Tony Gustafsson, August 19, 1999

Least Squares from a Linear Algebra Point of View

Problem Formulation

Suppose that we have collected measurements of a signal $y(n)$ for $n = 0, \cdots N-1$. Collect the measurements in a vector $\mathbf{Y}$:

$$\mathbf{Y} = \begin{bmatrix} y(0) \\ \vdots \\ y(N-1) \end{bmatrix}. \tag{1}$$

Furthermore, assume that we *model* the vector $\mathbf{Y}$ as

$$\mathbf{Y} = \mathbf{S}(\boldsymbol{\theta}) + \mathbf{E}, \tag{2}$$

where $\mathbf{E}$ denotes a noise vector, and $\mathbf{S}(\boldsymbol{\theta})$ denotes our model signal vector. Herein, consider the case of a linear relationship: $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta}$, where $\mathbf{A}$ is a *known* $N \times n$ matrix, and $\boldsymbol{\theta}$ is an unknown $n \times 1$ vector, $N \geq n$. The problem treated in these notes is that of estimating $\boldsymbol{\theta}$, given the measurement of $\mathbf{Y}$, and the knowledge of $\mathbf{A}$.

Example 1 Suppose that the signal $y(n)$ looks like the signal in Figure 1. Almost as if there is a straight line going through the measurements? Thus, a reasonable model is

$$y(n) = kn + m + e(n), n = 0, \cdots, N-1. \tag{3}$$

Then we can formulate a matrix description as

$$\underbrace{\begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(N-1) \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & N-1 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} m \\ k \end{bmatrix}}_{\boldsymbol{\theta}} + \underbrace{\begin{bmatrix} e(0) \\ e(1) \\ \vdots \\ e(N-1) \end{bmatrix}}_{\mathbf{E}}. \tag{4}$$

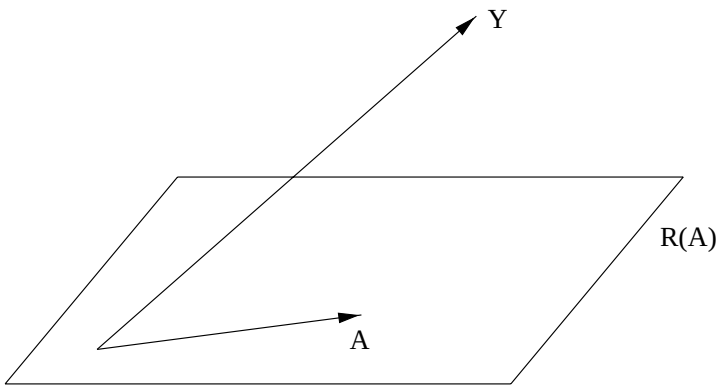

Figure 1: An example of a signal $y(n)$.

Exercise 1 In various applications, polynomial interpolation (curve fitting!) is common. That is, model the measured signal as

$$y(n) = b_0 + b_1 n + \cdots b_p n^p + e(n), \ n = 0, \cdots, N-1. \tag{5}$$

Let $\boldsymbol{\theta} = [b_0 \ b_1 \cdots b_p]^T$. Specify the matrix $\mathbf{A}$ in the matrix description $\mathbf{Y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{E}$.

The Least Squares (LS) Solution

To be able to really understand the LS solution, we have to refresh some linear algebra theory. For our present case, perhaps the most important term is the *range space* of a matrix, denoted $\mathcal{R}(\cdot)$. The definition of the range space of an $m \times n$ matrix $\mathbf{A}$ is:

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{y} : \ \mathbf{y} = \mathbf{A}\mathbf{x}, \ \forall \ \mathbf{x} \in \mathbb{R}^n\} \tag{6}$$

Try to think of this as follows: a) The matrix $\mathbf{A}$ is a linear mapping that takes the vector $\mathbf{x}$ from a point in the n-dimensional space, to a point in the m-dimensional space, b) Compute $\mathbf{y} = \mathbf{A}\mathbf{x}$ for all possible vectors $\mathbf{x}$, and remember the corresponding $\mathbf{y}$'s. c) The range space of $\mathbf{A}$ is defined as the part of the m-dimensional space that is filled by the computed $\mathbf{y}$'s. Too abstract? Try the following exercise:

Exercise 2 Draw the range space of

1. $\mathbf{A} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.
2. $\mathbf{A} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.
3. $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
4. $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.
5. $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$.

