Analytic Estimation

Modern Theory of Detection and Estimation. Block-1: Estimation

Emilio Parrado-Hernández, emilio.parrado@uc3m.es

September 20, 2021

uc3m | Universidad Carlos III de Madrid

Index

Index

Estimation problem definition

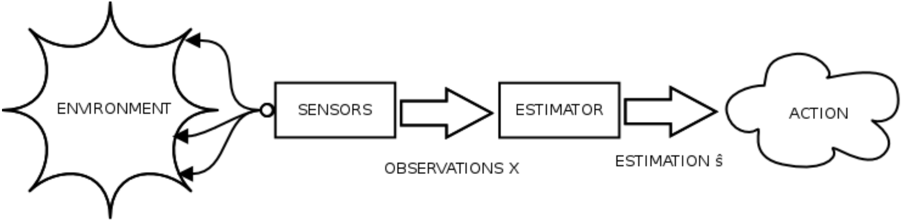
Goal in a estimation problem

You receive a scalar observation (x) or **a vector observation** ($\mathbf{x}$), and your task is to find out (**estimate**) the value of a real number s. The variable you want to find out can be either deterministic or random.

Examples:

- Estimate the number of stocks of a certain instrument that will be traded tomorrow.
- Estimate the daily energy consumption in a building as a function of the weather, season of year, day of week.
- Estimate the price of a house given the surface, number of rooms, size of the yard, distance to the centre of the town, etc.
- Guess the foot length of people based on their height.

Estimation problem setup



- Sensors sample information about the environment and construct observations $\mathbf{x}^k$, $k = 1, \ldots, l$.
- Estimator constructs estimated variable $\hat{s}$ as a function of the observations. There **must be** a statistical relationship between the observations and the variable we want to estimate, s.
- Then we use the estimation $\hat{s}$ instead of s to take actions. Remember we **never** have access to s (that's why we need to estimate it!!)

Notation

- $\mathbf{X}$: **observations**. Random in the analytical case setup. Could be scalars or vectors.
- S: **desired output**. This is the variable we want to estimate. It can be random (S) or deterministic (s). It can be a vector or an scalar, most of the time this course it will be an scalar.
- $\hat{S}$: **estimator**. $\hat{S} = f(\mathbf{X})$ is our approximation to the value of S given the observation. In the analytic estimation case $\hat{S}$ is random since it is a deterministic function ($f()$) of a random variable ($\mathbf{X}$), therefore its pdf can be obtained applying random variable transformation to the pdf of the observations.
- $p_{X,S}(\mathbf{x}, s)$: **statistical relationship between observations and estimated variable**. We can estimate S using $\mathbf{X}$ as input because they are related. We express this relationship as a joint pdf.

Index

Random variable vs. Deterministic parameter

The variable we want to estimate can be either random S, or deterministic s. Examples:

- Deterministic: I've thought the mean of a Gaussian distribution m. I draw K samples from the distribution and ask you to estimate m. Note that m is unknown but does not follow any prior distribution $p(m)$; m is a fixed [unknown] constant.
- Random: I pick at random one student from the class, I tell you his/her foot length ($X = x$) and you have to guess his/her height ($S = s$). S is random because I can pick any student in the class.

Analytic estimation vs. machine estimation

- **Machine estimation or regression**: the problem is defined in terms of a set of examples, $\{\mathbf{x}^k, s^k\}_{k=1}^l$. We use the examples to fit an estimation model, for example a linear regressor:

  $$\hat{s} = f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$

  we learn $\mathbf{w}$ and w_0 with the training set $\{\mathbf{x}^k, s^k\}_{k=1}^l$.
  Notice in the machine estimation setting the training observations $\{\mathbf{x}^k, s^k\}_{k=1}^l$ are not random (we do know them). The test observations can be considered random.
- **Analytic estimation**: the problem is defined in terms of a complete statistical characterisation: there exists a joint pdf $p_{\mathbf{X},S}(\mathbf{x}, s)$ and we know how to calculate it for any pair $(\mathbf{x}, s)$. In this setting $\mathbf{X}$ is random.
- **Semi-analytic estimation**: The problem is defined in terms of a data set, as in the regression case. We do not use this data to fit a regression model but to **learn the statistics**, i.e., estimate $p_{\mathbf{X},S}(\mathbf{x}, s)$. Once we have an estimation of the joint pdf (or any other useful pdf) we can apply analytic estimation.

Index

Index

Likelihood

We rely on different pdfs that capture the relationship between $\mathbf{x}$ and s.

Likelihood

$$p_{\mathbf{X}|S}(\mathbf{x}|s)$$

It's the key pdf. You will have a likelihood in both random and deterministic variable estimation cases.

The likelihood models the generation of observations for every possible value of the target variable S. If S changes, we get different statistics in the observations. That is why we can guess the value of S.

If s is deterministic there is no point in conditioning in the value of s since it is always the same. Anyway we use $p_{\mathbf{X}|s}(\mathbf{x}|s)$ to unify notations.

S is random

If S is a random variable we have three other pdfs:

- **Prior** or marginal of S: $p_S(s)$ gives information about how are the values of S distributed without access to the observations.
- **Joint distribution** of S and $\mathbf{X}$: $p_{\mathbf{X},S}(\mathbf{x}, s)$
- **Posterior** of S given the observation $\mathbf{X} = \mathbf{x}$: $p_{S|\mathbf{X}}(s|\mathbf{x})$ gives the values of S that concentrate a higher probability density for each particular observed value of $\mathbf{X}$. It is a key pdf in the design of estimators since to know $\mathbf{X} = \mathbf{x}$ **narrows** the uncertainty about S expressed in the prior (I mean $p_{S|\mathbf{X}}(s|\mathbf{x})$ as a function of s is “narrower” than $P_S(s)$).

