Chapter 2
Statistical Estimation Theory

Contents

2.1 Statistical Estimation Theory

2.1.1 General view of the estimation problem

The design of an estimator involves creating a real-valued function that, given an input vector, $\mathbf{x}$ of observational variables, makes predictions about a target variable, s.

We will assume that there exists some statistical dependency between the the observations and the target. To do so, we model the observations and the target by means of random variables $\mathbf{X}$ and S, respectively[1]. The observation is a sample from an *observation space* $\mathscr{X}$ which, in general, will be a subset of $\mathbb{R}^n$. In general, we will assume that the target variable is real, although the general formulation can be applied to multidimensional cases. A schematic view of the estimation problem is depicted in Fig. 2.1.

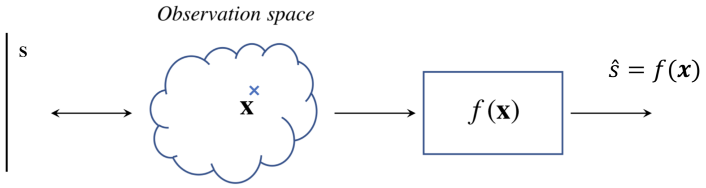

Fig. 2.1 Diagram block of estimation problems.

The estimation module applies a real output function $f(\cdot)$ that is commonly referred to as the *estimator*, and its output, $\hat{S} = f(\mathbf{X})$, as the *estimation* or *prediction*. The estimator is a deterministic function, meaning that for a given value $\mathbf{x}$, it will consistently produce the same output. Although $f(\cdot)$ is deterministic, if the input $\mathbf{X}$ is a random vector, the prediction $\hat{S}$ is a random variable.

The estimator is likely to incur in some estimation error that will be quantified by means of a cost (or, alternatively, a reward) function. Designing our estimator will require minimizing (or maximizing) the expected value of this cost (reward).

We identify two main types of problems related to estimation:

- Analysis of estimators: given an estimator, evaluate its performance using a specific measure (a cost or a reward function).
- Design of estimators: find a function $f(\mathbf{x})$ that optimizes a predefined goal.

2.1.2 Statistical information involved in estimation problems

The statistical relation between the observations and the target variable is described by the **joint** probability density function (pdf) of $\mathbf{X}$ and S: $p_{\mathbf{X},S}(\mathbf{x},s)$, or some distribution related to it.

[1] Note that we use capital letters to model the random variables, and lowercase letters to denote an arbitrary realization of them.

The join pdf can be factorized as products of conditional and marginal pdfs:

$$p_{\mathbf{X},S}(\mathbf{x},s) = p_{\mathbf{X}|S}(\mathbf{x}|s)\cdot p_S(s) = p_{S|\mathbf{X}}(s|\mathbf{x})\cdot p_{\mathbf{X}}(\mathbf{x}) \tag{2.1}$$

In the context of estimation theory, these factors receive specific names:

- The **likelihood** of $S = s$ for observation $\mathbf{x}$, $p_{\mathbf{X}|S}(\mathbf{x}|s)$: it characterizes the generation of observations for each value of the target variable.
- The **prior (or *a priori*) distribution** of S, $p_S(s)$: it describes how much is known (or unknown) about the target variable before observing $\mathbf{X}$.
- The **posterior (or *a posteriori*) distribution** of S given $\mathbf{X} = \mathbf{x}$, $p_{S|\mathbf{X}}(s|\mathbf{x})$: it describes the knowledge (or the uncertainty) about S after observing $\mathbf{X}$.
- The **evidence** or marginal distribution of $\mathbf{X}$, $p_{\mathbf{X}}(\mathbf{x})$.

The information available to design the estimator may depend on the application. A typical scenario, because it is related to the physical generative process of the observations, is the one in which the likelihood function is known, and the design of the estimator is based on it. If additionally, a prior distribution is available, the design can be grounded on the posterior distribution $p_{S|\mathbf{X}}(s|\mathbf{x})$, which can be calculated by means of Bayes' Theorem,

$$p_{S|\mathbf{X}}(s|\mathbf{x}) = \frac{p_{\mathbf{X},S}(\mathbf{x},s)}{p_{\mathbf{X}}(\mathbf{x})} = \frac{p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)}{\int p_{\mathbf{X}|S}(\mathbf{x}|s')p_S(s')ds'} \tag{2.2}$$

2.1.3 Cost functions for estimation problems

The evaluation and design of an estimator requires some objective criteria. In some cases, we will consider that this criterion materializes in the form of a **cost function** whose value we seek to minimize.

A cost function $c(s,\hat{s})$ is any measure of the discrepancy between the target variable and the estimation. In general, it is non-negative, $c(s,\hat{s}) \geq 0$, with equality for $s = \hat{s}$. In some cases, the cost function can be expressed as a function of the estimation error $e = s - \hat{s}$ and we will write[2] $c(s,\hat{s}) = c(s-\hat{s}) = c(e)$. Some frequently used cost functions are:

- Quadratic cost: $c(e) = e^2$.
- Absolute value of the error: $c(e) = |e|$.
- Relative quadratic error: $c(s,\hat{s}) = \frac{(s-\hat{s})^2}{s^2}$
- Cross Entropy: $c(s,\hat{s}) = -s\ln\hat{s} - (1-s)\ln(1-\hat{s})$, for $s,\hat{s} \in [0,1]$

El término "risk" no se usaba en estas lecture notes en la version anterior (que usa "mean cost"), pero yo creo que es importante, porque es el que se usa en machine learning. Y es más corto.

Since the target variable is unknown, the prediction cannot be computed by the direct minimization of the cost $c(s,\hat{s})$, and we have to work with expectations. The expected value of the cost is usually referred as the **risk** of an estimator $\hat{s} = f(\mathbf{x})$:

[2] Note that the cost function is denoted with a lowercase letter, c, because it is a deterministic function, i.e., for fixed values of s and $\hat{s}$ the cost always takes the same value. However, as with the estimation function, the application of that function to random variables will result in another random variable, i.e., $C = c(S,\hat{S})$.

Suelo contar este ejemplo en clase. Creo que ayuda a entender el riesgo como métrica de calidad de un estimador.

$$R_f = \mathbb{E}\{c(S,\hat{S})\} = \int_{\mathbf{x}}\int_{s} c(s,f(\mathbf{x}))p_{S,\mathbf{X}}(s,\mathbf{x})dsd\mathbf{x} \tag{2.3}$$

By the Law of Large Numbers, this is the average cost that we can expect from a given estimator, after a large number of predictions.

The **conditional risk** is the conditional mean for a given observation

$$R(\hat{s},\mathbf{x}) = \mathbb{E}\{c(S,\hat{s})|\mathbf{x}\} = \int_{s} c(s,\hat{s})p_{S|\mathbf{X}}(s|\mathbf{x})ds \tag{2.4}$$

Example 2.1 (Evaluation of estimators 1) Given the joint distribution

$$p_{S,X}(s,x) = \left[\begin{array}{ll} \frac{1}{x}, & 0 \le s \le x \text{ and } 0 < x \le 1 \\ 0, & \text{otherwise} \end{array}\right. , \tag{2.5}$$

consider the estimators $\hat{S}_1 = \frac{1}{2}X$ and $\hat{S}_2 = X$. Which is the best estimator from the point of view of the quadratic cost? To find out, we'll calculate the mean quadratic error for both estimators. Knowing that, for any w,

$$\begin{aligned}\mathbb{E}\{(S-wX)^2\} &= \int_0^1\int_0^x (s-wx)^2 p_{S,X}(s,x)dsdx = \int_0^1\int_0^x (s-wx)^2\frac{1}{x}dsdx \\ &= \int_0^1\left(\frac{1}{3}-w+w^2\right)x^2dx = \frac{1}{3}\left(\frac{1}{3}-w+w^2\right) \end{aligned} \tag{2.6}$$

Taking $w = 1/2$ and $w = 1$ we get, respectively,

$$\mathbb{E}\left\{(S-\hat{S}_1)^2\right\} = \mathbb{E}\left\{\left(S-\frac{1}{2}X\right)^2\right\} = \frac{1}{3}\left(\frac{1}{3}-\frac{1}{2}+\frac{1}{4}\right) = \frac{1}{36} \tag{2.7}$$

$$\mathbb{E}\{(S-\hat{S}_2)^2\} = \mathbb{E}\{(S-X)^2\} = \frac{1}{3}\left(\frac{1}{3}-1+1\right) = \frac{1}{9} \tag{2.8}$$

Therefore, from the point of view of the square error, $\hat{S}_1$ is a better estimator than $\hat{S}_2$.

