Physics 234 — Data Analysis

Spring 1997

Martin Siegert

Department of Physics

Simon Fraser University

Data Analysis

Extracting information from a set of data points is a problem almost every physicist has to deal with regardless of whether the data were obtained in a real experiment or a computer simulation. Let us assume that somebody gave us a set of data $x_1, x_2, \ldots, x_N$. The probability distribution $p(x)$ describes the occurrence of a particular value x in such a measurement: $p(x)dx$ is the probability of finding a value $\overline{x} \in [x, x+dx]$. From this probability distribution all quantities of interest can be calculated, e.g., the mean value $\mu = \langle x_i \rangle = \int_{-\infty}^{\infty} dx\, xp(x)$ or the variance $\sigma^2 = \langle x_i^2 \rangle - \langle x_i \rangle^2 = \int_{-\infty}^{\infty} dx\,(x^2 - \mu^2)p(x)$; in general $\langle f(x_i) \rangle = \int_{-\infty}^{\infty} dx\, f(x)p(x)$. Since the probability distribution is usually unknown, these quantities have to be estimated from the measured data. There are two mathematical theorems, the law of large numbers and the central limit theorem, that provide the necessary information.

1. Central Limit Theorem

Let $x_1, x_2, \ldots, x_N$ be a sequence of independent random variables with a common probability distribution $p(x)$ and expectation values $\langle x_i \rangle = \mu$. Let $S_N = \sum_{i=1}^{N} x_i$.
Law of large numbers: For all $\epsilon > 0$ the probability of $S_N/N - \mu$ being larger than ϵ approaches 0 in the limit $N \to \infty$.
In other words: The average S_N/N converges to the mean μ of the probability distribution $p(x)$ for large N.

An even stronger statement can be made, if we additionally assume that the variance σ^2 is finite:
Central Limit Theorem: In the limit $N \to \infty$ the probability

$$\Pr\left[\frac{S_N - N\mu}{\sigma\sqrt{N}} < X\right] \longrightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} dy\, e^{-y^2/2} \equiv P(X)$$

converges for all X to the normal or Gaussian probability function $P(X)$.
This theorem gives an estimate for the probability that the discrepancy $|S_N/N - \mu|$ is larger than $\sigma/\sqrt{N}$. Furthermore it states that for large N the mean μ and the variance σ^2 fully determine the probability distribution of the average $\overline{x} = S_N/N$

$$\overline{p}(\overline{x}) \xrightarrow{N\to\infty} \frac{\sqrt{N}}{\sigma\sqrt{2\pi}} e^{-\frac{N(\overline{x}-\mu)^2}{2\sigma^2}} \ , \tag{1}$$

i.e., regardless of the probability distribution $p(x)$ of the individual variables x_i, the average $\overline{x} = S_N/N$ follows in the limit of large N a Gaussian probability distribution with mean $\mu = \overline{x}$ and variance σ^2/N.
Examples:
(a) Rolling the die: $x_i = i$, $i = 1, \ldots, 6$

$$\Longrightarrow \langle x_i \rangle = \mu = 3.5 \ , \sigma^2 = (1^2 + \ldots + 6^2)/6 - (3.5)^2 = \tfrac{35}{12}.$$

Central limit theorem:

$$\Pr\left[|S_N - 3.5N| < \alpha\sqrt{35N/12}\right] \simeq P(\alpha) - P(-\alpha) = \frac{1}{2\pi} \int_{-\alpha}^{\alpha} dy\, e^{-y^2/2} \ .$$

$N = 1000$, $\alpha = 1$: $\Pr[3446 < S_N < 3554] \simeq 0.68$; $\alpha = 2$: $\Pr[3392 < S_N < 3608] \simeq 0.95$ (in other words: the measured value of $\overline{x} = S_N/N$ will fall within $\pm 2\sigma$ of the true value μ 95% of the time).

