Ch. 11
General Bayesian Estimators

Introduction

In Chapter 10 we:

- introduced the idea of a "a priori" information on θ

  $\Rightarrow$ use "prior" pdf: $p(\theta)$
- defined a new optimality criterion

  $\Rightarrow$ Bayesian MSE
- showed the Bmse is minimized by E $\{\theta|\mathbf{x}\}$

called:
- "mean of posterior pdf"
- "conditional mean"

In Chapter 11 we will:

- define a more general optimality criterion

  $\Rightarrow$ leads to several different Bayesian approaches

  $\Rightarrow$ includes Bmse as special case

Why? Provides flexibility in balancing:

- model,
- performance, and
- computations

11.3 Risk Functions

Previously we used Bmse as the Bayesian measure to minimize

$$Bmse = E\left\{\left(\theta - \hat{\theta}\right)^2\right\} \quad w.r.t.\ p(\mathbf{x},\theta)$$

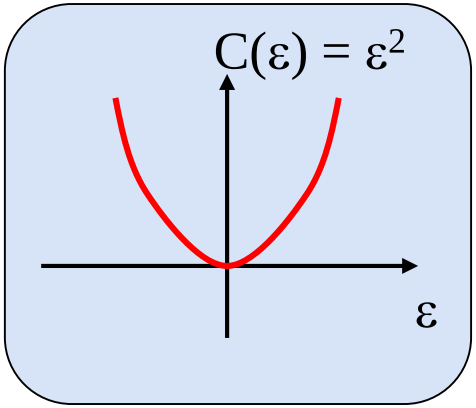
$$\theta - \hat{\theta} \triangleq \varepsilon$$

So, Bmse is… Expected value of square of error

Let's write this in a way that will allow us to generalize it.

Define a quadratic Cost Function: $C(\varepsilon) = \varepsilon^2 = \left(\theta - \hat{\theta}\right)^2$

Then we have that $Bmse = E\{C(\varepsilon)\}$



Why limit the cost function to just quadratic?

General Bayesian Criteria

1. Define a cost function: $C(\varepsilon)$

2. Define Bayes Risk: $\mathcal{R} = E\{C(\varepsilon)\}$ w.r.t. $p(\mathbf{x},\theta)$

$$\mathcal{R}(\hat{\theta}) = E\left\{C(\theta - \hat{\theta})\right\}$$

Depends on choice of estimator

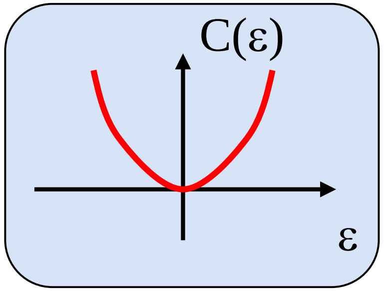
3. Minimize Bayes Risk w.r.t. estimate $\hat{\theta}$

The choice of the cost function can be tailored to:

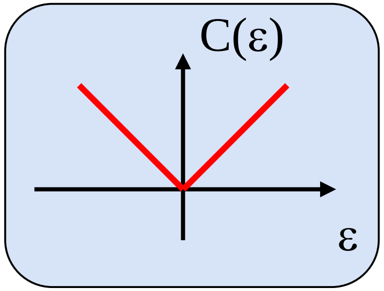
- Express importance of avoiding certain kinds of errors
- Yield desirable forms for estimates
  - e.g., easily computed
- Etc.

Three Common Cost Functions

1. Quadratic: $C(\varepsilon) = \varepsilon^2$

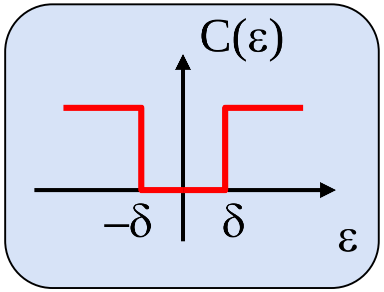


2. Absolute: $C(\varepsilon) = |\varepsilon|$



3. Hit-or-Miss: $C(\varepsilon) = \begin{cases} 0, & |\varepsilon| < \delta \\ 1, & |\varepsilon| \geq \delta \end{cases}$

   $\delta > 0$ and small



General Bayesian Estimators

Derive how to choose estimator to minimize the chosen risk:

$$\mathcal{R}(\hat{\theta}) = E\left\{C(\theta - \hat{\theta})\right\}$$

$$= \iint C(\theta - \hat{\theta}) \underbrace{p(x,\theta)}_{= p(\theta|x)\, p(x)} \, dx \, d\theta$$

$$= \int \underbrace{\left[\int C(\theta - \hat{\theta}) p(\theta|x) d\theta\right]}_{\triangleq g(\hat{\theta})} p(x) dx$$

must minimize this for each **x** value

So… for a given desired cost function…

you have to find the form of the optimal estimator

The Optimal Estimates for the Typical Costs

1. **Quadratic**: $\mathcal{R}(\hat{\theta}) = E\left\{\left(\theta - \hat{\theta}\right)^2\right\} = Bmse(\hat{\theta})$

