Nuclear Instruments and Methods in Physics Research A 346 (1994) 306–311
North-Holland

NUCLEAR
INSTRUMENTS
& METHODS
IN PHYSICS
RESEARCH
Section A

On the use of the covariance matrix to fit correlated data

G. D'Agostini
Dipartimento di Fisica, Università "La Sapienza" and INFN, Roma, Italy

(Received 10 December 1993; revised form received 18 February 1994)

Best fits to data which are affected by systematic uncertainties on the normalization factor have the tendency to produce curves lower than expected if the covariance matrix of the data points is used in the definition of the χ^2. This paper shows that the effect is a direct consequence of the hypothesis used to estimate the empirical covariance matrix, namely the linearization on which the usual error propagation relies. The bias can become unacceptable if the normalization error is large, or a large number of data points are fitted.

1. Introduction

It is frequently the case that one has to fit a theoretical curve through experimental data affected by overall systematic errors, often just a common uncertainty on the normalization factor. If the error matrix $\mathbf{V}$ of the data points is known, one can solve the problem by minimizing the χ^2, defined as

$$\chi^2 = \underline{\Delta}^{\mathrm{T}} \mathbf{V}^{-1} \underline{\Delta}, \tag{1}$$

where $\underline{\Delta}$ is the vector of the differences between the theoretical and the experimental values.

In performing this kind of fit it is not uncommon to obtain results that contradict expectations. To give a numerical example, let us consider the results of two measurements, $8.0 \pm 2\%$ and $8.5 \pm 2\%$, having a 10% common normalization error (see Fig. 1). Assuming that the two measurements refer to the same physical quantity, the best estimate of its true value can be obtained by fitting the points to a constant. Minimizing χ^2 as defined in Eq. (1), with $\mathbf{V}$ estimated empirically by the data, one obtains a value of 7.87 ± 0.81, which is at least surprising, since the most probable result is outside the interval determined by the two measured values.

A real example of this strange effect happened during the global analysis of the R ratio in e^+e^- performed by the CELLO collaboration [1], shown in Fig. 2. The data points represent the averages, in energy bins, of the results of the PETRA and PEP experiments. They are all correlated and the error bars show the total error (see ref. [1] for details). In particular, at the intermediate stage of the analysis shown in the figure, an overall 1% systematic error due to theoretical uncertainties was included in the covariance matrix. The R values above 36 GeV show the first hint of the rise of the e^+e^- cross section due to the Z^0 pole. It was at that time very interesting to prove that the observation was not just a statistical fluctuation. In order to test this, the data were fitted with a theoretical function having *no* Z^0 contributions and using only the data below a certain energy. The expectation was to observe a fast increase of χ^2/ν, where ν is the number of degrees of freedom, above 36 GeV, indicating that a theoretical prediction without Z^0 would be inadequate to describe the high energy data. The surprising result was a "repulsion" (see Fig. 2) between the experimental data and the fit: including the high energy points with larger R, a lower curve was obtained, while χ^2/ν remained almost constant.

It will be shown in this paper that such an effect, which appears if a sizeable normalization uncertainty is common to a data sample, originates from the standard way of performing the error propagation, where only first derivatives are considered. In order to get analytical results, the simple case of only two data points will be considered. Since the conclusions are based on the empirical covariance matrix of the experimental points, it will first be shown how to build it in the most general case, since this problem is usually not discussed in books of statistics [#1].

[#1] Apart from ref. [1], the only text book known to the author, where the construction of the covariance matrix from experimental data related by common errors is discussed, is the recent one by Barlow [2]. A more complete treatment is given in the DIN norms [3].

SSDI 0168-9002(94)00272-9

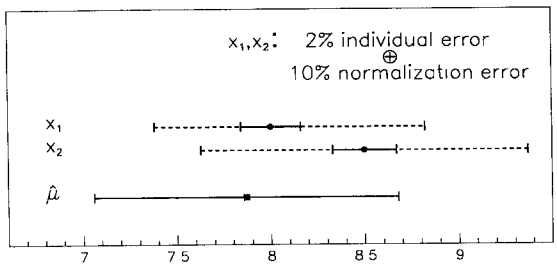

Fig. 1. Best estimate of the true value from two correlated data points, using in the χ^2 the empirical covariance matrix of the meaurements. The error bars show individual and total errors.

