Estimation Theory for Engineers

Roberto Togneri

30th August 2005

1 Applications

Modern estimation theory can be found at the heart of many electronic signal processing systems designed to extract information. These systems include:

Radar where the delay of the received pulse echo has to be estimated in the presence of noise

Sonar where the delay of the received signal from each sensor has to estimated in the presence of noise

Speech where the parameters of the speech model have to be estimated in the presence of speech/speaker variability and environmental noise

Image where the position and orientation of an object from a camera image has to be estimated in the presence of lighting and background noise

Biomedicine where the heart rate of a fetus has to be estimated in the presence of sensor and environmental noise

Communications where the carrier frequency of a signal has to be estimated for demodulation to the baseband in the presence of degradation noise

Control where the position of a powerboat for corrective navigation correction has to be estimated in the presence of sensor and environmental noise

Siesmology where the underground distance of an oil deposit has to be estimated from noisy sound reflections due to different densities of oil and rock layers

The majority of applications require estimation of an unknown parameter θ from a collection of observation data $x[n]$ which also includes "artificacts" due to to sensor inaccuracies, additive noise, signal distortion (convolutional noise), model imperfections, unaccounted source variability and multiple interfering signals.

2 Introduction

Define:

- $x[n] \equiv$ observation data at sample time n,
- $\mathbf{x} = (x[0]\, x[1] \ldots x[N-1])^T \equiv$ vector of N observation samples (N-point data set), and
- $p(\mathbf{x};\theta) \equiv$ mathemtical model (i.e. PDF) of the N-point data set parametrized by θ.

The problem is to find a function of the N-point data set which provides an *estimate* of θ, that is:

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

where $\hat{\theta}$ is an *estimate* of θ, and $g(\mathbf{x})$ is known as the estimator function.

Once a candidate $g(\mathbf{x})$ is found, then we usually ask:

1. How *close* will $\hat{\theta}$ be to θ (i.e. *how good* or *optimal* is our estimator)?
2. Are there better (i.e. *closer*) estimators?

A natural optimal criterion is minimisation of the *mean square error*:

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

But this does not yield realisable estimator functions which can be written as functions of the data only:

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

However although $\Big[E(\hat{\theta}) - \theta\Big]^2$ is a function of θ the variance of the estimator, $\text{var}(\hat{\theta})$, is only a function of data. Thus an alternative approach is to assume $E(\hat{\theta}) - \theta = 0$ and minimise $\text{var}(\hat{\theta})$. This produces the **Minimum Variance Unbiased** (**MVU**) estimator.

2.1 Minimum Variance Unbiased (MVU) Estimator

1. Estimator has to be unbiased , that is:

   $$E(\hat{\theta}) = \theta \;\; \text{for a} < \theta < \text{b}$$

   where [a,b] is the range of interest

2. Estimator has to have minimum variance:

   $$\hat{\theta}_{MVU} = \arg\min_{\hat{\theta}}\{\text{var}(\hat{\theta})\} = \arg\min_{\hat{\theta}}\{E(\hat{\theta} - E(\hat{\theta}))^2\}$$

2.2 EXAMPLE

Consider a fixed signal, A, embedded in a WGN (White Gaussian Noise) signal, $w[n]$:

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

where $\theta = A$ is the parameter to be estimated from the observed data, $x[n]$. Consider the *sample-mean* estimator function:

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

Is the sample-mean an MVU estimator for A?

Unbiased?

$E(\hat{\theta}) = E(\frac{1}{N} \sum x[n]) = \frac{1}{N} \sum E(x[n]) = \frac{1}{N} \sum A = \frac{1}{N} NA = A$

Minimum Variance?

$\text{var}(\hat{\theta}) = \text{var}(\frac{1}{N} \sum x[n]) = \frac{1}{N^2} \sum \text{var}(x[n]) = \frac{1}{N^2} \sum \sigma = \frac{N\sigma}{N} = \frac{\sigma^2}{N}$

But is the sample-mean $\hat{\theta}$ the MVU estimator? It is unbiased but is it minimum variance? That is, is $\text{var}(\tilde{\theta}) \geq \frac{\sigma^2}{N}$ for all other unbiased estimator functions $\tilde{\theta}$?

3 Cramer-Rao Lower Bound (CLRB)

The variance of any unbiased estimater $\hat{\theta}$ must be lower bounded by the CLRB, with the variance of the MVU estimator attaining the CLRB. That is:

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

and

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

Furthermore if, for some functions g and I:

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

then we can find the MVU estimator as: $\hat{\theta}_{MVU} = g(\mathbf{x})$ and the minimum variance is $1/I(\theta)$. For a p-dimensional vector parameter,$\boldsymbol{\theta}$, the equivalent condition is:

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} - \mathbf{I}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0}$$

i.e. $\mathbf{C}_{\hat{\boldsymbol{\theta}}} - \mathbf{I}^{-1}(\boldsymbol{\theta})$ is postitive semidefinite where $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = E[(\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}})^T(\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}}))$ is the covariance matrix. The **Fisher** matrix, $\mathbf{I}(\boldsymbol{\theta})$, is given as:

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

Furthermore if, for some p-dimensional function g and $p \times p$ matrix $\mathbf{I}$:

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

then we can find the MVU estimator as: $\boldsymbol{\theta}_{MVU} = \mathbf{g}(\mathbf{x})$ and the minimum covariance is $\mathbf{I}^{-1}(\boldsymbol{\theta})$.

3.1 EXAMPLE

Consider the case of a signal embedded in noise:

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

where $w[n]$ is a WGN with variance σ^2, and thus:

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

where $p(\mathbf{x};\theta)$ is considered a function of the parameter $\theta = A$ (for known $\mathbf{x}$) and is thus termed the <u>likelihood</u> function. Taking the first and then second derivatives:

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta} &= \frac{N}{\sigma^2}(\frac{1}{N}\sum x[n] - \theta) = \frac{N}{\sigma^2}(\hat{\theta} - \theta) \\ \frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2} &= -\frac{N}{\sigma^2} \end{aligned}$$

For a MVU estimator the lower bound has to apply, that is:

$$\text{var}(\hat{\theta}_{MVU}) = \frac{1}{-E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right]} = \frac{\sigma^2}{N}$$

but we know from the previous example that $\text{var}(\hat{\theta}) = \frac{\sigma^2}{N}$ and thus the sample-mean is a MVU estimator. Alternatively we can show this by considering the first derivative:

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

where $I(\theta) = \frac{N}{\sigma^2}$ and $g(\mathbf{x}) = \frac{1}{N}\sum x[n]$. Thus the MVU estimator is indeed $\hat{\theta}_{MVU} = \frac{1}{N}\sum x[n]$ with minimum variance $\frac{1}{I(\theta)} = \frac{\sigma^2}{N}$.

4 Linear Models

If N-point samples of data are observed and 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
$\boldsymbol{\theta}$	=	$p \times 1$	vector of parameters to be estimated
$\mathbf{w}$	=	$N \times 1$	noise vector with PDF $\mathbf{N}(0, \sigma^2\mathbf{I})$

then using the CRLB theorem $\boldsymbol{\theta} = \mathbf{g}(\mathbf{x})$ will be an MVU estimator if:

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

with $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{I}^{-1}(\boldsymbol{\theta})$. So we need to factor:

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

into the form $\mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta})$. When we do this the MVU estimator for $\boldsymbol{\theta}$ is:

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

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

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

4.1 EXAMPLES

4.1.1 Curve Fitting

Consider fitting the data, $x(t)$, by a p^{th} order polynomial function of t:

$$x(t) = \theta_0 + \theta_1 t + \theta_2 t^2 + \cdots + \theta_p t^p + w(t)$$

Say we have N samples of data, then:

$$\begin{aligned}
\mathbf{x} &= [x(t_0),\, x(t_1),\, x(t_2),\, \ldots,\, x(t_{N-1})]^T \\
\mathbf{w} &= [w(t_0),\, w(t_1),\, w(t_2),\, \ldots,\, w(t_{N-1})]^T \\
\boldsymbol{\theta} &= [\theta_0,\, \theta_1,\, \theta_2,\, \ldots.\theta_p]^T
\end{aligned}$$

so $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, where $\mathbf{H}$ is the $N \times p$ matrix:

$$\mathbf{H} = \begin{bmatrix} 1 & t_0 & t_0^2 & \cdots & t_0^p \\ 1 & t_1 & t_1^2 & \cdots & t_1^p \\ \vdots & \vdots & \vdots & \ddots & \cdots \\ 1 & t_{N-1} & t_{N-1}^2 & \cdots & t_{N-1}^p \end{bmatrix}$$

Hence the MVU estimate of the polynomial coefficients based on the N samples of data is:

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

4.1.2 Fourier Analysis

Consider fitting or representing the N samples of data, $x[n]$, by a linear combination of *sine* and *cosine* functions at different harmonics of the fundamental with period N samples. This implies that $x[n]$ is a periodic time series with period N and this type of analysis is known as *Fourier analysis*. We consider our model as:

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

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

$$\begin{aligned} \mathbf{x} &= [x[0],\, x[1],\, x[2],\, \ldots,\, x[N-1]]^T \\ \mathbf{w} &= [w[0],\, w[1],\, w[2],\, \ldots,\, w[N-1]]^T \\ \boldsymbol{\theta} &= [a_1,\, a_2,\, \ldots,\, a_M,\, b_1,\, b_2,\, \ldots,\, b_M]^T \end{aligned}$$

where $\mathbf{H}$ is the $N \times 2M$ matrix:

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

where:

$$\mathbf{h}_k^a = \begin{bmatrix} 1 \\ \cos\left(\frac{2\pi k}{N}\right) \\ \cos\left(\frac{2\pi k2}{N}\right) \\ \vdots \\ \cos\left(\frac{2\pi k(N-1)}{N}\right) \end{bmatrix}, \qquad \mathbf{h}_k^b = \begin{bmatrix} 0 \\ \sin\left(\frac{2\pi k}{N}\right) \\ \sin\left(\frac{2\pi k2}{N}\right) \\ \vdots \\ \sin\left(\frac{2\pi k(N-1)}{N}\right) \end{bmatrix}$$

Hence the MVU estimate of the Fourier co-efficients based on the N samples of data is:

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

After carrying out the simplification the solution can be shown to be:

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \frac{2}{N}(\mathbf{h}_1^a)^T\mathbf{x} \\ \vdots \\ \frac{2}{N}(\mathbf{h}_M^a)^T\mathbf{x} \\ \frac{2}{N}(\mathbf{h}_1^b)^T\mathbf{x} \\ \vdots \\ \frac{2}{N}(\mathbf{h}_M^b)^T\mathbf{x} \end{bmatrix}$$

which is none other than the standard solution found in signal processing textbooks, usually expressed directly as:

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

and $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \frac{2\sigma^2}{N}\mathbf{I}$.

