : A = 0\right\}$
     $\mathcal{H}_1$: $A \neq 0$, σ^2 unknown $\Rightarrow \mathcal{X}_1 = \left\{\left(A, \sigma^2\right) : A \neq 0\right\}$
  4. σ^2 is common to both hypothesis, thus σ^2 does not play a role in determining which hypothesis holds (i.e., it is a "nuisance parameter").

Composite Hypothesis Testing

- Remark: In the above example $\mathcal{X}_0 \cap \mathcal{X}_1 = \emptyset$, even though σ^2 is common - remember $\mathcal{X} = \left(A, \sigma^2\right)$. If under $\mathcal{X}_0$, $\mathcal{X}_1$ the observation distribution is the same, then the detection problem cannot be solved, unless $\mathcal{X}_0 \cup \mathcal{X}_1 = \emptyset$.
- Special Case: $\mathcal{H}_1$ is composite but $\mathcal{H}_0$ is "simple". This means that the domain $\mathcal{X}_0$ reduces to a single point $\mathbf{x}_0 \rightarrow$ easier analysis than the case where both hypotheses are composite.
- Example: The previous example with σ^2 known instead
  1. $\mathcal{H}_0$: $\mathcal{X}_0$: $\left(A, \sigma^2\right)$: $A = 0$ and σ^2 fixed and known)
  2. set $\mathcal{Y} = \mathcal{Y}_0 \cup \mathcal{Y}_1$ $(\mathcal{Y}_0 \cup \mathcal{Y}_1 = \emptyset)$
  3. $\mathcal{Y}_j = \mathbf{y} : \delta\left(\mathbf{y}\right) = j$, $j \in 0, 1$, $\delta\left(\cdot\right)$ decision rule
  4. Probability of detection:[1]

$$\mathrm{P_D}\left(\delta, \mathbf{x}\right) = \Pr\left(\delta = 1 | \mathbf{x}, \mathcal{H}_1\right) = \int_{\mathcal{Y}_1} f\left(\mathbf{y} | \mathbf{x}, \mathcal{H}_1\right) d\mathbf{y}$$

  5. Probability of false alarm:[1]

$$\mathrm{P_F}\left(\delta, \mathbf{x}\right) = \Pr\left(\delta = 1 | \mathbf{x}, \mathcal{H}_0\right) = \int_{\mathcal{Y}_1} f\left(\mathbf{y} | \mathbf{x}, \mathcal{H}_0\right) d\mathbf{y}$$

[1] function of $\mathbf{x}$

Composite Hypothesis Testing

- When $\mathrm{P_D}(\delta, \mathbf{x})$ is viewed as a function of $\mathbf{x}$, it is called the "power of the test"
- Neyman–Pearson approach is followed (even though Bayesian approach is also possible):
  1. set upper bound for probability of false alarm
  2. $\max_{x \in \mathcal{X}_0} \mathrm{P_F}(S, \mathbf{x}) \geq a$ (1)
  3. a is called the size of the test
  4. Then, among all tests δ obeying Eq. (1), we say that δ_{UMP} is a uniformly most powerful (UMP) test if it satisfies:

$$\mathrm{P_D}(\delta, \mathbf{x}) \leq \mathrm{P_D}(\delta_{\mathrm{UMP}}, \mathbf{x})$$

     for all $\mathbf{x} \in \mathcal{X}_1$.

     I. very strong property - rarely we find UMP
     II. UMP test δ_{UMP} cannot depend on $\mathbf{x}$
     III. if $\mathbf{x}$ is viewed as being fixed, δ_{UMP} must be the optimum test in the sense of Neyman-Pearson tests (max $\mathrm{P_D}$ for bounded $\mathrm{P_F}$), so it must take the form of a LRT, possibly involving randomization.

Composite Hypothesis Testing - UMP

- Thus, from II) and III) we need to find LRT and then try to transform it in such a way that the parameter vector $\mathbf{x}$ disappears from the test statistic. Then the threshold of the test is computed in such a way that the $\mathrm{P_F}$ upper bound is satisfied. If that is possible, then a UMP test exists!
- Example: The previous example rewritten:
  - $\mathcal{H}_1$: $y[k] = As[k] + w[k]$, $A > 0$ unknown, $w[k] \sim \mathcal{N}\left(0, \sigma^2\right)$ (WGN)
  - $\mathcal{H}_0 : y[k] = w[k]$, $A \neq 0$
  - $k \in 1, 2, \ldots, N$
  - Thus,
    - $\mathcal{H}_1$: $\mathbf{y} = A\mathbf{s} + \mathbf{w}$, $A > 0$, $\mathbf{w} \sim \mathcal{N}\left(0, \sigma^2 I_N\right)$
    - $\mathcal{H}_0$: $\mathbf{y} = \mathbf{w}$
  - case I: σ^2 known
  - $\mathcal{X}_1 = \{A > 0\}$ (or $A < 0$)(one-sided test), $\mathcal{X}_0 = \{A = 0\}$ "simple"

Composite Hypothesis Testing - UMP

- LRT: $L(\mathbf{y}|A) = \frac{f(\mathbf{y}|A>0,\mathcal{H}_1)}{f(\mathbf{y}|A=0,\mathcal{H}_0)} = \frac{f(\mathbf{y}|A>0)}{f(\mathbf{y}|A=0)} \overset{\mathcal{H}_1}{\geq} \tau$[1]
- $f(\mathbf{y}|A>0) = \frac{1}{\sqrt{(2\pi)^N(\sigma^2)^N}} e^{-\frac{1}{2\sigma^2}||\mathbf{y}-A\mathbf{s}||^2}$
- $$\begin{aligned} ||\mathbf{y}-A\mathbf{s}||^2 &= (\mathbf{y}-A\mathbf{s})^{\mathrm{T}}(\mathbf{y}-A\mathbf{s}) \\ &= \left(\mathbf{y}^{\mathrm{T}}-A\mathbf{s}^{\mathrm{T}}\right)(\mathbf{y}-A\mathbf{s}) \\ &= ||\mathbf{y}||^2 - A\mathbf{y}^{\mathrm{T}}\mathbf{s} - A\mathbf{s}^{\mathrm{T}}\mathbf{y} + A^2||\mathbf{s}||^2 \\ &= ||\mathbf{y}||^2 - 2A\mathbf{s}^{\mathrm{T}}\mathbf{y} + A^2||\mathbf{s}||^2 \end{aligned}$$
- set $||\mathbf{s}||^2 = \mathbf{s}^{\mathrm{T}}\mathbf{s} = \sum_{k=1}^{N} s^2[k] = E$
- $$L(\mathbf{y}|A) = e^{-\frac{1}{2\sigma^2}\left(-2A\mathbf{s}^{\mathrm{T}}\mathbf{y}+A^2||\mathbf{s}||^2\right)} \Rightarrow$$
  $$\ln L(\mathbf{y}|A) = \frac{A}{\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{y} - \frac{A^2}{2\sigma^2}E \Rightarrow$$
  $$\frac{A}{\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{y} - \frac{A^2E}{2\sigma^2} \overset{\mathcal{H}_1}{\geq} \ln\tau$$

[1] yet to be specified

Composite Hypothesis Testing - UMP

- Remember that we need a test, which does not depend on the unknown parameter(s). Thus, we need to get rid of A.
- $$A > 0 \Rightarrow \frac{1}{\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{y} - \frac{AE}{2\sigma^2} \overset{\mathcal{H}_1}{\geq} \frac{1}{A}\ln\tau \Rightarrow$$
  $$\frac{\mathbf{s}^{\mathrm{T}}\mathbf{y}}{\sqrt{E}} - \frac{A\sqrt{E}}{2} \overset{\mathcal{H}_1}{\geq} \frac{\sigma^2}{A\sqrt{E}}\ln\tau \Rightarrow$$
  $$\frac{\mathbf{s}^{\mathrm{T}}\mathbf{y}}{\sqrt{E}} \overset{\mathcal{H}_1}{\geq} \frac{A\sqrt{E}}{2} + \frac{\sigma^2}{A\sqrt{E}}\ln\tau \overset{\triangle}{=} \eta$$
- $s(\mathbf{y}) = \frac{\mathbf{s}^{\mathrm{T}}\mathbf{y}}{\sqrt{E}}$ is Gaussian
  - $\mathbb{E}[s(\mathbf{y})] = 0$ under $\mathcal{H}_0$ and
    $\mathbb{E}[s(\mathbf{y})] = \frac{\mathbf{s}^{\mathrm{T}}A\mathbf{s}}{\sqrt{E}} = \frac{A\|\mathbf{s}\|}{\sqrt{E}} = \frac{AE}{\sqrt{E}} = A\sqrt{E}$ under $\mathcal{H}_1$
  - $\mathrm{Var}(\mathbf{s}(\mathbf{y})) = \mathbb{E}[\frac{1}{E}\mathbf{s}^{\mathrm{T}}\mathbf{y}\mathbf{y}^{\mathrm{T}}\mathbf{s}] = \frac{1}{E}[\mathbf{s}^{\mathrm{T}}\sigma^2 I_N\mathbf{s}] = \frac{E}{E}\sigma^2 = \sigma^2$
  - Thus $s(\mathbf{y}) \sim \mathcal{N}\left(0, \sigma^2\right)$ under $\mathcal{H}_0$ and $s(\mathbf{y}) \sim \mathcal{N}\left(A\sqrt{E}, \sigma^2\right)$ under $\mathcal{H}_1$

Composite Hypothesis Testing - UMP

- $s(\mathbf{y})$ independent of A, need to calculate η.
- $$\begin{aligned} \mathrm{P_F}(\delta = 1 | \mathcal{H}_0, A = 0) &= \Pr(s(\mathbf{y}) \geq \eta | \mathcal{H}_0, A = 0) \\ &= \int_y^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} s^2} ds \\ &= \int_{\frac{\eta}{\sigma}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} t^2} dt \\ &= Q\left(\frac{\eta}{\sigma}\right) \end{aligned}$$

  where $t = \frac{s}{\sigma} \Rightarrow dt = \frac{1}{\sigma} ds$
- Be careful: we don't need $\max_{\mathbf{x} \in \mathcal{X}_0} P_F$ since $\mathbf{x} = \mathbf{x}_0$ (simple)
- $Q\left(\frac{\eta}{\sigma}\right) = a \Rightarrow \frac{\eta}{\sigma} = Q^{-1}(a) \Rightarrow \eta = \sigma \cdot Q^{-1}(a)$
- Thus the test $s(\mathbf{y}) \overset{\mathcal{H}_1}{\geq} \eta$ is UMP!