2. Covariance matrix of correlated data

In physics applications, it is rarely the case that the covariance between the best estimates of two physical quantities [#2], each given by the arithmetic average of direct measurements ($x_i = \bar{X}_i = 1/n\Sigma_{k=1}^{n} X_{ik}$), can be evaluated from the sample covariance of the two averages

$$\mathrm{Cov}(x_i, x_j) = \frac{1}{n(n-1)} \sum_{k=1}^{n} \left(X_{ik} - \bar{X}_i\right)\left(X_{jk} - \bar{X}_j\right). \tag{2}$$

More frequent is the well understood case in which the physical quantities are obtained as a result of a χ^2 minimization, and the terms of the inverse of the error matrix are related to the curvature of χ^2 at its minimum

$$(V^{-1})_{ij} = \frac{1}{2} \left.\frac{\partial^2 \chi^2}{\partial X_i \, \partial X_j}\right|_{x_i, x_j}.$$

In most cases one determines independent values of physical quantities with the same detector, and the correlation between them originates from the detector calibration errors. Conceptually, the use of Eq. (2) in this case would correspond to having a "sample of detectors", with each of which a measurement of all the physical quantities is to be performed.

A way to build the covariance matrix from the direct measurements is to consider the original measurements and the calibration constants as a common set of independent and uncorrelated measurements, and then to calculate corrected values that take into account the calibration constants. The error propagation will provide automatically the full covariance matrix of the set of results. Let us derive it for two cases that happen frequently, and then proceed to the general case.

2.1. Offset error

Let $x_i \pm \sigma_i$ be the $i = 1, 2, \ldots, n$ results of independent measurements and $\mathbf{V}_X$ the (diagonal) error matrix. Let assume that they are all affected by the same calibration constant c, having an error σ_c. The corrected results are then $y_i = x_i + c$. We can assume, for simplicity, that the most probable value of c is 0, i.e. the detector is well calibrated. One has to consider the calibration constant as the physical quantity X_{n+1}, the best estimate of which is $x_{n+1} = 0$. A term $V_{X_{n+1,n+1}} = \sigma_c^2$ must be added to the error covariance.

The covariance matrix of the corrected results is given by the transformation

$$\mathbf{V}_Y = \mathbf{M}\mathbf{V}_X\mathbf{M}^{\mathrm{T}},$$

where $M_{ij} = \partial Y_i / \partial X_j \,|_{x_j}$. The elements of $\mathbf{V}_Y$ are given by

$$V_{Y_{kl}} = \sum_{ij} \left|\frac{\partial Y_k}{\partial X_i}\right|_{x_i} \left|\frac{\partial Y_l}{\partial X_j}\right|_{x_j} V_{X_{ij}}.$$

In this case we get

$$\sigma^2(Y_i) = \sigma_i^2 + \sigma_c^2,$$

$$\mathrm{Cov}(Y_i, Y_j) = \sigma_c^2 \qquad (i \neq j).$$

The total error on the single measurement is given by the combination in quadrature of the individual and the common error, and all the covariances are equal to σ_c^2. To verify, in a simple case, that the result is reasonable, let us consider only two independent quan-

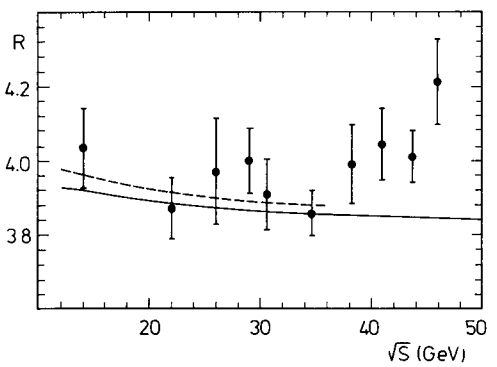

Fig. 2. R measurements from PETRA and PEP experiments with the best fits of QED + QCD to all the data (full line) and only below 36 GeV (dashed line). All data points are correlated (see text).

