Some Comments on χ^2 Minimisation Applications

V. Blobel
Institut für Experimentalphysik, Universität Hamburg, Germany

The determination of parameters in fits to measured data is a standard task of data analysis. The popular method called χ^2 minimization, as used in recent publications in a wide range of applications, is analysed and compared to standard statistical methods for parameter estimation, the method of least squares and the maximum likelihood method, which have certain optimal statistical properties.

1. INTRODUCTION

χ^2 minimisation. The determination of parameters in fits to measured data is a standard task of data analysis. The standard method of least squares is often referred to as χ^2 minimisation, which is a confusion in terminology; the minimum of the least squares sum follows often, but not always, the χ^2 distribution. In "χ^2 minimization" used in a wide range of applications from calorimeter calibration to complex analyses like fitting parton densities using data from different experiments ("*. . . to determine these parameters one must minimise a χ^2 which compares the measured values . . . to the calculated ones . . . .*") a variety of different non-standard concepts is used, often motivated by serious problems to handle the experimental data in a consistent way; these methods as used in recent applications may result in a bias of the fitted parameter values. Two examples showing common mistakes in the construction of χ^2-expressions are discussed below.

Calorimeter calibration. Calorimeters for energy measurements in a particle detector require a calibration, usually based on data taken with a fixed beam energy E. The measured data y_{jk} for calorimeter cell j in event k (total N events) have to be related to this known energy E. A method used in many experiments is based on the minimisation of the expression

$$\chi^2 = \frac{1}{N} \sum_{k=1}^{N} \left(a_1 y_{1,k} + a_2 y_{2,k} + \ldots + a_n y_{n,k} - E \right)^2$$

for the determination of the a_j, and this can produce biased results, as pointed out by D. Lincoln at al.[1]. To simplify the discussion single cell measurements y_k are assumed with standard deviation σ, with a mean value from N measurements of $\bar{y} = \sum_k y_k / N$; the intended result for the calibration factor is simply $a = E/\bar{y}$. The one-cell version of the above χ^2 definition

$$\chi^2 = \frac{1}{N} \sum_{k=1}^{N} \left(a \cdot y_k - E \right)^2$$

would produce the biased result $a = E \cdot \bar{y}/(\bar{y}^2 + \sigma^2) \neq E/\bar{y}$; the bias mimics a non-linear response of the calorimeter (a *known* bias can of course be corrected for). Using a fixed factor $1/\sigma^2$ instead of the unconventional $1/N$ would give the identical result. There would be no bias, if either a factor $1/(a \cdot \sigma)^2$ would be introduced, or if the inverse constant a_{inv} would have been determined from a modified χ^2 expression with $(y_k - a_{\text{inv}} \cdot E)$ instead of $(a \cdot y_k - E)$.

Normalisation errors. In several publications with χ^2 minimisation it is mentioned that normalisation errors can produce biased fit results ("*. . . that including normalisation errors in the correlation matrix will produce a fit which is biased towards smaller values . . .*"). This effect is described [2] for the data $y_1 = 8.0 \pm 2\%$ and $y_2 = 8.5 \pm 2\%$, with a common (relative) normalisation error of $\varepsilon = 10\%$. Assuming that the true values for both data values are identical the mean value $\bar{y}$ calculated by χ^2 minimisation in the paper is

$$\bar{y} = 7.87 \pm 0.81 \qquad \text{i.e. } \bar{y} < y_1 \text{ and } < y_2$$

– this is apparently wrong. This result has been obtained by minimising

$$\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta} \qquad \text{with} \quad \boldsymbol{\Delta} = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \end{pmatrix}$$

using the covariance matrix

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 + \varepsilon^2 y_1^2 & \varepsilon^2 y_1 y_2 \\ \varepsilon^2 y_1 y_2 & \sigma_2^2 + \varepsilon^2 y_2^2 \end{pmatrix} , \qquad (1)$$

which should include the common normalisation error.