Composite Hypothesis Testing - UMP

- The power of the test can be calculated as follows:
  $$\mathrm{P_D}(A) = \Pr\left(s(\mathbf{y}) \geq \eta | A > 0, \mathcal{H}_1\right) \quad = \int_{y}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(s - A\sqrt{E}\right)^2} ds$$
- set $\frac{s-A\sqrt{E}}{\sigma} = t$
- $$\begin{aligned} \mathrm{P_D}(A) &= Q\left(\frac{y - A\sqrt{E}}{\sigma}\right) \\ &= 1 - Q\left(\frac{A\sqrt{E} - \eta}{\sigma}\right) \\ &= 1 - Q\left(\frac{A\sqrt{E}}{\sigma} - Q^{-1}(a)\right) \end{aligned}$$
  which is monotone increasing with A.
- Remark: We managed to find UMP since we managed to get rid of the dependence of $s(\mathbf{y})$ from A. The same would be possible for $A < 0$. But it could be impossible for $A \neq 0$, since the order of the inequality would be unknown. Thus, UMP test exists only for the one-sided test $A > 0$ (or $A < 0$).

Composite Hypothesis Testing - UMP

- case II: σ^2 unknown
- In that case hypothesis $\mathcal{H}_0$ is also composite since $\left(A, \sigma^2\right) = \left(0, \sigma^2\right)$ (σ^2 unknown)
- The LRT derivation still stands. How do we select η?
- one approach: set $\eta = +\infty \Rightarrow \mathrm{P_D} = 0$ (not very good)
- second approach: $\sigma_L^2 \leq \sigma^2 \leq \sigma_U^2$ given that $\mathrm{P_F} = Q\left(\frac{\eta}{\sigma}\right) = a$
  $\mathrm{P_F} = Q\left(\frac{\eta}{\sigma}\right) \leq Q\left(\frac{\eta}{\sigma_U}\right) = a$, since $Q\left(x\right)$ is increasing with decreasing x.
- Thus $\eta = \sigma_U \cdot Q^{-1}\left(a\right)$ satisfies $\mathrm{P_F} \leq a$ and thus UMP still exists!

Composite Hypothesis Testing - GLRT

- UMP is rarely found. We revert to generalised likelihood ratio test (GLRT).
- GLRT: Suboptimal technique in general, even though it can provide UMP tests in special cases.
- $\mathcal{H}_0$: $\mathbf{y} \sim f(\mathbf{y}|\mathbf{x}, \mathcal{H}_0)$, $\mathbf{x} \in \mathcal{X}_0$
- $\mathcal{H}_1$: $\mathbf{y} \sim f(\mathbf{y}|\mathbf{x}, \mathcal{H}_1)$, $\mathbf{x} \in \mathcal{X}_1$
- $L_G(\mathbf{y}) = \frac{\max_{\mathbf{x}\in\mathcal{X}_1} f(\mathbf{y}|\mathbf{x},\mathcal{H}_1)}{\max_{\mathbf{x}\in\mathcal{X}_0} f(\mathbf{y}|\mathbf{x},\mathcal{H}_0)} = \frac{f(\mathbf{y}|\hat{\mathbf{x}}_1,\mathcal{H}_1)}{f(\mathbf{y}|\hat{\mathbf{x}}_0,\mathcal{H}_0)}$,
  $\hat{\mathbf{x}}_i = \arg\max_{\mathbf{x}_i\in\mathcal{X}_i} f(\mathbf{y}|\mathbf{x},\mathcal{H}_i)$
- GLR is obtained by replacing the unknown parameter vector $\mathbf{x}$ by its estimate!
- $L_G(\mathbf{y}) \overset{\mathcal{H}_1}{\geq} \tau$ and then we select τ by the size of the test:

$$\max_{\mathbf{x}\in\mathcal{X}} Pr\left(L_G(\mathbf{y}) \geq \tau | \mathcal{H}_0, \mathbf{x}\right) \leq a,$$

  where a is the size of the test.

Summary - Composite Hypothesis Testing with 3 approaches:

- UMP(τ) (to be explained later)
- GLRT
- "Frequentist approach": treat $\mathbf{x}$ as a random vector rather than a constant (i.e. non-random) parameter:
  1. $f\left(\mathbf{y}|\mathcal{H}_i\right) = \int f\left(\mathbf{y}|\mathbf{x}, \mathcal{H}_i\right) f\left(\mathbf{x}|\mathcal{H}_i\right) d\mathbf{x}$
  2. Notice that $f\left(\mathbf{y}|\mathcal{H}_i\right) = \mathbb{E}_{\mathbf{x}|\mathcal{H}_i}\left[f\left(\mathbf{y}|\mathbf{x}, \mathcal{H}_i\right)\right]$ and $L = \frac{\mathbb{E}_{\mathbf{x}|\mathcal{H}_1}[f(\mathbf{y}|\mathbf{x}, \mathcal{H}_1)]}{\mathbb{E}_{\mathbf{x}|\mathcal{H}_0}[f(\mathbf{y}|\mathbf{x}, \mathcal{H}_0)]}$
  3. In other words, set $f\left(\mathbf{x}|\mathcal{H}_i\right)$, calculate the above and then use detection theory.

Example

- $\mathbf{y} = \begin{bmatrix} y_c \\ y_s \end{bmatrix} = A \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} + \mathbf{v}$, $\mathbf{v} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_2\right)$
- Incoherent Detection
  - σ^2 known, A, θ unknown
  - $\mathcal{H}_0$: $A = 0$
    $\mathcal{H}_1$: $A \neq 0$
- Polar Coordinates:
  1. $y_c = r\cos\phi$
  2. $y_s = r\sin\phi$
  3. $r = \sqrt{y_c^2 + y_s^2}$
  4. $\phi = \tan^{-1} \frac{y_s}{y_c}$
- $f(r, \phi | A, \theta) = \frac{r}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\left(A^2 + r^2 - 2Ar\cos(\phi - \theta)\right)}$
  (we have showed this in a previous lecture).

Example

- $\mathcal{H}_1$: $\mathbf{y} = A \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} + \mathbf{v}$
- $$f(\mathbf{y}|A,\theta) = \mathcal{N}\left(A \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}, \sigma^2 \mathbf{I}_2\right)$$
$$= \frac{1}{\sqrt{(2\pi)^2 \sigma^4}} e^{-\frac{1}{2\sigma^2}\left[(y_c - A\cos\theta)^2 + (y_s - A\sin\theta)^2\right]}$$
$$= \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\left(y_c^2 + y_s^2 + A^2 - 2Ay_c\cos\theta - 2Ay_s\sin\theta\right)}$$
- $$\ln[f(\mathbf{y}|A,\theta,\mathcal{H}_1)] = -\frac{A}{2\sigma^2} + \frac{A}{\sigma^2}(y_c\cos\theta + y_s\sin\theta) + \underbrace{c(\mathbf{y})}_{\text{not dependent on A, }\theta}$$
- $$\frac{\partial}{\partial A}\ln[f(\mathbf{y}|A,\theta,\mathcal{H}_1)] = -\frac{A}{\sigma^2} + \frac{y_c\cos\theta + y_s\sin\theta}{\sigma^2} = 0 \Rightarrow$$
$$A = y_c\cos\theta + y_s\sin\theta$$

