Rapport BIPM-86/16

Correlations in successive differences and smoothings

by Jörg W. Müller

Bureau International des Poids et Mesures, F-92310 Sèvres

Abstract

There is a widespread belief that covariances - at least for practical applications - are too intricate to evaluate and have only effects so marginal that they can normally be neglected without danger. In order to counter this opinion, we show for two specific cases, which are successive differences and smoothing of originally uncorrelated data, that the evaluation of the respective variances, covariances and correlation factors is indeed very simple. A potentially useful application to the accurate measurement of the masses of radioactive sources by the pycnometer method is sketched.

1. Introduction

Whereas the evaluation of variances has become common practice in the handling of experimental data, since they are needed for indicating the uncertainty of measurements, covariances are still considered by many engineers and even physicists as quantities about which they normally will not have much to worry. In addition, since covariances have the reputation of being difficult to handle and to use, they are often even not mentioned - as few people look for unnecessary complications.

Such a practice causes little or no damage when applied to results of measurements which, for good reasons, may be considered independent of each other, but there are obviously many others, and for them neglecting a correlation can lead to serious errors.

It is the purpose of this note to illustrate by two examples which, although very simple, are still of a certain practical interest, how the respective variances and covariances can be readily determined. It will turn out that both examples, which at first sight seem to have little in common, lead to nearly identical formulae (at least for a special set of weighting factors).

2. Successive differences

Let us start with a series of uncorrelated results $x_1 \pm \sigma_1$, $x_2 \pm \sigma_2$, ..., $x_N \pm \sigma_N$, where the quantities σ_k denote the corresponding estimated (sample) standard deviations, and with $N \gg 1$. From these data we first form the quantities

$$_1D_k \equiv x_k - x_{k+1} , \qquad \text{with } k \geqslant 1 , \tag{1}$$

i.e. the successive differences of the original measurements. We can now go on forming differences of higher order, for instance

$$\begin{aligned} _2D_k &= {_1D_k} - {_1D_{k+1}} = (x_k - x_{k+1}) - (x_{k+1} - x_{k+2}) \\ &= x_k - 2x_{k+1} + x_{k+2} , \end{aligned}$$

and likewise (2)

$$\begin{aligned} _3D_k &= {_2D_k} - {_2D_{k+1}} = x_k - 3x_{k+1} + 3x_{k+2} - x_{k+3} , \\ _4D_k &= {_3D_k} - {_3D_{k+1}} = x_k - 4x_{k+1} + 6x_{k+2} - 4x_{k+3} + x_{k+4} , \end{aligned}$$

etc.

This suggests that the general formula for differences of order n is given by

$$_nD_k = \sum_{j=0}^{n} (-1)^j \binom{n}{j} x_{k+j} . \tag{3}$$

This basic expression can be readily proved by mathematical induction. Starting from (3), we evaluate the difference of order n+1 according to the definition $_{n+1}D_k = {_nD_k} - {_nD_{k+1}}$ and find

$$\begin{aligned} _{n+1}D_k &= \sum_{j=0}^{n} (-1)^j \binom{n}{j} x_{k+j} - \sum_{j=1}^{n+1} (-1)^{j-1} \binom{n}{j-1} x_{k+j} \\ &= \sum_{j=1}^{n} (-1)^j \left[\binom{n}{j} + \binom{n}{j-1}\right] x_{k+j} + x_k - (-1)^n x_{k+n+1} \\ &= \sum_{j=0}^{n+1} (-1)^j \binom{n+1}{j} x_{k+j} , \qquad \text{for} \quad n \geqslant 1 , \end{aligned}$$

as expected.

We note that by means of the so-called Blissard calculus [1], where indices are treated like powers during all intermediate formal operations, eq. (3) could be simplified to $_nD_k = x^k(1-x)^n$, but no use will be made of this symbolic notation in what follows.

