A9 Least-squares fitting

In this appendix *matrix notation* is used. A bold lower case letter is a column matrix (which is an $n \times 1$ matrix representing a vector); a bold capital letter is a matrix. A matrix product $\boldsymbol{C} = \boldsymbol{AB}$ is defined by $C_{ij} = \sum_k A_{ik}B_{kj}$. The transpose $\boldsymbol{A}^{\mathrm{T}}$ of $\boldsymbol{A}$ is defined by $(\boldsymbol{A}^{\mathrm{T}})_{ij} = A_{ji}$. The trace $\mathrm{Tr}\,(\boldsymbol{A})$ is the sum of diagonal elements of $\boldsymbol{A}$. The inverse $\boldsymbol{A}^{-1}$ fulfills $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{1}$ (unit matrix). Recall that $(\boldsymbol{AB})^{\mathrm{T}} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}$ and $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$. The trace of a matrix product is invariant for cyclic permutation of its terms: $\mathrm{Tr}\,(\boldsymbol{ABC}) = \mathrm{Tr}\,(\boldsymbol{CAB})$. Note that for a column matrix (vector) $\boldsymbol{a}$ the product $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}$ is a scalar equal to $\sum_i a_i^2$, while $\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$ is a square matrix with elements $a_i a_j$.

A9.1 How do you find the best parameters a and b in $y \approx ax + b$?

In order to find the values of a and b in the function $f(x) = ax + b$, such that

$$S = \sum_{i=1}^{n} w_i(y_i - f_i)^2 = \sum_{i=1}^{n} w_i(y_i - ax_i - b)^2 \text{minimal},$$

you simply solve for zero derivatives of S/w ($w = \sum_i w_i$) with respect to a and b:

$$\frac{1}{w}\frac{\partial S}{\partial a} = -\frac{2}{w}\sum_{i=1}^{n} w_i x_i(y_i - ax_i - b) = 0$$

$$\frac{1}{w}\frac{\partial S}{\partial b} = -\frac{2}{w}\sum_{i=1}^{n} w_i(y_i - ax_i - b) = 0.$$

From the second equation it follows that $b = \langle y\rangle - a\langle x\rangle$. Substitution of b in the first equation yields the solution for a, see (7.13). The averages are weighted averages such as

$$\langle y \rangle = \frac{1}{w} \sum_{i=1}^{n} w_i y_i.$$

A9.2 General linear regression

In general a set of equations linear in m parameters $\theta_k, k = 1, \ldots, m$ can be written as

$$f_i(\theta_1, \theta_2, \ldots, \theta_m) = \sum_{k=1}^{m} A_{ik}\theta_k, \quad \text{of } \boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{A}\boldsymbol{\theta}. \tag{A9.1}$$

Suppose that the "true" values y_i are given by

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\theta}_m + \boldsymbol{\epsilon}, \tag{A9.2}$$

where $\boldsymbol{\theta}_m$ are the "true" *model values* of the parameters and $\boldsymbol{\epsilon}$ the added stochastic variable or "noise" with properties

$$E[\boldsymbol{\epsilon}] = \boldsymbol{0} \tag{A9.3}$$

$$E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}] = \boldsymbol{\Sigma}. \tag{A9.4}$$

Here $\boldsymbol{\Sigma}$ is the *covariance matrix* of "errors" $\boldsymbol{\epsilon}$ in the measured values $\boldsymbol{y}$. This is a very general assumption allowing correlation between the data points. If $\boldsymbol{\Sigma}$ is diagonal, the data are not correlated.