4.1.3 System Identification

We assume a tapped delay line (TDL) or finite impulse response (FIR) model with p stages or "taps" of an unknown system with output, $x[n]$. To identify the system a known input, $u[n]$, is used to "probe" the system which produces output:

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

We assume N input samples are used to yield N output samples and our identification problem is the same as estimation of the linear model parameters for $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, where:

$$\begin{aligned} \mathbf{x} &= [x[0],\, x[1],\, x[2],\, \ldots,\, x[N-1]]^T \\ \mathbf{w} &= [w[0],\, w[1],\, w[2],\, \ldots,\, w[N-1]]^T \\ \boldsymbol{\theta} &= [h[0],\, h[1],\, h[2],\, \ldots,\, h[p-1]]^T \end{aligned}$$

and $\mathbf{H}$ is the $N \times p$ matrix:

$$\mathbf{H} = \begin{bmatrix} u[0] & 0 & \cdots & 0 \\ u[1] & u[0] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u[N-1] & u[N-2] & \cdots & u[N-p] \end{bmatrix}$$

The MVU estimate of the system model co-efficients is given by:

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

where $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}$. Since $\mathbf{H}$ is a function of $u[n]$ we would like to choose $u[n]$ to achieve minimum variance. It can be shown that the signal we need is a pseudorandom noise (PRN) sequence which has the property that the autocorrelation function is zero for $k \neq 0$, that is:

$$r_{uu}[k] = \frac{1}{N}\sum_{n=0}^{N-1-k} u[n]u[n+k] = 0 \qquad k \neq 0$$

and hence $\mathbf{H}^T\mathbf{H} = Nr_{uu}[0]\mathbf{I}$. Define the crosscorrelation function:

$$r_{ux}[k] = \frac{1}{N}\sum_{n=0}^{N-1-k} u[n]x[n+k]$$

then the system model co-efficients are given by:

$$\hat{h}[i] = \frac{r_{ux}[i]}{r_{uu}[0]}$$

4.2 General Linear Models

In a general linear model two important extensions are:

1. The noise vector, $\mathbf{w}$, is no longer white and has PDF $\mathbf{N}(0, \mathbf{C})$ (i.e. general Gaussian noise)
2. The observed data vector, $\mathbf{x}$, also includes the contribution of known signal components, $\mathbf{s}$.

Thus the general linear model for the observed data is expressed as:

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

where:

$\mathbf{s}$	$N \times 1$	vector of known signal samples
$\mathbf{w}$	$N \times 1$	noise vector with PDF $\mathbf{N}(0, \mathbf{C})$

Our solution for the simple linear model where the noise is assumed white can be used after applying a suitable whitening transformation. If we factor the noise covariance matrix as:

$$\mathbf{C}^{-1} = \mathbf{D}^T\mathbf{D}$$

then the matrix $\mathbf{D}$ is the required transformation since:

$$E[\mathbf{w}\mathbf{w}^T] = \mathbf{C} \Rightarrow E[(\mathbf{D}\mathbf{w})(\mathbf{D}\mathbf{w})^T] = \mathbf{D}\mathbf{C}\mathbf{D}^T = (\mathbf{D}\mathbf{D}^{-1})(\mathbf{D}^{T^{-1}}\mathbf{D}^T) = \mathbf{I}$$

that is $\mathbf{w}' = \mathbf{D}\mathbf{w}$ has PDF $\mathbf{N}(0, \mathbf{I})$. Thus by transforming the general linear model:

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

to:

$$\begin{aligned} \mathbf{x}' = \mathbf{D}\mathbf{x} &= \mathbf{D}\mathbf{H}\boldsymbol{\theta} + \mathbf{D}\mathbf{s} + \mathbf{D}\mathbf{w} \\ \mathbf{x}' &= \mathbf{H}'\boldsymbol{\theta} + \mathbf{s}' + \mathbf{w}' \end{aligned}$$

or:

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

we can then write the MVU estimator of $\boldsymbol{\theta}$ given the observed data $\mathbf{x}''$ as:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\mathbf{H}'^T\mathbf{H}')^{-1}\mathbf{H}'^T\mathbf{x}'' \\ &= (\mathbf{H}^T\mathbf{D}^T\mathbf{D}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}^T\mathbf{D}(\mathbf{x} - \mathbf{s}) \end{aligned}$$

That is:

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

and the covariance matrix is:

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

5 General MVU Estimation

5.1 Sufficient Statistic

For the cases where the CRLB cannot be established a more general approach to finding the MVU estimator is required. We need to first find a *sufficient statistic* for the unknown parameter θ :

Neyman-Fisher Factorization: If we can factor the PDF $p(\mathbf{x};\theta)$ as:

$$p(\mathbf{x};\theta) = g(T(\mathbf{x}),\theta)h(\mathbf{x})$$

where $g(T(\mathbf{x}),\theta)$ is a function of $T(\mathbf{x})$ and θ only and $h(\mathbf{x})$ is a function of $\mathbf{x}$ only, then $T(\mathbf{x})$ is a sufficient statistic for θ. Conceptually one expects that the PDF after the sufficient statistic has been observed, $p(\mathbf{x}|T(\mathbf{x}) = T_0;\theta)$, should not depend on θ since $T(\mathbf{x})$ is sufficient for the estimation of θ and no more knowledge can be gained about θ once we know $T(\mathbf{x})$.

5.1.1 EXAMPLE

Consider a signal embedded in a WGN signal:

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

Then:

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

where $\theta = A$ is the unknown parameter we want to estimate. We factor as follows:

$$\begin{aligned} p(\mathbf{x};\theta) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}(N\theta^2 - 2\theta\sum_{n=0}^{N-1}x[n]\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right] \\ &= g(T(\mathbf{x}),\theta)\cdot h(\mathbf{x}) \end{aligned}$$

where we define $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$ which is a sufficient statistic for θ.

5.2 MVU Estimator

There are two different ways one may derive the MVU estimator based on the sufficient statistic, $T(\mathbf{x})$:

1. Let $\breve{\theta}$ be *any* unbiased estimator of θ. Then $\hat{\theta} = E(\breve{\theta}|T(\mathbf{x})) = \int \breve{\theta} p(\breve{\theta}|T(\mathbf{x}))d\breve{\theta}$ is the MVU estimator.
2. Find some function g such that $\hat{\theta} = g(T(\mathbf{x}))$ is an unbiased estimator of θ, that is $E[g(T(\mathbf{x}))] = \theta$, then $\hat{\theta}$ is the MVU estimator.

The **Rao-Blackwell-Lehmann-Scheffe (RBLS)** theorem tells us that $\hat{\theta} = E(\breve{\theta}|T(\mathbf{x}))$ is:

1. A valid estimator for θ
2. Unbiased
3. Of lesser or equal variance that that of $\breve{\theta}$, for all θ.
4. The MVU estimator if the sufficient statistic, $T(\mathbf{x})$, is complete.

The sufficient statistic, $T(\mathbf{x})$, is complete if there is only one function $g(T(\mathbf{x}))$ that is unbiased. That is, if $h(T(\mathbf{x}))$ is another unbiased estimator (i.e. $E[h(T(\mathbf{x}))] = \theta$) then we must have that $g = h$ if $T(\mathbf{x})$ is complete.

5.2.1 EXAMPLE

Consider the previous example of a signal embedded in a WGN signal:

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

where we derived the sufficient statistic, $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$. Using method 2 we need to find a function g such that $E[g(T(\mathbf{x}))] = \theta = A$. Now:

$$E[T(\mathbf{x})] = E\left[\sum_{n=0}^{N-1} x[n]\right] = \sum_{n=0}^{N-1} E[x[n]] = N\theta$$

It is obvious that:

$$E\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \theta$$

and thus $\hat{\theta} = g(T(\mathbf{x})) = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$, which is the *sample mean* we have already seen before, is the MVU estimator for θ.

6 Best Linear Unbiased Estimators (BLUEs)

It may occur that the MVU estimator or a sufficient statistic cannot be found or, indeed, the PDF of the data is itself unknown (only the second-order statistics are known). In such cases one solution is to assume a functional model of the estimator, as being linear in the data, and find the linear estimator which is both *unbiased* and has *minimum variance,* i.e. the BLUE.

For the general vector case we want our estimator to be a linear function of the data, that is:

$$\hat{\boldsymbol{\theta}} = \mathbf{A}\mathbf{x}$$

Our first requirement is that the estimator be unbiased, that is:

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

which can only be satisfied if:

$$E(\mathbf{x}) = \mathbf{H}\boldsymbol{\theta}$$

i.e. $\mathbf{AH} = \mathbf{I}$. The BLUE is derived by finding the $\mathbf{A}$ which minimises the variance, $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{ACA}^T$, where $\mathbf{C} = E[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T]$ is the covariance of the data $\mathbf{x}$, subject to the constraint $\mathbf{AH} = \mathbf{I}$. Carrying out this minimisation yields the following for the BLUE:

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

where $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}$. The form of the BLUE is identical to the MVU estimator for the general linear model. The crucial difference is that the BLUE

does not make any assumptions on the PDF of the data (or noise) whereas the MVU estimator was derived assuming Gaussian noise. Of course, *if the data is truly Gaussian then the BLUE is also the MVU estimator.* The BLUE for the general linear model can be stated as follows:

Gauss-Markov Theorem Consider a general linear model of the form:

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

where $\mathbf{H}$ is known, and $\mathbf{w}$ is noise with covariance $\mathbf{C}$ (the PDF of $\mathbf{w}$ is otherwise arbitrary), then the BLUE of $\boldsymbol{\theta}$ is:

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

where $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}$ is the minimum covariance.

