Lecture 4: Worked Example

We saw in Lecture 1 that there are two kinds of uncertainties to consider when fitting functions to data. The first, σ_ν, comes from noise in the experimental measurements, when repeated experiments give different outcomes. This error is usually expressed by associating a *constant* error bar with all predictions: $y \pm \sigma_\nu$.

The second type of error which comes from the fitting uncertainty is not constant. This is illustrated in Fig. 1; there are many functions which can be fitted or extrapolated into uncertain regions of the input space, without unduly compromising the fit in adjacent regions which are rich in accurate data. Instead of calculating a unique set of weights, a probability distribution of sets of weights is used to define the fitting uncertainty. The error bars therefore become large when data are sparse or locally noisy.

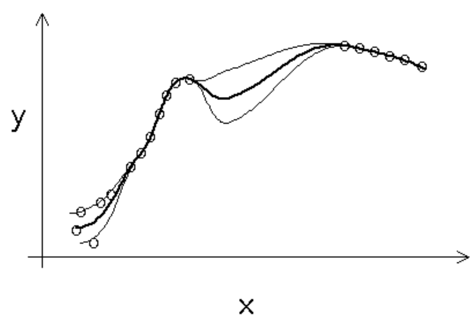

Fig. 1: Fitting uncertainties due to sparse or noisy data.

Example

Determine the best–fit straight line for the following data, and the fitting uncertainty associated with each datum and for values of inputs which lie beyond the given data.

$$\mathbf{T} = \begin{pmatrix} -2.8 \\ -0.9 \\ 0.3 \\ -0.2 \\ 2.2 \\ 2.8 \\ 4.2 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} -3 & 1 \\ -2 & 1 \\ -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix} \equiv \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \\ \mathbf{x}_6 \\ \mathbf{x}_7 \end{pmatrix}$$

The parameters α and β are in practice determined iteratively by minimising the function M (Lecture 3, equation 5). For the purposes of this exercise you may assume that $\alpha = 0.25$ and that there is a noise $\sigma_\nu = 0.026861$ so that $\beta = 1/\sigma_\nu^2 = 1386$.

Solution

The best–fit intercept on the y–axis is shown by linear regression to be $w_2 = 1.0$; the best–fit slope is similarly found to be $w_1 = 1.0821$ (Fig. 2) †

The mean vector is

$$\overline{\mathbf{x}}^T = (\,\overline{x}_1, \overline{x}_2\,) = (\,0 \quad 1\,)$$

---

† This assumes that the prior is flat, that is that the lines plotted in Fig. 9b are completely randomly distributed. If the prior belief implies a non–random distribution then most likely line will not correspond to that given by best–fitting in this way.

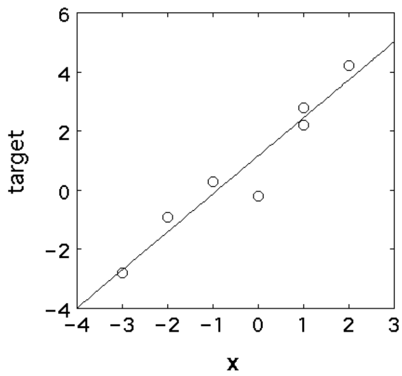

Fig. 2: Plot of target versus x.

The number of data is $n = 7$ so that

$$\sum_n \mathbf{x}\mathbf{x}^T = \sum_n \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{pmatrix} x_1 & x_2 \end{pmatrix}$$
$$= \sum_{i=1}^{7} \begin{pmatrix} (x_1^{(i)})^2 & x_1^{(i)} x_2^{(i)} \\ x_2^{(i)} x_1^{(i)} & (x_2^{(i)})^2 \end{pmatrix}$$
$$= \begin{pmatrix} 28 & 0 \\ 0 & 7 \end{pmatrix}$$

We need to determine the variance–covariance matrix $\mathbf{V}$ which is given by (Lecture 3, equation 5):

$$\mathbf{V}^{-1} = \left[\alpha \mathbf{I} + \beta \sum_n \mathbf{x}\mathbf{x}^T\right]$$
$$= \left[\begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{pmatrix} + \begin{pmatrix} 38808 & 0 \\ 0 & 9702 \end{pmatrix}\right]$$
$$= \begin{pmatrix} 38808.25 & 0 \\ 0 & 9702.25 \end{pmatrix}$$

It follows that

$$\mathbf{V} = \begin{pmatrix} 2.57677 \times 10^{-5} & 0 \\ 0 & 1.03069 \times 10^{-4} \end{pmatrix}$$

Suppose now that we wish to determine the fitting error associated with $\mathbf{x}^T = (\,-7 \quad 1\,)$, then the corresponding variance for that input vector is

$$\begin{aligned}\sigma_y^2 &= \mathbf{x}^T\mathbf{V}\mathbf{x} \\ &= (\,-7 \quad 1\,)\begin{pmatrix} 2.57677\times10^{-5} & 0 \\ 0 & 1.03069\times10^{-4}\end{pmatrix}\begin{pmatrix}-7\\1\end{pmatrix} \\ &= 0.001365687\end{aligned}$$

