S.I.D.A. review

F. A. Ramponi, 2013/5/27.

1 Least squares

1.1 Introduction and motivation

We know, suspect, or assume a relation between a variable $x \in \mathbb{R}^m$ and a variable $y \in \mathbb{R}$ in the form of a function that further depends on an unknown parameter θ: $y = f(x, \theta)$. *Interpolation problem*: given a sequence of pairs $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ that satisfy such a functional relation, find the particular θ^o that corresponds to the function. In general the problem does not admit an exact solution. The issue is that the functional relation is an *assumption*, but measures are always corrupted by measurement errors ε_i:

$$y_i = f(x_i, \theta) + \varepsilon_i, \quad i = 1, \cdots, N.$$

Hence, we search an *approximate* function $f(\cdot, \hat{\theta})$, instead of the "true" one $f(\cdot, \theta^o)$. There are at least three reasons why such an approximation is useful: 1) to investigate some internal properties of the mechanism that generates the data; 2) to measure the parameter θ^o, which may correspond to some physical quantity of interest; 3) to *predict* a future value y_{N+1}, when a future value x_{N+1} will be available. In approximating $f(\cdot, \theta^o)$ with a certain $f(\cdot, \theta)$, we commit approximation errors $\epsilon_i(\theta)$ called *residuals*:

$$\epsilon_i(\theta) := y_i - f(x_i, \theta), \quad i = 1, \cdots, N.$$

Least squares method: given $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$, find the estimate

$$\hat{\theta}_{\text{LS}} := \arg\min_{\theta} \sum_{i=1}^{N} \epsilon_i(\theta)^2 = \arg\min_{\theta} \sum_{i=1}^{N} (y_i - f(x_i, \theta))^2.$$

If f is *linear in the parameter* θ, then the least squares problem has an analytic solution.

1.2 Linearity in the parameter θ

Let $x_i = (x_i^1, \cdots, x_i^m) \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$ for $i = 1, \cdots, N$. If y_i is a function of x_i, we call x_i "explanatory variable", and we say that y_i are "explained" by x_i. y_i may not be linked to x_i directly, but through p functions $\varphi_1 : \mathbb{R}^m \to \mathbb{R}, \cdots, \varphi_p : \mathbb{R}^m \to \mathbb{R}$ called *regressors*, which can be nonlinear. The data generation model to which we will apply least squares is the following:

$$\begin{aligned} y_1 &= \theta_1 \varphi_1(x_1^1, \cdots, x_1^m) + \theta_2 \varphi_2(x_1^1, \cdots, x_1^m) + \cdots + \theta_p \varphi_p(x_1^1, \cdots, x_1^m) + \varepsilon_1; \\ &\vdots \\ y_N &= \theta_1 \varphi_1(x_N^1, \cdots, x_N^m) + \theta_2 \varphi_2(x_N^1, \cdots, x_N^m) + \cdots + \theta_p \varphi_p(x_N^1, \cdots, x_N^m) + \varepsilon_N. \end{aligned}$$

Defining the vectors $\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$ and $\varphi(x_i^1, \cdots, x_i^m) = \begin{bmatrix} \varphi_1(x_i^1, \cdots, x_i^m) \\ \vdots \\ \varphi_p(x_i^1, \cdots, x_i^m) \end{bmatrix}$ and recalling that $x_i = [x_i^1 \cdots x_i^m]^\top$, we write the model in compact form: $y_i = f(x_i, \theta) + \varepsilon_i = \varphi(x_i)^\top \theta + \varepsilon_i$ for $i = 1, \cdots, N$. Note that the function f is linear in the parameter θ, although not necessarily in the explanatory variable x. We shorten notation even more letting $\varphi_i := \varphi(x_i)$, and the model finally reads $y_i = \varphi_i^\top \theta + \varepsilon_i$, $i = 1, \cdots, N$. The least squares method asks to find

$$\hat{\theta}_{\text{LS}} := \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - \varphi_i^\top \theta)^2.$$

To find the minimum, we set its derivative w.r.t. θ equal to zero: $\frac{\partial}{\partial \theta} \sum_{i=1}^{N} (y_i - \varphi_i^\top \theta)^2 = -2 \sum_{i=1}^{N} \varphi_i (y_i - \varphi_i^\top \theta) = 0$. We obtain the *normal equations*:

$$\left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right) \theta = \sum_{i=1}^{N} \varphi_i y_i. \tag{1}$$

Any θ that solves (1) is a minimum point for the sum of squares and a solution to the least squares problem; in particular, if $R = \sum_{i=1}^{N} \varphi_i \varphi_i^\top$ is invertible (in practical applications this is generally the case), then the only solution to the normal equations reads explicitly:

$$\hat{\theta}_{\text{LS}} = \left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right)^{-1} \sum_{i=1}^{N} \varphi_i y_i. \tag{2}$$

1.3 Existence and uniqueness of a solution of the normal equations

Lemma 1.3.1

$$\text{range} \left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right) = \text{span } \{\varphi_1, \cdots, \varphi_N\}.$$

Proof. Let $R = \sum_{i=1}^{N} \varphi_i \varphi_i^\top$ and $S = \text{span } \{\varphi_1, \cdots, \varphi_N\}$. Since R is symmetric, range $R^\top$ = range R and null $R = (\text{range } R^\top)^\perp = (\text{range } R)^\perp$. Suppose that $v \in S^\perp$. Then $\varphi_i \perp v$ for all i. Then $Rv = \sum_{i=1}^{N} \varphi_i (\varphi_i^\top v) = 0$, hence $v \in \text{null } R = (\text{range } R)^\perp$. Suppose, on the other hand, that $v \in (\text{range } R)^\perp = \text{null } R$. Then $Rv = 0$, hence $0 = v^\top R v = \sum_{i=1}^{N} (\varphi_i^\top v)^2$. Since the last expression is a sum of nonnegative quantities, for it to be zero it must hold $\varphi_i^\top v = 0$, that is $\varphi_i \perp v$ for all i, hence $v \in S^\perp$. We conclude that $(\text{range } R)^\perp = S^\perp$. Since range R and S are subspaces of $\mathbb{R}^p$ (finite-dimensional), taking the orthogonal complement on both sides we obtain range $R = S$. □

Corollary 1.3.1 *The normal equations have at least one solution.*

Such solution may not be unique; this happens precisely when R is singular. Possible reasons: 1) there are not enough data ($N < p$, pathological); 2) the explanatory data x_i do not carry enough information.

1.4 Interpretation in terms of projections

Theorem 1.4.1 *Let V be a subspace of $\mathbb{R}^N$, and $y \in \mathbb{R}^N$. If there exists a vector $v_m \in V$ such that $\|y - v_m\| \leq \|y - v\|$ for all $v \in V$, then v_m is unique. Moreover, a necessary and sufficient condition for v_m to be the unique minimizing vector is that $y - v_m$ is orthogonal to V.*

Any such v_m is called the *orthogonal projection* of y on V; such orthogonality is precisely what the normal equations are asking for. Consider again the model $y_i = \varphi_i^\top \theta + \varepsilon_i$, $i = 1, \cdots, N$.

Let us define $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \Phi = \begin{bmatrix} \varphi_1^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix}, E = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}$, where $Y, E \in \mathbb{R}^N$ and $\Phi \in \mathbb{R}^{N \times p}$. The model then reads $Y = \Phi\theta + E$. Let $v_1, \cdots, v_p$ be the *columns* of Φ (whereas the regressors are its *rows*); then $V = \text{span } \{v_1, \cdots, v_p\}$ is a subspace of $\mathbb{R}^N$, and $v = \Phi\theta$ is a vector in V. The least squares problem asks to minimize $\|Y - \Phi\theta\|^2 = \|Y - v\|^2$, that is, to find $v_m = \Phi\hat{\theta}$ such that $\|Y - v_m\|^2$ is minimal. Now we apply Theorem 1.4.1: if $Y - v_m \perp V$, then v_m is a minimizing vector. Explicitly, if

$$\begin{aligned} Y - \Phi\hat{\theta} \perp v_i, &\quad \text{that is,} \\ v_i^\top (Y - \Phi\hat{\theta}) = 0 &\quad \text{for all } i = 1, \cdots, p, \end{aligned} \tag{3}$$

then $\Phi\hat{\theta}$ is a minimizing vector, and $\hat{\theta} = [\hat{\theta}_1, \cdots, \hat{\theta}_p]^\top$ is the vector of coefficients such that $\Phi\hat{\theta} = \hat{\theta}_1 v_1 + \cdots + \hat{\theta}_p v_p$ is the orthogonal projection of Y on the space spanned by the columns $v_1, \cdots, v_p$ of Φ. Stacking the rows $v_i^\top$ on each other we get $\Phi^\top$, hence stacking the equations (3) on each other we obtain:

$$\Phi^\top (Y - \Phi\hat{\theta}) = 0,$$

which finally yields

$$\Phi^\top \Phi \hat{\theta} = \Phi^\top Y. \tag{4}$$

Equation (4) is just another way to write the normal equations (1), because $\Phi^\top\Phi = \sum_{i=1}^N \varphi_i\varphi_i^\top$ and $\Phi^\top Y = \sum_{i=1}^N \varphi_i y_i$. Since we know that a solution $\hat\theta_{\text{LS}}$ to the normal equations exists, now we also know that at least one minimizer of $\|Y - \Phi\theta\|^2$ exists, or that there exists a solution to the least squares problem. Note that, according Theorem 1.4.1, the minimizer $v_m = \Phi\hat\theta_{\text{LS}}$ is unique. This does *not* mean that $\hat\theta_{\text{LS}}$ is also unique! The uniqueness of $\hat\theta_{\text{LS}}$ holds if Φ has full rank (linearly independent columns), and this in turn is the case if $R = \Phi^\top\Phi$ is invertible.

1.5 Goodness of fit

Given a collection of numbers $y_1, \cdots, y_N$, the minimization of its square deviation from a number θ:

$$\min_\theta \frac{1}{N}\sum_{i=1}^N (y_i - \theta)^2,$$

is a particular instance of least squares problem, with the only regressor $\varphi = 1$. The solution is the *sample average of the* $\{y_i\}$: $\hat\theta_{\text{LS}} = \frac{1}{N}\sum_{i=1}^N y_i = \mathsf{M}[y]$. Hence, the *sample average* is the number from which is minimal the squared deviation of the $\{y_i\}$. The attained minimum deviation is the *sample variance* of the $\{y_i\}$: $\mathsf{S.Var}[y] = \frac{1}{N}\sum_{i=1}^N (y_i - \mathsf{M}[y])^2$.
Since the regressor $\varphi(\cdot) = 1$ is always available irrespective of the nature of the explanatory data, it is often included in least squares problems. Now note that if $\varphi_1 = 1, \varphi_2(x_i), \cdots, \varphi_p(x_i)$ are the regressors, then there is always a parameter $\hat\theta_m$ such that the corresponding sum of squares is exactly the sample variance of $\{y_i\}$, namely $\hat\theta_m = [\mathsf{M}[y], 0, \cdots, 0]^\top$. It follows that if the explanatory data are really "explaining", the least squares solution must behave *better* than $\hat\theta_m$:

$$\frac{1}{N}\sum_{i=1}^N (y_i - \varphi_i^\top\hat\theta_{\text{LS}})^2 < \frac{1}{N}\sum_{i=1}^N (y_i - \varphi_i^\top\hat\theta_m)^2 = \frac{1}{N}\sum_{i=1}^N (y_i - \mathsf{M}[y])^2 = \mathsf{S.Var}[y],$$

We call *residual variance* the quantity $\mathsf{RV} := \frac{1}{N}\sum_{i=1}^N (y_i - \varphi_i^\top\hat\theta_{\text{LS}})^2$ and *explained variance* the quantity $\mathsf{EV} := \mathsf{S.Var}[y] - \mathsf{RV}$. Then $0 \le \mathsf{EV} \le \mathsf{S.Var}[y]$. Dividing by $\mathsf{S.Var}[y]$ and letting $\rho^2 := \frac{\mathsf{EV}}{\mathsf{S.Var}[y]}$, we obtain finally the relation $0 \le \rho^2 \le 1$. On one extreme, when $\rho^2 = 0$ (or equivalently $\mathsf{RV} = \mathsf{S.Var}[y]$), the explanatory variables are not adding any information, and the least squares solution is not any more useful than a sample average; on the other extreme, when $\rho^2 = 1$ (equivalently $\mathsf{RV} = 0$), the y_i are explained *perfectly* by the x_i, that is, the equations $y_i = \varphi_i^\top\hat\theta_{\text{LS}}$ are all verified *exactly*.
One usually expresses ρ^2 as a percentage, and says something like "the model explains 90% of the variance of the data". Thus, in general, the closer ρ^2 is to 100%, the better. But be warned: one possible reason why this happen is that *there are just too many regressors*, that is $p \simeq N$. This situation is called *over-fitting*. Remember: *the spirit of the least squares method is to use many measures to average out noise, not to identify many parameters!*

1.6 Consistency

So far, the functional relation $y = \varphi(x)^\top\theta^o + \varepsilon$ has been a *model* that we have used to explain the data $\{(x_i, y_i)\}_{i=1}^N$ through regressor functions, but we have not really pretended that a "true" θ^o actually exists. Up to now $\{x_i\}$, and consequently $\{\varphi_i\}$, have been just vectors, and the results were "algorithmic" in the sense that they guarantee the existence of a solution to the least squares problem, and tell us how to compute it; the numbers $\{\varepsilon_i\}$ have been there just to model "disturbances", without any particular requirement other than they be preferably small. Now we assume that there is indeed a θ^o, and that the data conform to the model $y = \varphi(x)^\top\theta^o + \varepsilon$. We assume, moreover, that $\{\varepsilon_i\}$ and possibly $\{\varphi_i\}$ are *random variables*, and examine the consequences on the asymptotic behavior of $\hat\theta_{\text{LS}}$ (its behavior "for large N"). Our results will justify the claim that $\hat\theta_{\text{LS}}$ is a good estimator of θ^o.

Let the data $\{(\varphi_i, y_i)\}$ be generated by the model

$$y_i = \varphi_i^\top\theta^o + \varepsilon_i \tag{5}$$

where $\theta^o \in \mathbb{R}^p$ is understood as the "true" parameter, deterministic but unknown. To see how $\hat{\theta}_{\text{LS}}$ is related to θ^o, substitute (5) in the normal equations (1) and divide by N:

$$\left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)\hat{\theta}_{\text{LS}} = \left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)\theta^o + \frac{1}{N}\sum_{i=1}^{N}\varphi_i\varepsilon_i$$

In practical situations, we expect that:

1. for big N, the matrix $\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top$ becomes invertible due to the fact that the φ_i carry more and more information; then the solution $\hat{\theta}_{\text{LS}}$ is unique, and $\hat{\theta}_{\text{LS}} = \theta^o + \left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varepsilon_i$;

2. that $\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varepsilon_i \to 0$ as $N \to \infty$, yielding $\hat{\theta}_{\text{LS}} \to \theta^o$ ($\hat{\theta}_{\text{LS}}$ is a *consistent* estimator of θ^o). This idea is in conformity with the principle that underlies least squares optimization: *noise tends to get averaged out and lose importance as the number of measures increases; consequently, it is usually better to take many noisy measures, than to take few precise measures.*

Theorem 1.6.1 *Suppose that $\{y_i\}_{i=1}^{\infty}$ are generated by the model $y_i = \varphi_i^\top\theta^o + \varepsilon_i$, and that*

1. *$\{\varphi_i\}_{i=1}^{\infty}$ are i.i.d. random vectors, with covariance matrix $\Sigma := \mathsf{E}\left[\varphi_i\varphi_i^\top\right] > 0$;*
2. *$\{\varepsilon_i\}_{i=1}^{\infty}$ are independent and identically distributed random variables, with mean $\mathsf{E}\left[\varepsilon_i\right] = 0$.*
3. *ε_i is independent from φ_i for all i.*

Then $\hat{\theta}_{\text{LS}} \to \theta^o$ almost surely as $N \to \infty$.

Sketch of proof:

- use a strong law of large numbers to show that, almost surely, $\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top$ becomes invertible for $N \geq \bar{N}$ (and converges to the invertible Σ);
- note that $\mathsf{E}\left[\varphi_i\varepsilon_i\right] = 0$;
- consequently, by a strong law of large numbers, $\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varepsilon_i \to 0$ almost surely.

Theorem 1.6.2 *Suppose that $\{y_i\}_{i=1}^{\infty}$ are generated by the model $y_i = \varphi_i^\top\theta^o + \varepsilon_i$, and that*

1. *$\{\varphi_i\}_{i=1}^{\infty}$ are deterministic vectors, such that $\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top \geq aI$ for all $N \geq \bar{N}$, and such that $\|\varphi_i\|^2 \leq A$ for all i, for certain positive constants a, A;*
2. *$\{\varepsilon_i\}_{i=1}^{\infty}$ are independent random variables, not necessarily identically distributed, but such that, for all i, $\mathsf{E}\left[\varepsilon_i\right] = 0$ and $\mathsf{E}\left[\varepsilon_i^2\right] \leq c$ for a certain constant $c \in \mathbb{R}$.*

Then $\hat{\theta}_{\text{LS}} \to \theta^o$ almost surely as $N \to \infty$.

Sketch of proof:

- use the first hypothesis to show that, for all $N \geq \bar{N}$, $\left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)^{-1}$ exists *and is bounded*;
- use the bound on $\|\varphi_i\|^2$, the second hypothesis, and a strong law of large numbers, to show that $\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varepsilon_i \to 0$ almost surely.

If the φ_i are deterministic and the ε_i are i.i.d. with mean 0 and variance σ^2, then $\hat{\theta}_{\text{LS}} \to \theta^o$ also in the mean-square sense. Indeed, with compact notation, $Y = \Phi\theta^o + E$; for big N, we expect that $\Phi^\top\Phi = \sum_{i=1}^{N}\varphi_i\varphi_i^\top$ becomes invertible; then

$$\hat{\theta}_{\text{LS}} = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top Y = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\left(\Phi\theta^o + E\right) = \theta^o + \left(\Phi^\top\Phi\right)^{-1}\Phi^\top E.$$

It follows

$$\mathsf{E}\left[\hat{\theta}_{\mathrm{LS}}\right] = \theta^o + \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\mathsf{E}\left[E\right] = \theta^o;$$
$$\mathsf{Var}\left[\hat{\theta}_{\mathrm{LS}}\right] = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\mathsf{E}\left[EE^\top\right]\Phi\left(\Phi^\top\Phi\right)^{-1} = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\,\sigma^2 I\,\Phi\left(\Phi^\top\Phi\right)^{-1} = \sigma^2\left(\Phi^\top\Phi\right)^{-1}.$$

If $\left(\Phi^\top\Phi\right)^{-1}$ tends to 0, as we expect, then $\mathsf{Var}\left[\hat{\theta}_{\mathrm{LS}}\right] \to 0$.

2 Linear-algebraic aspects of the least squares method

2.1 The Moore-Penrose pseudoinverse

2.1.1 Definition

Given any matrix $A \in \mathbb{R}^{m\times n}$, a matrix $A^+ \in \mathbb{R}^{n\times m}$ is called a *Moore/Penrose pseudo-inverse* of A, or a pseudoinverse of A for short, if it satisfies the following properties:

1. The matrix $AA^+ \in \mathbb{R}^{m\times m}$ is symmetric;
2. The matrix $A^+A \in \mathbb{R}^{n\times n}$ is symmetric;
3. $AA^+A = A$;
4. $A^+AA^+ = A^+$.

