Estimation: parts of Chapters 10-12

Bayesian Estimation

Lifang Feng
lffeng@ustb.edu.cn

北京科技大学
University of Science and Technology Beijing

Spring 2023

Summary

In *classical* estimation theory, the parameter θ to be estimated is considered to be a fixed but unknown constant. No prior knowledge of its value is assumed. In certain scenarios it may make sense to assume a prior distribution $p(\theta)$ on the unknown parameters. In contrast to the classical philosophy, in Bayesian estimation we consider the parameter to be an outcome of a random variable with some known prior distribution.

Prior knowledge should be exploited to obtain a better estimator.

Bayesian philosophy

Bayesianism

- If we play a coin toss game, we do not know the result, question: Head? Tail?

- A. Head B. Tail
- C. Head with 50% and Tail with 50%

- In classical statistics, A or B
- In Bayesian statistics, C is permited

Bayes Theorem

- True Bayesians actually consider conditional probabilities more than joint probabilities. It is easy to define P(A|B) without reference to the joint probability P(A,B). To see this note that we can rearrange the conditional probability formula to get:
- P(A|B) P(B) = P(A,B)

by symmetry:

- P(B|A) P(A) = P(A,B)
- It follows that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- which is the so-called ***Bayes Rule***.

An Example

- Suppose that we are interested in diagnosing cancer in patients who visit a chest clinic:
  - *A* represent the event "Person has cancer"
  - *B* represent the event "Person is a smoker"
  - The probability of the prior event P(*A*)=0.1 on the basis of past data (10% of patients entering the clinic turn out to have cancer).
- What is the probability of the posterior event P(*A*|*B*)?

An Example

- A: cancer B: smoker
- It is likely to know P(*B*) by considering the percentage of patients who smoke – suppose P(*B*)=0.5.
- It is also likely to know P(*B*|*A*) by checking from our record the proportion of smokers among those diagnosed- Suppose P(*B*|*A*)=0.8.
  - P(*A*|*B*) = (0.8 * 0.1)/0.5 = 0.16
- Thus, in the light of *evidence* that the person is a smoker we revise our prior probability from 0.1 to a posterior probability of 0.16.

Bayesian Reasoning

ASSUMPTIONS

1% of women aged forty who participate in a routine screening have breast cancer

80% of women with breast cancer will get positive tests

9.6% of women without breast cancer will also get positive tests

EVIDENCE

A woman in this age group had a positive test in a routine screening

PROBLEM

What's the probability that she has breast cancer?

Bayesian Reasoning

ASSUMPTIONS

100 out of 10,000 women aged forty who participate in a routine screening have breast cancer

80 of every 100 women with breast cancer will get positive tests

950 out of 9,900 women without breast cancer will also get positive tests

PROBLEM

If 10,000 women in this age group undergo a routine screening, about what fraction of women with positive tests will actually have breast cancer?

Compact Formulation

C = cancer present, T = positive test
p(A|B) = probability of A, given B, ~ = not

PRIORS

PRIOR PROBABILITY
p(C) = 1%

CONDITIONAL PROBABILITIES
p(T|C) = 80%
p(T|~C) = 9.6%

POSTERIOR PROBABILITY (or REVISED PROBABILITY)
p(C|T) = ?

Bayesian Reasoning

Before the screening:
100 women with breast cancer
9,900 women without breast cancer

After the screening:
A = 80 women with breast cancer and positive test
B = 20 women with breast cancer and negative test
C = 950 women without breast cancer and positive test
D = 8,950 women without breast cancer and negative test

Proportion of cancer patients with positive results, within the group of ALL patients with positive results:

A/(A+C) = 80/(80+950) = 80/1030 = 0.078 = 7.8%

Bayesian Reasoning

Prior Probabilities:

100/10,000 = 1/100 = 1% = p(C)

9,900/10,000 = 99/100 = 99% = p(~C)

Conditional Probabilities:

A = 80/10,000 = (80/100)*(1/100) = p(T|C)*p(C) = 0.008

B = 20/10,000 = (20/100)*(1/100) = p(~T|C)*p(C) = 0.002

C = 950/10,000 = (9.6/100)*(99/100) = p(T|~C)*p(~C) = 0.095

D = 8,950/10,000 = (90.4/100)*(99/100) = p(~T|~C) *p(~C) = 0.895

Rate of cancer patients with positive results, within the group of ALL patients with positive results:

A/(A+C) = 0.008/(0.008+0.095) = 0.008/0.103 = 0.078 = 7.8%

-----> Bayes' theorem

$$p(C|T) = \frac{p(T|C)*p(C)}{P(T|C)*p(C) + p(T|\sim C)*p(\sim C)}$$

A

A + C

-----> Bayes' theorem

$$p(A|X) = \frac{p(X|A)*p(A)}{P(X|A)*p(A) + p(X|\sim A)*p(\sim A)}$$

Given some phenomenon A that we want to investigate, and an observation X that is evidence about A, we can update the original probability of A, given the new evidence X.

Bayes' Theorem for a given parameter θ

$$p(\theta \mid data) = p(data \mid \theta)\, p(\theta) / p(data)$$

1/P (data) is basically a normalizing constant

Posterior $\propto$ likelihood x prior

The prior is the probability of the parameter and represents what was thought before seeing the data.

The likelihood is the probability of the data given the parameter and represents the data now available.

The posterior represents what is thought given both prior information and the data just seen.

It relates the conditional density of a parameter (posterior probability) with its unconditional density (prior, since depends on information present before the experiment).

Bayesian Theory

Bayesian statistical decision

(1) Get priori and conditional probabilities

(2) Get posteriori probabilities based on Bayes rule

(3) Decision according to the values of posteriori probabilities

How to get priori probabilities?

