System Identification and Data Analysis

Federico A. Ramponi
Università di Brescia, dipartimento di ingegneria dell'informazione

Written test
September 11th, 2013
Duration: 2 hours.

Question 1 [5%] Suppose that the variables $y_i \in \mathbb{R}$ are explained by the regressors $\varphi_i \in \mathbb{R}^p$ according to the following model:

$$y_i = \varphi_i^\top \theta^o + \varepsilon_i, \quad i = 1, \cdots, N.$$

Write the normal equations of the least squares method to estimate the parameter $\theta^o \in \mathbb{R}^p$. Discuss (max. 5 lines) the uniqueness of a solution to the normal equations; when the solution $\hat{\theta}_{\text{LS}}$ is unique, provide an explicit expression for it.

Question 2 [15%] Suppose that the variables $y_i \in \mathbb{R}$ are explained by the regressors $\varphi_i \in \mathbb{R}^p$ according to the model $y_i = \varphi_i^\top \theta^o + \varepsilon_i$. Suppose, moreover, that

1. $\{\varphi_i\}_{i=1}^\infty$ are deterministic vectors chosen in such a way that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \varphi_i \varphi_i^\top = \Sigma,$$

   where Σ is symmetric and *positive definite* ($\Sigma > 0$), hence *invertible*.

2. $\{\varepsilon_i\}_{i=1}^\infty$ are i.i.d. random variables, with mean $\mathsf{E}[\varepsilon_i] = 0$.

Show that $\hat{\theta}_{\text{LS}} \to \theta^o$ almost surely as $N \to \infty$. How is this property called?
Only a sketch of the proof is requested. Here are some hints:

- If a sequence of matrices $\{\Sigma_i\}$ converges to a matrix Σ, and Σ is invertible, then for large N the matrices $\{\Sigma_i\}$ are invertible as well.
- Strong law of large numbers: let $\{x_i\}_{i=1}^\infty$ be i.i.d. random variables, or vectors, with mean $\mathsf{E}[x_i] = \mu$. Then $\frac{1}{N}\sum_{i=1}^N x_i \to \mu$ almost surely.

Question 3 [20%] Explain briefly (max. $\frac{3}{4}$ of a page) what is the method of instrumental variables, and why we introduced it in class.
Hint. We started from this example: consider the dynamical model

$$y(t) = a^o y(t-1) + b^o u(t-1), \qquad y_m(t) = y(t) + e(t),$$

where $a^o, b^o \in \mathbb{R}$ are constants, $\{u(t)\}$ is an ("exogenous") input process, $\{e(t)\}$ are i.i.d. random variables with zero mean, independent from the past of $y(\cdot)$ and from $u(\cdot)$. The experimenter wants to estimate a^o and b^o; he/she can impose $u(t)$ and measure $y_m(t)$, but cannot measure $y(t)$. Substituting, we get

$$\begin{aligned} y_m(t) - e(t) &= a^o(y_m(t-1) - e(t-1)) + b^o u(t-1), \\ y_m(t) &= a^o y_m(t-1) + b^o u(t-1) + (e(t) - a^o e(t-1)), \\ y_m(t) &= a^o y_m(t-1) + b^o u(t-1) + r(t), \end{aligned}$$

which has the same structure of an ARX model. Hence one may think to apply the least squares method with $y_t = y(t)$, $\varphi_t = \begin{bmatrix} y_m(t-1) \\ u(t-1) \end{bmatrix}$, $\theta = \begin{bmatrix} a \\ b \end{bmatrix}$, and $\varepsilon_t = r(t)$ in order to get an estimate $\hat{\theta}_{\text{LS}}$ of $\theta^o = \begin{bmatrix} a^o \\ b^o \end{bmatrix}$. But this method does *not* work ($\hat{\theta}_{\text{LS}} \not\to \theta^o$), because...

Question 4 [20%] We have three measures of the variables u, y as follows:

u_i	1	2	3
y_i	3.2	5.8	9.1

Suppose that we want to explain the measures with a linear model $y_i = \theta u_i + \varepsilon_i$ comprising noise terms ε_i, which are independent, continuous random variables with density symmetric around 0. Use the LSCR method to provide a confidence interval for θ^o.

Hint:

	1	2	3
I_1	•	•	○
I_2	•	○	•
I_3	○	•	•
I_4	○	○	○

Question 5 [15%] Consider a binary classification problem in which 20 classifier functions (threshold, interval, ... whatever) are assigned:

$$\hat{f}_c : \mathbb{R} \to \{0, 1\} \qquad c = 1, \cdots, 20.$$

Suppose that $(U_1, Y_1), \cdots, (U_N, Y_N)$ are i.i.d. input-output pairs, where $U_i \in \mathbb{R}$ and $Y_i \in \{0, 1\}$. Recall that the *true* error function is $\bar{J}(c) = \mathsf{P}[\hat{f}_c(U_i) \neq Y_i]$, and that the *empirical* error function is $\hat{J}_N(c) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{\hat{f}_c(U_i) \neq Y_i\}}$. Using Hoeffding's inequality,

$$\mathsf{P}\left[|S_N - \mathsf{E}[S_N]| > \varepsilon\right] \leq 2e^{-2N\varepsilon^2},$$

compute how many (N) measures are needed to assert, with confidence $1 - 10^{-4}$, that the empirical error approximates the true one with accuracy $\varepsilon = 5\%$, for all the classifiers ($\forall c$) simultaneously.

Question 6 [5%] Suppose that $x_1, \cdots, x_{10}$ are independent and identically distributed random variables, distributed according to an unknown density. Consider the random interval

$$I = \left[\min_{i=1,\cdots,10} x_i, \quad \max_{i=1,\cdots,10} x_i\right].$$

What is the probability that a "future" independent observation x_{11}, distributed according to the same density, belongs to I?

