Lecture 3: Least Squares

Shi Pu

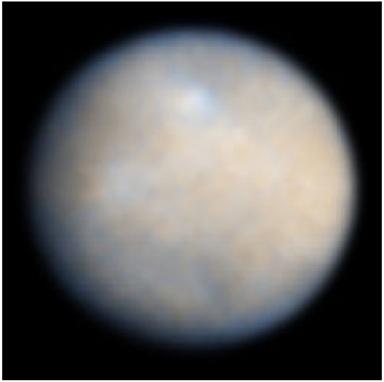
School of Data Science (SDS)
The Chinese University of Hong Kong, Shenzhen

Outline

Outline

- In January 1, 1801, an Italian monk Giuseppe Piazzi, discovered a faint, nomadic object through his telescope in Palermo, correctly believing it to reside in the orbital region between Mars and Jupiter.
- Piazzi watched the object for 41 days but then fell ill, and shortly thereafter the wandering star strayed into the halo of the Sun and was lost to observation.
- The newly-discovered planet had been lost, and astronomers had a mere 41 days of observation covering a tiny arc of the night from which to attempt to compute an orbit and find the planet again.

pages 1, 2 are from `http://www.keplersdiscovery.com/Asteroid.html`



- The dean of the French astrophysical establishment, Pierre-Simon Laplace (1749-1827), declared that it simply could not be done.
- In Germany, the 24 years old German mathematician Carl Friedrich Gauss had considered that this type of problem to determine a planet's orbit from a limited handful of observations - *"commended itself to mathematicians by its difficulty and elegance."*
- Gauss discovered a method for computing the planet's orbit using only three of the original observations and successfully predicted where Ceres might be found (now considered to be a dworf planet).
- The prediction catapulted him to worldwide acclaim.
- See `https://sites.math.rutgers.edu/~cherlin/History/Papers1999/weiss.html` for more details.

Formulation

- Consider the linear system

$$\mathbf{Ax} \approx \mathbf{b}, (\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^{m})$$

- Assumption: $\mathbf{A}$ has a full column rank, that is, $\text{rank}(\mathbf{A}) = n$.
- When $m > n$, the system is usually *inconsistent* and a common approach for finding an approximate solution is to pick the solution of the problem

$$\text{(LS)} \quad \min \|\mathbf{Ax} - \mathbf{b}\|^2.$$

The Least Squares Solution

- The LS problem is the same as

$$\min_{\mathbf{x}\in\mathbb{R}^n}\{f(\mathbf{x}) \equiv \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x} - 2\mathbf{b}^\top \mathbf{A}\mathbf{x} + \|\mathbf{b}\|^2\}.$$

- $\nabla^2 f(\mathbf{x}) = 2\mathbf{A}^\top \mathbf{A} \succ \mathbf{0}$
- Therefore, the unique optimal solution $\mathbf{x}_{\text{LS}}$ is the solution to $\nabla f(\mathbf{x}) = \mathbf{0}$, namely,

$$(\mathbf{A}^\top \mathbf{A})\mathbf{x}_{\text{LS}} = \mathbf{A}^\top \mathbf{b} \leftarrow \quad \text{normal equations}$$

- $\mathbf{x}_{\text{LS}} = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{b}$.

- Consider the inconsistent linear system

$$x_1 + 2x_2 = 0$$
$$2x_1 + x_2 = 1$$
$$3x_1 + 2x_2 = 1$$

- To find the least squares solution, solve the normal equations:

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 2 \end{pmatrix}^\top \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 2 \end{pmatrix}^\top \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix},$$

which is the same as

$$\begin{pmatrix} 14 & 10 \\ 10 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \end{pmatrix} \Rightarrow \mathbf{x}_{\mathrm{LS}} = \begin{pmatrix} 15/26 \\ -8/26 \end{pmatrix}.$$

- Note that $\mathbf{A}\mathbf{x}_{\mathrm{LS}} =$ (-0.038; 0.846; 1.115), so that the errors are $\mathbf{b} - \mathbf{A}\mathbf{x}_{\mathrm{LS}} = (0.0380.154 - 0.115)^\top \Rightarrow$ sq. err. $= 0.038^2 + 0.154^2 + (-0.115)^2 = 0.038.$