Index

Cost Functions

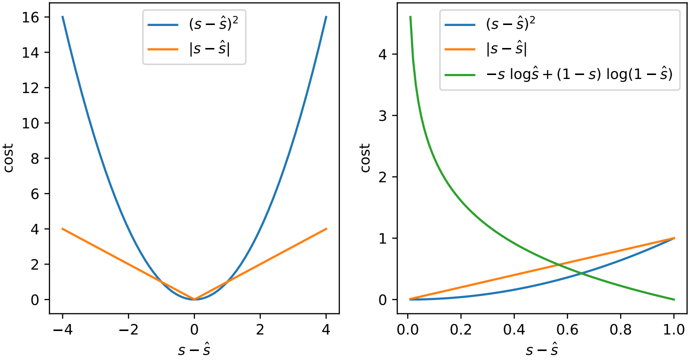
- An estimator is a function of the observations: $\hat{s} = f(\mathbf{x})$.
- The of design an estimator involves an optimisation to pick the best estimation function $f^*()$ from within a family or set of candidates.
- We need a criterion to select this best estimation function from the set of potential candidates. This criterion is the **cost function**.

Cost function

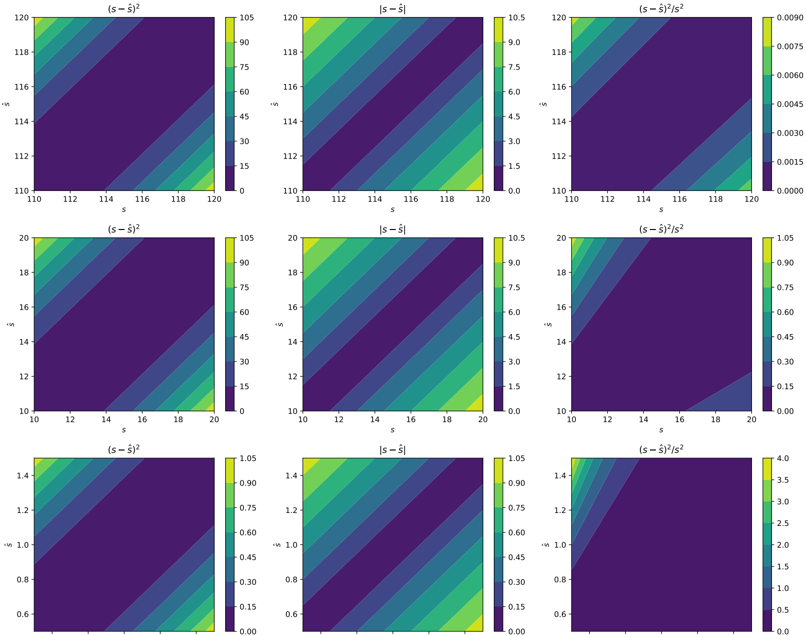
Function of the estimator and of the estimated variable: $C(S, \hat{S})$. It compares both quantities and determines the penalty in which we incur if we approximate $S = s$ with $\hat{S} = \hat{s}$. In most cases exact estimations ($\hat{s} = s$) yield zero cost: $C(s, \hat{s} = s) = 0$. The cost is always positive $C(S, \hat{S}) \geq 0$.

Examples of cost function

- Quadratic Cost: $c(e) = e^2$.
- Absolute Value of the error: $c(e) = |e|$.
- Relative Square Error: $c(s, \hat{s}) = \frac{(s-\hat{s})^2}{s^2}$
- Log loss: $c(s, \hat{s}) = -s \ln \hat{s} - (1-s)\ln(1-\hat{s})$, for $s \in \{0, 1\}$, and $\hat{s} \in [0, 1]$



Examples of cost function



Index

Expected Cost

Since we plan to use the estimator a large number of times, we are more interested in the expected cost, that takes into account all the possible values of S and $\hat{S}$.

Expected cost

$$\mathbb{E}\{c(S,\hat{S})\} = \int_{\mathbf{x}}\int_{s} c(s,\hat{s}(\mathbf{x}))\mathbf{p}_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x})\mathbf{dsdx}$$

Notice that $\hat{S} = \hat{S}(\mathbf{X})$ is a deterministic function of $\mathbf{X}$, therefore the statistics of $\hat{S}$ can be univocally defined in terms of the statistics of $\mathbf{X}$ (transformation of random variable).

Example

Let X be a noisy observation of S, such that

$$X = S + R$$

with S a random variable with mean 0 and variance 1, and R a Gaussian random variable, independent of S, with mean 0 and variance v. Consider the estimator $\hat{S} = X$, the expected quadratic cost is

Example

Let X be a noisy observation of S, such that

$$X = S + R$$

with S a random variable with mean 0 and variance 1, and R a Gaussian random variable, independent of S, with mean 0 and variance v. Consider the estimator $\hat{S} = X$, the expected quadratic cost is

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

The expected absolute cost is

Example

Let X be a noisy observation of S, such that

$$X = S + R$$

with S a random variable with mean 0 and variance 1, and R a Gaussian random variable, independent of S, with mean 0 and variance v. Consider the estimator $\hat{S} = X$, the expected quadratic cost is

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

The expected absolute cost is

$$\begin{aligned} \mathbb{E}\{|S - \hat{S}|\} = \mathbb{E}\{|R|\} &= \int_{-\infty}^{\infty} |r| \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{r^2}{2v}\right) dr \\ &= 2\int_0^{\infty} r \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{r^2}{2v}\right) dr = \sqrt{\frac{2v}{\pi}} \end{aligned}$$

Example: Calculation of Expected Quadratic Cost

Two random variables S and X follow a joint pdf

$$p_{S,X}(s,x) = \begin{cases} \frac{1}{x}, & 0 < s < x, \quad 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Consider two estimators $\hat{S}_1 = \frac{1}{2}X$ and $\hat{S}_2 = X$. Which is the best estimator from the point of view of minimising the quadratic cost?