The discussion in papers attributes the problem to the least squares method; however the origin of the wrong result is in fact the above definition (1) of the covariance matrix: the contribution to $\boldsymbol{V}$ from the normalization error was calculated from the *measured* values, which were different; the result is a covariance ellipse with axis different from 45° and this produces a biased mean value, as can be seen in Figure 1. According to the assumption both true data values are identical and then the normalisation error contribution has to be $\varepsilon^2 \bar{y}^2$ for all elements, and the correct mean value is obtained (axis of the covariance ellipse at 45°). Another method leading to the correct result is the introduction of the normalisation factor α as an additional measured value and using the χ^2 definition

$$\chi^2 = \sum_k \frac{(y_k - \alpha \cdot \bar{y})^2}{\sigma_k^2} + \frac{(\alpha - 1)^2}{\varepsilon^2} .$$

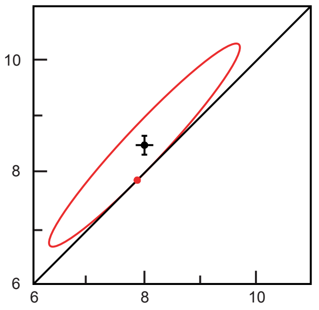

Figure 1: Measured point $(y_1, y_2) = (8.0, 8.5)$ and covariance ellipse according to the definition in equation (1). The slightly tilted ellipse touches the diagonal $y_1 = y_2$ at a point which is below both data points.

2. STANDARD METHODS

Least Squares. Doubts are raised in publications with χ^2 minimisation about the applicability of the method ("*the justification for using least squares lies in the assumption that the measurement errors are Gaussian distributed ... it is doubtful that Gaussian errors are realistic.*"). Arbitrary factors are often applied to increase the parameter errors (see section 3).

For the standard least squares method the properties of the result can be derived from certain conditions. The linear least squares problem is denoted by $\boldsymbol{A}\boldsymbol{a} \cong \boldsymbol{y}$. Given a $n \times p$ matrix $\boldsymbol{A}$ and given a n vector $\boldsymbol{y}$ with covariance matrix $\boldsymbol{V}_y$ the problem is to find the p vector $\boldsymbol{a}$ of parameters which minimises ($\boldsymbol{W} = \boldsymbol{V}_y^{-1}$)

$$S(\boldsymbol{a}) = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a})^T \; \boldsymbol{W} \; (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a})$$

with respect to $\boldsymbol{a}$. The solution (from $\partial S/\partial \boldsymbol{a} = 0$) is a linear transformation of the data vector $\boldsymbol{y}$

$$\hat{\boldsymbol{a}} = \left[\left(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} \right)^{-1} \; \boldsymbol{A}^T \boldsymbol{W} \right] \boldsymbol{y} \qquad = \boldsymbol{B}\,\boldsymbol{y} \; ,$$

the covariance matrix of vector $\hat{\boldsymbol{a}}$ is given by standard error propagation:

$$\boldsymbol{V}_a = \boldsymbol{B}\,\boldsymbol{V}_y\,\boldsymbol{B}^T = \left(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} \right)^{-1} \qquad (2)$$

and this relation does not depend on the "*quality of the fit*". Properties of the solution are derived under certain conditions: the data unbiased, i.e. $E[\boldsymbol{y}] = \boldsymbol{A}\,\bar{\boldsymbol{a}}$ ($\bar{\boldsymbol{a}}$ = true parameter vector), and the covariance matrix $\boldsymbol{V}_y$ of the data known (and correct). Distribution-free properties of least squares estimates in linear problems are: the estimated parameters are unbiased, and in the class of unbiased estimates, which are linear in the data, the least squares estimates $\hat{\boldsymbol{a}}$ have the smallest variance (Gauß-Markoff theorem). The expectation of the sum of squares of the residuals is $\hat{S} = (n - p)$. However the distribution of $\hat{S}$ follows the χ^2 distribution with $(n - p)$ degrees of freedom *only* in the case of Gaussian distributed data. For non-linear problems the above properties are only approximately valid.

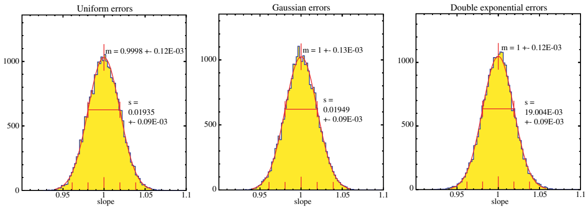

Figure 2: Distribution of the slope parameter in MC simulated least squares fits of straight lines. The data distributions were uniform (left), Gaussian (center) and double-exponential (right). The width of the parameter distributions are as expected from equation (2).