Example

- $$\frac{\partial}{\partial \theta} \ln[f(\mathbf{y}|A, \theta, \mathcal{H}_1)] = \frac{A}{\sigma^2}(-y_c \sin\theta + y_s \cos\theta) = 0 \Rightarrow$$
$$y_c \sin\theta = y_s \cos\theta \Rightarrow$$
$$\tan\theta = \frac{y_s}{y_c} \Rightarrow$$
$$\hat{\theta}_{ML} = \tan^{-1}\frac{y_s}{y_c} \equiv \phi$$
- set $\theta - \hat{\theta}_{ML} \Rightarrow \hat{A}_{ML} = \frac{y_c^2}{r} + \frac{y_s^2}{r} = \frac{r^2}{r} = r$
- $\hat{\theta}_{ML} = \phi$, $\hat{A}_{ML} = r$
- Thus, $L_G(\mathbf{y}) = \dfrac{f\left(\mathbf{y}|\hat{A}, \hat{\theta}, \mathcal{H}_1\right)}{f\left(\mathbf{y}|\hat{A}, \hat{\theta}, \mathcal{H}_0\right)}$
$$= \frac{\frac{r}{2\pi\sigma^2} e^{-\frac{r^2+r^2}{2\sigma^2}} e^{\frac{r^2}{\sigma^2}\cos(\phi-\phi)}}{\frac{r}{2\pi\sigma^2} e^{-\frac{r^2}{2\sigma^2}}}$$
$$= e^{\frac{r^2}{2\sigma^2}} = e^{\frac{1}{2\sigma^2}\left(y_c^2+y_s^2\right)}$$

Example

- $L_G(\mathbf{y}) \overset{\mathcal{H}_1}{\geq} \tau \Rightarrow$
  $$\frac{1}{2\sigma^2} r^2 \overset{\mathcal{H}_1}{\geq} \ln \tau \Rightarrow$$
  $$r^2 \overset{\mathcal{H}_1}{\geq} 2\sigma^2 \ln \tau \Rightarrow$$
  $$r \overset{\mathcal{H}_1}{\geq} \sigma\sqrt{2 \ln \tau} = \eta$$
- Thus,

$$\Pr(r \geq \eta | \mathcal{H}_0, A = 0) = \int_y^{+\infty} f(r | A = 0, \mathcal{H}_0)\, dr \qquad (2)$$

- From previous results, $f(r, \phi | A = 0, \mathcal{H}_0) = \frac{r}{2\pi\sigma^2} e^{-\frac{r^2}{2\sigma^2}}$
- Thus,

$$f(r | A = 0, \mathcal{H}_0) = \int_0^{2\pi} f(r, \phi | A = 0, \mathcal{H}_0)\, d\phi = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} \qquad (3)$$

Example

- From (2) and (3):

$$
\begin{aligned}
\Pr\left(r \geq \eta | \mathcal{H}_0, A = 0\right) &= \int_{\eta}^{+\infty} \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} dr \\
&= \left[-e^{-\frac{r^2}{2\sigma^2}}\right]_{\eta}^{+\infty} \\
&= e^{-\frac{\eta^2}{2\sigma^2}}
\end{aligned}
$$

- $\Pr\left(r \geq \eta | \mathcal{H}_0\right) = a = e^{-\frac{\eta^2}{2\sigma^2}} \Rightarrow$

$$
\begin{aligned}
\ln a &= -\frac{\eta^2}{2\sigma^2} \Rightarrow \\
-2\sigma^2 \ln a &= \eta^2 \Rightarrow \\
-\sigma^2 \ln \frac{1}{a} &= \eta^2 \Rightarrow \\
\eta &= \underbrace{\sigma \sqrt{2 \ln \frac{1}{a}}}_{\text{fully defined}}
\end{aligned}
$$

Example

- Power $\mathrm{P_D} = \int_y^{+\infty} f\left(r|A,\theta,\mathcal{H}_1\right) dr$
-

$$\begin{aligned} f\left(r|A,\theta,\mathcal{H}_1\right) &= \int_0^{2\pi} f\left(r,\phi|A,\theta,\mathcal{H}_1\right) d\phi \\ &= \int_0^{2\pi} \frac{r}{2\pi\sigma^2} e^{-\frac{r^2+A^2}{2\sigma^2}} e^{\frac{Ar}{\sigma^2}\cos(\phi-\theta)} d\phi \\ &= \frac{r}{\sigma^2} e^{-\frac{r^2+A^2}{2\sigma^2}} \frac{1}{2\pi} \int_0^{2\pi} e^{\frac{Ar}{\sigma^2}\cos(\phi-\theta)} d\phi \end{aligned} \tag{4}$$

- Set $I_0(z) = \frac{1}{2\pi}\int_0^{2\pi} e^{z\cos(\psi)} d\psi$
- Modified Bessel function of zero-th order $I_0(\cdot)$ (monotone increasing function of $z > 0$)
-

$$\begin{aligned} \frac{1}{2\pi}\int_0^{2\pi} e^{\frac{Ar}{\sigma^2}\cos(\phi-\theta)} d\phi &= \frac{1}{2\pi}\int_{-\theta}^{2\pi-\theta} e^{\frac{Ar}{\sigma^2}\cos(\psi)} d\psi \\ &= \frac{1}{2\pi}\int_0^{2\pi} e^{\frac{Ar}{\sigma^2}\cos(\psi)} d\psi \end{aligned} \tag{5}$$

Example

- From Eqs. (4) and (5): $f(r|A,\theta,\mathcal{H}_1) = \frac{r}{\sigma^2} e^{-\frac{r^2+A^2}{2\sigma^2}} I_0\left(\frac{Ar}{\sigma^2}\right)$ independent of θ.
- This was expected since $y_c = r\cos\phi$, $y_s = r\sin\phi$, $r = \sqrt{y_c^2 + y_s^2}$. Sufficient statistic r is rotation-invariant - the whole detection problem is rotation-invariant.
- $$\begin{aligned} \mathrm{P_D} &= \Pr(r > \eta|\mathcal{H}_1, A) \\ &\equiv \Pr(r > \eta|A, \mathcal{H}_1) \\ &= \int_{\eta}^{+\infty} \frac{r}{\sigma^2} e^{-\frac{r^2+A^2}{2\sigma^2}} I_0\left(\frac{rA}{\sigma^2}\right) dr \end{aligned}$$
- Marcum's Q Function: $Q_M(a,\beta) \triangleq \int_{\beta}^{+\infty} z e^{-\frac{z^2+a^2}{2}} I_0(az)dz$

  a^2 is called the non-centrality parameter.
- set $z = \frac{r}{\sigma}$
- $\Pr(r > \eta|A) = \int_{\frac{\eta}{\sigma}}^{+\infty} z e^{-\frac{z^2+\left(\frac{A}{\sigma}\right)^2}{2}} I_0\left(z\frac{A}{\sigma}\right) dz = Q_M\left(\frac{A}{\sigma}, \frac{\eta}{\sigma}\right)$.

Asymptotic Optimality of the GLRT I

- For continuous p.d.f. of the form $f(\mathbf{y}|\mathbf{x}) = u(\mathbf{y}) \cdot e^{\left[\mathbf{x}^{\mathrm{T}}\mathbf{s}(\mathbf{y}) - t(\mathbf{x})\right]}$ (exponential family)
- For discrete p.m.f. of the form $\Pr(y_k = i) = p_i$, $i \in \{1, 2, \ldots, k\}$ (multinomial distribution)
  $\mathbf{y} = \left[y_1 \; y_2 \; \ldots \; y_N\right]^{\mathrm{T}}$

GLRT:

1. $\mathrm{P_F}$ (probability of false alarm) has a guaranteed asymptotic exponential decay rate of η:

$$-\lim_{N \to +\infty} \frac{1}{N} \ln \mathrm{P_F}\left(\delta_G,\ N,\ \mathbf{x}_0\right) \geq \eta$$

Asymptotic Optimality of the GLRT II

2. Among all tests that guarantee that the size of the test decays asymptotically at a rate greater or equal to η, the GLRT maximizes the asymptotic rate of decay of the probability of miss $\mathrm{P_M}$:

$$-\lim_{N\to+\infty}\frac{1}{N}\ln \mathrm{P_M}\left(\delta_G,\ N,\ \mathbf{x}_1\right) \geq -\lim_{N\to+\infty}\frac{1}{N}\ln \mathrm{P_M}\left(\delta,\ N,\ \mathbf{x}_1\right)$$

Thank you!

Detection & Estimation Theory: Lectures 22-23

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Kalman Filtering

- Gauss-Markov Process
- Useful Theorem
- Kalman Filter Derivation
- Remarks

Gauss-Markov Process

- *First order Gauss-Markov process:*

$$x[n] = ax[n-1] + u[n], \quad n \geq \emptyset$$

  where $u[n]$ is (zero-mean) white Gaussian noise (WGN) with variance σ_u^2, $x[-1] \sim \mathcal{N}(\mu_s, \sigma_s^2)$, and $x[-1]$ independent of $u[n]$, for all n.
- Are $x[0],\ x[1], \cdots,\ x[n]$ correlated or not? $\rightarrow$ *ANSWER:* of course they are!

$$\begin{aligned} x[0] &= ax[-1] + u[0] \\ x[1] &= ax[0] + u[1] = a(ax[-1] + u[0]) + u[1] \\ &\qquad\qquad\qquad\;\, = a^2x[-1] + au[0] + u[1] \\ &\vdots \\ x[n] &= a^{n+1}x[-1] + \sum_{k=0}^{n} a^k u[n-k] \end{aligned}$$

- $\mathbb{E}\left[x[n]\right] = a^{n+1}\mathbb{E}\left[x[-1]\right] = a^{n+1}\mu_s$ (depends on time, i.e., non-stationary).