Linear Fitting:

- **Data:** $(\mathbf{s}_i, t_i), i = 1, 2, ..., m$, where $\mathbf{s}_i \in \mathbb{R}^n$ and $t_i \in \mathbb{R}$. Assume that an approximate linear relation holds:

$$t_i \approx \mathbf{s}_i^\top \mathbf{x}, i = 1, 2, ..., m$$

- The corresponding least squares problem is:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{m} (\mathbf{s}_i^\top \mathbf{x} - t_i)^2.$$

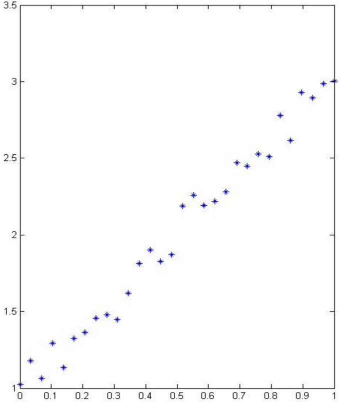
- Equivalent formulation:

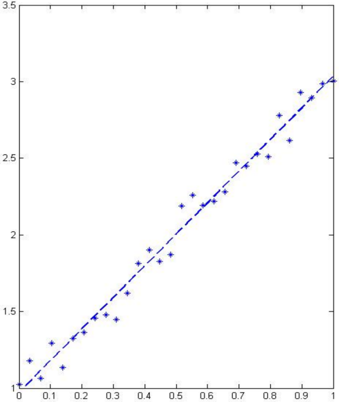
$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{S}\mathbf{x} - \mathbf{t}\|^2,$$

where

$$\mathbf{S} = \begin{pmatrix} ---\mathbf{s}_1^\top --- \\ ---\mathbf{s}_2^\top --- \\ \vdots \\ ---\mathbf{s}_m^\top --- \end{pmatrix}, \mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{pmatrix}.$$

Illustration





Example of Polynomial Fitting

- Given a set of points in $\mathbb{R}^2 : (u_i, y_i), i = 1, 2, ..., m$ for which the following approximate relation holds for some $a_0, ..., a_d$:

$$\sum_{j=0}^{d} a_j u_i^j \approx y_i, i = 1, 2, ..., m.$$

- The system is

$$\underbrace{\begin{pmatrix} 1 & u_1 & u_1^2 & \cdots & u_1^d \\ 1 & u_2 & u_2^2 & \cdots & u_2^d \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & u_m & u_m^2 & \cdots & u_m^d \end{pmatrix}}_{\mathbf{U}} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{pmatrix}.$$

- The least squares solution is of course well defined if the $m \times (d+1)$ matrix is of full column rank ($m \geq d+1$).
- This is true when all the u_i's are different from each other (why?)

Outline

Regularized Least Squares

- There are several situations in which the least squares solution does not give rise to a good estimate of the "true" vector $\mathbf{x}$.
- In these cases, a regularized problem (called regularized least squares (RLS)) is often solved:

$$\text{(RLS)} \quad \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda R(\mathbf{x}).$$

  Here λ is the *regularization parameter* and $R(\cdot)$ is the *regularization function* (also called a *penalty function*).
- *Quadratic regularization* is a specific choice of regularization function:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda \|\mathbf{D}\mathbf{x}\|^2.$$

- There are several situations in which the least squares solution does not give rise to a good estimate of the "true" vector $\mathbf{x}$.
- In these cases, a regularized problem (called regularized least squares (RLS)) is often solved:

$$\text{(RLS)} \quad \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda R(\mathbf{x}).$$

  Here λ is the *regularization parameter* and $R(\cdot)$ is the *regularization function* (also called a *penalty function*).
- *Quadratic regularization* is a specific choice of regularization function:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda \|\mathbf{D}\mathbf{x}\|^2.$$

- The optimal solution of the above problem is

$$\mathbf{x}_{\mathrm{RLS}} = (\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{D}^\top \mathbf{D})^{-1} \mathbf{A}^\top \mathbf{b}.$$

  How to assure that $\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{D}^\top \mathbf{D}$ is invertible? (answer: $\mathrm{Null}(\mathbf{A}) \cap \mathrm{Null}(\mathbf{D}) = \{\mathbf{0}\}$)

Application - Denoising

- Suppose that a noisy measurement of a signal $\mathbf{x} \in \mathbb{R}^n$ is given:

$$\mathbf{b} = \mathbf{x} + \mathbf{w}.$$

  $\mathbf{x}$ is the unknown signal, $\mathbf{w}$ is the unknown noise and $\mathbf{b}$ is the (known) measures vector.

