# Estimation Theory

Alireza Karimi

Laboratoire d'Automatique, MEC2 397,
email: alireza.karimi@epfl.ch

Spring 2013

## Extract information from noisy signals

**Parameter Estimation Problem :** Given a set of measured data

$$\{x[0], x[1], \ldots, x[N-1]\}$$

which depends on an unknown parameter vector $\boldsymbol{\theta}$, determine an estimator

$$\hat{\boldsymbol{\theta}} = g(x[0], x[1], \ldots, x[N-1])$$

where $g$ is some function.

**Applications :** Image processing, communications, biomedicine, system identification, state estimation in control, etc.

# Some Examples

**Range estimation :** We transmit a pulse that is reflected by the aircraft. An echo is received after $\tau$ second. Range $\theta$ is estimated from the equation $\theta = \tau c/2$ where $c$ is the light's speed.

**System identification :** The input of a plant is excited with $u$ and the output signal $y$ is measured. If we have :

$$y[k] = G(q^{-1}, \boldsymbol{\theta})u[k] + n[k]$$

where $n[k]$ is the measurement noise. The model parameters are estimated using $u[k]$ and $y[k]$ for $k = 1, \ldots, N$.

**DC level in noise :** Consider a set of data $\{x[0], x[1], \ldots, x[N-1]\}$ that can be modeled as :

$$x[n] = A + w[n]$$

where $w[n]$ is some zero mean noise process. $A$ can be estimated using the measured data set.

# Outline

**Classical estimation** ($\theta$ deterministic)

- Minimum Variance Unbiased Estimator (MVU)
- Cramer-Rao Lower Bound (CRLB)
- Best Linear Unbiased Estimator (BLUE)
- Maximum Likelihood Estimator (MLE)
- Least Squares Estimator (LSE)

**Bayesian estimation** ($\theta$ stochastic)

- Minimum Mean Square Error Estimator (MMSE)
- Maximum A Posteriori Estimator (MAP)
- Linear MMSE Estimator
- Kalman Filter

# References

**Main reference :**

## Fundamentals of Statistical Signal Processing Estimation Theory

by Steven M. KAY, Prentice-Hall, 1993 (available in Library de La Fontaine, RLC). We cover Chapters 1 to 14, skipping Chapter 5 and Chapter 9.

**Other references :**

- Lessons in Estimation Theory for Signal Processing, Communications and Control. By Jerry M. Mendel, Prentice-Hall, 1995.
- Probability, Random Processes and Estimation Theory for Engineers. By Henry Stark and John W. Woods, Prentice-Hall, 1986.

# Review of Probability and Random Variables

# Random Variables

**Random Variable :** A rule $X(\cdot)$ that assigns to every element of a sample space $\Omega$ a real value is called a RV. So $X$ is not really a variable that varies randomly but a function whose domain is $\Omega$ and whose range is some subset of the real line.

**Example :** Consider the experiment of flipping a coin twice. The sample space (the possible outcomes) is :

$$\Omega = \{HH, HT, TH, TT\}$$

We can define a random variable $X$ such that

$$X(HH) = 1, X(HT) = 1.1, X(TH) = 1.6, X(TT) = 1.8$$

Random variable $X$ assigns to each event (e.g. $E = \{HT, TH\} \subset \Omega$) a subset of the real line (in this case $B = \{1.1, 1.6\}$).

# Probability Distribution Function

For any element $\zeta$ in $\Omega$, the event $\{\zeta | X(\zeta) \leq x\}$ is an important event. The probability of this event

$$Pr[\{\zeta | X(\zeta) \leq x\}] = P_X(x)$$

is called the probability distribution function of $X$.

**Example :** For the random variable defined earlier, we have :

$$P_X(1.5) = Pr[\{\zeta | X(\zeta) \leq 1.5\}] = Pr[\{HH, HT\}] = 0.5$$

$P_X(x)$ can be computed for all $x \in R$. It is clear that $0 \leq P_X(x) \leq 1$.

**Remark :**

- For the same experiment (throwing a coin twice) we could define another random variable that would lead to a different $P_X(x)$.
- In most of engineering problems the sample space is a subset of the real line so $X(\zeta) = \zeta$ and $P_X(x)$ is a continuous function of $x$.

# Probability Density Function (PDF)

The Probability Density Function, if it exists, is given by :

$$p_X(x) = \frac{dP_X(x)}{dx}$$

When we deal with a single random variable the subscripts are removed :

$$p(x) = \frac{dP(x)}{dx}$$

**Properties :**

$$(i) \qquad \int_{-\infty}^{\infty} p(x)dx = P(\infty) - P(-\infty) = 1$$

$$(ii) \qquad Pr[\{\zeta | X(\zeta) \leq x\}] = Pr[X \leq x] = P(x) = \int_{-\infty}^{x} p(\alpha)d\alpha$$

$$(iii) \qquad Pr[x_1 < X \leq x_2] = \int_{x_1}^{x_2} p(x)dx$$

# Gaussian Probability Density Function

A random variable is distributed according to a Gaussian or normal distribution if the PDF is given by :

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The PDF has two parameters : $\mu$, the mean and $\sigma^2$ the variance.

We note $X \sim \mathcal{N}(\mu, \sigma^2)$ when the random variable $X$ has a normal (Gaussian) distribution with the mean $\mu$ and the standard deviation $\sigma$. Small $\sigma$ means small variability (uncertainty) and large $\sigma$ means large variability.

**Remark :** Gaussian distribution is important because according to the Central Limit Theorem the sum of $N$ independent RVs has a PDF that converges to a Gaussian distribution when $N$ goes to infinity.

# Some other common PDF

Uniform $(b > a)$ : $$p(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Exponential $(\mu > 0)$ : $$p(x) = \frac{1}{\mu} \exp(-x/\mu) u(x)$$

Rayleigh $(\sigma > 0)$ : $$p(x) = \frac{x}{\sigma^2} \exp(\frac{-x^2}{2\sigma^2}) u(x)$$

Chi-square $\chi^2$ : $$p(x) = \begin{cases} \frac{1}{\sqrt{2^n}\Gamma(n/2)} x^{n/2-1} \exp(-\frac{x}{2}) & \text{for } x > 0 \\ 0 & \text{for } x < 0 \end{cases}$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ and $u(x)$ is the unit step function.

# Joint, Marginal and Conditional PDF

**Joint PDF :** Consider two random variables $X$ and $Y$ then :

$$Pr[x_1 < X \leq x_2 \text{ and } y_1 < Y \leq y_2] = \int_{x_1}^{x_2} \int_{y_1}^{y_2} p(x,y)dxdy$$

**Marginal PDF :**

$$p(x) = \int_{-\infty}^{\infty} p(x,y)dy \qquad \text{and} \qquad p(y) = \int_{-\infty}^{\infty} p(x,y)dx$$

**Conditional PDF :** $p(x|y)$ is defined as the PDF of $X$ conditioned on knowing the value of $Y$.

**Bayes' Formula :** Consider two RVs defined on the same probability space then we have :

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x) \qquad \text{or} \qquad p(x|y) = \frac{p(x,y)}{p(y)}$$

# Independent Random Variables

Two RVs $X$ and $Y$ are independent if and only if :

$$p(x, y) = p(x)p(y)$$

A direct conclusion is that :

$$p(x|y) = \frac{p(x, y)}{p(y)} = \frac{p(x)p(y)}{p(y)} = p(x) \qquad \text{and} \qquad p(y|x) = p(y)$$

which means conditioning does not change the PDF.

**Remark :** For a joint Gaussian pdf the contours of constant density is an ellipse centered at $(\mu_x, \mu_y)$. For independent $X$ and $Y$ the major (or minor) axis is parallel to $x$ or $y$ axis.

# Expected Value of a Random Variable

The expected value, if it exists, of a random variable $X$ with PDF $p(x)$ is defined by :

$$E(X) = \int_{-\infty}^{\infty} xp(x)dx$$

**Some properties of expected value :**

- $E\{X + Y\} = E\{X\} + E\{Y\}$
- $E\{aX\} = aE\{X\}$
- The expected value of $Y = g(X)$ can be computed by :

$$E(Y) = \int_{-\infty}^{\infty} g(x)p(x)dx$$

**Conditional expectation :** The conditional expectation of X given that a specific value of $Y$ has occurred is :

$$E(X|Y) = \int_{-\infty}^{\infty} xp(x|y)dx$$

# Moments of a Random Variable

The $r$th moment of $X$ is defined as :

$$E(X^r) = \int_{-\infty}^{\infty} x^r p(x) dx$$

The first moment of X is its expected value or the mean ($\mu = E(X)$).

**Moments of Gaussian RVs :** A Gaussian RV with $\mathcal{N}(\mu, \sigma^2)$ has moments of all orders in closed form

$$\begin{aligned}
E(X) &= \mu \\
E(X^2) &= \mu^2 + \sigma^2 \\
E(X^3) &= \mu^3 + 3\mu\sigma^2 \\
E(X^4) &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \\
E(X^5) &= \mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4 \\
E(X^6) &= \mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6
\end{aligned}$$

# Central Moments

The $r$th central moment of $X$ is defined as :

$$E[(X-\mu)^r] = \int_{-\infty}^{\infty} (x-\mu)^r p(x) dx$$

The second central moment (variance) is denoted by $\sigma^2$ or var$(X)$.

**Central Moments of Gaussian RVs :**

$$E[(X-\mu)^r] = \begin{cases} 0 & \text{if } r \text{ is odd} \\ \sigma^r (r-1)!! & \text{if } r \text{ is even} \end{cases}$$

where $n!!$ denotes the double factorial that is the product of every odd number from n to 1.

# Some properties of Gaussian RVs

- If $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ then $Z = (X - \mu_x)/\sigma_x \sim \mathcal{N}(0, 1)$.
- If $Z \sim \mathcal{N}(0, 1)$ then $X = \sigma_x Z + \mu_x \sim \mathcal{N}(\mu_x, \sigma_x^2)$.
- If $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ then $Z = aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$.
- If $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ are two independent RVs, then

$$aX + bY \sim \mathcal{N}(a\mu_x + b\mu_y, a\sigma_x^2 + b\sigma_y^2)$$

- The sum of square of $n$ independent RV with standard normal distribution $\mathcal{N}(0, 1)$ has a $\chi_n^2$ distribution with $n$ degree of freedom. For large value of $n$, $\chi_n^2$ converges to $\mathcal{N}(n, 2n)$.
- The Euclidian norm $\sqrt{X^2 + Y^2}$ of two independent RVs with standard normal distribution has the Rayleigh distribution.

# Covariance

For two RVs $X$ and $Y$, the covariance is defined as

$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)]$$

$$\sigma_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) p(x, y) dx dy$$

- If $X$ and $Y$ are zero mean then $\sigma_{xy} = E\{XY\}$.
- $\text{var}(X + Y) = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$
- $\text{var}(aX) = a^2 \sigma_x^2$

**Important formula :** The relation between the variance and the mean of $X$ is given by

$$\begin{aligned} \sigma^2 &= E[(X - \mu)^2] = E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

The variance is the mean of the square minus the square of the mean.

# Independence, Uncorrelatedness and Orthogonality

- If $\sigma_{xy} = 0$, then $X$ and $Y$ are uncorrelated and

$$E\{XY\} = E\{X\}E\{Y\}$$

- $X$ and $Y$ called orthogonal if $E\{XY\} = 0$.
- If $X$ and $Y$ are independent then they are uncorrelated.

$$p(x, y) = p(x)p(y) \Rightarrow E\{XY\} = E\{X\}E\{Y\}$$

- Uncorrelatedness does not imply the independence. For example, if $X$ is a normal RV with zero mean and $Y = X^2$ we have $p(y|x) \neq p(y)$ but

$$\sigma_{xy} = E\{XY\} - E\{X\}E\{Y\} = E\{X^3\} - 0 = 0$$

  The correlation only shows the linear dependence between two RV so is weaker than independence.
- For Jointly Gaussian RVs, independence is equivalent to being uncorrelated.

# Random Vectors

**Random Vector :** is a vector of random variables[1] :

$$\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$$

**Expectation Vector :** $\boldsymbol{\mu}_x = E(\mathbf{x}) = [E(x_1), E(x_2), \ldots, E(x_n)]^T$
**Covariance Matrix :** $\mathbf{C}_x = E[(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^T]$

- $\mathbf{C}_x$ is an $n \times n$ symmetric matrix which is assumed to be positive definite and so invertible.
- The elements of this matrix are : $[\mathbf{C}_x]_{ij} = E\{[x_i - E(x_i)][x_j - E(x_j)]\}$.
- If the random variables are uncorrelated then $\mathbf{C}_x$ is a diagonal matrix.

**Multivariate Gaussian PDF :**

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{C}_x)}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_x)^T \mathbf{C}_x^{-1} (\mathbf{x} - \boldsymbol{\mu}_x)\right]$$

1. In some books (including our main reference) there is no distinction between random variable $X$ and its specific value $x$. From now on we adopt the notation of our reference.

# Random Processes

**Discrete Random Process :** $x[n]$ is a sequence of random variables defined for every integer $n$.
**Mean value :** is defined as $E(x[n]) = \mu_x[n]$.
**Autocorrelation Function (ACF) :** is defined as

$$r_{xx}[k, n] = E(x[n]x[n + k])$$

**Wide Sense Stationary (WSS) :** $x[n]$ is WSS if its mean and its autocorrelation function (ACF) do not depend on $n$.
**Autocovariance function :** is defined as

$$c_{xx}[k] = E[(x[n] - \mu_x)(x[n + k] - \mu_x)] = r_{xx}[k] - \mu_x^2$$

**Cross-correlation Function (CCF) :** is defined as

$$r_{xy}[k] = E(x[n]y[n + k])$$

**Cross-covariance function :** is defined as

$$c_{xy}[k] = E[(x[n] - \mu_x)(y[n + k] - \mu_y)] = r_{xy}[k] - \mu_x\mu_y$$

# Discrete White Noise

**Some properties of ACF and CCF :**

$$r_{xx}[0] \geq |r_{xx}[k]| \qquad r_{xx}[k] = r_{xx}[-k] \qquad r_{xy}[k] = r_{yx}[-k]$$

**Power Spectral Density :** The Fourier transform of ACF and CCF gives the Auto-PSD and Cross-PSD :

$$P_{xx}(f) = \sum_{k=-\infty}^{\infty} r_{xx}[k] \exp(-j2\pi fk)$$

$$P_{xy}(f) = \sum_{k=-\infty}^{\infty} r_{xy}[k] \exp(-j2\pi fk)$$

**Discrete White Noise :** is a discrete random process with zero mean and $r_{xx}[k] = \sigma^2\delta[k]$ where $\delta[k]$ is the Kronecker impulse function. The PSD of white noise becomes $P_{xx}(f) = \sigma^2$ and is completely flat with frequency.

# Introduction and Minimum Variance Unbiased Estimation

# The Mathematical Estimation Problem

**Parameter Estimation Problem :** Given a set of measured data

$$\mathbf{x} = \{x[0], x[1], \ldots, x[N-1]\}$$

which depends on an unknown parameter vector $\boldsymbol{\theta}$, determine an estimator

$$\hat{\boldsymbol{\theta}} = g(x[0], x[1], \ldots, x[N-1])$$

where $g$ is some function.

**The first step is to find the PDF of data as a function of $\boldsymbol{\theta}$ : $p(\mathbf{x}; \boldsymbol{\theta})$**

**Example :** Consider the problem of DC level in white Gaussian noise with one observed data $x[0] = \theta + w[0]$ where $w[0]$ has the PDF $\mathcal{N}(0, \sigma^2)$. Then the PDF of $x[0]$ is :

$$p(x[0]; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[0] - \theta)^2\right]$$

# The Mathematical Estimation Problem

**Example :** Consider a data sequence that can be modeled with a linear trend in white Gaussian noise

$$x[n] = A + Bn + w[n] \qquad n = 0, 1, \ldots, N-1$$

Suppose that $w[n] \sim \mathcal{N}(0, \sigma^2)$ and is uncorrelated with all the other samples. Letting $\boldsymbol{\theta} = [A\ B]$ and $\mathbf{x} = [x[0], x[1], \ldots, x[N-1]]$ the PDF is :

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{n=0}^{N-1} p(x[n]; \boldsymbol{\theta}) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A - Bn)^2\right]$$

The quality of any estimator for this problem is related to the assumptions on the data model. In this example, linear trend and WGN PDF assumption.

# The Mathematical Estimation Problem

## Classical versus Bayesian estimation

- If we assume $\boldsymbol{\theta}$ is deterministic we will have a classical estimation problem. The following method will be studied : MVU, MLE, BLUE, LSE.
- If we assume $\boldsymbol{\theta}$ is a random variable with a known PDF, then we will have a Bayesian estimation problem. In this case the data are described as the joint PDF

$$p(x, \theta) = p(x|\theta)p(\theta)$$

where $p(\theta)$ summarizes our knowledge about $\theta$ before any data is observed and $p(x|\theta)$ summarizes our knowledge provided by data $x$ conditioned on knowing $\theta$. The following methods will be studied : MMSE, MAP, Kalman Filter.

# Assessing Estimator Performance

Consider the problem of estimating a DC level $A$ in uncorrelated noise :

$$x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1$$

Consider the following estimators :

- $\hat{A}_1 = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$
- $\hat{A}_2 = x[0]$

Suppose that $A = 1, \hat{A}_1 = 0.95$ and $\hat{A}_2 = 0.98$. Which estimator is better ?