Theorem 2.1.1 *For any matrix $A \in \mathbb{R}^{m\times n}$, a pseudoinverse of A exists, it is unique, and it is uniquely determined by the above four properties.*

To show that a matrix B is the pseudoinverse of A it suffices to show that B satisfies the four properties. For example:

- If A is square and invertible, then $A^+ = A^{-1}$. Indeed, 1) $AA^{-1} = I$, which is symmetric; 2) $A^{-1}A = I$, which is symmetric; 3) $AA^{-1}A = IA = A$; 4) $A^{-1}AA^{-1} = IA^{-1} = A^{-1}$. Since A^{-1} satisfies the four properties, it is the pseudoinverse of A. This is where the name *pseudo*-inverse comes from: it is a generalization of the inverse of a square, full rank matrix.
- If A is a "tall" matrix, meaning that $m > n$, and if its columns are linearly independent, then A^+ is a *left inverse of A*, that is a matrix such that $A^+A = I$. More precisely, we have $A^+ = (A^\top A)^{-1}A^\top$. As before, one just needs to check that $(A^\top A)^{-1}A^\top$ satisfies the four properties. For example, the last one reads:

  $$((A^\top A)^{-1}A^\top)A((A^\top A)^{-1}A^\top) = (A^\top A)^{-1}(A^\top A)(A^\top A)^{-1}A^\top = (A^\top A)^{-1}A^\top.$$

  This is precisely the case of the least squares method, when Φ has full rank: $\Phi^+ = (\Phi^\top\Phi)^{-1}\Phi^\top$

In the general case (singular, non-square matrices), the pseudoinverse can be computed with the aid of a useful matrix decomposition, called the Singular Value Decomposition (SVD).

2.2 Least squares solution of a linear system

Given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, the linear equation $Ax = b$ may have no solution, and if a solution exists, it may not be unique. Note that, if x is a solution, then $Ax - b = 0$, hence $\|Ax - b\|_2 = 0$.

Definition 2.2.1 *Given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, consider the set*

$$S(A,b) = \{x \in \mathbb{R}^n \mid x \text{ minimizes } \|Ax - b\|_2\}$$

(A vector $x \in S(A,b)$ is "almost" a solution, in the sense that it attains the closest possible result to $\|Ax-b\|_2 = 0$.) Then any vector $x^* \in S(A,b)$ of minimum norm, that is

$$x^* = \arg\min_{x\in S(A,b)} \|x\|_2$$

is called a least squares solution *of the linear system $Ax = b$.*

Theorem 2.2.1 *A least squares solution of $Ax = b$ always exists, it is unique, and it is given by*

$$x^* = A^+ b.$$

In the least squares method, the problem is to minimize $\|\Phi\theta - Y\|_2$. The solution is $\hat{\theta}_{\text{LS}} = \Phi^+ Y$ (it is the solution with minimum norm $\|\theta\|_2$, if there are more than one). Example: interpolation with a third-degree polynomial $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$ given the measures $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$:

1. stack x_i and y_i in two separate columns X and Y of the same dimension;
2. build the "data matrix" Φ having $\varphi_j(x_i) = (x_i)^j$:

$$\Phi = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 \end{bmatrix}$$

3. compute $\hat{\theta}_{\text{LS}} = \Phi^+ Y$.

Matlab code:

```
Phi = [ones(length(X),1), X, X.^2, X.^3];
thetaLS = pinv(Phi)*Y;
```

3 Identification of dynamical systems

3.1 Model classes

An *autoregressive* (AR) process $\{y(t)\}_{t=0}^{\infty}$ is described by the following model:

$$y(t) + a_1 y(t-1) + \cdots a_n y(t-n) = e(t)$$

where $a_1, \cdots, a_n \in \mathbb{R}$, and $\{e(t)\}_{t=0}^{\infty}$ is a sequence of i.i.d. random variables ("process noise"), with mean 0 and unknown variance (not necessarily a disturbance: just an input invisible to the experimenter). We suppose that $e(t)$ is independent from $y(s)$ for all $s < t$. Let us denote with z and z^{-1} the anticipation operator and the delay operator respectively ($zy(t) = y(t+1)$ and $z^{-1}y(t) = y(t-1)$). Then the model reads $(1 + a_1 z^{-1} + \cdots + a_n z^{-n})y(t) = e(t)$; hence the process is the output of a linear system with transfer function

$$W(z) = \frac{z^n}{z^n + a_1 z^{n-1} + \cdots + a_n}$$

and with a white noise as an input. If the poles of $W(z)$, i.e. the roots of the polynomial $z^n + a_1 z^{n-1} + \cdots + a_n$, lie inside the interior of the unit circle ($\{z \in \mathbb{C} \mid |z| < 1\}$), then such system is BIBO-stable, and it can be shown that $\{y(t)\}$ is asymptotically wide-sense stationary.

The process $\{y(t)\}_{t=0}^{\infty}$ is *autoregressive with exogenous input* (ARX) if it conforms to the following model:

$$y(t) + a_1 y(t-1) + \cdots a_n y(t-n) = b_0 u(t) + b_1 u(t-1) + \cdots b_m u(t-m) + e(t)$$

where $m \leq n$, $\{e(t)\}$ is a white noise as above, and $\{u(t)\}_{t=0}^{\infty}$ is another signal, either random or deterministic, but known to the experimenter. We suppose that $e(t)$ is independent from $y(s)$ for all $s < t$, and independent from $u(s)$ for all s. If $u(t)$ is stationary and if the roots of $z^n + a_1 z^{n-1} + \cdots + a_n$ lie inside the unit circle, then $y(t)$ is asymptotically stationary as well.

With respect to both AR and ARX models, the system identification problem is the task of reconstructing $W(z)$, that is, finding the parameters $a_1, \cdots, a_n$ and possibly $b_0, \cdots, b_m$, from measures, respectively $(y(0), y(1), \cdots, y(N))$ or $((u(0), y(0)), (u(1), y(1)), \cdots, (u(N-1), y(N-1)), y(N))$. Examples: the identification of AR models is used in telecommunications to save bandwidth in the transmission of speech through digital channels. The identification of ARX models may be used to find the transfer function of a plant. AR and ARX models are interesting because their identification can be done with the least squares method.

3.2 Application of the least squares method

Consider the ARX model

$$y(t) + \sum_{i=1}^{n} a_i y(t-i) = \sum_{i=0}^{m} b_i u(t-i) + e(t) \tag{6}$$

A linear *predictor* of $y(t)$, given the past values $y(t-1)$ ad $u(t-1)$ is a function of the form

$$\begin{aligned}\hat{y}(t|t-1) &= -\sum_{i=1}^{n} \hat{a}_i y(t-i) + \sum_{i=0}^{m} \hat{b}_i u(t-i) \\ &= \begin{bmatrix} -y(t-1) & \cdots & -y(t-n) & u(t) & \cdots & u(t-m) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \cdots \\ \hat{a}_n \\ \hat{b}_0 \\ \cdots \\ \hat{b}_m \end{bmatrix} = \varphi_t^\top \hat{\theta}\end{aligned}$$

used to estimate $y(t)$ from past measures $\{y(t-1), \cdots, y(t-n), u(t), \cdots, u(t-m)\}$. Predicting $y(t)$, we commit an error $\tilde{y}(t|t-1) = y(t) - \hat{y}(t|t-1)$, a random quantity with unknown variance. The particular predictor

$$\hat{y}^o(t|t-1) = \begin{bmatrix} -y(t-1) & \cdots & -y(t-n) & u(t) & \cdots & u(t-m) \end{bmatrix} \begin{bmatrix} a_1 \\ \cdots \\ a_n \\ b_0 \\ \cdots \\ b_m \end{bmatrix} = \varphi_t^\top \theta^o$$

is optimal, in the sense that it minimizes such variance; the ideal goal of prediction is to compute $\hat{y}^o$, but we do not know θ^o, therefore we go for an approximation based on data. The so-called *prediction-error-minimization* (PEM) method prescribes to estimate θ^o finding a $\hat{\theta}$ that minimizes the sum of the squares of the prediction errors (the *residuals*):

$$V = \sum_{t=1}^{N} \tilde{y}(t|t-1)^2 = \sum_{t=1}^{N} \left(y(t) - \hat{y}(t|t-1)\right)^2.$$

We apply the method of least squares. Let

$$y_t := y(t), \qquad \varphi_t := \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-n) \\ u(t-1) \\ \vdots \\ u(t-m) \end{bmatrix}, \qquad \theta^o := \begin{bmatrix} a_1 \\ \vdots \\ a_n \\ b_0 \\ \vdots \\ b_m \end{bmatrix}, \qquad \varepsilon_t := e(t);$$

then (6) can be rewritten $y_t = \varphi_t^\top \theta^o + \varepsilon_t$ for $t = 1, \cdots, N$. Now, the method of least squares finds the $\hat{\theta}_{\text{LS}}$ such that the sum $\sum_{t=1}^{N} \left(y_t - \varphi_t^\top \hat{\theta}\right)^2$ is minimal, and this is precisely what the PEM method asks. Note that there is nothing strange in having, among the regressors that explain the sample $y(t)$, some samples of the same process itself ($y(t-1), y(t-2)$ etc.). Indeed this is precisely the reason for the name *autoregressive*: the process *regresses on itself.*

We pose now the question as whether or not, if (6) is the *true* model that generates $\{y(t)\}$, $\hat{\theta}_{\text{LS}}$ converges to $\theta^o := \begin{bmatrix} a_1 & \cdots & a_n & b_0 & \cdots & b_m \end{bmatrix}^\top$. Since $e(t)$ is independent from $u(\cdot)$ and from the past of $y(\cdot)$, ε_t is independent from φ_t, and you could be tempted to apply Theorem 1.6.1. This cannot be done directly, because here the regressors $\{\varphi_t\}$ are neither independent nor identically distributed. However, the crucial fact is that they are independent from ε_t.

Theorem 3.2.1 *Let $\{y(t)\}_{t=0}^{\infty}$ be an AR process, and suppose that*

1. *$e(t)$ is independent from $y(t-1), y(t-2), \cdots$;*
2. *the roots of the polynomial $z^n + a_1 z^{n-1} + \cdots + a_n$ lie inside the interior of the unit circle ($\{z \in \mathbb{C} \mid |z| < 1\}$);*

then the least squares-estimate $\hat{\theta}_{\text{LS}}$ convergences almost surely to 0^o as $N \to \infty$.

Theorem 3.2.2 *Let $\{y(t)\}_{t=0}^{\infty}$ be an ARX process, and suppose that*

1. *the process $\{u(t)\}_{t=0}^{\infty}$ is wide sense stationary, with correlation sequence $r(\tau) = \mathsf{E}[u(t)u(t+\tau)]$;*
2. *the Toeplitz matrix*

$$M = \begin{bmatrix} r(0) & r(1) & \cdots & r(m-1) \\ r(1) & r(0) & \cdots & r(m-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(m-1) & r(m-2) & \cdots & r(0) \end{bmatrix}$$

   is positive definite (this hypothesis is called persistent excitation *of the input, and ensures that it carries enough information. Recall that the only real-world issue that can prevent uniqueness in least squares was "the regressors φ_t do not carry enough information".);*

3. *$e(t)$ is independent from $y(t-1), y(t-2), \cdots$ and from $u(s)$ for all s (this implies that* there is no feedback between y and u*);*
4. *the roots of the polynomial $z^n + a_1 z^{n-1} + \cdots + a_n$ lie inside the interior of the unit circle;*

then the least squares-estimate $\hat{\theta}_{\text{LS}}$ convergences almost surely to 0^o as $N \to \infty$.

3.3 Instrumental variables

Consider now a different situation, in which the process $y(\cdot)$ is generated by the linear system

$$y(t) = ay(t-1) + bu(t-1) \tag{7}$$

without process noise, with $u(t)$ independent from $y(s)$ for all $s \leq t$ (no feedback). The output $y(\cdot)$ is not accessible anymore; instead, the experimenter has access to measures of the output corrupted by noise:

$$y_m(t) = y(t) + e(t) \tag{8}$$

where $\{e(t)\}$ are i.i.d. random variables with mean 0, independent from $u(s)$ and $y(s)$ for all $s \leq t$. Is it still possible to apply the method of least squares to estimate a and b? It would seem so, because substituting (8) into (7) we obtain

$$\begin{aligned} y_m(t) - e(t) &= a(y_m(t-1) - e(t-1)) + bu(t-1) \\ y_m(t) &= ay_m(t-1) + bu(t-1) + (e(t) - ae(t-1)) \end{aligned}$$

and letting

$$y_t := y_m(t), \qquad \varphi_t := \begin{bmatrix} y_m(t-1) \\ u(t-1) \end{bmatrix}, \qquad \theta^o := \begin{bmatrix} a \\ b \end{bmatrix}, \qquad \varepsilon_t := e(t) - ae(t-1)$$

the model becomes, as before, $y_t = \varphi_t^\top \theta^o + \varepsilon_t$. Nevertheless, *the regressor and the disturbance are correlated,* therefore none of the theorems about almost sure convergence of $\hat{\theta}_{\text{LS}}$ apply. There is no way out of this issue within the standard theory of least squares; but there is a remedy, called the method of *instrumental variables.* Consider the equation from which the least squares solution $\hat{\theta}_{\text{LS}}$ has been derived:

$$\frac{1}{N}\sum_{t=1}^{N} \varphi_t \left(y_t - \varphi_t^\top \hat{\theta}\right) = \frac{1}{N}\sum_{t=1}^{N} \varphi_t \left(\varphi_t^\top (\theta^o - \hat{\theta}) + \varepsilon_t\right) = 0.$$

The main idea of least squares is that, in the limit, this equation becomes $\mathsf{E}\left[\varphi_t\left(\varphi_t^\top(\theta^o-\hat\theta)+\varepsilon_t\right)\right]=0$, and if 1) $\mathsf{E}\left[\varphi_t\varphi_t^\top\right]$ is invertible and 2) $\mathsf{E}\left[\varphi_t\varepsilon_t\right]=0$, then its only solution is $\hat\theta=\theta^o$. In the above example this is not the case, because the second condition fails. But if we replace (somehow heuristically) the first occurrence of the regressor φ_t with *another* variable ψ_t such that 1) $\mathsf{E}\left[\psi_t\varphi_t^\top\right]$ is invertible, and 2) $\mathsf{E}\left[\psi_t\varepsilon_t\right]=0$, then the equation $\mathsf{E}\left[\psi_t\left(\varphi_t^\top(\theta^o-\hat\theta)+\varepsilon_t\right)\right]=0$ has again the only solution $\hat\theta=\theta^o$; therefore it makes sense to solve

$$\left(\sum_{t=1}^N\psi_t\varphi_t^\top\right)\hat\theta=\sum_{t=1}^N\psi_t y_t$$

instead of the normal equations. The variable ψ_t, correlated with φ_t but *not* with ε_t, is called *instrumental*; in our case the vector $\psi_t:=\begin{bmatrix}u(t-2)\\u(t-1)\end{bmatrix}$ is a good instrumental variable, provided that $\mathsf{E}\left[u(t)^2\right]\neq 0$ for all t.

4 Machine learning

4.1 Statement of the problem

A sequence of i.i.d. random pairs $(U_1,Y_1),\cdots,(U_N,Y_N)$ is observed, where $U_i\in\mathbb{R}$ and $Y_i\in\{0,1\}$. We wish to study the behavior of functions $\hat f:\mathbb{R}\to\{0,1\}$ in providing an estimate $\hat Y=\hat f(U)$. The objective is to find the "best" possible classifier. Following the least squares principle, "best" means that it minimizes

$$\bar J=\mathsf{E}\left[(Y_i-\hat f(U_i))^2\right]$$

The best classifier $\hat f$ is called *Bayesian*, but it cannot be computed since it requires knowledge of the data-generation mechanism. Moreover, we restrict the choice to a family of functions $\mathcal{F}=\{\hat f_c\}_{c\in C}$ parameterized by some set C of indices, and the Bayesian classifier does not necessarily belong to $\mathcal{F}$.

Definition 4.1.1 *Error function:*

$$\bar J(c):=\mathsf{E}\left[(Y_i-\hat f_c(U_i))^2\right]=\mathsf{E}\left[\mathbb{1}_{Y_i\neq\hat f_c(U_i)}\right]=\mathsf{P}\left[Y_i\neq\hat f_c(U_i)\right]$$

The best classifier in the family is the one that minimizes $\bar J$:

$$\bar c=\arg\min_{c\in C}\bar J(c)$$

Definition 4.1.2 *Empirical error function, based on the data* $(U_1,Y_1),\cdots,(U_N,Y_N)$*:*

$$\hat J_N(c):=\frac{1}{N}\sum_{i=1}^N(Y_i-\hat f_c(U_i))^2=\frac{1}{N}\sum_{i=1}^N\mathbb{1}(Y_i\neq\hat f_c(U_i))$$

The empirical error is our approximation of the "true" error $\bar J(c)$. With the sole knowledge of the data we can minimize only this one:

$$\hat c_N=\arg\min_{c\in C}\hat J_N(c)$$

Three quantities of interest:

- $\bar J(\bar c)$, the minimum *true* error choosing among all classifiers in $\mathcal{F}$;
- $\hat J_N(\hat c_N)$, obtained from the minimization of the empirical error;
- $\bar J(\hat c_N)$, the true error obtained from the choice that is optimal for the empirical error.

We expect that:

- For big N, $\hat J_N(c)$ is close to $\bar J(c)$ with high probability;

- For $N \to \infty$, $\hat{J}_N$ tends to $\bar{J}$;
- For big N, $\hat{J}_N(\hat{c}_N)$ is close to $\bar{J}(\hat{c}_N)$ and $\bar{J}(\bar{c})$ with high probability.

In general, however, these statements are false (example with needle functions). The problem arises when $\mathcal{F}$ is too complex, and it is advisable to stay away from complexity (recall pictures on blackboard). For these properties to hold, we need that as $N \to \infty$, *the function* $\hat{J}_N$ *tends to* $\bar{J}$ *uniformly with respect to* c. Program to follow:

1. If $\hat{J}_N \to \bar{J}$ uniformly, $\min \hat{J}_N \to \min \bar{J}$;
2. Convergence of the empirical distribution of a real random variable to its true distribution;
3. Apply results on empirical distributions to the family of threshold classifiers; show that $\hat{J}_N \to \bar{J}$ uniformly; conclude that the minimum of $\hat{J}_N$ indeed tends to the minimum of $\bar{J}$.

4.2 Lemma on uniform convergence

Lemma 4.2.1 *Suppose that*

1. $\hat{c}_N := \arg\min_{c \in C} \hat{J}_N(c)$ *exists for all* N;
2. $\bar{c} := \arg\min_{c \in C} \bar{J}(c)$ *exists;*
3. $\hat{J}_N \to \bar{J}$ *uniformly;*

Then $\lim_{N \to \infty} \bar{J}(\hat{c}_N) = \bar{J}(\bar{c})$.

4.3 Convergence of the empirical distribution

Let $X_1, \cdots X_N$ be i.i.d. random variables with distribution

$$F(x) := \mathsf{P}[X \leq x] = \mathsf{E}[\mathbb{1}_{X \leq x}]$$

The distribution F is unknown; estimate it with the *empirical distribution:*

$$\hat{F}_N(x) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{X_i \leq x}$$

$\hat{F}_N$ is a good distribution function: it is monotone non-decreasing, continuous from the right, and its limits at $-\infty$ and $+\infty$ are 0 and 1 respectively. $\hat{F}_N(x)$ is an unbiased estimator of $F(x)$. Moreover, it is a consistent estimator of $F(x)$, because the strong law of large numbers implies:

Lemma 4.3.1 *For each* x, $\hat{F}_N(x)$ *converges to* $F(x)$ *almost surely.*

More: $\hat{F}_N(x)$ converges in probability to $F(x)$ with exponential rate, due to the following

Theorem 4.3.1 *(Hoeffding's inequality). Let* $Z_i \in [a_i, b_i]$ *be independent, bounded random variables for* $i = 1 \cdots N$, *and let* $S_N = \frac{1}{N} \sum_{i=1}^{N} Z_i$. *Then*

$$\mathsf{P}[|S_N - \mathsf{E}[S_N]| \geq \varepsilon] \leq 2 \exp\left(\frac{-2N^2\varepsilon^2}{\sum_{i=1}^{N} (b_i - a_i)^2} \right).$$

(Example on tossing a coin 100 times: Hoeffding's bound is much tighter than Čebyšev's one.) Apply Hoeffding's inequality to the empirical distribution. For a fixed x, let

$$\begin{aligned} Z_i = \mathbb{1}_{X_i \leq x} \quad &\Rightarrow \quad \mathsf{E}[Z_i] = F(x) \\ S_N = \hat{F}_N(x) \quad &\Rightarrow \quad \mathsf{E}[S_N] = F(x) \\ [a_i, b_i] &= [0, 1] \end{aligned}$$

Hoeffding's inequality yields

$$\mathsf{P}\left[|\hat{F}_N(x) - F(x)| \geq \varepsilon\right] \leq 2e^{-2N\varepsilon^2}$$

Hence, at all x the empirical distribution converges *in probability* to the true distribution, with exponential rate of convergence. Exponential convergence is so strong that it is enough to re-establish *almost sure* convergence as well. Not yet enough. More: $\hat{F}_N$ converges *uniformly* to F, due to the following

Theorem 4.3.2 *(Glivenko/Cantelli). Let $\{X_i\}$ be i.i.d. variables with distribution $F(x)$. Then*

$$\lim_{N\to\infty} \sup_{x\in\mathbb{R}} |\hat{F}_N(x) - F(x)| = 0 \quad \textit{almost surely.}$$

Probabilistic bounds for fixed ε, N, applying Hoeffding's inequality to the proof of Glivenko/Cantelli:

Lemma 4.3.2 *Fix $\varepsilon > 0, N$. Then*

$$\mathsf{P}\left[\sup_{x\in\mathbb{R}} |\hat{F}_N(x) - F(x)| \geq \varepsilon\right] \leq \frac{4}{\varepsilon} e^{-N\varepsilon^2/2}$$

(The exponent now has $1/2$ instead of 2, because the proof of Glivenko/Cantelli's theorem is based on splitting in subintervals of length $\varepsilon/2$.)

