S.I.D.A. exercises

Revision 0.11 - March 18th, 2014

Federico Alessandro Ramponi
Dipartimento di ingegneria dell'informazione, Università degli studi di Brescia
Via Branze 38, 25123 Brescia, Italy
federico.ramponi@unibs.it

This document contains a collection of problems with more or less the same conceptual difficulty of those you may encounter in the final exam, although with possibly more demanding numerical computations (here you are invited to use Matlab or any other numerical software whenever it helps).

1 Problems

1.1 Exercises on least squares

Problem 1 (price of train tickets).
The following table contains the lengths of the railway connections between the Milano Centrale train station and the central station of other cities on the way from Milan to Venice[1], and the corresponding prices of a 'regional' train ticket[2]:

Connection	Length (km)	Ticket price (€)
Milano C.le → Brescia	82.842	7.00
Milano C.le → Verona P.N.	147.480	11.55
Milano C.le → Padova	229.408	15.65
Milano C.le → Venezia S.L.	266.341	18.35

Suppose that the prices are explained by a linear model comprising a fixed price due to administrative costs plus a price proportional to the length of the connection. The prices are "noisy" because they are quantized to multiples of 5 €-cents.

1. Write down the linear model for the ticket price and the normal equations of the least squares method, and find an estimate of the fixed price and of the proportionality coefficient.

2. Estimate the price of a ticket from Milano to Vicenza, knowing that the railway between these cities is 199.138 km long.

[1] Retrieved from http://it.wikipedia.org/wiki/Ferrovia_Milano-Venezia
[2] Retrieved from http://www.trenitalia.com on February 25, 2013.

Problem 2 (amplitude and phase of a sinusoid).
The following table contains 10 noisy measures of the values of a sinusoidal signal $y(t) = A\sin(2\pi Ft + \phi)$, taken at random times t_i:

t_i	2.188	3.043	4.207	4.937	5.675	6.104	6.260	7.150	8.600	9.655
y_i	-1.112	2.358	-1.807	1.202	-0.814	1.298	-2.520	-0.132	1.421	-0.302

We know the frequency $F = 2$ Hz of the signal, but we do not know its amplitude A and its phase ϕ.

1. Show how to apply the method of *linear* least squares in order to find an estimate of A and ϕ.
   Hint: recall trigonometry (!).
2. Compute the estimates $\hat{A}$ and $\hat{\phi}$.

Problem 3 (weighted least squares).
Given some measures $(\varphi_1, y_1), \cdots, (\varphi_N, y_N)$, the canonical Least Squares estimate is the vector in $\mathbb{R}^p$ minimizing the sum of the squared residuals:

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

Now suppose that we want to give more importance to some of the errors, and less to others, minimizing a *weighted* sum of squared residuals instead:

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

where $w_i \geq 0$ for $i = 1, \cdots, N$.

1. Find the corresponding version of the normal equations.
2. Find the new Weighted Least Squares estimate in terms of the matrices

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \qquad \Phi = \begin{bmatrix} \varphi_1^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix},$$

   assuming full rank whenever necessary.
   Hint: bring into the picture a new matrix containing the weights w_i.

Problem 4 (ranges).
Let Φ be a real matrix. Show that

$$\text{range } \Phi^\top \Phi = \text{range } \Phi^\top$$
$$\text{rank } \Phi^\top \Phi = \text{rank } \Phi^\top$$

Hint: rewrite the proof of a Lemma about the existence of a solution to the least squares problem, but keeping it in terms of Φ and $\Phi^\top$.

Problem 5 (systematic errors).
Let the measures $\{y_i\}_{i=1}^N$ be generated according to the model $y_i = \varphi_i^\top \theta^o + \varepsilon_i$, and suppose that:

1. ε_i are independent Gaussian variables with mean μ and variance σ^2;
2. φ_i are independent and identically distributed vectors with mean $\bar{\varphi}$ and second-order moment $\Sigma = \mathsf{E}\left[\varphi_i \varphi_i^\top\right] > 0$;
3. ε_i is independent of φ_i for all i.

Does the least square estimate converge almost surely as $N \to \infty$? If so, to what does it converge?

1.2 Exercises on algebraic aspects of least squares

Problem 1 (SVD and pseudo-inverse).

1. Verify that

$$A = U\Sigma V^\top = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} \sqrt{10} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

   is a singular value decomposition of $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$.