**An estimator is a random variable, so its performance can only be described by its PDF or statistically (e.g. by Monte-Carlo simulation).**

# Unbiased Estimators

An estimator that *on the average* yield the true value is unbiased.
Mathematically

$$E(\theta - \hat{\theta}) = 0 \qquad \text{for } a < \theta < b$$

Let's compute the expectation of the two estimators $\hat{A}_1$ and $\hat{A}_2$ :

$$E(\hat{A}_1) = \frac{1}{N}\sum_{n=0}^{N-1} E(x[n]) = \frac{1}{N}\sum_{n=0}^{N-1} E(A + w[n]) = \frac{1}{N}\sum_{n=0}^{N-1}(A + 0) = A$$

$$E(\hat{A}_2) = E(x[0]) = E(A + w[0]) = A + 0 = A$$

Both estimators are unbiased. Which one is better ?
Now, let's compute the variance of the two estimators :

$$\text{var}(\hat{A}_1) = \text{var}\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \frac{1}{N^2}\sum_{n=0}^{N-1}\text{var}(x[n]) = \frac{1}{N^2}N\sigma^2 = \frac{\sigma^2}{N}$$

$$\text{var}(\hat{A}_2) = \text{var}(x[0]) = \sigma^2 > var(\hat{A}_1)$$

# Unbiased Estimators

**Remark :** When several unbiased estimators of the same parameters from independent set of data are available, i.e., $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$, a better estimator can be obtained by averaging :

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i \quad \Rightarrow \quad E(\hat{\theta}) = \theta$$

Assuming that the estimators have the same variance, we have :

$$\text{var}(\hat{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(\hat{\theta}_i) = \frac{1}{n^2} n\,\text{var}(\hat{\theta}_i) = \frac{\text{var}(\hat{\theta}_i)}{n}$$

By increasing $n$, the variance will decrease (if $n \to \infty, \hat{\theta} \to \theta$).
It is not the case for biased estimators, no matter how many estimators are averaged.

# Minimum Variance Criterion

The most logical criterion for estimation is the Mean Square Error (MSE) :

$$\text{mse}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Unfortunately this type of estimators leads to unrealizable estimators (the estimator will depend on $\theta$).

$$\text{mse}(\hat{\theta}) = E\{[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2\} = E\{[\hat{\theta} - E(\hat{\theta}) + b(\theta)]^2\}$$

where $b(\theta) = E(\hat{\theta}) - \theta$ is defined as the bias of the estimator. Therefore :

$$\text{mse}(\hat{\theta}) = E\{[\hat{\theta} - E(\hat{\theta})]^2\} + 2b(\theta)E[\hat{\theta} - E(\hat{\theta})] + b^2(\theta) = \text{var}(\hat{\theta}) + b^2(\theta)$$

**Instead of minimizing MSE we can minimize the variance of the unbiased estimators :**
**Minimum Variance Unbiased Estimator**

# Minimum Variance Unbiased Estimator

**Existence of MVU Estimator :** In general MVU estimator does not always exist. There may be no unbiased estimator or non of unbiased estimators has a uniformly minimum variance.

**Finding the MVU Estimator :** There is no known procedure which always leads to the MVU estimator. Three existing approaches are :

1. Determine the Cramer-Rao lower bound (CRLB) and check to see if some estimator satisfies it.
2. Apply the Rao-Blackwell-Lehmann-Scheffe theorem (we will skip it).
3. Restrict to linear unbiased estimators.

# Cramer-Rao Lower Bound

# Cramer-Rao Lower Bound

CRLB is a lower bound on the variance of any unbiased estimator.

$$\text{var}(\hat{\theta}) \geq CRLB(\theta)$$

- Note that the CRLB is a function of $\theta$.
- It tells us what is the best performance that can be achieved (useful in feasibility study and comparison with other estimators).
- It may lead us to compute the MVU estimator.

# Cramer-Rao Lower Bound

## Theorem (scalar case)

*Assume that the PDF $p(\mathbf{x};\theta)$ satisfies the regularity condition*

$$E\left[\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right] = 0 \qquad \text{for all } \theta$$

*Then the variance of any unbiased estimator $\hat{\theta}$ satisfies*

$$var(\hat{\theta}) \geq \left[-E\left(\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right)\right]^{-1}$$

An unbiased estimator that attains the CRLB can be found iff :

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta} = I(\theta)(g(\mathbf{x}) - \theta)$$

for some functions $g(\mathbf{x})$ and $I(\theta)$. The estimator is $\hat{\theta} = g(\mathbf{x})$ and the minimum variance is $1/I(\theta)$.

# Cramer-Rao Lower Bound

**Example :** Consider $x[0] = A + w[0]$ with $w[0] \sim \mathcal{N}(0, \sigma^2)$.

$$p(x[0]; A) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[0] - A)^2\right]$$

$$\ln p(x[0], A) = -\ln\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x[0] - A)^2$$

Then

$$\frac{\partial \ln p(x[0]; A)}{\partial A} = \frac{1}{\sigma^2}(x[0] - A) \quad \Rightarrow \quad -\frac{\partial^2 \ln p(x[0]; A)}{\partial A^2} = \frac{1}{\sigma^2}$$

According to Theorem :

$$var(\hat{A}) \geq \sigma^2 \quad \text{and} \quad I(A) = \frac{1}{\sigma^2} \quad \text{and} \quad \hat{A} = g(x[0]) = x[0]$$

# Cramer-Rao Lower Bound

**Example :** Consider multiple observations for DC level in WGN :

$$x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1 \qquad \text{with } w[n] \sim \mathcal{N}(0, \sigma^2)$$

$$p(\mathbf{x}; A) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2\right]$$

Then

$$\begin{aligned}\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} &= \frac{\partial}{\partial A}\left[-\ln[(2\pi\sigma^2)^{N/2}] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A)^2\right] \\ &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n] - A) = \frac{N}{\sigma^2}\left(\frac{1}{N}\sum_{n=0}^{N-1}x[n] - A\right)\end{aligned}$$

According to Theorem :

$$var(\hat{A}) \geq \frac{\sigma^2}{N} \quad \text{and} \quad I(A) = \frac{N}{\sigma^2} \quad \text{and} \quad \hat{A} = g(x[0]) = \frac{1}{N}\sum_{n=0}^{N-1}x[n]$$

# Transformation of Parameters

If it is desired to estimate $\alpha = g(\theta)$, then the CRLB is :

$$\text{var}(\hat{\theta}) \geq \left[ -E\left( \frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2} \right) \right]^{-1} \left( \frac{\partial g}{\partial \theta} \right)^2$$

**Example :** Compute the CRLB for estimation of the power ($A^2$) of a DC level in noise :

$$\text{var}(\hat{A^2}) \geq \frac{\sigma^2}{N}(2A)^2 = \frac{4A^2\sigma^2}{N}$$

## Definition

**Efficient estimator :** An unbiased estimator that attains the CRLB is said to be efficient.

**Example :** Knowing that $\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$ is an efficient estimator for $A$, is $\bar{x}^2$ an efficient estimator for $A^2$ ?

# Transformation of Parameters

**Solution :** Knowing that $\bar{x} \sim \mathcal{N}(A, \sigma^2/N)$, we have :

$$E(\bar{x}^2) = E^2(\bar{x}) + \text{var}(\bar{x}) = A^2 + \frac{\sigma^2}{N} \neq A^2$$

So the estimator $\widehat{A^2} = \bar{x}^2$ is not even unbiased.
Let's look at the variance of this estimator :

$$\text{var}(\bar{x}^2) = E(\bar{x}^4) - E^2(\bar{x}^2)$$

but we have from the moments of Gaussian RVs (slide 15) :

$$E(\bar{x}^4) = A^4 + 6A^2\frac{\sigma^2}{N} + 3(\frac{\sigma^2}{N})^2$$

Therefore :

$$\text{var}(\bar{x}^2) = A^4 + 6A^2\frac{\sigma^2}{N} + \frac{3\sigma^4}{N^2} - \left(A^2 + \frac{\sigma^2}{N}\right)^2 = \frac{4A^2\sigma^2}{N} + \frac{2\sigma^4}{N^2}$$

# Transformation of Parameters

**Remarks :**

- The estimator $\widehat{A^2} = \bar{x}^2$ is biased and not efficient.
- As $N \to \infty$ the bias goes to zero and the variance of the estimator approaches the CRLB. This type of estimators are called **asymptotically efficient**.

**General Remarks :**

- If $g(\theta) = a\theta + b$ is an affine function of $\theta$, then $\widehat{g(\theta)} = g(\hat{\theta})$ is an efficient estimator. First, it is unbiased : $E(a\hat{\theta} + b) = a\theta + b = g(\theta)$, moreover :

$$\text{var}(\widehat{g(\theta)}) \geq \left(\frac{\partial g}{\partial \theta}\right)^2 \text{var}(\hat{\theta}) = a^2 \text{var}(\hat{\theta})$$

  but $\text{var}(\widehat{g(\theta)}) = \text{var}(a\hat{\theta} + b) = a^2\text{var}(\hat{\theta})$, so that the CRLB is achieved.
- If $g(\theta)$ is a nonlinear function of $\theta$ and $\hat{\theta}$ is an efficient estimator, then $g(\hat{\theta})$ is an asymptotically efficient estimator.

# Cramer-Rao Lower Bound

## Theorem (Vector Parameter)

*Assume that the PDF $p(\mathbf{x};\boldsymbol{\theta})$ satisfies the regularity condition*

$$E\left[\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0} \qquad \text{for all } \boldsymbol{\theta}$$

*Then the variance of any unbiased estimator $\hat{\boldsymbol{\theta}}$ satisfies $C_{\hat{\theta}} - \mathbf{I}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0}$ where $\geq \mathbf{0}$ means that the matrix is positive semidefinite. $\mathbf{I}(\boldsymbol{\theta})$ is called the Fisher information matrix and is given by :*

$$\mathbf{I}_{ij}(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right)$$

An unbiased estimator that attains the CRLB can be found iff :

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta})$$

# CRLB Extension to Vector Parameter

**Example :** Consider a DC level in WGN with $A$ and $\sigma^2$ unknown. Compute the CRLB for estimation of $\boldsymbol{\theta} = [A \quad \sigma^2]^T$.

$$\ln p(\mathbf{x};\boldsymbol{\theta}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2$$

The Fisher information matrix is :

$$\mathbf{I}(\boldsymbol{\theta}) = -E\begin{bmatrix} \frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2} & \frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial \sigma^2} \\ \frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial \sigma^2} & \frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}$$

The matrix is diagonal (just for this example) and can be easily inverted to yield :

$$\text{var}(\hat{\theta}) \geq \frac{\sigma^2}{N} \qquad \text{var}(\widehat{\sigma^2}) \geq \frac{2\sigma^4}{N}$$

Is there any unbiased estimator that achieves these bounds ?

# Transformation of Parameters

If it is desired to estimate $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$, and the CRLB for the covariance of $\hat{\boldsymbol{\theta}}$ is $\mathbf{I}^{-1}(\boldsymbol{\theta})$, then :

$$C_{\hat{\boldsymbol{\alpha}}} \geq \left(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}\right)\left[-E\left(\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right)\right]^{-1}\left(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}\right)^T$$

**Example :** Consider a DC level in WGN with $A$ and $\sigma^2$ unknown. Compute the CRLB for estimation of signal to noise ratio $\alpha = A^2/\sigma^2$.
We have $\boldsymbol{\theta} = [A \quad \sigma^2]^T$ and $\alpha = g(\boldsymbol{\theta}) = \theta_1^2/\theta_2$, then the Jacobian is :

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1} \quad \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_2}\right] = \left[\frac{2A}{\sigma^2} \quad -\frac{A^2}{\sigma^4}\right]$$

So the CRLB is :

$$\text{var}(\hat{\alpha}) \geq \left[\frac{2A}{\sigma^2} \quad -\frac{A^2}{\sigma^4}\right]\begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}^{-1}\begin{bmatrix} \frac{2A}{\sigma^2} \\ \frac{-A^2}{\sigma^4} \end{bmatrix} = \frac{4\alpha + 2\alpha^2}{N}$$

# Linear Models with WGN

If $N$ point samples of data observed can be modeled as

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

where

| | | |
|---|---|---|
| $\mathbf{x}$ | $= N \times 1$ | observation vector |
| $\mathbf{H}$ | $= N \times p$ | observation matrix (known, rank $p$) |
| $\boldsymbol{\theta}$ | $= p \times 1$ | vector of parameters to be estimated |
| $\mathbf{w}$ | $= N \times 1$ | noise vector with PDF $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ |

Compute the CRLB and the MVU estimator that achieves this bound.

Step 1 : Compute $\ln p(\mathbf{x}; \boldsymbol{\theta})$.

Step 2 : Compute $\mathbf{I}(\boldsymbol{\theta}) = -E\left[\dfrac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right]$ and the covariance matrix of $\hat{\boldsymbol{\theta}}$ : $C_{\hat{\theta}} = \mathbf{I}^{-1}(\boldsymbol{\theta})$.

Step 3 : Find the MVU estimator $\mathbf{g}(\mathbf{x})$ by factoring

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})[\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta}]$$

# Linear Models with WGN

Step 1 : $\ln p(\mathbf{x};\boldsymbol{\theta}) = -\ln(\sqrt{2\pi\sigma^2})^N - \frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})$.

Step 2 :

$$\begin{aligned}\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= -\frac{1}{2\sigma^2}\frac{\partial}{\partial \boldsymbol{\theta}}[\mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}] \\ &= \frac{1}{\sigma^2}[\mathbf{H}^T\mathbf{x} - \mathbf{H}^T\mathbf{H}\boldsymbol{\theta}]\end{aligned}$$

$$\text{Then} \qquad \mathbf{I}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right] = \frac{1}{\sigma^2}\mathbf{H}^T\mathbf{H}$$

Step 3 : Find the MVU estimator $\mathbf{g}(\mathbf{x})$ by factoring

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})[\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta}] = \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}[(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} - \boldsymbol{\theta}]$$

Therefore :

$$\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x}) = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \qquad C_{\hat{\theta}} = \mathbf{I}^{-1}(\boldsymbol{\theta}) = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}$$

# Linear Models with WGN

- For a linear model with WGN represented by $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ the MVU estimator is :

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

- This estimator is efficient and attains the CRLB.
- That the estimator is unbiased can be seen easily by :

$$E(\hat{\boldsymbol{\theta}}) = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T E(\mathbf{H}\boldsymbol{\theta} + \mathbf{w}) = \boldsymbol{\theta}$$

- The statistical performance of $\hat{\boldsymbol{\theta}}$ is completely specified because $\hat{\boldsymbol{\theta}}$ is a linear transformation of a Gaussian vector $\mathbf{x}$ and hence has a Gaussian distribution :

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$$

# Example (Curve Fitting)

Consider fitting the data $x[n]$ by a $p$-th order polynomial function of $n$ :

$$x[n] = \theta_0 + \theta_1 n + \theta_2 n^2 + \cdots + \theta_p n^p + w[n]$$

We have $N$ data samples, then :

$$\begin{aligned} \mathbf{x} &= [x[0], x[1], \ldots, x[N-1]]^T \\ \mathbf{w} &= [w[0], w[1], \ldots, w[N-1]]^T \\ \boldsymbol{\theta} &= [\theta_0, \theta_1, \ldots, \theta_p]^T \end{aligned}$$

so $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, where $\mathbf{H}$ is :

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 4 & \cdots & 2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & N-1 & (N-1)^2 & \cdots & (N-1)^p \end{bmatrix}_{N \times (p+1)}$$

Hence the MVU estimator is : $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$

# Example (Fourier Analysis)

Consider the Fourier analysis of the data $x[n]$ :

$$x[n] = \sum_{k=1}^{M} a_k \cos(\frac{2\pi kn}{N}) + \sum_{k=1}^{M} b_k \sin(\frac{2\pi kn}{N}) + w[n]$$

so we have $\boldsymbol{\theta} = [a_1, a_2, \ldots, a_M, b_1, b_2, \ldots, b_M]^T$ and $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ where :

$$\mathbf{H} = [\mathbf{h}_1^a, \mathbf{h}_2^a, \ldots, \mathbf{h}_M^a, \mathbf{h}_1^b, \mathbf{h}_2^b, \ldots, \mathbf{h}_M^b]$$

with

$$\mathbf{h}_k^a = \begin{bmatrix} 1 \\ \cos(\frac{2\pi k}{N}) \\ \cos(\frac{2\pi k2}{N}) \\ \vdots \\ \cos(\frac{2\pi k(N-1)}{N}) \end{bmatrix} \quad , \quad \mathbf{h}_k^b = \begin{bmatrix} 1 \\ \sin(\frac{2\pi k}{N}) \\ \sin(\frac{2\pi k2}{N}) \\ \vdots \\ \sin(\frac{2\pi k(N-1)}{N}) \end{bmatrix}$$

Hence the MVU estimate of the Fourier coefficients is : $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$

# Example (Fourier Analysis)

After simplification (noting that $(\mathbf{H}^T\mathbf{H})^{-1} = \frac{2}{N}\mathbf{I}$), we have :

$$\hat{\boldsymbol{\theta}} = \frac{2}{N}[(\mathbf{h}_1^a)^T\mathbf{x}, \ldots, (\mathbf{h}_M^a)^T\mathbf{x}, (\mathbf{h}_1^b)^T\mathbf{x}, \ldots, (\mathbf{h}_M^b)^T\mathbf{x}]^T$$

which is the same as the standard solution :

$$\hat{a}_k = \frac{2}{N}\sum_{n=0}^{N-1} x[n]\cos\left(\frac{2\pi kn}{N}\right), \quad \hat{b}_k = \frac{2}{N}\sum_{n=0}^{N-1} x[n]\sin\left(\frac{2\pi kn}{N}\right)$$

- From the properties of linear models the estimates are unbiased.
- The covariance matrix is :

$$\mathbf{C}_{\hat{\theta}} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1} = \frac{2\sigma^2}{N}\mathbf{I}$$

- Note that $\hat{\boldsymbol{\theta}}$ is Gaussian and $\mathbf{C}_{\hat{\theta}}$ is diagonal (the amplitude estimates are independent).