4.4 Extension to two variables, one of which is binary

Let $\{(X_i, Y_i)\}$ be i.i.d., where $X_i \in \mathbb{R}$ and $Y_i \in \{0, 1\}$. Let $\mathsf{P}[Y_i = 0] = \alpha$. Let

$$\begin{aligned}
F^0(x) &= \mathsf{P}[X_i \leq x, Y_i = 0] \quad (\text{tends to } \alpha \text{ for } x \to \infty) \\
F^1(x) &= \mathsf{P}[X_i \leq x, Y_i = 1] \quad (\text{tends to } 1-\alpha \text{ for } x \to \infty) \\
\hat{F}^0_N(x) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{X_i \leq x, Y_i = 0\} \quad (\text{empirical version of } F^0(x)) \\
\hat{F}^1_N(x) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{X_i \leq x, Y_i = 1\} \quad (\text{empirical version of } F^1(x))
\end{aligned}$$

Glivenko/Cantelli-like theorem:

$$\begin{aligned}
&\sup_{x\in\mathbb{R}} |\hat{F}^0_N(x) - F^0(x)| \to 0 \quad \text{almost surely;} \\
&\sup_{x\in\mathbb{R}} |\hat{F}^1_N(x) - F^1(x)| \to 0 \quad \text{almost surely.}
\end{aligned}$$

Hoeffding-like probabilistic bound for fixed ε, N, like Lemma 4.3.2:

Lemma 4.4.1 *Let $\{(X_i, Y_i)\}$ be i.i.d. where $X_i \in \mathbb{R}$ and $Y_i \in \{0, 1\}$. Let $\mathsf{P}[Y_i = 0] = \alpha$. Then, for fixed $\varepsilon > 0, N$,*

$$\begin{aligned}
\mathsf{P}\left[\sup_{x\in\mathbb{R}} |\hat{F}^0_N(x) - F^0(x)| \geq \varepsilon\right] &\leq \frac{4\alpha}{\varepsilon} e^{-N\varepsilon^2/2} \\
\mathsf{P}\left[\sup_{x\in\mathbb{R}} |\hat{F}^1_N(x) - F^1(x)| \geq \varepsilon\right] &\leq \frac{4(1-\alpha)}{\varepsilon} e^{-N\varepsilon^2/2}
\end{aligned}$$

4.5 Threshold classifiers

Consider the family of *threshold classifiers*:

$$\mathcal{F} = \{\mathbb{1}_{(-\infty,c]}(\cdot) \mid c \in \mathbb{R}\}.$$

Existence of $\hat{c}_N$: always. Uniqueness: never. Existence and uniqueness of $\bar{c}$: It depends on the data-generation rule. Assume as hypothesis.

Theorem 4.5.1 *Let $(U_1, Y_1), \cdots, (U_N, Y_N)$ be i.i.d. where U_i has continuous distribution $F(u)$ and $Y_i \in \{0, 1\}$. Assume that $\bar{c}$ exists. Then:*

1. *Almost surely $\hat{J}_N \to \bar{J}$ uniformly;*
2. *Almost surely $\bar{J}(\hat{c}_N) \to \bar{J}(\bar{c})$;*
3. *For fixed $\varepsilon > 0$ and N, it holds*

$$\mathsf{P}\left[\sup_{c \in C} |\hat{J}_N(c) - \bar{J}(c)| \geq \varepsilon\right] \leq \frac{8}{\varepsilon} e^{-N\varepsilon^2/8}.$$

Proof. For each U_i define $V_i = -U_i$ and denote $G(u)$ the distribution of V_i. Define as before

$$F^0(u) = \mathsf{P}[U_i \leq u, Y_i = 0] \qquad \hat{F}^0_N(u) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{U_i \leq u, Y_i = 0\}$$

$$G^1(u) = \mathsf{P}[V_i \leq u, Y_i = 1] \qquad \hat{G}^1_N(u) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{V_i \leq u, Y_i = 1\}$$

We can show that:

$$\bar{J}(c) = F^0(c) + G^1(-c)$$

And for its empirical counterpart:

$$\hat{J}_N(c) = \hat{F}^0_N(c) + \hat{G}^1_N(-c) - \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(V_i = -c, Y_i = 1)$$

It follows:

$$\left|\hat{J}_N(c) - \bar{J}(c)\right| \leq \left|\hat{F}^0_N(c) - F^0(c)\right| + \left|\hat{G}^1_N(-c) - G^1(-c)\right| + \left|\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(V_i = -c, Y_i = 1)\right|$$

As $N \to \infty$, the supremum over C of the first two terms tends almost surely to zero by Glivenko/Cantelli, and the third one tends almost surely to zero by the strong law of large numbers. This establishes point 1. Apply the lemma on uniform convergence. This establishes point 2. Now fix ε and N. By Lemma 4.4.1,

$$\mathsf{P}[\mathcal{A}] = \mathsf{P}\left[\sup_{c \in C}\left|\hat{F}^0_N(c) - F^0(c)\right| \geq \varepsilon\right] \leq \frac{4\alpha}{\varepsilon} e^{-N\varepsilon^2/2}$$

$$\mathsf{P}[\mathcal{B}] = \mathsf{P}\left[\sup_{c \in C}\left|\hat{G}^1_N(-c) - G^1(-c)\right| \geq \varepsilon\right] \leq \frac{4(1-\alpha)}{\varepsilon} e^{-N\varepsilon^2/2}$$

On the other hand, the event $\left|\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(V_i = -c, Y_i = 1)\right| \geq \varepsilon$ has probability 0 for all $\varepsilon > 0$, since V_i has continuous distribution. Putting together the inequalities,

$$\mathsf{P}\left[\sup_{c \in C} |J_N(c) - \bar{J}(c)| \geq 2\varepsilon\right] \leq \frac{4}{\varepsilon} e^{-N\varepsilon^2/2}$$

Point 3 of the claim follows from a change of variable $2\varepsilon = \varepsilon'$. □

One can establish bounds for the minima as well (they follow from the proof of the lemma on uniform convergence): for fixed $\varepsilon > 0$ and N, both the events $\bar{J}(\hat{c}_N) \leq J_N(\hat{c}_N) + \varepsilon$ and $\bar{J}(\hat{c}_N) \leq \bar{J}(\bar{c}) + 2\varepsilon$ have probability at least $1 - \frac{8}{\varepsilon} e^{-N\varepsilon^2/8}$.
Almost sure convergence: good. Probabilistic bounds coming from Hoeffding's inequality: not impressive unless there are many data available. Not the tightest bound available. In practice, the empirical error behaves better than how the bound says. However, this bound is remarkable because it is *distribution-free*: it does not depend on either the distribution of the data or the mechanism that links input to output.

5 The LSCR method

5.1 Introduction and motivation

Suppose that $\{y_1, \cdots, y_N\}$ are i.i.d. random variables drawn from a Gaussian distribution $N(\theta^o, \sigma^2)$ whose mean θ^o is not known. We want to extract information about θ^o from the data. The sample can be seen as N measures $y_i = \theta^o + \varepsilon_i, \quad i = 1, \cdots, N$, to be "explained" by the "true" parameter θ^o, corrupted by errors $\varepsilon_i \sim N(0, \sigma^2)$. LS approach: provide a *point estimate* of θ^o minimizing the sum of the squared deviation. Result: $\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} y_i$, the *sample average*, which is an unbiased estimator of θ^o ($\mathsf{E}[\hat{\theta}] = \theta^o$). A "guarantee" on the usefulness of $\hat{\theta}$ should be a "certificate" that $\hat{\theta}$ and θ^o are *probably* close. Such "certificate" usually comes in two forms: 1) the variance of $\hat{\theta}$ (the smaller, the better); 2) a statement like "the probability $\mathsf{P}[|\hat{\theta} - \theta^o| > d]$ is α", for a certain threshold d and significance α, say $\alpha = 5\%$. The second "certificate" can also be stated: the interval

$$I_\alpha = \left[\hat{\theta} - d,\ \hat{\theta} + d\right]$$

contains θ^o with *probability* $1 - \alpha$. Be careful about the use of the word "probability". *Before the sample is drawn*, the interval I_α is random, and whether or not it will contain θ^o depends on the outcome of the experiment; namely, it will happen with *probability* $1 - \alpha$. But *after the sample has been drawn* the interval is deterministic. Either it contains θ^o or it does not. Thus, $1 - \alpha$ is not anymore the "probability" of anything; the usual name for it is *confidence*. For example, in applied statistics the following interval

$$I_\alpha = \left[\hat{\theta} - \frac{\bar{s} t_\alpha}{\sqrt{N}},\ \hat{\theta} + \frac{\bar{s} t_\alpha}{\sqrt{N}}\right]$$

is often used as a $(1 - \alpha)$-*confidence interval* for θ^o. Here, $\bar{s}$ is the square root of the *sample variance* $\bar{s}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \hat{\theta})^2$, which is an unbiased estimator of σ^2, and t_α is a "percentile" such that $\int_{-t_\alpha}^{t_\alpha} f(x)\, dx = 1 - \alpha$, where $f(x)$ is the density of a Student's $t(N-1)$ random variable[1]. This is a beautiful *pivotal* result: it does not depend on the knowledge of any parameter θ^o or σ^2, only on the number of data N. But it has a fundamental drawback: *it still depends crucially on Gaussianity.* Is it possible to provide a confidence interval for θ^o without restricting the distribution of the errors to some parametric family?

5.2 Confidence intervals with the LSCR method

An *Abelian group* $(G, +)$ is a set G endowed with an operation $+$ such that: 1) for all $a, b, c \in G$, $((a+b)+c) = (a+(b+c))$ (associativity); 2) there exist an element $0 \in G$ (zero) such that $0 + a = a + 0 = a$ for all $a \in G$; 3) for all $a \in G$, there exist an element $-a \in G$ (inverse of a) such that $a + (-a) = (-a) + a = 0$; 4) for all $a, b \in G$, $a + b = b + a$ (commutativity). A *subgroup* of $(G, +)$ is a set $H \subseteq G$ which is closed with respect to the operation $+$ (if $a, b \in H$, then also $a + b \in H$). The simplest nontrivial *finite* Abelian group is $B = (\{\circ, \bullet\}, +)$, where addition is defined in the following way:

$+$	$\circ$	$\bullet$
$\circ$	$\circ$	$\bullet$
$\bullet$	$\bullet$	$\circ$

[1] The shape of a $t(N-1)$ density resembles that of a Gaussian $N(0,1)$, namely it is symmetric around 0 and is close to the Gaussian one, but it has *fatter tails*. As $N \to \infty$, the $t(N-1)$ density converges to the $N(0,1)$ density.

(its zero is ○; recall modulo 2 arithmetics). And here is a group that will serve us for a "canonical" example:

$G_7 =$

	1	2	3	4	5	6	7
I_1	●	●	○	●	●	○	○
I_2	●	○	●	●	○	●	○
I_3	○	●	●	○	●	●	○
I_4	●	●	○	○	○	●	●
I_5	●	○	●	○	●	○	●
I_6	○	●	●	●	○	○	●
I_7	○	○	○	●	●	●	●
I_8	○	○	○	○	○	○	○

It is a a subgroup of $(B^7, +)$ containing 8 elements, i.e. the rows $I_1, \cdots, I_8$; addition is defined component-wise. Any two rows in the table have another row as their sum; the zero is I_8, and the inverse of a row is the row itself.

Suppose that we are given 7 (very few) measures $y_i = \theta^o + \varepsilon_i$ for $i = 1, \cdots, 7$, where $\{\varepsilon_i\}$ are *continuous variables with a density centered around zero, independent but not necessarily identically distributed.* Consider for each measure an affine function: $f_i(\theta) = y_i - \theta$. For each of the first 7 elements $I_i = I_1, \cdots, I_7$ of the group, build a point-wise average of the 4 functions whose indices are marked • in the corresponding row:

$$g_1(\theta) = \frac{1}{4} \sum_{k\in\{1,2,4,5\}} f_k(\theta) = \frac{1}{4} \sum_{k\in\{1,2,4,5\}} y_k - \theta = \frac{1}{4} \sum_{k\in\{1,2,4,5\}} \theta^o + \varepsilon_k - \theta = (\theta^o - \theta) + \frac{1}{4} \sum_{k\in\{1,2,4,5\}} \varepsilon_k,$$

etc. Each of the g_i has exactly one intersection θ_i with the θ-axis. The θ_i's are *random* numbers coming in some order; since the errors ε_i are independent and have continuous distributions, almost surely no two of the θ_i coincide, and none of them equals θ^o. Furthermore, let $\bar{\theta}_1, \cdots, \bar{\theta}_7$ denote the same seven numbers, but *sorted*, i.e. such that $\bar{\theta}_1 < \bar{\theta}_2 < \cdots < \bar{\theta}_7$. They split the θ-axis in 8 intervals (the outermost ones are semi-infinite). Now note that at $\theta = \theta^o$ it holds

$$g_1(\theta^o) = \frac{1}{4} \sum_{k\in\{1,2,4,5\}} \varepsilon_k, \quad \cdots, \quad g_7(\theta^o) = \frac{1}{4} \sum_{k\in\{4,5,6,7\}} \varepsilon_k.$$

Each $g_i(\theta^o)$ is a sum of independent variables, *with density symmetric around* 0. *Hence, $g_i(\theta^o)$ has also a density symmetric around* 0, and it has equal probabilities of being positive or negative. The fundamental idea of LSCR is to compare the signs of each $g_i(\theta^o)$, depending on which one of the 8 intervals contains θ^o. If $\theta^o \in (-\infty, \bar{\theta}_1)$, then $g_1(\theta^o) > 0, \cdots, g_7(\theta^o) > 0$; if $\theta^o \in (\bar{\theta}_7, \infty)$, then $g_1(\theta^o) < 0, \cdots, g_7(\theta^o) < 0$; if θ^o belongs to the second interval $(\bar{\theta}_1, \bar{\theta}_2)$, then one and only one among $g_1(\theta^o), g_2(\theta^o), \cdots, g_7(\theta^o)$ has negative sign, etc. *What is the probability of these events?* The first one ($\theta^o \in (-\infty, \bar{\theta}_1)$, hence $g_1(\theta^o) > 0, \cdots, g_7(\theta^o) > 0$) happens when

$$\begin{array}{cccccccc}
\varepsilon_1 & +\varepsilon_2 & & +\varepsilon_4 & +\varepsilon_5 & & & > 0, \\
\varepsilon_1 & & +\varepsilon_3 & +\varepsilon_4 & & +\varepsilon_6 & & > 0, \\
 & \varepsilon_2 & +\varepsilon_3 & & +\varepsilon_5 & +\varepsilon_6 & & > 0, \\
\varepsilon_1 & +\varepsilon_2 & & & & +\varepsilon_6 & +\varepsilon_7 & > 0, \\
\varepsilon_1 & & +\varepsilon_3 & & +\varepsilon_5 & & +\varepsilon_7 & > 0, \\
 & \varepsilon_2 & +\varepsilon_3 & +\varepsilon_4 & & & +\varepsilon_7 & > 0, \\
 & & & \varepsilon_4 & +\varepsilon_5 & +\varepsilon_6 & +\varepsilon_7 & > 0.
\end{array} \tag{9}$$

If $\theta^o \in (\bar{\theta}_1, \bar{\theta}_2)$, then exactly one $g_i(\theta^o)$ among $g_1(\theta^o), \cdots, g_7(\theta^o)$ has negative sign, hence all the other values at θ^o (including $g_8(\theta^o) = 0$) are greater than it. Suppose, without loss of generality, that such g_i is g_1. Then

$$\begin{array}{cccccccccccc}
 & & & 0 & & & & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
\varepsilon_1 & & +\varepsilon_3 & +\varepsilon_4 & & +\varepsilon_6 & & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
 & \varepsilon_2 & +\varepsilon_3 & & +\varepsilon_5 & +\varepsilon_6 & & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
\varepsilon_1 & +\varepsilon_2 & & & & +\varepsilon_6 & +\varepsilon_7 & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
\varepsilon_1 & & +\varepsilon_3 & & +\varepsilon_5 & & +\varepsilon_7 & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
 & \varepsilon_2 & +\varepsilon_3 & +\varepsilon_4 & & & +\varepsilon_7 & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
 & & & \varepsilon_4 & +\varepsilon_5 & +\varepsilon_6 & +\varepsilon_7 & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5.
\end{array}$$

Simplifying and bringing every term to the left-hand side we obtain

$$\begin{array}{cccccccc}
-\varepsilon_1 & -\varepsilon_2 & & -\varepsilon_4 & -\varepsilon_5 & & & > 0, \\
 & -\varepsilon_2 & +\varepsilon_3 & & -\varepsilon_5 & +\varepsilon_6 & & > 0, \\
-\varepsilon_1 & & +\varepsilon_3 & -\varepsilon_4 & & +\varepsilon_6 & & > 0, \\
 & & & -\varepsilon_4 & -\varepsilon_5 & +\varepsilon_6 & +\varepsilon_7 & > 0, \\
 & -\varepsilon_2 & +\varepsilon_3 & -\varepsilon_4 & & & +\varepsilon_7 & > 0, \\
-\varepsilon_1 & & +\varepsilon_3 & & -\varepsilon_5 & & +\varepsilon_7 & > 0, \\
-\varepsilon_1 & -\varepsilon_2 & & & & +\varepsilon_6 & +\varepsilon_7 & > 0.
\end{array} \tag{10}$$

The sets of inequalities (9) and (10) are very similar, the differences being that the inequalities appear in different orders, and that some of the signs are changed. (9) *and* (10) *contain terms with the same indices due to the group structure, and they hold with the same probability because all the* ε_i *have symmetric densities.* The same reasoning applies to the entire sets of inequalities. Hence *the probabilities that* $\theta^o \in (-\infty, \bar{\theta}_1)$ *and that* $\theta^o \in (\bar{\theta}_1, \bar{\theta}_2)$ *are equal.* But the very same procedure could have been applied to any of the intervals; hence θ^o belongs to any of the intervals with the same probability $\frac{1}{8}$. And finally, the probability that θ^o belongs to any of the two outermost intervals is $\frac{1}{8} + \frac{1}{8} = \frac{1}{4}$, hence:

Theorem 5.2.1 *(Campi, Weyer). Let* $y_i = \theta^o + \varepsilon_i, \quad i = 1, \cdots, 7$*; suppose that* $\{\varepsilon_i\}$ *are continuous variables with a density centered around zero (with mean zero), independent but not necessarily identically distributed. Construct the functions* $g_1(\cdot), \cdots, g_7(\cdot)$ *and sort their intersections with the* θ*-axis,* $\bar{\theta}_1 < \cdots < \bar{\theta}_7$*, as above. Then* $I := [\bar{\theta}_1,\ \bar{\theta}_7]$ *is a* $\frac{3}{4} = 75\%$*-confidence interval for* θ^o*.*

The applicability of the method is not limited to 7 measures. To cope with the general, N-measures case one must construct a suitable subgroup of $(B^N, +)$, "balanced" in the sense that it should be fairly small, the number of bullets ($\bullet$) should be the same in each row, and it should be approximately half the size of a row. This construction is easy when $N = 2^n - 1$ for some n. For $N = 3 = 2^2 - 1$ and $N = 7 = 2^3 - 1$ we have:

$G_3 =$

	1	2	3
I_1	$\bullet$	$\bullet$	$\circ$
I_2	$\bullet$	$\circ$	$\bullet$
I_3	$\circ$	$\bullet$	$\bullet$
I_4	$\circ$	$\circ$	$\circ$

$G_7 =$

	1	2	3	4	5	6	7
I_1	$\bullet$	$\bullet$	$\circ$	$\bullet$	$\bullet$	$\circ$	$\circ$
I_2	$\bullet$	$\circ$	$\bullet$	$\bullet$	$\circ$	$\bullet$	$\circ$
I_3	$\circ$	$\bullet$	$\bullet$	$\circ$	$\bullet$	$\bullet$	$\circ$
I_4	$\bullet$	$\bullet$	$\circ$	$\circ$	$\circ$	$\bullet$	$\bullet$
I_5	$\bullet$	$\circ$	$\bullet$	$\circ$	$\bullet$	$\circ$	$\bullet$
I_6	$\circ$	$\bullet$	$\bullet$	$\bullet$	$\circ$	$\circ$	$\bullet$
I_7	$\circ$	$\circ$	$\circ$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
I_8	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$

$=$

	1	2	3	4	5	6	7
I_1							$\circ$
I_2		G_3			G_3		$\circ$
I_3							$\circ$
I_4							$\bullet$
I_5		G_3			$\bar{G}_3$		$\bullet$
I_6							$\bullet$
I_7	$\circ$	$\circ$	$\circ$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
I_8	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$

Balanced groups for the next powers of 2 are built recursively in this fashion. Since approximately $\frac{N}{2}$ noise terms are averaged in each intersection, the confidence interval "shrinks" towards θ^o as N increases. Since all of the $N+1$ intervals have the same probability $\frac{1}{N+1}$, some more intervals can be discarded, other than the two outermost ones, to obtain a smaller interval (with of course a smaller confidence). Thus, the confidence is "tunable", although not continuously as in the Gaussian case.