2. Compute the pseudo-inverse of $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$.

Problem 2 (orthogonal projector).
Consider $A \in \mathbb{R}^{m \times n}$, and let A^+ be its pseudo-inverse. Show that

$$\Pi_A := AA^+ \in \mathbb{R}^{m \times m}$$

is the *orthogonal projector* onto the subspace of $\mathbb{R}^m$ generated by the columns of A; more explicitly, that for any $v \in \mathbb{R}^m$, $\Pi_A v$ is the orthogonal projection of v on span {columns of A}.

1.3 Exercises on machine learning

Problem 1 (complaint telephone calls).
A big company receives N complaint calls $\{t_i\}$, $i = 1, \cdots, N$, and for each call it records the region $\{r_i\}$, $i = 1, \cdots, N$, of the caller. Suppose that the $\{r_i\}$ are independent and identically distributed random variables taking values, say, in the set of the 20 Italian regions $\{\text{Piedmont}, \text{Lombardy}, \cdots, \text{Sicily}\}$ with respective probabilities $P = \{p(\text{Piedmont}), p(\text{Lombardy}), \cdots, p(\text{Sicily})\}$ (which depend, in general, on the region's population, on the quality of service in the region, etc.). Using Hoeffding's inequality, compute how many telephone calls should be recorded in order to estimate the "mass distribution" P so that, with confidence at least $1 - 10^{-4}$, the estimation error of the probability is at most $\varepsilon = 1\%$ *at all the regions simultaneously.*

Problem 2 (finitely many classifiers).
Prove the following

Theorem 1.3.1 *Let $\mathcal{F}$ be a family of classifiers, parameterized by $c \in C$, where C* is a finite of $\mathbb{R}^p$*, namely $|C| = K$. Suppose that $(U_1, Y_1), \cdots, (U_N, Y_N)$ are independent and identically distributed, where U_i has continuous distribution $F(u)$ and $Y_i \in \{0, 1\}$. Define $\bar{J}(c)$, $\hat{J}_N(c)$, $\bar{c}$, and $\hat{c}_N$ as usual. Now the points $\hat{c}_N, \bar{c} \in C$ trivially exist, since C is finite; assume that $\bar{c}$ is unique. Then:*

1. *almost surely, $\hat{c}_N \to \bar{c}$;*
2. *for fixed $\varepsilon > 0$ and N, it holds*

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

1.4 Exercises on the LSCR method

Problem 1 (discrete distribution, wrong confidence).
Suppose that three measures are available:

$$\begin{aligned} y_1 &= \theta^o + \varepsilon_1, \\ y_2 &= \theta^o + \varepsilon_2, \\ y_3 &= \theta^o + \varepsilon_3, \end{aligned}$$

where ε_1, ε_2, and ε_3 are independent and identically distributed with *discrete*

distribution, each taking only the values 1 or -1 with equal probabilities:

$$\varepsilon_1 = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}, \end{cases}$$
$$\varepsilon_2 = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}, \end{cases}$$
$$\varepsilon_3 = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}. \end{cases}$$

Let $\theta^o = 1$. We employ LSCR method with the following group:

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

and select the interval $[\bar{\theta}_1, \bar{\theta}_3]$ which, according to the LSCR theory, should have a 50%-confidence interval. Show that the confidence of such interval is *not* 50% (this may be the case if the distribution is discrete). Hint: see what happens for every possible value of ε_1, ε_2, and ε_3.

Problem 2 (discrete distribution, correct confidence).
Suppose that three measures are available:

$$y_1 = \theta^o + \varepsilon_1,$$
$$y_2 = \theta^o + \varepsilon_2,$$
$$y_3 = \theta^o + \varepsilon_3,$$

where ε_1, ε_2, and ε_3 are independent but with discrete distribution, each taking only two possible values symmetrically around 0. Namely:

$$\varepsilon_1 = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}, \end{cases}$$
$$\varepsilon_2 = \begin{cases} \frac{1}{2} & \text{with probability } \frac{1}{2}, \\ -\frac{1}{2} & \text{with probability } \frac{1}{2}, \end{cases}$$
$$\varepsilon_3 = \begin{cases} 2 & \text{with probability } \frac{1}{2}, \\ -2 & \text{with probability } \frac{1}{2}. \end{cases}$$

Let $\theta^o = 1$, and verify that the LSCR method with the following group:

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

correctly provides a 50%-confidence interval even though the errors do not have a density. In order to do this, see what happens for every possible value of ε_1, ε_2, and ε_3.