- right priori probability→ estimator with good performance
- wrong priori probability→ estimator with bad performance
  - (1) Objective approach: history summary
  - (2) Subjective approach: personal understanding of the unknown parameter
  - (3) Equal unknown: uniform distribution—also called Bayes assumption
  - (4) conjugate distribution: the priori and posterior have the same distribution
  - (5) Jeffreys rule: θ and $g(\theta)$ has the same priori within a certain rule

Priori cannot be proved to be accurate or not in fact

Bayesian Risk

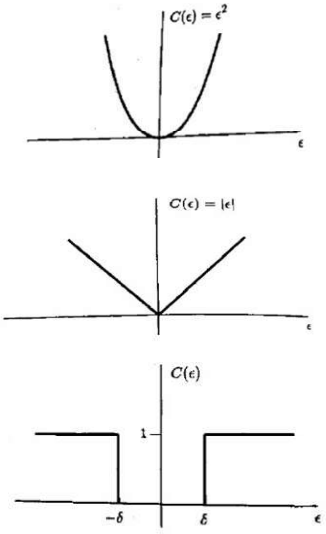
- We wish to estimate a parameter(s) θ from observations **x**.
- Estimation error: $\gamma=\theta-\hat{\theta}$

We assign a cost function $C(\gamma)$. Some examples of cost functions are:

(1) Quadratic cost $C(\gamma)= \gamma^2=(\theta- \hat{\theta})^2$.

(2) Absolute error cost $C(\gamma)=| \gamma |$.

(3) Hit-or-miss cost $C(\gamma)=1$ for $| \gamma |>\delta$
and $C(\gamma)=0$ for $| \gamma |< \delta$



Bayesian Risk

The goal in designing a Bayesian estimator $\hat{\theta}$ is the minimization of the Bayesian risk, defined as follows:

$$\begin{aligned}\mathcal{R} &= E[C(\epsilon)] \\ &= \int\int C(\epsilon)p(\mathbf{x};\theta)d\mathbf{x}d\theta \\ &= \int\int C(\theta-\hat{\theta})p(\mathbf{x};\theta)d\mathbf{x}d\theta \\ &= \int\left[\int C(\theta-\hat{\theta})p(\theta|\mathbf{x})d\theta\right]p(\mathbf{x})d\mathbf{x} \quad \text{by Bayes' rule}\end{aligned}$$

In general then, we want to minimize the inner integral for each value of the observed data **x**.

Quadratic error cost

Quadratic cost $C(\epsilon) = \epsilon^2 = (\theta - \hat{\theta})^2$:

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

The minimum mean squared error (MMSE) estimator is the *mean* of the posterior density $p(\theta|\mathbf{x})$.

We call it Minimal Mean Square Error (MMSE) estimator.

$$C(\hat{\theta}\,|\,\boldsymbol{x}) = \int_{-\infty}^{\infty} C(\theta-\hat{\theta})p(\theta\,|\,\boldsymbol{x})d\theta = \int_{-\infty}^{\infty} (\theta-\hat{\theta})^2\, p(\theta\,|\,\boldsymbol{x})d\theta$$

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial\hat{\theta}}\Big[(\theta-\hat{\theta})^2\Big] p(\theta\,|\,\boldsymbol{x})d\theta = \int_{-\infty}^{\infty} -2(\theta-\hat{\theta})p(\theta\,|\,\boldsymbol{x})d\theta$$

$$= -2\int_{-\infty}^{\infty} \theta p(\theta\,|\,\boldsymbol{x})d\theta + 2\hat{\theta}\int_{-\infty}^{\infty} p(\theta\,|\,\boldsymbol{x})d\theta$$

$$= -2\int_{-\infty}^{\infty} \theta p(\theta\,|\,\boldsymbol{x})d\theta + 2\hat{\theta}$$

Absolute error cost

Absolute error cost $C(\epsilon) = |\epsilon|$:

$$\hat{\theta} \text{ is such that } \int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{x})d\theta = \int_{\hat{\theta}}^{\infty} p(\theta|\mathbf{x})d\theta = 1/2.$$

The estimator is the *median* of the posterior distribution $p(\theta|\mathbf{x})$.

We call it conditional median estimator

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

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

$$\frac{\partial}{\partial\hat{\theta}} C(\hat{\theta}\,|\,\boldsymbol{x}) = \frac{\partial}{\partial\hat{\theta}}\int_{-\infty}^{\hat{\theta}} (\hat{\theta}-\theta)p(\theta\,|\,\boldsymbol{x})d\theta + \frac{\partial}{\partial\hat{\theta}}\int_{\hat{\theta}}^{\infty} (\theta-\hat{\theta})p(\theta\,|\,\boldsymbol{x})d\theta$$

$$= \int_{-\infty}^{\hat{\theta}} \frac{\partial}{\partial\hat{\theta}}\Big[(\hat{\theta}-\theta)p(\theta\,|\,\boldsymbol{x})\Big]d\theta + \int_{\hat{\theta}}^{\infty} \frac{\partial}{\partial\hat{\theta}}\Big[(\theta-\hat{\theta})p(\theta\,|\,\boldsymbol{x})\Big]d\theta$$

$$= \int_{-\infty}^{\hat{\theta}} p(\theta\,|\,\boldsymbol{x})d\theta - \int_{\hat{\theta}}^{\infty} p(\theta\,|\,\boldsymbol{x})d\theta \qquad (3$$

利用莱布尼兹（Leibnitz）准则

$$\frac{\partial}{\partial u}\int_{g_1(u)}^{g_2(u)} h(u,v)dv$$

$$= \int_{g_1(u)}^{g_2(u)} \frac{\partial h(u,v)}{\partial u}dv + \frac{dg_2(u)}{du}h(u,g_2(u)) - \frac{dg_1(u)}{du}h(u,g_1(u)) \qquad (3.$$

Hit-or-miss cost (MAP)

Hit-or-miss cost $C(\epsilon) = 1$ for $|\epsilon| > \delta$ and $C(\epsilon) = 0$ for $|\epsilon| < \delta$:

$$\hat{\theta} \text{ maximizes } p(\theta|\mathbf{x}).$$

The estimator is the *mode* of the posterior distribution $p(\theta|\mathbf{x})$ (also the Maximum a Posteriori Estimator (MAP) estimator).