Example 2.2 (Evaluation of estimators 2) Assume that S is a random variable of mean 0 and variance 1, and X is a noisy observation of S,

$$X = S + R \tag{2.9}$$

where R is a random Gaussian variable, independent of S, of mean 0 variance v. We will compute the risk for the estimator $\hat{S} = X$ and different cost functions. For the quadratic error:

$$\mathbb{E}\{(S-\hat{S})^2\} = \mathbb{E}\{(S-X)^2\} = \mathbb{E}\{R^2\} = v \tag{2.10}$$

For the mean absolute error

El sesgo y la varianza no se mencionaban en este capítulo, así que lo he añadido ahora, primero porque hay bastantes ejercicios que preguntan por el sesgo y la varianza en el boletín, y segundo porque quiero profundizar en esto en futuros cursos.

$$\begin{aligned}\mathbb{E}\{|S-\hat{S}|\} = \mathbb{E}\{|R|\} &= \int_{-\infty}^{\infty} |r| \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{r^2}{2v}\right) dr \\ &= 2\int_{0}^{\infty} r \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{r^2}{2v}\right) dr = \sqrt{\frac{2v}{\pi}} \end{aligned} \tag{2.11}$$

2.1.4 Bias and variance

The bias of estimator $\hat{S}$ for a true target $S = s$ is defined as

$$B = \mathbb{E}\{\hat{S}|S = s\} - s \tag{2.12}$$

and it accounts for the expected deviation of the estimator from the true value of the target variable.

The variance of estimator $\hat{S}$ for a true target $S = s$ is defined as

$$V = \mathrm{var}\{\hat{S}|S = s\} \tag{2.13}$$

The variance accounts for the spread of the sampling distribution, or in other words, it quantifies how much the estimates vary from one sample $\mathbf{x}$ to another, for the same realization of S. Unlike bias, which assesses systematic deviation from the true parameter value, variance captures the randomness inherent in the estimation process due to sampling variability.

2.2 Design of estimators

2.2.1 Maximum Likelihood (ML) estimation

The maximum likelihood estimator (ML) uses the likelihood as the reward function to be maximized:

$$\hat{s}_{\text{ML}} = \underset{s}{\operatorname{argmax}}\, p_{\mathbf{X}|S}(\mathbf{x}|s) = \underset{s}{\operatorname{argmax}} \ln(p_{\mathbf{X}|S}(\mathbf{x}|s)) \tag{2.14}$$

The ML estimator selects the value of the parameter s that maximizes the likelihood of observing $\mathbf{x}$ when $S = s$. Loosely speaking, observing $\mathbf{x}$ when $S = \hat{s}_{\text{ML}}$ is less unexpected than if S takes any other value. Note that $p_{\mathbf{X}|S}(\mathbf{x}|s)$, which is a density function over random variable $\mathbf{X}$, is not maximized with respect to $\mathbf{x}$, but with respect to s.

Este ejemplo, y algunos ejercicios, son útiles para mostrar que el sML o el sMAP no siempre están en un punto de derivada nula. Es importante hacerlo notar en algún momento en clase porque muchos alumnos, en el examen, van de cabeza a derivar, igualar a cero y despejar cada vez que tienen que maximizar o minimizar, sin tener en cuenta que el máximo o el mínimo pueden estar en los extremos del dominio.

Example 2.3 (ML Estimation)

We want to estimate the value of a random variable S from an observation X statistically related to it. For the design of the estimator, only the likelihood of S is known, which is given by

$$p_{X|S}(x|s) = \frac{2x}{(1-s)^2}, \quad 0 \le x \le 1-s, \quad 0 \le s \le 1 \tag{2.15}$$

Given the available statistical information, it is decided to construct the ML estimator of S. The likelihood is a probability density function of X with unit area, as represented in Figure 2.2(a). However, to carry out the maximization, representing this likelihood as a function[3] of s (Fig.2.2(b)) is more useful, as it shows that the estimator is

$$\hat{s}_{\text{ML}} = 1 - x$$

or, alternatively, if we consider the application of the estimation function on the random variable X instead of on a specific value of it,

$$\hat{S}_{\text{ML}} = 1 - X$$

Note that the second equality in (2.14) states that the maximization of the likelihood is equivalent to the maximization of its logarithm (the **log-likelihood** function). Since the logarithm function is strictly increasing, $p_{\mathbf{X}|S}(\mathbf{x}|s_1) > p_{\mathbf{X}|S}(\mathbf{x}|s_2)$ implies $\ln(p_{\mathbf{X}|S}(\mathbf{x}|s_1)) > \ln(p_{\mathbf{X}|S}(\mathbf{x}|s_2)))$ and, thus, the logarithm does not alter the outcome of the maximization. The logarithm is used by practical reasons when the likelihood is a product of several factors or an exponential function, as it will transform products into sums and an exponential into the exponents. In this way, the maximization process can be simplified considerably.

Note that the maximum likelihood does not need any probability model about the target variable, S, which is treated as a deterministic parameter. This is useful in situation where only the likelihood function is known.

[3] Note that the integral with respect to s of $p_{X|S}(x|s)$ will not generally be the unit, since this function does not constitute a probability density of S.

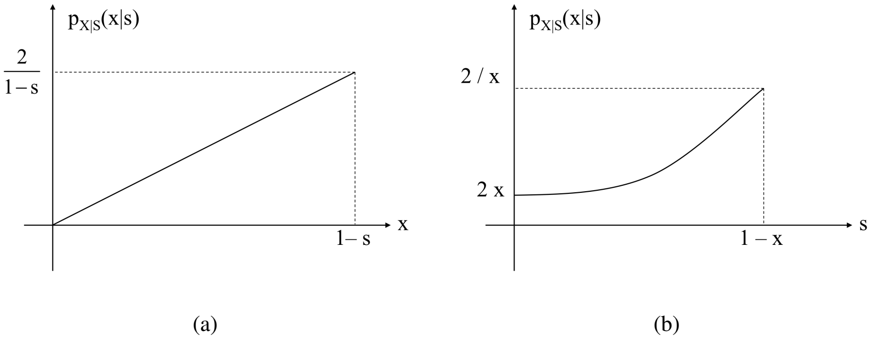

Fig. 2.2 Representation of the likelihood distribution of the example 2.3 as a function of x and s.

2.2.2 Maximum a posteriori (MAP) estimation

We define the maximum a posterior (MAP) estimator as the mode of the posterior distribution, that is

$$\hat{s}_{\text{MAP}} = \underset{s}{\operatorname{argmax}}\, p_{S|\mathbf{X}}(s|\mathbf{x}) = \underset{s}{\operatorname{argmax}}\ln(p_{S|\mathbf{X}}(s|\mathbf{x})) \tag{2.16}$$

Using the definition of conditional pdf and the Bayes' rule, it is easy to see that the MAP estimator can be computed as

$$\hat{s}_{\text{MAP}} = \underset{s}{\operatorname{argmax}}\, p_{S,\mathbf{X}}(s,\mathbf{x}) \tag{2.17}$$

$$= \underset{s}{\operatorname{argmax}}\{p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)\} \tag{2.18}$$

Eq. (2.18) demonstrates that the MAP estimator seeks to maximize the likelihood function, modulated by the prior distribution. This integration of the prior distribution allows the MAP estimator to incorporate existing knowledge or assumptions about the target S before observing the data, $\mathbf{x}$. When the prior distribution, is uniform across entire range of possible values of S, the distinction between the MAP and ML estimators vanishes, as the MAP estimator effectively reduces to the ML estimator.

However, when the prior distribution is not uniform, the MAP estimator is biased towards values of the target variable with higher prior probabilities. This shift illustrates the MAP estimator's sensitivity to prior knowledge.

Toda esta discusión sobre los paradigmas Bayesiano y frecuentista no estaba (salvo de pasada) en la versión anterior, pero yo la contaba en clase (de hecho, ya hablé sobre eso en la primera sesión del curso) y creo que es importante, no solo en SSP sino en todo el machine learning.

Beyond their mathematical formulations, the MAP and ML estimators embody fundamentally different inference philosophies. The ML estimator optimizes the likelihood function, which models the probability of the observed data under various values of the target, treating s as a fixed but unknown parameter. This approach aligns with the **frequentist paradigm**, which interprets probability as the long-run frequency of events and does not incorporate prior information about S.

Conversely, the MAP estimator embraces a **Bayesian framework**, treating the target as a random variable. This perspective allows the incorporation of prior knowledge or beliefs about S through the prior distribution, $p_S(s)$, and the posterior distribution, $p_{S|\mathbf{X}}(s|\mathbf{x})$, updates this knowledge based on new evidence from the data. The Bayesian approach, therefore, provides a probabilistic framework for updating beliefs about uncertain parameters in light of new data.

Example 2.4 (Estimation MAP) Considering that

$$p_{S|X}(s|x) = \frac{1}{x^2} s \exp\left(-\frac{s}{x}\right), \qquad x \geq 0, \quad s \geq 0 \tag{2.19}$$

the MAP estimator can be computed maximizing

$$\ln(p_{S|X}(s|x)) = -2\ln(x) + \ln(s) - \frac{s}{x}, \qquad x \geq 0, \quad s \geq 0, \tag{2.20}$$

Since $\ln(p_{S|X}(s|x))$ tends to $-\infty$ around $s = 0$ and $s = \infty$, its maximum must be at some intermediate point with zero derivative. Deriving respect to s results in

$$\frac{\partial}{\partial s} \ln p_{S|X}(s|x)\bigg|_{s=\hat{s}_{\text{MAP}}} = \frac{1}{\hat{s}_{\text{MAP}}} - \frac{1}{x} = 0, \qquad x \geq 0, \quad s \geq 0 \tag{2.21}$$

Thus,

$$\hat{s}_{\text{MAP}} = x \tag{2.22}$$

2.2.3 Minimum risk estimators

When a cost function is used to evaluate the quality of an estimation for a given estimation problem, one may wonder if we can find a mathematical expression for the estimator minimizing the mean value of the cost, that is, the risk.