1.5 Exercises on Interval Predictor Models

Problem 1 (support constraints).
Consider the problem:

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & f_1(\theta) \le \gamma \\ & \vdots \\ & f_N(\theta) \le \gamma \\ & \theta \in \mathbb{R}, \gamma \in \mathbb{R}, \end{aligned}$$

where $N \ge 2$, and each f_i is a convex parabola:

$$f_i(\theta) = a_i\theta^2 + b_i\theta + c_i, \quad a_i > 0.$$

Show that any such problem is feasible. For any of the following situations, tell whether it is possible, and if so, provide an example in which it holds:

- the problem has no support constraints;
- the problem has 1 support constraint;
- the problem has 2 support constraints;
- the problem has 3 support constraints.

2 Solutions

2.1 Solutions to the exercises on least squares

Solution 1 (price of train tickets).
The model reads

$$p_i = a + bd_i + \varepsilon_i,$$

where p_i is the price of the ticket in €(the "explained" variable), a the fixed price in €, b the proportionality coefficient in €/km, d_i is the distance in km (the "explanatory" variable), and ε_i is a quantization error (in €). The regressors are $\varphi_1(d) = 1$ and $\varphi_2(d) = d$, and a, b are the parameters to be estimated. To pose the problem in compact form, we let

$$Y = \begin{bmatrix} p_1 \\ \vdots \\ p_4 \end{bmatrix}, \quad \Phi = \begin{bmatrix} 1 & d_1 \\ \vdots & \vdots \\ 1 & d_4 \end{bmatrix}, \quad \theta = \begin{bmatrix} a \\ b \end{bmatrix},$$

the normal equations read

$$\Phi^\top \Phi \ \theta = \Phi^\top Y,$$

and the least squares solution is

$$\hat{\theta}_{\text{LS}} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \arg\min_\theta \|\Phi\theta - Y\|^2 = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top Y.$$

Once $\hat{a}$ and $\hat{b}$ are known, the estimated price of a ticket from Milan to Vicenza is $\hat{p} = \hat{a} + \hat{b} \cdot (199.138 \text{ km})$. The Matlab code

```
% Example: estimation of ticket prices
Y = [ 7.00 ;  11.55 ;  15.65 ;  18.35 ];
Phi = [ 1,   82.842 ;
        1,  147.480 ;
        1,  229.408 ;
        1,  266.341  ];
thetaLS = pinv(Phi)*Y
priceToVicenza = thetaLS(1) + thetaLS(2)*199.138
```

yields the estimates $\hat{a} = 2.254566$ €, $\hat{b} = 0.059955$ €/km, and $\hat{p} \simeq 14.20$ €. For comparison, the actual price of a ticket to Vicenza was 14.30 €[3].

Solution 2 (amplitude and phase of a sinusoid).
For brevity, let $\omega = 2\pi F$. The measurement model is then

$$y_i = A \sin(\omega t_i + \phi) + \varepsilon_i.$$

[3]Retrieved from `http://www.trenitalia.com` on February 25, 2013.

Since $\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y)$, we have

$$y_i = A\cos(\phi)\sin(\omega t_i) + A\sin(\phi)\cos(\omega t_i) + \varepsilon_i$$

Letting $a = A\cos(\phi)$ and $b = A\sin(\phi)$, this becomes a linear model:

$$y_i = \begin{bmatrix} \sin(\omega t_i) & \cos(\omega t_i) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \varepsilon_i,$$

where of course the explanatory data are the t_i, and the regressors are $\varphi_1(t) = \sin(\omega t), \varphi_2(t) = \cos(\omega t)$. To pose the problem in compact form, we let

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{10} \end{bmatrix}, \quad \Phi = \begin{bmatrix} \sin(\omega t_1) & \cos(\omega t_1) \\ \vdots & \vdots \\ \sin(\omega t_{10}) & \cos(\omega t_{10}) \end{bmatrix}, \quad \theta = \begin{bmatrix} a \\ b \end{bmatrix},$$

and the least squares solution is, as usual,

$$\hat{\theta}_{\mathrm{LS}} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \arg\min_{\theta} \|\Phi\theta - Y\|^2 = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top Y.$$

Note that

$$\sqrt{a^2 + b^2} = A\sqrt{\cos^2(\phi) + \sin^2(\phi)} = A;$$
$$\frac{b}{a} = \frac{A\sin(\phi)}{A\cos(\phi)} = \tan(\phi).$$