We call it Maximum A Posteriori (MAP) estimator.

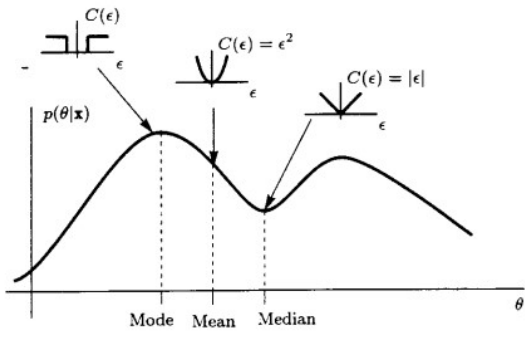
$$C(\hat{\theta}\,|\,\boldsymbol{x}) = \int_{-\infty}^{\hat{\theta}-\delta} p(\theta\,|\,\boldsymbol{x})d\theta + \int_{\hat{\theta}+\delta}^{\infty} p(\theta\,|\,\boldsymbol{x})d\theta \qquad C(\hat{\theta}\,|\,\boldsymbol{x}) = 1 - \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta\,|\,\boldsymbol{x})d\theta$$

MAP properties

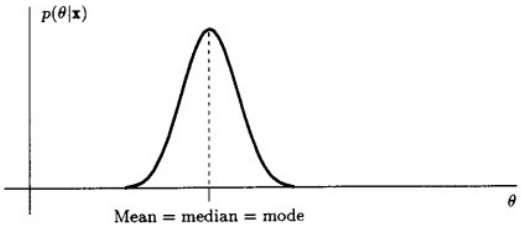
Some interesting properties of the MAP estimator:

- As $N \to \infty$ the MAP becomes the Bayesian ML estimator.
- If $\mathbf{x}$ and θ are jointly Gaussian then the MAP = MMSE estimator.
- The MAP commutes over invertible linear transformations, but in contrast to ML estimators, does not commute over non-linear functions in general.

For those three cost functions, quadratic error cost get the mean of the posterior density; absolute error cost get the median of the posterior density; hit-or-miss cost get the mode of the posteriori density.



(a) General posterior PDF



(b) Gaussian posterior PDF

MMSE→BMSE

We now focus on determining Bayesian MMSE estimators, i.e. estimators that minimize the Bayesian risk under the quadratic error/cost function. We saw that

$$\hat{\theta}_{MMSE} = E[\theta|\mathbf{x}].$$

What is the minimum mean squared error achieved by this estimator (called the $BMSE(\hat{\theta})$)? Unlike in classical estimation theory, this does not depend on the value of θ, which has been averaged out.

Classical MSE
$MSE(\hat{\theta}) = \int (\hat{\theta} - \theta)^2 p(\mathbf{x};\theta) d\mathbf{x}$
Depends on θ!

Bayesian MSE
$BMSE(\hat{\theta}) = E[(\theta - \hat{\theta})^2] = \int\int (\theta - \hat{\theta})^2 p(\mathbf{x};\theta) d\mathbf{x} d\theta$
Dependence on θ is averaged out!

MMSE estimators

The $BMSE(\hat{\theta})$ may be determined using the variance of the conditional distribution $p(\theta|\mathbf{x})$ as follows:

$$\begin{aligned} BMSE(\hat{\theta}) &= E[(\theta - \hat{\theta})^2] \\ &= \int\int (\theta - \hat{\theta})^2 p(\theta, \mathbf{x}) d\theta d\mathbf{x} \\ &= \int\int (\theta - E[\theta|\mathbf{x}])^2 \, p(\theta|\mathbf{x}) \, d\theta \; p(\mathbf{x}) d\mathbf{x} \\ &= \int var(\theta|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Example of MMSE

We observe x which relates to the parameter to be estimated, A, as follows:

$$p(A,x) = \begin{cases} 6A & 0 \le A \le x, 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the MMSE estimator of A.

Bayesian estimation on Gaussian priors/noise

We now state (proofs in Appendix A of Ch.10) some useful theorems on jointly Gaussian distributions.
(Theorem 10.1) If x, y are jointly Gaussian with mean vector $[E[x]\, E[y]]^T$ and covariance matrix

$$\mathbf{C} = \begin{bmatrix} var(x) & cov(x,y) \\ cov(y,x) & var(y) \end{bmatrix}$$

then the conditional pdf $p(y|x)$ is also Gaussian with mean and variance

$$E(y|x) = E(y) + \frac{cov(x,y)}{var(x)}(x - E(x))$$

$$var(y|x) = var(y) - \frac{cov(x,y)^2}{var(x)}.$$

This is crucial to know!

Bayesian estimation on Gaussian priors/noise

(Theorem 10.2) 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

$$E(\mathbf{y}|\mathbf{x}) = E(\mathbf{y}) + \mathbf{C_{yx}}\mathbf{C_{xx}^{-1}}(\mathbf{x} - E(\mathbf{x}))$$
$$\mathbf{C_{y|x}} = \mathbf{C_{yy}} - \mathbf{C_{yx}}\mathbf{C_{xx}^{-1}}\mathbf{C_{xy}}$$

This is crucial to know!

