Least Squares

Thao Tran Le
Student at Hanoi University of Science and Technology
Email: thao.tlp200604@sis.hust.edu.vn

In this report, I look at the powerful idea of finding approximate solutions of over-determined systems of linear equations by minimizing the sum of the squares of the errors in the equations.

From t.tranaflee with love ...

1. Least squares problem

Definition.
Suppose that the $m \times n$ matrix A is tall, so the system of linear equation $Ax = b$, where b is an m-vector, is over-determined.
For most choices of b, there is no n-vector x for which $Ax = b$.
Instead we seek an x for which $r = Ax - b$, which is the residual for the equations $Ax = b$ is as small as possible.
This suggests that we should choose x so as to minimize the sum of squares of the residuals

$$\text{minimize } \|Ax - b\|^2$$

Least squares problem is the problem of finding an n-vector $\hat{x}$ that minimize $\|Ax - b\|^2$. The quantity to be minimized $\|Ax - b\|^2$ is called the *objective function* of the least squares problem.
Any vector $\hat{x}$ that satisfies $\|A\hat{x} - b\|^2 \leq \|Ax - b\|^2$ for all x is a solution of the least squares problem. Such a vector is called a *least squares approximate solution* of $Ax = b$.
Column interpretation.
Suppose $a_1, ..., a_n$ are columns of A, then

$$\|Ax - b\|^2 = \|(x_1 a_1 + ... + x_n a_n) - b\|^2$$

If $\hat{x}$ is a solution of the least squares problem, then the vector

$$A\hat{x} = \hat{x}_1 a_1 + ... + \hat{x}_n a_n$$

is closest to the vector b, among all linear combinations o the vectors $a_1, ..., a_n$.
Row interpretation.
Suppose $\tilde{a_1}^T, ..., \tilde{a_m}^T$ are the rows of A, then the residual components are given by

$$r_i = \tilde{a_i}^T x - b_i \; i = 1, ..., m.$$

The least squares objective is then

$$\|Ax - b\|^2 = (\tilde{a_1}^T x - b_1)^2 + ... + (\tilde{a_m}^T x - b_m)^2$$

Minimizing this sum of squares of the residuals is a reasonable compromise if least squares attempts to make them all small.
Example.
We consider the least squares problem with data

$$A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

The over-determined set of three equations in two variables $Ax = b$,

$$2x_1 = 1, \qquad -x_1 + x_2 = 0, \qquad 2x_2 = -1$$

has no solution. The least squares problem is to choose x to minimize

$$\|Ax - b\|^2 = (2x_1 - 1)^2 + (-x_1 + x_2)^2 + (2x_2 + 1)^2$$

Then via calculus, it can solve the least squares approximate solution $\hat{x} = (1/3, 1/3)$, and this solution does not satisfy $Ax = b$ since the corresponding residuals are

$$\hat{r} = A\hat{x} - b = (-1/3, -2/3, 1/3)$$

with sum of squares value $\|A\hat{x} - b\|^2 = 2/3$.

2. Solutions

In this section, we derive several expressions for the solution of the least squares problem, under one assumption on the data matrix A: *The column of A are linearly independent.*

Derivation via calculus.
Rewriting the least squares objective out as a sum, we get

$$f(x) = \|Ax - b\|^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} x_j - b_j \right)^2$$

The minimizer $\hat{x}$ of the function $f(x) = \|Ax - b\|^2$ must satisfy

$$\frac{\partial f}{\partial x_i}(\hat{x}) = 0, \qquad i = 1, ..., n.$$

which we can express as the vector equation

$$\nabla f(\hat{x}) = 0$$

Any minimizer $\hat{x}$ of $\|Ax - b\|^2$ must satisfy

$$\nabla f(\hat{x}) = 2A^T(A\hat{x} - b) = 0$$

which can be written as

$$A^T A\hat{x} = A^T b$$

Our assumption that the columns of A are linearly independent implies that the Gram matrix $A^T A$ is invertible. This implies that

$$\hat{x} = (A^T A)^{-1} A^T b$$

This must be the unique solution of the least squares problem.

Direct verification.
Let $\hat{x} = (A^T A)^{-1} A^T b$, then $A^T(A\hat{x} - b) = 0$.
For any n-vector x we have

$$\begin{aligned} \|Ax - b\|^2 &= \|(Ax - A\hat{x}) + (A\hat{x} - b)\|^2 \\ &= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 + 2(A(x - \hat{x}))^T(A\hat{x} - b) \\ &= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 + 2(x - \hat{x})^T A^T(A\hat{x} - b) \\ &= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 \end{aligned} \tag{1}$$

So for any x, we obtain that $\|Ax - b\|^2 \geq \|A\hat{x} - b\|^2$.
If equality holds, $A(x - \hat{x}) = 0$, which implies $x = \hat{x}$ since columns of A are linearly independent.