# Example (System Identification)

Consider identification of a Finite Impulse Response (FIR) model, $h[k]$ for $k = 0, 1, \ldots, p-1$, with input $u[n]$ and output $x[n]$ provided for $n = 0, 1, \ldots, N-1$ :

$$x[n] = \sum_{k=0}^{p-1} h[k]u[n-k] + w[n] \qquad n = 0, 1, \ldots, N-1$$

FIR model can be represented by the linear model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ where

$$\boldsymbol{\theta} = \begin{bmatrix} h[0] \\ h[1] \\ \cdots \\ h[p-1] \end{bmatrix}_{p \times 1} \qquad \mathbf{H} = \begin{bmatrix} u[0] & 0 & \cdots & 0 \\ u[1] & u[0] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u[N-1] & u[N-2] & \cdots & u[N-p] \end{bmatrix}_{N \times p}$$

The MVU estimate is $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ with $\mathbf{C}_{\hat{\theta}} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}$.

# Linear Models with Colored Gaussian Noise

Determine the MVU estimator for the linear model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ with $\mathbf{w}$ a colored Gaussian noise with $\mathcal{N}(\mathbf{0}, \mathbf{C})$.

**Whitening approach :** Since $\mathbf{C}$ is positive definite, its inverse can be factored as $\mathbf{C}^{-1} = \mathbf{D}^T\mathbf{D}$ where $\mathbf{D}$ is an invertible matrix. This matrix acts as a whitening transformation for $\mathbf{w}$ :

$$E[(\mathbf{Dw})(\mathbf{Dw})^T] = E(\mathbf{Dww}^T\mathbf{D}) = \mathbf{DCD}^T = \mathbf{DD}^{-1}\mathbf{D}^{-T}\mathbf{D} = \mathbf{I}$$

Now if we transform the linear model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ to :

$$\mathbf{x}' = \mathbf{Dx} = \mathbf{DH}\boldsymbol{\theta} + \mathbf{Dw} = \mathbf{H}'\boldsymbol{\theta} + \mathbf{w}'$$

where $\mathbf{w}' = \mathbf{Dw} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is white and we can compute the MVU estimator as :

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}'^T\mathbf{H}')^{-1}\mathbf{H}'^T\mathbf{x}' = (\mathbf{H}^T\mathbf{D}^T\mathbf{DH})^{-1}\mathbf{H}^T\mathbf{D}^T\mathbf{Dx}$$

so, we have :

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} \quad \text{with} \quad \mathbf{C}_{\hat{\theta}} = (\mathbf{H}'^T\mathbf{H}')^{-1} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}$$

# Linear Models with known components

Consider a linear model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{s} + \mathbf{w}$, where $\mathbf{s}$ is a known signal. To determine the MVU estimator let $\mathbf{x}' = \mathbf{x} - \mathbf{s}$, so that $\mathbf{x}' = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ is a standard linear model. The MVU estimator is :

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T(\mathbf{x} - \mathbf{s}) \quad \text{with} \quad \mathbf{C}_{\hat{\theta}} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}$$

**Example :** Consider a DC level and exponential in WGN :
$x[n] = A + r^n + w[n]$ where $r$ is known. Then we have :

$$\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} A + \begin{bmatrix} 1 \\ r \\ \vdots \\ r^{N-1} \end{bmatrix} + \begin{bmatrix} w[0] \\ w[1] \\ \vdots \\ w[N-1] \end{bmatrix}$$

The MVU estimator is :

$$\hat{A} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T(\mathbf{x} - \mathbf{s}) = \frac{1}{N}\sum_{n=0}^{N-1}(x[n] - r^n) \quad \text{with} \quad \text{var}(\hat{A}) = \frac{\sigma^2}{N}$$

# Best Linear Unbiased Estimators (BLUE)

**Problems of finding the MVU estimators :**

- The MVU estimator does not always exist or impossible to find.
- The PDF of data may be unknown.

**BLUE is a suboptimal estimator that :**

- restricts estimates to be linear in data ; $\hat{\boldsymbol{\theta}} = \mathbf{Ax}$
- restricts estimates to be unbiased ; $E(\hat{\boldsymbol{\theta}}) = \mathbf{A}E(\mathbf{x}) = \boldsymbol{\theta}$
- minimizes the variance of the estimates ;
- needs only the mean and the variance of the data (not the PDF). As a result, in general, the PDF of the estimates cannot be computed.

**Remark :** The unbiasedness restriction implies a linear model for the data. However, it may still be used if the data are transformed suitably or the model is linearized.

# Finding the BLUE (Scalar Case)

1. Choose a linear estimator for the observed data

$$x[n] \quad , \quad n = 0, 1, \ldots, N-1$$

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] = \mathbf{a}^T \mathbf{x} \quad \text{where} \quad \mathbf{a} = [a_0, a_1, \ldots, a_{N-1}]^T$$

2. Restrict estimate to be unbiased :

$$E(\hat{\theta}) = \sum_{n=0}^{N-1} a_n E(x[n]) = \theta$$

3. Minimize the variance

$$\begin{aligned} \text{var}(\hat{\theta}) &= E\{[\hat{\theta} - E(\hat{\theta})]^2\} = E\{[\mathbf{a}^T \mathbf{x} - \mathbf{a}^T E(\mathbf{x})]^2\} \\ &= E\{\mathbf{a}^T [\mathbf{x} - E(\mathbf{x})][\mathbf{x} - E(\mathbf{x})]^T \mathbf{a}\} = \mathbf{a}^T \mathbf{C} \mathbf{a} \end{aligned}$$

# Finding the BLUE (Scalar Case)

Consider the problem of amplitude estimation of known signals in noise :

$$x[n] = \theta s[n] + w[n]$$

1. Choose a linear estimator : $\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] = \mathbf{a}^T \mathbf{x}$
2. Restrict estimate to be unbiased : $E(\hat{\theta}) = \mathbf{a}^T E(\mathbf{x}) = \mathbf{a}^T \mathbf{s}\theta = \theta$ then $\mathbf{a}^T \mathbf{s} = 1$ where $\mathbf{s} = [s[0], s[1], \ldots, s[N-1]]^T$
3. Minimize $\mathbf{a}^T \mathbf{C} \mathbf{a}$ subject to $\mathbf{a}^T \mathbf{s} = 1$.

The constrained optimization can be solved using Lagrangian Multipliers :

$$\text{Minimize} \qquad J = \mathbf{a}^T \mathbf{C} \mathbf{a} + \lambda(\mathbf{a}^T \mathbf{s} - 1)$$

The optimal solution is :

$$\hat{\theta} = \frac{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{x}}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}} \quad \text{and} \quad \text{var}(\hat{\theta}) = \frac{1}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}}$$

# Finding the BLUE (Vector Case)

## Theorem (Gauss–Markov)

*If the data are of the general linear model form*

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

*with* $\mathbf{w}$ *is a noise vector with zero mean and covariance* $\mathbf{C}$ *(the PDF of* $\mathbf{w}$ *is arbitrary), then the BLUE of* $\boldsymbol{\theta}$ *is :*

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$$

*and the covariance matrix of* $\hat{\boldsymbol{\theta}}$ *is*

$$\mathbf{C}_{\hat{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}$$

**Remark :** If noise is Gaussian then BLUE is MVU estimator.

# Finding the BLUE

**Example :** Consider the problem of DC level in noise : $x[n] = A + w[n]$, where $w[n]$ is of unspecified PDF with $\text{var}(w[n]) = \sigma_n^2$. We have $\theta = A$ and $\mathbf{H} = \mathbf{1} = [1, 1, \ldots, 1]^T$. The covariance matrix is :

$$\mathbf{C} = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{N-1}^2 \end{bmatrix} \quad \Rightarrow \quad \mathbf{C}^{-1} = \begin{bmatrix} \frac{1}{\sigma_0^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_{N-1}^2} \end{bmatrix}$$

and hence the BLUE is :

$$\hat{\theta} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x} = \left( \sum_{n=0}^{N-1} \frac{1}{\sigma_n^2} \right)^{-1} \sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2}$$

and the minimum covariance is :

$$\mathbf{C}_{\hat{\theta}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} = \left( \sum_{n=0}^{N-1} \frac{1}{\sigma_n^2} \right)^{-1}$$

# Maximum Likelihood Estimation

# Maximum Likelihood Estimation

**Problems :** MVU estimator does not often exist or cannot be found. BLUE is restricted to linear models.

**Maximum Likelihood Estimator (MLE) :**

- can always be applied if the PDF is known ;
- is optimal for large data size ;
- is computationally complex and requires numerical methods.

**Basic Idea :** Choose the parameter value that makes the *observed data*, the most likely data to have been observed.

**Likelihood Function :** is the PDF $p(x;\theta)$ when $\theta$ is regarded as a variable (not a parameter).

**ML Estimate :** is the value of $\theta$ that maximizes the likelihood function.

**Procedure :** Find log-likelihood function $\ln p(x;\theta)$ ; differentiate w.r.t $\theta$ and set to zero and solve for $\theta$.

# Maximum Likelihood Estimation

**Example :** Consider DC level in WGN with unknown variance $x[n] = A + w[n]$. Suppose that $A > 0$ and $\sigma^2 = A$. The PDF is :

$$p(\mathbf{x}; A) = \frac{1}{(2\pi A)^{\frac{N}{2}}} \exp\left[-\frac{1}{2A}\sum_{n=0}^{N-1}(x[n] - A)^2\right]$$

Taking the derivative of the log-likelihood function, we have :

$$\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} = -\frac{N}{2A} + \frac{1}{A}\sum_{n=0}^{N-1}(x[n] - A) + \frac{1}{2A^2}\sum_{n=0}^{N-1}(x[n] - A)^2$$

What is the CRLB ? Does an MVU estimator exist ?
MLE can be found by setting the above equation to zero :

$$\hat{A} = -\frac{1}{2} + \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}$$

# Stochastic Convergence

**Convergence in distribution :** Let $\{p(x_N)\}$ be a sequence of PDF. If there exists a PDF $p(x)$ such that

$$\lim_{N\to\infty} p(x_N) = p(x)$$

at every point $x$ at which $p(x)$ is continuous, we say that $p(x_N)$ converges in distribution to $p(x)$. We write also $x_N \xrightarrow{d} x$.

## Example

Consider $N$ independent RV $x_1, x_2, \ldots, x_N$ with mean $\mu$ and finite variance $\sigma^2$. Let $\bar{x}_N = \frac{1}{N}\sum_{n=1}^{N} x_n$, then according to the Central Limit Theorem (CLT), $z_N = \sqrt{N}\frac{\bar{x}_N - \mu}{\sigma}$ converges in distribution to $z \sim \mathcal{N}(0,1)$.

# Stochastic Convergence

**Convergence in probability :** Sequence of random variables $\{x_N\}$ converges in probability to the random variable $x$ if for every $\epsilon > 0$

$$\lim_{N\to\infty} Pr\{|x_N - x| > \epsilon\} = 0$$

**Convergence with probability 1 (almost sure convergence) :** Sequence of random variables $\{x_N\}$ converges with probability 1 to the random variable $x$ if and only if for all possible events

$$Pr\{\lim_{N\to\infty} x_N = x\} = 1$$

## Example

Consider $N$ independent RV $x_1, x_2, \ldots, x_N$ with mean $\mu$ and finite variance $\sigma^2$. Let $\bar{x}_N = \frac{1}{N}\sum_{n=1}^{N} x_n$. Then according to the Law of Large Numbers (LLN), $\bar{z}_N$ converges in probability to $\mu$.

# Asymptotic Properties of Estimators

**Asymptotic Unbiasedness :** Estimator $\hat{\theta}_N$ is an asymptotically unbiased estimator of $\theta$ if :

$$\lim_{N\to\infty} E(\hat{\theta}_N - \theta) = 0$$

**Asymptotic Distribution :** It refers to $p(x_n)$ as it evolves from $n = 1, 2, \ldots$, especially for large value of $n$ (it is not the ultimate form of distribution, which may be degenerate).

**Asymptotic Variance :** is not equal to $\lim_{N\to\infty} \text{var}(\hat{\theta}_N)$ (which is the limiting variance). It is defined as :

$$\text{asymptotic var}(\hat{\theta}_N) = \frac{1}{N} \lim_{N\to\infty} E\{N[\hat{\theta}_N - \lim_{N\to\infty} E(\hat{\theta}_N)]^2\}$$

**Mean-Square Convergence :** Estimator $\hat{\theta}_N$ converges to $\theta$ in a mean-squared sense, if :

$$\lim_{N\to\infty} E[(\hat{\theta}_N - \theta)^2] = 0$$

# Consistency

Estimator $\hat{\theta}_N$ is a consistent estimator of $\theta$ if for every $\epsilon > 0$

$$\text{plim}(\hat{\theta}_N) = \theta \quad \Leftrightarrow \quad \lim_{N\to\infty} Pr[|\hat{\theta}_N - \theta| > \epsilon] = 0$$

**Remarks :**

- If $\hat{\theta}_N$ is asymptotically unbiased and its limiting variance is zero then it converges to $\theta$ in mean-square.
- If $\hat{\theta}_N$ converges to $\theta$ in mean-square, then the estimator is consistent.
- Asymptotic unbiasedness does not imply consistency and vice versa.
- plim can be treated as an operator, e.g. :

$$\text{plim}(xy) = \text{plim}(x)\text{plim}(y) \quad ; \quad \text{plim}\left(\frac{x}{y}\right) = \frac{\text{plim}(x)}{\text{plim}(y)}$$

- The importance of consistency is that any continuous function of a consistent estimator is itself a consistent estimator.

# Maximum Likelihood Estimation

**Properties :** MLE may be biased and is not necessarily an efficient estimator. However :

- MLE is a consistent estimator meaning that

$$\lim_{N\to\infty} Pr[|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \epsilon] = 0$$

- MLE asymptotically attains the CRLB (asymptotic variance is equal to CRLB).
- Under some "regularity" conditions, the MLE is asymptotically normally distributed

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta}))$$

  even if the PDF of $\mathbf{x}$ is not Gaussian.
- If an MVU estimator exists then ML procedure will find it.

# Maximum Likelihood Estimation

## Example

Consider DC level in WGN with known variance $\sigma^2$.

Sol. Then

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]$$

$$\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) = 0$$

$$\Rightarrow \quad \sum_{n=0}^{N-1} x[n] - N\hat{A} = 0$$

which leads to $\hat{A} = \bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$

# MLE for Transformed Parameters (Invariance Property)

## Theorem (Invariance Property of the MLE)

*The MLE of the parameter $\alpha = g(\theta)$, where the PDF $p(\mathbf{x};\theta)$ is parameterized by $\theta$, is given by*

$$\hat{\alpha} = g(\hat{\theta})$$

*where $\hat{\theta}$ is the MLE of $\theta$.*

- It can be proved using the property of the consistent estimators.
- If $\alpha = g(\theta)$ is a one-to-one function, then

$$\hat{\alpha} = \arg\max_{\alpha} p(\mathbf{x}; g^{-1}(\alpha)) = g(\hat{\theta})$$

- If $\alpha = g(\theta)$ is not a one-to-one function, then

$$\bar{p}_T(\mathbf{x};\alpha) = \max_{\{\theta:\alpha=g(\theta)\}} p(\mathbf{x};\theta) \quad \text{and} \quad \hat{\alpha} = \arg\max_{\alpha} \bar{p}_T(\mathbf{x};\theta) = g(\hat{\theta})$$

# MLE for Transformed Parameters (Invariance Property)

## Example

Consider DC level in WGN and find MLE of $\alpha = \exp(A)$. Since $g(\theta)$ is a one-to-one function then :

$$\hat{\alpha} = \arg\max_{\alpha} p_T(\mathbf{x}, \alpha) = \arg\max_{\alpha} p(\mathbf{x}; \ln \alpha) = \exp(\bar{x})$$

## Example

Consider DC level in WGN and find MLE of $\alpha = A^2$. Since $g(\theta)$ is not a one-to-one function then :

$$\begin{aligned}
\hat{\alpha} &= \arg\max_{\alpha \geq 0} \{p(\mathbf{x}; \sqrt{\alpha}), p(\mathbf{x}; -\sqrt{\alpha})\} \\
&= \left[\arg\max_{\sqrt{\alpha} \geq 0} \{p(\mathbf{x}; \sqrt{\alpha}), p(\mathbf{x}; -\sqrt{\alpha})\}\right]^2 \\
&= \left[\arg\max_{-\infty < A < \infty} p(\mathbf{x}; A)\right]^2 = \hat{A}^2 = \bar{x}^2
\end{aligned}$$

# MLE (Extension to Vector Parameter)

## Example

Consider DC Level in WGN with unknown variance. The vector parameter $\boldsymbol{\theta} = [A \quad \sigma^2]^T$ should be estimated.
We have :

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)$$

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{n=0}^{N-1}(x[n]-A)^2$$

which leads to the following MLE :

$$\hat{A} = \bar{x} \qquad \hat{\sigma^2} = \frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\bar{x})^2$$

# MLE for General Gaussian Case

Consider the general Gaussian case where $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{C}(\boldsymbol{\theta}))$.
The partial derivative of the PDF is :

$$\begin{aligned}\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_k} &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_k}\right) + \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})^T}{\partial \theta_k}\mathbf{C}^{-1}(\boldsymbol{\theta})(\mathbf{x}-\boldsymbol{\mu}(\boldsymbol{\theta})) \\ &\quad -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T\frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \theta_k}(\mathbf{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))\end{aligned}$$

for $k = 1, \ldots, p$.