The chi-squared sum can now be written as

$$\chi^2 = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}). \tag{A9.5}$$

Now the case is quite common that $\boldsymbol{\Sigma}$ is not accurately known, and you only know something about the relative size and the mutual correlation of the data. So assume that – on the basis of your limited knowledge of the uncertainties – you can assign a *weight matrix* $\boldsymbol{W}$ that is proportional to the inverse of the covariance matrix of the random errors in the measured values:

$$\boldsymbol{W} = c\boldsymbol{\Sigma}^{-1}. \tag{A9.6}$$

For the moment the constant c is unknown, but – as we shall see below – under certain conditions c is derivable from the data themselves. Without correlations between the data points both $\boldsymbol{\Sigma}$ and $\boldsymbol{W}$ are diagonal, with σ_i^2, resp. $c\sigma_i^{-2}$, as diagonal elements.

We can now construct the SSQ: the Sum of (weighted) SQuare deviations S:

$$S = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}) = c\chi^2. \tag{A9.7}$$

The derivatives of S with respect to the parameters yields the following vector:

$$\frac{\partial S}{\partial \boldsymbol{\theta}} = -2\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}) = 0. \tag{A9.8}$$

The least-squares solution for $\boldsymbol{\theta}$, indicated by $\hat{\boldsymbol{\theta}}$, is the solution of the set of equations

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}. \tag{A9.9}$$

Thus the final solution for the best estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}. \tag{A9.10}$$

This equation solves any linear least-squares fit, including multiple explanatory variables and including any known correlations between data points. Note that the exact values of the individual inaccuracies are not needed to determine the minimum: if all values of $\boldsymbol{W}$ are multiplied by a constant, the solution $\hat{\boldsymbol{\theta}}$ does not change.

The least-squares solution $\hat{\boldsymbol{\theta}}$ is an *unbiased* estimate of $\boldsymbol{\theta}$, meaning that the expectation of the estimate equals the true value:

$$E[\hat{\boldsymbol{\theta}}] = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}E[\boldsymbol{y}] = \boldsymbol{\theta}_m, \tag{A9.11}$$

because, according to (A9.2) and (A9.3), $E[\boldsymbol{y}] = \boldsymbol{A}\boldsymbol{\theta}_m$.

A9.3 SSQ as a function of the parameters

The expression (A9.7) for $S(\boldsymbol{\theta})$ can be written as a quadratic function of the parameters. After relating S to χ^2, we find the likelihood $\exp[-\frac{1}{2}\chi^2]$ – see (7.4) on page 86 – as a quadratic function of the parameters and from that we can estimate the variances and covariances of the parameters.

Defining the *deviations* from the best estimates of the parameters:

$$\boldsymbol{\Delta\theta} \stackrel{\text{def}}{=} \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}, \tag{A9.12}$$

and the minimum of S:

$$S_0 = (\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{\theta}}), \tag{A9.13}$$

and inserting (A9.10) and (A9.12) into (A9.13), we find

$$S(\boldsymbol{\theta}) = S_0 + \boldsymbol{\Delta\theta}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A}\boldsymbol{\Delta\theta}. \tag{A9.14}$$

Here the gradient A9.8 has been used. So you see that S is a parabolic function in $\boldsymbol{\Delta\theta}$.

Since the likelihood depends on $\chi^2 = S/c$, we need to estimate c. This is straightforward as the expectation for χ_0^2 equals the number of degrees of freedom $n - m$:

$$\hat{\chi}_0^2 = \frac{S_0}{c} = n - m. \tag{A9.15}$$

Hence $c = S/(n - m)$ and

$$\hat{\chi}^2(\boldsymbol{\theta}) = n - m + \frac{n-m}{S_0}\boldsymbol{\Delta\theta}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A}\boldsymbol{\Delta\theta} \tag{A9.16}$$

$$= n - m + \boldsymbol{\Delta\theta}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Delta\theta}, \tag{A9.17}$$

where

$$\boldsymbol{B} \stackrel{\text{def}}{=} \frac{n-m}{S_0}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A}. \tag{A9.18}$$

From (A9.17) you see that the matrix of second derivatives of $\chi^2(\boldsymbol{\theta})$ is given by $2\boldsymbol{B}$.

