Detection - Estimation

Ali Mohammad-Djafari

A Graduated Course

Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46555, USA

Draft: May 1, 2001 [1]

[1] File: `nd1`

Contents

Preface

During my visit at the Electrical Engineering Department of the Notre Dame University, I had to teach a course on Detection-Estimation.

I could not find really a complete and convenient textbook to cover all the materials that an Electrical Engineer have to know on the subject. Some books are too mathematical, some others are too oriented on the applications and not enough rigorous on the mathematics.

The purpose of these notes is to fill this gap: to introduce the reader to the basic theory of hypothesis testing and estimation theory using the main probability and statistical tools and also to give him the basic theory of signal detection and estimation as used in practical applications of electrical engineering.

The contents of these notes are mainly covered by two books:

- Detection and estimation, by D. Kazakos and P. Papantoni-Kazakos, and
- An introduction to signal detection and estimation, by H. Vincent Poor.

Ali Mohammad-Djafari

Chapter 1

Introduction

Generally speaking, signal *detection* and *estimation* is the area of study that deals with information processing: conversion, transmission, observation and information extraction. The main area of applications of detection and estimation theory are radar, sonar, analog or digital communications, but detection and estimation theory becomes also the main tool in other area such as radioastronomy, geophysics, medical imaging, biological data processing, etc.

In general, detection and estimation applications involve making *inferences* from observations that are distorted or corrupted in some unknown way or too complicated to be modelled in a deterministic way. Moreover, sometimes even the information that one wishes to extract from such observations is not well determined. Thus, it is very useful to cast detection and estimation problems in a *probabilistic framework* and *statistical inference.* But using the probability theory and the statistical inference tools does not forcibly means that the corresponding physical phenomena are necessarily random.

In statistical inference, the goal is not to make an immediate decision, but is instead to provide a summary of the statistical evidence which the future users can easily incorporate into their decision process. The task of decision making is then given to the *decision theory.*

Signal detection is inherently a decision making task. In signal estimation also we need often to make decisions. So, for detection and estimation we need not only the probability theory and statistical inference tools but also the decision and hypothesis testing tools. The main common tool with which we have to start is then the probabilistic and stochastic description of the observations and the unknown quantities.

Once again a probablistic or stochastic description models the effect of causes whose origin and nature are either unknown or too complex to be described deterministically.

The simplest tool of a probabilistic model for a quantity is a scalar *random variable* X which is fully described by its probability distribution $F(x) = \Pr\{X \leq x\}$. The next simplest model is a *random vector* $\boldsymbol{X} = [X_1, \cdots, X_n]^t$, where $\{X_j\}$ are random variables. A random vector is fully described by its probability distribution $F(\boldsymbol{x}) = \Pr\{\boldsymbol{X} \leq \boldsymbol{x}\}$. The next and the most general stochastic model for a quantity is a *random function* $X(\boldsymbol{r})$, where $\boldsymbol{r}$ is a finite dimensional independent variable and where for every fixed values $\boldsymbol{r} = \boldsymbol{r}_j$, the scalar quantity $X_j = X(\boldsymbol{r}_j)$ is a scalar random variable. For example, when $\boldsymbol{r} = (x, y)$ represents the spatial coordinates in a plane, then $X(x, y)$ is called a *random field* and when $\boldsymbol{r} = t$ represents the time variable, then $X(t)$ is called a *stochastic process.* In the rest of these notes, we consider only this last model.

A stochastic process $X(t)$ is completely described by the probability distribution

$$F(x_1,\cdots,x_n;t_1,\cdots,t_n)=F(\boldsymbol{x};\boldsymbol{t})=\Pr\{X(t_j)\leq x_j;j=1,\cdots,n\}$$

for every n and every time instants $\{t_j\}$. The stochastic process is *discrete-time* if it is described only by its realizations on a countable set $\{t_j\}$ of time instants. Then, time is counted by the indices j, and the stochastic process is fully described by the random vectors $\boldsymbol{X}_j=[X_j,X_{j+1},\cdots,X_{j+n}]^t$.

A stochastic process $X(t)$ is said *well known*, if the distribution $F(x_1,\cdots,x_n;t_1,\cdots,t_n)$ is precisely known for all n, every set $\{t_j\}$ and every vector value $\boldsymbol{x}$. The process is instead said *parametrically known*, if there exists a finite dimensional parameter vector $\boldsymbol{\theta}=[\theta_1,\cdots,\theta_m]^t$ such that the conditional distribution $F(x_1,\cdots,x_n;t_1,\cdots,t_n|\boldsymbol{\theta})$ is precisely known for all n, every set $\{t_j\}$, every vector value $\boldsymbol{x}$ and a fixed given value of $\boldsymbol{\theta}$. A stochastic process $X(t)$ is *non parametrically* described, if there is no vector parameter $\boldsymbol{\theta}$ of finite dimensionality such that the distribution $F(\boldsymbol{x},\boldsymbol{t}|\boldsymbol{\theta})$ is completely described for all values of the vector $\boldsymbol{\theta}$ and for all $n,\boldsymbol{t}$ and $\boldsymbol{x}$. As an example, a stationary, discrete time process $\{X_i\}$ where the random variables X_i have finite variances is a nonparametrically described process. In fact, this description represents a whole class of stochastic processes. If we assume now that this process is also Gaussian, then it becomes *parametrically known*, since only its mean and spectral density functions are needed for its full description. When these two quantities are also provided, the process becomes *well known.*

From now, we have the main necessary ingredients to give a general scope of the detection and estimation theory. Let consider a case where the observed quantity is modelled by a stochastic process $X(t)$ and the observed signal $x(t)$ is considered as a realization of the process, *i.e.*, an observed waveform generated by $X(t)$.

1.1 Basic definitions

- Probability spaces:
  The probability theory starts by defining an *observation set* Γ and a class of subsets $\mathcal{G}$ of it, called *observation events*, to which we wish to assign probabilities. The pair $(\Gamma, \mathcal{G})$ is termed the *observation space.*

  For analytical reasons we will always assume that the collection $\mathcal{G}$ is a σ-algebra; that is, $\mathcal{G}$ contains all complements relative to Γ and denumerable unions of its members, *i.e.*;

$$\begin{array}{lcl} \text{if } A \in \mathcal{G} & \longrightarrow & A^c \in \mathcal{G} \\ \text{and} & & \\ \text{if } A_1, A_2, ... \in \mathcal{G} & \longrightarrow & \cup_i A_i \in \mathcal{G} \end{array} \tag{1.1}$$

  Two special cases are of interest:

  - Discrete case: $\Gamma = \{\gamma_1, \gamma_2, \cdots\}$
    In this case $\mathcal{G}$ is the set of all subsets of Γ which is usually denoted by 2^Γ and is called the *power set* of Γ.
    For this case, probabilities can be assigned to subsets of Γ in terms of a *probability mass function*, $p : \Gamma \longrightarrow [0,1]$, by

$$P(A) = \sum_{\gamma_i \in A} p(\gamma_i), \quad A \in 2^\Gamma. \tag{1.2}$$

    Any function mapping Γ to $[0,1]$ can be a probability mass function provided that it satisfies the condition of normality

$$\sum_i p(\gamma_i) = 1. \tag{1.3}$$

  - Continuous case: $\Gamma = \mathbb{R}^n$, the set of n-dimensional vectors with real components.
    In this case we want to assign the probabilities to the sets

$$\{\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n | a_1 < x_1 < b_1, \cdots, a_n < x_n < b_n\} \tag{1.4}$$

    where the a_i's and b_i's are arbitrary real numbers. So, in this case, $\mathcal{G}$ is the smallest σ-algebra containing all of these sets with the a_i's and b_i's ranging throughout the reals. This σ-algebra is usually denoted $\mathcal{B}^n$ and is called the class of *Borel sets* in $\mathbb{R}^n$.
    In this case the probabilities can be assigned in terms of a *probability density function*, $p : \mathbb{R}^n \longrightarrow \mathbb{R}^+$, by

$$P(A) = \int_A p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \quad A \in \mathcal{B}^n. \tag{1.5}$$

    Any integrable function mapping $\mathbb{R}^n$ to $\mathbb{R}^+$ can be a probability density function provided that it satisfies the condition

$$\int_{\mathbb{R}^n} p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1. \tag{1.6}$$

- For compactness, we may use the term *density* for both probability density function and probability mass function and use the following notation when necessary

$$P(A) = \int_A p(\boldsymbol{x})\mu(\,\mathrm{d}\boldsymbol{x}) \tag{1.7}$$

  for both the summation equation (1.2) and the integration equation (1.5).

- Random variable:
  $X = X(\omega)$ is a function $\omega \mapsto \mathbb{R}$, where ω represents elements on the probability space.

- Probability distribution:
  $F(x)$ is a function $\mathbb{R} \mapsto [0, 1]$ such that

$$F(x) = \Pr\{X \le x\} \tag{1.8}$$

- Probability density function:
  $f(x)$ is a function $\mathbb{R} \mapsto \mathbb{R}^+$ such that

$$\begin{aligned} f(x) &= \frac{\partial F(x)}{\partial x}, \\ F(x) &= \Pr\{X \le x\} = \int_{-infty}^{x} f(t)\,\mathrm{d}t \end{aligned} \tag{1.9}$$

- For a real function g of the random variable X, the *expected value* of $g(X)$, denoted $\mathrm{E}[g(X)]$, is defined by any of the followings:

$$\mathrm{E}[g(X)] = \sum_i g(\gamma_i)\,p(\gamma_i) \tag{1.10}$$

$$\mathrm{E}[g(X)] = \int_{\mathbb{R}} g(x)\,p(x)\,\mathrm{d}x \tag{1.11}$$

$$\mathrm{E}[g(X)] = \int_{\Gamma} g(x)\,p(x)\,\mu(\,\mathrm{d}x) \tag{1.12}$$

- Random vector or a vector of random variables:
  $X = [X_1, \cdots, X_n]$ where X_i are scalar random varaibles.

- Joint probability distribution:

$$\begin{aligned} F(x_1, \cdots, x_n) &= \Pr\{X_1 \le x_1, \cdots, X_n \le x_n\} \\ F(\boldsymbol{x}) &= \Pr\{\boldsymbol{X} \le \boldsymbol{x}\} \end{aligned} \tag{1.13}$$

- Stochastic process

- Stochastic process: $X(t) = X(t, \omega)$, where $X(t, \omega)$ is a scalar random variable for all t.

- A stochastic process is completely defined by

$$\begin{aligned} F(x_1,\cdots,x_n;t_1,\cdots,t_n) &= \Pr\{X_i(t_i) \leq x_i, i=1,\cdots,n\} \\ F(\boldsymbol{x};\boldsymbol{t}) &= \Pr\{\boldsymbol{X}(\boldsymbol{t}) \leq \boldsymbol{x}\} \end{aligned} \tag{1.14}$$

- A stochastic process is stationary if

$$\begin{aligned} F(x_1,\cdots,x_n;t_1,\cdots,t_n) &= F\left(x_1,\cdots,x_n;(t_1+\tau),\cdots,(t_n+\tau)\right) \\ F(\boldsymbol{x};\boldsymbol{t}) &= F(\boldsymbol{x};\boldsymbol{t}+\tau\mathbf{1}) \end{aligned} \tag{1.15}$$

- A stochastic process is memoryless or white if

$$F(x_1,\cdots,x_n;t_1,\cdots,t_n) = \prod_i F\left(x_i;t_i\right) \tag{1.16}$$

- Discrete time stochastic process:

$$F(\boldsymbol{x}) = \Pr\{\boldsymbol{X} \leq \boldsymbol{x}\} \tag{1.17}$$

- Memoryless discrete time stochastic process:

$$F(\boldsymbol{x}) = \Pr\{\boldsymbol{X} \leq \boldsymbol{x}\} = \prod_{i=1}^{n} F_i(x_i) \tag{1.18}$$

- Memoryless and stationary discrete time stochastic process:

$$F(\boldsymbol{x}) = \prod_{i=1}^{n} F_i(x_i) \quad \text{and} \quad F_i(x) = F_j(x) = F(x), \forall i,j \tag{1.19}$$

- A memoryless and stationary discrete time stochastic process generates in time *independent and identically distributed* (i.i.d.) random variables.

- Well known stochastic process:
  A stochastic process is well known if the distribution $F(\boldsymbol{x},\boldsymbol{t})$ is known for all $n,\boldsymbol{t}$ and $\boldsymbol{x}$.

- Parametrically well known stochastic process:
  A stochastic process is parametrically well known if there exists a finite dimentional vector parameter $\boldsymbol{\theta} = [\theta_1,\cdots,\theta_m]$ such that the conditional distribution $F(\boldsymbol{x},\boldsymbol{t}|\boldsymbol{\theta})$ is known for all $n,\boldsymbol{t}$ and $\boldsymbol{x}$.

- Non parametric description of a stochastic process:
  A stochastic process $X(t)$ is non parametrically described, if there is no vector parameter $\boldsymbol{\theta}$ of finite dimensionality such that the distribution $F(\boldsymbol{x},\boldsymbol{t}|\boldsymbol{\theta})$ is completely described for every given vector $\boldsymbol{\theta}$ and for all $n,\boldsymbol{t}$ and $\boldsymbol{x}$.

- Observed data:
  samples of $x(t)$ a realization of $X(t)$ in some time interval $[0,T]$.

1.2 Summary of notations

X	A random variable
x	A realization of a random variable
$\boldsymbol{x} = \{x_1, \cdots, x_n\}$	n samples (realizations) of a random variable
$X(t)$	A random function or a stochastic process
$x(t)$	A realization of a random function
$\boldsymbol{X}$	A random vector or a discrete-time stochastic process
$\boldsymbol{x}$	A realization of a random vector or a discrete-time stochastic process
$\boldsymbol{x}_n = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$	n samples (realizations) of a random vector or a discrete-time stochastic process
$F(x)$	A probability distribution of a scalar random variable
$f(x)$	A probability density function of a scalar random variable
$F_{\boldsymbol{\theta}}(x)$	A parametrical probability distribution
$f_{\boldsymbol{\theta}}(x\|\boldsymbol{\theta})$	A parametrical probability density function
$F(x\|\boldsymbol{\theta})$	A conditional probability distribution
$f(x\|\boldsymbol{\theta})$	A conditional probability density function
$F(\boldsymbol{\theta}\|\boldsymbol{x})$	A posterior probability distribution of $\boldsymbol{\theta}$ conditioned on the observations $\boldsymbol{x}$
$f(\boldsymbol{\theta}\|\boldsymbol{x})$	A posterior probability density function of $\boldsymbol{\theta}$ conditioned on the observations $\boldsymbol{x}$
Θ	A random scalar parameters
θ	A sample of Θ
$\boldsymbol{\Theta}$	A random vector of parameters
$\boldsymbol{\theta}$	A sample of $\boldsymbol{\Theta}$
Γ	The space of possible values of x
$\boldsymbol{\Gamma}$	The space of possible values of $\boldsymbol{x}$
$\mathcal{T}$	The space of possible values of θ
$\boldsymbol{\mathcal{T}}$	The space of possible values of $\boldsymbol{\theta}$

Chapter 2

Basic concepts of binary hypothesis testing

Most signal detection problems can be cast in the framework of M-ary hypothesis testing, where from some observations (data) we wish to decide among M possible situations. For example, in a communication system, the receiver observes an electric waveform that consists of one of the M possible signals corrupted by channel or receiver noise, and we wish to decide which of the M possible signals is present during the observation. Obviously, for any given decision problem, there are a number of possible decision strategies or rules that can be applied, however we would like to choose a decision rule that is optimal in some sense. There are several classical useful criteria of optimality for such problems. The main object of this chapter is to give all the necessary basic definitions to define these criteria and their practical signification. Before going to the general case of M-ary hypothesis testing problem, let us start by a particular problem of binary (M=2) hypothesis testing which allows us to introduce the main basis more easily.

2.1 Binary hypothesis testing

The primary problem that we consider as an introduction is the simple hypothesis testing problem in which we assume that the observed data belong only on two possible processes with well known probability distributions P_0 and P_1:

$$\left\{ \begin{array}{ll} H_0 & : X \sim P_0 \\ H_1 & : X \sim P_1 \end{array} \right. \tag{2.1}$$

where "$X \sim P$" denotes "X has distribution P" or "Data come from a stochastic process whose distribution is P". The hypotheses H_0 and H_1 are respectively referred to as *null* and *alternative* hypotheses. A decision rule δ for H_0 versus H_1 is any partition of the observation space Γ into Γ_1 and $\Gamma_0 = \Gamma_1^c$ such that we choose H_1 when $\boldsymbol{x} \in \Gamma_1$ and H_0 when $\boldsymbol{x} \in \Gamma_0$. The sets Γ_1 and Γ_0 are respectively called the *rejection* and *acceptance* regions. So, we can think of the decision rule δ as a function on Γ such that

$$\delta(\boldsymbol{x}) = \left\{ \begin{array}{ll} \delta_1 = 1 & \text{if } \boldsymbol{x} \in \Gamma_1 \\ \delta_0 = 0 & \text{if } \boldsymbol{x} \in \Gamma_0 = \Gamma_1^c \end{array} \right. \tag{2.2}$$

so that the value of δ for a given $\boldsymbol{x}$ is the index of the hypothesis accepted by the decision rule δ.

We can also think of the decision rule $\delta(\boldsymbol{x})$ as a probability distribution $\{\delta_0, \delta_1\}$ on the space $\mathcal{D}$ of all the possible decisions where δ_j is the probability of deciding H_j in the light of the data $\boldsymbol{x}$. In both cases $\delta_0 + \delta_1 = 1$.

We would like to choose H_0 or H_1 in some optimal way and, with this in mind, we may assign costs to our decisions. In particular we may assign costs $c_{ij} \geq 0$ to pay if we make the decision H_i while the true decision to make was H_j. With the partition $\Gamma = \{\Gamma_0, \Gamma_1\}$ of the observation set, we can then define the conditional probabilities

$$P_{ij} = \Pr\{\boldsymbol{X} = \boldsymbol{x} \in \Gamma_i | H = H_j\} = P_j(\Gamma_i) = \int_{\Gamma_i} p_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{2.3}$$

and then the average or expected *conditional risks* $R_j(\delta)$ for each hypothesis as

$$R_j(\delta) = \sum_{i=0}^{1} c_{ij} P_{ij} = c_{1j} P_j(\Gamma_1) + c_{0j} P_j(\Gamma_0), \quad j = 0, 1 \tag{2.4}$$

2.2 Bayesian binary hypothesis testing

Now assume that we can assign *prior* probabilities π_0 and $\pi_1 = 1 - \pi_0$ to the hypotheses H_0 and H_1, either to translate our preferences or to translate our prior knowledge about these hypotheses. Note that π_j is the probability that H_j is true unconditional (or independent) of the observation data $\boldsymbol{x}$ of $\boldsymbol{X}$. This is why they are called *prior* or *a priori* probabilities. For given priors $\{\pi_0, \pi_1\}$ we can define the *posterior* or *a posteriori* probabilities

$$\pi_j(\boldsymbol{x}) = \Pr\{H = H_j | \boldsymbol{X} = \boldsymbol{x}\} = \frac{p_j(\boldsymbol{x})\,\pi_j}{m(\boldsymbol{x})} \tag{2.5}$$

where

$$m(\boldsymbol{x}) = \sum_j P_j(\boldsymbol{x})\,\pi_j \tag{2.6}$$

is the overall density of $\boldsymbol{X}$.

We can also define an average or Bayes risk $r(\delta)$ as the overall average cost incurred by the decision rule δ:

$$r(\delta) = \sum_j \pi_j\, R_j(\delta) \tag{2.7}$$

We may now use this quantity to define an optimum decision rule as the one that minimizes, over all decision rules, the Bayes risk. Such a decision rule is known as a Bayes decision rule.

To go a little further in details, let combine (2.4) and (2.7) to give

$$\begin{aligned} r(\delta) &= \sum_j \pi_j\, R_j(\delta) = \sum_j \pi_j \sum_i c_{ij} P_j(\Gamma_i) \\ &= \sum_j \pi_j c_{0j} P_j(\Gamma_0) + \pi_j c_{1j} P_j(\Gamma_1) \\ &= \sum_j \pi_j c_{0j} (1 - P_j(\Gamma_1)) + \pi_j c_{1j} P_j(\Gamma_1) \end{aligned}$$

$$\begin{aligned} &= \sum_j \pi_j c_{0j} + \sum_j \pi_j (c_{1j} - c_{0j}) P_j(\Gamma_1) \\ &= \sum_j \pi_j c_{0j} + \int_{\Gamma_1} \sum_j \pi_j \left(c_{1j} - c_{0j}\right) p_j(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \qquad (2.8) \end{aligned}$$

Thus, we see that $r(\delta)$ is a minimum over all Γ_1 if we choose

$$\begin{aligned} \Gamma_1 &= \left\{ \boldsymbol{x} \in \Gamma | \sum_j (c_{1j} - c_{0j}) \pi_j p_j(\boldsymbol{x}) \leq 0 \right\} \qquad (2.9) \\ &= \{ \boldsymbol{x} \in \Gamma | (c_{11} - c_{01}) \pi_1 p_1(\boldsymbol{x}) \leq (c_{00} - c_{10}) \pi_0 p_0(\boldsymbol{x}) \} \qquad (2.10) \end{aligned}$$

In general, the costs $c_{jj} < c_{ij}$ which means that the cost of correctly deciding H_i is less than the cost of incorrectly deciding it. Then, (2.10) can be written

$$\begin{aligned} \Gamma_1 &= \left\{ \boldsymbol{x} \in \Gamma | \frac{\pi_1 p_1(\boldsymbol{x})}{\pi_0 p_0(\boldsymbol{x})} \geq \tau_1 = \frac{c_{10} - c_{00}}{c_{01} - c_{11}} \right\} \qquad (2.11) \\ &= \left\{ \boldsymbol{x} \in \Gamma | \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} \geq \tau_2 = \frac{\pi_0}{\pi_1} \frac{c_{10} - c_{00}}{c_{01} - c_{11}} \right\} \qquad (2.12) \\ &= \{ \boldsymbol{x} \in \Gamma | c_{10} \pi_0(\boldsymbol{x}) + c_{11} \pi_1(\boldsymbol{x}) \leq c_{00} \pi_0(\boldsymbol{x}) + c_{01} \pi_1(\boldsymbol{x}) \} \qquad (2.13) \end{aligned}$$

This decision rule is known as a *likelihood-ratio* test or *posterior probability ratio* test due to the fact that $L(\boldsymbol{x}) = \frac{\pi_1(\boldsymbol{x})}{\pi_0(\boldsymbol{x})}$ is the ratio of the likelihoods and $\frac{\pi_1 p_1(\boldsymbol{x})}{\pi_0 p_0(\boldsymbol{x})}$ is the ratio of the posterior probabilities.

Note also that the quantity $c_{i0}\pi_0(\boldsymbol{x}) + c_{i1}\pi_1(\boldsymbol{x})$ is the average cost incurred by choosing the hypothesis H_i given that $\boldsymbol{X} = \boldsymbol{x}$. This quantity is called the *posterior cost* of choosing H_i given the observation $\boldsymbol{X} = \boldsymbol{x}$. Thus, the Bayes rule makes its decision by choosing the hypothesis that yields the minimum posterior cost.

This test plays a central role in the theory of hypothesis testing. It computes the likelihood ratio $L(\boldsymbol{x})$ for a given observed value $\boldsymbol{X} = \boldsymbol{x}$ and then makes its decision by comparing this ratio to a threshold τ_1, *i.e.*;

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{if } L(\boldsymbol{x}) \geq \tau_1 \\ 0 & \text{if } L(\boldsymbol{x}) < \tau_1 \end{cases} \qquad (2.14)$$

A commonly cost assignment is the *uniform costs* given by

$$c_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases} \qquad (2.15)$$

The Bayes risk δ then becomes

$$r(\delta) = \pi_0 P_0(\Gamma_1) + \pi_1 P_1(\Gamma_0) = \pi_0 P_{01} + \pi_1 P_{10} \qquad (2.16)$$

Noting that $P_i(\Gamma_j)$ is the probability of choosing H_j when H_i is true, $r(\delta)$ in this case becomes the average *probability of error* incurred by the rule δ. This decision rule is then a minimum probability of error decision scheme.

Note also that with the uniform cost coefficients (2.15) the decision rule can be rewritten as

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \pi_1(\boldsymbol{x}) \geq \pi_0(\boldsymbol{x}) \\ 0 & \text{if } \pi_1(\boldsymbol{x}) < \pi_0(\boldsymbol{x}) \end{cases} \tag{2.17}$$

This test is called the *maximum a posteriori (MAP)* decision scheme.

Example : Detection of a constant signal in a Gaussian noise
Let consider

$$\begin{cases} H_0 & X = \epsilon \\ H_1 & X = \mu + \epsilon \end{cases} \tag{2.18}$$

where ϵ is a Gaussian random variable with zero mean and a known variance σ^2 and where $\mu > 0$ is a known constant. In terms of distributions we can rewrite these hypotheses as

$$\begin{cases} H_0 & X \sim \mathcal{N}\left(0, \sigma^2\right) \\ H_1 & X = \mathcal{N}\left(\mu, \sigma^2\right) \end{cases} \tag{2.19}$$

where $\mathcal{N}\left(\mu, \sigma^2\right)$ means

$$\mathcal{N}\left(\mu, \sigma^2\right) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \tag{2.20}$$

It is then easy to calculate the likelihood ratio $L(x)$

$$L(x) = \frac{p_1(x)}{p_0(x)} = \exp\left[\frac{\mu}{\sigma^2}(x - \mu/2)\right] \tag{2.21}$$

Thus, the Bayes test for these hypotheses becomes

$$\delta(x) = \begin{cases} 1 & \text{if } L(x) \geq \tau \\ 0 & \text{if } L(x) < \tau \end{cases} \tag{2.22}$$

where τ is an appropriate threshold. We can remark that $L(x)$ is a striclty increasing function of x, so comparing $L(x)$ to a threshold τ is equivalent to comparing x itself to another threshold $\tau' = L^{-1}(\tau) = \frac{\sigma^2}{\mu}\log(\tau) + \mu/2$:

$$\delta(x) = \begin{cases} 1 & \text{if } x \geq \tau' \\ 0 & \text{if } x < \tau' \end{cases} \tag{2.23}$$

where L^{-1} is the inverse function of L.

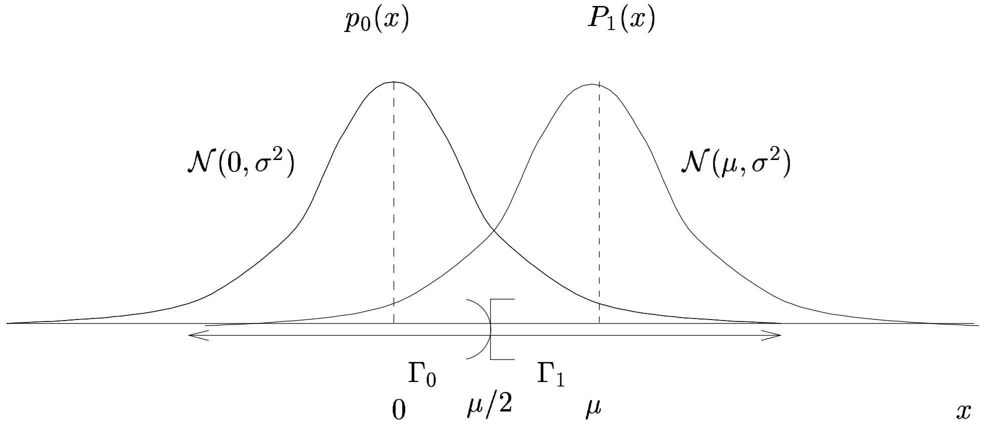


Figure 2.1: Location testing with Gaussian errors, uniform costs and equal priors.

In the special case of uniform costs and equal priors, we have $\tau = 1$ and so $\tau' = \mu/2$. Then, it is not difficult to show that the conditional probabilities are

$$P_{i1} = \Pr\{X = x \in \Gamma_1 | H = H_j\} = P_j(\Gamma_1) = \int_{\tau'}^{\infty} p_j(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \begin{cases} 1 - \Phi\left(\frac{\mu}{2\sigma}\right) & \text{for } j = 1 \\ 1 - \Phi\left(-\frac{\mu}{2\sigma}\right) & \text{for } j = 0 \end{cases} \tag{2.24}$$

and the minimum Bayes risk $r(\delta)$ is

$$r(\delta) = 1 - \Phi\left(\frac{\mu}{2\sigma}\right). \tag{2.25}$$

This is a decreasing function of $\frac{\mu}{\sigma}$, a quantity which is a simple version of the signal to noise ratio.

Summary of notations for binary hypothesis testing

Hypotheses H	H_0	H_1
Processes	P_0	P_1
Conditional densities or likelihood functions	$p_0(\boldsymbol{x})$	$p_1(\boldsymbol{x})$
Observation space Γ partition	Γ_0	Γ_1
Decisions $\delta(\boldsymbol{x})$	$\delta_0(\boldsymbol{x})$	$\delta_1(\boldsymbol{x})$
Conditional probabilities $P_{ij} = \int_{\Gamma_i} p_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$	P_{00}, P_{01}	P_{10}, P_{11}
Costs c_{ij}	c_{00}, c_{01}	c_{10}, c_{11}
Conditional risks $R_j = \sum_i c_{ij} P_{ij}$	$R_0 = c_{00}P_{00} + c_{10}P_{10}$	$R_1 = c_{01}P_{01} + c_{11}P_{11}$
Prior probabilities π_j	π_0	π_1
Posterior probabilities $\pi_j(\boldsymbol{x}) = \frac{p_j(\boldsymbol{x})\pi_j}{m(\boldsymbol{x})}$	$\pi_0(\boldsymbol{x})$	$\pi_1(\boldsymbol{x})$
Joint probabilities $Q_{ij} = \pi_j P_{ij}$	Q_{00}, Q_{01}	Q_{10}, Q_{11}
Posterior costs $\bar{c}_i(\boldsymbol{x}) = \sum_j c_{ij}\pi_j(\boldsymbol{x})$	$\bar{c}_0(\boldsymbol{x}) = c_{00}\pi_0(\boldsymbol{x}) + c_{01}\pi_1(\boldsymbol{x})$	$\bar{c}_1(\boldsymbol{x}) = c_{10}\pi_0(\boldsymbol{x}) + c_{11}\pi_1(\boldsymbol{x})$
Bayes risk $r(\delta)$	$r(\delta) = \sum_j \pi_j R_j(\boldsymbol{x})$	
Likelihoods ratio $L(\boldsymbol{x})$	$L(\boldsymbol{x}) = \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}$	
Posteriors ratio	$\frac{\pi_1(\boldsymbol{x})}{\pi_0(\boldsymbol{x})} = \frac{\pi_1 p_1(\boldsymbol{x})}{\pi_0 p_0(\boldsymbol{x})}$	
Posterior costs ratio	$\frac{\bar{c}_1(\boldsymbol{x})}{\bar{c}_0(\boldsymbol{x})} = \frac{c_{10}\pi_0(\boldsymbol{x})+c_{11}\pi_1(\boldsymbol{x})}{c_{00}\pi_0(\boldsymbol{x})+c_{01}\pi_1(\boldsymbol{x})}$	

The following equivalent tests minimize the Bayes risk $r(\delta)$

Likelihoods ratio test	$L(\boldsymbol{x}) = \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} > \tau_1 = \frac{\pi_0}{\pi_1}\frac{c_{10}-c_{00}}{c_{01}-c_{11}}$
Posteriors ratio test	$\frac{\pi_1(\boldsymbol{x})}{\pi_0(\boldsymbol{x})} > \tau_2 = \frac{c_{10}-c_{00}}{c_{01}-c_{11}}$
Posterior costs ratio test	$\frac{\bar{c}_1(\boldsymbol{x})}{\bar{c}_0(\boldsymbol{x})} > 1$

Special case of uniform costs binary hypothesis testing		
Costs $c_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$	$c_{00} = 0, \quad c_{01} = 1$	$c_{10} = 1, \quad c_{11} = 0$
Conditional risks $R_j = \sum_i c_{ij} P_{ij}$	$R_0 = P_{00}$	$R_1 = P_{11}$
Prior probabilities π_j	π_0	π_1
Posterior probabilities $\pi_j(\boldsymbol{x}) = \frac{p_j(\boldsymbol{x})\pi_j}{m(\boldsymbol{x})}$	$\pi_0(\boldsymbol{x})$	$\pi_1(\boldsymbol{x})$
Joint probabilities $Q_{ij} = \pi_j P_{ij}$	$Q_{00}, \quad Q_{01}$	$Q_{10}, \quad Q_{11}$
Posterior costs $\bar{c}_j(\boldsymbol{x}) = \sum_j c_{ij} \pi_j(\boldsymbol{x})$	$\bar{c}_0(\boldsymbol{x}) = \pi_1(\boldsymbol{x})$	$\bar{c}_1(\boldsymbol{x}) = \pi_0(\boldsymbol{x})$
Bayes risk $r(\delta)$	$r(\delta) = \sum_j \pi_j R_j(\boldsymbol{x}) = \sum_j \pi_j P_{jj}$	
Likelihoods ratio $L(\boldsymbol{x})$	$L(\boldsymbol{x}) = \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}$	
Posteriors ratio	$\frac{\pi_1(\boldsymbol{x})}{\pi_0(\boldsymbol{x})} = \frac{\pi_1 p_1(\boldsymbol{x})}{\pi_0 p_0(\boldsymbol{x})}$	
Posterior costs ratio	$\frac{\bar{c}_1(\boldsymbol{x})}{\bar{c}_0(\boldsymbol{x})} = \frac{\pi_1(\boldsymbol{x})}{\pi_0(\boldsymbol{x})}$	

The following equivalent tests minimize the Bayes risk $r(\delta)$	
Likelihoods ratio test	$L(\boldsymbol{x}) = \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} > \tau_1 = \frac{\pi_0}{\pi_1}$
Posteriors ratio test	$\frac{\pi_1(\boldsymbol{x})}{\pi_0(\boldsymbol{x})} > \tau_2 = 1$
Posterior costs ratio test	$\frac{\bar{c}_1(\boldsymbol{x})}{\bar{c}_0(\boldsymbol{x})} = \frac{\pi_1(\boldsymbol{x})}{\pi_0(\boldsymbol{x})} > 1$

2.3 Minimax binary hypothesis testing

In the previous section, we saw how the Bayesian hypothesis testing gives us a complete procedure to the hypothesis testing problems. However, in some applications, we may not be able to assign the prior probabilities $\{\pi_0, \pi_1\}$. Then, one approache is to choose arbitrarily $\pi_0 = \pi_1 = 1/2$ and continue all the Bayesian procedure as in the last section. An alternative approach is to choose another design criterion than the expected penalty $r(\delta)$. For example, we may use the conditional risks $R_0(\delta)$ and $R_1(\delta)$ and design a decision rule that minimizes, over all δ, the following criterion

$$\max\{R_0(\delta), R_1(\delta)\} \tag{2.26}$$

The decision rule based on this criterion is known as the *minimax rule.*

To design this decision rule, it is useful to consider the function $r(\pi_0, \delta)$, defined for a given prior $\pi_0 \in [0, 1]$ and a given decision rule δ as the average risk

$$r(\pi_0, \delta) = \pi_0 R_0(\delta) + (1 - \pi_0) R_1(\delta) \tag{2.27}$$

Noting that $r(\pi_0, \delta)$ is a linear function of π_0, then for fixed δ, its maximum occurs at either $\pi_0 = 0$ or $\pi_0 = 1$ with the maximum value respectively either $R_1(\delta)$ or $R_0(\delta)$. So, the optimization problem of minimizing the criterion (2.26) over δ is equivalent to minimizing the quantity

$$\max_{\pi_0 \in [0,1]} r(\pi_0, \delta) \tag{2.28}$$

over δ.

Now, for each prior π_0, let δ_{π_0} denote a Bayes rule corresponding to that prior and let $V(\pi_0) = r(\pi_0, \delta_{\pi_0})$; that is $V(\pi_0)$ is the minimum Bayes risk for the prior π_0. Then, it is not difficult to show that $V(\pi_0)$ is a concave function of π_0 with $V(0) = c_{11}$ and $V(1) = c_{00}$.

Now consider the function $r(\pi_0, \delta_{\pi_0'})$ which is a straight line tangent to $V(\pi_0)$ at $\pi_0 = \pi_0'$ and parallel to $r(\pi_0, \delta)$ (see figure 2.3).

From this figure, we can see that only Bayes rules can possibly be minimax rules. Indeed, we see that the minimax rule, in this case, is a Bayes rule for the prior value $\pi_0 = \pi_L$ that maximizes V, and for this prior $r(\pi_0, \delta_{\pi_L})$ is constant over π_0 and so $R_0(\delta_{\pi_L})) = R_1(\delta_{\pi_L}))$. This decision rule (with equal conditional risks) is called an *equalizer rule.* Because π_L maximizes the minimum Bayes risk, it is called the *least-favorable prior.* Thus, in this case, a minimax decision rule is the Bayes rule for the least-favorable prior π_L.

Even if we arrived at this conclusion through an example, it can be prouved that this fact is true in all practical situations. This result is stated as the following proposition:

Suppose that π_L is a solution to $V(\pi_L) = max_{\pi_0 \in [0,1]} V(\pi_0)$. Assume further that either $R_0(\delta_{\pi_L}) = R_1(\delta_{\pi_L})$ or $\pi_L = \{0, 1\}$. Then δ_{π_L} is a minimax rule. (see V. Poor for the proof). We will be back more in details on the minimax rule in chapter x.

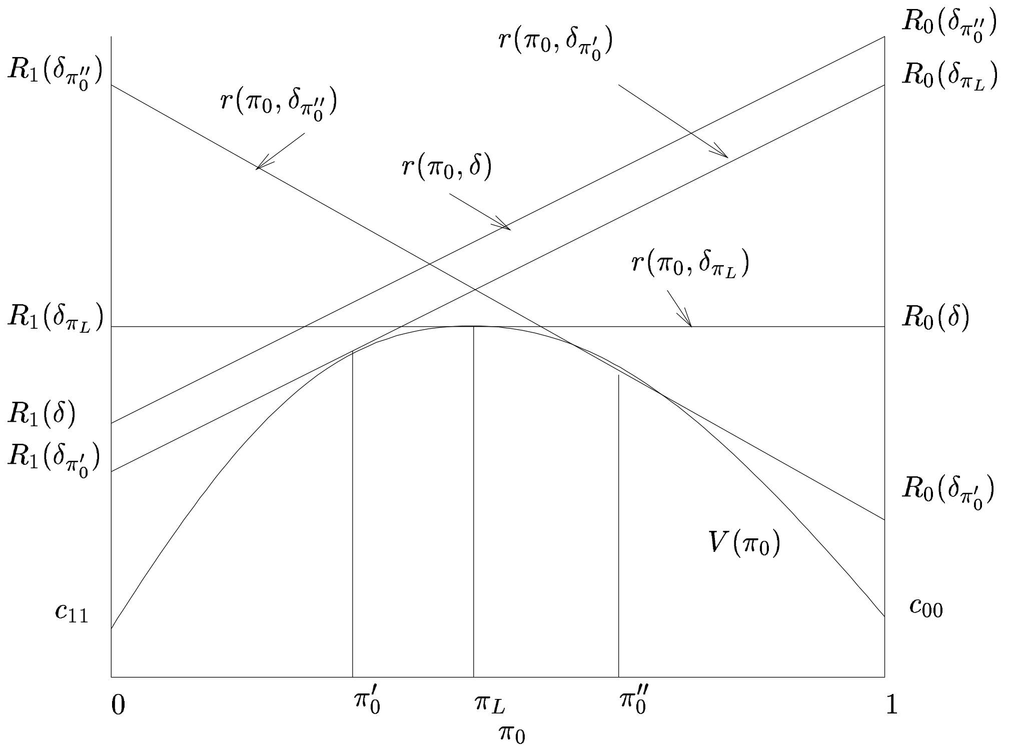


Figure 2.2: Illustration of minimax rule.

2.4 Neyman-Pearson hypothesis testing

In previous sections, we examined first the the Bayes hypothesis testing where the optimality was defined through the overall expected cost $r(\delta)$. Then, we considered the case where the prior probabilities $\{\pi_0, \pi_1\}$ are not available and described the minimax decision rule in terms of the maximum value of the conditional risks $R_0(\delta)$ and $R_1(\delta)$.

In both cases, we need to define the costs c_{ij}. In some applications, imposing a special cost structure on the decisions may not be available or not desirable. In such cases, an alternative criterion, known as the Neyman-Pearson criterion, is designed which is based on the probability of making a false decision. The main idea of this procedure is to choose one of the hypotheses as to be the main hypothesis and test other hypotheses against it. For example, in testing H_0 against H_1, two kinds of errors can be made:

- Falsely rejecting H_0 (or in this case falsely detecting H_1). This error is called either a *Type I error* or a *false alarm* or still a *false detection.*

- Falsely rejecting H_1 (or in this case falsely detecting H_0). This error is called either a *Type II error* or a *miss.*

The terms "false alarm" and "miss" come from radar applications in which H_0 and H_1 usually represent the absence and presence of a target.

For a decision rule δ, the probability of a Type I error is known as *false alarm probability* and denoted by $P_F(\delta)$. Similarly, the probability of a Type II error is called the *miss* *probability* and denoted by $P_M(\delta)$. The quantity $P_D(\delta) = 1 - P_M(\delta)$ is called as the *detection probability* or still the *power* of δ.

The Neyman-Pearson decision rule criterion is based on these quantities. It tries to place a bound on the *false alarm probability* and minimizes the *miss probability* within this constraint, *i.e.*;

$$\max P_D(\delta) \quad \text{subject to} \quad P_F(\delta) = 1 - P_D(\delta) \leq \alpha \tag{2.29}$$

where α is known as the *significance level* of the test. Thus the Neyman-Pearson decision rule criterion is to find the *most powerful α-level* test of H_0 against H_1.

Note that, in the Neyman-Pearson test, as opposed to the Bayesian and minimax tests, the two hypotheses are not considered symetrically.

The general form of the Neyman-Pearson decision rule takes the forme

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & \text{if } L(\boldsymbol{x}) > \tau \\ \gamma(\boldsymbol{x}) & \text{if } L(\boldsymbol{x}) = \tau \\ 0 & \text{if } L(\boldsymbol{x}) < \tau \end{cases} \tag{2.30}$$

where τ is a threshold.

The false alarm probability and the detection probability of a decision rule δ can be calculated respectively by

$$P_F(\delta) = \mathrm{E}_0\left\{\delta(\boldsymbol{x})\right\} = \int \delta(\boldsymbol{x}) p_0(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \tag{2.31}$$

$$P_D(\delta) = \mathrm{E}_1\left\{\delta(\boldsymbol{x})\right\} = \int \delta(\boldsymbol{x}) p_1(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \tag{2.32}$$

A parametric plot of $P_D(\delta)$ as a function of $P_F(\delta)$ is called the *receiver operation characterization* (ROCs).

Chapter 3

Basic concepts of general hypothesis testing

In previous chapters, we introduced the main basis of the simple binary hypothesis testing problem. In this chapter, we consider the more general case of the M-ary hypothesis testing. First, we give the basic definitions for the case of simple hypothesis testing for the well-known stochastic processes. Then, we consider the case of composite hypothesis testing where the stochastic processes are parametrically known. In each case, we try to make simple classification of different decision rules, describing their optimality criterion and their performances.

3.1 A general M-ary hypothesis testing problem

To consider a general M-ary hypothesis testing problem, let consider the necessary steps that any decision making procedure has to follow:

1. Get the data: observe $x(t)$ a realization of $X(t)$ in some time interval $[0, T]$.

2. Define a library of hypothesis $\{H_i, i = 1, \cdots, M\}$ where H_i is the hypothesis that the data $x(t)$ come from a stochastic processes $X_i(t)$ with either finite or infinite membership.

3. Define a performance criterion for evaluating the decisions $\{\delta_i, i = 1, \cdots, M\}$.

4. If possible, define a probability measure determining the *a priori* probabilities of stochastic processes in the library.

5. Use all the available assets to formulate a general optimization problem whose solution is a decision.

Evidently, the nature of the optimization problem and the subsequent decisions vary significantly with the specifities of the library of the stochastic processes, with the availability of the *a priori* probability distribution on these stochastic processes, the necessary performance criterion to optimize and the possibility of controling the observation time $[0, T]$ dynamically.

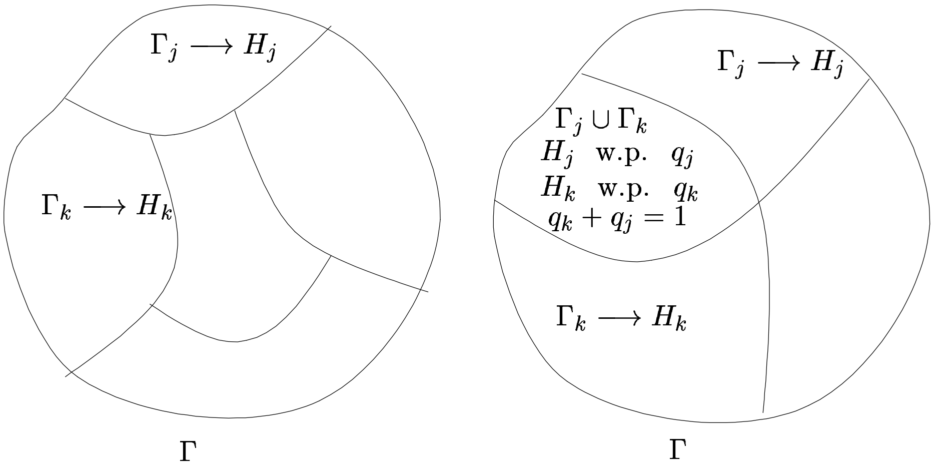

$\Gamma_j \longrightarrow H_j$
$\Gamma_k \longrightarrow H_k$
Γ
$\Gamma_j \longrightarrow H_j$
$\Gamma_j \cup \Gamma_k$
H_j w.p. q_j
H_k w.p. q_k
$q_k + q_j = 1$
$\Gamma_k \longrightarrow H_k$
Γ

Figure 3.1: Partition of the observation space and deterministic or probabilistic decision rules.

For any fixed specifications on the above issues, the decision then depends on the data. This is what is called the *decision rule* or *test*.

We can distinguish the following special cases:

- If the library has a finite number of members, the decision process is classified as *hypothesis testing*.

- If this number is only two, then the decision process is called *detection*.

- If the stochastic processes are well-known, the hypothesis testing is called *simple*. If they are defined parametrically, then the decision process is called *parametric*, if not, it is called *nonparametric*.

3.1.1 Deterministic or probabilistic decision rules

A decision rule $\delta = \{\delta_j(\boldsymbol{x}), j = 1, \cdots, M\}$ subdivides the space of the observations Γ into M subspaces $\{\Gamma_j, j = 1, \cdots, M\}$. One can distinguish two types of decision rules:

- *Deterministic decision rule*:
  If these subspaces are all disjoints and for a given data set $\boldsymbol{x}$, the hypothesis H_j is decided with probability one, the decision rule is called deterministic.

- *Probabilistic decision rule*:
  If some of theses subspaces overlap and for a given data set $\boldsymbol{x}$, none of the hypothesis can be decided with probability one, then the decision rule is called probabilistic. That is, given $\boldsymbol{x}$, the hypothesis H_k is decided with probability q_k and H_j with probability q_j with $\sum_j q_j = 1$.

3.1.2 Conditional, A priori and Joint Probability Distributions

For a given decision rule δ, we can define the following probability distributions:

- Conditional probability distribution:
  $P_{ki}(\delta)$ is the conditional probability that H_k is chosen given that H_i is true. These probabilities can be calculated from the probability distribution of the stochastic process:

$$\begin{aligned} P_{ki}(\delta) &= \Pr\{H_k \text{ decided by rule } \delta | H_i \text{ true}\} \\ &= \int_\Gamma \mathrm{d}\Pr\{H_k \text{ decided and } \boldsymbol{x} \text{ observed } |H_i \text{ true}\} \\ &= \int_\Gamma \Pr\{H_k \text{ decided}|\boldsymbol{x} \text{ observed}\}\, \mathrm{d}\Pr\{\boldsymbol{x} \text{ observed } |H_i \text{ true}\} \\ &= \int_\Gamma \delta_k(\boldsymbol{x})\, \mathrm{d}F_i(\boldsymbol{x}) = \int_\Gamma \delta_k(\boldsymbol{x}) f_i(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \qquad (3.1) \end{aligned}$$

  Note that, in this derivation, we have used the theorem of total probabilities and the Bayes rule and the fact that the decision induced by the decision rule δ is independent of the true hypothesis. Note also that the decision rule δ consists of probabilities such that

$$\sum_{j=1}^{n} \delta_j(\boldsymbol{x}) = 1 \quad \forall \boldsymbol{x} \in \Gamma \qquad (3.2)$$

  Using this fact, it is easy to show that

$$\sum_{k=1}^{M} P_{ki} = 1 \quad \forall i = 1, \cdots, M \qquad (3.3)$$

  This can be interpreted as: Given the true hypothesis H_i, the decision induced by the decision rule δ is restricted to one of the M hypotheses.

- A priori probability distribution:
  $\{\pi_i, i = 1, \cdots, M\}$ is a prior probability distribution on the hypothesis $\{H_i, i = 1, \cdots, M\}$. Naturally we have

$$\sum_{k=1}^{M} \pi_k = 1 \qquad (3.4)$$

- Joint probability distribution:
  Using the conditional probabilities $P_{kj}(\delta)$ and the prior probabilities π_i, we can calculate the joint probabilities $Q_{ki}(\delta)$, denoting the probabilities that H_k is decided by the decision rule δ while H_i is true. We then have:

$$\begin{aligned} Q_{ki}(\delta) &= \pi_i P_{ki}(\delta) = \pi_i \int_\Gamma \delta_k(\boldsymbol{x}) f_i(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \qquad (3.5) \\ \sum_{k=1}^{M} Q_{ki} &= \pi_i \quad \forall i = 1, \cdots, M \qquad (3.6) \end{aligned}$$

$$\sum_{k=1}^{M}\sum_{i=1}^{M} Q_{ki} = 1 \tag{3.7}$$

$$\tag{3.8}$$

3.1.3 Probabilities of false and correct detection

For a given decision rule δ, we can define the following probabilities:

- Probability of false decision:
  $P_e(\delta)$ is the probability that the decision induced by δ is erroneous, *i.e.*; H_k is decided while $H_{i\neq k}$ is true.

$$P_e(\delta) = \sum_{k\neq i} Q_{ki}(\delta) \tag{3.9}$$

- Probability of correct decision:
  $P_d(\delta)$ is the probability that the decision induced by δ is correct, *i.e.*H_k is decided while H_k is true.

$$P_d(\delta) = \sum_{k} Q_{kk}(\delta) = 1 - P_e(\delta) \tag{3.10}$$

- One can use these probabilities to define optimal decision procedures:

$$P_e(\delta^*) \leq P_e(\delta), \quad \forall \delta \in \mathcal{D} \tag{3.11}$$

$$P_d(\delta^*) \geq P_d(\delta), \quad \forall \delta \in \mathcal{D} \tag{3.12}$$

3.1.4 Penalty or costs coefficients

In addition to the M known hypotheses and their *a priori* probabilities, the analyst may be equipped with a set of real *cost* or *penalty* coefficients c_{ki} such that

$$c_{ki} \geq 0 \quad \text{and} \quad c_{ki} \geq c_{kk} \quad \forall k, i = 1, \cdots, M, \tag{3.13}$$

where c_{ki} is the penalty paied when H_k is decided while H_i is true. The implication behind the condition that each coefficient is nonnegative is that there is no gain associated with any decision, thus the term penalty and, in general, the penalties c_{ki} are chosen greater than c_{kk}.

3.1.5 Conditional and Bayes risks

For a given decision rule δ and a given set of cost functions c_{ki}, one can calculate the following quantities:

- Conditional expected penalties or conditional risks:

$$R_i(\delta) = \sum_{k=1}^{M} c_{ki} P_{ki}(\delta) \tag{3.14}$$

- Expected penalty or Bayes risk:

$$r(\delta) = \sum_{k=1}^{M}\sum_{i=1}^{M} c_{ki}\, Q_{ki}(\delta) \tag{3.15}$$

3.1.6 Bayesian and non Bayesian hypothesis testing

Different optimization problems which are classically used to define a decision process are:

- *Bayesian* hypothesis testing:
  - If a specific cost function that penalizes wrong decisions is provided, then the minimization of the *expected penalty* or the *Bayes risk* is chosen as the performance criterion.
  - If not, the *probability of making a decision error* is minimized instead. This decision process is called *ideal observation test.*
- *Non Bayesian* hypothesis testing:
  When an *a priori* probability distribution is unavailable, then
  - If a specific cost function is available, then first a least favorable *a priori* probability distribution is defined and then the expected penalty with this least favorable *a priori* probability distribution is minimized to obtain a decision rule. This decision process is what is called the *minimax* decision process.
  - If a cost function is not available then first one of the hypothesis is selected in advance as to be the most important and then the performance criterion used is the maximization of the probability of the detection of that hypothesis subject to the constraint that the probability of its false alarm does not exceed a given value α. This is what is called the *Neyman-Pearson* test procedure.

3.1.7 Admissible decision rules and stopping rule

- Admissible decision rules:
  It may happen that, for a given set of optimal criterions, there exist more than one best decision rule which satisfy these performances criterion, then these rules are called admissible.
- When a decision rule is designed, one may be intended to know how this decision rule performs with respect of the observed time interval $[0, T]$, *i.e.*; the number of data. The study of the behavior of the decision rule is called stoping rule.

All the above test procedures take a dynamic form if the observation time interval $[0, T]$ can be controlled dynamically.

3.1.8 Classification of simple hypothesis testing schemes

To summarize, let again list the richest possible set of assets available to the analyst:

i. A library of M distinct hypothesis $\{H_i, i = 1, \cdots, M\}$;

ii. A set of data $\boldsymbol{x}$ which is assumed to be a realization of a well known stochastic process under only one of these hypotheses;

iii. A prior probability distribution $\{\pi_i, i = 1, \cdots, M\}$ for the M hypotheses;

iv. A set of penalty coefficients $\{c_{ki}, k, i = 1, \cdots, M\}$;

The minimum set of assets that is (or must be) always available consists of those in i and ii and the performance criterion will suffer limitations as the number of remaining available assets decrease.

Now, to continue, first we assume that all assets in i to iv are available. Then an optimal rule δ^* is such that the expected penalty $r(\delta)$ is lower that any others, *i.e.*

$$\delta^* : r(\delta^*) \leq r(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.16}$$

This rule then guarantees a minimum average cost due to the wrong decisions. Note that this rule may not be unique. When the uniqueness is not satisfied, this means that there exist a number of admissible rules, among them, we can choose the one which is the simplest to implement.

If assets in i to iii are available, then an optimal rule δ^* can be defined by using the induced probability of error $P_e(\delta)$, or the probability of the detection,*i.e.*

$$\delta^* : P_e(\delta^*) \leq P_e(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.17}$$

or

$$\delta^* : P_d(\delta^*) \geq P_d(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.18}$$

Again, note that there may not exist a unique decision rule, but a set of admissible rules, among them, we can choose the one which is the simplest to implement.

The hypothesis testing rules based on the above criterions are called *Bayesian* due to the basic ingredient which is the availability of the prior probabilities π_i on the hypothesis space. When the asset iii is not available, then the decision rules are called *non Bayesian.*

Now assume all assets i, ii and iv are available, then the analyst can choose an arbitrary prior probability distribution $\pi = \{\pi_i\}$ and calculate the induced conditional expected penalty $R(\delta, \pi)$. Then, the decision rule

$$\delta^* : \sup_{\pi} R(\delta^*, \pi) \leq \sup_{\pi} R(\delta, \pi) \quad \forall \delta \in \mathcal{D} \tag{3.19}$$

defines admissible ones. The analyst then can choose between these admissible rules the one with the lowest complexity. This procedure, when successful, isolates the decision rule that induces the minimum maximum value of the conditional expected penalty and protects the analyst against the most costly case. This formalism and procedure is called *minimax.*

Finally, assume that only the assets in i and ii are available, Then, the main idea is to select one of the hypothesis as to be the principal and use the notion of the *power function* $P(\delta)$.

General Hypothesis Testing Schemes

Scheme	A priori	Cost	Decision rule
Bayesian	Yes	Yes	Minimization of expected penalty $r(\delta)$
	Yes	No	Minimization of error probability $P_e(\delta)$
Non Bayesian	No	Yes	Minimax test rule using conditional risks $R_j(\delta)$
	No	No	Neyman-Pearson test rule using $P_e(\delta)$ and $P_d(\delta)$

Classes of Hypothesis Testing Schemes

for Well Known Stochastic Processes.

Scheme	Assets used	Optimization function	Optimal Decision rule δ^*	Specific Name
Bayesian	i, ii, iii, iv, v	$r(\delta)$	$r(\delta^*) \leq r(\delta)$	Bayesian
	i, ii, iii, v	$P_e(\delta)$	$P_e(\delta^*) \leq P_e(\delta)$	Bayesian
Non Bayesian	i, ii, iii	$\sup_p r(\delta, \pi)$	$\sup\limits_p r(\delta^*, \pi) \leq \sup\limits_p r(\delta, \pi)$	Minimax
	ii, iii	$P_d(\delta)$ subject to $P_e(\delta) \leq \alpha$	$P_d(\delta^*) \geq P_d(\delta)$ and $P_e(\delta) \leq \alpha$	Neyman-Pearson

3.2 Composite hypothesis

Now assume that, the hypothesis H_i means that $\boldsymbol{x}$ is a realization of the process $\boldsymbol{X}_i$ but the process $\boldsymbol{X}_i$ is only parametrically known, i.e.; its probability distribution is known within a set of unknown parameters $\boldsymbol{\theta}$ so that, the prior probabilities $\{\pi_i\}$ depend on the parameter $\boldsymbol{\theta}$. Now, assume that the partitioning of the decision rule is due to the partition of the parameter space $\mathcal{T}$ of possible values of $\boldsymbol{\theta}$, i.e.;

$$\mathcal{T} = \{\mathcal{T}_1, \cdots, \mathcal{T}_M\}, \quad \cup_i \mathcal{T}_i = \mathcal{T} \tag{3.20}$$

Assume also that, for each value of $\boldsymbol{\theta}$, the stochastic process is defined through its probability distribution $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ and assume that we can define a probability density function $\pi(\boldsymbol{\theta})$ over the space $\mathcal{T}$. Then we have

$$\pi_i = \int_{\mathcal{T}_i} \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{3.21}$$

and

$$P_{k,\boldsymbol{\theta}}(\delta) = \int_{\Gamma} \delta_k(\boldsymbol{x}) f_{\boldsymbol{\theta}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{3.22}$$

where

$$\sum_{k=1}^{M} P_{k,\boldsymbol{\theta}}(\delta) = 1 \quad \forall \boldsymbol{\theta} \in \mathcal{T} \tag{3.23}$$

We can now calculate the conditional probabilities $P_{ki}(\delta)$

$$\begin{aligned} P_{ki}(\delta) &= \pi_i^{-1} \int_{\mathcal{T}_i} P_{k,\boldsymbol{\theta}}(\delta)\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\ &= \left[\int_{\mathcal{T}_i} \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}\right]^{-1} \int_{\Gamma} \mathrm{d}\boldsymbol{x}\, \delta_k(\boldsymbol{x}) \int_{\mathcal{T}_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \end{aligned} \tag{3.24}$$

and the joint probabilities Q_{ki} as follows:

$$\begin{aligned} Q_{ki}(\delta) &= \pi_i P_{ki}(\delta) = \int_{\mathcal{T}_i} P_{k,\boldsymbol{\theta}}(\delta)\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\Gamma} \mathrm{d}\boldsymbol{x}\, \delta_k(\boldsymbol{x}) \int_{\mathcal{T}_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \end{aligned} \tag{3.25}$$

3.2.1 Penalty or cost functions

In addition to the M parametrically known hypotheses, we may be equiped with a set of real *penalty* or *cost* functions $c_k(\boldsymbol{\theta})$.