[#2] Hereafter the symbol X_i will indicate the variable associated to the ith physical quantity and X_{ik} its kth direct measurement; x_i the best estimate of its value, obtained by an average over many direct measurements or indirect measurements, σ_i the standard deviation, and y_i the value corrected for the calibration constants. The weighted average of several values x_i will be denoted by $\bar{x}$.

tities X_1 and X_2, and a calibration constant $X_3 = c$, having an expected value equal to zero. From these we can calculate the correlated quantities Y_1 and Y_2 and finally their sum ($S \equiv Z_1$) and difference ($D \equiv Z_2$). The results are

$$\mathbf{V}_Y = \begin{pmatrix} \sigma_1^2 + \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_2^2 + \sigma_c^2 \end{pmatrix},$$

$$\mathbf{V}_Z = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 + 4\sigma_c^2 & \sigma_1^2 - \sigma_2^2 \\ \sigma_1^2 - \sigma_2^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix}.$$

It follows that

$$\sigma^2(S) = \sigma_1^2 + \sigma_1^2 + (2\sigma_c)^2,$$

$$\sigma^2(D) = \sigma_1^2 + \sigma_2^2,$$

as intuitively expected.

2.2. Normalization error

Let us consider now the case where the calibration constant is the scale factor f, known with an error σ_f. Also in this case, for simplicity and without losing generality, let us suppose that the most probable value of f is 1. Then $X_{n+1} = f$, i.e. $x_{n+1} = 1$, and $V_{X_{n+1,n+1}} = \sigma_f^2$. Then

$$\sigma^2(Y_i) = \sigma_i^2 + \sigma_f^2 x_i^2,$$

$$\mathrm{Cov}(Y_i, Y_j) = \sigma_f^2 x_i x_j \qquad (i \neq j).$$

To verify the results let us consider two independent measurements X_1 and X_2, let us calculate the correlated quantities Y_1 and Y_2, and finally their product ($P \equiv Z_1$) and their ratio ($R \equiv Z_2$):

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 + \sigma_f^2 x_1^2 & \sigma_f^2 x_1 x_2 \\ \sigma_f^2 x_1 x_2 & \sigma_2^2 + \sigma_f^2 x_2 \end{pmatrix},$$

$$\mathbf{V}_Z = \begin{pmatrix} \sigma_1^2 x_2^2 + \sigma_2^2 x_1^2 + 4\sigma_f^2 x_1^2 x_2^2 & \sigma_1^2 - \sigma_2^2 \dfrac{x_1^2}{x_2^2} \\ \sigma_1^2 - \sigma_2^2 \dfrac{x_1^2}{x_2^2} & \dfrac{\sigma_1^2}{x_2^2} + \sigma_2^2 \dfrac{x_1^2}{x_2^4} \end{pmatrix}.$$

It follows that

$$\sigma^2(P) = \sigma_1^2 x_2^2 + \sigma_2^2 x_1^2 + (2\sigma_f x_1 x_2)^2,$$

$$\sigma^2(R) = \frac{\sigma_1^2}{x_2^2} + \sigma_2^2 \frac{x_1^2}{x_2^4}.$$

Just as a common offset error cancels in differences and is enhanced in sums, a normalization error has a similar effect on the ratio and the product. It is also interesting to calculate the error on a difference in case of a normalization error:

$$\sigma^2(D) = \sigma_1^2 + \sigma_2^2 + \sigma_f^2 (x_1 - x_2)^2.$$

The contribution from the normalization error vanishes if the two values are equal.

2.3. General case

Let us assume there are n independent measured values x_i and m calibration constants c_j with their covariance matrix $\mathbf{V}_c$. The latter can also be theoretical parameters influencing the data, and moreover they may be correlated, as usually happens if they are parameters of a calibration fit. We can then include the c_j in the vector that contains the measurements and $\mathbf{V}_c$ in the error matrix $\mathbf{V}_x$:

$$\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \hline c_1 \\ \vdots \\ c_m \end{pmatrix}, \quad \mathbf{V}_X = \left(\begin{array}{cccc|c} \sigma_1^2 & 0 & \dots & 0 & \\ 0 & \sigma_2^2 & \dots & 0 & \\ \dots & \dots & \dots & \dots & \mathbf{0} \\ 0 & 0 & \dots & \sigma_n^2 & \\ \hline & & \mathbf{0} & & \mathbf{V}_c \end{array}\right).$$

The correlated quantities are obtained from the most general function

$$Y_i = Y_i(X_i, \underline{c}) \quad (i = 1, n),$$

and the covariance matrix $\mathbf{V}_Y$ from the error propagation $\mathbf{V}_Y = \mathbf{M}\mathbf{V}_X\mathbf{M}^{\mathrm{T}}$.