Example: Calculation of Expected Quadratic Cost

Two random variables S and X follow a joint pdf

$$p_{S,X}(s,x) = \begin{cases} \frac{1}{x}, & 0 < s < x, \quad 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Consider two estimators $\hat{S}_1 = \frac{1}{2}X$ and $\hat{S}_2 = X$. Which is the best estimator from the point of view of minimising the quadratic cost? To find it out, we'll compute the mean square error of both estimators.

For a general w,

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

Example: Calculation of Expected Quadratic Cost (ctd)

If $w = 1/2$

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

Alternatively, if $w = 1$

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

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

Index

Bayesian estimation

Given a cost function and a statistical relationship (joint distribution) between observations and estimated variable, the Bayesian Estimator is the one that minimizes the expected cost.

$$\hat{S}_{\text{Bayes}} = \arg\min_{\hat{S}} \mathbb{E}\{c(S, \hat{S})\} = \arg\min_{\hat{S}} \int_{\mathbf{x}} \int_{s} c(s, \hat{s}) p_{S,\mathbf{X}}(s, \mathbf{x}) ds d\mathbf{x}$$

$$= \arg\min_{\hat{S}} \int_{\mathbf{x}} \int_{s} c(s, \hat{s}) p_{S|\mathbf{X}}(s|\mathbf{x}) ds \; p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} =$$

$$\arg\min_{\hat{S}} \int_{\mathbf{x}} \mathbb{E}\{c(S, \hat{S})|\mathbf{X}\} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

Bayesian estimation

If we look carefully into the last integral, we see that $\mathbb{E}\{c(S,\hat{S})|\mathbf{X}\}$ is multiplied by a positive function $(p_{\mathbf{X}}(\mathbf{x}))$. Therefore, to minimize the expected global cost is to minimise the integral and it is achieved minimising $\mathbb{E}\{c(S,\hat{S})|\mathbf{X}\}$. Thus a first result is

$$\hat{S}_{\text{Bayes}} = \arg\min_{\hat{S}} \mathbb{E}\{c(S,\hat{S})|\mathbf{X}\}$$

In summary, to compute a Bayesian Estimator we need to take two steps:

1. Select a cost function $c(S,\hat{S})$.
2. Minimise $\mathbb{E}\{c(S,\hat{S})|\mathbf{X}\}$

Example: Find Bayesian estimation with Quadratic Cost

Two random variables S and X follow a joint pdf

$$p_{S,X}(s,x) = \begin{cases} \frac{1}{x}, & 0 < s < x, \quad 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Find the Bayesian estimator that minimizes the quadratic cost

Example: Find Bayesian estimation with Quadratic Cost

Two random variables S and X follow a joint pdf

$$p_{S,X}(s,x) = \begin{cases} \frac{1}{x}, & 0 < s < x, \quad 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Find the Bayesian estimator that minimizes the quadratic cost

First, find the posterior

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

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

Where you should apply that

$$p_X(x) = \int_{(s)} p_{S,X}(s,x)ds = \int_0^x \frac{1}{x} ds = 1$$

Example: Find Bayesian estimation with Quadratic Cost

The expected cost given x:

Example: Find Bayesian estimation with Quadratic Cost

The expected cost given x:

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

The expected cost given x, as a function of $\widehat{s}$, is a 2nd degree polynomial. We can minimize it by equating its derivative to 0.

Example: Find Bayesian estimation with Quadratic Cost

The expected cost given x:

$$\begin{aligned}\mathbb{E}\{c(S,\hat{S})|\mathbf{X}=\mathbf{x}\} &= \mathbb{E}\{(S-\widehat{s})^2)|\mathbf{X}=\mathbf{x}\} \\ &= \frac{1}{3}x^2 - \widehat{s}x + \widehat{s}^2\end{aligned}$$

The expected cost given x, as a function of $\widehat{s}$, is a 2nd degree polynomial. We can minimize it by equating its derivative to 0.

$$\frac{d}{d\widehat{s}}\mathbb{E}\{c(S,\hat{S})|\mathbf{X}=\mathbf{x}\} = -x + 2\widehat{s} = 0$$

And the solution is:

$$\widehat{s}_{\text{Bayes}} = \frac{1}{2}x$$

Index

Minimum Mean Square Error (MMSE) Estimator

This is the Bayesian Estimator when the cost function is the quadratic error: $c(s, \hat{s}) = e^2 = (s - \hat{s})^2$. Therefore the estimator is the solution to the optimisation problem given by

$$\hat{s}_{\text{MMSE}} = \arg\min \hat{s}\mathbb{E}\{c(S, \hat{s})|\mathbf{X} = \mathbf{x}\} = \arg\min \hat{s} \int_s (s - \hat{s})^2 p_{S|\mathbf{X}}(s|\mathbf{x})ds$$

Minimum Mean Square Error (MMSE) Estimator

To minimise, we take the derivative with respect to $\hat{s}$ and make it equal to zero.

$$\frac{\mathrm{d}}{\mathrm{d}\hat{s}} \int_s (s-\hat{s})^2 p_{S|\mathbf{X}}(s|\mathbf{x}) ds = \int_s \frac{\partial}{\partial \hat{s}} (s-\hat{s})^2 p_{S|\mathbf{X}}(s|\mathbf{x}) ds$$

$$= \int_{-\infty}^{\infty} -2(s-\hat{s}) p_{S|\mathbf{X}}(s|\mathbf{x}) ds$$

In the optimum the above integral is equal to zero:

$$\int_{-\infty}^{\infty} -2(s-\hat{s}_{\text{Bayes}}) p_{S|\mathbf{X}}(s|\mathbf{x}) ds = 0 \Rightarrow$$

$$\hat{s}_{\text{Bayes}} = \mathbb{E}\{s|\mathbf{X}\}$$

The intuition behind this result is that the knowledge of $\mathbf{X}$ determines as optimum estimator the one corresponding to choosing the expected value of S for that particular observed value of $\mathbf{X}$.