Bayesian linear model

The theorems relating the conditional pdfs of jointly Gaussian random variables/vectors to their joint distribution come into play in the Bayesian linear model, which assumes that the data $\mathbf{x}$ is related to the unknown parameters θ as follows:

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

$\mathbf{x}$: $N \times 1$ data
$\mathbf{H}$: $N \times p$ known matrix, not necessarily invertible!
θ : $p \times 1$ random parameters $\sim \mathcal{N}(\mu_\theta, \mathbf{C}_\theta)$
$\mathbf{w}$: $N \times 1$ noise $\sim \mathcal{N}(\mathbf{0}, \mathbf{C_w})$ independent of θ

$$E(\mathbf{y}|\mathbf{x}) = E(\mathbf{y}) + \mathbf{C_{yx}}\mathbf{C_{xx}^{-1}}(\mathbf{x} - E(\mathbf{x}))$$
$$\mathbf{C_{y|x}} = \mathbf{C_{yy}} - \mathbf{C_{yx}}\mathbf{C_{xx}^{-1}}\mathbf{C_{xy}}$$

This resembles the linear model for classical estimation we saw in Ch.4 except that now we have a prior on the parameters θ. In the Bayesian linear model this prior is assumed to be Gaussian.

Bayesian linear model

(Theorem 10.3) Under the Bayesian linear model described above θ and $\mathbf{x}$ are jointly Gaussian with a posterior pdf $p(\theta|\mathbf{x})$ that is also Gaussian with mean and covariance:

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

There are alternative forms of this mean and variance which may be more useful, depending on the application, given by:

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

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

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

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

$$\begin{aligned} \mathbf{C}_{yx} &= E\left[(\mathbf{y} - E(\mathbf{y}))(\mathbf{x} - E(\mathbf{x}))^T\right] \\ &= E\left[(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) + \mathbf{w})^T\right] \\ &= E\left[(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})(\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}))^T\right] \\ &= \mathbf{C}_{\boldsymbol{\theta}}\mathbf{H}^T. \end{aligned}$$

Examples: MMSE in Gaussian noise

Find the MMSE estimator of A when given $x(n) = A + w(n)$, for $n = 0, 1, \cdots N-1$, $A \sim \mathcal{N}(\mu_A, \sigma_A^2), \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

$$\mathbf{x} = \mathbf{1}A + \mathbf{w}.$$

$$E(A|\mathbf{x}) = \mu_A + \sigma_A^2\mathbf{1}^T(\mathbf{1}\sigma_A^2\mathbf{1}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{x} - \mathbf{1}\mu_A).$$

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

$$\begin{aligned} E(A|\mathbf{x}) &= \mu_A + \frac{\sigma_A^2}{\sigma^2}\mathbf{1}^T\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{N + \frac{\sigma^2}{\sigma_A^2}}\right)(\mathbf{x} - \mathbf{1}\mu_A) \\ &= \mu_A + \frac{\sigma_A^2}{\sigma^2}\left(\mathbf{1}^T - \frac{N}{N + \frac{\sigma^2}{\sigma_A^2}}\mathbf{1}^T\right)(\mathbf{x} - \mathbf{1}\mu_A) \\ &= \mu_A + \frac{\sigma_A^2}{\sigma^2}\left(1 - \frac{N}{N + \frac{\sigma^2}{\sigma_A^2}}\right)(N\bar{x} - N\mu_A) \\ &= \mu_A + \frac{N}{N + \frac{\sigma^2}{\sigma_A^2}}(\bar{x} - \mu_A) \\ &= \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}(\bar{x} - \mu_A). \end{aligned}$$

Properties of MMSE

- The MMSE estimator for the Bayesian model becomes the MVUE estimator for the classical linear model as the prior distribution becomes uninformative.
- The MMSE commutes over affine transformations.
- When $\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}]^{\mathbf{T}}$ and θ are jointly Gaussian, the MMSE is additive for independent data sets $\mathbf{x_1}, \mathbf{x_2}$, i.e.

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

Maximum a Posteriori (MAP) estimation

Maximum a posteriori (MAP) estimators are another commonly used form of Bayesian estimator. As we saw earlier, the MAP estimator is the mode (max) of the posterior distribution $p(\theta|\mathbf{x})$, and may be easier to compute that say the MMSE as no integration is needed, only maximization. The MAP estimator, in its various forms, is given by:

$$\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|\mathbf{x}) \\
&= \arg\max_{\theta} \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} \\
&= \arg\max_{\theta} p(\mathbf{x}|\theta)p(\theta) \\
&= \arg\max_{\theta} \left[\ln p(\mathbf{x}|\theta) + \ln p(\theta)\right]
\end{aligned}$$

Example

Determine the estimates of θ that minimize the expected Bayesian cost function $C(\epsilon) = C(\theta - \hat{\theta})$ for

$$C(\epsilon) = \begin{cases} 1 \text{ if } |\epsilon| > \delta, \delta > 0 \\ 0 \text{ if } |\epsilon| < \delta \end{cases},$$

where the joint distribution on $\mathbf{x}$ and θ is given by:

$$p(x[n]|\theta) = \begin{cases} \theta \exp(-\theta x[n]) & x[n] > 0 \\ 0 & x[n] < 0 \end{cases} \qquad p(\mathbf{x}|\theta) = \prod_{n=0}^{N-1} p(x[n]|\theta)$$

$$p(\theta) = \begin{cases} \lambda \exp(-\lambda\theta) & \theta > 0 \\ 0 & \theta < 0. \end{cases}$$