- By setting the above equations equal to zero, MLE can be found.
- A particular case is when $\mathbf{C}$ is known (the first and third terms become zero).
- In addition, if $\boldsymbol{\mu}(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, the general linear model is obtained.

# MLE for General Linear Models

Consider the general linear model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ where $\mathbf{w}$ is a noise vector with PDF $\mathcal{N}(\mathbf{0}, \mathbf{C})$ :

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{C})}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})\right]$$

Taking the derivative of $\ln p(\mathbf{x};\boldsymbol{\theta})$ leads to :

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial (\mathbf{H}\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} \mathbf{C}^{-1}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})$$

Then

$$\mathbf{H}^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}) = 0 \quad \Rightarrow \quad \hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$$

which is the same as MVU estimator. The PDF of $\hat{\boldsymbol{\theta}}$ is :

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1})$$

# MLE (Numerical Method)

**Newton–Raphson :** A closed form estimator cannot be always computed by maximizing the likelihood function. However, the maximum value can be computed by the numerical methods like the iterative Newton–Raphson algorithm.

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1} \left.\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k}$$

**Remarks :**

- The Hessian can be replaced by the negative of its expectation, the Fisher information matrix $\mathbf{I}(\theta)$.
- This method suffers from convergence problems (local maximum).
- Typically, for large data length, the log-likelihood function becomes more quadratic and the algorithm will produce the MLE.

# Least Squares Estimation

# The Least Squares Approach

In all the previous methods we assumed that the measured signal $x[n]$ is the sum of a true signal $s[n]$ and a measurement error $w[n]$ with known probabilistic model. In least squares method

$$x[n] = s[n, \theta] + e[n]$$

where $e(n)$ represents the modeling and measurement errors. The objective is to minimize the LS cost :

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - s[n, \theta])^2$$

**We do not need a probabilistic assumption but only a deterministic signal model.**

- It has a broader range of applications.
- No claim about the optimality can be made.
- The statistical performance cannot be assessed.

# The Least Squares Approach

## Example

Estimate the DC level of a signal. We observe $x[n] = A + \epsilon[n]$ for $n = 0, \ldots, N-1$ and the LS criterion is :

$$J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2$$

$$\frac{\partial J(A)}{\partial A} = -2 \sum_{n=0}^{N-1} (x[n] - A) = 0 \quad \Rightarrow \quad \hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

# Linear Least Squares

Suppose that the observation model is linear $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ then

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum_{n=0}^{N-1}(x[n] - s[n,\theta])^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \\ &= \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} \end{aligned}$$

where $\mathbf{H}$ is full rank. The gradient is

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} = 0 \quad \Rightarrow \hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

The minimum LS cost is :

$$J_{\text{min}} = (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})^T(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) = \mathbf{x}^T[\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T]\mathbf{x} = \mathbf{x}^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

where $\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ is an Idempotent matrix.

# Comparing Different Estimators for the Linear Model

Consider the following linear model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

| Estimator | Assumption | Estimate |
|---|---|---|
| LSE | No probabilistic assumption | $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ |
| BLUE | $\mathbf{w}$ is white with unknown PDF | $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ |
| MLE | $\mathbf{w}$ is white Gaussian noise | $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ |
| MVUE | $\mathbf{w}$ is white Gaussian noise | $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ |

**For MLE and MVUE the PDF of the estimate will be Gaussian.**

# Weighted Linear Least Squares

The LS criterion can be modified by including a positive definite (symmetric) weighting matrix $\mathbf{W}$ :

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{W} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$$

That leads to the following estimator :

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W} \mathbf{x}$$

and minimum LS cost :

$$J_{\min} = \mathbf{x}^T [\mathbf{W} - \mathbf{W}\mathbf{H}(\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W}] \mathbf{x}$$

**Remark :** If we take $\mathbf{W} = \mathbf{C}^{-1}$, where $\mathbf{C}$ is the covariance of noise then weighted least squares estimator is the BLUE. However, there is no true LS-based reason for this choice.

# Geometrical Interpretation

Recall the general signal model $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$. If we denote the column of $\mathbf{H}$ by $\mathbf{h}_i$ we have :

$$\mathbf{s} = [\mathbf{h}_1 \quad \mathbf{h}_2 \quad \cdots \quad \mathbf{h}_p] \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} = \sum_{i=1}^{p} \theta_i \mathbf{h}_i$$

- The signal model is a *linear combination* of the vectors $\{\mathbf{h}_i\}$.
- The LS minimizes the length of the error vector between the data and the signal model $\boldsymbol{\epsilon} = \mathbf{x} - \mathbf{s}$ :

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) = \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2$$

- The data vector can lie anywhere in $R^N$, while signal vectors must lie in a $p$-dimensional subspace of $R^N$, termed $S^p$, which is spanned by the column of $\mathbf{H}$.

# Geometrical Interpretation (Orthogonal Projection)

- Intuitively, it is clear that the LS error is minimized when $\hat{\mathbf{s}} = \mathbf{H}\hat{\boldsymbol{\theta}}$ is the **orthogonal projection** of $\mathbf{x}$ onto $S^p$.
- So the LS error vector $\boldsymbol{\epsilon} = \mathbf{x} - \hat{\mathbf{s}}$ is orthogonal to all columns of $\mathbf{H}$ :

$$\mathbf{H}^T\boldsymbol{\epsilon} = \mathbf{0} \quad \Rightarrow \quad \mathbf{H}^T(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) = \mathbf{0} \quad \Rightarrow \quad \hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

- The signal estimate is the projection of $\mathbf{x}$ onto $S^p$ :

$$\hat{\mathbf{s}} = \mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} = \mathbf{P}\mathbf{x}$$

  where $\mathbf{P}$ is the orthogonal projection matrix.
- Note that if $\mathbf{z} \in \text{Range}(\mathbf{H})$, then $\mathbf{P}\mathbf{z} = \mathbf{z}$. Recall that $\text{Range}(\mathbf{H})$ is the subspace spanned by the columns of $\mathbf{H}$.
- Now, since $\mathbf{P}\mathbf{x} \in S^p$ then $\mathbf{P}(\mathbf{P}\mathbf{x}) = \mathbf{P}\mathbf{x}$. Therefore any projection matrix is idempotent, i.e. $\mathbf{P}^2 = \mathbf{P}$.
- It can be verified that $\mathbf{P}$ is symmetric and singular (with rank $p$).

# Geometrical Interpretation (Orthonormal columns of $H$)

Recall that

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} <\mathbf{h}_1,\mathbf{h}_1> & <\mathbf{h}_1,\mathbf{h}_2> & \cdots & <\mathbf{h}_1,\mathbf{h}_p> \\ <\mathbf{h}_2,\mathbf{h}_1> & <\mathbf{h}_2,\mathbf{h}_2> & \cdots & <\mathbf{h}_2,\mathbf{h}_p> \\ \vdots & \vdots & \ddots & \vdots \\ <\mathbf{h}_p,\mathbf{h}_1> & <\mathbf{h}_p,\mathbf{h}_2> & \cdots & <\mathbf{h}_p,\mathbf{h}_p> \end{bmatrix}$$

- If the columns of $\mathbf{H}$ are orthonormal then $\mathbf{H}^T\mathbf{H} = \mathbf{I}$ and $\hat{\boldsymbol{\theta}} = \mathbf{H}^T\mathbf{x}$.
- In this case we have $\hat{\theta}_i = \mathbf{h}_i^T\mathbf{x}$, thus

$$\hat{\mathbf{s}} = \mathbf{H}\hat{\boldsymbol{\theta}} = \sum_{i=1}^{p}\hat{\theta}_i\mathbf{h}_i = \sum_{i=1}^{p}(\mathbf{h}_i^T\mathbf{x})\mathbf{h}_i$$

- If we increase the number of parameters (the order of the linear model), we can easily compute the new estimate.

# Choosing the Model Order

Suppose that you have a set of data and the objective is to fit a polynomial to data. What is the best polynomial order ?

**Remarks :**

- It is clear that by increasing the order $J_{\min}$ is monotonically non-increasing.
- By choosing $p = N$ we can perfectly fit the data to model. However, we fit the noise as well.
- We should choose the simplest model that adequately describes the data.
- We increase the order only if the cost reduction is significant.
- If we have an idea about the expected level of $J_{\min}$, we increase $p$ to approximately attain this level.
- There is an order-recursive LS algorithm to efficiently compute a $(p+1)$-th order model based on a $p$-th order one (See 8.6).

# Sequential Least Squares

Suppose that $\hat{\theta}[N-1]$ based on $\mathbf{x}[N-1] = \left[x[0], \ldots, x[N-1]\right]$ is available. If we get a new data $x[N]$, we want to compute $\hat{\theta}[N]$ as a function of $\hat{\theta}[N-1]$ and $x[N]$.

## Example

Consider LS estimate of DC level : $\hat{A}[N-1] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$. We have :

$$\begin{aligned}
\hat{A}[N] &= \frac{1}{N+1}\sum_{n=0}^{N} x[n] = \frac{1}{N+1}\left\{N\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] + x[N]\right\} \\
&= \frac{N}{N+1}\hat{A}[N-1] + \frac{1}{N+1}x[N] \\
&= \hat{A}[N-1] + \frac{1}{N+1}(x[N] - \hat{A}[N-1])
\end{aligned}$$

new estimate = old estimate + gain $\times$ prediction error

# Sequential Least Squares

## Example (DC level in uncorrelated noise)

$x[n] = A + w[n]$ and $\text{var}(w[n]) = \sigma_n^2$. The WLS estimate (or BLUE) is :

$$\hat{A}[N-1] = \frac{\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2}}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}}$$

Similar to the previous example we can obtain :

$$\begin{aligned} \hat{A}[N] &= \hat{A}[N-1] + \frac{1/\sigma_N^2}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}} \big(x[N] - \hat{A}[N-1]\big) \\ &= \hat{A}[N-1] + K[N]\big(x[N] - \hat{A}[N-1]\big) \end{aligned}$$

The gain factor $K[N]$ can be reformulated as :

$$K[N] = \frac{\text{var}(\hat{A}[N-1])}{\text{var}(\hat{A}[N-1]) + \sigma_N^2}$$

# Sequential Least Squares

Consider the general linear model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, where $\mathbf{w}$ is an uncorrelated noise with the covariance matrix $\mathbf{C}$. The BLUE (or WLS) is :

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} \quad \text{and} \quad \mathbf{C}_{\hat{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}$$

Let's define :

$$\begin{aligned} \mathbf{C}[n] &= \text{diag}(\sigma_0^2, \sigma_1^2, \ldots, \sigma_n^2) \\ \mathbf{H}[n] &= \begin{bmatrix} \mathbf{H}[n-1] \\ \mathbf{h}^T[n] \end{bmatrix} = \begin{bmatrix} n \times p \\ 1 \times p \end{bmatrix} \\ \mathbf{x}[n] &= \big[x[0], x[1], \ldots, x[n]\big] \end{aligned}$$

The objective is to find $\hat{\boldsymbol{\theta}}[n]$, based on $n+1$ data samples, as a function of $\hat{\boldsymbol{\theta}}[n-1]$ and new data $x[n]$. The batch estimator is :

$$\hat{\boldsymbol{\theta}}[n-1] = \boldsymbol{\Sigma}[n-1]\mathbf{H}^T[n-1]\mathbf{C}^{-1}[n-1]\mathbf{x}[n-1]$$

$$\text{with} \quad \boldsymbol{\Sigma}[n-1] = (\mathbf{H}^T[n-1]\mathbf{C}^{-1}[n-1]\mathbf{H}[n-1])^{-1}$$

# Sequential Least Squares

**Estimator Update :**

$$\hat{\theta}[n] = \hat{\theta}[n-1] + \mathbf{K}[n]\left(x[n] - \mathbf{h}^T[n]\hat{\theta}[n-1]\right)$$

where

$$\mathbf{K}[n] = \frac{\mathbf{\Sigma}[n-1]\mathbf{h}[n]}{\sigma_n^2 + \mathbf{h}^T[n]\mathbf{\Sigma}[n-1]\mathbf{h}[n]}$$

**Covariance Update :**

$$\mathbf{\Sigma}[n] = (\mathbf{I} - \mathbf{K}[n]\mathbf{h}^T[n])\mathbf{\Sigma}[n-1]$$

The following Lemma is used to compute the updates :

## Matrix Inversion Lemma

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}[\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B}]^{-1}\mathbf{DA}^{-1}$$

with $\mathbf{A} = \mathbf{\Sigma}^{-1}[n-1], \mathbf{B} = \mathbf{h}[n], \mathbf{C} = 1/\sigma_n^2, \mathbf{D} = \mathbf{h}^T[n]$.

**Initialization ?**

# Constrained Least Squares

Linear constraints in the form $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$ can be considered in LS solution.
The LS criterion becomes :

$$J_c(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) + \boldsymbol{\lambda}^T(\mathbf{A}\boldsymbol{\theta} - \mathbf{b})$$

$$\frac{\partial J_c}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta} + \mathbf{A}^T\boldsymbol{\lambda}$$

Setting the gradient equal to zero produces :

$$\begin{aligned} \hat{\boldsymbol{\theta}}_c &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} - \frac{1}{2}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\boldsymbol{\lambda} \\ &= \hat{\boldsymbol{\theta}} - (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\frac{\boldsymbol{\lambda}}{2} \end{aligned}$$

where $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ is the unconstrained LSE.
Now $\mathbf{A}\hat{\boldsymbol{\theta}}_c = \mathbf{b}$ can be solved to find $\boldsymbol{\lambda}$ :

$$\boldsymbol{\lambda} = 2[\mathbf{A}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T]^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{b})$$

# Nonlinear Least Squares

Many applications have nonlinear observation model : $\mathbf{s}(\boldsymbol{\theta}) \neq \mathbf{H}\boldsymbol{\theta}$. This leads to a *nonlinear optimization problem* that can be solved numerically.

**Newton-Raphson Method :** Find a zero of the gradient of the criterion by linearizing the gradient around the current estimate.

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k - \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{-1} \mathbf{g}(\boldsymbol{\theta})\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_k}$$

where $\mathbf{g}(\boldsymbol{\theta}) = \dfrac{\partial \mathbf{s}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}[\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})]$ and

$$\frac{\partial [\mathbf{g}(\boldsymbol{\theta})]_i}{\partial \theta_j} = \sum_{n=0}^{N-1} \left[ (x[n] - s[n]) \frac{\partial^2 s[n]}{\partial \theta_i \theta_j} - \frac{\partial \mathbf{s}[n]}{\partial \theta_j} \frac{\partial \mathbf{s}[n]}{\partial \theta_i} \right]$$

Around the solution $[\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})]$ is small so the first term in the Jacobian can be neglected. This makes this method equivalent to the Gauss-Newton algorithm which is numerically more robust.

# Nonlinear Least Squares

**Gauss-Newton Method :** Linearize signal model around the current estimate and solve the resulting linear problem.

$$\mathbf{s}(\boldsymbol{\theta}) \approx \mathbf{s}(\boldsymbol{\theta}_0) + \left.\frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \mathbf{s}(\boldsymbol{\theta}_0) + \mathbf{H}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

The solution to the linearized problem is :

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= [\mathbf{H}^T(\boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)]^{-1}\mathbf{H}^T(\boldsymbol{\theta}_0)[\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}_0) + \mathbf{H}(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0] \\ &= \boldsymbol{\theta}_0 + [\mathbf{H}^T(\boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)]^{-1}\mathbf{H}^T(\boldsymbol{\theta}_0)[\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}_0)] \end{aligned}$$

If we now iterate the solution, it becomes :

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + [\mathbf{H}^T(\hat{\boldsymbol{\theta}}_k)\mathbf{H}(\hat{\boldsymbol{\theta}}_k)]^{-1}\mathbf{H}^T(\hat{\boldsymbol{\theta}}_k)[\mathbf{x} - \mathbf{s}(\hat{\boldsymbol{\theta}}_k)]$$

**Remark :** Both the Newton-Raphson and the Gauss-Newton methods can have convergence problems.

# Nonlinear Least Squares

**Transformation of parameters :** Transform into a linear problem by seeking for an invertible function $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$ such that :

$$\mathbf{s}(\boldsymbol{\theta}) = \mathbf{s}(\mathbf{g}^{-1}(\boldsymbol{\alpha})) = \mathbf{H}\boldsymbol{\alpha}$$

So the nonlinear LSE is $\hat{\boldsymbol{\theta}} = \mathbf{g}^{-1}(\hat{\boldsymbol{\alpha}})$ where $\hat{\boldsymbol{\alpha}} = (\mathbf{H}^{\mathbf{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{x}$.

