MA580H Matrix Computations

Least-Squares Problem(LSP)

Lecture 11: Method of Normal Equation

Rafikul Alam
Department of Mathematics
IIT Guwahati

Outline

- Least squares problem
- Method of Normal Equation
- Tikhonov regularization

Overdetermined linear systems

- Overdetermined system: We have considered a linear system $Ax = b$ only when A is a square matrix. We now consider an $m \times n$ linear system $Ax = b$ when $m > n$. This is called an overdetermined linear system.

- Underdetermined system: The complementary underdetermined case $m < n$ turns up less frequently and will NOT be discussed.

- Least-squares solution: An overdetermined system is mostly inconsistent and hence does not have a solution. So, we need to define what is a best possible solution. There are multiple useful options but we consider only the least squares solution in which 2-norm of the residual vector is minimized.

Linear and nonlinear least squares problems

Problem statement: Given data points $(x_1, y_1), \ldots, (x_m, y_m)$, determine a function $f(x, \alpha)$ that best fit the data, where $\alpha := [\alpha_1, \ldots, \alpha_n]^\top$ is a parameter vector that solves the minimization problem

$$\min_{\alpha} \sum_{j=1}^{m} (f(x_j, \alpha) - y_j)^2.$$

Linear least-squares problem: In this case $f(x, \alpha)$ is linear with respect to α, that is, $f(x, \alpha) := \alpha_1\phi_1(x) + \cdots + \alpha_n\phi_n(x)$, where $\phi_1(x), \ldots, \phi_n(x)$ are given model functions.

For example, $f(x, \alpha) := \alpha_1 + \alpha_2 x$ leads to a linear regression problem (straightline fit).

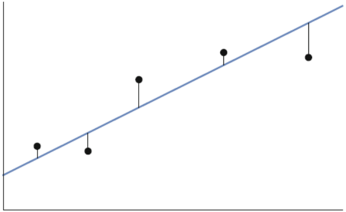
Nonlinear least-squares problem: In this case $f(x, \alpha)$ is nonlinear with respect to α.

For example, $f(x, \alpha) := \dfrac{1}{1 + e^{-(\alpha_1 x + \alpha_2)}}$ leads to a logistic regression problem.

Certain nonlinear least squares problem can be transformed to linear least squares problem by change of variables. For example, $f(x, \alpha) := \alpha_1 e^{\alpha_2 x} \implies \log f(x, \alpha) = \log \alpha_1 + \alpha_2 x = c_1 + c_2 x.$

Linear regression

Given data points $(t_1, b_1), \ldots, (t_m, b_m)$ in $\mathbb{R}^2$, find a straight line $f(t, \alpha) := \alpha_1 + \alpha_2 t$ that best fit the data. The task is to minimize the error $\sum_{j=1}^{m}(f(t_j, \alpha) - b_j)^2$ for all $\alpha_1, \alpha_2 \in \mathbb{R}$.



Setting $r_i := f(t_i, \alpha) - b_i \Longrightarrow f(t_i, \alpha) = b_i + r_i \Longrightarrow \alpha_1 + \alpha_2 t_i = b_i + r_i$ for $i = 1 : m$. This yields the LSP

$$\begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \approx \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}.$$

Linear regression

Now consider the LSP

$$Ax = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \approx \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = b.$$

Set $\mu_t := (t_1 + \cdots + t_m)/m, \sigma_t^2 := (t_1^2 + \cdots + t_m^2)/m, \mu_b := (b_1 + \cdots + b_m)/m$ and $\sigma_{tb} := (t_1 b_1 + \cdots + t_m b_m)/m.$ Then the normal equation $A^\top Ax = A^\top b$ gives

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_m \end{bmatrix} \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_m \end{bmatrix} b \Longrightarrow \begin{bmatrix} 1 & \mu_t \\ \mu_t & \sigma_t^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \mu_b \\ \sigma_{tb} \end{bmatrix}.$$

Hence $\alpha_1 = (\mu_b \sigma_t^2 - \mu_t \sigma_{tb})/(\sigma_t^2 - \mu_t^2)$ and $\alpha_2 = (\sigma_{tb} - \mu_t \mu_b)/(\sigma_t^2 - \mu_t^2).$