(b) Sampling: Suppose that in a population of M families there are M_k families with exactly k children, $k = 0, 1, \ldots$ and $\sum_k M_k = M$. For a family chosen at random, the number of children is a random variable which assumes the value ν with probability $p_\nu = M_\nu/M$. We now make a measurement by selecting randomly N families. Then the sample average S_N/N is likely to be near $\mu = \sum_{\nu=1}^{N} \nu p_\nu = \sum_{\nu=1}^{N} \nu M_\nu/M$, namely the population average. If it is desired that there be a probability of 0.99 or better that the sample average S_N/N differs from the unknown mean μ by less that $1/10$, then the sample size N should be such that $\Pr[|(S_N - N\mu)/N| < 1/10] \geq 0.99$. From $P(x) - P(-x) = 0.99$ we obtain $x \simeq 2.58$, hence N should satisfy $2.58\sigma/\sqrt{N} \leq 1/10$ or $N \geq 666\sigma^2$. Thus, this result depends crucially on an estimate of the variance σ^2.

How do we estimate μ and σ^2, and what is the error of the estimate of μ?
Result:

$$\overline{x} \simeq \mu \ , \tag{2a}$$

$$\langle \overline{x}^2 \rangle - \langle \overline{x} \rangle^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} \left\langle x_i^2 - \overline{x}^2 \right\rangle = \frac{1}{N}\sigma^2 \ . \tag{2b}$$

The first equation follows from the law of large numbers. The variance of $\overline{x}$ is obtained from

$$\begin{aligned}
\langle \overline{x}^2 \rangle &= \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_N p(x_1) \times \cdots \times p(x_N) \left(\frac{1}{N} \sum_{i=1}^{N} x_i \right)^2 \\
&= \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_N p(x_1) \times \cdots \times p(x_N) \left[\frac{1}{N^2} \sum_{i \neq j} x_i x_j + \frac{1}{N^2} \sum_i x_i^2 \right] \\
&= \frac{N(N-1)}{N^2} \langle x \rangle^2 + \frac{N}{N^2} \langle x^2 \rangle \\
&= \frac{N-1}{N} \mu^2 + \frac{1}{N}(\sigma^2 + \mu^2) = \mu^2 + \frac{1}{N}\sigma^2 \ ,
\end{aligned}$$

where we have used the identity $\langle x^2 \rangle = \langle (x - \langle x \rangle)^2 \rangle + \langle x \rangle^2$. Now we have

$$\sum_{i=1}^{N} \left\langle x_i^2 - \overline{x}^2 \right\rangle = N(\sigma^2 + \mu^2) - N\mu^2 - \sigma^2 = (N-1)\sigma^2 \quad \Rightarrow \quad \sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left\langle x_i^2 - \overline{x}^2 \right\rangle \ ,$$

and therefore

$$\mathrm{Var}(\overline{x}) = \langle \overline{x}^2 \rangle - \langle \overline{x} \rangle^2 = \mu^2 + \frac{1}{N}\sigma^2 - \mu^2 = \frac{1}{N}\sigma^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} \left\langle x_i^2 - \overline{x}^2 \right\rangle \ .$$

(Quote from Numerical Recipes: "If the difference between N and $N-1$ ever matters to you, then you are probably up to no good anyway — e.g., trying to substantiate a questionable hypothesis with marginal data."). The derivation above did not make use of the central limit theorem and is therefore valid for all probabilities $p(x)$ with mean μ and variance σ^2.

2. Least-Squares Fits

Assume that we made a measurement (experiment or computer simulation) and we want to compare the measured data with a theory or model that depends on some parameters. We want to verify whether the theory applies to our experiment and, if that is the case, what are the parameters? Since the data are not exact because of "noise" they will never fit the theory exactly. Thus, we need some criteria for answering the question "How good is the fit?". Hence, the fitting procedure should provide (i) the parameters, (ii) error estimates on the parameters, and (iii) a statistical measure of the goodness of the fit.

Suppose we are fitting N data points (x_i, y_i), $i = 1, \ldots, N$, to a model with M parameters, $a_1, \ldots, a_M$. Thus the model provides some function $f(x; a_1, \ldots, a_M)$ that predicts, in which way the measured data y_i should behave as a function of x. Now, suppose that each data point y_i has a measurement error that is independently random and is Gaussian distributed with variance σ_i^2 around the true model $f(x; \hat{\mathbf{a}})$, where all the parameters have the correct values $\hat{\mathbf{a}}$. Then the probability or likelihood that this data set can be described by the model $f(x; \mathbf{a})$ is the product of the probabilities of each data point,