5.3 The case with inputs

LSCR allows to find a confidence interval for θ^o also when an input is involved. Let $y_i = \theta^o u_i + \varepsilon_i$ for $i = 1, \cdots, 7$. Suppose that $\{\varepsilon_i\}$ are independent variables with a density symmetric around zero. The inputs u_i can be deterministic or random, but if we want the above construction to be useful and the confidence interval to "shrink" towards θ^o for big N, they should "stay away from zero". If they are deterministic, they should be nonzero, and should not tend to zero as $N \to \infty$; if they are random, they should be independent from ε_i, and have nonzero mean. Consider, for each measure, an affine function: $f_i(\theta) = y_i - \theta u_i$. Differently from before, these functions may have different slopes. Exactly as before, they intersect the θ-axis at one point (*unless* $u_i = 0$*, but we have excluded this case*). For each $I_i = I_1, \cdots, I_7$ build an average of 4 different f_k:

$$g_1(\theta) = \frac{1}{4} \sum_{k \in \{1,2,4,5\}} f_k(\theta) = \frac{1}{4} \sum_{k \in \{1,2,4,5\}} y_k - \theta u_k = (\theta^o - \theta) \frac{1}{4} \sum_{k \in \{1,2,4,5\}} u_k + \frac{1}{4} \sum_{k \in \{1,2,4,5\}} \varepsilon_k,$$

etc. The reasoning of the above section applies without other changes, and $[\bar{\theta}_1,\ \bar{\theta}_7]$ is a 75%-confidence interval for θ^o. What happens if the inputs $\{u_i\}$ are random, independent from the $\{\varepsilon_i\}$, but their mean is 0? Consider the intersection of (say) $g_1(\theta)$ with the θ-axis. It is the point θ_1 such that

$$0 = g_1(\theta_1) = (\theta^o - \theta_1)\left(\frac{1}{4}\sum_{k\in\{1,2,4,5\}} u_k\right) + \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_k; \qquad \theta_1 = \theta^o + \frac{\frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_k}{\frac{1}{4}\sum_{k\in\{1,2,4,5\}} u_k}.$$

As more measures come, the term $\frac{1}{N+1}\sum_k u_k$ at the denominator of the last expression, which is also the coefficient of θ in the expression of $g_1(\theta)$, tends to 0. The same happens for all the g_i; hence the straight lines corresponding to the g_i are approximately horizontal, their intersections tend to be far from each other, and the confidence interval is practically useless. Here is the remedy: suppose, to fix ideas, that the inputs are i.i.d., that $\mathsf{E}[u_i] = 0$, and $\mathsf{E}[u_i^2] > 0$. We consider, instead of the affine functions above, the following ones:

$$f_i(\theta) = (y_i - \theta u_i)u_i = y_i u_i - \theta u_i^2,$$

$$g_1(\theta) = \frac{1}{4}\sum_{k\in\{1,2,4,5\}} f_k(\theta) = \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \theta^o u_k^2 + \varepsilon_k u_k - \theta u_k^2 = (\theta^o - \theta)\left(\frac{1}{4}\sum_{k\in\{1,2,4,5\}} u_k^2\right) + \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_k u_k,$$

etc. Intersections with θ-axis:

$$0 = g_1(\theta_1) = (\theta^o - \theta_1)\left(\frac{1}{4}\sum_{k\in\{1,2,4,5\}} u_k^2\right) + \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_k u_k; \qquad \theta_1 = \theta^o + \frac{\frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_k u_k}{\frac{1}{4}\sum_{k\in\{1,2,4,5\}} u_k^2}.$$

The denominator is OK now: as more measures come, the term $\frac{1}{N+1}\sum_k u_k^2$ tends to $\mathsf{E}[u_i^2]$, and the functions g_i have slopes with comparable magnitude. On the other hand, as far as the $\{u_i\}$ are independent from the $\{\varepsilon_i\}$, *the terms $\varepsilon_k u_k$ still have densities symmetric around* 0, so that the fundamental idea of LSCR (sets of inequalities having same probability) applies. Thus, $[\bar{\theta}_1,\ \bar{\theta}_7]$ is once again a 75%-confidence interval for θ^o.

5.4 Leave-out Sign-dominant Correlation Regions

The LSCR method generalizes to system identification. Consider, for example, the simple AR process $y_i = \theta^o y_{i-1} + \varepsilon_i$. We cannot use anymore affine functions like $f_i(\theta) = y_i - \theta y_{i-1}$, because typically $\mathsf{E}[y_{i-1}] = 0$ (confidence intervals get larger and larger). Choose the following functions instead:

$$f_i(\theta) = (y_{i+1} - \theta y_i)(y_i - \theta y_{i-1})$$

$$g_1(\theta) = \frac{1}{4}\sum_{k\in\{1,2,4,5\}} f_k(\theta) = \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \left((\theta^o - \theta)y_k + \varepsilon_{k+1}\right)\left((\theta^o - \theta)y_{k-1} + \varepsilon_k\right)$$

$$= \frac{(\theta^o - \theta)^2}{4}\sum_{k\in\{1,2,4,5\}} y_k y_{k-1} + \frac{\theta^o - \theta}{4}\sum_{k\in\{1,2,4,5\}} y_k\varepsilon_k + \frac{\theta^o - \theta}{4}\sum_{k\in\{1,2,4,5\}} y_{k-1}\varepsilon_{k+1} + \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_{k+1}\varepsilon_k,$$

etc. The g_i are parabulas, their intersections with the θ-axis are not anymore unique. However, they split the θ-axis in 8 *regions*, that are unions of one or more disjoint intervals, where one out of 8 situations happen: 1) *all the g_i are positive*; 2) *exactly* 1 *of the g_i is negative*, 3) *exactly* 2 *of the g_i are negative*, ..., 8) *all the g_i are negative*. For each region we repeat the reasoning: *what happens if θ^o belongs to this region?* Note that since ε_{k+1} and ε_k are independent, their product $\varepsilon_{k+1}\varepsilon_k$ also has symmetric density. Hence $g_1(\theta^o) = \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_{k+1}\varepsilon_k$ is again *an average of terms with symmetric density*. Repeat the LSCR proof, which is based *only* on the symmetry and the group structure, to conclude that the regions have equal probability $\frac{1}{8}$. A function like

$$g_1(\theta) = \frac{1}{4}\sum_{k\in\{1,2,4,5\}} (y_{k+1} - \theta y_k)(y_k - \theta y_{k-1}) = \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \epsilon_k(\theta)\epsilon_{k-1}(\theta)$$

is an empirical 1-*lag correlation*; this explains the name LSCR: Leave-out Sign-dominant *Correlation* Regions.

6 Interval predictor models

6.1 Prediction intervals without explanatory data

The least squares method tells us how to construct a model $y = \hat{a}_{\rm LS} + \hat{b}_{\rm LS}\, x$ from measures $(x_1, y_1), \cdots, (x_N, y_N)$. One of the objectives of such a model is to *predict* the value of a future variable y_{N+1} when the corresponding explanatory variable x_{N+1} will be available: $\hat{y}_{N+1} = \hat{a}_{\rm LS} + \hat{b}_{\rm LS}\, x_{N+1}$. But what does it mean that $\hat{y}_{N+1}$ *predicts* the value of y_{N+1}? $\hat{y}_{N+1}$ is just a "plausible" value, that comes with the hope that the future observation be close to it. Our way to quantify "hope" is probability, hence the prediction is meaningful if we can show that the density of y_{N+1} is concentrated around $\hat{y}_{N+1}$. Usually, this means two things: 1) that the means of y_{N+1} and $\hat{y}_{N+1}$ are equal, and the variance of their difference is small; 2) that there exists an interval $[a, b]$, preferably small, that contains both $\hat{y}_{N+1}$ and y_{N+1} with known probability (the higher, the better). Here we will focus on the second one. Let us start without explanatory variables (x_i). Goal: given $y_1, \cdots, y_N$, construct an interval that contains y_{N+1} with certified probability. If $y_1, \cdots, y_N$ are Gaussian, it can be done with Student's t ($[a, b] = \left[\hat{\theta} - t_\alpha \bar{s}\sqrt{1 + \frac{1}{N}},\ \ \hat{\theta} + t_\alpha \bar{s}\sqrt{1 + \frac{1}{N}}\right]$). However, here we want a *pivotal* (distribution-free) result.

Let us consider a "chain" of examples, that get closer and closer to the main point.
Example. We have a list of $N + 1$ different numbers. We write each number on a ball, and mark the two balls with the minimum and the maximum number as "extreme point". Now we put the balls in an urn, shake, and extract one ball. What is the probability that the extracted ball is an extreme? Of course, $\frac{2}{N+1}$.
Example. We extract *all* the balls in random order. What is the probability that the *last* ball is an extreme? If all the different $(N + 1)!$ orders have the same probability, the selection of the last ball is equivalent to the extraction of a single ball. Hence the requested probability is again $\frac{2}{N+1}$. Put another way, the permutations of the balls that have either the minimum or the maximum at the last position are $N! + N! = 2N!$, hence the probability of extracting a permutation with an extreme at the last position is $\frac{2N!}{(N+1)!} = \frac{2}{N+1}$.
Example. We extract $N + 1$ independent samples $y_1, y_2, \cdots, y_{N+1}$ of a random variable having density $f(y)$. What is the probability P that y_{N+1} is either the maximum or the minimum? Denote the event "the last number y_{N+1} is an extreme" as $\mathcal{E}$, and the event "any two numbers among $y_1, \cdots, y_{N+1}$ are equal" as $\mathcal{N}$. Due to the hypothesis that the random variables have a density, the probability of $\mathcal{N}$ is 0, and the probability of $\mathcal{E}$ conditioned to the extraction of $N + 1$ different numbers is always the same ($\frac{2}{N+1}$). Hence

$$\begin{aligned}
P &= \int_{\mathbb{R}^{N+1}} \mathsf{P}\left[\mathcal{E} \mid y_1, \cdots, y_{N+1}\right] f(y_1) \cdots f(y_{N+1})\, dy_1 \ \cdots\ dy_{N+1} \\
&= \int_{\mathbb{R}^{N+1} \setminus \mathcal{N}} \mathsf{P}\left[\mathcal{E} \mid y_1, \cdots, y_{N+1}\right] f(y_1) \cdots f(y_{N+1})\, dy_1 \ \cdots\ dy_{N+1} \\
&= \int_{\mathbb{R}^{N+1} \setminus \mathcal{N}} \frac{2}{N+1}\, f(y_1) \cdots f(y_{N+1})\, dy_1 \ \cdots\ dy_{N+1} \\
&= \frac{2}{N+1} \int_{\mathbb{R}^{N+1} \setminus \mathcal{N}} f(y_1) \cdots f(y_{N+1})\, dy_1 \ \cdots\ dy_{N+1} = \frac{2}{N+1}.
\end{aligned}$$

Example. We extract N independent samples $y_1, y_2, \cdots, y_N$, from a random variable with density $f(y)$. Let a and b be their minimum and maximum. What is the probability that a "future" sample y_{N+1} falls outside the interval $[a, b]$? The probability is $\frac{2}{N+1}$, because y_{N+1} is outside $[a, b]$ exactly whenever it is either the maximum or the minimum of the whole sample. We have proven the following

Lemma 6.1.1 *Let $y_1, y_2, \cdots, y_N$ be independent and identically distributed variables with density $f(y)$. Let a and b be their minimum and maximum respectively. Then $[a, b]$ is a prediction interval for a future sample y_{N+1}, with probability $1 - \frac{2}{N+1} = \frac{N-1}{N+1}$.*

6.2 Convex problems and Helly's theorem

Definition 6.2.1 *We will call* convex problem *an optimization problem of the form:*

$$\begin{aligned} &\textit{minimize} && f(\theta) \\ &\textit{subject to} && f_1(\theta) \leq 0, \\ & && \vdots \\ & && f_n(\theta) \leq 0, \\ & && \theta \in \Theta \subseteq \mathbb{R}^d, \end{aligned}$$

where $f, f_1, \cdots, f_n$ are convex functions and Θ is a convex set. Each of the inequalities $f_i(\theta) \leq 0$ is called a constraint. *The problem is said to be* feasible *if there exists at least a point $\bar{\theta} \in \Theta$ that satisfies all the constraints, that is, such that $f_i(\bar{\theta}) \leq 0$ for all $i = 1, \cdots, n$.*

A convex problem may or may not have a solution, and a solution may not be unique. The fundamental fact about convex problems is that if a point is *locally* a minimum, then it is a minimum also *globally*, hence a solution to the problem. Convex problems are considered "easy", because we have readily available efficient and robust algorithms to solve them numerically. We consider $(d+1)$-dimensional problems of this form:

$$\begin{aligned} &\text{minimize} && \gamma \\ &\text{subject to} && g_1(\theta) - \gamma \leq 0 \\ & && \vdots \\ & && g_n(\theta) - \gamma \leq 0, \\ & && \theta \in \mathbb{R}^d, \gamma \in \mathbb{R}, \end{aligned} \tag{11}$$

where $g_1, \cdots, g_n$ are convex functions. The goal function $f(\theta, \gamma) = \gamma$ is trivially convex, and if $g_i(\theta)$ is convex, then the constraint function $f_i(\theta, \gamma) = g_i(\theta) - \gamma$ is also convex. We assume that all the problems under consideration are feasible and admit a unique solution (θ^*, γ^*).

Definition 6.2.2 *Let (θ^*, γ^*) be the solution of* (11)*, and consider one of its constraints, $g_i(\theta) - \gamma \leq 0$. We will call the latter a* support constraint *if the solution $(\theta^{**}, \gamma^{**})$ of the problem obtained by removing the constraint is strictly "better" than (θ^*, γ^*), meaning that it attains $\gamma^{**} < \gamma^*$.*

In other words, removing a support constraint the solution "falls". If a constraint is *not* a support constraint, then it can be removed from the problem without any consequence (the solution is the same). If we *add* a constraint to a problem, and it happens to become a support constraint in the new problem, then the new γ^* of the solution must *increase* with respect to the old solution (indeed, removing it again γ^* must fall).

Theorem 6.2.1 *(Helly). Let $S_1, \cdots, S_N$ be convex subsets of $\mathbb{R}^n$. If the intersection of any $n+1$ of these subsets is nonempty, then the intersection of* all of them *is nonempty.*

Theorem 6.2.2 *(Campi, Calafiore, Garatti; "scenario approach"). Any feasible convex problem like* (11) *has at most $d+1$ support constraints.*

Proof. Let (θ^*, γ^*) be the solution to the problem, and define

$$\begin{aligned} S_1 &= \left\{(\theta, \gamma) \in \mathbb{R}^{d+1} \;\middle|\; g_1(\theta) \leq \gamma\right\}, \\ &\vdots \\ S_n &= \left\{(\theta, \gamma) \in \mathbb{R}^{d+1} \;\middle|\; g_n(\theta) \leq \gamma\right\}, \\ Z &= \left\{(\theta, \gamma) \in \mathbb{R}^{d+1} \;\middle|\; \gamma < \gamma^*\right\}. \end{aligned}$$

The sets $S_1, \cdots S_n \subset \mathbb{R}^{d+1}$ are the epigraphs of the convex functions $g_1(\cdot), \cdots, g_n(\cdot)$, hence they are convex sets; a point belongs to S_i if and only if it satisfies the i-th constraint. The set $Z \subset \mathbb{R}^{d+1}$ is an open half-plane,

hence of course another convex set. Any point belonging to Z is "sub-optimal" ($\gamma < \gamma^*$); there cannot exist a point that satisfies all the constraints and belongs to Z, otherwise (θ^*, γ^*) would not be the minimizing solution. For the sake of contradiction, assume now that the support constraints of the problem are (without loss of generality) $d+2$. Extract an arbitrary collection of $d+2$ sets from $S_1, \cdots S_n, Z$. If Z happens to be in the collection, then the collection contains exactly $d+1$ (epigraphs of) constraints, of which at most $d+1$ are support constraints. Since the support constraints are assumed to be $d+2$, at least one of them has been "removed", and the solution "falls", namely there exists a point $(\theta^{**}, \gamma^{**})$ which satisfies the $d+1$ constraints (meaning that it belongs to the $d+1$ sets $\{S_i\}$) and attains $\gamma^{**} < \gamma^*$ (meaning that it belongs to Z). On the other hand, if Z is not in the collection, the latter contains just $d+2$ constraints, and since the problem is feasible, there exists at least a point satisfying all of them. Summing up, for any choice of $d+2$ sets in $S_1, \cdots S_n, Z$, their intersection is non-empty. Applying Helly's theorem we obtain that

$$S_1 \cap \cdots \cap S_n \cap Z \neq \varnothing,$$

but this is clearly in contradiction with the hypothesis that (θ^*, γ^*) is the solution to the problem. The contradiction stems from the assumption that the support constraints were more than $d+1$, and is enough to establish the claim. □

6.3 Prediction intervals revisited

Example. An urn contains $N+1$ balls, of which $d+1$ are labeled "support constraint" and the others are white. We extract a ball from the urn. What is the probability that it has the label? Of course, $\frac{d+1}{N+1}$.
Let $y_1, y_2, \cdots, y_N$ a random sample drawn from the density $f(y)$, and consider the following convex problem (the constraints are now random):

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & |\theta - y_1| - \gamma \leq 0 \\ & \vdots \\ & |\theta - y_N| - \gamma \leq 0 \\ & \theta \in \mathbb{R}, \gamma \in \mathbb{R}, \end{aligned}$$

Note that a constraint of the form $g(\theta, \gamma) = |\theta - y| - \gamma \leq 0$ is convex in the variables (θ, γ). Indeed, any set of the form $|\theta - y| - \gamma \leq 0$ is the intersection of two half-spaces: $\theta - y - \gamma \leq 0$, $\theta - y + \gamma \geq 0$. Hence it is the intersection of two convex sets, which is itself convex. The solution to the problem is the pair (θ^*, γ^*) attaining the *minimum* γ such that θ *has distance at most* γ *from all the points* $\{y_i\}$. It is such that $\theta^* - \gamma^* = \min_i\{y_i\}$, $\theta^* + \gamma^* = \max_i\{y_i\}$. Recall that the support constraints of the problem are at most $d+1 = 2$. It may actually be the case that they are *less* than 2; But since the random variables have a density, this happens with probability 0. Hence, *almost surely* the support constraints are exactly 2 (the ones with $\min_i\{y_i\}$ and $\max_i\{y_i\}$). The solution yields the prediction interval that we had obtained before:

$$[a, b] = [\theta^* - \gamma^*,\ \theta^* + \gamma^*] = \left[\min_i\{y_i\},\ \max_i\{y_i\}\right].$$

Does the solution to the above problem remain the same if another constraint, corresponding to a new measure y_{N+1} drawn from the same density, is added?