The best fit is given by the line $y = \beta(t - \mu_t) + \mu_b,$ where $\beta = (\sigma_{tb} - \mu_t \mu_b)/(\sigma_t^2 - \mu_t^2).$

Linear fit for temperature data

Data for the 5-year temperature averages.

```
year = (1955:5:2000)';
y = [ -0.0480; -0.0180; -0.0360; -0.0120; -0.0040;
       0.1180; 0.2100; 0.3320; 0.3340; 0.4560 ];

t = (year - 1955) / 10;    % better matrix conditioning later
A = [ t.^0 t ];            % coefficient matrix
c = A \ y;
f = @(year) polyval(c(end:-1:1), (year - 1955) / 10);

clf
scatter(year, y), axis tight
xlabel('year'), ylabel('anomaly ({\circ}C)')
hold on
fplot(f, [1955, 2000])
```

Linear fit for temperature data

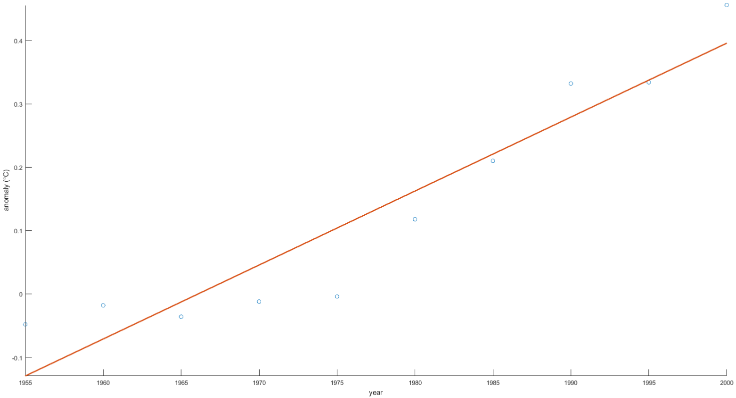


Figure : St-line fit for 5-year average temperature data points.

Polynomial data fitting problem

For $(n-1)$ degree polynomial $p(t) = x_1 + x_2 t + \cdots + x_n t^{n-1}$ fitting the data $(t_1, b_1), \ldots, (t_m, b_m)$, we have $p(t_i) = b_i + r_i$ for $i = 1 : m$. This yields the LSP

$$\begin{bmatrix} 1 & t_1 & \cdots & t_1^{n-1} \\ 1 & t_2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & t_m & \cdots & t_m^{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & t_1 & \cdots & t_1^{n-1} \\ 1 & t_2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & t_m & \cdots & t_m^{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \approx \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}.$$

In practice, one considers $n = 2, 3, 4$ which correspond to straight-line, quadratic and cubic polynomial fit, respectively.

The matrix in the LSP has full rank. Hence the LSP $Ax \approx b$ has a unique solution which is obtained by solving the normal equation

$$A^\top Ax = A^\top b.$$

However, for large n the matrix becomes highly ill-conditioned.

Example

Consider the data points

t	-1.0	-0.5	0.0	0.5	0.1
b	1.0	0.5	0.0	1.5	2.0

For the quadratic polynomial fit, we have the LSP

$$\begin{bmatrix} 1 & -1.0 & 1.0 \\ 1 & -0.5 & 0.25 \\ 1 & 0.0 & 0.0 \\ 1 & 0.5 & 0.25 \\ 1 & 0.1 & 1.0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \approx \begin{bmatrix} 1.0 \\ 0.5 \\ 0.0 \\ 1.5 \\ 2.0 \end{bmatrix}.$$

Solving the LSP we have $x = \begin{bmatrix} 0.086 & 0.40 & 1.4 \end{bmatrix}^\top$ which yields the polynomial $p(t) = 0.086 + 0.4t + 1.4t^2$.