Index

Expected absolute cost

The minimum absolute error estimator corresponds to the cost function $c(e) = |e| = |s - \widehat{s}|$. Therefore:

$$\begin{aligned}\widehat{s}_{\mathrm{MAD}} &= \arg\min_{\widehat{s}} \mathbb{E}\{c(S,\widehat{s})|\mathbf{X}=\mathbf{x}\} \\ &= \arg\min_{\widehat{s}} \int_s |s-\widehat{s}| p_{S|\mathbf{X}}(s|\mathbf{x}) ds\end{aligned}$$

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

Minimization of the average cost a posteriori

The Fundamental Theorem of Calculus enables to get the derivative of the average cost a posteriori as:

$$\frac{d\mathbb{E}\{|S-\widehat{s}||\mathbf{X}=\mathbf{x}\}}{d\widehat{s}} =$$

$$\frac{d}{d\widehat{s}}\left(\widehat{s}\left[\int_{-\infty}^{\widehat{s}} p_{S|\mathbf{X}}(s|\mathbf{x})ds - \int_{\widehat{s}}^{\infty} p_{S|\mathbf{X}}(s|\mathbf{x})ds\right] + \right.$$

$$\left.\int_{\widehat{s}}^{\infty} s\, p_{S|\mathbf{X}}(s|\mathbf{x})ds - \int_{-\infty}^{\widehat{s}} s\, p_{S|\mathbf{X}}(s|\mathbf{x})ds\right)$$

and then make

$$\frac{d\mathbb{E}\{|S-\widehat{s}||\mathbf{X}=\mathbf{x}\}}{d\widehat{s}} = 0$$

MAD estimator

MAD estimator

$$\widehat{s}_{\text{MAD}}(\mathbf{x}) = \text{median}\{S|\mathbf{X} = \mathbf{x}\}$$

Note: It is usually computed with the following expression:

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

Example: Find Bayesian estimation with Absolute Error Cost

Two random variables S and X follow a joint pdf

$$p_{S,X}(s,x) = \begin{cases} \frac{1}{x}, & 0 < s < x, \quad 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Find the Bayesian estimator that minimizes the absolute error cost
From previous example retrieve the posterior

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

Example: Find Bayesian estimation with Absolute Error Cost

$$\widehat{s}_{\mathrm{MAD}} \Rightarrow \int_0^{\widehat{s}_{\mathrm{MAD}}} p_{S|\mathbf{x}}(s|\mathbf{x})ds = \frac{1}{2} = \int_{\widehat{s}_{\mathrm{MAD}}}^{x} p_{S|\mathbf{x}}(s|\mathbf{x})ds$$

$$\int_0^{\widehat{s}_{\mathrm{MAD}}} \frac{1}{x}ds = \frac{1}{2} \quad \Rightarrow \quad \widehat{s}_{\mathrm{MAD}}(x) = \frac{x}{2}$$

Index

MAP estimator

MAP Estimator

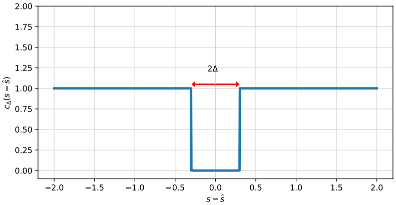
$$\begin{aligned}\widehat{s}_{\mathrm{MAP}}(\mathbf{x}) &= \arg\max_{s} p_{S|\mathbf{X}}(s|\mathbf{x}) \\ &= \arg\max_{s} \ln\left[p_{S|\mathbf{X}}(s|\mathbf{x})\right]\end{aligned}$$

When the distribution a posteiori presents several global maxima, the MAP estimator is not unique!

MAP estimator from a Bayesian point of view

Strictly speaking, the MAP estimator is not Bayesian because it is not minimizing any expected cost. However, if we consider the following cost function:

$$c_\Delta(s - \widehat{s}) = \begin{cases} 1, & \text{if } |s - \widehat{s}| > \Delta \\ 0, & \text{if } |s - \widehat{s}| < \Delta \end{cases}$$



and $\widehat{s}_\Delta$ as the bayesian estimator corresponding to c_Δ:

$$\widehat{s}_{\text{MAP}} = \lim_{\Delta \to 0} \widehat{s}_\Delta$$

Therefore the MAP estimator can be considered a limit case in a family of bayesian estimators.

Index

Maximum Likelihood (ML) Estimator

ML estimator

$$\hat{s}_{ML} = \arg\max_{s} p_{\mathbf{x}|s}(\mathbf{x}|s) = \arg\max_{s} \ln(p_{\mathbf{x}|s}(\mathbf{x}|s))$$

- No associated cost function, therefore **in general the ML is not a Bayesian estimator**. It could happen that the $\hat{s}_{ML}$ coincides with a Bayesian estimator, but by definition the $\hat{s}_{ML}$ does not optimise a cost function
- $\hat{s}_{ML}$ does not take into account the prior distribution $p_S(s)$. Therefore it can be used when s is a **deterministic parameter**.