Question 7 [20%] Consider a convex problem of this form:

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & |\theta - x_1| \leq \gamma \\ & \vdots \\ & |\theta - x_{10}| \leq \gamma \\ & \theta \in \mathbb{R}, \gamma \in \mathbb{R}, \end{aligned}$$

where x_i are random observations as in Question 6. Suppose that its solution is (θ^*, γ^*). Write down the definition of a *support constraint*. Discuss briefly (max. $\frac{1}{2}$ of a page) how this problem is linked to the interval I of Question 6, and to the probability requested therein.

Hint: make sure that you understand *clearly* what is the (epi-)graph of the function $\gamma = g_i(\theta) = |\theta - x_i|$. Now draw an instance of the optimization problem on the (θ, γ)-plane.

System Identification and Data Analysis

Federico A. Ramponi
Università di Brescia, dipartimento di ingegneria dell'informazione

Written test
September 3rd, 2013
Duration: 2 hours.

Question 1 [5%] Suppose that the variables $y_i \in \mathbb{R}$ are explained by the regressors $\varphi_i \in \mathbb{R}^p$ according to the following model:

$$y_i = \varphi_i^\top \theta^o + \varepsilon_i, \quad i = 1, \cdots, N.$$

1. Write the normal equations of the least squares method to estimate the parameter $\theta^o \in \mathbb{R}^p$.
2. Discuss (max. 5 lines) the uniqueness of a solution to the normal equations; when the solution $\hat{\theta}_{\text{LS}}$ is unique, provide an explicit expression for it.

Question 2 [15%] Suppose that the variables $y_i \in \mathbb{R}$ are explained by the regressors $\varphi_i \in \mathbb{R}^p$ according to the model $y_i = \varphi_i^\top \theta^o + \varepsilon_i$. Suppose, moreover, that

1. $\{\varphi_i\}_{i=1}^\infty$ are i.i.d. random vectors, with a covariance matrix $\Sigma = \mathsf{E}\left[\varphi_i \varphi_i^\top\right]$ which is ***positive definite***, hence ***invertible***;
2. $\{\varepsilon_i\}_{i=1}^\infty$ are i.i.d. random variables, with mean $\mathsf{E}[\varepsilon_i] = 0$;
3. ε_i is independent from φ_i for all i (it follows that $\mathsf{E}[\varphi_i \varepsilon_i] = \mathsf{E}[\varphi_i]\mathsf{E}[\varepsilon_i]$).

Show that $\hat{\theta}_{\text{LS}} \to \theta^o$ almost surely as $N \to \infty$. How is this property called?
Only a sketch of the proof is requested. Here are some hints:

- if a sequence of matrices $\{\Sigma_i\}$ converges to a matrix Σ, and Σ is invertible, then for large N the matrices $\{\Sigma_i\}$ are invertible as well;
- Strong law of large numbers: Let $\{x_i\}_{i=1}^\infty$ be i.i.d. random variables (or vectors) with mean $\mathsf{E}[x_i] = \mu$. Then $\frac{1}{N}\sum_{i=1}^N x_i \to \mu$ almost surely.

Question 3 [~~15%~~ 20] Consider the following dynamical model:

$$y(t) = ay(t-1) + bu(t-1) + e(t),$$

where $a, b \in \mathbb{R}$ are constants, $\{u(t)\}$ is an ("exogenous") input process, $\{e(t)\}$ are i.i.d. random variables with zero mean, and $e(t)$ is independent from $y(t-1), y(t-2), \cdots$.

1. Show how to apply the method of least squares in order to estimate the constants a and b, as seen in class.
2. Suppose now that $u(t) = ky(t)$ for all t (output feedback). Show that the solution to the corresponding least squares problem cannot be unique. Draw some conclusion about the identifiability of a and b.

Question 4 [15%] We have three measures of the variables u, y as follows:

u_i	1	2	3
y_i	1.2	3.9	6.1

Suppose that we want to explain the measures with a linear model comprising a noise term:

$$y_i = \theta u_i + \varepsilon_i,$$

where ε_i are independent, continuous random variables with density symmetric around 0. Use the LSCR method to provide a confidence interval for θ^o.

Hint:

	1	2	3
I_1	•	•	○
I_2	•	○	•
I_3	○	•	•
I_4	○	○	○

Question 5 [5%] Suppose that $X_1, \cdots, X_N, \cdots$ is a sequence of i.i.d. random variables with common distribution $F(x) = \mathsf{P}[X_i \leq x]$. Write down the definition of their *empirical distribution* $\hat{F}_N(x)$ and show, using the strong law of large numbers, that *for any fixed* $x \in \mathbb{R}$, $\hat{F}_N(x) \to F(x)$ *almost surely.*

Question 6 [15%] Suppose that N students compile a questionnaire and reply to a question that allows for 6 possible answers. Assume that the respective answers are i.i.d. random variables (taking 6 possible values). We estimate the probability p_i of a certain answer A_i as

$$\hat{p}_i = \frac{\text{number of answers equal to } A_i}{N}.$$

Using Hoeffding's inequality,

$$\mathsf{P}\left[|S_N - \mathsf{E}[S_N]| > \varepsilon\right] \leq 2e^{-2N\varepsilon^2},$$

compute how many questionnaires are needed to claim, with confidence 99%, that the true "mass distribution" $\{p_1, \cdots, p_6\}$ has been estimated with an uniform accuracy of 10%.