Therefore, once $\hat{a}$ and $\hat{b}$ are known, we can recover an estimate of A and ϕ as follows:

$$\hat{A} = \sqrt{\hat{a}^2 + \hat{b}^2}$$
$$\hat{\phi} = \arctan(\hat{b}/\hat{a})$$

(or $\hat{\phi} = \arctan(\hat{b}/\hat{a}) + \pi$, depending on the signs of $\hat{a}$ and $\hat{b}$). The Matlab code

```
% Example: amplitude and phase of a sinusoid
F = 2;
omega = 2*pi*F;
T = [  2.188; 3.043;  4.207; 4.937;  5.675; 6.104;  6.260;  7.150; 8.600;  9.655 ];
Y = [ -1.112; 2.358; -1.807; 1.202; -0.814; 1.298; -2.520; -0.132; 1.421; -0.302 ];
Phi = [sin(omega*T), cos(omega*T)];
thetaLS = pinv(Phi)*Y;
Ahat    = sqrt(thetaLS(1)^2 + thetaLS(2)^2)
phihat  = atan2(thetaLS(2), thetaLS(1))
```

yields the estimates $\hat{A} = 2.5036$ and $\hat{\phi} = 1.2938$. For comparison, the true values were $A = 2.5$ and $\phi = 1.3$ radians, and ε_i were Gaussian with mean 0 and variance 0.01 (i.e. standard deviation 0.1).

Solution 3 (weighted least squares).

1. To find $\arg\min_{\theta\in\mathbb{R}^p} \sum_{i=1}^N w_i\,(y_i-\varphi_i^\top\theta)^2$, we set equal to zero the derivative with respect to θ in the very same way as we do for ordinary least squares:

$$\begin{aligned}\frac{\partial}{\partial\theta}\sum_{i=1}^N w_i\,(y_i-\varphi_i^\top\theta)^2 &= \sum_{i=1}^N w_i\,2(y_i-\varphi_i^\top\theta)(-\varphi_i^\top)\\ &= \sum_{i=1}^N 2\,w_i\,(y_i-\theta^\top\varphi_i)(-\varphi_i^\top) = 0,\end{aligned}$$

After grouping terms and transposing, we find

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

This is the weighted version of the normal equations. Another way, but nicer, to get to the same result, is to bring the weights inside the squares before minimizing:

$$\sum_{i=1}^N w_i\,(y_i-\varphi_i^\top\theta)^2 = \sum_{i=1}^N(\sqrt{w_i}y_i-\sqrt{w_i}\varphi_i^\top\theta)^2;$$

defining $\bar{y}_i = \sqrt{w_i}y_i$ and $\bar{\varphi}_i = \sqrt{w_i}\varphi_i$, the problem becomes

$$\arg\min_{\theta\in\mathbb{R}^p}\ \sum_{i=1}^N(\bar{y}_i-\bar{\varphi}_i^\top\theta)^2, \tag{2}$$

which is a least squares problem in standard form. The corresponding normal equations are

$$\left(\sum_{i=1}^N \bar{\varphi}_i\bar{\varphi}_i^\top\right)\theta = \sum_{i=1}^N \bar{\varphi}_i\bar{y}_i,$$

which of course are the same as (1), once the coefficients $\sqrt{w_i}$ are extracted back from $\bar{y}_i$ and $\bar{\varphi}_i$.

2. In the same spirit as in (2), we define

$$\bar{Y} = \begin{bmatrix} \sqrt{w_1} y_1 \\ \vdots \\ \sqrt{w_N} y_N \end{bmatrix} = \begin{bmatrix} \sqrt{w_1} & & \\ & \ddots & \\ & & \sqrt{w_N} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = W^{1/2} Y,$$

$$\bar{\Phi} = \begin{bmatrix} \sqrt{w_1} \varphi_1^\top \\ \vdots \\ \sqrt{w_1} \varphi_N^\top \end{bmatrix} = \begin{bmatrix} \sqrt{w_1} & & \\ & \ddots & \\ & & \sqrt{w_N} \end{bmatrix} \begin{bmatrix} \varphi_1^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix} = W^{1/2} \Phi,$$

where $W = \operatorname{diag}(w_1, \cdots, w_N) \in \mathbb{R}^{N \times N}$, and $W^{1/2}$ denotes its square root. The problem then reads

