Estimation and Detection

Lecture 7: General Bayesian Estimators and Linear MMSE Estimators

Previous Lecture

Classical approach

Bayesian approach

Knowledge on how \( A \) is chosen!
Previous Lecture - Minimum Mean Square Error Estimator

We know that \( p(x, \theta) = p(\theta|x)p(x) \), so that

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

Since \( p(x) \geq 0 \) for all \( x \), we have to minimize the inner integral for each \( x \).

Problem:
\[
\min_{\theta} \int (\hat{\theta} - \theta)^2 p(\theta|x) d\theta
\]

Solution: mean of posterior pdf of \( \theta \):
\[
\hat{\theta} = E(\theta|x) = \int \theta p(\theta|x) d\theta
\]

Previous Lecture - The Bayesian approach

Bayesian approach:

- Parameter to be estimated is assumed to be a realisation of a random variable \( \theta \).
- Our knowledge about the unknown parameter is summarised in the density

\[
p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)}.
\]

- The Bayesian MSE minimises the MSE over all realisations of \( \theta \) and \( x \).
- Bayesian MMSE: \( \hat{\theta} = E(\theta|x) = \int \theta p(\theta|x) d\theta \).
- The MMSE depends thus on both the prior knowledge (via \( p(\theta) \)) and the data (via \( p(x;\theta) \)).
- Remember: Classical approaches only use \( p(x;\theta) \).
Previous Lecture

For a linear Gaussian model:

\[ x = H\theta + w, \quad w \sim \mathcal{N}(0, C) \]

with \( \theta \) a random vector with distribution \( \mathcal{N}(\mu_\theta, C_\theta) \).

In that case \( p(\theta|x) \) is also Gaussian with mean and covariance matrix

\[
\begin{align*}
E(\theta|x) &= \mu_\theta + C_\theta H^T (HC_\theta H^T + C)^{-1} (x - H\mu_\theta) \\
C_{\theta|x} &= C_\theta - C_\theta H^T (HC_\theta H^T + C)^{-1} H C_\theta
\end{align*}
\]

Today:

- MMSE estimators for vector processes.
- Extension of the "Bayesian Philosophy" to other cost functions.
- Linear MMSE estimators.
MMSE Estimators – Estimation of a vector (1)

- In the previous lecture it was shown that the MMSE estimator is given by $E[\theta | x]$. 
- What if $\theta$ is a $p \times 1$ vector?

For a linear Gaussian model we already know:

$$x = H\theta + w, \quad w \sim \mathcal{N}(0, C)$$

with $\theta$ a random vector with distribution $\mathcal{N}(\mu_\theta, C_\theta)$.

In that case $p(\theta | x)$ is also Gaussian with mean and covariance matrix

$$E(\theta | x) = \mu_\theta + C_\theta H^T (HC_\theta H^T + C)^{-1} (x - H\mu_\theta)$$

$$C_{\theta|x} = C_\theta - C_\theta H^T (HC_\theta H^T + C)^{-1} HC_\theta$$


MMSE Estimators – Estimation of a vector (2)

- In the previous lecture it was shown that the MMSE estimator is given by $E[\theta | x]$. 
- What if $\theta$ is a $p \times 1$ vector?

In general: In the case of a vector we can write for the $i$th element:

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

$$= \int \theta_i \left[ \int \cdots \int p(\theta | x) d\theta_1 \cdots d\theta_{i-1} d\theta_{i+1} \cdots d\theta_p \right] d\theta_i$$

$$= \int \theta_i p(\theta | x) d\theta.$$

So in vector form we have:

$$\hat{\theta} = \int \theta p(\theta | x) d\theta = E[\theta | x].$$
MMSE Estimators – Estimation of a vector (3)