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & |\theta - y_1| - \gamma \leq 0 \\ & \vdots \\ & |\theta - y_N| - \gamma \leq 0 \\ & |\theta - y_{N+1}| - \gamma \leq 0 \\ & \theta \in \mathbb{R}, \gamma \in \mathbb{R}, \end{aligned}$$

If $y_{N+1} \in [a,b]$, then the last constraint does not change anything, and the solutions to the two problems are identical. If $y_{N+1} \notin [a,b]$, the last constraint $|\theta - y_{N+1}| - \gamma \leq 0$ becomes a support constraint for the new problem, and the solution must change.

What is the probability? Let the event $\mathcal{E} = \{\text{the new constraint } |\theta - y_{N+1}| - \gamma \leq 0 \text{ is a support constraint}\}$. Since the support constraints are almost surely $d+1$, conditioning to the extraction $y_1, y_2, \cdots, y_N, y_{N+1}$,

$$\begin{aligned} P &= \int_{\mathbb{R}^{N+1}} \mathsf{P}\left[\mathcal{E} \mid y_1, \cdots, y_{N+1}\right] f(y_1) \cdots f(y_{N+1})\, dy_1 \, \cdots \, dy_{N+1} \\ &= \int_{\mathbb{R}^{N+1} \setminus \mathcal{N}} \mathsf{P}\left[\mathcal{E} \mid y_1, \cdots, y_{N+1}\right] f(y_1) \cdots f(y_{N+1})\, dy_1 \, \cdots \, dy_{N+1} \\ &= \int_{\mathbb{R}^{N+1} \setminus \mathcal{N}} \frac{d+1}{N+1} \, f(y_1) \cdots f(y_{N+1})\, dy_1 \, \cdots \, dy_{N+1} \\ &= \frac{d+1}{N+1} \int_{\mathbb{R}^{N+1} \setminus \mathcal{N}} f(y_1) \cdots f(y_{N+1})\, dy_1 \, \cdots \, dy_{N+1} = \frac{d+1}{N+1}. \end{aligned}$$

Summing up: we have considered a convex problem with N constraints, and its solution has yielded the prediction interval that we knew from before. A new measure y_{N+1} falls outside this prediction interval exactly when, adding to the problem a new constraint corresponding to y_{N+1}, this becomes a support constraint, in other words it violates the old solution. By virtue of the same reasoning on the ordering of measurements that we have considered in Section 6.1, the probability of violation is $\frac{d+1}{N+1}$. Finally, since in this example $d =$ dimension of $\theta = 1$, the violation probability is $\frac{2}{N+1}$, as we had found before.

6.4 Interval predictor models

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N), (x_{N+1}, y_{N+1})$ be random pairs, independent and identically distributed according to an unknown density, where $x_i, y_i \in \mathbb{R}$. Consider the optimization problem with the constraints corresponding to the first N measures:

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & |\theta_1 + \theta_2 x_1 - y_1| - \gamma \leq 0 \\ & \vdots \\ & |\theta_1 + \theta_2 x_N - y_N| - \gamma \leq 0 \\ & (\theta_1, \theta_2) \in \mathbb{R}^2, \gamma \in \mathbb{R}. \end{aligned}$$

A constraint like $g(\theta_1, \theta_2, \gamma) = |\theta_1 + \theta_2 x - y| - \gamma \leq 0$ is convex in the variables $(\theta_1, \theta_2, \gamma)$, because any set of the form $|\theta_1 + \theta_2 x - y| - \gamma \leq 0$ is the intersection of two half-spaces, as before. The solution to the problem is a certain triple $(\theta_1^*, \theta_2^*, \gamma^*)$. The parameters θ_1^*, θ_2^* yield a linear model $y = \theta_1^* + \theta_2^* \, x$ which is "closest to the data $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$" in the sense that it attains $\min \max_{i=1,\cdots,N} |\theta_1^* + \theta_2^* \, x_i - y_i| = \gamma^*$. It is of course *not* the least squares model (the LS model attains $\min \sum_{i=1}^N (\theta_1^* + \theta_2^* \, x_i - y_i)^2$ instead), although similar for well-behaved data. As the "next" explanatory variable x_{N+1} comes, we can *predict* the value of y_{N+1}:

$$\hat{y}_{N+1} = \theta_1^* + \theta_2^* \, x_{N+1}$$

and append a "certificate" to this prediction: it will hold $|\hat{y}_{N+1} - y_{N+1}| = |\theta_1^* + \theta_2^* \, x_{N+1} - y_{N+1}| > \gamma^*$ with the same probability with which the new constraint $|\theta_1 + \theta_2 x_{N+1} - y_{N+1}| - \gamma \leq 0$ would violate the solution $(\theta_1^*, \theta_2^*, \gamma^*)$ if added to the problem. Since here $d =$ dimension of $\theta = 2$, the maximum number of support constraints in this problem is $d + 1 = 3$. We conclude that, given x_{N+1}, the interval

$$[a,b] = [\theta_1^* + \theta_2^* \, x_{N+1} - \gamma^*, \;\; \theta_1^* + \theta_2^* \, x_{N+1} + \gamma^*]$$

is a prediction interval for y_{N+1} with probability $1 - \frac{3}{N+1}$.

Generalization:

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & \left|\varphi(x_1)^\top \theta - y_1\right| - \gamma \leq 0 \\ & \vdots \\ & \left|\varphi(x_N)^\top \theta - y_N\right| - \gamma \leq 0 \\ & \theta \in \mathbb{R}^d, \gamma \in \mathbb{R}; \end{aligned}$$

The support constraints are $\leq d+1$. If the data (x_i, y_i) are distributed according to a density, then in general the support constraint are *almost surely* $d+1$, hence the violation probability is *exactly* $\frac{d+1}{N+1}$. On the other hand, if the data are not distributed according to a density the support constraints may be *less* than $d+1$, and the violation probability satisfies an inequality in the "good" direction: $\mathsf{P}\left[\left|\varphi(x_{N+1})^\top \theta^* - y_{N+1}\right| > \gamma^*\right] \leq \frac{d+1}{N+1}$. The following:

$$[a, b] = \left[\varphi(x_{N+1})^\top \theta^* - \gamma^*,\ \varphi(x_{N+1})^\top \theta^* + \gamma^*\right]$$

is a prediction interval for y_{N+1} with probability *at least* $1 - \frac{d+1}{N+1}$.

7 Essential concepts to recall from maths

7.1 Linear algebra

Prototypical scalar product in $\mathbb{R}^p$: $\langle x, y\rangle = x^\top y$.
Corresponding norm (Euclidean norm): $\|x\|_2 = \sqrt{\langle x, x\rangle}$.
Two vectors $v, w \in \mathbb{R}^p$ are *orthogonal* if $\langle v, w\rangle = 0$. This is denoted $v \perp w$.
Given a subset $S \subset \mathbb{R}^p$, the *orthogonal complement* of S in $\mathbb{R}^p$ is the set $S^\perp = \{v \in \mathbb{R}^p \mid v \perp w \text{ for all } w \in S\}$.
Let V be a subspace of $\mathbb{R}^p$. Then $\left(V^\perp\right)^\perp = V$.
The *range* of a matrix $A \in \mathbb{R}^{n\times p}$ is the set range $A = \{v \in \mathbb{R}^n \mid \text{there exists } w \in \mathbb{R}^p \text{such that } v = Aw\}$.
Note that range $A =$ span {columns of A}.
The *null space* of a matrix $A \in \mathbb{R}^{n\times p}$ is the set null $A = \{v \in \mathbb{R}^p \mid Av = 0\}$.
Let $A \in \mathbb{R}^{n\times p}$, understood as a linear mapping from $\mathbb{R}^p$ to $\mathbb{R}^n$. Then 1) range $A = \left(\text{null } A^\top\right)^\perp$; 2) null $A = \left(\text{range } A^\top\right)^\perp$; 3) range $A^\top = (\text{null } A)^\perp$; 4) null $A^\top = (\text{range } A)^\perp$.
Corollary: for any matrix $A \in \mathbb{R}^{m\times n}$, range $A =$ range $AA^\top$.
Let $A \in \mathbb{R}^{n\times p}$. The *rank* of A is the maximum dimension of a square matrix, obtained from A by suppressing some rows and/or columns, with nonzero determinant.
The rank of A is equal to: 1) the dimension of the subspace of $\mathbb{R}^n$ generated by its columns: 2) the dimension of the subspace of $\mathbb{R}^p$ generated by its rows:
If $A \in \mathbb{R}^{n\times p}$, where $n \geq p$ ("tall" matrix, i.e. more rows than columns), we say that A has *full rank* if rank $A =$ number of columns. Then the columns of A are linearly independent, and the subspace of $\mathbb{R}^n$ generated by them has dimension p (the maximum possible).
Conversely, if $n \leq p$ ("flat" matrix, i.e. more columns than rows), we say that A has *full rank* if rank $A =$ number of rows. Then the *rows* of A are linearly independent, and their span has dimension n.
In particular, if $A \in \mathbb{R}^{p\times p}$ (square), the following statements are equivalent:

- A has full rank $(= p)$;
- the columns of A are linearly independent, and form a basis of $\mathbb{R}^p$;
- the columns of $A^\top$ (transposes of the rows of A) are linearly independent, and form a basis of $\mathbb{R}^p$;
- A is invertible, i.e. non-singular, its determinant is nonzero, etc.

7.2 Probability

Lemma 7.2.1 *(Markov inequality). Let X be a nonnegative random variable (that is, $X \geq 0$ almost surely). Then for all $\varepsilon > 0$*

$$\mathsf{P}[X \geq \varepsilon] \leq \frac{\mathsf{E}[X]}{\varepsilon}.$$

Proof. Consider the function $\mathbb{1}_{[\varepsilon,\infty)}(x)$ that takes the values 1 when $x \geq \varepsilon$, and 0 otherwise. Then, for all $x \geq 0$, $\mathbb{1}_{[\varepsilon,\infty)}(x) \leq x/\varepsilon$ (verify this!). Consequently,

$$\mathsf{P}[X \geq \varepsilon] = \mathsf{E}\left[\mathbb{1}_{[\varepsilon,\infty)}(X)\right] \leq \mathsf{E}[X/\varepsilon] = \frac{\mathsf{E}[X]}{\varepsilon}.$$

□

Lemma 7.2.2 *(Čebyšev inequality). Let X be any real random variable with mean μ and variance σ^2. Then*

$$\mathsf{P}[|X - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. Let us apply the Markov inequality to the nonnegative random variable $(X - \mu)^2$:

$$\mathsf{P}[|X - \mu| \geq \varepsilon] = \mathsf{P}\left[(X - \mu)^2 \geq \varepsilon^2\right] \leq \frac{\mathsf{E}\left[(X - \mu)^2\right]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}.$$

□

Let $\{X_i\}_{i=1}^{\infty}$ be random variables.

Definition 7.2.1 *X_i is said to converge to X* in probability *if for all $\varepsilon > 0$*

$$\lim_{i\to\infty} \mathsf{P}[|X_i - X| \geq \varepsilon] = 0.$$

Definition 7.2.2 *X_i is said to converge to X* almost surely *if the event*

$$\left\{\lim_{i\to\infty} X_i = X\right\} = \left\{\omega \mid \lim_{i\to\infty} X_i(\omega) = X(\omega)\right\}$$

has probability 1. *In other terms, the set of trajectories $X_1(\omega), X_2(\omega), X_3(\omega), \cdots$ which do* not *converge to X has probability* 0.

Among these, almost sure convergence is the strongest form of convergence, and the most desirable; convergence in probability is an intermediate property, and convergence in distribution is the weakest one. Here, "strong" and "weak" are meant in the sense that the stronger property implies the weaker:

$$\begin{array}{c} X_i \text{ converges almost surely to } X \\ \Downarrow \\ X_i \text{ converges in probability to } X \end{array}$$

Definition 7.2.3 *X_i is said to converge to X* in the mean-square *if*

$$\lim_{i\to\infty} \mathsf{E}\left[|X_i - X|^2\right] = 0.$$

Mean-square convergence is also a kind-of-strong form of convergence, in the sense that

$$\begin{array}{c} X_i \text{ converges in mean-square to } X \\ \Downarrow \\ X_i \text{ converges in probability to } X \end{array}$$

However, no implication exists between almost sure convergence and convergence in the mean-square.

Lemma 7.2.3 *(Weak law of large numbers, Čebyšev). Let $\{X_i\}_{i=1}^{\infty}$ be independent random variables such that for all i*

$$\begin{aligned}\mathsf{E}[X_i] &= \mu;\\ \mathsf{Var}[X_i] &= \mathsf{E}\left[(X_i-\mu)^2\right] = \sigma^2;\end{aligned}$$

then

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to \mu \quad \textit{in probability.}$$

Proof. We have

$$\begin{aligned}\mathsf{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] &= \frac{1}{N}\sum_{i=1}^{N}\mathsf{E}[X_i] = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu;\\ \mathsf{Var}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] &= \frac{1}{N^2}\mathsf{Var}\left[\sum_{i=1}^{N} X_i\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathsf{Var}[X_i] = \frac{1}{N^2}N\sigma^2 = \frac{\sigma^2}{N}.\end{aligned}$$

Therefore for all $\varepsilon > 0$, by Čebyšev's inequality,

$$\mathsf{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \geq \varepsilon\right] \leq \frac{\sigma^2/N}{\varepsilon^2} \to 0$$

□

Theorem 7.2.1 *(Strong law of large numbers, Kolmogorov). Let $\{X_i\}_{i=1}^{\infty}$ be independent and identically distributed random variables with mean $\mathsf{E}[X_i] = \mu$. Then*

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to \mu \quad \textit{almost surely.}$$

Theorem 7.2.2 *(Strong law of large numbers, also by Kolmogorov). Let $\{X_i\}_{i=1}^{\infty}$ be independent random variables with arbitrary distributions but having the same mean $\mathsf{E}[X_i] = \mu$. If*

$$\sum_{i=1}^{\infty}\frac{1}{i^2}\mathsf{E}\left[X_i^2\right] < \infty,$$

then

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to \mu \quad \textit{almost surely.}$$

Definition 7.2.4 *$f(X_1,\cdots,X_N)$ is called an* unbiased *estimator of θ if*

$$\mathsf{E}[f(X_1,\cdots,X_N)] = \theta.$$

Definition 7.2.5 *$f(X_1,\cdots,X_N)$ is called a* consistent *estimator of θ if*

$$\lim_{N\to\infty} f(X_1,\cdots,X_N) = \theta \quad \textit{almost surely.}$$

Example. The sample average $\bar{X} = \mathsf{M}[X] = \frac{1}{N}\sum_{i=1}^{N} X_i$ is a good estimator of the mean $\mathsf{E}[X_i] = \mu$, being

- unbiased: indeed
$$\mathsf{E}\left[\bar{X}\right] = \mathsf{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathsf{E}[X_i] = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu;$$
- consistent: this is precisely the statement of Theorem 7.2.1.

As another example, the sample variance $s^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2$ is a consistent estimator of the variance $\sigma^2 = \mathsf{E}\left[(X_i-\mu)^2\right]$, but not an unbiased one: It is not difficult to show that $\mathsf{E}\left[s^2\right] = \frac{N-1}{N}\sigma^2$. Hence, usually the estimator $\bar{s}^2 = \frac{N}{N-1}s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2$, which enjoys both properties, is preferred. □

7.3 Analysis

Definition 7.3.1 *We say that f_n converges to f* uniformly *if*

$$\lim_{n\to\infty} \sup_{x\in C} |f_n(x) - f(x)| = 0$$

This means that for all $\varepsilon > 0$ there exists $N \geq 0$ such that, for all $n \geq N$, $|f_n(x) - f(x)| \leq \varepsilon$ over all the domain C (in words, the f_n are all "uniformly ε-close" to f after a certain index N).

Definition 7.3.2 *A subset C of a metric space X is called* compact *if from any sequence of points $x_n \subseteq C$ it is possible to extract a sub-sequence that converges to a point belonging to C.*

Theorem 7.3.1 *(Heine/Borel). A subset C of a* finite-dimensional *vector space is compact if and only if it is closed and bounded.*

Definition 7.3.3 *A subset S of a vector space is called* convex *if, whenever the points x, y belong to S, the point $z_\lambda = \lambda x + (1-\lambda)y$ also belongs to S for all $\lambda \in [0,1]$. (Any such z_λ is called a* convex combination *of x and y.)*

Examples: subspaces and their translations, hyperplanes etc.; closed and open balls (i.e. sets of the form $\{x \in \mathbb{R}^p \mid \|x - c\| \leq r\}$ or $\{x \in \mathbb{R}^p \mid \|x - c\| < r\}$).

Lemma 7.3.1 *An arbitrary intersection of convex sets is itself convex.*

Definition 7.3.4 *A function $f : S \to \mathbb{R}$ defined on a convex set S is called* convex *if, for all $x, y \in S$ and all $\lambda \in [0,1]$, it holds*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

Examples: linear and affine functions ($f(x) = c^\top x + d$); norms.

