About a procedure to solve the equations to which the method of least squares leads, as well as linear equations in general, by successive approximation

Ludwig Seidel

Aus den Abhandlungen der k. bayer. Akademie der Wissenschaften II Cl. XI Bd. III Abth. München February 1874.

This is only a partial translation of Seidel's paper.

1.

Following the notation which is common in the theory of adjustment of observations, let us assume that the results of the individual observations according to the unknown quantities $x, y, z, \ldots$ satisfy the linear equations

$$\begin{array}{rrcl} & a_1x + b_1y + c_1z + \cdots + n_1 & = & 0 \\ (A) & a_2x + b_2y + c_2z + \cdots + n_2 & = & 0 \\ & a_3x + b_3y + c_3z + \cdots + n_3 & = & 0 \\ & & \vdots & \end{array}$$

It is assumed that the number of observations, and consequently also of equations in (A), is larger than that of the unknowns $x, y, z, \ldots$ but that the individual equations are affected by observation errors or (may be) that the expressions on the left would generally not exactly vanish even when substituting the true values $x, y, z, \ldots$.

At the same time it is assumed that the weights of different observations have already been taken into account in the equations according to known rules, so that a priori there is no longer any reason for a major error in one of the observation equations (i.e. a larger absolute value of the expression, which should be zero according to this equation) rather than in the others.

It is also assumed that, according to the arrangement of the observations, those prerequisites are applicable, or at least must be regarded as applicable according to our understanding, under which the application of the "least squares method" is rationally justified.

After this, the most probable[1] of all sets of values $x, y, z, \ldots$ is derived from the condition that its substitution in the observation equations (A) makes the sum of squares of the numerical values on the left as small as possible. If, as is usually the case in this theory, the [ ] bracket is used as a summation sign,

$$[aa] = a_1^2 + a_2^2 + a_3^2 + \cdots$$

$$[ab] = [ba] = a_1 b_1 + a_2 b_2 + a_3 b_3 + \cdots$$

(where every sum extends over many terms, such that all observations are present), then the sum of the squares is

$$\begin{aligned} Q &= [aa]^2 x^2 + [bb]^2 y^2 + [cc]^2 z^2 + \cdots \\ &+ 2[ab]xy + 2[ac]xz + 2[bc]yz + \cdots \\ &+ 2[an]x + 2[bn]y + 2[cn]z + \cdots \\ &+ [nn] \end{aligned}$$

[Note from the translator: this must probably be $[nn]^2$.]

and it becomes a minimum when the unknowns are calculated from the following normal equations, the number of which is just sufficient and which are therefore strictly satisfied:

$$\begin{aligned} 0 &= [aa]x + [ab]y + [ac]z + \cdots + [an] \\ (B) \quad 0 &= [ab]x + [bb]y + [bc]z + \cdots + [bn] \\ 0 &= [ac]x + [bc]y + [cc]z + \cdots + [cn] \\ &\vdots \end{aligned}$$

As simple as the task is, according to its mathematical nature of calculating any number of unknown quantities from an equal number of linear equations equations, its numerical implementation becomes very tedious when the number of unknowns becomes considerably large, and for this reason one feels compelled, in cases of the kind mentioned, such as, for example, the compensation of a large geodetic network, to form partial set of unknowns at the sacrifice of strictly systematic execution and, after each has been calculated separately, to connect them as well as possible.

I do not know whether a problem of more than seventy unknowns has ever been uniformly calculated.

[1] The advantage of this most probable set is not justified by the fact that the infinitely small probability that it will be exactly right is larger than for any other set, - but by the fact that the finite probability that the real values of the unknown differ from the presupposed only within arbitrarily set but tight bounds, is larger in the particular system than the one relating to equally tights bounds (likewise finite) for every differently chosen system. This does not seem to have been understood with the necessary clarity everywhere.

The number 70 is reached in the measurement of the East Prussian network[2] (and this in a case, where between the unknowns there are still 31 equations to be strictly satisfied, which circumstance, however, according to the usual way of treatment, only complicates the matter), and when, with 72 unknowns, I had to do with the calculation of the most probable values for the logarithms of the brightness of stars, which were obtained from the photometric network[3] - As is well known, the usual method of solution given by Gauss is based on the fact that one takes the value of any unknown, expressed by the others, from that equation in which this unknown is multiplied by a sum of squares, or, according to the usual and immediately understandable way of speaking and writing, from the equation in which this unknown occurs in the diagonal of the system - thus, for example, the value of x from the first equation in (B), - and substituting this value into the remaining equations whereby one obtains the first transformed system of normal equations