- The least squares problem:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{b}\|^2.$$

  MEANINGLESS.

- Regularization is performed by exploiting some a priori information. For example, if the signal is "smooth" in some sense, then $R(\cdot)$ can be chosen as

$$R(\mathbf{x}) = \sum_{i=1}^{n-1} (x_i - x_{i+1})^2.$$

- $R(\cdot)$ can also be written as $R(\mathbf{x}) = \|\mathbf{L}\mathbf{x}\|^2$ where $\mathbf{L} \in \mathbb{R}^{(n-1)\times n}$ is given by

$$\mathbf{L} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}.$$

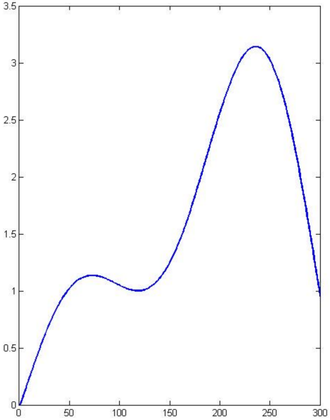
- The resulting regularized least squares problem is

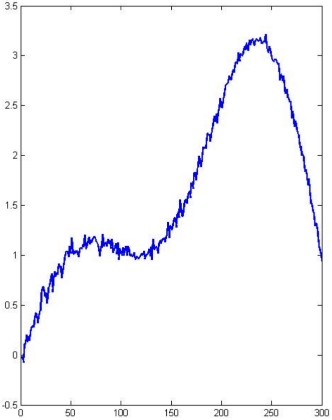
$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{b}\|^2 + \lambda \|\mathbf{L}\mathbf{x}\|^2$$

- Hence,

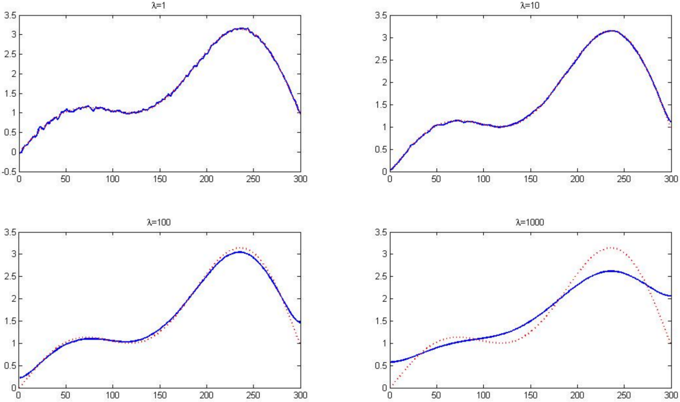
$$\mathbf{x}_{\mathrm{RLS}}(\lambda) = (\mathbf{I} + \lambda \mathbf{L}^\top \mathbf{L})^{-1} \mathbf{b}.$$

Example - true and noisy signals





RLS reconstructions



Outline

- The least squares problem $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ is often called *linear least squares*.
- In some applications we are given a set of nonlinear equations:

$$f_i(\mathbf{x}) \approx b_i, i = 1, 2, ..., m.$$

- The *nonlinear least squares (NLS) problem* is the one of finding an $\mathbf{x}$ solving the problem

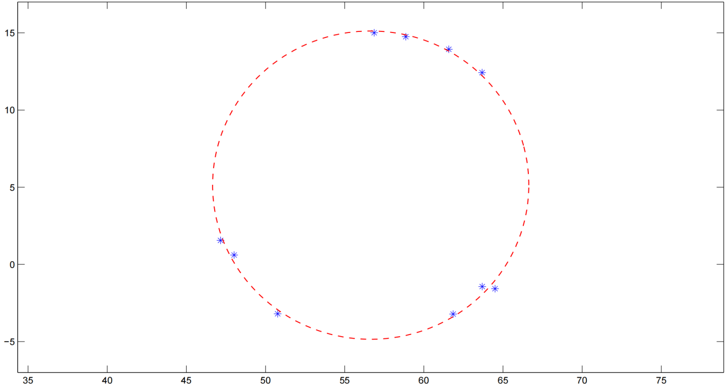
$$\min_{\mathbf{x}} \sum_{i=1}^{m} (f_i(\mathbf{x}) - b_i)^2.$$

- As opposed to linear least squares, no easy way to to solve NLS problems. However, there are some dedicated algorithms for this problem, which we will explore later on.