Relationship between the MAP estimator and the ML estimator

$$\hat{s}_{MAP} = \arg\max_s p_{S|\mathbf{X}}(s|\mathbf{x}) = \arg\max_s \frac{p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)}{p_{\mathbf{X}}(\mathbf{x})}$$

If we don't know $p_S(s)$, we can assume it is **uniform** (no preference for any particular value of s). Therefore

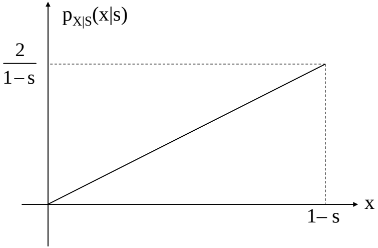
$$\arg\max_s \frac{p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)}{p_{\mathbf{X}}(\mathbf{x})} \sim \arg\max_s p_{\mathbf{X}|S}(\mathbf{x}|s) = \hat{s}_{ML}$$

Example: ML estimation of a random variable

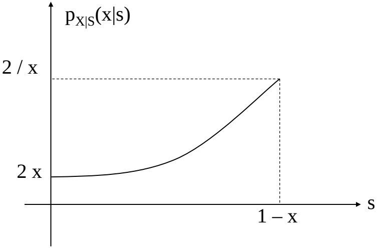
One wishes to estimate the value of a random variable S from the observed value of another variable X, statistically related to S through:

$$p_{X|S}(x|s) = \frac{2x}{(1-s)^2}, \;\; 0 < x < 1-s, \;\; 0 < s < 1$$

We need to maximise $p_{X|S}(x|s)$ with respect to s, (not to x)



$p_{X|S}(x|s)$ as a function of x



$p_{X|S}(x|s)$ as a function of s

Solution: $\hat{s}_{ML} = 1 - X$

Example: Estimation of the parameters of a Gaussian

We have access to a collection of l data samples $\{X^{(k)}\}_{k=1}^{l}$ drawn **independently** from a Gaussian pdf. Construct estimators for the mean m and the variance v of the Gaussian.

- The likelihood of each observation given m and v is

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

- Since we observe l samples, the **joint** pdf will be the product of the individual pdfs.

$$p_{\{X^{(k)}\}|m,v}(\{x^{(k)}\}|m,v) = \prod_{k=1}^{l} p_{X|m,v}(x^{(k)}|m,v)$$

$$= \frac{1}{(2\pi v)^{l/2}} \prod_{k=1}^{l} \exp\left[-\frac{(x^{(k)}-m)^2}{2v}\right]$$

Example: Estimation of the parameters of a Gaussian

- Taking logs and optimising

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

- Using $\hat{m}_{\mathrm{ML}}$ in the expression for the estimator of the variance

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

Gaussian case. Linear Mean Squared Error Estimation. Quality of estimators.

Modern Theory of Detection and Estimation. Block-1: Estimation

Emilio Parrado-Hernández, emilio.parrado@uc3m.es

September 27, 2022

uc3m | Universidad Carlos III de Madrid

Index

Estimation with Gaussian pdfs

- Estimation of **random variables** when the joint distribution of all the involved variables is a multivariate Gaussian.
- This is a very frequent case in real world problems.
- It is straightforward to proof that in this case all the marginals and all the conditionals will be Gaussian.
- Particularly, $p_{S|\mathbf{X}}(s|\mathbf{x})$ being Gaussian implies that the mean, the median and the mode of the posterior distribution coincide. Therefore $\hat{S}_{\mathrm{MMSE}} = \hat{S}_{\mathrm{MAD}} = \hat{S}_{\mathrm{MAP}}$. We focus on $\hat{S}_{\mathrm{MMSE}}$.

Index

Review of a Gaussian joint pdf for 1D variables

Assume $\mathbb{E}\{X\} = 0$ and $\mathbb{E}\{S\} = 0$. Then

$$p_{S,X}(s,x) \sim G\left(\left[\begin{array}{c} 0 \\ 0 \end{array} \right], \left[\begin{array}{cc} v_S & \rho \\ \rho & v_X \end{array} \right] \right)$$

where v_S is the variance of S, v_X is the variance of X and $\rho = \mathbb{E}\{SX\} - \mathbb{E}\{S\}\mathbb{E}\{X\}$ is their covariance.
Then the posterior is given by

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

$$= \frac{\frac{1}{2\pi\sqrt{v_X v_S - \rho^2}} \exp\left[-\frac{1}{2(v_X v_S - \rho^2)} \left[\begin{array}{cc} s & x \end{array} \right] \left[\begin{array}{cc} v_X & -\rho \\ -\rho & v_S \end{array} \right] \left[\begin{array}{c} s \\ x \end{array} \right] \right]}{\frac{1}{\sqrt{2\pi v_X}} \exp\left[-\frac{x^2}{2v_X} \right]}$$

MMSE in 1D

We need the mean of $p_{S|X}(s|x)$. We exploit that $p_{S|X}(s|x)$ is Gaussian:

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

After developing both expressions for the same Gaussian and making term identification:

$$\frac{m_{S|X}^2}{v_{S|X}} = \frac{v_S x^2}{v_X v_S - \rho^2} - \frac{x^2}{v_X}$$

$$\frac{s\ m_{S|X}}{v_{S|X}} = \frac{\rho x s}{v_X v_S - \rho^2}$$

$$\frac{s^2}{v_{S|X}} = \frac{v_X s^2}{v_X v_S - \rho^2}$$

Therefore $\hat{s}_{\mathrm{MMSE}} = m_{S|X} = \frac{\rho}{v_X} x$. Notice it is a **linear function** of x.

Example: Estimation of a Gaussian signal with additive Gaussian noise

$X = S + N$, S is a Gaussian signal with zero mean and variance v_S. N is a Gaussian noise with zero mean and variance v_N independent of S.

We need to construct an estimator for S given X.

Example: Estimation of a Gaussian signal with additive Gaussian noise

$X = S + N$, S is a Gaussian signal with zero mean and variance v_S. N is a Gaussian noise with zero mean and variance v_N independent of S.

We need to construct an estimator for S given X.

According to the previous result $\hat{S} = \rho X / v_X$

$$\rho = \mathbb{E}\{(X - m_X)(S - m_S)\} = \mathbb{E}\{X\ S\} = \mathbb{E}\{(S+N)S\} = \mathbb{E}\{S^2\} + \mathbb{E}\{S\ N\} = v_S$$

$$v_X = v_S + v_N \text{ since they are independent}$$

Therefore $\hat{S} = \frac{v_S X}{v_S + v_N}$.