Example

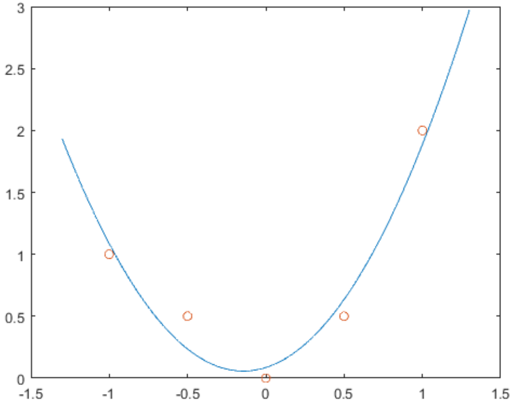


Figure : The plot of $p(t) = 0.086 + 0.4t + 1.4t^2$ and the data points.

Linear and cubic polynomial fit for temperature data

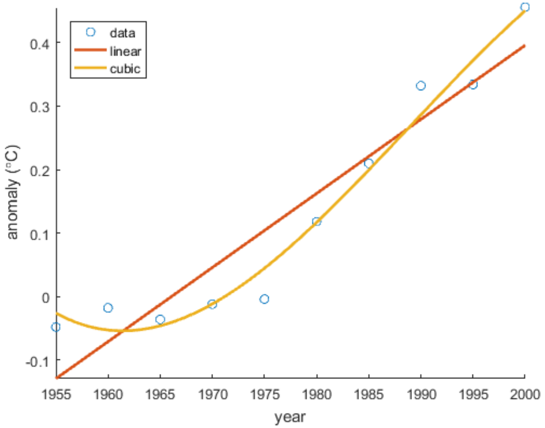


Figure : Linear and cubic fit for 5-year average temperature data points.

Linear parameterized model

Consider the data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ in $\mathbb{R}^n \times \mathbb{R}$ and the regression model given by

$$f(\mathbf{x}, \alpha, v) = \mathbf{x}^\top \alpha + v = \alpha_1 \phi_1(\mathbf{x}) + \cdots + \alpha_n \phi_n(\mathbf{x}) + v$$

where $\phi_j(\mathbf{x}) = \mathbf{e}_j^\top \mathbf{x} = x_j$ for $j = 1 : n$.

- The components of $\mathbf{x}$ are called regressors.
- The regressor model is parameterized by the weight vector α and the intercept v.
- The prediction error $\mathbf{r} \in \mathbb{R}^n$ is given by $r_j = f(\mathbf{x}_j, \alpha, v) - y_j = \mathbf{x}_j^\top \alpha + v - y_j$ for $j = 1 : n$.

The minimization of $\|\mathbf{r}\|_2$ yields the LSP

$$\begin{bmatrix} 1 & \phi_1(\mathbf{x}_1) & \cdots & \phi_n(\mathbf{x}_1) \\ 1 & \phi_1(\mathbf{x}_2) & \cdots & \phi_n(\mathbf{x}_2) \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \phi_1(\mathbf{x}_m) & \cdots & \phi_n(\mathbf{x}_m) \end{bmatrix} \begin{bmatrix} v \\ \alpha \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{x}_1^\top \\ 1 & \mathbf{x}_2^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_m^\top \end{bmatrix} \begin{bmatrix} v \\ \alpha \end{bmatrix} \approx \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Least-squares classification

A data fitting problem where the outcome $y = g(\mathbf{x})$ can take only two values 1 and -1 gives a classification algorithm. The values of y represent two categories and $y = g(\mathbf{x})$ is called Boolean classifier.

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ be training data in $\mathbb{R}^n \times \mathbb{R}$. Determine the regressor model

$$f(\mathbf{x}, \alpha, v) = \mathbf{x}^\top \alpha + v$$

that best fit the data.

Define

$$g(\mathbf{x}) := \mathrm{sign}(f(\mathbf{x}, \alpha, v)) = \begin{cases} +1 & \text{if } f(\mathbf{x}, \alpha, v) \geq 0 \\ -1 & \text{if } f(\mathbf{x}, \alpha, v) < 0 \end{cases}$$

Then $y = g(\mathbf{x})$ is a least-squares Boolean classifier.