We can then calculate:

- Conditional expected penalty or conditional risk function:

$$R(\delta, \boldsymbol{\theta}) = \sum_{k=1}^{M} c_k(\boldsymbol{\theta}) P_{k,\boldsymbol{\theta}}(\delta) \tag{3.26}$$

$$= \int \mathrm{d}\boldsymbol{x} \sum_{k=1}^{M} \delta_k(\boldsymbol{x}) c_k(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{3.27}$$

- Expected penalty or Bayes risk:
  If $\pi(\boldsymbol{\theta})$ is available, then the expected penaly can be calculated by

$$r(\delta) = \int_{\mathcal{T}} R(\delta,\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = \sum_{k=1}^{M} \int c_k(\boldsymbol{\theta}) P_{k,\boldsymbol{\theta}}(\delta)\,\mathrm{d}\boldsymbol{\theta} \tag{3.28}$$

$$= \int \mathrm{d}\boldsymbol{x} \sum_{k=1}^{M} \delta_k(\boldsymbol{x}) \int c_k(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{x}) \pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{3.29}$$

3.2.2 Case of binary hypothesis testing

Now, consider the particular case of *binary hypothesis testing*, where there are only hypotheses, and assume that they are determined through a single parametrically known stochastic process and two disjoint subdivisions of the parameter space $\mathcal{T}$. Let note these two hypotheses H_0 and H_1 and the decision rules $\delta_0(\boldsymbol{x})$ and $\delta_1(\boldsymbol{x})$ with $\delta_0(\boldsymbol{x}) + \delta_1(\boldsymbol{x}) = 1 \quad \forall \boldsymbol{x} \in \Gamma$. Now, if we emphasize the hypothesis H_1 (detection), then $\delta_0(\boldsymbol{x}) = 1 - \delta_1(\boldsymbol{x})$, so that we can drop the indices in the decision rule and denote by $\delta(\boldsymbol{x}) = \delta_1(\boldsymbol{x})$. Now, we have

$$P_{\boldsymbol{\theta}}(\delta) = \int_{\Gamma} \delta(\boldsymbol{x})\, f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{3.30}$$

This expression represents the probability that the emphasized hypothesis H_1 is decided, conditioned on the value $\boldsymbol{\theta}$ of the vector parameter. This function is called *the power function of the decision rule.* This is due to the fact that it provides the probability with which the emphatic hypothesis is decided for each fixed parameter value $\boldsymbol{\theta}$.

3.2.3 Classification of hypothesis testing schemes for a parametrically known stochastic process

As before, we now summarize the richest possible set of assets available for a composite hypothesis testing problem:

1. A library of M distinct hypotheses $\{H_i, i = 1, \cdots, M\}$
2. A set of data $\boldsymbol{x}$ which is assumed to be a realization of a parametrically known stochastic process, the hypotheses $\{H_i, i = 1, \cdots, M\}$ corresponding to the M disjoint subdivisions of the parameter space $\mathcal{T}$.
3. A prior probability distribution $\pi(\boldsymbol{\theta})$ on the parameter space $\mathcal{T}$
4. A set of penalty functions $\{c_k(\boldsymbol{\theta}), k = 1, \cdots, M\}$ defined on $\mathcal{T}$.

First we assume that all assets in i to iv are available. Then an optimal rule δ^* is such that the expected penalty $r(\delta)$ is lower that any others, *i.e.*

$$\delta^* : r(\delta^*) \le r(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.31}$$

If assets in i to iii are available, then an optimal rule δ^* can be defined by using the induced probability of error $P_e(\delta)$, or the probability of the detection, *i.e.*;

$$\delta^* : P_e(\delta^*) \le P_e(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.32}$$

or

$$\delta^* : P_d(\delta^*) \geq P_d(\delta) \quad \forall \delta \in \mathcal{D} \tag{3.33}$$

Again, note that there may not exist a unique decision rule, but a set of admissible rules, among which we can choose the one which is the simplest to implement.

Now assume that the assets i, ii and iv are available, then the analyst can use the induced conditional expected penalty $R(\delta, \boldsymbol{\theta})$. An optimal rule would induce relatively low $R(\delta, \boldsymbol{\theta})$ values for all values of $\boldsymbol{\theta} \in \mathcal{T}$. So, if there exist two rules $\delta^{(1)}$ and $\delta^{(2)}$ such that

$$R(\delta^{(1)}, \boldsymbol{\theta}) \leq R(\delta^{(2)}, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \mathcal{T} \tag{3.34}$$

then $\delta^{(2)}$ should be rejected in the presence of $\delta^{(1)}$. The rule $\delta^{(1)}$ is said to be *uniformly superior* than the rule $\delta^{(2)}$. But, it may happen that $R(\delta^{(1)}, \boldsymbol{\theta}) \leq R(\delta^{(2)}, \boldsymbol{\theta})$ for some values of $\boldsymbol{\theta}$ and $R(\delta^{(1)}, \boldsymbol{\theta}) > R(\delta^{(2)}, \boldsymbol{\theta})$ for other values of $\boldsymbol{\theta}$. In this case, we may ask to prefer $\delta^{(1)}$ to $\delta^{(2)}$ if

$$\sup_{\boldsymbol{\theta} \in \mathcal{T}} R(\delta^{(1)}, \boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in \mathcal{T}} R(\delta^{(2)}, \boldsymbol{\theta}) \tag{3.35}$$

Thus, the selection procedure has, in general, two steps: first reject all the *uniformly inadmissible* rules, and then between the remaining ones, define the optimal rules:

$$\delta^* : \sup_{\boldsymbol{\theta} \in \mathcal{T}} R(\delta^*, \boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in \mathcal{T}} R(\delta, \boldsymbol{\theta}) \quad \forall \delta \in \mathcal{D} \tag{3.36}$$

which are admissible. Finally, the analyst then can choose between these admissible rules the one with the lowest complexity. This procedure, when successful, isolates the decision rule that induces the minimum maximum value of the conditional expected penalty and protects the analyst against the most costly case. This formalism and procedure is called *minimax*.

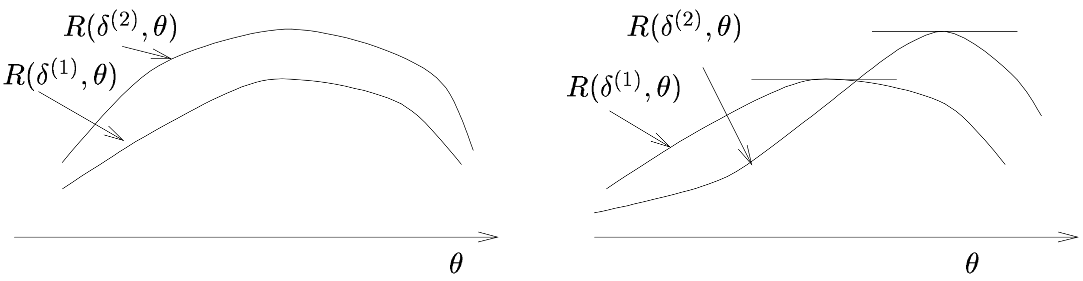

$R(\delta^{(2)}, \theta)$
$R(\delta^{(1)}, \theta)$
θ
$R(\delta^{(2)}, \theta)$
$R(\delta^{(1)}, \theta)$
θ

Figure 3.2: Two decision rules $\delta^{(1)}$ and $\delta^{(2)}$ and thier respectives risk functions. In both cases $\delta^{(2)}$ is rejected in presence of $\delta^{(1)}$.

Finally, assume that only the assets in i and ii are available. Then, the main idea is to select one of the hypotheses as to be the principal and use the notion of the *power function* $P_{\boldsymbol{\theta}}(\delta)$. Let note H_1 the emphasized hypothesis, $\mathcal{T}_1$ its associated region in $\mathcal{T}$ and $P_{\boldsymbol{\theta}}(\delta)$ the power function associated to it. It is then desirable that $P_{\boldsymbol{\theta}}(\delta)$ for any $\boldsymbol{\theta} \in \mathcal{T}_1$ has a value higher that its value for other hypotheses, *i.e.*;

$$P_{\boldsymbol{\theta} \in \mathcal{T}_1}(\delta) \geq P_{\boldsymbol{\theta} \in \mathcal{T}_j}(\delta) \tag{3.37}$$

The quantity $\sup_{\boldsymbol{\theta}\in\mathcal{T}_0} P_{\boldsymbol{\theta}}(\delta)$ is the false alarm induced by δ. The value of $P_{\boldsymbol{\theta}}(\delta)$ for a given value of $\boldsymbol{\theta} \in \mathcal{T}_1$ is the power induced by the decision rule δ. If the subspaces $\mathcal{T}_0$ and $\mathcal{T}_1$ are fixed, then the best decision rule δ^* is the one that induces the highest power subject to a false alarm constraint, *i.e.*;

$$\delta^* : P_{\boldsymbol{\theta}}(\delta^*) \leq P_{\boldsymbol{\theta}}(\delta) \quad \forall \boldsymbol{\theta} \in \mathcal{T}_1 \quad \text{subject to} \quad \sup_{\boldsymbol{\theta}\in\mathcal{T}_0} P_{\boldsymbol{\theta}}(\delta^*) \leq \alpha, \quad \forall \delta \in \mathcal{D} \tag{3.38}$$

The procedure, as in the minimax scheme, may have more than one solution.

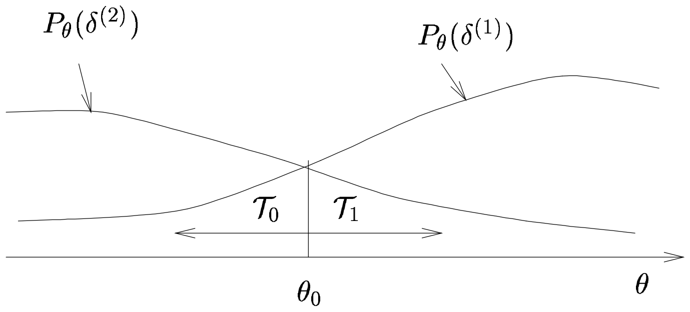


Figure 3.3: Two decision rules $\delta^{(1)}$ and $\delta^{(2)}$ and thier respectives power functions. In this case $\delta^{(1)}$ is prefered to $\delta^{(2)}$.

Classes of Hypothesis Testing Schemes

for Parametrically Known Stochastic Processes.

Scheme	Assets used	Optimization function	Optimal Estimate δ^*	Specific Name
Bayesian	i, ii, iii, iv, v	$r(\delta)$	$r(\delta^*) \leq r(\delta)$	Bayesian
	i, ii, iii, v	$P_e(\delta)$	$P_e(\delta^*) \leq P_e(\delta)$	Bayesian
Non Bayesian	i, ii, iii	$\sup_{\boldsymbol{\theta}\in\mathcal{T}} R(\delta,\boldsymbol{\theta})$	$\sup_{\boldsymbol{\theta}\in\mathcal{T}} R(\delta^*,\boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta}\in\mathcal{T}} R(\delta,\boldsymbol{\theta})$	Minimax
	ii, iii	$P_{\theta\in\Theta_1}(\delta)$ subject to $\sup_{\theta\in\Theta_0} P_\theta(\delta) \leq \alpha$	$P_{\theta\in\Theta_1}(\delta^*) \geq P_{\theta\in\Theta_1}(\delta)$ and $\sup_{\theta\in\Theta_0} P_\theta(\delta) \leq \alpha$	Neyman-Pearson

3.3 Classification of parameter estimation schemes

The basic ingredient that distinguishes the parameter estimation from the hypothesis testing is the dimension of the hypothesis space and the nature of the stochastic process corresponding to each alternative. In hypothesis testing the dimension of the hypothesis space is finite and any of the M alternatives are represented by one stochastic process. In parameter estimation, we are face to an infinite number of alternatives represented by some m dimentional vector parameter $\boldsymbol{\theta}$ that takes its values in $\mathcal{T}$.

The basic elements of parameter estimation are then the vector parameter $\boldsymbol{\theta}$ and a stochastic process $X(t)$ which is parameterized by $\boldsymbol{\theta}$ and we still can distinguish two cases:

- If for a fixed $\boldsymbol{\theta}$ the stochastic process is well-known, then we have a *parametric* parameter estimation scheme.
- If for a fixed $\boldsymbol{\theta}$ the stochastic process is a member of some class $\mathcal{F}_{\boldsymbol{\theta}}$ of processes, then we have a *non parametric* or *robust parameter estimation scheme.*

In both cases, the main assumption is that the value of the parameter and so the nature of the stochastic process remains unchanged during the observation time $[0, T]$. The main objective is then to determine the active value of the parameter $\boldsymbol{\theta}$. Given a set of data $\boldsymbol{x}$, the solution is noted $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ and is called *parameter estimate.*

Between the different criteria to measure the performances of an estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$, one can mention the following:

- *Bias* :
  For a real valued parameter vector $\boldsymbol{\theta}$, the Euclidean norm
  $$\left\| \boldsymbol{\theta} - \mathrm{E}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{X})\right] \right\|^{1/2}$$
  is called the *bias* of the estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ at the process. If the bias is zero for all $\boldsymbol{\theta} \in \mathcal{T}$, then the estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ is called *unbiased* at the process.
- *Conditional variance* :
  The quantity
  $$\mathrm{E}\left[\left\| \widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \mathrm{E}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{X})\right] \right\| \mid \boldsymbol{\theta}\right]$$
  is called the *Conditional variance* of the estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$.

In general, the bias and the conditional variance present a tradeoff. Indeed, an unbiased estimate may induce a relatively large variance, and very often, admitting a small bias may result in a significant reduction of the conditional variance.

A parameter estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ is called *efficient* at the process, if the conditional variance equals a lower bound known as the *Cramer-Rao bound.*

A more general criterion is here also the expected penalty, if we define a penalty function $c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}]$– a scalar, non negative function whose values vary as $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ vary in $\mathcal{T}$. We can then define:

- Conditional expected penalty or conditional risk function:

$$R(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}) = \mathrm{E}\left[c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{\theta}]\,|\,\boldsymbol{\theta}\right] = \int_{\Gamma} c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{\theta}]\, f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{3.39}$$

  where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the probability density function of the stochastic process defined by $\boldsymbol{\theta}$ at the point $\boldsymbol{x}$.

- Expected penalty or Bayes risk function:
  When an *a priori* probability density function $\pi(\boldsymbol{\theta})$ is available, we can calculate the expected value of $R(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathcal{T}$, and thus definec the total expected penalty or the Bayes risk function by

$$r(\widehat{\boldsymbol{\theta}}) = r(\widehat{\boldsymbol{\theta}},\pi) = \int_{\mathcal{T}} R(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{3.40}$$

Now, let try to make a classification of parameter estimation schemes. For this, we list the richest possible set of assets:

i. A parametric or nonparametric description of a stochastic process depending on a finite dimensional parameter vector $\boldsymbol{\theta}$.

ii. A set of data $\boldsymbol{x}$ which is assumed to be a realization of one of the active stochastic processes with the implicite assumption that this process remains unchanged during the observation time.

iii. A parameter space $\mathcal{T}$ where $\boldsymbol{\theta}$ takes its values.

iv. An *a priori* probability distribution $\pi(\boldsymbol{\theta})$ defined on the parameter space $\mathcal{T}$.

v. A penalty function $c[\widehat{\theta}(\boldsymbol{x}),\boldsymbol{\theta}]$ defined for each data sequence $\boldsymbol{x}$, parameter vector $\boldsymbol{\theta}$ and the estimated parameter vector $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$.

Here also, some of the assets listed above may not be available and we will see how different schemes come out from partial availability of these assets.

The minimum set of assets that is (or must be) always available consists of those in i, ii and iii and the performance criterion of the estimation will suffer limitations as the number of remaining available assets decrease.

First we assume that a parametric description of the stochastic process is available. When all the assets are available, we will have the *Bayesian parameter estimation scheme* where the performance criterion is the expected penalty or the Bayes risk

$$r(\widehat{\boldsymbol{\theta}}) = \mathrm{E}\left[c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{\theta}]\right] = \int_{\mathcal{T}}\int_{\Gamma} c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}),\boldsymbol{\theta}]\, f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\theta} \tag{3.41}$$

with respect to the estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ for all $\boldsymbol{x}$ in the observation space Γ.

The Bayesian optimal estimate is then defined as the estimate $\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x})$ which minimizes the expected penalty function, *i.e.*

$$\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}): \quad r(\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x})) \le r(\boldsymbol{\theta}(\boldsymbol{x})) \quad \forall \boldsymbol{\theta} \in \mathcal{T} \tag{3.42}$$

If assets in i to iv are available, then we can calculate the posterior probability density function of $\boldsymbol{\theta}$, using the Bayes rule:

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})} = \frac{f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})} \tag{3.43}$$

where

$$m(\boldsymbol{x}) = \int_{\mathcal{T}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{3.44}$$

and define an estimate, called *maximum a posteriori (MAP)* estimate by

$$\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}): \quad \pi(\widehat{\boldsymbol{\theta}}^*|\boldsymbol{x}) \geq \pi(\boldsymbol{\theta}|\boldsymbol{x}) \quad \forall \boldsymbol{\theta} \in \mathcal{T} \tag{3.45}$$

or written differently

$$\widehat{\boldsymbol{\theta}}^* = \arg\max_{\boldsymbol{\theta}\in\mathcal{T}} \{\pi(\boldsymbol{\theta}|\boldsymbol{x})\} \tag{3.46}$$

If assets in i to iii and v are available, then an optimal estimate $\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x})$ can be defined by using the expected conditional penalty

$$R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathrm{E}\left[c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}]|\boldsymbol{\theta}]\right] = \int_{\Gamma} c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}] f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{3.47}$$

where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the probability density function of the stochastic process defined by $\boldsymbol{\theta}$ at the point $\boldsymbol{x}$. We are then in the *minimax parameter estimation* scheme which is based on the saddle-point game formalization, with payoff function the expected penalty $r(\widehat{\boldsymbol{\theta}}, \pi)$ and with variables the parameters estimate $\widehat{\boldsymbol{\theta}}$ and the *a priori* probability density function π.

In summary, we can say that, if a minimax estimate $\widehat{\boldsymbol{\theta}}^*$ exists, it is an optimal Bayesian estimate at some least favorable *a priori* probability distribution p_0, *i.e.*

$$\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}) : \exists \pi_0 : \quad r[\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}), \pi] \leq r[\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}), \pi_0] \leq r[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \pi_0] \quad \forall \boldsymbol{\theta} \in \mathcal{T} \text{ and } \forall \pi \tag{3.48}$$

When only the assets i to iii are available, then the analyst can use the induced conditional probability density function $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. The scheme is called *maximum likelihood* and the main idea is to use the induced conditional probability density function $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ as a function $l(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\boldsymbol{x})$, called *likelihood* of the vector parameter $\boldsymbol{\theta}$ and define the maximum likelihood (ML) estimate $\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x})$ as the one who maximizes the likelihood $l(\boldsymbol{\theta})$, *i.e.*

$$\widehat{\boldsymbol{\theta}}^*(\boldsymbol{x}) : \quad l(\widehat{\boldsymbol{\theta}}^*) \geq l(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \mathcal{T} \tag{3.49}$$

All the above schemes comprise the class of parametric parameter estimation procedures with the common characteristic of the assumtion that the stochastic process that generates the data is parametrically well-known.

When, for a given vector parameter $\boldsymbol{\theta}$, the stochastic process is nonparametrically described, then the parameter estimation is called *nonparametric* or sometimes *robust.* As in minimax scheme, the robust estimation scheme uses a saddle-point game procedure, but here the payoff function originates from the likelihood. So, in this scheme, in addition to the nonparametric description of the stochastic process, the only assets ii and iii are used to define a performance criterion using the likelihood function. We will be back more in details on this scheme in future chapters.

Classes of Parameter Estimation Schemes

for Parametrically Known Stochastic Processes.

Scheme	A priori	Cost	Decision rule
Bayesian	Yes	Yes	Minimization of expected penalty $r(\widehat{\boldsymbol{\theta}})$
	Yes	No	Maximization of the posterior probability $\pi(\boldsymbol{\theta}\|\boldsymbol{x})$
Non Bayesian	No	Yes	Minimax estimation using $r(\widehat{\boldsymbol{\theta}}, \pi)$
	No	No	Maximum likelihood tests using $l(\boldsymbol{\theta})$

Classes of Parameter Estimation Schemes

Assets used	Optimization function	Optimal estimate $\widehat{\boldsymbol{\theta}}^*$	Scheme
i, ii, iii, iv, v	$r(\widehat{\boldsymbol{\theta}})$	$\widehat{\boldsymbol{\theta}}^* : \quad r(\widehat{\boldsymbol{\theta}}^*) \leq r(\widehat{\boldsymbol{\theta}}) \quad \forall \boldsymbol{\theta} \in \mathcal{T}$	Bayesian
i, ii, iii, iv	$\pi(\boldsymbol{\theta}\|\boldsymbol{x})$	$\pi(\widehat{\boldsymbol{\theta}}^*\|\boldsymbol{x}) \leq \pi(\widehat{\boldsymbol{\theta}}\|\boldsymbol{x}) \quad \forall \boldsymbol{\theta} \in \mathcal{T}$	MAP
i, ii, iv	$R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$	$\sup_{\boldsymbol{\theta}\in\mathcal{T}} R(\widehat{\boldsymbol{\theta}}^*, \boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta}\in\mathcal{T}} R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$	Minimax
i, ii	$l(\boldsymbol{\theta})$	$\widehat{\boldsymbol{\theta}}^* : \quad l(\widehat{\boldsymbol{\theta}}^*) \geq l(\widehat{\boldsymbol{\theta}}) \quad \forall \boldsymbol{\theta} \in \mathcal{T}$	Maximum likelihood
i, ii nonparametric description of the stochastic process	Based on $l(\boldsymbol{\theta})$	Appropriate saddle point optimization	Robust estimation

3.4 Summary of notations and abbreviations

- $\delta = \{\delta_k, k = 1, \cdots, M\}$
  A decision rule (or a set of possible actions)

- $\Gamma = \{\Gamma_k, k = 1, \cdots, M\}$
  The partitions of the observation space Γ corresponding to the hypotheses $\{H_k\}$ and the decision rule δ

- $\mathcal{T} = \{\mathcal{T}_k, k = 1, \cdots, M\}$
  The partitions of the parameter space $\mathcal{T}$ corresponding to the hypotheses $\{H_k\}$ and the decision rule δ

- $\{\pi_i\}$
  A prior probability distribution for the hypotheses $\{H_i\}$

- $\pi(\theta)$
  A prior probability density function for a scalar parameter θ

- $\pi(\boldsymbol{\theta})$
  A prior probability density function for a vector parameter $\boldsymbol{\theta}$

- $\{\pi_i(\boldsymbol{\theta})\}$
  Conditional prior probability density functions for the vector parameter $\boldsymbol{\theta}$ under the hypothesis H_i

- $\{r_i(\boldsymbol{\theta})\} = \{\pi_i \, \pi_i(\boldsymbol{\theta})\}$
  Unconditional prior probability density functions for the vector parameter $\boldsymbol{\theta}$ under the hypothesis H_i

- $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})$
  Conditional probability density function of the observations for a given $\boldsymbol{\theta}$

- $l(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})$
  Likelihood function of $\boldsymbol{\theta}$ for a given data $\boldsymbol{x}$

- $\pi_(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \pi(\boldsymbol{\theta})}{m(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}) \, \pi(\boldsymbol{\theta})}{m(\boldsymbol{x})}$
  Posterior probability density function of $\boldsymbol{\theta}$ given the observations $\boldsymbol{x}$

- $m(\boldsymbol{x}) = \displaystyle\int p(\boldsymbol{x}|\boldsymbol{\theta}) \, \pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$
  Marginal distribution of the observations $\boldsymbol{x}$

- $\{c_{ki}\}$
  Penalty coefficients

- $\{c_{ki}(\boldsymbol{\theta})\}$ or $\{c_k(\boldsymbol{\theta})\}$
  Penalty functions

- $\{P_{ki}(\delta)\}$
  Conditional probabilities of the decision rule δ for a well known stochastic process

- $\{P_{ki,\boldsymbol{\theta}}(\delta)\}$ or $\{P_{k,\boldsymbol{\theta}}((\delta)\}$
  Conditional probabilities of the decision rule δ for a parametrically known stochastic process

- $\{Q_{ki}(\delta) = \pi_i P_{ki}(\delta)\}$
  Probabilities of the decisions in the decision rule δ

- $P_e(\delta) = \sum_{k \neq i} Q_{ki}(\delta)$
  Probability of the error due to the decision rule δ

- $P_d(\delta) = \sum_{k} Q_{kk}(\delta) = 1 - P_e(\delta)$
  Probability of the correct detection due to the decision rule δ

- $P_{fa}(\delta) = Q_{10}(\delta)$
  Probability of false alarm in a binary hypothesis testing

- $P_{fd}(\delta) = Q_{01}(\delta)$
  Probability of false detection in a binary hypothesis testing

- $P_e(\delta) = Q_{01}(\delta) + Q_{10}(\delta)$
  Probability of the error due to the decision rule δ in a binary hypothesis testing

- $P_d(\delta) = Q_{00}(\delta) + Q_{11}(\delta)$
  Probability of the correct detection due to the decision rule δ in a binary hypothesis testing

- Conditional expected penalty or Risk function

  $$R_i(\delta) = \sum_{k=1}^{M} c_{ki} P_{ki}(\delta)$$

  for a well known stochastic process

  $$R_{\boldsymbol{\theta}}(\delta) = \sum_{k=1}^{M} c_k(\boldsymbol{\theta}) P_{k,\boldsymbol{\theta}}(\delta)$$

  for a parametrically known stochastic process

- Expected penalty or Bayes risk function

  $$r_i(\delta) = \sum_{k=1}^{M} c_{ki} Q_{ki}(\delta)$$

  for a well known stochastic process

  $$r_{\boldsymbol{\theta}}(\delta) = \sum_{k=1}^{M} c_k(\boldsymbol{\theta}) Q_{k,\boldsymbol{\theta}}(\delta)$$

  for a parametrically known stochastic process

Chapter 4

Bayesian hypothesis testing

4.1 Introduction

Let start by reminding and precising the notations and definitions. First we consider a general case and we assume that there exists a parameter vector $\boldsymbol{\theta}$ of finite dimensionality m and M stochastic processes, such that for every fixed value of $\boldsymbol{\theta} \in \mathcal{T}$, the conditional distributions $\{F_{\boldsymbol{\theta},i}(\boldsymbol{x}) = F_i(\boldsymbol{x}|\boldsymbol{\theta}),\, i = 1, \cdots, M\}$ and their coresponding densities $\{f_{\boldsymbol{\theta},i}(\boldsymbol{x}) = f_i(\boldsymbol{x}|\boldsymbol{\theta}),\, i = 1, \cdots, M\}$ are well known, for all values $\boldsymbol{x} \in \Gamma$.

We also assume to know the conditional prior probability distributions

$$\{\pi_i(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|H_i), \quad i = 1, \cdots, M\},$$

their coresponding unconditional prior distributions

$$\{r_i(\boldsymbol{\theta}) = P(H = H_i)\, \pi_i(\boldsymbol{\theta}) = \pi_i\, \pi_i(\boldsymbol{\theta}), \quad i = 1, \cdots, M\}$$

and the penalty functions $\{c_{ki}(\boldsymbol{\theta})\}$.

For a given decision rule $\delta(\boldsymbol{x}) = \{\delta_j(\boldsymbol{x}), j = 1, \cdots, M\}$ we define the expected penalty

$$r(\delta) = \int_{\mathbf{\Gamma}} \mathrm{d}\boldsymbol{x} \sum_{k=1}^{M} \delta_k(\boldsymbol{x}) \int_{\mathcal{T}} \sum_{i=1}^{M} c_{ki}(\boldsymbol{\theta}) f_{\boldsymbol{\theta},i}(\boldsymbol{x})\, r_i(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{4.1}$$

Note that, this general case reduces to the two following special cases:

- When the M stochastic processes coresponding to the M hypotheses are described through a single parametric stochastic process with M disjoint subdivisions $\{\mathcal{T}_1, \cdots, \mathcal{T}_M\}$ of the parameter space $\mathcal{T}$, then the quantities $\pi_i(\boldsymbol{\theta})$ and $r_i(\boldsymbol{\theta})$, both reduce to $\pi(\boldsymbol{\theta})$, and the quantities $\{f_{\boldsymbol{\theta},i}(\boldsymbol{x})\}$ and $\{c_{ki}(\boldsymbol{\theta})\}$ reduce to $\{f_{\boldsymbol{\theta}}(\boldsymbol{x})\}$ and $\{c_k(\boldsymbol{\theta})\}$. We then have
$$r(\delta) = \int_{\mathbf{\Gamma}} \mathrm{d}\boldsymbol{x} \sum_{k=1}^{M} \delta_k(\boldsymbol{x}) \int_{\mathcal{T}} c_k(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{x}) \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{4.2}$$

- When the M stochastic processes coresponding to the M hypotheses are all well known then we can eliminate $\boldsymbol{\theta}$ from the above equations and the quantities $f_{\boldsymbol{\theta},i}(\boldsymbol{x})$,

$r_i(\boldsymbol{\theta})$ and $\{c_{ki}(\boldsymbol{\theta})\}$ reduce respectively to $\{f_i(\boldsymbol{x})\}$, $\pi_i = \Pr\{H_i\}$ and $\{c_{ki}\}$. We then have

$$r(\delta) = \int_{\boldsymbol{\Gamma}} \mathrm{d}\boldsymbol{x} \sum_{k=1}^{M} \delta_k(\boldsymbol{x}) \sum_{i=1}^{M} c_{ki}\, f_i(\boldsymbol{x})\, \pi_i \, \mathrm{d}\boldsymbol{x} \tag{4.3}$$

Note that, in all the three above cases we can write

$$r(\delta) = \int_{\boldsymbol{\Gamma}} \mathrm{d}\boldsymbol{x} \sum_{k=1}^{M} \delta_k(\boldsymbol{x}) g_k(\boldsymbol{x}) \tag{4.4}$$

where $g_k(\boldsymbol{x})$ is given by one of the following equations:

$$g_k(\boldsymbol{x}) = \int_{\mathcal{T}} \sum_{i=1}^{M} c_{ki}(\boldsymbol{\theta}) f_{\boldsymbol{\theta},i}(\boldsymbol{x})\, r_i(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{4.5}$$

$$g_k(\boldsymbol{x}) = \int_{\mathcal{T}} c_k(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{x}) \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{4.6}$$

$$g_k(\boldsymbol{x}) = \sum_{i=1}^{M} c_{ki}\, f_i(\boldsymbol{x})\, \pi_i \, \mathrm{d}\boldsymbol{x} \tag{4.7}$$

4.2 Optimization problem

Now, we have all the ingredients to write down the optimization problem of the Bayesian hypothesis testing. Before starting, remember that for any decision rule $\delta = \{\delta_j(\boldsymbol{x}), j = 1, \cdots, M\}$, we have

$$\delta_k(\boldsymbol{x}) \geq 0, \; k = 1, \cdots, M \quad \text{and} \quad \sum_{k=1}^{M} \delta_k(\boldsymbol{x}) = 1 \quad \forall \boldsymbol{x} \in \boldsymbol{\Gamma} \tag{4.8}$$

Now, consider the following optimization problem:

$$\text{Minimize} \qquad r(\delta) = \int_{\boldsymbol{\Gamma}} \mathrm{d}\boldsymbol{x} \sum_{k=1}^{M} \delta_k(\boldsymbol{x})\, g_k(\boldsymbol{x}) \tag{4.9}$$

$$\text{subject to} \qquad \left\{ \begin{array}{ll} \delta_k(\boldsymbol{x}) \geq 0, & k = 1, \cdots, M, \\ \sum_{k=1}^{M} \delta_k(\boldsymbol{x}) = 1, \quad \forall \boldsymbol{x} \in \boldsymbol{\Gamma} & \end{array} \right. \tag{4.10}$$

where $\{g_k(\boldsymbol{x}), \; k = 1, \cdots, M\}$ are non negative functions defined on $\boldsymbol{\Gamma}$. Their expression can be given by either (4.5), (4.6) or (4.7).

This optimization problem does not have, in general, a unique solution and there may exist a whole class $\mathcal{D}^*$ of equivalent decision rules. we remember that two decision rules δ_1^* and δ_2^* are equivalent if

$$r(\delta_1^*) = r(\delta_2^*) \; \leq \; r(\delta) \quad \forall \delta \in \mathcal{D} \tag{4.11}$$

The calss $\mathcal{D}^*$ includes both random and determinist decision rules. For the reason of simplicity, in general, one chooses the non random decision rules. Noting that δ_k are then either 0 or 1 depending on the conditions $\boldsymbol{x} \in \Gamma_i$ or $\boldsymbol{x} \notin \Gamma_i$. The expression (4.9) of $r(\delta)$ becomes

$$r(\delta) = \sum_k \int_{\Gamma_k} g_k(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \tag{4.12}$$

Now, assuming that $g_k(\boldsymbol{x}) \geq 0 \; \forall \boldsymbol{x} \in \Gamma_k$, the Bayesian hypothesis testing scheme consistes of the following steps:

- Given the observations $\boldsymbol{x}$, compute

$$t(\boldsymbol{x}) = \min_k g_k(\boldsymbol{x}); \tag{4.13}$$

- Select a single $k^* = k(\boldsymbol{x})$ such that $g_{k^*(\boldsymbol{x})}(\boldsymbol{x}) = t(\boldsymbol{x})$;
- Define

$$\delta_j^*(\boldsymbol{x}) = \left\{ \begin{array}{ll} 1 & j = k^*(\boldsymbol{x}) \\ 0 & j \neq k^*(\boldsymbol{x}) \end{array} \right. \tag{4.14}$$

The function $t(\boldsymbol{x})$ together with the index $k(\boldsymbol{x})$ are called the *test* and the statistic behavior of the pair $[t(\boldsymbol{X}), k(\boldsymbol{X})]$ is called the *test statistics.*

To go more in details we consider some examples from general cases to more specific ones.

Let start with a general case. We here we assume that during any n observations the transmitted signal is exactly one of the M possibles $\{\boldsymbol{s}_i, \; i = 0, \cdots, M-1\}$. Then we have M hypotheses:

$$H_i: \quad \boldsymbol{x} \sim f_i(\boldsymbol{x}), \quad i = 0, \cdots, M-1 \tag{4.15}$$

Then, we have

$$g_k(\boldsymbol{x}) = \sum_{i=1}^{M} c_{ki} \; f_i(\boldsymbol{x}) \; \pi_i \tag{4.16}$$

and

$$t(\boldsymbol{x}) = \min_k g_k(\boldsymbol{x}) = \min_k \sum_{i=1}^{M} c_{ki} \; f_i(\boldsymbol{x}) \; \pi_i \tag{4.17}$$

Given the observation $\boldsymbol{x}$, the search for some index $k(\boldsymbol{x})$ that satisfies (4.17) can be realized via the differences

$$\sum_{i=1}^{M} (c_{ki} - c_{li}) \; f_i(\boldsymbol{x}) \; \pi_i \tag{4.18}$$

The optimal index $k^*(\boldsymbol{x})$ is such that

$$k^*(\boldsymbol{x}): \quad \sum_{i=1}^{M} (c_{k^*i} - c_{li}) \; f_i(\boldsymbol{x}) \; \pi_i \leq 0 \quad \forall l \tag{4.19}$$

If $\boldsymbol{x}$ is such that $f_0(\boldsymbol{x}) > 0$ and if $\pi_0 > 0$, then we have

$$k^*(\boldsymbol{x}): \quad \sum_{i=1}^{M} (c_{k^*i} - c_{li}) \; \frac{\pi_i}{\pi_0} \; \frac{f_i(\boldsymbol{x})}{f_0(\boldsymbol{x})} \leq 0 \quad \forall l \tag{4.20}$$

The ratio $\left\{\frac{f_i(\boldsymbol{x})}{f_0(\boldsymbol{x})}\right\}$ are called the *likelihood ratios*, so that, the procedure to obtain the decisions now consists in comparing the likelihood ratios against some thresholds. The test induced by (4.20) thus consists of a weighted sum of the likelihood ratios. This weigthed sum is called the *test function*. So, in general, the test function is compared with the threshold zero which is independent of the observation.

Note also that we can rewrite (4.20) in the two following other forms:

$$k^*(\boldsymbol{x}) : \quad \sum_{i=1}^{M} (c_{k^*i} - c_{li}) \, \frac{\pi_i(\boldsymbol{x})}{\pi_0(\boldsymbol{x})} \leq 0 \quad \forall l \tag{4.21}$$

or still

$$k^*(\boldsymbol{x}) : \quad \sum_{i=1}^{M} c_{k^*i} \, \pi_i(\boldsymbol{x}) \leq \sum_{i=1}^{M} c_{li} \, \pi_i(\boldsymbol{x}) \quad \forall l \tag{4.22}$$

The fractions $\dfrac{\pi_i(\boldsymbol{x})}{\pi_0(\boldsymbol{x})}$ are the *posterior likelihood ratios* and $\bar{c}_l(\boldsymbol{x}) = \sum_{i=1}^{M} c_{li} \, \pi_i(\boldsymbol{x})$ are the expected posterior penalties.

Further simplifications can be acheived with uniform cost functions

$$c_{ki} = \begin{cases} 0 & \text{if } k = i \\ 1 & \text{if } k \neq i \end{cases} \tag{4.23}$$

and with uniform priors

$$\pi_1 = \pi_2 = \cdots = \pi_M = \frac{1}{M}. \tag{4.24}$$

With these assumptions, the Bayesian decision rule becomes

$$k^*(\boldsymbol{x}) : \quad L_{k^*}(\boldsymbol{x}) \geq L_l(\boldsymbol{x}) \quad \forall l \tag{4.25}$$

where

$$L_i(\boldsymbol{x}) = \frac{f_i(\boldsymbol{x})}{f_0(\boldsymbol{x})} \tag{4.26}$$

Now, let consider some special cases.

4.3 Examples

4.3.1 Radar applications

One of the oldest area where the detection-estimation theory has been used is the radar applications. The main problem is to detect the presence of M knwon signals transmitted through the atmosphere. The transmission chanel is the atmosphere and it is assumed to be statistically well knwon. A simple model for the received signal is then

$$X(t) = S(t) + N(t) \tag{4.27}$$

where $N(t)$ is the additive noise due to the chanel. In the discrete case, where we assume to observe n samples of the received signal in the time period $[0, T]$, this model becomes

$$X_j = S_j + N_j, \quad j = 1, \cdots, n \tag{4.28}$$

or still

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}. \tag{4.29}$$

Consider now the case where one of the signals $\boldsymbol{s}_i$ is null. Then the M hypotheses become:

$$\begin{cases} H_0 & : \quad \boldsymbol{x} = \boldsymbol{n} \\ H_i & : \quad \boldsymbol{x} = \boldsymbol{s}_i + \boldsymbol{n}, \quad i = 1, \ldots, M-1 \end{cases} \tag{4.30}$$

Assume that we know also the conditional probability density functions $f_0(\boldsymbol{x})$ and $f_i(\boldsymbol{x})$ of the received signal under the hypotheses H_0 and H_i. The likelihood ratios $L_i(\boldsymbol{x})$ then become

$$L_i(\boldsymbol{x}) = \frac{f_i(\boldsymbol{x})}{f_0(\boldsymbol{x})} = \frac{\exp\left[\frac{-1}{2\sigma^2}(\boldsymbol{x} - \boldsymbol{s}_i)^t(\boldsymbol{x} - \boldsymbol{s}_i)\right]}{\exp\left[\frac{-1}{2\sigma^2}\boldsymbol{x}^t\boldsymbol{x}\right]} = \exp\left[\frac{-1}{2\sigma^2}(-2\boldsymbol{s}_i^t\boldsymbol{x} + \boldsymbol{s}_i^t\boldsymbol{s}_i)\right] \tag{4.31}$$

and $k^*(\boldsymbol{x})$ satisfies:

$$k^*(\boldsymbol{x}): \quad \boldsymbol{s}_{k^*}^t\boldsymbol{x} + \frac{1}{2}\boldsymbol{s}_{k^*}^t\boldsymbol{s}_{k^*} > \boldsymbol{s}_l^t\boldsymbol{x} + \frac{1}{2}\boldsymbol{s}_l^t\boldsymbol{s}_l \quad \forall l \tag{4.32}$$

Figure 4.3.1 shows the structure of this optimal test.

Indeed, if we assume that all the signals have the same energies $|\boldsymbol{s}_i|^2 = \boldsymbol{s}_i^t\boldsymbol{s}_i$, then we have

$$k^*(\boldsymbol{x}): \quad \boldsymbol{s}_{k^*}^t\boldsymbol{x} > \boldsymbol{s}_l^t\boldsymbol{x} \quad \forall l \tag{4.33}$$

Figure 4.3.1 shows the structure of this optimal test.

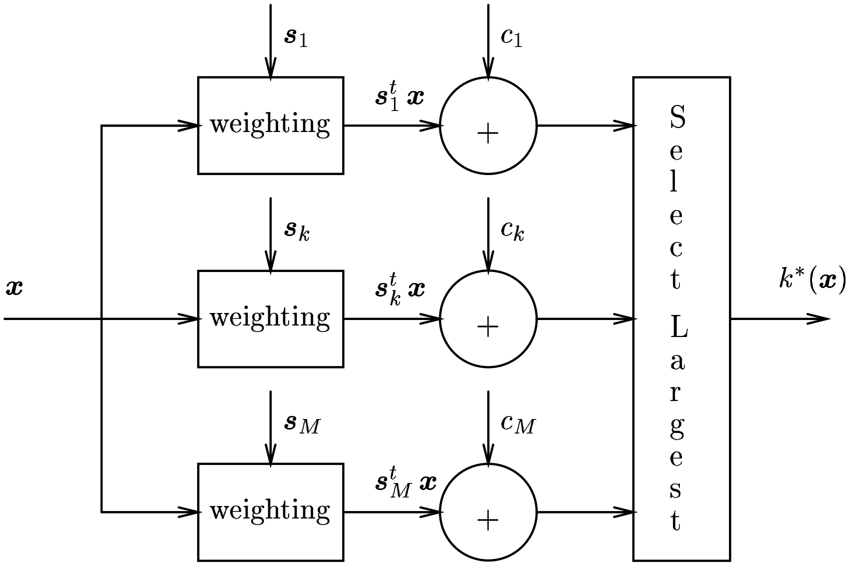


Figure 4.1: General structure of a Bayesian optimal detector.

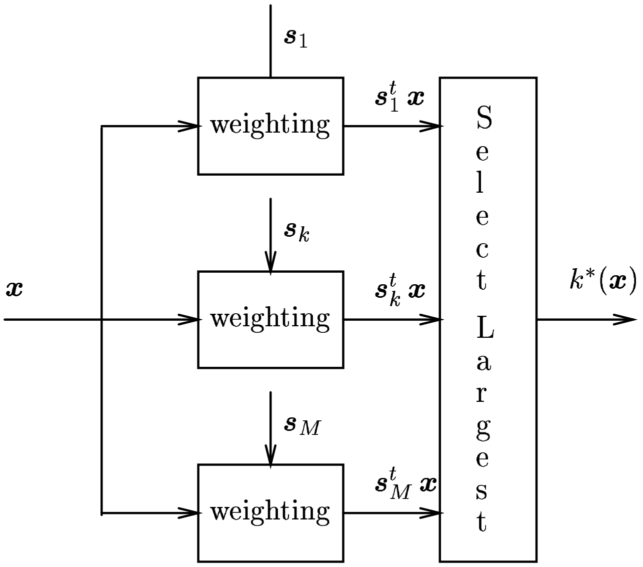


Figure 4.2: Simplified structure of a Bayesian optimal detector.

4.3.2 Detection of a known signal in an additive noise

Consider now the case where there is only one signal. So that, we have a binary detection problem:

$$\left\{\begin{array}{lcll} H_0 & : & \boldsymbol{x}=\boldsymbol{n} & \longrightarrow f_0(\boldsymbol{x})=f(\boldsymbol{x}) \\ H_1 & : & \boldsymbol{x}=\boldsymbol{s}+\boldsymbol{n} & \longrightarrow f_1(\boldsymbol{x})=f(\boldsymbol{x}-\boldsymbol{s}) \end{array}\right. \tag{4.34}$$

Then, we have:

$$L(\boldsymbol{x})=\frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} \tag{4.35}$$

General case:

The general optimal Bayesian detector structure becomes:

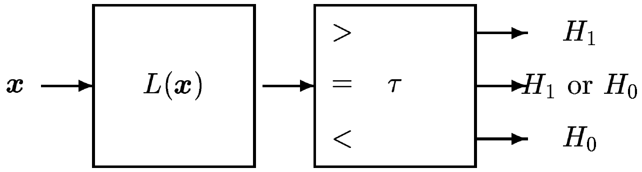


Figure 4.3: General Bayesian detector.

Case of white noise:

Now, if we assume that the noise is white, we have:

$$\left\{\begin{array}{lcll} H_0 & : & \boldsymbol{x}=\boldsymbol{n} & \longrightarrow f_0(\boldsymbol{x})=\prod_j f(x_j) \\ H_1 & : & \boldsymbol{x}=\boldsymbol{s}+\boldsymbol{n} & \longrightarrow f_1(\boldsymbol{x})=\prod_j f(x_j-s_j) \end{array}\right. \tag{4.36}$$

and consequently

$$L(\boldsymbol{x})=\prod_j L_j(x_j)=\prod_j \frac{f(x_j-s_j)}{f(x_j)} \tag{4.37}$$

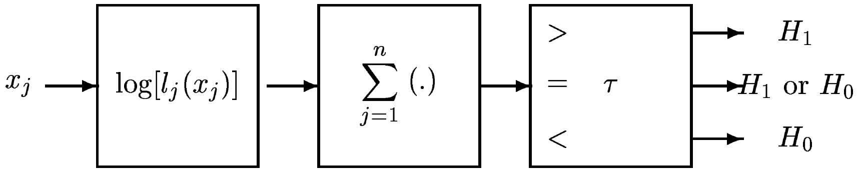


Figure 4.4: Bayesian detector in the case of i.i.d. data.

Case of Gaussian noise:

In this case we have

$$\left\{\begin{array}{lcll} H_0 & : & \boldsymbol{x}=\boldsymbol{n} & \longrightarrow f_0(\boldsymbol{x})\propto \exp\left[-\frac{1}{2\sigma^2}\boldsymbol{x}^t\boldsymbol{x}\right] \\ H_1 & : & \boldsymbol{x}=\boldsymbol{s}+\boldsymbol{n} & \longrightarrow f_1(\boldsymbol{x})\propto \exp\left[-\frac{1}{2\sigma^2}[\boldsymbol{x}-\boldsymbol{s}]^t[\boldsymbol{x}-\boldsymbol{s}]\right] \end{array}\right. \tag{4.38}$$

$$L(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} = \frac{\exp\left[\frac{-1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{s})^t(\boldsymbol{x}-\boldsymbol{s})\right]}{\exp\left[\frac{-1}{2\sigma^2}\boldsymbol{x}^t\boldsymbol{x}\right]} = \exp\left[\frac{-1}{2\sigma^2}(-2\boldsymbol{s}^t\boldsymbol{x}+\boldsymbol{s}^t\boldsymbol{s})\right] \tag{4.39}$$

We then have

$$\delta_B(\boldsymbol{x}) = \begin{cases} 1 & > \\ 0/1 \quad \text{if} \quad \boldsymbol{s}^t(\boldsymbol{x}-\boldsymbol{s}) & = \quad \tau \\ 1 & < \end{cases} \tag{4.40}$$

The detector has the following structure:

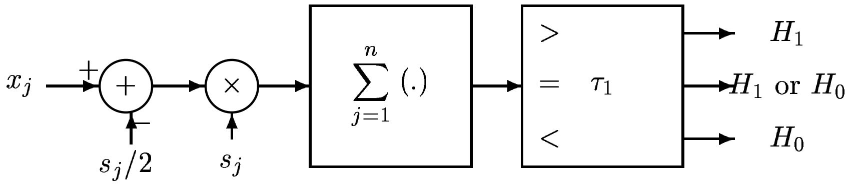


Figure 4.5: General Bayesian detector in the case of i.i.d. Gaussian data.

Note that we can rewrite (4.32) as:

$$\delta_B(\boldsymbol{x}) = \begin{cases} 1 & > \\ 0/1 \quad \text{if} \quad \boldsymbol{s}^t\boldsymbol{x} & = \quad \tau' \\ 1 & < \end{cases} \tag{4.41}$$

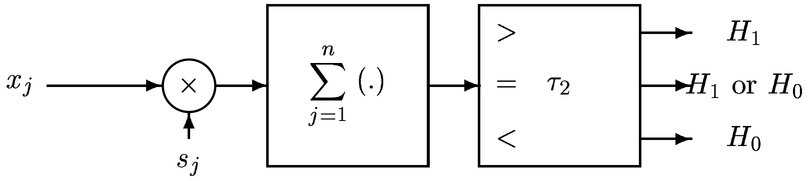


Figure 4.6: Simplified Bayesian detector in the case of i.i.d. Gaussian data.

Case of Laplacian noise:

$$H_0 : \boldsymbol{x} = \boldsymbol{n} \quad \longrightarrow \quad f_0(\boldsymbol{x}) = \prod_{j=1}^{n} \frac{\alpha}{2} \exp\left[-\alpha \sum_{j=1}^{n} |x_j|\right] \tag{4.42}$$

$$H_i : \boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n} \quad \longrightarrow \quad f_1(\boldsymbol{x}) = \prod_{j=1}^{n} \frac{\alpha}{2} \exp\left[-\alpha \sum_{j=1}^{n} |x_j - s_j|\right] \tag{4.43}$$

$$L(\boldsymbol{x}) = \prod_{j=1}^{n} L_j(x_j) \tag{4.44}$$

with

$$
\begin{aligned}
L_j(x_j) \quad &= \quad = \frac{f_1(x_j)}{f_0(x_j)} = \exp\left[\alpha|x_j - s_j| + \alpha|x_j|\right] \\
&= \left\{ \begin{array}{llrcl}
\exp\left[-\alpha|s_j|\right] & \text{if} & \operatorname{sgn}(s_j)x_j & < & 0 \\
expf\,\alpha|2x_j - s_j| & \text{if} & 0 < \operatorname{sgn}(s_j)x_j & < & |s_j| \\
expf{-\alpha}|s_j| & \text{if} & \operatorname{sgn}(s_j)x_j & > & |s_j|
\end{array} \right.
\end{aligned}
\tag{4.45}
$$

where

$$
\operatorname{sgn}(x) = \begin{cases} +1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 0 \end{cases}
\tag{4.46}
$$

Considering then two cases of $s_j < 0$ and $s_j > 0$ we obtain

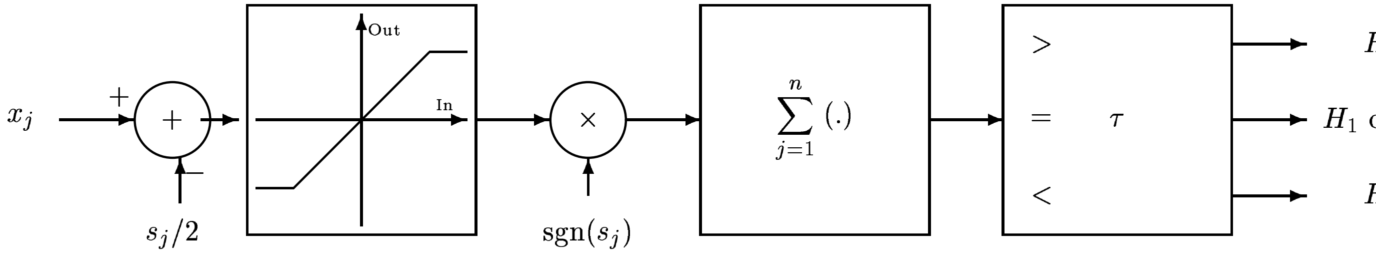


Figure 4.7: Bayesian detector in the case of i.i.d. Laplacian data.

4.4 Binary chanel transmission

In numeric signal transmission, in general, we have to transmit binary sequences. If we assume that the chanel transmits each bit separately in a memoryless fashion and that each bit s_j is transmitted correctly with probability q, then we can describe this chanel graphically as follows

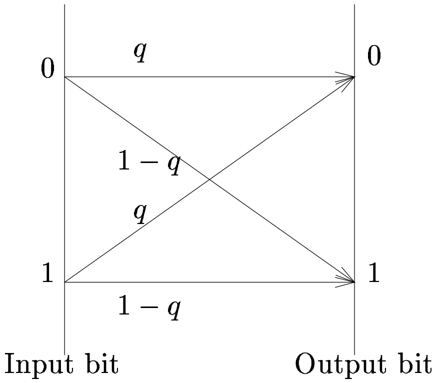


Figure 4.8: A binary chanel.

A useful binary, memoryless and symetric chanel should have a probability of correct transmission higher than the probability of incorrect transmission, $i.e.q > 0.5$ and $1-q < 0.5$.

With these assumptions on the chanel we can easily calculate the probability of observing $\boldsymbol{x}$ conditional to the transmitted sequence $\boldsymbol{s}$

$$\begin{aligned}\Pr\{\boldsymbol{x}|\boldsymbol{s}\} &= \prod_{j=1}^{n} \Pr\{x[j]|s[j]\} = \prod_{i=1}^{n} q^{1-(x[j]\oplus s[j])}(1-q)^{(x[j]\oplus s[j])} \\ &= q^n \left(\frac{1-q}{q}\right)^{\sum_{i=1}^{M}(x[j]\oplus s[j])} \qquad (4.47)\end{aligned}$$

where $\oplus$ signifies binary sum.

Now, assume that, during each observation period, only one of the M well knwon binary sequences $\boldsymbol{s}_k$ (called codewords) are transmitted. Now, we have received the binary sequence $\boldsymbol{x}$ and we want to know which one of them has been transmitted.

Indeed, if we assume $p_1 = p_2 = \cdots = p_M = 1/M$ and if we note by

$$H(\boldsymbol{s}_k, \boldsymbol{s}_l) = \frac{1}{n}\sum_{j=1}^{n} s_k[j] \oplus s_l[j] \qquad (4.48)$$

the Haming distance between the two binary words $\boldsymbol{s}_k$ and $\boldsymbol{s}_l$, then, the likelihood ratios have the following form:

$$\begin{aligned}\frac{\Pr\{\boldsymbol{x}|\boldsymbol{s}_k\}}{\Pr\{\boldsymbol{x}|\boldsymbol{s}_l\}} &= \prod_{j=1}^{n} q^{1-(x[j]\oplus s_k[j])}(1-q)^{(x[j]\oplus s_k[j])} \\ &= q^n \left(\frac{1-q}{q}\right)^{\sum_{i=1}^{M}(x[j]\oplus s_k[j])} \qquad (4.49)\end{aligned}$$

and the Bayesian optimal test becomes:

$$k^*(\boldsymbol{x}) : \quad \left(\frac{1-q}{q}\right)^{\sum_{j=1}^{n}(x[j]\oplus s_{k*}[j])-\sum_{j=1}^{n}(x[j]\oplus s_l[j])} \geq 1 \quad \forall l = 1,\ldots,M \tag{4.50}$$

Taking the logarithm of both parts we obtain the following condition on $k^*(\boldsymbol{x})$:

$$k^*(\boldsymbol{x}) : \quad \sum_{j=1}^{n}(x[j] \oplus s_{k*}[j]) - \sum_{j=1}^{n}(x[j] \oplus s_l[j]) \, \log\left(\frac{1-q}{q}\right) \geq 0 \quad \forall l = 1,\ldots,M \tag{4.51}$$

We can then discriminate two cases:

- Case 1: Let $q > .5$, which means that the transmission chanel has a higher probability of transmitting correctly than incorrectly. Then $\frac{1-q}{q} < 1$ and $k^*(\boldsymbol{x})$ satisfies

  $$k^*(\boldsymbol{x}) : \quad \sum_{j=1}^{n}(x[j] \oplus s_{k*}[j]) \leq \sum_{j=1}^{n}(x[j] \oplus s_l[j]), \quad \forall l = 1,\ldots,M \tag{4.52}$$

  or still

  $$k^* : \quad H(\boldsymbol{x}, \boldsymbol{s}_{k*}) \leq H(\boldsymbol{x}, \boldsymbol{s}_l), \quad \forall l = 1,\ldots,M \tag{4.53}$$

  The test clearly decides in favor of the codeword $\boldsymbol{s}_{k*}$ whose Hamming distance from the observed sequence $\boldsymbol{x}$ is the minimum one. This is why this detector is called *the minimum distance decoding scheme.*

  Let now the M codewords be designed so that the Haming distance between any two such codewords equals $(2d+1)/n$, where d is a positive integer, i.e.

  $$H(\boldsymbol{s}_k, \boldsymbol{s}_l) = (2d+1)/n, \quad \forall k \neq l, \; k, l = 1,\ldots,M \tag{4.54}$$

  and d such that

  $$\sum_{i=0}^{d}\binom{n}{i} \leq 2^n/M \tag{4.55}$$

  Then, via the minimum distance decoding scheme, if the distance between the received word $\boldsymbol{x}$ and the codeword $\boldsymbol{s}_k$ is at most d/n, then the codeword $\boldsymbol{s}_k$ is correctly detected and we have

  $$\begin{aligned} P_d(\boldsymbol{s}_k) &\geq \sum_{i=0}^{d}\binom{n}{i} q^{n-i}(1-q)^i && (4.56)\\ P_d = \sum_{k=1}^{M}(1/M)P_d(\boldsymbol{s}_k) &\geq \sum_{i=0}^{d}\binom{n}{i} q^{n-i}(1-q)^i && (4.57)\\ P_e = 1 - Pd &\leq 1 - \sum_{i=0}^{d}\binom{n}{i} q^{n-i}(1-q)^i && (4.58)\\ & && (4.59) \end{aligned}$$

- Case 2: In the case of $q < 0.5$, by the same analysis, the Bayesian detection scheme decides in favor of the codeword whose Hamming distance from the observed sequence is the maximum. This is not surprising because if $q < 0.5$, then with probability $1 - q > 0.5$, more than half of the codeword bits are changed in the transmission.

Chapter 5

Signal detection and structure of optimal detectors

In previous chapters we discussed some basic optimality criteria and design methods for general hypothesis testing problems. In this chapter we apply them to derive optimal procedures for the detection of signals corrupted by some noise. We consider only the discrete case.

First, we summarize the Bayesian composite hypothesis testing and focus on the binary case. Then, we describe other related tests in this particular case. Finally, through some examples with different models for the signal and the noise, we derive the optimum detector structures.

At the end, we give some basic elements of robust, sequential and non parametric detection.