For the evaluation of the variances we take advantage of the fact that the original quantities x_k are uncorrelated; simple error propagation then leads to

$$\mathrm{Var}\ ({}_nD_k) \quad = \quad \sum_{j=0}^{n} \binom{n}{j}^2 \sigma_{k+j}^2 \ . \tag{4}$$

In order to evaluate the covariances, we first form the sum

$$\begin{aligned} {}_nS_r \ \equiv \ {}_nD_k + {}_nD_{k+r} \ &= \ \sum_{j=0}^{n} (-1)^j \binom{n}{j} x_{k+j} + \sum_{j=0}^{n} (-1)^j \binom{n}{j} x_{k+r+j} \\ &= \sum_{j=0}^{n+r} \left[(-1)^j \binom{n}{j} + (-1)^{j-r} \binom{n}{j-r}\right] x_{k+j} \ . \end{aligned} \tag{5}$$

Its variance is easily seen to be

$$\mathrm{Var}\ ({}_nS_r) \quad = \quad \sum_{j=0}^{n+r} \left[\binom{n}{j}^2 + \binom{n}{j-r}^2 + 2(-1)^r \binom{n}{j} \binom{n}{j-r}\right] \sigma_{k+j}^2 \ . \tag{6}$$

On the other hand, the definition of ${}_nS_r$ also allows us to write

$$\begin{aligned} \mathrm{Var}\ ({}_nS_r) \quad &= \quad \mathrm{Var}\ ({}_nD_k) + \mathrm{Var}\ ({}_nD_{k+r}) + 2\ \mathrm{Cov}\ ({}_nD_k,\ {}_nD_{k+r}) \\ &= \sum_{j=0}^{n} \binom{n}{j}^2 \sigma_{k+j}^2 + \sum_{j=0}^{n} \binom{n}{j}^2 \sigma_{k+r+j}^2 + 2\ \mathrm{Cov}\ (\ldots) \\ &= \quad \ldots \quad + \sum_{j=r}^{n+r} \binom{n}{j-r}^2 \sigma_{k+j}^2 + \ldots \ , \end{aligned} \tag{7}$$

and hence, with (6),

$$\begin{aligned} \mathrm{Cov}\ ({}_nD_k,\ {}_nD_{k+r}) \ &= \ \frac{1}{2} \left\{\mathrm{Var}\ ({}_nS_r) - \sum_{j=0}^{n+r} \left[\binom{n}{j}^2 + \binom{n}{j-r}^2\right] \sigma_{k+j}^2\right\} \\ &= \frac{1}{2} \left\{ \sum_{j=0}^{n+r} \left[\binom{n}{j}^2 + \binom{n}{j-r}^2 + 2(-1)^r \binom{n}{j} \binom{n}{j-r}\right] \sigma_{k+j}^2 - \sum_{j=0}^{n+r} [\ldots]\ \sigma_{k+j}^2 \right\} \\ &= (-1)^r \sum_{j=r}^{n} \binom{n}{j} \binom{n}{j-r} \sigma_{k+j}^2 \ . \end{aligned} \tag{8}$$

Two special cases may be of interest, namely

$$\begin{aligned} &\text{- for } r > n: && \mathrm{Cov}\,({}_nD_k,\ {}_nD_{k+r}) = 0, \\ &\text{- for } r = n: && \mathrm{Cov}\,({}_nD_k,\ {}_nD_{n+k}) = (-1)^n\,\sigma^2_{n+k}\ . \end{aligned} \qquad (9)$$

For the frequent situation where all initial data may be assumed to have the same precision, thus $\sigma_k = \sigma$, for any value of k, the relations (4) and (8) can be simplified. This is achieved by means of the identity

$$\sum_{j=r}^{n} \binom{n}{j} \binom{n}{j-r} = \binom{2\,n}{n-r}\ , \qquad \text{for } 0 \leqslant r \leqslant n\ , \qquad (10)$$

which can be readily deduced from formulae given in [2].

We then obtain

$$\mathrm{Var}\,({}_nD_k) = \sigma^2 \sum_{j=0}^{n} \binom{n}{j}^2 = \binom{2n}{n}\,\sigma^2 \qquad (11)$$

and

$$\mathrm{Cov}\,({}_nD_k,\ {}_nD_{k+r}) = (-1)^r\,\sigma^2 \sum_{j=r}^{n} \binom{n}{j}\binom{n}{j-r} = (-1)^r \binom{2\,n}{n-r}\,\sigma^2, \qquad (12)$$

independently of k. For numerical values see Table 1.