$$\arg\min_{\theta \in \mathbb{R}^p} \left\| \bar{\Phi}\theta - \bar{Y} \right\|_2^2,$$

the corresponding normal equations are

$$\begin{gathered} \bar{\Phi}^\top \bar{\Phi} \theta = \bar{\Phi}^\top \bar{Y}; \\ \Phi^\top \left(W^{1/2} \right)^\top W^{1/2} \Phi \theta = \Phi^\top \left(W^{1/2} \right)^\top W^{1/2} Y; \\ \Phi^\top W \Phi \theta = \Phi^\top W Y, \end{gathered}$$

and finally

$$\hat{\theta}_{\mathrm{WLS}} = \left(\Phi^\top W \Phi \right)^{-1} \Phi^\top W Y.$$

Solution 4 (ranges).
Suppose that $v \in \operatorname{null} \Phi$. This means $\Phi v = 0$, hence also $\Phi^\top \Phi v = 0$ and $v \in \operatorname{null} \Phi^\top \Phi$. Suppose, on the other hand, that $v \in \operatorname{null} \Phi^\top \Phi$. Then $\Phi^\top \Phi v = 0$, hence also $\|\Phi v\|_2^2 = (\Phi v)^\top \Phi v = v^\top \Phi^\top \Phi v = 0$. This implies that $\Phi v = 0$ and $v \in \operatorname{null} \Phi$. Hence $\operatorname{null} \Phi^\top \Phi = \operatorname{null} \Phi$.
Now, since all the spaces in consideration are subspaces of finite-dimensional vector spaces,

$$\operatorname{range} \Phi^\top \Phi = \left(\operatorname{null} \left(\Phi^\top \Phi \right)^\top \right)^\perp = \left(\operatorname{null} \Phi^\top \Phi \right)^\perp = (\operatorname{null} \Phi)^\perp = \operatorname{range} \Phi^\top,$$

and consequently

$$\operatorname{rank} \Phi^\top \Phi = \dim \operatorname{range} \Phi^\top \Phi = \dim \operatorname{range} \Phi^\top = \operatorname{rank} \Phi^\top.$$

Solution 5 (systematic errors).
Since ε_i are Gaussian with mean μ and variance σ^2, we can write $\varepsilon_i = \bar{\varepsilon}_i + \mu$, where $\bar{\varepsilon}_i$ are Gaussian with mean zero. The variables $\bar{\varepsilon}_i$ are still independent of each other and independent of φ_i.
Consider the normal equations, with the substitution $y_i = \varphi_i^\top \theta^o + \varepsilon_i$ and divided by N:

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

By a strong law of large numbers, $\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top \to \Sigma$ almost surely, hence for big N the matrix $\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top$ is invertible, and

$$\begin{aligned}\hat{\theta}_{\mathrm{LS}} &= \theta^o + \left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varepsilon_i \\ &= \theta^o + \left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\varphi_i\bar{\varepsilon}_i + \left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\varphi_i\mu\end{aligned}$$

Now, since $\bar{\varepsilon}_i$ and φ_i are independent, and $\mathsf{E}[\bar{\varepsilon}_i] = 0$, the second term converges to zero almost surely by a strong law of large numbers. And since μ is a constant (we can bring it outside the sum), the third term also converges almost surely, namely to $\Sigma^{-1}\bar{\varphi}\mu$. Hence,

$$\hat{\theta}_{\mathrm{LS}} \to \theta^o + \Sigma^{-1}\bar{\varphi}\mu \quad \text{almost surely.}$$

The take-home message is that, in general, you cannot pretend the least squares method to be consistent in the presence of a *systematic error* (μ).

2.2 Solutions to the exercises on algebraic aspects

Solution 1 (SVD and pseudo-inverse).

1. We check that U and V are orthogonal; indeed

$$\begin{aligned}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} &= \begin{bmatrix}1/\sqrt{5} & -2/\sqrt{5}\\ 2/\sqrt{5} & 1/\sqrt{5}\end{bmatrix}\begin{bmatrix}1/\sqrt{5} & 2/\sqrt{5}\\ -2/\sqrt{5} & 1/\sqrt{5}\end{bmatrix} \\ &= \begin{bmatrix}1/\sqrt{5} & 2/\sqrt{5}\\ -2/\sqrt{5} & 1/\sqrt{5}\end{bmatrix}\begin{bmatrix}1/\sqrt{5} & -2/\sqrt{5}\\ 2/\sqrt{5} & 1/\sqrt{5}\end{bmatrix};\\ \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} &= \begin{bmatrix}1/\sqrt{2} & -1/\sqrt{2}\\ 1/\sqrt{2} & 1/\sqrt{2}\end{bmatrix}\begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ -1/\sqrt{2} & 1/\sqrt{2}\end{bmatrix} \\ &= \begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ -1/\sqrt{2} & 1/\sqrt{2}\end{bmatrix}\begin{bmatrix}1/\sqrt{2} & -1/\sqrt{2}\\ 1/\sqrt{2} & 1/\sqrt{2}\end{bmatrix}.\end{aligned}$$