The likelihood P (proportional to $\exp[-\frac{1}{2}\chi^2]$) is of the form:

$$P \propto \exp\left[-\frac{1}{2}\boldsymbol{\Delta\theta}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Delta\theta}\right]. \tag{A9.19}$$

In case you have reliable knowledge on the uncertainties $\boldsymbol{\Sigma}$, so that you can take the weight matrix *exactly* equal to $\boldsymbol{\Sigma}^{-1}$, the likelihood is

$$P \propto \exp\left[-\frac{1}{2}\boldsymbol{\Delta\theta}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Delta\theta}\right]. \tag{A9.20}$$

Both forms are multivariate normal distributions. With this knowledge we can derive the (co)variances of the parameters.

A9.4 Covariances of the parameters

A multivariate normal distribution (see data sheet NORMAL DISTRIBUTION on page 205) has the form

$$P \propto \exp\left[-\frac{1}{2}\boldsymbol{\Delta\theta}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{\Delta\theta}\right], \tag{A9.21}$$

where $\boldsymbol{C}$ is the covariance matrix:

$$\boldsymbol{C} = E[(\boldsymbol{\Delta\theta})(\boldsymbol{\Delta\theta})^{\mathrm{T}}] \tag{A9.22}$$

$$\boldsymbol{C}_{kl} = \mathrm{cov}\,(\Delta\theta_k, \Delta\theta_l). \tag{A9.23}$$

Comparing this to the likelihood expressions (A9.19) and (A9.20), the expressions for the covariance matrix are found. For the common case that S_0 is used to estimate χ^2:

$$\boldsymbol{C} = \boldsymbol{B}^{-1}; \quad \boldsymbol{B} \text{ defined in (A9.18)} \tag{A9.24}$$

and for the case that uncertainties $\boldsymbol{\Sigma}$ are accurately known:

$$\boldsymbol{C}' = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right)^{-1}. \tag{A9.25}$$

These are our main results. Practical equations are simplifications of (A9.24) and (A9.25).

In order to simplify the presentation, consider the case that there is no correlation between data points, and their variances are σ_i^2 so that $\boldsymbol{\Sigma} = \mathrm{diag}\,(\sigma_i^2)$ and $\boldsymbol{W} = c\,\mathrm{diag}\,(\sigma_i^{-2})$. Then the covariance matrix (A9.24) simplifies to

$$\boldsymbol{C} = \boldsymbol{B}^{-1}; \quad B_{kl} = \frac{n-m}{S_0}\sum_i w_i A_{ik} A_{il} \tag{A9.26}$$

and (A9.25) simplifies to

$$\boldsymbol{C}' = \boldsymbol{B}'^{-1}; \quad B'_{kl} = \sum_i \sigma_i^{-2} A_{ik} A_{il}. \tag{A9.27}$$

The equations for the parameter (co)variances for linear regression of $f(x) = ax + b$, given in Chapter 7 on page 89 in (7.18), (7.19) and (7.20), are easily recovered from these equations. For $\theta_1 = a$ and $\theta_2 = b$, the $n \times 2$ matrix $\boldsymbol{A}$ is given by

$$A_{i1} = x_i; \quad A_{i2} = 1. \tag{A9.28}$$

For example, the element B_{11} of the 2×2 matrix $\boldsymbol{B}$ (A9.26) can be written as

$$B_{11} = \frac{n-m}{S_0}\sum w_i x_i^2 = \frac{n-m}{S_0}\frac{1}{w}\langle x^2\rangle, \tag{A9.29}$$

where w is the total sum of w_i. The rest of the derivation is straightforward and left to the reader.

Why is the s.d. of a parameter given by the projection of the ellipsoid $\Delta\chi^2 = 1$?