## Example (Estimate the amplitude and phase of a sinusoidal signal)

$$s[n] = A\cos(2\pi f_0 n + \phi) \qquad n = 0, 1, \ldots, N-1$$

The LS problem is nonlinear, however, we have :

$$A\cos(2\pi f_0 n + \phi) = A\cos\phi\cos 2\pi f_0 n - A\sin\phi\sin 2\pi f_0 n$$

if we let $\alpha_1 = A\cos\phi$ and $\alpha_2 = -A\sin\phi$ then the signal model becomes linear $\mathbf{s} = \mathbf{H}\boldsymbol{\alpha}$. The LSE is $\hat{\boldsymbol{\alpha}} = (\mathbf{H}^{\mathbf{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{x}$ and $\mathbf{g}^{-1}(\boldsymbol{\alpha})$ is

$$A = \sqrt{\alpha_1^2 + \alpha_2^2} \qquad \text{and} \qquad \phi = \arctan\left(\frac{-\alpha_2}{\alpha_1}\right)$$

# Nonlinear Least Squares

**Separability of parameters :** If the signal model is linear in some of the parameters, try to write it as $\mathbf{s} = \mathbf{H}(\boldsymbol{\alpha})\boldsymbol{\beta}$. The LS error can be minimized wrt $\boldsymbol{\beta}$.

$$\hat{\boldsymbol{\beta}} = [\mathbf{H}^{\mathbf{T}}(\boldsymbol{\alpha})\mathbf{H}(\boldsymbol{\alpha})]^{-1}\mathbf{H}^{\mathbf{T}}(\boldsymbol{\alpha})\mathbf{x}$$

The cost function is a function of $\boldsymbol{\alpha}$ that can be minimized by a numerical method (e.g. brute force).

$$J(\boldsymbol{\alpha}, \hat{\boldsymbol{\beta}}) = \mathbf{x}^{\mathbf{T}}\left[\mathbf{I} - \mathbf{H}(\boldsymbol{\alpha})[\mathbf{H}^{\mathbf{T}}(\boldsymbol{\alpha})\mathbf{H}(\boldsymbol{\alpha})]^{-1}\mathbf{H}^{\mathbf{T}}(\boldsymbol{\alpha})\right]\mathbf{x}$$

Then

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \quad \mathbf{x}^{\mathbf{T}}\left[\mathbf{H}(\boldsymbol{\alpha})[\mathbf{H}^{\mathbf{T}}(\boldsymbol{\alpha})\mathbf{H}(\boldsymbol{\alpha})]^{-1}\mathbf{H}^{\mathbf{T}}(\boldsymbol{\alpha})\right]\mathbf{x}$$

**Remark :** This method is interesting if the dimension of $\boldsymbol{\alpha}$ is much less than the dimension of $\boldsymbol{\beta}$.

## Example (Damped Exponentials)

Consider the following signal model :

$$s[n] = A_1 r^n + A_2 r^{2n} + A_3 r^{3n} \qquad n = 0, 1, \ldots, N-1$$

with $\boldsymbol{\theta} = [A_1, A_2, A_3, r]^T$. It is known that $0 < r < 1$. Let's take $\boldsymbol{\beta} = [A_1, A_2, A_3]^T$ so we get $\mathbf{s} = \mathbf{H}(r)\boldsymbol{\beta}$ with :

$$\mathbf{H}(r) = \begin{bmatrix} 1 & 1 & 1 \\ r & r^2 & r^3 \\ \vdots & \vdots & \vdots \\ r^{N-1} & r^{2(N-1)} & r^{3(N-1)} \end{bmatrix}$$

Step 1 : Maximize $\mathbf{x}^T \left[\mathbf{H}(r)[\mathbf{H}^T(r)\mathbf{H}(r)]^{-1}\mathbf{H}^T(r)\right]\mathbf{x}$ to obtain $\hat{r}$.

Step 2 : Compute $\hat{\boldsymbol{\beta}} = [\mathbf{H}^T(\hat{r})\mathbf{H}(\hat{r})]^{-1}\mathbf{H}^T(\hat{r})\mathbf{x}$.

# The Bayesian Philosophy

# Introduction

**Classical Approach :**

- Assumes $\boldsymbol{\theta}$ is unknown but *deterministic*.
- Some prior knowledge on $\boldsymbol{\theta}$ cannot be used.
- Variance of the estimate may depend on $\boldsymbol{\theta}$.
- MVUE may not exist.
- In Monte-Carlo Simulations, we do $M$ runs for each fixed $\boldsymbol{\theta}$ and then compute sample mean and variance for each $\boldsymbol{\theta}$ (no averaging over $\boldsymbol{\theta}$).

**Bayesian Approach :**

- Assumes $\boldsymbol{\theta}$ is *random* with a known prior PDF, $p(\boldsymbol{\theta})$.
- We estimate a *realization* of $\boldsymbol{\theta}$ based on the available data.
- Variance of the estimate does not depend on $\boldsymbol{\theta}$.
- A Bayesian estimate always exists
- In Monte-Carlo simulations, we do $M$ runs for randomly chosen $\boldsymbol{\theta}$ and then we compute sample mean and variance over all $\boldsymbol{\theta}$ values.

# Mean Square Error (MSE)

## Classical MSE

Classical MSE is a function of the unknown parameter $\theta$ and cannot be used for constructing the estimators.

$$\text{mse}(\hat{\theta}) = E\{[\theta - \hat{\theta}(\mathbf{x})]^2\} = \int [\theta - \hat{\theta}(\mathbf{x})]^2 p(\mathbf{x};\theta) d\mathbf{x}$$

Note that the $E\{\cdot\}$ is wrt the PDF of $\mathbf{x}$.

## Bayesian MSE :

Bayesian MSE is not a function of $\theta$ and can be minimized to find an estimator.

$$\text{Bmse}(\hat{\theta}) = E\{[\theta - \hat{\theta}(\mathbf{x})]^2\} = \int \int [\theta - \hat{\theta}(\mathbf{x})]^2 p(\mathbf{x},\theta) d\mathbf{x} d\theta$$

Note that the $E\{\cdot\}$ is wrt the joint PDF of $\mathbf{x}$ and $\theta$.

# Minimum Mean Square Error Estimator

Consider the estimation of $A$ that minimizes the Bayesian MSE, where $A$ is a random variable with uniform PDF $p(A) = \mathcal{U}[-A_0, A_0]$ and independent of $w[n]$.

$$\text{Bmse}(\hat{\theta}) = \int \int [A - \hat{A}]^2 p(\mathbf{x}, \theta) d\mathbf{x} dA = \int \left[ \int [A - \hat{A}]^2 p(A|\mathbf{x}) dA \right] p(\mathbf{x}) d\mathbf{x}$$

where we used Bayes' theorem : $p(\mathbf{x}, \theta) = p(A|\mathbf{x})p(\mathbf{x})$.
Since $p(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, we need only minimize the integral in brackets.
We set the derivative equal to zero :

$$\frac{\partial}{\partial \hat{A}} \int [A - \hat{A}]^2 p(A|\mathbf{x}) dA = -2 \int A p(A|\mathbf{x}) dA + 2\hat{A} \int p(A|\mathbf{x}) dA$$

which results in

$$\hat{A} = \int A p(A|\mathbf{x}) dA = E(A|\mathbf{x})$$

**Bayesian MMSE Estimate :** is the conditional mean of $A$ given data $\mathbf{x}$ or the mean of *posterior PDF* $p(A|\mathbf{x})$.

# Minimum Mean Square Error Estimator

**How to compute the Bayesian MMSE estimate :**

$$\hat{A} = E(A|\mathbf{x}) = \int A p(A|\mathbf{x}) dA$$

The posteriori PDF can be computed using Bayes' Rule :

$$p(A|\mathbf{x}) = \frac{p(\mathbf{x}|A)p(A)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|A)p(A)}{\int p(\mathbf{x}|A)p(A)dA}$$

- $p(\mathbf{x})$ is the marginal PDF defined as $p(\mathbf{x}) = \int p(\mathbf{x}, A)dA$.
- The integral in the denominator acts as a "normalization" of $p(\mathbf{x}|A)p(A)$ such the integral of $p(A|\mathbf{x})$ be equal to 1.
- $p(\mathbf{x}|A)$ has exactly the same form as $p(\mathbf{x}; A)$ which is used in the classical estimation.
- The MMSE estimator always exists and can be computed by :

$$\hat{A} = \frac{\int A p(\mathbf{x}|A)p(A)dA}{\int p(\mathbf{x}|A)p(A)dA}$$

# Minimum Mean Square Error Estimator

**Example :** For $p(A) = \mathcal{U}[-A_0, A_0]$ we have :

$$\hat{A} = \frac{\frac{1}{2A_0}\int_{-A_0}^{A_0} A \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[\frac{-1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] dA}{\frac{1}{2A_0}\int_{-A_0}^{A_0} \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[\frac{-1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] dA}$$

- Before collecting the data the mean of the prior PDF $p(A)$ is the best estimate while after collecting the data the best estimate is the mean of posterior PDF $p(A|\mathbf{x})$.
- Choice of $p(A)$ is crucial for the quality of the estimation.
- Only Gaussian prior PDF leads to a closed-form estimator.

**Conclusion :** For an accurate estimator choose a prior PDF that can be physically justified. For a closed-form estimator choose a Gaussian prior PDF.

# Properties of the Gaussian PDF

## Theorem (Conditional PDF of Bivariate Gaussian)

*If $x$ and $y$ are distributed according to a bivariate Gaussian PDF :*

$$p(x,y) = \frac{1}{2\pi\sqrt{\det(\mathbf{C})}} \exp\left[\frac{-1}{2}\begin{bmatrix} x-E(x) \\ y-E(y) \end{bmatrix}^T \mathbf{C}^{-1} \begin{bmatrix} x-E(x) \\ y-E(y) \end{bmatrix}\right]$$

*with covariance matrix :* $\mathbf{C} = \begin{bmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{bmatrix}$

*then the conditional PDF $p(y|x)$ is also Gaussian and :*

$$\begin{aligned} E(y|x) &= E(y) + \frac{cov(x,y)}{var(x)}[x - E(x)] \\ var(y|x) &= var(y) - \frac{cov^2(x,y)}{var(x)} \end{aligned}$$

# Properties of the Gaussian PDF

## Example (DC level in WGN with Gaussian prior PDF)

Consider the Bayesian model $x[0] = A + w[0]$ (just one observation) with a prior PDF of $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$ independent of noise.
First we compute the covariance $\text{cov}(A, x[0])$ :

$$\begin{aligned}
\text{cov}(A, x[0]) &= E\{(A - \mu_A)(x[0] - E(x[0])\} \\
&= E\{(A - \mu_A)(A + w[0] - \mu_A - 0)\} \\
&= E\{(A - \mu_A)^2 + (A - \mu_A)w[0]\} = \text{var}(A) + 0 = \sigma_A^2
\end{aligned}$$

The Bayesian MMSE estimate is also Gaussian with :

$$\hat{A} = \mu_{A|x} = \mu_A + \frac{\text{cov}(A, x[0])}{\text{var}(x)}[x[0] - \mu_A] = \mu_A + \frac{\sigma_A^2}{\sigma^2 + \sigma_A^2}(x[0] - \mu_A)$$

$$\text{var}(\hat{A}) = \sigma_{A|x}^2 = \sigma_A^2 - \frac{\text{cov}^2(A, x[0])}{\text{var}(x)} = \sigma_A^2(1 - \frac{\sigma_A^2}{\sigma^2 + \sigma_A^2})$$

# Properties of the Gaussian PDF

## Theorem (Conditional PDF of Multivariate Gaussian)

*If* $\mathbf{x}$ *(with dimension* $k \times 1$*) and* $\mathbf{y}$ *(with dimension* $l \times 1$*) are jointly Gaussian with PDF :*

$$p(x,y) = \frac{1}{(2\pi)^{\frac{k+l}{2}}\sqrt{\det(\mathbf{C})}} \exp\left[\frac{-1}{2}\begin{bmatrix}\mathbf{x}-E(\mathbf{x})\\ \mathbf{y}-E(\mathbf{y})\end{bmatrix}^T \mathbf{C}^{-1}\begin{bmatrix}\mathbf{x}-E(\mathbf{x})\\ \mathbf{y}-E(\mathbf{y})\end{bmatrix}\right]$$

*with covariance matrix :* $\mathbf{C} = \begin{bmatrix}\mathbf{C}_{xx} & \mathbf{C}_{xy}\\ \mathbf{C}_{yx} & \mathbf{C}_{yy}\end{bmatrix}$

*then the conditional PDF* $p(\mathbf{y}|\mathbf{x})$ *is also Gaussian and :*

$$\begin{aligned} E(\mathbf{y}|\mathbf{x}) &= E(\mathbf{y}) + \mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}[\mathbf{x}-E(\mathbf{x})] \\ \mathbf{C}_{y|x} &= \mathbf{C}_{yy} - \mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy} \end{aligned}$$

# Bayesian Linear Model

Let the data be modeled as

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

where $\boldsymbol{\theta}$ is a $p \times 1$ random vector with prior PDF $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$ and $\mathbf{w}$ is a noise vector with PDF $\mathcal{N}(0, \mathbf{C}_w)$.
Since $\boldsymbol{\theta}$ and $\mathbf{w}$ are independent and Gaussian, they are jointly Gaussian then the posterior PDF is also Gaussian. In order to find the MMSE estimator we should compute the covariance matrices :

$$\begin{aligned}
\mathbf{C}_{xx} &= E\{[\mathbf{x} - E(\mathbf{x})][\mathbf{x} - E(\mathbf{x})]^T\} \\
&= E\{(\mathbf{H}\boldsymbol{\theta} + \mathbf{w} - \mathbf{H}\boldsymbol{\mu}_{\boldsymbol{\theta}})(\mathbf{H}\boldsymbol{\theta} + \mathbf{w} - \mathbf{H}\boldsymbol{\mu}_{\boldsymbol{\theta}})^T\} \\
&= E\{(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) + \mathbf{w})(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) + \mathbf{w})^T\} \\
&= \mathbf{H}E\{(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T\}\mathbf{H}^T + E(\mathbf{w}\mathbf{w}^T) = \mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T + \mathbf{C}_w \\
\mathbf{C}_{\theta x} &= E\{(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})[\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) + \mathbf{w}]^T\} \\
&= E\{(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T\}\mathbf{H}^T = \mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T
\end{aligned}$$

# Bayesian Linear Model

## Theorem (Posterior PDF for the Bayesian General Linear Model)

*If the observed data can be modeled as*

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

*where* $\boldsymbol{\theta}$ *is a* $p \times 1$ *random vector with prior PDF* $\mathcal{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$ *and* $\mathbf{w}$ *is a noise vector with PDF* $\mathcal{N}(0, \mathbf{C}_w)$*, then the posterior PDF* $p(\mathbf{x}|\boldsymbol{\theta})$ *is Gaussian with mean :*

$$E(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\mu}_\theta + \mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^{\mathbf{T}}(\mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^{\mathbf{T}} + \mathbf{C}_{\mathbf{w}})^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_{\boldsymbol{\theta}})$$

*and covariance*

$$\mathbf{C}_{\theta|x} = \mathbf{C}_\theta - \mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^{\mathbf{T}}(\mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^{\mathbf{T}} + \mathbf{C}_{\mathbf{w}})^{-1}\mathbf{H}\mathbf{C}_{\boldsymbol{\theta}}$$

**Remark :** In contrast to the classical general linear model, $\mathbf{H}$ need not be full rank.

# Bayesian Linear Model

## Example (DC Level in WGN with Gaussian Prior PDF)

$$x[n] = A + w[n], \quad n = 0, \ldots, N-1, \quad A \sim \mathcal{N}(\mu_A, \sigma_A^2), \quad w[n] \sim \mathcal{N}(0, \sigma^2)$$

We have the general Bayesian linear model $\mathbf{x} = \mathbf{HA} + \mathbf{w}$ with $\mathbf{H} = \mathbf{1}$. Then

$$\begin{aligned} E(A|\mathbf{x}) &= \mu_A + \sigma_A^2 \mathbf{1}^T (\mathbf{1}\sigma_A^2 \mathbf{1}^T + \sigma^2 \mathbf{I})^{-1} (\mathbf{x} - \mathbf{1}\mu_A) \\ \text{var}(A|\mathbf{x}) &= \sigma_A^2 - \sigma_A^2 \mathbf{1}^T (\mathbf{1}\sigma_A^2 \mathbf{1}^T + \sigma^2 \mathbf{I})^{-1} \mathbf{1}\sigma_A^2 \end{aligned}$$

We use the Matrix Inversion Lemma to get :

$$(\mathbf{I} + \mathbf{1}\frac{\sigma_A^2}{\sigma^2}\mathbf{1}^T)^{-1} = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{\frac{\sigma^2}{\sigma_A^2} + \mathbf{1}^T\mathbf{1}} = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{\frac{\sigma^2}{\sigma_A^2} + N}$$

$$E(A|\mathbf{x}) = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}(\bar{x} - \mu_A) \quad \text{and} \quad \text{var}(A|\mathbf{x}) = \frac{\frac{\sigma^2}{N}\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}$$

# Nuisance Parameters

## Definition (Nuisance Parameter)

Suppose that $\theta$ and $\alpha$ are unknown parameters but we are only interested in estimating $\theta$. In this case $\alpha$ is called a *nuisance parameter*.

- In the classical approach we have to estimate both but in the Bayesian approach we can *integrate it out*.
- Note that in the Bayesian approach we can find $p(\theta|\mathbf{x})$ from $p(\theta, \alpha|\mathbf{x})$ as a marginal PDF :

$$p(\theta|\mathbf{x}) = \int p(\theta, \alpha|\mathbf{x}) d\alpha$$

- We can also express it as

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta} \quad \text{where} \quad p(\mathbf{x}|\theta) = \int p(\mathbf{x}|\theta, \alpha)p(\alpha|\theta)d\alpha$$

- Furthermore if $\alpha$ is independent of $\theta$ we have :

$$p(\mathbf{x}|\theta) = \int p(\mathbf{x}|\theta, \alpha)p(\alpha)d\alpha$$

# General Bayesian Estimators

# General Bayesian Estimators

## Risk Function

A general Bayesian estimator is obtained by minimizing the Bayes Risk

$$\hat{\theta} = \arg\min_{\hat{\theta}} \mathbb{R}(\hat{\theta})$$

where $\mathbb{R}(\hat{\theta}) = E\{C(\varepsilon)\}$ is the Bayes Risk, $C(\varepsilon)$ is a cost function and $\varepsilon = \theta - \hat{\theta}$ is the estimation error.