$$L = \prod_{i=1}^{N} p(y_i) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{(y_i - f(x_i; \mathbf{a}))^2}{2\sigma_i^2}\right] = c_\sigma \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i - f(x_i; \mathbf{a})}{\sigma_i}\right)^2\right] \quad (3)$$

with $c_\sigma = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}}$. By maximizing the probability (3) we obtain the most likely set of parameters. Since c_σ does not depend on $\mathbf{a}$, maximizing L is equivalent to minimizing the argument of the exponential,

$$\chi^2 = \sum_{i=1}^{N}\left[\frac{y_i - f(x_i; a_1, \ldots, a_M)}{\sigma_i}\right]^2 , \quad (4)$$

with respect to the parameters $a_1, \ldots, a_M$. The minimization of (4) is called a *least-squares fit*. From the previous section we know that any probability distribution of an average will converge to a gaussian distribution, *if the number of measurements is large.* However, in real life the assumption of a Gaussian distribution may not be justified, e.g., sometimes you have outliers in your experimental data that are way off the theoretical curve. These outliers will completely spoil the least-squares fit. This must be kept in mind when estimating the size of errorbars for parameters obtained from a least-squares fit.

This brings us to the questions "what is a good fit?" and "what are the errorbars on the calculated parameters?" If each of the data points y_i is distributed according to a Gaussian probability distribution with variances σ_i^2, it can be shown that the value $\chi^2_{\min}$ obtained from a minimization of (4) is distributed according to the so-called χ^2-probability distribution with ν degrees of freedom

$$p(\chi^2|\nu) = \frac{(\chi^2)^{\frac{\nu}{2}-1} e^{-\chi^2/2}}{2^{\nu/2}\Gamma(\nu/2)} , \quad (5)$$

where $\nu = N - M$ and $\Gamma(z)$ is the Gamma-function which can be regarded as a generalization of the factorial function $n!$ for nonintegers: $\Gamma(z+1) = z\Gamma(z)$, $\Gamma(n+1) = n!$. The statement that the probability distribution of χ^2 is given by (5) is correct under the assumption that the measurement errors are sufficiently small so that at the minimum of χ^2 the deviations of the parameters $\mathbf{a}$ from the true values $\hat{\mathbf{a}}$ are so small that the fitting function can be replaced by the Taylor expansion of f around the true values $\hat{\mathbf{a}}$ truncated after the linear term,

$$f(x_i;\mathbf{a}) = f(x_i;\hat{\mathbf{a}}) + \sum_{n=1}^{M}(a_n - \hat{a}_n)\left.\frac{\partial f(x_i;\mathbf{a})}{\partial a_n}\right|_{\mathbf{a}=\hat{\mathbf{a}}}$$

The probability distribution $p(\chi^2|\nu)$ depends on only one parameter ν. In particular, the mean of this distribution is ν itself and the variance is 2ν. Since the mean of the χ^2 distribution function is $\nu = N - M$, this fact can be used as a criterion for the goodness of the fit: a reasonable fit should have $\chi^2_{\text{min}} \simeq N - M$. However, the validity of this criterion depends crucially on the estimates of the variances σ_i^2. If the σ_i^2 are poorly estimated or even arbitrarily set to 1, the value of χ^2_{min} is meaningless. In these cases you can use the relation $\chi^2_{\text{min}} = N - M$ to determine an overall factor of the σ_i^2: If the least-squares fit using variances $\tilde{\sigma}_i^2$ yielded a result $\tilde{\chi}^2$, the variances $\sigma_i^2 = [\tilde{\chi}^2/(N-M)]\tilde{\sigma}_i^2$ would give a result of $\chi^2 = N - M$. If these values for σ_i^2 turn out to be unreasonable, this is an indication that the quality of the fit is not very good. Clearly, this method prohibits a direct evaluation of the goodness of the fit.

The probability to measure a data set that yields a value χ'^2 that is larger than the value χ^2 that was actually obtained from the minimization procedure for the current data set is given by the integral over the probability distribution,

$$Q(\chi^2|\nu) = \int_{\chi^2}^{\infty} dy\, \frac{y^{\frac{\nu}{2}-1}e^{-y/2}}{2^{\nu/2}\Gamma(\nu/2)}\ . \tag{6}$$

These integrals can be calculated numerically or found in tables, e.g., P. R. Bevington and D. K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, or M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. If the fitting function is a good approximation to your data, you expect a value of $Q(\chi^2|\nu) \simeq 1/2$. Values of $Q \gtrsim 10^{-3}$ may still be acceptable, in particular if the variances σ_i^2 have been underestimated. Even smaller values for Q indicate a problem with the fit.