Moreover, it holds

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} \sqrt{10} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

Thus, $A = U\Sigma V^\top$ as required. Note that the eigenvalues of both $AA^\top$ and $A^\top A$ are 10 and 0.

2. From the previous point, we have

$$\begin{aligned} A^+ = V\Sigma^+ U^\top &= \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{10} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \\ &= \begin{bmatrix} 1/10 & 1/5 \\ 1/10 & 1/5 \end{bmatrix}. \end{aligned}$$

Solution 2 (orthogonal projector).
Any $v \in \mathbb{R}^m$ can be decomposed in a unique way as

$$v = v^c + v^\perp,$$

where $v^c \in$ span {columns of A} = range A (v^c is the requested orthogonal projection) and $v^\perp \in$ span {columns of A}$^\perp$ = (range A)$^\perp$ = null $A^\top$. Specifically,

$$\begin{aligned} v^c &= Ax \quad \text{for some } x \in \mathbb{R}^n; \\ A^\top v^\perp &= 0. \end{aligned}$$

Therefore, recalling the defining properties of the pseudo-inverse,

$$\begin{aligned} \Pi_A v &= AA^+(v^c + v^\perp) \\ &= \left(AA^+A\right)x + \left(AA^+\right)v^\perp \\ &= Ax + \left(AA^+\right)^\top v^\perp \\ &= v^c + \left(A^+\right)^\top A^\top v^\perp \\ &= v^c. \end{aligned}$$

2.3 Solutions to the exercises on machine learning

Solution 1 (complaint telephone calls).

Let $r \in R = \{\text{Piedmont}, \text{Lombardy}, \cdots, \text{Sicily}\}$ denote regions. We estimate the probability $p(r) = \mathsf{P}[r_i = r]$ of a call from r with the "empirical mass distribution" (i.e. frequency)

$$\hat{p}(r) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{r_i = r\}} = \frac{\text{number of calls received from } r}{\text{total calls}}$$

Of course, since $\mathsf{E}[\mathbb{1}_{\{r_i=r\}}] = \mathsf{P}[r_i = r] = p(r)$,

$$\mathsf{E}[\hat{p}(r)] = p(r),$$

hence, using Hoeffding's inequality, the probability that $|\hat{p}(r) - p(r)| > \varepsilon$ at *any* of the 20 regions is

$$\begin{aligned}\mathsf{P}\left[\bigcup_{r\in R}\{|\hat{p}(r) - p(r)| > \varepsilon\}\right] &\leq \sum_{r\in R}\mathsf{P}\left[|\hat{p}(r) - p(r)| > \varepsilon\right] \\ &\leq \sum_{r\in R} 2e^{-2N\varepsilon^2} \\ &= 40e^{-2N\varepsilon^2}.\end{aligned}$$

The problem asks precisely to find N such that $40e^{-2N\varepsilon^2} \leq 10^{-4}$. Solving for N,

$$\begin{aligned}e^{-2N\left(\frac{1}{100}\right)^2} &\leq 25 \cdot 10^{-7}; \\ -2N\frac{1}{100^2} &\leq \log(25) - 7\log(10) \simeq -12.9; \\ N &\geq 5000 \cdot 12.9 = 64500.\end{aligned}$$

Thus, any $N \geq 64500$ will do.