5.1 Bayesian composite hypothesis testing

Consider the following composite hypothesis testing:

$$\boldsymbol{X} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{5.1}$$

and define the decision $\boldsymbol{\delta}(\boldsymbol{x})$, its associated cost function $c[\boldsymbol{\delta}(\boldsymbol{x}), \boldsymbol{\theta}]$. Then the conditional risk function is given by

$$\begin{aligned} R_{\boldsymbol{\theta}}(\boldsymbol{\delta}) &= \mathrm{E}_{\boldsymbol{\theta}}\left\{c[\boldsymbol{\delta}(\boldsymbol{X}), \boldsymbol{\theta}]\right\} = \mathrm{E}\left[c[\boldsymbol{\delta}(\boldsymbol{X}), \Theta]|\Theta = \boldsymbol{\theta}\right] \\ &= \int_{\boldsymbol{\Gamma}} c[\boldsymbol{\delta}(\boldsymbol{x}), \boldsymbol{\theta}] f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \end{aligned} \tag{5.2}$$

and the Bayes risk by

$$\begin{aligned} r(\boldsymbol{\delta}) &= \mathrm{E}\left[R_{\boldsymbol{\Theta}}(\boldsymbol{\delta}(\boldsymbol{X}))\right] = \mathrm{E}\left[\mathrm{E}\left[c[\boldsymbol{\delta}(\boldsymbol{x}), \Theta]|\Theta = \boldsymbol{\theta}\right]\right] \\ &= \int_{\boldsymbol{\mathcal{T}}} \int_{\boldsymbol{\Gamma}} c[\boldsymbol{\delta}(\boldsymbol{x}), \boldsymbol{\theta}] f_{\boldsymbol{\theta}}(\boldsymbol{x}) \pi(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Gamma}} \int_{\boldsymbol{\mathcal{T}}} c[\boldsymbol{\delta}(\boldsymbol{x}), \boldsymbol{\theta}] \pi(\boldsymbol{\theta}|\boldsymbol{x}) \,\mathrm{d}\boldsymbol{\theta} \,\mathrm{d}\boldsymbol{x} \\ &= \mathrm{E}\left[\mathrm{E}\left[c[\boldsymbol{\delta}(\boldsymbol{X}), \Theta]|\boldsymbol{X} = \boldsymbol{x}\right]\right] \end{aligned} \tag{5.3}$$

From this relation, and the fact that in general the cost function is a positive function, we can deduce that minimizing $r(\boldsymbol{\delta})$ over $\boldsymbol{\delta}$ is equivalent to minimize, for any $\boldsymbol{x} \in \boldsymbol{\Gamma}$, the mean posterior cost

$$\bar{c}[\boldsymbol{x}|\boldsymbol{\theta}] = \mathrm{E}\left[c[\boldsymbol{\delta}(\boldsymbol{X}), \boldsymbol{\Theta}]|\boldsymbol{X} = \boldsymbol{x}\right] = \int_{\boldsymbol{\mathcal{T}}} c[\boldsymbol{\delta}(\boldsymbol{x}), \boldsymbol{\theta}]\pi(\boldsymbol{\theta}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{\theta}. \tag{5.4}$$

5.1.1 Case of binary composite hypothesis testing

In this case, we have

$$\delta_B(\boldsymbol{x}) = \begin{cases} 1 & & > & \\ 0/1 & \text{if } \mathrm{E}\left[c[1, \boldsymbol{\theta}]|\boldsymbol{X} = \boldsymbol{x}\right] & = & \mathrm{E}\left[c[0, \boldsymbol{\theta}]|\boldsymbol{X} = \boldsymbol{x}\right] \\ 0 & & < & \end{cases} \tag{5.5}$$

If the two hypotheses correspond to two disjoint partitions of the parameter space $\boldsymbol{\mathcal{T}} = \{\mathcal{T}_0, \mathcal{T}_1\}$, we have

$$c[i, \boldsymbol{\theta}] = c_{ij} \quad \text{if} \quad \boldsymbol{\theta} \in \mathcal{T}_j \tag{5.6}$$

and if we consider the uniform cost function, then we have

$$\delta_B(\boldsymbol{x}) = \begin{cases} 1 & & > & \\ 0/1 & \text{if } \frac{\Pr\{\boldsymbol{\theta} \in \mathcal{T}_1|\boldsymbol{X}=\boldsymbol{x}\}}{\Pr\{\boldsymbol{\theta} \in \mathcal{T}_0|\boldsymbol{X}=\boldsymbol{x}\}} & = & \frac{c_{10}-c_{00}}{c_{01}-c_{11}} \\ 0 & & < & \end{cases} \tag{5.7}$$

Now, using the Bayes' rule, we have

$$\Pr\{\boldsymbol{\theta} \in \mathcal{T}_i|\boldsymbol{X} = \boldsymbol{x}\} = \frac{\Pr\{\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\theta} \in \mathcal{T}_i\}\Pr\{\boldsymbol{\theta} \in \mathcal{T}_i\}}{\sum_{i=0}^{1}\Pr\{\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\theta} \in \mathcal{T}_i\}\ \Pr\{\boldsymbol{\theta} \in \mathcal{T}_i\}}, \quad i = 0, 1. \tag{5.8}$$

The decision rule (5.7) becomes

$$\delta_B(\boldsymbol{x}) = \begin{cases} 1 & & > & \\ 0/1 & \text{if } L(\boldsymbol{x}) = \frac{\Pr\{\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{\theta} \in \mathcal{T}_1\}}{\Pr\{\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{\theta} \in \mathcal{T}_0\}} & = & \frac{\pi_0}{\pi_1}\frac{c_{10}-c_{00}}{c_{01}-c_{11}} \\ 0 & & < & \end{cases} \tag{5.9}$$

where

$$\pi_i = \Pr\{\boldsymbol{\theta} \in \mathcal{T}_i\}, \quad i = 0, 1. \tag{5.10}$$

The conditional probability density functions of $\boldsymbol{\theta}$ are noted $r_i(\boldsymbol{\theta})$ and given by

$$r_i(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } \boldsymbol{\theta} \notin \mathcal{T}_i \\ p(\boldsymbol{\theta})/\pi_i & \text{if } \boldsymbol{\theta} \in \mathcal{T}_i \end{cases} \tag{5.11}$$

The expression of $\Pr\{\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\theta} \in \mathcal{T}_i\}$ is then given by

$$\begin{aligned} \Pr\{\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\theta} \in \mathcal{T}_i\} &= \int_{\boldsymbol{\mathcal{T}}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\, r_i(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{\pi_i}\int_{\boldsymbol{\mathcal{T}}_i} f_{\boldsymbol{\theta}}(\boldsymbol{x})\, p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{\pi_i}\mathrm{E}_i\left\{f_{\boldsymbol{\Theta}}(\boldsymbol{x})\right\} \end{aligned} \tag{5.12}$$

where $\mathrm{E}_i\left\{f_{\boldsymbol{\Theta}}(\boldsymbol{x})\right\}$ stands for the expectation under the hypothesis H_i. We can then rewrite (5.9) as:

$$\delta_B(\boldsymbol{x}) = \begin{cases} 1 & > \\ 0/1 \quad \text{if} \quad L(\boldsymbol{x}) = \dfrac{\mathrm{E}_1\left\{f_{\boldsymbol{\Theta}}(\boldsymbol{x})\right\}}{\mathrm{E}_0\left\{f_{\boldsymbol{\Theta}}(\boldsymbol{x})\right\}} & = \quad \frac{c_{10}-c_{00}}{c_{01}-c_{11}} \\ 0 & < \end{cases} \tag{5.13}$$

With such hypotheses, the false alarm and correct detection probabilities become respectively

$$P_F(\boldsymbol{\delta},\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left\{\boldsymbol{\delta}(\boldsymbol{X})\right\} \quad \text{for} \quad \boldsymbol{\theta} \in \mathcal{T}_0 \tag{5.14}$$

$$P_D(\boldsymbol{\delta},\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left\{\boldsymbol{\delta}(\boldsymbol{X})\right\} \quad \text{for} \quad \boldsymbol{\theta} \in \mathcal{T}_1 \tag{5.15}$$

5.2 Uniform most powerful (UMP) test

The α-level uniform most powerful (UMP) test is defined as:

$$\max P_D(\boldsymbol{\delta},\boldsymbol{\theta}) \qquad \text{s.t.} \qquad P_F(\boldsymbol{\delta},\boldsymbol{\theta}) \leq \alpha \tag{5.16}$$

Unfortunately, this optimization problem may not have a solution. In those cases, we can try to design a test by following the Neyman-Pearson scheme which is summarized below.

Neyman-Pearson lemma:
Suppose the hypothesis H_0 is simple ($\mathcal{T}_0$ has only one component $\boldsymbol{\theta}_0$), the hypothesis H_1 is composite and we have a parametrically defined probability density function $p(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathcal{T}_1$. Then the most powerful α-level test for H_0 against H_1 has a unique critical region given by

$$\boldsymbol{\Gamma}_{\boldsymbol{\theta}} = \left\{\boldsymbol{x} \in \boldsymbol{\Gamma} \,|\, f_{\boldsymbol{\theta}}(\boldsymbol{x}) > \tau\, f_{\boldsymbol{\theta}_0}(\boldsymbol{x})\right\} \tag{5.17}$$

where τ depends on α. The corresponding test is given by

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & > \\ 0/1 \quad \text{if} \quad f_\theta(\boldsymbol{x}) & = \quad \tau\, f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) \\ 0 & < \end{cases} \tag{5.18}$$

5.3 Locally most powerful (LMP) test

The UMP test is too strong and the optimization problem may not have a unique solution in some more general cases. To illustrate this test, let consider the following case:

$$\begin{cases} H_0 & : \quad \theta = \theta_0 \\ H_1 & : \quad \theta = \theta > \theta_0 \end{cases} \tag{5.19}$$

The α-level locally most powerful (LMP) test is based on the development of $P_D(\boldsymbol{\delta},\theta)$ in Taylor series around the simple hypothesis parameter value θ_0

$$P_D(\delta,\theta) \simeq P_D(\delta,\theta_0) + (\theta-\theta_0)\left.\frac{\partial P_D(\delta,\theta)}{\partial\theta}\right|_{\theta=\theta_0} + \mathcal{O}(\theta-\theta_0)^2 \tag{5.20}$$

Noting that $P_F(\delta,\theta) = P_D(\delta,\theta_0)$, then the Neyman-Pearson test

$$\max\ P_D(\delta,\theta) \qquad \text{s.t.} \qquad P_F(\delta,\theta) \leq \alpha \tag{5.21}$$

becomes

$$\max\ \left.\frac{\partial P_D(\delta,\theta)}{\partial\theta}\right|_{\theta=\theta_0} \qquad \text{s.t.} \qquad P_F(\delta,\theta) \leq \alpha \tag{5.22}$$

Noting that

$$P_D(\delta,\theta) = \mathrm{E}_\theta\left\{\delta(\boldsymbol{X})\right\} = \int_{\boldsymbol{\Gamma}} \delta(\boldsymbol{x}) f_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{5.23}$$

and assuming that $f_\theta(\boldsymbol{x})$ is sufficiently regular in the neighbourhood of θ_0, we can calculate

$$P_D'(\delta,\theta_0) = \left.\frac{\partial P_D(\delta,\theta)}{\partial\theta}\right|_{\theta=\theta_0} = \int_{\boldsymbol{\Gamma}} \delta(\boldsymbol{x}) \left.\frac{\partial f_\theta(\boldsymbol{x})}{\partial\theta}\right|_{\theta=\theta_0} \mathrm{d}\boldsymbol{x} \tag{5.24}$$

In conclusion, the α-level locally most powerful (LMP) test is obtained in the same way that the α-level most powerful test by replacing $f_\theta(\boldsymbol{x})$ by $f_{\theta_0}'(\boldsymbol{x}) = \left.\frac{\partial f_\theta(\boldsymbol{x})}{\partial\theta}\right|_{\theta=\theta_0}$. The critical region of H_0 against H_1 is then given by

$$\boldsymbol{\Gamma}_{\boldsymbol{\theta}} = \left\{\boldsymbol{x}\in\boldsymbol{\Gamma} \mid f_{\theta_0}'(\boldsymbol{x}) > \tau\ f_{\boldsymbol{\theta}_0}(\boldsymbol{x})\right\} \tag{5.25}$$

and the test becomes

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & & > & \\ 0/1 & \text{if}\ \ f_{\theta_0}'(\boldsymbol{x}) & = & \tau f_{\boldsymbol{\theta}_0}(\boldsymbol{x}) \\ 0 & & < & \end{cases} \tag{5.26}$$

where τ and η depend on α.

5.4 Maximum likelihood test

In the absence of the aformentionned optimal tests, we can design a test just based on the likelihood ratios

$$\delta(\boldsymbol{x}) = \begin{cases} 1 & & > & \\ \eta & \text{if}\ \ \frac{\max_{\theta\in\mathcal{T}_1} f_\theta(\boldsymbol{x})}{\max_{\theta\in\mathcal{T}_0} f_\theta(\boldsymbol{x})} & = & \tau \\ 0 & & < & \end{cases} \tag{5.27}$$

5.5 Examples of signal detection schemes

To illustrate the common structure of these test designs, consider the following signal detection test:

$$\left\{\begin{array}{lll} H_0 & : & X_j = N_j, \\ H_1 & : & X_j = S_j + N_j \end{array}\right., \quad j = 1, \ldots, n \tag{5.28}$$

Assume that the noise is i.i.d. with known probability density function $f(x)$

$$\left\{\begin{array}{lllll} H_0 & : & X_j = N_j, & \longrightarrow & f_0(\boldsymbol{x}) = \prod_j f(x_j) \\ H_1 & : & X_j = S_j + N_j & \longrightarrow & f_1(\boldsymbol{x}) = \prod_j f(x_j - s_j) \end{array}\right., \quad j = 1, \ldots, n \tag{5.29}$$

The likelihood ratio becomes

$$L(\boldsymbol{x}) = \prod_j L_j(x_j) \quad \text{with} \quad L_j(x_j) = \frac{f(x_j - s_j)}{f(x_j)} \tag{5.30}$$

and the test becomes

$$\delta(\boldsymbol{x}) = \left\{\begin{array}{llll} 1 & & & > \\ 0/1 & \text{if} & \sum_j \log L_j(x_j) & = & \log\tau \\ 0 & & & < \end{array}\right. \tag{5.31}$$

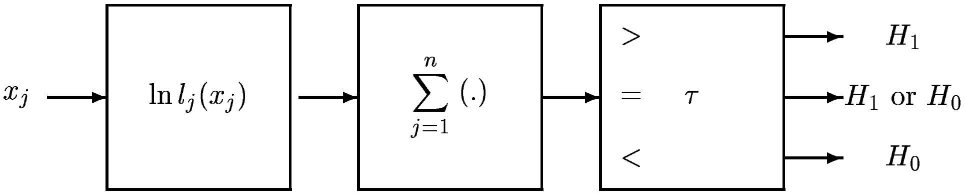


Figure 5.1: The structure of the optimal detector for an i.i.d. noise model.

5.5.1 Case of Gaussian noise

In this case we have

$$f(x_j) \propto \exp\left[-\frac{1}{2\sigma^2} x_j^2\right] \tag{5.32}$$

and the log likelihood $L_j(x_j)$ ratios become

$$\log L_j(x_j) = \log \frac{f(x_j - s_j)}{f(x_j)} = \frac{1}{\sigma^2}\left[s_j(x_j - \frac{s_j}{2})\right] \tag{5.33}$$

Noting by $\tau_1 = \sigma^2 \log\tau$, we have the following structure for the optimal detector:

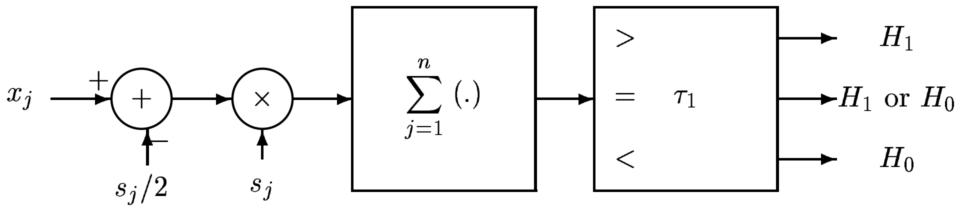


Figure 5.2: The structure of the optimal detector for an i.i.d. Gaussian noise model.

In reporting the constant value $\sum_j s_j^2$ in the treshold, we note $\tau_2 = \tau_1 + \frac{1}{2}\sum_{j=1}^{n} s_j^2$ and obtain the following scheme

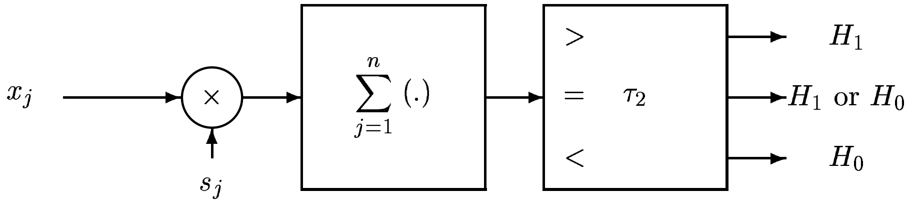


Figure 5.3: The simplified structure of the optimal detector for an i.i.d. Gaussian noise model.

5.5.2 Laplacian noise

In this case we have

$$f(x_j) \propto \exp\left[-\alpha|x_j|\right] \tag{5.34}$$

and the log likelihood $L_j(x_j)$ ratios become

$$\begin{aligned} \log L_j(x_j) &= \log \frac{f(x_j - s_j)}{f(x_j)} = -\alpha\,|x_j - s_j| + \alpha\,|x_j| \\ &= \left\{ \begin{array}{cccccc} -\alpha|s_j| & \text{if} & & \operatorname{sgn}(s_j)\,x_j & < & 0 \\ \alpha \operatorname{sgn}(x_j)\,|2x_j - s_j| & \text{if} & 0 < & \operatorname{sgn}(s_j)\,x_j & < & |s_j| \\ \alpha|s_j| & \text{if} & & \operatorname{sgn}(s_j)\,x_j & > & |s_j| \end{array} \right. \end{aligned} \tag{5.35}$$

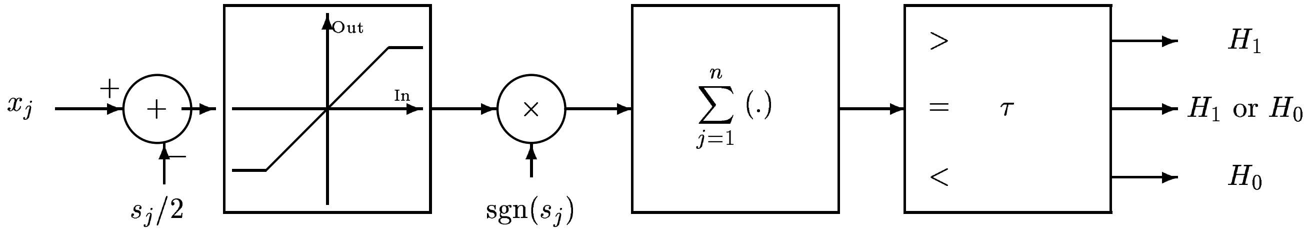


Figure 5.4: Bayesian detector in the case of i.i.d. Laplacian data.

5.5.3 Locally optimal detectors

Consider the following problem:

$$\left\{ \begin{array}{lll} H_0 & : & X_j = N_j \\ H_1 & : & X_j = N_j + \theta s_j, \quad \theta > 0 \end{array} \right. \tag{5.36}$$

We remember that the α-level uniformly optimal test for this problem is:

$$\delta(\boldsymbol{x}) = \left\{ \begin{array}{lll} 1 & & > \\ \eta & \text{if } \; L_\theta(\boldsymbol{x}) & = \; \tau \\ 0 & & < \end{array} \right. \tag{5.37}$$

and the α-level locally optimal test for this problem is:

$$\delta(\boldsymbol{x}) = \left\{ \begin{array}{lll} 1 & & > \\ \eta & \text{if } \; \left.\frac{\partial L_\theta(\boldsymbol{x})}{\partial \theta}\right|_{\theta=\theta_0} & = \; \tau \\ 0 & & < \end{array} \right. \tag{5.38}$$

where τ and η depend on α. For the case of i.i.d. noise model we have

$$L_\theta(\boldsymbol{x}) = \prod_{j=1}^{n} \frac{f(x_j - \theta s_j)}{f(x_j)} \tag{5.39}$$

Then, it is easy to show that

$$\left.\frac{\partial \log L_\theta(\boldsymbol{x})}{\partial \theta}\right|_{\theta=\theta_0} = \sum_{j=1}^{n} s_j \, g_{lo}(x_j) \tag{5.40}$$

where

$$g_{lo}(x) = -\frac{\frac{\partial f(x)}{\partial x}}{f(x)} = -\frac{f'(x)}{f(x)} \tag{5.41}$$

and the structure of a locally optimal detector is given in the following figure.

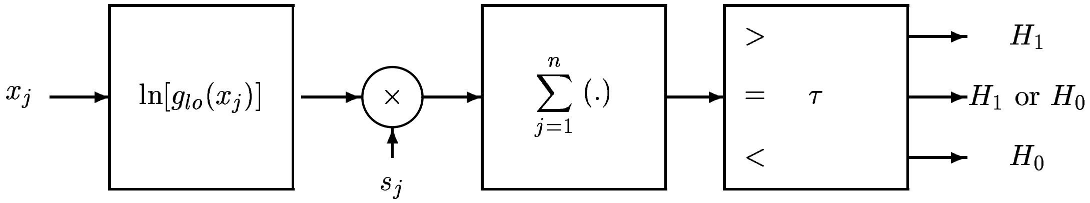


Figure 5.5: The structure of a locally optimal detector for an i.i.d. noise model.

It is clear from (5.41) that, if we know the expression of the probability density function of the noise $f(x)$, we can easily determine the expression of $g_{lo}(x)$. For example:

- For a Gaussian noise model we have

$$f(x) \propto \exp\left[-\frac{1}{2\sigma^2}x^2\right] \longrightarrow g_{lo}(x) = \frac{1}{\sigma^2}\,x \tag{5.42}$$

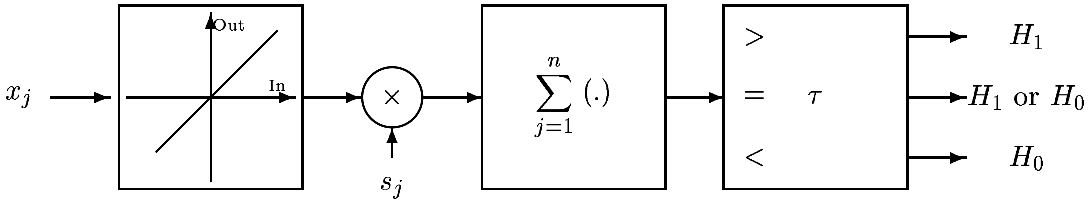


Figure 5.6: The structure of a locally optimal detector for an i.i.d. Gaussian noise model.

- For a Laplacian noise model we have

$$f(x) \propto \exp\left[-\alpha\,|x|\right] \longrightarrow g_{lo}(x) = \alpha\,\mathrm{sgn}(x) \tag{5.43}$$

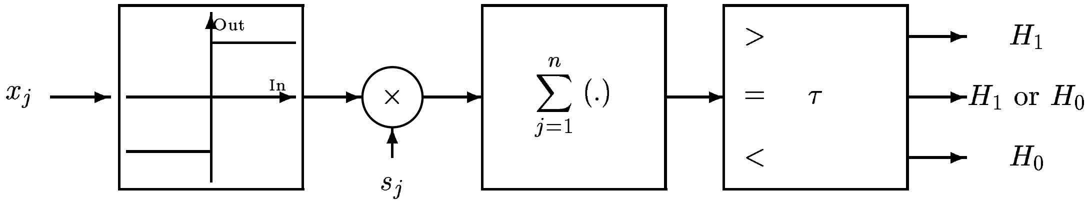


Figure 5.7: The structure of a locally optimal detector for an i.i.d. Laplacian noise model.

- For a Cauchy noise model we have

$$f(x) \propto \frac{1}{1+x^2} \longrightarrow g_{lo}(x) = \frac{2x}{1+x^2} \tag{5.44}$$

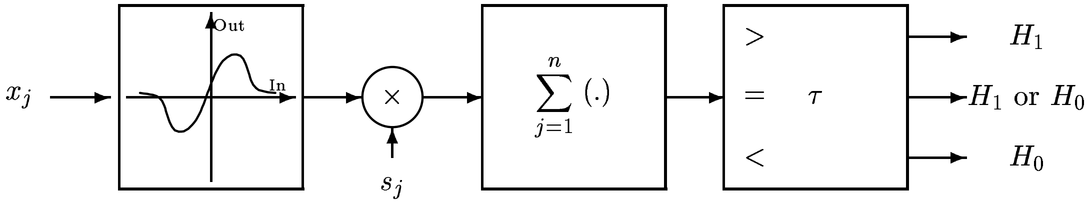


Figure 5.8: The structure of a locally optimal detector for an i.i.d. Cauchy noise model.

5.6 Detection of signals with unknown parameter

Here, we consider the case where the transmitted signals depend on some unknown parameter θ:

$$\left\{\begin{array}{lll} H_0 & : & X_j = N_j + S_{0_j}(\theta) \\ H_1 & : & X_j = N_j + S_{1_j}(\theta) \end{array}\right. , \quad \boldsymbol{N} \sim f(\boldsymbol{n}) \tag{5.45}$$

The main quantity that we need to calculate is

$$L(\boldsymbol{x}) = \frac{\mathrm{E}_1\left\{f(\boldsymbol{x} - s_{1_j}(\theta))\right\}}{\mathrm{E}_0\left\{f(\boldsymbol{x} - s_{0_j}(\theta))\right\}} \tag{5.46}$$

In the following we consider the particular case of $\boldsymbol{s}_0 = \boldsymbol{0}$, $\boldsymbol{s}_1 = \boldsymbol{s}(\theta)$ and i.i.d. noise. Then we have

$$\left\{\begin{array}{lll} H_0 & : & X_j = N_j \\ H_1 & : & X_j = N_j + S_j(\theta) \end{array}\right. , \quad N_j \sim f(n_j) \tag{5.47}$$

$$\begin{aligned} L(\boldsymbol{x}) &= \int_\Gamma \frac{f(\boldsymbol{x} - \boldsymbol{s}(\theta))}{f(\boldsymbol{x})}\, p(\theta)\, \mathrm{d}\theta \\ &= \int_\Gamma \prod_{j=1}^n \frac{f(x_j - s_j(\theta)}{f(x_j)}\, p(\theta)\, \mathrm{d}\theta \\ &= = \int L_\theta(\boldsymbol{x})\, p(\theta)\, \mathrm{d}\theta \end{aligned} \tag{5.48}$$

Example: Non coherent detection:
Consider an amplitude modulated signal

$$s_j(\theta) = a_j \sin[(j-1)\omega_c T_s + \theta], \quad j = 1, \ldots, n, \quad \omega_c T_s = m\frac{2\pi}{n} \tag{5.49}$$

where ω_c is the career frequency, T_s is the sampling rate, m is the number of periods in the observation time $[0, T = nT_s]$ and n/m is the number of samples per cycle. The carrier phase is unknown. With a uniform prior $p(\theta) = \frac{1}{2\pi}$ and the i.i.d. noise assumption we have

$$L(\boldsymbol{x}) = \frac{1}{2\pi}\int_0^{2\pi} \exp\left[\frac{1}{\sigma^2}\left[\sum_{j=1}^n x_j s_j(\theta) - \frac{1}{2}\sum_{j=1}^n s_j^2(\theta)\right]\right] \mathrm{d}\theta \tag{5.50}$$

Using the trigonometric relations:

$$\begin{aligned} \sin(a+b) &= \cos a \sin b + \sin a \cos b \quad (5.51) \\ \sin^2(a) &= \frac{1}{2} - \frac{1}{2}\cos(2a) \quad (5.52) \end{aligned}$$

we obtain:

$$\sum_{j=1}^n x_j s_j(\theta) = x_c \sin\theta + x_s \cos\theta \tag{5.53}$$

with

$$x_c \overset{\text{def}}{=} \sum_{j=1}^{n} a_j x_j \cos[(j-1)\omega_c T_s] \tag{5.54}$$

$$x_s \overset{\text{def}}{=} \sum_{j=1}^{n} a_j x_j \sin[(j-1)\omega_c T_s] \tag{5.55}$$

From (5.49) we have

$$\sum_{j=1}^{n} s_j^2(\theta) = \frac{1}{2}\sum_{j=1}^{n} a_j^2 + \frac{1}{2}\sum_{j=1}^{n} a_j^2 \cos[2(j-1)\omega_c T_s + 2\theta] \tag{5.56}$$

The second term is, in general either equal to zero or negligeable with respect to the first term.

Noting $\overline{a^2} = \frac{1}{n}\sum_{j=1}^{n} a_j^2$, we obtain

$$\begin{aligned} L(\boldsymbol{x}) &= \exp\left[-\frac{n\overline{a^2}}{4\sigma^2}\right] \frac{1}{2\pi} \int_0^{2\pi} \exp\left[\frac{1}{\sigma^2}\left[x_c \sin\theta + x_s \cos\theta\right]\right] \mathrm{d}\theta && (5.57) \\ &= \exp\left[-\frac{n\overline{a^2}}{4\sigma^2}\right] \frac{1}{2\pi} I_0\left(\frac{r}{\sigma^2}\right) && (5.58) \end{aligned}$$

where I_0 is the zeroth-order Bessel function, which is monotone. Then the detection rule becomes

$$\left\{\begin{array}{lcc} 1 & & > \\ \gamma & \text{if } L(\boldsymbol{x}) & = & 1 \\ 0 & & < \end{array}\right. \longrightarrow \left\{\begin{array}{lcc} 1 & & > \\ \gamma & \text{if } r & = & \tau' = \sigma^2 I_0^{-1}\left(\tau \exp\left[\frac{n\overline{a^2}}{4\sigma^2}\right]\right) \\ 0 & & < \end{array}\right. \tag{5.59}$$

The structure of this detector is given in the following figure.

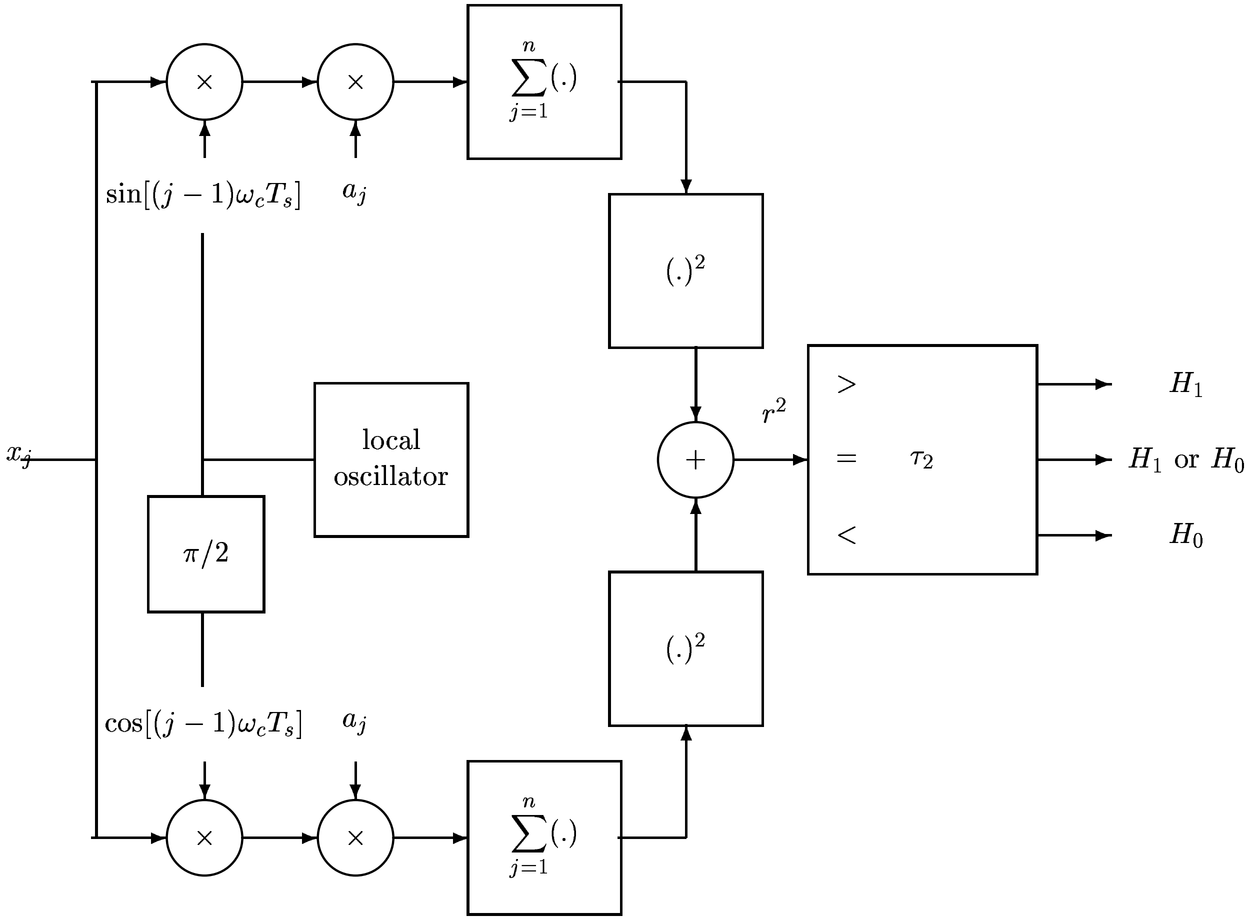


Figure 5.9: Coherent detector.

Performance analysis:

We need to calculate the probabilities such as

$$P_j(R > \tau') = P_j(R^2 > \tau'^2), \quad j = 0, 1 \tag{5.60}$$

with $R^2 = X_c^2 + X_s^2$.

Note that X_c and X_s are linear combinations of X_j. So, if X_j are Gaussian, X_c and X_s are Gaussian too.

Under the hypothesis H_0, we have

$$\begin{aligned}
\mathrm{E}\left[X_c|H_0\right] &= \mathrm{E}\left[X_s|H_0\right] = 0 && (5.61)\\
\mathrm{Var}\left[X_s|H_0\right] &= \mathrm{Var}\left[X_s|H_0\right] = \frac{n\sigma^2\overline{a^2}}{2} && (5.62)\\
\mathrm{Cov}\left\{X_s, X_c|H_0\right\} &= 0 && (5.63)
\end{aligned}$$

and

$$P_0(\Gamma_1) = \iint_{r^2=x_c^2+x_s^2\geq\tau'^2} \frac{1}{n\pi\sigma^2\overline{a^2}} \exp\left[-\frac{1}{n\pi\sigma^2\overline{a^2}}(x_c^2+x_s^2)\right] \mathrm{d}x_c\,\mathrm{d}x_s \tag{5.64}$$

With the cartesian to polar coordinate change $(x_c, x_s) \longrightarrow (r, \theta)$ we obtain

$$\begin{aligned}
P_0(\Gamma_1) &= \frac{1}{n\pi\sigma^2\overline{a^2}} \int_0^{2\pi}\int_{\tau'}^{\infty} r\exp\left[-\frac{r^2}{n\sigma^2\overline{a^2}}\right] \mathrm{d}r\,\mathrm{d}\theta \\
&= \frac{1}{n\pi\sigma^2\overline{a^2}} \exp\left[-\frac{\tau'^2}{n\sigma^2\overline{a^2}}\right] && (5.65)
\end{aligned}$$

Under the hypothesis H_1, noting that for a given value of θ $\boldsymbol{x}|\theta \sim \mathcal{N}\left(\boldsymbol{s}(\theta), \sigma^2\boldsymbol{I}\right)$ we have

$$\begin{aligned}
\mathrm{E}\left[X_c|H_1, \Theta=\theta\right] &= \frac{n\overline{a^2}}{2}\sin\theta && (5.66)\\
\mathrm{E}\left[X_s|H_1, \Theta=\theta\right] &= \frac{n\overline{a^2}}{2}\cos\theta && (5.67)\\
\mathrm{Var}\left[X_s|H_1, \Theta=\theta\right] &= \mathrm{Var}\left[X_s|H_1, \Theta=\theta\right] = \frac{n\sigma^2\overline{a^2}}{2} && (5.68)\\
\mathrm{Cov}\left\{X_s, X_c|H_1, \Theta=\theta\right\} &= 0 && (5.69)
\end{aligned}$$

and

$$\begin{aligned}
p(x_c, x_s|H_1) &= \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{n\pi\sigma^2\overline{a^2}} \exp\left[-\frac{1}{n\sigma^2\overline{a^2}}\,q(x_c, x_s; n\overline{a^2}/2, \theta)\right] \mathrm{d}\theta && (5.70)\\
&= p(x_c, x_s|H_0)\exp\left[-\frac{n\overline{a^2}}{4\sigma^2}\right] I_0\left(\frac{r}{\sigma^2}\right) && (5.71)
\end{aligned}$$

The detection probability becomes then

$$\begin{aligned}
P_D(\delta) = P_1(\Gamma_1) &= \iint_{r^2=x_c^2+x_s^2\geq\tau'^2} p(x_c, x_s|H_1)\,\mathrm{d}x_c\,\mathrm{d}x_s && (5.72)\\
& \frac{\exp\left[-\frac{n\overline{a^2}}{4\sigma^2}\right]}{n\pi\sigma^2\overline{a^2}} \int_0^{2\pi}\int_{r'}^{\infty} r\exp\left[-\frac{r^2}{n\sigma^2\overline{a^2}}\right] I_0\left(\frac{r}{\sigma^2}\right) \mathrm{d}r\,\mathrm{d}\theta && (5.73)
\end{aligned}$$

Noting $b^2 = \frac{n\overline{a^2}}{2\sigma^2}$ and $\tau_0 = \frac{\tau}{b\sigma^2}$ and changing the variable $x = \frac{r}{b\sigma^2}$ we obtain

$$P_D(\delta) = P_1(\Gamma_1) \quad = \quad \int_{r_0}^{\infty} x \exp\left[-\frac{1}{2}(x^2+b^2)\right] I_0(bx)\,\mathrm{d}x \overset{\text{def}}{=} Q(b,\tau_0) \tag{5.74}$$

$Q(b,\tau_0)$ is called Marcum's Q-function.

Note also that $P_f(\delta) = Q(0,\tau_0)$. So for a α-level Neyman-Pearson detection test we have $\tau' = \left[n\,\sigma^2\overline{a^2}\log(1/\alpha)\right]^{\frac{1}{2}}$ and the probability of detection is given by

$$P_D(\delta) = Q\left[b,\ 2[\log(1/\alpha)]^{\frac{1}{2}}\right] \tag{5.75}$$

Note also that

$$\mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n} s_j^2(\theta)\right] = \overline{a^2}/2 \tag{5.76}$$

so, $b^2 = \frac{n\overline{a^2}}{2\sigma^2}$ is a measure of S/N ratio.

5.7 sequential detection

$$\Gamma = \{X_j, j = 1, 2, \ldots\} \tag{5.77}$$

$$\left\{ \begin{array}{ll} H_0 & : X_j \sim P_0 \\ H_1 & : X_j \sim P_1 \end{array} \right. \tag{5.78}$$

A sequential decision rule is a pair of sequences $(\boldsymbol{\Delta}, \boldsymbol{\delta})$ where :

- $\boldsymbol{\Delta} = \{\Delta_j, j = 0, 1, 2, \ldots\}$ is the sequence of stopping rules ;
- $\boldsymbol{\delta} = \{\delta_j, j = 0, 1, 2, \ldots\}$ is the sequence of decision rules ;
- $\Delta_j(x_1, \ldots, x_j)$ is a function from $\mathbb{R}^j$ to $\{0, 1\}$;
- $\delta_j(x_1, \ldots, x_j)$ is a decision rule on $(\mathbb{R}^j, \mathcal{B}^j)$;
- If $\boldsymbol{\Delta}_n(x_1, \ldots, x_n) = 0$ we take another sample ;
- If $\boldsymbol{\Delta}_n(x_1, \ldots, x_n) = 1$ we stop sampling and make a decision.
- $N = \min\{n | \Delta_n(x_1, \ldots, x_n) = 1\}$ is the stopping time ;
- $\boldsymbol{\Delta}_N(x_1, \ldots, x_N)$ is the terminal decision rule.
- (Δ_0, δ_0) correspond to the situation where we have not yet observed any data. $\Delta_0 = 0$ means take at least one sample before making a decision. $\Delta_0 = 1$ means make a decision without taking any sample.

Note that N is a random variable depending on the data sequence. The terminal decision rule $\boldsymbol{\delta}_N(x_1, \ldots, x_N)$ tells us which decision to make when we stop sampling.

The fixed-sample-size N decision rule can be defined as the following sequential detection rule:

$$\boldsymbol{\Delta}_j(x_1, \ldots, x_j) = \left\{ \begin{array}{ll} 0 & \text{if} \quad j \neq N \\ 1 & \text{if} \quad j = N \end{array} \right. \tag{5.79}$$

$$\boldsymbol{\delta}_j(x_1, \ldots, x_j) = \left\{ \begin{array}{ll} \boldsymbol{\delta}(x_1, \ldots, x_n) & \text{if} \quad j = N \\ \text{arbitrary} & \text{if} \quad j \neq N \end{array} \right. \tag{5.80}$$

In the following we consider only the binary hypothesis testing and we analyse the Bayesian approach with the prior distribution $\{\pi_0 = 1 - \pi_1, \pi_1\}$ and the uniform cost function. We assume that we can have an infinite number of i.i.d. observations at our disposal. However, we should assign a cost $c > 0$ to each sample, so that the cost of taking n samples is nc.

With these assumptions, the conditional risks for a given sequential decision rule are:

$$R_0(\boldsymbol{\Delta}, \boldsymbol{\delta}) = \mathrm{E}_0\left\{\boldsymbol{\delta}(x_1, \ldots, x_n)\right\} + \mathrm{E}_0\left\{cN\right\} \tag{5.81}$$

$$R_1(\boldsymbol{\Delta}, \boldsymbol{\delta}) = 1 - \mathrm{E}_1\left\{\boldsymbol{\delta}(x_1, \ldots, x_n)\right\} + \mathrm{E}_1\left\{cN\right\} \tag{5.82}$$

where the subscripts denote the hypotheses under which the expectation is computed and N is the stopping time. The Bayes risk is thus given by

$$r(\boldsymbol{\Delta}, \boldsymbol{\delta}) = (1 - \pi_1)R_0(\boldsymbol{\Delta}, \boldsymbol{\delta}) + \pi_1 R_1(\boldsymbol{\Delta}, \boldsymbol{\delta}) \tag{5.83}$$

and the sequential Bayesian rule is the one which minimizes $r(\boldsymbol{\Delta}, \boldsymbol{\delta})$.

To analyse the structure of this optimal rule we define

$$V^*(\pi_1) \overset{\text{def}}{=} \min_{\substack{\boldsymbol{\Delta}, \boldsymbol{\delta} \\ \Delta_0 = 0}} r(\boldsymbol{\Delta}, \boldsymbol{\delta}), \quad 0 \leq \pi_1 \leq 1. \tag{5.84}$$

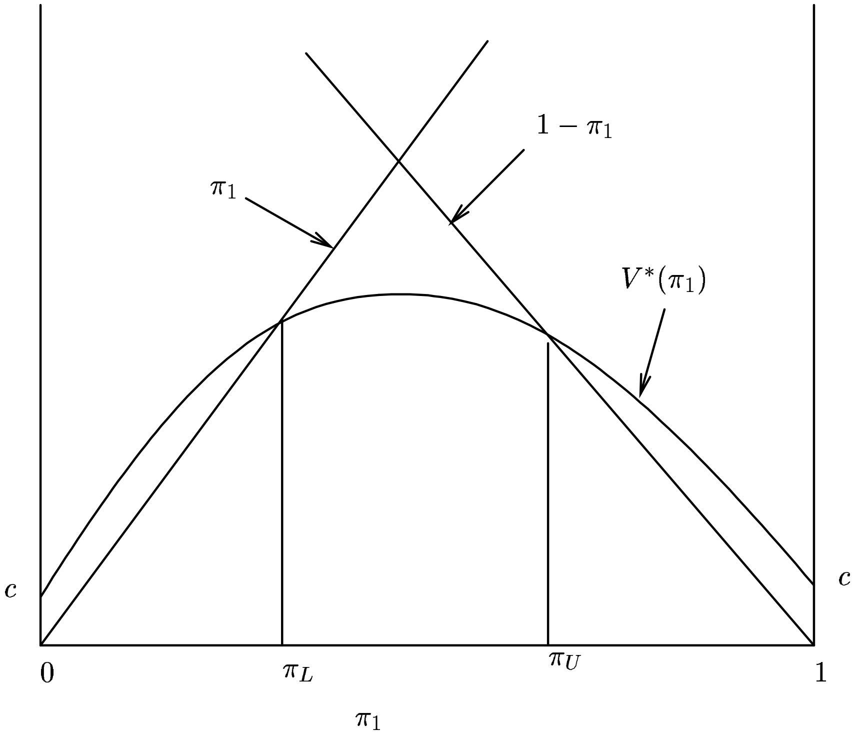


Figure 5.10: Sequential detection.

Since $\Delta_0 = 0$ means that the test does not stop with zero observation, $V^*(\pi_1)$ corresponds then to the minimum Bayes risk over all sequential tests that take at least one sample. $V^*(\pi_1)$ is in general concave and continuous and $V^*(0) = V^*(1) = c$. Now, let plot this function as well as these two specific sequential tests:

- Take no sample and decide H_0, i.e., $\Delta_0 = 1, \delta_0 = 0$ and
- Take no sample and decide H_1, i.e., $\Delta_0 = 1, \delta_0 = 1$.

Note that the Bayes risks for these tests are

$$r(\boldsymbol{\Delta}, \boldsymbol{\delta})|_{\Delta_0=1,\delta_0=0} = 1 - \pi_1$$

$$r(\boldsymbol{\Delta}, \boldsymbol{\delta})|_{\Delta_0=1,\delta_0=1} = \pi_1$$

These tests are the only two Bayesian tests that are not included in the minimization of (5.84). We note, respectively by π_U and π_L the abscisses of the intersections of the lines $r(\boldsymbol{\Delta}, \boldsymbol{\delta})|_{\Delta_0=1,\delta_0=0}$ and $r(\boldsymbol{\Delta}, \boldsymbol{\delta})|_{\Delta_0=1,\delta_0=1}$ with $V^*(\pi_1)$.

Now, by inspection of these plots, we see that the Bayes rule with a fixed given prior π_1 is:

- $(\Delta_0 = 1, \delta_0 = 0)$ if $\pi_1 \leq \pi_L$;
- $(\Delta_0 = 1, \delta_0 = 1)$ if $\pi_1 \geq \pi_U$;
- The decision rule with minimizes the Bayes risk among all the tests such that $(\Delta_0 = 0)$ corresponds to a point such that $\pi_L \leq \pi_1 \leq \pi_U$.

In the two first cases the test is stopped. In the third one, we know that the optimal test takes at least one more sample. After doing so, we are faced to a similar situation as before except that we now have more information due to the additional sample. In particular, the prior π_1 is replaced by $\pi_1(x_1) = \Pr\{H = H_1 | X_1 = x_1\}$ which is the posterior probability of H_1 given the observation $X_1 = x_1$. We can apply this method to any arbitrary number of samples. We then have the following rules:

- **Stopping rule:**

$$\boldsymbol{\Delta}_n(x_1, \ldots, x_n) = \begin{cases} 0 & \text{if} \quad \pi_L < \pi_1(x_1, \ldots, x_n) < \pi_U \\ 1 & \quad\text{otherwise.} \end{cases} \tag{5.85}$$

- **Terminal decision rule:**

$$\boldsymbol{\delta}_n(x_1, \ldots, x_n) = \begin{cases} 0 & \text{if} \quad \pi_1(x_1, \ldots, x_n) \leq \pi_L \\ 1 & \text{if} \quad \pi_1(x_1, \ldots, x_n) \geq \pi_U. \end{cases} \tag{5.86}$$

It has been proved that under mild conditions the posterior probability $\pi_1(x_1, \ldots, x_n)$ converges almost surely to 1 under H_1 and to 0 under H_0. Thus the test terminates with probability one. The only knowledge of the probabilities π_L and π_U and an algorithm to compute $\pi_1(x_1, \ldots, x_n)$ are sufficient to define this rule. The computation of $\pi_1(x_1, \ldots, x_n)$ is quite easy, but unfortunately, it is very difficult to obtain exactly π_L and π_U.

Now consider the case where the two processes P_0 and P_1 have densities f_0 and f_1. Then the Baye fomula yields

$$\begin{aligned} \pi_1(x_1, \ldots, x_n) &= \frac{\pi_1 \prod_{j=1}^n f_1(x_j)}{\pi_0 \prod_{j=1}^n f_0(x_j) + \pi_1 \prod_{j=1}^n f_1(x_j)} \\ &= \frac{\pi_1 \lambda_n(x_1, \ldots, x_n)}{\pi_0 + \pi_1 \lambda_n(x_1, \ldots, x_n)} \end{aligned} \tag{5.87}$$

where

$$\lambda_n(x_1, \ldots, x_n) = \prod_{j=1}^n \frac{f_1(x_j)}{f_0(x_j)} \tag{5.88}$$

is the likelihood ratio based on n samples.

Noting that $\pi_1(x_1, \ldots, x_n)$ is an increasing function of $\lambda_n(x_1, \ldots, x_n)$ we can rewrite (5.85) and (5.86) as:

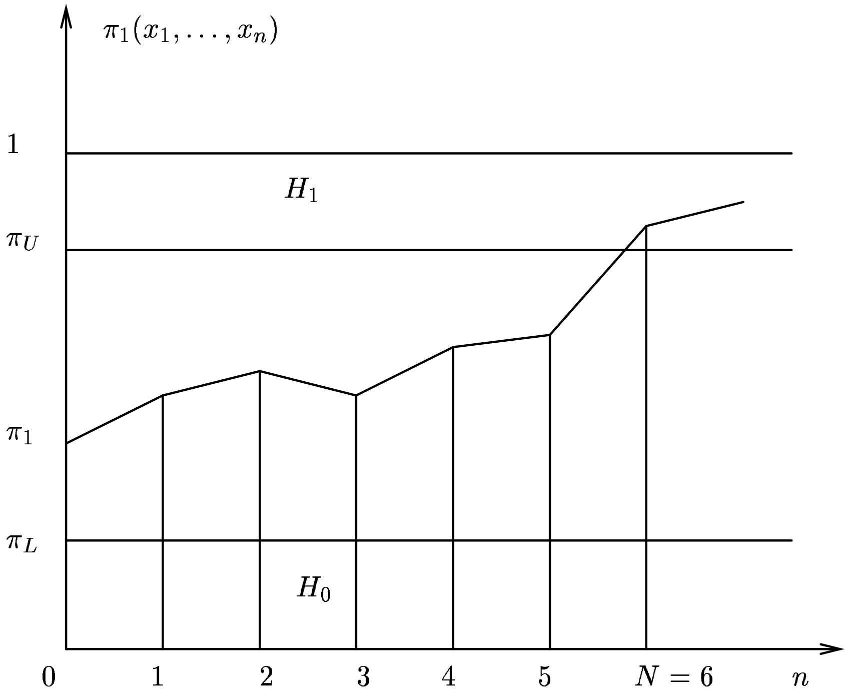


Figure 5.11: Stopping rule in sequential detection.

- **Stopping rule:**

$$\boldsymbol{\Delta}_n(x_1,\ldots,x_n) = \begin{cases} 0 & \text{if} \quad \underline{\pi} < \lambda_n(x_1,\ldots,x_n) < \overline{\pi} \\ 1 & \quad \text{otherwise.} \end{cases} \tag{5.89}$$

- **Terminal decision rule:**

$$\boldsymbol{\delta}_n(x_1,\ldots,x_n) = \begin{cases} 0 & \text{if} \quad \lambda_n(x_1,\ldots,x_n) \le \underline{\pi} \\ 1 & \text{if} \quad \lambda_n(x_1,\ldots,x_n) \ge \overline{\pi}. \end{cases} \tag{5.90}$$

where

$$\underline{\pi} \overset{\text{def}}{=} \frac{\pi_0\pi_L}{\pi_1(1-\pi_L)} \quad \text{and} \quad \overline{\pi} \overset{\text{def}}{=} \frac{\pi_0\pi_U}{\pi_1(1-\pi_U)}. \tag{5.91}$$

In conclusion, the Bayesian sequential test takes samples until the likelihood ratio falls outside the interval $[\underline{\pi},\overline{\pi}]$ and decides H_0 or H_1 if $\lambda_n(x_1,\ldots,x_n)$ falls outside of this interval.

The main problem in practical situations is to fix the values of the boundaries $a=\underline{\pi}$ and $b=\overline{\pi}$. This test is called the *sequential probability ratio test* with the boundaries a and b and is noted $SPART(a,b)$.

The following theorem gives some of the optimality properties of $SPART(a,b)$.

Wald-Wolfowitz theorem :
Note by

$$N(\boldsymbol{\Delta}) \quad = \quad \min\{n|\Delta_n(x_1,\ldots,x_n)=1\}$$

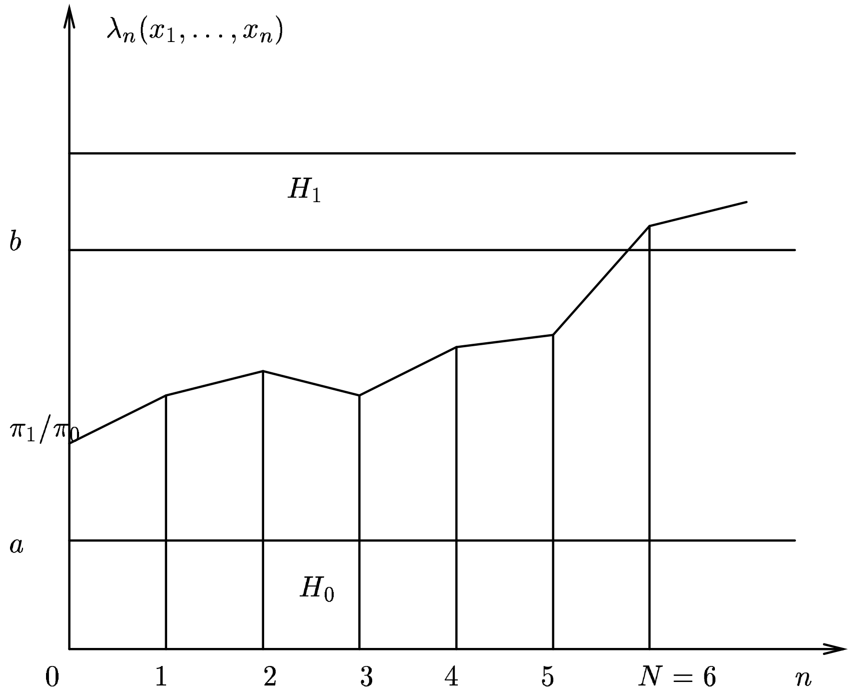


Figure 5.12: Stopping rule in $SPART(a,b)$.

$$\begin{aligned} P_F(\boldsymbol{\Delta},\boldsymbol{\delta}) &= \Pr\{\delta_N(x_1,\ldots,x_N)=1|H=H_0\} \\ P_M(\boldsymbol{\Delta},\boldsymbol{\delta}) &= \Pr\{\delta_N(x_1,\ldots,x_N)=0|H=H_1\} \end{aligned}$$

and $(\boldsymbol{\Delta}^*,\boldsymbol{\delta}^*)$ the $SPART(a,b)$. Then, for any sequential decision rule $(\boldsymbol{\Delta},\boldsymbol{\delta})$ for which

$$\begin{aligned} P_F(\boldsymbol{\Delta},\boldsymbol{\delta}) &\le P_F(\boldsymbol{\Delta}^*,\boldsymbol{\delta}^*) \\ P_M(\boldsymbol{\Delta},\boldsymbol{\delta}) &\le P_M(\boldsymbol{\Delta}^*,\boldsymbol{\delta}^*) \end{aligned}$$

we have

$$\mathrm{E}\left[N(\boldsymbol{\Delta})|H=H_j\right] \ge \mathrm{E}\left[N(\boldsymbol{\Delta}^*)|H=H_j\right], \quad j=0,1$$

The validity of Wald-Wolfowitz theorem is a consequence of the Bayes optimality of $SPART(a,b)$. The results of this theorem and other related theorems are sumarized in the following items:

- For a given performance, there is no other sequential decision rule with a smaller expected sample size than the $SPART(a,b)$ with the same performance.
- The average sample size of $SPART(a,b)$ is not greater than the sample size of a fixed-sample-size test with the same performance.
- For a given expected sample size, no sequential decision rule has smaller error probabilities than the $SPART(a,b)$.

Two main questions remains:

- How to choose a and b to yield a desired level of performance?
- How to evaluate the expected sample size of a sequential detector?

The following result gives an answer to the first one.

Let $(\boldsymbol{\Delta}, \boldsymbol{\delta}) = SPART(a, b)$ with $a < 1 < b$ and $\alpha = P_F(\boldsymbol{\Delta}, \boldsymbol{\delta})$, $\gamma = 1 - \beta = P_M(\boldsymbol{\Delta}, \boldsymbol{\delta})$ and $N = N(\boldsymbol{\Delta})$. Then the rejection region of $(\boldsymbol{\Delta}, \boldsymbol{\delta})$ is

$$\Gamma_1 = \left\{ \boldsymbol{x} \in \mathbb{R}^\infty \middle| \lambda_N(x_1, \ldots, x_N) \geq b \right\} = \cup_{n=1}^{\infty} Q_n \tag{5.92}$$

with

$$Q_n = \left\{ \boldsymbol{x} \in \mathbb{R}^\infty \middle| N = n, \; \lambda_n(x_1, \ldots, x_N) \geq b \right\} = \cup_{n=1}^{\infty} Q_n \tag{5.93}$$

Since Q_n and Q_m are mutually exclusive sets for $m \neq n$, we have

$$\alpha = \Pr\{\lambda_n(x_1, \ldots, x_N) \geq b | H = H_0\} = \sum_{n=1}^{\infty} \int_{Q_n} \prod_{j=1}^{n} f_0(x_j) \, \mathrm{d}x_j \tag{5.94}$$

On Q_n we have

$$\prod_{j=1}^{n} f_0(x_j) \, \mathrm{d}x_j \leq \frac{1}{b} \prod_{j=1}^{n} f_1(x_j) \, \mathrm{d}x_j \tag{5.95}$$

So, we have

$$\alpha \leq \frac{1}{b} \sum_{n=1}^{\infty} \int_{Q_n} \prod_{j=1}^{n} f_1(x_j) \, \mathrm{d}x_j = \frac{1}{b} \Pr\{\lambda_n(x_1, \ldots, x_N) \geq b | H = H_0\} = \frac{1}{b}(1 - \gamma) \tag{5.96}$$

and in the same manner we obtain

$$\gamma = \Pr\{\lambda_n(x_1, \ldots, x_N) \leq a | H = H_1\} \leq a\,(1 - \alpha) \tag{5.97}$$

From these two relations we deduce

$$\begin{cases} b < \frac{1-\gamma}{\alpha} \\ a > \frac{\gamma}{1-\alpha} \end{cases} \tag{5.98}$$

The following choice is called the Wald's approximation:

$$\begin{cases} b \simeq \frac{1-\gamma}{\alpha} \\ a \simeq \frac{\gamma}{1-\alpha} \end{cases} \tag{5.99}$$

`To be completed later`

5.8 Robust detection

```
To be completed later
```

Chapter 6

Elements of parameter estimation

6.1 Bayesian parameter estimation

Throughout this chapter we assume that the data are samples of a parametrically known process $\{\mathcal{P}_\theta; \theta \in \boldsymbol{\tau}\}$, where $\mathcal{P}_\theta$ denotes a distribution on the observation space $(\Gamma, \mathcal{G})$:

$$\boldsymbol{X} \sim \mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{6.1}$$

The goal of the parameter estimation problem is to find a function $\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) : \Gamma \mapsto \boldsymbol{\tau}$ such that $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ is the best guess of the true value of $\boldsymbol{\theta}$. Of course, the solution depends on a goodness criterion. As in the hypothesis testing problems, we have to define a cost function $c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}] : \boldsymbol{\tau} \times \boldsymbol{\tau} \mapsto \mathbb{R}^+$ such that $c[\boldsymbol{a}, \boldsymbol{\theta}]$ is the cost of estimating the true value of $\boldsymbol{\theta}$ by $\boldsymbol{a}$.

Then, as in the hypothesis testing problems, we can define the conditional risk function

$$\begin{aligned} R_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) &= \mathrm{E}_{\boldsymbol{\theta}}\left\{c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}]\right\} = \mathrm{E}\left[c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \Theta] \,|\, \Theta = \boldsymbol{\theta}\right] \\ &= \int_{\boldsymbol{\Gamma}} c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}] f_{\boldsymbol{\theta}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \end{aligned} \tag{6.2}$$

and the Bayes risk

$$\begin{aligned} r(\widehat{\boldsymbol{\theta}}) &= \mathrm{E}\left[R_{\boldsymbol{\Theta}}(\widehat{\boldsymbol{\theta}}(\boldsymbol{X}))\right] \\ &= \int_{\boldsymbol{\tau}} \int_{\boldsymbol{\Gamma}} c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}] f_{\boldsymbol{\theta}}(\boldsymbol{x}) \pi(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Gamma}} \int_{\boldsymbol{\tau}} c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}] \pi(\boldsymbol{\theta}|\boldsymbol{x}) \,\mathrm{d}\boldsymbol{\theta} \,\mathrm{d}\boldsymbol{x} \\ &= \mathrm{E}\left[\mathrm{E}\left[c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \Theta] \,|\, \boldsymbol{X} = \boldsymbol{x}\right]\right] \end{aligned} \tag{6.3}$$

From this relation, and the fact that in general the cost function is positive, we see that $r(\widehat{\boldsymbol{\theta}})$ is minimized over $\widehat{\boldsymbol{\theta}}$, when for any $\boldsymbol{x} \in \boldsymbol{\Gamma}$, the mean posterior cost

$$\bar{c}[\boldsymbol{x}] = \mathrm{E}\left[c[\widehat{\boldsymbol{\theta}}(\boldsymbol{X}), \Theta] \,|\, \boldsymbol{X} = \boldsymbol{x}\right] = \int_{\boldsymbol{\tau}} c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}] \pi(\boldsymbol{\theta}|\boldsymbol{x}) \,\mathrm{d}\boldsymbol{\theta} \tag{6.4}$$

is minimized.

It is clear that the resulting estimate depends on the choice of the cost function. In the following section we first consider the case of a scalar parameter and then extend it to the vector parameter case.

6.1.1 Minimum-Mean-Squared-Error

In case where $\boldsymbol{\tau} = \mathbb{R}$, a commonly used cost function is

$$c[a,\theta] = c(a-\theta) = (a-\theta)^2, \quad (a,\theta) \in \mathbb{R}^2 \tag{6.5}$$

The corresponding Bayes risk is $\mathrm{E}\left[(\widehat{\theta}(\boldsymbol{X}) - \Theta)^2\right]$, a quantity which is known as the *Mean-Squared-Error* (MSE). The corresponding Bayes estimate is called the *Minimum-Mean-Squared-Error* (MMSE) *estimator.*

The posterior cost is given by

$$\begin{aligned}
\mathrm{E}\left[(\widehat{\theta}(\boldsymbol{X}) - \Theta)^2 \,|\, \boldsymbol{X} = \boldsymbol{x}\right] &= \mathrm{E}\left[\widehat{\theta}^2(\boldsymbol{X}) \,|\, \boldsymbol{X} = \boldsymbol{x}\right] - 2\,\mathrm{E}\left[\widehat{\theta}(\boldsymbol{X})\Theta \,|\, \boldsymbol{X} = \boldsymbol{x}\right] + \mathrm{E}\left[\Theta^2 \,|\, \boldsymbol{X} = \boldsymbol{x}\right] \\
&= [\widehat{\theta}(\boldsymbol{X})]^2 - 2\,\widehat{\theta}(\boldsymbol{X})\,\mathrm{E}\left[\Theta \,|\, \boldsymbol{X} = \boldsymbol{x}\right] + \mathrm{E}\left[\Theta^2 \,|\, \boldsymbol{X} = \boldsymbol{x}\right] \qquad (6.6)
\end{aligned}$$

This expression is a quadratic function of $\widehat{\theta}(\boldsymbol{X})$ and its minimum is obtained for

$$\widehat{\theta}_{MMSE}(\boldsymbol{X}) = \mathrm{E}\left[\Theta \,|\, \boldsymbol{X} = \boldsymbol{x}\right] \tag{6.7}$$

Thus the MMSE estimate is the mean of the posterior probability density function. This estimate is also called *posterior mean* (PM) estimate.