   As we saw in Ch. 10

   $$\hat{\theta} = E\{\theta \mid \mathbf{x}\}$$
   $$= \text{mean of } p(\theta \mid \mathbf{x})$$

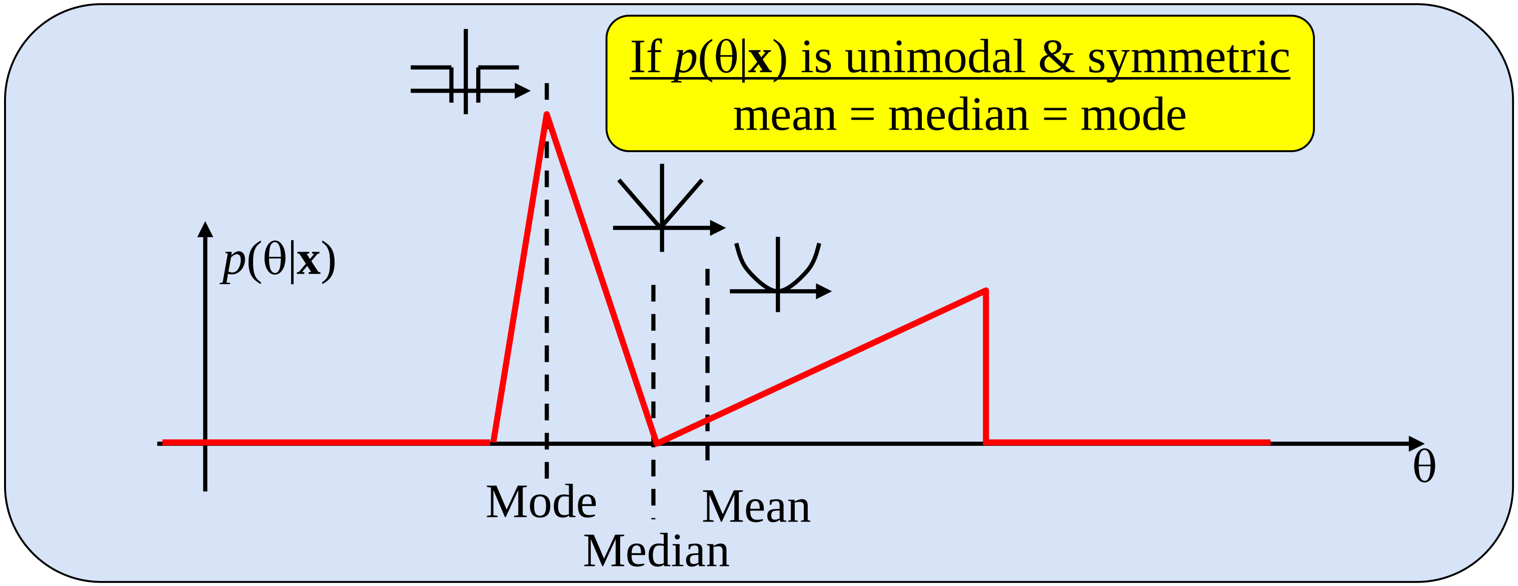
2. **Absolute**: $\mathcal{R}(\hat{\theta}) = E\left\{\left|\theta - \hat{\theta}\right|\right\}$

   $$\hat{\theta} = \text{median of } p(\theta \mid \mathbf{x})$$

3. **Hit-or-Miss**:

   $$\hat{\theta} = \text{mode of } p(\theta \mid \mathbf{x})$$

   "Maximum A Posteriori" or MAP



Derivation for Absolute Cost Function

Writing out the function to be minimized gives:

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

$$= \underbrace{\int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) p(\theta | \mathbf{x}) d\theta}_{\text{region where } |\theta-\hat{\theta}|=\hat{\theta}-\theta} + \underbrace{\int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) p(\theta | \mathbf{x}) d\theta}_{\text{region where } |\theta-\hat{\theta}|=\theta-\hat{\theta}}$$

Now set $\dfrac{\partial g(\hat{\theta})}{\partial \hat{\theta}} = 0$ and use Leibnitz's rule for $\dfrac{\partial}{\partial u} \int_{\phi_1(u)}^{\phi_2(u)} h(u,v)\, dv$

$$\Rightarrow \int_{-\infty}^{\hat{\theta}} p(\theta | \mathbf{x})\, d\theta - \int_{\hat{\theta}}^{\infty} p(\theta | \mathbf{x}) d\theta = 0$$

which is satisfied if… (area to the left) = (area to the right)