Table 1 - Some numerical values for the variances and covariances of successive differences (for data of equal precision)

n	$\mathrm{Var}\,({}_nD_k)/\sigma^2$	$\mathrm{Cov}\,({}_nD_k,\ {}_nD_{k+r})/\sigma^2$					
		r = 1	r = 2	r = 3	r = 4	r = 5	r = 6
1	2	-1					
2	6	-4	1				
3	20	-15	6	-1			
4	70	-56	28	-8	1		
5	252	-210	120	-45	10	-1	
6	924	-792	495	-220	66	-12	1

It is also straightforward to evaluate the correlation coefficient, which is defined here by

$$\rho_{n,r} = \frac{\mathrm{Cov}\,({}_nD_k,\ {}_nD_{k+r})}{\sqrt{\mathrm{Var}\,({}_nD_k)\ \mathrm{Var}\,({}_nD_{k+r})}} \,. \tag{13}$$

Substitution of (11) and (12) first leads to

$$\rho_{n,r} = \frac{(-1)^r \binom{2\,n}{n-r}}{\binom{2\,n}{n}} \,.$$

Since

$$\binom{2\,n}{n-r} = \binom{2\,n}{n+r} = \frac{2n\ (2n-1)\ \ldots\ (n+1)\ n\ \ldots\ (n-r+1)}{(n+r)\ \ \ldots\ (n+1)\ n!}$$
$$= \binom{2\,n}{n} \frac{n(n-1)\ \ldots\ (n-r+1)}{(n+r)\ (n+r-1)\ \ldots\ (n+1)} \,,$$

we then obtain

$$\rho_{n,r} = (-1)^r \prod_{j=1}^{r} \frac{n+1-j}{n+j} \,. \tag{14}$$

The simplest cases are

$$\begin{aligned} &- \text{ for } r = 1: \quad & \rho_{n,1} &= - \frac{n}{n+1} \,, \\ &- \text{ for } r = 2: \quad & \rho_{n,2} &= \frac{n\ (n-1)}{(n+1)\ (n+2)} \,, \\ &- \text{ for } r = 3: \quad & \rho_{n,3} &= - \frac{n\ (n-1)\ (n-2)}{(n+1)\ (n+2)\ (n+3)} \,. \end{aligned} \tag{15}$$

For $n \gg 1$ the limiting value of the correlation coefficient is thus $(-1)^r$, as might have been expected.

We may note, in passing, that the differences ${}_nD_k$ can also serve as a starting point for the evaluation of the variance of the original series x_k, for instance by forming the quantity

$${}_ns_A^2 = \frac{\langle {}_nD_k^2 \rangle}{\mathrm{Var}\,({}_nD_k)/\sigma^2} \,,$$

where $\langle \ldots \rangle$ denotes the mean for the sample of values x_k available. The numerical value of the denominator can be found in Table 1. While n = 1 corresponds to the so-called Allan variance, the case n = 2 would lead to the alternative form

$${}_2s_A^2 = \frac{\sum\limits_{k=1}^{N-2} (x_k - 2\ x_{k+1} + x_{k+2})^2}{6\ (N-2)} \,.$$

3. Repeated smoothings

The smoothing of experimental data is a rather common, but also quite controversial, procedure: smoothed data may look much "better" to the naive eye, but this apparent improvement has been obtained at the price of irreversible systematic distortions. In what follows we shall study only one aspect of this deformation, and no doubt the simplest one, as it can be expressed in a quantitative way by traditional statistical concepts. It will turn out that the reduction of the variance is obtained at the price of increased correlation between the smoothed values.

Smoothings can be performed in many ways. One of the most popular approaches, outlined in [3], consists of the local adjustment of a suitable polynomial. Apart from the general shortcomings mentioned above, this method suffers from the arbitrariness in choosing range and order of the polynomial. In addition, it leads to rather awkward numerical smoothing factors which require extensive tabulations (for the correction of many errors in [3] see [4]*). The evaluation of variances and covariances of measured values that have been smoothed in this way is possible, but leads to cumbersome expressions which cannot be simplified in general. Therefore we shall not illustrate this approach here. Instead, there exists an alternative with a much simpler algebraic structure; the virtues of this so-called "binomial smoothing" have been well described recently in [5] to which we refer the interested reader for all details.

Our short presentation consists of two parts. We first develop a simple scheme for successive averages of adjacent values, and for this we can take advantage of the close analogy that exists with successive differences, the subject treated in the previous section. In a second step we shall establish the link with binomial smoothing.