However, quite often you will be able to obtain a good fit to your data, but the parameters $\mathbf{a}$ still fluctuate quite strongly without changing the value of χ^2 significantly. This occurs if the minimum of χ^2 sits in a very flat valley, so that by changing the parameters in a particular way you wander around in that valley without changing χ^2. This kind of degeneracy is an indication that the model contains too many parameters, or, conversely, your data set is not sufficient to determine all the parameters independently. A trivial example: Assume that somebody tries to fit a set of data y_i to the function $f(x_i;a_1,a_2,a_3) = a_1\exp[-a_2x_i + a_3]$. Naturally, he will not be able to determine a_1 and a_3 independently, since all fits with the same value for $a_1e^{a_3}$ will give the same χ^2. Unfortunately, these degeneracies will occur even if they are not built into the fitting function $f(x;\mathbf{a})$ in such an obvious way. In these cases the number of parameters must be reduced, or you have to concede that the available data do not allow to determine the parameters with the desired accuracy.

Mathematical predictions for the errorbars on the parameters a_i can only be derived, if several preconditions apply (see the following section). Most importantly, it is assumed that only statistical errors are present. Quite often this is incorrect. Systematic errors appear because the "true fitting function" is unknown. For example the chosen fitting function f

is not a perfect model for your data y_i over the full range of the x_i. So-called crossover effects are a typical example that cause such systematic errors. Also the assumption that the measurement data y_i are Gaussian distributed may not be valid. If systematic errors are more important than statistical errors, the standard deviations of the parameters that are derived in the following section are meaningless. Indications for systematic errors can be obtained from the residuals $R_i = y_i - f(x_i; \mathbf{a})$. If only statistical errors are present, the R_i are randomly distributed around zero. If systematic errors are dominant, there are intervals in x, where *all* R_i are positive, followed by an interval, where all R_i are negative, i.e., the R_i have an oscillatory behavior as a function of x.

For practical purposes you can estimate the errorbars on your parameters by changing the initial guesses for $\mathbf{a}$ and compare the results after the changes in χ^2 are no longer significant. You can also fix some of the parameters and minimize χ^2 only with respect to the remaining parameters. To test how sensitively the parameters depend on the model one can make "supposedly insignificant" changes to the function f and repeat the least-squares fit. The obtained differences in the parameter values yield in many cases by far better (most often larger!) error estimates of the parameters than the formulae presented in Sec. 2.1. In fact, if this leads to large changes in the parameters, your model is either ill-defined or your data set is too small. In any case it is a good idea to perform the least-squares fit under various conditions in order to get a feeling for the accuracy of the parameter estimates.

Although the formulae for the variances of the parameters derived in the following section may not be valid, e.g., if systematic errors are present, it is nevertheless helpful to calculate the so-called correlation coefficients to judge the quality of your fit. The Taylor expansion of χ^2 about some set of parameters $\tilde{\mathbf{a}}$ has the form

$$\chi^2(\mathbf{a}) = \chi^2(\tilde{\mathbf{a}}) + \sum_{j=1}^{M} \frac{\partial \chi^2}{\partial a_j}\Big|_{\mathbf{a}=\tilde{\mathbf{a}}} (a_j - \tilde{a}_j) + \frac{1}{2} \sum_{j,k=1}^{M} \frac{\partial^2 \chi^2}{\partial a_j \partial a_k}\Big|_{\mathbf{a}=\tilde{\mathbf{a}}} (a_j - \tilde{a}_j)(a_k - \tilde{a}_k) + \mathcal{O}((\mathbf{a} - \tilde{\mathbf{a}})^3) \ , \quad (7)$$

where $\mathcal{O}((\mathbf{a} - \tilde{\mathbf{a}})^3)$ denotes terms of higher than second order in $\mathbf{a} - \tilde{\mathbf{a}}$. The first and second derivatives of χ^2 with respect to the parameters $\mathbf{a}$ are easily expressed in terms of the fitting function:

$$\beta_j(\mathbf{a}) \ := \ -\frac{1}{2} \frac{\partial \chi^2}{\partial a_j} = \sum_{i=1}^{N} \frac{[y_i - f(x_i; \mathbf{a})]}{\sigma_i^2} \frac{\partial f(x_i; \mathbf{a})}{\partial a_j} \ , \qquad j = 1, \ldots, M \ , \quad (8a)$$

and

$$\alpha_{jk}(\mathbf{a}) \ := \ \frac{1}{2} \frac{\partial \chi^2}{\partial a_j \partial a_k} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial f(x_i; \mathbf{a})}{\partial a_j} \frac{\partial f(x_i; \mathbf{a})}{\partial a_k} - [y_i - f(x_i; \mathbf{a})] \frac{\partial^2 f(x_i; \mathbf{a})}{\partial a_j \partial a_k} \right] \ . \quad (8b)$$

At the minimum $\tilde{\mathbf{a}} = \mathbf{a}_{\min}$, $\chi^2(\mathbf{a}_{\min}) = \chi^2_{\min}$ the first derivatives of χ^2 vanish:

$$\beta_j(\mathbf{a}_{\min}) = 0 \quad j = 1, \ldots, M \ . \quad (9)$$

These are M equations that determine the values of the parameters $\mathbf{a}_{\min}$. The symmetric $M \times M$ matrix α of the second derivatives is called the curvature matrix. The terms containing the second derivatives of the fitting function in (8b) are usually neglected in the definition of the matrix α. The reason is that these terms are multiplied by $y_i - f(x_i; \mathbf{a})$, i.e., for a reasonable model this is just the measurement error of the i-th data point. Since these errors should be random, the terms containing the second derivatives should cancel when summed over i. Thus, we can use the simpler form

$$\alpha_{jk} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial f(x_i;\mathbf{a})}{\partial a_j} \frac{\partial f(x_i;\mathbf{a})}{\partial a_k} \ . \tag{10}$$

At the minimum $\mathbf{a} = \mathbf{a}_{\min}$ the curvature matrix α contains information about the uncertainties of the parameters $\mathbf{a}_{\min}$. If only statistical errors are present, it can be shown (see Section 2.1) that the variances of the parameters of the parameter a_k are

$$\left\langle (a_k - \hat{a}_k)^2 \right\rangle = \frac{\chi^2}{N-M} c_{kk} \ . \tag{11}$$

Here $\hat{\mathbf{a}}$ are the true values of the parameters and the so-called covariance matrix c is the inverse of the curvature matrix α, i.e., $\sum_{k=1}^{M} \alpha_{jk} c_{kl} = \delta_{jl}$ with the Kronecker delta $\delta_{jl} = 1$ for $j = l$ and $\delta_{jl} = 0$ for $j \neq l$. Consequently the standard deviations are

$$\Delta a_k = \sqrt{\frac{\chi^2}{N-M} c_{kk}} \ . \tag{12}$$

It should be emphasized, however, that the standard deviations (12) *do not* correspond to 68% confidence intervals of the parameters, since the parameters are not Gaussian distributed. The offdiagonal elements of the matrix c are related to the covariances of the parameters

$$\langle (a_j - \hat{a}_j)(a_k - \hat{a}_k) \rangle = \frac{\chi^2}{N-M} c_{jk} \ . \tag{13}$$

The covariances contain information about correlations between different parameters. The correlation coefficients r_{jk} defined as

$$r_{jk} = \frac{c_{jk}}{\sqrt{c_{jj} c_{kk}}} \tag{14}$$

take values between -1 and 1. $|r_{jk}| = 1$ implies a linear relation between the uncertainties of the parameters a_j and a_k. If $r_{kl} = 1$ Δa_j increases linearly with Δa_k, whereas for $r_{jk} = -1$ it decreases. Values of $|r_{jk}| \gtrsim 0.9$ indicate a problem with your fit. Such large correlations between parameters are obtained, if the minimum of χ^2 sits in a flat valley: If you wander along the flat direction, it is possible to change the values of the parameters without significantly changing the value of χ^2. Therefore large values of $|r_{jk}|$ always indicate large uncertainties in the corresponding parameters. In such a case the number of parameters must be reduced, or, in other words: you cannot determine all parameters a_j, $j = 1, \ldots, N$ independently from your data set.

