University of Maryland College Park

Institute for Advanced Computer Studies TR–92–42
Department of Computer Science TR–2877

On a New Way of Solving the Linear Equations that Arise in the Method of Least Squares*

C. G. J. Jacobi

Translated by
G. W. Stewart†

April 1992

ABSTRACT

This report contains a translation of a paper of C. G. J. Jacobi, "Ueber eine neue Auflösungsart der bei der Methode der kleinsten Quadrate vorkommenden lineären Gleichungen," which appeared in *Astronomische Nachrichten* **22** (1845). In the paper Jacobi shows how to use rotations to increase the diagonal dominance of symmetric linear systems, which he then solves by what we today call the point Jacobi method. This preconditioner is none other than Jacobi's method for diagonalizing a symmetric matrix. Although Jacobi points out his method can be used to find eigenvalues, he reserves a fuller exposition for a later paper [Journal für die reine und angewandte Mathematik, **30** (1846), 51–s94], which is now generally cited as the source of the method. A variant for unsymmetric equations is also considered.

*This report is available by anonymous ftp from `thales.cs.umd.edu` in the directory `pub/reports`.

†Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742.

ASTRONOMISCHE NACHRICHTEN.
No. 523

On a New Way of Solving the Linear Systems that Arise in the Method of Least Squares

The burden of solving exactly a large number of linear equations — the method of least squares leads to such equations in many cases — has caused people to consider the use of iterative methods. One such method naturally presents itself when a different variable in each equation is multiplied by a pre-eminently large coefficient. Specifically, let the equations be

$$\begin{aligned}(00)x + (01)x_1 + (02)x_2 \text{ etc.} &= (0m)\\(10)x + (11)x_1 + (12)x_2 \text{ etc.} &= (1m)\\(20)x + (21)x_1 + (22)x_2 \text{ etc.} &= (2m)\\ \text{etc.} \quad \text{etc.} \quad \text{etc.},\end{aligned}$$

and let all the coefficients (ik) be small compared to the diagonal coefficient (ii). Then we can get approximations to the unknowns x, x_1, x_2, etc. from the equations

$$(00)x = (0m), \quad (11)x_1 = (1m), \quad (22)x_1 = (2m), \quad \text{etc.}$$

If we denote these values by a, a_1, a_2, etc., we obtain their first corrections, which I denote by Δ, Δ_1, Δ_2, etc., from the equations

$$\begin{aligned}(00)\Delta\ &= -\{(01)a_2 + (02)a_2 \text{ etc.}\}\\(11)\Delta_1 &= -\{(10)a\ + (12)a_2 \text{ etc.}\}\\ &\quad \text{etc.} \qquad \text{etc.}\end{aligned}$$

In general, if we set

$$\begin{aligned}x\ &= a\ + \Delta\ + \Delta^2 + \Delta^3 \text{ etc.}\\x_1 &= a_1 + \Delta_1 + \Delta_1^2 + \Delta_1^3 \text{ etc.}\\x_2 &= a_2 + \Delta_2 + \Delta_2^2 + \Delta_2^3 \text{ etc.}\\ &\quad \text{etc.} \qquad \text{etc.},\end{aligned}$$

where the superscripts index the sequence of ever decreasing corrections, then we get the Δ^{i+1} from the Δ^i according to the equations

$$\begin{aligned}(00)\Delta^{i+1} &= -\{(01)\Delta_1^i + (02)\Delta_2^i \text{ etc.}\}\\(11)\Delta_1^{i+1} &= -\{(10)\Delta^i\ + (12)\Delta_2^i \text{ etc.}\}\\(22)\Delta_1^{i+1} &= -\{(20)\Delta^i\ + (21)\Delta_1^i + (23)\Delta_3^i \text{ etc.}\},\\ &\quad \text{etc.} \qquad \text{etc.},\end{aligned}$$

Now in the equations that come from the method of least squares, the diagonals *usually* dominate; for they are sums of squares, while the remaining coefficients arise from the addition of positive and negative numbers, which tend to cancel each other. Nonetheless, some of the off-diagonal elements will, as a rule, assume values large enough to affect the success of the iterative method just given. However, as I will show in what follows, we can transform the equations by the repetition of an easy calculation into other equations for which the this problem grows less and less, until eventually the equations assume a form that permits the application of the above iterative method.