Let us again start with a series of uncorrelated measurements x_1, x_2, ..., x_N, the respective variances σ_1^2, σ_2^2, ... of which are supposed to be known. In close analogy with (1) successive mean values are formed, first from the original data with

$$ {}_1y_k \equiv \frac{1}{2}\,(x_k + x_{k+1}) \ , $$

and then similarly for "higher" averages ($n \geqslant 2$)

$$ {}_ny_k = \frac{1}{2}\,({}_{n-1}y_k + {}_{n-1}y_{k+1}) \ . \qquad (16) $$

* We take this opportunity to correct a misprint in [4], where in Table 5, for p = 3 or 4 and m = 4, one should read ${}_1^1N$ = 2 220, instead of 2 200.

When written more explicitly, this results in

$$_2y_k = \frac{1}{4}(x_k + 2x_{k+1} + x_{k+2}) ,$$

$$_3y_k = \frac{1}{8}(x_k + 3x_{k+1} + 3x_{k+2} + x_{k+3}) ,$$

$$_4y_k = \frac{1}{16}(x_k + 4x_{k+1} + 6x_{k+2} + 4x_{k+3} + x_{k+4}) ,$$

etc.

It is easy to see that the general expression for the n^{th} averaging is

$$_ny_k = \frac{1}{2^n}\sum_{j=0}^{n} \binom{n}{j} x_{k+j} , \tag{17}$$

with variance

$$\mathrm{Var}\ (_ny_k) = \frac{1}{4^n}\sum_{j=0}^{n} \binom{n}{j}^2 \sigma^2_{k+j} . \tag{18}$$

Since (17) has essentially the same structure as (3), we can readily take advantage of the previously established formula (12) for the covariance, with the result that

$$\mathrm{Cov}\ (_ny_k,\ _ny_{k+r}) = \frac{1}{4^n}\sum_{j=r}^{n} \binom{n}{j}\binom{n}{j-r} \sigma^2_{k+j} . \tag{19}$$

For original data x_k of equal precision $\sigma_k = \sigma$ this leads to the simple expressions

$$\mathrm{Var}\ (_ny_k) = \frac{1}{4^n}\binom{2\ n}{n}\sigma^2 \quad \text{and} \tag{20}$$

$$\mathrm{Cov}\ (_ny_k,\ _ny_{k+r}) = \frac{1}{4^n}\binom{2\ n}{n-r}\sigma^2 . \tag{21}$$

Likewise we now obtain for the correlation coefficient, with the help of (14),

$$\rho_{n,r} = \prod_{j=1}^{r} \frac{n+1-j}{n+j} . \tag{22}$$

As schematically represented in the upper part of Fig. 1, each averaging process results in a "shift to the right". Hence, it is clearly not possible to consider all data ${}_n y_k$ as smoothings of some original value x_j. In order to reestablish this important link, we have to put for the smoothed values

$$ {}_m S_k = {}_{2m} y_{k-m} , \tag{23} $$

as shown in the lower part of Fig. 1. With this labeling it is possible to consider ${}_m S_k$ as the smoothed value, of order m, which corresponds to the original result x_k.

Since for a finite set of measurements x_k, with $1 \leqslant k \leqslant N$, smoothed values ${}_m S_k$ are only available for the range

$$ m+1 \leqslant k \leqslant N-m , \tag{24} $$

various recipes have been proposed for handling the situation at the extremes. Perhaps the simplest prescription consists in taking, instead of ${}_m S_k$, the values based on the smoothing of the highest order available, i.e. (with ${}_o S_k \equiv x_k$)

$$ \begin{aligned} &{}_{k-1}S_k , \qquad \text{for} \quad 1 \leqslant k \leqslant m \\ \text{and} \\ &{}_{N-k}S_k , \qquad \text{for} \quad N-m+1 \leqslant k \leqslant N . \end{aligned} \tag{25} $$

With the notation introduced in (23), the previous formulae (17), (18) and (19) now become

$$ {}_m S_k = \frac{1}{4^m} \sum_{j=0}^{2m} \binom{2m}{j} x_{k-m+j} , \tag{26} $$

$$ \mathrm{Var}\,({}_m S_k) = \frac{1}{16^m} \sum_{j=0}^{2m} \binom{2m}{j}^2 \sigma^2_{k-m+j} , \tag{27} $$

$$ \mathrm{Cov}\,({}_m S_k, {}_m S_{k+r}) = \frac{1}{16^m} \sum_{j=r}^{2m} \binom{2m}{j} \binom{2m}{j-r} \sigma^2_{k-m+j} . \tag{28} $$