Then, the MAP estimator is found by maximizing

$$\begin{aligned} g(\theta) &= \ln p(\mathbf{x}|\theta) + \ln p(\theta) \\ &= \ln \left[\theta^N \exp\left(-\theta \sum_{n=0}^{N-1} x[n] \right) \right] + \ln \left[\lambda \exp(-\lambda\theta) \right] \\ &= N \ln \theta - N\theta\bar{x} + \ln \lambda - \lambda\theta \end{aligned}$$

for $\theta > 0$. Differentiating with respect to θ produces

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

and setting it equal to zero yields the MAP estimator

$$\hat{\theta} = \frac{1}{\bar{x} + \frac{\lambda}{N}}.$$

Note that as $N \to \infty$, $\hat{\theta} \to 1/\bar{x}$. Also, recall that $E(x[n]|\theta) = 1/\theta$ (see Example 9.2), so that

$$\theta = \frac{1}{E(x[n]|\theta)}$$

confirming the reasonableness of the MAP estimator. Also, if $\lambda \to 0$ so that the prior PDF is nearly uniform, we obtain the estimator $1/\bar{x}$. In fact, this is the *Bayesian MLE* (the estimator obtained by maximizing $p(\mathbf{x}|\theta)$) since as $\lambda \to 0$ we have the situation in Figure 11.3 in which the conditional PDF *dominates* the prior PDF. The maximum of g is then unaffected by the prior PDF. ◇

MAP properties

Some interesting properties of the MAP estimator:

- As $N \to \infty$ the MAP becomes the Bayesian ML estimator.
- If $\mathbf{x}$ and θ are jointly Gaussian then the MAP = MMSE estimator.
- The MAP commutes over invertible linear transformations, but in contrast to ML estimators, does not commute over non-linear functions in general.

Linear MMSE (LMMSE)

When the data and parameters to be estimated are jointly Gaussian the MAP and MMSE estimators are easy to obtain and coincide. However, in general optimal Bayesian estimators may be difficult to obtain: both MMSE and MAP requires multi-dimensional integration. We are thus motivated to look at a sub-optimal estimator which is in general easier to obtain and does not make any assumptions about the pdfs of the data or the noise. The sub-optimal estimator is the Linear Minimum Mean-Squared Error (LMMSE) estimator, which seeks to minimize the mean squared error as before, but constraints the estimator to be linear. In many ways, the LMMSE is the Bayesian version of the Best Linear Unbiased Estimator (BLUE) in classical estimation theory.

LMMSE definition

When estimating θ from data samples $x(n), n = 0, 1, \cdots, N-1$, the LMMSE estimator is of the form

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

and is chosen to minimize the Bayesian MMSE

$$BMMSE(\hat{\theta}) = E[(\theta - \hat{\theta})^2].$$

We need to determine the constants a_n that minimize the BMMSE. This may be done in 2 ways: algebraically (sub-in $\hat{\theta}$ and minimize the BMMSE by taking partial derivatives and setting them to zero) or geometrically, by viewing the parameters and data as elements of a suitably defined vector space.

Geometric LMMSE

The random variables of interest are $\theta, x(0), x(1), \cdots, x(N-1)$, which we think of as elements of a vector space whose inner product is defined as $(x, y) = E(xy)$, making the length of a vector $||x|| = \sqrt{E(x, x)} = \sqrt{E(x^2)}$. You can check that this has all the properties of an inner product (and vector space). For now we assume θ and $\mathbf{x}$ are all zero mean. Two vectors are orthogonal if $E(x, y) = 0$, i.e. if they are uncorrelated.

Taking this geometric view, the BMMSE is just:

$$\begin{aligned} BMMSE(\hat{\theta}) &= E[(\theta - \hat{\theta})^2] \\ &= E[(\theta - \sum_{n=0}^{N-1} a_n x(n))^2] \\ &= ||\theta - \sum_{n=0}^{N-1} a_n x(n)||^2 \\ &= ||\epsilon||^2, \end{aligned}$$

the squared length of the error vector ϵ.

Geometric LMMSE

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

$$\sum_{n=0}^{N-1} a_m E(x[m]x[n]) = E(\theta x[n]) \qquad \mathbf{C}_{xx}\mathbf{a} = \mathbf{C}_{x\theta}$$

Using our intuition/knowledge of geometry, we realize that this is minimized if the error vector ϵ is orthogonal to the subspace spanned by $\{x(0), x(1), \cdots, x(N-1)\}$. This leads to $E(\theta - \hat{\theta})x(n)) = 0$ for all $n = 0, 1, \cdots N-1$. Substituting in $\hat{\theta}$ and writing everything in matrix form, we see that the optimal $\mathbf{a} = [a_0, a_1, \cdots, a_{N-1}]^T$ is given by $\epsilon \perp x[0], x[1], \ldots, x[N-1]$

$$\mathbf{a} = \mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{x}\theta} \quad \Rightarrow \quad \begin{array}{l} \hat{\theta}_{LMMSE} = \mathbf{a}^{\mathbf{T}}\mathbf{x} = \mathbf{C}_{\theta\mathbf{x}}\mathbf{C}_{\mathbf{xx}}^{-1}\,\mathbf{x} \\ BMMSE(\hat{\theta}) = \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta\mathbf{x}}\mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{x}\theta} \end{array}.$$

We can extend the LMMSE to the more general case when θ and $\mathbf{x}$ are not zero mean, and when they are vectors as follows:

General vector LMMSE: $\hat{\theta} = E(\theta) + \mathbf{C}_{\theta\mathbf{x}}\mathbf{C}_{\mathbf{xx}}^{-1}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$

Bayesian mean squared error matrix: $\mathbf{M}_{\hat{\theta}} = \mathbf{C}_{\theta\theta} - \mathbf{C}_{\theta\mathbf{x}}\mathbf{C}_{\mathbf{xx}}^{-1}\mathbf{C}_{\mathbf{x}\theta}$.