6.1 EXAMPLE

Consider a signal embedded in noise:

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

Where $w[n]$ is of *unspecified PDF* with $\text{var}(w[n]) = \sigma_n^2$ and the unknown parameter $\theta = A$ is to be estimated. We assume a BLUE estimate and we derive $\mathbf{H}$ by noting:

$$E[\mathbf{x}] = \mathbf{1}\theta$$

where $\mathbf{x} = [x[0], x[1], x[2], \ldots, x[N-1]]^T$, $\mathbf{1} = [1, 1, 1, \ldots, 1]^T$and we have $\mathbf{H} \equiv \mathbf{1}$. Also:

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

and hence the BLUE is:

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

and the minimum covariance is:

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

and we note that in the case of white noise where $\sigma_n^2 = \sigma^2$ then we get the sample mean:

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

and miminum variance $\text{var}(\hat{\theta}) = \frac{\sigma^2}{N}$.

7 Maximum Likelihood Estimation (MLE)

7.1 Basic MLE Procedure

In some cases the MVU estimator may not exist or it cannot be found by any of the methods discussed so far. The MLE approach is an alternative method in cases where the PDF is known. With MLE the unknown parameter is estimated by maximising the PDF. That is define $\hat{\theta}$ such that:

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

where $\mathbf{x}$ is the vector of observed data (of N samples). It can be shown that $\hat{\theta}$ is asymptotically unbiased:

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

and asymptotically efficient:

$$\lim_{N \to \infty} \text{var}(\hat{\theta}) = \text{CRLB}$$

An important result is that i**f an MVU estimator exists, then the MLE procedure will produce it**. An important observation is that unlike the previous estimates the **MLE does not require an explicit expression for** $p(x; \theta)$! Indeed given a histogram plot of the PDF as a function of θ one can numerically search for the θ that maximises the PDF.

7.1.1 EXAMPLE

Consider the signal embedded in noise problem:

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

where $w[n]$ is WGN with zero mean but unknown variance which is also A, that is the unknown parameter, $\theta = A$, manifests itself both as the unknown signal and the variance of the noise. Although a highly unlikely scenario, this simple example demonstrates the power of the MLE approach since finding the MVU estimator by the previous procedures is not easy. Consider the PDF:

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

We consider $p(\mathbf{x};\theta)$ as a function of θ, thus it is a likelihood function and we need to maximise it wrt to θ. For Gaussian PDFs it easier to find the maximum of the log-likelihood function:

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

Differentiating we have:

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

and setting the derivative to zero and solving for θ, produces the MLE estimate:

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

where we have assumed $\theta > 0$. It can be shown that:

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

and:

$$\lim_{N\to\infty} \text{var}(\hat{\theta}) = \text{CRLB} = \frac{\theta^2}{N(\theta+\frac{1}{2})}$$

7.1.2 EXAMPLE

Consider a signal embedded in noise:

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

where $w[n]$ is WGN with zero mean and known variance σ^2. We know the MVU estimator for θ is the sample mean. To see that this is also the MLE, we consider the PDF:

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

and maximise the log-likelihood function be setting it to zero:

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

thus $\hat{\theta} = \frac{1}{N}\sum_{n=0}^{N-1}x[n]$ which is the sample-mean.

7.2 MLE for Transformed Parameters

The MLE of the transformed parameter, $\alpha = g(\theta)$, is given by:

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

where $\hat{\theta}$ is the MLE of θ. If g is not a one-to-one function (i.e. not invertible) then $\hat{\alpha}$ is obtained as the MLE of the transformed likelihood function, $p_T(\mathbf{x}; \alpha)$, which is defined as:

$$p_T(\mathbf{x}; \alpha) = \max_{\{\theta: \alpha = g(\theta)\}} p(\mathbf{x}; \theta)$$

7.3 MLE for the General Linear Model

Consider the general linear model of the form:

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

where $\mathbf{H}$ is a known $N \times p$ matrix, $\mathbf{x}$ is the $N \times 1$ observation vector with N samples, and $\mathbf{w}$ is a noise vector of dimension $N \times 1$ with PDF $\mathbf{N}(\mathbf{0}, \mathbf{C})$. The PDF is:

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

and the MLE of $\boldsymbol{\theta}$ is found by differentiating the log-likelihood which can be shown to yield:

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

which upon simplification and setting to zero becomes:

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

and this we obtain the MLE of $\boldsymbol{\theta}$ as:

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

which is the same as the MVU estimator.

7.4 EM Algorithm

We wish to use the MLE procedure to find an estimate for the unknown parameter $\boldsymbol{\theta}$ which requires maximisation of the log-likelihood function, $\ln p_x(\mathbf{x}; \boldsymbol{\theta})$. However we may find that this is either too difficult or there arediffículties in finding an expression for the PDF itself. In such circumstances where direct expression and maxmisation of the PDF in terms of the observed data $\mathbf{x}$ is

difficult or intractable, an iterative solution is possible if another data set, $\mathbf{y}$, can be found such that the PDF in terms of $\mathbf{y}$ set is much easier to express in closed form and maximise. We term the data set $\mathbf{y}$ the *complete* data and the original data $\mathbf{x}$ the *incomplete* data. In general we can find a mapping from the complete to the incomplete data:

$$\mathbf{x} = \mathbf{g}(\mathbf{y})$$

however this is usually a many-to-one transformation in that a subset of the *complete* data will map to the same *incomplete* data (e.g. the incomplete data may represent an accumulation or sum of the complete data). This explains the terminology: $\mathbf{x}$ is *incomplete* (or is "missing" something) relative to the data, $\mathbf{y}$, which is complete (for performing the MLE procedure). This is usually not evident, however, until one is able to define what the complete data set. Unfortuneately defining what constitutes the complete data is usually an arbitrary procedure which is highly problem specific.

7.4.1 EXAMPLE

Consider spectral analysis where a known signal, $x[n]$, is composed of an unknown summation of harmonic components embedded in noise:

$$x[n] = \sum_{i=1}^{p} \cos\, 2\pi f_i n + w[n] \qquad n = 0, 1, \ldots, N-1$$

where $w[n]$ is WGN with known variance σ^2 and the unknown parameter vector to be estimated is the group of frequencies: $\boldsymbol{\theta} = \mathbf{f} = [f_1\, f_2\, \ldots\, f_p]^T$. The standard MLE would require maximisation of the log-likelihood of a multivariate Gaussian distribution which is equivalent to minimising the argument of the exponential:

$$J(\mathbf{f}) = \sum_{n=0}^{N-1} \left(x[n] - \sum_{i=1}^{p} \cos\, 2\pi f_i n \right)^2$$

which is a p-dimensional minimisation problem (hard!). On the other hand if we had access to the individual harmonic signal embedded in noise:

$$y_i[n] = \cos\, 2\pi f_i n + w_i[n] \qquad \begin{matrix} i = 1, 2, \ldots, p \\ n = 0, 1, \ldots, N-1 \end{matrix}$$

where $w_i[n]$ is WGN with known variance σ_i^2 then the MLE procedure would result in minimisation of:

$$J(f_i) = \sum_{n=0}^{N-1} (y_i[n] - \cos\, 2\pi f_i n)^2 \qquad i = 1, 2, \ldots, p$$

which are p independent one-dimensional minimisation problems (easy!). Thus $\mathbf{y}$ is the complete data set that we are looking for to facilitate the MLE procedure. However we do not have access to $\mathbf{y}$. The relationship to the known data $\mathbf{x}$ is:

$$\mathbf{x} = \mathbf{g}(\mathbf{y}) \quad \Rightarrow \quad x[n] = \sum_{i=1}^{p} y_i[n] \qquad n = 0, 1, \ldots, N-1$$

and we further assume that:

$$\begin{aligned} w[n] &= \sum_{i=1}^{p} w_i[n] \\ \sigma^2 &= \sum_{i=1}^{p} \sigma_i^2 \end{aligned}$$

However since the mapping from $\mathbf{y}$ to $\mathbf{x}$ is many-to-one we cannot directly form an expression for the PDF $p_y(\mathbf{y};\boldsymbol{\theta})$ in terms of the known $\mathbf{x}$ (since we can't do the obvious substitution of $\mathbf{y} = \mathbf{g}^{-1}(\mathbf{x})$).

♢ ♢ ♢ ♢ ♢ ♢ ♢ ♢ ♢

Once we have found the complete data set, $\mathbf{y}$, even though an expression for $\ln\, p_y(\mathbf{y};\boldsymbol{\theta})$ can now easily be derived we can't directly maxmise $\ln\, p_y(\mathbf{y};\boldsymbol{\theta})$ wrt $\boldsymbol{\theta}$ since $\mathbf{y}$ is unavailable. However we know $\mathbf{x}$ and if we further assume that we have a good "guess" estimate for $\boldsymbol{\theta}$ then we can consider the expected value of $\ln\, p_y(\mathbf{y};\boldsymbol{\theta})$ conditioned on what we know:

$$E[\ln\, p_y(\mathbf{y};\boldsymbol{\theta})|\mathbf{x};\boldsymbol{\theta}] = \int \ln\, p_y(\mathbf{y};\boldsymbol{\theta}) p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) d\mathbf{y}$$

and we attempt to maximise this expectation to yield, not the MLE, but next "best-guess" estimate for $\boldsymbol{\theta}$.

What we do is iterate through both an **E-step** (find the expression for the Expectation) and **M-step** (Maxmisation of the expectation), hence the name EM algorithm. Specifically:

Expectation(E): Determine the average or expectation of the log-likelihood of the complete data given the known incomplete or observed data and current estimate of the parameter vector

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k) = \int \ln\, p_y(\mathbf{y};\boldsymbol{\theta}) p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_k) d\mathbf{y}$$

Maxmisation (M): Maximise the average log-likelihood of the complete date or "Q-function" to obtain the next estimate of the parameter vector

$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k)$$

Convergence of the EM algorithm is guaranteed (under mild conditions) in the sense that the average log-likelihood of the complete data does not decrease at each iteration, that is:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{k+1}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k)$$

with equality when $\boldsymbol{\theta}_k$ is the MLE. The three main attributes of the EM algorithm are:

1. An initial value for the unknown parameter is needed and as with most iterative procedures a good initial estimate is required for good convergence
2. The selection of the complete data set is arbitrary
3. Although $\ln p_y(\mathbf{y}; \boldsymbol{\theta})$ can usually be easily expressed in closed form finding the closed form expression for the expectation is usually harder.

7.4.2 EXAMPLE

Applying the EM algorithm to the previous example requires finding a closed expression for the average log-likelihood.