For the vector case, we thus have

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

$$\text{Bmse}(\hat{\theta}) = \int [C_{\theta|x}]_{ii} p(x) dx$$

$$C_{\theta|x} = E_{\theta|x}[(\theta - E(\theta|x))(\theta - E(\theta|x))^T]$$

MMSE Estimators – Affine Transformations

Imagine that we want to estimate a linear function of $\theta$, e.g.,

$$\alpha = A\theta + b$$

Using the linearity of $E[\cdot]$ it is easy to show that the MMSE estimator for $\alpha$ is then given by

$$\hat{\alpha} = A\hat{\theta} + b$$

where $\hat{\theta}$ is the MMSE estimator for $\theta$. 
General Bayesian Estimators – Risk Functions

The Bayesian MSE defined as

$$B_{\text{mse}}(\hat{\theta}) = E[(\hat{\theta}(x) - \theta)^2] = \int \int (\hat{\theta}(x) - \theta)^2 p(x, \theta) dx d\theta.$$ 

Notice that

- $(\hat{\theta}(x) - \theta)^2$ can be seen as an example of a cost function, say $C(\epsilon)$ with $\epsilon = \hat{\theta}(x) - \theta$ the error.

Although mathematically nice, there is no need to restrict ourselves to such quadratic cost functions.

General Bayesian Estimators – Risk Functions

We can more generally minimize the Bayes risk $R$:

$$R = E[C(\epsilon)]$$

where $\epsilon = \theta - \hat{\theta}$ and $C$ is a cost function that can take many different forms:

$$C(\epsilon) = \epsilon^2, \quad C(\epsilon) = |\epsilon|, \quad C(\epsilon) = \begin{cases} 0 & |\epsilon| \leq \delta \\ 1 & |\epsilon| > \delta \end{cases}, \text{ with } \delta \to 0$$

As for the MMSE, we now have to minimize

$$g(\tilde{\theta}) = \int C(\theta - \tilde{\theta}) p(\theta|x) d\theta.$$ 

Taking $C(\epsilon) = \epsilon^2$ we thus get $\hat{\theta} = E[\theta|x]$. 
General Bayesian Estimators – Risk Functions

For the "proportional" cost function $C(\varepsilon) = |\varepsilon|$, this leads to the median of $p(\theta|x)$:

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

Differentiation with respect to $\hat{\theta}$, setting the result to zero we get:

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

that is, $\hat{\theta}$ is exactly the Median.

---

General Bayesian Estimators – Risk Functions

The "hit-or-miss" cost function:

$$C(\varepsilon) = \begin{cases} 
0 & |\varepsilon| \leq \delta \\
1 & |\varepsilon| > \delta 
\end{cases}, \text{ with } \delta \to 0$$

With this cost function

$$g(\hat{\theta}) = \int C(\varepsilon)p(\theta|x)d\theta.$$ 

becomes

$$g(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}-\delta} 1p(\theta|x)d\theta + \int_{\hat{\theta}+\delta}^{\infty} 1p(\theta|x)d\theta.$$ 

Minimizing $g(\hat{\theta})$ is obtained by maximizing

$$\int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|x)d\theta$$

As $\delta$ is arbitrarily small, this implies $\hat{\theta}$ corresponds to the location of the maximum of $p(\theta|x)$. 
General Bayesian Estimators – Summary…

- For the squared-error, absolute error and hit-or-miss cost functions, the estimators that minimise the Bayes risk are the mean, median, and mode (maximum) of the a posteriori pdf.
- The latter one is therefore often called the maximum a posteriori (MAP) estimator.
- For symmetric a posteriori pdfs (like the Gaussian) these three are identical.

The MAP estimator

The MAP estimator corresponds thus to

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

Using Bayes’ rule, this is thus identical to

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

and

$$\hat{\theta} = \arg\max_\theta \log(p(x|\theta)) + \log(p(\theta)).$$

Often, the MAP is much easier to calculate than the MMSE, as no integration is involved.
The MAP estimator – Example 1

\[ x[n] = A + w[n], \quad n = 0, \ldots, N-1, \quad w[n] \sim \mathcal{N}(0, \sigma^2), \quad A \sim \mathcal{U}[-A_0, A_0] \]