## Three common risk functions

1. Quadratic : $\mathbb{R}(\hat{\theta}) = E\{\varepsilon^2\} = E\{(\theta - \hat{\theta})^2\}$

2. Absolute : $\mathbb{R}(\hat{\theta}) = E\{|\varepsilon|\} = E\{|\theta - \hat{\theta}|\}$

3. Hit-or-Miss : $\mathbb{R}(\hat{\theta}) = E\{C(\varepsilon)\}$ where $C(\varepsilon) = \begin{cases} 0 & |\varepsilon| < \delta \\ 1 & |\varepsilon| \geq \delta \end{cases}$

# General Bayesian Estimators

The Bayes risk is

$$\mathbb{R}(\hat{\theta}) = E\{C(\varepsilon)\} = \int\int C(\theta - \hat{\theta})p(\mathbf{x}, \theta)d\mathbf{x}d\theta = \int g(\hat{\theta})p(\mathbf{x})d\mathbf{x}$$

$$\text{where} \quad g(\hat{\theta}) = \int C(\theta - \hat{\theta})p(\theta|\mathbf{x})d\theta \quad \text{should be minimized.}$$

## Quadratic

For this case $\mathbb{R}(\hat{\theta}) = E\{(\theta - \hat{\theta})^2\} = \text{Bmse}(\hat{\theta})$ which leads to the MMSE estimator with

$$\hat{\theta} = E(\theta|\mathbf{x})$$

so $\hat{\theta}$ is the mean of the posterior PDF $p(\theta|\mathbf{x})$.

# General Bayesian Estimators

## Absolute

In this case we have :

$$g(\hat{\theta}) = \int |\theta - \hat{\theta}| p(\theta|\mathbf{x}) d\theta = \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) p(\theta|\mathbf{x}) d\theta + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) p(\theta|\mathbf{x}) d\theta$$

By setting the derivative of $g(\hat{\theta})$ equal to zero and the use of Leibnitz's rule, we get :

$$\int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{x}) d\theta = \int_{\hat{\theta}}^{\infty} p(\theta|\mathbf{x}) d\theta$$

Then $\hat{\theta}$ is the median (area to the left=area to the right) of $p(\theta|\mathbf{x})$

# General Bayesian Estimators

## Hit-or-Miss

In this case we have

$$\begin{aligned} g(\hat{\theta}) &= \int_{-\infty}^{\hat{\theta}-\delta} 1 \cdot p(\theta|\mathbf{x})d\theta + \int_{\hat{\theta}+\delta}^{\infty} 1 \cdot p(\theta|\mathbf{x})d\theta \\ &= 1 - \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|\mathbf{x})d\theta \end{aligned}$$

For $\delta$ arbitrarily small the optimal estimate is the location of the maximum of $p(\theta|\mathbf{x})$ or the mode of the posterior PDF.

**This estimator is called the maximum a posteriori (MAP) estimator.**

**Remark :** For unimodal and symmetric posterior PDF (e.g. Gaussian PDF), the mean and the mode and the median are the same.

# Minimum Mean Square Error Estimators

**Extension to the vector parameter case :** In general, like the scalar case, we can write :

$$\hat{\theta}_i = E(\theta_i|\mathbf{x}) = \int \theta_i p(\theta_i|\mathbf{x}) d\theta_i \qquad i = 1, 2, \ldots, p$$

Then $p(\theta_i|\mathbf{x})$ can be computed as a marginal conditional PDF. For example :

$$\hat{\theta}_1 = \int \theta_1 \left[ \int \cdots \int p(\boldsymbol{\theta}|\mathbf{x}) d\theta_2 \cdots d\theta_p \right] d\theta_1 = \int \theta_1 p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$$

In vector form we have :

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \int \theta_1 p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \\ \int \theta_2 p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \\ \vdots \\ \int \theta_p p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \end{bmatrix} = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = E(\boldsymbol{\theta}|\mathbf{x})$$

Similarly $\qquad \text{Bmse}(\hat{\theta}_i) = \int [\mathbf{C}_{\theta|x}]_{ii} p(\mathbf{x}) d\mathbf{x}$

# Properties of MMSE Estimators

- *For Bayesian Linear Model, poor prior knowledge leads to MVU estimator.*

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{x}) = \boldsymbol{\mu}_\theta + (\mathbf{C}_\theta^{-1} + \mathbf{H}^\mathbf{T}\mathbf{C}_\mathbf{w}^{-1}\mathbf{H})^{-1}\mathbf{H}^\mathbf{T}\mathbf{C}_\mathbf{w}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\theta)$$

For no prior knowledge $\boldsymbol{\mu}_\theta \to \mathbf{0}$ and $\mathbf{C}_\theta^{-1} \to \mathbf{0}$, and therefore :

$$\hat{\boldsymbol{\theta}} \to [\mathbf{H}^\mathbf{T}\mathbf{C}_\mathbf{w}^{-1}\mathbf{H}]^{-1}\mathbf{H}^\mathbf{T}\mathbf{C}_\mathbf{w}^{-1}\mathbf{x}$$

- *Commutes over affine transformations.*
Suppose that $\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\theta} + \mathbf{b}$ then the MMSE estimator for $\boldsymbol{\alpha}$ is

$$\hat{\boldsymbol{\alpha}} = E(\boldsymbol{\alpha}|\mathbf{x}) = E(\mathbf{A}\boldsymbol{\theta} + \mathbf{b}|\mathbf{x}) = \mathbf{A}E(\boldsymbol{\theta}|\mathbf{x}) + \mathbf{b} = \mathbf{A}\hat{\boldsymbol{\theta}} + \mathbf{b}$$

- *Enjoys the additive property for independent data sets.*
Assume that $\boldsymbol{\theta}, \mathbf{x_1}, \mathbf{x_2}$ are jointly Gaussian with $\mathbf{x_1}$ and $\mathbf{x_2}$ independent :

$$\hat{\boldsymbol{\theta}} = E(\theta) + \mathbf{C}_{\theta x_1}\mathbf{C}_{x_1 x_1}^{-1}[\mathbf{x_1} - E(\mathbf{x_1})] + \mathbf{C}_{\theta x_2}\mathbf{C}_{x_2 x_2}^{-1}[\mathbf{x_2} - E(\mathbf{x_2})]$$

# Maximum A Posteriori Estimators

In the MAP estimation approach we have :

$$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathbf{x})$$

where

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

- An equivalent maximization is :

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{x}|\theta)p(\theta)$$

or

$$\hat{\theta} = \arg\max_{\theta}[\ln p(\mathbf{x}|\theta) + \ln p(\theta)]$$

- If $p(\theta)$ is uniform or is approximately constant around the maximum of $p(\mathbf{x}|\theta)$ (for large data length), we can remove $p(\theta)$ to obtain the *Bayesian Maximum Likelihood* :

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{x}|\theta)$$

# Maximum A Posteriori Estimators

## Example (DC Level in WGN with Uniform Prior PDF)

The MMSE estimator cannot be obtained in explicit form due to the need to evaluate the following integrals :

$$\hat{A} = \frac{\frac{1}{2A_0}\int_{-A_0}^{A_0} A \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[\frac{-1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] dA}{\frac{1}{2A_0}\int_{-A_0}^{A_0} \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[\frac{-1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] dA}$$

The MAP estimator is given as : $\hat{A} = \begin{cases} -A_0 & \bar{x} < -A_0 \\ \bar{x} & -A_0 \leq \bar{x} \leq A_0 \\ A_0 & \bar{x} > A_0 \end{cases}$

**Remark :** The main advantage of MAP estimator is that for non jointly Gaussian PDFs it may lead to explicit solutions or less computational effort.

# Maximum A Posteriori Estimators

## Extension to the vector parameter case

$$\hat{\theta}_1 = \arg\max_{\theta_1} p(\theta_1|\mathbf{x}) = \arg\max_{\theta_1} \int \cdots \int p(\boldsymbol{\theta}|\mathbf{x}) d\theta_2 \ldots d\theta_p$$

This needs integral evaluation of the marginal conditional PDFs.

An alternative is the following vector MAP estimator that maximizes the joint conditional PDF :

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

This corresponds to a circular Hit-or-Miss cost function.

- In general, as $N \to \infty$, the MAP estimator $\to$ the Bayesian MLE.
- If the posterior PDF is Gaussian, the mode is identical to the mean, therefore the MAP estimator is identical to the MMSE estimator.
- The invariance property in ML theory does not hold for the MAP estimator.

# Performance Description

- In the classical estimation the mean and the variance of the estimate (or its PDF) indicates the performance of the estimator.
- In the Bayesian approach the PDF of the estimate is different for each realization of $\theta$. So a good estimator should perform well for every possible value of $\theta$.
- The estimation error $\varepsilon = \theta - \hat{\theta}$ is a function of two random variables ($\theta$ and $\mathbf{x}$).
- The mean and the variance of the estimation error, wrt two random variables, indicates the performance of the estimator.

The mean value of the estimation error is zero. So the estimates are unbiased (in the Bayesian sense).

$$E_{x,\theta}(\theta - \hat{\theta}) = E_{x,\theta}(\theta - E(\theta|\mathbf{x})) = E_x[E_{\theta|x}(\theta) - E_{\theta|x}(\theta|\mathbf{x})] = E_x(0) = 0$$

The variance of the estimate is Bmse :

$$\text{var}(\varepsilon) = E_{x,\theta}(\varepsilon^2) = E_{x,\theta}[(\theta - \hat{\theta})^2] = \text{Bmse}(\hat{\theta})$$

# Performance Description (Vector Parameter Case)

- The vector of the estimation error is $\varepsilon = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ and has a zero mean.
- Its covariance matrix is :

$$\begin{aligned}\mathbf{M}_{\hat{\theta}} &= E_{x,\theta}(\varepsilon\varepsilon^T) = E_{x,\theta}\{[\boldsymbol{\theta} - E(\boldsymbol{\theta}|\mathbf{x})][\boldsymbol{\theta} - E(\boldsymbol{\theta}|\mathbf{x})]^T\} \\ &= E_x\left[E_{\boldsymbol{\theta}|\mathbf{x}}\left\{[\boldsymbol{\theta} - E(\boldsymbol{\theta}|\mathbf{x})][\boldsymbol{\theta} - E(\boldsymbol{\theta}|\mathbf{x})]^T\right\}\right] = E_x(\mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}})\end{aligned}$$

- if $\mathbf{x}$ and $\boldsymbol{\theta}$ are jointly Gaussian, we have :

$$\mathbf{M}_{\hat{\theta}} = \mathbf{C}_{\theta|x} = \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}$$

  since $\mathbf{C}_{\theta|x}$ does not depend on $\mathbf{x}$.
- For a Bayesian linear model we have :

$$\mathbf{M}_{\hat{\theta}} = \mathbf{C}_{\theta|x} = \mathbf{C}_\theta - \mathbf{C}_\theta\mathbf{H}^\mathbf{T}(\mathbf{H}\mathbf{C}_\theta\mathbf{H}^\mathbf{T} + \mathbf{C}_w)^{-1}\mathbf{H}\mathbf{C}_\theta$$

- In this case $\varepsilon$ is a linear transformation of $\mathbf{x}$ and $\boldsymbol{\theta}$ and thus is Gaussian. Therefore :

$$\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{M}_{\hat{\theta}})$$

# Signal Processing Example

**Deconvolution Problem :** Estimate a signal transmitted through a channel with known impulse response :

$$x[n] = \sum_{m=0}^{n_s-1} h[n-m]s[m] + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $s[n]$ is a WSS Gaussian process with known ACF, and $w[n]$ is WGN with variance $\sigma^2$.

$$\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} = \begin{bmatrix} h[0] & 0 & \cdots & 0 \\ h[1] & h[0] & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ h[N-1] & h[N-2] & \cdots & h[N-n_s] \end{bmatrix} \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[n_s-1] \end{bmatrix} + \begin{bmatrix} w[0] \\ w[1] \\ \vdots \\ w[N-1] \end{bmatrix}$$

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \qquad \Rightarrow \qquad \hat{\mathbf{s}} = \mathbf{C_s}\mathbf{H}^T(\mathbf{H}\mathbf{C_s}\mathbf{H}^T + \sigma^2\mathbf{I})^{-1}\mathbf{x}$$

where $\mathbf{C_s}$ is a symmetric Toeplitz matrix with $[\mathbf{C_s}]_{ij} = r_{ss}[i-j]$

# Signal Processing Example

Consider that $\mathbf{H} = \mathbf{I}$ (no filtering), so the Bayesian model becomes

$$\mathbf{x} = \mathbf{s} + \mathbf{w}$$

Classical Approach The MVU estimator is $\hat{\mathbf{s}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} = \mathbf{x}$.
Bayesian Approach The MMSE estimator is $\hat{\mathbf{s}} = \mathbf{C_s}(\mathbf{C_s} + \sigma^2\mathbf{I})^{-1}\mathbf{x}$.

- **Scalar case :** We estimate $s[0]$ based on $x[0]$ :

$$\hat{s}[0] = \frac{r_{ss}[0]}{r_{ss}[0] + \sigma^2}x[0] = \frac{\eta}{\eta + 1}x[0]$$

  where $\eta = r_{ss}[0]/\sigma^2$ is the SNR.
- **Signal is a Realization of an Auto Regressive Process :** Consider a first order AR process : $s[n] = -a\,s[n-1] + u[n]$, where $u[n]$ is WGN with variance $\sigma_u^2$. The ACF of $s$ is :

$$r_{ss}[k] = \frac{\sigma_u^2}{1-a^2}(-a)^{|k|} \quad \Rightarrow \quad [\mathbf{C_s}]_{ij} = r_{ss}[i-j]$$

# Linear Bayesian Estimators

# Introduction

**Problems with general Baysian estimators :**

- difficult to determine in closed form,
- need intensive computations,
- involve multidimensional integration (MMSE) or multidimensional maximization (MAP),
- can be determined only under the jointly Gaussian assumption.

**What can we do if the joint PDF is not Gaussian or unknown ?**

- Keep the MMSE criterion ;
- Restrict the estimator to be linear.

**This leads to the Linear MMSE Estimator :**

- Which is a suboptimal estimator that can be easily implemented.
- It needs only the first and second moments of the joint PDF.
- It is analogous to BLUE in classical estimation.
- In practice, this estimator is termed *Wiener filter*.

# Linear MMSE Estimators (scalar case)

**Goal :** Estimate $\theta$, given the data vector $\mathbf{x}$. Assume that only the first two moments of the joint PDF of $\mathbf{x}$ and $\theta$ are available :

$$\begin{bmatrix} E(\theta) \\ E(\mathbf{x}) \end{bmatrix} \quad , \quad \begin{bmatrix} C_{\theta\theta} & \mathbf{C}_{\theta x} \\ \mathbf{C}_{x\theta} & \mathbf{C}_{xx} \end{bmatrix}$$

**LMMSE Estimator :** Take the class of all affine estimators

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] + a_N = \mathbf{a}^T \mathbf{x} + a_N$$

where $\mathbf{a} = [a_0, a_1, \ldots, a_{N-1}]^T$. Then minimize the Bayesian MSE :

$$\text{Bmse}(\hat{\theta}) = E[(\theta - \hat{\theta})^2]$$

**Computing** $a_N$ : Let's differentiate the Bmse with respect to $a_N$ :

$$\frac{\partial}{\partial a_N} E\left[(\theta - \mathbf{a}^T \mathbf{x} - a_N)^2\right] = -2E\left[(\theta - \mathbf{a}^T \mathbf{x} - a_N)\right]$$

Setting this equal to zero gives : $a_N = E(\theta) - \mathbf{a}^T E(\mathbf{x})$.

# Linear MMSE Estimators (scalar case)

**Minimize the Bayesian MSE :**

$$\begin{aligned}
\text{Bmse}(\hat{\theta}) &= E[(\theta-\hat{\theta})^2] = E[(\theta-\mathbf{a}^T\mathbf{x}-E(\theta)+\mathbf{a}^TE(\mathbf{x}))^2] \\
&= E\left\{[\mathbf{a}^T(\mathbf{x}-E(\mathbf{x}))-(\theta-E(\theta))]^2\right\} \\
&= E[\mathbf{a}^T(\mathbf{x}-E(\mathbf{x}))(\mathbf{x}-E(\mathbf{x}))^T\mathbf{a}]-E[\mathbf{a}^T(\mathbf{x}-E(\mathbf{x}))(\theta-E(\theta))] \\
&\quad -E[(\theta-E(\theta))(\mathbf{x}-E(\mathbf{x}))^T\mathbf{a}]+E[(\theta-E(\theta))^2] \\
&= \mathbf{a}^T\mathbf{C}_{xx}\mathbf{a}-\mathbf{a}^T\mathbf{C}_{x\theta}-\mathbf{C}_{\theta x}\mathbf{a}+C_{\theta\theta}
\end{aligned}$$

This can be minimized by setting the gradient to zero :

$$\frac{\partial \text{Bmse}(\hat{\theta})}{\partial \mathbf{a}} = 2\mathbf{C}_{xx}\mathbf{a}-2\mathbf{C}_{x\theta}=0$$

which results in $\mathbf{a}=\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}$ and leads to (Note that : $\mathbf{C}_{\theta x}=\mathbf{C}_{x\theta}^T$) :

$$\begin{aligned}
\hat{\theta} &= \mathbf{a}^T\mathbf{x}+a_N = \mathbf{C}_{x\theta}^T\mathbf{C}_{xx}^{-1}\mathbf{x}+E(\theta)-\mathbf{C}_{x\theta}^T\mathbf{C}_{xx}^{-1}E(\mathbf{x}) \\
&= E(\theta)+\mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}(\mathbf{x}-E(\mathbf{x}))
\end{aligned}$$

# Linear MMSE Estimators (scalar case)

**The minimum Bayesian MSE is obtained :**

$$\begin{aligned} \text{Bmse}(\hat{\theta}) &= \mathbf{C}_{x\theta}^T \mathbf{C}_{xx}^{-1} \mathbf{C}_{xx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} - \mathbf{C}_{x\theta}^T \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} \\ &\quad -\mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} + C_{\theta\theta} \\ &= C_{\theta\theta} - \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} \end{aligned}$$

**Remarks :**

- The results are identical to those of MMSE estimators for jointly Gaussian PDF.
- The LMMSE estimator relies on the correlation between random variables.
- If a parameter is uncorrelated with the data (but nonlinearly dependent), it cannot be estimated with an LMMSE estimator.