6.1.2 Minimum-Mean-Absolute-Error

In case where $\boldsymbol{\tau} = \mathbb{R}$, another commonly used cost function is

$$c[a,\theta] = c(a-\theta) = |a-\theta|, \quad (a,\theta) \in \mathbb{R}^2 \tag{6.8}$$

The corresponding Bayes risk is $\mathrm{E}\left[|\widehat{\theta}(\boldsymbol{X}) - \Theta|\right]$, a quantity which is known as the *Mean-Absolute-Error* (MAE). The corresponding Bayes estimate is called the *Minimum-Mean-Absolute-Error* (MMAE) *estimator.*

The posterior cost is given by

$$\begin{aligned}
\mathrm{E}\left[|\widehat{\theta}(\boldsymbol{X}) - \Theta| \,|\, \boldsymbol{X} = \boldsymbol{x}\right] &= \int_0^\infty \Pr\left\{|\widehat{\theta}(\boldsymbol{x}) - \Theta| > z \,|\, \boldsymbol{X} = \boldsymbol{x}\right\} \mathrm{d}z \\
&= \int_0^\infty \Pr\left\{\Theta > z + \widehat{\theta}(\boldsymbol{x}) \,|\, \boldsymbol{X} = \boldsymbol{x}\right\} \mathrm{d}z \\
&\quad + \int_0^\infty \Pr\left\{\Theta < -z + \widehat{\theta}(\boldsymbol{x}) \,|\, \boldsymbol{X} = \boldsymbol{x}\right\} \mathrm{d}z
\end{aligned} \tag{6.9}$$

Doing the variable change $t = z + \widehat{\theta}(\boldsymbol{x})$ in the first integral and $t = -z + \widehat{\theta}(\boldsymbol{x})$ in the second one we obtain

$$\begin{aligned}
\mathrm{E}\left[|\widehat{\theta}(\boldsymbol{x}) - \Theta| \,|\, \boldsymbol{X} = \boldsymbol{x}\right] &= \int_{\widehat{\theta}(\boldsymbol{x})}^\infty \Pr\left\{\Theta > t \,|\, \boldsymbol{X} = \boldsymbol{x}\right\} \mathrm{d}t \\
&\quad + \int_{-\infty}^{\widehat{\theta}(\boldsymbol{x})} \Pr\left\{\Theta < t \,|\, \boldsymbol{X} = \boldsymbol{x}\right\} \mathrm{d}t \qquad (6.10)
\end{aligned}$$

This expression is differentiable with respect to $\widehat{\theta}(\boldsymbol{x})$ and

$$\begin{aligned}\frac{\partial \mathrm{E}\left[|\widehat{\theta}(\boldsymbol{X})-\Theta| \,|\, \boldsymbol{X}=\boldsymbol{x}\right]}{\partial \widehat{\theta}(\boldsymbol{x})} &= \Pr\left\{\Theta < \widehat{\theta}(\boldsymbol{x}) \,|\, \boldsymbol{X}=\boldsymbol{x}\right\} \\ &\quad -\Pr\left\{\Theta > \widehat{\theta}(\boldsymbol{x}) \,|\, \boldsymbol{X}=\boldsymbol{x}\right\} \qquad (6.11)\end{aligned}$$

This derivative is a nondecreasing function of $\widehat{\theta}(\boldsymbol{x})$ which approaches -1 as $\widehat{\theta}(\boldsymbol{x}) \longrightarrow -\infty$ and $+1$ as $\widehat{\theta}(\boldsymbol{x}) \longrightarrow +\infty$. The minimum of (6.11) is achieved at the point $\widehat{\theta}(\boldsymbol{x})$ where the derivative vanishes. Consequently, the Bayes estimate satisfies

$$\begin{aligned}\Pr\{\Theta < t \,|\, \boldsymbol{X}=\boldsymbol{x}\} &\leq \Pr\{\Theta > t \,|\, \boldsymbol{X}=\boldsymbol{x}\}, \quad t < \widehat{\theta}(\boldsymbol{x}) \\ &\text{and} \\ \Pr\{\Theta < t \,|\, \boldsymbol{X}=\boldsymbol{x}\} &\geq \Pr\{\Theta > t \,|\, \boldsymbol{X}=\boldsymbol{x}\}, \quad t > \widehat{\theta}(\boldsymbol{x})\end{aligned}$$

or

$$\Pr\left\{\Theta < \widehat{\theta}(\boldsymbol{x}) \,|\, \boldsymbol{X}=\boldsymbol{x}\right\} = \Pr\left\{\Theta > \widehat{\theta}(\boldsymbol{x}) \,|\, \boldsymbol{X}=\boldsymbol{x}\right\}. \qquad (6.12)$$

$\widehat{\theta}(\boldsymbol{x})$ is the *median* of the posterior distribution of Θ given $\boldsymbol{X}=\boldsymbol{x}$:

$$\widehat{\theta}_{MMAE}(\boldsymbol{X}) = \text{median of} \quad \pi(\theta \,|\, \boldsymbol{X}=\boldsymbol{x}) \qquad (6.13)$$

6.1.3 Maximum *A Posteriori* (MAP) estimation

Another commonly used cost function in the cases where $\boldsymbol{\tau} = \mathbb{R}$ is

$$c[a,\theta] = c(a-\theta) = \begin{cases} 0 & \text{if} \quad |a-\theta| \leq \Delta \\ 1 & \text{if} \quad |a-\theta| > \Delta \end{cases} \qquad (6.14)$$

where Δ is a positive real number. The corresponding Bayes risk is given by

$$\begin{aligned}\mathrm{E}\left[c[\widehat{\theta}(\boldsymbol{X})-\Theta] \,|\, \boldsymbol{X}=\boldsymbol{x}\right] &= \Pr\left\{|\widehat{\theta}(\boldsymbol{x})-\Theta| > \Delta \,|\, \boldsymbol{X}=\boldsymbol{x}\right\} \\ &= 1-\Pr\left\{|\widehat{\theta}(\boldsymbol{x})-\Theta| \leq \Delta \,|\, \boldsymbol{X}=\boldsymbol{x}\right\} \qquad (6.15)\end{aligned}$$

To minimize this expression we consider two cases:

- Θ is a discrete random variable taking its values in a finite set $\boldsymbol{\tau} = \{\theta_1, \ldots, \theta_M\}$ such that $|\theta_i - \theta_j| > \Delta$ for any $i \neq j$. Then we have

  $$\mathrm{E}\left[c[\widehat{\theta}(\boldsymbol{X}),\Theta] \,|\, \boldsymbol{X}=\boldsymbol{x}\right] = 1-\Pr\left\{\Theta = \widehat{\theta}(\boldsymbol{x}) \,|\, \boldsymbol{X}=\boldsymbol{x}\right\} = 1-\pi(\widehat{\theta}(\boldsymbol{x}) \,|\, \boldsymbol{x}) \qquad (6.16)$$

  where $\pi(\theta \,|\, \boldsymbol{x})$ is the posterior distribution of Θ given $\boldsymbol{X}=\boldsymbol{x}$. The estimate is the value of Θ which has the maximum *a posteriori* probability:

  $$\widehat{\theta}_{MAP} = \arg\max_{\theta \in \mathcal{T}} \{\pi(\theta \,|\, \boldsymbol{x})\} \qquad (6.17)$$

- Θ is a continuous random variable. In this case, we have

$$\mathrm{E}\left[c[\widehat{\theta}(\boldsymbol{x}),\Theta]\,|\,\boldsymbol{X}=\boldsymbol{x}\right]=1-\int_{\widehat{\theta}(\boldsymbol{x})-\Delta}^{\widehat{\theta}(\boldsymbol{x})+\Delta}\pi(\theta\,|\,\boldsymbol{X}=\boldsymbol{x})\,\mathrm{d}\theta \tag{6.18}$$

If we assume that the posterior probability distribution $\pi(\theta\,|\,\boldsymbol{x})$ is a continuous and smooth function and Δ is sufficiently small, then we can write

$$\mathrm{E}\left[c[\widehat{\theta}(\boldsymbol{x}),\Theta]\,|\,\boldsymbol{X}=\boldsymbol{x}\right]=1-2\Delta\;\pi(\widehat{\theta}(\boldsymbol{x})\,|\,\boldsymbol{X}=\boldsymbol{x}) \tag{6.19}$$

and again we have

$$\widehat{\theta}_{MAP}=\arg\max_{\theta\in\mathcal{T}}\{\pi(\theta\,|\,\boldsymbol{x})\} \tag{6.20}$$

Example 1: Estimation of the parameter of an exponential distribution.
Suppose both distributions $f_\theta(x)$ and $\pi(\theta)$ are exponential:

$$f_\theta(x)=\begin{cases}\theta\exp[-\theta x] & \text{if } x>0\\ 0 & \text{otherwise}\end{cases} \tag{6.21}$$

and

$$\pi(\theta)=\begin{cases}\alpha\exp[-\alpha\theta] & \text{if } \theta>0\\ 0 & \text{otherwise}\end{cases} \tag{6.22}$$

Note that

$$\begin{cases}\mathrm{E}[X]=\theta\\ \mathrm{Var}\{X\}=\mathrm{E}\left[(X-\theta)^2\right]=\theta^2\end{cases} \tag{6.23}$$

and

$$\begin{cases}\mathrm{E}[\Theta]=\alpha\\ \mathrm{Var}\{\Theta\}=\mathrm{E}\left[(\Theta-\alpha)^2\right]=\alpha^2\end{cases} \tag{6.24}$$

Then, we can calculate the joint distribution $\phi(x,\theta)$

$$\phi(x,\theta)=\begin{cases}\alpha\,\theta\,\exp[-\theta x-\alpha\theta] & \text{if } \theta>0, x>0\\ 0 & \text{otherwise.}\end{cases} \tag{6.25}$$

The marginal distribution $m(x)$ is given by

$$m(x)=\begin{cases}\int_0^\infty\alpha\,\theta\,\exp[-(\alpha+x)\theta]\,\mathrm{d}\theta=\frac{\alpha}{(\alpha+x)^2} & \text{if } x>0\\ 0 & \text{otherwise.}\end{cases} \tag{6.26}$$

and the posterior distribution $\pi(\theta|x)$ is given by

$$\pi(\theta|x)=\begin{cases}\frac{\alpha\theta\exp[-(\alpha+x)\theta]}{m(x)}=(\alpha+x)^2\theta\exp[-(\alpha+x)\theta] & \text{if } \theta>0\\ 0 & \text{otherwise.}\end{cases} \tag{6.27}$$

Note that we have

$$\begin{cases}\mathrm{E}[\Theta\,|\,X=x]=\frac{2}{\alpha+x}\\ \mathrm{Var}\{\Theta\,|\,X=x\}=\frac{2}{(\alpha+x)^2}\end{cases} \tag{6.28}$$

The MMSE estimator is given by

$$\begin{aligned}\widehat{\theta}_{MMSE}(x) &= \mathrm{E}\left[\Theta \mid X = x\right] \\ &= \int_0^{\infty} \theta\pi(\theta|x)\,\mathrm{d}\theta \\ &= \int_0^{\infty} (\alpha+x)^2\theta^2 \exp\left[-(\alpha+x)\theta\right]\,\mathrm{d}\theta \\ &= \frac{2}{\alpha+x} \qquad (6.29)\end{aligned}$$

and the corresponding MMSE is:

$$\begin{aligned}MMSE &= r(\widehat{\theta}_{MMSE}) = \mathrm{E}\left[\mathrm{Var}\left\{\Theta \mid X\right\}\right] \\ &= \int_0^{\infty} \frac{2}{(\alpha+x)^2}\, m(x)\,\mathrm{d}x \\ &= \int_0^{\infty} \frac{2\alpha}{(\alpha+x)^4}\,\mathrm{d}x \\ &= \frac{2}{3\alpha^2} \qquad (6.30)\end{aligned}$$

The MMAE estimate $\widehat{\theta}_{ABS}(x)$ is such that

$$\int_{\widehat{\theta}_{ABS}}^{\infty} \pi(\theta|x)\,\mathrm{d}\theta = \left[1 + (\alpha+x)\widehat{\theta}_{ABS}\right] \exp\left[-(\alpha+x)\widehat{\theta}_{ABS}\right] = \frac{1}{2} \qquad (6.31)$$

It is easily shown that

$$\widehat{\theta}_{ABS}(x) = \frac{T_0}{\alpha+x} \qquad (6.32)$$

where T_0 is the solution of

$$(1+T_0)\exp\left[-T_0\right] = \frac{1}{2} \longrightarrow T_0 \simeq 1.68$$

To calculate $\widehat{\theta}_{MAP}(x)$, we can remark that

$$\left\{\begin{aligned}\frac{\partial \log \pi(\theta \mid x)}{\partial\theta} &= \frac{1}{\theta} - (\alpha+x) \\ \frac{\partial^2 \log \pi(\theta \mid x)}{\partial\theta^2} &= -\frac{1}{\theta^2} < 0\end{aligned}\right. \qquad (6.33)$$

So, we have

$$\widehat{\theta}_{MAP}(x) = \frac{1}{\alpha+x} \qquad (6.34)$$

The following table sumarizes theses estimates:

$$\left\{\begin{aligned}\widehat{\theta}_{MMSE}(x) &= \frac{2}{\alpha+x} \\ \widehat{\theta}_{ABS}(x) &= \frac{T_0}{\alpha+x} \\ \widehat{\theta}_{MAP}(x) &= \frac{1}{\alpha+x}\end{aligned}\right. \qquad (6.35)$$

Example 2: Estimation of the location parameter of a Gaussian distribution.

Assume $X|\Theta$ and Θ have both Gaussian distributions:

$$\begin{array}{lcccl} X|\Theta & \sim & f_\theta(x) & = & \mathcal{N}\left(\theta,\sigma^2\right) \\ \Theta & \sim & \pi(\theta) & = & \mathcal{N}\left(\theta_0,\sigma_\theta^2\right) \end{array}$$

Then, the joint distribution $\phi(x,\theta)$, the marginal distribution $m(x)$ and the posterior distribution $\pi(\theta\,|\,x)$ are

$$\begin{array}{lcccl} X,\Theta & \sim & \phi(x,\theta) & = & \mathcal{N}\left((\theta,\theta_0),\begin{pmatrix}\sigma_x^2 & 0\\ 0 & \sigma_\theta^2\end{pmatrix}\right) \\ X & \sim & m(x) & = & \mathcal{N}\left(\theta,\sigma_x^2=\sigma^2+\sigma_\theta^2\right) \\ \Theta\,|\,X & \sim & \pi(\theta\,|\,x) & = & \mathcal{N}\left(\widehat{\theta},\widehat{\sigma^2}\right) \end{array}$$

with

$$\left\{\begin{array}{lcl} \widehat{\theta} & = & \frac{\sigma_\theta^2}{\sigma_x^2}\,\theta_0+\frac{\sigma^2}{\sigma_x^2}\,x \\ \widehat{\sigma^2} & = & \frac{\sigma\,\sigma_\theta}{\sigma_x^2} \end{array}\right.$$

In this case all the estimators are equal and we have

$$\widehat{\theta}_{MMSE}(x)=\widehat{\theta}_{ABS}(x)=\widehat{\theta}_{MAP}(x) = \widehat{\theta}=\frac{\sigma_\theta^2}{\sigma_x^2}\,\theta_0+\frac{\sigma^2}{\sigma_x^2}\,x \tag{6.36}$$

$$= \frac{\sigma_\theta^2}{\sigma^2+\sigma_\theta^2}\,\theta_0+\frac{\sigma^2}{\sigma^2+\sigma_\theta^2}\,x \tag{6.37}$$

$$\tag{6.38}$$

They also have the same posterior variance

$$\widehat{\sigma^2}=\frac{\sigma\,\sigma_\theta}{\sigma_x^2}=\frac{\sigma\,\sigma_\theta}{\sigma^2+\sigma_\theta^2} \tag{6.39}$$

Note also the following limit cases:

$$\begin{array}{lll} \text{When } \sigma^2\longrightarrow 0 & \text{then} & \widehat{\theta}\longrightarrow\theta_0 \\ \text{When } \sigma_\theta^2\longrightarrow 0 & \text{then} & \widehat{\theta}\longrightarrow x \end{array}$$

Example 3: Estimation of the parameter of a Binomial distribution.

Assume that

$$\begin{array}{lcll}
X\,|\,\theta & \sim & f(x\,|\,\theta) & = \mathbf{Bin}(x|\theta,n) = C_n^x\,\theta^x\,(1-\theta)^{n-x} \\[2ex]
\Theta & \sim & \pi(\theta) & = \mathbf{Beta}(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}
\end{array}$$

Then, the joint distribution $\phi(x,\theta)$, the marginal distribution $m(x)$ and the posterior distribution $\pi(\theta\,|\,x)$ are

$$\begin{array}{lcll}
(X,\Theta) & \sim & \phi(x,\theta) & = \frac{C_n^x}{\mathcal{B}(\alpha,\beta)}\,\theta^{\alpha+x-1}\,(1-\theta)^{\beta+n-x-1} \\[2ex]
X & \sim & m(x) & = \frac{C_n^x}{\mathcal{B}(\alpha,\beta)}\,\mathcal{B}(\alpha+x,n+\beta-x) \\[2ex]
 & & & = C_n^x\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\frac{\Gamma(\alpha+x)\Gamma(n+\beta-x)}{\Gamma(\alpha+\beta+n)} \\[2ex]
\Theta\,|\,X & \sim & \pi(\theta|x) & = \frac{1}{B(\alpha+x,\beta+n-x)}\,\theta^{\alpha+x-1}\,(1-\theta)^{\beta+n-x-1} \\[2ex]
 & & & = \mathbf{Beta}(\alpha+x,n+\beta-x)
\end{array}$$

6.2 Other cost functions and related estimators

Here are some other cost functions and corresponding Bayesian estimator expressions.

Name	$C[a,\theta]$	$\widehat{\theta}$
Quadratic	$q(a-\theta)^2$	$\mathrm{E}\left[\Theta\|\boldsymbol{x}\right] = \int \theta\pi(\theta\|\boldsymbol{x})\,\mathrm{d}\theta$
Weighted Quadratic	$\omega(\theta)(a-\theta)^2$	$\frac{\mathrm{E}[\omega(\Theta)\,\Theta\|\boldsymbol{x}]}{\mathrm{E}[\omega(\Theta)\|\boldsymbol{x}]} = \dfrac{\int \omega(\theta)\;\theta\;\pi(\theta\|\boldsymbol{x})\,\mathrm{d}\theta}{\int \omega(\theta)\;\pi(\theta\|\boldsymbol{x})\,\mathrm{d}\theta}.$
Absolute	$\|a-\theta\|$	$\int_{-\infty}^{\widehat{\theta}} \pi(\theta\|\boldsymbol{x})\,\mathrm{d}\theta = \int_{\widehat{\theta}}^{+\infty} \pi(\theta\|\boldsymbol{x})\,\mathrm{d}\theta = \frac{1}{2}$
Nonsymetric Absolute	$\begin{cases} k_2(\theta-a) & \text{si } \theta \leq a, \\ k_1(a-\theta) & \text{si } \theta \geq a. \end{cases}$	$k_1\ \int_{-\infty}^{\widehat{\theta}} \pi(\theta\|\boldsymbol{x})\,\mathrm{d}\theta = k_2\ \int_{\widehat{\theta}}^{+\infty} \pi(\theta\|\boldsymbol{x})\,\mathrm{d}\theta$
Linex	$\beta\left[\exp\left[-\alpha(a-\theta)\right] - \alpha(\alpha-\theta) - 1\right],$ $\alpha \neq 0, \beta > 0$	$-\frac{1}{\alpha}\log\left(\mathrm{E}\left[\exp\left[-\alpha\Theta\right]\ \|\boldsymbol{x}\right]\right)$ $= -\frac{1}{\alpha}\log\left[\int \exp\left[-\alpha\theta\right] p(\theta\|\boldsymbol{x})\,\mathrm{d}\theta\right]$
	$\frac{(a-\theta)^2}{\theta}$	$\frac{1}{\mathrm{E}[1/\Theta\|\boldsymbol{x}]} = \frac{1}{\int \frac{1}{\theta}\pi(\theta\|\boldsymbol{x})\,\mathrm{d}\theta}$
	$\frac{(a-\theta)^2}{a}$	$\sqrt{\mathrm{E}\left[\Theta^2\|\boldsymbol{x}\right]} = \sqrt{\int \theta^2\,\pi(\theta\|\boldsymbol{x})\,\mathrm{d}\theta}$
	$-\ln[a(\theta)]$	$\pi(\theta\|\boldsymbol{x})$

Table 6.1: Relations between the data, *a priori*, marginal and *a posteriori* distributions.

6.3 Examples of posterior calculation

In previous sections, we saw that the computation of Bayesian estimates requires the posterior probability distribution. The following table gives a summary of the expressions of the posterior probability distributions in some classical cases.

Observation law $f(x\mid\theta)$	Prior law $\pi(\theta)$	Marginal law $m(x) = \int f(x\mid\theta)\,\pi(\theta)\,\mathrm{d}\theta$	Posterior law $\pi(\theta\mid x) = \dfrac{f(x\mid\theta)\,\pi(\theta)}{m(x)}$
Discrete variables			
Binomial $\mathbf{Bin}(x\mid n,\theta)$	Beta $\mathbf{Bet}(\theta\mid\alpha,\beta)$	Binomial-Beta $\mathbf{BinBet}(x\mid\alpha,\beta,n)$	Beta $\mathbf{Bet}(\theta\mid\alpha+x,\beta+n-x)$
Negative Binomial $\mathbf{NegBin}(x\mid n,\theta)$	Beta $\mathbf{Bet}(\theta\mid\alpha,\beta)$	Negative Binomial-Beta $\mathbf{NegBinBet}(x\mid\alpha,\beta,\theta)$	Beta $\mathbf{Bet}(\theta\mid\alpha+n,\beta+x)$
Poisson $\mathbf{Pn}(x\mid\theta)$	Gamma $\mathbf{Gam}(\theta\mid\alpha,\beta)$	Poisson-Gamma $\mathbf{PnGam}(x\mid\alpha,\beta,1)$	Gamma $\mathbf{Gam}(\theta\mid\alpha+x,\beta+1)$
Continuous variables			
Gamma $\mathbf{Gam}(x\mid\nu,\theta)$	Gamma $\mathbf{Gam}(\theta\mid\alpha,\beta)$	Gamma-Gamma $\mathbf{GamGam}(x\mid\alpha,\beta,\nu)$	Gamma $\mathbf{Gam}(\theta\mid\alpha+\nu,\beta+x)$
Exponential $\mathbf{Ex}(x\mid\theta)$	Gamma $\mathbf{Gam}(\theta\mid\alpha,\beta)$	Pareto $\mathbf{Par}(x\mid\alpha,\beta)$	Gamma $\mathbf{Gam}(\theta\mid\alpha+1,\beta+x)$
Normal $\mathbf{N}(x\mid\theta,\sigma^2)$	Normal $\mathbf{N}(\theta\mid\mu,\tau^2)$	Normal $\mathbf{N}(x\mid\mu+\theta,\tau^2)$?	Normal $\mathbf{N}\left(\mu\mid\frac{\mu\sigma^2+\tau^2x}{\sigma^2+\tau^2},\frac{\sigma^2\,\tau^2}{\sigma^2+\tau^2}\right)$
Normal $\mathbf{N}(x\mid\mu,\lambda\theta)$	Gamma $\mathbf{Gam}(\theta\mid\frac{\alpha}{2},\frac{\alpha}{2})$	Student (t) $\mathbf{St}(x\mid\mu,\lambda,\alpha)$	Gamma $\mathbf{Gam}\left(\theta\mid\frac{\alpha+1}{2},\frac{\alpha}{2}+\frac{1}{2}(\mu-x)^2\right)$

Table 6.2: Relation between the data, *a priori*, marginal and *a posteriori* distributions.

6.4 Estimation of vector parameters

In the case where we have a vector parameter $\boldsymbol{\theta} = [\theta_1, \dots, \theta_m]^t$ we have to define a cost function $c[\boldsymbol{a}, \boldsymbol{\theta}] : \mathbb{R}^m \times \mathbb{R}^m \mapsto \mathbb{R}^+$. Then it is again possible to define the Bayes risk. In many cases the cost function is of the form

$$c[\boldsymbol{a}, \boldsymbol{\theta}] = \sum_{i=1}^{m} c_i[a_i, \theta_i] \tag{6.40}$$

We then have

$$\mathrm{E}\left[c[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}] \,|\, \boldsymbol{X} = \boldsymbol{x}\right] = \sum_{i=1}^{m} \mathrm{E}\left[c_i[\widehat{\theta}_i(\boldsymbol{x}), \theta_i] \,|\, \boldsymbol{X} = \boldsymbol{x}\right] \tag{6.41}$$

Here after, we consider some common cost functions:

6.4.1 Minimum-Mean-Squared-Error

In case where $\boldsymbol{\tau} = \mathbb{R}^m$, a commonly used cost function is

$$c[\boldsymbol{a}, \boldsymbol{\theta}] = \|\boldsymbol{a} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{m} (a_i - \theta_i)^2 \tag{6.42}$$

The corresponding Bayes risk is $\mathrm{E}\left[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\Theta}\|^2\right]$ and the corresponding Bayes estimate is the *Minimum-Mean-Squared-Error* (MMSE) *estimator* or the Bayes estimate:

$$\widehat{\boldsymbol{\theta}}_{MMSE}(\boldsymbol{X}) = \mathrm{E}\left[\boldsymbol{\Theta} \,|\, \boldsymbol{X} = \boldsymbol{x}\right] \tag{6.43}$$

Thus the MMSE estimate is the mean of the posterior probability density function. It is also called *posterior mean* (PM) estimate.

Note that, as in the scalar case, the following weighted quadratic cost function

$$c[\boldsymbol{a}, \boldsymbol{\theta}] = \|\boldsymbol{a} - \boldsymbol{\theta}\|^2_{\boldsymbol{Q}} = [\boldsymbol{a} - \boldsymbol{\theta}]^t \boldsymbol{Q} [\boldsymbol{a} - \boldsymbol{\theta}] \tag{6.44}$$

gives the same estimate as in (6.43), *i.e.* the MMSE estimate does not depend on the weighting matrix $\boldsymbol{Q}$. However, the corresponding minimum Bayes risks are different and we have

$$\mathrm{E}\left[\|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\Theta}\|^2_{\boldsymbol{Q}}\right] = \mathrm{tr}\left\{\boldsymbol{Q}\mathrm{E}\left[\mathrm{Cov}\left\{\boldsymbol{\Theta} \,|\, \boldsymbol{X} = \boldsymbol{x}\right\}\right]\right\} \tag{6.45}$$

.

6.4.2 Minimum-Mean-Absolute-Error

In case where $\boldsymbol{\tau} = \mathbb{R}$, another commonly used cost function is

$$c[\boldsymbol{a}, \boldsymbol{\theta}] = \sum_{i=1}^{m} |a_i - \theta_i|, \tag{6.46}$$

The corresponding estimate is such that

$$\Pr\left\{\Theta_i < \widehat{\theta}_i(\boldsymbol{x}) \,|\, \boldsymbol{X} = \boldsymbol{x}\right\} = \Pr\left\{\Theta_i > \widehat{\theta}_i(\boldsymbol{x}) \,|\, \boldsymbol{X} = \boldsymbol{x}\right\}, \tag{6.47}$$

which means that $\widehat{\theta}_i(\boldsymbol{x})$ is the *median* of the marginal posterior distribution of Θ_i given $\boldsymbol{X} = \boldsymbol{x}$, *i.e.* $\pi(\theta_i \,|\, \boldsymbol{X} = \boldsymbol{x})$

6.4.3 Marginal Maximum *A Posteriori* (MAP) estimation

Another commonly used cost function in the cases where $\boldsymbol{\mathcal{T}} = \mathbb{R}^m$ is

$$c[\boldsymbol{a}, \boldsymbol{\theta}] = \sum_{i=1}^{m} c[a_i - \theta_i] \quad \text{with} \quad c[a_i - \theta_i] = \begin{cases} 0 & \text{if} \quad |a_i - \theta_i| \leq \Delta \\ 1 & \text{if} \quad |a_i - \theta_i| > \Delta \end{cases} \tag{6.48}$$

where Δ is a positive real number. The corresponding estimate is given by :

$$\widehat{\theta}_i = \arg\max_{\theta_i \in \mathcal{T}} \{\pi(\theta_i \,|\, \boldsymbol{x})\} \tag{6.49}$$

if Δ is sufficiently small.

6.4.4 Maximum *A Posteriori* (MAP) estimation

Two other cost functions which give the same estimates are :

$$c[\boldsymbol{a}, \boldsymbol{\theta}] = \begin{cases} 0 & \text{if} \quad \max_i |a_i - \theta_i| \leq \Delta \\ 1 & \text{if} \quad \max_i |a_i - \theta_i| > \Delta \end{cases} \tag{6.50}$$

and

$$c[\boldsymbol{a}, \boldsymbol{\theta}] = \begin{cases} 0 & \text{if} \quad \|\boldsymbol{a} - \boldsymbol{\theta}\|^2 \leq \Delta \\ 1 & \text{if} \quad \|\boldsymbol{a} - \boldsymbol{\theta}\|^2 > \Delta \end{cases} \tag{6.51}$$

where Δ is a positive real number.

In both cases, if the posterior distribution $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ is continuous and smooth enough, we obtain the MAP estimate.

The corresponding estimate is given by :

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\mathcal{T}}} \{\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\} \tag{6.52}$$

Note that the MMAP estimate in (6.49) and the estimate (6.52) may be very different.

6.4.5 Estimation of a Gaussian vector parameter from jointly Gaussian observation

The case of the estimation of a Gaussian vector parameter $\boldsymbol{\theta} \in \mathbb{R}^m$ from a jointly Gaussian observation $\boldsymbol{x} \in \mathbb{R}^n$ is a very useful example and is used in many applications.

Suppose $\boldsymbol{\Theta}$ and $\boldsymbol{X}$ have the following *a priori* distributions:

$$\begin{aligned} \boldsymbol{\Theta} &\sim \mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{R}_\Theta) \\ \boldsymbol{X} &\sim \mathcal{N}(\boldsymbol{x}_0, \boldsymbol{R}_X) \end{aligned}$$

and

$$\begin{pmatrix} \boldsymbol{\Theta} \\ \boldsymbol{X}_0 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\theta}_0 \\ \boldsymbol{x}_0 \end{pmatrix} \ , \ \begin{pmatrix} \boldsymbol{R}_\Theta & \boldsymbol{R}_{\Theta X} \\ \boldsymbol{R}_{X\Theta} & \boldsymbol{R}_X \end{pmatrix} \right)$$

with $\boldsymbol{R}_{\Theta X} = \boldsymbol{R}_{X\Theta}^t$.

It is easy to show that the posterio law is also Gaussian and is given by

$$\boldsymbol{\Theta}|\boldsymbol{X} \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{R}}) \tag{6.53}$$

with

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \boldsymbol{R}_{\Theta X}\boldsymbol{R}_X^{-1}(\boldsymbol{x} - \boldsymbol{x}_0) \tag{6.54}$$

$$\widehat{\boldsymbol{R}} = \boldsymbol{R}_\Theta - \boldsymbol{R}_{\Theta X}\boldsymbol{R}_X^{-1}\boldsymbol{R}_{X\Theta} \tag{6.55}$$

We also have

$$\mathrm{E}\left[\boldsymbol{\Theta} \,|\, \boldsymbol{X} = \boldsymbol{x}\right] = \widehat{\boldsymbol{\theta}} \tag{6.56}$$

$$\mathrm{Cov}\left\{\boldsymbol{\Theta} \,|\, \boldsymbol{X} = \boldsymbol{x}\right\} = \widehat{\boldsymbol{R}} \tag{6.57}$$

The corresponding minimum Bayes risk is

$$r(\widehat{\boldsymbol{\theta}}) = \mathrm{tr}\left\{\boldsymbol{Q}\widehat{\boldsymbol{R}}\right\} = \mathrm{tr}\left\{\boldsymbol{Q}\boldsymbol{R}_\Theta\right\} - \mathrm{tr}\left\{\boldsymbol{Q}\boldsymbol{R}_{\Theta X}\boldsymbol{R}_X^{-1}\boldsymbol{R}_{X\Theta}\right\} \tag{6.58}$$

6.4.6 Case of linear models

When the observation vector is related to the vector parameter $\boldsymbol{\theta}$ by a linear model we have

$$X_i = \sum_{j=1}^{m} h_{i,j}\,\Theta_j + N_i, \quad i = 1, \ldots, n \tag{6.59}$$

or in a matrix form

$$\boldsymbol{X} = \boldsymbol{H}\boldsymbol{\Theta} + \boldsymbol{N} \tag{6.60}$$

with

$$\begin{aligned} \boldsymbol{\Theta} &\sim \mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{R}_\Theta) \\ \boldsymbol{N} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_N) \end{aligned}$$

Then we have

$$\begin{aligned} \boldsymbol{X}|\boldsymbol{\Theta} = \boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{H}\boldsymbol{\theta}, \boldsymbol{R}_N) \\ \\ \boldsymbol{R}_X &= \boldsymbol{H}\boldsymbol{R}_\Theta\boldsymbol{H}^t + \boldsymbol{H}\boldsymbol{R}_{\Theta N} + \boldsymbol{R}_{N\Theta}\boldsymbol{H}^t + \boldsymbol{R}_N \\ \boldsymbol{R}_{X\Theta} &= \boldsymbol{H}\boldsymbol{R}_\Theta + \boldsymbol{R}_{\Theta N}, \quad \boldsymbol{R}_{\Theta X} = \boldsymbol{R}_{X\Theta}^t \end{aligned}$$

If we assume that the noise $\boldsymbol{N}$ and the vector parameter $\boldsymbol{\Theta}$ are independant, we have

$$\boldsymbol{R}_X = \boldsymbol{H}\boldsymbol{R}_\Theta\boldsymbol{H}^t + \boldsymbol{R}_N \tag{6.61}$$

$$\boldsymbol{R}_{X\Theta} = \boldsymbol{H}\boldsymbol{R}_\Theta, \quad \boldsymbol{R}_{\Theta X} = \boldsymbol{R}_\Theta^t\boldsymbol{H}^t \tag{6.62}$$

$$\boldsymbol{R}_X^{-1} = \boldsymbol{R}_N^{-1} - \boldsymbol{R}_N^{-1}\boldsymbol{H}\left(\boldsymbol{R}_\Theta^{-1} + \boldsymbol{H}^t\boldsymbol{R}_N^{-1}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^t\boldsymbol{R}_\Theta^{-1} \tag{6.63}$$

and

$$\boldsymbol{\Theta}|\boldsymbol{X} \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{R}}) \tag{6.64}$$

with

$$\left\{\begin{aligned} \widehat{\boldsymbol{\theta}} &= \mathrm{E}\left[\boldsymbol{\Theta} \,|\, \boldsymbol{x}\right] = \boldsymbol{\theta}_0 + \boldsymbol{R}_{\Theta X}\boldsymbol{R}_X^{-1}(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}_0) \\ \widehat{\boldsymbol{R}} &= \mathrm{E}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \,|\, \boldsymbol{x}\right] = \boldsymbol{R}_\Theta - \boldsymbol{R}_{\Theta X}\boldsymbol{R}_X^{-1}\boldsymbol{R}_{X\Theta} \end{aligned}\right.$$

$$\left\{\begin{aligned} \widehat{\boldsymbol{\theta}} &= \boldsymbol{\theta}_0 + \boldsymbol{R}_\Theta\boldsymbol{H}^t\left[\boldsymbol{H}\boldsymbol{R}_\Theta\boldsymbol{H}^t + \boldsymbol{R}_N\right]^{-1}(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}_0) \\ &= \boldsymbol{\theta}_0 + \left[\boldsymbol{H}^t\boldsymbol{R}_N^{-1}\boldsymbol{H} + \boldsymbol{R}_\Theta^{-1}\right]^{-1}\boldsymbol{H}^t\boldsymbol{R}_N^{-1}(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}_0) \\ \widehat{\boldsymbol{R}} &= \boldsymbol{R}_\Theta - \boldsymbol{R}_\Theta\boldsymbol{H}^t\left[\boldsymbol{H}\boldsymbol{R}_\Theta\boldsymbol{H}^t + \boldsymbol{R}_N\right]^{-1}\boldsymbol{H}\boldsymbol{R}_\Theta \\ &= \left[\boldsymbol{H}^t\boldsymbol{R}_N^{-1}\boldsymbol{H} + \boldsymbol{R}_\Theta^{-1}\right]^{-1} \end{aligned}\right. \tag{6.65}$$

Consider now the particular case of $\boldsymbol{R}_N = \sigma_b^2\boldsymbol{I}$, $\boldsymbol{R}_\Theta = \sigma_x^2(\boldsymbol{D}^t\boldsymbol{D})^{-1}$ and $\boldsymbol{\theta}_0 = \boldsymbol{0}$. We then have

$$\left\{\begin{aligned} \widehat{\boldsymbol{\theta}} &= \left(\boldsymbol{H}^t\boldsymbol{H} + \boldsymbol{\lambda}\boldsymbol{D}^t\boldsymbol{D}\right)^{-1}\boldsymbol{H}^t\boldsymbol{x}, \\ \widehat{\boldsymbol{R}} &= \sigma_b^2\left(\boldsymbol{H}^t\boldsymbol{H} + \lambda\boldsymbol{D}^t\boldsymbol{D}\right)^{-1}, \quad \text{with} \quad \lambda = \sigma_b^2/\sigma_b^2 \end{aligned}\right. \tag{6.66}$$

6.5 Examples

6.5.1 curve fitting

We consider here a classical problem of curve fitting that any engineer is almost anytime faced to. We analyse this problem as a parameter estimation : Given a set of data $\{(x_i, t_i), i = 1, \ldots, n\}$ estimate the parameters of an algebraic curve to fit the best these data. Among different curves, the polynomials are used very commonly.

A polynomial model of degree p relating $x_i = x(t_i)$ to t_i is

$$x_i = x(t_i) = \theta_0 + \theta_1 t_i + \theta_2 t_i^2 + \cdots + \theta_p t_i^p, \quad i = 1, \ldots, n \tag{6.67}$$

Noting that this relation is linear in θ_i, we can rewrite it in the following

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & \cdots & t_1^p \\ 1 & t_2 & t_2^2 & \cdots & \cdots & t_2^p \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & t_n & t_n^2 & \cdots & \cdots & t_n^p \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} \tag{6.68}$$

or

$$\boldsymbol{x} = \boldsymbol{H\theta} \tag{6.69}$$

The matrix $\boldsymbol{H}$ is called the *Vandermond* matrix. It is entirely determined by the vector $\boldsymbol{t} = [t_1, t_2, \ldots, t_n]^t$.

In the case where $n = p + 1$, this matrix is invertible iff $t_i \neq t_j, \forall i \neq j$. In general, however we have more data than unknowns, *i.e.* $n > p + 1$.

Note that the matrix $\boldsymbol{H}^t\boldsymbol{H}$ is a *Hankel* matrix:

$$[\boldsymbol{H}^t\boldsymbol{H}]_{kl} = \sum_{i=1}^n t_i^{k-1}\, t_i^{l-1} = \sum_{i=1}^n t_i^{k+l-2}, \quad k, l = 1, \ldots, p + 1 \tag{6.70}$$

and the vector $\boldsymbol{H}^t\boldsymbol{x}$ is such that

$$[\boldsymbol{H}^t\boldsymbol{x}]_k = \sum_{i=1}^n t_i^{k-1}\, x_i, \quad k = 1, \ldots, p + 1 \tag{6.71}$$

Line fitting is the following particular case

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} \tag{6.72}$$

In this case we have

$$\boldsymbol{H}^t\boldsymbol{H} = \begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix} \tag{6.73}$$

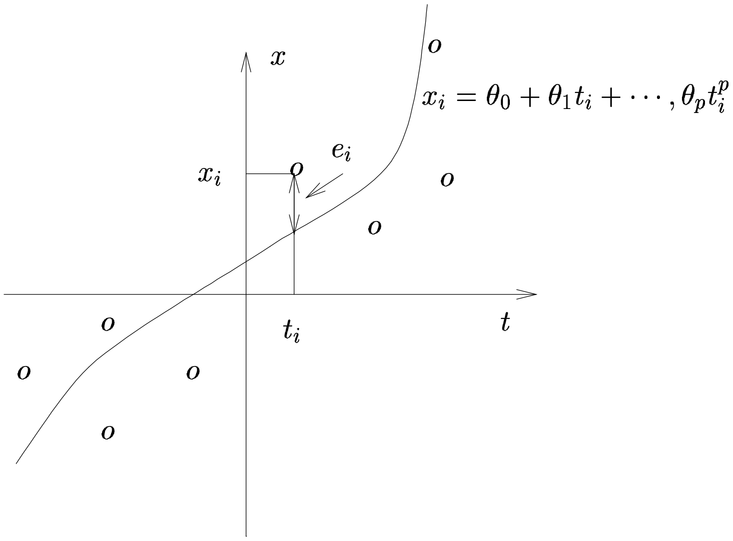


Figure 6.1: Curve fitting.

and

$$\boldsymbol{H}^t\boldsymbol{x} = \begin{pmatrix} \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} t_i\, x_i \end{pmatrix} \tag{6.74}$$

In the following we consider the line fitting case and will see how different assumptions about the problem can give different solutions.

Model 1:
The easiest model is to assume that t_i are perfectly known, and we only have uncertainties on x_i, *i.e.*

$$x_i = x(t_i) = \theta_0 + \theta_1 t_i + e_i, \quad i = 1, \ldots, n \tag{6.75}$$

where e_i represents the error on x_i. In a geometric language, e_i is the signed distance between the point (t_i, x_i) and the point $(t_i, \theta_0 + \theta_1 t_i)$ (see figure 6.5.1).

Here, we have $\boldsymbol{x} = \boldsymbol{H\theta} + \boldsymbol{e}$ with $\boldsymbol{\theta} = [\theta_0, \theta_1]^t$. The matrix $\boldsymbol{H}$ is perfectly known. Note that in this model, if we assume that e_i are zero mean, white and Gaussian

$$e_i = x_i - \theta_0 - \theta_1 t_i \sim \mathcal{N}\left(0, \sigma_e^2\right) \tag{6.76}$$

then the likelihood function becomes

$$f(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}\right) \propto \exp\left[-\frac{1}{2\sigma_e^2}\|\boldsymbol{x} - \boldsymbol{H\theta}\|^2\right] \tag{6.77}$$

and the maximum likelihood estimate is

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{\|\boldsymbol{x} - \boldsymbol{H\theta}\|^2\right\} \tag{6.78}$$

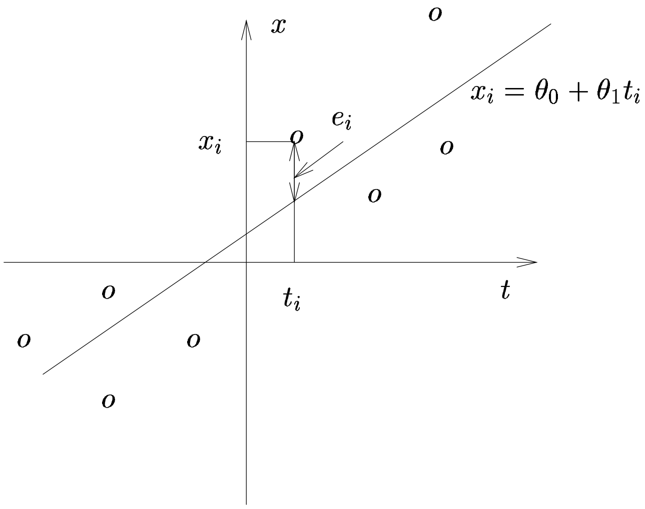


Figure 6.2: Line fitting: model 1: $e_i = x_i - (\theta_0 + \theta_1 t_i)$

If the $\boldsymbol{H}^t\boldsymbol{H}$ is invertible, $\widehat{\boldsymbol{\theta}}$ is given by

$$\widehat{\boldsymbol{\theta}} = [\boldsymbol{H}^t\boldsymbol{H}]^{-1}\boldsymbol{H}^t\boldsymbol{x} \tag{6.79}$$

To define any Bayesian estimate, we have to assign a prior probability law to $\boldsymbol{\theta}$. Let assume that θ_0 and θ_1 are independent and

$$\theta_0 \sim \mathcal{N}\left(0, \sigma_0^2\right), \qquad \theta_1 \sim \mathcal{N}\left(1, \sigma_1^2\right) \tag{6.80}$$

or

$$\boldsymbol{\theta} = \begin{pmatrix}\theta_0\\ \theta_1\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & 0\\ 0 & \sigma_1^2\end{pmatrix}\right) = \mathcal{N}\left(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_\theta\right) \tag{6.81}$$

Exercise 1:

- Write the complete expressions of $f(x_i \,|\, \boldsymbol{\theta})$, $f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$
- Show that the posterior law $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ is Gaussian, *i.e.*

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \sim \mathcal{N}\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Sigma}}\right) \tag{6.82}$$

  and give the expressions of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Sigma}}$.
- Show that the MAP estimate is obtained by

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{J(\boldsymbol{\theta})\} \tag{6.83}$$

  with

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{\sigma_e^2}\|\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta}\|^2 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^t\boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= \frac{1}{\sigma_e^2}\sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i)^2 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^t\boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \end{aligned}$$

- Show that, in this case an explicite expression of $\widehat{\boldsymbol{\theta}}$ is available and is given by

$$\widehat{\boldsymbol{\theta}} = \left[\frac{1}{\sigma_e^2}\boldsymbol{H}^t\boldsymbol{H} + \boldsymbol{\Sigma}_\theta^{-1}\right]^{-1}\left[\frac{1}{\sigma_e^2}\boldsymbol{H}^t\boldsymbol{x} + \boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\theta}_0\right] \tag{6.84}$$

- Compare this solution to the ML solution (6.79) which is equal to least square (LS) solution.

Model 2:
A little more complex model is

$$x_i = x(t_i) = \theta_0 + \theta_1 t_i + e_i, \quad i = 1, \ldots, n \tag{6.85}$$

with the assumption that

$$r_i = e_i \cos\phi = \frac{e_i}{\sqrt{1+\theta_1^2}} = \frac{x_i - \theta_0 - \theta_1 t_i}{\sqrt{1+\theta_1^2}}$$

the distance of the point (t_i, x_i) to the line $x(t_i) = \theta_0 + \theta_1 t_i$ is zero mean, white and Gaussian with known variance σ_r^2 (See figure 6.5.1.)

Note that r_i is no more a linear function of θ_1.

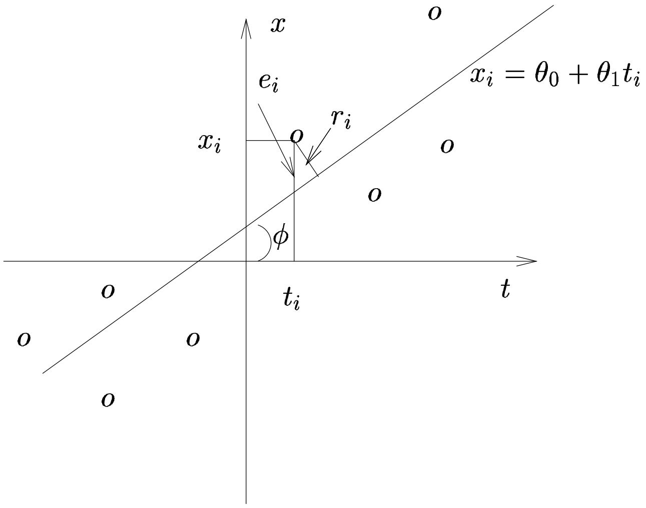


Figure 6.3: Line fitting: model 2: $r_i = e_i \cos\phi = \frac{e_i}{\sqrt{1+\theta_1^2}} = \frac{x_i-\theta_0-\theta_1 t_i}{\sqrt{1+\theta_1^2}}$

Exercise 2: With this model and assuming that r_i are zero mean, white and Gaussian with known variance $\sigma^2 = 1$:

- Write the expressions of $f(x_i \,|\, \boldsymbol{\theta})$, $f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$
- Show that the MAP estimate is obtained by

  $$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{J(\boldsymbol{\theta})\} \tag{6.86}$$

  with

  $$J(\boldsymbol{\theta}) = n \ln\left(2\pi(1+\theta_1^2)\sigma_r^2\right) + \frac{1}{(1+\theta_1^2)\sigma_r^2} \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i)^2 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^t \Sigma_\theta^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \tag{6.87}$$

  where $\boldsymbol{\theta} = [\theta_0, \theta_1]^t$.
- Is it possible to obtain explicit expressions for $\widehat{\theta}_0$ and $\widehat{\theta}_1$?

Model 3:

A little different model assumes that t_i are also uncertain, *i.e.*

$$x_i = x(t_i) = \theta_0 + \theta_1(t_i + \epsilon_i) + e_i, \quad i = 1, \ldots, n \tag{6.88}$$

where ϵ_i represents the error on t_i. Here also, we have $\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{e}$ with $\boldsymbol{\theta} = [\theta_0, \theta_1]^t$, but the matrix $\boldsymbol{H}$ is now uncertain.

Note that we have

$$x_i = x(t_i) = \theta_0 + \theta_1 t_i + \theta_1 \epsilon_i + e_i, \quad i = 1, \ldots, n \tag{6.89}$$

which can also be written as

$$\boldsymbol{x} = \boldsymbol{H}_0\boldsymbol{\theta} + \boldsymbol{H}_\epsilon\boldsymbol{\theta} + \boldsymbol{e}$$

with

$$\boldsymbol{H}_0 = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ & \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix}, \quad \boldsymbol{H}_\epsilon = \begin{pmatrix} 0 & \epsilon_1 \\ \vdots & \vdots \\ & \\ \vdots & \vdots \\ 0 & \epsilon_n \end{pmatrix},$$

Exercise 3: With this model and assuming that ϵ_i are zero mean, white and Gaussian with known variance σ_ϵ^2 and that e_i are also zero mean, white and Gaussian with known variance σ_e^2:

- Write the expressions of $f(x_i \,|\, \boldsymbol{\theta})$, $f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$
- Give the expressions of the ML and the MAP estimators.
- Compare them to the solutions of the previous cases.

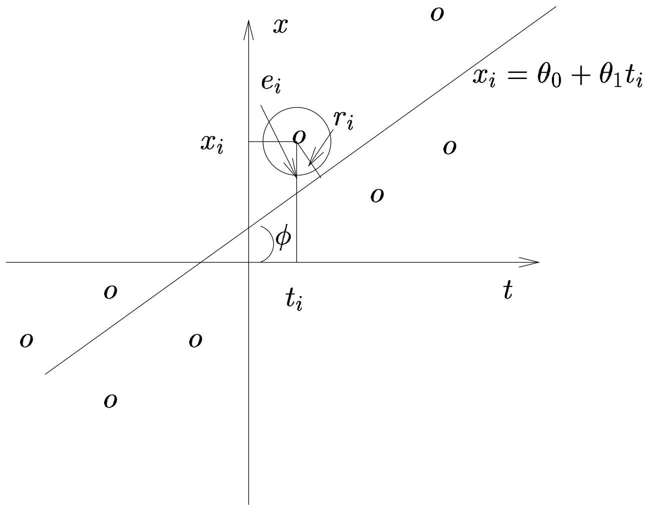


Figure 6.4: Line fitting: model 3

Model 4:

This is the combination of cases 2 and 3, *i.e.*

$$x_i = x(t_i) = \theta_0 + \theta_1 (t_i + \epsilon_i) + e_i, \quad i = 1, \ldots, n \tag{6.90}$$

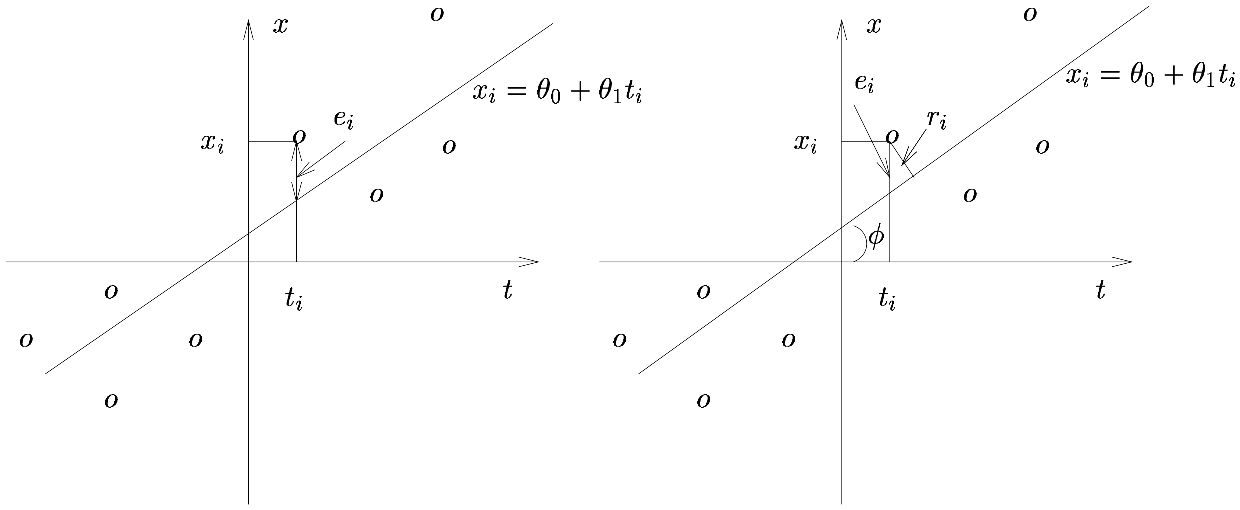
where

$$r_i = e_i \cos\phi = \frac{e_i}{\sqrt{1+\theta_1^2}} = \frac{x_i - \theta_0 - \theta_1 t_i}{\sqrt{1+\theta_1^2}}$$

the distance of the point (t_i, x_i) to the line $x(t_i) = \theta_0 + \theta_1 t_i$ are assumed zero mean, white and Gaussian.

Exercise 4: With this model and assuming that ϵ_i are zero mean, white and Gaussian with known variance σ_ϵ^2 and that r_i are also zero mean, white and Gaussian with known variance σ_r^2:

- Write the expressions of $f(x_i \,|\, \boldsymbol{\theta})$, $f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$
- Give the expressions of the ML and the MAP estimators.
- Compare them to the solutions in previous examples.


x
e_i
x_i
$x_i = \theta_0 + \theta_1 t_i$
t_i
t
x
e_i
r_i
x_i
$x_i = \theta_0 + \theta_1 t_i$
ϕ
t_i
t

Model 1: $e_i = x_i - (\theta_0 + \theta_1 t_i)$ Model 2: $r_i = [x_i - (\theta_0 + \theta_1 t_i)] \cos \phi = \frac{x_i - (\theta_0 + \theta_1 t_i)}{\sqrt{1+\theta_1^2}}$

Figure 6.5: Line fitting: models 1 and 2.

$$\begin{aligned}
\boldsymbol{x}|\boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{H}\boldsymbol{\theta},\, \boldsymbol{R}_N)\\
\boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{\theta}_0,\, \boldsymbol{R}_\Theta)\\
\boldsymbol{\theta}|\boldsymbol{x} &\sim \mathcal{N}(\widehat{\boldsymbol{\theta}},\, \widehat{\boldsymbol{R}})\\
\widehat{\boldsymbol{\theta}} &= \boldsymbol{\theta}_0 + \boldsymbol{R}_\Theta\boldsymbol{H}^t\left[\boldsymbol{H}\boldsymbol{R}_\Theta\boldsymbol{H}^t + \boldsymbol{R}_N\right]^{-1}(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta}_0)\\
&= \boldsymbol{\theta}_0 + \left[\boldsymbol{H}^t\boldsymbol{R}_N^{-1}\boldsymbol{H} + \boldsymbol{R}_\Theta^{-1}\right]^{-1}\boldsymbol{H}^t\boldsymbol{R}_N^{-1}(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta}_0)\\
\widehat{\boldsymbol{R}} &= \boldsymbol{R}_\Theta - \boldsymbol{R}_\Theta\boldsymbol{H}^t\left[\boldsymbol{H}\boldsymbol{R}_\Theta\boldsymbol{H}^t + \boldsymbol{R}_N\right]^{-1}\boldsymbol{H}\boldsymbol{R}_\Theta\\
&= \left[\boldsymbol{H}^t\boldsymbol{R}_N^{-1}\boldsymbol{H} + \boldsymbol{R}_\Theta^{-1}\right]^{-1}
\end{aligned}$$

Model 1:

$$\begin{aligned}
e_i &\sim \mathcal{N}(0,\, \sigma_e^2)\\
x_i|\boldsymbol{\theta} &\sim \mathcal{N}(\theta_0+\theta_1 t_i,\, \sigma_e^2)\\
\boldsymbol{x}|\boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{H}\boldsymbol{\theta},\, \sigma_e^2\boldsymbol{I})\\
\widehat{\boldsymbol{\theta}} &= \boldsymbol{\theta}_0 + \left[\boldsymbol{H}^t\boldsymbol{H} + \sigma_e^2\boldsymbol{R}_\Theta^{-1}\right]^{-1}\boldsymbol{H}^t(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta}_0)\\
\widehat{\boldsymbol{R}} &= \sigma_e^2\left[\boldsymbol{H}^t\boldsymbol{H} + \sigma_e^2\boldsymbol{R}_\Theta^{-1}\right]^{-1}
\end{aligned}$$

Model 2:

$$\begin{aligned}
r_i &= \frac{e_i}{\sqrt{1+\theta_1^2}} \sim \mathcal{N}(0,\, \sigma_r^2)\\
e_i &\sim \mathcal{N}\left(0,\, (1+\theta_1^2)\sigma_r^2\right)\\
x_i|\boldsymbol{\theta} &\sim \mathcal{N}\left(\theta_0+\theta_1 t_i,\, (1+\theta_1^2)\sigma_r^2\right)\\
\boldsymbol{x}|\boldsymbol{\theta} &\sim \mathcal{N}\left(\boldsymbol{H}\boldsymbol{\theta},\, (1+\theta_1^2)\sigma_r^2\boldsymbol{I}\right)\\
\pi(\theta|\boldsymbol{x}) &= \tfrac{1}{m(\boldsymbol{x})}\left[2\pi(1+\theta_1^2)\sigma_e^2\right]^{-n/2}\exp\left[-\tfrac{1}{2(1+\theta_1^2)\sigma_e^2}\|\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta}\|^2 - \tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^t\boldsymbol{R}_\Theta^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right]\\
\widehat{\boldsymbol{\theta}}_{MAP} &= \arg\min_{\boldsymbol{\theta}}\left\{J(\boldsymbol{\theta})\right\}\\
J(\boldsymbol{\theta}) &= -\tfrac{n}{2}\ln[2\pi(1+\theta_1^2)\sigma_e^2] - \tfrac{1}{2(1+\theta_1^2)\sigma_e^2}\|\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta}\|^2 - \tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^t\boldsymbol{R}_\Theta^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)
\end{aligned}$$

Model 3:

$$\begin{aligned}
\epsilon_i &\sim \mathcal{N}(0,\, \sigma_\epsilon^2)\\
e_i &\sim \mathcal{N}(0,\, \sigma_e^2)\\
x_i|\boldsymbol{\theta} &\sim \mathcal{N}\left(\theta_0+\theta_1 t_i,\, \theta_1^2\sigma_\epsilon^2+\sigma_e^2\right)\\
\boldsymbol{x}|\boldsymbol{\theta} &\sim \mathcal{N}\left(\boldsymbol{H}\boldsymbol{\theta},\, (\theta_1^2\sigma_\epsilon^2+\sigma_e^2)\boldsymbol{I}\right)\\
\pi(\theta|\boldsymbol{x}) &= \tfrac{1}{m(\boldsymbol{x})}\left[2\pi(\theta_1^2\sigma_\epsilon^2+\sigma_e^2)\right]^{-n/2}\exp\left[-\tfrac{1}{2(\theta_1^2\sigma_\epsilon^2+\sigma_e^2)}\|\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta}\|^2 - \tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^t\boldsymbol{R}_\Theta^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right]\\
\widehat{\boldsymbol{\theta}}_{MAP} &= \arg\min_{\boldsymbol{\theta}}\left\{J(\boldsymbol{\theta})\right\}\\
J(\boldsymbol{\theta}) &= -\tfrac{n}{2}\ln[2\pi(\theta_1^2\sigma_\epsilon^2+\sigma_e^2)] - \tfrac{1}{2(\theta_1^2\sigma_\epsilon^2+\sigma_e^2)}\|\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta}\|^2 - \tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^t\boldsymbol{R}_\Theta^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)
\end{aligned}$$

Model 4:

$$\begin{aligned}
\epsilon_i &\sim \mathcal{N}(0,\, \sigma_\epsilon^2)\\
e_i &\sim \mathcal{N}(0,\, (1+\theta_1)^2\sigma_e^2)\\
x_i|\boldsymbol{\theta} &\sim \mathcal{N}\left(\theta_0+\theta_1 t_i,\, \theta_1^2\sigma_\epsilon^2+(1+\theta_1)^2\sigma_e^2\right)\\
&= \mathcal{N}\left(\theta_0+\theta_1 t_i,\, \theta_1^2(\sigma_\epsilon^2+\sigma_e^2)+\sigma_e^2\right)\\
\boldsymbol{x}|\boldsymbol{\theta} &\sim \mathcal{N}\left(\boldsymbol{H}\boldsymbol{\theta},\, (\theta_1^2(\sigma_\epsilon^2+\sigma_e^2)+\sigma_e^2)\boldsymbol{I}\right)\\
\pi(\theta|\boldsymbol{x}) &= \tfrac{1}{m(\boldsymbol{x})}\left[2\pi(\theta_1^2(\sigma_\epsilon^2+\sigma_e^2)+\sigma_e^2)\right]^{-n/2}\exp\left[-\tfrac{1}{2(\theta_1^2(\sigma_\epsilon^2+\sigma_e^2)+\sigma_e^2)}\|\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta}\|^2 - \tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^t\boldsymbol{R}_\Theta^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right]\\
\widehat{\boldsymbol{\theta}}_{MAP} &= \arg\min_{\boldsymbol{\theta}}\left\{J(\boldsymbol{\theta})\right\}\\
J(\boldsymbol{\theta}) &= -\tfrac{n}{2}\ln[(\theta_1^2(\sigma_\epsilon^2+\sigma_e^2)+\sigma_e^2)] - \tfrac{1}{2(\theta_1^2(\sigma_\epsilon^2+\sigma_e^2)+\sigma_e^2)}\|\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta}\|^2 - \tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^t\boldsymbol{R}_\Theta^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)
\end{aligned}$$

Remark: To do these calculations easily we need the following relations:

- If $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{A}+\boldsymbol{B}$ are invertible, then we have

$$\begin{aligned} \left[\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right]^{-1} &= \boldsymbol{A}\left[\boldsymbol{A}+\boldsymbol{B}\right]^{-1}\boldsymbol{B}=\boldsymbol{B}\left[\boldsymbol{A}+\boldsymbol{B}\right]^{-1}\boldsymbol{A} \qquad (6.91)\\ \left[\boldsymbol{A}+\boldsymbol{B}\right]^{-1} &= \boldsymbol{A}^{-1}\left[\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right]^{-1}\boldsymbol{B}^{-1}=\boldsymbol{B}^{-1}\left[\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right]^{-1}\boldsymbol{A}^{-1} \end{aligned}$$

- If $\boldsymbol{A}$ and $\boldsymbol{C}$ are invertible matrices, then we have

$$\left[\boldsymbol{A}+\boldsymbol{B}\boldsymbol{C}\boldsymbol{D}\right]^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}\left[\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B}+\boldsymbol{C}^{-1}\right]^{-1}\boldsymbol{D}\boldsymbol{A}^{-1} \qquad (6.92)$$

- A special case very useful in system theory

$$\left[\boldsymbol{I}+\boldsymbol{B}(s\boldsymbol{I}-\boldsymbol{C})^{-1}\boldsymbol{D}\right]^{-1}=\boldsymbol{I}-\boldsymbol{B}\left[s\boldsymbol{I}-\boldsymbol{C}+\boldsymbol{D}\boldsymbol{B}\right]^{-1}\boldsymbol{D} \qquad (6.93)$$

- If $\boldsymbol{A}$ is invertible then,

$$\left[\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^t\right]^{-1}=\boldsymbol{A}^{-1}-\frac{(\boldsymbol{A}^{-1}\boldsymbol{u})(\boldsymbol{v}^t\boldsymbol{A}^{-1})}{1+\boldsymbol{v}^t\boldsymbol{A}^{-1}\boldsymbol{u}} \qquad (6.94)$$

- If $\boldsymbol{A}$ is a bloc matrix

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{pmatrix}$$

then $\boldsymbol{B}=\boldsymbol{A}^{-1}$ is also a bloc matrix

$$\boldsymbol{B}=\begin{pmatrix}\boldsymbol{B}_{11} & \boldsymbol{B}_{12}\\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22}\end{pmatrix}$$

and

  - If $\boldsymbol{A}_{22}^{-1}$ exists, then

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{I} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\\ 0 & \boldsymbol{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & 0\\ 0 & \boldsymbol{A}_{22}\end{pmatrix}\begin{pmatrix}\boldsymbol{I} & 0\\ \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{I}\end{pmatrix}$$

and we have

$$\operatorname{rank}\{\boldsymbol{A}\}=\operatorname{rank}\left\{\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\right\}+\operatorname{rank}\{\boldsymbol{A}_{22}\}$$

$\boldsymbol{A}^{-1}$ exists iff the matrix $\boldsymbol{T}=\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$ is invertible. Then we have

$$\begin{aligned} \boldsymbol{B}_{11}=\boldsymbol{T}^{-1} &= \left(\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\right)^{-1}\\ \boldsymbol{B}_{22}= &= \boldsymbol{A}_{22}^{-1}+\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{B}_{11}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\\ \boldsymbol{B}_{12} &= -\boldsymbol{B}_{11}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\\ \boldsymbol{B}_{21} &= -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{B}_{11} \end{aligned}$$

Written differently, we have

$$\begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{pmatrix}^{-1}=\begin{pmatrix}(\boldsymbol{A}_{11}-\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} & -\boldsymbol{B}_{11}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{B}_{11} & \boldsymbol{A}_{22}^{-1}+\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{B}_{11}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\end{pmatrix}$$

– If $\boldsymbol{A}_{22}^{-1}$ exists, then we have

$$\begin{aligned}
\boldsymbol{B}_{11} = &= \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{B}_{22}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} \\
\boldsymbol{B}_{22} = \boldsymbol{D}^{-1} &= (\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})^{-1} \\
\boldsymbol{B}_{12} &= -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{B}_{22} \\
\boldsymbol{B}_{21} &= -\boldsymbol{B}_{22}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}
\end{aligned}$$

The matrices $\boldsymbol{T}$ and $\boldsymbol{D}$ are called the *Shur's complement* of the matrix $\boldsymbol{A}$.