Taking back the formula of the risk, and applying the total expectation theorem, we find

$$R_f = \mathbb{E}\{c(S,\hat{S})\} = \int_{\mathbf{x}} \mathbb{E}\{c(S,\hat{s})|\mathbf{X} = \mathbf{x}\} p_{\mathbf{X}}(\mathbf{x})\mathbf{dx} \tag{2.23}$$

where $\hat{s} = f(\mathbf{x})$. That is, the risk is the integral of the conditional risk, and, thus, the estimator minimizing the risk will be such that it minimizes the conditional risk for each observation $\mathbf{x}$,

$$\hat{s}^* = \underset{\hat{s}}{\text{argmin}}\ \mathbb{E}\{c(S,\hat{s})|\mathbf{X} = \mathbf{x}\} \tag{2.24}$$

We will refer to this estimator as the **Bayesian estimator** associated to cost function $c()$.

Example 2.5 (Calculation of a minimum mean square error estimator) Following the example 2.1, we can calculate the posterior distribution of S through

$$p_{S|X}(s|x) = \frac{p_{S,X}(s,x)}{p_X(x)}. \tag{2.25}$$

Knowing that

$$p_X(x) = \int_0^1 p_{S,X}(s,x)ds = \int_0^x \frac{1}{x} ds = 1, \qquad 0 \leq x \leq 1 \tag{2.26}$$

we obtain

$$p_{S|X}(s|x) = \begin{bmatrix} \frac{1}{x}, & 0 < s < x < 1 \\ 0, & \text{otherwise} \end{bmatrix} \tag{2.27}$$

The conditional risk will be given by

$$\begin{aligned} \mathbb{E}\{c(S,\hat{s})|X=x\} &= \mathbb{E}\{(S-\hat{s})^2|X=x\} \\ &= \int_0^1 (s-\hat{s})^2 p_{S|X}(s|x)ds \\ &= \frac{1}{x}\int_0^x (s-\hat{s})^2 ds = \frac{1}{x}\left(\frac{(x-\hat{s})^3}{3}+\frac{\hat{s}^3}{3}\right) \\ &= \frac{1}{3}x^2 - \hat{s}x + \hat{s}^2. \end{aligned} \tag{2.28}$$

As a function of $\hat{s}$, conditional risk is a second-degree polynomial, whose minimum can be calculated through differentiation. Since

$$\frac{d}{d\hat{s}}\mathbb{E}\{c(S,\hat{s})|X=x\} = -x+2\hat{s}, \tag{2.29}$$

the Bayesian estimator associated to the quadratic error is

$$\hat{s}^* = \frac{1}{2}x, \tag{2.30}$$

which matches the estimator $\hat{S}_1$ from the example 2.1. Therefore, $\hat{S}_1$ is the best possible estimator from the point of view of the mean square error.

Based on (2.24) we can conclude that, regardless of the cost to be minimized, the knowledge of the posterior distribution of S given $\mathbf{X}$, $p_{S|\mathbf{X}}(s|\mathbf{x})$, is sufficient to design the Bayesian estimator for a given cost. As mentioned above, this distribution is often calculated from the likelihood of S and its a prior distribution using the Bayes Theorem, which is in fact the origin of the denomination of these estimators.

2.3 Common Bayesian estimators

This section presents some of the most commonly used Bayesian estimators. For their calculation, we will proceed to minimize the mean cost given $\mathbf{X}$ (posterior mean cost) for different cost functions.

2.3.1 Minimum Mean Squared Error estimator (MSE)

The minimum mean squared error (MSE) estimator is the Bayesian estimator associated with the cost function $c(e) = e^2 = (s-\hat{s})^2$, and therefore is given by

$$\begin{aligned}\hat{s}_{\text{MSE}} &= \arg\min_{\hat{s}} \mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\} \\ &= \arg\min_{\hat{s}} \mathbb{E}\{(S-\hat{s})^2|\mathbf{X}=\mathbf{x}\} \qquad (2.31)\end{aligned}$$

Figure 2.3 illustrates the minimum MSE estimation problem. The risk can be obtained by integrating (with respect to s) the product of the square error and the posterior pdf of S. The argument for minimization is $\hat{s}$, which allows to shift the graph corresponding to the cost function (represented with discontinuous stroke) so that the result of that integral is minimal.

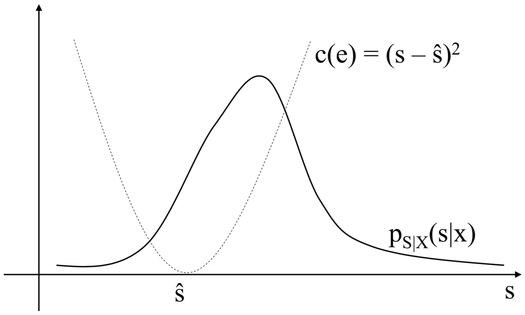

Fig. 2.3 Graphical representation of the process of calculating the posterior mean for a generic value $\hat{s}$.

For the square error, the conditional risk in (2.31) becomes

$$\mathbb{E}\{(S-\hat{s})^2|\mathbf{X}=\mathbf{x}\} = \mathbb{E}\{S^2|\mathbf{X}=\mathbf{x}\} - 2\mathbb{E}\{S|\mathbf{X}=\mathbf{x}\}\hat{s} + \hat{s}^2 \qquad (2.32)$$

This is a second-degree polynomial that can be minimized by differentiation to get

$$\boxed{\hat{s}_{\text{MSE}} = \mathbb{E}\{S|\mathbf{X}=\mathbf{x}\} = \int s\, p_{S|\mathbf{X}}(s|x)ds \qquad (2.33)}$$

In other words, the minimum MSE estimator of S is the posterior mean of S given $\mathbf{X}$.

Example 2.6 (Straightforward calculation of the MSE estimator) According to (2.33), minimum mean squared error estimator obtained in 2.1 can alternatively be derived as follows

$$\hat{s}_{\text{MSE}} = \int_0^1 s p_{S|X}(s|x) ds = \int_0^x \frac{s}{x} ds = \frac{1}{2} x \tag{2.34}$$

which is consistent with (2.30).

2.3.2 Minimum Mean Absolute Deviation Estimator (MAD)

In the same way as we have proceeded in the case of the estimator $\hat{s}_{\text{MSE}}$, we can calculate the estimator associated with the absolute deviation of the estimation error, $c(e) = |e| = |s - \hat{s}|$. This estimator, which we will refer to as the Mean Absolute Deviation (MAD), is characterized by

$$\begin{aligned} \hat{s}_{\text{MAD}} &= \arg\min_{\hat{s}} \ \mathbb{E}\{|S - \hat{s}| \ |\mathbf{X} = \mathbf{x}\} = \\ &= \arg\min_{\hat{s}} \int_s |s - \hat{s}| \ p_{S|\mathbf{X}}(s|\mathbf{x}) ds \end{aligned} \tag{2.35}$$

Again, it is simple to illustrate the process of calculating the posterior mean cost by overlapping on the same axes the cost expressed as a function of s and the posterior distribution of the variable to be estimated (see Fig. 2.4). This representation also suggests the convenience of splitting the integral into two parts corresponding to the two slopes of the cost function:

Este desarrollo no lo hago en clase. Respecto al MAD, lo defino con (2.35) y les digo la solución (2.38, 2.39), diciendo que los detalles los tienen en las notes. El desarrollo en (2.36-37) lleva tiempo, y no aporta mucho.

$$\begin{aligned} \mathbb{E}\{|S - \hat{s}| \ |\mathbf{X} = \mathbf{x}\} &= \int_{-\infty}^{\hat{s}} (\hat{s} - s) \ p_{S|\mathbf{X}}(s|\mathbf{x}) ds + \int_{\hat{s}}^{\infty} (s - \hat{s}) \ p_{S|\mathbf{X}}(s|\mathbf{x}) ds \\ &= \hat{s} \left[\int_{-\infty}^{\hat{s}} p_{S|\mathbf{X}}(s|\mathbf{x}) ds - \int_{\hat{s}}^{\infty} p_{S|\mathbf{X}}(s|\mathbf{x}) ds \right] + \\ &\quad + \int_{\hat{s}}^{\infty} s \ p_{S|\mathbf{X}}(s|\mathbf{x}) ds - \int_{-\infty}^{\hat{s}} s \ p_{S|\mathbf{X}}(s|\mathbf{x}) ds \end{aligned} \tag{2.36}$$

The fundamental theorem of calculus[4] allows us to obtain the derivative of the conditional risk as

$$\frac{d\mathbb{E}\{|S - \hat{s}| \ |\mathbf{X} = \mathbf{x}\}}{d\hat{s}} = 2F_{S|\mathbf{X}}(\hat{s}|\mathbf{x}) - 1 \tag{2.37}$$

where $F_{S|\mathbf{X}}(s|\mathbf{x})$ is the posterior distribution function of S given $\mathbf{X}$. Since this derivative must vanish at the minimum, we get $F_{S|\mathbf{X}}(\hat{s}_{\text{MAD}}|\mathbf{x}) = 1/2$. In other words, the absolute minimum error estimator is given by the median of $p_{S|\mathbf{X}}(s|\mathbf{x})$:

[4] $\frac{d}{dx} \int_{t_0}^{x} g(t) dt = g(x)$.