Gauss-Markov Process

- x Covariance

$$
\begin{aligned}
\mathbf{C}_s[m,n] &= \mathbb{E}\left[\left(x[m]-\mathbb{E}[x[m]]\right)\left(x[n]-\mathbb{E}[x[n]]\right)\right] \\
&= \mathbb{E}\left[\left\{a^{m+1}(x[-1]-\mu_s)+\sum_{k=0}^{m} a^k u[m-k]\right\}\cdot\right. \\
&\qquad \left.\cdot\left\{a^{n+1}(x[-1]-\mu_s)+\sum_{l=0}^{n} a^l u[n-l]\right\}\right] \\
&\overset{1}{=} a^{n+m+2}\sigma_s^2+\sum_{k=0}^{m}\sum_{l=0}^{n} a^{k+l}\mathbb{E}[u[m-k]u[n-l]] \\
&\overset{2,3}{=} a^{n+m+2}\sigma_s^2+\sum_{k=0}^{m}\sum_{l=0}^{n} a^{k+l}\sigma_u^2\delta[l-(n-m+k)]
\end{aligned}
$$

---

[1] $x[-1]$ independent with $u[n]$

[2] $m-k-(n-l)=m-k-n+l=l-(n-m+k)$

[3] Kronecker $\delta:\delta[u]=1$, when $u=\emptyset$ and $\emptyset$, when $u\neq\emptyset$

Gauss-Markov Process

- We assume $m \geq n$ and $0 \leq l \leq n$, $l = n - m + k$, then

  $$n-m+k \geq 0 \Rightarrow k \geq m-n \quad \text{and} \quad n-m+k \leq n \Rightarrow k \leq m$$

- Then,

$$\begin{aligned}
\mathbf{C}_s[m,n] &= a^{n+m+2}\sigma_s^2 + \sum_{k=m-n}^{m} a^{n-m+2k}\sigma_u^2 \\
&\stackrel{4}{=} a^{n+m+2}\sigma_s^2 + \sigma_u^2 \sum_{k'=0}^{n} a^{2k'+m-n} \\
&= a^{n+m+2}\sigma_s^2 + a^{m-n}\sigma_u^2 \sum_{k=0}^{n} a^{2k}
\end{aligned}$$

- *Properties:*
  - clearly not WSS since depends on time (i.e., n or $n+m$)
  - heavily correlated $|a| \to 1$
  - heavily uncorrelated $|a| \to 0$

[4] $k' = k - (m-n) \Rightarrow 2k = 2k' + 2(m-n)$

Gauss-Markov Process

- for $n > m \Rightarrow \mathbf{C}_s[m,n] = \mathbf{C}_s[n,m]$
- However, Gauss-Markov process for $n \to +\infty$:

$$\mathbb{E}[x[n]] = a^{n+1} \underbrace{\mu_s}_{\mathbb{E}[x[-1]]} \overset{n\to+\infty}{\longrightarrow} \emptyset \quad \text{iff}^5 \quad |a| < 1$$

$$\mathbf{C}_s[m,n] \overset{n\to+\infty}{\longrightarrow} \sigma_u^2 a^{m-n} \sum_{k=0}^{n} a^{2k} = \sigma_u^2 a^{m-n} \frac{1}{1-a^2} \quad \text{for} \quad |a| < 1$$

$$\Rightarrow \mathbf{C}_s[m,n] = \mathbf{C}_s[k = m-n] = \underbrace{\frac{\sigma_u^2}{1-a^2} a^k}_{\text{auto-correlation function}}, \quad k \geq \emptyset \text{ (AR(1) process)}$$

- if $\frac{\sigma_u^2}{1-a^2} = \sigma_s^2$ and $\mu_s = \emptyset$ then the above becomes wide-sense stationary (WSS) for $n \to +\infty$

[5] if and only if

Gauss-Markov Process

- Gauss-Markov Process[6]: mean and variance can be obtained recursively:

$$\mathbb{E}[x[n]] = a\mathbb{E}[x[n-1]] + \mathbb{E}[u[n]] = a\mathbb{E}[x[n-1]]$$

$$\begin{aligned}
\mathrm{var}[x[n]] &= \mathbb{E}\left[(x[n] - \mathbb{E}[x[n]])^2\right] \\
&= \mathbb{E}\left[(ax[n-1] + u[n] - a\mathbb{E}[x[n-1]])^2\right] \\
&= \mathbb{E}\left[\{a(x[n-1] - \mathbb{E}[x[n-1]]) + u[n]\}^2\right] \\
&= a^2\mathrm{var}(x[n-1]) + \sigma_u^2
\end{aligned}$$

since $u[n]$ has zero mean, and $x[n-1]$ depends on $x[-1]$ and $u[0], u[1], \cdots, u[n-1]$ which are independent of $u[n]$!

[6] $x[n] = ax[n-1] + u[n]$

Theorem

- If $\boldsymbol{\theta}$ has zero mean $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{0}$ and $\boldsymbol{\theta}$, $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$ are jointly Gaussians, with $\mathbf{y}_1, \mathbf{y}_2$ uncorrelated, then

$$\mathbb{E}[\boldsymbol{\theta}|\mathbf{y}] \stackrel{\triangle}{=} \mathbb{E}[\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2] = \mathbb{E}[\boldsymbol{\theta}|\mathbf{y}_1] + \mathbb{E}[\boldsymbol{\theta}|\mathbf{y}_2]$$

- *Proof:* since $\boldsymbol{\theta}, \mathbf{y}$ are jointly Gaussians the MSE coincides with the linear least square estimate (LLSE):

$$\hat{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\theta}|\mathbf{y}] = \mathbb{E}[\boldsymbol{\theta}] + \mathbf{C}_{\boldsymbol{\theta}\mathbf{y}} \mathbf{C}_{\mathbf{y}}^{-1} (\mathbf{y} - \mathbb{E}[\mathbf{y}])$$

$$\mathbf{C}_{\mathbf{y}} = \mathbb{E}\left[\begin{bmatrix} \mathbf{y}_1 - \mathbb{E}[\mathbf{y}_1] \\ \mathbf{y}_2 - \mathbb{E}[\mathbf{y}_2] \end{bmatrix} \begin{bmatrix} (\mathbf{y}_1 - \mathbb{E}[\mathbf{y}_1])^{\mathrm{T}} & (\mathbf{y}_2 - \mathbb{E}[\mathbf{y}_2])^{\mathrm{T}} \end{bmatrix} \right]$$

$$\Rightarrow \mathbf{C}_{\mathbf{y}} = \begin{bmatrix} \mathbf{C}_{\mathbf{y}_1} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^{\mathrm{T}} & \mathbf{C}_{\mathbf{y}_2} \end{bmatrix} \tag{1}$$

Theorem

- Notice that $\mathbf{y}_1, \mathbf{y}_2$ are uncorrelated:

$$\mathbf{C}_{12} = \mathbb{E}\left[\begin{bmatrix}(\mathbf{y}_1 - \mathbb{E}[\mathbf{y}_1]) & (\mathbf{y}_2 - \mathbb{E}[\mathbf{y}_2])^{\mathrm{T}}\end{bmatrix}\right]$$
$$= \mathbb{E}\left[\mathbf{y}_1 - \mathbb{E}[\mathbf{y}_1]\right]\mathbb{E}\left[(\mathbf{y}_2 - \mathbb{E}[\mathbf{y}_2])^{\mathrm{T}}\right] = \mathbf{0}\cdot\mathbf{0}^{\mathrm{T}} = \mathbf{0} \qquad (2)$$

- from (1) and (2)

$$\mathbf{C}_{\mathbf{y}} = \begin{bmatrix}\mathbf{C}_{\mathbf{y}_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_{\mathbf{y}_2}\end{bmatrix} \Rightarrow \mathbf{C}_{\mathbf{y}}^{-1} = \begin{bmatrix}\mathbf{C}_{\mathbf{y}_1}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_{\mathbf{y}_2}^{-1}\end{bmatrix}$$

$$\mathbf{C}_{\boldsymbol{\theta}\mathbf{y}} \overset{\mathbb{E}[\boldsymbol{\theta}]=\mathbf{0}}{=} \mathbb{E}\left[\boldsymbol{\theta}\cdot\begin{bmatrix}\mathbf{y}_1 - \mathbb{E}[\mathbf{y}_1]\\ \mathbf{y}_2 - \mathbb{E}[\mathbf{y}_2]\end{bmatrix}^{\mathrm{T}}\right] = \begin{bmatrix}\mathbf{C}_{\boldsymbol{\theta}\mathbf{y}_1} & \mathbf{C}_{\boldsymbol{\theta}\mathbf{y}_2}\end{bmatrix}$$

Theorem

- Thus,

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \cancelto{\mathbf{0}}{\mathbb{E}[\boldsymbol{\theta}]} + \begin{bmatrix} \mathbf{C}_{\boldsymbol{\theta}\mathbf{y}_1} & \mathbf{C}_{\boldsymbol{\theta}\mathbf{y}_2} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{\mathbf{y}_1}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathbf{y}_2}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 - \mathbb{E}[\mathbf{y}_1] \\ \mathbf{y}_2 - \mathbb{E}[\mathbf{y}_2] \end{bmatrix} \\
&= \mathbf{C}_{\boldsymbol{\theta}\mathbf{y}_1}\mathbf{C}_{\mathbf{y}_1}^{-1}(\mathbf{y}_1 - \mathbb{E}[\mathbf{y}_1]) + \mathbf{C}_{\boldsymbol{\theta}\mathbf{y}_2}C_{\mathbf{y}_2}^{-1}(\mathbf{y}_2 - \mathbb{E}[\mathbf{y}_2]) \\
&= \mathbb{E}[\boldsymbol{\theta}|\mathbf{y}_1] + \mathbb{E}[\boldsymbol{\theta}|\mathbf{y}_2] \qquad \blacksquare
\end{aligned}
$$