The condition $\Delta\chi^2 = 1$ describes a surface (an ellipsoid) in the m-dimensional parameter space. In Fig. 7.5 on page 103 tangents to the ellipse $\Delta\chi^2 = 1$ indicate that the *projection* of this figure on one of the axes

(e.g. θ_1) occurs within the limits $\hat{\theta}_1 \pm \sigma_1$. The tangent touches the ellipse in a point where χ^2 is minimal with respect to all *other* parameters $\theta_2, \ldots, \theta_m$, i.e., where the gradient of χ^2 points in the direction of θ_1:

$$\mathbf{grad}\ \chi^2 = (a, 0, \ldots, 0)^{\mathrm{T}},$$

where a is a constant resulting from $\Delta\chi^2 = \boldsymbol{\Delta\theta}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Delta\theta} = 1$: because[1] $\mathbf{grad}\ \chi^2 = 2\boldsymbol{B}\boldsymbol{\Delta\theta}$,

$$\boldsymbol{\Delta\theta}^{\mathrm{T}}\frac{1}{2}(a, 0, \ldots, 0)^{\mathrm{T}} = \frac{1}{2}a\Delta\theta_1 = 1.$$

Hence

$$\boldsymbol{B}\boldsymbol{\Delta\theta} = \frac{1}{2}(2/\Delta\theta_1, 0, \ldots, 0)^{\mathrm{T}}$$

and

$$\boldsymbol{\Delta\theta} = \boldsymbol{C}(1/\Delta\theta_1, 0, \ldots, 0)^{\mathrm{T}} \text{ or } \Delta\theta_1 = \pm\sqrt{C_{11}} = \pm\sigma_1. \tag{A9.30}$$

This is what we wished to prove.[2]

Nonlinear least-squares fit

When the functions $f_i(\theta_1, \ldots, \theta_m)$ are not linear in all parameters, but $S = (\boldsymbol{y} - \boldsymbol{f})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{f})$ does have a minimum $S_0 = S(\hat{\boldsymbol{\theta}})$, then $S(\boldsymbol{\theta})$ can be expanded around that minimum in a Taylor series with zero linear term, just as in (A9.14) in the linear case. In terms of the expectation of χ^2 (which equals $n - m$ at the minimum):

$$\hat{\chi}^2(\boldsymbol{\theta}) = \frac{n-m}{S_0}S(\boldsymbol{\theta}) = n - m + \boldsymbol{\Delta\theta}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Delta\theta} + \cdots, \tag{A9.31}$$

After a redefinition of the matrix $\boldsymbol{A}$:

$$A_{ik} = \left(\frac{\partial f_i}{\partial\theta_k}\right)_{\hat{\theta}}, \tag{A9.32}$$

all equations for the parameters and their (co)variances remain approximately valid. The inverse of $\boldsymbol{B} = \frac{n-m}{S_0}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A}$ is still (but approximately) equal to the covariance matrix of the parameters. See Press *et al.* (1992)[3] for

[1] The gradient of a quadratic form $\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{x}$, with $\boldsymbol{G}$ symmetric, equals $\boldsymbol{G}\boldsymbol{x}$.
[2] The proof can be found in Press *et al.* (1992), see the reference section on page 124.
[3] See reference list on page 124.

a discussion on this point. For uncorrelated data, Equation (A9.26) on page 164 is still valid:

$$B_{kl} = \frac{n-m}{S_0} \sum_{i=1}^{n} w_i \frac{\partial f_i}{\partial \theta_k} \frac{\partial f_i}{\partial \theta_l}. \tag{A9.33}$$

The covariance matrix is approximately equal to the inverse of $\boldsymbol{B}$.

For functions that are not linear in the parameters, the likelihood function is only approximately equal to a multivariate normal distribution. Especially the tails of the distribution may differ and the blind derivation of confidence limits based on normal distributions may be erroneous in the tail regions of the distribution. More accurate estimations can be done using the likelihood function

$$p(\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2}\chi^2(\boldsymbol{\theta})\right]. \tag{A9.34}$$

As the practical implications are of little importance, we don't pursue this point here any further.