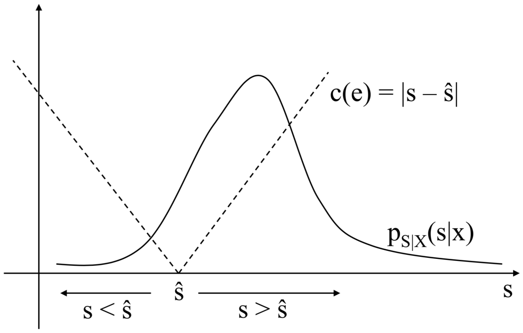

Fig. 2.4 Calculation of the posterior mean absolute error for a generic value $\hat{s}$.

$$\boxed{\hat{s}_{\mathrm{MAD}} = \mathrm{median}\{S|\mathbf{X} = \mathbf{x}\} \qquad (2.38)}$$

Remember that the median of a distribution is the point that separates that distribution into two regions that have the same probability, so the minimum mean absolute error estimator will verify that

$$P\{S > \hat{s}_{\mathrm{MAD}}\} = P\{S < \hat{s}_{\mathrm{MAD}}\}$$

In practice, this can be computed as the solution of

$$\int_{-\infty}^{\hat{s}_{\mathrm{MAD}}} p_{S|\mathbf{X}}(s|\mathbf{x})ds = \frac{1}{2} \tag{2.39}$$

Example 2.7 (Design of a Minimum Mean Absolute Deviation Estimator)

In the scenario of the example 2.1, the posterior distribution of S given X is uniform between 0 and x, the median of which is $x/2$. Thus,

$$\hat{s}_{\mathrm{MAD}} = \frac{1}{2}x \tag{2.40}$$

Note that, in this case, the MAD estimator matches the MSE obtained at (2.30). This is a consequence of the symmetry of the a posterior distribution.

2.4 Estimation with constrains

2.4.1 General principles

Occasionally, it might be beneficial to prescribe a specific parametric form for the estimator, denoted as $\hat{S} = f_{\mathbf{w}}(\mathbf{X})$, where $\mathbf{w}$ represents a vector of parameters. For instance, in scenarios involving two observations, $\mathbf{X} = [X_1, X_2]^T$, a design constraint might necessitate limiting the search for an estimator to the family of quadratic estimators, characterized by $\hat{S} = w_0 + w_1 X_1^2 + w_2 X_2^2$. In such situations, the task of designing an estimator consists on identifying the optimal parameter vector $\mathbf{w}^*$ that minimizes the risk, while adhering to the specified constraints on the estimator structure:

$$\begin{aligned}\mathbf{w}^* &= \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbb{E}\{c(S,\hat{S})\} = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbb{E}\{c(S, f_{\mathbf{w}}(\mathbf{X}))\} \\ &= \underset{\mathbf{w}}{\operatorname{argmin}} \int_{\mathbf{x}} \int_s c(s, f_{\mathbf{w}}(\mathbf{x})) p_{S,\mathbf{X}}(s,\mathbf{x}) ds d\mathbf{x}. \end{aligned} \tag{2.41}$$

Restricting the estimator to a specific analytic form typically results in a higher risk than what could be achieved with a Bayesian estimator tailored to the same cost function. The exception to this rule occurs when the imposed constraints align with the optimal estimator form, essentially when the Bayesian estimator inherently fits within the specified constraints. Despite potentially incurring higher costs, practical considerations may justify opting for such constrained estimators, such as simplification in design or implementation. An exploration of this concept is presented in Section 2.4.2, focusing on linear estimators that achieve minimum MSE.

Example 2.8 (Calculating an Estimator with Constrains)

Continuing the example 2.5, we want to calculate the minimum MSE estimator that has the form $\hat{s} = wx^2$. Starting from the conditional risk calculated in (2.28), the expression of the global average cost can be obtained as

$$\mathbb{E}\{c(S,\hat{S})\} = \int_x \mathbb{E}\{c(S,\hat{s})|X = x\}\, p_X(x) dx \tag{2.42}$$

$$= \int_x \left(\frac{1}{3}x^2 - \hat{s}x + \hat{s}^2\right) p_X(x) dx \tag{2.43}$$

Forcing $\hat{s} = wx^2$ and taking into account that $p_X(x) = 1$ for $0 < x < 1$, we get the MSE as a function of w.

$$\mathbb{E}\{c(S, wX^2)\} = \int_0^1 \left(\frac{1}{3}x^2 - wx^3 + w^2x^4\right) dx \tag{2.44}$$

$$= \frac{1}{9} - \frac{1}{4}w + \frac{1}{5}w^2 \tag{2.45}$$

The value w^* that optimizes (2.45) can be calculated by differentiation:

$$\frac{d}{d\hat{w}} \mathbb{E}\{c(S, w\mathbf{X}^2)\}\Big|_{w=w^*} = -\frac{1}{4} + \frac{2}{5}w^* = 0, \tag{2.46}$$

Este matiz es importante. Los estimadores que se discuten aquí son lineales en los parámetros, pero pueden ser no lineales en la observación. El término “estimación lineal” se puede referir a una cosa o la otra. Creo que aquí queda claro por el contexto a qué se refiere en cada momento, pero en todo caso es bueno que al final del tema les quede clara la dualidad.

$$w^* = \frac{5}{8}, \tag{2.47}$$

and therefore the estimator is $\hat{s} = \frac{5}{8}x^2$.

2.4.2 Linear (in the parameters) estimation of minimum MSE

Esta sección la he cambiado bastante. Ahora uso la tranformación genérica en (2.49) que permite una formulación más compacta, más breve (menos fórmulas) y (para mí), más clara, a un coste bajo (es un poco más abstracta, quizás)

In this section we will focus on the study of estimators that are a linear combination of variables related to the observations, using the minimization of the MSE as design criterion. More specifically, we will consider estimators given by the general expression

$$\hat{S} = \mathbf{w}^\intercal \mathbf{Z} \tag{2.48}$$

where

$$\mathbf{Z} = \phi(\mathbf{X}) \tag{2.49}$$

is some known transformation of the observations. The nature of this transformation may depend on the application, but there are some cases of particular interest:

- **Linear estimation**: in this case, ϕ is the identity function, so that $\mathbf{Z} = \mathbf{X}$, and the estimator a linear function of the observations

$$\hat{S} = w_0 X_0 + w_1 X_1 + \cdots + w_{N-1} X_{N-1} \tag{2.50}$$

- **Linear estimation with independent term**: in this case, $\mathbf{Z} = (1, \mathbf{X}^\intercal)^\intercal$ and, thus the estimator includes a constant term w_0

$$\hat{S} = w_0 + w_1 X_0 + \cdots + w_N X_{N-1} \tag{2.51}$$

- **Polynomial estimation**: in this case, the components of $\mathbf{Z}$ are monomials of the observational variables. For instance, for a scalar observation $X \in \mathbb{R}$, we can take $\mathbf{Z} = (1, X, X^2, \ldots X^M)$, for some $M > 1$, and the estimation becomes a polynomial of the observation with degree M.

$$\hat{S} = w_0 + w_1 X + w_2 X^2 \cdots + w_M X^M \tag{2.52}$$

Note that in general, the dimensions of $\mathbf{X}$ and $\mathbf{Z}$ may be different. Note, also, that, despite the estimator may be a non-linear function of the observation, all estimators given by (2.48) are linear functions of the parameters. For this reason we will refer to them as estimators that are *linear in the parameters*.

By imposing a restriction on the analytic form of the estimator, linear-in-the-parameter estimators will generally obtain lower performance than the optimal Bayesian estimator. However, the interest of linear estimators is justified by their simplicity and ease of design. As we shall see, the linear estimator of minimum MSE depends exclusively on first and second order statistical moments (means and covariances) associated with the target variable and the transformed observation, $\mathbf{Z}$.