What we made use of in this derivation was the *orthogonality principle: in estimating the realization of a random variable by a linear combination of data samples, the optimal estimator is obtained when the error is orthogonal to each data sample.*

LMMSE example

Find the LMMSE estimates of $\hat{A}$ if for $n = 0, 1, 2 \cdots, N-1$,

$$x(n) = A + w(n), \qquad A \sim \mathcal{U}[-A_0, A_0],\ \mathbf{w} \sim \mathcal{N}(0, \sigma^2 I).$$

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

$$\begin{aligned} \hat{A} &= \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{x} \\ &= \sigma_A^2\mathbf{1}^T(\sigma_A^2\mathbf{1}\mathbf{1}^T + \sigma^2\mathbf{I})^{-1}\mathbf{x} \end{aligned}$$

$$\hat{A} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}}\bar{x}.$$

Since $\sigma_A^2 = E(A^2) = (2A_0)^2/12 = A_0^2/3$, the LMMSE estimator of A is

$$\hat{A} = \frac{\frac{A_0^2}{3}}{\frac{A_0^2}{3} + \frac{\sigma^2}{N}}\bar{x}.$$

As opposed to the original MMSE estimator which required integration, we have obtained the LMMSE estimator in closed form. Also, note that we did not really need to know that A was uniformly distributed but only its mean and variance, or that $w[n]$ was Gaussian but only that it is white and its variance. Likewise, independence of A and $\mathbf{w}$ was not required, only that they were uncorrelated. In general, all that is required to determine the LMMSE estimator are the first two moments of $p(\mathbf{x}, \theta)$ or

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

However, we must realize that the LMMSE of (12.9) will be *suboptimal* since it has been *constrained to be linear*. The optimal estimator for this problem is given by (10.9).

LMMSE properties

- We only need the first and second order moment of the parameter and the data: $E(\theta), E(\mathbf{x})$, first order statistics and $\mathbf{C}_{\theta\theta}, \mathbf{C}_{\mathbf{x}\theta}\mathbf{C}_{\theta\mathbf{x}}, \mathbf{C}_{\mathbf{xx}}$, the covariance matrices of the parameters and data.
- The LMMSE yields the same form of the estimator as the Gaussian MMSE estimator except that we do not assume the noise or prior parameter pdf is Gaussian.
- The LMMSE is sub-optimal except when $E[\theta|\mathbf{x}]$ happens to be linear (which is the case for jointly Gaussian $\mathbf{x}, \theta$). This means that under the jointly Gaussian assumption, LMMSE = MMSE.
- LMMSE, like the MMSE estimator, commutes over affine transformations.
- If $x(0)$ and $x(1)$ are orthogonal (uncorrelated) observations, then $\hat{\theta}_{LMMSE} = \hat{\theta}_{LMMSE}(x(0)) + \hat{\theta}_{LMMSE}(x(1))$.

LMMSE properties

- (Theorem 12.1: Bayesian Gauss-Markov Theorem) If the data are described by the Bayesian linear model form

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

where $\mathbf{x}$ is an $N \times 1$ data vector, $\mathbf{H}$ is a known $N \times p$ observation matrix, θ is a $p \times 1$ random vector of parameters whose realization is to be estimated and has mean $E(\theta)$ and covariance matrix $\mathbf{C}_{\theta\theta}$ and $\mathbf{w}$ is an $N \times 1$ random vector with mean zero and covariance matrix $\mathbf{C}_{\mathbf{w}}$ that is uncorrelated with θ (and the joint distribution of $p(\theta, \mathbf{w})$ is otherwise arbitrary), then the LMMSE estimator of θ is

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

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

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

whose diagonal elements yield the minimum BMSE of the individual parameters θ_i.

Sequential LMMSE

Let's consider the estimation of A from data $x(n) = A + w(n)$, where $w(n)$ is white Gaussian noise of variance σ^2 and A is uniform on $[-A_0, A_0]$. Say we have N data points $x(0), x(1), \cdots x(N-1)$ from which we form an LMMSE estimate of the parameter at time $N-1$,

$$\hat{A}(N-1) = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \overline{x}.$$

Now we obtain a new data point $x(N)$ and wish to update our estimate of A accordingly. For each new data point thereafter we wish to update the estimator; this procedure is called sequential LMMSE. We consider only the simple example of estimating the DC level in noise, a general form of the sequential LMMSE for the Bayesian linear model form is given in the textbook pg. 397-399.

Algebraic approach Geometric approach

Sequential LMMSE: algebraic

Since we know the general form of the LMMSE estimator given N data points, we try to write $\hat{A}(N)$ as a linear combination of the previous estimate $\hat{A}(N-1)$ and some linear function of the new data point $x(N)$. Just playing around with the equations, we can see that

$$\begin{aligned}\hat{A}(N) &= \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N+1}} \frac{1}{N+1} \sum_{n=0}^{N} x(n) \\ &= \cdots \\ &= \hat{A}(N-1) + \frac{\sigma_A^2}{(N+1)\sigma_A^2 + \sigma^2}(x(N) - \hat{A}(N-1)).\end{aligned}$$

Recall that

$$BMMSE(\hat{A}(N-1)) = \frac{\sigma^2}{N}\left(\frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/N}\right) = \frac{\sigma_A^2\sigma^2}{N\sigma_A^2 + \sigma^2}.$$

If we set $K(N) = \frac{\sigma_A^2}{(N+1)\sigma_A^2+\sigma^2} = \frac{BMSE(\hat{A}(N-1)}{BMSE(\hat{A}(N-1))+\sigma^2}$ then we see that the new estimate is just the old estimate plus some scaled (by $K(N)$) version of the prediction error $x(N) - \hat{A}(N-1)$. In summary, the LMMSE iterates the steps:

- *Estimator update:*

$$\hat{A}(N) = \hat{A}(N-1) + K(N)(x(N) - \hat{A}(N-1)),$$

  where

$$K(N) = \frac{BMSE(\hat{A}(N-1))}{BMSE(\hat{A}(N-1)) + \sigma^2}.$$

- *Minimum MSE update:*

$$BMSE(\hat{A}(N)) = (1 - K(N))BMSE(\hat{A}(N-1)).$$

Sequential LMMSE: geometric approach

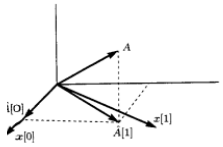
An alternative approach is to use the vector space analogy. There, the main idea when looking at new data $x(N)$ is to look at what information is "new" for the estimate, i.e. what information is orthogonal to the previous data. This direction/new information is called the *innovation*. Sequential LMMSE basically builds up a sequence of innovations, which are all orthogonal to each other. This makes estimation easy: the LMMSE estimator is the sum of the projections of the parameter θ onto the orthogonal set of innovations. More concretely,

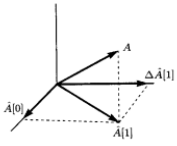
1. Find the LMMSE estimator of A based on $x(0)$, yielding $\hat{A}(0)$
2. Find the LMMSE estimator of $x(1)$ based on $x(0)$, yielding $\hat{x}(1|0)$.
3. Find the innovation of the new datum, $x(1) - \hat{x}(1|0)$
4. Add to $\hat{A}(0)$ the LMMSE estimator of A based on the innovation, yielding $\hat{A}(1)$
5. Repeat the process

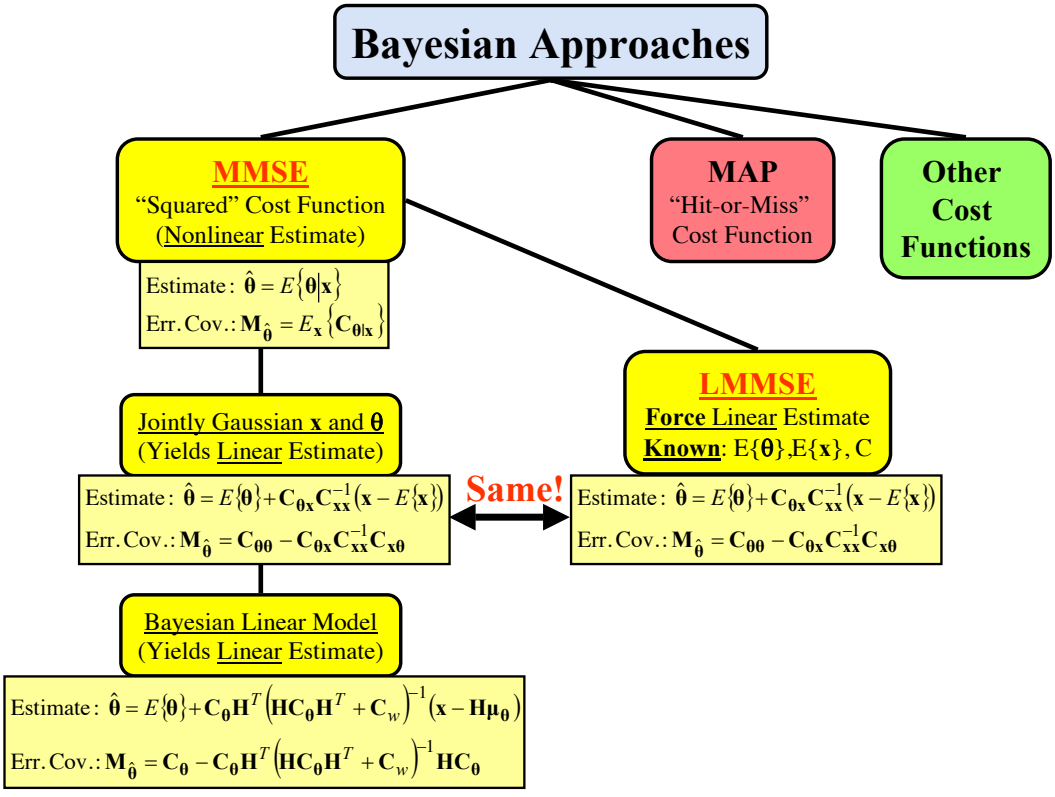
We are generating a set of uncorrelated or orthogonal random variables called the *innovations:*

$$\{x(0), x(1) - \hat{x}(1|0), x(2) - \hat{x}(2|1,0), x(3) - \hat{x}(3|2,1,0), \cdots\}$$

This is/ressembles the Gram-Schmidt process for finding an orthogonal basis for the span of a set of generally non-orthogonal vectors. The same equations as before for $\hat{A}(N)$ and $K(N)$ may be found.







Signal process example

We look at two examples of sequential estimation:

- Wiener filtering: filtering, smoothing and prediction *(wide-sense stationary signals)* in sequential LMMSE framework
- Kalman filtering: generalization of Wiener filtering to *(non-stationary signals)*, i.e. sequential MMSE estimator of a signal in noise, where signal characterized by a dynamical model (i.e. tracking)

Signal model: $x[n] = s[n] + w[n], \quad n = 0, 1, 2, \cdots N-1$ where noise $w[n]$ is WSS, zero-mean with $\mathbf{C}_{ww} = \mathbf{R}_{ww}$

Problem: Process $x[n]$ using a *linear* filter to obtain a "de-noised" version of the signal that has *minimum mean square error* relative to the desired signal $s[n]$.

Wiener filtering

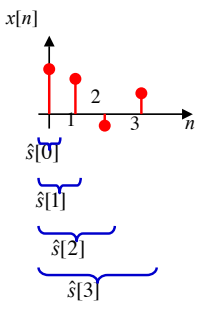
Signal model: $x[n] = s[n] + w[n], \quad n = 0, 1, 2, \cdots N-1$ where:

- noise $w[n]$ is WSS, zero-mean with $\mathbf{C}_{ww} = \mathbf{R}_{ww}$
- desired signal $s[n]$ is WSS, zero mean with $\mathbf{C}_{ss} = \mathbf{R}_{ss}$
- observed noisy signal $x[n]$ is WSS, zero mean with $\mathbf{C}_{xx} = R_{xx}$

Filtering

Given: $x[0], x[1], \ldots, x[n]$

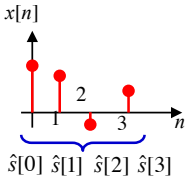
Find: $\hat{s}[n]$



Smoothing

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

Find: $\hat{s}[0], \hat{s}[1], \ldots, \hat{s}[N-1]$

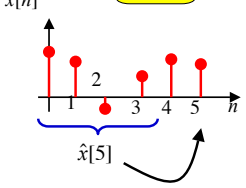


Prediction

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

+ $\hat{s}[N+l], \quad l > 0$

Note!!