Solution 2 (finitely many classifiers).
First, we prove that almost surely, $\hat{J}_N \to \bar{J}$ uniformly. Remembering the proof of Glivenko/Cantelli's theorem, this is quite easy; indeed by the strong law of large numbers, at all points $c \in C$ it holds $\hat{J}_N(c) \to \bar{J}(c)$ almost surely. Hence, almost surely for all $\varepsilon > 0$ there exists N_c such that $|\hat{J}_N(c) - \bar{J}(c)| \leq \varepsilon$ for all $N \geq N_c$. Since the c are finitely many, it is well defined the index $\bar{N} := \max_{c\in C} N_c$, such that for all $N \geq \bar{N}$ the inequalities

$$|\hat{J}_N(c) - \bar{J}(c)| \leq \varepsilon, \quad c \in C$$

hold simultaneously; this is enough to establish uniform convergence.
Now we can invoke the lemma on uniform convergence, *exploiting all the hypotheses*, because

- any *finite* subset $C \subset \mathbb{R}^p$ is automatically compact (because it is closed and bounded — recall the Heine/Borel theorem);
- any function defined on a finite set $C \subset \mathbb{R}^p$ is automatically continuous;
- $\bar{c}$ is unique by assumption.

(If you are not at ease with the second claim, see the remarks after Definition 4.5 in [1]; alternatively, you may develop an ad-hoc version of the lemma on uniform convergence.)
It follows that $\hat{c}_N \to \bar{c}$ almost surely.
Finally, recalling the definitions of $\hat{J}_N$ and $\bar{J}$, and exploiting Hoeffding's inequality,

$$\begin{aligned}
\mathsf{P}\left[\max_{c\in C} |\hat{J}_N(c) - \bar{J}(c)| \geq \varepsilon\right] &= \mathsf{P}\left[\bigcup_{c\in C} \left\{|\hat{J}_N(c) - \bar{J}(c)| \geq \varepsilon\right\}\right] \\
&\leq \sum_{c\in C} \mathsf{P}\left[|\hat{J}_N(c) - \bar{J}(c)| \geq \varepsilon\right] \\
&\leq \sum_{c\in C} 2e^{-2N\varepsilon^2} \\
&= 2Ke^{-2N\varepsilon^2}.
\end{aligned}$$

2.4 Solutions to the exercises on LSCR

Problem 1 (discrete distribution, wrong confidence).
The LSCR method works by considering 3 partial-average functions:

$$\begin{aligned}
g_1(\theta) &= \frac{y_1 + y_2}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_1 + \varepsilon_2}{2}; \\
g_2(\theta) &= \frac{y_1 + y_3}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_1 + \varepsilon_3}{2}; \\
g_3(\theta) &= \frac{y_2 + y_3}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_2 + \varepsilon_3}{2}.
\end{aligned}$$

Their respective intersections with the θ-axis are

$$\begin{aligned}
\theta_1 &= \frac{y_1 + y_2}{2}, \\
\theta_2 &= \frac{y_1 + y_3}{2}, \\
\theta_3 &= \frac{y_2 + y_3}{2},
\end{aligned}$$

and they split it in 4 intervals (the outermost two being semi-infinite), where θ^o falls with equal probability. Thus, a reasonable choice is to choose $[\bar{\theta}_1, \bar{\theta}_3]$ as a 50%-confidence interval, where $\bar{\theta}_1 = \min(\theta_1, \theta_2, \theta_3)$ and $\bar{\theta}_3 = \max(\theta_1, \theta_2, \theta_3)$.

Let us tabulate y_i and θ_i for each possible value of ε_1, ε_2, and ε_3:

ε_1	ε_2	ε_3	y_1	y_2	y_3	θ_1	θ_2	θ_3	interval	contains θ^o?
1	1	1	2	2	2	2	2	2	[2, 2]	no
1	1	−1	2	2	0	2	1	1	[1, 2]	yes
1	−1	1	2	0	2	1	2	1	[1, 2]	yes
1	−1	−1	2	0	0	1	1	0	[0, 1]	yes
−1	1	1	0	2	2	1	1	2	[1, 2]	yes
−1	1	−1	0	2	0	1	0	1	[0, 1]	yes
−1	−1	1	0	0	2	0	1	1	[0, 1]	yes
−1	−1	−1	0	0	0	0	0	0	[0, 0]	no

(Of course, an "interval" like $[2, 2]$ means the set $\{2\}$.) As you can see, in 6 cases out of 8 the interval $[\bar\theta_1, \bar\theta_3]$ computed in the last-but-one column contains θ^o; the confidence of the interval $[\bar\theta_1, \bar\theta_3]$ is 75%, not 50% (thus, in this case the conclusions of the LSCR theory are *conservative*).