The posterior PDF was earlier shown to be given by

\[
p(A|x) = \begin{cases} 
\frac{1}{\sqrt{2\pi \sigma^2/N}} \exp \left[ -\frac{1}{2\sigma^2/N} (A - \bar{x})^2 \right], & |A| \leq A_0 \\
\int_{-A_0}^{A_0} \frac{1}{\sqrt{2\pi \sigma^2/N}} \exp \left[ -\frac{1}{2\sigma^2/N} (A - \bar{x})^2 \right] dA, & |A| > A_0
\end{cases}
\]

It is then clear that the MAP estimator is given by

\[
\hat{A} = \begin{cases} 
-A_0, & \bar{x} < -A_0 \\
\bar{x}, & -A_0 \leq \bar{x} \leq A_0 \\
A_0, & \bar{x} > A_0
\end{cases}
\]

Estimate depends on mean of conditional Gaussian \( p(x|A) \)

The MAP estimator – Example 2

Assume that

\[
p(x[n]|\theta) = \begin{cases} 
\theta \exp(-\theta x[n]), & x[n] > 0 \\
0, & x[n] < 0
\end{cases}
\]

and

\[
p(\theta) = \begin{cases} 
\lambda \exp(-\lambda \theta), & \theta > 0 \\
0, & \theta < 0
\end{cases}
\]

\[
g(\theta) = \ln p(x|\theta) + \ln p(\theta)
\]

\[
\frac{dg(\theta)}{d\theta} = \frac{N}{\theta} - N\bar{x} - \lambda
\]

leading to \( \hat{\theta} = \frac{1}{\bar{x} + \lambda} \)
The MAP estimator => Bayesian MLE

The MAP estimator corresponds to

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

Notice that if $p(\theta)$ is uniform (and maximum of $p(x|\theta)$ falls in the uniform range), the MAP becomes

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

which is essentially the Bayesian MLE.

Generally, it holds that if $N \to \infty$, the pdf $p(x|\theta)$ becomes dominant over $p(\theta)$ and the MAP becomes thus identical to the Bayesian MLE.

The MAP estimator – the Vector Case

- The direct extension of the MAP estimator would be

$$\hat{\theta}_i = \arg \max_{\theta_i} p(\theta_i|x) = \arg \max_{\theta_i} \int \ldots \int p(\theta|x)d\theta_1 \ldots d\theta_{i-1}d\theta_{i+1} \ldots d\theta_p.$$  

Note that this approach still requires integration. However, it minimizes a similar cost function as in the scalar case: namely $R_i = E = [C(\theta_i - \hat{\theta}_i)]$ for each $i$.

- An easier yet different approach is the so-called vector MAP

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

This estimator actually minimizes the following average "hit-or-miss" function:

$$C(\epsilon) = \begin{cases} 
0 & \|\epsilon\| \leq \delta \\
1 & \|\epsilon\| > \delta 
\end{cases}, \text{ with } \epsilon = \theta - \hat{\theta} \text{ and } \delta \to 0.$$
The Linear MMSE Estimator

- Under jointly Gaussian assumptions, the Optimal Bayesian estimators are easy to determine.

- However, in general, the optimal Bayesian estimators are often difficult to determine in closed form:
  - MMSE estimator: Generally involves multidimensional integration.
  - MAP estimator: Generally involves multidimensional maximization.

A useful solution is to constrain the MMSE estimator to be linear.
The Linear MMSE Estimator

Remember that for the classical (deterministic) estimators, we introduced the BLUE as a simpler alternative to the MVU estimator (having a larger MSE, though).

For the Bayesian estimators we can do something similar: Let's constrain our estimator to be of the form \( \hat{\theta} = a^T x \).