$$\begin{array}{rrcl} & [bb.1]y & + & [bc.1]z + \cdots = [bn.1] \\ (C) & [bc.1]y & + & [cc.1]z + \cdots = [cn.1] \\ & & \vdots & \end{array}$$

which shares with the original system (B) the property that the coefficients of the unknowns are symmetric with respect to the diagonal, and in which

$$[bc.1] = [cb.1] = [bc] - \frac{[ab]\,[ac]}{[aa]}$$

By proceeding with a second unknown in the same way, etc., successive transformed systems are obtained, each of which with one less unknown than the preceding one, until the last unknown stands alone and, after determining its value, all the others are calculated successively, each from the equation that was used for its elimination, according to the reverse order.

Jacobi devised another method and applied it to the 7 equations which were set up by Leverrier for the calculation of part of the secular perturbations in the planetary system according to Laplace[4]; As a student, I had the honor of doing the numerical calculations for him.

According to this, the greatest of the coefficients outside the diagonal are successively made to vanish, by using a suitable linear substitution which corresponds perfectly to the rotation of a right-angled coordinate system, instead of the two unknowns which are multiplied by such coefficients, two new unknowns are introduced, and the symmetry of the whole is preserved;

[2]See the work of von Bessel and Baeyer, Section III.

[3]See my treatise, "Resultate photometrischer Messungen etc." in Denkschriften der Münchener Akademie, 1862, Paragraph 8.

[4]See Crelle's Journal, Band 30, p. 51 "Ueber ein leichtes Verfahren etc."

in the place of the canceled coefficient, a non-vanishing coefficient appears again in the course of the calculation through the later substitutions, but the sum of the squares of the coefficients outside the diagonal is steadily reduced in favor of those in the diagonal, and in a way whose convergence has been strictly proved, one approaches the final goal as much as one wants, where (apart from the purely constant members) only the diagonal ones are left, and the last unknowns result immediately.

As useful as this method is, adapted to the special difficulties of the case for which Jacobi initiated its application and in which the diagonal coefficients themselves are still linear functions of a supernumerary unknown, it seems to me that it is by no means more advantageous for the usually occurring simpler case than the general one; also I doubt whether it has been applied in any other case so far.

I have taken a third route in my above cited photometric paper; its choice was a particularly obvious one for the task treated there in view of the simple form of the individual observation equations. In my present paper, I intend to present and justify this method of resolution in the form in which it is quite generally applicable, and at the same time to discuss some details which are connected with it.

2.

First imagine some set of unknowns $x, y, z, \ldots$ assumed with numerical values which do not yet satisfy the "normal equations" (B) of the most probable system, but rather are such that

$$\begin{aligned} N_1 &= [aa]x + [ab]y + \ldots + [an] \\ N_2 &= [ab]x + [bb]y + \ldots + [bn] \\ \vdots &= \vdots \end{aligned}$$

The sum of the squares of the errors Q (see above) can be written like this after an identical transformation

$$\begin{aligned} Q &= \frac{1}{[aa]}\{[aa]x + [ab]y + [ac]z + \cdots + [an]\}^2 \\ &+ [bb.1]y^2 + [cc.1]z^2 + \cdots \\ &+ 2[bc.1]yz + \cdots + 2[bn.1]y + 2[cn.1]z + \cdots \\ &+ [nn.1] \\ &= \frac{1}{[aa]}N_1^2 + [bb.1]y^2 + \cdots + [nn.1] \end{aligned}$$

In this form of the expression, the unknown x occurs only in the first term (namely in N_1); from this it is immediately clear that the sum Q of the squares of the errors is reduced by the amount

$$\frac{1}{[aa]}N_1^2$$

if, while $y, z, \ldots$ keep their first assumed values, one changes that of x in such a way that the expression which was previously N_1 is made zero. This is caused by such a change Δx to be applied to x,

$$\Delta x = -\frac{N_1}{[aa]}$$

and the thus improved value $x + \Delta x$ is now obviously that of the first unknown, which fits best with the initially assumed values of the other unknowns, and which would then be the most plausible one for them if the initial values of the remaining unknowns were already known as their true values.

The change in x, which replaces N_1 with the value $N_1' = 0$ will, at the same time, change the values of $N_2, N_3, \ldots$ into

$$\begin{aligned} N_2' &= N_2 + [ab]\Delta x, \\ N_3' &= N_3 + [ac]\Delta x, \\ \vdots &= \vdots \end{aligned}$$

If, instead of holding $y, z, \ldots$ and correcting x by $-\frac{N_1}{[aa]}$, one had rather leave $x, z, \ldots$ with their initial values, and corrected y by $-\frac{N_2}{[bb]}$, then one would obviously have reduced the sum Q from its initial value by the size $\frac{N_2^2}{[bb]}$; in the same way one would have reduced the same sum by $\frac{N_3^2}{[cc]}$ if z alone had been corrected by $-\frac{N_3}{[cc]}$ from its initial value.