Physical interpretation when $v_S \gg v_N$ or $v_N \gg v_S$

Index

Joint multivariate Gaussian pdf

In a general multivariate case $\mathbf{S}$ and $\mathbf{X}$ are random vectors of dimensions N and M, respectively. Their joint pdf is

$$p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x}) \sim G\left(\begin{bmatrix} \mathbf{m_S} \\ \mathbf{m_X} \end{bmatrix}, \begin{bmatrix} \mathbf{V_S} & \mathbf{V_{SX}} \\ \mathbf{V_{SX}^T} & \mathbf{V_X} \end{bmatrix} \right)$$

where $\mathbf{m_S}$ and $\mathbf{m_X}$ are the mean vectors and the covariances are

$$\mathbf{V_S} = \mathbb{E}\{(\mathbf{S}-\mathbf{m_S})(\mathbf{S}-\mathbf{m_S})^T\}$$
$$\mathbf{V_X} = \mathbb{E}\{(\mathbf{X}-\mathbf{m_X})(\mathbf{X}-\mathbf{m_X})^T\}$$
$$\mathbf{V_{SX}} = \mathbb{E}\{(\mathbf{S}-\mathbf{m_S})(\mathbf{X}-\mathbf{m_X})^T\}$$

Posterior distribution, multivariate case

The posterior $p_{\mathbf{S}|\mathbf{X}}(\mathbf{S}|\mathbf{X})$ is also a multivariate Gaussian with parameters

- mean:
$$\mathbf{m}_{\mathbf{S}|\mathbf{X}} = \mathbf{m}_{\mathbf{S}} + \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}(\mathbf{x} - \mathbf{m}_{\mathbf{X}})$$
- and covariance:
$$\mathbf{V}_{\mathbf{S}|\mathbf{X}} = \mathbf{V}_{\mathbf{S}} - \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}\mathbf{V}_{\mathbf{SX}}^{T}$$

Therefore the MMSE estimator is given by

$$\hat{\mathbf{s}}_{\mathrm{MMSE}} = \mathbb{E}\{\mathbf{s}|\mathbf{x}\} = \mathbf{m}_{\mathbf{S}} + \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}(\mathbf{x} - \mathbf{m}_{\mathbf{X}})$$

Index

Estimator with a fixed shape

- Sometimes you need to add a priori information about the estimation problem in the design of the estimator.
- Most of the times it means to fix a parametric shape for the estimation function $f_{\mathbf{w}}(\mathbf{X})$, with $\mathbf{w}$ a vector of parameters.
- Example: Observations in 2D and fix a shape for the estimator $\hat{S} = w_0 + w_1 X_1^2 + w_2 X_2^2$ (non-linear terms). The design task involves finding appropriate values for w_0, w_1 and w_2
- Minimize expected cost, but introducing the shape of the estimator as a constraint in the optimization

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \; \mathbb{E}\{c(S, \hat{S})\} = \arg\min_{\mathbf{w}} \; \mathbb{E}\{c(S, f_{\mathbf{w}}(\mathbf{X}))\}$$

$$= \arg\min_{\mathbf{w}} \int_{\mathbf{x}} \int_{s} c(s, f_{\mathbf{w}}(\mathbf{x})) p_{S,\mathbf{X}}(s, \mathbf{x}) ds d\mathbf{x}$$

Example: Estimation with constraints

Two random variables S and X follow a joint pdf

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

Find the estimator of the form $\hat{s} = wx^2$ that minimizes the quadratic cost.

Example: Estimation with constraints

Two random variables S and X follow a joint pdf

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

Find the estimator of the form $\hat{s} = wx^2$ that minimizes the quadratic cost.

$$\hat{s} = \arg\min_{w} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (s - wx^2)^2 p_{S,X}(s,x) ds dx$$

Taking derivatives and making them equal to zero

$$\hat{s} = \frac{5}{8} x^2$$

Index

LMMSE Estimator motivation

- Bayesian estimation: minimize expected cost. Leads to
  - MSE: $\mathbb{E}\{s|\mathbf{X}\}$. You need $p_{S|\mathbf{X}}(s|\mathbf{x})$ and compute an integral
  - MAP: $\arg\max_s p_{S|\mathbf{X}}(s|\mathbf{x})$. You need also $p_{S|\mathbf{X}}(s|\mathbf{x})$ and maximize.
- Under Gaussian joint pdfs MAP and MSE estimators coincide and they are **linear**
- What if we can't access the complete $p_{S|\mathbf{X}}(s|\mathbf{x})$?

LMMSE essence

- Assume linearity $\hat{s} = \mathbf{w}^T\mathbf{x} + w_0$
- Minimize MSE

LMMSE properties

- Depends only on first and second order statistics
- Easy to evaluate
- In general LMMSE is suboptimal
- ... but optimal in the ubiquitous Gaussian case

Index

1D case

1. MMSE:
$$\hat{s} = \arg\min_{\hat{s}} \mathbb{E}\{(s - \hat{s})^2\}$$
2. Linearity
$$\hat{s} = w_0 + w_1 x \rightarrow \hat{s} = \arg\min_{w_0, w_1} \mathbb{E}\{(s - w_0 - w_1 x)^2\}$$
3. Optimizing
   - $\frac{\partial}{\partial w_0}\mathbb{E}\{(s - w_0 - w_1 x)^2\} = 0 \rightarrow \mathbb{E}\{(s - w_0 - w_1 x)\} = \mathbb{E}\{e\} = 0$: **Error with zero mean**
   - $\frac{\partial}{\partial w_1}\mathbb{E}\{(s - w_0 - w_1 x)^2\} = 0 \rightarrow \mathbb{E}\{(xe)\} = 0$ **Error orthogonal to observations**
4. Solution
$$w_1 = \frac{\text{Covariance}(x, s)}{\text{Variance}(x)}, \qquad w_0 = \mathbb{E}\{s\} - \frac{\text{Covariance}(x, s)}{\text{Variance}(x)}\mathbb{E}\{x\}$$
   Notice $w_0 = 0$ if x and s are zero mean.
5. MSE: $\mathbb{E}\{e^2\} = \mathbb{E}\{s^2\} - w_0\mathbb{E}\{s\} - w_1\mathbb{E}\{sx\}$