E-step We start with finding an expression for $\ln p_y(\mathbf{y}; \boldsymbol{\theta})$ in terms of the complete data:

$$\begin{aligned} \ln p_y(\mathbf{y}; \boldsymbol{\theta}) &\approx h(\mathbf{y}) + \sum_{i=1}^{p} \frac{1}{\sigma_i^2} \sum_{n=0}^{N-1} y_i[n] \cos 2\pi f_i n \\ &\approx h(\mathbf{y}) + \sum_{i=1}^{p} \frac{1}{\sigma_i^2} \mathbf{c}_i^T \mathbf{y}_i \end{aligned}$$

where the terms in $h(\mathbf{y})$ do not depend on $\boldsymbol{\theta}$, $\mathbf{c}_i = [1, \cos 2\pi f_i, \cos 2\pi f_i(2), \ldots, \cos 2\pi f_i(N-1)]^T$ and $\mathbf{y}_i = [y_i[0], y_i[1], y_i[2], \ldots, y_i[N-1]]^T$. We write the conditional expectation as:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k) &= E[\ln p_y(\mathbf{y}; \boldsymbol{\theta}) | \mathbf{x}; \boldsymbol{\theta}_k] \\ &= E(h(\mathbf{y}) | \mathbf{x}; \boldsymbol{\theta}_k) + \sum_{i=1}^{p} \frac{1}{\sigma_i^2} \mathbf{c}_i^T E(\mathbf{y}_i | \mathbf{x}; \boldsymbol{\theta}_k) \end{aligned}$$

Since we wish to maximise $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k)$ wrt to $\boldsymbol{\theta}$, then this is equivalent to maximing:

$$Q'(\boldsymbol{\theta}, \boldsymbol{\theta}_k) = \sum_{i=1}^{p} \mathbf{c}_i^T E(\mathbf{y}_i | \mathbf{x}; \boldsymbol{\theta}_k)$$

We note that $E(\mathbf{y}_i | \mathbf{x}; \boldsymbol{\theta}_k)$ can be thought as as an estimate of the $y_i[n]$ data set given the observed data set $x[n]$ and current estimate $\boldsymbol{\theta}_k$. Since $\mathbf{y}$ is Gaussian

then $\mathbf{x}$ is a sum of Gaussians and thus $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian and one of the standard results is:

$$E(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_k) = E(\mathbf{y}) + \mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}(\mathbf{x} - E(\mathbf{x}))$$

and application of this yields:

$$\hat{\mathbf{y}}_i = E(\mathbf{y}_i|\mathbf{x};\boldsymbol{\theta}_k) = \mathbf{c}_i + \frac{\sigma_i^2}{\sigma^2}(\mathbf{x} - \sum_{i=1}^{p}\mathbf{c}_i)$$

and

$$\hat{y}_i[n] = \cos 2\pi f_{i_k} n + \frac{\sigma_i^2}{\sigma^2}\left(x[n] - \sum_{i=1}^{p}\cos 2\pi f_{i_k} n\right)$$

Thus:

$$Q'(\boldsymbol{\theta}, \boldsymbol{\theta}_k) = \sum_{i=1}^{p}\mathbf{c}_i^T\hat{\mathbf{y}}_i$$

M-step Maximisation of $Q'(\boldsymbol{\theta}, \boldsymbol{\theta}_k)$ consists of maximising each term in the sum separately or:

$$f_{i_{k+1}} = \arg\max_{f_i} \mathbf{c}_i^T\hat{\mathbf{y}}_i$$

Furthermore since we assumed $\sigma^2 = \sum_{i=1}^{p}\sigma_i^2$ we still have the problem that we don't know what the σ_i^2 are. However as long as:

$$\sigma^2 = \sum_{i=1}^{p}\sigma_i^2 \quad \Rightarrow \quad \sum_{i=1}^{p}\frac{\sigma_i^2}{\sigma^2} = 1$$

then we can chose these values arbitrarily.

8 Least Squares Estimation (LSE)

8.1 Basic LSE Procedure

The MVU, BLUE and MLE estimators developed previously required an expression for the PDF $p(\mathbf{x};\theta)$ in order to estimate the unknown parameter θ in some optimal fashion. An alternative approach is to assume a signal model (rather than probabilistic assumptions about the data) and achieve a design goal assuming this model. With the Least Squares (LS) approach we assume that the signal model is a function of the unknown parameter θ and produces a signal:

$$s[n] \equiv s(n;\theta)$$

where $s(n;\theta)$ is a function of n and parameterised by θ. Due to noise and model inaccuracies, $w[n]$, the signal $s[n]$ can only be observed as:

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

Unlike previous approaches no statement is made on the probabilistic distribution nature of $w[n]$. We only state that what we have is an "error": $e[n] = x[n] - s[n]$ which with the appropriate choice of θ should be minimised in a least-squares sense. That is we choose $\theta = \hat{\theta}$ so that the criterion:

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

is minimised over the N observation samples of interest and we call this the LSE of θ. More precisely we have:

$$\hat{\theta} = \arg\min_{\theta} J(\theta)$$

and the minimum LS error is given by:

$$J_{\min} = J(\hat{\theta})$$

An important assumption to produce a meaningful unbiassed estimate is that the noise and model inaccuracies, $w[n]$, have *zero-mean*. However no other probabilistic assumptions about the data are made (i.e. LSE is valid for both Gaussian and non-Gaussian noise), although by the same token we also cannot make any optimality claims with LSE. (since this would depend on the distribution of the noise and modelling errors).

A problem that arises from assuming a signal model function $s(n;\theta)$ rather than knowledge of $p(\mathbf{x};\theta)$ is the need to choose an appropriate signal model. Then again in order to obtain a closed form or "parameteric" expression for $p(\mathbf{x};\theta)$ one usually needs to know what the underlying model and noise characteristics are anyway.

8.1.1 EXAMPLE

Consider observations, $x[n]$, arising from a DC-level signal model, $s[n] = s(n;\theta) = \theta$:

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

where θ is the unknown parameter to be estimated. Then we have:

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

The differentiating wrt θ and setting to zero:

$$\left.\frac{\partial J(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0 \quad \Rightarrow \quad -2\sum_{n=0}^{N-1}(x[n]-\hat{\theta}) = 0 \quad \Rightarrow \quad \sum_{n=0}^{N-1} x[n] - N\hat{\theta} = 0$$

and hence $\hat{\theta} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$ which is the *sample-mean*. We also have that:

$$J_{\min} = J(\hat{\theta}) = \sum_{n=0}^{N-1}\left(x[n] - \frac{1}{N}\sum_{n=0}^{N-1} x[n]\right)^2$$

8.2 Linear Least Squares

We assume the signal model is a linear function of the estimator, that is:

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$$

where $\mathbf{s} = [s[0], s[1], \ldots, s[N-1]]^T$and $\mathbf{H}$ is a known $N \times p$ matrix with $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]$. Now:

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

and with $\mathbf{x} = [x[0], x[1], \ldots, x[N-1]]^T$ we have:

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

Differentiating and setting to zero:

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

yields the required LSE:

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

which, surprise, surprise, is the identical functional form of the MVU estimator for the linear model.

An interesting extension to the linear LS is the *weighted* LS where the contribution to the error from each component of the parameter vector can be weighted in importance by using a different from of the error criterion:

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

where $\mathbf{W}$ is an $N \times N$ postive definite (symmetric) weighting matrix.

8.3 Order-Recursive Least Squares

In many cases the signal model is unknown and must be assumed. Obviously we would like to choose the model, $s(\boldsymbol{\theta})$, that minimises $J_{\min}$, that is:

$$s_{\text{best}}(\boldsymbol{\theta}) = \arg\min_{s(\hat{\boldsymbol{\theta}})} J_{\min}$$

We can do this arbitrarily by simply choosing models, obtaining the LSE $\hat{\theta}$, and then selecting the model which provides the smallest $J_{\min}$. However, models are not arbitrary and some models are more "complex" (or more precisely have a larger number of parameters or degrees of freedom) than others. The more complex a model the lower the $J_{\min}$ one can expect but also the more likely the model is to *overfit* the data or be *overtrained* (i.e. fit the noise and not generalise to other data sets).

8.3.1 EXAMPLE

Consider the case of "line fitting" where we have observations $x(t)$ plotted against the sample time index t and we would like to fit the "best" line to the data. But what line do we fit: $s(t;\boldsymbol{\theta}) = \theta_1$, constant? $s(t;\boldsymbol{\theta}) = \theta_1 + \theta_2 t$, a straight line? $s(t;\boldsymbol{\theta}) = \theta_1 + \theta_2 t + \theta_3 t^2$, a quadratic? etc. Each case represents an increase in the *order* of the model (i.e. order of the polynomial fit), or the number of parameters to be estimated and consequent increase in the "modelling power". A polynomial fit represents a linear model, $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$, where $\mathbf{s} = [s(0), s(1), \ldots, s(N-1)]^T$ and:

Constant $\boldsymbol{\theta} = [\theta_1]^T$ and $\mathbf{H} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ is a $N \times 1$ matrix

Linear $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$ and $\mathbf{H} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & N-1 \end{bmatrix}$ is a $N \times 2$ matrix

Quadratic $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]^T$ and $\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ \vdots & \vdots & \vdots \\ 1 & N-1 & (N-1)^2 \end{bmatrix}$ is a $N \times 3$ matrix

and so on. If the underlying model is indeed a straight line then we would expect not only that the minimum $J_{\min}$ result with a straight line model but also that higher order polynomial models (e.g. quadratic, cubic, etc.) will yield the same $J_{\min}$ (indeed higher-order models would "degenerate" to a straight line model, except in cases of overfitting). Thus the straight line model is the "best" model to use.

$\diamondsuit \quad \diamondsuit \quad \diamondsuit \quad \diamondsuit \quad \diamondsuit \quad \diamondsuit \quad \diamondsuit \quad \diamondsuit \quad \diamondsuit$

An alternative to providing an independent LSE for each possible signal model a more efficient *order-recursive* LSE is possible if the models are different orders of the same base model (e.g. polynomials of different degree). In this method the LSE is updated in order (of increasing parameters). Specifically define $\hat{\boldsymbol{\theta}}_k$ as the LSE of order k (i.e. k parameters to be estimated). Then for a linear model we can derive the order-recursive LSE as:

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \text{UPDATE}_k$$

However the success of this approach depends on proper formulation of the linear models in order to facilitate the derivation of the recursive update. For example, if $\mathbf{H}$ has othonormal column vectors then the LSE is equivalent to projecting the observation $\mathbf{x}$ onto the space spanned by the orthonormal column vectors of $\mathbf{H}$. Since increasing the order implies increasing the dimensionality of the space by just adding another column to $\mathbf{H}$ this allows a recursive update relationship to be derived.