I will assume that the two off-diagonal coefficients (ik) and (kj) are equal, something that is always true of the equations that come from the method of least squares. I will also suppose that the coefficient (01) has a significant value, whose effect is to slow the iterative method. To annihilate this coefficient, I set

$$\begin{aligned}x\ &= cos\,\alpha.\eta + sin\,\alpha.\eta_1,\\x_1 &= sin\,\alpha.\eta - cos\,\alpha.\eta_1,\end{aligned}$$

whence

$$\begin{aligned}(00)x + (01)x_1 &= \{(00)cos\,\alpha + (01)sin\,\alpha\}\eta\\&\quad + \{(00)sin\,\alpha - (01)cos\,\alpha\}\eta_1,\\(10)x + (11)x_1 &= \{(10)cos\,\alpha + (11)sin\,\alpha\}\eta\\&\quad + \{(10)sin\,\alpha - (11)cos\,\alpha\}\eta_1.\end{aligned}$$

I then replace the two equations

$$\begin{aligned}u\ &= (00)x + (01)x_1 + (02)x_2 \text{ etc.} - (0m) = 0,\\u_1 &= (10)x + (11)x_1 + (12)x_2 \text{ etc.} - (1m) = 0\end{aligned}$$

with two other equations:

$$\begin{aligned}\nu\ &= cos\,\alpha.u + sin\,\alpha.u_1 = 0,\\\nu_1 &= sin\,\alpha.u - cos\,\alpha.u_1 = 0.\end{aligned}$$

If we now determine the angle α so that

$$\{(0,0) - (1,1)\}cos\,\alpha\, sin\,\alpha = (01)\{cos\,\alpha^2 - sin\,\alpha^2\}$$

or

$$\frac{1}{2}tang\,2\alpha = \frac{(01)}{(00) - (11)},$$

then the two new equations become

$$\{(00)\cos\alpha^2 + 2(01)\cos\alpha\,\sin\alpha + (11)\sin\alpha^2\}\eta$$
$$+\{\cos\alpha.(02)+\sin\alpha.(12)\}x_2+\text{etc.} = \cos\alpha.(0m)+\sin\alpha(1m),$$
$$\{(00)\sin\alpha^2 - 2(01)\sin\alpha\,\cos\alpha + (11)\cos\alpha^2\}\eta$$
$$+\{\sin\alpha.(02)-\cos\alpha.(12)\}x_2+\text{etc.} = \sin\alpha.(0m)-\cos\alpha(1m).$$

One can easily determine the coefficients of x_2, x_3, etc. trigonometrically by means of auxiliary angles whose tangents are equal to $\dfrac{(12)}{(02)}$, $\dfrac{(13)}{(03)}$, etc. Here one must pay close attention to the correctness of the signs of the coefficients. In this regard an effective check may be obtained by assuming that in u and u_1 and ν and ν_1 we have

$$x = \cos\alpha + \sin\alpha, \quad x_1 = \sin\alpha - \cos\alpha,$$
$$\eta = \eta_1 = x_2 = x_3 \text{ etc.} = 1$$

and testing the equality of the values

$$\nu = \cos\alpha.u + \sin\alpha.u_1,$$
$$\nu_1 = \sin\alpha.u - \cos\alpha.u_1.$$

The coefficients of η and η_1 can be represented in the following forms:

$$\frac{(00)+(11)}{2} + \sqrt{R},$$
$$\frac{(00)+(11)}{2} - \sqrt{R},$$

where

$$R = \left\{\frac{(00)-(01)}{2}\right\}^2 + (01)^2.$$

The sign of $\sqrt{R}$ depends on the the quadrant in which 2α is taken according to the two formulas

$$\sqrt{R} = \frac{(00)-(01)}{2\cos 2\alpha} = \frac{(01)}{\sin 2\alpha},$$

which at the same time provides a check.

When x and x_1 are replaced by η and η_1, each one of the remaining equations, such as

$$(20)x + (21)x_1 + (22)x_2 + \text{etc.} = (2m),$$

is transformed as follows:

$$\{(20)\cos\alpha + (21)\sin\alpha\}\eta + \{(20)\sin\alpha - (21)\cos\alpha\}\eta_1$$
$$(22)x_2 + (23)x_3 + \text{etc.} = (2m).$$

Since here the coefficients of η and η_1 are the same as the coefficients of x_2 in the the first two transformed equations, we see that the transformed equations retain their symmetry about the diagonal. Therefore, to get the coefficients of η and η_2 in the remaining equations we need only calculate the coefficients of x_2, x_3, etc. in the first two equations. The coefficients of x_2, x_3, etc. in the remaining equations are unchanged, as are the constant terms.