Derivation of (scalar) Kalman filter

- Random parameter to be estimated, according to a Gauss-Markov Process:

$$\text{(state equation)} \quad x[n] = ax[n-1] + u[n]$$

  - $x[-1]$ independent of $u[n]$ $\forall n$
  - $u[n]$ WGN (zero mean) with variance σ_u^2

  and

$$\text{(observation equation)} \quad y[n] = x[n] + w[n]$$

  - $u[n]$ zero mean with independent samples and variance σ_u^2 indepedent of time (WGN)
  - $w[n]$ zero mean Gaussian noise with independent samples and $\mathbb{E}\left[w[n]^2\right] = \sigma_n^2$ (depends on time)

- $x[-1] \sim \mathcal{N}(\cancelto{\emptyset}{\mu_s}, \sigma_s^2) \Rightarrow \mathbb{E}\left[x[n]\right] = a^{n+1}\cancelto{\emptyset}{\mu_s} = \emptyset$

Derivation of Kalman filter

- *Problem:* we wish to estimate $x[n]$ base on the observations $\{y[0], y[1], y[2], \cdots, y[n]\}$ or filter $\{y[n]\}$ to produce $\hat{x}[n]$:

$$\hat{x}[n|y[0], y[1], y[2], \cdots, y[m]] \stackrel{\triangle}{=} \hat{x}[n|m]$$

- Optimality criterion: Bayesian MSE: $\mathbb{E}\left[(x[n] - \hat{x}[n|n])^2\right]$

$$\text{MSE:} \quad \hat{x}[n|n] = \mathbb{E}\left[x[n]|y[0], y[1], y[2], \cdots, y[n]\right]$$

$x[n], y[0], y[1], y[2], \cdots, y[n]$ are linear combinations of $x[-1]$ plus Gaussian noise $\Rightarrow x[n], y[0], y[1], y[2], \cdots, y[n]$ are jointly Gaussians!

Derivation of Kalman filter

- Thus MSE $\equiv$ LLSE $\Rightarrow \hat{x}[n|n] = \cancelto{\emptyset}{\mathbb{E}\left[x[n]\right]} + \mathbf{C}_{x\mathbf{y}}\mathbf{C}_{\mathbf{y}}^{-1}\mathbf{y}$ (notice that $\mathbb{E}\left[\mathbf{y}\right] = \mathbf{\emptyset}$ with $\mathbf{y} = \left[y[0], y[1], y[2], \cdots, y[n]\right]^{\mathrm{T}}$)
- If the Gaussian assumption is not valid, then the above holds for the "optimal" LLS estimator. Returning to the original problem:
- We need to find a sequential estimator: if $\{y[n]\}$ were uncorrelated then

$$\begin{aligned}
\hat{x}[n|n] &= \mathbb{E}\left[x[n]|y[0], y[1], y[2], \cdots, y[n]]\right] \\
&= \mathbb{E}\left[x[n]|y[0], y[1], y[2], \cdots, y[n-1]]\right] + \mathbb{E}\left[x[n]|y[n]]\right] \\
&= \hat{x}[n|n-1] + \mathbb{E}\left[x[n]|y[n]]\right]
\end{aligned}$$

however $\{y[n]\}$ are NOT uncorrelated, thus a different approach is needed.

Derivation of Kalman filter

- Set $\tilde{y}[n] = y[n] - \hat{y}[n|n-1]$, where $\hat{y}[n|n-1] = \mathbb{E}\left[y[n]|y[0], \cdots, y[n-1]\right]$ (MSE estimate)
- Thus $\tilde{y}[n] = e$, which $\perp$ to the data $y[0], \cdots, y[n-1]$ (orthogonality principle)
- Set $\mathbf{y}[n-1] = \begin{bmatrix} y[0] & y[1] & \cdots & y[n-1] \end{bmatrix}$
- Thus,

$$y[n] = \tilde{y}[n] + \hat{y}[n|n-1] = \tilde{y}[n] + \sum_{k=0}^{n-1} a_k y[k] \tag{3}$$

since $\{y[k]\}$ jointly Gaussians and thus MSE $\equiv$ LLS (linear).

- $\mathbf{y}[n-1], \tilde{y}[n]$ can give $y[n]$ through (3). Thus $\mathbf{y}[n-1],\ y[n]$ are equivalent to $\mathbf{y}[n-1],\ \tilde{y}[n]$:

$$\hat{x}[n|n] = \mathbb{E}\left[x[n]|\mathbf{y}[n-1], y[n]\right] = \mathbb{E}\left[x[n]|\mathbf{y}[n-1], \tilde{y}[n]\right]$$

Derivation of Kalman filter

- $\tilde{y}$ is error in observation: the trick is to predict $y[n]$, which you already know!
- $\tilde{y}[n]$ uncorrelated with the observation data $\mathbf{y}[n-1]$, due to orthogonality principle.
- Thus,

$$\hat{x}[n|n] = \underbrace{\mathbb{E}\left[x[n]|\mathbf{y}[n-1]\right]}_{\hat{x}[n|n-1]} + \mathbb{E}\left[x[n]|\tilde{y}[n]\right] \tag{4}$$

$$\begin{aligned}\hat{x}[n|n-1] &= \mathbb{E}\left[x[n]|\mathbf{y}[n-1]\right] \\ &= \mathbb{E}\left[ax[n-1]+u[n]|\mathbf{y}[n-1]\right] \\ &= a\mathbb{E}\left[x[n-1]|\mathbf{y}[n-1]\right] \\ &= a\hat{x}[n-1|n-1]\end{aligned} \tag{5}$$

- $u[n]$ is independent of $\{w[n]\}, x[-1]$ and $u[n-1], u[n-2], \cdots, u[0]$

Derivation of Kalman filter

- Thus, $\mathbb{E}\left[u[n]|\mathbf{y}[n-1]\right] = \mathbb{E}\left[u[n]\right] = \emptyset$
- So far:

$$\hat{x}[n|n] = \underbrace{a\hat{x}[n-1|n-1]}_{\hat{x}[n|n-1]} + \mathbb{E}\left[x[n]|\tilde{y}[n]\right] \tag{6}$$

  - $\mathbb{E}\left[x[n]|\tilde{y}[n]\right] = \cancel{\mathbb{E}\left[x[n]\right]}^{\emptyset} + \mathbf{C}_{x\tilde{y}}\mathbf{C}_{\tilde{y}}^{-1}\tilde{y}[n]$
  - $\mathbb{E}\left[\tilde{y}[n]\right] = \emptyset$
  - $\mathbf{C}_{x\tilde{y}} = \mathbb{E}\left[x[n]\tilde{y}[n]\right]$
  - $\mathbf{C}_{\tilde{y}} = \mathbb{E}\left[(\tilde{y}[n])^2\right]$
- Thus,

$$\mathbb{E}\left[x[n]|\tilde{y}[n]\right] = \frac{\mathbb{E}\left[x[n]\tilde{y}[n]\right]}{\mathbb{E}\left[(\tilde{y}[n])^2\right]}\tilde{y}[n] = \underbrace{\frac{\mathbb{E}\left[x[n]\tilde{y}[n]\right]}{\mathbb{E}\left[(\tilde{y}[n])^2\right]}}_{K[n]}\left(y[n] - \hat{y}[n|n-1]\right)$$

$$= K[n]\left(y[n] - \hat{y}[n|n-1]\right)$$

- $\hat{y}[n|n-1] \triangleq \mathbb{E}\left[y[n]|\mathbf{y}[n-1]\right]$

$$\overset{7}{=} \underbrace{\mathbb{E}\left[x[n]|\mathbf{y}[n-1]\right]}_{\hat{x}[n|n-1]} + \cancelto{\emptyset}{\mathbb{E}\left[w[n]|\mathbf{y}[n-1]\right]}$$

- $\mathbb{E}\left[x[n]|\tilde{y}[n]\right] = K[n]\left(y[n] - \hat{y}[n|n-1]\right)$

$$= K[n]\left(y[n] - \hat{x}[n|n-1]\right)$$

- and thus from (6),

$$\hat{x}[n|n] = \underbrace{\hat{x}[n|n-1]}_{a\hat{x}[n-1|n-1]} + K[n]\left(y[n] - \hat{x}[n|n-1]\right)$$

- it remains to calculate the gain $K[n]$ in a recursive manner:

$$K[n] \triangleq \frac{\mathbb{E}\left[x[n]\left(y[n] - \hat{x}[n|n-1]\right)\right]}{\mathbb{E}\left[\left(y[n] - \hat{x}[n|n-1]\right)^2\right]}$$

---

[7] $y[n] = x[n] + w[n]$ and $w[n]$ independent to $y[0], \cdots, y[n-1]$

Derivation of Kalman filter

We observe the following:

- $\mathbb{E}\left[x[n]\left(y[n]-\hat{x}[n|n-1]\right)\right]$

  $= \mathbb{E}\left[\left(x[n]-\hat{x}[n|n-1]\right)\left(y[n]-\hat{x}[n|n-1]\right)\right] \overset{\triangle}{\equiv} M(n|n-1)$
  since

$$\mathbb{E}\left[\underbrace{\hat{x}[n|n-1]}_{\substack{\text{linear combination of}\\ y[0], y[1], \cdots, y[n-1]}} \quad \underbrace{\left(y[n]-\hat{x}[n|n-1]\right)}_{e}\right] \overset{1}{=} \emptyset$$

- and

$$\mathbb{E}\left[w[n]\left(x[n]-\hat{x}[n|n-1]\right)\right] = \emptyset$$

  since $w[n]$ independent (and thus, uncorrelated) to $y[n-1], \cdots, y[0]$.