8.4 Sequential Least Squares

In most signal processing applications the observations samples arrive as a stream of data. All our estimation strategies have assumed a *batch* or block mode of processing whereby we wait for N samples to arrive and then form our estimate based on these samples. One problem is the delay in waiting for the N samples before we produce our estimate, another problem is that as more data arrives we have to repeat the calculations on the larger blocks of data (N increases as more data arrives). The latter not only implies a growing computational burden but also the fact that we have to buffer all the data we have seen, both will grow linearly with the number of samples we have. Since in signal processing applications samples arise from sampling a continuous process our computational and memory burden will grow linearly with time! One solution is to use a *sequential* mode of processing where the parameter estimate for n samples, $\hat{\theta}[n]$, is derived from the previous parameter estimate for $n-1$ samples, $\hat{\theta}[n-1]$. For linear models we can represent sequential LSE as:

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

where $s[n|n-1] \equiv s(n;\theta[n-1])$. The $K[n]$ is the correction gain and $(x[n] - s[n|n-1])$ is the prediction error. The magnitude of the correction gain $K[n]$ is

usually directly related to the value of the estimator error variance, $\text{var}(\hat{\theta}[n-1])$, with a larger variance yielding a larger correction gain. This behaviour is reasonable since a larger variance implies a poorly estimated parameter (which should have minumum variance) and a larger correction gain is expected. Thus one expects the variance to decrease with more samples and the estimated parameter to converge to the "true" value (or the LSE with an infinite number of samples).

8.4.1 EXAMPLE

We consider the specific case of linear model with a vector parameter:

$$\mathbf{s}[n] = \mathbf{H}[n]\boldsymbol{\theta}[n]$$

The most interesting example of sequential LS arises with the weighted LS error criterion with $\mathbf{W}[n] = \mathbf{C}^{-1}[n]$ where $\mathbf{C}[n]$ is the covariance matrix of the zero-mean noise, $\mathbf{w}[n]$, which is assumed to be uncorrelated. The argument $[n]$ implies that the vectors are based on n sample observations. We also consider:

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

where $\mathbf{h}^T[n]$ is the n^{th}row vector of the $n \times p$ matrix $\mathbf{H}[n]$. It should be obvious that:

$$\mathbf{s}[n-1] = \mathbf{H}[n-1]\boldsymbol{\theta}[n-1]$$

We also have that $s[n|n-1] = \mathbf{h}^T[n]\boldsymbol{\theta}[n-1]$. So the **estimator update** is:

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

Let $\Sigma[n] = \mathbf{C}_{\hat{\boldsymbol{\theta}}}[n]$ be the covariance matrix of $\hat{\boldsymbol{\theta}}$ based on n samples of data and the it can be shown that:

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

and furthermore we can also derive a **covariance update**:

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

yielding a wholly recursive procedure requiring only knowledge of the observation data $x[n]$ and the initialisation values: $\hat{\boldsymbol{\theta}}[-1]$ and $\Sigma[-1]$, the initial estimate of the parameter and a initial estimate of the parameter covariance matrix.

8.5 Constrained Least Squares

Assume that in the vector parameter LSE problem we are aware of constraints on the individual parameters, that is p-dimensional $\boldsymbol{\theta}$ is subject to $r < p$ independent linear constraints. The constraints can be summarised by the condition that $\boldsymbol{\theta}$ satisfy the following system of linear "constraint" equations:

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$$

Then using the technique of Lagrangian multipliers our LSE problem is that of miminising the following Lagrangian error criterion:

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

Let $\hat{\boldsymbol{\theta}}$ be the unconstrained LSE of a linear model, then the expression for $\hat{\boldsymbol{\theta}}_c$ the constrained estimate is:

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

where $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ for unconstrained LSE.

8.6 Nonlinear Least Squares

So far we have assumed a linear signal model: $\mathbf{s}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta}$ where the notation for the signal model, $\mathbf{s}(\boldsymbol{\theta})$, explicitly shows its dependence on the parameter $\boldsymbol{\theta}$. In general the signal model will be an N-dimensional nonlinear function of the p-dimensional parameter $\boldsymbol{\theta}$. In such a case the minimisation of:

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

becomes much more difficult. Differentiating wrt $\boldsymbol{\theta}$ and setting to zero yields:

$$\frac{\partial J}{\partial \boldsymbol{\theta}} = \mathbf{0} \quad \Rightarrow \quad \frac{\partial \mathbf{s}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}(\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})) = \mathbf{0}$$

which requires solution of N nonlinear simultaneous equations. Approximate solutions based on linearization of the problem exist which require iteration until convergence.

8.6.1 Newton-Rhapson Method

Define:

$$\mathbf{g}(\boldsymbol{\theta}) = \frac{\partial \mathbf{s}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}(\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))$$

So the problem becomes that of finding the zero of a nonlinear function, that is: $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}$. The Newton-Rhapson method linearizes the function about the

initialised value $\boldsymbol{\theta}_k$ and directly solves for the zero of the function to produce the next estimate:

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

Of course if the function was linear the next estimate would be the correct value, but since the function is nonlinear this will not be the case and the procedure is iterated until the estimates converge.

8.6.2 Gauss-Newton Method

We linearize the signal model about the known (i.e. initial guess or current estimate) $\boldsymbol{\theta}_k$:

$$\mathbf{s}(\boldsymbol{\theta}) \approx \mathbf{s}(\boldsymbol{\theta}_k) + \frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k} (\boldsymbol{\theta} - \boldsymbol{\theta}_k)$$

and the LSE minimisation problem which can be shown to be:

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

where $\hat{\mathbf{x}}(\boldsymbol{\theta}_k) = \mathbf{x} - \mathbf{s}(\boldsymbol{\theta}_k) + \mathbf{H}(\boldsymbol{\theta}_k)\boldsymbol{\theta}_k$ is known and:

$$\mathbf{H}(\boldsymbol{\theta}_k) = \frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k}$$

Based on the standard solution to the LSE of the linear model, we have that:

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

9 Method of Moments

Although we may not have an expression for the PDF we assume that we can use the natrual estimator for the k^{th} moment, $\mu_k = E(x^k[n])$, that is:

$$\hat{\mu}_k = \frac{1}{N}\sum_{n=0}^{N-1} x^k[n]$$

If we can write an expression for the k^{th} moment as a function of the unknown parameter θ:

$$\mu_k = h(\theta)$$

and assuming h^{-1} exists then we can derive the an estimate by:

$$\hat{\theta} = h^{-1}(\hat{\mu}_k) = h^{-1}\left(\frac{1}{N}\sum_{n=0}^{N-1} x^k[n]\right)$$

If $\boldsymbol{\theta}$ is a p-dimensional vector then we require p equations to solve for the p unknowns. That is we need some set of p moment equations. Using the lowest order p moments what we would like is:

$$\hat{\boldsymbol{\theta}} = \mathbf{h}^{-1}(\hat{\boldsymbol{\mu}})$$

where:

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \frac{1}{N}\sum_{n=0}^{N-1} x[n] \\ \frac{1}{N}\sum_{n=0}^{N-1} x^2[n] \\ \vdots \\ \frac{1}{N}\sum_{n=0}^{N-1} x^p[n] \end{bmatrix}$$

9.1 EXAMPLE

Consider a 2-mixture Gaussian PDF:

$$p(x;\theta) = (1-\theta)g_1(x) + \theta g_2(x)$$

where $g_1 = \mathrm{N}(\mu_1, \sigma_1^2)$ and $g_2 = \mathrm{N}(\mu_2, \sigma_2^2)$ are two different Gaussian PDFs and θ is the unknown parameter that has to be estimated. We can write the second moment as a function of θ as follows:

$$\mu_2 = E(x^2[n]) = \int x^2 p(x;\theta)dx = (1-\theta)\sigma_1^2 + \theta\sigma_2^2 = h(\theta)$$

and hence:

$$\hat{\theta} = \frac{\hat{\mu}_2 - \sigma_1^2}{\sigma_2^2 - \sigma_1^2} = \frac{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] - \sigma_1^2}{\sigma_2^2 - \sigma_1^2}$$

10 Bayesian Philosophy

10.1 Minimum Mean Square Estimator (MMSE)

The *classic approach* we have been using so far has assumed that the parameter θ is unknown but deterministic. Thus the optimal estimator $\hat{\theta}$ is optimal irrespective and independent of the actual value of θ. But in cases where the actual value or prior knowledge of θ could be a factor (e.g. where an MVU estimator does not exist for certain values or where prior knowledge would improve the estimation) the classic approach would not work effectively.

In the Bayesian philosophy the θ is treated as a random variable with a known *prior pdf*, $p(\theta)$. Such prior knowledge concerning the distribution of the estimator should provide better estimators than the deterministic case.

In the classic approach we derived the MVU estimator by first considering mimisation of the *mean square eror,* i.e. $\hat{\theta} = \arg\min_{\hat{\theta}} \text{mse}(\hat{\theta})$ where:

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

and $p(\mathbf{x}; \theta)$ is the pdf of $\mathbf{x}$ parametrised by θ. In the Bayesian approach we similarly derive an estimator by minimising $\hat{\theta} = \arg\min_{\hat{\theta}} \text{Bmse}(\hat{\theta})$ where:

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

is the Bayesian mse and $p(\mathbf{x}, \theta)$ is the joint pdf of $\mathbf{x}$ and θ (since θ is now a random variable). It should be noted that the Bayesian squared error $(\theta - \hat{\theta})^2$ and classic squared error $(\hat{\theta} - \theta)^2$ are the same. The minimum $\text{Bmse}(\hat{\theta})$ estimator or MMSE is derived by differentiating the expression for $\text{Bmse}(\hat{\theta})$ with respect to $\hat{\theta}$ and setting this to zero to yield:

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

where the *posterior pdf,* $p(\theta|\mathbf{x})$, is given by:

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

Apart from the computational (and analytical!) requirements in deriving an expression for the posterior pdf and then evaluating the expectation $E(\theta|\mathbf{x})$ there is also the problem of finding an appropriate prior pdf . The usual choice is to assume the joint pdf, $p(\mathbf{x}, \theta)$, is Gaussian and hence both the prior pdf, $p(\theta)$, and posterior pdf, $p(\theta|\mathbf{x})$, are also Gaussian (this property imples the Gaussian pdf is a conjugate prior distribution). Thus the form of the pdfs remains the same and all that changes are the means and variances.