Let us return to our original problem. Then we note that the model signal $\mathbf{S}(\boldsymbol{\theta})$ is constrained to lie in the range space of $\mathbf{A}$: $\mathbf{S}(\boldsymbol{\theta}) \in \mathcal{R}(\mathbf{A})$, see Figure 2. Due to measurement noise

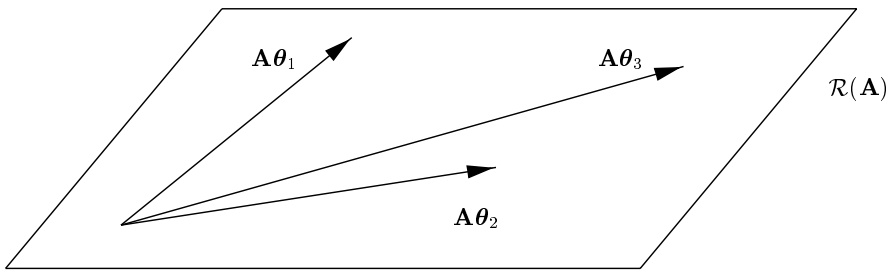

Figure 2: Illustration of $\mathcal{R}(\mathbf{A})$.

and model imperfections, the vector $\mathbf{Y}$ will typically not lie in $\mathcal{R}(\mathbf{A})$, see Figure 3. As a consequence, there will typically **NOT** exist an exact solution to the equation $\mathbf{Y} = \mathbf{A}\boldsymbol{\theta}$. Accepting the fact that there generally (at least when $N > n$, i.e an overdetermined system of equations) is no solution to $\mathbf{Y} = \mathbf{A}\boldsymbol{\theta}$, it is natural to look for approximative solutions. One popular way to do this is to apply the LS principle:
Find $\hat{\boldsymbol{\theta}}$ such that the Euclidean distance between $\mathbf{Y}$ and $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta}$ is minimized!

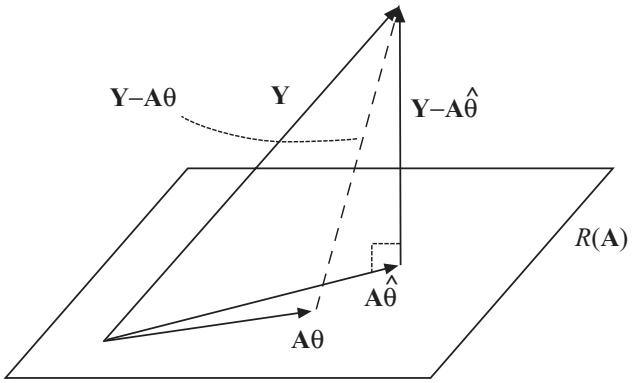

Figure 3: Illustration of $\mathbf{Y} \notin \mathcal{R}(\mathbf{A})$.

Exercise 3 Let $\mathbf{Y} = [1\ 1]^T$, and $\mathbf{A} = [1\ 0]^T$, see Figure 4. How would you select θ such that the length of $\mathbf{E} = \mathbf{Y} - \mathbf{A}\theta$ is minimized? We are now getting close to the LS solution.

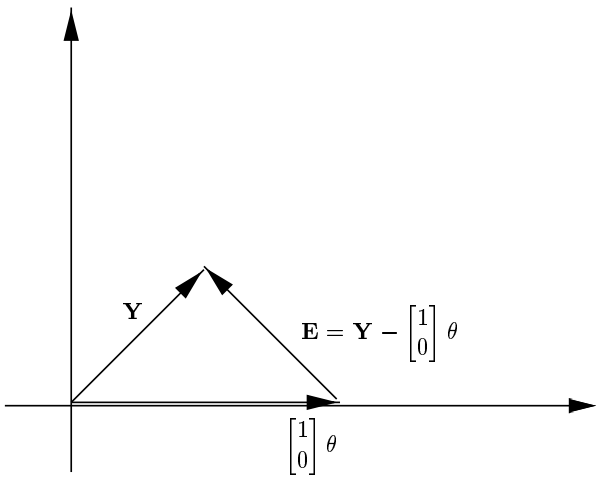

Figure 4: Illustration of Example

Define the LS solution of $\boldsymbol{\theta}$ as the value of $\boldsymbol{\theta}$ that minimizes $\|\mathbf{Y} - \mathbf{A}\boldsymbol{\theta}\|^2$ and denote it with $\hat{\boldsymbol{\theta}}$.