The Figure 2 shows the distribution of the slope parameter in 25 000 MC simulations of straight-line fits ($n = 20$ data points) with different data distributions: the uniform distribution (left), the Gaussian distribution (center) and the double-exponential distribution(right). In all these cases (*same* standard deviation of the data) a Gaussian distribution of the slope parameter (and the intercept) is observed (central limit theorem!), although the input data distribution are different and especially have very different tails. In addition the mean value of $\hat{S}$ is $(n-p) = 20-2$ in all three cases, but the distribution of the corresponding P-values (calculated from the observed χ^2 and $(n - p)$) is uniform only in the case of Gaussian-distributed data, as expected.

Likelihood function and Information. The maximum likelihood method can be used, if the details about the distribution of the data are known and the likelihood function $\mathcal{L}(\boldsymbol{a})$ of the problem can be constructed. In the case of several parameters $a_1, a_2 \ldots a_p$ a $p \times p$ symmetric information matrix $\boldsymbol{I}$ with elements determined by the expectation values

$$I_{jk} = E\left[\frac{\partial \ln \mathcal{L}}{\partial a_j} \frac{\partial \ln \mathcal{L}}{\partial a_k} \right] = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial a_j \partial a_k} \right]$$

is defined, and it can be shown that the minimal variance $\boldsymbol{V}_{\hat{\boldsymbol{a}}}$ of an estimate $\hat{\boldsymbol{a}}$ is given by the inverse of the information matrix: $\boldsymbol{V}_{\hat{\boldsymbol{a}}} = \boldsymbol{I}^{-1}$. In practice the negative log likelihood function is defined as the objective function $F(\boldsymbol{a}) = -\ln \mathcal{L}(\boldsymbol{a})$ and the minimum w.r.t. $\boldsymbol{a}$ is determined by the condition $\boldsymbol{g} = \partial F/\partial a_j = 0$. In case of good statistic the Hessian $\boldsymbol{H}$, the matrix of second derivatives of $F(\boldsymbol{a})$, is almost constant in the region around the minimum and is a good estimate for the information matrix $\boldsymbol{I}$; the inverse $\boldsymbol{H}^{-1}$ is a good estimate of the covariance matrix $\boldsymbol{V}_{\hat{\boldsymbol{a}}}$ of the parameters $\hat{\boldsymbol{a}}$. This corresponds (like eq. (2)) to standard error propagation from the data errors to the parameter errors, especially there is no freedom to introduce additional factors. Objective functions from the maximum likelihood and the least squares method can be combined (e.g. $F(\boldsymbol{a}) + 1/2\, S(\boldsymbol{a})$).

The minimisation and the calculation of the covariance matrix require the inversion of the Hessian. The introduction of redundant parameters in χ^2 minimisation ("*. . . we found our input parameterisation was sufficiently flexible to accommodate data and indeed there is a certain redundancy evident . . .*") in a fit problem can be rather dangerous, because with redundant parameters the Hessian is singular and without special techniques neither the minimum of the objective function can be found nor the inverse can be calculated.

3. DATA AND PARAMETER ERRORS

Systematic errors. The statistical and systematic uncertainties of the data can only be correctly taken into account in a fit if there is a clear model describing all aspects of the uncertainties. For statistical errors the strategy is usually well-defined: they can be described by the standard deviations (uncorrelated data points), or by a covariance matrix. The latter is necessary if e.g. corrections for finite resolution are applied to the data. For contributions due to systematic errors there are two alternative models: multiplicative effects due to normalisations errors and additive effects due to offset errors have to be treated differently.

In general the normalisation uncertainty is given by a relative error, and in fits with data from more than one experiment the treatment of the normalisation error may be important. Instead of adding a contribution to the covariance matrix one additional parameter α can be introduced for each experiment (measured value $\alpha = 1 \pm \varepsilon$) and the expectation $f(x_i, \boldsymbol{a})$ for y_i is modified to $\alpha \cdot f(x_i, \boldsymbol{a})$, leaving the measured data point y_i unchanged:

$$S(\boldsymbol{a}) = \sum_i \frac{(y_i - \alpha \cdot f(x_i, \boldsymbol{a}))^2}{\sigma_i^2} \ + \ \Delta S^{\text{norm}}$$

with $\Delta S^{\text{norm}} = (\alpha - 1)^2 / \varepsilon^2$. The normalisation factor determined in an experiment is more the *product* than the sum of random variables. According to the *multiplicative* central limit theorem the product of positive random variables follows the log-normal distribution, i.e. the logarithm of the normalisation factor follows the normal distribution. For a log-normal distribution of a random variable α with $E[\alpha] = 1$ and standard deviation of ε the contribution to $S(\boldsymbol{a}, \alpha)$ is (from the likelihood function)

$$\Delta S^{\text{norm}} = \ln \alpha \left(3 + \frac{\ln \alpha}{\ln (1 + \varepsilon^2)} \right)$$

for each α, which reduces to the previous term for small deviations of the value α from 1.