Assuming that \( x \) and \( \theta \) have zero mean for simplicity we can then write the Bayesian MSE as

\[
\text{Bmse}(\hat{\theta}) = E[(a^T x - \theta)^2] = a^T E(xx^T) a - 2a^T E(x\theta) + E(\theta^2)
\]

\[\begin{align*}
&= a^T C_{x,x} a - 2a^T C_{x,\theta} + \text{var}(\theta)
\end{align*}\]

Setting the derivative with respect to \( a \) to zero, we obtain

\[ a = C_{x,x}^{-1} C_{x,\theta} \]

The LMMSE estimator is therefore given by

\[ \hat{\theta} = C_{x,\theta}^T C_{x,x}^{-1} x \]

The Bayesian MSE is given by

\[ \text{Bmse}(\hat{\theta}) = \text{var}(\theta) - C_{x,\theta}^T C_{x,x}^{-1} C_{x,\theta} \]
The Linear MMSE Estimator – Example 1 (1)

Remember from the previous lecture:

Example: estimation of the mean

\[ x[n] = A + w[n], \quad n = 0, \ldots, N - 1, \quad w[n] \sim \mathcal{N}(0, \sigma^2), \quad A \sim U(-A_0, A_0) \]

Conditional pdf: \( p(x[n] | A) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x[n] - A)^2\right) \)

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

and \( p(A) = \begin{cases} 
\frac{1}{2A_0}, & |A| \leq A_0 \\
0, & |A| > A_0 
\end{cases} \)

The Linear MMSE Estimator – Example 1 (2)

Altogether, the a posteriori pdf becomes:

\[ p(A|x) = \begin{cases} 
\frac{1}{2A_0(2\pi\sigma^2)^{N/2}} \exp\left[ \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right], & |A| \leq A_0 \\
\int_{-A_0}^{A_0} \frac{1}{2A_0(2\pi\sigma^2)^{N/2}} \exp\left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right] dA, & |A| > A_0 
\end{cases} \]
The Linear MMSE Estimator – Example 1 (3)

This can be simplified to

\[ p(A|x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma^2/N} \exp\left[ -\frac{1}{2\sigma^2/N} (A - \bar{x})^2 \right], & |A| \leq A_0 \\ \int_{-A_0}^{A_0} \frac{1}{\sqrt{2\pi}\sigma^2/N} \exp\left[ -\frac{1}{2\sigma^2/N} (A - \bar{x})^2 \right] dA, & |A| > A_0 \end{cases} \]

This looks like a truncated Gaussian, and the final MMSE estimate is given by

\[ \hat{A} = \int_{-\infty}^{\infty} A p(A|x) dA = \int_{-A_0}^{A_0} \frac{1}{\sqrt{2\pi}\sigma^2/N} \exp\left[ -\frac{1}{2\sigma^2/N} (A - \bar{x})^2 \right] dA \]

The Linear MMSE Estimator – Example 1 (4)

Let us now look at the linear MMSE:

\[ C_{xx} = E[xx^T] = E[A^2]11^T + \sigma^2 I \]

and

\[ C_{ax} = E[Ax^T] = E[A^2]1^T. \]

The linear MMSE is then given by

\[ \hat{A} = E[A^2]1^T \left( E[A^2]11^T + \sigma^2 I \right)^{-1} x. \]

Using the matrix inversion lemma (MIL)

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

it then follows that

\[ \hat{A} = \frac{\sigma^2 A}{\sigma^2 A + \frac{\bar{x}}{\sigma^2}} \]

Remember that this is identical to the (non-linear) MMSE estimator assuming \( \theta \) is Gaussian!
Geometrical Interpretation – Orthogonality Principle

Remember that to get the linear MMSE, we minimized the Bmse:

\[ \frac{\partial}{\partial a^T} E[(\hat{\theta} - \theta)^2] = 2E[x(\hat{\theta} - \theta)] = 0 \]

The LMMSE leads to the following orthogonality condition:

\[ E[x(\hat{\theta} - \theta)] = 0 \iff (\hat{\theta} - \theta) \perp \text{stat } x \]

Geometrical Interpretation – Orthogonality Principle

The orthogonality principle can be used directly to find the LMMSE:

\[ E[x(\hat{\theta} - \theta)] = 0 \Rightarrow E[x(x^T a - \theta)] = 0 \]
\[ \Rightarrow E(xx^T) a - E(x\theta) = 0 \]
\[ \Rightarrow C_{xx} a - c_{x\theta} = 0 \]
\[ \Rightarrow a = C_{xx}^{-1} c_{x\theta} \]