2.4.2.1 Minimization of the mean squared error.

We will consider as design criteria the squared error, $c(e) = (s - \hat{s})^2$, so the optimal weight vector will be the one that minimizes the MSE risk:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\ \mathbb{E}\{(S - \hat{S})^2\} = \underset{\mathbf{w}}{\operatorname{argmin}}\ \mathbb{E}\{(S - \mathbf{w}^\intercal \mathbf{Z})^2\} \tag{2.53}$$

and we will refer to the estimator associated with weight vector as $\hat{S}_{\text{LMSE}}$:

$$\hat{S}_{\text{LMSE}} = \mathbf{w}^{*T}\mathbf{Z}$$

The MSE can be expanded as

$$\begin{aligned} MSE &= \mathbb{E}\{(S - \hat{S})^2\} = \mathbb{E}\{(S - \mathbf{w}^\intercal \mathbf{Z})^2\} \\ &= \mathbb{E}\{S^2\} - 2\mathbb{E}\{\mathbf{w}^\intercal \mathbf{Z} S\} + \mathbb{E}\{(\mathbf{w}^\intercal \mathbf{Z})^2\} \\ &= \mathbb{E}\{S^2\} - 2\mathbb{E}\{S\mathbf{Z}\}^\intercal \mathbf{w} + \mathbf{w}^\intercal \mathbb{E}\{\mathbf{Z}\mathbf{Z}^\intercal\}\mathbf{w} \\ &= \mathbb{E}\{S^2\} - 2\mathbf{r}_{S\mathbf{Z}}^\intercal \mathbf{w} + \mathbf{w}^\intercal \mathbf{R}_{\mathbf{Z}} \mathbf{w} \end{aligned} \tag{2.54}$$

where

- $\mathbf{r}_{S\mathbf{Z}} = \mathbb{E}\{S\mathbf{Z}\}$ is the **cross-correlation vector** between S and $\mathbf{Z}$
- $\mathbf{R}_{\mathbf{Z}} = \mathbb{E}\{\mathbf{Z}\mathbf{Z}^\intercal\}$ is the **autocorrelation matrix** of $\mathbf{Z}$.

Thus, the MSE is a second-degree polynomial in $\mathbf{w}$. This is illustrated in Figure 2.5, which depicts the error surface in a case with two observations. Being the function to minimize quadratic in weights (minimization argument), the error surface will take the form of a N dimensional paraboloid.

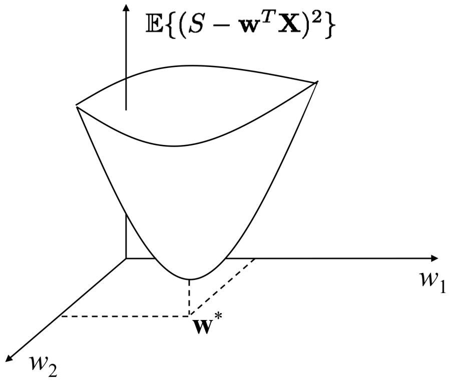

Fig. 2.5 Surface of the MSE for the linear estimator with $\mathbf{Z} = \mathbf{X}$, as a function of the parameters.

Since the MSE is non-negative, it is guaranteed to be a convex function of **w** and, thus, its minimum must be located at a point with zero gradient[5] (with respect to **w**):

$$\nabla_{\mathbf{w}} MSE|_{\mathbf{w}=\mathbf{w}^*} = -2\mathbf{r}_{S\mathbf{Z}} + 2\mathbf{R}_{\mathbf{Z}}\mathbf{w}^* = \mathbf{0} \tag{2.55}$$

therefore, the optimal weight vector is any solution of

$$\mathbf{R}_{\mathbf{Z}}\mathbf{w}^* = \mathbf{r}_{S\mathbf{Z}} \tag{2.56}$$

If the autocorrelation matrix in invertible, we get

$$\boxed{\mathbf{w}^* = \mathbf{R}_{\mathbf{Z}}^{-1}\mathbf{r}_{S\mathbf{Z}} \qquad (2.57)}$$

2.4.2.2 Alternative expression for the linear case

We can obtain an alternative expression for the optimal weights in the case of the linear estimator with independent term in (2.51), which is given by $\mathbf{Z} = (1, \mathbf{X}^\intercal)^\intercal$ and can thus be written as

$$\hat{S} = w_0 + \mathbf{w}_{1:}^\intercal \mathbf{X} \tag{2.58}$$

where $\mathbf{w}_{1:}$ is the weight vector after removing the first component, w_0. To do so, we can re-express (2.56) as the equivalent pair of equations

Este desarrollo no lo cuento en clase. A lo sumo, se puede contar la fórmula final (2.69), sin deducir, porque vuelve a aparecer en el caso Gaussiano, y permite conectar el estimador Gaussiano con el lineal (sec. 2.5.3).

$$\mathbb{E}\{w_0^* + \mathbf{w}_{1:}^{*\intercal}\mathbf{X}\} = \mathbb{E}\{S\} \tag{2.59}$$

$$\mathbb{E}\{\mathbf{X}\}w_0^* + \mathbb{E}\{\mathbf{X}\mathbf{X}^\intercal\}\mathbf{w}_{1:}^* = \mathbb{E}\{S\mathbf{X}\} \tag{2.60}$$

that is

$$w_0^* = m_S - \mathbf{w}_{1:}^{*\intercal}\mathbf{m}_{\mathbf{x}} \tag{2.61}$$

$$\mathbb{E}\{\mathbf{X}\mathbf{X}^\intercal\}\mathbf{w}_{1:}^* = \mathbb{E}\{S\mathbf{X}\} - w_0^*\mathbf{m}_{\mathbf{X}} \tag{2.62}$$

where $m_S = \mathbb{E}\{S\}$, $m_{\mathbf{X}} = \mathbb{E}\{\mathbf{X}\}$. Now using the expressions that relate the correlation and covariance of two variables:

$$\mathbb{E}\{S\mathbf{X}\} = \mathbf{v}_{S\mathbf{X}} + m_S\mathbf{m}_{\mathbf{X}} \tag{2.63}$$

$$\mathbb{E}\{\mathbf{X}\mathbf{X}^\intercal\} = \mathbf{V}_{\mathbf{X}} + \mathbf{m}_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}^\intercal \tag{2.64}$$

we get

$$w_0^* = m_S - \mathbf{w}_{1:}^{*\intercal}\mathbf{m}_{\mathbf{x}} \tag{2.65}$$

$$\mathbf{v}_{S\mathbf{X}} + m_S\mathbf{m}_{\mathbf{X}} = (\mathbf{V}_{\mathbf{X}} + \mathbf{m}_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}^T)\mathbf{w}_{1:}^* + w_0^*\mathbf{m}_{\mathbf{X}} \tag{2.66}$$

[5] The gradient of a function scale $f(\mathbf{w})$ with respect to the vector **w** is defined as a vector formed by the derivatives of the function with respect to each one of the components of **w**: $\nabla_{\mathbf{w}} f(\mathbf{w}) = \left[\frac{\partial f}{\partial w_1}, \ldots \frac{\partial f}{\partial w_N}\right]^T$.

Solving these equations for w_0^* and $w_{1:}^*$ we get

$$
\begin{aligned}
w_0^* &= m_S - \mathbf{w}_{1:}^{*\,T}\mathbf{m}_{\mathbf{x}} && (2.67)\\
\mathbf{w}_{1:}^* &= \mathbf{V}_{\mathbf{X}}^{-1}\,\mathbf{v}_{S,\mathbf{X}} && (2.68)
\end{aligned}
$$

The optimal estimator is, thus,

$$\hat{S}_{\text{LMSE}} = m_S + \mathbf{v}_{S\mathbf{X}}^{\intercal}\mathbf{V}_{\mathbf{X}}(\mathbf{X} - \mathbf{m}_{\mathbf{X}}) \tag{2.69}$$

The bias term w_0 compensates for differences between the means of the target variable and the observations. Therefore, when all the variables involved have zero mean, $w_0^* = 0$, and the estimator is purely linear in the observations.

2.4.2.3 Minimum squared mean error

The minimum MSE can be computed by replacing the optimal weights (2.57) into (2.70)

$$
\begin{aligned}
MSE^* &= \mathbb{E}\{S^2\} - 2\mathbf{r}_{S\mathbf{Z}}^{\intercal}\mathbf{R}_{\mathbf{Z}}^{-1}\mathbf{r}_{S\mathbf{Z}} + (\mathbf{R}_{\mathbf{Z}}^{-1}\mathbf{r}_{S\mathbf{Z}})^{\intercal}\mathbf{R}_{\mathbf{Z}}\mathbf{R}_{\mathbf{Z}}^{-1}\mathbf{r}_{S\mathbf{Z}}\\
&= \mathbb{E}\{S^2\} - \mathbf{r}_{S\mathbf{Z}}^{\intercal}\mathbf{R}_{\mathbf{Z}}^{-1}\mathbf{r}_{S\mathbf{Z}} && (2.70)
\end{aligned}
$$

For the linear estimator (2.69)

$$MSE^* = v_S - \mathbf{w}^{*T}\mathbf{v}_{S\mathbf{X}} \tag{2.71}$$

2.5 Estimation with Gaussian distributions

In this section, we delve into the estimation of random variables within the context where the combined distribution of all involved variables (the target variable along with the observational variables) is a multidimensional Gaussian. This scenario is particularly significant due to the prevalent occurrence of these distributions in signal processing, communications and various other fields.

