With the definitions

$$g_m = \sum_i y_i \, f_m(x_i)/\sigma_i^2 \tag{17.13}$$

and

$$\left(V_{\widehat{\alpha}}^{-1}\right)_{mn} = \sum_i f_n(x_i)\, f_m(x_i)/\sigma_i^2 \,, \tag{17.14}$$

the k-element column vector of solutions $\widehat{\boldsymbol{\alpha}}$, for which $\partial\chi^2/\partial\alpha_m = 0$ for all m, is given by

$$\widehat{\boldsymbol{\alpha}} = V_{\widehat{\alpha}}\, \boldsymbol{g} \,. \tag{17.15}$$

Non-independent y_i's

More generally, the measured y_i's are not independent. Then the set of σ_i^2's must be replaced by the $N \times N$ covariance matrix V_y. Then, if H is the $N \times k$ matrix with element $H_{in} = f_n(x_i)$, the solution $\widehat{\boldsymbol{\alpha}}$ is given by the solution to the *normal equation*

$$(H^T\, V_y^{-1}\, H)\, \widehat{\boldsymbol{\alpha}} = H^T\, V_y^{-1}\, \boldsymbol{y} \,, \tag{17.16a}$$

or, formally,

$$\widehat{\boldsymbol{\alpha}} = (H^T\, V_y^{-1}\, H)^{-1}\, H^T\, V_y^{-1}\, \boldsymbol{y} \equiv D\boldsymbol{y} \,, \tag{17.16b}$$

where $\boldsymbol{y}$ is the N-element vector of measured y_i's. The normal equations may be solved by numerical methods much more computationally efficient than brute application of Eq. (17.16b). In particular, $H^T\, V_y^{-1}\, H$ is sometimes singular or nearly singular. In such cases there is at least one f_n which may be expressed as a linear combination of others (or nearly so) when evaluated at the data points. The best procedure is usually to drop such functions from the expansion (or set $\widehat{\alpha}_n = 0$). See Press [8], Maindonald [9], or Basilevsky [10] for discussions.

In terms of the $k \times N$ matrix D, the standard covariance matrix for the $\widehat{\boldsymbol{\alpha}}$ is estimated by

$$V_{\widehat{\alpha}} = D\, V_y\, D^T \,. \tag{17.17}$$

If the measured y_i's are independent, V_y is diagonal with ii^{th} element σ_i^2 and $V_{\widehat{\alpha}}$ is obtained from Eq. (17.14) above.

The expected covariance [see Eq. (16.9)] of $\widehat{\alpha}_n$ and $\widehat{\alpha}_m$ is estimated by

$$E\left[(\alpha_n - \widehat{\alpha}_n)(\alpha_m - \widehat{\alpha}_m)\right] = (V_{\widehat{\alpha}})_{nm} \,. \tag{17.18}$$

Even when the y_i's are independent (diagonal V_y), $\widehat{\alpha}_n$ and $\widehat{\alpha}_m$ may not be (nondiagonal $V_{\widehat{\alpha}}$).

For the model function $y = \sum \alpha_n\, f_n(x)$, the estimated variance of an interpolated or extrapolated value of y at a point x is

$$E\left[(y-\widehat{y}\,)^2\right] = \sigma^2(y)$$

$$= \sum_{n,m} (V_{\widehat{\alpha}})_{nm}\, f_n(x)\, f_m(x) \,. \tag{17.19}$$

Example: straight-line fit

For the case of a straight-line fit, $y(x) = \alpha_1 + \alpha_2\, x$, one obtains, for independent measurements y_i, the following estimates of α_1 and α_2,

$$\widehat{\alpha}_1 = (g_1\, S_{22} - g_2\, S_{12})/D \,, \tag{17.20}$$

$$\widehat{\alpha}_2 = (g_2\, S_{11} - g_1\, S_{12})/D \,, \tag{17.21}$$

where

$$(S_{11},\, S_{12},\, S_{22}) = \sum (1,\, x_i,\, x_i^2)/\sigma_i^2 \,, \tag{17.22a}$$

$$(g_1,\, g_2) = \sum (1,\, x_i) y_i/\sigma_i^2 \,. \tag{17.22b}$$

respectively, and

$$D = S_{11}\, S_{22} - S_{12}^2 \,. \tag{17.23}$$

The covariance matrix of the fitted parameters is:

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{pmatrix} = \frac{1}{D} \begin{pmatrix} S_{22} & -S_{12} \\ -S_{12} & S_{11} \end{pmatrix} \,. \tag{17.24}$$

The estimated variance of an interpolated or extrapolated value of y at point x is:

$$\left.(\widehat{y} - y_{\rm true})^2\right|_{\rm est} = \frac{1}{S_{11}} + \frac{S_{11}}{D}\left(x - \frac{S_{12}}{S_{11}}\right)^2 \,. \tag{17.25}$$

