# Stochastic Processes

## Session 16

## Minimum Mean Square Error Estimation

1) Recap Estimation, Errors, Bias & MSE
2) MMSE
3) LMMSE
4) Vector LMMSE

# Estimation Problem

Given $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$ estimate $\theta$.

$\boxed{\text{Exercise: State a few estimation problems relevant to your field of study}}$

# Estimator

**Def:** An estimator is a function $\hat{\theta} = g(X)$ of data $X$ used to guess the value of some unknown entity $\theta$.

Recall the following defs.:

- Estimation error : $\theta - \hat{\theta}$
- Bias (mean error) : $E[\theta - \hat{\theta}]$
- Mean Squared Error (MSE) :

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

> Exercise: Suppose we ignore the data $X$ and just guess on the mean: $\hat{\theta} = E[\theta] = \mu_\theta$.
> What is the bias and the MSE of this estimator?

# Bayesian estimation

- We assume $\theta$ to be a random variable with a priori pdf $p(\theta)$
- $\theta$ is jointly distributed with the data $X$ according to $p(\theta, X)$.

- A priori knowledge may be obtained from past ~~experements~~ or information on the context.

- In general the a priori information may be full (i.e. $p(\theta)$) or partial (e.g. moments of $p(\theta)$).

Exercise: Can you phrase your estimation problem from before as a Bayesian Estimation problem? What a priori information can be used in this case?

# Minimum MSE estimator (MMSE)

<u>Def</u>.: The MMSE of $\theta$ given the data $X$ is the function $\hat{\Theta} = g(X)$ such that the MSE $E[(\theta - \hat{\Theta})^2]$ is minimum.

Notice that $E[(\theta-\hat{\Theta})^2] = E\left[E[(\theta-\hat{\Theta})^2 \mid X]\right]$.
Therefore, $\hat{\Theta}$ should be chosen to minimize the conditional MSE $E[(\theta-\hat{\Theta})^2 \mid X]$ for <u>each</u> $X$.

We find $\hat{\Theta}$ by "differentiating and equating to zero". The chain-rule of differentiation simplifies this task

# Derivation of the MMSE

It suffices to minimize the conditional MSE wrt $\hat{\theta}$:

$$\frac{\partial E[(\theta-\hat{\theta})^2 | X]}{\partial \hat{\theta}} = E\left[ \frac{\partial (\theta-\hat{\theta})^2}{\partial \hat{\theta}} \,\middle|\, X \right]$$

$$= E\left[ 2(\theta-\hat{\theta}) \cdot \frac{\partial(\theta-\hat{\theta})}{\partial \hat{\theta}} \,\middle|\, X \right]$$

$$= -2\, E[\theta - \hat{\theta} | X]$$

$$= -2\, E[\theta | X] + 2\hat{\theta}$$

(Since $\hat{\theta} = g(X)$ is "deterministic" cond. on $X$)

Equating to zero, we obtain:

$$\boxed{\hat{\theta} = E[\theta | X]}$$

(Double diff. shows that this critical point is indeed minimum)

## Properties of the MMSE:

- The MMSE is unbiased:
  $E[\theta-\hat{\theta}] = E[\theta - E[\theta|X]] = E[\theta]-E[\theta]=0$

- The MMSE has the lowest MSE among <u>all</u> estimators.

- The MMSE fulfills the "orthogonality principle", i.e the error is uncorr. to <u>any</u> function of the data:

$$E\left[(\theta-\hat{\theta})\cdot h(X)\right] = 0$$

<u>Proof</u>: $E[(\theta-\hat{\theta})\cdot h(X)] = E[\,E[(\theta-\hat{\theta})h(X)\,|\,X]\,]$

$$= E[\,\underbrace{E[\theta-\hat{\theta}\,|X]}_{=0 \text{ (se deriv of MMSE)}}\cdot h(X)] = 0.$$

## Remarks to the MMSE

- Although simple to derive, the MMSE is impractical unless $E[\theta|X]$ can be computed

- Analytic solutions are available only in (very) special cases.

- Alternatively $E[\theta|X]$ may be computed numerically (e.g. via Monte Carlo simulation), or using other approximation techniques.

- Instead of the MMSE, a linear MMSE (LMMSE) may be applied.

# Linear MMSE (LMMSE)

<u>Def</u>: The LMMSE is the estimator of the form

$$\hat{\theta} = h_0 + \sum_{n=1}^{N} h_n X_n = h_0 + h^T X$$

where $h_0, h_1, \ldots, h_N$ are chosen to minimize $E[(\theta - \hat{\theta})^2]$.

<u>Remark</u>: The estimator is actually not linear, but "affine" in cases where $h_0 \neq 0$. However, the name is used widely in the literature so we stick to the convention.

- The coeficients $h_0, \ldots, h_N$ can be obtained directly by minimizing the MSE.