When the joint distribution $p_{S,\mathbf{X}}(s,\mathbf{x})$ is Gaussian, all marginal and conditional distributions retain a Gaussian form. In particular, the posterior distribution, $p_{S|\mathbf{X}}(s|\mathbf{x})$ is Gaussian. Since the mean, median, and mode of the Gaussian distribution align, $\hat{s}_{\text{MSE}} = \hat{s}_{\text{MAD}} = \hat{s}_{\text{MAP}}$. Thus, our discussion in this section will primarily concentrate on deriving the estimator that minimizes the MSE.

Besides, we will demonstrate that the minimum MSE estimator and, consequently, the MAP and MAD estimators are linear, which will allow us to use the results shown in the previous section for minimum MSE estimation.

Este desarrollo tampoco lo hago en clase. De hecho, esta seccion me la salto, y solamente les pongo el caso multidimensional.

2.5.1 One dimensional case

We will consider as a starting point a case with one-dimensional random variables with zero means, in which the joint distribution of X and S has the following form:

$$p_{S,X}(s,x) \sim G\left(\begin{bmatrix} s \\ x \end{bmatrix}, \begin{bmatrix} v_X & \rho \\ \rho & v_S \end{bmatrix}\right) \tag{2.72}$$

where ρ is the covariance between the two random variables.

From this joint distribution we can obtain any other distribution involving the variables S and X; specifically, the posterior distribution can be obtained as:

$$p_{S|X}(s|x) = \frac{p_{S,X}(s,x)}{p_X(x)} = \frac{\frac{1}{2\pi\sqrt{v_X v_S - \rho^2}} \exp\left[-\frac{1}{2(v_X v_S - \rho^2)} \begin{bmatrix} s \\ x \end{bmatrix}^\intercal \begin{bmatrix} v_X & -\rho \\ -\rho & v_S \end{bmatrix} \begin{bmatrix} s \\ x \end{bmatrix}\right]}{\frac{1}{\sqrt{2\pi v_X}} \exp\left[-\frac{x^2}{2v_X}\right]} \tag{2.73}$$

where it has been necessary to calculate the inverse of the covariance matrix of S and X.

Noting that, as a function of s, $p_{S|X}(s|x)$ differs from $p_{S,X}(s,x)$ in the scale factor $p_X(x)$, which does not depend on s, $p_{S|X}(s|x)$ should be a Gaussian pdf too. Therefore, it must be expressed as

$$p_{S|X}(s|x) = \frac{1}{\sqrt{2\pi v_{S|X}}} \exp\left[-\frac{(s - m_{S|X})^2}{2v_{S|X}}\right] \tag{2.74}$$

where $m_{S|X}$ and $v_{S|X}$ are the posterior mean and variance, respectively, to be determined.

Joining (2.73) and (2.74), we can write

$$\frac{2\pi\sqrt{v_Xv_S-\rho^2}}{\sqrt{2\pi v_{S|X}}\sqrt{2\pi v_X}}\exp\left[-\frac{(s-m_{S|X})^2}{2v_{S|X}}+\frac{1}{2(v_Xv_S-\rho^2)}\begin{bmatrix}s\\x\end{bmatrix}^T\begin{bmatrix}v_X&-\rho\\-\rho&v_S\end{bmatrix}\begin{bmatrix}s\\x\end{bmatrix}-\frac{x^2}{2v_X}\right]=1 \tag{2.75}$$

which can be simplified to

$$\frac{\sqrt{v_Xv_S-\rho^2}}{\sqrt{v_{S|X}v_X}}\exp\left[-\frac{s^2-2m_{S|X}s+m_{S|X}^2}{2v_{S|X}}+\frac{v_Xs^2-2\rho xs+v_Sx^2}{2(v_Xv_S-\rho^2)}-\frac{x^2}{2v_X}\right]=1 \tag{2.76}$$

Note that the equation above must be satisfied for any $s\in\mathbb{R}$. Since the right-hand side is constant, and the exponent on the left-hand side is a polynomial function of s, the coefficients multiplying s^2 and s must be zero. Thus

$$\frac{m_{S|X}}{v_{S|X}}=\frac{\rho x}{v_Xv_S-\rho^2} \tag{2.77}$$

$$\frac{1}{v_{S|X}}=\frac{v_X}{v_Xv_S-\rho^2} \tag{2.78}$$

From (2.78) we get the posterior variance

$$\boxed{v_{S|X}=v_S-\frac{\rho^2}{v_X} \qquad (2.79)}$$

and, replacing (2.79) into (2.77) we get the posterior mean, which is the minimum MSE estimate

$$\boxed{\hat{s}_{\text{MSE}}=\hat{s}_{\text{MAD}}=\hat{s}_{\text{MAP}}=m_{S|X}=\frac{\rho}{v_X}x \qquad (2.80)}$$

Note that the estimator is a <u>linear</u> function of the observation.

Exercise 2.1 Generalize the above result for the case where the variables S and X have (non-zero) means m_S and m_X, respectively. Demonstrate that in such a case, the estimator is

$$\hat{s}_{\text{MSE}}=m_S+\frac{\rho}{v_X}(x-m_X) \tag{2.81}$$

Example 2.9 (Estimation of a Gaussian signal contaminated by Gaussian noise)

In this example we will consider the case in which the observation is the sum of the target variable and an independent noise component: $X=S+R$. Both the target and the noise are zero-mean Gaussian random variables with variances v_S and v_R, respectively.

Figure (2.6) represents the situation described for a case with $v_S<v_R$.

According to (2.80), for the resolution of the problem we must find the variance of X and the covariance between S and X (ρ). The variance v_X is obtained simply as the sum of v_S and v_R because both are independent variables. For the covariance calculation,

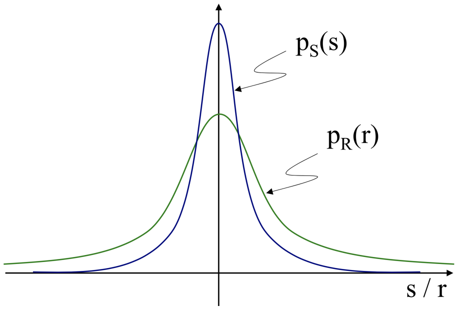

Fig. 2.6 Estimation of Gaussian random variable S contaminated by Gaussian noise R.

$$\rho = \mathbb{E}\{(X-m_X)(S-m_S)\} = \mathbb{E}\{X\,S\} = \mathbb{E}\{(S+R)S\} = \mathbb{E}\{S^2\} + \mathbb{E}\{S\,R\} = v_S \tag{2.82}$$

where independence of S and R has been used, and the fact that all variables (including X) have zero means.

Replacing these results in (2.80) we get

$$\hat{s}_{\text{MSE}} = \frac{v_S}{v_S + v_R}x \tag{2.83}$$

This result can be interpreted quite intuitively: when the variance of the noise is much lower than that of the signal (high Signal to Noise Ratio (SNR), $v_S \gg v_R$) we get $\hat{s}_{\text{MSE}} \to x$, which makes sense since the effect of the noise component in this case is not very significant; on the contrary, when the SNR is very small ($v_S \ll v_R$), the observation barely provides information about the S value in each experiment, so the estimator keeps the mean value of the signal component, $\hat{s}_{\text{MSE}} \to 0$.

2.5.2 Case with multidimensional variables

In a general multidimensional case, $\mathbf{S}$ and $\mathbf{X}$ can be random vectors of dimensions N and M, respectively, with joint Gaussian distribution.

$$p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x}) \sim G\left(\begin{bmatrix}\mathbf{m}_\mathbf{S}\\ \mathbf{m}_\mathbf{X}\end{bmatrix}, \begin{bmatrix}\mathbf{V}_\mathbf{S} & \mathbf{V}_{\mathbf{SX}}\\ \mathbf{V}_{\mathbf{SX}}^T & \mathbf{V}_\mathbf{X}\end{bmatrix}\right) \tag{2.84}$$

being $\mathbf{m}_\mathbf{S}$ and $\mathbf{m}_\mathbf{X}$ the means of $\mathbf{S}$ and $\mathbf{X}$, respectively, $\mathbf{V}_\mathbf{S}$ and $\mathbf{V}_\mathbf{X}$ the covariance matrix of $\mathbf{S}$ and $\mathbf{X}$, respectively, and $\mathbf{V}_{\mathbf{SX}}$ the matrix of crossed covariances of $\mathbf{S}$ and $\mathbf{X}$, that is,

$$\mathbf{V}_\mathbf{S} = \mathbb{E}\{(\mathbf{S}-\mathbf{m}_\mathbf{S})(\mathbf{S}-\mathbf{m}_\mathbf{S})^\intercal\} \tag{2.85}$$

$$\mathbf{V}_\mathbf{X} = \mathbb{E}\{(\mathbf{X}-\mathbf{m}_\mathbf{X})(\mathbf{X}-\mathbf{m}_\mathbf{X})^\intercal\} \tag{2.86}$$

$$\mathbf{V}_{\mathbf{SX}} = \mathbb{E}\{(\mathbf{S}-\mathbf{m}_\mathbf{S})(\mathbf{X}-\mathbf{m}_\mathbf{X})^\intercal\} \tag{2.87}$$

Esto es lo que cuento. En clase les pongo la fórmula de la gaussiana multidimensional (que aquí no está) porque muchos no la conocen. Lo importante es (2.90) y que vean la relación con el estimador lineal.