Some last comments: If your data y_i are spread out over a large interval of several orders of magnitude it may be a good idea to fit the logarithms $\ln y_i$ to the fitting function $\ln f(x;\mathbf{a})$. A direct fit of y_i to f would be dominated by the largest values of y_i since already small *relative* errors $|(y_i - f(x_i;\mathbf{a}))/y_i|$ will give large contributions to χ^2. Typical examples are power-law fits with $f(x;a_1,a_2,a_3) = a_1 x^{a_2} + a_3$. Thus, if the absolute errors Δy_i are all of the same order of magnitude then the y_i should be directly fitted to the fitting function f, if the relative errors $\Delta y_i / y_i$ are of the same order of magnitude you may consider fitting $\ln y_i$ to $\ln f$. These considerations also elucidate that it is very important to have good estimates of the variances σ_i^2 that serve as weight factors in order to get a reliable fit to your data.

2.1 Estimating the Uncertainties of Least-Squares Fit Parameters

(see, J. R. Wolberg, Prediction Analysis, Van Nostrand Company, Princeton, 1967, ch. 3.10)

In this section we assume that the following preconditions apply:

(1) The measured values y_i are Gaussian distributed around the true values $\widehat{y}_i$ with variances σ_i^2.

(2) The measurement errors $\Delta y_i = y_i - \widehat{y}_i$ are uncorrelated.

(3) The data are fitted to a model $y = f(x; \mathbf{a})$. The true model obeys $\widehat{y}_i - f(x_i; \widehat{\mathbf{a}}) = 0$. It is assumed that $f(x_i; \mathbf{a})$ can be approximated around $\widehat{\mathbf{a}}$ by the following expansion:

$$y_i - f(x_i; \mathbf{a}) = y_i - \widehat{y}_i - \sum_{k=1}^{M} \frac{\partial f_i}{\partial a_k}(a_k - \widehat{a}_k) + \mathcal{O}((a_k - \widehat{a}_k)^2) \ , \tag{15}$$

where the notation $\partial f_i / \partial a_k$ has been used as a short hand for $\partial f(x_i; \mathbf{a})/\partial a_k|_{\mathbf{a}=\widehat{\mathbf{a}}}$.

(4) Only statistical errors are present.

At the minimum of χ^2 we have $\beta_k = 0$. Thus, setting the left-hand side of Eq. (15) to zero, multiplying the equation by $\frac{1}{\sigma_i^2}\frac{\partial f_i}{a_l}$ and summing over all data points we obtain

$$\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\frac{\partial f_i}{\partial a_l}\Delta y_i = \sum_{i=1}^{N}\sum_{k=1}^{M} \frac{1}{\sigma_i^2}\frac{\partial f_i}{\partial a_l}\frac{\partial f_i}{\partial a_k}(a_k - \widehat{a}_k) = \sum_{k=1}^{M} \alpha_{lk}(a_k - \widehat{a}_k) \ .$$

By introducing the inverse $c = \alpha^{-1}$ of the matrix α so that $\sum_{m=1}^{M} \alpha_{km} c_{ml} = \delta_{ij}$, we can solve for the errors of the parameters

$$a_k - \widehat{a}_k = \sum_{l=1}^{M} c_{kl} \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\frac{\partial f_i}{\partial a_l}\Delta y_i \ .$$

From this equation we obtain

$$(a_k - \widehat{a}_k)(a_l - \widehat{a}_l) = \sum_{i=1}^{N} \left(\frac{\Delta y_i}{\sigma_i^2}\right)^2 \left[\sum_{m=1}^{M} c_{km}\frac{\partial f_i}{\partial a_m}\right]\left[\sum_{n=1}^{M} c_{ln}\frac{\partial f_i}{\partial a_n}\right] + \begin{matrix}\text{terms containing} \\ \Delta y_i \Delta y_j \text{ with } i \neq j\end{matrix} \ .$$