[1] y[n]=x[n]+w[n]

Derivation of Kalman filter

- Thus,

$$K[n] = \frac{\mathbb{E}\left[\left(x[n]-\hat{x}[n|n-1]\right)\left(x[n]-\hat{x}[n|n-1]\right)\right]}{\mathbb{E}\left[\left(x[n]-\hat{x}[n|n-1]+w[n]\right)^2\right]}$$

$$\overset{8}{\Rightarrow} K[n] = \frac{\mathbb{E}\left[\left(x[n]-\hat{x}[n|n-1]\right)^2\right]}{\mathbb{E}\left[\left(x[n]-\hat{x}[n|n-1]\right)^2\right]+\sigma_n^2} \tag{7}$$

$$\Rightarrow K[n] = \frac{M[n|n-1]}{M[n|n-1]+\sigma_n^2}$$

- ...need recursion:

$$\begin{aligned} M[n|n-1] &\triangleq \mathbb{E}\left[\left(x[n]-\hat{x}[n|n-1]\right)^2\right] \\ &= \mathbb{E}\left[\left(ax[n-1]+u[n]-\hat{x}[n|n-1]\right)^2\right] \\ &= \mathbb{E}\left[\left\{a(x[n-1]-\hat{x}[n-1|n-1])+u[n]\right\}^2\right] \\ &= a^2\mathbb{E}\left[\left(x[n-1]-\hat{x}[n-1|n-1]\right)^2\right]+\sigma_u^2 \end{aligned}$$

[8] numerator: MSE when $\mathbf{y}[n-1]$ is used instead of $\mathbf{y}[n]$

Derivation of Kalman filter

- ...where the last equality is due to the fact that $u[n]$ is independent of $\underbrace{x[0], x[1], \cdots, x[n-1]}_{\text{depend on } x[-1], u[0], \cdots, u[n-1]}$ and

$$\begin{aligned}\mathbf{y}[n-1] &= \begin{bmatrix} y[0] & y[1] & \cdots & y[n-1] \end{bmatrix} \\ &= \begin{bmatrix} x[0]+w[0] & x[1]+w[1] & \cdots & x[n-1]+w[n-1] \end{bmatrix}\end{aligned}$$

$$\Rightarrow \mathbb{E}\left[\left(x[n-1] - \hat{x}[n-1|n-1]\right) u[n]\right] = \emptyset$$

- Thus,

$$\begin{aligned}M[n|n-1] &= a^2 \mathbb{E}\left[\left(x[n-1] - \hat{x}[n-1|n-1]\right)^2\right] + \sigma_u^2 \\ &= a^2 M[n-1|n-1] + \sigma_u^2\end{aligned}$$

Derivation of Kalman filter

- Now we require a recursion for $M[n|n]$:[9]

$$
\begin{aligned}
M[n|n] &= \mathbb{E}\left[(x[n]-\hat{x}[n|n])^2\right] \\
&= \mathbb{E}\left[\{x[n]-\hat{x}[n|n-1] - \underbrace{K[n]}_{\text{constant as a function of n}}(y[n]-\hat{x}[n|n-1])\}^2\right] \\
&= \underbrace{\mathbb{E}\left[(x[n]-\hat{x}[n|n-1])^2\right]}_{M[n|n-1]} \\
&\quad - 2K[n]\underbrace{\mathbb{E}\Big[(x[n]-\hat{x}[n|n-1])(y[n]-\hat{x}[n|n-1])\Big]}_{\text{from (7), numerator of } K[n] \Rightarrow M[n|n-1]} \\
&\quad + K^2[n]\underbrace{\mathbb{E}\left[(y[n]-\hat{x}[n|n-1])^2\right]}_{\text{denominator of } K[n] \Rightarrow \frac{M[n|n-1]}{K[n]}}
\end{aligned}
$$

[9] $\hat{x}[n|n] = \hat{x}[n|n-1] + K[n](y[n] - \hat{x}[n|n-1]$

- Thus,

$$\begin{aligned} M[n|n] &= M[n|n-1] - 2K[n]M[n|n-1] + K[n]M[n|n-1] \\ &= (1 - K[n])\, M[n|n-1]. \end{aligned}$$

[9] $\hat{x}[n|n] = \hat{x}[n|n-1] + K[n](y[n] - \hat{x}[n|n-1]$

Derivation of Kalman filter: Summary

- *Prediction:*
$$\hat{x}[n|n-1] = a\hat{x}[n-1|n-1]$$
- *Prediction MSE:*
$$M[n|n-1] = a^2 M[n-1|n-1] + \sigma_u^2$$
- *Kalman Gain:*
$$K[n] = \frac{M[n|n-1]}{\sigma_n^2 + M[n|n-1]}$$
- *Correction:*
$$\hat{x}[n|n] = \hat{x}[n|n-1] + K[n](y[n] - \hat{x}[n|n-1]$$
- *Minimum MSE:*
$$M[n|n] = (1 - K[n])M[n|n-1]$$
- ...the same for $\mu_s \neq \emptyset$
- *Initialization:*
$$\hat{x}[-1|-1] = \mathbb{E}[x[-1]] = \mu_s \quad \text{and} \quad M[-1|-1] = \sigma_s^2.$$

Remarks

- *Remark 0:*
  Derived equations hold for $\mu_s \neq \emptyset$ too.
- *Remark 1:*
  - LLS estimates: $\text{LLS}(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2) = \text{LLS}(\boldsymbol{\theta}|\mathbf{y}_1) + \text{LLS}(\boldsymbol{\theta}|\mathbf{y}_1)$, where $\mathbf{y}_1,\ \mathbf{y}_2$ uncorrelated
  - orthogonality principle: $\mathbf{e} \perp$ linear combination of the data thus, Kalman filter is the "optimal" (in MSE sense) recursive LINEAR estimator
  - if Gaussian statistics are employed $\Rightarrow$ Kalman is the "optimal" (in MSE sense) estimator!
- *Remark 2:*
  we used Gauss-Markon process for the parameter to be estimated
  - (state equation) $\quad x[n] = ax[n-1] + u[n]$
  - (observation equation) $\quad y[n] = x[n] + w[n]$
  - $\mathbb{E}[(w[n])^2] = \sigma_n^2 \rightarrow$ function of n and $\mathbb{E}[x[n]] = a^{n+1}\mathbb{E}[x[-1]] = a^{n+1}\mu_s$

  thus, Kalman filer holds for non-WSS processes (we haven't seen that so far)!

Remarks

- *Remark 3:*
  Set $a = 1$ and $\sigma_u^2 = \emptyset \Rightarrow x[n] = x[n-1]$ (prediction: last estimate of $x[n]$).
  In that case:

$$\hat{x}[n|n-1] = \hat{x}[n-1]$$
$$M[n|n-1] = M[n-1]$$
$$\hat{x}[n] = \hat{x}[n-1] + K[n](y[n] - \hat{x}[n-1])$$
$$K[n] = \frac{M[n-1]}{M[n-1] + \sigma_n^2}$$
$$M[n] = (1 - K[n])M[n-1]$$

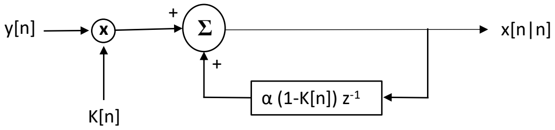
...can omit the prediction stage of Kalman.

Remarks

- *Remark 4:*
  Kalman filter is a time varying linear filter:

$$\hat{x}[n|n] = a\hat{x}[n-1|n-1] + K[n](y[n] - \underbrace{a\hat{x}[n-1|n-1]}_{\hat{x}[n|n-1]})$$

$$= \underbrace{K[n]}_{\text{time varying-constant}} y[n] + \underbrace{a(1-K[n])}_{\text{time varying-constant}} \hat{x}[n-1|n-1]$$



- For $n \to +\infty$ the filter becomes time-invariant (steady-state).

Remarks

- *Remark 5:*
  no-matrix inversion is needed (true only here). For the vector Kalman filter, this is not true.
- *Remark 6:*
  Minimum prediction MSE:

$$\begin{aligned} M[n|n-1] &= \mathbb{E}\left[(x[n]-\hat{x}[n|n-1])^2\right] \\ &= a^2 M[n-1|n-1] + \sigma_u^2 \end{aligned}$$

Kalman Gain:

$$K[n] = \frac{M[n|n-1]}{\sigma_n^2 + M[n|n-1]}$$

Minimum MSE:

$$M[n|n] = (1-K[n])M[n|n-1] \stackrel{\triangle}{=} (1-K[n])\mathbb{E}\left[(x[n]-\hat{x}[n|n-1])^2\right]$$

Thus, $M[n|n]$ can be computed independently of the observation data $\{y[n]\} \Rightarrow$ can be computed offline!