So if one thinks that after x has been corrected by $\Delta x = -\frac{N_1}{[aa]}$, and after the quantities

$$\begin{aligned} N_1' &= 0, \\ N_2' & \\ N_3' & \\ \vdots & \end{aligned}$$

have taken the place of $N_1, N_2, N_3, \ldots$ now, in the second instance, such a correction

$$\Delta y = -\frac{N_2'}{[bb]}$$

is added the variable y, by means of which $y + \Delta y$ receives the numerical value that best fits the set of values

$$x + \Delta x, z, \ldots$$

of the remaining unknowns, - so this second change will further reduce the sum of the squared errors Q by

$$\frac{(N_2')^2}{[bb]}$$

after it had already been reduced by $\frac{N_1^2}{[aa]}$ from its first value.

This change in the value of y replaces the quantities $N_1' = 0, N_2', N_3', \ldots$ again with new quantities $N_1'', N_2'' = 0, N_3'',$, etc., where one has

$$\begin{aligned} N_1'' &= N_1' + [ab]\Delta y = [ab]\Delta y \\ N_2'' &= N_2' + [bb]\Delta y = 0 \\ N_3'' &= N_3' + [bc]\Delta y \\ \vdots &= \vdots \end{aligned}$$

If one were to correct the third variable z in such a way that the new value $z + \Delta z$ would fit as well as possible to the set that would be formed from $x + \Delta x$, $y + \Delta y$ and the initial values of the other unknowns, one would reduce the sum of the squares of the error again, namely by the size

$$\frac{(N_3'')^2}{[cc]};$$

on the other hand they would be reduce to

$$\frac{(N_1'')^2}{[aa]}$$

if one comes back to the variable x, and (since $x + \Delta x$ is no longer the value that best fits the system $y + \Delta y, z, \ldots$) wants to apply a second correction $-\frac{N_1''}{[aa]}$ to x.

If one therefore, starting from any system of initial values, and in any sequence of the unknowns (whereby it is not exactly necessary to go through the whole cycle of them before one comes back to an already improved one), applies successive corrections to the unknowns, taking care to always determine the improvement of each one in such a way that through it the normal equation is satisfied, in which the unknown in question occupies the prominent position in the diagonal, - so one reduces step by step the sum of the error squares (each time by an immediately definable amount of the form $\frac{N^2}{[aa]}$ or $[aa]\Delta x^2$) as long as there is is still something to be reduced. Because the reductions to be achieved from Q and the corrections to be applied to the unknowns (the latter of the form $-\frac{N}{[aa]}$) only become imperceptible when all N have been reduced to infinitesimally small values. But if this final goal has been reached, then all normal equations (B) are satisfied at the same time by the repeatedly improved values of the unknowns, and the unknowns therefore have their most probable values.

It should be noted that the proof of the constant decrease in Q and thus of the convergence of this approximation method is entirely based on the assumption that, for each variable, each of its successive improvements is formed according to the equation in which this variable is on the diagonal;

for if, for instance, in any system of linear equations with an equal number of unknowns, one proceeds from any set of values and applies successive improvements in such a way that the correction of x would always be determined according to a certain, arbitrarily selected, equation, as well as the correction of y according to another arbitrarily selected equation, etc., - it would by no means be possible to prove that one approaches the true set of the values of the unknowns without end - it could rather (as one can be easily convinced) that the successive values of the unknowns finally grow to infinity, or oscillate constantly between finite boundaries.

However, any system of linear equations with the same number of unknowns can be brought into the normal form (B), exactly according to the rule applied to the equations (A), and from this form it can be solved according to our method in exactly the same way as the normal equations derived from observations.

The endless fluctuation or infinite growth of the successive values of the variables is cut off from the beginning in a system of equations of this special kind, and in our way of treating it, because it is proved that here the goal of reducing the sum of squares Q is approached with every step, and that it is reduced by nothing noticeable only when all equations (B) are fulfilled up to infinitesimally small quantities, thus when all unknowns have reached their definite values.

Of course, the sum Q exists just as well in a strictly satisfiable system of equations with the allowed number of unknowns as in observation equations; the only difference is that in the first case the minimum value, which it finally get, is equal to zero.