Multivariate case

1. $\hat{s} = w_0 + \sum_{j=1}^{d} w_j x_j = w_0 + \mathbf{w}^T \mathbf{x}$
2. $\hat{s} = \arg\min_{w_0,\mathbf{w}} \mathbb{E}\{(s - w_0 - \mathbf{w}^T \mathbf{x})^2\}$
3. Optimization
   - $\frac{\partial}{\partial w_0} \mathbb{E}\{(s - w_0 - \mathbf{w}^T \mathbf{x})^2\} = 0 \rightarrow \mathbb{E}\{e\} = 0$: **Error with zero mean** $w_0 = \mathbb{E}\{s\} - \mathbf{w}^T \mathbb{E}\{\mathbf{x}\}$
   - $\nabla_w \mathbb{E}\{(s - w_0 - \mathbf{w}^T \mathbf{x})^2\} = \mathbf{0} \rightarrow \mathbb{E}\{(\mathbf{x}e)\} = \mathbf{0}$ **Error orthogonal to each of the observed variables**
   - Consequently: **Error orthogonal to the estimator** $\mathbb{E}\{\hat{s}e\} = 0$
4. **Normal Equations** We develop the principle of orthogonality for the ith variable x_i:

$$\mathbb{E}\{(x_i e)\} = 0 \Rightarrow \mathbb{E}\{x_i s\} - w_0 \mathbb{E}\{x_i\} - \sum_j w_j \mathbb{E}\{x_i x_j\} = 0$$

Substituting w_0

$$\mathbb{E}\{x_i s\} - \mathbb{E}\{s\}\mathbb{E}\{x_i\} + \sum_j w_j \mathbb{E}\{x_j\}\mathbb{E}\{x_i\} - \sum_j w_j \mathbb{E}\{x_i x_j\} = 0$$

Multivariate case

Remember

$$\mathrm{cov}(s, x_i) = \mathbb{E}\{x_i s\} - \mathbb{E}\{x_i\}\mathbb{E}\{s\}; \quad \mathrm{cov}(x_i, x_j) = \mathbb{E}\{x_i x_j\} - \mathbb{E}\{x_i\}\mathbb{E}\{x_j\}$$

Substituting covariances yields

$$\mathrm{cov}(s, x_i) - \sum_j w_j \mathrm{cov}(x_i, x_j) = 0, \quad \forall i = 1, \ldots, d$$

Stack the d equations (one for each x_i) and write in matrix form:

$$\mathbf{c}_{\mathbf{x},s} = C_{\mathbf{x},\mathbf{x}}\mathbf{w}$$

5. Solution:

$$\mathbf{w} = C_{\mathbf{x},\mathbf{x}}^{-1}\mathbf{c}_{\mathbf{x},s}$$

$$w_0 = w_0 = \mathbb{E}\{s\} - \mathbf{c}_{\mathbf{x},s}^T C_{\mathbf{x},\mathbf{x}}^{-1}\mathbb{E}\{\mathbf{x}\}$$

6. MSE: $\mathbb{E}\{e^2\} = \mathrm{var}(s) - \mathbf{c}_{\mathbf{x},s}^T C_{\mathbf{x},\mathbf{x}}^{-1}\mathbf{c}_{\mathbf{x},s}$

Example

The statistical relationship between S and the observed variables X_1 and X_2 is given by:

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

Determine the LMMSE estimator of S given the observations.

Example

The statistical relationship between S and the observed variables X_1 and X_2 is given by:

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

Determine the LMMSE estimator of S given the observations.

$$\hat{s} = \mathbf{c}_{S,\mathbf{X}}^T C_{\mathbf{x},\mathbf{x}}^{-1}\mathbf{x} + \mathbb{E}\{s\} - \mathbf{c}_{\mathbf{x},s}^T C_{\mathbf{x},\mathbf{x}}^{-1}\mathbb{E}\{\mathbf{x}\}$$

$$\hat{s} = \begin{bmatrix} .5 & 2 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 2 \end{bmatrix}^{-1} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + 0.5 - \begin{bmatrix} .5 & 2 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Index

Index

Assessing the quality of an estimator

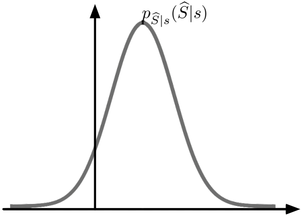
- One can design different estimators to work in a same scenario (problem).
- Fix **criteria** that enable a **fair comparison** between estimators.
- **Expected cost** for a determined cost function leads to always choose the Bayesian estimator as optimum.
- **Other measures** that can be of interest in different scenarios you may come accross.
  - **Bias** ($\sim$ systematic error)
  - **Variance** (concentration of the estimations around their expected value)

  Important! Bias and variance depend on if the variable to be estimated is random or deterministic.

Index

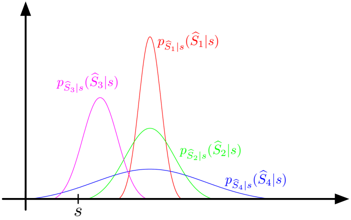
Probability density of the estimator

- $p_{\widehat{S}|s}(\widehat{s}|s)$ provides a complete characterization of the behaviour of the estimator.
  (Since $\widehat{S}$ is a function of the observations, $\widehat{S} = f(\mathbf{X})$, $p_{\widehat{S}|s}(\widehat{s}|s)$ can be obtained from $p_{\mathbf{X}|s}(\mathbf{X}|s)$ with a transformation of random variable.