An example for an additive error is the error of a calorimeter constant – a change of the constant will change *all* data values y_i, because events are moved between bins. Here one has to determine shifts s_i of data values y_i, for a one-standard deviation change of the calorimeter constant; the shifts s_i will carry a relative sign. The error could be taken into account by adding the rank=1 matrix $\boldsymbol{s}\boldsymbol{s}^T$ to the covariance matrix. Alternatively one additional parameter β can be introduced for each error contribution (measured value $\beta = 0 \pm 1$), and the expectation can be modified to $f(x_i, \boldsymbol{a}) + \beta \cdot s_i$:

$$S(\boldsymbol{a}) = \sum_i \frac{(y_i - (f(x_i, \boldsymbol{a}) + \beta s_i))^2}{\sigma_i^2} \ + \ \beta^2 \ .$$

The introduction of additional parameters α and β allow to see the effect of the systematic errors in the fit, including the correlation of the parameter to other parameters in the fit, and of the pull, which has an expected mean of zero and variance of 1. The pull is the ratio of the shift of a parameter value in the fit, divided by the standard deviation of the *shift* (not the original standard deviation). In the case of a parameter β introduced above the pull is $\hat{\beta}/\sqrt{1 - V_{\beta\beta}}$. Rather useful for checks of parameter correlations is the global correlation coefficient ρ_k,

$$\rho_k = \sqrt{1 - \frac{1}{(\boldsymbol{V})_{kk} \cdot \left(\boldsymbol{V}^{-1}\right)_{kk}}} \ ,$$

which is a measure of the total amount of correlation between the k-th parameter and *all* the other variables. It is the largest correlation between the k-th parameter and every possible linear combination of all the other variables.

Different expressions are used in publications using χ^2 minimisation. One example is called the offset method, where systematic errors are ignored in the fit ("*. . . forces the theory prediction to be as close as possible to the data . . .*"), but later added in quadrature in the error calculation. It is clear that the fit result must be biased, if incomplete error information is used.

Parameter errors. With the given definition of the fit expression and the error contributions there is no freedom in standard methods in the calculation of the parameter errors; they are the result of error propagation and the parameter covariance matrix is the inverse of the Hessian $\boldsymbol{H}$.

In publications using χ^2 minimisation the following statements are found: "*Notice that the covariance matrix*

$$V_{ij}^p = \langle \Delta_i \Delta_j \rangle = \Delta\chi^2 \cdot H_{ij}^{-1}$$

depends on the choice of $\Delta\chi^2$ which usually, but not always, is taken to be $\Delta\chi^2 = 1$. This choice

Table I The values of the parameter $\alpha_S(M_Z^2)$, obtained in different structure function analyses and the value of $\Delta\chi^2$, used in the error calculation. The column marked # gives the number of experiments used in the analysis.

Group	$\Delta\chi^2$	Ref.	#	Value of $\alpha_S(M_Z^2)$
H1	1	[4]	2	0.115 ± 0.0017 (exp) $^{+0.0009}_{-0.0005}$ (model) ± 0.005 (theory)
GKK	1	[5]	3	0.112 ± 0.001 (exp)
MRST02	20	[6]	many	0.1195 ± 0.002 (exp) ± 0.003 (theory)
ZEUS	50	[7]	several	0.1166 ± 0.0049 (exp) ± 0.0018 (model)
CTEQ6	100	[8]	several	0.1165 ± 0.0065 (exp)

... corresponds to the definition of the width of a Gaussian distribution." [9] "*Ideally $\Delta\chi^2 = 1$, but unrealistic.*" " *... and $\Delta\chi^2$ is the allowed deterioration in fit quality for the error determination.*" [6]. The freedom taken in this unconventional error definition is shown in the overview of table I, which shows the tendency to use a value $\Delta\chi^2$ larger than 1 in analyses where a large number of different experiments is combined. A value of $\Delta\chi^2 = 100$ is equivalent to multiplying all input errors by a factor of 10, and this procedure is not justified by the "χ^2-value" of the fit alone. In one of the cases this value is 2328 for 2097-15 = 2082 degrees of freedom; in standard methods all data errors would be increased by $\sqrt{\chi^2/n_{\mathrm{df}}} = 1.06$. The large, artificial and arbitrary magnification of errors points to severe problem of the whole data analysis.

