# CO327 (2021Spring) Assignment 5

Lecturer: Andersen Ang

June 27, 2021

## 1 Introduction

The goals of this assignment are

- to familiarize yourself with calling library/solver of the MATLAB programming language, and
- to have a hands-on experience in a few simple machine learning / data science applications.

This assignment is about using LP/QP to solve some machine learning/data science problems, therefore the focus is not on obtaining the correct solution, but the knowledge and the process on "how to build up a computer program to solve problem". It is therefore possible that your solutions obtained at the end of your program do not seem to be correct, which is okay. The majority of the marking for this assignment is on how you solve the problem, but not on obtaining the correct solution. You get points as long as your whole program / method / logic are sort of correct.

Download the data file here: `https://angms.science/doc/teaching/CO327/a5data.zip`

**On the assignment and important remarks**

- Submission requirement: Several mfiles that are error-free (no bug, run-able) that solves the problems in the next sections, and a PDF file that explains these MATLAB code as well as how you approach these problems.
- PDF requirement: you have to use LaTeX to generate the PDF. If you submit handwritten document as your PDF, you get **zero points** for the PDF (but you will still get thee points that corresponds to the MATLAB files).
- Scores distribution: 70% goes to the MATLAB mfiles and 30% goes to the PDF.
- How to submit: submit your zipped file (including a single PDF and all the MATLAB m.files) to the course dropbox in WATERLOO LEARN.
- Assignment deadline: July-31 23:55 (Eastern daylight time).
- Ask for help when needed.
- Work on this assignment early. Do not start to work on it the day before the deadline.

# 2 Robust curve fitting (20 points)

**Introduction** This part refers to `Q2_main.m` and `Q2.mat` in the data file.

One of the task in data analysis is to find the pattern(s) in data. In this question you will perform a simple simple data analysis known as "Regression". More specifically, "Least Sqaures" and "Robust regression'.

**About the problem** You are given $n$ points $(x_i, y_i)$, $i = 1, 2, \ldots, n$ which you believe they are generated under the model $y = mx + c$. However, the data points are corrupted by noise and therefore the points are not exactly lying on a straight line.

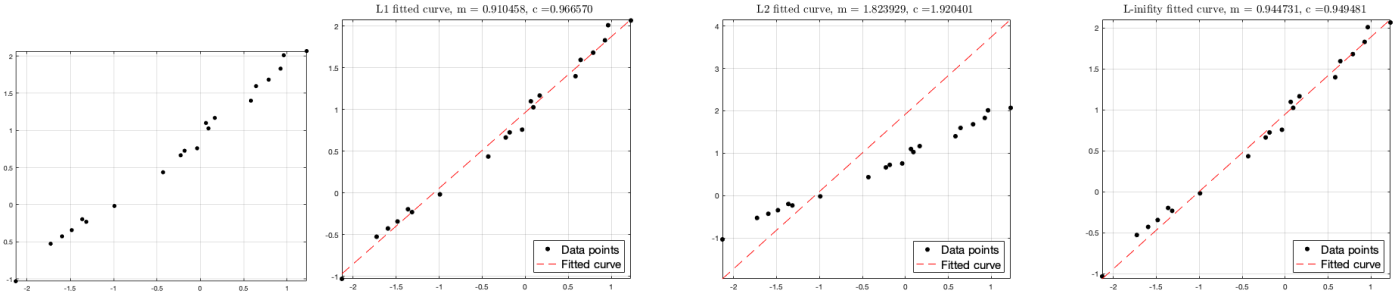



Figure 1: The input data and what you should get for different types of fitting.

For the data points, you propose the following model $y_i = mx_i + c + \epsilon_i$, $i = 1, 2, \ldots, n$ where $\epsilon$ refers to the noise. Refer to the lecture on curve fitting, a way to find $(m, c)$ is to minimize the norm $\mathbf{e} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]$. Mathematically, we have the following so-called "linear model":

$$\mathbf{Ax} = \mathbf{b} + \mathbf{e}, \quad \text{where} \quad \mathbf{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} c \\ m \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

We can see that minimizing the norm of $\mathbf{e}$ is equivalent to minimizing $\|\mathbf{Ax} - \mathbf{b}\|_p$ for $p \in \{1, 2, \infty\}$. Your task: write 3 MATLAB codes that minimizes this function for $p = 1, 2, \infty$. Especially,