For measurements x_k of equal precision σ, the relations (27) and (28) can be simplified to

$$ \mathrm{Var}\,({}_m S_k) = \frac{1}{16^m} \binom{4m}{2m} \sigma^2 \qquad \text{and} \tag{29} $$

$$ \mathrm{Cov}\,({}_m S_k, {}_m S_{k+r}) = \frac{1}{16^m} \binom{4m}{2m-r} \sigma^2 , \tag{30} $$

and for the correlation coefficient we find

$$\rho_{m,r} = \prod_{j=1}^{r} \frac{2m+1-j}{2m+j} . \tag{31}$$

These are the basic relations for the binomial smoothing filter. While (26) is equivalent to eq. 9 in [5], the expressions for the variance and the covariance seem to be new.

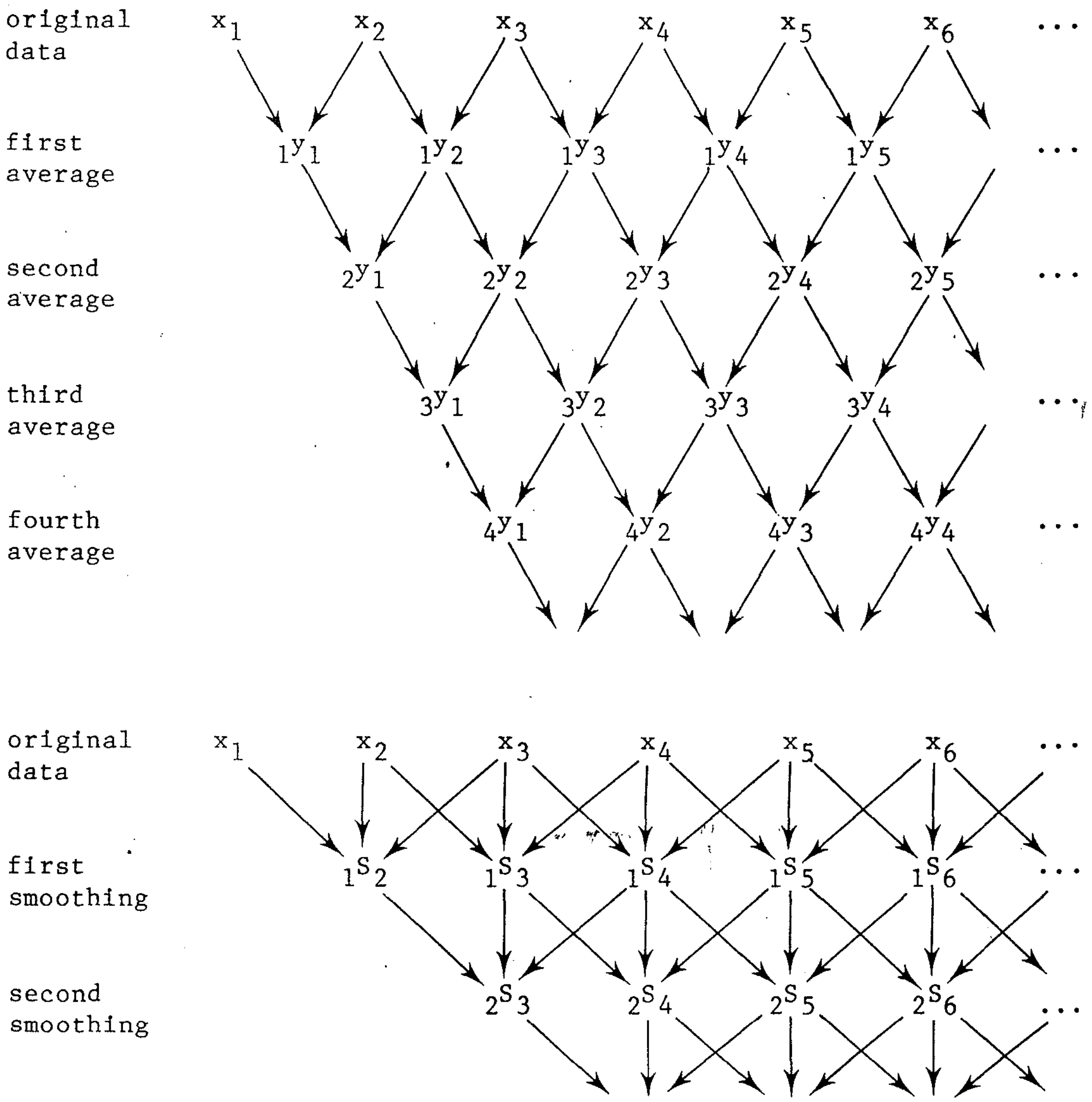
original data
x_1 x_2 x_3 x_4 x_5 x_6 ...
first average
$_1y_1$ $_1y_2$ $_1y_3$ $_1y_4$ $_1y_5$ ...
second average
$_2y_1$ $_2y_2$ $_2y_3$ $_2y_4$ $_2y_5$ ...
third average
$_3y_1$ $_3y_2$ $_3y_3$ $_3y_4$ ...
fourth average
$_4y_1$ $_4y_2$ $_4y_3$ $_4y_4$ ...
original data
x_1 x_2 x_3 x_4 x_5 x_6 ...
first smoothing
$_1S_2$ $_1S_3$ $_1S_4$ $_1S_5$ $_1S_6$ ...
second smoothing
$_2S_3$ $_2S_4$ $_2S_5$ $_2S_6$ ...