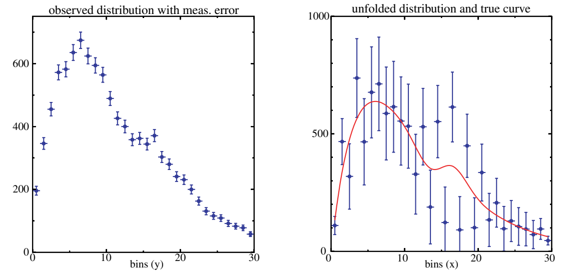

Figure 3: The measured distribution (left) and the result of unfolding by matrix inversion (right), showing large errors due to the negative correlations between adjacent data points.

4. STATISTICAL DATA PROPERTIES

Where is the origin of the problems apparent in the combined analysis of many experiments (see table I)? "*Indeed, we have always believed the theory, rather than experiment, will provide the dominant source of error.*" [6] However the origin seems to be on both the theoretical and the experimental side. Recent publications from experiments contain a lot of information on various types of errors, but this may be still insufficient for global fits.

The finite resolution in the measurement of kinematical variables requires in principle an unfolding procedure. Instead of measuring the true distribution $f(x)$ of a kinematical quantity x the distribution $g(y)$ of a quantity y is measured, which is related to the true variable x by a resolution function $A(y,x)$, known only implicitly by a sample of MC generated events:

$$g(y) = \int_\Omega A(y,x) f(x) \mathrm{d}x \qquad \text{or short } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} ,$$

The "correction factor"-method used by most of the experiments seems to introduce a hidden positive correlation between the data points; the method is usually not described by giving mathematical formulas, but in words like the following text taken from an early structure function measurement: " *... The main problem of the analysis is the correction for measurement errors (unsmearing corrections), which are large at large x where the structure functions vary rapidly with x. We proceed by assuming a true structure function and calculate by Monte Carlo simulation, on the basis of the known experimental resolution functions, the result to be expected in the apparatus. By iteration a true distribution which reproduces the experimental result is found. The unsmearing factor is the ratio of Monte Carlo events for any particular (x, Q^2) bin in the true distribution divided by those in the resolution smeared distribution. If this factor differs from unity by more than 30 %, the bin is not retained ...*". [3] Often the result from a fit to a previous measurement is taken for the MC simulation, which may introduce a bias and this is certainly not a *blind analysis.* The unavoidable correlation between corrected data points is usually neglected, thus giving a too large weight to the experimental data in a later fit.

The method quoted above is usually applied because attempts to solve the problem by standard methods fail. This is illustrated in the simulated data of Figure 3. Figure 3a shows the 30-bins histogram of a distribution measured using 10 000 events, assuming a migration parameter $\varepsilon = 0.24$, which is the probability for the migration into both adjacent bins (the probability of measuring the entry in the correct bin is $(1-2\varepsilon)$). Using the symmetric migration matrix $\boldsymbol{A}$ the unfolded result $\boldsymbol{x}$ can be obtained by the solution of the equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$; the result in Figure 3b shows large fluctuations of the unfolded distribution.

An improved solution, derived from the properties

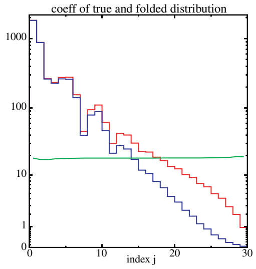

Figure 4: Absolute values of the elements of the transformed vectors $\boldsymbol{b} = \boldsymbol{U}^T\boldsymbol{x}$ and $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y}$ (without measurement errors) (left) and of the vector $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y}$ with measurement errors (right). The error level is shown as a line in both figures.