Definition 7.3.5 *The* epigraph *of a function $f : S \to \mathbb{R}$ is the subset of $S \times \mathbb{R}$ defined as follows:*

$$\text{Epi } f = \left\{(x, y) \in S \times \mathbb{R} \;\middle|\; f(x) \leq y\right\}.$$

Lemma 7.3.2 *A function $f : S \to \mathbb{R}$ is convex if and only if* Epi f *is a convex set.*

Definition 7.3.6 *The k-*sublevel set *of a function $f : S \to \mathbb{R}$, where $k \in \mathbb{R}$, is the subset of S defined as follows:*

$$S_k = \left\{x \in S \;\middle|\; f(x) \leq k\right\}.$$

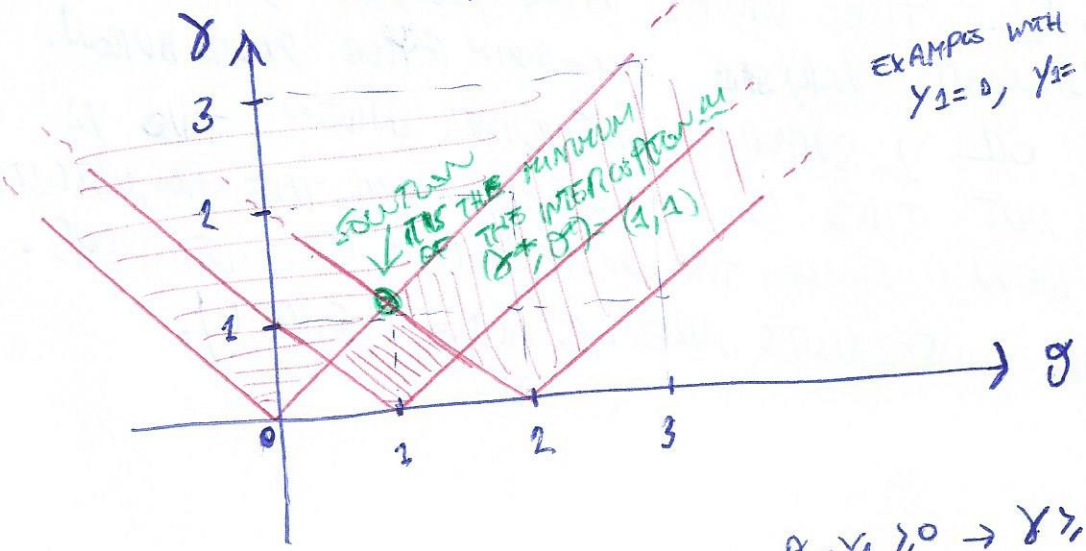
Lemma 7.3.3 *The sublevel sets of a convex function $f : S \to \mathbb{R}$ are convex sets.*

MINIMIZE γ

SUBJECT TO $|\vartheta - y_1| - \gamma \leq 0$

$|\vartheta - y_2| - \gamma \leq 0$

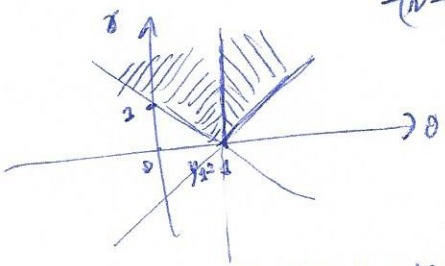
$\vartheta \in \mathbb{R}, \gamma \in \mathbb{R}$



EXAMPLE WITH $y_1 = 0$, $y_2 = 1$, $y_3 = 2$

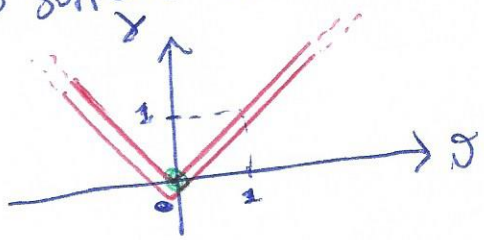
$|\vartheta - y_1| \leq \gamma \rightarrow \vartheta - y_1 \leq \gamma$ IF $\vartheta - y_1 \geq 0 \rightarrow \gamma \geq \vartheta - y_1$ IF $\vartheta \geq y_1$

$-(\vartheta - y_1) \leq \gamma$ IF $\vartheta - y_1 < 0 \rightarrow \gamma \geq -\vartheta + y_1$ IF $\vartheta < y_1$

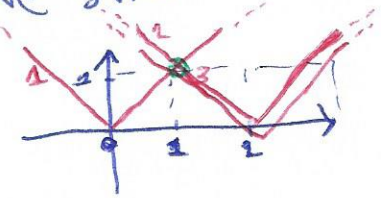


A) NO SUPPORT CONSTRAINTS: YES FOR EXAMPLE TWO $y = 0$

IF I REMOVE ONE CONSTRAINT THE SOLUTION DOESN'T CHANGE



B) ONE SUPPORT CONSTRAINT: YES

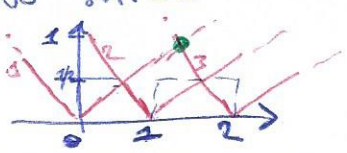


IF I REMOVE 1 THE SOLUTION GETS BETTER ($\gamma^* = 0$ INSTEAD OF $\gamma^* = 1$) SO IT'S A SUPPORT CONSTRAINT

IF I REMOVE 2 THE SOLUTION IS THE SAME SO IT'S NOT A SUPPORT CONSTRAINT

IF I REMOVE 3 THE SOLUTION IS THE SAME SO IT'S NOT A SUPPORT CONSTRAINT

C) TWO SUPPORT CONSTRAINTS: YES



IF I REMOVE 1 THE SOLUTION GOES BETTER, IT'S A SUPPORT CONSTRAINT

IF I REMOVE 2 THE SOLUTION DOESN'T CHANGE

IF I REMOVE 3 THE SOLUTION GETS BETTER, IT'S A SUPPORT CONSTRAINT

D) 3 SUPPORT CONSTRAINTS IS IMPOSSIBLE DUE TO THE CAMPI, CARATHEODORY THEOREM. THE SUPPORT CONSTRAINTS ARE AT MOST $d+1$ AND HERE $d=1$ SO 2.

IF $Y_1, \dots, Y_N$ ARE RANDOM I.I.D. CONTINUOUS WITH UNIFORM DENSITY IN $[0,1]$. THE VALUES ARE RANDOM SO THE POSITIONS OF THE CONSTRAINTS VARY ACCORDING TO THE DISTRIBUTION. WE HAVE 1 OR 2 SUPPORT CONSTRAINTS WHEN TWO Y_i ARE EQUAL BUT THIS IS IMPOSSIBLE FROM THE PROBABILISTIC POINT OF VIEW SINCE THE DISTRIBUTION IS CONTINUOUS. SO THE SUPPORT CONSTRAINTS ARE 2 ALMOST SURELY.

TRUE FUNCTION $f(\cdot, \theta^0)$

APPROXIMATE FUNCTION $f(\cdot, \hat{\theta})$

GENERIC FUNCTION $f(\cdot, \theta)$
("AGEBRAIN FUNCTION")

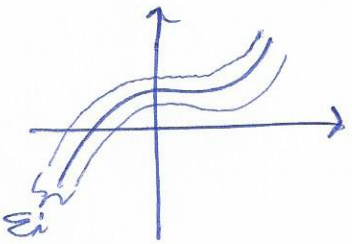
$$y_i = f(x_i, \theta^0) + \varepsilon_i$$

$$y_i = f(x_i, \hat{\theta}) + \varepsilon_i(\hat{\theta})$$

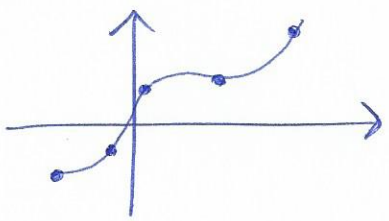
$$y_i = f(x_i, \theta) + \varepsilon_i(\theta)$$

BE CAREFUL:

ε_i ARE DISTURBANCES DUE TO THE RANDOMNESS



$\varepsilon_i(\theta)$ ARE ERRORS WE COMMIT SELECTING A FUNCTION EXPLOITING THE ONLY DATA AT HAND



$$\hat{\theta}_{LS} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - \varphi_i^T \theta)^2$$

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{N} (y_i - \varphi_i^T \theta)^2 = 2 \sum_{i=1}^{N} (y_i - \varphi_i^T \theta)(-\varphi_i^T) = 0$$

$$-\sum_{i=1}^{N} y_i \varphi_i^T + \sum_{i=1}^{N} \varphi_i^T \theta \varphi_i^T = 0$$

TRANSPOSE THIS.

$$\sum_{i=1}^{N} \theta^T \varphi_i \varphi_i^T = \sum_{i=1}^{N} y_i \varphi_i^T$$

$$\theta^T \left(\sum_{i=1}^{N} \varphi_i \varphi_i^T \right) = \sum_{i=1}^{N} y_i \varphi_i^T$$

TRANSPOSING EVERYTHING

$$\left(\sum_{i=1}^{N} \varphi_i \varphi_i^T \right) \theta = \sum_{i=1}^{N} \varphi_i y_i$$

NORMAL EQUATION

y_i □

θ ▯

φ_i = ▯

$R = \sum_{i=1}^{N} \varphi_i \varphi_i^T \in \mathbb{R}^{p \times p}$

ANY θ THAT SOLVES THE NORMAL EQUATION IS A SOLUTION FOR THE LEAST SQUARES PROBLEM.

IF $R = \sum_{i=1}^{N} \varphi_i \varphi_i^T$ IS INVERTIBLE THEN THE ONLY SOLUTION TO THE L.S. PROBLEM IS:

$$\theta_{LS} = \left(\sum_{i=1}^{N} \varphi_i \varphi_i^T \right)^{-1} \sum_{i=1}^{N} \varphi_i y_i$$

DEF: Let $v_1, \dots, v_m \in V$

- $W = \operatorname{span}\{x_1, \dots, x_m\} = \{v \in V \mid v = a_1 x_1 + \dots + a_m x_m$ FOR SOME COEFFICIENTS $a_1, \dots a_m\}$

1) W IS A SUBSPACE OF V
2) BY DEFINITION $x_1, \dots, x_m$ GENERATE SPAN $\{x_1, \dots x_m\}$

$L: V \to W$ IS A LINEAR MAPPING IF $\forall x, y \in V$

$$L(\alpha x + \beta y) = \alpha L(x) + \beta L(y)$$

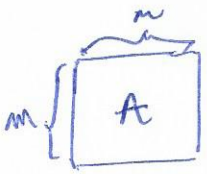
LINEAR MAPPINGS FROM $\mathbb{R}^n$ TO $\mathbb{R}^m$ ARE REPRESENTED BY MATRICES IN $\mathbb{R}^{m \times n}$

- LET $A \in \mathbb{R}^{m+m}$ A LINEAR MAP FROM $\mathbb{R}^n$ (DOMAIN) TO $\mathbb{R}^m$ (CODOMAIN)

RANGE $A = \{v \in \mathbb{R}^m \mid v = Aw$ for some $w \in \mathbb{R}^n\}$ ← SUBSPACE OF CO-DOMAIN
(IMAGE OF A)

NULL $A = \{v \in \mathbb{R}^n \mid Av = 0\}$ ← SUBSPACE OF DOMAIN
(KERNEL OF A)

- $A \in \mathbb{R}^{m \times n}$



$A: \mathbb{R}^n \to \mathbb{R}^m$

HAVE TWO SUBSPACES:

RANGE A: $\{v \in \mathbb{R}^m \mid \exists w \in \mathbb{R}^n, Aw = v\} \subseteq \mathbb{R}^m$ (CO-DOMAIN OF A)

NULL A: $\{w \in \mathbb{R}^n \mid Aw = 0\} \subseteq \mathbb{R}^n$ (DOMAIN OF A)

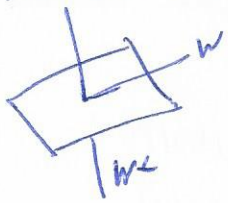
LET V BE A REAL VECTOR SPACE AND $S \subset V$ THEN THE ORTHOGONAL COMPLEMENT OF S IS DEFINED AS FOLLOWS:

- $S^\perp := \{v \in V \mid v \perp s \ \forall s \in S\}$

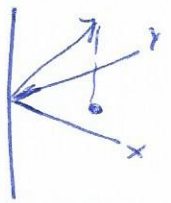
FACT: LET W BE A SUBSPACE OF V $\forall \vec{v} \in V$ CAN BE WRITTEN UNIQUELY AS: $\vec{v} = w + w^\perp$

WHERE $w \in W$ AND $w^\perp \in W^\perp$

FACT: IF V IS FINITE DIMENSIONAL ($\mathbb{R}^n$) AND W IS A SUBSPACE OF V THEN $(W^\perp)^\perp = W$



EXAMPLE: A PROJECTION ON THE XY PLANE



$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\text{RANGE } A = \text{SPAN}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}\right\}$$

$$\dim \text{RANGE } A = 2$$

$$\text{NULL } A = \text{SPAN}\left\{\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right\} = \left\{\begin{bmatrix} 0 \\ 0 \\ \alpha \end{bmatrix}\right\}$$

R SYMMETRIC $\Rightarrow$ RANGE $R = (\text{NULL } R^T)^\perp = (\text{NULL } R)^\perp$

DIM 1) $v \in \text{NULL } R \rightarrow Rv = 0$

$$v^T R v = v^T \sum_{\lambda=1}^{N} \varphi_\lambda \varphi_\lambda^T v = 0$$

$$\sum_{\lambda=1}^{N} (v^T \varphi_\lambda)(\varphi_\lambda^T v) = 0 \rightarrow \sum_{\lambda=1}^{N} (\varphi_\lambda^T v)^2 = 0$$

$\varphi_i^T v = 0 \quad \forall i = 1, \dots, N \quad v \perp \varphi_i \ \forall i$

$v \perp \alpha_1 \varphi_1 + \alpha_2 \varphi_2 + \dots + \alpha_N \varphi_N \quad \forall \alpha_1, \dots, \alpha_N \in \mathbb{R}$

$v \perp \text{SPAN}\{\varphi_1, \dots \varphi_N\}$

$v \in \text{SPAN}\{\varphi_1, \dots \varphi_N\}^\perp$

2) $v \in \text{SPAN}\{\varphi_1, \dots \varphi_N\}^\perp$

$\Downarrow$

$v \perp \alpha_1 \varphi_1 + \dots + \alpha_N \varphi_N \quad \forall \alpha_1, \dots, \alpha_N \in \mathbb{R}$

$\Downarrow$

$v \perp \varphi_1, v \perp \varphi_2, \dots, v \perp \varphi_i \quad \forall i = 1, \dots, N$

$\Downarrow$

$\varphi_i v = 0 \ \forall i \Rightarrow \left(\sum_{\lambda=1}^{N} \varphi_\lambda \varphi_\lambda^T\right) v = 0 \rightarrow Rv = 0 \Rightarrow v \in \text{NULL } R$

NORMAL EQUATIONS

$$R\vartheta = \sum_{i=1}^{N} \varphi_i y_i$$

RANGE $R = \text{SPAN}\{\varphi_i\}$

$\forall \tau \in \text{SPAN}\{\varphi_i\} \rightarrow \tau \in \text{RANGE } R \rightarrow \exists \vartheta \in \mathbb{R}^p$ SUCH THAT $\tau = R\vartheta$

COROLLARY: THE NORMAL EQUATIONS HAVE AT LEAST ONE SOLUTION

SYSTEM OF LINEAR EQUATIONS $Ax = b$

$$R\vartheta = \sum_{i=1}^{N} \varphi_i y_i$$

EXACTLY ONE SOLUTION $\Leftrightarrow$ R IS INVERTIBLE (AND THE SOLUTION IS OF COURSE $\vartheta_{LS} = R^{-1}r$)

INFINITELY MANY SOLUTIONS $\Leftrightarrow$ R IS NOT INVERTIBLE

R INVERTIBLE
$\Updownarrow$
R HAS FULL RANK
$\Updownarrow$
RANGE $R = \mathbb{R}^p$
$\Updownarrow$
$\boxed{\text{SPAN}\{\varphi_1 \dots \varphi_N\} = \mathbb{R}^p}$

$\varphi_1, \varphi_2, \dots \varphi_i, \dots, \varphi_{N-1}, \varphi_N \in \mathbb{R}^p$

p LINEARLY INDEPENDENT VECTORS IN $\mathbb{R}^p$ FORM A BASIS OF $\mathbb{R}^p$

p VECTORS IN $\mathbb{R}^p$ THAT GENERATE $\mathbb{R}^p$ ARE LINEARLY INDEPENDENT

$\varphi_1, \dots \varphi_N$ DO NOT GENERATE $\mathbb{R}^p$: WHY?

1) $N < p$ (PATHOLOGICAL)
2) THE REGRESSORS ARE FLAWED ($\varphi_1 = x_1$, $\varphi_2 = x_2 + x_3$, $\varphi_3 = x_1 + x_2 + x_3$)
3) x_i DO NOT CONTAIN ENOUGH INFORMATION

MEASUREMENT MODEL

$$y_1 = \varphi_1^T\vartheta + \varepsilon_1$$
$$y_2 = \varphi_2^T\vartheta + \varepsilon_2$$
$$\vdots$$
$$y_i = \varphi_i^T\vartheta + \varepsilon_i$$
$$\vdots$$
$$y_N = \varphi_N^T\vartheta + \varepsilon_N$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \Phi = \begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_N^T \end{bmatrix} \quad \vartheta \in \mathbb{R}^p \quad \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \in \mathbb{R}^N$$

$$Y = \Phi\vartheta + E$$

$$\min_{\vartheta} \frac{1}{N}\sum_{i=1}^{N}(y_i-\vartheta)^2 \qquad \hat{\vartheta}_{LS} = \frac{1}{N}\sum_{i=1}^{N} y_i = M[y]$$

ATTAINED MINIMUM: $\frac{1}{N}\sum_{i=1}^{N}(y_i - M[y])^2$ = SAMPLE VARIANCE OF y
S.VAR. $\{y_1 \dots y_N\}$

REGRESSORS OFTEN ARE: $\varphi_0 = 1,\ \varphi_2(x),\ \varphi_3(x),\ \dots,\ \varphi_p(x)$

$$\varphi_i = \begin{bmatrix} 1 \\ \varphi_2(x_i) \\ \vdots \\ \varphi_p(x_i) \end{bmatrix}$$

PROBLEM: (*) $\min_{\vartheta \in \mathbb{R}^p} \frac{1}{N}\sum_{i=1}^{N}(y_i - \varphi_i^T\vartheta)^2$

IF I CHOOSE $\vartheta_m = \begin{bmatrix} M[y] \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ THEN (*) = S.VAR[y]

THEREFORE LET: R.V. = RESIDUAL VARIANCE

$$:= \frac{1}{N}\sum_{i=1}^{N}(y_i - \varphi_i^T\hat{\vartheta}_{LS})^2$$

AND IT FOLLOWS

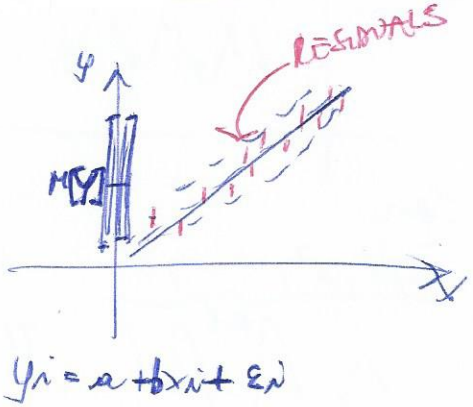
$$RV \le S.VAR[y]$$

$$EV = \text{EXPLAINED VARIANCE} := S.VAR[y] - RV$$

$$0 \le RV \le S.VAR[y]$$

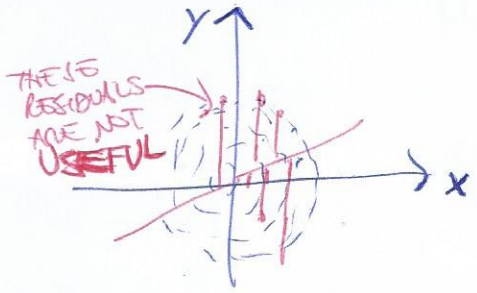
$$0 \le EV \le S.VAR[y]$$

$$S.VAR = EV + RV$$

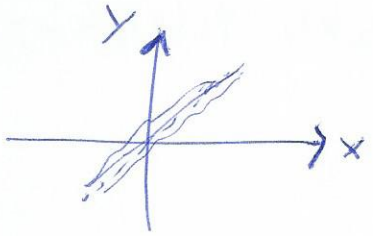


$$y_i = a + bx_i + \varepsilon_i$$

COULD I ALSO OBTAIN THIS PICTURE



TYPICAL CLOUD OF TWO INDEPENDENT VARIABLES



TYPICAL CLOUD OF TWO POSITIVELY CORRELATED VARIABLES

$$\rho^2 := \frac{EV}{S.Var[y]} \qquad 0 \le \rho^2 \le 1$$

} IT'S A WAY TO QUANTIFY HOW GOOD THE MODEL IS

IN STATISTICAL LITERATURE, ρ = CORRELATION COEFFICIENT BETWEEN $\{x_i\}$ AND $\{y_i\}$

$$\rho = \frac{Cov\{x_i, y_i\}}{\sqrt{Var\{x_i\}} \cdot \sqrt{Var\{y_i\}}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - M[x])(y_i - M[y])}{\sqrt{\frac{1}{N}\sum(x_i - M[x])^2}\ \sqrt{\ \ }} =$$

$E\{(x - E[x])(Y - E[y])\}$

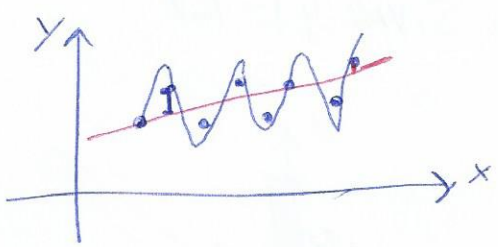
$\sqrt{E[(X - E[x])^2]}$

=

(*) IF IT'S NEXT TO 1 DATA DESCRIBES THE MODEL IN A GOOD WAY ($\rho^2 \simeq 1$ GOOD!)

BUT THERE IS AN EXCEPTION: $N \simeq p$

Ex:



$N = 8$ } NUMBER OF READINGS

$p = 2$ —

$p = 8$ ∿∿ } NUMBER OF REGRESSORS

$\underbrace{1, X, X^2, \dots, X^7}_{p=8}$

IF I USE LS IFIND THE POLYNOMIAL THAT IS COMPLETELY USELESS

BECAUSE I TRY TO FOLLOW THE NOISE TOO

IF I HAD $N = 100$ AND I USE A 90 ORDER POLYNOMIAL THE SOLUTION IS COMPLETELY USELESS

$$\hat{\vartheta}_{LS} = \theta^\circ + \left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^T\right)^{-1}\cdot\frac{1}{N}\cdot\sum_{i=1}^{N}\varphi_i\varepsilon_i$$

TENDS TO A MATRIX Σ (SIGMA) — TENDS TO $\emptyset$

WE EXPECT (HOPE) THAT $\hat{\vartheta}_{LS} \to \vartheta_\circ$ AS $N \to \infty$

(WE ARE DEALING WITH RANDOM VARIABLES!!)