- **Particular case 1 :**
  If $\boldsymbol{A}$ is a superior bloc-triangular, *i.e.* $\boldsymbol{A}_{21} = \boldsymbol{0}$, then $\boldsymbol{B}$ is also superior bloc-triangular, *i.e.*

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{A}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} \end{pmatrix}$$

- **Particular case 2 :**
  If $\boldsymbol{A}$ is an inferior bloc-triangular matrix, *i.e.* $\boldsymbol{A}_{12} = \boldsymbol{0}$, then $\boldsymbol{B}$ is also an inferior bloc-triangular matrix, *i.e.*

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{A}_{11}^{-1} & \boldsymbol{0} \\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{A}_{22}^{-1} \end{pmatrix}$$

- **Particular case 3 :**
  If $\boldsymbol{A}_{22}$ is a sclar and $\boldsymbol{A}_{21}$ and $\boldsymbol{A}_{12}$ are vectors, we have

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{x} \\ \boldsymbol{z}^t & y \end{pmatrix}, \quad \boldsymbol{B} = \boldsymbol{A}^{-1} = \frac{1}{\alpha}\begin{pmatrix} \alpha\boldsymbol{A}_{11}^{-1} + \boldsymbol{w}\boldsymbol{v}^t & \boldsymbol{w} \\ \boldsymbol{v}^t & 1 \end{pmatrix}$$

  where α, $\boldsymbol{w}$, and $\boldsymbol{v}$ are given by:

$$\alpha = \frac{1}{(y - \boldsymbol{z}^t\boldsymbol{A}_{11}^{-1}\boldsymbol{x})} = \frac{|\boldsymbol{A}|}{|\boldsymbol{A}_{11}|}, \quad \boldsymbol{w} = -\boldsymbol{A}_{11}^{-1}\boldsymbol{x}, \quad \boldsymbol{v} = -\boldsymbol{A}_{11}^{-t}\boldsymbol{z}$$

- If $\boldsymbol{A}$ is a $[N, P]$ matrix

$$\boldsymbol{I}_N \pm \boldsymbol{A}\boldsymbol{A}^t = [\boldsymbol{I}_N \pm \boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{A}^t].[\boldsymbol{I}_N \pm \boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{A}^t]^t$$

  where $\boldsymbol{R} = \boldsymbol{I}_{\boldsymbol{P}} + [\boldsymbol{I}_{\boldsymbol{P}} \pm \boldsymbol{A}^t\boldsymbol{A}]^{1/2}$.

- If $\boldsymbol{x}$ is a vector and $u(\boldsymbol{x})$ a scalar function of $\boldsymbol{x}$ and if we define the gradient vector $\nabla u = \frac{\partial u}{\partial \boldsymbol{x}} = \left[\frac{\partial u}{\partial x_i}\right]$, then we have the following relations
  – If $u = \boldsymbol{\theta}^t\boldsymbol{x}$ then $\nabla u = \frac{\partial u}{\partial \boldsymbol{x}} = \boldsymbol{\theta}$
  – If $u = \boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}$ then $\nabla u = \frac{\partial u}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$

Generalized Gaussian

$$p(x) = \frac{p^{1-1/p}}{2\sigma\Gamma(1/p)} \exp\left[-\frac{1}{p}\frac{|x-x_0|^p}{\sigma^p}\right]$$

$$p=1 \longrightarrow p(x) = \frac{1}{2\sigma}\exp\left[-\frac{|x-x_0|}{\sigma}\right]$$

$$p=2 \longrightarrow p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{|x-x_0|^2}{\sigma^2}\right]$$

$$p=\infty \longrightarrow p(x) = \begin{cases} 1/2\sigma & \text{if} \quad |x-x_0| < \sigma \\ 0 & \text{otherwise} \end{cases}$$

Centered case $x_0 = 0$.

$$p(x) = \frac{p^{1-1/p}}{2\sigma\Gamma(1/p)} \exp\left[-\frac{1}{p}\frac{|x|^p}{\sigma^p}\right]$$

$$p=1 \longrightarrow p(x) = \frac{1}{2\sigma}\exp\left[-\frac{|x|}{\sigma}\right]$$

$$p=2 \longrightarrow p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{|x|^2}{\sigma^2}\right]$$

$$p=\infty \longrightarrow p(x) = \begin{cases} 1/2\sigma & \text{if} \quad |x| < \sigma \\ 0 & \text{otherwise} \end{cases}$$

Multivariable case:
Separable:

$$p(\boldsymbol{x}) = \frac{p^{n-n/p}}{2^n\sigma^n\Gamma^n(1/p)} \exp\left[-\frac{1}{p\sigma^p}\sum_{i=1}^{n}|x_i|^p\right]$$

Correlated: Markov models

$$p(\boldsymbol{x}) = Z(\alpha)\exp\left[-\alpha\sum_{i=1}^{n}\sum_{j\ i}\phi(x_i - x_j)\right]$$

$$Z(\alpha) = \int \exp\left[-\alpha\sum_{i=1}^{n}\sum_{j\ i}\phi(x_i - x_j)\right] \mathrm{d}\boldsymbol{x}$$

Example: $\phi(x) = x^2, \quad j\ i = i-1$

$$p(\boldsymbol{x}) = Z(\alpha)\exp\left[-\alpha x_1^2 - \alpha\sum_{i=2}^{n}(x_i - x_{i-1})^2\right]$$

This can be written as

$$p(\boldsymbol{x}) = Z(\alpha)\exp\left[-\alpha\boldsymbol{x}^t\boldsymbol{D}^t\boldsymbol{D}\boldsymbol{x}\right]$$

with

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0 & & & & 0 \\ 1 & -1 & & & & \vdots \\ 0 & 1 & -1 & & & \\ \vdots & & 1 & -1 & & \\ & & & & & \\ 0 & \cdots & & 0 & 1 & -1 \end{pmatrix}$$

$$Z(\alpha) = (2\pi)^{-n/2}\,(2\alpha)^{n/2}|\boldsymbol{D}^t\boldsymbol{D}|$$

Extension :

$$p(\boldsymbol{x}) = Z(\alpha)\exp\left[-\alpha\phi(x_1) - \alpha\sum_{i=2}^{n}\phi(x_i - x_{i-1})\right] = Z(\alpha)\exp\left[-\alpha\sum_{i=1}^{n}\phi([\boldsymbol{D}\boldsymbol{x}]_i)\right]$$

with $\phi(x) = |x|^p$

The questions are:

$Z(\alpha)$ exists ?

Can we obtain an analytical expression for it ?

Chapter 7

Elements of signal estimation

In the previous chapters we discussed the methods for designing estimators for static parameter estimation. In this chapter we consider the case of dynamic or time varying parameters (signal estimation).

7.1 Introduction

In many time-varying systems, the physical quantities of interest $\boldsymbol{x}$ can be modeled as obeying a dynamic equation

$$\boldsymbol{x}_{n+1} = \boldsymbol{f}_n(\boldsymbol{x}_n, \boldsymbol{u}_n) \tag{7.1}$$

where

- $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$, is a sequence of vectors in $\mathbb{R}^N$, called the *state* of the system, representing the unknown quantities of interest;
- $\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots$, is a sequence of vectors in $\mathbb{R}^M$, called the *state input* of the system, representing the influencing quantities acting on $\boldsymbol{x}_n$;
- $\boldsymbol{f}_0, \boldsymbol{f}_1, \ldots$, is a sequence of functions mapping $\mathbb{R}^N \times \mathbb{R}^M$ to $\mathbb{R}^M$, called the *state equation* of the system, representing the dynamic model relating $\boldsymbol{x}_n$ and $\boldsymbol{u}_n$;

A dynamic system is such that, for any fixed k and l, $\boldsymbol{x}_k$ is completely determined from the state at time l and the inputs from times l up to $k-1$. So, complete determination of $\boldsymbol{x}_n, n = 1, 2, \ldots$ requires not only the inputs $\boldsymbol{u}_n, n = 0, 1, 2, \ldots$ but also the initial condition $\boldsymbol{x}_0$.

The equation (7.1) is called the *state equation.* Associated to this equation is the *observation equation*

$$\boldsymbol{z}_n = \boldsymbol{h}_n(\boldsymbol{x}_n, \boldsymbol{v}_n) \tag{7.2}$$

where

- $\boldsymbol{z}_0, \boldsymbol{z}_1, \ldots$, is a sequence of vectors in $\mathbb{R}^P$ representing the observable quantities;
- $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots$, is a sequence of vectors in $\mathbb{R}^P$ representing the errors on the observations;
- $\boldsymbol{h}_0, \boldsymbol{h}_1, \ldots$, is a sequence of functions mapping $\mathbb{R}^N \times \mathbb{R}^P$ to $\mathbb{R}^P$ representing the *observation model.*

The main problem then is to estimate the state vector $\boldsymbol{x}_k$ from the observations $\boldsymbol{z}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_l$.

Example 1: One-dimensional motion

Consider a moving target subjected to an acceleration A_t for $t > 0$. Its position X_t and its velocity V_t at time t satisfy

$$\left\{ \begin{array}{l} X_t = \frac{\mathrm{d}P_t}{\mathrm{d}t} \\ A_t = \frac{\mathrm{d}V_t}{\mathrm{d}t} \end{array} \right. \tag{7.3}$$

Assume that we can measure the position V_t at time instants $t_n = nT$ and we wish to write a model of type (7.1) describing its motion. Assuming T is small, a Taylor series approximation allows us to write

$$\left\{ \begin{array}{l} X_{n+1} \simeq X_n + TV_n \\ V_{n+1} \simeq V_n + TA_n \end{array} \right. \tag{7.4}$$

From these equations we see that two quantities X_n and V_n are necessary to describe the motion. So, defining

$$\left\{ \begin{array}{l} \boldsymbol{x} = \begin{pmatrix} X \\ V \end{pmatrix} \\ U = A \\ Z = X + V \end{array} \right. \longrightarrow \left\{ \begin{array}{l} \boldsymbol{x}_n = \begin{pmatrix} X_n \\ V_n \end{pmatrix} \\ U_n = A_n \\ Z_n = X_n + V_n \end{array} \right. \tag{7.5}$$

we can write

$$\left\{ \begin{array}{ll} \boldsymbol{x}_{n+1} & = \boldsymbol{F}\boldsymbol{x}_n + \boldsymbol{G}U_n \\ Z_n & = \boldsymbol{H}\boldsymbol{x}_n + V_n \end{array} \right. \text{with} \quad \boldsymbol{F} = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix}, \boldsymbol{G} = \begin{pmatrix} 0 \\ T \end{pmatrix}, \boldsymbol{H} = (1 \quad 0) \tag{7.6}$$

$$\left\{ \begin{array}{l} \boldsymbol{f}_n(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{F}\boldsymbol{x} + \boldsymbol{G}\boldsymbol{u} \\ \boldsymbol{h}_n(\boldsymbol{x}, \boldsymbol{v}) = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{v} \end{array} \right. \tag{7.7}$$

In this example, we assumed that we can measure directly the position of the moving target. In general, however, we may observe a quantity $z(n)$ related to the unknown quantity $x(n)$ by a linear transformation:

$$\begin{array}{ccccc} x(n) & \longrightarrow & \boxed{\text{Linear System}} & \longrightarrow & z(n) \\ \text{non observable} & & & & \text{observable} \end{array}$$

and we want to estimate $x(n)$ from the observed values of $\{z(n), n = 1, \ldots, k\}$. The estimate $\widehat{x}(n)$ is then a function of the data $\{z(n), n = 1, \ldots, k\}$ and we note

$$\widehat{x}(n \,|\, z(1), z(2), \ldots, z(k)) \overset{\text{def}}{=} \widehat{x}(n \,|\, k)$$

Three cases may occur:

- we may want to estimate $x(n+k)$ from the past observations. The estimate $\widehat{x}(n+k|n)$ is called the k-th *order prediction* of $z(n)$ and the estimation procedure is called *prediction.*

- we may want to estimate $x(n)$ from present and past observations. The estimate $\widehat{x}(n|n)$ is the *filtered value* of $z(n)$ and the estimation procedure is called *filtering.*

- we may want to estimate $x(n)$ from past, present and future observations. The estimate $\widehat{x}(n|n+l)$ is the *smoothed value* of $z(n)$ and the estimation procedure is called *smoothing.*

7.2 Kalman filtering : General linear case

In this section we consider the linear systems with finite dimensions described by the following equations:

$$\begin{cases} \boldsymbol{x}_{k+1} & = \boldsymbol{F}_k\,\boldsymbol{x}_k + \boldsymbol{G}_k\,\boldsymbol{u}_k & \text{state equation,} \\ \boldsymbol{z}_k & = \boldsymbol{H}_k\;\boldsymbol{x}_k + \boldsymbol{v}_k & \text{observation equation} \end{cases}$$

where

- $k = 0, 1, 2, \ldots$ represents the discrete time ;

- $\boldsymbol{x}_k$ is a N-dimensional vector called *state vector* of the system ;

- $\boldsymbol{z}_k$ is a P-dimensional vector containing the observations (output of the system) ;

- $\boldsymbol{v}_k$ is a P-dimensional vector containing the observations errors (output noise of the system) ;

- $\boldsymbol{u}_k$ is a M-dimensional vector representing the state representation error (state space noise process) ;

- $\boldsymbol{F}_k$, $\boldsymbol{G}_k$ and $\boldsymbol{H}_k$ with respective dimensions of (N, N), (N, M) and (P, N) are the state transition, the state input and the observation matrices and are assumed to be known.

- The noise sequences $\{\boldsymbol{u}_k\}$ and $\{\boldsymbol{v}_k\}$ are assumed to be centered, white and jointly Gaussian.

- The initial state $\boldsymbol{x}_0$ is also assumed to be Gaussian and independent of $\{\boldsymbol{u}_k\}$ and $\{\boldsymbol{v}_k\}$:

  $$\mathrm{E}\left[\begin{pmatrix} \boldsymbol{v}_k \\ \boldsymbol{x}_0 \\ \boldsymbol{u}_k \end{pmatrix} \left(\, \boldsymbol{v}_l^t, \boldsymbol{x}_0^t, \boldsymbol{u}_l^t \,\right)\right] = \begin{pmatrix} \boldsymbol{R}_k & 0 & 0 \\ 0 & \boldsymbol{P}_0 & 0 \\ 0 & 0 & \boldsymbol{Q}_k \end{pmatrix} \delta_{kl}$$

  where $\boldsymbol{R}_k$ is the covariance matrix of the observation noise vector $\boldsymbol{v}_k$, $\boldsymbol{Q}_k$ is the covariance matrix of the state noise vector $\boldsymbol{Q}_k$ and $\boldsymbol{P}_0$ is the covariance matrix of the initial state $\boldsymbol{x}_0$.

Remember that the aim is to find a best estimate $\widehat{\boldsymbol{x}}_{k|l}$ of $\boldsymbol{x}_k$ from the observations $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_l$. Depending on the relative position of k with respect to i we have:

- If $k > l$ prediction

- If $k = l$ filtering
- If $k < l$ smoothing.

Of course, for fixed k and l, this problem is not different from the vector parameter estimation of the last chapter. However, we are usually interested in producing estimates eith in real time or at least on-line for increasing k.

Three different approaches can be used to obtain the Kalman filtering equations.

- Linear Mean Square (LMS) estimation :

  $$\widehat{\boldsymbol{x}}_{k|l} \stackrel{\text{def}}{=} \text{LMS}(\boldsymbol{x}_k \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_l)$$

  which minimizes

  $$\text{E}\left[[\boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|l}]^t \, \boldsymbol{W}_k \, [\boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|l}]\right]$$

- Maximum A posteriori (MAP) estimate :

  $$\widehat{\boldsymbol{x}}_{k|l} = \arg\max_{\boldsymbol{x}} \{p(\boldsymbol{x}_k \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_l)\}$$

- Bayesian MSE estimate :

  $$\widehat{\boldsymbol{x}}_{k|l} = \text{E}\left[\boldsymbol{x}_k \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\right]$$

We know that, for linear relations and Gaussian assumption all these estimates are equivalent. We consider here the last approach.

The main procedure is to apply the Bayes rule recursively to find the expression of the posterior law $p(\boldsymbol{x}_k \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_l)$. Note that, we can obtain easily this expression thanks to the following facts :

1. All variables are assumed Gaussian ;
2. $\{\boldsymbol{u}_k\}$ and $\{\boldsymbol{v}_k\}$ are assumed white, Gaussian and mutually independent ;
3. The state and the observation models are linear ;
4. All the conditional laws such as $p(\boldsymbol{z}_{k+1}|\boldsymbol{x}_{k+1})$ and $p(\boldsymbol{z}_{k+1}|\boldsymbol{z}_{1:k})$ are Gaussian. So, the posterior law

   $$p(\boldsymbol{x}_{k+1}|\boldsymbol{z}_{1:k+1}) = p(\boldsymbol{x}_{k+1}|\boldsymbol{z}_{1:k}) \, \frac{p(\boldsymbol{z}_{k+1}|\boldsymbol{x}_{k+1})}{p(\boldsymbol{z}_{k+1}|\boldsymbol{z}_{1:k})}$$

   is also Gaussian.

To obtain the equations in the general case we note

- $\widehat{\boldsymbol{x}}_{k|k}$ the estimate of the state vector at time k from the observations up to time k ;
- $\widehat{\boldsymbol{x}}_{k+1|k}$ the estimate of the state vector at time $k+1$ from the observations up to the instant k ;

- $\boldsymbol{e}_{k+1} = \boldsymbol{z}_{k+1} - H_{k+1}\,\widehat{\boldsymbol{x}}_{k+1|k}$ the *innovation process* of the observations at the instant $k+1$

- The covariance matrix of the innovation by

$$\boldsymbol{R}^e_{k+1} = \mathrm{E}\left[\boldsymbol{e}_{k+1|k}\,\boldsymbol{e}^t_{k+1|k}\right]$$

  which is diagonal;

- The covariance matrix of the prediction error by

$$\boldsymbol{P}_{k+1|k} = \mathrm{E}\left[\left[\boldsymbol{x}_{k+1} - \widehat{\boldsymbol{x}}_{k+1|k}\right]\left[\boldsymbol{x}_{k+1} - \widehat{\boldsymbol{x}}_{k+1|k}\right]^t\right],$$

- The posterior covariance matrix of the estimation error by

$$\boldsymbol{P}_{k+1|k+1} = \mathrm{E}\left[\left[\boldsymbol{x}_{k+1} - \widehat{\boldsymbol{x}}_{k+1|k+1}\right]\left[\boldsymbol{x}_{k+1} - \widehat{\boldsymbol{x}}_{k+1|k+1}\right]^t\right]$$

  which is also called the covariance matrix of the filtering error.

With these definitions, we obtain easily:

$$\begin{array}{lcl}
\mathrm{E}\left[\boldsymbol{x}_k|\boldsymbol{z}_{1:k}\right] & \stackrel{\text{def}}{=} & \widehat{\boldsymbol{x}}_{k|k} \\
\mathrm{Cov}\left\{\boldsymbol{x}_k|\boldsymbol{z}_{1:k}\right\} & \stackrel{\text{def}}{=} & \boldsymbol{P}_{k|k} \\
\mathrm{E}\left[\boldsymbol{x}_{k+1}|\boldsymbol{z}_{1:k}\right] & \stackrel{\text{def}}{=} & \widehat{\boldsymbol{x}}_{k+1|k} \\
\mathrm{Cov}\left\{\boldsymbol{x}_{k+1}|\boldsymbol{z}_{1:k}\right\} & \stackrel{\text{def}}{=} & \boldsymbol{P}_{k+1|k}
\end{array}$$

$$\begin{array}{lcl}
\mathrm{E}\left[\boldsymbol{z}_{k+1}|\boldsymbol{x}_{k+1}\right] & = & \boldsymbol{H}_{k+1}\,\widehat{\boldsymbol{x}}_{k+1} \\
\mathrm{Cov}\left\{\boldsymbol{z}_{k+1}|\boldsymbol{x}_{k+1}\right\} & = & \boldsymbol{R}_{k+1} \\
\\
\mathrm{E}\left[\boldsymbol{x}_{k+1}|\boldsymbol{z}_{1:k}\right] & = & \boldsymbol{F}_k\,\widehat{\boldsymbol{x}}_{k|k} \\
\mathrm{Cov}\left\{\boldsymbol{x}_{k+1}|\boldsymbol{z}_{1:k}\right\} & = & \boldsymbol{F}_k\,\boldsymbol{P}_k\boldsymbol{F}^t_k + \boldsymbol{G}_k\,\boldsymbol{Q}_k\,\boldsymbol{G}^t_k \\
\\
\mathrm{E}\left[\boldsymbol{z}_{k+1}|\boldsymbol{z}_{1:k}\right] & = & \boldsymbol{H}_{k+1}\,\boldsymbol{F}_k\,\widehat{\boldsymbol{x}}_{k|k} \\
\mathrm{Cov}\left\{\boldsymbol{z}_{k+1}|\boldsymbol{z}_{1:k}\right\} & = & \boldsymbol{H}_{k+1}\,\boldsymbol{P}_{k+1|k}\boldsymbol{H}^t_{k+1} + \boldsymbol{R}_{k+1}
\end{array}$$

Replacing the expressions of $p(\boldsymbol{z}_{k+1}|\boldsymbol{x}_{k+1})$ and $p(\boldsymbol{z}_{k+1}|\boldsymbol{z}_{1:k})$ we obtain :

$$p(\boldsymbol{x}_{k+1}|\boldsymbol{z}_{1:k+1}) = A\exp\left[-\frac{1}{2}[\boldsymbol{x}_{k+1} - \widehat{\boldsymbol{x}}_{k+1|k}]^t\boldsymbol{P}^{-1}_{k+1|k+1}[\boldsymbol{x}_{k+1} - \widehat{\boldsymbol{x}}_{k+1|k}]\right]$$

with

$$A = \frac{1}{(2\pi)^{n/2}}\left\{\left|\boldsymbol{H}_{k+1}\,\boldsymbol{P}_{k+1|k}\,\boldsymbol{H}^t_{k+1} + \boldsymbol{R}_k\right|^{1/2}\left|\boldsymbol{R}_k\right|^{-1/2}\left|\boldsymbol{P}_{k+1|k}\right|\right\}^{-1/2}$$

$$\begin{array}{rcl}
\widehat{\boldsymbol{x}}_{k+1|k} & = & \boldsymbol{F}_k\,\widehat{\boldsymbol{x}}_{k|k} + \boldsymbol{P}_{k+1|k}\,\boldsymbol{H}^t_{k+1}\left[\boldsymbol{R}_{k+1} + \boldsymbol{H}_{k+1}\,\boldsymbol{P}_{k+1|k}\,\boldsymbol{H}^t_{k+1}\right]^{-1}\left[\boldsymbol{z}_{k+1} - \boldsymbol{H}_{k+1}\,\boldsymbol{F}_k\,\widehat{\boldsymbol{x}}_{k|k}\right] \\
\boldsymbol{P}_{k+1|k} & = & \boldsymbol{F}_k\,\boldsymbol{P}_{k|k}\,\boldsymbol{F}^t_k + \boldsymbol{G}_k\,\boldsymbol{Q}_k\,\boldsymbol{G}^t_k \\
\boldsymbol{P}^{-1}_{k+1|k+1} & = & \boldsymbol{P}^{-1}_{k+1|k} + \boldsymbol{H}^t_{k+1}\boldsymbol{R}^{-1}_{k+1}\,\boldsymbol{H}_{k+1}
\end{array}$$

These equations can be rewritten in many different ways. Here are two of them:

- **Prediction-Correction form :**

  - **Prediction (Time update) :**

$$\begin{aligned}
\widehat{\boldsymbol{x}}_{k+1|k} &= \boldsymbol{F}_k\, \widehat{\boldsymbol{x}}_{k|k} \\
\boldsymbol{P}_{k+1|k} &= \boldsymbol{F}_k\, \boldsymbol{P}_{k|k}\, \boldsymbol{F}_k^t + \boldsymbol{G}_k\, \boldsymbol{Q}_k\, \boldsymbol{G}_k^t
\end{aligned}$$

  - **Correction (measurement update) :**

$$\begin{aligned}
\widehat{\boldsymbol{x}}_{k+1|k+1} &= \widehat{\boldsymbol{x}}_{k+1|k} + \boldsymbol{K}_{k+1}^g[\boldsymbol{z}_{k+1} - \boldsymbol{H}_{k+1}\, \widehat{\boldsymbol{x}}_{k+1|k}] \\
\boldsymbol{K}_{k+1}^g &= \boldsymbol{P}_{k+1|k}\, \boldsymbol{H}_{k+1}^t(\boldsymbol{R}_{k+1}^e)^{-1} \\
\boldsymbol{R}_{k+1}^e &= \boldsymbol{R}_{k+1} + \boldsymbol{H}_{k+1}\, \boldsymbol{P}_{k+1|k}\, \boldsymbol{H}_{k+1}^t \\
\boldsymbol{P}_{k+1|k+1} &= [\boldsymbol{I} - \boldsymbol{K}_{k+1}^f \boldsymbol{H}_{k+1}]\boldsymbol{P}_{k+1|k}
\end{aligned}$$

- **Compact form for prediction :**

$$\begin{aligned}
\widehat{\boldsymbol{x}}_{k+1|k} &= \boldsymbol{F}_k\, \widehat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{R}_k^e)^{-1}[\boldsymbol{z}_k - \boldsymbol{H}_k\, \widehat{\boldsymbol{x}}_{k|k-1}] \\
\boldsymbol{R}_k^e &= \boldsymbol{R}_k + \boldsymbol{H}_k\, \boldsymbol{P}_{k|k-1}\, \boldsymbol{H}_k^t \\
\boldsymbol{K}_k &= \boldsymbol{F}_k\, \boldsymbol{P}_{k|k-1}\, \boldsymbol{H}_k^t \\
\boldsymbol{P}_{k+1|k} &= \boldsymbol{F}_k\, \boldsymbol{P}_{k|k-1}\, \boldsymbol{F}_k^t + \boldsymbol{G}_k\, \boldsymbol{Q}_k\, \boldsymbol{G}_k^t - \boldsymbol{K}_k(\boldsymbol{R}_k^e)^{-1}\, \boldsymbol{K}_k^t
\end{aligned}$$

- **Compact form for filtering :**

$$\begin{aligned}
\widehat{\boldsymbol{x}}_{k|k} &= \widehat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k^g[\boldsymbol{z}_k - \boldsymbol{H}_k\, \widehat{\boldsymbol{x}}_{k|k-1}] \\
\boldsymbol{K}_k^g &= \boldsymbol{P}_{k|k-1}\, \boldsymbol{H}_k^t[\boldsymbol{R}_k + \boldsymbol{H}_k\, \boldsymbol{P}_{k|k-1}\, \boldsymbol{H}_k^t]^{-1} \\
\boldsymbol{P}_{k|k} &= \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k^g \boldsymbol{H}_k \boldsymbol{P}_{k|k-1} \\
\boldsymbol{P}_{k+1|k} &= \boldsymbol{F}_k \boldsymbol{P}_{k|k}\, \boldsymbol{F}_k^t + \boldsymbol{G}_k\, \boldsymbol{Q}_k\, \boldsymbol{G}_k^t
\end{aligned}$$

- **Very compact form for prediction :**

$$\begin{aligned}
\widehat{\boldsymbol{x}}_{k+1|k} &= \boldsymbol{F}_k \widehat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k^g[\boldsymbol{z}_k - \boldsymbol{H}_k\, \widehat{\boldsymbol{x}}_{k|k-1}] \\
\boldsymbol{K}_k^g &= \boldsymbol{P}_{k|k-1}\, \boldsymbol{H}_k^t[\boldsymbol{R}_k + \boldsymbol{H}_k\, \boldsymbol{P}_{k|k-1}\, \boldsymbol{H}_k^t]^{-1} \\
\boldsymbol{P}_{k+1|k} &= \boldsymbol{F}_k \boldsymbol{P}_{k|k-1}\, \boldsymbol{F}_k^t - \boldsymbol{F}_k \boldsymbol{K}_k^g \boldsymbol{H}_k \boldsymbol{P}_{k|k-1} \boldsymbol{F}_k^t + \boldsymbol{G}_k\, \boldsymbol{Q}_k\, \boldsymbol{G}_k^t
\end{aligned}$$

  where $\boldsymbol{K}_k$ is called the *Kalman filter gain* and $\boldsymbol{K}_k^g = \boldsymbol{K}_k(\boldsymbol{R}_k^e)^{-1}$ the generalized *Kalman filter gain.*

In all cases the initialization is :

$$\widehat{\boldsymbol{x}}_{0|-1} = 0 \quad \boldsymbol{P}_{0|-1} = \boldsymbol{P}_0$$

7.3 Examples

7.3.1 1D case:

$$\left\{ \begin{array}{lcl} x_{n+1} & = & f\,x_n + u_n \\ z_n & = & h\,x_n + v_n \end{array} \right. \tag{7.8}$$

where u_n and v_n are assumed independent, zero-mean, white and Gaussian with known variance q and r respectively. x_0 is also assumed Gaussian with known mean m_0 and known variance p_0

$$\left\{ \begin{array}{lcl} u_n & \sim & \mathcal{N}(0, q) \\ v_n & \sim & \mathcal{N}(0, r) \\ x_0 & \sim & \mathcal{N}(m_0, p_0) \end{array} \right. \tag{7.9}$$

The equations in this case reduce to

$$\left\{ \begin{array}{lcl} \widehat{x}_{n+1|n} & = & f\,\widehat{x}_{n|n} \\ \widehat{x}_{n|n} & = & \widehat{x}_{n|n-1} + k_n\left(z_n - h\,\widehat{x}_{n|n-1}\right) \\ k_n & = & \frac{p_{n|n-1}h}{h^2 p_{n|n-1}+r} = \frac{1}{h}\,\frac{p_{n|n-1}}{p_{n|n-1}+r/h^2} \end{array} \right. \tag{7.10}$$

The role of the Kalman gain in the measurement update is easily seen from these expressions. $p_{n|n-1}$ is the MSE incurred in the estimation of x_n from $z_{0:n-1}$, and the ratio r/h^2 is a measure of noisiness of the observations. It is interesting to compare these equations with the Bayesian estimation of the signal amplitude as described in Example ().

For this particular time-invariant model, we have

$$\left\{ \begin{array}{lcl} p_{n+1|n} & = & f^2\,p_{n|n} + q \\ p_{n|n} & = & \frac{1}{h}\,\frac{p_{n|n-1}}{\frac{r}{h^2}\,p_{n|n-1+1}} \end{array} \right. \tag{7.11}$$

We can eliminate the coupling between these equations and obtain

$$p_{n+1|n} = \frac{f^2\,p_{n|n-1}}{\frac{h^2}{r}\,p_{n|n-1} + 1} + q \tag{7.12}$$

We see here that as n increases, $p_{n+1|n}$ and so the gain k_n approaches a constant. Note that if $p_{n+1|n}$ does approach a constant, say p_∞, then p_∞ must satisfy

$$p_\infty = \frac{f^2\,p_\infty}{\frac{h^2}{r}\,p_\infty + 1} + q \tag{7.13}$$

This equation is quadratic and has a unique positive solution

$$p_\infty = \frac{1}{2}\left\{\left[\frac{r}{h^2}(1-f^2) - q\right]^2 + \frac{4rq}{h^2}\right\}^{1/2} - \frac{r}{2h^2}(1-f^2) + q \tag{7.14}$$

$$\begin{array}{rcl} \left|p_{n+1|n} - p_\infty\right| & = & f^2\left|\frac{p_{n|n-1}}{\frac{h^2}{r}\,p_{n|n-1}+1} - \frac{p_\infty}{\frac{h^2}{r}\,p_\infty + 1}\right| \\ & \leq & f^2\left|p_{n|n-1} - p_\infty\right| \end{array}$$

$$\begin{aligned}\left|p_{n+1|n} - p_\infty\right| \leq f^{2(n+1)} \left|p_0 - p_\infty\right| \\ \leq f^2 \left|p_{n|n-1} - p_\infty\right|\end{aligned}$$

This means that if $|f| < 1$ then $p_{n+1|n}$ converges to p_∞. So $|f| < 1$ is a sufficient condition for Kalman-Bucy filter to approach a steady state.

7.3.2 Track-While-Scan (TWS) Radar

Let consider the example of one dimentional moving target and assume that the target is subject to random acceleration A_n. Then we have

$$\begin{aligned}\begin{pmatrix} X_{n+1} \\ V_{n+1}\end{pmatrix} &= \begin{pmatrix} 1 & T \\ 0 & 1\end{pmatrix}\begin{pmatrix} X_n \\ V_n\end{pmatrix} + \begin{pmatrix} 0 \\ T\end{pmatrix} A_n \\ Z_n &= \begin{pmatrix} 1 & 0\end{pmatrix}\begin{pmatrix} X_n \\ V_n\end{pmatrix} + e_n\end{aligned}$$

For a more general case in 3D we have a state vector with 6 components (3 positions and 3 velocities). But, if we assume that the measurement noise in 3 dimensions are independent of one another and independent to the components of the acceleration, the problem can be treated as 3 independent one-dimensional moving target.

The Kalman equations for this simple model become

$$\begin{aligned}\begin{pmatrix} \widehat{X}_{n+1|n} \\ \widehat{V}_{n+1|n}\end{pmatrix} &= \begin{pmatrix} \widehat{X}_{n|n} + T\widehat{V}_{n|n} \\ \widehat{V}_{n|n}\end{pmatrix} \\ \begin{pmatrix} \widehat{X}_{n|n} \\ \widehat{V}_{n|n}\end{pmatrix} &= \begin{pmatrix} \widehat{X}_{n|n-1} \\ \widehat{V}_{n|n-1}\end{pmatrix} + \begin{pmatrix} K_{n,1} \\ K_{n,2}\end{pmatrix} (Z_n - \widehat{X}_{n|n-1}) \\ \begin{pmatrix} K_{n,1} \\ K_{n,2}\end{pmatrix} &= \begin{pmatrix} \frac{P(1,1)}{P(1,1)+r} \\ \frac{P(2,1)}{P(1,1)+r}\end{pmatrix}\end{aligned}$$

where $P(k,l)$ is the $(k-l)$th component of the matrix $P_{n|n-1}$.

To reduce the computation, the time varying elements of the Kalman gain vector can be replaced with some constants (the steady states values) to obtain

$$\begin{pmatrix} \widehat{X}_{n|n} \\ \widehat{V}_{n|n}\end{pmatrix} = \begin{pmatrix} \widehat{X}_{n|n-1} \\ \widehat{V}_{n|n-1}\end{pmatrix} + \begin{pmatrix} \alpha \\ \beta/T\end{pmatrix} (Z_n - \widehat{X}_{n|n-1}) \tag{7.15}$$

with constatnt values for α and β.

7.3.3 Track-While-Scan (TWS) Radar with dependent acceleration sequences

The simple model of the previous example is not very realistic, because there was assumed that the target is subjected to random acceleration. For a heavy target, we can do a little better by assuming that the acceleration A_n is modeled as

$$A_{n+1} = \rho A_n + W_n, \quad n = 0, 1, \ldots$$

a first order autoregressive (AR) model. The value of ρ can be choosed in accordance of the target. ρ near to 0 means a very low inertia target and ρ near to 1 means a high inertia target.

To account for this equation, we can extend the state vector and we obtain

$$\begin{aligned}\begin{pmatrix} X_{n+1} \\ V_{n+1} \\ A_{n+1} \end{pmatrix} &= \begin{pmatrix} 1 & T & 0 \\ 0 & 1 & T \\ 0 & 0 & \rho \end{pmatrix} \begin{pmatrix} X_n \\ V_n \\ A_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} W_n \\ Z_n &= \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} X_n \\ V_n \\ A_n \end{pmatrix} + e_n\end{aligned}$$

We again can apply the Kalman equations to this model and obtain:

$$\begin{aligned}\begin{pmatrix} \widehat{X}_{n+1|n} \\ \widehat{V}_{n+1|n} \\ \widehat{A}_{n+1|n} \end{pmatrix} &= \begin{pmatrix} \widehat{X}_{n|n} + T\widehat{V}_{n|n} \\ \widehat{V}_{n|n} + T\widehat{A}_{n|n} \\ \rho\widehat{A}_{n|n} \end{pmatrix} \\ \begin{pmatrix} \widehat{X}_{n|n} \\ \widehat{V}_{n|n} \\ \widehat{A}_{n|n} \end{pmatrix} &= \begin{pmatrix} \widehat{X}_{n|n-1} \\ \widehat{V}_{n|n-1} \\ \widehat{A}_{n|n-1} \end{pmatrix} + \begin{pmatrix} K_{n,1} \\ K_{n,2} \\ K_{n,3} \end{pmatrix} (\, Z_n - \widehat{X}_{n|n-1}\,) \\ \begin{pmatrix} K_{n,1} \\ K_{n,2} \\ K_{n,3} \end{pmatrix} &= \begin{pmatrix} \frac{P(1,1)}{P(1,1)+r} \\ \frac{P(2,1)}{P(1,1)+r}\, \frac{P(3,1)}{P(1,1)+r} \end{pmatrix}\end{aligned}$$

where $P(k,l)$ is the $(k-l)$th component of the matrix $P_{n|n-1}$.

Again to reduce the computation, the time varying elements of the Kalman gain vector can be replaced with some constants related to its steady state value and obtain

$$\begin{pmatrix} \widehat{X}_{n|n} \\ \widehat{V}_{n|n} \\ \widehat{A}_{n|n} \end{pmatrix} \begin{pmatrix} \widehat{X}_{n|n-1} \\ \widehat{V}_{n|n-1} \\ \widehat{A}_{n|n-1} \end{pmatrix} + \begin{pmatrix} \alpha \\ \beta/T \\ \gamma/T^2 \end{pmatrix} (\, Z_n - \widehat{X}_{n|n-1}\,) \tag{7.16}$$

with constant values for α, β and γ.

7.4 Fast Kalman filter equations

The general equations of the Kalman filtering do not assume any stationarity of the system and all the matrices of the system $\boldsymbol{F}$, $\boldsymbol{G}$ and $\boldsymbol{H}$, and also $\boldsymbol{R}$ and $\boldsymbol{Q}$ may depend on the index k.

Through the simple example in previous section, we saw that, when these quantities are independent of k, *i.e.*

$$\begin{aligned}
\widehat{\boldsymbol{x}}_{k+1|k} &= \boldsymbol{F}\widehat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{R}^e_k)^{-1}[\boldsymbol{z}_k - \boldsymbol{H}\widehat{\boldsymbol{x}}_{k|k-1}] \\
\boldsymbol{R}^e_k &= \boldsymbol{R} + \boldsymbol{H}\boldsymbol{P}_{k|k-1}\,\boldsymbol{H}^t \\
\boldsymbol{K}_k &= \boldsymbol{F}\boldsymbol{P}_{k|k-1}\,\boldsymbol{H}^t \\
\boldsymbol{P}_{k+1|k} &= \boldsymbol{F}\boldsymbol{P}_{k|k-1}\,\boldsymbol{F}^t + \boldsymbol{G}\boldsymbol{Q}\boldsymbol{G}^t - \boldsymbol{K}_k(\boldsymbol{R}^e_k)^{-1}\,\boldsymbol{K}^t_k
\end{aligned}$$

the system can reach a stationnary point and we can use the steady state values of the gain or an approximation to it to reduce the calculation cost. But doing so does not give an optimal estimate and may not give a very satisfactory solution. Here, we present shortly a slightly better way to obtain fast algorithms without loosing too much its optimality.

Let assume a constatnt system and note by $\delta\boldsymbol{P}_k$, $\delta\boldsymbol{K}^g_k$ and $\delta\boldsymbol{R}^e_k$ the increments

$$\begin{aligned}
\delta\boldsymbol{P}_k &= \boldsymbol{P}_{k|k-1} - \boldsymbol{P}_{k-1|k-2} \\
\delta\boldsymbol{K}^g_k &= \boldsymbol{K}^g_k - \boldsymbol{K}^g_{k-1} \\
\delta\boldsymbol{R}^e_k &= \boldsymbol{R}^e_k - \boldsymbol{R}^e_{k-1}
\end{aligned}$$

Then, it can be shown that the $\delta\boldsymbol{P}_{k+1}$ can be factorized by

$$\begin{aligned}
\delta\boldsymbol{P}_{k+1} &= [\boldsymbol{F} - \boldsymbol{K}^g_{k-1}\boldsymbol{H}][\delta\boldsymbol{P}_k - \delta\boldsymbol{P}_k\,\boldsymbol{H}^t(\boldsymbol{R}^e_k)^{-1}\,\boldsymbol{H}\boldsymbol{P}_k][\boldsymbol{F} - \boldsymbol{K}^g_{k-1}\boldsymbol{H}]^t \\
&= [\boldsymbol{F} - \boldsymbol{K}^g_k\boldsymbol{H}][\delta\boldsymbol{P}_k + \delta\boldsymbol{P}_k\,\boldsymbol{H}^t(\boldsymbol{R}^e_{k-1})^{-1}\,\boldsymbol{H}\boldsymbol{P}_k][\boldsymbol{F} - \boldsymbol{K}^g_k\boldsymbol{H}]^t
\end{aligned}$$

Then, it is possible to reduce the cost of the calculations by noting that if $\delta\boldsymbol{P}_k$ can be factorized as:

$$\delta\boldsymbol{P}_1 = \boldsymbol{z}_0\boldsymbol{M}_0\boldsymbol{z}^t_0,$$

then $\delta\boldsymbol{P}_{k+1}$ can also be factorized as

$$\delta\boldsymbol{P}_{k+1} = \boldsymbol{z}_k\,\boldsymbol{M}_k\,\boldsymbol{z}^t_k$$

and we obtain then the following equations:

$$\begin{aligned}
\delta\boldsymbol{P}_{k+1} &= \boldsymbol{z}_k\,\boldsymbol{M}_k\,\boldsymbol{z}^t_k \\
\boldsymbol{z}_k &= [\boldsymbol{F} - \boldsymbol{K}^g_k\boldsymbol{H}]\boldsymbol{z}_{k-1} \\
\boldsymbol{M}_k &= \boldsymbol{M}_{k-1} + \boldsymbol{M}_{k-1}\,\boldsymbol{z}^t_{k-1}\boldsymbol{H}^t(\boldsymbol{R}^e_{k-1})^{-1}\,\boldsymbol{H}\boldsymbol{z}_{k-1}\,\boldsymbol{M}_{k-1} \\
\boldsymbol{R}^e_{k-1} &= \boldsymbol{R}^e_k + \boldsymbol{H}\boldsymbol{z}_k\,\boldsymbol{M}_k\,\boldsymbol{z}^t_k\boldsymbol{H}^t \\
\boldsymbol{K}_{k+1} &= \boldsymbol{K}^g_{k+1}\,\boldsymbol{R}^e_{k+1} = \boldsymbol{K}_k + \boldsymbol{F}\boldsymbol{z}_k\,\boldsymbol{M}_k\,\boldsymbol{z}^t_k\boldsymbol{H}^t \\
\boldsymbol{P}_{k|k-1} &= \boldsymbol{P}_0 + \sum_{j=0}^{k-1}\boldsymbol{z}_j\boldsymbol{M}_j\boldsymbol{z}^t_j
\end{aligned}$$

which are called *Chandrasekhar* equations.

Note that if $\alpha = \text{rang}\,\{\delta \boldsymbol{P}_1\}$ where

$$\delta \boldsymbol{P}_1 = \boldsymbol{F}\boldsymbol{P}_0\boldsymbol{F}^t + \boldsymbol{G}\boldsymbol{Q}\boldsymbol{G}^t - \boldsymbol{K}_0(\boldsymbol{R}_0^e)^{-1}\boldsymbol{K}_0^t - \boldsymbol{P}_0$$

then $\boldsymbol{z}_k$ has dimensions (N, α), $\boldsymbol{M}_k$ has dimensions (α, α). So, in place of updating, at each time k the matrix $\boldsymbol{P}_k$ with dimensions (N, N) we only have to update the matrixes $\boldsymbol{z}_k$ and $\boldsymbol{M}_k$ with dimensions (N, α) and (α, α).

Note also that $\boldsymbol{M}_0$ is the signature matrix of $\delta \boldsymbol{P}_1$ and the value of α depends on the choice of the initial covariance matrix $\boldsymbol{P}_0$. It is not unusual to have $\alpha = 1$ which greately reduces the computation cost.

7.5 Kalman filter equations for signal deconvolution

Starting by the convolution equation:

$$z(k) = \sum_{i=0}^{p-1} h(i)x(k-i) + b(k)$$

rewritten in matrix form

$$\begin{pmatrix} z(1) \\ \vdots \\ z(k) \\ \vdots \\ z(M) \end{pmatrix} = \begin{pmatrix} h_{(p-1)} & \cdots & h_{(0)} & \cdots & \cdots \\ \vdots & & & & \vdots \\ 0 & h_{(p-1)} & \cdots & h_{(0)} & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & h_{(p-1)} & \cdots & h_{(0)} \end{pmatrix} \begin{pmatrix} x(-p) \\ \vdots \\ x(0) \\ \vdots \\ x(M) \end{pmatrix} + \begin{pmatrix} b(1) \\ \vdots \\ b(k) \\ \vdots \\ z(M) \end{pmatrix}$$

we can propose the following models:

- **Constant state vector model**

$$\begin{cases} \boldsymbol{x}_{k+1} & = & \boldsymbol{x}_k = \boldsymbol{x} = [x_{-p}, \ldots, x_{-1}, x_0, x_1, \ldots, x_n]^t \\ z_k & = & \boldsymbol{h}_k^t \cdot \boldsymbol{x}_k + v_k \end{cases}$$

$$\boldsymbol{h}_k = (0 \quad 0 \quad 0 \quad h_{p-1} \quad \ldots \quad h_0 \quad 0 \quad 0)^t$$

where coefficient h_0 is in the k-th position. Then we have

$$\begin{cases} \boldsymbol{u}_k & = & \boldsymbol{0} \\ \boldsymbol{F}_k & = & \boldsymbol{G}_k = \boldsymbol{I} \quad \text{with } \boldsymbol{D} = \begin{pmatrix} 0 & \ldots & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \vdots & \vdots \\ \vdots & & & \vdots & \vdots \\ 0 & & \ldots & 1 & 0 \end{pmatrix} \\ \boldsymbol{h}_{k+1}^t & = & \boldsymbol{D}\boldsymbol{h}_k^t \end{cases}$$

If we note by

$$\mathrm{E}\left[\boldsymbol{x}\right] = \boldsymbol{x}_0 \qquad \mathrm{E}\left[[\boldsymbol{x} - \boldsymbol{x}_0][\boldsymbol{x} - \boldsymbol{x}_0]^t\right] = \boldsymbol{P}_0$$

$$\mathrm{E}\left[v_k\right] = 0 \qquad \mathrm{E}\left[v_k v_j\right] = r\,\delta_{kj}$$

we obtain

$$\begin{aligned} \widehat{\boldsymbol{x}}_{k|k} &= \widehat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k (r_k^e)^{-1} [y_k - \boldsymbol{h}_k^t \cdot \widehat{\boldsymbol{x}}_{k|k-1}] \\ r_k^e &= r + \boldsymbol{h}_k^t \boldsymbol{P}_{k|k-1}\, \boldsymbol{h}_k \\ \boldsymbol{K}_k &= \boldsymbol{P}_{k|k-1}\, \boldsymbol{h}_k^t \\ \boldsymbol{P}_{k+1|k} &= \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k (r_k^e)^{-1}\, \boldsymbol{K}_k^t \end{aligned}$$

Note that the observations y_k are scalar. So, r_k^e is also a scalar quantity, but $\boldsymbol{x}$ is a N-dimensional vector and so the covariance matrix $\boldsymbol{P}$ has the dimensions $(N \times N)$.

- **Non constant state space model**

  we can choose

$$\boldsymbol{h} = [h_0, \ldots, h_{p-1}]^t, \quad \boldsymbol{x}_k = [x_k, x_{k-1}, \ldots, x_{k-p+1}]^t$$

$$z(k) = \boldsymbol{h}^t \boldsymbol{x}_k + b(k), \quad \text{dimension of } \boldsymbol{x}_k = p \leq N$$

But now we need to introduce a generating state space model for $\boldsymbol{x}_k$.

One of such models is an AR model :

$$x(n+1) = \sum_{i=1}^{q} a(i)\, x(n-i+1) + u(n+1)$$

where

$$\mathrm{E}\left[u_n\right] = 0, \quad \mathrm{E}\left[|u_n|^2\right] = \beta^2, \quad \mathrm{E}\left[u_m u_n\right] = 0, \quad m \neq n$$

It is easy then to see that we can write

$$\begin{pmatrix} x(n+1) \\ x(n) \\ \vdots \\ \vdots \\ x(n-q+2) \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & \ldots & \ldots & a_q \\ 1 & 0 & \ldots & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & 1 & 0 \end{pmatrix} \begin{pmatrix} x(n) \\ x(n-1) \\ \vdots \\ \vdots \\ x(n-q+1) \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix} u(n+1)$$

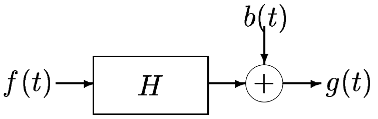
Thus we have

$$\begin{cases} \boldsymbol{x}_{k+1} & = & \boldsymbol{F}\boldsymbol{x}_k + \boldsymbol{G}\boldsymbol{u}_{k+1} \\ y_k & = & \boldsymbol{h}^t \cdot \boldsymbol{x}_k + b_k \end{cases}$$

$$\boldsymbol{F} = \begin{pmatrix} a_1 & a_2 & \ldots & \ldots & a_q \\ 1 & 0 & \ldots & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & 1 & 0 \end{pmatrix}, \quad \boldsymbol{G} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

This model has the advantage that $\boldsymbol{F}$, $\boldsymbol{G}$ and $\boldsymbol{H}$ are constant and we can use fast algorithms, but the main drawback in practical applications is the determination of q and a_k, $k = 1, \ldots, q$.

Consider the following convolution model :



$$g(t) = \int f(t')h(t-t')\,\mathrm{d}t' + b(t) = \int h(t')f(t-t')\,\mathrm{d}t' + b(t)$$

and assume the following hypotheses :

- The signals $f(t)$, $g(t)$, $h(t)$ are discretized with the same sampling period $\Delta T = 1$,
- The impulse response is finite (FIR) : $h(t) = 0$, for t such that $t < -q\Delta T$ or $\forall t > p\Delta T$.

Then we have :

$$g(m) = \sum_{k=-q}^{p} h(k)f(m-k) + b(m), \quad m = 0, \cdots, M$$

or in a matrix form

$$\begin{pmatrix} g(0) \\ g(1) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ g(M) \end{pmatrix} = \begin{pmatrix} h(p) & \cdots & h(0) & \cdots & h(-q) & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & & \ddots & & \ddots & & & & \vdots \\ \vdots & & & & & & & & & \vdots \\ \vdots & & h(p) & \cdots & h(0) & \cdots & h(-q) & & & \vdots \\ \vdots & & & & & & & & & \vdots \\ \vdots & & & \ddots & & & \ddots & & & \vdots \\ \vdots & & & & \ddots & & & & & 0 \\ 0 & \cdots & \cdots & & 0 & h(p) & \cdots & h(0) & \cdots & h(-q) \end{pmatrix} \begin{pmatrix} f(-p) \\ \vdots \\ f(0) \\ f(1) \\ \vdots \\ f(M) \\ f(M+1) \\ \vdots \\ f(M+q) \end{pmatrix}$$

or

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{v}$$

Note that $\boldsymbol{g}$ is a $(M+1)$-dimensional vector, $\boldsymbol{f}$ has dimension $M+p+q+1$, $\boldsymbol{h} = [h(p), \cdots, h(0), \cdots, h(-q)]$ has dimension $(p+q+1)$ and matrix $\boldsymbol{H}$ has dimensions $(M+1) \times (M+p+q+1)$.

Now, if we assume that the system is causal $(q = 0)$ we obtain

$$\begin{pmatrix} g(0) \\ g(1) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ g(M) \end{pmatrix} = \begin{pmatrix} h(p) & \cdots & h(0) & 0 & \cdots & \cdots & 0 \\ 0 & & & & & & \vdots \\ \vdots & & & & & & \vdots \\ \vdots & & h(p) & \cdots & h(0) & & \vdots \\ \vdots & & & & & & \vdots \\ \vdots & & & & & & 0 \\ 0 & \cdots & \cdots & 0 & h(p) & \cdots & h(0) \end{pmatrix} \begin{pmatrix} f(-p) \\ \vdots \\ f(0) \\ f(1) \\ \vdots \\ \vdots \\ f(M) \end{pmatrix}$$

If the input signal is also assumed to be causal, we obtain :

$$\begin{pmatrix} g(0) \\ g(1) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ g(M) \end{pmatrix} = \begin{pmatrix} h(0) & & & & & \\ h(1) & \ddots & & & & \\ \vdots & & & & & \\ h(p) & \cdots & & h(0) & & \\ 0 & \ddots & & & \ddots & \\ \vdots & & & & & \\ 0 & \cdots & 0 & h(p) & \cdots & h(0) \end{pmatrix} \begin{pmatrix} f(0) \\ f(1) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ f(M) \end{pmatrix} \tag{7.17}$$

and finally if $p = M$ we have :

$$\begin{pmatrix} g(0) \\ g(1) \\ \vdots \\ g(M) \end{pmatrix} = \begin{pmatrix} h(0) & & & \\ h(1) & h(0) & & \\ \vdots & & \ddots & \\ h(M) & \cdots & h(1) & h(0) \end{pmatrix} \begin{pmatrix} f(0) \\ f(1) \\ \vdots \\ f(M) \end{pmatrix}$$

Remark that, in all cases matrix $\boldsymbol{H}$ is TOEPLITZ.

In the case where the input signal and the system are both causal, (7.17) can be rewritten as

$$\begin{pmatrix} g(0) \\ g(1) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ g(M) \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} h(0) & 0 & \cdots & & & 0 & h(p) & \cdots & h(1) \\ h(1) & \ddots & & & & & & \ddots & \vdots \\ \vdots & & & & & & & & h(p) \\ h(p) & \cdots & & h(0) & 0 & & & & 0 \\ 0 & \ddots & & & \ddots & & & & \vdots \\ \vdots & & & & & & & & \\ 0 & \cdots & 0 & h(p) & \cdots & & h(0) & & 0 \\ \vdots & & & & & & & & \vdots \\ & & & & \ddots & & & \ddots & 0 \\ 0 & \cdots & & \cdots & 0 & h(p) & \cdots & & h(0) \end{pmatrix} \begin{pmatrix} f(0) \\ f(1) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ f(M) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

where $\boldsymbol{f}$ and $\boldsymbol{g}$ have been completed artificially by some zeros. This operation is called zero-filling and the main advantage to do so is that the matrix $\boldsymbol{H}$ is now a circulant matrix.