Remarks

- *Remark 7:*
  Kalman filter is a filter: transient response and steady-state response. At steady-state (or $n \to +\infty$) it can be proved that $M[n|n-1] > M[n-1|n-1]$
  Thus, error increases at prediction stage and decreases at correction stage ($K[n] < 1$)

$$M[n|n] = (1 - K[n])M[n|n-1] < M[n|n-1]$$

- *Remark 8:*
  Infinite-length causal Wiener filter:

$$\hat{x}[n] = \sum_{k=0}^{+\infty} h[k]y[n-k]$$

solved through Wiener-Hopf equations.

Remarks

- *Remark 8 continued:*

  (A) We showed that Gauss-Markov $x[n]$ as $n \to +\infty$ could become WSS

  (B) if $\mathbb{E}\left[(w[n])^2\right] = \sigma_n^2 = \sigma^2$ (time independent)

  if (**A**), (**B**) hold then Kalman $\equiv$ Wiener

- *Remark 9:*

  steady-state: time invariant filter for conditions (**A**), (**B**)

$$K[n] = K[\infty]$$

$$\begin{aligned}
\hat{x}[n|n] &= \hat{x}[n|n-1] + K[n](y[n] - \hat{x}[n|n-1]) \\
&= a\hat{x}[n-1|n-1] + K[n](y[n] - a\hat{x}[n-1|n-1]) \\
&= a(1 - K[\infty])\hat{x}[n-1|n-1] + K[\infty]y[n]
\end{aligned}$$

Remarks

- *Remark 9 continued:*

$$\hat{x}[n|n] - a(1-K[\infty])\hat{x}[n-1|n-1] = K[\infty]y[n]$$
$$\Rightarrow \hat{X}(z) - a(1-K[\infty])\hat{X}(z)z^{-1} = K[\infty]Y(z)$$
$$\Rightarrow H(z) = \frac{K[\infty]}{1-a(1-K[\infty])z^{-1}} = H(z=j\omega) = H(z=j2\pi f)$$

- *Remark 10:*
  Same properties for vector Kalman filter (apart from matrix invertibility).

Remarks

- *Remark 10:*
  Equations of the vector Kalman filter can be found in any estimation theory textbook.
- Derivation of the (scalar or vector) Kalman filter equations can be performed in an elegant, simplified way, using (modern) inference theory!
- Kalman filter = Gaussian Belief Propagation in HMMs!
- Pls take the graduate course on ***Probabilistic Graphical Models and Inference Algorithms*** to see this.

Thank you!

Detection & Estimation Theory: Lecture 24

Prof. Aggelos Bletsas

Technical University of Crete
School of Electrical and Computer Engineering

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Acknowledgment of Support

Technical University of Crete

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education - Technical University of Crete" (EDULLL 153).

The project is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

- Problem Definition & Basic Assumptions
- Prediction/Correction Equations
- Particle Filtering Derivation
  - Importance Sampling
- Remarks

Problem Definition

We limit discussion on scalar case - same reasoning holds for the vector case.

- denote as $p_{n|m} \equiv p_{x_n|y_0,y_1,\cdots,y_m}(x_n|y_0, y_1, \cdots, y_m)$
- again, we follow the same Detection/Estimation notation:
  - y_m is the m-th measurement/observation
  - x_l is the l-th random variable to be estimated ("hidden state")
- General problem:
  estimate $x_0, x_1, \cdots, x_n$ given observations $y_0, y_1, \cdots, y_n$, i.e,

$$\text{find} \quad p_{x_i|y_0,y_1,\cdots,y_n}(x_i|y_0, y_1, \cdots, y_n), \ \underline{0 \leq i \leq n}$$

Prediction/Correction Equations (& Assumptions)

General equations, written in an iterative manner.

- *Prediction Equation:*

$$
\begin{aligned}
&p_{n+1|n}(x_{n+1}|y_0, y_1, \cdots, y_n) = \int_{x_n} p(x_{n+1}, x_n|y_0, y_1, \cdots, y_n)dx_n \\
&= \int_{x_n} p(x_{n+1}|x_n, y_0, y_1, \cdots, y_n)p(x_n|y_0, y_1, \cdots, y_n)dx_n \\
&\overset{1}{=} \int_{x_n} p(x_{n+1}|x_n)p(x_n|y_0, y_1, \cdots, y_n)dx_n \\
&= \int_{x_n} p(x_{n+1}|x_n) \underbrace{p_{n|n}(x_n|y_0, y_1, \cdots, y_n)}_{\text{previous iteration}} dx_n
\end{aligned}
$$

[1] $x_{n+1} \perp y_0, y_1, \cdots, y_n | x_n$

Prediction/Correction Equations (& Assumptions)

- *Update/Correction Equation:*

$$p_{n+1|n+1}(x_{n+1}|y_0, y_1, \cdots, y_{n+1}) = \frac{p(x_{n+1}, y_{n+1}|y_0, y_1, \cdots, y_n)}{p(y_{n+1}|y_0, y_1, \cdots, y_n)}$$

$$= \frac{p(y_{n+1}|x_{n+1}, y_0, y_1, \cdots, y_n)p(x_{n+1}|y_0, y_1, \cdots, y_n)}{\underbrace{p(y_{n+1}|y_0, y_1, \cdots, y_n)}_{z}}$$

$$\stackrel{2}{=} \frac{1}{z} p(y_{n+1}|x_{n+1}) p_{n+1|n}(x_{n+1}|y_0, y_1, \cdots, y_n)$$

- Notice that the above prediction/correction equations hold for:

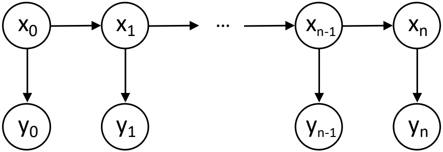
$$x_{n+1} \perp y_0, y_1, \cdots, y_n | x_n \tag{1}$$

$$y_{n+1} \perp y_0, y_1, \cdots, y_n | x_{n+1} \tag{2}$$

[2] $y_{n+1} \perp y_0, y_1, \cdots, y_n | x_{n+1}$

Particle Filter for HMMs with General Continuous

- Eqs. (1), (2) are satisfied by hidden Markov model (HMM):



- HMM satisfies the following:

$$p(x_n|x_{n-1}, x_{n-2}, \cdots, x_0) = p(x_n|x_{n-1}) \tag{3}$$

- Thus, one can produce $(x_0(s), x_1(s), \cdots, x_n(s))$ that adhere to $p_{x_0,x_1,\cdots,x_n}(x_0^n) \triangleq p_{x_0,x_1,\cdots,x_n}(x_0, x_1, \cdots, x_n)$ as follows:

$$x_0(s) \sim p_{x_0}(\cdot),\ \ x_i(s) \sim p_{x_i|x_{i-1}}(\cdot|x_{i-1}(s)),\ \ i = 1, \cdots, n \tag{4}$$

- Note that due to (3),

$$p_{x_0,x_1,\cdots,x_n}(x_0^n) = p_{x_0}(x_0) \prod_{i=1}^{n} p_{x_i|x_{i-1}}(x_i|x_{i-1})$$

Particle Filter Derivation

- HMM also satisfies the following:

$$p(y_n|x_n, x_{n-1}, y_{n-1}, x_{n-2}, y_{n-2}, \cdots, x_0, y_0) = p(y_n|x_n)$$

thus,

$$\begin{aligned}
p(y_n, y_{n-1}, \cdots, y_0|x_n, x_{n-1}, \cdots, x_0) = \\
= p(y_n|y_{n-1}, \cdots, y_0, x_n, x_{n-1}, \cdots, x_0) \\
\cdot\, p(y_{n-1}, \cdots, y_0|x_n, x_{n-1}, \cdots, x_0) \\
\overset{(6)}{=} p(y_n|x_n)p(y_{n-1}, \cdots, y_0|x_n, x_{n-1}, \cdots, x_0)
\end{aligned}$$

Working inductively,

$$p(y_n, y_{n-1}, \cdots, y_0|x_n, x_{n-1}, \cdots, x_0) = \prod_{i=0}^{n} p_{y_i|x_i}(y_i|x_i) \quad (5)$$

Particle Filter Derivation

- x_{n-1} separates $y_{n-1}, \cdots, y_0$ from x_n and thus (6):

$$y_{n-1}, \cdots, y_0 \perp x_n | x_{n-1}, x_{n-2}, \cdots, x_0 \tag{6}$$

- In HMM-particle filtering, we are given $p(x_i|x_{i-1})$ and $p(y_i|x_i),\ 0 \leq i \leq n$
- Thus,[3]

$$p_{x_0,x_1,\cdots,x_n|y_0,y_1,\cdots,y_n}(x_0^n|y_0^n) = \frac{p(y_0^n|x_0^n)p(x_0^n)}{p(y_0^n)}$$

$$= \frac{p_{x_0,x_1,\cdots,x_n}(x_0^n)\prod_{i=0}^{n} p_{y_i|x_i}(y_i|x_i)}{\underbrace{p_{y_0,y_1,\cdots,y_n}(y_0^n)}_{z \,\rightarrow\, \text{unknown (i.e., hard to compute) constant}}} \tag{7}$$