In the transformed equation the coefficient corresponding to (01) is zero. The sum of the diagonal coefficients remains unchanged; i.e., $(00)+(11)$. On the other hand, the sum of their squares increases by $2.(01)^2$. From this it follows that these coefficients spread apart: the larger becomes larger and the smaller becomes smaller. However, provided the coefficients of the equations are formed as they are in applications of the method of least squares, the smaller coefficient can never vanish. Specifically, the product of the two coefficients is

$$\left\{\frac{(00)-(01)}{2}\right\}^2 - R = (00)(11) - (01)^2.$$

Hence if we set

$$(00) = \alpha\alpha \quad + \beta\beta \quad + \gamma\gamma \quad + \delta\delta \quad \text{etc.}$$
$$(11) = \alpha_1\alpha_1 + \beta_1\beta_1 + \gamma_1\gamma_1 + \delta_1\delta_1 \text{ etc.}$$
$$(01) = \alpha\alpha_1 \quad + \beta\beta_1 \quad + \gamma\gamma_1 \quad + \delta\delta_1 \quad \text{etc.}$$

the product is always the quantity

$$(00)(11) - (01)^2 = \Sigma(\alpha\beta_1 - \beta\alpha_1)^2,$$

which is positive. For the above sum consists of all squares formed pairwise from the elements α, β, γ, δ, etc. and cannot be zero unless the quantities α, β, γ, δ, etc. are all proportional to the quantities α_1, β_1, γ_1, δ_1, etc.

The sum of squares of the coefficients of x_2, of x_3, etc. — $(02)^2+(12)^2$, $(03)^2+(13)^2$, etc. — are also unchanged in the two transformed equations. Likewise each of the remaining equations the sum of squares of the the coefficients of η and η_1 are the same as those of x and x_1 in the original system. The sum of squares of the off-diagonal coefficients thus decreases by $2(01)^2$, which is the same quantity by which the sum of squares of the two diagonal coefficients increases. Hence the sum of squares of all the coefficients remains unchanged, which is also true of the sum of squares of the constant terms. From this we see the following. Suppose that we treat the transformed system in a similar way, applying transformations several times one after the other while each time removing the most influential off-diagonal coefficient. Then in the last system so obtained

1) the sum of the diagonal coefficients, the sum of squares of all the coefficients, and the sum of squares of the constant terms are all the same as in the original system;

2) the sum of squares of the diagonal coefficients increases, and the sum of squares of the off-diagonal coefficients

decreases by the same amount: namely, twice the sum of squares of of the coefficients that were annihilated by the individual transformations.

In this way we can transform the equations to be solved in the application of the method of least squares into other equations that permit the use of the iterative method given at the outset. In fact, it is easy to show that if we keep transforming indefinitely, always annihilating the largest off-diagonal coefficient, we can make the off-diagonal coefficients smaller than any given quantity. However, at a certain point, which is best left to the judgement of the calculator, it will be profitable to switch to the iterative method. If this is done too early, the iteration method itself will show which coefficient is making the results uncertain and hence must be annihilated by new transformations.

Since $\eta\eta + \eta_1\eta_1 = xx + x_1x_1$ and the remaining unknowns x_2, x_3 etc. remain unaltered by the transformation, the sum of squares of all the unknowns retain the value throughout the successive transformations. If we denote by s, s_1, s_2 etc. the unknowns of the system we arrive at after several successive transformations, then

$$1)\quad xx + x_1x_1 + x_2x_2 \text{ etc.} = ss + s_1s_1 + s_2s_2 \text{ etc.}$$

If we collect all the successive substitutions into a single one, so that the original unknowns x, x_1, x_2 etc. are expressed in terms of s, s_1, s_2 etc. from the last equation, then this same formula immediately gives the values of s, s_1, s_2 etc. in terms of x, x_1, x_2 etc. Namely, if we have

$$\begin{aligned} 2)\quad x &= a\ s + b\ s_1 + c\ s_2 \text{ etc.,}\\ x_1 &= a_1s + b_1s_1 + c_1s_2 \text{ etc.,}\\ x_2 &= a_2s + b_2s_1 + c_2s_2 \text{ etc.,}\\ &\text{etc.} \qquad \text{etc.} \end{aligned}$$

it follows from equation 1), which must must remain an identity under this substitution, that

$$\begin{aligned} &s\ (ax + a_1x_1 + a_2x_2 \text{ etc.})\\ &+s_1(bx + b_1x_1 + b_2x_2 \text{ etc.})\\ &+s_2(cx + c_1x_1 + c_2x_2 \text{ etc.})\\ &\qquad \text{etc.} \qquad \text{etc.} \qquad = ss + s_1s_1 + s_2s_2 + \text{etc.} \end{aligned}$$