Starting by the Kalman filter equations:

$$\begin{cases} \boldsymbol{z}_k & = \boldsymbol{H}_k \, \boldsymbol{x}_k + \boldsymbol{v}_k \quad \text{observation equation} \\ \boldsymbol{x}_{k+1} & = \boldsymbol{F}_k \, \boldsymbol{x}_k + \boldsymbol{G}_k \, \boldsymbol{u}_k \quad \text{state equation} \end{cases}$$

$$\begin{aligned} \widehat{\boldsymbol{x}}_{k+1|k} &= \boldsymbol{F}_k \, \widehat{\boldsymbol{x}}_{k|k} \\ \boldsymbol{P}_{k+1|k} &= \boldsymbol{F}_k \, \boldsymbol{P}_{k|k} \, \boldsymbol{F}_k^t + \boldsymbol{G}_k \, \boldsymbol{Q}_k \, \boldsymbol{G}_k^t \\ \widehat{\boldsymbol{x}}_{k+1|k+1} &= \widehat{\boldsymbol{x}}_{k+1|k} + \boldsymbol{K}_{k+1}^f [\boldsymbol{z}_{k+1} - \boldsymbol{H}_{k+1} \, \widehat{\boldsymbol{x}}_{k+1|k}] \\ \boldsymbol{K}_{k+1}^f &= \boldsymbol{P}_{k+1|k} \, \boldsymbol{H}_{k+1}^t (\boldsymbol{R}_{k+1}^e)^{-1} \\ \boldsymbol{R}_{k+1}^e &= \boldsymbol{R}_{k+1} + \boldsymbol{H}_{k+1} \, \boldsymbol{P}_{k+1|k} \, \boldsymbol{H}_{k+1}^t \\ \boldsymbol{P}_{k+1|k+1} &= [\boldsymbol{I} - \boldsymbol{K}_{k+1}^f \boldsymbol{H}_{k+1}] \boldsymbol{P}_{k+1|k} \end{aligned}$$

7.5.1 AR, MA and ARMA Models

AR model

$$u(n) = \sum_{k=1}^{p} a(k)\, u(n-k) + \epsilon(n), \quad \forall n$$

$$\mathrm{E}\left[\epsilon(n)\right] = 0, \quad \mathrm{E}\left[|\epsilon(n)|^2\right] = \beta^2,$$

$$\mathrm{E}\left[\epsilon(n)\, u(m)\right] = 0, \quad m \neq n$$

$$\epsilon(n) \longrightarrow \boxed{H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{p} a(k) z^{-k}}} \longrightarrow u(n)$$

MA model

$$u(n) = \sum_{k=0}^{q} b(k)\, \epsilon(n-k), \quad \forall n$$

$$\epsilon(n) \longrightarrow \boxed{B(z) = \sum_{k=0}^{q} b(k) z^{-k}} \longrightarrow u(n)$$

ARMA model

$$u(n) = \sum_{k=1}^{p} a(k)\, u(n-k) + \sum_{l=0}^{q} b(l)\, \epsilon(n-l)$$

$$\epsilon(n) \longrightarrow \boxed{H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{k=0}^{q} b(k) z^{-k}}{1 + \sum_{k=1}^{p} a(k) z^{-k}}} \longrightarrow u(n)$$

$$\epsilon(n) \longrightarrow \boxed{H(z) = B_q(z)} \longrightarrow \boxed{H(z) = \frac{1}{A_p(z)}} \longrightarrow u(n)$$

In a dynamic system, in general, we are interested in a physical quantity $\boldsymbol{x}$ through the observation of a quantity $\boldsymbol{z}$ related to $\boldsymbol{x}$ by the following system of equations

$$\left\{ \begin{array}{ll} \boldsymbol{x}_{n+1} & = \boldsymbol{f}_n(\boldsymbol{x}_n, \boldsymbol{u}_n) \\ \boldsymbol{z}_n & = \boldsymbol{h}_n(\boldsymbol{x}_n, \boldsymbol{v}_n) \end{array} \right. \tag{7.18}$$

Chapter 8

Some complements to Bayesian estimation

8.1 Choice of a prior law in the Bayesian estimation

One of the main difficulties in the application of Bayesian theory in practice is the choice or the attribution of the direct probabilities $f(x|\theta)$ and $\pi(\theta)$. In general, $f(x|\theta)$ is obtained via an appropriate model relating the observable quantity X to the parameters θ and is well accepted. The choice or the attribution of the prior $\pi(\theta)$ has been, and still is, the main subject of discussion and controversy between the Bayesian and orthodox statisticians.

Here, I will try to give a brief summary of different approaches and different tools that can be used to attribute a prior probability distribution. There are mainly four tools:

- use of some invariance principles
- use of maximum entropy (ME) principle
- use of conjugate and reference priors
- use of other information criteria

8.1.1 Invariance principles

Définition 1 [Group invariance] A probability model $f(x|\theta)$ is said to be invariant (or closed) under the action of a group of transformations $\mathcal{G}$ if, for every $g \in \mathcal{G}$, there exists a unique $\theta^* = \bar{g}(\theta) \in \mathcal{T}$ such that $y = g(x)$ is distributed according to $f(y|\theta^*)$.

Exemple 1 Any probability density function in the form $f(x|\theta) = f(x - \theta)$ is invariant under the *translation* group

$$\mathcal{G} : \{g_c(x) : g_c(x) = x + c, \quad c \in \mathbb{R}\} \tag{8.1}$$

This can be verified as follows

$$x \sim f(x - \theta) \longrightarrow y = x + c \sim f(y - \theta^*) \quad \text{with} \quad \theta^* = \theta + c$$

Exemple 2 Any probability density function in the form $f(x|\theta) = \frac{1}{\theta} f(\frac{x}{\theta})$ is invariant under the *multiplicative* or *scale* transformation group

$$\mathcal{G} : \{g_s(x) : g_s(x) = s\,x, \quad s > 0\} \tag{8.2}$$

This can be verified as follows

$$x \sim \frac{1}{\theta} f(\frac{x}{\theta}) \longrightarrow y = s\,x \sim \frac{1}{\theta^*} f(\frac{y}{\theta^*}) \quad \text{with} \quad \theta^* = s\,\theta$$

Exemple 3 Any probability density function in the form $f(x|\theta_1, \theta_2) = \frac{1}{\theta_2} f(\frac{x-\theta_1}{\theta_2})$ is invariant under the *affine* transformation group

$$\mathcal{G} : \{g_{a,b}(x) : g_{a,b}(x) = a\,x + b, \quad a > 0, b \in \mathbb{R}\} \tag{8.3}$$

This can be verified as follows

$$x \sim \frac{1}{\theta_2} f(\frac{x-\theta_1}{\theta_2}) \longrightarrow y = a\,x + b \sim \frac{1}{\theta_2^*} f(\frac{y-\theta_1^*}{\theta_2^*}) \quad \text{with} \quad \theta_2^* = a\,\theta_2, \quad \theta_1^* = a\,\theta_1 + b.$$

Exemple 4 Any multi variable probability density function in the form $f(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x}-\boldsymbol{\theta})$ is invariant under the translation group

$$\mathcal{G} : \{g_{\boldsymbol{c}}(\boldsymbol{x}) : g_{\boldsymbol{c}}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{c}, \quad \boldsymbol{c} \in \mathbb{R}^n\} \tag{8.4}$$

Exemple 5 Any multi variable probability density function in the form $f(\boldsymbol{x}) = f(\|\boldsymbol{x}\|)$ is invariant under the orthogonal transformation group

$$\mathcal{G} : \left\{g_{\boldsymbol{A}}(\boldsymbol{x}) : g_{\boldsymbol{A}}(\boldsymbol{x}) = \boldsymbol{A}\,\boldsymbol{x}, \quad \boldsymbol{A}^t\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^t = \boldsymbol{I}\right\} \tag{8.5}$$

Exemple 6 Any multi variable probability density function in the form $f(\boldsymbol{x}|\theta) = \frac{1}{\theta} f(\frac{\|\boldsymbol{x}\|}{\theta})$ is invariant under the following transformation group

$$\mathcal{G} : \left\{g_{\boldsymbol{A},s}(\boldsymbol{x}) : g_{\boldsymbol{A},s}(\boldsymbol{x}) = s\,\boldsymbol{A}\,\boldsymbol{x}, \quad \boldsymbol{A}^t\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^t = \boldsymbol{I}, \quad s > 0.\right\} \tag{8.6}$$

This can be verified as follows

$$\boldsymbol{x} \sim \frac{1}{\theta} f(\frac{\|\boldsymbol{x}\|}{\theta}) \longrightarrow \boldsymbol{y} = s\,\boldsymbol{A}\,\boldsymbol{x} \sim \frac{1}{\theta^*} f(\frac{\|\boldsymbol{y}\|}{\theta^*}) \quad \text{with} \quad \theta^* = s\,\theta.$$

From these examples we see also that any invariance transformation group $\mathcal{G}$ on $x \in \mathcal{X}$ induces a corresponding transformation group $\bar{\mathcal{G}}$ on $\theta \in \mathcal{T}$. For example for the translation invariance $\mathcal{G}$ on $x \in \mathcal{X}$ induces the following translation group on $\theta \in \mathcal{T}$

$$\bar{\mathcal{G}} : \{\bar{g}_c(\theta) : \bar{g}_c(\theta) = \theta + c, \quad c \in \mathbb{R}\} \tag{8.7}$$

and the scale invariance $\mathcal{G}$ on $x \in \mathcal{X}$ induces the following translation groupe on $\theta \in \mathcal{T}$

$$\bar{\mathcal{G}} : \{\bar{g}_s(\theta) : \bar{g}_s(\theta) = s\,\theta, \quad s > 0\} \tag{8.8}$$

We just see that for an invariant family of $f(x|\theta)$ we have a corresponding invariant family of prior laws $\pi(\theta)$. To be complete, we have also to consider the cost function to be able to define the Bayesian estimate.

Définition 2 [Invariant cost functions] Assume a probability model $f(x|\theta)$ is invariant under the action of the group of transformations $\mathcal{G}$. Then the cost function $c[\theta,\widehat{\theta}]$ is said to be invariant under the group of transformations $\tilde{\mathcal{G}}$ if, for every $g \in \mathcal{G}$ and $\widehat{\theta} \in \mathcal{T}$, there exists a unique $\widehat{\theta}^* = \tilde{g}(\widehat{\theta}) \in \mathcal{T}$ with $\tilde{g} \in \tilde{\mathcal{G}}$ such that

$$c[\theta,\widehat{\theta}] = c[\bar{g}(\theta),\widehat{\theta}^*] \qquad \text{for every } \theta \in \mathcal{T}.$$

Définition 3 [Invariant estimate] For an invariant probability model $f(x|\theta)$ under the group of transformation Gc and an invariant cost function $c[\theta,\widehat{\theta}]$ under the corresponding group of transformation $\bar{\mathcal{G}}$, an estimate $\widehat{\theta}$ is said to be invariant or *equivariant* if

$$\widehat{\theta}(g(x)) = \tilde{g}\left(\widehat{\theta}(x)\right)$$

Exemple 7 Estimation of θ from the data coming from any model of the kind $f(x|\theta) = f(x-\theta)$ with a quadratic cost function $c[\theta,\widehat{\theta}] = (\theta-\widehat{\theta})^2$ is equivariant and we have

$$\mathcal{G} = \bar{\mathcal{G}} = \tilde{\mathcal{G}} = \{g_c(x) : g_c(x) = x - c, \quad c \in \mathbb{R}\}$$

Exemple 8 Estimation of θ from the data coming from any model of the kind $f(x|\theta) = \frac{1}{\theta}f(\frac{1}{\theta})$ with the entropy cost function

$$c[\theta,\widehat{\theta}] = \frac{\theta}{\widehat{\theta}} - \ln(\frac{\theta}{\widehat{\theta}}) - 1$$

is equivariant and we have

$$\begin{aligned} \mathcal{G} &= \{g_s(x) : g_s(x) = s\,x, \quad s > 0\} \\ \bar{\mathcal{G}} = \tilde{\mathcal{G}} &= \{g_s(\theta) : g_s(\theta) = s\,\theta, \quad s > 0\} \end{aligned}$$

Proposition 1 [Invariant Bayesian estimate] Suppose that a probability model $f(x|\theta)$ is invariant under the group of transformations $\mathcal{G}$ and that there exists a probability distribution $\pi^*(\theta)$ on $\mathcal{T}$ which is invariant under the group of transformations $\bar{\mathcal{G}}$, *i.e.*,

$$\pi^*(\bar{g}(A)) = \pi^*(A)$$

for any measurable set $A \in \mathcal{T}$. Then the Bayes estimator associated with π^*, noted $\widehat{\theta}^*$ minimizes

$$\int R\left(\theta,\widehat{\theta}\right)\pi^*(\theta)\,\mathrm{d}\theta = \int R\left(\theta,\bar{g}(\widehat{\theta})\right)\pi^*(\theta)\,\mathrm{d}\theta = \int \mathrm{E}\left[c\left[\theta,\bar{g}\left(\widehat{\theta}^{(}X)\right)\right]\right]\pi^*(\theta)\,\mathrm{d}\theta \quad \text{over } \widehat{\theta}.$$

If this Bayes estimator is unique, it satisfies

$$\widehat{\theta}^*(x) = \tilde{g}^{-1}\left(\widehat{\theta}^*(g(x))\right)$$

Therefore, a Bayes estimator associated with an invariant prior and a strictly convex invariant cost function is almost equivariant.

Actually, invariant probability distributions are rare. The following are some examples:

Exemple 9 If $\pi(\theta)$ is invariant under the translation group $\mathcal{G}_c$, it satisfies $\pi(\theta) = \pi(\theta + c)$ for every θ and for every c, which implies that $\pi(\theta) = \pi(0)$ uniformly on $\mathbb{R}$ and this leads to the Lebesgue measure as an invariant measure.

Exemple 10 If $\theta > 0$ and $\pi(\theta)$ is invariant under the scale group $\mathcal{G}_s$, it satisfies $\pi(\theta) = s\,\pi(s\theta)$ for every $\theta > 0$ and for every $s > 0$, which implies that $\pi(\theta) = 1/\theta$.

Note that in both cases the invariant laws are improper.

8.2 Conjugate priors

The conjugate prior concept is tightly related to the sufficient statistic and exponential families.

Définition 4 [Sufficient statistics] When $X \sim P_\theta(x)$, a function $h(X)$ is said to be a sufficient statistic for $\{P_\theta(x), \theta \in \mathcal{T}\}$ if the distribution of X conditioned on $h(X)$ does not depend on θ for $\theta \in \mathcal{T}$.

Définition 5 [Minimal sufficiency] A function $h(X)$ is said to be minimal sufficient for $\{P_\theta(x), \theta \in \mathcal{T}\}$ if it is a function of every other sufficient statistic for $P_\theta(x)$.

A minimal sufficient statistic contains the whole information brought by the observation $X = x$ about θ.

Proposition 2 [Factorization theorem] Suppose that $\{P_\theta(x), \theta \in \mathcal{T}\}$ has a corresponding family of densities $\{p_\theta(x), \theta \in \mathcal{T}\}$. A statistic T is sufficient for θ if and only if there exist functions g_θ and h such that

$$p_\theta(x) = g_\theta(T(x))\, h(x) \tag{8.9}$$

for all $x \in \Gamma$ and $\theta \in \mathcal{T}$.

Exemple 11 If $X \sim \mathcal{N}(\theta, 1)$ then $T(x) = x$ can be chosen as a sufficient statistic.

Exemple 12 If $\{X_1, X_2, \ldots, X_n\}$ are i.i.d. and $X_i \sim \mathcal{N}(\theta, 1)$ then

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= (2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right] \\ &= \exp\left[-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right] (2\pi)^{-n/2} \exp\left[-\frac{n}{2}\theta^2\right] \exp\left[\theta \sum_{i=1}^{n} x_i\right] \end{aligned}$$

and we have $T(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$.

Note that, in this case, we need to know n and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Note also that we can write

$$f(\boldsymbol{x}|\theta) = a(\boldsymbol{x})\, g(\theta) \exp\left[\theta T(\boldsymbol{x})\right]$$

where

$$g(\theta) = (2\pi)^{-n/2} \exp\left[-\frac{n}{2}\theta^2\right] \quad \text{and} \quad a(\boldsymbol{x}) = \exp\left[-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right]$$

Exemple 13 If $X \sim \mathcal{N}(0, \theta)$ then $T(x) = x^2$ can be chosen as a sufficient statistic.

Exemple 14 If $X \sim \mathcal{N}(\theta_1, \theta_2)$ then $T_1(x) = x^2$ and $T_2(x) = x$ can be chosen as a set of sufficient statistics.

Exemple 15 If $\{X_1, X_2, \ldots, X_n\}$ are i.i.d. and $X_i \sim \mathcal{N}(\theta_1, \theta_2)$ then

$$\begin{aligned} f(\boldsymbol{x}|\theta_1,\theta_2) &= (2\pi)^{-n/2}\,\theta_2^{-1/2}\,\exp\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right] \\ &= (2\pi)^{-n/2}\,\theta_2^{-1/2}\,\exp\left[-\frac{n\theta_1^2}{2\theta_2}\right]\,\exp\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}x_i^2+\frac{\theta_1}{\theta_2}\sum_{i=1}^{n}x_i\right] \end{aligned}$$

and we have $T_1(\boldsymbol{x}) = \sum_{i=1}^n x_i$ and $T_2(\boldsymbol{x}) = \sum_{i=1}^n x_i^2$.

Note also that we can write

$$f(\boldsymbol{x}|\theta) = a(\boldsymbol{x})\,g(\theta_1,\theta_2)\exp\left[\frac{\theta_1}{\theta_2}T_1(\boldsymbol{x}) - \frac{1}{2\theta_2}T_2(\boldsymbol{x})\right]$$

where

$$g(\theta_1,\theta_2) = (2\pi)^{-n/2}\theta_2^{-1/2}\exp\left[-\frac{n\theta_1^2}{2\theta_2}\right] \quad \text{and} \quad a(\boldsymbol{x}) = 1.$$

In this case, $\frac{\theta_1}{\theta_2}$ and $\frac{-1}{2\theta_2}$ are called canonical parametrization. It is also usual to use n, $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2$ as the sufficient statistics.

Exemple 16 If $X \sim \mathbf{Gam}(\alpha,\theta)$ then $T(x) = x$ can be chosen as a sufficient statistic.

Exemple 17 If $X \sim \mathbf{Gam}(\theta,\beta)$ then $T(x) = \ln x$ can be chosen as a sufficient statistic.

Exemple 18 If $X \sim \mathbf{Gam}(\theta_1,\theta_2)$ then $T_1(x) = \ln x$ and $T_2(x) = x$ can be chosen as a set of sufficient statistics.

Exemple 19 If $\{X_1, X_2, \ldots, X_n\}$ are i.i.d. and $X_i \sim \mathbf{Gam}(\theta_1,\theta_2)$ then it is easy to show that $T_1(\boldsymbol{x}) = \sum_{i=1}^n \ln x_i$ and $T_2(\boldsymbol{x}) = \sum_{i=1}^n x_i$.

Définition 6 [Exponential family] A class of distributions $\{P_\theta(\boldsymbol{x}), \theta \in \mathcal{T}\}$ is said to be an exponential family if there exist: $a(\boldsymbol{x})$ a function of Γ on $\mathbb{R}$, $g(\boldsymbol{\theta})$ a function of $\mathcal{T}$ on $\mathbb{R}^+$, $\phi_k(\boldsymbol{\theta})$ functions of $\mathcal{T}$ on $\mathbb{R}$, and $h_k(\boldsymbol{x})$ functions of Γ on $\mathbb{R}$ such that

$$\begin{aligned} p_\theta(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta}) &= a(\boldsymbol{x})\,g(\boldsymbol{\theta})\exp\left[\sum_{k=1}^{K}\phi_k(\boldsymbol{\theta})\,h_k(\boldsymbol{x})\right] \\ &= a(\boldsymbol{x})\,g(\boldsymbol{\theta})\exp\left[\boldsymbol{\phi}^t(\boldsymbol{\theta})\boldsymbol{h}(\boldsymbol{x})\right] \end{aligned}$$

for all $\theta \in \mathcal{T}$ and $x \in \Gamma$. This family is entirely determined by $a(\boldsymbol{x})$, $g(\boldsymbol{\theta})$, and $\{\phi_k(\boldsymbol{\theta}), h_k(\boldsymbol{x}),\, k = 1, \cdots, K\}$ and is noted $\mathbf{Exfn}(\boldsymbol{x}|a, g, \boldsymbol{\phi}, \boldsymbol{h})$

Particular cases:

- When $a(\boldsymbol{x}) = 1$ and $g(\boldsymbol{\theta}) = \exp[-b(\boldsymbol{\theta})]$ we have

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left[\boldsymbol{\phi}^t(\boldsymbol{\theta})\boldsymbol{h}(\boldsymbol{x}) - b(\boldsymbol{\theta})\right]$$

  and is noted $\mathbf{CExf}(\boldsymbol{x}|b, \boldsymbol{\phi}, \boldsymbol{h})$.

- Natural exponential family:
  When $a(\boldsymbol{x}) = 1$, $g(\boldsymbol{\theta}) = \exp[-b(\boldsymbol{\theta})]$, $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$ and $\boldsymbol{\phi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ we have

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \exp\left[\boldsymbol{\theta}^t\boldsymbol{x} - b(\boldsymbol{\theta})\right] \mathbf{Exf}(\boldsymbol{x}|b).$$

  and is noted $\mathbf{NExf}(\boldsymbol{x}|b)$.

- Scalar random variable with a vector parameter:

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= \mathbf{Exf}(x|a, g, \boldsymbol{\phi}, \boldsymbol{h}) \\ &= a(x)g(\boldsymbol{\theta}) \exp\left[\sum_{k=1}^{K} \phi_k(\boldsymbol{\theta})h_k(x)\right] \\ &= a(x)g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\phi}^t(\boldsymbol{\theta})\boldsymbol{h}(x)\right] \end{aligned}$$

  and is noted $\mathbf{Exfk}(\boldsymbol{x}|a, g, \boldsymbol{\phi}, \boldsymbol{h})$.

- Scalar random variable with a scalar parameter:

$$p(x|\theta) = \mathbf{Exf}(x|a, g, \phi, h) = a(x)g(\theta) \exp\left[\phi(\theta)h(x)\right]$$

  and is noted $\mathbf{Exf}(\boldsymbol{x}|a, g, \phi, h)$.

- Simple scalar exponential family:

$$p(x|\theta) = \theta \exp\left[-\theta x\right] = \exp\left[-\theta x + \ln\theta\right], \quad x \geq 0, \quad \theta \geq 0.$$

Définition 7 [Conjugate distributions] A family $\mathcal{F}$ of probability distributions $\pi(\boldsymbol{\theta})$ on $\mathcal{T}$ is said to be conjugate (or closed under sampling) if, for every $\pi(\boldsymbol{\theta}) \in \mathcal{F}$, the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ also belongs to $\mathcal{F}$.

The main argument for the development of the conjugate priors is the following: When the observation of a variable X with a probability law $f(x|\theta)$ modifies the prior $\pi(\theta)$ to a posterior $\pi(\theta|x)$, the information conveyed by x about θ is obviously limited, therefore it should not lead to a modification of the whole structure of $\pi(\theta)$, but only of its parameters.

Définition 8 [Conjugate priors] Assume that $f(\boldsymbol{x}|\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\boldsymbol{x}) = l(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{x}))$ where $\boldsymbol{t} = \{n, \boldsymbol{s}\} = \{n, s_1, \ldots, s_k\}$ is a vector of dimension $k+1$ and is sufficient statistic for $f(\boldsymbol{x}|\boldsymbol{\theta})$. Then, if there exists a vector $\{\tau_0, \boldsymbol{\tau}\} = \{\tau_0, \tau_1, \ldots, \tau_k\}$ such that

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}) = \frac{f(\boldsymbol{s} = (\tau_1, \cdots, \tau_k)|\boldsymbol{\theta}, n = \tau_0)}{\displaystyle\int f(\boldsymbol{s} = (\tau_1, \cdots, \tau_k)|\boldsymbol{\theta}', n = \tau_0)\, \mathrm{d}\boldsymbol{\theta}'}$$

exists and defines a family $\mathcal{F}$ of distributions for $\boldsymbol{\theta} \in \mathcal{T}$, then the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\tau})$ will remain in the same family $\mathcal{F}$. The prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\tau})$ is then a conjugate prior for the sampling distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$.

Proposition 3 [Sufficient statistics for the exponential family] For a set of n i.i.d. samples $\{x_1, \cdots, x_n\}$ of a random variable $X \sim \mathbf{Exf}(x|a, g, \boldsymbol{\theta}, \boldsymbol{h})$ we have

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{j=1}^{n} f(x_j|\boldsymbol{\theta}) &= [g(\boldsymbol{\theta})]^n \left(\prod_{j=1}^{n} a(x_j)\right) \exp\left[\sum_{k=1}^{K} \phi_k(\boldsymbol{\theta}) \sum_{j=1}^{n} h_k(x_j)\right] \\ &= g^n(\boldsymbol{\theta})\, a(\boldsymbol{x}) \exp\left[\boldsymbol{\phi}^t(\boldsymbol{\theta}) \sum_{j=1}^{n} \boldsymbol{h}(x_j)\right], \end{aligned}$$

where $a(\boldsymbol{x}) = \prod_{j=1}^{n} a(x_j)$. Then, using the factorization theorem it is easy to see that

$$\boldsymbol{t} = \left\{n, \sum_{j=1}^{n} h_1(x_j), \cdots, \sum_{j=1}^{n} h_K(x_j)\right\}$$

is a sufficient statistic for $\boldsymbol{\theta}$.

Proposition 4 [Conjugate priors of the Exponential family] A conjugate prior family for the exponential family

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = a(\boldsymbol{x})\, g(\boldsymbol{\theta}) \exp\left[\sum_{k=1}^{K} \phi_k(\boldsymbol{\theta})\, h_k(\boldsymbol{x})\right]$$

is given by

$$\pi(\boldsymbol{\theta}|\tau_0, \boldsymbol{\tau}) = z(\boldsymbol{\tau})[g(\boldsymbol{\theta})]^{\tau_0} \exp\left[\sum_{k=1}^{K} \tau_k \phi_k(\boldsymbol{\theta})\right]$$

The associated posterior law is

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}, \tau_0, \boldsymbol{\tau}) \propto [g(\boldsymbol{\theta})]^{n+\tau_0} a(\boldsymbol{x}) z(\boldsymbol{\tau}) \exp\left[\sum_{k=1}^{K} \left(\tau_k + \sum_{j=1}^{n} h_k(x_j)\right) \phi_k(\boldsymbol{\theta})\right].$$

We can rewrite this in a more compact way:
If

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \mathbf{Exfn}(\boldsymbol{x}|a(\boldsymbol{x}), g(\boldsymbol{\theta}), \boldsymbol{\phi}, \boldsymbol{h}),$$

then a conjugate prior family is

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}) = \mathbf{Exfn}(\boldsymbol{\theta}|g^{\tau_0}, z(\boldsymbol{\tau}), \boldsymbol{\tau}, \boldsymbol{\phi}),$$

and the associated posterior law is

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\tau}) = \mathbf{Exfn}(\boldsymbol{\theta}|g^{n+\tau_0}, a(\boldsymbol{x})\, z(\boldsymbol{\tau}), \boldsymbol{\tau}', \boldsymbol{\phi})$$

where

$$\tau_k' = \tau_k + \sum_{j=1}^{n} h_k(x_j)$$

or

$$\boldsymbol{\tau}' = \boldsymbol{\tau} + \bar{\boldsymbol{h}}, \quad \text{with} \quad \bar{h}_k = \sum_{j=1}^{n} h_k(x_j).$$

Définition 9 [Conjugate priors of natural exponential family] If

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = a(\boldsymbol{x}) \exp\left[\boldsymbol{\theta}^t \boldsymbol{x} - b(\boldsymbol{\theta})\right]$$

Then a conjugate prior family is

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}_0) = g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\tau}_0^t \boldsymbol{\theta} - d(\boldsymbol{\tau}_0)\right]$$

and the corresponding posterior is

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\tau}_0) = g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\tau}_n^t \boldsymbol{\theta} - d(\boldsymbol{\tau}_n)\right] \quad \text{with} \quad \boldsymbol{\tau}_n = \boldsymbol{\tau}_0 + \bar{\boldsymbol{x}}$$

where

$$\bar{\boldsymbol{x}}_n = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_j$$

A slightly more general notation which gives some more explicit properties of the conjugate priors of the natural exponential family is the following:
If

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = a(\boldsymbol{x}) \exp\left[\boldsymbol{\theta}^t \boldsymbol{x} - b(\boldsymbol{\theta})\right]$$

Then a conjugate prior family is

$$\pi(\boldsymbol{\theta}|\alpha_0, \boldsymbol{\tau}_0) = g(\alpha_0, \boldsymbol{\tau}_0) \exp\left[\alpha_0 \, \boldsymbol{\tau}_0^t \boldsymbol{\theta} - \alpha_0 b(\boldsymbol{\tau}_0)\right]$$

The posterior is

$$\pi(\boldsymbol{\theta}|\alpha_0, \boldsymbol{\tau}_0, \boldsymbol{x}) = g(\alpha, \boldsymbol{\tau}) \exp\left[\alpha \, \boldsymbol{\tau}^t \boldsymbol{\theta} - \alpha b(\boldsymbol{\tau})\right]$$

with

$$\alpha = \alpha_0 + n \quad \text{and} \quad \boldsymbol{\tau} = \frac{\alpha_0 \boldsymbol{\tau}_0 + n\bar{\boldsymbol{x}}}{(\alpha_0 + n)})$$

and we have the following properties:

$$\mathrm{E}\left[\boldsymbol{X}|\boldsymbol{\theta}\right] = \mathrm{E}\left[\bar{\boldsymbol{X}}|\boldsymbol{\theta}\right] = \nabla b(\boldsymbol{\theta})$$

$$\mathrm{E}\left[\nabla b(\boldsymbol{\Theta})|\alpha_0, \boldsymbol{\tau}_0\right] = \boldsymbol{\tau}_0$$

$$\mathrm{E}\left[\nabla b(\boldsymbol{\theta})|\alpha_0, \boldsymbol{\tau}_0, \boldsymbol{x}\right] = \frac{n\bar{\boldsymbol{x}} + \alpha_0 \boldsymbol{\tau}_0}{\alpha_0 + n} = \pi \bar{\boldsymbol{x}}_n + (1 - \pi)\boldsymbol{\tau}_0, \quad \text{with} \quad \pi = \frac{n}{\alpha_0 + n}$$

Conjugate priors

Observation law $p(x\|\theta)$	Prior law $p(\theta\|\boldsymbol{\tau})$	Posterior law $p(\theta\|x,\boldsymbol{\tau}) \propto p(\theta\|\boldsymbol{\tau})p(x\|\theta)$
Discrete variables		
Binomial $\mathbf{Bin}(x\|n,\theta)$	Beta $\mathbf{Bet}(\theta\|\alpha,\beta)$	Beta $\mathbf{Bet}(\theta\|\alpha+x,\beta+n-x)$
Negative Binomial $\mathbf{NegBin}(x\|n,\theta)$	Beta $\mathbf{Bet}(\theta\|\alpha,\beta)$	Beta $\mathbf{Bet}(\theta\|\alpha+n,\beta+x)$
Multinomial $\mathbf{M}_k(x\|\theta_1,\cdots,\theta_k)$	Dirichlet $\mathbf{Di}_k(\theta\|\alpha_1,\cdots,\alpha_k)$	Dirichlet $\mathbf{Di}_k(\theta\|\alpha_1+x_1,\cdots,\alpha_k+x_k)$
Poisson $\mathbf{Pn}(x\|\theta)$	Gamma $\mathbf{Gam}(\theta\|\alpha,\beta)$	Gamma $\mathbf{Gam}(\theta\|\alpha+x,\beta+1)$
Gamma $\mathbf{Gam}(x\|\nu,\theta)$	Gamma $\mathbf{Gam}(\theta\|\alpha,\beta)$	Gamma $\mathbf{Gam}(\theta\|\alpha+\nu,\beta+x)$
Beta $\mathbf{Bet}(x\|\alpha,\theta)$	Exponential $\mathbf{Ex}(\theta\|\lambda)$	Exponential $\mathbf{Ex}(\theta\|\lambda-\log(1-x))$
Normal $\mathbf{N}(x\|\theta,\sigma^2)$	Normal $\mathbf{N}(\theta\|\mu,\tau^2)$	Normal $\mathbf{N}\left(\mu\|\frac{\mu\sigma^2+\tau^2x}{\sigma^2+\tau^2},\frac{\sigma^2\,\tau^2}{\sigma^2+\tau^2}\right)$
Continuous variables		
Normal $\mathbf{N}(x\|\mu,1/\theta)$	Gamma $\mathbf{Gam}(\theta\|\alpha,\beta)$	Gamma $\mathbf{Gam}\left(\theta\|\alpha+\frac{1}{2},\beta+\frac{1}{2}(\mu-x)^2\right)$
Normal $\mathbf{N}(x\|\theta,\theta^2)$	Generalized inverse Normal $\mathbf{INg}(\theta\|\alpha,\mu,\sigma) \propto$ $\|\theta\|^{-\alpha}\exp\left[-\frac{1}{2\sigma^2}\left(\frac{1}{\theta}-\mu\right)^2\right]$	Generalized inverse Normal $\mathbf{INg}(\theta\|\alpha_n,\mu_n,\sigma_n)$

Table 8.1: Relation between the sampling distributions, their associated conjugate priors and their corresponding posteriors

8.3 Non informative priors based on Fisher information

Another notion of information related to the maximum likelihood estimation is the Fisher information. In this section, first we give some definitions and results related to this notion and we see how this is used to define non informative priors.

Proposition 5 [Information Inequality] Let $\widehat{\theta}$ be an estimate of the parameter θ in a family $\{P_\theta; \theta \in \mathcal{T}\}$ and assume that the following conditions hold:

1. The family $\{P_\theta; \theta \in \mathcal{T}\}$ has a corresponding family of densities $\{p_\theta(x); \theta \in \mathcal{T}\}$, all with the same support.
2. $p_\theta(x)$ is differentiable for all $\theta \in \mathcal{T}$ and all x in its support.
3. The integral
$$g(\theta) = \int_\Gamma h(x)\, p_\theta(x)\, \mu(\,\mathrm{d}x)$$
exists and is differentiable for $\theta \in \mathcal{T}$, for $h(x) = \widehat{\theta}(x)$ and for $h(x) = 1$ and
$$\frac{\partial g(\theta)}{\partial \theta} = \int_\Gamma h(x)\, \frac{\partial p_\theta(x)}{\partial \theta}\, \mu(\,\mathrm{d}x)$$

Then
$$\mathrm{Var}_\theta[\widehat{\theta}(X)] \geq \frac{\left[\frac{\partial}{\partial\theta}\mathrm{E}_\theta\left\{\widehat{\theta}(X)\right\}\right]^2}{I_\theta} \tag{8.10}$$
where
$$I_\theta \overset{\text{def}}{=} \mathrm{E}_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln p_\theta(X)\right]^2\right\} \tag{8.11}$$

Furthermore, if $\frac{\partial^2}{\partial\theta^2}p_\theta(x)$ exists for all $\theta \in \mathcal{T}$ and all x in the support of $p_\theta(x)$, and if
$$\int \frac{\partial^2}{\partial\theta^2}p_\theta(x)\, \mu(\,\mathrm{d}x) = \frac{\partial^2}{\partial\theta^2}\int p_\theta(x)\, \mu(\,\mathrm{d}x)$$
then I_θ can be computed via
$$I_\theta = -\mathrm{E}_\theta\left\{\frac{\partial^2}{\partial\theta^2}\ln p_\theta(X)\right\} \tag{8.12}$$

The quantity defined in (8.11) is known as *Fisher's information* for estimating θ from X, and (8.10) is called the *information inequality.*

For the particular case in which $\widehat{\theta}$ is unbiased $\mathrm{E}_\theta\left\{\widehat{\theta}(X)\right\} = \theta$, the information inequality becomes
$$\mathrm{Var}_\theta[\widehat{\theta}(X)] \geq \frac{1}{I_\theta} \tag{8.13}$$

Expression $\frac{1}{I_\theta}$ is known as the *Cramer-Rao lower bound* (CRLB).

Exemple 20 [The information Inequality for exponential families] Assume that $\mathcal{T}$ is open and p_θ is given by

$$p_\theta(x) = a(x)\, g(\theta) \exp\left[g(\theta)\, h(x)\right]$$

Then it can be shown that

$$I_\theta \stackrel{\text{def}}{=} \mathrm{E}_\theta \left\{ \left[\frac{\partial}{\partial \theta} \ln p_\theta(X) \right]^2 \right\} = \left|g'(\theta)\right|^2 \mathrm{Var}_\theta(h(X)) \tag{8.14}$$

and

$$\frac{\partial}{\partial \theta} \mathrm{E}_\theta \left\{h(X)\right\} = g'(\theta)\, \mathrm{Var}_\theta(h(X)) \tag{8.15}$$

and thus, if we choose $\widehat{\theta}(x) = h(x)$ we obtain the lower bound in the information inequality (8.10)

$$\mathrm{Var}_\theta[\widehat{\theta}(X)] = \frac{\left[\frac{\partial}{\partial \theta} \mathrm{E}_\theta \left\{\widehat{\theta}(X)\right\}\right]^2}{I_\theta} \tag{8.16}$$

Définition 10 [Non informative priors] Assume $X \sim f(x|\theta) = p_\theta(x)$ and assume that

$$I_\theta \stackrel{\text{def}}{=} \mathrm{E}_\theta \left\{ \left[\frac{\partial}{\partial \theta} \ln p_\theta(X) \right]^2 \right\} = -\mathrm{E}_\theta \left\{ \frac{\partial^2}{\partial \theta^2} \ln p_\theta(X) \right\} \tag{8.17}$$

Then, a non informative prior $\pi(\theta)$ is defined as

$$\pi(\theta) \propto I_\theta^{1/2} \tag{8.18}$$

Définition 11 [Non informative priors, case of vector parameters]
Assume $X \sim f(x|\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(x)$ and assume that

$$I_{ij}(\theta) \stackrel{\text{def}}{=} -\mathrm{E}_\theta \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p_{\boldsymbol{\theta}}(X) \right\} \tag{8.19}$$

Then, a non informative prior $\pi(\boldsymbol{\theta})$ is defined as

$$\pi(\boldsymbol{\theta}) \propto |\boldsymbol{I}(\boldsymbol{\theta})|^{1/2} \tag{8.20}$$

where $\boldsymbol{I}(\theta)$ is the Fisher information matrix with the elements $I_{ij}(\boldsymbol{\theta})$.

Exemple 21 If

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = a(\boldsymbol{x}) \exp\left[\boldsymbol{\theta}^t \boldsymbol{x} - b(\boldsymbol{\theta})\right]$$

then

$$\boldsymbol{I}(\boldsymbol{\theta}) = \nabla \nabla^t b(\boldsymbol{\theta})$$

and

$$\pi(\boldsymbol{\theta}) \propto |\boldsymbol{I}(\boldsymbol{\theta})|^{1/2} = \left| \prod_{i=1}^{n} \frac{\partial^2 \theta_i}{\partial b(\boldsymbol{\theta})^2} \right|^{1/2}$$

Exemple 22 If

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{N}\left(\mu, \sigma^2\right), \quad \boldsymbol{\theta} = (\mu, \sigma^2)$$

then

$$I(\boldsymbol{\theta}) = \mathrm{E}_\theta\left\{\begin{pmatrix} \frac{1}{\sigma^2} & \frac{2(X-\mu)}{\sigma^3} \\ \frac{2(X-\mu)}{\sigma^3} & \frac{3(X-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2} \end{pmatrix}\right\} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$$

and

$$\pi(\boldsymbol{\theta}) = \pi(\mu, \sigma^2) \propto \frac{1}{\sigma^4}$$

Chapter 9

Linear Estimation

In previous chapters we saw that the optimum estimation, in a MMSE sense, of an unknown signal X_t, given the observations $Y_{a:b} = \{Y_a, \ldots, Y_b\}$ of a related quantity, is given by $\widehat{X}_t = \mathrm{E}\left[X_t | Y_{a:b}\right]$. This estimate is not, in general, a linear function of the data and its computation needs the knowledge of the joint distribution of $\{X_t, Y_{a:b}\}$. Only when this joint distribution is Gaussian and when X_t is a related to $Y_{a:b}$ by a linear relation, this optimal estimate is a linear function of the data. Even in this case, its computation needs the inversion of the covariance matrix of the data $\mathbf{\Sigma}_Y$ whose dimensions increase with the number of data.

One way to circumvent these drawbacks is, from the first step, to constraint the estimate to be a linear function of the data. Doing so, as we will see below, we do not need anymore the joint distribution of $\{X_t, Y_{a:b}\}$ but only its second order statistics. Furthermore, we will see that, in this case, we can develop real time or on-line algorithms with lower complexity and lower cost, if we assume data to be stationary.

9.1 Introduction

Assume that we want to obtain an estimate $\widehat{X}_t$ of a quantity X_t which is a linear (or more precisely an affine) function of the data $Y_{a:b} = \{Y_a, \ldots, Y_b\}$, *i.e.*

$$\widehat{X}_t = \sum_{n=a}^{b} h_{t,n} Y_n + c_t \tag{9.1}$$

where, in general, a can be either $-\infty$ or finite and b can also be either finite or ∞. When a and b are finite the meaning of the summation is clear. For the cases where $a = -\infty$ or $b = \infty$, these and all the following summations have to be understood in the MMSE sense, for example for the case $a = -\infty$

$$\lim_{m \mapsto -\infty} \mathrm{E}\left[\left(\sum_{n=m}^{b} h_{t,n} Y_n + c_t - \widehat{X}_t\right)^2\right] = 0 \tag{9.2}$$

In these cases we need also to assume that

$$\mathrm{E}\left[X_n^2\right] < \infty \quad \text{and} \quad \mathrm{E}\left[Y_n^2\right] < \infty.$$

The following propositions resume all we need for developing linear estimation theory.

Proposition 1 *Assume $\widehat{X}_t \in \mathcal{H}_a^b$ where $\mathcal{H}_a^b$ is the Hilbert space generated by the affine transform (9.1). Then $E\left[\widehat{X}_t^2\right] < \infty$ and if Z is a random variable satisfying $E\left[Z^2\right] < \infty$, then*

$$E\left[Z\,\widehat{X}_t\right] = \sum_{n=a}^{b} h_{t,n} E[Z\,Y_n] + c_t\,E[Z]$$

Proposition 2 (Orthogonality principle) *$\widehat{X}_t \in \mathcal{H}_a^b$ solves*

$$\min_{\widehat{X}_t \in \mathcal{H}_a^b} E\left[(\widehat{X}_t - X_t)^2\right] \tag{9.3}$$

if and only if

$$E\left[(\widehat{X}_t - X_t)\,Z\right] = 0 \quad \forall Z \in \mathcal{H}_a^b. \tag{9.4}$$

In other words, $\widehat{X}_t$ is a MMSE linear estimate of X_t given $Y_{a:b}$ if and only if the estimation error $(\widehat{X}_t - X_t)$ is orthogonal to every linear function of the observation $Y_{a:b}$.

Considering the particular cases of $Z = 1$ and $Z = Y_l$, $a \leq l \leq b$ we can rewrite this proposition in the following way

Proposition 3 *$\widehat{X}_t \in \mathcal{H}_a^b$ solves (9.3) if and only if*

$$E\left[\widehat{X}_t\right] = E[X_t] \tag{9.5}$$

and

$$E\left[(\widehat{X}_t - X_t)\,Y_l\right] = 0 \quad \forall a \leq l \leq b. \tag{9.6}$$

Now replacing (9.1) in (9.6) we obtain

$$\mathrm{E}\left[(X_t - \widehat{X}_t)\,Y_l\right] = \mathrm{E}\left[(X_t - \sum_{n=a}^{b} h_{t,n} Y_n - c_t)\,Y_l\right] = 0 \quad \forall a \leq l \leq b. \tag{9.7}$$

To go further in details more easily and without any loss of generality, we assume $\mathrm{E}\left[Y_l\right] = 0, \forall a \leq l \leq b$. Then, since $\mathrm{E}\left[X_t\right] = c_t$, the previous equation becomes

$$\mathrm{Cov}\left\{X_t, Y_l\right\} = \sum_{n=a}^{b} h_{t,n} \mathrm{Cov}\left\{Y_n, Y_l\right\} \quad \forall a \leq l \leq b. \tag{9.8}$$

which is known as the *Wiener-Hopft* equation.

Writing this in a matrix form we have

$$\boldsymbol{\sigma}_{XY}(t) = \boldsymbol{\Sigma}_Y \boldsymbol{h}_t \tag{9.9}$$

where

$$\begin{aligned}
\boldsymbol{\sigma}_{XY}(t) &\overset{\mathrm{def}}{=} \left[\mathrm{Cov}\left\{X_t, Y_a\right\}, \ldots, \mathrm{Cov}\left\{X_t, Y_b\right\}\right]^t \\
\boldsymbol{\Sigma}_{XY}(t) &\overset{\mathrm{def}}{=} \left[\mathrm{Cov}\left\{Y_n, Y_l\right\}\right] \\
\boldsymbol{h}_t &\overset{\mathrm{def}}{=} \left[h_{t,a}, \ldots, h_{t,b}\right]^t
\end{aligned}$$

So, theoretically we have

$$\boldsymbol{h}_t = \boldsymbol{\Sigma}_Y^{-1} \boldsymbol{\sigma}_{XY}(t) \tag{9.10}$$

The main difficulty is however the computation of $\boldsymbol{\Sigma}_Y^{-1}$. Note that this matrix is symmetric and positive definite. So, theoretically, it is not singular. However, its inversion cost increases exponentially with the number of data. In the following we will see how the stationary assumption will help to reduce this cost.

9.2 One step prediction

Consider the case where $a = 0$, $b = t$ and $X_t = Y_{t+1}$ and assume that Y_l is wide sense stationary, *i.e.*, $\mathrm{E}[Y_l] = 0$ and $\mathrm{Cov}\{Y_l, Y_m\} = C_Y(l-m)$. Then we have

$$\mathrm{Cov}\{X_t, Y_l\} = \mathrm{Cov}\{Y_{t+1}, Y_l\} = C_Y(t+1-l) \tag{9.11}$$

and the Wiener-Hopft equation becomes

$$\begin{pmatrix} C_Y(t+1) \\ C_Y(t) \\ \vdots \\ \\ \vdots \\ C_Y(1) \end{pmatrix} = \begin{pmatrix} C_Y(0) & C_Y(1) & & C_Y(t) \\ C_Y(1) & \ddots & \ddots & \\ & \ddots & & \\ & & & \\ & & & C_Y(1) \\ C_Y(t) & & C_Y(1) & C_Y(0) \end{pmatrix} \begin{pmatrix} h_{t,0} \\ h_{t,1} \\ \vdots \\ \\ {:}h_{t,t} \end{pmatrix} \tag{9.12}$$

called *Yule-Walker* equation.

Note that the w.s.s. hypothesis of the data Y_n leads to a covariance matrix which is Toeplitz. Unlike the general case, the cost of the inversion of this matrix is only $O(n^2)$ against $O(n^3)$ for the general case, where n is the number of data. Thus, in any linear MMSE estimation problem, the w.s.s. assumption can reduce the complexity of the computation of the coefficients by a factor equal to the number of the data.

In the following, we will see that, we can still go further and use the specific structure of the Yule-Walker equation to keep on reducing this cost.

9.3 Levinson algorithm

Levinson algorithm uses the special structure of the Yule-Walker equation for the one step prediction problem where the left hand side vector of this equation is equal to the last column of the covariance matrix shifted by one time unit.

Rewriting this equation

$$\widehat{Y}_{t+1} = \sum_{n=0}^{t} h_{t,n} Y_n = -\sum_{n=0}^{t} a_{t+1,t+1-n} Y_n \tag{9.13}$$

the coefficients $a_{t,1}, \ldots, a_{t,t}$ can be updated recursively in t through the Levinson algorithm :

$$\begin{aligned} a_{t+1,k} &= a_{t,k} - k_t \, a_{t,t+1-k}, \quad k = 1, \ldots, t \\ a_{t+1,t+1} &= -k_t \end{aligned}$$

where k_t itself, is generated recursively with $\epsilon_t \overset{\text{def}}{=} \mathrm{E}\left[(Y_t - \widehat{Y}_t)^2\right]$ via

$$k_t = \frac{1}{\epsilon_t}[C_Y(t+1) + \sum_{k=1}^{t} a_{t,k}\, C_Y(t+1-k)]$$

$$\epsilon_{t+1} \quad = \quad (1-k_t^2)\epsilon_t$$

with the initialization $k_0 = \frac{C_Y(1)}{C_Y(0)}$ and $\epsilon_0 = C_Y(0)$.

The coefficients a_k are called *reflection coefficients* or still *partial correlation coefficients* (PARCOR).

9.4 Vector observation case

The linear estimation can be extended to the case where both the observation sequence and the quantity to be estimated are vectors. This extension is straight forward and we have:

$$\widehat{\underline{X}}_t = \sum_{n=a}^{b} \boldsymbol{H}_{t,n}\underline{Y}_n + \underline{c}_t \tag{9.14}$$

where $\boldsymbol{H}_{t,n}$ is a sequence of matrices. When a or b are infinite, the summations have the MSE sense. For example when $a = -\infty$, we have

$$\lim_{m \mapsto -\infty} \mathrm{E}\left[\left\|\sum_{n=m}^{b} \boldsymbol{H}_{t,n}\underline{Y}_n + \underline{c}_t - \widehat{\underline{X}}_t\right\|^2\right] = 0 \tag{9.15}$$

where $\|\boldsymbol{x}\| \overset{\text{def}}{=} \boldsymbol{x}^t\boldsymbol{x}$.

The orthogonality principle becomes:

Proposition 4 (Orthogonality principle) *$\widehat{\underline{X}}_t \in \mathcal{H}_a^b$ solves*

$$\min_{\widehat{\underline{X}}_t \in \mathcal{H}_a^b} E\left[\left\|\widehat{\underline{X}}_t - \underline{X}_t\right\|^2\right] \tag{9.16}$$

if and only if

$$E\left[(\widehat{\underline{X}}_t - \underline{X}_t)^t\, \underline{Z}\right] = 0 \quad \forall Z \in \mathcal{H}_a^b. \tag{9.17}$$

Writing this last equation for $\underline{Z} = \underline{1}$ and for $\underline{Z} = \underline{Y}_l^t$ we obtain

$$\begin{aligned} \mathrm{E}\left[\underline{X}_t\right] &= \mathrm{E}\left[\widehat{\underline{X}}_t\right] \\ \mathrm{E}\left[(\underline{X}_t - \widehat{\underline{X}}_t)\,\underline{Y}_l^t\right] &= 0, \quad \forall a \le l \le b. \end{aligned}$$

Using these relations we obtain

$$\mathrm{E}\left[(\underline{X}_t - \widehat{\underline{X}}_t)\,\underline{Y}_l^t\right] = \mathrm{E}\left[\left(\underline{X}_t - \sum_{n=a}^{b} \boldsymbol{H}_{t,n}\underline{Y}_n - \underline{c}_t\right)\,\underline{Y}_l^t\right] = [\mathbf{0}], \quad \forall a \le l \le b. \tag{9.18}$$

where $[\mathbf{0}]$ means a matrix whose all elements are equal to zero. The Wiener-Hopft equation becomes:

$$\boldsymbol{C}_{XY}(t,l) = \sum_{n=a}^{b} \boldsymbol{H}_{t,n} \boldsymbol{C}_Y(n,l), \quad \forall a \leq l \leq b$$

where $\boldsymbol{C}_{XY}(t,l) \overset{\text{def}}{=} \operatorname{Cov}\{\underline{X}_t, \underline{Y}_l\}$ is the cross-covariance of $\{\underline{X}_n\}_{n=-\infty}^{\infty}$ and $\{\underline{Y}_l\}_{l=-\infty}^{\infty}$ and $\boldsymbol{C}_Y(n,l) \overset{\text{def}}{=} \operatorname{Cov}\{\underline{Y}_n, \underline{Y}_l\}$ is the auto-covariance of $\{\underline{Y}_n\}_{n=-\infty}^{\infty}$. Note that $\boldsymbol{C}_{XY}(t,l)$ and $\boldsymbol{C}_Y(n,l)$ are $(m \times k)$ and $(k \times k)$ matrices respectively, where k and m are respectively the dimensions of the vectors $\underline{Y}_n$ and $\underline{X}_t$.

9.5 Wiener-Kolmogorov filtering

We assume here that Y_n is wide sense stationarity (w.s.s.) and that there is an infinite number of observations.

Two cases are of interest: Non causal where $(a = -\infty, b = t)$ and Causal where $(a = -\infty, b = \infty)$.

9.5.1 Non causal Wiener-Kolmogorov

Without losing any generality, we assume $\operatorname{E}[Y_n] = \operatorname{E}[X_n] = 0$. Then we have

$$\widehat{X}_t = \sum_{n=-\infty}^{\infty} h_{t,n} Y_n \tag{9.19}$$

The Wiener-Hopft equation becomes

$$C_{XY}(t-l) = \sum_{n=-\infty}^{\infty} h_{t,n} C_Y(n-l) \tag{9.20}$$

Changing variable $\tau = t - l$ we obtain

$$C_{XY}(\tau) = \sum_{n=-\infty}^{\infty} h_{t,n} C_Y(n+\tau-t) \tag{9.21}$$

Changing now the summation variable $n = t - \alpha$ we obtain

$$C_{XY}(\tau) = \sum_{\alpha=-\infty}^{\infty} h_{t,t-\alpha} C_Y(\tau-\alpha) \tag{9.22}$$

In this summation t appears only in $h_{t,t-\alpha}$. This means that we can choose it to be independent of t, *i.e.* if this equation has a solution, it can be chosen to be time-invariant with coefficients $h_{t,t-\alpha} = h_{0,0-\alpha} \overset{\text{def}}{=} h_\alpha$. Then we have

$$C_{XY}(\tau) = \sum_{\alpha=-\infty}^{\infty} h_\alpha C_Y(\tau-\alpha) \tag{9.23}$$

which is a convolution equation. Using then the following DFTs

$$\begin{aligned} H(\omega) &\overset{\text{def}}{=} \sum_{n=-\infty}^{\infty} h_n \exp\left[-j\omega n\right], \quad -\pi < \omega < \pi \\ S_{XY}(\omega) &\overset{\text{def}}{=} \sum_{n=-\infty}^{\infty} C_{XY}(n) \exp\left[-j\omega n\right] \quad -\pi < \omega < \pi \\ S_Y(\omega) &\overset{\text{def}}{=} \sum_{n=-\infty}^{\infty} C_Y(n) \exp\left[-j\omega n\right] \quad -\pi < \omega < \pi \end{aligned}$$

we have

$$S_{XY}(\omega) = H(\omega)\, S_Y(\omega)$$

and finally

$$H(\omega) = \frac{S_{XY}(\omega)}{S_Y(\omega)} \tag{9.24}$$

The coefficients h_n can then be obtained by inverse FT

$$h_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S_{XY}(\omega)}{S_Y(\omega)} \exp\left[j\omega n\right] \,\mathrm{d}\omega, \quad n \in \mathbf{Z} \tag{9.25}$$

It is interesting to see that we have

$$\begin{aligned} \mathrm{E}\left[\widehat{X}_t X_t\right] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{S_{XY}(\omega)}{S_Y(\omega)} \,\mathrm{d}\omega \\ \mathrm{E}\left[X_t^2\right] &= C_X(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_X(\omega)\,\mathrm{d}\omega \\ \mathrm{E}\left[(\widehat{X}_t - X_t)^2\right] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_X(\omega) - \frac{S_{XY}(\omega)}{S_Y(\omega)} \,\mathrm{d}\omega \end{aligned}$$

This last equation can be written

$$MMSE = \mathrm{E}\left[(\widehat{X}_t - X_t)^2\right] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left[1 - \frac{|S_{XY}(\omega)|^2}{S_X(\omega) S_Y(\omega)}\right] S_X(\omega)\,\mathrm{d}\omega \tag{9.26}$$

Noting that we have $|S_{XY}(\omega)|^2 \leq S_X(\omega)\, S_Y(\omega)$, with equality if $\{X_t\}_{t=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$ are perfectly correlated, we can conclude that the MMSE ranges from $\mathrm{E}\left[X_t^2\right]$ to zero as the relationship between $\{X_t\}_{t=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$ ranges from independence to perfect correlation.

Example 1 (Noise filtering) *Consider the model*

$$Y_n = S_n + N_n, \quad n \in \boldsymbol{Z}$$

where S_n and N_n are assumed uncorrelated, zero mean and w.s.s. Suppose that we want to estimate $X_t = S_{t+\lambda}$ for some integer λ. The problem represents filtering, prediction *and smoothing respectively when $\lambda = 0$, $\lambda > 0$ and when $\lambda < 0$. To obtain the necessary equations, it is straightforward to show*

$$\begin{aligned} S_Y(\omega) &= S_X(\omega) + S_N(\omega) \\ S_{XY}(\omega) &= \exp[j\omega\lambda]\; S_X(\omega) \\ S_X(\omega) &= S_S(\omega) \end{aligned}$$

So, the transfer function of the optimum non causal filter is

$$H(\omega) = \frac{\exp[j\omega\lambda]\; S_S(\omega)}{S_S(\omega) + S_N(\omega)}$$

Example 2 (Deconvolution) *Consider the model*

$$Y_n = \sum_{k=0}^{p} h_k S_{n-k} + N_n, \quad n \in \boldsymbol{Z}$$

where S_n and N_n are assumed uncorrelated, zero mean and w.s.s. Suppose that we want to estimate $X_t = S_{t+\lambda}$ for some integer λ. Again here, it is straightforward to show

$$\begin{aligned} S_Y(\omega) &= |H(\omega)|^2 S_X(\omega) + S_N(\omega) \\ S_{XY}(\omega) &= \exp[j\omega\lambda]\; H^*(\omega)\, S_X(\omega) \\ S_X(\omega) &= S_S(\omega) \end{aligned}$$

where

$$H(\omega) = \sum_{n=0}^{p} h_n \exp[-jn\omega], -\pi < \omega < \pi$$

So, the transfer function of the optimum non causal filter is

$$H(\omega) = \frac{\exp[j\omega\lambda]\; H(\omega)^* S_S(\omega)}{|H(\omega)|^2 S_S(\omega) + S_N(\omega)}$$

For $\lambda = 0$ we have

$$H(\omega) = \frac{H(\omega)^* S_S(\omega)}{|H(\omega)|^2 S_S(\omega) + S_N(\omega)} = \frac{1}{H(\omega)} \frac{|H(\omega)|^2}{|H(\omega)|^2 + \frac{S_N(\omega)}{S_S(\omega)}}$$

9.6 Causal Wiener-Kolmogorov

To develop the causal Wiener-Kolmogorov filtering, let note by $\widetilde{X}_t$ the non causal Wiener-Kolmogorov solution and by $\widehat{X}_t$ the causal one, *i.e.*

$$\begin{aligned} \widehat{X}_t &= \sum_{n=-\infty}^{\infty} h_{t-n} Y_n \\ \widetilde{X}_t &= \sum_{n=-\infty}^{\infty} \tilde{h}_{t-n} Y_n \end{aligned}$$

Note also that $\mathcal{H}_{-\infty}^{t}$ is a subset of $\mathcal{H}_{-\infty}^{\infty}$. So, if the solution to the noncausal Wiener-Kolmogorov problem happens to be causal, it also solves the causal Wiener-Kolmogorov problem. But unfortunately, this is not the case excepted very special cases. However, there is surely a relation between these two solutions.

To obtain this relation we start by writing

$$(X_t - \widehat{X}_t) = (\widetilde{X}_t - \widehat{X}_t) + (X_t - \widetilde{X}_t)$$

So, for any $Z \in \mathcal{H}_{-\infty}^{t}$ we have

$$\mathrm{E}\left[(X_t - \widehat{X}_t)Z\right] = \mathrm{E}\left[(\widetilde{X}_t - \widehat{X}_t)Z\right] + \mathrm{E}\left[(X_t - \widetilde{X}_t)Z\right]$$

The left hand term $\mathrm{E}\left[(X_t - \widehat{X}_t)Z\right]$ is zero due to the orthogonality principle applied to $\widehat{X}_t$. The second right hand term $\mathrm{E}\left[(X_t - \widetilde{X}_t)Z\right]$ is zero due to the orthogonality principle applied to $\widetilde{X}_t$. So we have

$$\mathrm{E}\left[(\widetilde{X}_t - \widehat{X}_t)Z\right] = 0, \quad \forall Z \in \mathcal{H}_{-\infty}^{t}$$

which means that $\widehat{X}_t$ is the MMSE estimate of $\widetilde{X}_t$ among all estimates in $\mathcal{H}_{-\infty}^{t}$. In other words, $\widehat{X}_t$ which is the projection of X_t on $\mathcal{H}_{-\infty}^{t}$ can be obtained by first projecting X_t on $\mathcal{H}_{-\infty}^{\infty}$ to get $\widetilde{X}_t$ and then projecting $\widetilde{X}_t$ onto $\mathcal{H}_{-\infty}^{t}$.