Question 7 [5%] Consider a convex problem of this form:

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

Suppose that its solution is (θ^*, γ^*). Write down the definition of a *support constraint.*

Question 8 [15%] Describe briefly (max. $\frac{3}{4}$ of a page) what is the purpose of the IPM (Interval Predictor Models) method, and how it works. You may provide an example of your choice. If you decide to state some results, no proofs are requested.

System Identification and Data Analysis, Spring 2013

Federico A. Ramponi
Università di Brescia, dipartimento di ingegneria dell'informazione

Written test
July 8th, 2013
Duration: 2 hours.

Question 1 [5%] Suppose that the variables $y_i \in \mathbb{R}$ are explained by the regressors $\varphi_i = \varphi(x_i) \in \mathbb{R}^p$ according to the following model:

$$y_i = \varphi_i^\top \theta^o + \varepsilon_i, \quad i = 1, \cdots, N. \tag{1}$$

Write down the normal equations of the least squares method to estimate the parameter $\theta^o \in \mathbb{R}^p$. Assuming some more hypotheses if necessary, provide an explicit expression for the estimate $\hat{\theta}_{\text{LS}}$.

Question 2 [15%] Assume that (1) is the "true" model according to which the $\{y_i\}$ are generated. Suppose that the regressors $\{\varphi_i\}$ are deterministic, and that the errors $\{\varepsilon_i\}$ are i.i.d. random variables with mean 0 and variance σ^2. Under the further hypotheses that you have assumed in Question 1, show that $\mathsf{E}\left[\hat{\theta}_{\text{LS}}\right] = \theta^o$ (how is this property of $\hat{\theta}_{\text{LS}}$ called?) and that $\mathsf{Var}\left[\hat{\theta}_{\text{LS}}\right] = \sigma^2 \left(\sum_{i=1}^n \varphi_i \varphi_i^\top\right)^{-1}$.

Question 3 [15%] We have four measures of the variables x, y as follows:

x_i	1	2	3	4
y_i	2.1	0.9	0.7	0.5

Say that we have good reasons to suspect an inverse proportionality between them, as follows:

$$y_i = \frac{\theta}{x_i};$$

on the other hand, we also know that the measures y_i are corrupted by noise. Compute the least squares estimate of the parameter $\theta \in \mathbb{R}$.

Question 4 [15%] Suppose that the variables y_1, y_2, y_3, are generated by the model

$$y_i = \theta^o u_i + \varepsilon_i,$$

where $u_i \in \mathbb{R}$, $u_i > 0$, and $\theta^o \in \mathbb{R}$ is a parameter. Explain briefly what is a *confidence interval* for θ^o. Make some assumptions on $\varepsilon_1, \varepsilon_2, \varepsilon_3$, and show how to obtain a confidence interval for θ^o (having *guaranteed* confidence) with the LSCR method.

Hint:

	1	2	3
I_1	•	•	○
I_2	•	○	•
I_3	○	•	•
I_4	○	○	○

Question 5 [20%] Describe *briefly* (do not exceed $\frac{3}{4}$ of a page) what are, in a binary classification problem in which the classifier is indexed by the parameter c, the "true" error $\bar{J}(c)$, the empirical error $\hat{J}_N(c)$, and what is the "hope", in the machine learning theory seen in class, when we minimize the latter with respect to c (or vice-versa, what can go wrong if the family of classifiers indexed by c is too complex).

Question 6 [5%] Write down the statement (*not* the proof) of the Glivenko/Cantelli theorem. (Provide a definition of the functions involved in the statement.)

Question 7 [5%] Consider a convex problem of this form:

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

Suppose that its solution is (θ^*, γ^*). Write down the definition of a *support constraint.*

Question 8 [15%] Consider now a convex problem of this form:

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

where $y_1, \cdots, y_n \in \mathbb{R}$. Sketch a picture of such a problem in the (θ, γ)-plane and mark its solution on the picture. Say if, *in general*, it is possible that: A) the problem has no support constraints at all; B) the problem has 1 support constraint; C) the problem has 2 support constraints; D) the problem has 3 support constraints. (If you believe that a situation is actually possible, sketch an example; otherwise justify your belief.)

Question 9 [5%] Consider the same problem as in Question 8, and suppose that $y_1, \cdots, y_n$ are i.i.d. continuous random variables with uniform density in the interval $[0, 1]$. (Thus, the constraints are now random themselves.) What can you deduce, from the probabilistic standpoint, about the number of support constraints?

System Identification and Data Analysis

Federico A. Ramponi
Università di Brescia, dipartimento di ingegneria dell'informazione

Written test
July 29th, 2013
Duration: 2 hours.

Question 1 [15%] Let $y_i \in \mathbb{R}$ and $\varphi_i \in \mathbb{R}^p$ for $i = 1, \cdots, N$, and let $\theta \in \mathbb{R}^p$.

1. Show that any solution to the LS problem, that is to find the $\arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^N \left(y_i - \varphi_i^\top \theta\right)^2$, must satisfy the normal equations:

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

   In other words, show how the normal equations were obtained in class. Hint: *pay attention to the compatibility of multiplications!* φ_i and θ are supposed to be *column vectors*, hence expressions like "$\theta\varphi_i$" do not make sense.

2. Discuss (max. 5 lines) the uniqueness of a solution of (1).

Question 2 [15%] We have five measures of the variables x, y as follows:

x_i	1	2	3	4	5
y_i	1.2	3.9	9.1	15.6	25.3

We have good reasons to suspect a quadratic relation between them, as follows:

$$y_i = \theta x_i^2;$$

on the other hand, we also know that the measures y_i are corrupted by noise. Compute the least squares estimate of the parameter $\theta \in \mathbb{R}$.