[3] notation: $x_0^n = x_0, x_1, \cdots, x_n$ and $y_0^n = y_0, y_1, \cdots, y_n$

Particle Filter Idea

- Particle filtering idea:
  1) Produce samples that adhere to $p_{x_0,x_1,\cdots,x_n|y_0,y_1,\cdots,y_n}(x_0^n|y_0^n)$ without knowing z.
  2) These samples can be used to estimate any $\underset{p_{x_0^n|y_0^n}}{\mathbb{E}}[f(x_0^n)]$, for any function $f(\cdot)$ as if one perfectly knew the p.d.f. $p_{x_0^n|y_0^n}(\cdot|y_0^n)$.
- We do not know z; can only produce as many samples $\{(x_0(s), x_1(s), \cdots, x_n(s)\},\ s = 1, \cdots, \mathcal{S}$ according to known $\prod_{i=0}^{n} p_{y_i|x_i}(y_i|x_i)$
- Particle filtering is a a case of non-parametric modeling, since we do not know the posterior p.d.f. in closed form $p_{x_0^n|y_0^n}(\cdot|y_0^n)$.

Example

- In localization research, we need the conditional mean (i.e., MMSE estimator):

$$\mathop{\mathbb{E}}_{p_{x_0^n | y_0^n}} \left[f(x_0^n) \right] \equiv \mathbb{E} \left[x_n | y_0, y_1, \cdots, y_n \right]$$

Importance Sampling

- Theory to solve the above problem $\Rightarrow$ *Importance Sampling:*

  $$\mu(x) = \frac{q(x)}{z} \begin{array}{l} \leftarrow \text{ known function} \\ \leftarrow \text{ unknown function} \end{array}$$

  We want samples from $\mu(\cdot)$ so that we can compute the following:

  $$\mathbb{E}_{\mu}\left[f(x)\right] \triangleq \int f(x)\mu(x)dx$$

Important Sampling Algorithm

- *Important Sampling algorithm:*
  1) Produce samples $x(1), x(2), \cdots, x(s), \cdots, x(\mathcal{S})$ from known distribution $v(x)$, called the "proposal" distribution;
  2) Compute as many weights as the samples, according to

$$w(s) = \frac{q(x(s))}{v(x(s))}, \quad s = 1, 2, \cdots, \mathcal{S}$$

  Calculate $\hat{E}(\mathcal{S})$ instead of $\mathbb{E}_\mu [f(x)]$:

$$\mathbb{E}_\mu [f(x)] \to \hat{E}(\mathcal{S}) = \frac{\frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} w(s) f(x(s))}{\frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} w(s)}$$

- *Definition:* support of function $p(x)$(supp(p))

$$\text{supp}(p) = \{x : p(x) > 0\}$$

Important Sampling Algorithm: Theorem

- *Theorem 1:* Let $\text{supp}(\mu) \subseteq \text{supp}(v)$. Then for $\mathcal{S} \to +\infty$

$$\hat{E}(\mathcal{S}) \to \mathbb{E}_{\mu}\left[f(x)\right], \quad \text{with probability 1}$$

- *Proof:*

$$\lim_{\mathcal{S}\to+\infty} \frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} w(s) f(x(s)) \xrightarrow[\text{Large numbers}]{\text{strong Law of}} \underbrace{\mathbb{E}_v\left[w \cdot f(x)\right]}_{\substack{\text{expected value in terms of} \\ v(x) \text{ because we have samples from} \\ \text{the "proposal" pdf } v(x)}}$$

$$\mathbb{E}_v\left[w \cdot f(x)\right] = \mathbb{E}_v\left[\frac{q(x)}{v(x)} f(x)\right] = \int_{\text{supp}(v)} \frac{q(x)}{\cancel{v(x)}} f(x) \cancel{v(x)} dx =$$

$$= \int_{\text{supp}(v)} q(x) f(x) dx \stackrel{4}{=} z \int_{\text{supp}(\mu)} \mu(x) f(x) dx = z\mathbb{E}_{\mu}[f(x)] \quad (8)$$

[4] $q(x) = 0 \ \forall \ x \notin \text{supp}(\mu)$

Important Sampling Algorithm: Theorem

- *Proof continued:*
  similarly,

$$\lim_{\mathcal{S}\to+\infty}\frac{1}{\mathcal{S}}\sum_{s=1}^{\mathcal{S}} w(s) \xrightarrow[\text{Large numbers}]{\text{strong Law of}} \mathbb{E}_v[w] = \int_{\mathrm{supp}(v)} \frac{q(x)}{\cancel{v(x)}}\cancel{v(x)}dx \overset{5}{=}$$

$$= \int_{\mathrm{supp}(\mu)} q(x)dx =$$

$$= z\cancelto{1}{\int_{\mathrm{supp}(\mu)} \mu(x)dx} = z \tag{9}$$

from (8) and (9) $\Rightarrow$ proof is completed. ■

[5] $q(x) = 0 \;\forall\; x \notin \mathrm{supp}(\mu)$

Remarks

- It is worth noting that as long as $\mathrm{supp}(\mu)$ is contained in $\mathrm{supp}(v)$, the estimation converges irrespective of the choice of v.

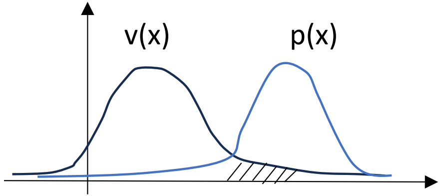


Figure 1: Area where proposal distribution obtains small values, as opposed to true distribution; that amplifies the number of required samples.

- However, the choice of v determines the variance of the estimator and hence the number of samples $\mathcal{S}$ required to obtain good estimation.

Back to HMM-PF

- Need to estimate $\underset{p_{x_0^n|y_0^n}}{\mathbb{E}}[f(x_0^n)|y_0^n]$
- Set

$$q(x_0^n) = p_{x_0,x_1,\cdots,x_n}(x_0^n)\prod_{i=0}^{n} p_{y_i|x_i}(y_i|x_i)$$

- Set

$$v(x_0^n) = \text{ prior } = p_{x_0,x_1,\cdots,x_n}(x_0^n)$$

- For any given $(x_0, x_1, \cdots, x_n) \equiv x_0^n$ sample, calculate the corresponding weight as follows:

$$w(x_0^n) \equiv w_0^n = \frac{q(x_0^n)}{v(x_0^n)} \overset{(7)}{=} \prod_{i=0}^{n} p_{y_i|x_i}(y_i|x_i)$$

HMM-PF Summary

- Thus, for given $y_0, y_1, \cdots, y_n = y_0^n$
  1) Obtain $\mathcal{S}$ iid samples $x_0^n(s),\ 1 \leq s \leq \mathcal{S}$ according to $p_{x_0, x_1, \cdots, x_n}(\cdot)$
  2) Compute $\mathcal{S}$ corresponding weights
  $$w_0^n(s) = \prod_{i=0}^{n} p(y_i | x_i(s)) \quad \forall s,\ 1 \leq s \leq \mathcal{S}$$
  3) Output estimation of $\underset{p_{x_0^n | y_0^n}}{\mathbb{E}} [f(x_0, x_1, \cdots, x_n) | y_0, y_1, \cdots, y_n]$

$$\text{as} \quad \frac{\sum_{s=1}^{\mathcal{S}} w_0^n(s) f(x_0^n(s))}{\sum_{s=1}^{\mathcal{S}} w_0^n(s)}.$$

Remark 1

- The above estimator can be implemented in a sequential fashion, exploiting the HMM properties:
  - for given s in $\{1, 2, \cdots, \mathcal{S}\}$ sample $x_{k+1}(s) \sim p_{x_{k+1}|x_k}(\cdot|x_k(s))$
  - compute

$$\underbrace{w_0^{k+1}(s)}_{\substack{\text{corresponding}\\\text{weight for}\\\text{particle } x_{k+1}(s)}} = \underbrace{w_0^{k+1}(s)}_{\substack{\text{weight for}\\\text{sample } x_k(s)}} \; p_{y_{k+1}|x_{k+1}}(y_{k+1}|x_{k+1}(s))$$

  - repeat for all k up to n and all s up to $\mathcal{S}$
  - re-sample the weights/particles before you proceed to next $k+1$ (particle depletion problem)

Remark 2

- Can we extend particle filtering to other probabilistic graphical models (PGM), other than HMMs?
- *Answer:* YES, using mixture of Gaussians and producing samples based on specific Markov chain Monte Carlo (MCMC) techniques: say $p(x) = \frac{q(x)}{z}$ ($z \rightarrow$ unknown constant)
  You can craft a Markov chain (MC) that produces samples according to $p(x)$, even though the MC was built using $q(x) \rightarrow$ Metropolis-Hastings technique (Gibbs sampling is a special case).
- ...you can take graduate class ***Probabilistic Graphical Models and Inference Algorithms*** to see the above!

Thank you!