# Linear MMSE Estimators (scalar case)

## Example (DC Level in WGN with Uniform Prior PDF)

The data model is : $x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1$
where $A \sim \mathcal{U}[-A_0, A_0]$ is independent from $w[n]$ (WGN with variance $\sigma^2$).
We have $E(A) = 0$ and $E(\mathbf{x}) = \mathbf{0}$. The covariances are :

$$\begin{aligned} \mathbf{C}_{xx} &= E(\mathbf{x}\mathbf{x}^T) = E[(A\mathbf{1} + \mathbf{w})(A\mathbf{1} + \mathbf{w})^T] = E(A^2)\mathbf{1}\mathbf{1}^T + \sigma^2\mathbf{I} \\ \mathbf{C}_{\theta x} &= E(A\mathbf{x}^T) = E[A(A\mathbf{1} + \mathbf{w})^T] = E(A^2)\mathbf{1}^T \end{aligned}$$

Therefore the LMMSE estimator is :

$$\hat{A} = \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{x} = \sigma_A^2\mathbf{1}^T(\sigma_A^2\mathbf{1}\mathbf{1}^T + \sigma^2\mathbf{I})^{-1}\mathbf{x} = \frac{A_0^2/3}{A_0^2/3 + \sigma^2/N}\bar{x}$$

where

$$\sigma_A^2 = E(A^2) = \int_{-A_0}^{A_0} A^2 \frac{1}{2A_0} dA = \left.\frac{A^3}{6A_0}\right|_{-A_0}^{A_0} = \frac{A_0^2}{3}$$

# Linear MMSE Estimators (scalar case)

## Example (DC Level in WGN with Uniform Prior PDF)

Comparison of different Bayesian estimators :

$$\text{MMSE}: \hat{A} = \frac{\int_{-A_0}^{A_0} A \exp\left[\frac{-1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] dA}{\int_{-A_0}^{A_0} \exp\left[\frac{-1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] dA}$$

$$\text{MAP}: \hat{A} = \begin{cases} -A_0 & \bar{x} < -A_0 \\ \bar{x} & -A_0 \le \bar{x} \le A_0 \\ A_0 & \bar{x} > A_0 \end{cases}$$

$$\text{LMMSE}: \hat{A} = \frac{A_0^2/3}{A_0^2/3 + \sigma^2/N}\bar{x}$$

# Geometrical Interpretation

**Vector space of random variables :** The set of scalar zero mean random variables is a *vector space*.

- The zero length vector of the set is a RV with zero variance.
- For any real scalar $a$, $ax$ is another zero mean RV in the set.
- For two RVs $x$ and $y$, $x + y = y + x$ is a RV in the set.
- For two RVs $x$ and $y$, the inner product is : $< x, y >= E(xy)$
- The length of $x$ is defined as : $\|x\| = \sqrt{< x, x >} = \sqrt{E(x^2)}$
- Two RVs $x$ and $y$ are orthogonal if : $< x, y >= E(xy) = 0.$
- For RVs, $x_1, x_2, y$ and real numbers $a_1$ and $a_2$ we have :

$$< a_1x_1 + a_2x_2, y >= a_1 < x_1, y > +a_2 < x_2, y >$$

$$E[(a_1x_1 + a_2x_2)y] = a_1E(x_1y) + a_2E(x_2y)$$

- The projection of $y$ on $x$ is : $\frac{< y, x >}{\|x\|^2}x = \frac{E(yx)}{\sigma_x^2}x$

# Geometrical Interpretation

The LMMSE estimator can be determined using the vector space viewpoint. We have :

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n]$$

where $a_N$ is zero, because of zero mean assumption.

- $\hat{\theta}$ belongs to the subspace spanned by $x[0], x[1], \ldots, x[N-1]$.
- $\theta$ is not in this subspace.
- A good estimator will minimize the MSE :

$$E[(\theta - \hat{\theta})^2] = \|\epsilon\|^2$$

  where $\epsilon = \theta - \hat{\theta}$ is the error vector.
- Clearly the length of the vector error is minimized when $\epsilon$ is orthogonal to the subspace spanned by $x[0], x[1], \ldots, x[N-1]$ or to each data sample :

$$E\Big[(\theta - \hat{\theta})x[n]\Big] = 0 \quad \text{for} \quad n = 0, 1, \ldots, N-1$$

# Geometrical Interpretation

The LMMSE estimator can be determined by solving the following equations :

$$E\left[\left(\theta - \sum_{m=0}^{N-1} a_m x[m]\right) x[n]\right] = 0 \qquad n = 0, 1, \ldots, N-1$$

or

$$\sum_{m=0}^{N-1} a_m E(x[m]x[n]) = E(\theta x[n]) \qquad n = 0, 1, \ldots, N-1$$

$$\begin{bmatrix} E(x^2[0]) & E(x[0]x[1]) & \cdots & E(x[0]x[N-1]) \\ E(x[1]x[0]) & E(x^2[1]) & \cdots & E(x[1]x[N-1]) \\ \vdots & \vdots & \ddots & \vdots \\ E(x[N-1]x[0]) & E(x[N-1]x[1]) & \cdots & E(x^2[N-1]) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{N-1} \end{bmatrix} = \begin{bmatrix} E(\theta x[0]) \\ E(\theta x[1]) \\ \vdots \\ E(\theta x[N-1]) \end{bmatrix}$$

Therefore

$$\mathbf{C}_{xx}\mathbf{a} = \mathbf{C}_{\theta x}^T \quad \Rightarrow \quad \mathbf{a} = \mathbf{C}_{xx}^{-1}\mathbf{C}_{\theta x}^T$$

The LMMSE estimator is :

$$\hat{\theta} = \mathbf{a}^T\mathbf{x} = \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{x}$$

# The vector LMMSE Estimator

We want to estimate $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^T$ with a linear estimator that minimize the Bayesian MSE for each element.

$$\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in} x[n] + a_{iN} \qquad \text{Minimize} \qquad \text{Bmse}(\hat{\theta}_i) = E\left[(\theta_i - \hat{\theta}_i)^2\right]$$

Therefore :

$$\hat{\theta}_i = E(\theta_i) + \underbrace{\mathbf{C}_{\theta_i x}}_{1\times N} \underbrace{\mathbf{C}_{xx}^{-1}}_{N\times N} \underbrace{\big(\mathbf{x} - E(\mathbf{x})\big)}_{N\times 1} \qquad i = 1, 2, \ldots, p$$

The scalar LMMSE estimators can be combined into a vector estimator :

$$\hat{\boldsymbol{\theta}} = \underbrace{E(\boldsymbol{\theta})}_{p\times 1} + \underbrace{\mathbf{C}_{\theta x}}_{p\times N} \underbrace{\mathbf{C}_{xx}^{-1}}_{N\times N} \underbrace{\big(\mathbf{x} - E(\mathbf{x})\big)}_{N\times 1}$$

and similarly

$$\text{Bmse}(\hat{\theta}_i) = [\mathbf{M}_{\hat{\theta}}]_{ii} \qquad \text{where} \qquad \underbrace{\mathbf{M}_{\hat{\theta}}}_{p\times p} = \underbrace{\mathbf{C}_{\theta\theta}}_{p\times p} - \underbrace{\mathbf{C}_{\theta x}}_{p\times N} \mathbf{C}_{xx}^{-1} \underbrace{\mathbf{C}_{x\theta}}_{N\times p}$$

# The vector LMMSE Estimator

## Theorem (Bayesian Gauss–Markov Theorem)

*If the data are described by the Bayesian linear model form*

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

*where $\boldsymbol{\theta}$ is a $p \times 1$ random vector with mean $E(\boldsymbol{\theta})$ and covariance $\mathbf{C}_{\theta\theta}$, $\mathbf{w}$ is a noise vector with zero mean and covariance $\mathbf{C}_w$ and is uncorrelated with $\boldsymbol{\theta}$ (the joint PDF of $p(\boldsymbol{\theta}, \mathbf{w})$ is arbitrary), then the LMMSE estimator of $\boldsymbol{\theta}$ is :*

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}) + \mathbf{C}_{\theta\theta}\mathbf{H}^T(\mathbf{H}\mathbf{C}_{\theta\theta}\mathbf{H}^T + \mathbf{C}_w)^{-1}\big(\mathbf{x} - \mathbf{H}E(\boldsymbol{\theta})\big)$$

*The performance of the estimator is measured by $\boldsymbol{\epsilon} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ whose mean is zero and whose covariance matrix is*

$$\mathbf{C}_\epsilon = \mathbf{M}_{\hat{\theta}} = \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta\theta}\mathbf{H}^T(\mathbf{H}\mathbf{C}_{\theta\theta}\mathbf{H}^T + \mathbf{C}_w)^{-1}\mathbf{H}\mathbf{C}_{\theta\theta} = \left(\mathbf{C}_{\theta\theta}^{-1} + \mathbf{H}^T\mathbf{C}_w^{-1}\mathbf{H}\right)^{-1}$$

# Sequential LMMSE Estimation

**Objective :** Given $\hat{\theta}[n-1]$ based on $\mathbf{x}[n-1]$, update the new estimate $\hat{\theta}[n]$ based on the new sample $x[n]$.

## Example (DC Level in White Noise)

Assume that both $A$ and $w[n]$ have zero mean :

$$x[n] = A + w[n] \qquad \Rightarrow \qquad \hat{A}[N-1] = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/N}\bar{x}$$

**Estimator Update :** $\hat{A}[N] = \hat{A}[N-1] + K[N](x[N] - \hat{A}[N-1])$

$$\text{where} \qquad K[N] = \frac{\text{Bmse}(\hat{A}[N-1])}{\text{Bmse}(\hat{A}[N-1]) + \sigma^2}$$

**Minimum MSE Update :**

$$\text{Bmse}(\hat{A}[N]) = (1 - K[N])\text{Bmse}(\hat{A}[N-1])$$

# Sequential LMMSE Estimation

**Vector space view :** *If two observations are orthogonal, the LMMSE estimate of $\theta$ is the sum of the projection of $\theta$ on each observation.*

1. Find the LMMSE estimator of $A$ based on $x[0]$, yielding $\hat{A}[0]$ :

$$\hat{A}[0] = \frac{E(Ax[0])}{E(x^2[0])}x[0] = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2}x[0]$$

2. Find the LMMSE estimator of $x[1]$ based on $x[0]$, yielding $\hat{x}[1|0]$

$$\hat{x}[1|0] = \frac{E(x[0]x[1])}{E(x^2[0])}x[0] = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2}x[0]$$

3. Determine the *innovation* of the new data :$\tilde{x}[1] = x[1] - \hat{x}[1|0]$.

   This error vector is orthogonal to $x[0]$

4. Add to $\hat{A}[0]$ the LMMSE estimator of $A$ based on the innovation :

$$\hat{A}[1] = \hat{A}[0] + \underbrace{\frac{E(A\tilde{x}[1])}{E(\tilde{x}^2[1])}\tilde{x}[1]}_{\text{the projection of } A \text{ on } \tilde{x}[1]} = \hat{A}[0] + K[1](x[1] - \hat{x}[1|0])$$

# Sequential LMMSE Estimation

**Basic Idea :** Generate a sequence of orthogonal RVs, namely, the innovations :

$$\Big\{ \underbrace{\tilde{x}[0]}_{x[0]}, \quad \underbrace{\tilde{x}[1]}_{x[1]-\hat{x}[1|0]} \quad , \quad \underbrace{\tilde{x}[2]}_{x[2]-\hat{x}[2|0,1]} \quad , \ldots , \quad \underbrace{\tilde{x}[n]}_{x[n]-\hat{x}[n|0,1,\ldots,n-1]} \Big\}$$

Then, add the individual estimators to yield :

$$\hat{A}[N] = \sum_{n=0}^{N} K[n]\tilde{x}[n] = \hat{A}[N-1]+K[N]\tilde{x}[N] \qquad \text{where} \qquad K[n] = \frac{E(A\tilde{x}[n])}{E(\tilde{x}^2[n])}$$

It can be shown that :

$$\tilde{x}[N] = x[N] - \hat{A}[N-1] \qquad \text{and} \qquad K[N] = \frac{\text{Bmse}(\hat{A}[N-1])}{\sigma^2 + \text{Bmse}(\hat{A}[N-1])}$$

the minimum MSE be updated as :

$$\text{Bmse}(\hat{A}[N]) = (1 - K[N])\text{Bmse}(\hat{A}[N-1])$$

# General Sequential LMMSE Estimation

Consider the general Bayesian linear model $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, where $\mathbf{w}$ is an uncorrelated noise with the diagonal covariance matrix $\mathbf{C}_w$. Let's define :

$$\begin{aligned} \mathbf{C}_w[n] &= \text{diag}(\sigma_0^2, \sigma_1^2, \ldots, \sigma_n^2) \\ \mathbf{H}[n] &= \begin{bmatrix} \mathbf{H}[n-1] \\ \mathbf{h}^T[n] \end{bmatrix} = \begin{bmatrix} n \times p \\ 1 \times p \end{bmatrix} \\ \mathbf{x}[n] &= \big[x[0], x[1], \ldots, x[n]\big] \end{aligned}$$

**Estimator Update :**

$$\hat{\boldsymbol{\theta}}[n] = \hat{\boldsymbol{\theta}}[n-1] + \mathbf{K}[n]\big(x[n] - \mathbf{h}^T[n]\hat{\boldsymbol{\theta}}[n-1]\big)$$

where

$$\mathbf{K}[n] = \frac{\mathbf{M}[n-1]\mathbf{h}[n]}{\sigma_n^2 + \mathbf{h}^T[n]\mathbf{M}[n-1]\mathbf{h}[n]}$$

**Minimum MSE Matrix Update :**

$$\mathbf{M}[n] = (\mathbf{I} - \mathbf{K}[n]\mathbf{h}^T[n])\mathbf{M}[n-1]$$

# General Sequential LMMSE Estimation

**Remarks :**

- For the initialization of the sequential LMMSE estimator, we can use the prior information :

$$\hat{\boldsymbol{\theta}}[-1] = E(\boldsymbol{\theta}) \qquad \mathbf{M}[-1] = \mathbf{C}_{\theta\theta}$$

- For no prior knowledge about $\boldsymbol{\theta}$ we can let $\mathbf{C}_{\theta\theta} \to \infty$. Then we have the same *form* as the sequential LSE, although the approaches are *fundamentally* different.
- No matrix inversion is required.
- The gain factor $\mathbf{K}[n]$ weights confidence in the new data (measured by $\sigma_n^2$) against all previous data (summarized by $\mathbf{M}[n-1]$).

# Wiener Filtering

Consider the signal model : $x[n] = s[n] + w[n] \qquad n = 0, 1, \ldots, N-1$ where the data and noise are zero mean with known covariance matrix.

There are three main problems concerning the Wiener Filters :

Filtering : Estimate $\theta = s[n]$ (scalar) based on the data set $\mathbf{x} = [x[0], x[1], \ldots, x[n]]$. The signal sample is estimated based on the *the present and past data only*.

Smoothing : Estimate $\boldsymbol{\theta} = \mathbf{s} = [s[0], s[1], \ldots, s[N-1]]$ (vector) based on the data set $\mathbf{x} = \big[x[0], x[1], \ldots, x[N-1]\big]$. The signal sample is estimated based on the *the present, past and future data*.

Prediction : Estimate $\theta = x[N-1+\ell]$ for $\ell$ positive integer based on the data set $\mathbf{x} = [x[0], x[1], \ldots, x[N-1]]$.

**Remarks :** All these problems are solved using the LMMSE estimator :

$\hat{\boldsymbol{\theta}} = \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{x}$ with the minimum Bmse : $\mathbf{M}_{\hat{\theta}} = \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\theta}$

# Wiener Filtering (Smoothing)

Estimate $\theta = \mathbf{s} = [s[0], s[1], \ldots, s[N-1]]$ (vector) based on the data set $\mathbf{x} = \left[x[0], x[1], \ldots, x[N-1]\right]$.

$$\mathbf{C}_{xx} = \mathbf{R}_{xx} = \mathbf{R}_{ss} + \mathbf{R}_{ww} \qquad \text{and} \qquad \mathbf{C}_{\theta x} = E(\mathbf{s}\mathbf{x}^T) = E(\mathbf{s}(\mathbf{s}+\mathbf{w})^T) = \mathbf{R}_{ss}$$

Therefore :

$$\hat{\mathbf{s}} = \mathbf{C}_{sx}\mathbf{C}_{xx}^{-1}\mathbf{x} = \mathbf{R}_{ss}(\mathbf{R}_{ss} + \mathbf{R}_{ww})^{-1}\mathbf{x} = \mathbf{W}\mathbf{x}$$

**Filter Interpretation :** The *Wiener Smoothing Matrix* can be interpreted as an FIR filter. Let's define :

$$\mathbf{W} = [\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{N-1}]$$

where $\mathbf{a}_n^T$ is the $(n+1)$th raw of $\mathbf{W}$. Let's define also :
$\mathbf{h}_n = [h^{(n)}[0], h^{(1)}[1], \ldots, h^{(n)}[N-1]]$ which is just the vector $\mathbf{a}_n$ when flipped upside down. Then

$$\hat{s}[n] = \mathbf{a}_n^T\mathbf{x} = \sum_{k=0}^{N-1} h^{(n)}[N-1-k]x[k]$$

This represents a time-varying, non causal FIR filter.