Therefore, for $\mathbf{x} = (\,-7 \quad 0.8\,)$, $y = -6.7747$, $\sigma_y^2 = 0.001365687$, $\sigma_y = 0.037$ so the prediction may be stated, with 67% confidence to be

$$y \pm \sigma_y = -6.7747 \pm 0.037.$$

Calculations like these done for a variety of input vectors are listed in Table 1 and plotted in Fig. 3.

x_1	x_2	y	σ_y^2	σ_y
-7	1	-6.7747	0.001365687	0.036955203
-3	1	-2.4463	0.000334978	0.018302413
-2	1	-1.3642	0.00020614	0.014357567
-1	1	-0.2821	0.000128837	0.011350621
0	1	0.8	0.000103069	0.010152284
1	1	1.8821	0.000128837	0.011350621
1	1	1.8821	0.000128837	0.011350621
2	1	2.9642	0.00020614	0.014357567
7	1	8.3747	0.001365687	0.036955203

Table 1: Some predictions

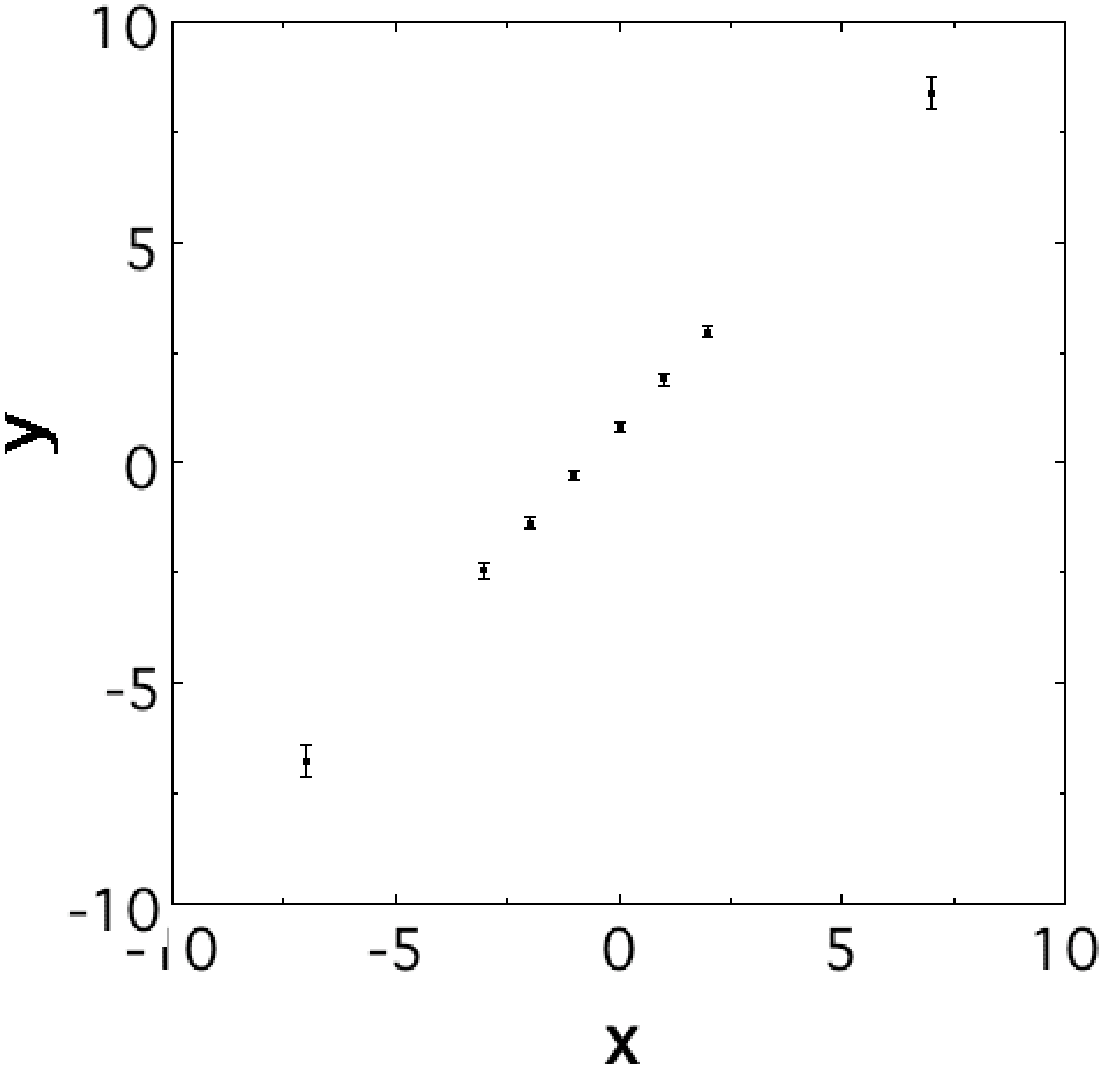

Fig. 3: Plot of the predicted values of $y \pm 10\sigma_y$ for the data in Table 1. Notice that the uncertainty is largest when making predictions out of the range of the training data.

In this simple example, the values of $\alpha = 1/\sigma_w^2$ and σ_ν were given at the outset. In general, these would have to be inferred using the techniques described in MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.