Wiener filtering

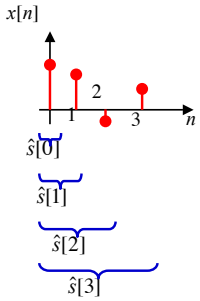
Solve all three using general LMMSE estimation

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

Filtering

Given: $x[0], x[1], \ldots, x[n]$

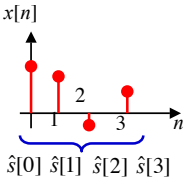
Find: $\hat{s}[n]$



Smoothing

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

Find: $\hat{s}[0], \hat{s}[1], \ldots, \hat{s}[N-1]$

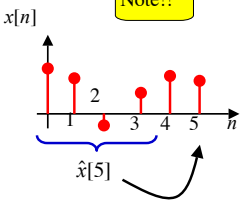


Prediction

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

Find: $\hat{s}[N-1+l]$, $l > 0$

Note!!

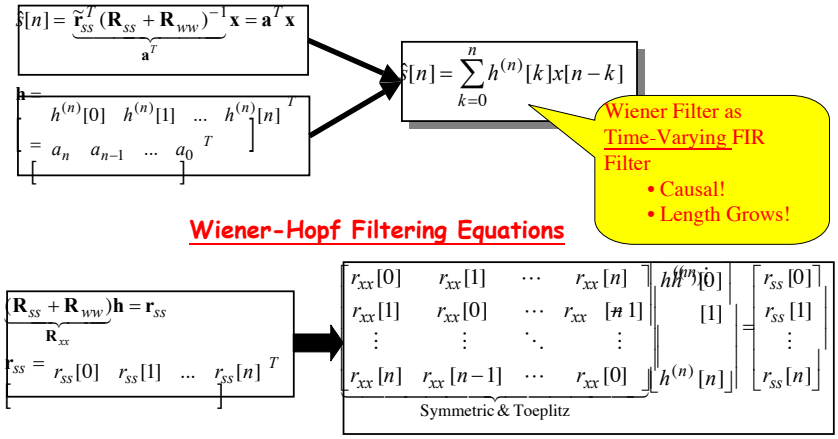


Wiener filtering

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

Filtering	**Smoothing**	**Prediction**
$\theta = s[n]$ (scalar)	$\boldsymbol{\theta} = \mathbf{s}$ (vector)	$\theta = x[N-1+l]$ (scalar)
		***x* not *s*!**
$\mathbf{C}_{\theta\mathbf{x}} = E\{s[n]\mathbf{x}^T\}$	$\mathbf{C}_{\boldsymbol{\theta}\mathbf{x}} = E\{\mathbf{s}\mathbf{x}^T\}$	$\mathbf{C}_{\theta x} = E\{x[N-1+l]\mathbf{x}^T\}$
$= E\{s[n]\mathbf{s}^T\}$	$= E\{\mathbf{s}(\mathbf{s}+\mathbf{w})^T\}$	$= [r_{xx}[N-1+l]\cdots r_{xx}[l]$
$= [r_{ss}[n]\cdots r_{ss}[0]]$	$= E\{\mathbf{s}\mathbf{s}^T + \mathbf{s}\mathbf{w}^T\}$	$]$
$= \tilde{\mathbf{r}}_{ss}^{\mathrm{T}}$ (vector!)	$= \mathbf{R}_{ss}$ (Matrix!)	$= \tilde{\mathbf{r}}_{xx}^{\mathrm{T}}$ (vector!)
$\mathbf{C}_{xx} = E\{(\mathbf{s}+\mathbf{w})(\mathbf{s}+\mathbf{w})^T\}$	$\mathbf{C}_{xx} = E\{(\mathbf{s}+\mathbf{w})(\mathbf{s}+\mathbf{w})^T\}$	$\mathbf{C}_{xx} = E\{\mathbf{x}\mathbf{x}^T\}$
$= E\{\mathbf{s}\mathbf{s}^T + \mathbf{w}\mathbf{w}^T\}$	$= E\{\mathbf{s}\mathbf{s}^T + \mathbf{w}\mathbf{w}^T\}$	$= \mathbf{R}_{xx}$
$= \mathbf{R}_{ss} + \mathbf{R}_{ww}$	$= \mathbf{R}_{ss} + \mathbf{R}_{ww}$	
$\hat{s}[n] = \tilde{\mathbf{r}}_{ss}^T(\mathbf{R}_{ss}+\mathbf{R}_{ww})^{-1}\mathbf{x}$	$\hat{\mathbf{s}} = \mathbf{R}_{ss}(\mathbf{R}_{ss}+\mathbf{R}_{ww})^{-1}\mathbf{x}$	$\hat{x}[N-1+l] = \tilde{\mathbf{r}}_{xx}^T\mathbf{R}_{xx}^{-1}\mathbf{x}$
$[1\times(n+1)][(n+1)\times(n+1)][(n+1)\times 1]$	$[N\times N][N\times N][N\times 1]$	$[1\times N][N\times N][N\times 1]$

Wiener filtering



In Principle: Solve WHF Eqs for filter **h** at each *n*

In Practice: Use Levinson Recursion to Recursively Solve