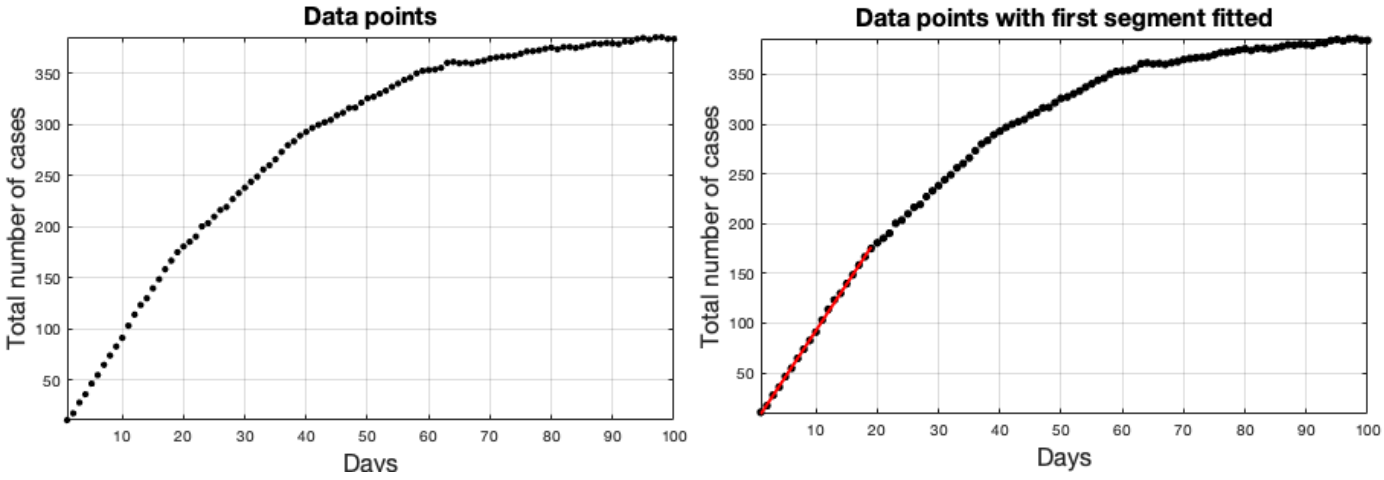
- For $p = 1$ or $\infty$, turn the problem as a LP and then solve it using `linprog`.
- For $p = 2$, consider minimizing $\|\mathbf{Ax} - \mathbf{b}\|_2^2$, turn the problem as a QP and then solve it using `quadprog`. Note that
$$\|\mathbf{Ax} - \mathbf{b}\|_2^2 = \mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} - 2\mathbf{b}^\top \mathbf{Ax} + \mathbf{b}^\top \mathbf{b}$$

After you obtain $(m, c)$, plot the result. For details, see `Q2_main.m` in the data file.

# 3 Change point detection (40 points)

This part refers to `Q3_main.m` in the data file.

You are given 100 days of record of total infected cases of an illness of a pandemic in a country. Mathematically, you are given 100 points $(x_i, y_i)$, where $x_i$ are integers from 1 to 100, representing the number of days since the outbreak of the pandemic, and $y_i$ refer to the number of total infected cases.




The data is suspected to be generated under a piecewise-linear model as

$$\mathbf{y} = \min\{m_1\mathbf{x} + c_1, m_2\mathbf{x} + c_2, \ldots, m_r\mathbf{x} + c_r\}$$

with $r$ mode (different $m_j, c_j$). Or in other words, you assume the data is generated by the model $y_i = m_j x_i + c_j$ with multiple sets of $(m_j, c_j)$ that correspond to different period of time (different range of $x_i$). Your tasks are:

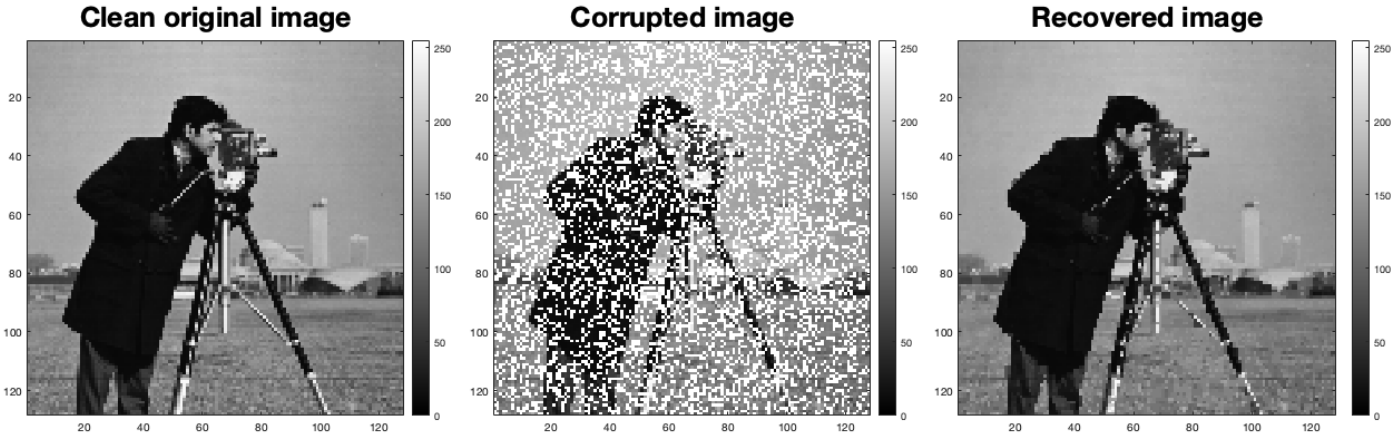
- Identify $r$, the number of mode in the data
- Identify these $(m, c)$
- Identify the change point (the day when the model changed).