10.1.1 EXAMPLE

Consider signal embedded in noise:

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

where as before $w[n] = \text{N}(0, \sigma^2)$ is a WGN process and the unknown parameter $\theta = A$ is to be estimated. However in the Bayesian approach we also assume the parameter A is a random variable with a prior pdf which in this case is the Gaussian pdf $p(A) = \text{N}(\mu_A, \sigma_A^2)$. We also have that $p(\mathbf{x}|A) = \text{N}(A, \sigma^2)$ and we can assume that $\mathbf{x}$ and A are jointly Gaussian. Thus the posterior pdf:

$$p(A|\mathbf{x}) = \frac{p(\mathbf{x}|A) p(A)}{\int p(\mathbf{x}|A) p(A) dA} = \text{N}(\mu_{A|\mathbf{x}}, \sigma_{A|\mathbf{x}}^2)$$

is also a Gaussian pdf and after the required simplification we have that:

$$\sigma^2_{A|\mathbf{x}} = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma^2_A}} \qquad \text{and} \qquad \mu_{A|\mathbf{x}} = \left(\frac{N}{\sigma^2}\bar{x} + \frac{\mu_A}{\sigma^2_A}\right)\sigma^2_{A|\mathbf{x}}$$

and hence the MMSE is:

$$\begin{aligned} \hat{A} = E[A|\mathbf{x}] &= \int Ap(A|\mathbf{x})dA = \mu_{A|\mathbf{x}} \\ &= \alpha\bar{x} + (1-\alpha)\mu_A \end{aligned}$$

where $\alpha = \frac{\sigma^2_A}{\sigma^2_A + \frac{\sigma^2}{N}}$. Upon closer examination of the MMSE we observe the following (assume $\sigma^2_A \ll \sigma^2$):

1. With few data (N is small) then $\sigma^2_A \ll \sigma^2/N$ and $\hat{A} \to \mu_A$, that is the MMSE tends towards the mean of the prior pdf (and effectively ignores the contribution of the data). Also $p(A|\mathbf{x}) \approx \mathrm{N}(\mu_A, \sigma^2_A)$.

2. With large amounts of data (N is large)$\sigma^2_A \gg \sigma^2/N$ and $\hat{A} \to \bar{x}$, that is the MMSE tends towards the sample mean $\bar{x}$ (and effectively ignores the contribution of the prior information). Also $p(A|\mathbf{x}) \approx \mathrm{N}(\bar{x}, \sigma^2/N)$.

If $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, where $\mathbf{x}$ is $k \times 1$ and $\mathbf{y}$ is $l \times 1$, with mean vector $[E(\mathbf{x})^T\ , E(\mathbf{y})^T]^T$ and partitioned covariance matrix

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} = \begin{bmatrix} k \times k & k \times l \\ l \times k & l \times l \end{bmatrix}$$

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

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

10.2 Bayesian Linear Model

Now consider the Bayesian Llinear Model:

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

where $\boldsymbol{\theta}$ is the unknown parameter to be estimated with prior pdf $\mathbf{N}(\boldsymbol{\mu}_\theta, \mathbf{C}_\theta)$ and $\mathbf{w}$ is a WGN with $\mathbf{N}(\mathbf{0}, \mathbf{C}_w)$. The MMSE is provided by the expression for $E(\mathbf{y}|\mathbf{x})$ where we identify $\mathbf{y} \equiv \boldsymbol{\theta}$. We have that:

$$\begin{aligned} E(\mathbf{x}) &= \mathbf{H}\boldsymbol{\mu}_\theta \\ E(\mathbf{y}) &= \boldsymbol{\mu}_\theta \end{aligned}$$

and we can show that:

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

and hence since $\mathbf{x}$ and $\boldsymbol{\theta}$ are *jointly Gaussian* we have that:

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

♢ ♢ ♢ ♢ ♢ ♢ ♢ ♢ ♢

10.3 Relation with Classic Estimation

In classical estimation we cannot make any assumptions on the prior, thus all possible θ have to be considered. The equivalent prior pdf would be a flat distribution, essentially $\sigma_\theta^2 = \infty$. This so-called *noninformative* prior pdf will yield the classic estimator where such is defined.

10.3.1 EXAMPLE

Consider the signal embedded in noise problem:

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

where we have shown that the MMSE is:

$$\hat{A} = \alpha\bar{x} + (1-\alpha)\mu_A$$

where $\alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}$. If the prior pdf is noninformative then $\sigma_A^2 = \infty$ and $\alpha = 1$ with $\hat{A} = \bar{x}$ which is the classic estimator.

10.4 Nuisance Parameters

Suppose that both θ and α were unknown parameters but we are only interested in θ. Then α is a nuisance parameter. We can deal with this by "integrating α out of the way". Consider:

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

Now $p(\mathbf{x}|\theta)$ is, in reality, $p(\mathbf{x}|\theta,\alpha)$, but we can obtain the true $p(\mathbf{x}|\theta)$ by:

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

and if α and θ are independent then:

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

11 General Bayesian Estimators

The $\text{Bmse}(\hat{\theta})$ given by:

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

is one specific case for a general estimator that attempts to minimise the average of the cost function, $\mathcal{C}(\epsilon)$, that is the Bayes risk $\mathcal{R} = E[\mathcal{C}(\epsilon)]$ where $\epsilon = (\theta - \hat{\theta})$. There are three different cost functions of interest:

1. **Quadratic:** $\mathcal{C}(\epsilon) = \epsilon^2$ which yields $\mathcal{R} = \text{Bmse}(\hat{\theta})$. We already know that the estimate to minimise $\mathcal{R} = \text{Bmse}(\hat{\theta})$ is:

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

   which is the mean of the *posterior pdf*.

2. **Absolute:** $\mathcal{C}(\epsilon) = |\epsilon|$. The estimate, $\hat{\theta}$, that minimises $\mathcal{R} = E[|\theta - \hat{\theta}|]$ satisfies:

   $$\int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{x}) d\theta = \int_{\hat{\theta}}^{\infty} p(\theta|\mathbf{x}) d\theta \text{ or } \Pr\{\theta \leq \hat{\theta}|\mathbf{x}\} = \frac{1}{2}$$

   that is, the median of the *posterior pdf*.

3. **Hit-or-miss:** $\mathcal{C}(\epsilon) = \begin{cases} 0 & |\epsilon| < \delta \\ 1 & |\epsilon| > \delta \end{cases}$. The estimate that minimises the Bayes risk can be shown to be:

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

   which is the mode of the *posterior pdf* (the value that maxmises the pdf).

For the Gaussian posterior pdf it should be noted that the mean, mediam and mode are identical. Of most interest are the quadratic and hit-or-miss cost functions which, together with a special case of the latter, yield the following three important classes of estimator:

1. **MMSE** (***Minimum Mean Square Error***) estimator which we have already introduced as the mean of the *posterior pdf*:

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

2. **MAP** (***Maximum A Posteriori***) estimator which is the mode or maximum of the *posterior pdf*:

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} p(\theta|\mathbf{x}) \\ &= \arg\max_{\theta} p(\mathbf{x}|\theta)p(\theta)\end{aligned}$$

3. **Bayesian ML** (***Bayesian Maximum Likelihood***) estimator which is the special case of the MAP estimator where the prior pdf, $p(\theta)$, is uniform or noninformative:

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

Noting that the conditional pdf of $\mathbf{x}$ given θ, $p(\mathbf{x}|\theta)$, is essentially equivalent to the pdf of $\mathbf{x}$ parametrized by θ, $p(\mathbf{x};\theta)$, the Bayesian ML estimator is equivalent to the classic MLE.

Comparing the three estimators:

- The MMSE is preferred due to its least-squared cost function but it is also the most difficult to derive and compute due to the need to find an expression or measurements of the posterior pdf $p(\theta|\mathbf{x})$in order to integrate $\int \theta p(\theta|\mathbf{x})d\theta$
- The MAP hit-or-miss cost function is more "crude" but the MAP estimate is easier to derive since there is no need to integrate, only find the maximum of the posterior pdf $p(\theta|\mathbf{x})$ either analytically or numerically.
- The Bayesian ML is equivalent in performance to the MAP only in the case where the prior pdf is noninformative, otherwise it is a sub-optimal estimator. However, like the classic MLE, the expression for the conditional pdf $p(\mathbf{x}|\theta)$ is usually easier to obtain then that of the posterior pdf, $p(\theta|\mathbf{x})$. Since in most cases knowledge of the the prior pdf is unavailable so, not surprisingly, ML estimates tend to be more prevalent. However it may not always be prudent to assume the prior pdf is uniform, especially in cases where prior knowledge of the estimate is available even though the exact pdf is unknown. In these cases a MAP estimate may perform better even if an "artificial" prior pdf is assumed (e.g. a Gaussian prior which has the added benefit of yielding a Gaussian posterior).