Hence,

$$\begin{aligned} s &= ax + a_1x_1 + a_2x_2 + \text{etc.}\\ s_2 &= bx + b_1x_1 + b_2x_2 + \text{etc.}\\ s_3 &= cx + c_1x_1 + c_2x_2 + \text{etc.}\\ &\text{etc.} \qquad \text{etc.} \end{aligned}$$

In order to have a check, we can derive the last system of equations all at once from the original by the single substitution 2). Namely, denote the original system of equations as above by

$$u = 0, \quad u_1 = 0, \quad , u_2 = 0, \text{ etc.,}$$

By means of 2), introduce the quantities s, s_1, s_2 etc. in place of x, x_1, x_2 etc., and then form the equations

$$\begin{aligned} 3)\quad au + a_1u_1 + a_2u_2 \text{ etc.} &= 0,\\ bu + b_1u_1 + b_2u_2 \text{ etc.} &= 0,\\ cu + c_1u_1 + c_2u_2 \text{ etc.} &= 0,\\ \text{etc.} \qquad \text{etc.} & \end{aligned}$$

which are the equations finally obtained from the successive transformations. Alternatively one can first form the equation 3) from the original equations and then by means of 2) introduce the quantities s, s_1, s_2 etc. as unknowns. Relations holding between the coefficients, such as

$$\begin{aligned} aa + a_1a_1 + a_2a_2 \text{ etc.} &= 1,\\ aa + a_1a_1 + a_2a_2 \text{ etc.} &= 0,\\ \text{etc.} \qquad \text{etc.} & \end{aligned}$$

$$\begin{aligned} aa + bb + cc \quad \text{etc.} &= 1,\\ aa_1 + bb_1 + cc_1 \quad \text{etc.} &= 0,\\ \text{etc.} \qquad \text{etc.} & \end{aligned}$$

can also serve as checks that can be applied everywhere and in very many ways. In any case one will do well not to start applying the iterative method before convincing himself of the of the correspondence between last equations and the original. And it is a good idea to carry out the calculations necessary to form the equations in higher precision. If the equations divide into several groups that are connected to each other by only a few unknowns, as is case for large triangular networks, the substitution 3) will also divide into corresponding groups.

I will now briefly sketch how the method followed here can be extended to linear systems that are not symmetric about the diagonal; i.e., systems for which $(ik) = (ki)$ does not hold. However, it is essential for the success of the method that the two coefficients (ik) and (ki) not differ too greatly from one another — or rather that when they have significant values they at least have the *same signs*. I will content myself with writing down the results.

The system of equations is once again

$$\begin{aligned} u &= (00)x + (01)x_1 + (02)x_2 \text{ etc.} - (0m) = 0\\ u_1 &= (10)x + (11)x_1 + (12)x_2 \text{ etc.} - (1m) = 0\\ u_2 &= (20)x + (21)x_1 + (22)x_2 \text{ etc.} - (2m) = 0\\ &\qquad \text{etc.} \qquad\qquad \text{etc.} \end{aligned}$$

If the coefficients (01) and (10) have significant values, I set

$$\begin{aligned} cos\,2\Delta.x &= cos(\alpha+\Delta).\eta + sin(\alpha-\Delta).\eta_1,\\ cos\,2\Delta.x_1 &= sin(\alpha+\Delta).\eta - cos(\alpha-\Delta).\eta_1,\end{aligned}$$

where the angles α and Δ are determined by the equations

$$\begin{aligned} \rho\, cos\,2\alpha &= (00)-(11),\\ \rho\, sin\,2\alpha &= (01)+(10),\\ \rho\, sin\,2\Delta &= (10)-(01).\end{aligned}$$

Setting

$$\begin{aligned} \nu &= cos(\alpha-\Delta).u + sin(\alpha-\Delta).u_1,\\ \nu_1 &= sin(\alpha+\Delta).u - cos(\alpha+\Delta).u_1,\end{aligned}$$

I replace the first two equations with $\nu=0$, $\nu_1=0$, so that the transformed system is the following:

$$\nu=0,\quad \nu_1=0,\quad ,u_2=0,\quad ,u_3=0,\quad \text{etc.}$$

In the equation $\nu=0$ the coefficient of η_1 vanishes; in the equation $\nu_1=0$ the coefficient of η vanishes. If we set

$$\begin{aligned} \nu &= [00]\eta + \quad * \quad + [02]x_2 + [03]x_3 \text{ etc.},\\ \nu_1 &= \quad * \quad + [11]\eta_1 + [12]x_2 + [13]x_3 \text{ etc.},\\ u_2 &= [20]\eta + [21]\eta_1 + [22]x_2 + [23]x_3 \text{ etc.},\\ &\qquad \text{etc.} \qquad\qquad \text{etc.}\end{aligned}$$