DELICATO WHEN WE USE PROCESSES

PROBABILITY SPACE

$(\Omega, \tilde{F}, P)$ $\quad P: \tilde{F} \to [0,1]$ SUCH THAT

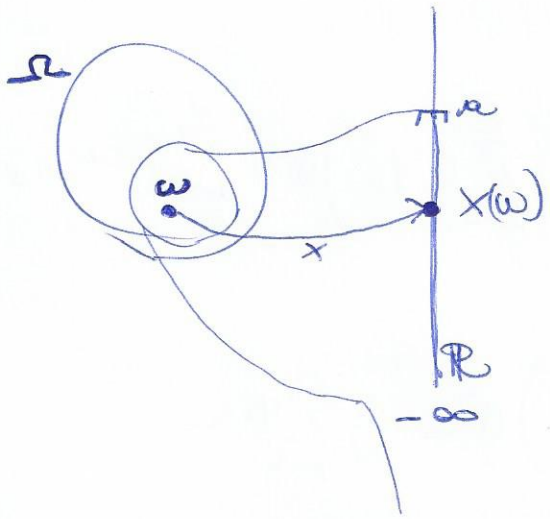
1) $P\left[\bigcup_{i=1}^{+\infty} A_i\right] = \sum_{i=1}^{N} P[A_i]$ (DISJOINT) — IT'S A GENERALIZATION OF AN AREA

2) $P[\Omega] = 1$

Ω: SET CALLED SAMPLES SPACE

$\tilde{F}$: FAMILY OF SUBSETS → = EVENTS OF Ω SUCH THAT

1) $\Omega \in \tilde{F}$
2) $\{A_i\}_{i=1}^{\infty} \in \tilde{F} \Rightarrow \bigcup_{i=1}^{\infty} \in \tilde{F}$ } CLOSE TO RESPECT OF UNION
3) $A \in \tilde{f} \Rightarrow A^c \in f$ } CLOSE TO RESPECT OF COMPLEMENT



IF THE FUNCTION IS THAT

$X^{-1}((-\infty, a]) = \{\omega \in \Omega \mid$

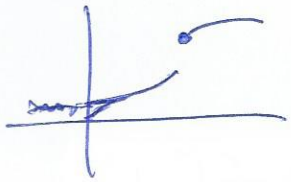
$\ldots X(\omega) \in (-\infty, a]\} \in \tilde{F}\ \forall a$

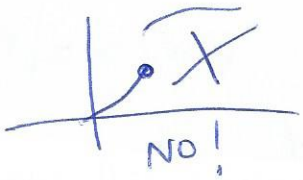
THEN X IS CALLED A RANDOM VARIABLE

$$P\left[\underbrace{X^{-1}((-\infty, a])}_{\in \tilde{F}}\right] = P[\{\omega \mid X(\omega) \le a\}] = P[X \le a] := F(a)$$

$F_X: \mathbb{R} \to [0,1]$ DISTRIBUTION of X

PROPOSITION : F_X IS CONTINUOUS FROM THE RIGHT


ok



- $\lim_{a \to -\infty} F(a) = 0$
- $\lim_{a \to +\infty} F(a) = 1$

IF IN PARTICULAR F_X IS DIFFERENTIABLE (ALL $\mathbb{R}$), THEN

$f_X(a) = \frac{\partial}{\partial a} F_X(a)$ IS CALLED THE DENSITY OF X AND X IS CALLED CONTINUOUS VARIABLE

$$F_X(a) = \int_{-\infty}^{a} f_X(\xi)\, d\xi$$

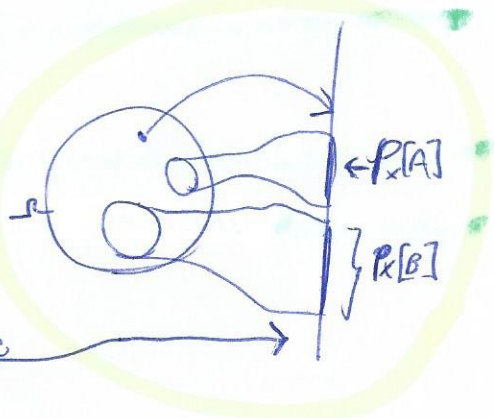
$$X^{-1}(A) = \{\omega \in \Omega \mid X(\omega) \in A\} \qquad X : \Omega \to \mathbb{R}$$

$$\uparrow$$
$$\mathbb{R}$$

$$P\left[X^{-1}(A)\right] = P\left[X \in A\right] := P_X[A]$$

$(\Omega, \mathcal{F}, P)$ $\xrightarrow{X}$ $(\mathbb{R}, \mathcal{B}, P_X)$

I USE THIS MAPPING



FOR A CONTINUOUS VARIABLE

$$E[X] := \int_{-\infty}^{+\infty} x \cdot f_X(x)\, dx = \text{MEAN VALUE} \qquad VAR[X] = E\left[(X - E[X])^2\right]$$

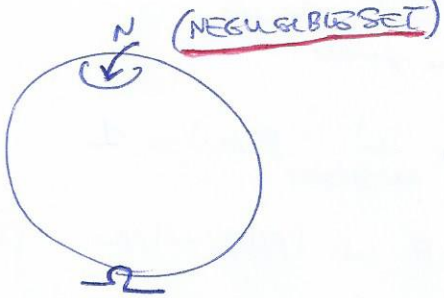
PROPERTIES

1) $E[aX + bY] = a \cdot E[X] + b \cdot E[Y]$ $\quad a, b \in \mathbb{R}$, X, Y = R.V. ON THE SAME Ω

2) $|E[X]| \leq E[|X|]$

3) IF $X \leq Y$ (ALMOST SURELY) THEN $E[X] \leq E[Y]$

Def: IF A PROPERTY HOLDS FOR ALL $\omega \in \Omega$ EXCEPT FOR $\omega \in N$ AND $P[N] = 0$, THEN WE SAY THAT THE PROPERTY HOLDS ALMOST SURELY



4) IF $X = Y$ ALMOST SURELY THEN THE $E[X] = E[Y]$

$X \sim N(0,1)$

$$Y = \begin{cases} X \text{ if } X \neq 0 \\ 12 \text{ if } X = 0 \end{cases}$$

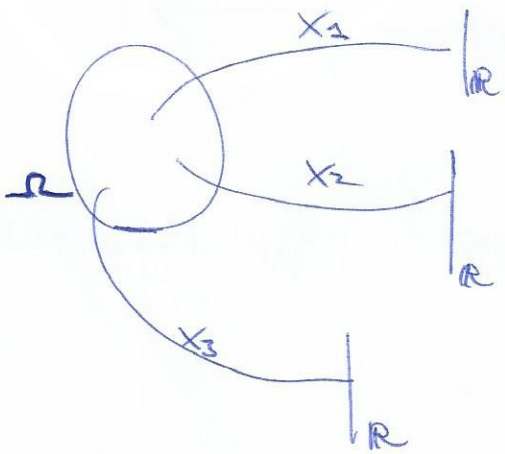
$X = Y$ A.S. $\Rightarrow E[X] = E[Y]$

LET X, Y BE R.V. [on the same (Ω, F, P)]

X, Y ARE INDEPENDENT IF $P[X \in A, Y \in B] = P[X \in A] \cdot P[X \in B]$ (JOINT DISTRIBUTION)

IF X, Y ARE INDEPENDENT, THEN $E[XY] = E[X]E[Y]$

A SEQUENCE $X_1, X_2, X_3, \ldots$ OF R.V. ON THE SAME (Ω, F, P) IS CALLED A RANDOM PROCESS:

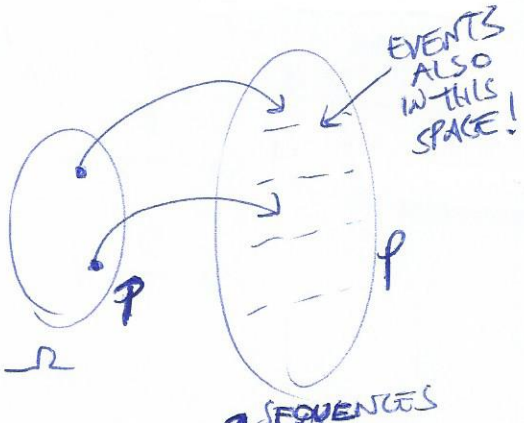


FOR FIXED i, X_i IS A RANDOM VARIABLE

FOR FIXED $\omega \in \Omega$, $X_1(\omega), X_2(\omega), \dots, X_i(\omega), \dots$ IS A SEQUENCE OF REAL NUMBERS. $\omega \mapsto$ SEQUENCE

$(\Omega, \mathcal{F}, P) \longrightarrow (\{\text{Sequences}\}, A, P_S)$

STOCH. PROCESS



$\{X_1 \le 5\}$ IS AN EVENT ON $\mathbb{R}$ AXIS

$\{\lim_{i\to\infty} X_i = 0\}$ IS AN EVENT ON THIS SPACE

IF I USE THE LIMIT I'M LOOKING AT THE PROCESS

– CONVERGENCES –

Def: Let $X_1, X_2, \dots$ BE A STOCHASTIC PROCESS ON $(\Omega, \mathcal{F}, P)$ AND X A R.V. ON $(\Omega, \mathcal{F}, P)$

IF $P\left[\{\omega \in \Omega \mid \lim_{i\to\infty} X_i(\omega) = X(\omega)\}\right] = 1$

THEN $X_i \to X$ ALMOST SURELY AS $i \to +\infty$

Def: If $\forall \varepsilon > 0 \quad \lim_{i\to\infty} P\left[|X_i - X| > \varepsilon\right] = 0$

THEN $X_i \to X$ IN PROBABILITY

Def: IF $\lim_{i\to\infty} E\left[|X_i - X|^2\right] = 0$ THEN $X_i \to X$

IN THE MEAN-SQUARES

Def: IF $\lim_{i\to\infty} F_{X_i}(a) = F_X(a)$ AT ALL THE POINTS AT WHICH F IS CONTINUOUS, THEN $X_i \to X$ IN DISTRIBUTION

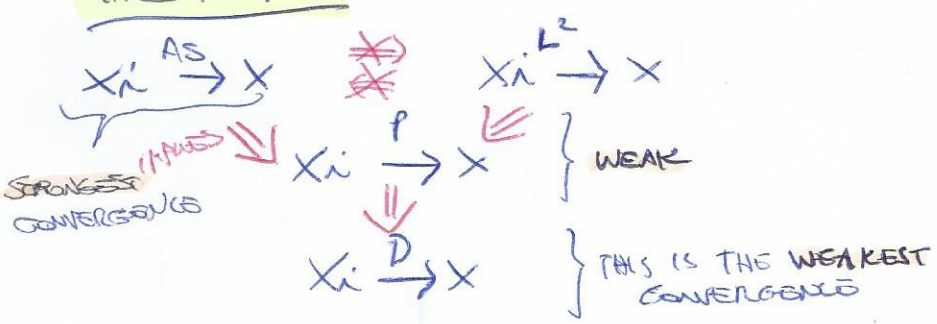
$X_i \xrightarrow{AS} X$ (AS = ALMOST SURELY)

$X_i \xrightarrow{P} X$ (IN PROBABILITY = P)

$X_i \xrightarrow{L^2} X$ (IN MEAN-SQUARES = L^2)

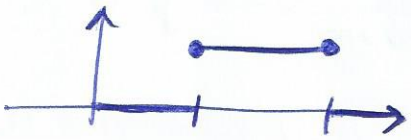
$X_i \xrightarrow{D} X$ (IN DISTRIBUTION = D)

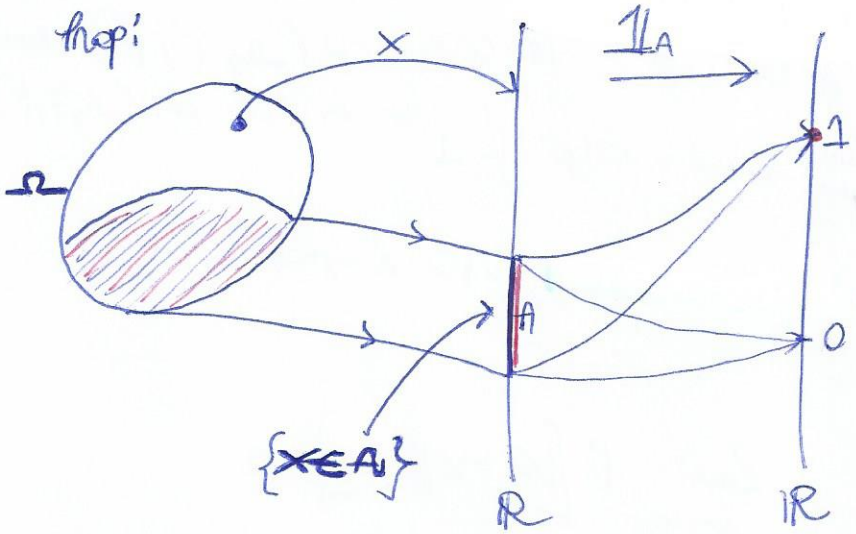
THEOREMS



INDICATOR FUNCTION of $A \subset \mathbb{R}$

$$\mathbb{1}(x): \begin{cases} 1, & x \in A \\ 0, & \text{OTHERWISE} \end{cases}$$





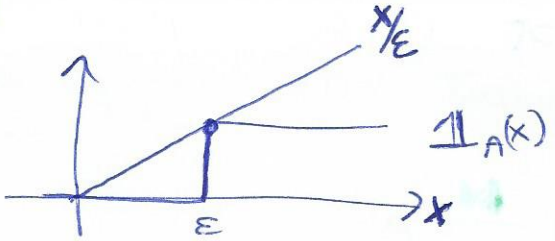
$$P[\mathbb{1}_A(x) = 1] = P[X \in A]$$

$$P[\quad = 0] = P[X \notin A]$$

$$E[\mathbb{1}_A(x)] = 1 \cdot P[X \in A] + 0 \cdot P[\notin] = P[X \in A]$$

MARKOV'S INEQUALITY

$A = \{X \geq \varepsilon\}$



$\forall X \in \mathbb{R}$

For $X \geq 0$, $\mathbb{1}_A(X) \leq \frac{X}{\varepsilon}$

Let X be a positive random variable ≥ 0.

Then $P[X \in A] = P[X \geq \varepsilon]$

$$E[\mathbb{1}_A(X)] \leq E\left[\frac{X}{\varepsilon}\right] = \frac{E[X]}{\varepsilon}$$

$$\boxed{\forall \varepsilon > 0,\ X \text{ R.V.} \geq 0 \qquad P[X \geq \varepsilon] \leq \frac{E[X]}{\varepsilon}}$$

ČEBYŠEV'S INEQUALITY

Let X be a R.V. with values: mean μ and variance σ^2

Then $\boxed{P[|X-\mu| \geq \varepsilon] \leq \frac{\sigma^2}{\varepsilon^2}}$

Proof: $(X-\mu)^2 \geq 0$ ← APPLY MARKOV'S INEQUALITY TO THIS

$$P[(X-\mu)^2 \geq \varepsilon^2] = P[|X-\mu| \geq \varepsilon] \leq E\left[\frac{(X-\mu)^2}{\varepsilon^2}\right] = \frac{\sigma^2}{\varepsilon^2} \quad \square$$

A theorem is called a LAW OF LARGE NUMBERS if it has the form:

HYPOTHESIS

$$\Downarrow$$

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to E[X_i] \quad (\mu)$$

SOME KIND OF CONVERGENCE

(THE LAW CAN BE WEAK / STRONG ACCORDING TO THE KIND OF CONVERGENCE)

1) WEAK LAW OF LARGE NUMBERS

Let x_i independent with mean μ and variance σ^2

Then $\frac{1}{N}\sum_{i=1}^{N} x_i \xrightarrow{P} \mu$

Proof:

$$P\left[\left|\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\right| \geq \varepsilon\right] = \quad \text{Fix AN } \varepsilon > 0$$

BUT $E\left[\frac{1}{N}\sum x_i\right] = \frac{1}{N}\sum E[x_i] = \mu$ so:

$$= P\left[\underbrace{\left|\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)\right|}_{\text{THIS QUANTITY HAS AVERAGE } = 0} \geq \varepsilon\right] \leq \frac{E\left[\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)\right)^2\right]}{\varepsilon^2} = \frac{1}{N^2}\frac{E\left[\left(\sum_{i=1}^{N}(x_i-\mu)\right)^2\right]}{\varepsilon^2} =$$

$$= \frac{1}{N^2}\frac{\sum_{i=1}^{N} E\left[(x_i-\mu)^2\right]}{\varepsilon^2} = \frac{N\sigma^2}{N^2\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}\cdot\frac{1}{N} \xrightarrow{N\to\infty} 0$$

(IT'S THE DEFINITION OF CONVERGENCE IN PROBABILITY!!!)

2) STRONG L.L.N. [KOLMOGOROV]

Let $\{x_i\}$ be i.i.d. random variables with mean

$E[x] = \mu$

Then $\frac{1}{N}\sum_{i=1}^{N} x_i \xrightarrow{A.S.} \mu$

3) STRONG L.L.N. [KOLMOGOROV]

Let $\{x_i\}$ be independent R.V. with arbitrary distributions but with the same $E[x_i] = \mu$.

If $\sum_{i=1}^{+\infty} \frac{E[x_i^2]}{i^2} < +\infty$, THEN $\frac{1}{N}\sum_{i=1}^{N} x_i \xrightarrow{A.S.} \mu$

ADVANTAGE OF 3) IS ARBITRARY DISTRIBUTION

Ex: $f_x(x) = \frac{c}{1+x^2}$ DOESN'T HAVE A MEAN VALUE !!!

$$\int_{-\infty}^{+\infty} \frac{c}{1+x^2}\,dx \sim \left[\arctg\right]_{-\infty}^{+\infty}$$

[redacted]

STATISTICAL JARGON

ϑ IS A PARAMETER TO BE ESTIMATED

$\hat{\vartheta}(x_1, \dots, x_N)$ IS A FUNCTION OF THE RANDOM VARIABLES $X_1, \dots, X_N$

IF I USE $\hat{\vartheta}(\dots)$ TO "ESTIMATE" ϑ, I CALL IT AN ESTIMATOR OF $\hat{\vartheta}$.

UNBIASED ESTIMATOR (GOOD!)

$$E\left[\hat{\vartheta}(X_1, \dots, X_N)\right] = \vartheta$$

CONSISTENT ESTIMATOR (GOOD!)