Now, let define

$$\bar{X}_t \overset{\text{def}}{=} \sum_{n=-\infty}^{t} \tilde{h}_{t-n} Y_n$$

and consider the error

$$\widetilde{X}_t - \bar{X}_t = \sum_{n=-\infty}^{\infty} \tilde{h}_{t-n} Y_n - \sum_{n=-\infty}^{t} \tilde{h}_{t-n} Y_n = \sum_{n=t+1}^{\infty} \tilde{h}_{t-n} Y_n$$

If this error could be orthogonal to Y_m for all $m \leq t$, we could consider $\bar{X}_t$ as the projection of $\widetilde{X}_t$ on $\mathcal{H}_{-\infty}^{t}$. But this is not the case in general. This could be the case if $\{Y_n\}_{n=-\infty}^{\infty}$ was a sequence of uncorrelated random variables, because in that case we would have

$$\begin{aligned}
\mathrm{E}\left[(\widetilde{X}_t - \bar{X}_t)Y_m\right] &= \mathrm{E}\left[\left(\sum_{n=t+1}^{\infty} \tilde{h}_{t-n} Y_n\right) Y_m\right] \\
&= \sum_{n=t+1}^{\infty} \tilde{h}_{t-n} \mathrm{E}\left[Y_n, Y_m\right] \\
&= \sigma^2 \sum_{n=t+1}^{\infty} \tilde{h}_{t-n} \delta_{n,m} = 0, \quad \forall\, m \leq t
\end{aligned}$$

where $\delta_{n,m}$ is the Kronecker delta ($\delta_{n,m} = 1$ if $n = m$ and $\delta_{n,m} = 0$ if $n \neq m$ and where $\sigma^2 = \mathrm{E}\left[Y_n^2\right]$

From the above discussion we see that, if we transform first the data $\{Y_n\}_{n=-\infty}^{\infty}$ into an equivalent w.s.s. and uncorrelated sequence $\{Z_n\}_{n=-\infty}^{\infty}$, then the causal Wiener-Kolmogorov filter coefficients can be obtained by simple truncation on the non causal Wiener-Kolmogorov filter, *i.e.*

$$\widehat{X}_t = \sum_{n=-\infty}^{t} \widehat{h}_{t-n} Z_n$$

where

$$\widehat{h}_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{XZ}(\omega) \exp[jn\omega]\, \mathrm{d}\omega = C_{XZ}(n), \quad n \geq 0$$

where $S_{XZ}(\omega)$ and $C_{XZ}(n)$ are, respectively, the cross spectrum and the cross covariance of the sequences $\{X_n\}_{n=-\infty}^{\infty}$ and $\{Z_n\}_{n=-\infty}^{\infty}$.

Now the aim is to obtain $\{Z_n\}_{n=-\infty}^{\infty}$ from $\{Y_n\}_{n=-\infty}^{\infty}$. The analyse of the one step prediction problem in previous section in this chapter gives us the solution. If we note by $\widehat{Y}_{t|t-1}$ the one step prediction of Y_t from the data $\{Y_n\}_{n=-\infty}^{t-1}$ and by $\sigma_t^2 = \mathrm{E}\left[(Y_t - \widehat{Y}_{t|t-1})\right]$, then define

$$Z_n = \frac{Y_n - \widehat{Y}_{n|n-1}}{\sigma_n}, \quad n \in \mathbf{Z}$$

we can verify that $\mathrm{E}[Z_n] = 0$, $\mathrm{E}[Z_n^2] = 1$ and $\mathrm{Cov}\{Z_n, Z_m\} = 0$. So, Z_n has all the necessary properties that we need. Now, still we have to show that $\{Z_n\}_{n=-\infty}^{\infty}$ is equivalent to $\{Y_n\}_{n=-\infty}^{\infty}$ for the purpose of MMSE. To do this we need the result of the following theorem.

Theorem 1 (Spectral Factorization) *Assume Y_n has a power spectral density satisfying the Paley-Wiener condition*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln S_Y(\omega)\, d\omega > -\infty.$$

Then

$$S_Y(\omega) = S_Y^{-}(\omega)\, S_Y^{+}(\omega), \quad -\pi < \omega < \pi \tag{9.27}$$

where

$$S_Y(\omega) = |S_Y^{-}(\omega)|^2 = |S_Y^{+}(\omega)|^2 \tag{9.28}$$

and

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\pi}^{\pi} S_Y^{-}(\omega) \exp[jn\omega]\, d\omega &= 0 \quad n < 0 \\
\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1}{S_Y^{-}(\omega)} \exp[jn\omega]\, d\omega &= 0 \quad n < 0 \\
\frac{1}{2\pi}\int_{-\pi}^{\pi} S_Y^{+}(\omega) \exp[jn\omega]\, d\omega &= 0 \quad n > 0 \\
\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1}{S_Y^{+}(\omega)} \exp[jn\omega]\, d\omega &= 0 \quad n > 0
\end{aligned}$$

Now consider the time invariant filter with $H(\omega) = \frac{1}{S_Y^+(\omega)}$. This filter, by definition, is causal. The output of this filter with the wide sense stationnary input sequence Y_n will be another wide sense stationnary sequence Z_n with

$$S_Z(\omega) = \left| \frac{1}{S_Y^-(\omega)} \right|^2 S_Y(\omega) = 1, \quad -\pi < \omega < \pi \tag{9.29}$$

Since $S_Z(\omega) = 1$ corresponds to a white sequence, the filter is called a *whitenning filter.*

Note also that the input sequence Y_n can also be obtained causally from the output Z_n by

$$Y_t = \sum_{n=-\infty}^{\infty} f_{t-n} Z_n, \quad t \in \mathbf{Z} \tag{9.30}$$

where

$$f_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_Y^+(\omega) \exp\left[jn\omega\right] \, \mathrm{d}\omega, \quad n \geq 0$$

Now consider the λ-step prediction of Y_n from Z_n:

$$\begin{aligned} Y_{t+\lambda} &= \sum_{n=-\infty}^{t+\lambda} f_{t+\lambda-n} Z_n \\ &= \sum_{n=-\infty}^{t} f_{t+\lambda-n} Z_n + \sum_{n=t+1}^{t} f_{t+\lambda-n} Z_n \end{aligned}$$

But Z_n is white, so $Z_{n+1}, \ldots, Z_{t+\lambda}$ are orthogonal to $\{Z_n\}_{n=-\infty}^{t}$. So the best estimate $\widehat{Y}_{t+\lambda}$ of $Y_{t+\lambda}$ from $\{Y_n\}_{n=-\infty}^{t}$ is

$$\widehat{Y}_{t+\lambda} = \sum_{n=-\infty}^{t} f_{t+\lambda-n} Z_n$$

Combining the whitening filter and this prediction we obtain

$$\cdots, Y_{t-2}, Y_{t-1}, Y_t \longrightarrow \boxed{\frac{1}{S_Y^+(\omega)}} \longrightarrow Z_n \longrightarrow \boxed{\left[\exp\left[j\omega\lambda\right] S_Y^+(\omega)\right]_+} \longrightarrow \widehat{Y}_{t+\lambda}$$

where the operator $[H(\omega)]_+$ is defined as

$$[H(\omega)]_+ = \sum_{n=0}^{\infty} h_n \exp\left[-j\omega n\right]$$

where

$$h_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\omega) \exp\left[jn\omega\right] \, \mathrm{d}\omega$$

Now we obtain a procedure for causally prewhitenning a stationary sequence of observations. Thus the solution of the general causal Wiener-Kolmogorov problem follows immediately. Assuming that $\{Y_n\}_{n=-\infty}^{\infty}$ satisfies the Paley-Winner condition, it can be transformed equivalently into $\{Z_n\}_{n=-\infty}^{t}$ by passing it through the causal filter $1/S_Y^+(\omega)$

$$Y_n \longrightarrow \boxed{\frac{1}{S_Y^+(\omega)}} \longrightarrow Z_n$$

Then we need to find the cross spectrum $S_{XZ}(\omega)$ and pass $\{Z_n\}_{n=-\infty}^{t}$ through the causal filter $[S_{XZ}(\omega)]_+$ to obtain the required result.

$$Z_n \longrightarrow \boxed{S^+_{XZ}(\omega)} \longrightarrow \widehat{Y}_t$$

Knowing that

$$S_{XZ}(\omega) = \frac{S_{XY}(\omega)}{[S^+_Y(\omega)]^*} = \frac{S_{XY}(\omega)}{S^-_Y(\omega)}$$

we obtain

$$Y_n \longrightarrow \boxed{\frac{1}{S^+_Y(\omega)}} \longrightarrow Z_n \longrightarrow \boxed{\left[\frac{S_{XY}(\omega)}{S^-_Y(\omega)}\right]_+} \longrightarrow \widehat{Y}_t$$

It is interesting to compare this filter with the non causal Wiener-Kolmogorov filter

$$\frac{S_{XY}(\omega)}{S_Y(\omega)} = \frac{1}{S^+_Y(\omega)}\left[\frac{S_{XY}(\omega)}{S^-_Y(\omega)}\right]$$

The following diagrams summarize this comparison:

$$Y_n \longrightarrow \boxed{\frac{1}{S^+_Y(\omega)}} \longrightarrow \boxed{\left[\frac{S_{XY}(\omega)}{S^-_Y(\omega)}\right]_+} \longrightarrow \widehat{Y}_{t+\lambda} \qquad Y_n \longrightarrow \boxed{\frac{1}{S^+_Y(\omega)}} \longrightarrow \boxed{\frac{S_{XY}(\omega)}{S^-_Y(\omega)}} \longrightarrow \widehat{Y}_{t+\lambda}$$

Causal Non Causal

9.7 Rational spectra

Definition 1 (Rational spectra) *$S_y(\omega)$ is said rational if it can be written as the ratio of two real trigonometric polynomials*

$$S_y(\omega) = \frac{n_0 + 2\sum_{k=1}^{p} n_k \cos k\omega}{d_0 + 2\sum_{k=1}^{m} d_k \cos k\omega}, \qquad n_k, d_k \in \mathbb{R} \tag{9.31}$$

Using the relation $\cos k\omega = \exp[jk\omega] + \exp[-jk\omega]$ we deduce

$$S_y(\omega) = \frac{N(\exp[jk\omega])}{D(\exp[jk\omega])}.$$

Noting $z = \exp[jk\omega]$ we have

$$\begin{aligned} N(z) &= \sum_{k=-p}^{p} n_{|k|} z^{-k} \\ D(z) &= \sum_{k=-p}^{p} d_{|k|} z^{-k} \end{aligned}$$

$z^p N(z)$ is a $(2p)$-th order polynomial. So we can write

$$N(z) = n_p\, z^{-p} \prod_{k=1}^{2p} (z - z_k)$$

Since $N(z) = N(1/z)$ the roots z_k are reciprocal pairs. If we order them in such a way that $|z_1| > |z_2| > \cdots > |z_{2p}|$ we have

$$|z_{2p}| = \frac{1}{|z_1|}, \quad |z_{2p-1}| = \frac{1}{|z_2|}, \cdots, |z_{p+1}| = \frac{1}{|z_p|}.$$

We deduce that all the roots are outside or over the unit circle. Due to the reciprocity of the roots we can write

$$N(z) = B(z)\, B(1/Z) \tag{9.32}$$

where

$$B(z) = \sqrt{(-1)^p np / \prod_{k=1}^{p} z_k} \prod_{k=1}^{p} (z^{-1} - z_k) \tag{9.33}$$

So $B(z)$ is a polynomial of degree p and can be extended as

$$B(z) = \sum_{k=0}^{p} b_k\, z^{-k} \tag{9.34}$$

Similarly, we can do exactly the same analysis for the denominator $D(z)$:

$$D(z) = A(z)\, A(1/Z) \tag{9.35}$$

where

$$A(z) = \sum_{k=0}^{m} a_k\, z^{-k} \tag{9.36}$$

Putting these together we have

$$S_y(\omega) = \frac{B\left(\exp\left[jk\omega\right]\right)\, B\left(\exp\left[-jk\omega\right]\right)}{A\left(\exp\left[jk\omega\right]\right)\, A\left(\exp\left[-jk\omega\right]\right)} = \frac{B\left(\exp\left[jk\omega\right]\right)}{A\left(\exp\left[jk\omega\right]\right)}\, \frac{B\left(\exp\left[-jk\omega\right]\right)}{A\left(\exp\left[-jk\omega\right]\right)}$$

Assuming that none of the roots of B or A is on the unit circle $|z| = 1$ we have

$$\begin{aligned} S_y^+(\omega) &= \frac{B\left(\exp\left[jk\omega\right]\right)}{A\left(\exp\left[jk\omega\right]\right)} \\ S_y^-(\omega) &= \frac{B\left(\exp\left[-jk\omega\right]\right)}{A\left(\exp\left[-jk\omega\right]\right)} \end{aligned}$$

Now consider the whitenning filter of the last section and assume that the power spectrum of the data $S_Y(\omega)$ is a rational fraction.

$$Y_n \longrightarrow \boxed{\frac{1}{S_Y^+(\omega)}} \longrightarrow Z_n \longrightarrow Y_n \longrightarrow \boxed{\frac{A(z)}{B(z)} = \frac{\sum_{k=0}^{m} a_k\, z^{-k}}{\sum_{k=0}^{p} b_k\, z^{-k}}} \longrightarrow Z_n$$

Then, we can see easily that we have

$$\sum_{k=0}^{m} a_k\, Y_{n-k} = \sum_{k=0}^{p} b_k\, Z_{n-k} \tag{9.37}$$

We can rewrite this equation in two other equivalent forms

$$\begin{aligned} b_0 z_n &= -\sum_{k=1}^{p} b_k\, Z_{n-k} + \sum_{k=0}^{m} a_k\, Y_{n-k} \\ a_0 Y_n &= -\sum_{k=1}^{m} a_k\, Y_{n-k} + \sum_{k=0}^{p} b_k\, Z_{n-k} \end{aligned}$$

Autoregressive, moving Average (ARMA) sequence of order (m, p). For $p = 0$, we have an Autoregressive (AR) and for $m = 0$ we have a moving average (MA) sequence.

Example 3 (Wide-Sense Markov sequences) *A simple and useful model for the correlation structure of a stationary random sequence is the so-called* wide-sense Markov model*:*

$$C_Y(n) = \sigma^2 r^{|n|}, \quad n \in \boldsymbol{Z} \tag{9.38}$$

where $|r| < 1$. The power spectrum of such a sequence is

$$S_Y(\omega) = \frac{\sigma^2(1 - r^2)}{1 - 2r\cos(\omega) + r^2}$$

which is a rational fraction, and we can see easily that we can write it

$$S_Y(\omega) = \frac{\sigma^2(1-r^2)}{(1 - r\exp[-j\omega])\,(1 - r\exp[+j\omega])} = \frac{1}{A(\exp[-j\omega])\,A(\exp[+j\omega])} = \frac{1}{A(Z)\,A(z^{-1})}$$

where

$$A(z) = a_0 + a_1 z^{-1}$$

with $a_0 = \sqrt{\sigma^2(1-r^2)}$, $a_1 = -r\,a_0$.

We can conclude here that a wide-sense Markov sequence with the covariance structure (9.38) is an AR(1) sequence.

Example 4 (Prediction of a Wide-Sense Markov sequences) *Consider now the prediction problem where we wish to predict $Y_{n+\lambda}$ from the sequence $Y_k{}_{k=-\infty}^{n}$. Using the relations we obtained in the last section, this can be done through a causal filter whose transfer function is*

$$Y_n \longrightarrow \boxed{H(\omega) = \frac{1}{S_Y^+(\omega)} \left[\frac{S_{XY}(\omega)}{S_Y^-(\omega)}\right]_+} \longrightarrow \widehat{Y}_{t+\lambda}$$

Here we have

$$H(\omega) = A(\exp[j\omega]) \left[\frac{\exp[j\omega\lambda]}{A(\exp[j\omega])}\right]_+$$

Using the following geometric series relations

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}, \quad \textit{and} \quad \sum_{k=1}^{\infty} x^k = \frac{x}{1-x}, \quad |x| < 1$$

we obtain easily

$$\frac{1}{A(z)} = \frac{1}{a_0} \sum_{k=0}^{\infty} r^n \, z^{-1}$$

and

$$\begin{aligned}
\left[\frac{\exp\left[j\omega\lambda\right]}{A(\exp\left[j\omega\right])}\right]_+ &= \left[\frac{1}{a_0} \sum_{n=0}^{\infty} r^n \, \exp\left[-j\omega(n-\lambda)\right]\right]_+ \\
&= \frac{1}{a_0} \sum_{n=\lambda}^{\infty} r^n \, \exp\left[-j\omega(n-\lambda)\right] \\
&= \frac{1}{a_0} \sum_{l=0}^{\infty} r^{l+\lambda} \, \exp\left[-jl\omega\right] \\
&= \frac{r^\lambda}{A(\exp\left[j\omega\right])}
\end{aligned}$$

Finally, we obtain

$$H(\omega) = A(\exp\left[j\omega\right]) \, \frac{r^\lambda}{A(\exp\left[j\omega\right])} = r^\lambda$$

which is a pure pure gain

$$\widehat{Y}_{t+\lambda} = r^\lambda \, Y_t$$

It is also easy to show that, in this case we have

$$MSE = E\left[(\widehat{Y}_{t+\lambda} - Y_{t+\lambda})^2\right] = \sigma^2(1 - r^{2\lambda})$$

which means that the prediction error increases monotonically from $\sigma^2(1-r^{2\lambda})$ to σ^2 as λ increases from 1 to ∞.

Appendix A

Annexes

A.1 Summary of Bayesian inference

Observation model:

$$X_i \sim f(x_i|\boldsymbol{\theta})$$

$$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad \boldsymbol{z}_n = \{x_1, \cdots, x_n\}, \quad \boldsymbol{z}_{n+1} = \{x_1, \cdots, x_n, x_{n+1}\},$$

Likelihood and sufficient statistics:

$$l(\boldsymbol{\theta}|\boldsymbol{z}) = f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

$$l(\boldsymbol{\theta}|\boldsymbol{z}) = l(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{z}))$$

$$p(\boldsymbol{t}(\boldsymbol{z})|\boldsymbol{\theta}) = \frac{l(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{z}))}{\displaystyle\int l(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{z}))\,\mathrm{d}\boldsymbol{\theta}}$$

Inference with any prior law:

$$\pi(\boldsymbol{\theta})$$

$$f(x_i) = \int f(x_i|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \quad \text{and} \quad f(\boldsymbol{x}) = \int f(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

$$p(\boldsymbol{t}(\boldsymbol{z})) = \int p(\boldsymbol{t}(\boldsymbol{z})|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

$$p(\boldsymbol{z}, \boldsymbol{\theta}) = p(\boldsymbol{z}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$$

$$p(\boldsymbol{z}) = \int p(\boldsymbol{z}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}, \qquad \text{prior predictive}$$

$$\pi(\boldsymbol{\theta}|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|\boldsymbol{\theta})\,\pi(\theta)}{p(\boldsymbol{z})} \quad \mathrm{E}\left[\boldsymbol{\theta}|\boldsymbol{z}\right] = \int \boldsymbol{\theta}\,\pi(\boldsymbol{\theta}|\boldsymbol{z})\,\mathrm{d}\boldsymbol{\theta}$$

$$f(x|\boldsymbol{z}) = \frac{p(\boldsymbol{z}, x)}{p(\boldsymbol{z})} = \frac{p(\boldsymbol{z}_{n+1})}{p(\boldsymbol{z}_n)}, \qquad \text{posterior predictive}$$

$$\mathrm{E}\left[x|\boldsymbol{z}\right] = \int x\, f(x|\boldsymbol{z})\,\mathrm{d}x$$

Inference with conjugate priors:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}_0) = \frac{p(\boldsymbol{t} = \boldsymbol{\tau}_0|\boldsymbol{\theta})}{\displaystyle\int p(\boldsymbol{t} = \boldsymbol{\tau}_0|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \in \mathcal{F}_{\boldsymbol{\tau}_0}(\boldsymbol{\theta})$$

$$\pi(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{\tau}) \in \mathcal{F}_{\boldsymbol{\tau}}(\boldsymbol{\theta}), \quad \text{with} \quad \boldsymbol{\tau} = g(\boldsymbol{\tau}_0, n, \boldsymbol{z})$$

Inference with conjugate priors and generalized exponential family:

if $f(x_i|\boldsymbol{\theta}) = a(x_i)\, g(\boldsymbol{\theta}) \exp\left[\sum_{k=1}^{K} c_k \phi_k(\boldsymbol{\theta})\, h_k(x_i)\right]$

then

$$t_k(\boldsymbol{x}) = \sum_{j=1}^{n} h_k(x_j), \quad k = 1, \cdots, K$$

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}_0) = [g(\boldsymbol{\theta})]^{\tau_0}\, z(\boldsymbol{\tau})\, \exp\left[\sum_{k=1}^{K} \tau_k \phi_k(\boldsymbol{\theta})\right]$$

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\tau}) = [g(\boldsymbol{\theta})]^{n+\tau_0}\, a(\boldsymbol{x})\, Z(\boldsymbol{\tau})\, \exp\left[\sum_{k=1}^{K} c_k \phi_k(\boldsymbol{\theta})\, (\tau_k + t_k(\boldsymbol{x}))\right].$$

Inference with conjugate priors and natural exponential family:

if $f(x_i|\boldsymbol{\theta}) = a(x_i) \exp[\theta x_i - b(\theta)]$

then

$$t(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$$

$$\pi(\theta|\tau_0) = c(\theta) \exp[\tau_0 \theta - d(\tau_0)]$$

$$\pi(\theta|\boldsymbol{x}, \tau_0) = c(\theta) \exp[\tau_n \theta - d(\tau_n)] \quad \text{with} \quad \tau_n = \tau_0 + \bar{x}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$

Inference with conjugate priors and natural exponential family Multivariable case:

if $f(\boldsymbol{x}_i|\boldsymbol{\theta}) = a(\boldsymbol{x}_i) \exp\left[\boldsymbol{\theta}^t \boldsymbol{x}_i - b(\boldsymbol{\theta})\right]$

then

$$t_k(\boldsymbol{x}) = \sum_{i=1}^{n} x_{ki}, \quad k = 1, \ldots, K$$

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}_0) = c(\boldsymbol{\theta}) \exp[\boldsymbol{\tau}_0 \boldsymbol{\theta} - d(\boldsymbol{\tau}_0)]$$

$$\pi(\theta|\boldsymbol{x}, \tau_0) = c(\theta) \exp[\boldsymbol{\tau}_n \boldsymbol{\theta} - d(\boldsymbol{\tau}_n)] \quad \text{with} \quad \boldsymbol{\tau}_n = \boldsymbol{\tau}_0 + \bar{\boldsymbol{x}}$$

where $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i,$

Bernouilli model:

$z = \{x_1, \cdots, x_n\}, \quad x_i \in \{0, 1\}, \quad r = \sum x_i$: number of 1, $\quad n - r$: number of 0

$f(x_i|\theta) = \mathbf{Ber}(x_i|\theta), \quad 0 < \theta < 1$

Likelihood and sufficient statistics:

$l(\theta|z) = \prod_{i=1}^{n} \mathbf{Ber}(x_i|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} = \theta^r(1-\theta)^{n-r}$

$t(z) = r = \sum_{i=1}^{n} x_i, \qquad l(\theta|r) = \theta^r(1-\theta)^{1-r}$

$p(r|\theta) = \mathbf{Bin}(r|\theta, n)$

Inference with conjugate priors:

$\pi(\theta) = \mathbf{Bet}(\theta|\alpha, \beta)$

$f(x) = \mathbf{BinBet}(x|\alpha, \beta, 1)$

$p(r) = \mathbf{BinBet}(r|\alpha, \beta, n)$

$\pi(\theta|z) = \mathbf{Bet}(\theta|\alpha + r, \beta + n - r), \qquad \mathrm{E}\left[\theta|z\right] = \dfrac{\alpha + r}{\beta + n - r}$

$f(x|z) = \mathbf{BinBet}(x|\alpha + r, \beta + n - r, 1), \qquad \mathrm{E}\left[x|z\right] = \dfrac{\alpha + r}{\beta + n - r}$

Inference with reference priors:

$\pi(\theta) = \mathbf{Bet}(\theta|\frac{1}{2}, \frac{1}{2})$

$\pi(x) = \mathbf{BinBet}(x|\frac{1}{2}, \frac{1}{2}, 1)$

$\pi(r) = \mathbf{BinBet}(r|\frac{1}{2}, \frac{1}{2}, n)$

$\pi(\theta|z) = \mathbf{Bet}(\theta|\frac{1}{2} + r, \frac{1}{2} + n - r)$

$\pi(x|z) = \mathbf{BinBet}(x|\frac{1}{2} + r, \frac{1}{2} + n - r, 1)$

Binomial model:

$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad x_i = 0, 1, 2, \cdots, m$

$f(x_i|\theta, m) = \mathbf{Bin}(x_i|\theta, m), \quad 0 < \theta < 1, \quad m = 0, 1, 2, \cdots$

Likelihood and sufficient statistics:

$$l(\theta|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{Bin}(x_i|\theta, m)$$

$$t(\boldsymbol{z}) = r = \sum_{i=1}^{n} x_i$$

$p(r|\theta) = \mathbf{Bin}(r|\theta, nm)$

Inference with conjugate priors:

$\pi(\theta) = \mathbf{Bet}(\theta|\alpha, \beta)$

$f(x) = \mathbf{BinBet}(x|\alpha, \beta, m)$

$p(r) = \mathbf{BinBet}(r|\alpha, \beta, nm)$

$$\pi(\theta|\boldsymbol{z}) = \mathbf{Bet}(\theta|\alpha + r, \beta + n - r), \qquad \mathrm{E}[\theta|\boldsymbol{z}] = \frac{\alpha + r}{\beta + n - r}$$

$f(x|\boldsymbol{z}) = \mathbf{BinBet}(x|\alpha + r, \beta + n - r, m)$

Inference with reference priors:

$\pi(\theta) = \mathbf{Bet}(\theta|\frac{1}{2}, \frac{1}{2})$

$\pi(x) = \mathbf{BinBet}(x|\frac{1}{2}, \frac{1}{2}, 1)$

$\pi(r) = \mathbf{BinBet}(r|\frac{1}{2}, \frac{1}{2}, n)$

$\pi(\theta|\boldsymbol{z}) = \mathbf{Bet}(\theta|\frac{1}{2} + r, \frac{1}{2} + n - r)$

$\pi(x|\boldsymbol{z}) = \mathbf{BinBet}(x|\frac{1}{2} + r, \frac{1}{2} + n - r, m)$

Poisson:

$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad x_i = 0, 1, 2, \cdots$

$f(x_i|\lambda) = \mathbf{Pn}(x_i|\lambda), \quad \lambda \geq 0$

Likelihood and sufficient statistics:

$$l(\lambda|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{Pn}(x_i|\lambda)$$

$$t(\boldsymbol{z}) = r = \sum_{i=1}^{n} x_i$$

$p(r|\lambda) = \mathbf{Pn}(r|n\lambda)$

Inference with conjugate priors:

$p(\lambda) = \mathbf{Gam}(\lambda|\alpha, \beta)$

$f(x) = \mathbf{PnGam}(x|\alpha, \beta, 1)$

$p(r) = \mathbf{PnGam}(r|\alpha, \beta, n)$

$$p(\lambda|\boldsymbol{z}) = \mathbf{Gam}(\lambda|\alpha + r, \beta + n), \qquad \mathrm{E}[\lambda|\boldsymbol{z}] = \frac{\alpha + r}{\beta + n}$$

$f(x|\boldsymbol{z}) = \mathbf{PnGam}(x|\alpha + r, \beta + n, 1)$

Inference with reference priors:

$$\pi(\lambda) \propto \lambda^{-1/2} = \mathbf{Gam}(\lambda|\frac{1}{2}, 0)$$

$$\pi(x) = \mathbf{PnGam}(x|\frac{1}{2}, 0, 1)$$

$$\pi(r) = \mathbf{PnGam}(r|\frac{1}{2}, 0, n)$$

$$\pi(\lambda|\boldsymbol{z}) = \mathbf{Gam}(\lambda|\frac{1}{2} + r, n)$$

$$\pi(x|\boldsymbol{z}) = \mathbf{PnGam}(x|\frac{1}{2} + r, n, 1)$$

Negative Binomial model:

$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad x_i = 0, 1, 2, \cdots$

$f(x_i|\theta, r) = \mathbf{NegBin}(x_i|\theta, r), \quad 0 < \theta < 0,\ r = 1, 2, \cdots$

Likelihood and sufficient statistics:

$l(\theta|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{NegBin}(x_i|\theta, r)$

$t(\boldsymbol{z}) = s = \sum_{i=1}^{n} x_i$

$p(s|\theta) = \mathbf{NegBin}(s|\theta, nr)$

Inference with conjugate priors:

$\pi(\theta) = \mathbf{Bet}(\theta|\alpha, \beta)$

$f(x) = \mathbf{NegBinBet}(x|\alpha, \beta, r)$

$p(s) = \mathbf{NegBinBet}(s|\alpha, \beta, nr)$

$\pi(\theta|\boldsymbol{z}) = \mathbf{Bet}(\theta|\alpha + nr, \beta + s), \qquad \mathrm{E}\,[\theta|\boldsymbol{z}] = \dfrac{\alpha + nr}{\beta + s}$

$f(x|\boldsymbol{z}) = \mathbf{NegBinBet}(x|\alpha + nr, \beta + s, nr)$

Inference with reference priors:

$\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1/2} = \mathbf{Bet}(\theta|0, \frac{1}{2})$

$\pi(x) = \mathbf{NegBinBet}(x|0, \frac{1}{2}, r)$

$\pi(s) = \mathbf{NegBinBet}(s|0, \frac{1}{2}, nr)$

$\pi(\theta|\boldsymbol{z}) = \mathbf{Bet}(\theta|nr, s + \frac{1}{2})$

$\pi(x|\boldsymbol{z}) = \mathbf{NegBinBet}(x|nr, s + \frac{1}{2}, nr)$

Exponential model:

$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad 0 < x_i < \infty$

$f(x_i|\lambda) = \mathbf{Ex}(x_i|\lambda), \quad \lambda > 0$

Likelihood and sufficient statistics:

$l(\lambda|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{Ex}(x_i|\lambda)$

$t(\boldsymbol{z}) = t = \sum_{i=1}^{n} x_i$

$p(t|\lambda) = \mathbf{Gam}(t|n, \lambda)$

Inference with conjugate priors:

$p(\lambda) = \mathbf{Gam}(\lambda|\alpha, \beta)$

$f(x) = \mathbf{GamGam}(x|\alpha, \beta, 1)$

$p(t) = \mathbf{GamGam}(t|\alpha, \beta, n)$

$p(\lambda|\boldsymbol{z}) = \mathbf{Gam}(\lambda|\alpha + n, \beta + t) \qquad \mathrm{E}[\lambda|\boldsymbol{z}] = \dfrac{\alpha + n}{\beta + t}$

$f(x|\boldsymbol{z}) = \mathbf{GamGam}(x|\alpha + n, \beta + t, 1)$

Inference with reference priors:

$\pi(\lambda) \propto \lambda^{-1} = \mathbf{Gam}(\lambda|0, 0)$

$\pi(x) = \mathbf{GamGam}(x|0, 0, 1)$

$\pi(t) = \mathbf{GamGam}(t|0, 0, n)$

$\pi(\lambda|\boldsymbol{z}) = \mathbf{Gam}(\lambda|n, t)$

$\pi(x|\boldsymbol{z}) = \mathbf{GamGam}(x|n, t, 1)$

Uniform model:

$$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad 0 < x_i < \theta$$

$$f(x_i|\theta) = \mathbf{Uni}(x_i|0, \theta), \quad \theta > 0$$

Likelihood and sufficient statistics:

$$l(\theta|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{Uni}(x_i|0, \theta)$$

$$t(\boldsymbol{z}) = t = \max\{x_1, \cdots, x_n\}$$

$$p(t|\theta) = \mathbf{IPar}(t|n, \theta^{-1})$$

Inference with conjugate priors:

$$\pi(\theta) = \mathbf{Par}(\theta|\alpha, \beta)$$

$$f(x) = \begin{cases} \frac{\alpha}{\alpha+1}\mathbf{Uni}(x|0, \beta), & \textbf{if} \;\; x \leq \beta, \\ \frac{1}{\alpha+1}\mathbf{Par}(x|\alpha, \beta), & \textbf{if} \;\; x > \beta \end{cases}$$

$$p(t) = \begin{cases} \frac{\alpha}{\alpha+n}\mathbf{IPar}(t|n, \beta^{-1}), & \textbf{if} \;\; t \leq \beta, \\ \frac{n}{\alpha+n}\mathbf{Par}(t|\alpha, \beta), & \textbf{if} \;\; x > \beta \end{cases}$$

$$\pi(\theta|\boldsymbol{z}) = \mathbf{Par}(\theta|\alpha + n, \beta_n), \quad \beta_n = \max\{\beta, t\}$$

$$f(x|\boldsymbol{z}) = \begin{cases} \frac{\alpha+n}{\alpha+n+1}\mathbf{Uni}(x|0, \beta_n), & \textbf{if} \;\; t \leq \beta_n, \\ \frac{1}{\alpha+n+1}\mathbf{Par}(x|\alpha, \beta_n), & \textbf{if} \;\; x > \beta_n \end{cases}$$

Inference with reference priors:

$$\pi(\theta) \propto \theta^{-1} = \mathbf{Par}(\theta|0, 0)$$

$$\pi(\theta|\boldsymbol{z}) = \mathbf{Par}(\theta|n, t)$$

$$\pi(x|\boldsymbol{z}) = \begin{cases} \frac{n}{n+1}\mathbf{Uni}(x|0, t), & \textbf{if} \;\; x \leq t, \\ \dfrac{1}{n+1}\mathbf{Par}(x|n, t), & \textbf{if} \;\; x > t \end{cases}$$

Normal with known precision $\lambda = \frac{1}{\sigma^2} > 0$
(Estimation of μ):

$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad x_i \in \mathbb{R}, \quad x_i = \mu + b_i, \quad b_i \sim \mathbf{N}(b_i|0, \lambda)$
$f(x_i|\mu, \lambda) = \mathbf{N}(x_i|\mu, \lambda), \quad \mu \in \mathbb{R}$

Likelihood and sufficient statistics:

$$l(\mu|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}(x_i|\mu, \lambda)$$
$$t(\boldsymbol{z}) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$p(\bar{x}|\mu, \lambda) = \mathbf{N}(\bar{x}|\mu, n\lambda)$$

Inference with conjugate priors:

$$p(\mu) = \mathbf{N}(\mu|\mu_0, \lambda_0)$$
$$f(x) = \mathbf{N}\left(x|\mu_0, \frac{\lambda\lambda_0}{\lambda + \lambda_0}\right)$$
$$f(\boldsymbol{x}) = f(x_1, \cdots, x_n) = \mathbf{N}_n\left(\boldsymbol{x}|\mu_0\mathbf{1}, \left(\frac{1}{\lambda}\boldsymbol{I} + \frac{1}{\lambda_0}\mathbf{1}.\mathbf{1}^t\right)^{-1}\right)$$

$$p(\bar{x}) = \mathbf{N}\left(\bar{x}|\mu_0, \tfrac{n\lambda\lambda_0}{\lambda_n}\right), \quad \lambda_n = \lambda_0 + n\lambda$$
$$p(\mu|\boldsymbol{z}) = \mathbf{N}\left(\mu|\mu_n, \lambda_n\right), \quad \mu_n = \frac{\lambda_0\mu_0 + n\lambda\bar{x}}{\lambda_n}$$
$$f(x|\boldsymbol{z}) = \mathbf{N}\left(x|\mu_n, \tfrac{\lambda\lambda_n}{\lambda+\lambda_n}\right)$$

Inference with reference priors:

$$\pi(\mu) = \mathbf{constant}$$
$$\pi(\mu|\boldsymbol{z}) = \mathbf{N}(\mu|\bar{x}, n\lambda)$$
$$\pi(x|\boldsymbol{z}) = \mathbf{N}\left(x|\bar{x}, \frac{n\lambda}{n+1}\right)$$

Inference with other prior laws:

$\pi(\mu) = \mathbf{St}(\mu|0, \tau^2, \alpha) = \pi_1(\mu|\rho)\pi_2(\rho|\alpha)$
with

$\quad \pi_1(\mu|\rho) = \mathbf{N}(\mu|0, \tau^2\rho),$
$\quad \pi_2(\rho|\alpha) = \mathbf{IGam}(\rho|\alpha/2, \alpha/2),$

$$\pi(\mu|\boldsymbol{z}, \rho) = \mathbf{N}\left(\mu|\frac{1}{1+\tau^2\rho}\bar{x}, \frac{\tau^2\rho}{1+\tau^2\rho}\right)$$
$$\pi(\rho|\boldsymbol{z}) \propto (1+\tau^2\rho)^{-1/2}\exp\left[\frac{-1}{2(1+\tau^2\rho)}\boldsymbol{x}^t\boldsymbol{x}\right]\pi_2(\rho)$$

Normal with known variance $\sigma^2 > 0$
(Estimation of μ):

$$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad x_i \in \mathbb{R}, \quad x_i = \mu + b_i, \quad b_i \sim \mathbf{N}(b_i|0, \sigma^2)$$
$$f(\boldsymbol{x}) = f(x_i|\mu, \sigma^2) = \mathbf{N}(x_i|\mu, \sigma^2), \quad \mu \in \mathbb{R}$$

Likelihood and sufficient statistics:

$$l(\mu|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}(x_i|\mu, \sigma^2)$$
$$t(\boldsymbol{z}) = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
$$p(\bar{x}|\mu, \sigma^2) = \mathbf{N}(\bar{x}|\mu, \frac{1}{n}\sigma^2)$$

Inference with conjugate priors:

$$p(\mu) = \mathbf{N}(\mu|\mu_0, \sigma_0^2)$$
$$f(x) = \mathbf{N}\left(x|\mu_0, \sigma_0^2 + \sigma^2\right)$$
$$f(x_1, \cdots, x_n) = \mathbf{N}_n\left(\boldsymbol{x}|\mu_0\mathbf{1}, \sigma^2\boldsymbol{I} + \sigma_0^2\mathbf{1}.\mathbf{1}^t\right)$$
$$p(\bar{x}) = \mathbf{N}\left(\bar{x}|\mu_0, \sigma_0^2 + \frac{1}{n}\sigma^2\right),$$
$$p(\mu|\boldsymbol{z}) = \mathbf{N}\left(\mu|\mu_n, \sigma_n^2\right), \quad \mu_n = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}\left(\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}\right), \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$
$$f(x|\boldsymbol{z}) = \mathbf{N}\left(x|\mu_n, \sigma^2 + \sigma_n^2\right)$$

Inference with reference priors:

$$\pi(\mu) = \textbf{constant}$$
$$\pi(\mu|\boldsymbol{z}) = \mathbf{N}(\mu|\bar{x}, \frac{1}{n}\sigma^2)$$
$$\pi(x|\boldsymbol{z}) = \mathbf{N}\left(x|\bar{x}, \frac{n+1}{n\sigma^2}\right)$$

Inference with other prior laws:

$$\pi(\mu) = \mathbf{St}(\mu|0, \tau^2, \alpha) = \pi_1(\mu|\rho)\pi_2(\rho|\alpha)$$

with

$$\pi_1(\mu|\rho) = \mathbf{N}(\mu|0, \tau^2\rho),$$
$$\pi_2(\rho|\alpha) = \mathbf{IGam}(\rho|\alpha/2, \alpha/2),$$
$$\pi(\mu|\boldsymbol{z}, \rho) = \mathbf{N}\left(\mu|\frac{1}{1+\tau^2\rho}\bar{x}, \frac{\tau^2\rho}{1+\tau^2\rho}\right)$$
$$\pi(\rho|\boldsymbol{z}) \propto (1+\tau^2\rho)^{-1/2} \exp\left[\frac{-1}{2(1+\tau^2\rho)}\boldsymbol{x}^t\boldsymbol{x}\right] \pi_2(\rho)$$

Normal with known mean $\mu \in \mathbb{R}$
(Estimation of λ):

$$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad x_i \in \mathbb{R}, \quad x_i = \mu + b_i, \quad b_i \sim \mathbf{N}(b_i|0, \lambda)$$
$$f(x_i|\mu, \lambda) = \mathbf{N}(x_i|\mu, \lambda), \quad \lambda > 0$$

Likelihood and sufficient statistics:

$$l(\lambda|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}(x_i|\mu, \lambda)$$
$$t(\boldsymbol{z}) = t = \sum_{i=1}^{n} (x_i - \mu)^2$$
$$p(t|\mu, \lambda) = \mathbf{Gam}(t|\frac{n}{2}, \lambda/2), \quad p(\lambda t|\mu, \lambda) = \mathbf{Chi}^2(\lambda t|n)$$

Inference with conjugate priors:

$$p(\lambda) = \mathbf{Gam}(\lambda|\alpha, \beta)$$
$$f(x) = \mathbf{St}\,(x|\mu, \alpha/\beta, 2\alpha)$$
$$p(t) = \mathbf{GamGam}\left(t|\alpha, 2\beta, \frac{n}{2}\right)$$
$$p(\lambda|\boldsymbol{z}) = \mathbf{Gam}\left(\lambda|\alpha + \frac{n}{2}, \beta + \frac{t}{2}\right)$$
$$f(x|\boldsymbol{z}) = \mathbf{St}\left(x|\mu, \tfrac{\alpha+\frac{n}{2}}{\beta+\frac{t}{2}}, 2\alpha + n\right)$$

Inference with reference priors:

$$\pi(\lambda) \propto \lambda^{-1} = \mathbf{Gam}(\lambda|0, 0)$$
$$\pi(\lambda|\boldsymbol{z}) = \mathbf{Gam}(\lambda|\frac{n}{2}, \frac{t}{2})$$
$$\pi(x|\boldsymbol{z}) = \mathbf{St}\left(x|\mu, \frac{n}{t}, n\right)$$

Normal with known mean $\mu \in \mathbb{R}$
(Estimation of σ^2):

$$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad x_i \in \mathbb{R}, \quad x_i = \mu + b_i, \quad b_i \sim \mathbf{N}(b_i|0, \sigma^2)$$
$$f(x_i|\mu, \sigma^2) = \mathbf{N}(x_i|\mu, \sigma^2), \quad \sigma^2 > 0$$

Likelihood and sufficient statistics:

$$l(\sigma^2|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}(x_i|\mu, \sigma^2)$$
$$t(\boldsymbol{z}) = t = \sum_{i=1}^{n} (x_i - \mu)^2$$
$$p(t|\mu, \sigma^2) = \mathbf{Gam}\left(t|\frac{n}{2}, \frac{\sigma^2}{2}\right), \quad p(\frac{t}{\sigma^2}|\mu, \sigma^2) = \mathbf{Chi}^2\left(\frac{t}{\sigma^2}|n\right)$$

Inference with conjugate priors:

$$p(\sigma^2) = \mathbf{IGam}(\sigma^2|\alpha, \beta)$$
$$f(x) = \mathbf{St}\,(x|\mu, \alpha/\beta, 2\alpha)$$
$$p(t) = \mathbf{GamGam}\left(t|\alpha, 2\beta, \frac{n}{2}\right)$$
$$p(\sigma^2|\boldsymbol{z}) = \mathbf{IGam}\left(\sigma^2|\alpha + \frac{n}{2}, \beta + \frac{t}{2}\right)$$
$$f(x|\boldsymbol{z}) = \mathbf{St}\left(x|\mu, \frac{\alpha+\frac{n}{2}}{\beta+\frac{t}{2}}, 2\alpha + n\right)$$

Inference with reference priors:

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} = \mathbf{IGam}(\sigma^2|0, 0)$$
$$\pi(\sigma^2|\boldsymbol{z}) = \mathbf{IGam}\left(\sigma^2|\frac{n}{2}, \frac{t}{2}\right)$$
$$\pi(x|\boldsymbol{z}) = \mathbf{St}\left(x|\mu, \frac{n}{t}, n\right)$$

Normal with both unknown parameters
Estimation of mean and precision (μ, λ)**:**

$$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad x_i \in \mathbb{R}, \quad x_i = \mu + b_i, \quad b_i \sim \mathbf{N}(b_i|0, \lambda)$$
$$f(x_i|\mu, \lambda) = \mathbf{N}(x_i|\mu, \lambda), \quad \mu \in \mathbb{R}, \, \lambda > 0$$

Likelihood and sufficient statistics:

$$l(\mu, \lambda|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}(x_i|\mu, \lambda)$$
$$t(\boldsymbol{z}) = (\bar{x}, s), \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$p(\bar{x}|\mu, \lambda) = \mathbf{N}(\bar{x}|\mu, n\lambda),$$
$$p(ns^2|\mu, \lambda) = \mathbf{Gam}(ns^2|(n-1)/2, \lambda/2), \quad p(\lambda ns^2|\mu, \lambda) = \mathbf{Chi}^2(\lambda ns^2|n-1)$$

Inference with conjugate priors:

$$p(\mu, \lambda) = \mathbf{NGam}(\mu, \lambda|\mu_0, n_0, \alpha, \beta) = \mathbf{N}(\mu|\mu_0, n_0\lambda)\, \mathbf{Gam}(\lambda|\alpha, \beta)$$
$$p(\mu) = \mathbf{St}\left(\mu|\mu_0, n_0\frac{\alpha}{\beta}, 2\alpha\right)$$
$$p(\lambda) = \mathbf{Gam}(\lambda|\alpha, \beta)$$
$$f(x) = \mathbf{St}\left(x|\mu_0, \frac{n_0}{n_0+1}\frac{\alpha}{\beta}, 2\alpha\right)$$
$$p(\bar{x}) = \mathbf{St}\left(\bar{x}|\mu_0, \frac{n_0 n}{n_0+n}\frac{\alpha}{\beta}, 2\alpha\right)$$
$$p(ns^2) = \mathbf{GamGam}\left(ns^2|\alpha, 2\beta, \frac{n-1}{2}\right)$$
$$p(\mu|\boldsymbol{z}) = \mathbf{St}\left(\mu|\mu_n, (n+n_0)(\alpha_n)\beta_n^{-1}, 2\alpha_n\right),$$
$$\quad \alpha_n = \alpha + \tfrac{n}{2},$$
$$\quad \mu_n = \tfrac{n_0\mu_0 + n\bar{x}}{n_0+n},$$
$$\quad \beta_n = \beta + ns^2/2 + \tfrac{1}{2}\tfrac{n_0 n}{n_0+n}(\mu_0 - \bar{x})^2$$
$$p(\lambda|\boldsymbol{z}) = \mathbf{Gam}\left(\lambda|\alpha_n, \beta_n\right)$$
$$f(x|\boldsymbol{z}) = \mathbf{St}\left(x|\mu_n, \tfrac{n+n_0}{n+n_0+1}\tfrac{\alpha_n}{\beta_n}, 2\alpha_n\right)$$

Inference with reference priors:

$$\pi(\mu, \lambda) = \pi(\lambda, \mu) \propto \lambda^{-1}, \quad n > 1$$
$$\pi(\mu|\boldsymbol{z}) = \mathbf{St}(\mu|\bar{x}, (n-1)s^2, n-1)$$
$$\pi(\lambda|\boldsymbol{z}) = \mathbf{Gam}(\lambda|(n-1)/2, ns^2/2)$$
$$\pi(x|\boldsymbol{z}) = \mathbf{St}\left(x|\bar{x}, \frac{n-1}{n+1}s^{-2}, n-1\right)$$

Normal with both unknown parameters mean and variance
Estimation of (μ, σ^2)**:**

$$\boldsymbol{z} = \{x_1, \cdots, x_n\}, \quad x_i \in \mathbb{R}, \quad x_i = \mu + b_i, \quad b_i \sim \mathbf{N}(b_i|0, \sigma^2)$$
$$f(x_i|\mu, \sigma^2) = \mathbf{N}(x_i|\mu, \sigma^2), \quad \mu \in \mathbb{R},\ \sigma^2 > 0$$

Likelihood and sufficient statistics:

$$l(\mu, \sigma^2|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}(x_i|\mu, \sigma^2)$$
$$t(\boldsymbol{z}) = (\bar{x}, s), \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$p(\bar{x}|\mu, \sigma^2) = \mathbf{N}(\bar{x}|\mu, \frac{1}{n}\sigma^2),$$
$$p(ns^2|\mu, \sigma^2) = \mathbf{Gam}(ns^2|(n-1)/2, \tfrac{\sigma^2}{2}), \quad p(\sigma^2 ns^2|\mu, \sigma^2) = \mathbf{Chi}^2(\sigma^2 ns^2|n-1)$$

Inference with conjugate priors:

$$p(\mu, \sigma^2) = \mathbf{NIGam}(\mu, \sigma^2|\mu_0, n_0, \alpha, \beta) = \mathbf{N}(\mu|\mu_0, n_0\sigma^2)\,\mathbf{IGam}(\sigma^2|\alpha, \beta)$$
$$p(\mu) = \mathbf{St}\left(\mu|\mu_0, n_0\frac{\alpha}{\beta}, 2\alpha\right)$$
$$p(\sigma^2) = \mathbf{IGam}(\sigma^2|\alpha, \beta)$$
$$f(x) = \mathbf{St}\left(x|\mu_0, \frac{n_0}{n_0+1}\frac{\alpha}{\beta}, 2\alpha\right)$$
$$p(\bar{x}) = \mathbf{St}\left(\bar{x}|\mu_0, \frac{n_0 n}{n_0+n}\frac{\alpha}{\beta}, 2\alpha\right)$$
$$p(ns^2) = \mathbf{GamGam}\left(ns^2|\alpha, 2\beta, \frac{n-1}{2}\right)$$
$$p(\mu|\boldsymbol{z}) = \mathbf{St}\left(\mu|\mu_n, (n+n_0)(\alpha_n)\beta_n^{-1}, 2\alpha_n\right),$$
$$\alpha_n = \alpha + \tfrac{n}{2},$$
$$\mu_n = \tfrac{n_0\mu_0 + n\bar{x}}{n_0+n},$$
$$\beta_n = \beta + ns^2/2 + \tfrac{1}{2}\tfrac{n_0 n}{n_0+n}(\mu_0 - \bar{x})^2$$
$$p(\sigma^2|\boldsymbol{z}) = \mathbf{IGam}\left(\sigma^2|\alpha_n, \beta_n\right)$$
$$f(x|\boldsymbol{z}) = \mathbf{St}\left(x|\mu_n, \tfrac{n+n_0}{n+n_0+1}\tfrac{\alpha_n}{\beta_n}, 2\alpha_n\right)$$

Inference with reference priors:

$$\pi(\mu, \sigma^2) = \pi(\sigma^2, \mu) \propto \frac{1}{\sigma^2}, \quad n > 1$$
$$\pi(\mu|\boldsymbol{z}) = \mathbf{St}(\mu|\bar{x}, (n-1)s^2, n-1)$$
$$\pi(\sigma^2|\boldsymbol{z}) = \mathbf{IGam}(\sigma^2|(n-1)/2, ns^2/2)$$
$$\pi(x|\boldsymbol{z}) = \mathbf{St}\left(x|\bar{x}, \frac{n-1}{n+1}s^{-2}, n-1\right)$$

Multinomial:

$$\boldsymbol{z} = \{r_1, \cdots, r_k, n\}, \quad r_i = 0, 1, 2, \cdots, \quad \sum_{i=1}^{k} r_i \leq n,$$

$p(r_i|\theta_i, n) = \mathbf{Bin}(r_i|\theta_i, n),$

$p(\boldsymbol{z}|\boldsymbol{\theta}, n) = \mathbf{Mu}_k(\boldsymbol{z}|\boldsymbol{\theta}, n), \quad 0 < \theta_i < 1, \quad \sum_{i=1}^{k} \theta_i \leq 1$

Likelihood and sufficient statistics:

$l(\theta|\boldsymbol{z}) = \mathbf{Mu}_k(\boldsymbol{z}|\boldsymbol{\theta}, n)$

$t(\boldsymbol{z}) = (\boldsymbol{r}, n), \quad \boldsymbol{r} = \{r_1, \cdots, r_k\}$

$p(\boldsymbol{r}|\boldsymbol{\theta}) = \mathbf{Mu}_k(\boldsymbol{r}|\boldsymbol{\theta}, n)$

Inference with conjugate priors:

$\pi(\boldsymbol{\theta}) = \mathbf{Di}_k(\boldsymbol{\theta}|\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_{k+1}\}$

$p(\boldsymbol{r}) = \mathbf{Mu}_k(\boldsymbol{r}|\boldsymbol{\alpha}, n)$

$$\pi(\boldsymbol{\theta}|\boldsymbol{z}) = \mathbf{Di}_k\left(\boldsymbol{\theta}|\alpha_1 + r_1, \cdots, \alpha_k + r_k, \alpha_{k+1} + n - \sum_{i=1}^{k} r_k\right)$$

$$f(\boldsymbol{x}|\boldsymbol{z}) = \mathbf{Di}_k\left(\boldsymbol{\theta}|\alpha_1 + r_1, \cdots, \alpha_k + r_k, \alpha_{k+1} + n - \sum_{i=1}^{k} r_k\right)$$

Inference with reference priors:

$\pi(\boldsymbol{\theta}) \propto ??$

$\pi(\boldsymbol{\theta}|\boldsymbol{z}) = ??$

$\pi(\boldsymbol{x}|\boldsymbol{z}) = ??$

Multi-variable Normal with known precision matrix $\mathbf{\Lambda}$ (Estimation of the mean $\boldsymbol{\mu}$):

$\boldsymbol{z} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}, \quad \boldsymbol{x}_i \in \mathbb{R}^k, \quad \boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{b}_i, \quad \boldsymbol{b}_i \sim \mathbf{N}_k(\boldsymbol{b}_i|\mathbf{0}, \mathbf{\Lambda})$

$f(\boldsymbol{x}_i|\boldsymbol{\mu}, \mathbf{\Lambda}) = \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \mathbf{\Lambda}), \quad \boldsymbol{\mu} \in \mathbb{R}^k$, $\mathbf{\Lambda}$ matrix d.p. of dimensions $k \times k$

Likelihood and sufficient statistics:

$$l(\boldsymbol{\mu}|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \mathbf{\Lambda})$$

$$t(\boldsymbol{z}) = \bar{\boldsymbol{x}}, \quad \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i,$$

$$p(\bar{\boldsymbol{x}}|\boldsymbol{\mu}, \mathbf{\Lambda}) = \mathbf{N}_k(\bar{\boldsymbol{x}}|\boldsymbol{\mu}, n\mathbf{\Lambda})$$

Inference with conjugate priors:

$$p(\boldsymbol{\mu}) = \mathbf{N}_k(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \mathbf{\Lambda}_0)$$

$$f(\boldsymbol{x}) = \mathbf{N}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_0, (\mathbf{\Lambda}_0\mathbf{\Lambda})\mathbf{\Lambda}_1^{-1}\right), \quad \mathbf{\Lambda}_1 = \mathbf{\Lambda}_0 + \mathbf{\Lambda}$$

$$p(\boldsymbol{\mu}|\boldsymbol{z}) = \mathbf{N}_k\left(\boldsymbol{\mu}|\boldsymbol{\mu}_n, \mathbf{\Lambda}_n\right)$$

$$\mathbf{\Lambda}_n = \mathbf{\Lambda}_0 + n\mathbf{\Lambda},$$

$$\boldsymbol{\mu}_n = \mathbf{\Lambda}_n^{-1}(\mathbf{\Lambda}_0\boldsymbol{\mu}_0 + n\mathbf{\Lambda}\bar{\boldsymbol{x}})$$

$$f(\boldsymbol{x}|\boldsymbol{z}) = \mathbf{N}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_n, \mathbf{\Lambda}_n\right)$$

Inference with reference priors:

$\pi(\boldsymbol{\mu}) =??$

$\pi(\boldsymbol{\mu}|\boldsymbol{z}) =??$

$\pi(\boldsymbol{x}|\boldsymbol{z}) =??$

Multi-variable Normal with known covariance matrix $\boldsymbol{\Sigma}$ (Estimation of the mean $\boldsymbol{\mu}$):

$\boldsymbol{z} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}, \quad \boldsymbol{x}_i \in \mathbb{R}^k, \quad \boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{b}_i, \quad \boldsymbol{b}_i \sim \mathbf{N}_k(\boldsymbol{b}_i|\mathbf{0}, \boldsymbol{\Sigma})$

$f(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} \in \mathbb{R}^k$, $\boldsymbol{\Sigma}$ p.d. matrix of dimensions $k \times k$

Likelihood and sufficient statistics:

$$l(\boldsymbol{\mu}|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$t(\boldsymbol{z}) = \bar{\boldsymbol{x}}, \quad \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i,$$

$$p(\bar{\boldsymbol{x}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbf{N}_k(\bar{\boldsymbol{x}}|\boldsymbol{\mu}, n\boldsymbol{\Sigma})$$

Inference with conjugate priors:

$p(\boldsymbol{\mu}) = \mathbf{N}_k(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$

$f(\boldsymbol{x}) = \mathbf{N}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_1\right), \qquad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}$

$p(\boldsymbol{\mu}|\boldsymbol{z}) = \mathbf{N}_k\left(\boldsymbol{\mu}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right)$

$\quad \boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma},$

$\quad \boldsymbol{\mu}_n = \boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\Sigma}_0\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}\bar{\boldsymbol{x}})$

$f(\boldsymbol{x}|\boldsymbol{z}) = \mathbf{N}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right)$

Inference with reference priors:

$\pi(\boldsymbol{\mu}) =??$

$\pi(\boldsymbol{\mu}|\boldsymbol{z}) =??$

$\pi(\boldsymbol{x}|\boldsymbol{z}) =??$

Multi-variable Normal with known mean $\boldsymbol{\mu}$
(Estimation of precision matrix $\boldsymbol{\Lambda}$):

$\boldsymbol{z} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}, \quad \boldsymbol{x}_i \in \mathbb{R}^k, \quad \boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{b}_i, \quad \boldsymbol{b}_i \sim \mathbf{N}_k(\boldsymbol{b}_i|\mathbf{0}, \boldsymbol{\Lambda})$

$f(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\lambda}), \quad \boldsymbol{\mu} \in \mathbb{R}^k$, $\boldsymbol{\lambda}$ matrix d.p. of dimensions $k \times k$

Likelihood and sufficient statistics:

$$l(\boldsymbol{\lambda}|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\lambda})$$

$$t(\boldsymbol{z}) = \boldsymbol{S}, \quad \boldsymbol{S} = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^t$$

$$p(\boldsymbol{S}|\boldsymbol{\lambda}) = \mathbf{Wi}_k(\boldsymbol{S}|(n-1)/2, \boldsymbol{\lambda}/2),$$

Inference with conjugate priors:

$$p(\boldsymbol{\lambda}) = \mathbf{Wi}_k(\boldsymbol{\lambda}|\alpha, \boldsymbol{\beta})$$

$$f(\boldsymbol{x}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_0, \frac{n_0}{n_0+1}(\alpha - \frac{k-1}{2})\boldsymbol{\beta}^{-1}, 2\alpha - k + 1\right)$$

$$p(\boldsymbol{\lambda}|\boldsymbol{z}) = \mathbf{Wi}_k\left(\boldsymbol{\lambda}|\alpha_n, \boldsymbol{\beta}_n\right)$$

$$\alpha_n = \alpha + \tfrac{n}{2} - \tfrac{k-1}{2},$$

$$\boldsymbol{\mu}_n = \tfrac{n_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{n_0+n},$$

$$\boldsymbol{\beta}_n = \boldsymbol{\beta} + \tfrac{1}{2}\boldsymbol{S} + \tfrac{1}{2}\tfrac{n_0 n}{n_0+n}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})^t$$

$$f(\boldsymbol{x}|\boldsymbol{z}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_n, \tfrac{n+n_0}{n+n_0+1}\alpha_n\boldsymbol{\beta}_n^{-1}, 2\alpha_n\right)$$

Inference with reference priors:

$\pi(\boldsymbol{\lambda}) =??$

$\pi(\boldsymbol{\lambda}|\boldsymbol{z}) =??$

$\pi(\boldsymbol{x}|\boldsymbol{z}) =??$

Multi-variable Normal with known mean $\boldsymbol{\mu}$
(Estimation of covariance matrix $\boldsymbol{\Sigma}$):

$$\boldsymbol{z} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}, \quad \boldsymbol{x}_i \in \mathbb{R}^k, \quad \boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{b}_i, \quad \boldsymbol{b}_i \sim \mathbf{N}_k(\boldsymbol{b}_i|\mathbf{0}, \boldsymbol{\Sigma})$$

$f(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} \in \mathbb{R}^k$, $\boldsymbol{\Sigma}$ matrix d.p. of dimensions $k \times k$

Likelihood and sufficient statistics:

$$l(\boldsymbol{\Sigma}|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$t(\boldsymbol{z}) = \boldsymbol{S}, \quad \boldsymbol{S} = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^t$$

$$p(\boldsymbol{S}|\boldsymbol{\Sigma}) = \mathbf{Wi}_k(\boldsymbol{S}|(n-1)/2, \boldsymbol{\Sigma}/2),$$