Question 3 [15%] Consider the following dynamical model:

$$y(t) = ay(t-1) + bu(t-1) + e(t),$$

where $a, b \in \mathbb{R}$ are constants, $\{u(t)\}$ is an ("exogenous") input process, $\{e(t)\}$ are i.i.d. random variables with zero mean, and $e(t)$ is independent from $y(t-1), y(t-2), \cdots$. Show how to apply the method of least squares in order to estimate the constants a and b. Describe a pathological situation in which it is *not possible* to estimate a and b (because the least squares solution is intrinsically not unique).

Question 4 [15%] Describe *briefly* (max. $\frac{3}{4}$ of a page) what the purpose of the LSCR method is, the hypotheses under which it is usually applied, and how it works. You may provide an example of your choice. If you decide to state some results, no proofs are requested.

Hint:

	1	2	3
I_1	•	•	○
I_2	•	○	•
I_3	○	•	•
I_4	○	○	○

Question 5 [5%] Suppose that $X_1, \cdots, X_N, \cdots$ is a sequence of i.i.d. random variables with common distribution $F(x) = \mathsf{P}[X_i \leq x]$. Write down the definition of their *empirical distribution* $\hat{F}_N(x)$ and show, using the strong law of large numbers, that *for any fixed* $x \in \mathbb{R}$, $\hat{F}_N(x) \to F(x)$ *almost surely.*

Question 6 [15%] Consider a binary classification problem in which 20 classifier functions (threshold, interval, ... whatever) are assigned:

$$\hat{f}_c : \mathbb{R} \to \{0, 1\} \qquad c = 1, \cdots, 20.$$

Suppose that $(U_1, Y_1), \cdots, (U_N, Y_N)$ are i.i.d. input-output pairs, where $U_i \in \mathbb{R}$ and $Y_i \in \{0, 1\}$. Recall that the *true* error function is $\bar{J}(c) = \mathsf{P}[\hat{f}_c(U_i) \neq Y_i]$, and that the *empirical* error function is $\hat{J}_N(c) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{\hat{f}_c(U_i) \neq Y_i\}}$. Using Hoeffding's inequality,

$$\mathsf{P}\left[|S_N - \mathsf{E}[S_N]| > \varepsilon\right] \leq 2e^{-2N\varepsilon^2},$$

compute how many (N) measures are needed to assert, with confidence $1 - 10^{-4}$, that the empirical error approximates the true one with accuracy $\varepsilon = 5\%$, for all the classifiers ($\forall c$) simultaneously.

Question 7 [5%] Consider a convex problem of this form:

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

Suppose that its solution is (θ^*, γ^*). Write down the definition of a *support constraint.*

Question 8 [15%] Recall that an optimization problem is called *feasible* if there exists at least one point (θ, γ) that satisfies all the constraints. Consider now a 2-dimensional convex problem with four linear constraints:

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & a_1\theta + b_1 \leq \gamma \\ & \vdots \\ & a_4\theta + b_4 \leq \gamma \\ & \theta \in \mathbb{R}, \gamma \in \mathbb{R}, \end{aligned}$$

where $a_1, b_1, \cdots, a_4, b_4 \in \mathbb{R}$. Say if, *in general*, it is possible that:
A) the problem is not feasible;
B) the problem is feasible, but does not admit a solution (θ^*, γ^*);
C) the problem admits a solution, but such solution is not unique;
D) the problem has a unique solution and no support constraints at all;
E) the problem has a unique solution and 1 support constraint;
F) the problem has a unique solution and 2 support constraints;
G) the problem has a unique solution and 3 support constraints.
If you believe that a situation is actually possible, sketch an example; otherwise justify your belief.

Exam of 18-06-2012

Exercise 1

1. Consider the process $\{y(t)\}_{t=0}^{\infty}$ generated by the following dynamic model:
$$y(t+1)=a_0 y(t)+a_1 y(t-1)+w(t)$$
where $\{w(t)\}_{t=0}^{\infty}$ are independent and identically distributed random variables, with zero mean and with a certain unknown variance $\sigma^2>0$, and $w(t)$ is independent from $y(s)$ for each $s\leq t$. Given $N+2$ measures $y(0),\dots,y(N+1)$ of the process, write the normal equations to estimate the parameters a_0 and a_1 with the least squares method.
2. Explain briefly what does it mean that, under appropriate hypothesis, the least squares method is consistent; say if the consistence is verified in the previous point, justifying briefly the answer.

EXERCISE 2

Show briefly (using maximum half of a page) what are, in a binary classification problem where the classifier is indexed by the parameter c, the "true" error $\bar{J}(c)$, the empirical error $\hat{J}_N(c)$ and which problems can arise minimizing the second with the idea of approximating the minimum of the first.

EXERCISE 3

A big company receives N complaint phone calls $t_{1,..},t_N$ and for each call t_i, the company records the origin region r_i. Suppose that r_i are independent and identically distributed random variables; each of them can assume values in the set of the 20 regions (Piemonte, Lombardia, ..., Sicilia) with the respectives probabilities P=(p(Piemonte), p(Lombardia), ..., p(Sicilia)) (that depend on the population of the regions, quality of service, etc.). Using Hoeffding inequality:
$$P\left[\left|S_N-E[S_N]\right|\geq\epsilon\right]\leq 2e^{-2N\epsilon^2}$$
compute how many phone calls should be collected in order to estimate the probability "distribution" P in order to be sure, with confidence at least $1-10^{-10}$, that for all the regions simultaneously the probability estimation error is at most $\epsilon=1$ %.