then we have

$$\begin{aligned} [00] &= \frac{(00)+(11)}{2} + \frac{\rho}{2} cos\,2\Delta,\\ [11] &= \frac{(00)+(11)}{2} - \frac{\rho}{2} cos\,2\Delta,\\ [02] &= cos(\alpha-\Delta)(02) + sin(\alpha-\Delta)(12),\\ [12] &= sin(\alpha+\Delta)(02) - cos(\alpha+\Delta)(12),\\ cos\,2\Delta\,[20] &= cos(\alpha+\Delta)(20) + sin(\alpha+\Delta)(21),\\ cos\,2\Delta\,[21] &= sin(\alpha-\Delta)(20) - cos(\alpha-\Delta)(21).\end{aligned}$$

From these formulas it follows that

$$\begin{aligned} &[00]+[11] = (00)+(11),\\ &[00]^2+[11]^2 = (00)^2+(11)^2+2(01)(10),\\ &[02][20]+[12][21] = (02)(20)+(12)(21)\end{aligned}$$

These equations show that however many times the transformation is successively applied the sums

$$\Sigma[ii], \qquad \Sigma\{(ii)(ii)+2(ik)(ki)\}$$

remain unchanged. Moreover the second sum remains unchanged in such a way that $\Sigma.(ii)^2$ is always larger and $2\Sigma(ik)(ik)$ is always smaller by twice the product of the two coefficients annihilated by the transformation in question. Having reducted the off-diagonal coefficients sufficiently by repeating the transformation, we can apply the iterative method I described at the outset.

The method given here can be used with even greater profit when the equations to be solved have the following form:

$$\begin{aligned} &\{(00)-G\}x + (01)x_1 + (02)x_2 \text{ etc.} = 0,\\ &(10)x + \{(11)-G\}x_1 + (12)x_2 \text{ etc.} = 0,\\ &(20)x + (21)x_1 + \{(22)-G\}x_2 \text{ etc.} = 0.\end{aligned}$$

As is well known, by eliminating the unknowns x, x_1, x_2, etc. one gets a higher equation whose roots are the various values of G. For each of these values we must determine the ratios of x, x_1, x_2, etc. In this case the preliminary transformations turn out to be the same for all the systems corresponding to the various values of G, and they give these values directly and with increasing accuracy, without the necessity of forming the the higher equation. Thereafter a method similar to the one described at the outset gives the small corrections in the values of G and the ratios of the unknowns corresponding to these values. Here I will content myself with the above sketch, since the method and its application to the secular perturbations of the *seven* chief planets will be presented in another paper. There it will be seen from the calculations my learned friend Herr Dr. *Seidl* has so carefully performed that owing to the speed and stability with which one arrives at an accurate approximation to the final results the method has noteworthy advantages over the one used by Herr *Leverrier*.

The application of the method to the equations given in the Theoria motus p. 219 will serve as an example here. The original equations are

$$\begin{aligned} 27p + 6q + *r - 88 &= 0\\ 6p + 15q + r - 70 &= 0\\ *p + q + 54 - 107 &= 0.\end{aligned}$$

If the coefficient 6 of q in the first equation is eliminated, then $\alpha = 22°30'$. Hence

$$\begin{aligned} p &= 0,92390y + 0,38268y'\\ q &= 0,38268y - 0,92390y'\end{aligned}$$

and the new equations are

$$\begin{aligned} 29,4853y + \quad * \quad y' + 0,38268r - 108,0901 &= 0\\ * \quad y + 12,5147y' - 0,92390r + 30,9967 &= 0\\ 0,38268y - 0,92390y' + 54r \qquad - 107 \qquad &= 0\end{aligned}$$

From them comes the first approximation

$$\begin{aligned} log\,y &= 0,56419\\ log\,y' &= 0,39389n\\ log\,r &= 0,29699.\end{aligned}$$

The second approximation

$$\begin{aligned} \log y &= 0,56114 \\ \log y' &= 0,36746n \\ \log r &= 0,28174. \end{aligned}$$

After two more easy corrections, we get the exact values

$$\begin{aligned} \log y &= 0,56125 \\ \log y' &= 0,36836n \\ \log r &= 0,28233, \end{aligned}$$

from which come the values

$$\begin{aligned} \log p &= 0,39276 \\ \log q &= 0,55036 \end{aligned}$$

The weights of y, y', and r are very near the coefficients on the diagonal. In fact, from them one obtains the logarithms of the weights of

p 1,39092
q 1,13565
r 1,73239

which are very near the true weights.

Berlin Nov. 17 1844 *C. G. J. Jacobi.*