Least-squares problem (LSP)

Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. Usually $m \gg n$. Find $x \in \mathbb{C}^n$ that minimizes

$$\|Ax - b\|_2^2 = \sum_{i=1}^{m} \Big| \Big(\sum_{j=1}^{n} a_{ij} x_j - b_i \Big) \Big|^2.$$

This is called least-squares problem because we minimize the sum of the squares of the errors

$$|r_1|^2 + \cdots + |r_m|^2 \text{ where } r := Ax - b.$$

The vector $r := Ax - b$ is called residual vector and $\|r\|_2$ is called residual error of the least squares problem. We write a solution x of the LSP as

$$x = \arg\min_{y \in \mathbb{C}^n} \|Ay - b\|_2.$$

The LSP is called a linear least squares problem and is written as

$$\text{solve } Ax \approx b \text{ or } \text{LSP } Ax \approx b.$$

Remark: If x is a solution of the LSP $Ax \approx b$ then so is $x + z$ for any $z \in N(A)$.

Normal equation

If $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, then define

$$f(x) := \|Ax - b\|_2^2 = \sum_{i=1}^{m} \big(\sum_{j=1}^{n} a_{ij}x_j - b_i\big)^2.$$

Then gradient $\nabla f(x) = 2A^\top(Ax - b)$ and Hessian $H_f(x) = A^\top A$. For a minimum $\nabla f(x) = 0$ yields the normal equation $A^\top Ax = A^\top b$. Since Hessian is symmetric positive semidefinite, f has a minimum at x.

Note that $A^\top A$ is positive semidefinite and the normal equation

$$A^\top Ax = A^\top b$$

is always consistent and has a solution. If $\mathrm{rank}(A) = n$ then $A^\top A$ is positive definite and $x = (A^\top A)^{-1}A^\top b = A^+ b$ is a unique solution of the LSP $Ax \approx b$.

Remark: If $\mathrm{rank}(A) = n$ then $A^\top Ax = A^\top b$ can be solved by Cholesky factorization. However, $A^\top A$ may be highly ill-conditioned.

Geometry of Least-squares problem

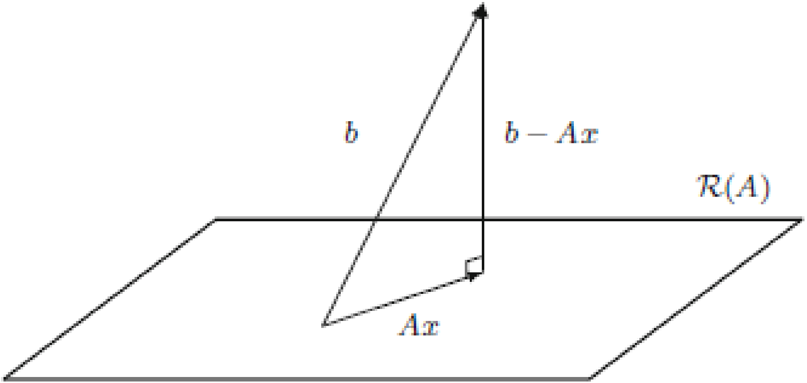


Figure : Relationships among b, r and $R(A)$.

Normal equation

Consider $R(A) := \{Ax : x \in \mathbb{C}^n\} \subset \mathbb{C}^m$ and $N(A) := \{x \in \mathbb{C}^n : Ax = 0\} \subset \mathbb{C}^n$ of $A \in \mathbb{C}^{m\times n}$. Then we have

$$\begin{aligned} \mathbb{C}^m &= R(A) \oplus N(A^*) \text{ and } R(A) \perp N(A^*) \\ \mathbb{C}^n &= N(A) \oplus R(A^*) \text{ and } N(A) \perp R(A^*). \end{aligned}$$

Theorem: The LSP $Ax = b$ has a solution and

$$x = \underset{y\in\mathbb{C}^n}{\operatorname{argmin}} \|Ay - b\|_2 \iff (Ax - b) \perp R(A) \iff A^*Ax = A^*b.$$

Proof: $(Ax - b) \perp R(A) \iff b - Ax \in N(A^*) \iff A^*(Ax - b) = 0 \iff A^*Ax = A^*b.$

Let $b = b_1 + b_2$ with $b_1 \in R(A)$ and $b_2 \in N(A^*)$. Then

$$\begin{aligned} \|Ax - b\|_2^2 &= \|Ax - b_1 - b_2\|_2^2 = \|Ax - b_1\|_2^2 + \|b_2\|_2^2 = \|b_2\|_2^2 \\ &\Leftrightarrow Ax = b_1 \iff b_2 = b - Ax \iff (Ax - b) \perp R(A). \blacksquare \end{aligned}$$

The system $A^*Ax = A^*b$ is called the normal equation for $Ax \approx b$.