EXERCISE 4

1. Consider a convex optimization problem like the following:

$$min_{x,y}\, y \quad \text{such that} \quad g_1(x) \le y, \ldots, g_n(x) \le y, x \in \mathbb{R}, y \in \mathbb{R}$$

and (x^*, y^*) is its solution. Explain briefly what does it mean that a certain constraint $g_1(x) \le y$ is *of support* for the solution.

2. Consider an optimization problem of the same kind, with three constraints delimited by parables facing upwards:

$$min_{x,y}\, y \quad \text{such that} \quad a_1 x^2 + b_1 x + c_1 \le y$$
$$a_2 x^2 + b_2 x + c_2 \le y$$
$$a_3 x^2 + b_3 x + c_3 \le y$$
$$x \in \mathbb{R}, y \in \mathbb{R}$$

where $a_i > 0$ for $i = 1,2,3$. Consider the following situations:

- no constraints are of support
- only one support constraint is present
- two support constraints are present
- all the constraints are of support

in each of the upper cases say if the situation is possible and if it is so, draw the constraints graph of an instance of the problem in which the situation exists (is not required the millimetric precision, a sketch is enough).

EXERCISE OPTIONAL

Note: this exercise will be corrected in order to improve the mark <u>if and only if</u> all the previous exercises have been solved; it's suggested so not to spend time for this before having solved the others.

Below are reported the lengths of some trades of the Milano–Venezia railway and the ticket price of a "fast regional" train that covers them, approximated at the nearest multiple of 1€.

Trade	Length	Ticket Price
Milano–Brescia	83 Km	7 €
Milano–Verona	148 Km	11 €
Milano–Padova	229 Km	15 €

Give a reasonable model that explains the price according to the length of the trade and of a fixed price of the ticket; estimate the price of a ticket from Milano to Vicenza, knowing that the relative trade length is 199 Km.

SIDA

Exam of 28-06-2011

EXERCISE 1

We have some observations $(u_i, y_i), i=1,\dots,N$, where $u_i \in \mathbb{R}, y_i \in \mathbb{R}$.
In order to describe the phenomena that relates u to y , it's used a model with the structure $\hat{y}=\theta u^2$ and the parameter θ is estimated at the least squares.

1. Write the normal equations that allow to determine θ .
2. Given the three observations $(1,1),(1.5,1),(1.8,1.5)$ determine the parameter θ_{LS} that solves the normal equations. Draw the model overlapped to the data in the (u,y) plane.
3. Suppose that the data are generated with the mechanism $y=0.6u^2+n$ where u is uniformly distributed in $[0,2]$ and n is Gaussian with zero average and unitary variance. The different data are generated with inputs u independents one from each other and the noise n is independent from u . Say what the least square estimation tends to when the number of observations N tends to infinite.

EXERCISE 2

1. Describe briefly (using maximum 1/3 of a page) what the LSCR method is.
2. Consider again the 3 observations at the point 2 of the previous exercise. Determine a region Θ for the parameter θ of the model $\hat{y}=\theta u^2$ using the LSCR method.

EXERCISE 3

1. Hoeffding inequality is written: $\hat{P}(A)-P(A)\leq\epsilon$, with probability $1-\beta$ where $\beta=e^{-2Nt^2}$, where $\hat{P}(A)$ is the sample estimate of the probability of an event A . Explain the the meaning of ϵ and β (using maximum ¼ of page).
2. Over a 1 meter long bar are extracted N points according to an unknown distribution where each extraction is independent from the others. Starting from the extracted points is selected the smallest interval quantized to a millimeter (that is, the infimum of the interval is x mm and the supremum is y mm, where x and y are integers) that contains all the extractions. Using Hoeffding inequality say how many extractions is sufficient to have in order to to be sure with confidence $\beta=10^{-10}$ that the selected interval contains at least the 95% of the probability distribution.

EXERCISE 4 (optional)

Two investments funds A and B give a yearly financial return r^A and r^B ($financial\,return = final\,value / initial\,value$) a priori unknown and variable according to the investment year. So, if an investor invests 1$, x$ on A and (1−x)$ on B, the money that the investor will have at the end of the year will be $xr^A+(1-x)r^B$.

1. You have N independent observations $r_i^A, r_i^B, i=1,\dots,N$, of the financial returns of A and B in the past years (historical data). Consider the following optimization program:

   max of r such that: $r \le xr_i^A+(1-x)r_i^B, i=1,\dots,N$

   and denote with x°, r° its solution.

   Give an interpretation of the meaning of x°, r° .
2. Say what is maximum number of the support constraints for the problem at point 1.
3. Suppose to invest the next year 1$, where x°$ on A and (1−x°)$ on B. Supposing that the financial returns are independent random variables in different years and that their distribution doesn't vary over the years, determine the probability that the money that the investor will have at the end of the next year is less than r°.

ES. 1

Si hanno a disposizione alcune osservazioni (u_i, y_i) $i = 1 \dots N$ reali $(u_i, y_i \in \mathbb{R})$. Al fine di descrivere il fenomeno che lega u a y si usa il modello

$$\hat{y} = \theta u^2$$, e il parametro θ viene stimato ai minimi quadrati.