The advantage of certain convergence, which the normal form (B) of the equations offers for our kind of solution over any other forms, is naturally founded in the distinctive properties of its coefficient system, which are by no means exhausted in the symmetry around the diagonal, and whose most important consists in the fact that every sub-determinant extracted from the system is positive, if it has, as diagonal, a piece of the diagonal of the whole system.

It goes without saying that the more the assumed initial values of $x, y, z, \ldots$ come closer to the truth, the faster our method will lead to the goal. For the certainty of finite convergence of the calculation, however, the possession of even approximately correct initial values is in no way necessary.

Strictly speaking, at every stage of the calculation it would be the most rational to first improve the variable whose correction reduces the sum of the squares of the errors the most. However, as easy as it is to recognize the amount by which this sum is reduced in one case or the other, in practice it will usually lead to the goal faster if one proceeds quickly after a mere survey rather than if one proceeds every time systematically chooses according to this principle.

It has already been emphasized that it is by no means necessary to

correct all the unknowns one after the other, but that one can very well come back to one variable before all the others have been corrected the same number of times. But that one cannot generally arrive at the definitive set of values without having improved all the unknowns, is evident, and would manifest itself in our procedure also by the fact that nothing would have happened for the reduction of the value of the quantity N corresponding to the neglected unknown, or the same means, that nothing would have happened for the fulfillment of the normal equation in which this unknown occupies the prominent place[5].

3.

The rule to determine at any time the improved value of an unknown in such a way as it results from the equation in which this unknown takes the diagonal, by substitution of the values derived for the other unknowns up to that point, can also be put into words differently, namely in such a way that one does not presuppose the normal equation for x

$$[aa]x + [ab]y + \cdots + [an] = 0$$

as formed at all, but acts only with the individual observation equations in which the unknown x occurs, which have the form

$$\begin{aligned} a_1x + b_1y + c_1z + \cdots + n_1 &= 0 \\ a_2x + b_2y + c_2z + \cdots + n_2 &= 0 \\ &\vdots \end{aligned}$$

The value of x, which "best fits" the assumed values of the other unknowns $y, z, \ldots$ or which results from the above normal equation, is none other than the arithmetic mean, with respect to weights, of all the individual determinations which, given the assumed values of the other unknowns $y, z, \ldots$ as their true values, result for x from the various observation equations in which this quantity occurs.

Because, as in the beginning, it is assumed that the individual observation equations have already been provided with such factors, by virtue of which they all have the same "probable error" v or one and the same weight

[5]The method of solution by Jacobi, mentioned in §1, also runs according to all its linear substitutions in such a way that the values of the last unknowns are determined by successive corrections such as those proposed here. But the whole preparatory calculation, through which the coefficients are decreased outside the diagonal by Jacobi, so that they become small, where the focus of his method lies and which is very tedious to use, is omitted in our calculation process, because the proof has been given that one does not need such a preparation of the normal equations in order to approach the goal with certainty. Incidentally, the transformation of the normal equations in the special case for which Jacobi devise his method, according to the special conditions, was much more indicated than it would be in the ordinary case.

1, and if one imagines the values of the $y, z, \ldots$ are known to be correct, then the determinations for x are obtained from the individual observations

$$\begin{aligned} x &= -\frac{b_1}{a_1}y - \frac{c_1}{a_1}z - \cdots - \frac{n_1}{a_1} \\ x &= -\frac{b_2}{a_2}y - \frac{c_2}{a_2}z - \cdots - \frac{n_2}{a_2} \\ &\vdots \end{aligned}$$

which in turn have the probable errors $\frac{v}{a_1}, \frac{v}{a_2}, \ldots$ or the weights $a_1^2, a_2^2, \ldots$ Therefore, if one imagines each of the values of x, which are to be taken together to form the arithmetic mean, is applied as often as corresponding to its weight, or in other words, before the addition of the last equation, each is multiplied by its weight, one obtains for the arithmetic mean by using the brackets as the sum symbol

$$[aa]x = -[ab]y - [ac]z - \cdots - [an]$$

i.e. the mean value formed according to our rule is the same which is found for x from its normal equation.

In the manner justified by this, without forming the normal equations for the amounts of light of the stars compared with one another, I have applied the procedure in question to derive their most probable values in my photometric treatise cited above. In the ordinary cases, however, where the coefficients a, b, etc. are not as simple numbers as they were in that example, and where therefore the derivation of the arithmetic mean of the individual determinations of the unknowns is not as simple as there, the actual formation of the normal equations should be preferred.

4.

[. . .]