Fig. 1 - Schematic representation of the interdependencies between successive averaging processes, for the two schemes described in the text. Note, for instance, that $_2S_5$, the second smoothed value of the original x_5, is identical with $_4y_3$.

4. Note on a possible application

Since smoothings are so commonly done (and mostly without the necessary precautions), there is no need here for a particular illustration. On the other hand, differences are less often formed and one might therefore suspect that the developments outlined in section 2 are at most of theoretical interest. The practical example sketched in what follows will show that this is not the case.

For any absolute measurement of the activity of radioactive substances, the mass of the source prepared for its determination has to be known accurately. The seemingly obvious approach would consist in measuring directly the mass of a source freshly deposited on an appropriate support, but this technique has been completely abandoned for many years because the initial evaporation of a drop cannot be observed, and simple extrapolations to time zero are known to be biased. Therefore, all metrological laboratories nowadays use the "pycnometer method", in which the source mass (m) is obtained from the difference in the weight of the pycnometer (M) before and after deposition of the source, i.e.

$$m_k = M_k - M_{k+1} , \qquad \text{with} \quad 1 \leqslant k \leqslant N . \tag{32}$$

This is completely analogous to (1) and it therefore follows that "successive" masses m_k and m_{k+1} are correlated. This must also apply to the actually measured quantity "specific activity", defined by

$$z_k \equiv a_k/m_k , \tag{33}$$

where a_k is the activity for source number k. Uncertainties may also be associated with a_k. Since the measuring techniques, at least for small and moderate count rates, are assumed to be well under control, it is unlikely that the intrinsic measurement of a_k will be responsible for a significant contribution to the total uncertainty; on the other hand, possible inhomogeneities in the solution, which are very difficult to detect, cannot be safely excluded, but they would produce random deviations with no correlation between the sources. The different statistical behaviour of the uncertainties according to their origin should make it possible to separate them, essentially by forming combinations of values of z_k with their immediate or remote "neighbours" $z_{k'}$. The technical details of the suggested approach to detect in this way the uncertainty component due to weighing will be given in a subsequent report.

References

[1] J. Riordan: "An Introduction to Combinatorial Analysis", (Wiley, New York, 1958), p. 27 ff.

[2] I.S. Gradshteyn, I.M. Ryzhik: "Table of Integrals, Series, and Products" (Academic Press, New York, 1980^4), eqs. 0.156/7

[3] A. Savitzky, M.J.E. Golay: "Smoothing and differentiation of data by simplified least squares procedures", Anal. Chemistry 36, 1627-1639 (1964)

[4] J.W. Müller: "On the smoothing of empirical spectra", Rapport BIPM-74/1 (1974), 37 p.

[5] P. Marchand, L. Marmet: "Binomial smoothing filter: a way to avoid some pitfalls of least-squares polynomial smoothing", Rev. Sci. Instr. 54, 1034-1041 (1983)

(December 1986)