1) Si scrivono le equazioni normali per la determinazione di θ.

$$\sum \varphi_i \varphi_i^T \theta = \sum \varphi_i y_i \quad \Rightarrow \quad \sum_{i=1}^{N} u_i^4 \theta = \sum_{i=1}^{N} u_i^2 y_i$$

Qui $\varphi_i = u^2$, $\varphi_i^T = u^2$

2) Sono date le 3 osservazioni

(1, 1)
(1.5, 1)
(1.8, 1.5)

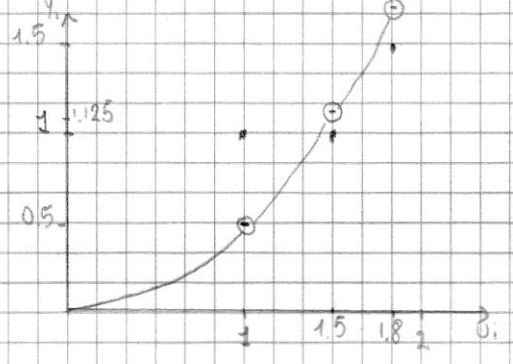
Determinare la stima ai minimi quadrati di θ e disegnare il modello corrispondente (cioè la curva $\rightarrow$ è una parabola).

u_i	y_i	u_i^4	u_i^2	$u_i^2 y_i$
1	1	1	1	1
1.5	1	5.06	2.25	2.25
1.8	1.5	10.5	3.24	4.86

$$\sum_{i=1}^{N} u_i^2 y_i = 8.11$$

$$\sum_{i=1}^{N} u_i^4 = 16.56$$

$$\theta = \frac{8.11}{16.56} \simeq 0.49$$



modello: $\hat{y} = 0.49 u^2$

Con $u = 1$ $\hat{y} \approx 0.5$

con $u = 1.5$ $\hat{y} \approx 1.12$

...

3) Si supponga che i dati siano generati dal meccanismo $y = 0.6 u^2 + n$, dove u è distribuito uniformemente in $[0, 2]$, n è gaussiano $N(0, 1)$. I diversi n sono generati con ingressi u indipendenti gli uni dagli altri e il rumore n è indipendente da u. Si dica a cosa tende la stima ai minimi quadrati quando # osservazioni $\rightarrow \infty$.

1) Se il modello è $y_i = 0.6 u_i^2 + n_i$

Converge a 0.6 quasi certamente. $n_i \rightarrow$ rumore bianco
n_i a media nulla e indipendente dalle u_i
Per la legge forte dei grandi numeri $\rightarrow 0.6$

2) Se n_i non è una costante, ma è una variabile casuale

~~$\hat{\theta}_{LS} = (\sum u_i^4)^{-1}$~~

$$y_i = \varphi^T \theta_0 + n_i$$

$$\hat{y}_i = \varphi_i^T \theta$$

$$y_i - \hat{y}_i = \varphi_i^T (\theta - \theta_0) + n_i$$

L'errore di polarizzazione è: $\left(\sum \varphi_i \varphi_i^T\right)^{-1} \sum \varphi_i n$ (n è una costante)

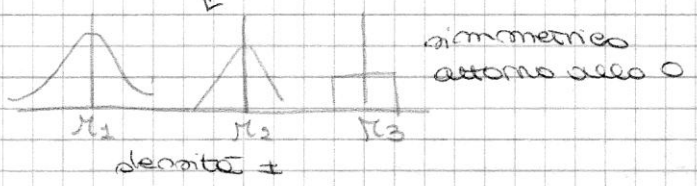
Non c'è più la media
$\sum \varphi_i \rightarrow 1$ perché uniformi
$\sum \varphi_i \varphi_i^T \rightarrow$ quadratura di positivo
$\Rightarrow$ polarizza la stima

Es. 2

1) Si descriva brevemente (≤ 1/3 di pagina) a parole a cosa serve il metodo LSCR (Leave out 1/3 Sign-dominant Correlation Region).

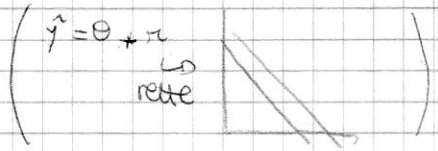
SCOPO: Fornire un intervallo di confidenza con probabilità certificata per il parametro θ in un modello lineare in θ nelle ipotesi che il rumore sia additivo e costituito da variabili indipendenti dotate di densità simmetrica rispetto all'origine (non necessariamente identicamente distribuite).

$y_i = u_i\theta + r_i$

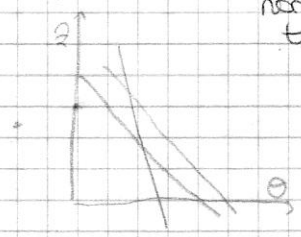


2) Si consideri ancora ee 3 osservazioni del punto 2) dell'es. precedente. Si determini una regione Θ per il parametro θ del modello $\hat{y}_i = u_i^2\theta$ con il metodo LSCR.

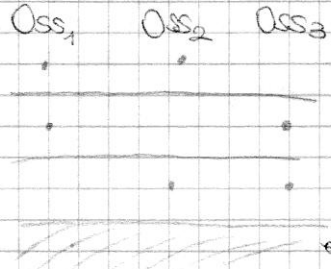
- Costruisco un gruppo



Qui: $z = y_i - u_i^2\theta$ rette con una asse y_i e una asse u_i non hanno + tutte = pendenza



Scelgo 2 rette:



Nell'es. in aula

1-2-6-5	somma
1-3-6-6	2-3-5-6
2-3-5-6	
1-2-6-7	ogni riga ha = no di elementi
1-3-5-7	
2-3-4-7	
6-5-6-7	
elemento neutro	

$z = y_1 - u_1^2\theta$

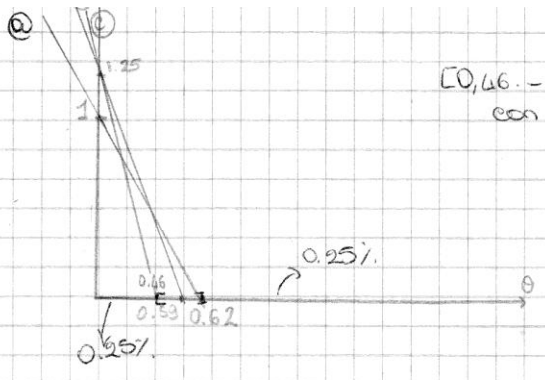
$z = y_2 - u_2^2\theta$

$z = y_3 - u_3^2\theta$

ⓐ 1ª MEDIA: $z = \frac{y_1 + y_2}{2} - \frac{u_1^2 + u_2^2}{2}\theta$ $\quad z = 1 - 1.62\theta$

