Chapter 6.1

Linear Least Squares Problems

Uri M. Ascher and Chen Greif
Department of Computer Science
The University of British Columbia
{ascher,greif}@cs.ubc.ca

Slides for the book
A First Course in Numerical Methods (published by SIAM, 2011)
`http://bookstore.siam.org/cs07/`

Some slides are from lecture notes of Dr. Peter Arbenz, ETH.

Goal: introduce and solve linear least squares problem, ubiqiutous in data fitting applications

- We discuss how to solve overdetermined linear systems of equations

$$A\mathbf{x} \approx \mathbf{b}, \qquad \mathbf{b} \in \mathbb{R}^m,\ \mathbf{x} \in \mathbb{R}^n, \quad m > n.$$

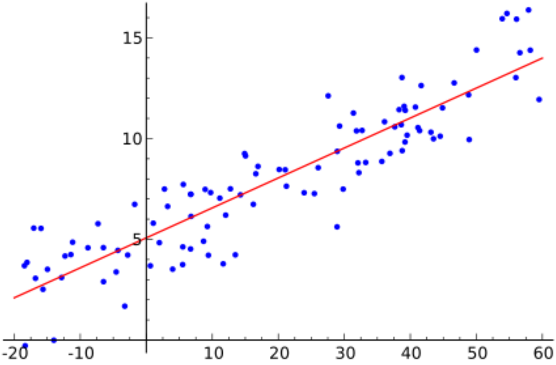
- These systems do in general not have a solution.

Reference

- Ascher–Greif, Chapter 6.

Origins of linear least squares problems

Data fitting



http://en.wikipedia.org/wiki/Least_squares

Linear least-squares

- Throughout this chapter we consider the problem

$$\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|_2,$$

where A is $m \times n$, with $m > n$.

- So, it is an overdetermined system of equations: we have more rows, for instance corresponding to data measurements, than columns, where $\mathbf{x}$ corresponds to unknown model parameters.
- In general, there is no $\mathbf{x}$ satisfying $A\mathbf{x} = \mathbf{b}$, hence we seek to minimize a norm of the residual $\mathbf{r} = \mathbf{b} - A\mathbf{x}$. The ℓ_2 norm is the most convenient to work with, although it is not suitable for all purposes, and it enjoys rich theory.
- Assume A has linearly independent columns. Then there is a unique solution to this problem, as we'll soon see.

Normal equations

- Drop the index 2: $\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|$.
- Equivalent to minimizing
  $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{b} - A\mathbf{x}\|^2 = \frac{1}{2}\sum_{i=1}^{m}\left(b_i - \sum_{j=1}^{n} a_{ij}x_j\right)^2$.
- Necessary conditions: $\frac{\partial}{\partial x_k}\psi(\mathbf{x}) = 0, \quad k = 1, \ldots, n$.
- So,
$$\sum_{i=1}^{m}\left[\left(b_i - \sum_{j=1}^{n} a_{ij}x_j\right)(-a_{ik})\right] = 0.$$
- In matrix-vector form this expression looks much simpler:
$$A^T A\mathbf{x} = A^T\mathbf{b}.$$
- Also *sufficient* for minimum because $\nabla^2\psi = A^T A$ is positive definite.

$$\mathbf{r} = \mathbf{b} - A\mathbf{x}$$

$$B = A^T A$$

Normal equations algorithm

$$\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|.$$

- Assume A has linearly independent columns. Then for an optimum it is necessary and sufficient to satisfy the **normal equations**

$$A^T A\mathbf{x} = A^T\mathbf{b}.$$

B Use LU decomp on B to solve

- Simple, efficient, classical.

Example

- Consider the least-squares problem $\min_{\mathbf{x}} \|\mathbf{b} - A\mathbf{x}\|$ for

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 3 & 5 \\ 5 & 3 & -2 \\ 3 & 5 & 4 \\ -1 & 6 & 3 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 4 \\ -2 \\ 5 \\ -2 \\ 1 \end{pmatrix}.$$

- Solving via normal equations: form

$$B = A^T A = \begin{pmatrix} 40 & 30 & 10 \\ 30 & 79 & 47 \\ 10 & 47 & 55 \end{pmatrix}, \quad \mathbf{y} = A^T\mathbf{b} = \begin{pmatrix} 18 \\ 5 \\ -21 \end{pmatrix};$$

solve $B\mathbf{x} = \mathbf{y}$ obtaining $\mathbf{x} = (.3472, .3990, -.7859)^T$.

- The optimal residual (rounded) is

$$\mathbf{r} = \mathbf{b} - A\mathbf{x} = (4.4387, .0381, .495, -1.893, 1.311)^T.$$

This vector is orthogonal to each column of A.