As a frequently encountered example, we can think of several normalization constants, each affecting a subsample of the data – as is the case where each of several detectors each measures a set of physical quantities. For simplicity we can consider only three quantities (X_i) and three uncorrelated normalization errors (σ_{f_j}), the first one common to X_1 and X_2, the second to X_2 and X_3 and the third to all three. We get the following covariance matrix:

$$\begin{pmatrix} \sigma_1^2 + (\sigma_{f_1}^2 + \sigma_{f_3}^2)x_1^2 & (\sigma_{f_1}^2 + \sigma_{f_3}^2)x_1 x_2 & \sigma_{f_3}^2 x_1 x_3 \\ (\sigma_{f_1}^2 + \sigma_{f_3}^2)x_1 x_2 & \sigma_2^2 + (\sigma_{f_1}^2 + \sigma_{f_2}^2 + \sigma_{f_3}^2)x_2^2 & (\sigma_{f_2}^2 + \sigma_{f_3}^2)x_2 x_3 \\ \sigma_{f_3}^2 x_1 x_3 & (\sigma_{f_2}^2 + \sigma_{f_3}^2)x_2 x_3 & \sigma_3^2 + (\sigma_{f_2}^2 + \sigma_{f_3}^2)x_3^2 \end{pmatrix}$$

3. Best estimate of the true value from two correlated values

Once the covariance matrix is built, one can make use of Eq. (1) to estimate the parameters of interest. Let us consider the simple case in which two results of the same physical quantity are available, and the individual and the common errors are known. The best estimate of the true value of the physical quantity is then obtained by fitting the constant $Y = k$ through the data points. In this simple case the χ^2 minimization can be performed easily. We will consider the two

cases of offset and normalization error. As before, we assume that the detector is well calibrated, i.e. the most probable value of the calibration constant is, respectively for the two cases, 0 and 1, and hence $y_i = x_i$.

3.1. Offset error

Let $x_1 \pm \sigma_1$ and $x_2 \pm \sigma_2$ be the two measured values, and σ_c the common error. The χ^2 is

$$\chi^2 = \frac{1}{D}\Big[(x_1-k)^2(\sigma_2^2+\sigma_c^2) + (x_2-k)^2(\sigma_1^2+\sigma_c^2) - 2(x_1-k)(x_2-k)\sigma_c^2\Big],$$

where $D = \sigma_1^2\sigma_2^2 + (\sigma_1^2+\sigma_2^2)\sigma_c^2$ is the determinant of the covariance matrix.

Minimizing χ^2 and using the second derivative calculated at the minimum we obtain the best value of k and its error:

$$\hat{k} = \frac{x_1\sigma_2^2 + x_2\sigma_1^2}{\sigma_1^2+\sigma_2^2} \; (=\bar{x}),$$

$$\sigma^2(\hat{k}) = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} + \sigma_c^2.$$

The most probable value of the physical quantity is exactly what one obtains from the average $\bar{x}$ weighted with the inverse of the individual variances. Its error is the quadratic sum of the error of the weighted average and the common one. The result coincides with the simple expectation.

3.2. Normalization error

Let $x_1 \pm \sigma_1$ and $x_2 \pm \sigma_2$ be the two measured values, and σ_f the common error on the scale. The χ^2 is

$$\chi^2 = \frac{1}{D}\Big[(x_1-k)^2(\sigma_2^2+x_2^2\sigma_f^2) + (x_2-k)^2(\sigma_1^2+x_1^2\sigma_f^2) - 2(x_1-k)(x_2-k)x_1x_2\sigma_f^2\Big],$$

where $D = \sigma_1^2\sigma_2^2 + (x_1^2\sigma_2^2 + x_2^2\sigma_1^2)\sigma_f^2$. We obtain in this case the following result:

$$\hat{k} = \frac{x_1\sigma_2^2 + x_2\sigma_1^2}{\sigma_1^2+\sigma_2^2+(x_1-x_2)^2\sigma_f^2},$$

$$\sigma^2(\hat{k}) = \frac{\sigma_1^2\sigma_2^2 + (x_1^2\sigma_2^2 + x_2^2\sigma_1^2)\sigma_f^2}{\sigma_1^2+\sigma_2^2+(x_1-x_2)^2\sigma_f^2}.$$