ⓑ 2ª " $z = \frac{y_1 + y_3}{2} - \frac{u_1^2 + u_3^2}{2}\theta$ $\quad z = 1.25 - 2.12\theta$

ⓒ 3ª " $z = \frac{y_2 + y_3}{2} - \frac{u_2^2 + u_3^2}{2}\theta$ $\quad z = 1.25 - 2.74\theta$



[0,46 - 0.62] intervallo di confidenza
con probab. = 50%.
(ho 4 regioni in cui suddivido l'asse θ, tutte equiprobabili)

prima (es. 1) $\theta = 0.49$ $\in$ all'intervallo

ES. 3

1. La disuguaglianza di Hoeffding si scrive $\hat{P}(\hat{A}) - P(A) \leq \varepsilon$ (evento)

con probabilità $(\beta = e^{-2N\varepsilon^2})$ aumento (è < bound!!) $\rightarrow 1-\beta$ (in realtà)

dove $\beta = e^{-2N\varepsilon^2}$ (Hoeffding: $P[\dots \leq \varepsilon] \geq$ area di piccolo).

$$P[S_N - E[S_N] \geq \varepsilon] \leq \exp\left(\frac{-2N^2\varepsilon^2}{\sum(b_i-a_i)^2}\right)$$

$$P[S_N - E[S_N] \leq \varepsilon] \leq \exp(\dots)$$

dove $\hat{P}(A)$ è la stima campionaria della probabilità di un evento A. Si spieghi a parole il significato di ε e β. ($\leq \frac{1}{4}$ pagina)

Qui $S_N = \frac{1}{N}\sum_{i=1}^{N} I_A(x)$ funzione indicatrice
↳ numero di casi in cui A si verifica / numero totale di casi ($= N$)

$E[S_N] = P[x \in A]$ probabilità dell'evento

$\hat{P}(A)$ ex. Stima che un italiano sia biondo → quando si pensa e se vedo 1.

ε = numero deciso a priori x stabilire quanto la probabilità campionaria si avvicina a quella vera → accuratezza con cui voglio approssimare la distribuzione empirica. ε è piccolo, la differenza è piccola.

$1-\beta$: prima di fare l'esperimento è una probabilità. Dopo l'esperimento ho numeri, quindi è una confidenza.

ε è piccolo con probab. $1-\beta$ = confidenza

$$P[S_N - E[S_N] \leq \varepsilon] \geq 1 - \exp\left(\frac{-2N^2\varepsilon^2}{\sum(b_i-a_i)}\right)$$

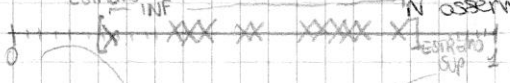
↑ β

È un bound sulla confidenza, non è una vera confidenza, su cui le SN. ruotano.
Se N non è abbastanza grande le bound esce dall'intervallo (0,1) → occorrono sufficienti osservazioni, altrimenti può venire ben > 1.

Chebyshev:

$$P[|x-\mu| \geq \varepsilon] \leq \frac{\sigma^2}{\varepsilon^2 N}$$

2. Vengono estratti N punti su un'asta di lunghezza 1 m, in accordo ad una distribuzione non nota, dove ogni estrazione è indipendente dalle altre. A partire dalle osservazioni, si desidera il più piccolo intervallo quantizzato sul misurimetro, che li contiene tutti.
[intervallo x misurimetri - y misurimetri, dove x e y sono interi].
Utilizzando la disuguaglianza di Hoeffding si dica quante estrazioni sono sufficienti per essere sicuri con confidenza $1-\beta$, dove $\beta = 10^{-10}$, che l'intervallo selezionato contenga almeno il 95% della distribuzione di probabilità.



tocche in accordo ad una distribuzione non nota

Identità su J ≥ 0.95

Quante estrazioni = quanto è grande N

Ovvio, ma sbagliato:

$$e^{-2N\,0.05^2} = 10^{-10} \qquad \varepsilon = 5\%$$

$\hat{P}(A) = 1$ se mi costruisco apposta l'intervallo

$$1 - P(A) \leq \varepsilon \qquad P(A) \geq 1 - \varepsilon \qquad \Rightarrow \varepsilon = 0.05\%$$

$$P(A) \geq 95\%$$

Risolvo per N:

$$-2N \cdot 0.05^2 = \ln 10^{-10} \quad \text{ricavo } N$$

È sbagliato xché nella disuguaglianza di Hoeffding l'evento deve essere fisso, così come il suo intervallo. Qui l'intervallo MIN-MAX è casuale, non fisso.

$$S_N = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_A(x)$$

↳ intervallo fisso, evento fisso!

Qui l'intervallo fissato a priori sono tutti gli intervalli possibili.

sono $\frac{1000 \cdot 1001}{2} \approx 1.000.000 = 10^6$ intervalli possibili (estremo inf fra tutti i 1000 pt possibili, così come l'estremo sup.).

i_1 |——|

i_2 |——|

⋮

i_{10^6} |————|

Nello spazio campionario, ∀ ciascun intervallo ho una certa probabilità di sbagliare. (vado sotto una certa soglia $\bar{\beta}$ confidenza) P(A) vale qualcosa

Per l'intervallo i_2 la disuguaglianza $\hat{P}(A) - P(A)$ non vale per $\bar{\beta}$.