Inference with conjugate priors:

$$p(\boldsymbol{\Sigma}) = \mathbf{IWi}_k(\boldsymbol{\Sigma}|\alpha, \boldsymbol{\beta})$$

$$f(\boldsymbol{x}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_0, \frac{n_0}{n_0+1}(\alpha - \frac{k-1}{2})\boldsymbol{\beta}^{-1}, 2\alpha - k + 1\right)$$

$$p(\boldsymbol{\Sigma}|\boldsymbol{z}) = \mathbf{IWi}_k\left(\boldsymbol{\Sigma}|\alpha_n, \boldsymbol{\beta}_n\right)$$

$$\alpha_n = \alpha + \tfrac{n}{2} - \tfrac{k-1}{2},$$

$$\boldsymbol{\mu}_n = \tfrac{n_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{n_0+n},$$

$$\boldsymbol{\beta}_n = \boldsymbol{\beta} + \tfrac{1}{2}\boldsymbol{S} + \tfrac{1}{2}\tfrac{n_0 n}{n_0+n}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})^t$$

$$f(\boldsymbol{x}|\boldsymbol{z}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_n, \tfrac{n+n_0}{n+n_0+1}\alpha_n\boldsymbol{\beta}_n^{-1}, 2\alpha_n\right)$$

Inference with reference priors:

$\pi(\boldsymbol{\Sigma}) =??$

$\pi(\boldsymbol{\Sigma}|\boldsymbol{z}) =??$

$\pi(\boldsymbol{x}|\boldsymbol{z}) =??$

Multi-variable Normal with both unknown parameters
Estimation of mean and precision matrix $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$:

$\boldsymbol{z} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}, \quad \boldsymbol{x}_i \in \mathbb{R}^k,$
$\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{b}_i, \quad \boldsymbol{b}_i \sim \mathbf{N}_k(\boldsymbol{b}_i|\mathbf{0}, \boldsymbol{\Lambda}) \quad \boldsymbol{\mu} \sim \mathbf{N}_k(\boldsymbol{\mu}|\boldsymbol{\mu}_0, n_0\boldsymbol{\Lambda}) \quad \boldsymbol{\Lambda} \sim \mathbf{Wi}_k(\boldsymbol{\Lambda}|\alpha, \boldsymbol{\beta})$
$f(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Lambda}), \quad \boldsymbol{\mu} \in \mathbb{R}^k,$ $\boldsymbol{\Lambda}$ matrix d.p. of dimensions $k \times k$

Likelihood and sufficient statistics:

$$l(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Lambda})$$
$$t(\boldsymbol{z}) = (\bar{\boldsymbol{x}}, \boldsymbol{S}), \quad \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \quad \boldsymbol{S} = \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t$$
$$p(\bar{\boldsymbol{x}}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathbf{N}_k(\bar{\boldsymbol{x}}|\boldsymbol{\mu}, n\boldsymbol{\Lambda})$$
$$p(\boldsymbol{S}|\boldsymbol{\Lambda}) = \mathbf{Wi}_k(\boldsymbol{S}|(n-1)/2, \boldsymbol{\Lambda}/2),$$

Inference with conjugate priors:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathbf{NWi}_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, n_0, \alpha, \boldsymbol{\beta}) = \mathbf{N}_k(\boldsymbol{\mu}|\boldsymbol{\mu}_0, n_0\boldsymbol{\Lambda})\mathbf{Wi}_k(\boldsymbol{\Lambda}|\alpha, \boldsymbol{\beta})$$
$$p(\boldsymbol{\mu}) = \mathbf{St}_k\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, n_0\alpha\boldsymbol{\beta}^{-1}, 2\alpha\right) \quad ??$$
$$p(\boldsymbol{\Lambda}) = \mathbf{Wi}_k(\boldsymbol{\Lambda}|\alpha, \boldsymbol{\beta}) \quad ??$$
$$f(\boldsymbol{x}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_0, \frac{n_0}{n_0+1}(\alpha - \frac{k-1}{2})\boldsymbol{\beta}^{-1}, 2\alpha - k + 1\right)$$
$$p(\boldsymbol{\mu}|\boldsymbol{z}) = \mathbf{St}_k\left(\boldsymbol{\mu}|\boldsymbol{\mu}_n, (n+n_0)\alpha_n\boldsymbol{\beta}_n^{-1}, 2\alpha_n\right)$$
$$\alpha_n = \alpha + \tfrac{n}{2} - \tfrac{k-1}{2},$$
$$\boldsymbol{\mu}_n = \tfrac{n_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{n_0+n},$$
$$\boldsymbol{\beta}_n = \boldsymbol{\beta} + \tfrac{1}{2}\boldsymbol{S} + \tfrac{1}{2}\tfrac{n_0 n}{n_0+n}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})^t$$
$$p(\boldsymbol{\Lambda}|\boldsymbol{z}) = \mathbf{Wi}_k\left(\boldsymbol{\Lambda}|\alpha_n, \boldsymbol{\beta}_n\right)$$
$$f(\boldsymbol{x}|\boldsymbol{z}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_n, \tfrac{n+n_0}{n+n_0+1}\alpha_n\boldsymbol{\beta}_n^{-1}, 2\alpha_n\right)$$

Inference with reference priors:

$\pi(\boldsymbol{\mu}, \boldsymbol{\Lambda}) =??$
$\pi(\boldsymbol{\mu}|\boldsymbol{z}) =??$
$\pi(\boldsymbol{\Lambda}|\boldsymbol{z}) =??$
$\pi(\boldsymbol{x}|\boldsymbol{z}) =??$

Mult-variable Normal with both unknown parameters
Estimation of mean and covariance matrix $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$\boldsymbol{z} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}, \quad \boldsymbol{x}_i \in \mathbb{R}^k, \quad \boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{b}_i, \quad \boldsymbol{b}_i \sim \mathbf{N}_k(\boldsymbol{b}_i|\mathbf{0}, \boldsymbol{\Sigma})$
$f(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} \in \mathbb{R}^k$, $\boldsymbol{\Sigma}$ matrix d.p. of dimensions $k \times k$

Likelihood and sufficient statistics:

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{z}) = \prod_{i=1}^{n} \mathbf{N}_k(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$t(\boldsymbol{z}) = (\bar{\boldsymbol{x}}, \boldsymbol{S}), \quad \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \quad \boldsymbol{S} = \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t$$
$$p(\bar{\boldsymbol{x}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbf{N}_k(\bar{\boldsymbol{x}}|\boldsymbol{\mu}, n\boldsymbol{\Sigma})$$
$$p(\boldsymbol{S}|\boldsymbol{\Sigma}) = \mathbf{Wi}_k(\boldsymbol{S}|(n-1)/2, \boldsymbol{\Sigma}/2),$$

Inference with conjugate priors:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbf{NWi}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\mu}_0, n_0, \alpha, \boldsymbol{\beta}) = \mathbf{N}_k(\boldsymbol{\mu}|\boldsymbol{\mu}_0, n_0\boldsymbol{\Sigma})\mathbf{Wi}_k(\boldsymbol{\Sigma}|\alpha, \boldsymbol{\beta})$$
$$p(\boldsymbol{\mu}) = \mathbf{St}_k\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, n_0\alpha\boldsymbol{\beta}^{-1}, 2\alpha\right) \quad ??$$
$$p(\boldsymbol{\Sigma}) = \mathbf{IWi}_k(\boldsymbol{\Sigma}|\alpha, \boldsymbol{\beta}) \quad ??$$
$$f(\boldsymbol{x}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_0, \frac{n_0}{n_0+1}(\alpha - \frac{k-1}{2})\boldsymbol{\beta}^{-1}, 2\alpha - k + 1\right)$$
$$p(\boldsymbol{\mu}|\boldsymbol{z}) = \mathbf{St}_k\left(\boldsymbol{\mu}|\boldsymbol{\mu}_n, (n+n_0)\alpha_n\boldsymbol{\beta}_n^{-1}, 2\alpha_n\right)$$
$$\alpha_n = \alpha + \tfrac{n}{2} - \tfrac{k-1}{2},$$
$$\boldsymbol{\mu}_n = \tfrac{n_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{n_0+n},$$
$$\boldsymbol{\beta}_n = \boldsymbol{\beta} + \tfrac{1}{2}\boldsymbol{S} + \tfrac{1}{2}\tfrac{n_0 n}{n_0+n}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})^t$$
$$p(\boldsymbol{\Sigma}|\boldsymbol{z}) = \mathbf{Wi}_k\left(\boldsymbol{\Sigma}|\alpha_n, \boldsymbol{\beta}_n\right)$$
$$f(\boldsymbol{x}|\boldsymbol{z}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{\mu}_n, \tfrac{n+n_0}{n+n_0+1}\alpha_n\boldsymbol{\beta}_n^{-1}, 2\alpha_n\right)$$

Inference with reference priors:

$\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) =??$
$\pi(\boldsymbol{\mu}|\boldsymbol{z}) =??$
$\pi(\boldsymbol{\Sigma}|\boldsymbol{z}) =??$
$\pi(\boldsymbol{x}|\boldsymbol{z}) =??$

Linear regression:

$$\boldsymbol{z} = (\boldsymbol{y}, \boldsymbol{X}),\ \boldsymbol{y} = \{y_1, \cdots, y_n\} \in \mathbb{R}^n,$$
$$\boldsymbol{x}_i = \{\boldsymbol{x}_{i_1}, \cdots, \boldsymbol{x}_{i_k}\} = \{x_{i1}, \cdots, x_{ik}\} \in \mathbb{R}^k,\ \boldsymbol{X} = (x_{i,j})$$
$$\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_k\} \in \mathbb{R}^k,\ y_i = \boldsymbol{x}_i^t \boldsymbol{\theta} = \boldsymbol{\theta}^t \boldsymbol{x}_i$$
$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}, \lambda) = \mathbf{N}_n(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\theta}, \lambda \boldsymbol{I}_n), \quad \boldsymbol{\theta} \in \mathbb{R}^k,\ \lambda > 0$$

Likelihood and sufficient statistics:

$$l(\boldsymbol{\theta}|\boldsymbol{z}) = \mathbf{N}_n(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\theta}, \lambda \boldsymbol{I}_n)$$
$$t(\boldsymbol{z}) = (\boldsymbol{X}^t \boldsymbol{X}, \boldsymbol{X}^t \boldsymbol{y})$$

Inference with conjugate priors:

$$\pi(\boldsymbol{\theta}, \lambda) = \mathbf{NGam}_k(\boldsymbol{\theta}, \lambda|\boldsymbol{\theta}_0, \boldsymbol{\Lambda}_0, \alpha, \beta) = \mathbf{N}_k(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \lambda\boldsymbol{\Lambda}_0)\mathbf{Gam}(\lambda|\alpha, \beta)$$
$$\pi(\boldsymbol{\theta}|\lambda) = \mathbf{N}_k(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \lambda\boldsymbol{\Lambda}_0), \quad \mathrm{E}\left[\boldsymbol{\theta}|\lambda\right] = \boldsymbol{\theta}_0, \quad \mathrm{Var}\left[\boldsymbol{\theta}|\lambda\right] = (\lambda\boldsymbol{\Lambda}_0)^{-1}$$
$$\pi(\lambda|\alpha, \beta) = \mathbf{Gam}(\lambda|\alpha, \beta)$$
$$\pi(\boldsymbol{\theta}) = \mathbf{St}_k\left(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \frac{\alpha}{\beta}\boldsymbol{\Lambda}_0, 2\alpha\right), \quad \mathrm{E}\left[\boldsymbol{\theta}\right] = \boldsymbol{\theta}_0, \quad \mathrm{Var}\left[\boldsymbol{\theta}\right] = \frac{\alpha}{\alpha - 2}\boldsymbol{\Lambda}_0^{-1}$$
$$p(y_i|\boldsymbol{x}_i) = \mathbf{St}\left(y_i|\boldsymbol{x}_i^t\boldsymbol{\theta}_0, \frac{\alpha}{\beta}f(\boldsymbol{x}_i), 2\alpha\right), \quad \text{with} \quad f(\boldsymbol{x}_i) = 1 - \boldsymbol{x}_i^t(\boldsymbol{\Lambda}_0 + \boldsymbol{x}_i\boldsymbol{x}_i^t)^{-1}\boldsymbol{x}_i,$$

$$\pi(\boldsymbol{\theta}, \lambda|\boldsymbol{z}) = \mathbf{NGam}_k(\boldsymbol{\theta}, \lambda|\boldsymbol{\theta}_n, \boldsymbol{\Lambda}_n, \alpha_n, \beta_n) = \mathbf{N}_k(\boldsymbol{\theta}|\boldsymbol{\theta}_n, \lambda\boldsymbol{\Lambda}_n)\mathbf{Gam}(\lambda|\alpha_n, \beta_n)$$
$$\pi(\boldsymbol{\theta}|\boldsymbol{z}) = \mathbf{St}_k\left(\boldsymbol{\theta}|\boldsymbol{\theta}_n, (\boldsymbol{\Lambda}_0 + \boldsymbol{X}^t\boldsymbol{X})\frac{\alpha_n}{\beta_n}, 2\alpha_n\right)$$
$$\alpha_n = \alpha + \tfrac{n}{2},$$
$$\boldsymbol{\theta}_n = (\boldsymbol{\Lambda}_0 + \boldsymbol{X}^t\boldsymbol{X})^{-1}(\boldsymbol{\Lambda}_0\boldsymbol{\theta}_0 + \boldsymbol{X}^t\boldsymbol{y}) = (\boldsymbol{I} - \boldsymbol{\Lambda}_n)\boldsymbol{\theta}_0 + \boldsymbol{\Lambda}_n\widetilde{\boldsymbol{\theta}},$$
$$\beta_n = \beta + \tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}^t\boldsymbol{\theta}_n)^t\boldsymbol{y} + \tfrac{1}{2}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_n)^t\boldsymbol{\Lambda}_0\boldsymbol{\theta}_0 = \beta + \tfrac{1}{2}\boldsymbol{y}^t\boldsymbol{y} + \tfrac{1}{2}\boldsymbol{\theta}_0^t\boldsymbol{\Lambda}_0\boldsymbol{\theta}_0 - \tfrac{1}{2}\boldsymbol{\theta}_n\boldsymbol{\Lambda}_n\boldsymbol{\theta}_n$$
$$\widetilde{\boldsymbol{\theta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}, \quad \boldsymbol{\Lambda}_n = (\boldsymbol{\Lambda}_0 + \boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{X}$$

$$\mathrm{E}\left[\boldsymbol{\theta}|\boldsymbol{z}\right] = \boldsymbol{\theta}_n, \quad \mathrm{Var}\left[\boldsymbol{\theta}|\boldsymbol{z}\right] = (\boldsymbol{\Lambda}_0 + \boldsymbol{X}^t\boldsymbol{X})^{-1}$$
$$\pi(\lambda|\boldsymbol{z}) = \mathbf{Gam}\left(\lambda|\alpha_n, \beta_n\right)$$
$$p(y_i|\boldsymbol{x}_i, \boldsymbol{z}) = \mathbf{St}\left(y_i|\boldsymbol{x}_i^t\boldsymbol{\theta}_n, f_n(\boldsymbol{x}_i)\tfrac{\alpha_n}{\beta_n}, 2\alpha_n\right)$$
$$f_n(\boldsymbol{x}_i) = 1 - \boldsymbol{x}_i^t(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{\Lambda}_0 + \boldsymbol{x}_i\boldsymbol{x}_i^t)^{-1}\boldsymbol{x}_i,$$

Inference with reference priors:

$$\pi(\boldsymbol{\theta}, \lambda) = \pi(\lambda, \boldsymbol{\theta}) \propto \lambda^{-(k+1)/2}$$
$$\pi(\boldsymbol{\theta}|\boldsymbol{z}) = \mathbf{St}_k\left(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}_n, \frac{n-k}{2\widehat{\beta}_n}\boldsymbol{X}^t\boldsymbol{X}, n-k\right)$$
$$\widetilde{\boldsymbol{\theta}}_n = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y},$$
$$\widehat{\beta}_n = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}^t\widetilde{\boldsymbol{\theta}}_n)^t(\boldsymbol{y} - \boldsymbol{X}^t\widetilde{\boldsymbol{\theta}}_n)$$
$$\pi(\lambda|\boldsymbol{z}) = \mathbf{Gam}\left(\lambda|\frac{n-k}{2}, \widehat{\beta}_n\right)$$
$$p(y_i|\boldsymbol{x}_i, \boldsymbol{z}) = \mathbf{St}\left(y_i|\boldsymbol{x}_i^t\widetilde{\boldsymbol{\theta}}_n, \frac{n-k}{2\widehat{\beta}_n}f_n(\boldsymbol{x}_i), n-k\right),$$
$$f_n(\boldsymbol{x}_i) = 1 - \boldsymbol{x}_i^t(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{x}_i\boldsymbol{x}_i^t)^{-1}\boldsymbol{x}_i,$$

Inverse problems:

$\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{b}$, $\boldsymbol{z} = \{z_1, \cdots, z_n\} \in \mathbb{R}^n$, $\boldsymbol{h}_i = \{h_{i_1}, \cdots, h_{i_k}\} \in \mathbb{R}^k$, $\boldsymbol{H} = (h_{i,j})$

$\boldsymbol{x} = \{x_1, \cdots, x_k\} \in \mathbb{R}^k$,

$p(\boldsymbol{z}|\boldsymbol{H}, \boldsymbol{x}, \lambda) = \mathbf{N}_n(\boldsymbol{z}|\boldsymbol{H}\boldsymbol{x}, \lambda \boldsymbol{I}_n), \quad \boldsymbol{x} \in \mathbb{R}^k,\ \lambda > 0$

Likelihood and sufficient statistics:

$l(\boldsymbol{x}|\boldsymbol{z}) = \mathbf{N}_n(\boldsymbol{z}|\boldsymbol{H}\boldsymbol{x}, \lambda \boldsymbol{I}_n)$

$t(\boldsymbol{z}) = (\boldsymbol{H}^t\boldsymbol{z}, \boldsymbol{H}^t\boldsymbol{x}^t\boldsymbol{x}\boldsymbol{H})$

Inference with conjugate priors:

$\pi(\boldsymbol{x}, \lambda) = \mathbf{NGam}_k(\boldsymbol{x}, \lambda|\boldsymbol{x}_0, \boldsymbol{\Lambda}_0, \alpha, \beta) = \mathbf{N}_k(\boldsymbol{x}|\boldsymbol{x}_0, \lambda\boldsymbol{\Lambda}_0)\mathbf{Gam}(\lambda|\alpha, \beta)$

$\pi(\boldsymbol{x}|\lambda) = \mathbf{N}_k(\boldsymbol{x}|\boldsymbol{x}_0, \lambda\boldsymbol{\Lambda}_0), \quad \mathrm{E}\left[\boldsymbol{x}|\lambda\right] = \boldsymbol{x}_0, \quad \mathrm{Var}\left[\boldsymbol{x}|\lambda\right] = (\lambda\boldsymbol{\Lambda}_0)^{-1}$

$f(\boldsymbol{x}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{x}_0, \dfrac{\alpha}{\beta}\boldsymbol{\Lambda}_0, 2\alpha\right)$

$\pi(\lambda|\alpha, \beta) = \mathbf{Gam}(\lambda|\alpha, \beta)$

$\pi(\boldsymbol{x}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{x}_0, \dfrac{\alpha}{\beta}\boldsymbol{\Lambda}_0, 2\alpha\right), \quad \mathrm{E}\left[\boldsymbol{x}\right] = \boldsymbol{x}_0, \quad \mathrm{Var}\left[\boldsymbol{x}\right] = \dfrac{\alpha}{\alpha - 2}\boldsymbol{\Lambda}_0^{-1}$

$\pi(\lambda|\alpha, \beta) = \mathbf{Gam}(\lambda|\alpha, \beta)$

$p(z_i|\boldsymbol{x}) = \mathbf{St}\left(z_i|\boldsymbol{x}^t\boldsymbol{x}_0, \dfrac{\alpha}{\beta}f(\boldsymbol{x}), 2\alpha\right)$

$\quad f(\boldsymbol{x}) = 1 - \boldsymbol{x}^t(\boldsymbol{\Lambda}_0 + \boldsymbol{x}^t\boldsymbol{x})^{-1}\boldsymbol{x}$

$\pi(\boldsymbol{x}, \lambda|\boldsymbol{z}) = \mathbf{NGam}_k(\boldsymbol{x}, \lambda|xb_n, \boldsymbol{\Lambda}_n, \alpha_n, \beta_n) = \mathbf{N}_k(\boldsymbol{x}|\boldsymbol{x}_n, \lambda\boldsymbol{\Lambda}_n)\mathbf{Gam}(\lambda|\alpha_n, \beta_n)$

$f(\boldsymbol{x}|\boldsymbol{z}) = \mathbf{St}_k\left(\boldsymbol{x}|\boldsymbol{x}_n, (\boldsymbol{\Lambda}_0 + \boldsymbol{H}^t\boldsymbol{H})\dfrac{\alpha_n}{\beta_n}, 2\alpha_n\right)$

$\quad \alpha_n = \alpha + \frac{n}{2},$

$\quad \boldsymbol{x}_n = (\boldsymbol{\Lambda}_0 + \boldsymbol{H}^t\boldsymbol{H})^{-1}(\boldsymbol{\Lambda}_0\boldsymbol{x}_0 + \boldsymbol{H}^t\boldsymbol{z}) = (\boldsymbol{I} - \boldsymbol{\Lambda}_n)\boldsymbol{x}_0 + \boldsymbol{\Lambda}_n\widetilde{\boldsymbol{\theta}}$

$\quad \beta_n = \beta + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{H}^t\boldsymbol{x}_n)^t\boldsymbol{z} + \frac{1}{2}(\boldsymbol{x}_0 - \boldsymbol{x}_n)^t\boldsymbol{\Lambda}_0\boldsymbol{x}_0 = \beta + \frac{1}{2}\boldsymbol{z}^t\boldsymbol{z} + \frac{1}{2}\boldsymbol{x}_0^t\boldsymbol{\Lambda}_0\boldsymbol{x}_0 - \frac{1}{2}\boldsymbol{x}_n^t\boldsymbol{\Lambda}_n\boldsymbol{\theta}_n$

$\quad \widetilde{\boldsymbol{x}} = (\boldsymbol{H}^t\boldsymbol{H})^{-1}\boldsymbol{H}^t\boldsymbol{y}, \quad \boldsymbol{\Lambda}_n = (\boldsymbol{\Lambda}_0 + \boldsymbol{H}^t\boldsymbol{H})^{-1}\boldsymbol{H}^t\boldsymbol{H}$

$\quad \mathrm{E}\left[\boldsymbol{x}|\boldsymbol{z}\right] = \boldsymbol{x}_n, \quad \mathrm{Var}\left[\boldsymbol{x}|\boldsymbol{z}\right] = (\boldsymbol{\Lambda}_0 + \boldsymbol{H}^t\boldsymbol{H})^{-1}$

$\pi(\lambda|\boldsymbol{z}) = \mathbf{Gam}\left(\lambda|\alpha_n, \beta_n\right)$

$p(z_i|\boldsymbol{h}_i, \boldsymbol{z}) = \mathbf{St}\left(z_i|\boldsymbol{h}_i^t\boldsymbol{z}_n, f_n(\boldsymbol{h}_i)\frac{\alpha_n}{\beta_n}, 2\alpha_n\right)$

$\quad f_n(\boldsymbol{h}_i) = 1 - \boldsymbol{h}_i^t(\boldsymbol{H}^t\boldsymbol{H} + \boldsymbol{\Lambda}_0 + \boldsymbol{h}_i\boldsymbol{h}_i^t)^{-1}\boldsymbol{h}_i,$

Inference with reference priors:

$\pi(\boldsymbol{x}, \lambda) = \pi(\lambda, \boldsymbol{x}) \propto \lambda^{-(k+1)/2}$

$\pi(\boldsymbol{x}|\boldsymbol{z}) = \mathbf{St}_k\left(\boldsymbol{x}|\widehat{\boldsymbol{x}}_n, \dfrac{n-k}{2\widehat{\beta}_n}\boldsymbol{H}^t\boldsymbol{H}, n-k\right)$

$\quad \widehat{\boldsymbol{x}}_n = (\boldsymbol{H}^t\boldsymbol{H})^{-1}\boldsymbol{H}^t\boldsymbol{z},$

$\quad \widehat{\beta}_n = \dfrac{1}{2}(\boldsymbol{z} - \boldsymbol{H}^t\widehat{\boldsymbol{x}}_n)^t\boldsymbol{z}$

$\pi(\lambda|\boldsymbol{z}) = \mathbf{Gam}\left(\lambda|\dfrac{n-k}{2}, \widehat{\beta}_n\right)$

$p(z_i|\boldsymbol{h}_i, \boldsymbol{z}) = \mathbf{St}\left(z_i|\boldsymbol{h}_i^t\boldsymbol{z}_n, f_n(\boldsymbol{h}_i)\frac{\alpha_n}{\beta_n}, 2\alpha_n\right)$

$\quad f_n(\boldsymbol{h}_i) = 1 - \boldsymbol{h}_i^t(\boldsymbol{H}^t\boldsymbol{H} + \boldsymbol{\Lambda}_0 + \boldsymbol{h}_i\boldsymbol{h}_i^t)^{-1}\boldsymbol{h}_i,$

A.2 Summary of probability distributions

Probability laws for discrete random variables

Bernoulli	$\mathbf{Ber}\,(x\vert\theta) = \theta^x\,(1-\theta)^{1-x}$, $0<\theta<1$, $x=\{0,1\}$
Binomial	$\mathbf{Bin}\,(x\vert\theta,n) = c\,\theta^x\,(1-\theta)^{n-x}$, $c=\binom{n}{x}$, $0<\theta<1,\ n=1,2,\cdots$, $x=0,1,\cdots,n$
Hypergeometric	$\mathbf{HypGeo}\,(x\vert N,M,n) = c\binom{N}{x}\binom{M}{n-x}$, $c=\binom{N+M}{n}^{-1}$, $N,M=1,2,\cdots,\ n=1,\cdots,N+M$, $x=a,a+1,\cdots,b,$ avec $a=\max\{0,n-M\},\ b=\min\{N,n\}$
Negative-Binomial	$\mathbf{NegBin}\,(x\vert\theta,r) = c\binom{r+x-1}{r-1}(1-\theta)^x$, $c=\theta^r$, $0<\theta<1,\ r=1,2,\cdots$, $x=0,1,2,\cdots$
Poisson	$\mathbf{Pn}\,(x\vert\lambda) = \frac{\lambda^x}{x!}$, $c=\exp\left[-\lambda\right]$, $\lambda>0$, $x=0,1,2,\cdots$
Binomial-Beta	$\mathbf{BinBet}\,(x\vert\alpha,\beta,n) = c\binom{n}{x}\Gamma(\alpha+x)\Gamma(\beta+n-x)$, $c=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}$, $\alpha,\beta>0,\ n=1,2,\cdots$, $x=0,\cdots,n$

Probability laws for discrete random variables (cont.)

Negative Binomial-Beta	$\mathbf{NegBinBet}\,(x\mid\alpha,\beta,r) = c \binom{r+x-1}{r-1} \frac{\Gamma(\beta+x)}{\Gamma(\alpha+\beta+x+\alpha)}$, $c = \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+r)}{\Gamma(\alpha)\Gamma(\beta)}$, $\alpha,\beta > 0,\ r = 1,2,\cdots,$ $x = 0,1,2,\cdots$
Poisson-Gamma	$\mathbf{PnGam}\,(x\mid\alpha,\beta,n) = c\,\frac{\Gamma(\alpha+x)}{x\,!}\frac{n^x}{(\beta+n)^{\alpha+x}}$, $c = \frac{\beta^\alpha}{\Gamma(\alpha)}$, $\alpha,\beta > 0,\ n = 0,1,2,\cdots,$ $x = 0,1,2,\cdots$
Composite Poisson	$\mathbf{Pnc}\,(x\mid\lambda,\mu) = \exp\left[-\lambda\right]\sum_{n=0}^{\infty}\frac{(n\mu)^x\exp\left[-n\mu\right]}{x!}\frac{\lambda^N}{n!}$, $\lambda,\mu > 0, \quad x = 0,1,2\cdots$
Geometric	$\mathbf{Geo}\,(x\mid\theta) = c\,(1-\theta)^{x-1}$, $c = \theta$, $\theta > 0, \quad x = 0,1,2\cdots$
Pascal	$\mathbf{Pas}\,(x\mid m,\theta) = C_{x-1}^{m-1}\theta^m(1-\theta)^{x-m}$, $m > 0, 0 < \theta < 1 \quad x = 0,1,2\cdots$

Probability laws for real random variables

Beta	$\mathbf{Bet}\,(x\mid\alpha,\beta) = c\,x^{\alpha-1}(1-x)^{\beta-1}$, $c = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$, $\alpha,\beta > 0$, $0 < x < 1$
Gamma	$\mathbf{Gam}\,(x\mid\alpha,\beta) = c\,x^{\alpha-1}\exp\left[-\beta x\right]$, $c = \dfrac{\beta^\alpha}{\Gamma(\alpha)}$, $\alpha,\beta > 0$, $x > 0$
Inverse Gamma	$\mathbf{IGam}\,(x\mid\alpha,\beta) = c\,x^{-(\alpha+1)}\exp\left[-\dfrac{\beta}{x}\right]$, $c = \dfrac{\beta^\alpha}{\Gamma(\alpha)}$, $\alpha,\beta > 0$, $x > 0$
Gamma–Gamma	$\mathbf{GamGam}\,(x\mid\alpha,\beta,n) = c\,\dfrac{x^{n-1}}{(\beta+x)^{\alpha+n}}$, $c = \dfrac{\beta^\alpha}{\Gamma(\alpha)}\dfrac{\Gamma(\alpha+n)}{\Gamma(n)}$, $\alpha,\beta > 0,\ n = 0,1,2,\cdots$, $x = 0,1,2,\cdots$
Pareto	$\mathbf{Par}\,(x\mid\alpha,\beta) = c\,x^{-(\alpha+1)}$, $c = \alpha\beta^\alpha$, $\alpha,\beta > 0$, $x \geq \beta$
Normal	$\mathbf{N}\,(x\mid\mu,\lambda) = c\,\exp\left[-\dfrac{1}{2}\lambda(x-\mu)^2\right]$, $c = \sqrt{\dfrac{\lambda}{2\pi}}$, $\mu \in \mathbb{R},\ \lambda > 0$, $x \in \mathbb{R}$
Normal	$\mathbf{N}\,(x\mid\mu,\sigma) = c\,\exp\left[-\dfrac{1}{2\sigma^2}(x-\mu)^2\right]$, $c = \dfrac{1}{\sqrt{2\pi\sigma^2}}$, $\mu \in \mathbb{R},\ \sigma > 0$, $x \in \mathbb{R}$

Probability laws for real random variables (cont.)

Logistic	$\mathbf{Lo}\,(x \mid \alpha, \beta) = c\,\frac{\exp\left[-\beta^{-1}(x-\alpha)\right]}{\left(1+\exp\left[\beta^{-1}(x-\alpha)\right]\right)^2}$, $c = \beta^{-1}$, $\alpha \in \mathbb{R}$, $\beta > 0$, $x \in \mathbb{R}$
Student (t)	$\mathbf{St}\,(x \mid \mu, \lambda, \alpha) = c\,\left[1 + \frac{\lambda}{\alpha}(x-\mu)^2\right]^{-(\alpha+1)/2}$, $c = \frac{\Gamma(\frac{\alpha+1}{2})}{\Gamma(\alpha/2)\Gamma(1/2)}\left(\frac{\lambda}{\alpha}\right)^{1/2}$ $\mu \in \mathbb{R}$, $\lambda, \alpha > 0$, $x \in \mathbb{R}$
Fisher-Snedecor	$\mathbf{FS}\,(x \mid \alpha, \beta) = c\,\frac{x^{\alpha/2-1}}{(\beta+\alpha x)^{(\alpha+\beta)/2}}$, $c = \frac{\Gamma((\alpha+\beta)/2)}{\Gamma(\alpha/2)\Gamma(\beta/2)}\alpha^{\alpha/2}\beta^{\beta/2}$, $\alpha, \beta > 0$, $x > 0$
Uniform	$\mathbf{Uni}\,(x \mid \theta_1, \theta_2) = c$ $c = \frac{1}{\theta_2 - \theta_1}$, $\theta_2 > \theta_1$, $\theta_1 < x < \theta_2$
Exponential	$\mathbf{Ex}\,(x \mid \lambda) = c\,\exp\left[-\lambda x\right]$, $c = \lambda$, $\lambda > 0, \quad x > 0$
Inverse Gamma	$\mathbf{IGam}^{-1/2}\,(x \mid \alpha, \beta) = c\,x^{-(2\alpha+1)}\,\exp\left[-\frac{\beta}{x^2}\right]$, $c = \frac{2\beta^{\alpha}}{\Gamma(\alpha)}$, $\alpha, \beta > 0$, $x > 0$
Inverse Pareto	$\mathbf{IPar}\,(x \mid \alpha, \beta) = c\,x^{\alpha-1}$, $c = \alpha\beta^{\alpha}$, $\alpha, \beta > 0$, $0 < x < \beta^{-1}$

Probability laws for real random variables (cont.)

Cauchy	$\mathbf{Cau}\,(x\vert\lambda) = \dfrac{1/(\pi\lambda)}{1+(x/\lambda)^2}$, $\lambda \in \mathbb{R}, \quad x \in \mathbb{R}$
Rayleight	$\mathbf{Ray}\,(x\vert\theta) = c\,x\,\exp\left[-\dfrac{x^2}{\theta}\right]$, $c = \dfrac{2}{\theta}$, $\theta > 0, \quad x > 0$
Log-Normal	$\mathbf{LogN}\,(x\vert\mu,\Lambda) = c\,\exp\left[-\dfrac{(\ln x-\mu)^2}{2\Lambda^2}\right]$, $c = \dfrac{1}{\Lambda\sqrt{2\pi x}}$,
Generalized Normal	$\mathbf{Ngen}\,(x\vert\alpha,\beta) = c\,x^{\alpha-1}\exp\left[-\beta x^2\right]$, $c = \dfrac{2\beta^{\alpha/2}}{\Gamma(\alpha/2)}$,
Weibull	$\mathbf{Wei}\,(x\vert\alpha) = c\,x^{\alpha-1}\exp\left[-x^{\beta}/\alpha\right]$, $c = \dfrac{\beta}{\alpha}$, $\alpha,\beta > 0, \quad x > 0$
Double Exponential	$\mathbf{Exd}\,(x\vert\lambda) = c\,\exp\left[-\vert\lambda\vert x\right]$, $c = \dfrac{\lambda}{2}$, $\lambda > 0, \quad x \in \mathbb{R}$
Truncated Exponential	$\mathbf{Ext}\,(x\vert\lambda) = \exp\left[-(x-\lambda)\right]$, $\lambda > 0, \quad x > \lambda$

Probability laws for real random variables (cont.)

Khi	$\mathbf{Chi}\,(x\vert n) = c\,x^{n/2-1}\exp\left[-\frac{x}{2}\right]$, $c = \frac{\frac{1}{2}^{n/2}}{\Gamma(n/2)}$, $n > 0, \quad x > 0$
Khi-squared	$\mathbf{Chi}^2\,(x\vert n) = c\,x^{n/2-1}\exp\left[-\frac{x}{2}\right]$, $c = \frac{\frac{1}{2}^{n/2}}{\Gamma(n/2)}$, $n > 0, \quad x > 0$
Non centered Khi-squared	$\mathbf{Chi}^2\,(x\vert \nu, \lambda) = \sum_{i=0}^{\infty} \mathbf{Pn}\left(x\vert \frac{\lambda}{2}\right)\,\chi^2\,(x\vert \nu + 2i)$, $\nu, \lambda > 0$, $x > 0$
Inverse Khi	$\mathbf{IChi}\,(x\vert \nu) = c\,x^{-(\nu/2+1)}\exp\left[-\frac{1}{2x^2}\right]$, $c = \frac{\frac{1}{2}^{\nu/2}}{\Gamma(\nu/2}$, $\nu > 0, \quad x > 0$

Probability laws for real random variables (cont.)

Generalized Exponential with one parameters	$\mathbf{Exf}\,(x\vert a, g, \phi, h, \theta) = a(x)\,g(\theta)\exp\left[h(\theta)\phi(x)\right]$,
Generalized Exponential with K parameters	$\mathbf{Exfk}\,(x\vert a, g, \boldsymbol{\phi}, \boldsymbol{h}, \boldsymbol{\theta}) = a(x)\,g(\boldsymbol{\theta})\exp\left[\sum_{k=1}^{K} h_k(\boldsymbol{\theta})\phi_k(x)\right]$,

Probability laws for two real random variables

Normal-Gamma	$\mathbf{NGam}\,(x, y\vert \mu, \lambda, \alpha, \beta) = \mathbf{N}(x\vert \mu, \lambda y)\,\mathbf{Gam}(y\vert \alpha, \beta)$, $\mu \in \mathbb{R}, \lambda, \alpha, \beta > 0$, $x \in \mathbb{R}, y > 0$
Pareto bi-variable	$\mathbf{Par}_2\,(x, y\vert \alpha, \beta_0, \beta_1) = (y - x)^{(\alpha+2)}$, $c = \alpha(\alpha+1)(\beta_1 - \beta_0)^{\alpha}$, $(\beta_0, \beta_1) \in \mathbb{R}^2$, $\beta_0 < \beta_1, \alpha > 0$, $(x, y) \in \mathbb{R}^2$, $x < \beta_0, y > \beta_1$

Probability laws with n discrete variables

Multinomial	$\mathbf{Mu}_k\left(\boldsymbol{x}\mid\boldsymbol{\theta},n\right)=\frac{n!}{\prod_{l=1}^{k+1}x_l!}\prod_{l=1}^{k+1}\theta_l^{x_l},$ $x_{k+1}=n-\sum_{l=1}^{k}x_l,\quad \theta_{k+1}=1-\sum_{l=1}^{k}\theta_l,$ $0<\theta_l<1,\quad \sum\theta_l<1,\quad n=1,2,\cdots,$ $x_l=0,1,2,\cdots,\quad \sum x_l\le n$
Dirichlet	$\mathbf{Di}_k\left(\boldsymbol{x}\mid\boldsymbol{\alpha}\right)=c\prod_{l=1}^{k+1}x_l^{\alpha_l-1},$ $c=\frac{\Gamma\left(\sum_{l=1}^{k+1}\alpha_l\right)}{\prod_{l=1}^{k+1}\Gamma(\alpha_l)},$ $\alpha_l>0,l=1,\cdots k+1$ $0<x_l<1,l=1,\cdots k+1\quad x_{l+1}=1-\sum_{l=1}^{k}x_l$
Multinomial–Dirichlet	$\mathbf{MuDi}_k\left(\boldsymbol{x}\mid\boldsymbol{\alpha},n\right)=c\prod_{l=1}^{k+1}\frac{\alpha_l^{[x_l]}}{x_l!},$ $c=\frac{n!}{\sum_{l=1}^{k+1}\alpha_l^{[n]}},$ $\alpha^{[s]}=\prod_{l=1}^{s}(\alpha+l-1),\quad x_{k+1}=n-\sum_{l=1}^{k}x_l,$ $\alpha_l>0,n=1,2,\cdots,$ $x_l=0,1,2,\cdots,\quad \sum_{l=1}^{k}x_l<n$

Probability laws for n real variables

Canonical Exponential	$(x\mid b, \phi, h, \theta)$
Generalized Exponential	$\mathbf{Exf}_n\left(\boldsymbol{x}\mid a, g, \phi, h, \boldsymbol{\theta}\right)$
Normal	$\mathbf{N}_k\left(\boldsymbol{x}\mid\boldsymbol{\mu}, \boldsymbol{\Lambda}\right) = c \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu})\right],$ $c = \lvert\boldsymbol{\Lambda}\rvert^{1/2}\left(2\pi\right)^{-\frac{k}{2}},$ $\boldsymbol{\mu}\in\mathbb{R}^k, \boldsymbol{\Lambda} > 0,$ $\boldsymbol{x}\in\mathbb{R}^k$
Normal	$\mathbf{N}_k\left(\boldsymbol{x}\mid\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = c \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right],$ $c = \lvert\boldsymbol{\Sigma}\rvert^{-1/2}\left(2\pi\right)^{-\frac{k}{2}},$ $\boldsymbol{\mu}\in\mathbb{R}^k, \boldsymbol{\Sigma} > 0,$ $\boldsymbol{x}\in\mathbb{R}^k$
Student	$\mathbf{St}_k\left(\boldsymbol{x}\mid\boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha\right) = c\left[1+\frac{1}{\alpha}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-(\alpha+k)/2},$ $c = \frac{\Gamma\left((\alpha+k)/2\right)}{\Gamma(\alpha/2)(\alpha\pi)^{k/2}}\left(\frac{\Lambda}{\alpha}\right)^{1/2},$ $\boldsymbol{\mu}\in\mathbb{R}^k, \boldsymbol{\Lambda} > 0, \alpha > 0,$ $\boldsymbol{x}\in\mathbb{R}^k$
Wishart	$\mathbf{Wi}_k\left(\boldsymbol{X}\mid\alpha, \boldsymbol{\Lambda}\right) = c\,\lvert\boldsymbol{X}\rvert^{\alpha-(k+1)/2}\exp\left[-\mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{X})\right],$ $c = \frac{\lvert\boldsymbol{\Lambda}\rvert^{\alpha}}{\Gamma_k(\alpha)},$ $\boldsymbol{\Lambda}$ une matrice de dimensions $k\times k$, $\boldsymbol{X}$ une matrice symétrique d.p. de dimensions $k\times k$, $X_{i,j} = X_{j,i}, \quad i,j = 1,\cdots,k,$ $2\alpha > k-1$

Probability laws for $n+1$ real variables

Normal-Gamma	$\mathbf{NGam}_k\left(\boldsymbol{x}, y\mid\boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta\right) = \mathbf{N}_k(\boldsymbol{x}\mid\boldsymbol{\mu}, y\boldsymbol{\Lambda})\,\mathbf{Gam}(y\mid\alpha, \beta),$ $\boldsymbol{\mu}\in\mathbb{R}^k, \boldsymbol{\Lambda} > 0, \alpha, \beta > 0,$ $\boldsymbol{x}\in\mathbb{R}^k, y > 0$
Normal-Wishart	$\mathbf{NWi}_k\left(\boldsymbol{x}, \boldsymbol{Y}\mid\boldsymbol{\mu}, \lambda, \alpha, \boldsymbol{B}\right) = \mathbf{N}_k(\boldsymbol{x}\mid\boldsymbol{\mu}, \lambda\boldsymbol{Y})\,\mathbf{Wi}_k(\boldsymbol{Y}\mid\alpha, \boldsymbol{B}),$ $\boldsymbol{\mu}\in\mathbb{R}^k, \lambda > 0, 2\alpha > k-1, \boldsymbol{B} > 0,$ $\boldsymbol{x}\in\mathbb{R}^k, Y_{i,j} = Y_{j,i}, i,j = 1,\cdots,k$

Link between different distributions

$\mathbf{Bin}(x\|\theta, 1)$	$\mathbf{Ber}(x\|\theta)$
$\mathbf{NegBin}(x\|\theta, 1)$	$\mathbf{Geo}(x\|\theta) = \mathbf{Pas}(x\|\theta)$
$\mathbf{BinBet}(x\|1, 1, n)$	$\mathbf{Unid}(x\|n) = \frac{1}{n+1}$, $x = 0, 1, \cdots, n$
$\mathbf{Bet}(x\|1, 1)$	$\mathbf{Uni}(x\|0, 1)$
$\mathbf{Gam}(x\|0, \beta)$	$\mathbf{Ex}(x\|\beta)$
$\mathbf{Gam}(x\|\alpha, 1)$	$\mathbf{Erl}(x\|\alpha)$
$\mathbf{Gam}(x\|\frac{\nu}{2}, 1/2)$	$\mathbf{Chi}^2(x\|\nu)$
$\mathbf{IGam}(x\|\frac{\nu}{2}, 1/2)$	$\mathbf{IChi}^2(x\|\nu)$
$\mathbf{St}(x\|\mu, \lambda, 1)$	$\mathbf{Cau}(x\|\mu, \lambda)$
$\mathbf{Mu}_1(x\|\theta, n)$	$\mathbf{Bin}(x\|\theta, n)$
$\mathbf{Di}_1(x\|\alpha_1, \alpha_2)$	$\mathbf{Bet}(x\|\alpha_1, \alpha_2)$
$\mathbf{Wi}_1(x\|\alpha, \beta)$	$\mathbf{Gam}(x\|\alpha, \beta)$
$\mathbf{St}_1(x\|\mu, \lambda, \alpha)$	$\mathbf{St}(x\|\mu, \lambda, \alpha)$
$\mathbf{N}_1(x\|\mu, \lambda)$	$\mathbf{N}(x\|\mu, \lambda)$
$\mathbf{BinBet}(x\|\alpha, \beta, n)$	$\int_0^1 \mathbf{Bin}(x\|\theta, n)\, \mathbf{Beta}(\theta\|\alpha, \beta)\, \mathrm{d}\theta$
$\mathbf{NegBinBet}(x\|\alpha, \beta, r)$	$\int_0^1 \mathbf{NegBin}(x\|\theta, r)\, \mathbf{Beta}(\theta\|\alpha, \beta)\, \mathrm{d}\theta$
$\mathbf{PnGam}(x\|\alpha, \beta, n)$	$\int_0^\infty \mathbf{Pn}(x\|n\lambda)\, \mathbf{Gam}(\lambda\|\alpha, \beta)\, \mathrm{d}\lambda$
$\mathbf{Par}(x\|\alpha, \beta)$	$\int_0^\infty \mathbf{Ex}(x - \beta\|\lambda)\, \mathbf{Gam}(\lambda\|\alpha, \beta)\, \mathrm{d}\lambda$
$\mathbf{St}(x\|\mu, \lambda, \alpha)$	$\int_0^\infty \mathbf{N}(x\|\mu, \lambda y)\, \mathbf{Gam}(y\|\alpha/2, \beta/2)\, \mathrm{d}y$
$\mathbf{Chi}^2_{nc}(x\|\mu, \lambda)$	$\sum_{i=1}^{\infty} \mathbf{Pn}(i\|\lambda/2)\, \mathbf{Chi}^2(x\|\nu + 2i)$
$\lim_{\alpha \mapsto \infty} \mathbf{St}(x\|\mu, \lambda, \alpha)$	$\mathbf{N}(x\|\mu, \lambda)$
$\mathbf{St}(x\|0, 1, \alpha)$	Standard Student$\mathbf{St}(x\|\alpha)$
if $x \sim \mathbf{Gam}(x\|\alpha, \beta)$	then $y = \frac{1}{x} \sim \mathbf{IGam}(y\|\alpha, \beta)$
if $x_i \sim \mathbf{Gam}(x_i\|\alpha_i, \beta)$	then $y = \sum_{i=1}^{n} x_i \sim \mathbf{Gam}\left(y\| \sum_{i=1}^{n} \alpha_i, \beta\right)$
if $x \sim \mathbf{N}(x\|0, 1)$ and $y \sim = \mathbf{Chi}^2(x\|\nu)$	then $z = \frac{x}{\sqrt{x/\nu}} \sim \mathbf{St}(z\|0, 1, \nu)$
if $x \sim \mathbf{Chi}^2(x\|\nu_1)$ and $y \sim \mathbf{Chi}^2(y\|\nu_2)$	then $z = \frac{x/\nu_1}{y/\nu_2} \sim \mathbf{FS}(z\|\nu_1, \nu_2)$

Family of invariant position-scale distribution			
$p(x\|\mu,\beta) = \frac{1}{\beta}f(t)$ with $t = \frac{x-\mu}{\beta}$		$\mathrm{Var}\,[x] = \beta^2 v$	$-\int p(x)\ln p(x)\,\mathrm{d}x = \log\beta + h$
Family	$f(t)$	v	h
Normal	$(2\pi)^{-\frac{1}{2}}\exp\left[-\frac{t^2}{2}\right]$	1	$\frac{1}{2}\log(2\pi e)$
Gumbel	$\exp[-t]\exp[-\exp[-t]]$	$\pi^2/6$	$1+\gamma$
Laplace	$\frac{1}{2}\exp[-\|t\|]$	2	$1+\log 2$
Logistic	$\frac{\exp[-t]}{(1+\exp[-t])^2}$	$\pi^2/3$	2
Exponential	$\exp[-t], \quad x > 0$	1	1
Uniform	$1, \quad \mu - \frac{\beta}{2} < x < \mu + \frac{\beta}{2}$	$1/12$	0

Family of invariant shape-scale distributions		
$p(x\|\alpha,\beta) = \frac{1}{\beta}f(t;\alpha)$ with $t = \frac{x}{\beta}$		$\begin{cases}\mathrm{Var}\,[x] = \beta^2 v(\alpha)\\ -\int p(x)\ln p(x)\,\mathrm{d}x = \log\beta + h(\alpha)\end{cases}$
Family	$f(t)$	$\begin{cases}v\\ h\end{cases}$
Generalized Gaussian $\alpha = 2$: **Rayleigh**, $\alpha = 3$: **Maxwell-Boltzmann**, $\alpha = \nu$: **khi**	$\frac{2}{\Gamma(\frac{\alpha}{2})}t^{\alpha-1}\exp[-t^2]$	$\begin{cases}\frac{\alpha-2\Gamma^2(\frac{\alpha+1}{2})}{\Gamma^2(\frac{\alpha}{2})}\\ \log[\Gamma(\frac{\alpha}{2})/2] + \frac{1-\alpha}{2}\psi(\frac{\alpha}{2}) + \frac{\alpha}{2}\end{cases}$
Inverse Generalized Gaussian $\alpha = \nu, \beta = \sqrt{2}$: **khi inverse**	$\frac{2}{\Gamma(\frac{\alpha}{2})}t^{-\alpha-1}\exp[-t^{-2}]$	$\begin{cases}\frac{1}{\alpha-2} - \alpha\frac{\Gamma^2(\alpha-1)}{\Gamma^2(\frac{\alpha}{2})}\\ \log[\Gamma(\frac{\alpha}{2})/2] + \frac{-\alpha}{2}\psi(\frac{\alpha}{2}) + \frac{\alpha}{2}\end{cases}$
Gamma $\alpha = 1$: **Exponential**, $\alpha = \frac{\nu}{2}, \beta = 2$: **khi-squared**, $\alpha = \nu$: **Erlang**	$\frac{1}{\Gamma(\alpha)}t^{\alpha-1}\exp[-t]$	$\begin{cases}\alpha\\ \log[\Gamma(\alpha)] + (1-\alpha)\psi(\alpha) + \alpha\end{cases}$
Inverse Gamma $\alpha = \frac{\nu}{2}, \beta = \frac{1}{2}$: **khi-squared inverse**	$\frac{1}{\Gamma(\alpha)}t^{-\alpha+1}\exp[-t^{-1}]$	$\begin{cases}\frac{1}{(\alpha-1)^2(\alpha-2)},\ \alpha > 2\\ \log[\Gamma(\alpha)] - (1+\alpha)\psi(\alpha) + \alpha\end{cases}$
Pareto	$\alpha t^{-\alpha-1},\ x > \beta$	$\begin{cases}\frac{1}{(\alpha-1)^2(\alpha-2)},\ \alpha > 2\\ \frac{1}{\alpha} - \log\alpha + 1\end{cases}$
Weibull	$\alpha t^{\alpha-1}\exp[-t]^{\alpha}$	$\begin{cases}\Gamma(1+\frac{2}{\alpha})\Gamma^2(1+\frac{1}{\alpha})\\ \frac{\gamma(\alpha-1)}{\alpha} - \log\alpha + 1\end{cases}$

Appendix A

Exercises

A.1 Exercise 1: Signal detection and parameter estimation

Assume $z_i = s_i + e_i$ where $s_i = a\cos(\omega t_i + \phi)$ is the transmitted signal, e_i is the noise and z_i is the received signal. Assume that we have received n independent samples $z_i, i = 1, \ldots, n$. Assume also that $\sum_{i=1}^{n} z_i = 0$.

1. Assume that $e_i \sim \mathcal{N}(0, \theta)$ and that s_i is perfectly known. Design a Bayesian optimal detector with the uniform prior and the uniform cost coefficients.

2. Repeat (1) with the assumption that $e_i \sim \frac{1}{2\theta}\exp\left[-|e_i|/\theta\right], \theta > 0$.

3. Repeat (1) but assume now that θ is unknown and distributed as $\theta \sim \pi(\theta) \propto \theta^{\alpha}$.

4. Assume that $e_i \sim \mathcal{N}(0, \theta)$, but now a is not known.
   - Give the expression of the ML estimate $\widehat{a}_{ML}(\boldsymbol{z})$ of a.
   - Give the expressions of the MAP estimate $\widehat{a}_{MAP}(\boldsymbol{z})$ and the Bayes optimal estimate $\widehat{a}_B(\boldsymbol{z})$ if we assume that $a \sim \mathcal{N}(0, 1)$.
   - Give the expressions of the MAP estimate $\widehat{a}_{MAP}(\boldsymbol{z})$ and the Bayes optimal estimate $\widehat{a}_B(\boldsymbol{z})$ if we assume that $\frac{1}{2\alpha}\exp\left[-|a|/\alpha\right], \alpha > 0$.

5. Repeat (4) but assume now that θ is unknown and distributed as $\theta \sim \pi(\theta) \propto \theta^{\alpha}$.

6. Assume that $e_i \sim \mathcal{N}(0, \theta)$ with known θ and that a is known, but now ω is unknown.
   - Give the expression of the ML estimate $\widehat{\omega}_{ML}(\boldsymbol{z})$ of ω.
   - Give the expressions of the MAP estimate $\widehat{\omega}_{MAP}(\boldsymbol{z})$ and the Bayes optimal estimate $\widehat{\omega}_B(\boldsymbol{z})$, if we assume that $\omega \sim \mathbf{Uni}(0, 1)$.

7. Assume that $e_i \sim \mathcal{N}(0, \theta)$ with known θ and that a and ω are known, but now ϕ is unknown.
   - Give the expression of the ML estimate $\widehat{\phi}_{ML}(\boldsymbol{z})$ of ϕ.
   - Give the expressions of the MAP estimate $\widehat{\phi}_{MAP}(\boldsymbol{z})$ and the Bayes optimal estimate $\widehat{\phi}_B(\boldsymbol{z})$, if we assume that $\phi \sim \mathbf{Uni}(0, 2\pi)$.

8. Assume that $e_i \sim \mathcal{N}(0, \theta)$ with known θ, but now both a and ω are unknown.
   - Give the expressions of the ML estimates $\widehat{a}_{ML}(\boldsymbol{z})$ of a and $\widehat{\omega}_{ML}(\boldsymbol{z})$ of ω.
   - Give the expressions of the Bayes optimal estimates $\widehat{a}_B(\boldsymbol{z})$ of a and $\widehat{\omega}_B(\boldsymbol{z})$ of ω if we assume that $a \sim \mathcal{N}(0, 1)$ and $\omega \sim \mathbf{Uni}(0, 1)$.
   - Give the expressions of the MAP estimates $\widehat{a}_{MAP}(\boldsymbol{z})$ of a and $\widehat{\omega}_{MAP}(\boldsymbol{z})$ of ω if we assume that $a \sim \mathcal{N}(0, 1)$ and $\omega \sim \mathbf{Uni}(0, 1)$.

9. Assume now that ω is known and we know that a can only take the values $\{-1, 0, +1\}$. Design an optimal detector for a.

10. Assume now that

$$s_i = \sum_{k=1}^{K} a_k \cos(\omega_k t + \phi_k), \quad \omega_k \neq \omega_l, \forall k \neq l$$

- Give the expression of the ML estimates $\widehat{a}_k(\boldsymbol{z})$ assuming ω_k and ϕ_k known.
- Give the expressions of the ML estimates $\widehat{\omega}_k(\boldsymbol{z})$ assuming a_k and ϕ_k known.
- Give the expressions of the ML estimates $\widehat{\phi}_k(\boldsymbol{z})$ assuming a_k and ω_k known.
- Give the expressions of the joint ML estimates $\widehat{a}_k(\boldsymbol{z})$ and $\widehat{\omega}_k(\boldsymbol{z})$ assuming ϕ_k unknown.
- Discuss the pssibilty of the joint estimation of $\widehat{a}_k(\boldsymbol{z})$, $\widehat{\omega}_k(\boldsymbol{z})$ and a_k.

A.2 Exercise 2: Discrete deconvolution

Assume $z_i = s_i + e_i$ where

$$s_i = \sum_{k=0}^{p} h_k \, x(i-k)$$

and where

- $\boldsymbol{h} = [h_0, h_1, \ldots, h_p]^t$ represents the finite impulse response of a chanel,
- $\boldsymbol{x} = [x(0), \ldots, x(n)]^t$ the finite input sequence,
- $\boldsymbol{z} = [z(p), \ldots, z(p+n)]^t$ the received signal, and
- $\boldsymbol{e} = [e(p), \ldots, e(p+n)]^t$ the chanel noise.

1. Construct the matrixes $\boldsymbol{H}$ and $\boldsymbol{X}$ in such a way that $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{e}$ and $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{h} + \boldsymbol{e}$.
2. Assume that $e_i \sim \mathcal{N}(0, \theta)$ and that we know perfectly $\boldsymbol{h}$ and the input sequence $\boldsymbol{x}$. Design a Bayesian optimal detector with the uniform prior and the uniform cost coefficients.
3. Repeat (2) with the assumption that $e_i \sim \frac{1}{2\theta} \exp\left[-|e_i|/\theta\right], \theta > 0$.
4. Repeat (2) but assume now that θ is unknown and distributed as $\theta \sim \pi(\theta) \propto \theta^\alpha$.
5. Assume that $e_i \sim \mathcal{N}(0, \theta)$, but now $\boldsymbol{x}$ is unknown.
   - Give the expression of the ML estimate $\widehat{\boldsymbol{x}}_{ML}(\boldsymbol{z})$ of $\boldsymbol{x}$.
   - Give the expressions of the MAP $\widehat{\boldsymbol{x}}_{MAP}(\boldsymbol{z})$ and the Bayes optimal estimate $\widehat{\boldsymbol{x}}_B(\boldsymbol{z})$ if we assume that $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.
6. Repeat (5) but assume now that θ is unknown and distributed as $\theta \sim \pi(\theta) \propto \theta^\alpha$.
7. Assume that $e_i \sim \mathcal{N}(0, \theta)$ with known θ, and that $\boldsymbol{x}$ is known but $\boldsymbol{h}$ is unknown.
   - Give the expression of the ML estimate $\widehat{\boldsymbol{h}}_{ML}(\boldsymbol{z})$ of $\boldsymbol{h}$.
   - Give the expression of the MAP estimate $\widehat{\boldsymbol{h}}_{MAP}(\boldsymbol{z})$ and the Bayes optimal estimate $\widehat{\boldsymbol{h}}_B(\boldsymbol{z})$ if we assume that $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.
8. Repeat (7) but assume now that θ is unknown and distributed as $\theta \sim \pi(\theta) \propto \theta^\alpha$.
9. Assume that $e_i \sim \mathcal{N}(0, \theta)$ with known θ, but now both $\boldsymbol{h}$ and $\boldsymbol{x}$ are unknown.
   - Give the expressions of the ML estimates $\widehat{\boldsymbol{h}}_{ML}(\boldsymbol{z})$ of $\boldsymbol{h}$ and $\widehat{\boldsymbol{x}}_{ML}(\boldsymbol{z})$ of $\boldsymbol{x}$.
   - Give the expressions of the MAP and the Bayes optimal estimates $\widehat{\boldsymbol{x}}_{MAP}(\boldsymbol{z})$ and $\widehat{\boldsymbol{x}}_B(\boldsymbol{z})$ of $\boldsymbol{x}$ and $\widehat{\boldsymbol{h}}_{MAP}(\boldsymbol{z})$ and $\widehat{\boldsymbol{h}}_B(\boldsymbol{z})$ of $\boldsymbol{h}$ if we assume that $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.
   - Give the expressions of the MAP and the Bayes optimal estimates $\widehat{\boldsymbol{x}}_{MAP}(\boldsymbol{z})$ and $\widehat{\boldsymbol{x}}_B(\boldsymbol{z})$ of $\boldsymbol{x}$ and $\widehat{\boldsymbol{h}}_{MAP}(\boldsymbol{z})$ and $\widehat{\boldsymbol{h}}_B(\boldsymbol{z})$ of $\boldsymbol{h}$ if we assume that $\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_x^2 \boldsymbol{D}_x^t \boldsymbol{D}_x\right)$ and $\boldsymbol{h} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_h^2 \boldsymbol{D}_h^t \boldsymbol{D}_h\right)$.

10. Assume now that $\boldsymbol{h}$ is known and we know that $x(k)$ can only take the values $\{0,1\}$.

    - First assume that $x(k)$ are independent and $\pi_0 = P(x(k) = 0)$ and $\pi_1 = P(x(k) = 1) = 1 - \pi_0$, with known π_0. Design an optimal detector for $x(k)$.
    - Now assume $x(k)$ can be modelled as a first order Markov chaine and that we know the probabilites

      $$\begin{array}{ll} \pi_{00} = P(x(k)=0, x(k+1)=0) & \pi_{01} = P(x(k)=0, x(k+1)=1) = 1 - \pi_{00} \\ \pi_{11} = P(x(k)=1, x(k+1)=1) & \pi_{10} = P(x(k)=1, x(k+1)=0) = 1 - \pi_{11} \end{array}$$

      Design an optimal detector for $x(k)$.
    - Repeat these two last items, assuming now that $\boldsymbol{h}$ is unknown and $\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0}, \sigma_h^2 \boldsymbol{D}_h^t \boldsymbol{D}_h\right)$.

List of Figures

Bibliography

Index