of the resolution matrix $\boldsymbol{A}$ alone, is based on the orthogonal decomposition $\boldsymbol{A} = \boldsymbol{UDU}^T$ of the migration matrix $\boldsymbol{A}$. Multiplying the original matrix equation

$$\boldsymbol{y} \cong \boldsymbol{A}\boldsymbol{x} = \boldsymbol{UDU}^T\boldsymbol{x}$$

by $\boldsymbol{U}^T$ from the left, one obtains ($\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{1}$)

$$\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y} \cong \boldsymbol{D}\left(\boldsymbol{U}^T\boldsymbol{x}\right) = \boldsymbol{D}\boldsymbol{b} \qquad \boldsymbol{b} = \boldsymbol{D}^{-1}\boldsymbol{c}\,.$$

The transformed measurement vector $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y}$ allows one to calculate, by $\boldsymbol{b} = \boldsymbol{D}^{-1}\,\boldsymbol{c}$, the elements of the transformed result vector $\boldsymbol{b} = \boldsymbol{U}^T\boldsymbol{x}$ element by element, because the matrix $\boldsymbol{D}$ is diagonal (the diagonal elements are the eigenvalues of the symmetric matrix $\boldsymbol{A}$). Figure 4a shows the spectrum of elements for the true distribution before and after folding with the resolution matrix; the first elements are almost not affected by the resolution, but the *smoothing* effect of folding is clearly visible in the second half of the elements in a reduction of the size. The horizontal line shows the 1-standard deviation level of the measurement with the given number of events. It is clear that the second half of the elements can not be measured. The actual measurement is shown in Figure 4b. Only the first half of the elements represent a real measurement; for the second half of the elements the measured values corresponds to the average error level.

Taking the first 15 elements only and transforming the vector $\boldsymbol{b}$ back to the original bins by $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{b}$ the result of Figure 5a with 30 bins is obtained, which is close to the curve representing the original dependence. Since this rather smooth result has been obtained from 15 measured elements, the rank of the covariance matrix can only be 15, and the correlations between adjacent bins are positive and large, like in the “correction factor”-method. A more acceptable result shown in Figure 5b is obtained by averaging each two (positively correlated) neighbor bins; the standard deviations of the data points are almost unchanged, but now the covariance matrix is of full-rank with small correlations. The severe problems observed in globals fits may be caused by retaining far too many data points with hidden large positive correlations from the “correction factor”-method, which appear to be more precise than they are.

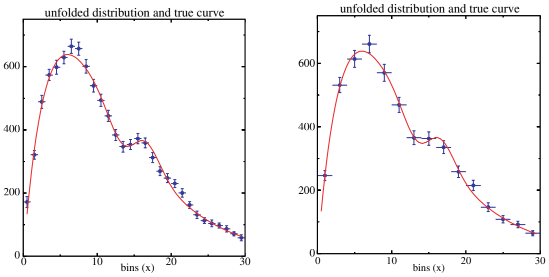

Figure 5: Result of unfolding with a cut-off after 15 elements. The left figure shows all 30 data points, which have a singular (rank-15) covariance matrix. The right figure show the result after combining pairs of two data points to one point; the covariance matrix has full rank.

References

[1] D. Lincoln, G. Morrow and P. Kaspar, A hidden bias in a common calorimeter calibration scheme, NIM **A 345**, 449 (1994)

[2] G. D’Agostini, On the use of the covariance matrix to fit correlated data, NIM **A 346**, 306 (1994)

[3] J. G. H. de Groot et al., Inclusive interactions of high-energy neutrinos and antineutrinos in iron, Zeitschrift für Physik **C** 1, 143 (1979)

[4] C. Adloff et al. (H1 Collaboration), Deep-inelastic inclusive ep-scattering at low x and a determination of α_s, Eur. Phys. J. C 21, 33 (2001)

[5] W. T. Giele and S. Keller, Implications of hadron collider observables on parton distributtion function uncertainties, Phys. Rev. D **58**, 094023 (1998), `hep-ph/9803393`

[6] A.D. Martin, R.G. Roberts, W.J. Stirling and R.S. Thorne, Uncertainties of predictions from parton distributions I: Experimental errors, Cavendish-HEP-2002/10, `hep-ph/0211080` (2002)

[7] S. Chekanov et al. (ZEUS Collaboration), ZEUS next-to-leading-order QCD analysis of data on deep-inelastic scattering, Phys. Rev. D **67**, 012007 (2003)

[8] J. Pumplin et al., New generation of parton distributions with uncertainties from global QCD analysis, JHEP 0207:012 (2002)

[9] M. Botje, Error Estimates on Parton Density Distributions, NIKHEF-01-014, `hep-ph/0110123` (2001)