With respect to the previous case, $\hat{k}$ has a new term $(x_1-x_2)^2\sigma_f^2$ in the denominator. As long as this is negligible with respect to the individual variances we still get the the weighted average $\bar{x}$, otherwise a smaller value is obtained. Calling r the ratio between $\hat{k}$ and $\bar{x}$, we obtain

$$r = \hat{k}/\bar{x} = \frac{1}{1 + \dfrac{(x_1-x_2)^2}{\sigma_1^2+\sigma_2^2}\sigma_f^2}.$$

Written in this way, one can see that the deviation from the simple average value depends on the compatibility of the two values and on the normalization error. This can be understood in the following way: as soon as the two values are in some disagreement, the fit starts to vary – in a hidden way – the normalization factor and to squeeze the scale, by an amount allowed by σ_f, in order to minimize the χ^2. The advantage for the fit to prefer, under these conditions, normalization factors smaller than 1 finds its deep reason in the standard formalism of the error propagation, where only first derivatives are considered. This implies that the individual errors are not rescaled by lowering the normalization factor, while the points get closer.

To see the source of this effect more explicitly, let us consider an alternative way often used to take into account the normalization uncertainty. A scale factor f, by which all data points are multiplied, is introduced in the expression of χ^2:

$$\chi^2_{\mathrm{A}} = \frac{(fx_1-k)^2}{(f\sigma_1)^2} + \frac{(fx_2-k)^2}{(f\sigma_2)^2} + \frac{(f-1)^2}{\sigma_f^2}. \qquad (3)$$

Let us consider also the same expression when the individual errors are not rescaled:

$$\chi^2_{\mathrm{B}} = \frac{(fx_1-k)^2}{\sigma_1^2} + \frac{(fx_2-k)^2}{\sigma_2^2} + \frac{(f-1)^2}{\sigma_f^2}. \qquad (4)$$

The use of χ^2_{A} always gives the result $\hat{k} = \bar{x}$, because the term $(f-1)^2/\sigma_f^2$ is harmless [#3] as far as the value of the minimum χ^2 and the determination on $\hat{k}$ are concerned. Its only influence is on $\sigma(\hat{k})$, which turns out to be equal to quadratic combination of the weighted average error with $\sigma_f\bar{x}$, the normalization uncertainty on the average. This result corresponds to the usual one, when the normalization factor in the definition of χ^2 is not included, and the overall uncertainty is added at the end.

The use of χ^2_{B} instead is equivalent to the covariance matrix: the same values of the minimum χ^2, of $\hat{k}$ and of $\sigma(\hat{k})$ are obtained, and $\hat{f}$ at the minimum turns

[#3] A simple way to see it is to rewrite Eq. (3) as:

$$\frac{(x_1-k/f)^2}{\sigma_1^2} + \frac{(x_2-k/f)^2}{\sigma_2^2} + \frac{(f-1)^2}{\sigma_f^2}.$$

For any f, the first two terms determine the value of k, and the third one constrains f to 1.

out to be exactly the r ratio defined above. This demonstrates that the effect happens when the data values are rescaled independently of their errors. The effect can become huge in the case where the data show mutual disagreement. The equality of the results obtained with χ^2_B with those obtained with the covariance matrix allows us to study, in a simpler way, the behaviour of $r\ (=\hat{f})$ when an arbitrary amount of data points are analysed. The fitted value of the normalization factor is

$$\hat{f} = \frac{1}{1 + \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma_i^2}\sigma_f^2}.$$

In the case where the values of x_i are consistent with coming from a common true value, it can be shown directly that the expected value of f is

$$\langle f \rangle = \frac{1}{1 + (n-1)\sigma_f^2}.$$

There is, hence, a bias on the result when, for a non-vanishing σ_f, a large number of data points are fitted. In particularly, the fit produces on average a bias larger than the normalization error itself if $\sigma_f > 1/(n-1)$. One can also see that $\sigma^2(\hat{k})$ and the minimum of χ^2 obtained with the covariance matrix, or with χ^2_B, are smaller by the same factor r than those obtained with χ^2_A.