... Ho 10^6 intervalli, in cui ognuno esclude una parte di probabilità sempre con = confidenza.

P complessiva che si sbagli $\leq$ somma delle rispettive probabilità ($= \bar{\beta}$).
(unione di eventi)

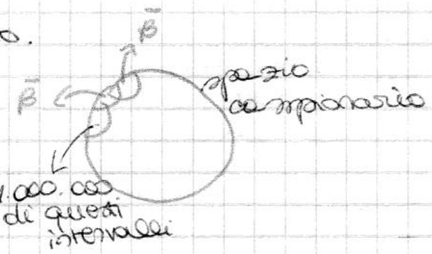
↓

Area complessiva $\leq 10^{-10}$

A deve essere fissato a priori.

$\bar{\beta}$ deve essere preso in modo che la somma sia al max 10^{-10}

$$\hookrightarrow \bar{\beta} = \frac{10^{-10}}{10^6} = 10^{-16} \quad \text{corretto.}$$



poi $e^{-2N\,0.05^2} = 10^{-16}$ e risolvo per N

VEDI DIM. LUNGA SIMILE ← ∀ intervallo, esiste una parte dello spazio campionario con probab. 10^{-10} che la disuguaglianza di Hoeffding della parte 1. non valga.

⇒ Anche la somma delle aree $\leq 10^{-10}$

ES. 4

2 fondi di investimento A e B danno un ritorno finanziario annuale R^A, R^B. R^B dà rendimenti + bassi. R^A e R^B sono casuali.

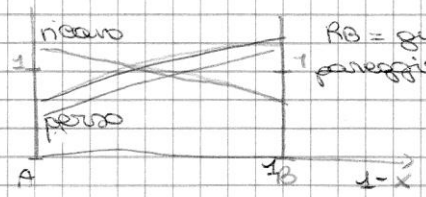
Si hanno N osservazioni dei ritorni finanziari di A e B negli anni passati indipendenti R^A_i, R^B_i, $i = 1 \dots N$ (ad es. su anni +).

1 dollaro lo posso dividere per i 2 fondi: $\begin{cases} x \text{ in } A \\ 1-x \text{ in } B \end{cases}$ dove x sono centesimi.

Se un investitore investe 1 dollaro, di cui x su A e $1-x$ su B ricevono il ritorno

$$R_A x + (1-x) R_B$$

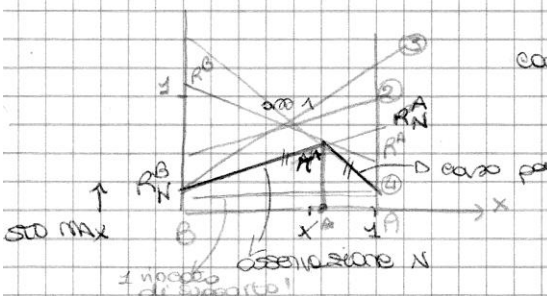
progetto: investimento x (A) + $(1-x)$ (B)



R_B = guadagno peggiore

In + periodi storici ho osservato + coppie.
→ sono tutte rette

se $x = 1$ $1-x = 0$ investo tutto in A



combinazione convessa di R^A e R^B

caso peggiore: sono i 2 vincoli di supporto

Si considerino il problema di ottimizzazione

$$\max r$$

$$\text{t.c. } r \le x r_i^A + (1-x) r_i^B \qquad \forall i = 1 \dots N$$

Sol. (x^*, r^*)

2 variabili r e x
x quota che investo nel fondo A
r = quello che ne ottengo

x^*, r^*? Spiegazione a parole:

x^* quota per cui in passato avrei guadagnato di + investendo nei suoi peggiori. Anzi massimizzato il minimo ritorno possibile in tutti gli anni (eliminando le code).

Ho caso peggiore rispetto a tutte le combo possibili (curva nera) e lo sto massimizzando (per minimizzare la perdita)

$\max r$ al variare di r e x (dim = 2, vincoli di supporto di $\max 2$)

Se investo in x^* guadagno almeno r^* in ogni anno passato.

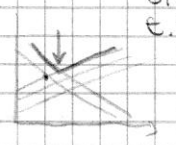
x^* = proporzione di investimento che garantisce il miglior ritorno r^* negli anni passati.

Si consideri il problema (= uno ribaltato)

$$\min_{\theta, \gamma} \gamma$$

$$\text{t.c. } \theta r_i^A + (1-\theta) r_i^B - \gamma \le 0$$

$$\theta \in [0, 1]$$



Il problema ha dimensione 2.
max n° vincoli di supporto = 2

$\dim_\theta = 1$

3) Si supponga di investire il prossimo anno 1 dollaro, di cui x^* su A e $1-x^*$ su B. Supponendo che R^A e R^B siano indipendenti su anni + e la loro distribuzione non varia di anno in anno, probabilità che il ritorno $< r^*$.

$P[N+1\text{-esimo diventi di supporto}]$

Nell'ipotesi $P[MAX, MIN] \neq 2$ $P[MAX, MIN] = 0$.

↑ potrei avere 1 solo vincolo di supporto (con probabilità $\neq 0$, ma nulla) (a causa di crisi finanziarie)

Sono in x^* → è l'anno prossimo ho una nuova retta) Se guadagno meno di prima ($< r^*$) il nuovo vincolo è di supporto.

$P[\text{ultimo vincolo osservato}] \le \frac{2}{N+1}$ è di 2 di supporto

Non posso + dire che i vincoli di supporto sono 2.

Al MAX - 1 o 2 vincoli di supporto