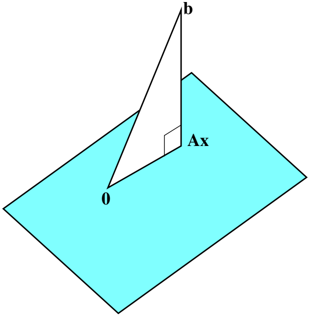
Geometrical interpretation

From

$$\mathbf{grad}\,\psi(\mathbf{x}) = A^T A\mathbf{x} - A^T\mathbf{b} = A^T(A\mathbf{x} - \mathbf{b}) = \mathbf{0}$$

we see that $A^T\mathbf{r} = \mathbf{0}$.

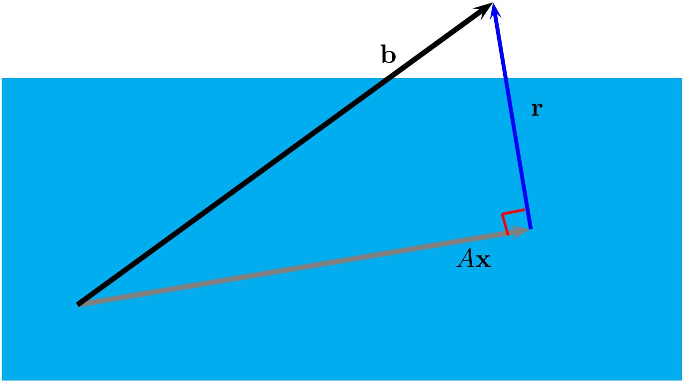
The blue plane shows $\mathcal{R}(A)$. $A\mathbf{x}$ is the orthogonal projection of $\mathbf{b}$ onto $\mathcal{R}(A)$.



Orthogonality of the residual

A^T (b-Ax) = A^T r = 0

Hence the residual is orthogonal to the column space of A.



Theorem: Least squares

The least squares problem

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2,$$

where A has full column rank, has a unique solution that satisfies the normal equations

$$(A^T A)\mathbf{x} = A^T \mathbf{b}.$$

We have $\mathbf{x} = \left(A^T A\right)^{-1} A^T \mathbf{b}$. The matrix multiplying $\mathbf{b}$ is called the pseudo-inverse of A:

$$A^+ = (A^T A)^{-1} A^T \in \mathbb{R}^{n \times m}.$$

Normal equations facts

- The residual vector $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ is *orthogonal* to the columns of A: $A^T\mathbf{r} = \mathbf{0}$.
- Thus, $\mathbf{b}$ is orthogonally projected to the space range(A).
- Define pseudo-inverse of A by

$$A^{\dagger} = B^{-1}A^T.$$

- For $m \gg n$, most of the algorithm cost is in the formation of $B = A^TA$.
- This is the way to solve many data fitting problems.
- But, difficulties arise when A has (almost) linearly dependent columns.

In MATLAB: backslash \ operator does a least squares fit if the matrix is m x n, m > n.

Outline

- Normal equations
- Application: data fitting

Data fitting

Generally, data fitting problems arise as follows:

- We have *observed data* $\mathbf{b}$ and a *model function* that for any candidate model $\mathbf{x}$ provides *predicted data*.
- The task is to find $\mathbf{x}$ such that the predicted data match the observed data to the extent possible.
- We want to minimize the difference of predicted and observed data in the least squares sense.
- Here, we study the linear case where predicted data are given by $A\mathbf{x}$.
  (The condition that A has maximal rank means that there is no redundancy in the representation of the predicted data.)

Example 6.2 from Ascher–Greif: Linear regression

Consider fitting a given data set of m pairs (t_i, b_i) by a straight line:

$$v(t) = x_1 + x_2 t \qquad \Longrightarrow \qquad v(t_i) \approx b_i, \quad i = 1, \dots, m.$$

$$A = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix} \qquad B = \begin{pmatrix} m & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & \sum_{i=1}^m t_i^2 \end{pmatrix}$$

i	1	2	3
t_i	0.0	1.0	2.0
b_i	0.1	0.9	2.0

$$\Longrightarrow \qquad \mathbf{x} = \begin{pmatrix} 0.05 \\ 0.95 \end{pmatrix}$$

Example: linear regression

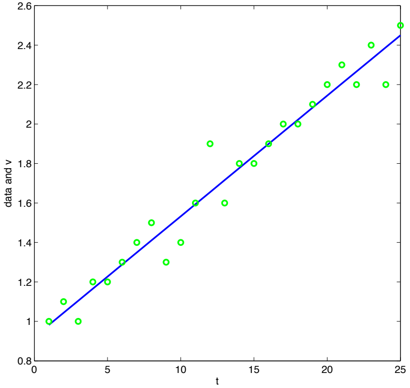


FIGURE : Linear regression curve (in blue) through green data points. Here $m = 25$ and $n = 2$.