Regularized Least-squares problem

Linear measurement model estimation: $y = Ax + v$

- x is a n-vector containing parameter that we wish to estimate
- v is an m-vector containing unknown measurement error or noise
- y is an m-vector containing measurement and A is an $m \times n$ matrix

Regularized LSP (Tikhonov regularization): $\min_x \left(\|Ax - y\|_2^2 + \lambda \|Dx\|_2^2\right)$ where $\lambda > 0$.

The goal is to make both $\|Ax - y\|_2$ and $\|Dx\|_2$ small. The solution is unique when $\mathrm{rank}(D) = n$. Indeed,

$$\|Ax - y\|_2^2 + \lambda \|Dx\|_2^2 = \left\| \begin{bmatrix} A \\ \sqrt{\lambda} D \end{bmatrix} x - \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|_2^2 \Longrightarrow (A^*A + \lambda D^*D)x = A^*y.$$

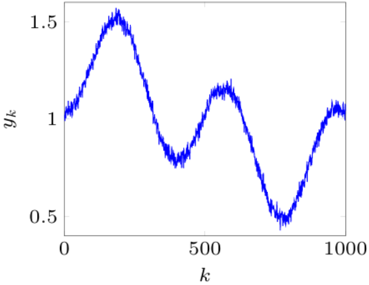
This shows that $x = (A^*A + \lambda D^*D)^{-1} A^*y$ is the unique solution of the regularized LSP.

Signal denoising

- observed signal is n-vector

$$y = x + v$$

- x is unknown signal
- v is noise



Least squares denoising

$$\text{minimize} \quad \|x - y\|^2 + \lambda \sum_{i=1}^{n-1} (x_i - x_{i-1})^2$$

goal is to find slowly varying signal $\hat{x}$, close to observed signal y

Matrix formulation

$$\text{minimize} \quad \left\| \begin{bmatrix} I \\ \sqrt{\lambda} D \end{bmatrix} x - \begin{bmatrix} y \\ 0 \end{bmatrix} \right\|^2$$

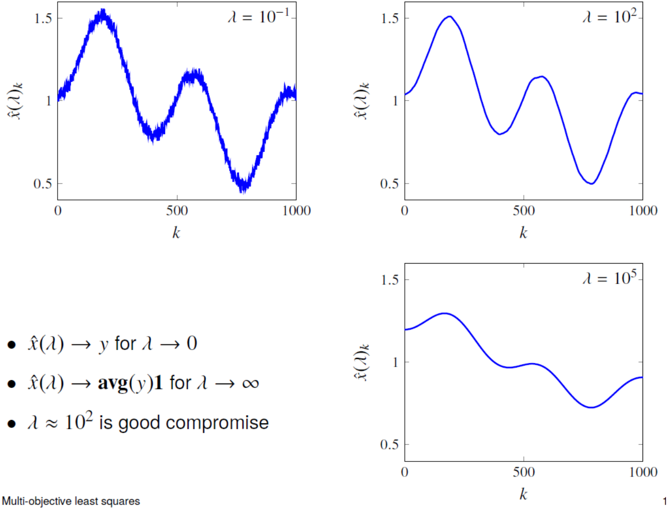
- D is $(n-1) \times n$ finite difference matrix

$$D = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix}$$

- equivalent to linear equation

$$(I + \lambda D^T D)x = y$$

Three solutions



- $\hat{x}(\lambda) \rightarrow y$ for $\lambda \rightarrow 0$
- $\hat{x}(\lambda) \rightarrow \mathbf{avg}(y)\mathbf{1}$ for $\lambda \rightarrow \infty$
- $\lambda \approx 10^2$ is good compromise