Exercise 4 Assuming that $\mathbf{A}$ is full rank ($\mathbf{A}^T\mathbf{A}$ invertible), show that

$$\begin{aligned} \|\mathbf{Y} - \mathbf{A}\boldsymbol{\theta}\|^2 &= \left(\boldsymbol{\theta} - (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}\right)^T \mathbf{A}^T\mathbf{A}\left(\boldsymbol{\theta} - (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}\right) \\ &+ \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}. \end{aligned} \tag{7}$$

Hint: $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x}$.
Note, only the first term of (7) depends on $\boldsymbol{\theta}$ and this term is clearly greater than or equal to zero. Thus, $\|\mathbf{Y} - \mathbf{A}\boldsymbol{\theta}\|^2$ is minimized if we let

$$\hat{\boldsymbol{\theta}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y} \tag{8}$$

which is the desired solution. At the same time we get a measure of how large the error is; the norm of the error vector is

$$\|\mathbf{Y} - \mathbf{A}\hat{\boldsymbol{\theta}}\|^2 = \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}. \tag{9}$$

Note, in the calculations above it is assumed that $\mathbf{A}^T\mathbf{A}$ is invertible, which is a necessary assumption for the above to be valid.

Remark 1 In the literature, the matrix $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is called the "pseudo-inverse"[1] of $\mathbf{A}$. Note, if $\mathbf{A}$ is square and invertible, the pseudo-inverse and the usual inverse coincide:

$$(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{A}^{-1}\mathbf{A}^{-T}\mathbf{A}^T = \mathbf{A}^{-1}. \tag{10}$$

Remark 2 In Matlab, the LS solution is computed as

```
>> TH=A\Y;
```

Remark 3 Consider the error vector

$$\mathbf{E} = \mathbf{Y} - \mathbf{A}\hat{\boldsymbol{\theta}} = \left(\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\right)\mathbf{Y}. \tag{11}$$

This error vector is orthogonal to $\mathbf{A}$:

$$\mathbf{A}^T\mathbf{E} = \mathbf{A}^T\left(\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\right)\mathbf{Y} = \mathbf{A}^T\mathbf{Y} - \mathbf{A}^T\mathbf{Y} = 0, \tag{12}$$

which may be interpreted as an example of the principle of orthogonality. Finally, compare this orthogonality relation with your answer to exercise !

Remark 4 In these notes we derived the LS solution as the minimizing argument of the criterion in Exercise . Another way to arrive at the presented LS solution, is to use the concept of projection matrices. This kind of matrices is discussed in accompanying notes on linear algebra.

Please run the m-file polyreg in Matlab.

[1] The pseudo-inverse is defined also when $\mathbf{A}$ is not full rank ($\mathbf{A}^T\mathbf{A}$ not invertible).

```
% polyreg.m
% to investigate the least squares approach the polynomial regressions.
clear

% generate a coefficient vector, generaty noisy data
degree=3;
sigma=0.2;
coeff=rand(1,degree+1);
x=-1:0.02:1;
truedata=coeff*[(x.^3)' (x.^2)' (x.^1)' (x.^0)']';
data=truedata+sigma*randn(1,length(truedata));

% estimate the coefficients in the least squares sense
coeffest=polyfit(x,data,degree);

% plot the result
figure(1)
clf
plot(x,truedata,'r', x,data,'g', x, coeffest*[(x.^3)' (x.^2)' (x.^1)' (x.^0)']','b')
hold on
title(' red-true, blue-estimate ')
hold off

% Maybe you will note, particularly if you increase the noise, that polynomial
% fitting is not a very stable method. You can do better if you choose other
% basis functions than x^0, x^1, x^2, ... . Try to obtain a set of orthogonal
% basis functions, such as Legendre polynomials if you insist on using powers.
% A choise noted for its stability is sines and cosines, a k a the Fourier transform.
```