The calculation of the posterior distribution of $\mathbf{S}$ given $\mathbf{X}$ is more complex than in the one-dimensional case, but it follows a similar procedure, which we will omit here. It can be shown that the posterior distribution is gaussian with mean

$$\mathbf{m_{S|X}} = \mathbf{m_S} + \mathbf{V_{SX}}\mathbf{V_X}^{-1}(\mathbf{x} - \mathbf{m_X}) \tag{2.88}$$

and covariance

$$\mathbf{V_{S|X}} = \mathbf{V_S} - \mathbf{V_{SX}}\mathbf{V_X}^{-1}\mathbf{V}_{\mathbf{SX}}^{T} \tag{2.89}$$

Since the minimum MSE estimator of **S** given **X** is precisely the posterior mean, we can write

$$\boxed{\hat{\mathbf{s}}_{\mathrm{MSE}} = \hat{\mathbf{s}}_{\mathrm{MAD}} = \hat{\mathbf{s}}_{\mathrm{MAP}} = \mathbf{m_{S|X}} = \mathbf{m_S} + \mathbf{V_{SX}}\mathbf{V_X}^{-1}(\mathbf{x} - \mathbf{m_X}) \qquad (2.90)}$$

2.5.3 Linear estimation and Gaussian estimation

Note that, if the target variable is scalar, (2.90) is identical to (2.69). This is not coincidental: if the minimum MSE estimator in the Gaussian case is linear, it must be equal to the best linear estimator, which is given by (2.90).

Esta sección no la he tocado. También tiene que crecer un poco en el futuro (con una discusión sobre sesgos, varianzas y consistencia en función de l) pero por ahora se queda así.

A veces explico el ejemplo (2.10) pero otras explico esta sección haciendo algún ejercicio del boletín (si haces el ejemplo, no cuenta como “ejercicio hecho en clase”, pero si lo explicas con un ejercicio, sí ;-)

2.6 ML estimation of probability distributions parameters

Sometimes we may be interested in estimating the parameters of a probability distribution, such as the mean or variance of a Gaussian distribution, the decay parameter that characterizes an exponential distribution, or values a and b delimiting the interval in which a uniform distribution is defined.

In these cases, the prior distribution of these variables is not usually known, even more, in many cases, these parameters are said to be deterministic and they are not treated them as random parameters. However, if a set of observations generated from these distributions is available, we can obtain the likelihood of these variables and estimate their values with a maximum likelihood criteria.

Example 2.10 (ML estimate of the mean and variance of a one-dimensional Gaussian distribution)

The weight of individuals of a family of mollusks is known to obey a Gaussian distribution, but the mean and variance are unknown. The weight of l individuals taken independently, $\{X^{(k)}\}_{k=1}^{l}$, is available.

The likelihood of the mean and the variance for a single observation x, in this case, is given by:

$$p_X(x) = p_{X|m,v}(x|m,v) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-m)^2}{2v}\right] \tag{2.91}$$

Since we must construct the estimator based on the joint observation of l observations, we will need to calculate the joint distribution of all of them which, being independent observations, is obtained as the product of individual observations:

$$\begin{aligned} p_{\{X^{(k)}\}|m,v}(\{x^{(k)}\}|m,v) &= \prod_{k=1}^{l} p_{X|m,v}(x^{(k)}|m,v) \\ &= \frac{1}{(2\pi v)^{l/2}} \prod_{k=1}^{l} \exp\left[-\frac{(x^{(k)}-m)^2}{2v}\right] \end{aligned} \tag{2.92}$$

The ML estimators of m and v will be the values maximizing (2.92). The analytical form of this function suggests the use of the logarithm to simplify the maximization:

$$L = \ln\left[p_{\{X^{(k)}\}|m,v}(\{x^{(k)}\}|m,v)\right] = -\frac{l}{2}\ln(2\pi v) - \frac{1}{2v}\sum_{k=1}^{l}(x^{(k)}-m)^2 \tag{2.93}$$

Differentiating (2.93) with respect to m and v, we get the system of equations to solve

$$\begin{aligned} \left.\frac{d\,L}{d\,m}\right|_{\substack{m=\hat{m}_{\mathrm{ML}}\\ v=\hat{v}_{\mathrm{ML}}}} &= \left.-\frac{1}{v}\sum_{k=1}^{l}(x^{(k)}-m)\right|_{\substack{m=\hat{m}_{\mathrm{ML}}\\ v=\hat{v}_{\mathrm{ML}}}} = 0 \\ \left.\frac{d\,L}{d\,v}\right|_{\substack{m=\hat{m}_{\mathrm{ML}}\\ v=\hat{v}_{\mathrm{ML}}}} &= \left.-\frac{l}{2v} + \frac{1}{2v^2}\sum_{k=1}^{l}(x^{(k)}-m)^2\right|_{\substack{m=\hat{m}_{\mathrm{ML}}\\ v=\hat{v}_{\mathrm{ML}}}} = 0 \end{aligned} \tag{2.94}$$

2.7 Problems

2.1 The posterior distribution of S given X is

$$p_{S|X}(s|x) = x^2 \exp(-x^2 s), \qquad s \geq 0$$

Compute estimators $\hat{S}_{\mathrm{MMSE}}$, $\hat{S}_{\mathrm{MAD}}$ y $\hat{S}_{\mathrm{MAP}}$.

2.2 Consider an estimation problem given by the following posterior distribution:

$$p_{S|X}(s|x) = x \exp(-xs), \;\; s > 0 \tag{2.99}$$

Compute estimators $\hat{S}_{\mathrm{MMSE}}$, $\hat{S}_{\mathrm{MAD}}$ y $\hat{S}_{\mathrm{MAP}}$.

2.3 A r.v. S must be estimated from the observation of another r.v. X by means of a linear mean square error estimator given by:

$$\hat{S}_{\mathrm{LMSE}} = w_0 + w_1 X$$

Knowing that $\mathbb{E}\{X\} = 1$, $\mathbb{E}\{S\} = 0$, $\mathbb{E}\{X^2\} = 2$, $\mathbb{E}\{S^2\} = 1$ y $\mathbb{E}\{SX\} = 1/2$, compute:

a) The values for w_0 y w_1.
b) The mean square error of the estimator, $\mathbb{E}\left\{\left(S - \hat{S}_{\mathrm{LMSE}}\right)^2\right\}$.

2.4 (Linear estimation of minimum mean squared error) We want to construct a linear estimator of minimum mean squared error that will allow us to estimate the random variable S from the random variables X_1 and X_2. Knowing that

$$\begin{array}{lll}
\mathbb{E}\{S\} = 1/2 & \mathbb{E}\{X_1\} = 1 & \mathbb{E}\{X_2\} = 0 \\
\mathbb{E}\{S^2\} = 4 & \mathbb{E}\{X_1^2\} = 3/2 & \mathbb{E}\{X_2^2\} = 2 \\
\mathbb{E}\{SX_1\} = 1 & \mathbb{E}\{SX_2\} = 2 & \mathbb{E}\{X_1X_2\} = 1/2
\end{array}$$

get the weights from the desired estimator and calculate its squared mean error. Calculate the estimated value for the observation vector: $[X_1, X_2] = [3, 1]$.

2.5 Let X and S be two random variables with joint pdf

$$p_{X,S}(x,s) \begin{cases} 2 \; 0 < x < 1, 0 < s < x \\ 0 \text{ resto} \end{cases}$$

a) Compute the minimum mean square error estimate of S given X, $\hat{S}_{\mathrm{MMSE}}$.
b) Compute the risk of estimator $\hat{S}_{\mathrm{MMSE}}$.

2.6 A digitized image of dimensions 8x8 is available, whose luminance values are statistically independent and evenly distributed between 0 (white) and 1 (black); the image has been modified by applying a transformation of the form $Y = X^r$ on each pixel; $r > 0$, where X is the r.v. associated with the pixels of the original image and Y is associated with the transformed image. Obtain the expression that allows to estimate r by maximum likelihood given the 64 pixel values of the transformed image $\{y^{(k)}\}_{k=1}^{64}$, without knowing the original image.