Computing least squares approximate solutions.
Let $A = QR$ be the QR factorization of A (which exists since the columns of A are linearly independent). [$2mn^2$ flops]
We have that the pseudo-inverse $A^* = (A^T A)^{-1} A^T$ can be expressed as $A^* = R^{-1}Q^T$, then

$$\hat{x} = R^{-1}Q^T b$$

To compute $\hat{x}$,
(1) form $Q^T b$ [$2mn$ flops]
(2) compute $\hat{x} = R^{-1}(Q^T b)$ via back substitution [n^2 flops]
The total complexity is $2mn^2 + 2mn + n^2$ flops.
Remarks: *This computation is identical to algorithm for solving $Ax = b$ for square invertible A. But when A is tall, it gives least squares approximate solution.*

3. Application in Advertising Purchases

Problem.
We have m demographic groups we want to advertise tom, with m-vector of target views or impressions, v^{des}. To reach these audiences, we purchase advertising in n different channels in amounts that we give as an n-vector s.

Example.
Considering a simple numerical example, with $n = 3$ channels and $m = 10$ demographic groups, and matrix

$$R = \begin{bmatrix} 0.97 & 1.86 & 0.41 \\ 1.23 & 2.18 & 0.53 \\ 0.80 & 1.24 & 0.62 \\ 1.29 & 0.98 & 0.51 \\ 1.10 & 1.23 & 0.69 \\ 0.67 & 0.34 & 0.54 \\ 0.87 & 0.26 & 0.62 \\ 1.10 & 0.16 & 0.48 \\ 1.92 & 0.22 & 0.71 \\ 1.29 & 0.12 & 0.62 \end{bmatrix}$$

with units od 1000 views per dollar. The entries of R range over an 18:1 range, so the 3 channels are quite different in terms of their audience reach, see figure **3.1**.
We take $v^{des} = [10^3]_{10\times1}$, this means our goal is to reach one million views in each of the 10 demographic groups.
Least squares gives the advertising budget allocation,

$$\hat{s} = (62, 100, 1443)$$

which achieves a views vector with RMS error 132 ($RMS = \|v^{des} - Rs\|/\sqrt{m}$).
The views vector is shown in figure **3.2**.

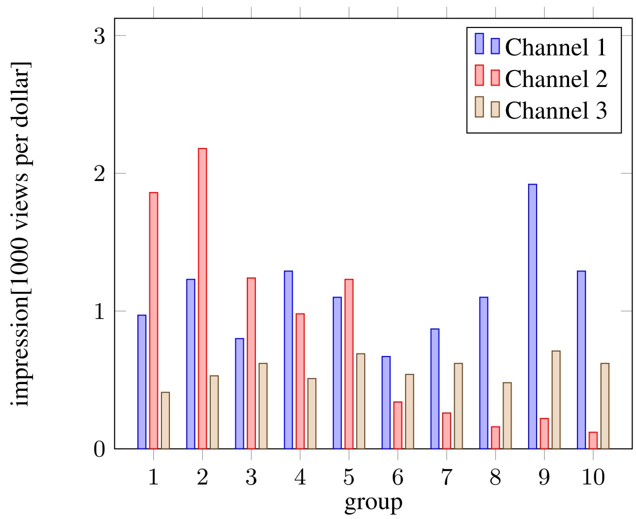

Figure 3.1. Number of impressions in 10 demographic groups, per dollar spent on advertising in three channels.

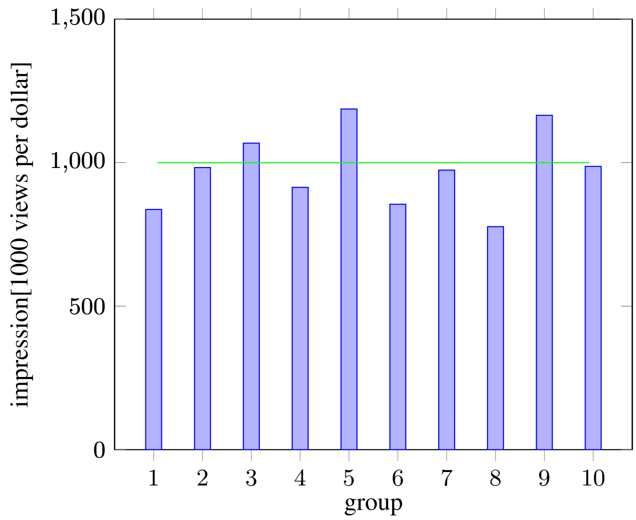

Figure 3.2. Views vector that best approximates the target of one million impressions in each group.

References

[1] Gilbert Strang. *Introduction to Linear Algebra. 5th Edition* 2016. Printed in the United States of America.

[2] Stephen Boyd, Lieven Vandenberghe. *Introduction to Applied Linear Algebra. 1st Edition* 2018. Printed in Cambridge University Press.