IF AS $N \to \infty$

$$\hat{\vartheta}(X_1, \dots, X_N) \xrightarrow{\text{A.S.}} \vartheta$$

$$y_i = \varphi_i^T \vartheta^0 + \varepsilon_i$$

(y_i: random; $\varphi_i^T \vartheta^0$; ε_i)

$$\hat{\vartheta}_{LS} = \vartheta^0 + \left(\sum_{i=1}^{N} \varphi_i \varphi_i^T\right)^{-1} \sum_{i=1}^{N} \varphi_i \varepsilon_i$$

MUST BE CLAIMED !!! ← WE NEED THIS MATRIX INVERTIBLE

SINGULAR VALUE DECOMPOSITION (SVD)

For matrices $A \in \mathbb{R}^{n \times p}$, $\boxed{A \geq 0}$ (SYMMETRIC) WE HAVE A BEAUTIFUL DECOMPOSITION:

$$A = M \Lambda M^T$$

(Λ: DIAGONAL; M: ORTHOGONAL: $MM^T = M^TM = I$ (IDENTITY MATRIX))

$$AM = M\Lambda \qquad \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \end{bmatrix} \qquad M = \begin{bmatrix} | & | & \\ v_1 & v_2 & \cdots \\ | & | & \end{bmatrix}$$

$$Av_1 = \lambda_1 v_1$$
$$Av_2 = \lambda_2 v_2$$
$$\vdots$$

(λ: eigenvalues)

(DECOMPOSIZIONE AI VALORI SINGOLARI = FATTORIZZAZIONE BASATA SULL'USO DI AUTOVETT. E AUTOVALORI, UTILIZZATA PER RISOLVERE UN'APPROSSIMAZIONE DELLA MATRICE ORIGINARIA CON RANGO MINORE)

APPLICAZIONI:
- FORNISCE INFO IMPORTANTI (RANGO)
- (KERNEL / IMMAGINE)
- USATA PER DEFINIRE LA PSEUDOINVERSA
- APPROSSIMA LA MATRICE CON UNA DI RANGO INFERIORE

BE CAREFUL: I HAVE $A \geq 0$ SO $M^T = M^{-1}$ OR $A = A^T \geq 0$

IF IT'S SYMMETRIC IT'S DIAGONALIZABLE

There are matrices in $\mathbb{R}^{n \times p}$

- NORMAL BUT NOT SYMMETRIC

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

so can be decomposed as $M \Lambda M$ (LAMBDA λ; $\sim$ TR. CONJ; COMPLEX; COMPLEX)

$s^2 + 1 = 0 \to \pm i$

- NOT NORMAL BUT DIAGONALIZABLE ANYWAY

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1} =$

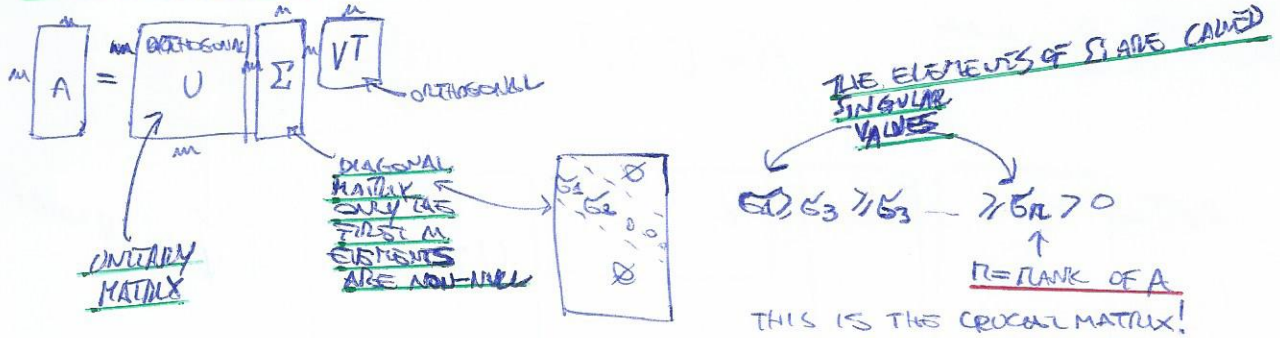
- NOT SQUARE!

- SQUARE BUT NOT DIAGONALIZABLE!

Jordan decomposition diagonal by blocks

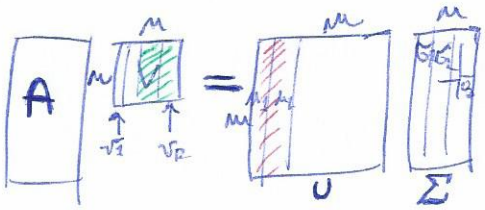
Let $A \in \mathbb{R}^{m \times n}$. I call an SVD of A a decomposition like

$$A = U \Sigma V^T$$



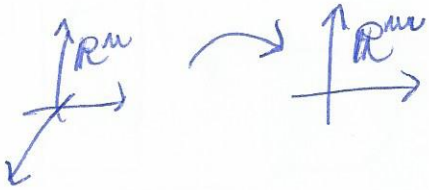
THEOREM

Every matrix in $\mathbb{R}^{m \times n}$ admits an SVD.
Furthermore, Σ is uniquely determined by A.



$$\left.\begin{aligned} A v_1 &= \sigma_1 u_1 \\ A v_2 &= \sigma_2 u_2 \\ &\vdots \\ A v_r &= \sigma_r u_r \end{aligned}\right\} \in \mathbb{R}^m$$

$$\left.\begin{aligned} A v_{r+1} &= 0 \\ A v_{r+2} &= 0 \\ &\vdots \\ A v_n &= 0 \end{aligned}\right\}$$



(*) $\{u_1, u_2, \dots, u_r\} \subseteq \mathbb{R}^m$ — IT'S A BASIS OF RANGE A

$\{v_{r+1}, \dots, v_n\} \in \mathbb{R}^n$ — IT'S A ORTHONORMAL BASIS OF NULL A

A maps the subspace to zero

FACT: for $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times m}$

AB AND BA HAVE THE SAME NONZERO EIGENVALUES AND WITH THE SAME MULTIPLICITY

$\Downarrow$

AA^T AND A^TA HAVE THE SAME NON-ZERO EIGENVALUES

$AA^T \in \mathbb{R}^{m \times m}$
$A^TA \in \mathbb{R}^{n \times n}$
} BOTH SYMMETRIC, POSITIVE DEFINITE (≥ 0)

$\downarrow$

CAN BE DIAGONALIZED WITH ORTHOGONAL MATRIX

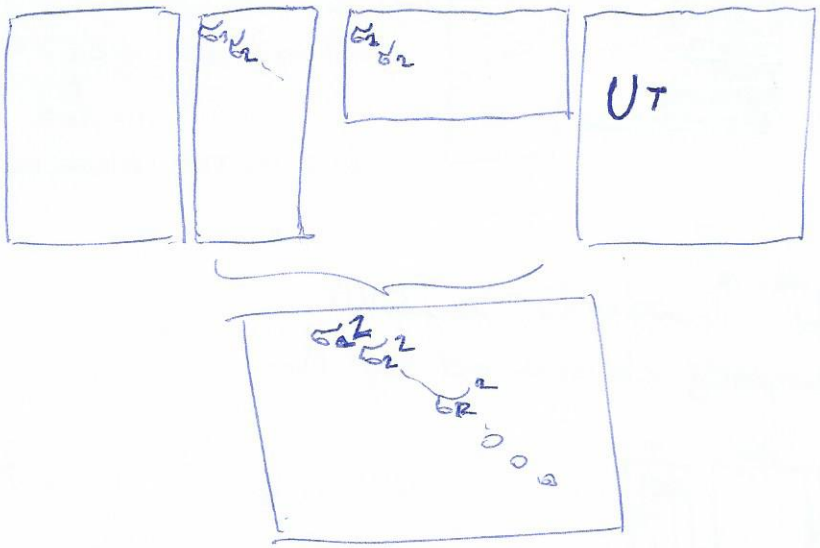
Indeed recall the definition:

$$\forall x \quad x^T(AA^T)x = (x^TA)(A^Tx) = (A^Tx)^T(A^Tx) = \|A^Tx\|^2 \geq 0$$

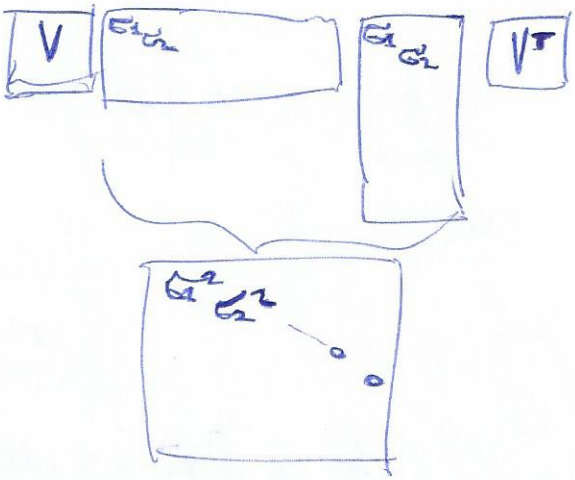
Let $A \in \mathbb{R}^{m \times n}$, $A = U\Sigma V^T$, $AA^T = U\Sigma \cancel{V^T}(V\Sigma^T U^T)$

V is orthogonal !!!

So $AA^T =$



$A^TA = V\Sigma^T U^T U \Sigma V^T$



IMPORTANT THINGS:

1) THERE EXISTS AN SVD $A = U\Sigma V^T$
2) (*)
3) The columns of U form a basis of range AA^T (orthogonal matrix in diagonalization of AA^T)
4) The columns of V form a basis of range A^TA (U = orthogonal matrix in diagonalization of A^TA)
5) $\sigma_1 \dots \sigma_r$ are the square roots of the non-zero eigenvalues of both A^TA and AA^T

LINEAR EQUATION (System of linear equations)

$Ax = b$ → UNIQUE SOLUTION; this means A^{-1} exists

$(A^TA)x = A^Tb \Rightarrow x = (A^TA)^{-1}A^Tb$

$A \in \mathbb{R}^{m \times n}$ (TALL MATRIX FULL RANK)

$b \in \mathbb{R}^m$

find $x \in \mathbb{R}^n$

→ INFINITELY MANY SOLUTIONS; we must choose one (singular matrix) NULL $A \neq \{0\}$

$\forall v \in \text{Nul}\, A$, if $\bar{x}$ is a solution ($A\bar{x} = b$) then $\bar{x} + v$ is also ($A(\bar{x}+v) = A\bar{x} + Av = b$)

↳ many x → CHOOSE THE ONE WITH MINIMUM NORM

→ NO SOLUTIONS; must find an approximate one

find x that attains $\min_x \|Ax - b\|_2$

(Remember: $Ax = b \Rightarrow \|Ax - b\|_2 = 0$)

↓

there can be many x that minimize

$\arg\min_x$ not unique

← So, I REFER TO THE CASE 2) WITH INFINITELY MANY SOLUTIONS

Def: $S(A,b) = \{x \in \mathbb{R}^n \mid x \text{ minimizes } \|Ax - b\|_2\}$

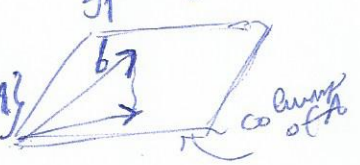
Any $x \in S(A,b)$ is "almost" or really a solution.

Def: $x^* := \arg\min_{x \in S(A,b)} \|x\|_2$ is called the LEAST SQUARES SOLUTION of the equation $Ax = b$

PROPOSITION: A LS SOLUTION of $Ax = b$ always exists and is unique

We have to minimize: $\min \|Ax - b\|_2 \Rightarrow \|Ax - b_c\| = 0$ (b_c = PROJECTION of b ONTO Span{columns of A})

x^* = projection of any such x onto Span{rows of A}



$$E\left[y(t)^2\right] = a_1^2 E\left[y(t-1)^2\right] + \sigma_e^2$$

A process $\{y(t)\}$ is wide-sense STATIONARY if

$t \in \mathbb{Z}$

1) $E[y(t)] = \mu$ DOES NOT DEPEND ~~on t~~

2) $E[y(t)\, y(t+\tau)] = r(\tau)$ ← CORRELATION SIGNAL, DOES NOT DEPEND ON t

(τ: TIME LAG)

$\{y(t)\}$, $t \in \{0, 1, 2, \dots\}$ IS QUASI-STATIONARY

1) $|E[y(t)]| \leq c$
2) $|E[y(t)\, y(t+\tau)]| \leq c$

} bounded

3) $$\lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} E[y(t)\, y(t+\tau)] = r(\tau) \quad (\exists \text{ A LIMIT})$$

THE AVERAGES OF THE CORRELATIONS EXISTS

PROPERTY: if $y(t)$ is a quasi-stationary process then

$\lim_{t\to\infty} E[y(t)]$ } ~~exists~~

$\lim_{t\to\infty} E[y(t)^2]$ }

So: $$E[y(t)^2] = a_1^2 E[y(t-1)^2] + \sigma_e^2$$

$$\sigma_y^2 = a_1^2 \sigma_y^2 + \sigma_e^2 \rightarrow \sigma_y^2 = \frac{\sigma_e^2}{1 - a_1^2}$$

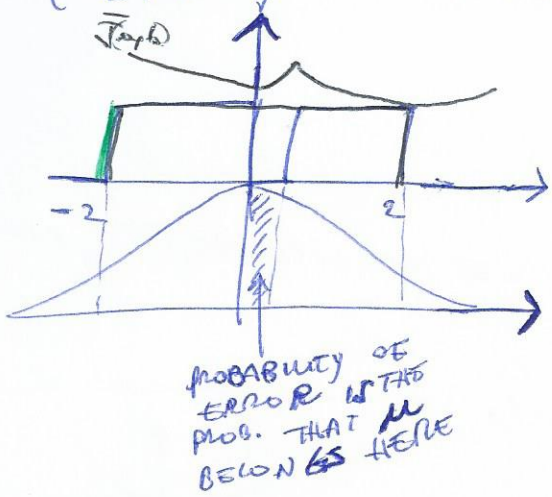
Prop: if $W(z)$ is BIBO-stable 1) $y(t)$ is QUASI-STATIONARY

2) $\frac{1}{N} \sum_{t=0}^{N-1} y(t) \xrightarrow{A.S.} \mu = \lim_{t\to\infty} E[y(t)]$ } ERGODICITY

3) $\frac{1}{N} \sum_{t=0}^{N-1} y(t) y(t+\tau) \xrightarrow{A.S.} r(\tau)$

FIRST: WE CHOOSE

$\mathcal{F}_1 = \{\mathbb{1}_{[a,b]}(u)\}$ $\quad C = \{(a,b) \in \mathbb{R}^2 \mid a < b\}$



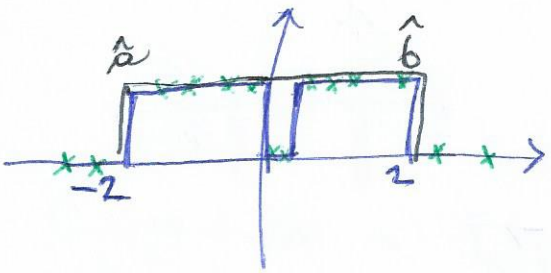
IFIX $a = -2$ AND I ENLARGE b

$\bar{C} = (-2, 2)$

$\bar{J}(\bar{C})$ = probability that $U \in [0, 1/3]$

$= \frac{1}{\sqrt{2\pi}} \int_0^{1/2} e^{-t^2/2} dt \simeq 0.13$

What about $\hat{J}_N, \hat{C}_N$?



$\hat{C}_N = [-2-\varepsilon, 2+\varepsilon]$

SAY $\hat{J}_N(\hat{C}_N) = 0.125$

IN THIS CASE THE APPROXIMATION IS GOOD $\bar{J}(\hat{C}_N) \simeq \bar{J}(\bar{C})$

SECOND: WE CHOOSE

$\mathcal{F}_2 = \mathcal{F}_1 \cup \{\mathbb{1}_{\text{finite sets}}(u)\}$

$C_1 = \{(a,b)\} \cup \{S = \text{finite set}\}$

$\mathcal{F}_2 = \{\text{⊓}\} \cup \{\text{⁞⁞⁞}\}$ (N (ANY N))

NEEDLE FUNCTIONS — THE PROBABILITY TO BE A POINT $\leq \varepsilon$

$\hat{f}_c$ minimizing $\bar{J}$: the same AS BEFORE

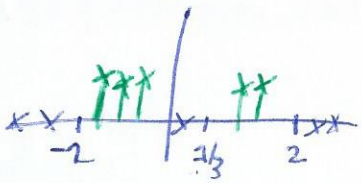
Repeating $\bar{J}$: the optimal $\hat{f}_c$ in $\mathcal{F}_1$ is the same or before

$(\bar{J}(\bar{C}) = 0.13)$ But for any needle function $\hat{f}_c$, $\bar{J}(C) = \int_{-2}^{-1/2} \frac{1}{\pi} e^{-t^2/2} dt +$

$+ \int_{1/2}^{2} \frac{1}{\pi} dt = 0.82$

IT'S BAD! WE DON'T HAVE TO CHOOSE OUR FUNCTION AMONG "NEEDLE FUNCTIONS"

What about $\hat{J}_N$, $\hat{C}_N$?

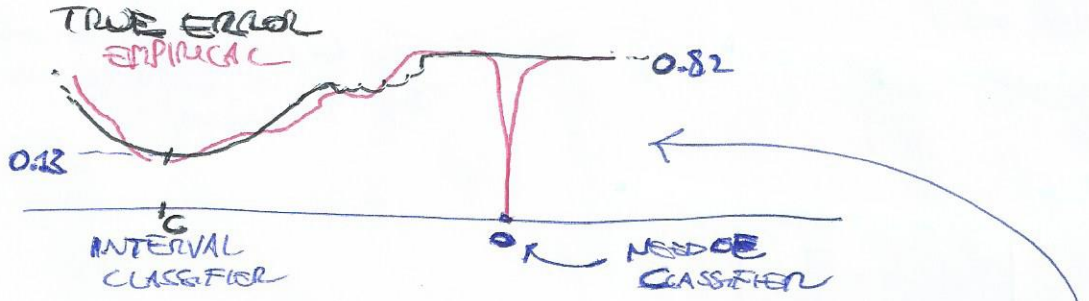


choosing $\hat{f}_C = \{\mathbb{1}\{u_i | Y_i = 1\}\}$

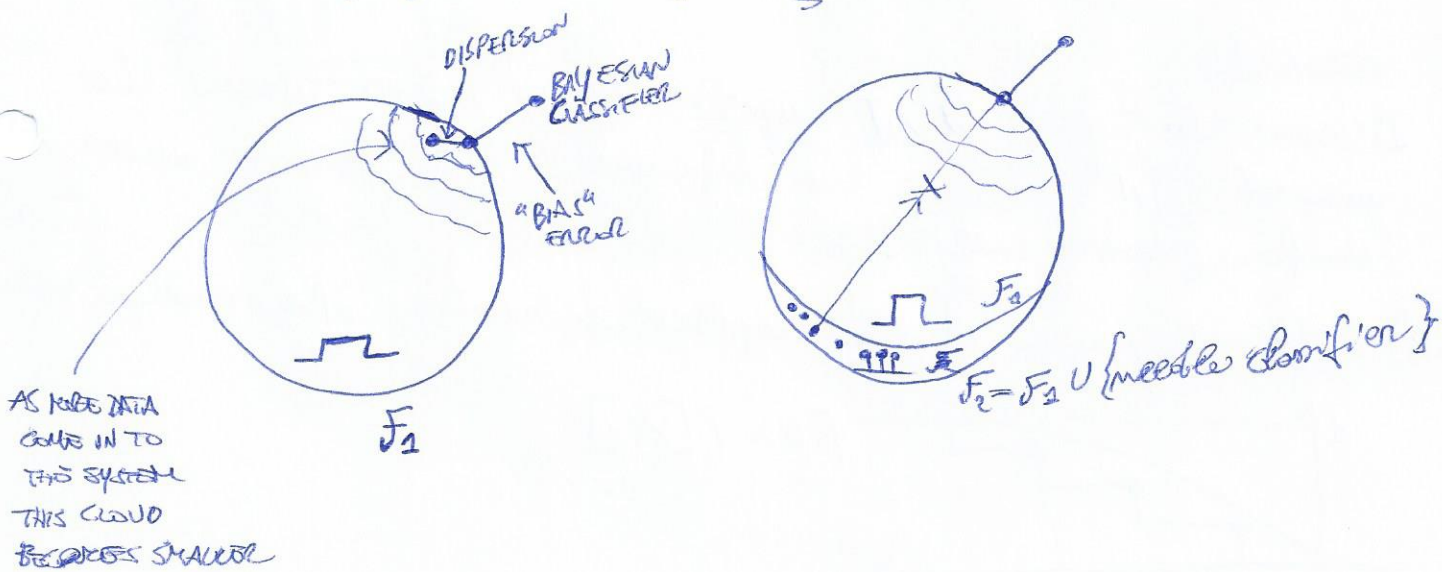
$$\sum_{i=1}^{N} \underbrace{\mathbb{1}(Y_i \neq \hat{f}_C(U_i))}_{0} \Rightarrow \hat{J}_N(\hat{C}_N) = 0$$

$$\bar{J}(\hat{C}_N) = 0.82$$

BECAUSE THE SET $\{\mathbb{1}_{\text{finite set}}(u)\}$ IT'S TOO COMPLEX!!!



$\hat{J}_N \xrightarrow{N \to \infty} \bar{J}$ UNIFORMLY? NO!!! (BECAUSE OF THIS POINT)



$\hat{J}_N$ "ε-close" to $\bar{J}$ if $\sup_{c \in C} \left| \hat{J}_N(c) - \bar{J}(c) \right| \leq \varepsilon$