Solving for m the first equation we get

$$\hat{m}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l} x^{(k)} \tag{2.95}$$

which is the sample average of the observations. On the other hand, we can solve the second equation for v to get

$$\hat{v}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l} (x^{(k)} - \hat{m}_{\mathrm{ML}})^2 \tag{2.96}$$

which is the sample variance. Note that, if instead of applying the estimation function (of m or v) on some specific observations we did it on generic values $\{X^{(k)}\}$, the estimators could be treated as random variables, i.e.,

$$\hat{M}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l} X^{(k)} \tag{2.97}$$

$$\hat{V}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l} [X^{(k)} - \hat{M}_{\mathrm{ML}}]^2 \tag{2.98}$$

2.7 For the design of a communication system it is desired to estimate the signal attenuation between the transmitter and the receiver, as well as the noise power introduced by the channel when this noise is Gaussian of zero mean and independent of the transmitted signal. For this, the transmitter sends a signal with a constant amplitude of 1 and the receiver collects a set of K observations available at its input.

a) Estimate the channel attenuation, α, and the noise variance, v_r, by maximum likelihood, when the available observations on the receiver are

$$\{0.55, 0.68, 0.27, 0.58, 0.53, 0.37, 0.45, 0.53, 0.86, 0.78\}.$$

b) If the system is to be used for the transmission of digital signals with unipolar coding (a A signal level is used to transmit a bit 1 and the signal level is maintained at 0 for the transmission of bit 0), considering equiprobability between symbols, indicate the minimum level of signal that should be used in the coding, $A_{\min}$, to guarantee a SNR level in the receiver of 3 dB.

2.8 Company *Like2Call* offers hosting services for call centers. In order to dimension the staff of operators the company is designing a statistical model to characterize the activity in the hosted call centers. One of the components of such model relies on the well-known fact that the times between incoming calls follow an exponential distribution

$$p_{X|S}(x|s) = s\ \exp(-s\,x), \qquad x > 0$$

where random variable X represents the time before a new call arrives, and S is the parameter of such distribution, that depends on the time of the day and each particular call-center service (e.g., attention to the clients of an insurance company, customers of an on-line bank, etc).

For random variable S, the following *a priori* model is assumed:

$$p_S(s) = \exp(-s), \qquad s > 0.$$

With this information, we would like to design an estimator of S that is based on the first K incoming calls for each implemented service and time interval, i.e., the observation vector is given by $\mathbf{x} = \left[x^{(0)}, x^{(1)}, \cdots, x^{(K-1)}\right]$, where all observations in the vector are assumed i.i.d.

a) Obtain the maximum likelihood estimator or S based on the observation vector $\mathbf{X}$, and verify that it depends just on the sum of all observations, $z = \sum_{k=0}^{K-1} x^{(k)}$.
b) Calculate the posterior distribution of S given $\mathbf{X}$, $p_{S|\mathbf{X}}(s|\mathbf{x})$.
c) Obtain the maximum *a posteriori* estimator of S given $\mathbf{X}$, $\hat{s}_{\mathrm{MAP}}$.
d) Obtain the minimum mean square error estimator of S given $\mathbf{X}$, $\hat{s}_{\mathrm{MSE}}$.
e) Calculate the mean square error given $\mathbf{X}$ of a generic estimator $\hat{S}$, and particularize the result for estimators of the following analytical shape $\hat{s}_c = \frac{c}{z+1}$.
f) Find expressions for the following probability density functions: $p_{Z|S}(z|s)$, $p_{Z,S}(z,s)$, and $p_Z(z)$.
g) Calculate the mean square error of a generic estimator $\hat{s}_c = \frac{c}{z+1}$. Study how the result changes with c and K.

You can use the following results:

i.

$$\int_0^\infty x^N \exp(-x)dx = N!$$

ii. If $f(x) = a\,exp(-a\,x),\ x > 0$ then

$$\underbrace{f(x) * f(x) * \cdots * f(x)}_{N \text{ times}} = \frac{a^N\,x^{N-1}}{(N-1)!} exp(-a\,x),\ x > 0$$

iii. For K an integer

$$\int_0^\infty \frac{K\,x^{K-1}}{(x+1)^{K+3}} dx = \frac{2}{(K+2)(K+1)}$$

Solution 2.1

a)

$$\begin{aligned}
&p_{\mathbf{X}|S}(\mathbf{x}|s) = s^K\ \exp(-s\,z), \qquad z > 0 \\
&\ln p_{\mathbf{X}|S}(\mathbf{x}|s) = K \ln s - s\,z \\
&\frac{d}{ds} \ln p_{\mathbf{X}|S}(\mathbf{x}|s) = \frac{K}{s} - z \\
&\hat{s}_{\mathrm{ML}} = \frac{K}{z}
\end{aligned}$$

b)

$$p_{\mathbf{X},S}(\mathbf{x},s) = p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s) = s^K\ \exp[-s(z+1)]$$

(note the expression above is not the joint pdf of Z and S)

$$p_{\mathbf{X}}(\mathbf{x}) = \int p_{\mathbf{X},S}(\mathbf{x},s)\ ds = \int_0^\infty s^K\ \exp[-s(z+1)]\ ds$$

With the change of variable $s' = s(z+1)$ the previous integral can be simplified using expression (i), and we get

$$p_{S|\mathbf{X}}(s|\mathbf{x}) = \frac{p_{\mathbf{X},S}(\mathbf{x},s)}{p_{\mathbf{X}}(\mathbf{x})} = \frac{(z+1)^{K+1}\ p_{\mathbf{X},S}(\mathbf{x},s)}{K!} = \frac{s^K (z+1)^{K+1}\ \exp[-s(z+1)]}{K!}$$

c)

$$\begin{aligned}
&\hat{s}_{\mathrm{MAP}} = \arg\max_s p_{S|\mathbf{X}}(s|\mathbf{x}) = \arg\max_s p_{\mathbf{X},S}(\mathbf{x},s) \\
&\ln p_{\mathbf{X},S}(\mathbf{x},s) = K \ln s - s\,(z+1) \\
&\frac{d}{ds} \ln p_{\mathbf{X},S}(\mathbf{x},s) = \frac{K}{s} - (z+1) \\
&\hat{s}_{\mathrm{MAP}} = \frac{K}{z+1}
\end{aligned}$$

d)

$$\hat{s}_{\text{MSE}} = \mathbb{E}\{S|\mathbf{x}\} = \int s\, p_{S|\mathbf{X}}(s|\mathbf{x})\, ds = \frac{(z+1)^{K+1}}{K!} \int_0^\infty s^{K+1}\ \exp[-s(z+1)]\, ds$$

Replacing again $s' = s(z+1)$ and using expression (i), we get

$$\hat{s}_{\text{MSE}} = \frac{K+1}{z+1}$$

e) The calculation is somehow tedious, but can be summarized as follows:

$$\begin{aligned}\mathbb{E}\left\{(S-\hat{s})^2|X\right\} &= \int_0^\infty (s-\hat{s})^2\ p_{S|X}(s|x)ds \\ &= \frac{(z+1)^{K+1}}{K!}\left[\frac{(K+2)!}{(z+1)^{K+3}} + \hat{s}^2\frac{K!}{(z+1)^{K+1}} - 2\hat{s}\frac{(K+1)!}{(z+1)^{K+2}}\right] \\ &= \frac{(K+2)(K+1)+c^2-2c(K+1)}{(z+1)^2}\end{aligned}$$

For the MAP and MSE estimators the expressions are substantially simplified:

$$\begin{aligned}\mathbb{E}\left\{(S-\hat{s}_{MAP})^2|z\right\} &= \frac{K+2}{(z+1)^2} \\ E\left\{(S-\hat{s}_{MSE})^2|z\right\} &= \frac{K+1}{(z+1)^2}\end{aligned}$$

f) Using the fact that Z is the sum of K i.i.d. variables (given S):

$$p_{Z|S}(z|s) = \underbrace{[s\ \exp(-s\,z)] * \cdots * [s\ \exp(-s\,z)]}_{K \text{ times}} = \frac{s^K\ z^{K-1}}{(K-1)!}\exp(-s\,z), \qquad z > 0$$

The joint pdf of Z and S can now be obtained as

$$p_{Z,S}(z,s) = p_{Z|S}(z|s)p_S(s) = \frac{s^K\ z^{K-1}}{(K-1)!}\exp[-s\ (z+1)], \qquad s,z > 0$$

Finally, integrating s out, we have

$$p_Z(z) = \int p_{Z,S}(z,s)ds = \frac{z^{K-1}}{(K-1)!}\int_0^\infty s^K \exp[-s\ (z+1)] = \frac{K\ z^{K-1}}{(z+1)^{K+1}},\ \ z > 0$$

g)

$$\mathbb{E}\left\{(S-\hat{S}_c)^2\right\} = \int \mathbb{E}\left\{(S-\hat{s}_c)^2|z\right\}\ p_Z(z)dz$$

Using the results from the previous two sections we can obtain an expression that depends on the value of an integral over z:

$$\mathbb{E}\left\{(S-\hat{S}_c)^2\right\} = \left[(K+2)(K+1)+c^2-2c(K+1)\right]\int_0^\infty \frac{K\ z^{K-1}}{(z+1)^{K+3}}dz$$

The value of the integral is given in (iii). Simplifying also for the MAP and MSE estimators:

$$\mathbb{E}\left\{(S-\hat{S}_{MAP})^2\right\} = \frac{2}{K+1}$$
$$E\left\{(S-\hat{S}_{MSE})^2\right\} = \frac{2}{K+2}$$