12 Linear Bayesian Estimators

12.1 Linear Minimum Mean Square Error (LMMSE) Estimator

We assume that the parameter θ is to be estimated based on the data set $\mathbf{x} = [x[0]\, x[1]\, \ldots\, x[N-1]]^T$ rather than assume any specific form for the joint

pdf $p(\mathbf{x}, \theta)$. We consider the class of all affine estimators of the form:

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

where $\mathbf{a} = [a_1\, a_2\, \ldots\, a_{N-1}]^T$ and choose the weight co-efficients $\{\mathbf{a}, a_N\}$ to minimize the Bayesian MSE:

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

The resultant estimator is termed the *linear minimum mean square error* (**LMMSE**) estimator. The LMMSE will be sub-optimal unless the MMSE is also linear. Such would be the case if the Bayesian linear model applied:

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

The weight co-efficients are obtained from $\frac{\partial \text{Bmse}(\hat{\theta})}{\partial a_i} = 0 \quad \text{for } i = 1, 2, \ldots, N$ which yields:

$$a_N = E(\theta) - \sum_{n=0}^{N-1} a_n E(x[n]) = E(\theta) - \mathbf{a}^T \mathbf{E}(\mathbf{x}) \quad \text{and} \quad \mathbf{a} = \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta}$$

where $\mathbf{C}_{xx} = E\left[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T\right]$ is the $N \times N$ covariance matrix and $\mathbf{C}_{x\theta} = E\left[(\mathbf{x} - E(\mathbf{x}))(\theta - E(\theta))\right]$ is the $N \times 1$ cross-covariance vector. Thus the LMMSE estimator is:

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

where we note $\mathbf{C}_{\theta x} = \mathbf{C}_{x\theta}^T = E\left[(\theta - E(\theta))(\mathbf{x} - E(\mathbf{x}))^T\right]$. For the $1 \times p$ vector parameter $\boldsymbol{\theta}$ an equivalent expression for the LMMSE estimator is derived:

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

where now $\mathbf{C}_{\theta x} = E\left[(\boldsymbol{\theta} - E(\boldsymbol{\theta}))(\mathbf{x} - E(\mathbf{x}))^T\right]$ is the $p \times N$ cross-covariance matrix. And the Bayesian MSE matrix is:

$$\begin{aligned} \mathbf{M}_{\hat{\theta}} &= E\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T\right] \\ &= \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta x} \mathbf{C}_{xx}^{-1} \mathbf{C}_{x\theta} \end{aligned}$$

where $\mathbf{C}_{\theta\theta} = E\left[(\boldsymbol{\theta} - E(\boldsymbol{\theta}))(\boldsymbol{\theta} - E(\boldsymbol{\theta}))^T\right]$ is the $p \times p$ covariance matrix.

12.1.1 Bayesian Gauss-Markov Theorem

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

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

where $\mathbf{x}$ is the $N \times 1$ data vector, $\mathbf{H}$ is a known $N \times p$ observation matrix, $\boldsymbol{\theta}$ is a $p \times 1$ random vector of parameters with mean $E(\boldsymbol{\theta})$ and covariance matrix $\mathbf{C}_{\theta\theta}$ and $\mathbf{w}$ is an $N \times 1$ random vector with zero mean and covariance matrix $\mathbf{C}_w$ which is uncorrelated with $\boldsymbol{\theta}$ (the joint pdf $p(\mathbf{w}, \boldsymbol{\theta})$ and hence also $p(\mathbf{x}, \boldsymbol{\theta})$ are otherwise arbitrary), then noting that:

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

the LMMSE estimator of $\boldsymbol{\theta}$ is:

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

and the covariance of the error which is the Bayesian MSE matrix is:

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

12.2 Wiener Filtering

We assume N samples of time-series data $\mathbf{x} = [x[0]\, x[1]\, \ldots\, x[N-1]]^T$which are wide-sense stationary (WSS). As such the $N \times N$ covariance matrix takes the symmetric Toeplitz form:

$$\mathbf{C}_{xx} = \mathbf{R}_{xx} \quad \text{where} \quad [\mathbf{R}_{xx}]_{ij} = r_{xx}[i-j]$$

where $r_{xx}[k] = E(x[n]x[n-k])$ is the autocorrelation function (ACF) of the $x[n]$ process and $\mathbf{R}_{xx}$ denotes the autocorrelation matrix. Note that since $x[n]$ is WSS the expectation $E(x[n]x[n-k])$ is independent of the absolute time index n. In signal processing the estimated ACF is used:

$$\hat{r}_{xx}[k] = \begin{cases} \frac{1}{N}\sum_{n=0}^{N-1-|k|} x[n]x[n+|k|] & |k| \leq N-1 \\ 0 & |k| \geq N \end{cases}$$

Both the data $\mathbf{x}$ and the parameter to be estimated $\hat{\boldsymbol{\theta}}$ are assumed zero mean. Thus the LMMSE estimator is:

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

Application of the LMMSE estimator to the three signal processing estimation problems of filtering, smoothing and prediction gives rise to the *Wiener filtering* equation solutions.

12.2.1 Smoothing

The problem is to estimate the signal $\boldsymbol{\theta} = \mathbf{s} = [s[0]\, s[1] \ldots s[N-1]]^T$based on the noisy data $\mathbf{x} = [x[0]\, x[1] \ldots x[N-1]]^T$ where:

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

and $\mathbf{w} = [w[0]\, w[1] \ldots w[N-1]]^T$ is the noise process. An important difference between smoothing and filtering is that the signal estimate $s[n]$ can use the entire data set: the *past* values ($x[0]$, $x[1]$, $\ldots$ $x[n-1]$), the *present* $x[n]$ and *future* values ($x[n+1]$, $x[n+2]$, $\ldots, x[N-1]$). This means that the solution cannot be cast as "filtering" problem since we cannot apply a causal filter to the data.

We assume that the signal and noise processes are uncorrelated. Hence:

$$r_{xx}[k] = r_{ss}[k] + r_{ww}[k]$$

and thus:

$$\mathbf{C}_{xx} = \mathbf{R}_{xx} = \mathbf{R}_{ss} + \mathbf{R}_{ww}$$

also:

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

Hence the LMMSE estimator (also called the Wiener estimator) is:

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

and the $N \times N$ matrix:

$$\mathbf{W} = \mathbf{R}_{ss}(\mathbf{R}_{ss} + \mathbf{R}_{ww})^{-1}$$

is referred to as the *Wiener smoothing matrix*.

12.2.2 Filtering

The problem is to estimate the signal $\theta = s[n]$ based only on the *present and past* noisy data $\mathbf{x} = [x[0]\, x[1] \ldots x[n]]^T$. As n increases this allows us to view the estimation process as an application of a causal filter to the data and we need to cast the LMMSE estimator expression in the form of a filter.

Assuming the signal and noise processes are uncorrelated we have:

$$\mathbf{C}_{xx} = \mathbf{R}_{ss} + \mathbf{R}_{ww}$$

where $\mathbf{C}_{xx}$ is an $(n+1) \times (n+1)$ autocorrelation matrix. Also:

$$\begin{aligned} \mathbf{C}_{\theta x} = E(s\mathbf{x}^T) = E(s(\mathbf{s}+\mathbf{w})^T) &= E(s\mathbf{s}^T) \\ &= E(s[n]\,[s[0]\, s[1] \ldots s[n]]) \\ &= [r_{ss}[n]\, r_{ss}[n-1] \ldots r_{ss}[0]] = \check{\mathbf{r}}_{ss}^T \end{aligned}$$

is a $1 \times (n+1)$ row vector. The LMMSE estimator is:

$$\hat{s}[n] = \check{\mathbf{r}}_{ss}^T (\mathbf{R}_{ss} + \mathbf{R}_{ww})^{-1} \mathbf{x} = \mathbf{a}^T \mathbf{x}$$

where $\mathbf{a} = [a_0\, a_1\, \ldots\, a_n]^T = (\mathbf{R}_{ss} + \mathbf{R}_{ww})^{-1} \check{\mathbf{r}}_{ss}^T$ is the $(n+1) \times 1$ vector of weights. Note that the "check" subscript is used to denote *time-reversal*. Thus:

$$\mathbf{r}_{ss}^T = [r_{ss}[0]\, r_{ss}[1]\, \ldots\, r_{ss}[n]]$$

We interpret the process of forming the estimator as time evolves (n increases) as a filtering operation. Specifically we let $h^{(n)}[k]$, the time-varying impulse response, be the response of the filter at time n to an impulse applied k samples before (i.e. at time $n-k$). We note that a_i can be intrepreted as the response of the filter at time n to the signal (or impulse) applied at time $i = n-k$. Thus we can make the following correspondence:

$$h^{(n)}[k] = a_{n-k} \qquad k = 0,\, 1,\, \ldots,\, n.$$

Then:

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

We define the vector $\mathbf{h} = \left[h^{(n)}[0]\, h^{(n)}[1]\, \ldots\, h^{(n)}[n]\right]^T$. Then we have that $\mathbf{h} = \check{\mathbf{a}}$, that is, $\mathbf{h}$ is a time-reversed version of $\mathbf{a}$. To explicitly find the impulse response $\mathbf{h}$ we note that since:

$$(\mathbf{R}_{ss} + \mathbf{R}_{ww})\mathbf{a} = \check{\mathbf{r}}_{ss}^T$$

then it also true that:

$$(\mathbf{R}_{ss} + \mathbf{R}_{ww})\mathbf{h} = \mathbf{r}_{ss}$$

When written out we get the *Wiener-Hopf filtering equations*:

$$\begin{bmatrix} r_{xx}[0] & r_{xx}[1] & \cdots & r_{xx}[n] \\ r_{xx}[1] & r_{xx}[0] & \cdots & r_{xx}[n-1] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[n] & r_{xx}[n-1] & \cdots & r_{xx}[0] \end{bmatrix} \begin{bmatrix} h^{(n)}[0] \\ h^{(n)}[1] \\ \vdots \\ h^{(n)}[n] \end{bmatrix} = \begin{bmatrix} r_{ss}[0] \\ r_{ss}[1] \\ \vdots \\ r_{ss}[n] \end{bmatrix}$$

where $r_{xx}[k] = r_{ss}[k] + r_{ww}[k]$. A computationally efficient solution for solving the equations is the *Levinson recursion* which solves the equations recursively to avoid resolving them for each value of n.

12.2.3 Prediction

The problem is to estimate $\theta = x[N-1+l]$ based on the current and past $\mathbf{x} = [x[0]\, x[1]\, \ldots\, x[N-1]]^T$at sample $l \geq 1$ in the future. The resulting estimator is termed the *l-step linear predictor*.