**What to do** Write a function to perform a robust curve fitting on a set of points. then test on different segment of the data to identify each linear function. In other words, use the code you have developed in Q2. Here, instead of using all the points as in Q2, use a subset of points that you think they are belong to the same model.

**Example** The red line in the figure is obtained using the first 20 points in the data points. As the $\|\mathbf{A}_{\text{subset}}\mathbf{x}_{\text{subset}} - \mathbf{b}_{\text{subset}}\|_p$ is very low for these 20 points, hence we have high confidence that the first 20 points fall into the model of $m = 9.3238$ and $c = -0.67$. That is, we have $y = 9.3238x - 0.67$ for the period $1 \leq x_i \leq 20$.

# 4 Image completion (80 points)

In this question you are going to perform a "dark magic" in image processing called image inpainting (or mathematically called image completion). The task is simple: you are given a "broken image", the goal is to repair the image. Here an image (a $n$-by-$n$ matrix or a $n^2$ vectorized vector) contains numbers from 0 to 254, and 255 represents a broken pixel value.




Refer to the lecture on image inpainting, the mathematics of such "magic" is the following optimization problem

$$\min \ \|\mathbf{Ex}\|_1 \quad \text{s.t.} \quad \mathbf{Sx} = \hat{\mathbf{x}}_\Omega, \ \mathbf{x} \geq \mathbf{0}, \tag{*}$$

where $\mathbf{x} \in \mathbb{R}^{n^2}$ is the vectorized target image, $\|\mathbf{Ex}\|_1$ is called the "Total Variation" (TV) of the vector $\mathbf{x}$, the matrix $\mathbf{S} \in \mathbb{R}_+^{|\Omega| \times n^2}$ is a sub-matrix of $\mathbf{I}_{n^2}$ consists of rows labeled in the set $\Omega$, and $\hat{\mathbf{x}}_\Omega \in \mathbb{R}_+^{|\Omega|}$ is the clean part of the observed image, with $|\Omega| < n^2$ number of entries.

**LP** Given $\mathbf{E} \in \mathbb{R}^{2n(n-1) \times n^2}, \mathbf{S} \in \mathbb{R}_+^{|\Omega| \times n^2}$, and $\hat{\mathbf{x}}_\Omega \in \mathbb{R}_+^{|\Omega|}$, write a program to solve Problem (*). Plot the recovered image. See `Q4_main.m` in the data file for details.

**Relaxations to QP** Now instead of solving Problem (*), we consider solving the following problem:

$$\min \|\mathbf{Ex}\|_1 + \frac{1}{2}\|\mathbf{Sx} - \hat{\mathbf{x}}_\Omega\|_2^2 \ \text{s.t.} \ \mathbf{x} \geq \mathbf{0} \tag{**}$$

Write a program to solve Problem (**), plot the recovered image, and compare the recovered image to the one from solving Problem (*). Hint: Make use of `quadprog(Q,p,A,b,Aeq,beq,l)`.

**General formulation** Solve the following

$$\min \|\mathbf{Ex}\|_p + \|\mathbf{Sx} - \hat{\mathbf{x}}_\Omega\|_q \ \text{s.t.} \ \mathbf{x} \geq \mathbf{0} \tag{***}$$

for $p, q \in \{1, \infty\}$. Write a series of programs to solve Problem (***), plot these recovered images, compare the recovered image to the one from solving Problem (*) and (**).

**General hint** If you implemented everything correct, the recovered image should looks like the original clean image. There are 2 images in the data file, try with `mario.m` (the smaller one) first.

END of assignment.