

Feb 18: Least squares fitting via the normal equations

In simple least squares fitting of a set of observations to a linear function, or linear regression, what is assumed is that a set of observations y_i can be described by a “model”

$$\begin{aligned}y(t) &= \theta(t) + n(t) \\ &= a + bt + n(t)\end{aligned}$$

Here, $n(t)$ is the measurement noise and is the source of the misfit between the observations and the “model”.

We can write this as a matrix equation:

$$\mathbf{E}\mathbf{x} + \mathbf{n} = \mathbf{y}$$

where

$$\mathbf{E} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_M \end{pmatrix} \quad \mathbf{x} = \begin{pmatrix} a \\ b \end{pmatrix} \quad \mathbf{n} = \begin{pmatrix} n_1 \\ \vdots \\ n_M \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix}$$

Having zero error would be exceptional, so in general the parameters a and b will represent a best possible fit of the model to the observations.

We have many more data points than the 2 parameters a and b , so the problem is said to be “over determined”.

Typically, the measure of best fit is the parameter choice that minimizes the mean squared misfit of model and data.

$$\begin{aligned}\min J &= \sum_{i=1}^M n_i^2 = \mathbf{n}^T \mathbf{n} = (\mathbf{E}\mathbf{x} - \mathbf{y})^T (\mathbf{E}\mathbf{x} - \mathbf{y}) \\ &= (\mathbf{E}\mathbf{x})^T (\mathbf{E}\mathbf{x}) - \mathbf{y}^T \mathbf{E}\mathbf{x} - (\mathbf{E}\mathbf{x})^T \mathbf{y} + \mathbf{y}^T \mathbf{y}\end{aligned}$$

Each of these terms is a scalar, so each is its own transpose.

$$\text{So, } \mathbf{y}^T \mathbf{E}\mathbf{x} = (\mathbf{y}^T \mathbf{E}\mathbf{x})^T = (\mathbf{E}\mathbf{x})^T \mathbf{y}$$

$$\text{Also, } (\mathbf{E}\mathbf{x})^T \mathbf{y} = \mathbf{x}^T \mathbf{E}^T \mathbf{y} \quad \text{so we have}$$

$$J = \mathbf{x}^T \mathbf{E}^T \mathbf{E} \mathbf{x} - 2\mathbf{x}^T \mathbf{E}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}$$

To minimize, we differentiate with respect to \mathbf{x} and set to zero, anticipating a minimum.

$$\frac{\partial J}{\partial \mathbf{x}} = 2(\mathbf{E}^T \mathbf{E}) \mathbf{x} - 2\mathbf{E}^T \mathbf{y} = 0$$

This leads to the set of normal equations

$$(\mathbf{E}^T \mathbf{E}) \mathbf{x} = \mathbf{E}^T \mathbf{y}$$

Assuming the inverse of the normal equations matrix exists, the solution is

$$\mathbf{x} = (\mathbf{E}^T \mathbf{E})^{-1} \mathbf{E}^T \mathbf{y}$$

No assumptions have been made about the statistical probability density functions of the errors n_i .

Let's give some consideration to how the estimated parameters of the model fit, a and b denoted by \mathbf{x} , are affected by the random elements of the observations.

Assume the estimated values are unbiased, then the $\langle \mathbf{x} \rangle = \langle \mathbf{x}^{\text{true}} \rangle$ (expected values).

The uncertainty in the estimated values is described by their variance about the true mean.

$$\begin{aligned} P &= \langle \langle \mathbf{x} - \mathbf{x}^{\text{true}} \rangle \langle \mathbf{x} - \mathbf{x}^{\text{true}} \rangle^T \rangle \\ &= (\mathbf{E}^T \mathbf{E})^{-1} \mathbf{E}^T \langle \mathbf{n} \mathbf{n}^T \rangle \mathbf{E} (\mathbf{E}^T \mathbf{E})^{-1} \end{aligned}$$

In the special case that we have uncorrelated errors, $\langle \mathbf{n} \mathbf{n}^T \rangle = \sigma_n^2 \mathbf{I}$, i.e. all the observations are known with an uncertainty $\pm \sigma_n$

So the uncertainty in the parameter estimates is $P = \sigma_n^2 (\mathbf{E}^T \mathbf{E})^{-1}$

and the uncertainty in the estimated derived from these parameters is

$$\begin{aligned} \mathbf{y}^{\text{est}} &= \mathbf{E}\mathbf{x} \\ \mathbf{y}^{\text{est}} - \mathbf{y} &= \mathbf{n} \\ P_{mm} &= (\mathbf{n}^{\text{est}} - \mathbf{n})(\mathbf{n}^{\text{est}} - \mathbf{n})^T = \sigma_n^2 (\mathbf{I} - \mathbf{E}(\mathbf{E}^T \mathbf{E})^{-1} \mathbf{E}^T) \end{aligned}$$

(see Wunsch section 3.3 for details).

You should, at the very least, examine the residuals of the “model” fit compared to the data to see if they are randomly distributed.

Weighted least squares and solutions via Singular Value Decomposition

An example Matlab session to demonstrate least squares fitting using the normal equations, singular value decomposition, and matrix left divide is given in the m-file `jw_lecture_least_squares.m`.

The least squares minimization problem described by

$$\min J = \sum_{i=1}^M n_i^2 = \mathbf{n}^T \mathbf{n} = (\mathbf{E}\mathbf{x} - \mathbf{y})^T (\mathbf{E}\mathbf{x} - \mathbf{y})$$

had the solution, via the normal equations, of $\mathbf{x} = (\mathbf{E}^T \mathbf{E})^{-1} \mathbf{E}^T \mathbf{y}$

and can be thought of as the “best” solution \mathbf{x} (in a least squares sense) to the over-determined set of equations:

$$\mathbf{E}\mathbf{x} = \mathbf{y}$$

We can find the least squares best fit solution to this equation by Singular Value Decomposition:

Suppose that a Singular Value Decomposition of matrix \mathbf{E} exists:

$$\mathbf{E} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

$$M \times N : [M \times N][N \times N][N \times N]$$

where \mathbf{U} and \mathbf{V} are orthonormal matrices (each column orthogonal to the others).

\mathbf{S} is diagonal.

In Matlab:

```
>> [U,S,V] = svd(E,0);
```

The least squares solution to:

$$\mathbf{E}\mathbf{a} = \mathbf{d}$$

follows from

$$\begin{aligned}\mathbf{E}\mathbf{a} &= \mathbf{d} \\ (\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T)(\mathbf{U}^T\mathbf{S}\mathbf{V}^T)\mathbf{a} &= (\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T)\mathbf{d} \\ \mathbf{a} &= (\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T)\mathbf{d}\end{aligned}$$

The inversion \mathbf{S}^{-1} is straightforward because \mathbf{S} is diagonal.

For an over-determined system such as we have in fitting a model described by a small number of parameters to a large number of data, the *Matrix Left Divide* in Matlab finds this solution directly:

```
>> a=E\d;
```

Wunsch chapter 3.4 gives a thorough description of how the SVD works.

Its relation to the solution by normal equations can be seen by direct substitution.

Wunsch points out that the SVD can be computed for over-determined and under-determined systems, and when $(\mathbf{E}^T\mathbf{E})^{-1}$ does not exist (the design matrix is singular).