# Coeffts of the LMMSE

As with the MMSE, we make use of the chain-rule to minimize $E[(\theta-\hat{\theta})^2|X]$:

- $$\frac{\partial E[(\theta-\hat{\theta})^2|X]}{\partial h_0} = \underbrace{\frac{\partial\hat{\theta}}{\partial h_0}}_{=1}\cdot\underbrace{\frac{\partial E[(\theta-\hat{\theta})^2|X]}{\partial\hat{\theta}}}_{-2E[\theta-\hat{\theta}|X]}$$

$$= -2\,E[\theta|X] + 2(h_0 + h^T X)$$

Equating to zero and taking the expectation w.r.t. X gives

$$0 = -E[\theta] + h_0 + h^T E[X]. \Rightarrow \underline{\underline{h_0 = E[\theta] - h^T E[X]}}.$$

- $$\frac{\partial E[(\theta-\hat{\theta})^2|X]}{\partial h^T} = \underbrace{\frac{\partial\hat{\theta}}{\partial h^T}}_{=X}\cdot\underbrace{\frac{\partial E[(\theta-\hat{\theta})^2|X]}{\partial\hat{\theta}}}_{-2E[\theta-\hat{\theta}|X]}$$

$$= -2\,E[(\theta-\hat{\theta})\cdot X|X]$$

Equating to zero and taking the expectation w.r.t. X gives

$$E[(\theta-(h_0+h^T X))\cdot X] = 0$$

$\rightarrow$

Inserting for $h_0$ $\Rightarrow$

$$E\Big[\underbrace{\Big(\theta - E[\theta] - h^T\big(X - E[X]\big)\Big)}\cdot X\Big] = 0$$

we notice that this is zero mean, and thus we see that

$$E\Big[\Big(\theta - E[\theta] - h^T\big(X - E[X]\big)\Big)\cdot\big(X - E[X]\big)\Big] = 0$$

$$\Updownarrow$$

$$\underline{\underline{C}}_{\theta X} - h^T \underline{\underline{C}}_{XX} = 0$$

$$\Updownarrow$$

$$h^T = \underline{\underline{C}}_{\theta X}\,\underline{\underline{C}}_{XX}^{-1}$$

$$\Updownarrow$$

$$\underline{\underline{h = \underline{\underline{C}}_{XX}^{-1}\, C_{X\theta}}} \qquad \left(C_{\theta X}^T = C_{X\theta}\right)$$

The LMMSE now reads

$$\boxed{\begin{aligned}\hat{\theta} = E[\theta] - C_{\theta X} C_{XX}^{-1} E[X] \\ + C_{\theta X} C_{XX}^{-1} X.\end{aligned}}$$

Alternative form:

$$\boxed{\hat{\theta} = E[\theta] + C_{\theta X} C_{XX}^{-1} (X - E[X])}$$

Remarks

- The LMMSE requires a priori knowledge of $E[\theta]$, $E[X]$, $C_{\theta X}$ and $C_{XX}$.
- Simple to compute provided $C_{XX}$ is invertible

# Properties of the LMMSE

- The LMMSE is unbiased
  (se derivation of $h_0$)

- The LMMSE minimizes the MSE among linear (or rather affine) estimators.
  - Nonlinear estimators may perform better!

- The LMMSE fulfills the "orthogonality principle":

  $$E[(\theta-\hat{\theta})f(X)]=0$$

  where $f(x)$: is <u>any affine</u> function of $X$, i.e $f(x)=f_0+\underline{f}^T\underline{X}$.

<u>Proof:</u> $E[(\theta-\hat{\theta})f(x)] = \underbrace{E[(\theta-\hat{\theta})f_0]}_{=0,\ \hat{\theta}\text{ unbiased.}} + E[(\theta-\hat{\theta})f^T x]$

$$= f^T \underbrace{E[(\theta-\hat{\theta})X]}_{=0 \text{ see deriv of } \underline{h}.}$$

# Residual MSE of LMMSE

The residual MSE can be computed by application of the orthogonality principle. (o.p.):

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

$$= E\left[(\theta-\hat{\theta})\,\theta\right] \qquad (\text{o.p.})$$

$$= E\left[(\theta-\hat{\theta})\left(\theta-E[\theta]\right)\right] \quad (\hat{\theta}\ \text{unbiased})$$

$$= E\left[\left((\theta-E[\theta]) - C_{\theta x}C_{xx}^{-1}(x-E[\theta])\right)(\theta-E[\theta])\right]$$

$$(\text{alt. form f. } \hat{\theta})$$

$$= \underline{\underline{Var(\theta) - C_{\theta x}C_{xx}^{-1}C_{x\theta}.}}$$

- So the MSE of the LMMSE is known upfront!

# LMMSE for multiple variables

("Vector LMMSE")

The unknown parameter $\underline{\theta}$ is now a vector with $K$ entries:

$$\underline{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}$$

- We seek the linear (or affine) estimator $\hat{\underline{\theta}}$ which minimizes the total MSE:

$$MSE_{tot} = MSE_1 + MSE_2 + \dots + MSE_K$$

where $MSE_k = E[(\theta_k - \hat{\theta}_k)^2]$.

- $MSE_{tot}$ is minimized if and only if $MSE_1, \dots, MSE_K$ are all minimum.

$\Rightarrow$ We simply have to use parallel scalar LMMSEs!

# Expression for "Vector LMMSE"

We stack scalar LMMSEs in a vector:

$$\hat{\underline{\theta}} = \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_K \end{bmatrix} = \begin{bmatrix} E[\theta_1] + C_{\theta_1 X} C_{xx}^{-1}(X - E[X]) \\ \vdots \\ E[\theta_K] + C_{\theta_K X} C_{xx}^{-1}(X - E[X]) \end{bmatrix}$$

$$= E[\theta] + \underline{\underline{C}}_{\theta X} C_{XX}^{-1} (X - E[X])$$

The total MSE is then:

$$MSE_{tot} = \sum_{k=1}^{K} MSE_k$$

$$= \sum_{k=1}^{K} \left( Var(\theta_k) - C_{\theta_k X} C_{xx}^{-1} C_{X\theta_k} \right)$$

$$= trace \left( C_{\theta\theta} - C_{\theta X} C_{XX}^{-1} C_{X\theta} \right).$$