As before we have $\mathbf{C}_{xx} = \mathbf{R}_{xx}$ where $\mathbf{R}_{xx}$ is the $N \times N$ autocorrelation matrix, and:

$$\begin{aligned}\mathbf{C}_{\theta x} &= E(x\mathbf{x}^T) \\ &= E(x[N-1+l]\,[x[0]\,x[1]\,\ldots x[N-1]]) \\ &= [r_{xx}[N-1+l]\,r_{xx}[N-2+l]\,\ldots\,r_{xx}[l]] = \check{\mathbf{r}}_{xx}^T\end{aligned}$$

Then the LMMSE estimator is:

$$\hat{x}[N-1+l] = \check{\mathbf{r}}_{xx}^T \mathbf{R}_{xx}^{-1}\mathbf{x} = \mathbf{a}^T\mathbf{x}$$

where $\mathbf{a} = [a_0\,a_1\,\ldots a_{N-1}]^T = \mathbf{R}_{xx}^{-1}\check{\mathbf{r}}_{xx}$. We can interpret the process of forming the estimator as a filtering operation where $h^{(N)}[k] = h[k] = a_{n-k} \;\Rightarrow\; h[N-k] = a_k$ and then:

$$\hat{x}[N-1+l] = \sum_{k=0}^{N-1} h[N-k]x[k] = \sum_{k=1}^{N} h[k]x[N-k]$$

Defining $\mathbf{h} = [h[1]\,h[2]\,\ldots\,h[N]]^T = [a_{N-1}\,a_{N-2}\,\ldots a_0]^T = \check{\mathbf{a}}$ as before we can find an explicit expression for $\mathbf{h}$ by noting that:

$$\mathbf{R}_{xx}\mathbf{a} = \mathbf{r}_{xx} \;\Rightarrow \mathbf{R}_{xx}\mathbf{h} = \mathbf{r}_{xx}$$

where $\mathbf{r}_{xx} = [r_{xx}[l]\,r_{xx}[l+1]\,\ldots\,r_{xx}[l+N-1]]^T$ is the time-reversed version of $\check{\mathbf{r}}_{xx}$. When written out we get the *Wiener-Hopf prediction equations*:

$$\begin{bmatrix} r_{xx}[0] & r_{xx}[1] & \cdots & r_{xx}[N-1] \\ r_{xx}[1] & r_{xx}[0] & \cdots & r_{xx}[N-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[N-1] & r_{xx}[N-2] & \cdots & r_{xx}[0] \end{bmatrix} \begin{bmatrix} h[1] \\ h[2] \\ \vdots \\ h[N] \end{bmatrix} = \begin{bmatrix} r_{xx}[l] \\ r_{xx}[l+1] \\ \vdots \\ r_{xx}[l+N-1] \end{bmatrix}$$

A computationally efficient solution for solving the equations is the *Levinson recursion* which solves the equations recursively to avoid resolving them for each value of n. The special case for $l = 1$, the one-step linear predictor, covers two important cases in signal processing:

- the values of $-h[n]$ are termed the *linear prediction coefficients* which are used extensively in speech coding, and
- the resulting equations are identical to the *Yule-Walker equations* used to solve the AR filter parameters of an AR(N) process.

12.3 Sequential LMMSE

We can derive a recursive re-estimation of $\hat{\theta}[n]$ (i.e. $\hat{\theta}$ based on n data samples) from:

- $\hat{\theta}[n-1]$, the estimate based on $n-1$ data samples,
- $x[n]$, the n^{th} data sample, and
- $\hat{x}[n|n-1]$, an estimate of $x[n]$ based on $n-1$ data samples.

Using a vector space analogy where we define the following inner product:

$$(x, y) = E(xy)$$

the procedure is as follows:

1. From the previous iteration we have the estimate, $\hat{\theta}[n-1]$, or we have an initial estimate $\hat{\theta}[0]$. We can consider $\hat{\theta}[n-1]$ as the true value of θ projected on the subspace spanned by $\{x[0], , x[1], \ldots, x[n-1]\}$.
2. We find the LMMSE estimator of $x[n]$ based on the previous $n-1$ samples, that is the one-step linear predictor, $\hat{x}[n|n-1]$, or we have an initial estimate, $\hat{x}[0|-1]$. We can consider $\hat{x}[n|n-1]$ as the true value $x[n]$ projected on the subspace spanned by $\{x[0], , x[1], \ldots, x[n-1]\}$.
3. We form the *innovation* $x[n] - \hat{x}[n|n-1]$ which we can consider as being orthogonal to the space spanned by $\{x[0], , x[1], \ldots, x[n-1]\}$ and representing the direction of correction based exclusively on the new data sample $x[n]$.
4. We calculate the *correction gain* $K[n]$ by the normalised projection of θ on the innovation $x[n] - \hat{x}[n|n-1]$, that is:
$$K[n] = \frac{E[\theta(x[n] - \hat{x}[n|n-1])]}{E[(x[n] - \hat{x}[n|n-1])^2]}$$
5. Finally we update the estimator by adding the correction:
$$\hat{\theta}[n] = \hat{\theta}[n-1] + K[n](x[n] - \hat{x}[n|n-1])$$

12.3.1 Sequential LMMSE for the Bayesian Linear Model

For the vector parameter-scalar data form of the Bayesian linear model:

$$x[n] = \mathbf{h}^T\boldsymbol{\theta} + w[n]$$

we can derive the sequential vector LMMSE:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}[n] &= \hat{\boldsymbol{\theta}}[n-1] + \mathbf{K}[n](x[n] - \mathbf{h}^T[n]\hat{\boldsymbol{\theta}}[n-1]) \\
\mathbf{K}[n] &= \frac{\mathbf{M}[n-1]\mathbf{h}[n]}{\sigma_n^2 + \mathbf{h}^T[n]\mathbf{M}[n-1]\mathbf{h}[n]} \\
\mathbf{M}[n] &= (\mathbf{I} - \mathbf{K}[n]\mathbf{h}^T[n])\mathbf{M}[n-1]
\end{aligned}$$

where $\mathbf{M}[n]$ is the error covariance estimate:

$$\mathbf{M}[n] = E[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}[n])(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}[n])^T]$$

13 Kalman Filters

The Kalman filter can be viewed as an extension of the sequential LMMSE to the case where the parameter estimate, or state, *varies with time* in a known but stochastic way and unknown initialisation. We consider the vector state-vector observation case and consider the unknown state (to be estimated) at time n, $\mathbf{s}[n]$, to vary according to the Gauss-Markov signal model *process* equation:

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

where $\mathbf{s}[n]$ is the $p \times 1$ state vector with unknown initialisation $\mathbf{s}[-1] = \mathbf{N}(\boldsymbol{\mu}_s, \mathbf{C}_s)$, $\mathbf{u}[n] = \mathbf{N}(\mathbf{0}, \mathbf{Q})$ is the WGN $r \times 1$ driving noise vector and $\mathbf{A}, \mathbf{B}$ are known $p \times p$ and $p \times r$ matrices. The observed data, $\mathbf{x}[n]$, is then assumed a linear function of the state vector with the following *observation* equation:

$$\mathbf{x}[n] = \mathbf{H}[n]\mathbf{s}[n] + \mathbf{w}[n] \quad n \geq 0$$

where $\mathbf{x}[n]$ is the $M \times 1$ observation vector, $\mathbf{w}[n] = \mathbf{N}(\mathbf{0}, \mathbf{R})$ is the WGN $M \times 1$ observation noise sequence and $\mathbf{H}$ is a known $M \times p$ matrix.

We wish to estimate $\mathbf{s}[n]$ based on the observations $\mathbf{X}[n] = \left[\mathbf{x}^T[0]\, \mathbf{x}^T[1] \ldots \mathbf{x}^T[n]\right]$. Our criterion of optimality is the Bayesian MMSE:

$$E[(\mathbf{s}[n] - \hat{\mathbf{s}}[n|n])^2]$$

where $\hat{\mathbf{s}}[n|n] = E(\mathbf{s}[n]|\mathbf{X}[n])$ is the estimate of $\mathbf{s}[n]$ based on the observation sequence $\mathbf{X}[n]$. We define the *innovation* sequence:

$$\tilde{\mathbf{x}}[n] = \mathbf{x}[n] - \hat{\mathbf{x}}[n|n-1]$$

where $\hat{\mathbf{x}}[n|n-1] = E(\mathbf{x}[n]|\mathbf{X}[n-1])$ is the predictor of $\mathbf{x}[n]$ based on the observation sequence $\mathbf{X}[n-1]$. We can consider $\hat{\mathbf{s}}[n|n]$ as being derived from two separate components:

$$\hat{\mathbf{s}}[n|n] = E(\mathbf{s}[n]|\mathbf{X}[n-1]) + E(\mathbf{s}[n]|\tilde{\mathbf{x}}[n])$$

where $E(\mathbf{s}[n]|\mathbf{X}[n-1])$ is the "state prediction" and $E(\mathbf{s}[n]|\tilde{\mathbf{x}}[n])$ is the "innovation correction". By definition $\hat{\mathbf{s}}[n|n-1] = E(\mathbf{s}[n]|\mathbf{X}[n-1])$ and from the process equation we can show that:

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

From our sequential LMMSE results we also have that :

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

and:

$$\mathbf{K}[n] = E(\mathbf{s}[n]\tilde{\mathbf{x}}^T[n])E(\tilde{\mathbf{x}}[n]\tilde{\mathbf{x}}^T[n])^{-1} = \mathbf{M}[n|n-1]\mathbf{H}^T[n](\mathbf{H}[n]\mathbf{M}[n|n-1]\mathbf{H}^T[n]+\mathbf{R})^{-1}$$

where:

$$\mathbf{M}[n|n-1] = E\left[(\mathbf{s}[n] - \hat{\mathbf{s}}[n|n-1])(\mathbf{s}[n] - \hat{\mathbf{s}}[n|n-1])^T\right]$$

is the state error covariance predictor at time n based on the observation sequence $\mathbf{X}[n-1]$. From the process equation we can show that:

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

and by appropriate substitution we can derive the following expression for the state error covariance estimate at time n:

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

Kalman Filter Procedure

Prediction

$$\begin{aligned} \hat{\mathbf{s}}[n|n-1] &= \mathbf{A}\hat{\mathbf{s}}[n-1|n-1] \\ \mathbf{M}[n|n-1] &= \mathbf{A}\mathbf{M}[n-1|n-1]\mathbf{A}^T + \mathbf{B}\mathbf{Q}\mathbf{B}^T \end{aligned}$$

Gain

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

Correction

$$\begin{aligned} \hat{\mathbf{s}}[n|n] &= \hat{\mathbf{s}}[n|n-1] + \mathbf{K}[n](\mathbf{x}[n] - \hat{\mathbf{x}}[n|n-1]) \\ \mathbf{M}[n|n] &= (\mathbf{I} - \mathbf{K}[n]\mathbf{H}[n])\mathbf{M}[n|n-1] \end{aligned}$$

14 MAIN REFERENCE

Steven M. Kay, "*Fundamentals of Statistical Signal Processing: Estimation Theory*", Prentice-Hall, 1993