# Wiener Filtering (Filtering)

Estimate $\theta = s[n]$ based on the data set $\mathbf{x} = \big[x[0], x[1], \ldots, x[n]\big]$. We have again $\mathbf{C}_{xx} = \mathbf{R}_{xx} = \mathbf{R}_{ss} + \mathbf{R}_{ww}$ and

$$\mathbf{C}_{\theta x} = E(s[n]\big[x[0], x[1], \ldots, x[n]\big]) = \big[r_{ss}[n], r_{ss}[n-1], \ldots, r_{ss}[0]\big] = (\mathbf{r}'_{ss})^T$$

Therefore :

$$\hat{s}[n] = (\mathbf{r}'_{ss})^T(\mathbf{R}_{ss} + \mathbf{R}_{ww})^{-1}\mathbf{x} = \mathbf{a}^T\mathbf{x}$$

**Filter Interpretation :** If we define $\mathbf{h}_n = [h^{(n)}[0], h^{(1)}[1], \ldots, h^{(n)}[N-1]]$ which is just the vector $\mathbf{a}$ when flipped upside down. Then :

$$\hat{s}[n] = \mathbf{a}^T\mathbf{x} = \sum_{k=0}^{n} h^{(n)}[n-k]x[k]$$

This represents a time-varying, causal FIR filter. The impulse response can be computed as :

$$(\mathbf{R}_{ss} + \mathbf{R}_{ww})\mathbf{a} = \mathbf{r}'_{ss} \qquad \Rightarrow \qquad (\mathbf{R}_{ss} + \mathbf{R}_{ww})\mathbf{h}_n = \mathbf{r}_{ss}$$

# Wiener Filtering (Prediction)

Estimate $\theta = x[N-1+\ell]$ based on the data set $\mathbf{x} = [x[0], x[1], \ldots, x[N-1]]$. We have again $\mathbf{C}_{xx} = \mathbf{R}_{xx}$ and

$$\begin{aligned} \mathbf{C}_{\theta x} &= E(x[N-1+\ell][x[0], x[1], \ldots, x[N-1]]) \\ &= \left[r_{xx}[N-1+\ell], r_{xx}[N-2+\ell], \ldots, r_{xx}[\ell]\right] = (\mathbf{r}'_{xx})^T \end{aligned}$$

Therefore :

$$\hat{x}[N-1+\ell] = (\mathbf{r}'_{xx})^T \mathbf{R}_{xx}^{-1} \mathbf{x} = \mathbf{a}^T \mathbf{x}$$

**Filter Interpretation :** If we let again $h[N-k] = a_k$, we have

$$\hat{x}[N-1+\ell] = \mathbf{a}^T \mathbf{x} = \sum_{k=0}^{N-1} h[N-k] x[k]$$

This represents a time-invariant, causal FIR filter. The impulse response can be computed as :

$$\mathbf{R}_{xx} \mathbf{h} = \mathbf{r}_{xx}$$

For $\ell = 1$, this is equivalent to an Auto-Regressive process.

# Kalman Filters

# Introduction

- Kalman filter is a generalization of the Wiener filter.
- In Wiener filter we estimate $s[n]$ based on the noisy observation vector $\mathbf{x}[n]$. We assume that $s[n]$ is a stationary random process with known mean and covariance matrix.
- In Kalman filter we assume that $s[n]$ is a non stationary random process whose mean and covariance matrix vary according to a known dynamical model.
- Kalman filter is a sequential MMSE estimator of $s[n]$. If the signal and noise are jointly Gaussian, then the Kalman filter is an optimal MMSE estimator, if not, it is the optimal LMMSE estimator.
- Kalman filter has many applications in Control theory for state estimation.
- It can be generalized to vector signals and noise (in contrast to Wiener filter).

# Dynamical Signal Models

Consider a first order dynamical model :

$$s[n] = as[n-1] + u[n] \qquad n \geq 0$$

where $u[n]$ is WGN with variance $\sigma_u^2$, $s[-1] \sim \mathcal{N}(\mu_s, \sigma_s^2)$ and $s[-1]$ is independent of $u[n]$ for all $n \geq 0$. It can be shown that :

$$s[n] = a^{n+1}s[-1] + \sum_{k=0}^{n} a^k u[n-k] \quad \Rightarrow \quad E(s[n]) = a^{n+1}\mu_s$$

and

$$\begin{aligned} C_s[m,n] &= E\Big[\big(s[m] - E(s[m])\big)\big(s[n] - E(s[n])\big)\Big] \\ &= a^{m+n+2}\sigma_s^2 + \sigma_u^2 a^{m-n}\sum_{k=0}^{n} a^{2k} \qquad \text{for } m \geq 0 \end{aligned}$$

and $C_s[m,n] = C_s[n,m]$ for $m < n$.

## Theorem (Vector Gauss-Markov Model)

*The Gauss-Markov model for a $p \times 1$ vector signal $\mathbf{s}[n]$ is :*

$$\mathbf{s}[n] = \mathbf{A}\mathbf{s}[n-1] + \mathbf{B}\mathbf{u}[n] \qquad n \geq 0$$

*$\mathbf{A}$ is $p \times p$ and $\mathbf{B}$ $p \times r$ and all eigenvalues of $\mathbf{A}$ are less than 1 in magnitude. The $r \times 1$ vector $\mathbf{u}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ and the initial condition is a $p \times 1$ random vector with $\mathbf{s}[n] \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{s}, \mathbf{C}_\mathbf{s})$ and is independent of $\mathbf{u}[n]$'s. Then, the signal process is Gaussian with mean $E(\mathbf{s}[n]) = \mathbf{A}^{n+1}\boldsymbol{\mu}_\mathbf{s}$ and covariance :*

$$\mathbf{C}_\mathbf{s}[m,n] = \mathbf{A}^{m+1}\mathbf{C}_\mathbf{s}\left[\mathbf{A}^{n+1}\right]^T + \sum_{k=m-n}^{m} \mathbf{A}^k \mathbf{B}\mathbf{Q}\mathbf{B}^T \left[\mathbf{A}^{n-m+k}\right]^T$$

*for $m \geq n$ and $\mathbf{C}_\mathbf{s}[m,n] = \mathbf{C}_\mathbf{s}[n,m]$ for $m < n$. The covariance matrix $\mathbf{C}[n] = \mathbf{C}_\mathbf{s}[n,n]$ and the mean and covariance propagation equations are :*
$E(\mathbf{s}[n]) = \mathbf{A}E(\mathbf{s}[n-1]) \quad , \quad \mathbf{C}[n] = \mathbf{A}\mathbf{C}[n-1]\mathbf{A}^T + \mathbf{B}\mathbf{Q}\mathbf{B}^T$

# Scalar Kalman Filter

**Data model :**

$$\begin{aligned} \text{State equation} \quad & s[n] = as[n-1] + u[n] \\ \text{Observation equation} \quad & x[n] = s[n] + w[n] \end{aligned}$$

**Assumptions :**

- $u[n]$ is zero mean Gaussian noise with independent samples and $E(u^2[n]) = \sigma_u^2$.
- $w[n]$ is zero mean Gaussian noise with independent samples and $E(w^2[n]) = \sigma_n^2$ (time-varying variance).
- $s[-1] \sim \mathcal{N}(\mu_s, \sigma_s^2)$ (for simplicity we suppose that $\mu_s = 0$).
- $s[-1]$, $u[n]$ and $w[n]$ are independent.

**Objective :** Develop a sequential MMSE estimator to estimate $s[n]$ based on the data $\mathbf{X}[n] = \big[x[0], x[1], \ldots, x[n]\big]$. This estimator is the mean of posterior PDF :

$$\hat{s}[n|n] = E\big(s[n]\big|x[0], x[1], \ldots, x[n]\big)$$

# Scalar Kalman Filter

To develop the equations of a Kalman filter we need the following properties :

Property 1 : For two uncorrelated jointly Gaussian data vector, the MMSE estimator $\theta$ (if it is zero mean) is given by :

$$\hat{\theta} = E(\theta|\mathbf{x}_1, \mathbf{x}_2) = E(\theta|\mathbf{x}_1) + E(\theta|\mathbf{x}_2)$$

Property 2 : If $\theta = \theta_1 + \theta_2$, then the MMSE estimator of $\theta$ is :

$$\hat{\theta} = E(\theta|\mathbf{x}) = E(\theta_1 + \theta_2|\mathbf{x}) = E(\theta_1|\mathbf{x}) + E(\theta_2|\mathbf{x})$$

**Basic Idea :** Generate the innovation $\tilde{x}[n] = x[n] - \hat{x}[n|n-1]$ which is uncorrelated with previous samples $\mathbf{X}[n-1]$. Then use $\tilde{x}[n]$ instead of $x[n]$ for estimation ($\mathbf{X}[n]$ is equivalent to $\left[\mathbf{X}[n-1], \tilde{x}[n]\right]$).

# Scalar Kalman Filter

- From Property 1, we have :

$$\begin{aligned}\hat{s}[n|n] &= E(s[n]\big|\mathbf{X}[n-1], \tilde{x}[n]) = E(s[n]\big|\mathbf{X}[n-1]) + E(s[n]\big|\tilde{x}[n]) \\ &= \hat{s}[n|n-1] + E(s[n]\big|\tilde{x}[n])\end{aligned}$$

- From Property 2, we have :

$$\begin{aligned}\hat{s}[n|n-1] &= E(as[n-1] + u[n]\big|\mathbf{X}[n-1]) \\ &= aE(s[n-1]\big|\mathbf{X}[n-1]) + E(u[n]\big|\mathbf{X}[n-1]) \\ &= a\hat{s}[n-1|n-1]\end{aligned}$$

- The MMSE estimator of $s[n]$ based on $\tilde{x}[n]$ is :

$$E(s[n]\big|\tilde{x}[n]) = \frac{E(s[n]\tilde{x}(n))}{E(\tilde{x}^2[n])}\tilde{x}[n] = K[n](x[n] - \hat{x}[n|n-1])$$

where $\hat{x}[n|n-1] = \hat{s}[n|n-1] + \hat{w}[n|n-1] = \hat{s}[n|n-1]$.

- Finally we have :

$$\hat{s}[n|n] = a\hat{s}[n-1|n-1] + K[n](x[n] - \hat{s}[n|n-1])$$

# Scalar State – Scalar Observation Kalman Filter

**State Model :** $s[n] = as[n-1] + u[n]$

**Observation Model :** $x[n] = s[n] + w[n]$

Prediction :

$$\hat{s}[n|n-1] = a\hat{s}[n-1|n-1]$$

Minimum 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 :

$$\hat{s}[n|n] = \hat{s}[n|n-1] + K[n](x[n] - \hat{s}[n|n-1])$$

Minimum MSE :

$$M[n|n] = (1 - K[n])M[n|n-1]$$

# Properties of Kalman Filter

- The Kalman filter is an extension of the sequential MMSE estimator but applied to time-varying parameters that are represented by a dynamic model.
- No matrix inversion is required.
- The Kalman filter is a time-varying linear filter.
- The Kalman filter computes (and uses) its own performance measure, the Bayesian MSE $M[n|n]$.
- The prediction stage increases the error, while the correction stage decreases it.
- The Kalman filter generates an uncorrelated sequence from the observations so can be viewed as a whitening filter.
- The Kalman filter is optimal in the Gaussian case and the optimal LMMSE estimator if the Gaussian assumption is not valid.
- In KF $M[n|n], M[n|n-1]$ and $K[n]$ do not depend on the measurement and can be computed off line. At steady state ($n \to \infty$) the Kalman filter becomes an LTI filter ($M[n|n], M[n|n-1]$ and $K[n]$ become constant).

# Vector State – Scalar Observation Kalman Filter

**State Model :** $\mathbf{s}[n] = \mathbf{A}\mathbf{s}[n-1] + \mathbf{B}\mathbf{u}[n]$ with $\mathbf{u}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$

**Observation Model :** $x[n] = \mathbf{h}^T[n]\mathbf{s}[n] + w[n]$ with $\mathbf{s}[-1] \sim \mathcal{N}(\boldsymbol{\mu}_s, \mathbf{C}_s)$

Prediction :

$$\hat{\mathbf{s}}[n|n-1] = \mathbf{A}\hat{\mathbf{s}}[n-1|n-1]$$

Minimum Prediction MSE ($p \times p$) :

$$\mathbf{M}[n|n-1] = \mathbf{A}\mathbf{M}[n-1|n-1]\mathbf{A}^T + \mathbf{B}\mathbf{Q}\mathbf{B}^T$$

Kalman Gain ($p \times 1$) :

$$\mathbf{K}[n] = \frac{\mathbf{M}[n|n-1]\mathbf{h}[n]}{\sigma_n^2 + \mathbf{h}^T[n]\mathbf{M}[n|n-1]\mathbf{h}[n]}$$

Correction :

$$\hat{\mathbf{s}}[n|n] = \hat{\mathbf{s}}[n|n-1] + \mathbf{K}[n](x[n] - \mathbf{h}^T[n]\hat{s}[n|n-1])$$

Minimum MSE Matrix ($p \times p$) :

$$\mathbf{M}[n|n] = (I - \mathbf{K}[n]\mathbf{h}^T[n])\mathbf{M}[n|n-1]$$

# Vector State – Vector Observation Kalman Filter

**State Model :** $\mathbf{s}[n] = \mathbf{A}\mathbf{s}[n-1] + \mathbf{B}\mathbf{u}[n]$ with $\mathbf{u}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$

**Observation Model :** $\mathbf{x}[n] = \mathbf{H}[n]\mathbf{s}[n] + \mathbf{w}[n]$ with $\mathbf{w}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{C}[n])$

Prediction :

$$\hat{\mathbf{s}}[n|n-1] = \mathbf{A}\hat{\mathbf{s}}[n-1|n-1]$$

Minimum Prediction MSE ($p \times p$) :

$$\mathbf{M}[n|n-1] = \mathbf{A}\mathbf{M}[n-1|n-1]\mathbf{A}^T + \mathbf{B}\mathbf{Q}\mathbf{B}^T$$

Kalman Gain Matrix ($p \times M$) : (Need matrix inversion !)

$$\mathbf{K}[n] = \mathbf{M}[n|n-1]\mathbf{H}^T[n]\left(\mathbf{C}[n] + \mathbf{H}[n]\mathbf{M}[n|n-1]\mathbf{H}^T[n]\right)^{-1}$$

Correction :

$$\hat{\mathbf{s}}[n|n] = \hat{\mathbf{s}}[n|n-1] + \mathbf{K}[n](x[n] - \mathbf{H}[n]\hat{s}[n|n-1])$$

Minimum MSE Matrix ($p \times p$) :

$$\mathbf{M}[n|n] = (I - \mathbf{K}[n]\mathbf{H}[n])\mathbf{M}[n|n-1]$$

# Extended Kalman Filter

**The extended Kalman filter** is a sub-optimal approach when the state and the observation equations are nonlinear.

**Nonlinear data model :** Consider the following nonlinear data model

$$\begin{aligned} \mathbf{s}[n] &= \mathbf{f}(\mathbf{s}[n-1]) + \mathbf{B}\mathbf{u}[n] \\ \mathbf{x}[n] &= \mathbf{g}(\mathbf{s}[n]) + \mathbf{w}[n] \end{aligned}$$

where $\mathbf{f}(\cdot)$ and $\mathbf{g}(\cdot)$ are nonlinear function mapping (vector to vector).

**Model linearization :** We linearize the model around the estimated value of $\mathbf{s}$ using a first order Taylor series.

$$\begin{aligned} \mathbf{f}(\mathbf{s}[n-1]) &\approx \mathbf{f}(\hat{\mathbf{s}}[n-1|n-1]) + \frac{\partial \mathbf{f}}{\partial \mathbf{s}[n-1]}\Big[\mathbf{s}[n-1] - \hat{\mathbf{s}}[n-1|n-1]\Big] \\ \mathbf{g}(\mathbf{s}[n]) &\approx \mathbf{g}(\hat{\mathbf{s}}[n|n-1]) + \frac{\partial \mathbf{g}}{\partial \mathbf{s}[n]}\Big[\mathbf{s}[n] - \hat{\mathbf{s}}[n|n-1]\Big] \end{aligned}$$

We denote the Jacobians by :

$$\mathbf{A}[n-1] = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{s}[n-1]}\right|_{\hat{s}[n-1|n-1]} \qquad \mathbf{H}[n] = \left.\frac{\partial \mathbf{g}}{\partial \mathbf{s}[n]}\right|_{\hat{s}[n|n-1]}$$

# Extended Kalman Filter

Now we use the linearized models :

$$\begin{aligned}\mathbf{s}[n] &= \mathbf{A}[n-1]\mathbf{s}[n-1] + \mathbf{B}\mathbf{u}[n] \\ &\quad + \mathbf{f}(\hat{\mathbf{s}}[n-1|n-1]) - \mathbf{A}[n-1]\hat{\mathbf{s}}[n-1|n-1] \\ \mathbf{x}[n] &= \mathbf{H}[n]\mathbf{s}[n] + \mathbf{w}[n] \\ &\quad + \mathbf{g}(\hat{\mathbf{s}}[n|n-1]) - \mathbf{H}[n]\hat{\mathbf{s}}[n|n-1]\end{aligned}$$

- There are two differences with the standard Kalman filter :
  1. $\mathbf{A}$ is now, time varying.
  2. Both equations have known terms added to them. This will not change the derivation of the Kalman filter.
- The extended Kalman filter has exactly the same equations where $\mathbf{A}$ is replaced with $\mathbf{A}[n-1]$.
- $\mathbf{A}[n-1]$ and $\mathbf{H}[n]$ should be computed at each sampling time.
- MSE matrices and Kalman gain can no longer be computed off line.