17.4.1. *General comments*:

If y is not linear in the fitting parameters α_n, or if the errors σ_i depend upon y and therefore on α_n, the solution vector may have to be found by iteration of Eqs. (17.13)–(17.15) or Eq. (17.16b). The same results may be obtained by numerical techniques from the sum of squares, χ^2, directly, if we have a reasonable first guess $\boldsymbol{\alpha}_0$ for the solution vector:

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0 - \left(\frac{\partial^2\chi^2}{\partial\alpha^2}\right)^{-1}_{\alpha_0} \cdot \left.\frac{\partial\chi^2}{\partial\alpha}\right|_{\alpha_0} \tag{17.26a}$$

and

$$V_{\widehat{\alpha}} = 2\left(\frac{\partial^2\chi^2}{\partial\alpha^2}\right)^{-1}_{\widehat{\alpha}} \,, \tag{17.26b}$$

where $\partial\chi^2/\partial\alpha$ is a k-element vector whose n^{th} element is $\partial\chi^2/\partial\alpha_n$, $\partial^2\chi^2/\partial\alpha^2$ is a $k \times k$ matrix with mn^{th} element $\partial^2\chi^2/(\partial\alpha_m \cdot \partial\alpha_n)$, and all derivatives are to be evaluated at the points indicated. If "χ^2" is a true χ^2, the second-derivative matrix is independent of $\boldsymbol{\alpha}$; therefore the shape of the χ^2 as a function of $\boldsymbol{\alpha}$ is a paraboloid and Eq. (17.26a) will give the solution immediately. Otherwise one may need to iterate Eq. (17.26a) to arrive at a solution (Newton-Raphson method). The CERN program MINUIT [11] offers several iteration schemes for solving such problems.

Note that in Eq. (17.16b), one needs only a matrix proportional to V_y to find $\widehat{\boldsymbol{\alpha}}$. Hence, for example, if the variances σ_i^2 of the errors are unknown but assumed equal and independent, and $E(\epsilon_i) = 0$, one can still solve for $\widehat{\boldsymbol{\alpha}}$. One cannot, however, solve for $V_{\widehat{\alpha}}$ or evaluate goodness-of-fit. These can be estimated from the *residuals*, $r_i = \widehat{y}\,(x_i) - y_i$, where $\widehat{y}\,(x_i)$ is the fitted curve at x_i, because study of the r_i enables one to estimate V_y. In addition, the residuals can be used to look for evidence of bias such as trends in the data not incorporated in the model [4].

The errors on the solution $\widehat{\boldsymbol{\alpha}}$ are independent of the value of χ^2 at minimum—they depend only upon the shape about the minimum. Eq. (17.26b) implies that s-standard-deviation limits on the elements of $\widehat{\boldsymbol{\alpha}}$ are given by the set of $\boldsymbol{\alpha}'$ such that

$$\chi^2(\boldsymbol{\alpha}') = \chi^2_{\rm min} + s^2 \,; \tag{17.27}$$

(This is a special case of Eq. (17.8).) This equation, which defines a contour in $\boldsymbol{\alpha}$-space, is often convenient for estimating errors in applications of least-squares techniques to nonlinear cases, where the second derivative [Eq. (17.26b)] may be a rapidly varying function of $\boldsymbol{\alpha}$. If the problem is highly nonlinear, all such contours are, at best, only approximations to desired exact confidence regions which would have some given probability of covering the true value of $\boldsymbol{\alpha}$. It may be that Eq. (17.27) will define a set of disjoint regions. In addition, iteration of Eq. (17.26a) may require sophisticated techniques [8,11] to reach convergence in a practical amount of computation. For example, in cases involving many variables in $\boldsymbol{\alpha}$, especially if the correlations are large, simplex or other techniques which do not involve explicit calculation of derivatives are often to be preferred. Such techniques are designed to find their way through complicated nonlinear problems without diverging to infinite $\boldsymbol{\alpha}$ (unless the minimum is actually at infinity).

The method of least squares is sometimes used in cases where the distribution is not Gaussian or not known to be Gaussian. In such cases it can still be used, but it is then not a special case of the maximum-likelihood method, and the theorems having to do with that approach no longer apply. However, if (a) the distribution of $y_i - \sum \alpha_k f_k(x_i)$ has an expectation value of zero (unbiased) and (b) has a finite, known, fixed variance σ_i^2 (does not depend on $\boldsymbol{\alpha}$), then estimates of $\boldsymbol{\alpha}$ obtained by minimizing χ^2 will be unbiased and have the smallest possible variance of all linear unbiased estimates (Gauss-Markov theorem). This statement is more general than the least-squares method as a special case of the maximum likelihood method in that the distributions do not have to be Gaussian, but more restrictive in that it applies only when the fitting function is linear in the α_k's.