Since the measurement errors are assumed to be uncorrelated the terms $\sim \Delta y_i \Delta y_j$ with $i \neq j$ will be zero on average. Thus, if the measurement would be repeated many times the "mean value" or, more accurately, the most probable value

$$\langle (a_k - \widehat{a}_k)(a_l - \widehat{a}_l) \rangle = \sum_{i=1}^{N} \left\{ \left\langle \left(\frac{\Delta y_i}{\sigma_i}\right)^2 \right\rangle \sum_{n,m=1}^{M} c_{km} c_{kn} \frac{1}{\sigma_i^2}\frac{\partial f_i}{\partial a_k}\frac{\partial f_i}{\partial a_l} \right\} \tag{16}$$

is obtained. If σ_i^2 is the variance of the measurement of y_i then the ratio $\langle (\Delta y_i / \sigma_i)^2 \rangle \equiv 1$. However, in practice σ_i can be estimated at most, quite often the factors σ_i^2 are nothing but weight factors that have no direct relation to the variances. Thus, the expectation value $\langle (\Delta y_i / \sigma_i)^2 \rangle$ may not be a constant but depend on i. In order to proceed we approximate this expectation value by its average over i,

$$\left\langle \left(\frac{\Delta y_i}{\sigma_i}\right)^2 \right\rangle \approx \left\langle \frac{1}{N}\sum_{i=1}^{N} \left(\frac{\Delta y_i}{\sigma_i}\right)^2 \right\rangle =: q \ .$$

Since we allow $\langle(\Delta y_i)^2\rangle$ to differ from σ_i^2 the average q must be estimated in a different way later on. Using the definition of the matrix α (10) we obtain the following result for the correlations of the a_k

$$\langle(a_k - \widehat{a}_k)(a_l - \widehat{a}_l)\rangle = qc_{kl} \tag{17}$$

and, specifically, the estimated variances of the parameters

$$\left\langle(a_k - \widehat{a}_k)^2\right\rangle = qc_{kk} \ . \tag{18}$$

It remains to estimate q. Using Eq. (15) and the definition $R_i = y_i - f(x_i; \mathbf{a})$ of the residuals we find

$$(\Delta y_i)^2 = R_i^2 + 2R_i \sum_{k=1}^{M} \frac{\partial f_i}{\partial a_k}(a_k - \widehat{a}_k) + \sum_{k,l=1}^{M} \frac{\partial f_i}{\partial a_k}\frac{\partial f_i}{\partial a_l}(a_k - \widehat{a}_k)(a_l - \widehat{a}_l) \ .$$

The term linear in $a_k - \widehat{a}_k$ will be zero on average. By summing over i we obtain

$$Nq = \sum_{i=1}^{N} \left\langle \frac{R_i^2}{\sigma_i^2} \right\rangle + \sum_{i=1}^{N} \left\langle \frac{1}{\sigma_i^2} \sum_{k,l=1}^{M} \frac{\partial f_i}{\partial a_k}\frac{\partial f_i}{\partial a_l}(a_k - \widehat{a}_k)(a_l - \widehat{a}_l) \right\rangle \ .$$

The first term on the right-hand side is nothing but the minimized χ^2, the second term can be evaluated using (35) and (17) with the result Mq. Hence, we finally obtain

$$q \equiv \left\langle \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\Delta y_i}{\sigma_i}\right)^2 \right\rangle = \frac{\chi^2}{N-M} \tag{19}$$

and the variances

$$\left\langle(a_k - \widehat{a}_k)^2\right\rangle = \frac{\chi^2}{N-M} c_{kk} \tag{20}$$

as well as the covariances

$$\langle(a_k - \widehat{a}_k)(a_l - \widehat{a}_l)\rangle = \frac{\chi^2}{N-M} c_{kl} \ . \tag{21}$$

The results (20), (21) explain why the matrix c is called the covariance matrix. From Eq. (20) one obtaines the standard deviation of the parameters $\Delta a_k = \sqrt{c_{kk}\chi^2/(N-M)}$.

Some final remarks: As already mentioned in Sec. 7.2 χ^2 is distributed according to the χ^2-probability function $P(\chi^2|\nu)$ that has a mean value of $\nu = N - M$. Thus, *if σ_i^2 is the correct variance of the measurement of y_i* Eq. (19) gives 1 on both sides of the equation. Otherwise, one can use the condition $\chi^2 \overset{!}{=} N - M$ to estimate an overall common factor of the σ_i^2 as already mentioned in the previous section. The final result (20) depends on several assumptions and approximations and for a particular experiment it may not be clear whether these assumptions are valid. Therefore, an independent estimate of the errors along the lines indicated at the end of the previous section is essential in almost every case.