Given m points $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m \in \mathbb{R}^n$, the *circle fitting problem* seeks to find a circle

$$C(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{x}\| = r\}$$

that best fits the m points.



- Approximate equations:

$$\|\mathbf{x} - \mathbf{a}_i\| \approx r, \quad i = 1, 2, ..., m.$$

- To avoid nondifferentiability, consider the squared version:

$$\|\mathbf{x} - \mathbf{a}_i\|^2 \approx r^2, \quad i = 1, 2, ..., m.$$

- Nonlinear least squares formulation:

$$\min_{\mathbf{x}\in\mathbb{R}^n, r\in\mathbb{R}_+} \sum_{i=1}^{m} \left(\|\mathbf{x} - \mathbf{a}_i\|^2 - r^2\right)^2.$$

- $$\min_{\mathbf{x},r}\{\sum_{i=1}^{m}(-2\mathbf{a}_i^\top\mathbf{x}+\|\mathbf{x}\|^2-r^2+\|\mathbf{a}_i\|^2)^2:\mathbf{x}\in\mathbb{R}^n, r\in\mathbb{R}\}.$$

- Making the change of variables $R=\|\mathbf{x}\|^2-r^2$, the above problem reduces to

$$\min_{\mathbf{x}\in\mathbb{R}^n,R\in\mathbb{R}}\{f(\mathbf{x},R)\equiv\sum_{i=1}^{m}(-2\mathbf{a}_i^\top\mathbf{x}+R+\|\mathbf{a}_i\|^2)^2:\|\mathbf{x}\|^2\geq R\}.$$

- The constraint $\|\mathbf{x}\|^2\geq R$ can be dropped (will be shown soon), and therefore the problem is equivalent to the LS problem

$$\text{(CF-LS)}\ \min_{\mathbf{x},R}\{\sum_{i=1}^{m}(-2\mathbf{a}_i^\top\mathbf{x}+R+\|\mathbf{a}_i\|^2)^2:\mathbf{x}\in\mathbb{R}^n, R\in\mathbb{R}\}.$$

Redundancy of the Constraint $\|\mathbf{x}\|^2 \geq R$

- We will show that any optimal solution $(\hat{\mathbf{x}}, \hat{R})$ of (CF-LS) automatically satisfies $\|\hat{\mathbf{x}}\|^2 \geq \hat{R}$.
- Otherwise, if $\|\hat{\mathbf{x}}\|^2 < \hat{R}$, then

$$-2\mathbf{a}_i^\top \hat{\mathbf{x}} + \hat{R} + \|\mathbf{a}_i\|^2 > -2\mathbf{a}_i^\top \hat{\mathbf{x}} + \|\hat{\mathbf{x}}\|^2 + \|\mathbf{a}_i\|^2 = \|\hat{\mathbf{x}} - \mathbf{a}_i\|^2 \geq 0, i = 1, \dots, m.$$

- Thus,

$$f(\hat{\mathbf{x}}, \hat{R}) = \sum_{i=1}^{m} \left(-2\mathbf{a}_i^\top \hat{\mathbf{x}} + \hat{R} + \|\mathbf{a}_i\|^2\right)^2$$

$$> \sum_{i=1}^{m} \left(-2\mathbf{a}_i^\top \hat{\mathbf{x}} + \|\hat{\mathbf{x}}\|^2 + \|\mathbf{a}_i\|^2\right)^2 = f(\hat{\mathbf{x}}, \|\hat{\mathbf{x}}\|^2),$$

- Contradiction to the optimality of $(\hat{\mathbf{x}}, \hat{R})$.