One can think of a different approach [4] which in principle would offer an alternative to Eq. (3) for solving the problem. In the hypothesis that the measurements come from the same physical value, the best estimate of the covariance matrix is

$$\sigma^2(Y_i) = \sigma_i^2 + \sigma_f^2\hat{k}^2,$$

$$\mathrm{Cov}(Y_i, Y_j) = \sigma_f^2\hat{k}^2 \qquad (i \neq j),$$

where $\hat{k}$ is the result of the fit. One obtains then a covariance matrix equal to that of a common offset error with $\sigma_c = \sigma_f\hat{k}$. Since we have shown that, in this case, the best value of k does not depend on the normalization uncertainty and that its error is the quadratic combination of the weighted average error and of the normalization one, we reach exactly the same results obtained using χ^2_A. One may be tempted to conclude that this is the best obtained solution, in the sense that one can still work with the covariance matrix leading to unbiased results. In reality it is clear that (apart from the simple case of the fit to a constant) the definition of a covariance matrix having in the elements functions of the fitted parameters instead of numbers leads to many complications, and one has to use iterative methods to evaluate the covariance matrix. The attempt to solve a non-linear problem with linear methods offered by this approach clearly yields a more complicated situation than that previously discussed.

4. Conclusions

The knowledge of the best estimate of the covariance matrix of the data points is recognized as a powerful tool in treating complex problems with correlations between the data. In general, for the case of measurements related by a common uncertainty in calibration constants or in theoretical corrections, the covariance matrix is derived from standard error propagation. It has been shown in simple cases that, if one considers a new physical quantity as a function of the measured ones, the use of the appropriately built covariance matrix gives the correct error on the new quantity.

In the case that one has only an overall systematic error and the covariance matrix is used to define χ^2, the behaviour of the best fit is different depending on whether the uncertainty is on the offset or on the scale. In the first case the best estimates of the function parameters are exactly those obtained without systematic errors, and only the parameters' errors are affected. In the case of *normalization* errors, biased results can be obtained instead. The size of the bias depends on the fitted function, on the magnitude of the overall error and on the number of data points. It has been shown that, in the case of a fit to a constant – the result can be qualitatively extended to other functions – a negative bias is obtained, the absolute size of which is proportional to the number of degrees of freedom and to the square of the normalization error. It has also been shown that this bias comes from the linearization performed in the usual error propagation. This means that, even though the use of the covariance matrix can be very useful in analysing the data in a compact way using available computer algorithms, attention is required if there is one large normalization uncertainty which affects all the data. In this case it is preferable [#4] not to include the overall error in the covariance matrix for several reasons. Firstly, one avoids the problem just discussed. Moreover – and this argument holds also in case of an offset global error – it is generally preferred to give separately the system-

[#4] This is the way how CELLO [1] finally presented the result of the analysis (see ref. [5] for details). A check was also done using a χ^2 definition similar to Eq. (3), where the individual normalization factors of the experiments were fitted. The fitted f values were distributed around 1 with a standard deviation compatible with the normalization error declared by each of the experiments. Moreover, the size of a possible bias that an overall systematic error would have produced on the results was also estimated.

atic errors. In particular, if also the variation of the result for a given variation of the normalization factor around unity is provided, one can correct the results when a better knowledge of the systematics is available.

Acknowledgements

This work is partially based on old notes from the time when I had been a member of the CELLO Collaboration. Remembering the friendly experience we had, I would like to acknowledge all the debates on the subject with my colleagues. A discussion with Guido Martinelli convinced me to reelaborate those notes. I would like to thank Peter Bussey for the many useful discussions which led me to extend the analysis, and Günter Wolf for comments on the manuscript.

References

[1] CELLO Collaboration, H.J. Behrend et al., Phys. Lett. B 183 (1987) 400.
[2] R.J. Barlow, Statistics (Wiley, Chichester, 1989).
[3] DIN Deutsches Institut für Normung e.V., DIN 1319, Teil 4, Grundbegriffe der Messtechnik – Behandlung von Unsicherheiten bei der Auswertung von Messungen (Beuth Verlag GmbH, Berlin, Germany, 1985).
[4] P.J. Bussey, private communication.
[5] G. D'Agostini, Proc. XXII Rencontre de Moriond, Les Arcs, France, March 15–21 1987, ed. J. Tran Thanh Van (Editions Frontière) p. 325.