$\Rightarrow$ Median of conditional PDF

Derivation for Hit-or-Miss Cost Function

Writing out the function to be minimized gives:

$$g(\hat{\theta}) = \int_{-\infty}^{\infty} C(\theta - \hat{\theta})\, p(\theta \mid \mathrm{x}) d\theta$$

$$= \int_{-\infty}^{\hat{\theta}-\delta} 1 \cdot p(\theta \mid \mathrm{x})\, d\theta \ + \int_{\hat{\theta}+\delta}^{\infty} 1 \cdot p(\theta \mid \mathrm{x})\, d\theta$$

Almost all the probability = 1 – left out

$$= 1 - \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta \mid \mathrm{x})\, d\theta$$

Maximize this integral

So… center the integral around peak of integrand

$\Rightarrow$ Mode of conditional PDF

11.4 MMSE Estimators

We've already seen the solution for the scalar parameter case

$$\hat{\theta} = E\{\theta \mid \mathbf{x}\}$$

$$= \text{mean of } p(\theta \mid \mathbf{x}$$

Here we'll look at:

- Extension to the vector parameter case
- Analysis of Useful Properties

Vector MMSE Estimator

The criterion is… minimize the MSE for each component

Vector Parameter: $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_p \end{bmatrix}^T$

Vector Estimate: $\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_2 & \cdots & \hat{\theta}_p \end{bmatrix}^T$

is chosen to minimize each of the MSE elements:

$$E\{(\theta_i - \hat{\theta}_i)^2\} = \int (\theta_i - \hat{\theta}_i)^2 \ p(\mathbf{x}, \theta_i)\, d\mathbf{x}\, d\theta_i$$

= $p(\mathbf{x}, \boldsymbol{\theta})$ integrated over all other θ_j's

From the scalar case we know the solution is:

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

$$= \int \cdots \int \theta_i p(\theta_1, \ldots, \theta_p \,|\, \mathbf{x})\, d\theta_1 \cdots d\theta_p$$

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

$$\Rightarrow \quad \hat{\theta}_i = E\{\theta_i \,|\, \mathbf{x}\}$$

So… putting all these into a vector gives:

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_2 & \cdots & \hat{\theta}_p \end{bmatrix}^T$$

$$= \begin{bmatrix} E\{\theta_1 | \mathbf{x}\} & E\{\theta_2 | \mathbf{x}\} & \cdots & E\{\theta_p | \mathbf{x}\} \end{bmatrix}^T$$

$$= E\left\{ \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_p \end{bmatrix}^T | \mathbf{x} \right\}$$

$$\Rightarrow \quad \hat{\boldsymbol{\theta}} = E\{\boldsymbol{\theta} | \mathbf{x}\}$$

Vector MMSE Estimate
= Vector Conditional Mean

Similarly… $Bmse(\hat{\theta}_i) = \int \left[\mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}} \right]_{ii} p(\mathbf{x})\, d\mathbf{x} \qquad i = 1, \ldots, p$

where $\quad \mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}} = E_{\boldsymbol{\theta}|\mathbf{x}} \left\{ \left[\boldsymbol{\theta} - E\{\boldsymbol{\theta} | \mathbf{x}\}\right]\left[\boldsymbol{\theta} - E\{\boldsymbol{\theta} | \mathbf{x}\}\right]^T \right\}$

Ex. 11.1 Bayesian Fourier Analysis

Signal model is: $x[n] = a\cos(2\pi f_o n) + b\sin(2\pi f_o n) + w[n]$

AWGN
w/ zero mean and σ^2

$$\boldsymbol{\theta} = \begin{bmatrix} a \\ b \end{bmatrix} \sim N(\mathbf{0}, \sigma_\theta^2 \mathbf{I})$$

$\boldsymbol{\theta}$ and $w[n]$ are independent for each n

This is a common propagation model called Rayleigh Fading

Write in matrix form: $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ Bayesian Linear Model

$$\mathbf{H} = \begin{bmatrix} \uparrow & \uparrow \\ \text{cosine} & \text{sine} \\ \downarrow & \downarrow \end{bmatrix}$$

Results from Ch. 10 show that

$$\hat{\boldsymbol{\theta}} = E\{\boldsymbol{\theta} \mid \mathbf{x}\} = \left[\frac{1}{\sigma_\theta^2}\mathbf{I} + \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\right]^{-1} \frac{\mathbf{H}^T\mathbf{x}}{\sigma^2} \qquad \mathbf{C}_{\boldsymbol{\theta}|\mathbf{x}} = \left[\frac{1}{\sigma_\theta^2}\mathbf{I} + \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\right]^{-1}$$