Problem 2 (discrete distribution, correct confidence).
The LSCR method works by considering 3 partial-average functions:

$$g_1(\theta) = \frac{y_1 + y_2}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_1 + \varepsilon_2}{2};$$
$$g_2(\theta) = \frac{y_1 + y_3}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_1 + \varepsilon_3}{2};$$
$$g_3(\theta) = \frac{y_2 + y_3}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_2 + \varepsilon_3}{2}.$$

Their respective intersections with the θ-axis are

$$\theta_1 = \frac{y_1 + y_2}{2},$$
$$\theta_2 = \frac{y_1 + y_3}{2},$$
$$\theta_3 = \frac{y_2 + y_3}{2},$$

and they split it in 4 intervals (the outermost two being semi-infinite), where θ^o falls with equal probability. Thus, a reasonable choice is to choose $[\bar\theta_1, \bar\theta_3]$ as a 50%-confidence interval, where $\bar\theta_1 = \min(\theta_1, \theta_2, \theta_3)$ and $\bar\theta_3 = \max(\theta_1, \theta_2, \theta_3)$.
We tabulate y_i and θ_i for each possible value of ε_1, ε_2, and ε_3 (all the θ_i are written as multiples of $\frac{1}{4}$ for ease of reading):

ε_1	ε_2	ε_3	y_1	y_2	y_3	θ_1	θ_2	θ_3	interval	contains θ^o?
1	1/2	2	2	3/2	3	7/4	10/4	9/4	[7/4, 10/4]	no
1	1/2	−2	2	3/2	−1	7/4	2/4	1/4	[1/4, 7/4]	yes
1	−1/2	2	2	1/2	3	5/4	10/4	7/4	[5/4, 10/4]	no
1	−1/2	−2	2	1/2	−1	5/4	2/4	−1/4	[−1/4, 5/4]	yes
−1	1/2	2	0	3/2	3	3/4	6/4	9/4	[3/4, 9/4]	yes
−1	1/2	−2	0	3/2	−1	3/4	−2/4	1/4	[−2/4, 3/4]	no
−1	−1/2	2	0	1/2	3	1/4	6/4	7/4	[1/4, 7/4]	yes
−1	−1/2	−2	0	1/2	−1	1/4	−2/4	−1/4	[−2/4, 1/4]	no

As you can see, in 4 cases out of 8 the interval $[\bar{\theta}_1, \bar{\theta}_3]$ computed in the last-but-one column contains θ^o; thus, $[\bar{\theta}_1, \bar{\theta}_3]$ is indeed a 50%-confidence interval for θ^o. In this case, the result is exactly what the LSCR theory claims, despite the fact that the distributions are discrete (in this particular example, no two intersections θ_i coincide, whatever the values of $\varepsilon_1, \varepsilon_2, \varepsilon_3$).

2.5 Solutions to the exercises on Interval Predictor Models

Solution 1 (support constraints).
Any problem like the following:

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & f_1(\theta) \le \gamma \\ & \vdots \\ & f_N(\theta) \le \gamma \\ & \theta \in \mathbb{R}, \gamma \in \mathbb{R}, \end{aligned}$$

where each f_i is a convex parabola is feasible, because each f_i is defined over the whole of $\mathbb{R}$. Indeed, fix an arbitrary $\bar{\theta}$, and let

$$M = \max\{f_1(\bar{\theta}), \cdots, f_N(\bar{\theta})\}.$$

Then

$$f_i(\bar{\theta}) < M + 1$$

for all $i = 1, \cdots, N$; hence $(\bar{\theta}, M + 1)$ satisfies all the constraints, and the problem is feasible.
Now we consider the requested situations:

- The problem has no support constraints. This is possible. Example with 2 constraints: $f_1(\theta) = \theta^2$, $f_2(\theta) = 2\theta^2$.
- The problem has just one support constraint. This is possible. Example with 2 constraints: $f_1(\theta) = \theta^2$, $f_2(\theta) = \theta^2 + 1$ (the support constraint is of course $f_2(\theta) \le \gamma$).

- The problem has two support constraints. This is also possible. Example with 2 constraints: $f_1(\theta) = (\theta - 1)^2 = \theta^2 - 2\theta + 1$, $f_2(\theta) = (\theta + 1)^2 = \theta^2 + 2\theta + 1$.
- The problem has three support constraints. This is not possible, since the dimension of the problem is 2 (see the Theorem by Campi, Calafiore, Garatti).

References

[1] Walter Rudin, *Principles of mathematical analysis 3rd ed.*, McGraw-Hill, 1976.