And the Bayesian MSE also follows from this:

\[ \text{Bmse}(\hat{\theta}) = E[(\theta - a^T x)^2] = E[(\theta - a^T x)(\theta - a^T x)] = E[(\theta - a^T x)\theta] - \frac{E[(\theta - a^T x)x^T]}{\sigma^2} a \]
\[ = \text{var}(\theta) - a^T C_{xx}^{-1} c_{x\theta} = \text{var}(\theta) - C_{x\theta}^T C_{xx}^{-1} c_{x\theta} \]
General Linear Gaussian Model (from previous lecture)

The above example can be generalized to the linear Gaussian model:

\[ x = H\theta + w, \quad w \sim \mathcal{N}(0, C) \]

with \( \theta \) a random vector with distribution \( \mathcal{N}(\mu_{\theta}, C_{\theta}) \).

In that case \( p(\theta|x) \) is also Gaussian with mean and covariance matrix

\[
E(\theta|x) = \mu_{\theta} + C_{\theta}H^T(HC_{\theta}H^T + C)^{-1}(x - H\mu_{\theta}) \\
C_{\theta|x} = C_{\theta} - C_{\theta}H^T(HC_{\theta}H^T + C)^{-1}HC_{\theta}
\]

Linear MMSE – Linear Data Model

For the linear model

\[ x = H\theta + w \]

with \( w \) is zero-mean with covariance matrix \( C \) and with \( \theta \) a random vector with mean \( \mu_{\theta} \) and covariance \( C_{\theta} \).

The LMMSE estimator is

\[
\hat{\theta} = \mu_{\theta} + C_{\theta}H^T(HC_{\theta}H^T + C)^{-1}(x - H\mu_{\theta}) = \mu_{\theta} + (H^T C^{-1}H + C_{\theta}^{-1})^{-1}H^T C^{-1}(x - H\mu_{\theta})
\]

where the last equality is again due to the matrix inversion lemma (MIL).

Remark: Notice the identical form as for the MMSE estimator for the linear Gaussian model presented during last lecture!

The LMMSE estimator is equivalent to the MMSE estimator when the noise and the unknown parameter are Gaussian.
<table>
<thead>
<tr>
<th></th>
<th>linear model</th>
<th>linear Gaussian model</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta )</td>
<td>deterministic</td>
<td>deterministic</td>
</tr>
<tr>
<td>MVU</td>
<td>( \hat{\theta} ) ( \doteq ) ( (H^TH)^{-1}H^T ) ( x )</td>
<td>( \hat{\theta} ) ( \doteq ) ( (H^TH)^{-1}H^T ) ( C ) ( x )</td>
</tr>
<tr>
<td>BLUE</td>
<td>( \hat{\theta} ) ( \doteq ) ( (H^TH)^{-1}H^T ) ( C ) ( x )</td>
<td>same as linear model</td>
</tr>
<tr>
<td>MLE</td>
<td>?</td>
<td>( \hat{\theta} ) ( \doteq ) ( (H^TH)^{-1}H^T ) ( C ) ( x )</td>
</tr>
<tr>
<td>LSE</td>
<td>?</td>
<td>same as linear model</td>
</tr>
<tr>
<td></td>
<td>( \theta ) stochastic with cov. ( C_{\theta} )</td>
<td>( \theta ) Gaussian with cov. ( C_{\theta} )</td>
</tr>
<tr>
<td>MMSE</td>
<td>?</td>
<td>( \theta = \mu_{\theta} + (H^T ) ( C^{-1} ) ( H + C_{\theta}^{-1} ) ( H^T ) ( C^{-1} ) ( C ) ( x ) ( - ) ( H \mu_{\theta} )</td>
</tr>
<tr>
<td>LMMSE</td>
<td>( \hat{\theta} = \mu_{\theta} + (H^T ) ( C^{-1} ) ( H + C_{\theta}^{-1} ) ( H^T ) ( C^{-1} ) ( C ) ( x ) ( - ) ( H \mu_{\theta} )</td>
<td>same as linear model</td>
</tr>
</tbody>
</table>