For f_o chosen such that **H** has orthogonal columns then

$$\hat{\boldsymbol{\theta}} = E\{\boldsymbol{\theta} \mid \mathbf{x}\} = \left[\frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_\theta^2} + \frac{1}{\sigma^2}}\right]\mathbf{H}^T\mathbf{x}$$

$$\hat{a} = \beta\left[\frac{2}{N}\sum_{n=0}^{N-1} x[n]\cos(2\pi f_o n)\right] \qquad \beta = \frac{1}{1 + \frac{2\sigma^2/N}{\sigma_\theta^2}}$$

$$\hat{b} = \beta\left[\frac{2}{N}\sum_{n=0}^{N-1} x[n]\sin(2\pi f_o n)\right]$$

Fourier Coefficients in the Brackets

Recall: Same form as classical result, except there $\beta = 1$

Note: $\beta \approx 1$ if $\sigma_\theta^2 >> 2\sigma^2/N$

$\Rightarrow$ if prior knowledge is poor, this degrades to classical

Impact of Poor Prior Knowledge

Conclusion: For poor prior knowledge in Bayesian Linear Model
MMSE Est. → MVU Est.

Can see this holds in general: Recall that

$$\hat{\boldsymbol{\theta}} = E\{\boldsymbol{\theta} \mid \mathbf{x}\} = \boldsymbol{\mu}_{\boldsymbol{\theta}} + \left[\mathbf{C}_{\boldsymbol{\theta}}^{-1} + \mathbf{H}^T \mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}\right]^{-1} \mathbf{H}^T \mathbf{C}_{\mathbf{w}}^{-1}\left[\mathbf{x} + \mathbf{H}\boldsymbol{\mu}_{\boldsymbol{\theta}}\right]$$

For no prior information: $\mathbf{C}_{\boldsymbol{\theta}}^{-1} \rightarrow \mathbf{0}$ and $\boldsymbol{\mu}_{\boldsymbol{\theta}} \rightarrow \mathbf{0}$

$$\hat{\boldsymbol{\theta}} \rightarrow \underbrace{\left[\mathbf{H}^T \mathbf{C}_{\mathbf{w}}^{-1}\mathbf{H}\right]^{-1} \mathbf{H}^T \mathbf{C}_{\mathbf{w}}^{-1}\mathbf{x}}$$

MVUE for General Linear Model

Useful Properties of MMSE Est.

Will be used for Kalman Filter

1. **Commutes over affine mappings:**

   If we have $\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\theta} + \mathbf{b}$ then $\hat{\boldsymbol{\alpha}} = \mathbf{A}\hat{\boldsymbol{\theta}} + \mathbf{b}$

2. **Additive Property for independent data sets**

   Assume $\boldsymbol{\theta}$, $\mathbf{x}_1$, $\mathbf{x}_2$ are jointly Gaussian w/ $\mathbf{x}_1$ and $\mathbf{x}_2$ independent

$$\hat{\boldsymbol{\theta}} = E\{\boldsymbol{\theta}\} + \mathbf{C}_{\boldsymbol{\theta}\mathbf{x}_1}\mathbf{C}_{\mathbf{x}_1}^{-1}[\mathbf{x}_1 - E\{\mathbf{x}_1\}] + \mathbf{C}_{\boldsymbol{\theta}\mathbf{x}_2}\mathbf{C}_{\mathbf{x}_2}^{-1}[\mathbf{x}_2 - E\{\mathbf{x}_2\}]$$

a priori Estimate | Update due to $\mathbf{x}_1$ | Update due to $\mathbf{x}_2$

Proof: Let $\mathbf{x} = [\mathbf{x}_1^T \ \mathbf{x}_2^T]^T$. The jointly Gaussian assumption gives:

$$\hat{\boldsymbol{\theta}} = E\{\boldsymbol{\theta}\} + \mathbf{C}_{\theta x}\mathbf{C}_{x}^{-1}[\mathbf{x} - E\{\mathbf{x}\}]$$

Indep. $\Rightarrow$ Block Diagonal

$$= E\{\boldsymbol{\theta}\} + \begin{bmatrix}\mathbf{C}_{\theta x_1} & \mathbf{C}_{\theta x_2}\end{bmatrix}\begin{bmatrix}\mathbf{C}_{x_1}^{-1} & 0 \\ 0 & \mathbf{C}_{x_2}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 - E\{\mathbf{x}_1\} \\ \mathbf{x}_2 - E\{\mathbf{x}_2\}\end{bmatrix}$$

Simplify to get the result

3. **Jointly Gaussian case leads to a linear estimator:** $\hat{\boldsymbol{\theta}} = \mathbf{P}\mathbf{x} + \mathbf{m}$