Motivation

Imagine a case of estimation of a parameter s with 4 different estimators ($\widehat{S}_1 = f_1(\mathbf{X})$, $\widehat{S}_2 = f_2(\mathbf{X})$, $\widehat{S}_3 = f_3(\mathbf{X})$ and $\widehat{S}_4 = f_4(\mathbf{X})$) Examining their pdfs, which estimator seems more appropriate?



Perhaps $\widehat{S}_3$ as the probability of getting estimations close to the true s is significantly larger.

Bias and variance when s is deterministic

- **Bias**:
  - Expresses how far is the mean of $p_{\widehat{S}|s}(\widehat{s}|s)$ from the true value of s

$$\mathrm{Bias}\left\{\widehat{S}|s\right\} = \mathbb{E}\left\{s - \widehat{S}|s\right\} = \mathbb{E}\{s|s\} - \mathbb{E}\left\{\widehat{S}|s\right\} = s - \mathbb{E}\left\{\widehat{S}|s\right\}$$

- **Variance**:

$$\mathrm{Variance}\left\{\widehat{S}|s\right\} = \mathbb{E}\left\{\left(\widehat{S} - \mathbb{E}\left\{\widehat{S}\right\}\right)^2\Big|s\right\} = \mathbb{E}\left\{\widehat{S}^2|s\right\} - \mathbb{E}^2\left\{\widehat{S}|s\right\}$$

  If s is a deterministic parameter, the variance of the estimator coincides with the estimation error.

$$\mathrm{Variance}\left\{\widehat{S} - s\Big|s\right\} = \mathrm{Variance}\left\{\widehat{S}|s\right\} - \mathrm{Variance}\{s|s\} = \mathrm{Variance}\left\{\widehat{S}|s\right\}$$

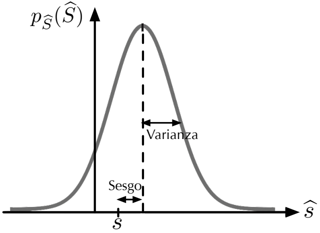
Important! The bias and the variance of the estimators of a deterministic parameter are a function of its true value (s).

Properties

Bias and variance of the estimador of a deterministic paramter

$\text{Bias}\{\widehat{S}|s\} = s - \mathbb{E}\{\widehat{S}|s\}$

$\text{Variance}\left\{\widehat{S}|s\right\} = \mathbb{E}\{\widehat{S}^2|s\} - \mathbb{E}^2\{\widehat{S}|s\}$



- The estimators with **zero bias** are known as **unbiased estimators**.
- If the estimator operates on a number K of observations of a random variable and $\text{Variance}(\widehat{S}) \to 0$ if $K \to \infty$, the estimator is **consistent in variance**.

Mean squared error

Mean squared error of the estimator of a deterministic parameter

$$\begin{aligned}\mathbb{E}\Big\{(s-\widehat{S})^2\big|s\Big\} &= \text{Variance}\Big\{\widehat{S}-s\big|s\Big\}+\mathbb{E}^2\{s-\widehat{S}|s\}\\ &= \text{Variance}\{\widehat{S}|s\}+\Big[\text{Bias}(\widehat{S}|s)\Big]^2\end{aligned}$$

Example

Calculate the bias and variance of the sample estimation of the mean of a random variable.

Index

General vision

- As in the deterministic case, one could directly use $p_{\widehat{S}|S}(\widehat{s}|s)$ to study the performance of an estimador.
- However, when S is a random variable, the true value $(S = s)$ changes every time we repeat the experiment of drawing the observations and running the estimator.

Bias and variance

Bias and variance of the estimator of a random variable

$$\text{Bias}\{\widehat{S}\} = \mathbb{E}\{S - \widehat{S}\} = \mathbb{E}\{S\} - \mathbb{E}\{\widehat{S}\}$$
$$\text{Variance}\left\{\widehat{S}\right\} = \mathbb{E}\left\{\left(\widehat{S} - \mathbb{E}\{\widehat{S}\}\right)^2\right\} = \mathbb{E}\{\widehat{S}^2\} - \mathbb{E}^2\{\widehat{S}\}$$

The calculation of these expectations uses the joint pdf of S and $\mathbf{X}$.

Important! The bias and variance of the estimators of a random variable **are not** a **function** of the true value of S (s).

Mean squared error

The mean squared error of the estimator of a deterministic parameter was deterministic

$$\mathbb{E}\left\{(S-\widehat{S})^2\right\} = \text{Variance}\left\{\widehat{S}-S\right\} + \mathbb{E}^2\{S-\widehat{S}\}$$
$$= \text{Variance}\{\widehat{E}\} + \left[\text{Bias}(\widehat{S})\right]^2$$

If $(\widehat{S})$ is random, the variance of the error in general will not be equal to the variance of the estimator.

Properties

- The **unconstrained minimum mean square error estimator is unbiased**:

$$
\begin{aligned}
\text{Bias}\{\widehat{S}_{\text{MMSE}}\} &= \mathbb{E}\left\{S - \widehat{S}_{\text{MMSE}}\right\} = \mathbb{E}\{S\} - \mathbb{E}\left\{\widehat{S}_{\text{MMSE}}\right\} \\
&= \mathbb{E}\{S\} - \int \mathbb{E}\ \{S|\mathbf{X} = \mathbf{x}\} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&= \mathbb{E}\{S\} - \mathbb{E}\{S\} \\
&= 0
\end{aligned}
$$

- The **linear MMSE** is **unbiased**.
  $(\mathbb{E}(E^*) = 0)$