ACTA CHEMICA SCANDINAVICA 18 (1964) 1085–1098

High-speed Computers as a Supplement to Graphical Methods

III. Twist Matrix Methods for Minimizing the Error-square Sum in Problems with Many Unknown Constants

LARS GUNNAR SILLÉN

Department of inorganic chemistry, Royal Institute of Technology (KTH), Stockholm 70, Sweden

Matrix equations are given for use in the general minimizing computer program LETAGROP VRID. Like the earlier program LETAGROP,[1,2] it finds the position of the minimum for a function $U(k_1...k_N)$ — which may be an error-square sum — by calculating U for $\frac{1}{2}(N+1)\ (N+2)$ systematically chosen sets **k** and assuming $U(k_1...k_N)$ to be a second-degree surface. In cases with "skew pits" (covariation of the k_i), the coordinates are transformed by means of a triangular "twist matrix" **S**.

In many problems, some of the constants (*e.g.* equilibrium constants) cannot be negative. If any such constants are found to be negative at the minimum, they are set equal to zero, and the minimum for U is calculated for the "reduced pit", which means the section in (U,**k**) space where the "minus" constants have been eliminated (MIKO). The necessary matrix equations for this operation are given.

These methods, in conjunction with more rapid computers, seem to open a path for a more systematic trial of all possible combinations than has been possible to date. To avoid self-deception in a mechanized data treatment it seems advisable to treat all conceivable systematic errors as unknown constants to be determined.

Part I of this series [1] deals with the principles of LETAGROP, a series of computer programs designed for finding the values for a set of unknown constants $k_1k_2...k_N$ that will minimize an error-square sum defined by (w = weight)

$$U = \Sigma w(y_{exp} - y_{calc})^2 \qquad (1)$$

Here, y_{exp} is a measured quantity, and y_{calc} is obtained from a functional relationship

$$y = f(k_1, k_2...k_N; a_1, a_2...) \qquad (2)$$

where a_1, etc., are quantities assumed to be known. With a computer it is easy enough to calculate U for various sets, $\mathbf{k}(k_1,k_2,k_3...k_N)$, even if the rela-

tionship between y and the k in (2) is not linear (as required by the standard "least-squares method") and not even explicit. Parts I[1] and II[2] indicate a number of applications to chemical problems.

Whatever the original problem, we have reduced it to finding the minimum for a function

$$U(\mathbf{k}) = U(k_1 \ldots k_N) \tag{3}$$

The clue to the solution is that around the minimum, U approximates a second-degree surface (a paraboloid) in $(N+1)$-dimensional space. We start with a "central" set $\mathbf{k}_c$ (estimated or guessed), and calculate U for $\mathbf{k}_c$ and for sets where one or two elements in $\mathbf{k}_c$ have been changed by given steps h_i. From the U values for $\frac{1}{2}(N+1)\,(N+2)$ systematically chosen points, we may calculate the coefficients of the equation for a second-degree surface through these points, and hence the position $\mathbf{k}_0$, of the minimum point of that surface. For brevity, the procedure described until now will be referred to as a "shot".

The $\mathbf{k}_0$ obtained by the first shot may be used as the central value for the next shot, etc.

The original program LETAGROP has been applied to a considerable number of chemical systems and in general has performed well, as seen for instance in Refs. [4-8] (A recent paper by Tobias and Yasuda [9] probably also refers to an early edition of LETAGROP).

Some difficulties were, however, met with in cases which may be described as "skew pits", or strong covariation of the k_i. In Fig. 1 as in most other figures we shall restrict ourselves for simplicity to a case with two constants k_1 and k_2. The reader's imagination may help him to extend the idea to $(N+1)$-dimensional space, $N>2$.

Suppose that the region of the surface (3) close to the minimum, the "pit", has its main axes at angles to the coordinate axes, k_i, and moreover is narrow and steep in some directions, so that it looks like a cleft (Fig. 1a). In the first

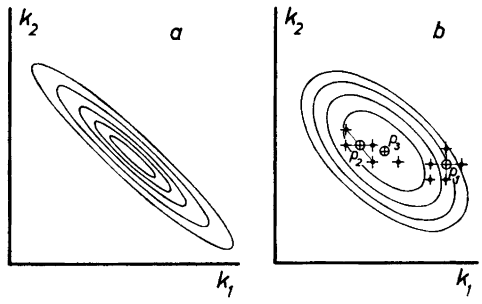

Fig. 1. a) Schematic map of surface $U(k_1, k_2)$, giving error square sum U as a function of constants k_1 and k_2. The pit is skew and comparatively narrow. b) Same type of map, pit for clarity drawn somewhat broader. P_1 is a first guess. The crosses around P_1 indicate the sets for which U is calculated in first "shot"; like in the original LETAGROP the variation is made parallel to axes. The calculated minimum is at P_2. In the second "shot", the twist matrix is available and variation is made along twisted axes. The next calculated minimum is indicated at P_3.

LETAGROP, $\mathbf{k}$ was varied by steps parallel to the coordinate axes. When the steps were chosen too large, and the pit was steep and skew, the points fell high up on the walls of the "cleft", and terms of higher degree than the second became important. When we used small steps, the rounding errors in the computer (which after all has a limited precision) caused rounding errors

in the calculation of $\mathbf{k}_0$. In either case the computer took aim badly, as compared with well-behaving systems, where in general two or three shots were enough to give a satisfactory minimum.

It was not difficult to program the computer so that it always used, for the next calculation, the point with the lowest U value it had found thus far; so, the computer was always working toward the minimum. On the other hand, it was thought (and later confirmed by experience) that we could greatly increase the speed and accuracy of aiming if we varied $\mathbf{k}$ along the main axes of the pit, instead of parallel to the coordinate axes. (Fig. 1 b). The programs with this device were called LETAGROP VRID, (from Swedish vrida = twist, turn); examples of successful application are Refs.[10-13]

A further improvement was the procedure MIKO for eliminating "minus" complexes, which will be described in the following.

The equations needed for these operations are very conveniently derived and expressed by means of matrices. They are given here since some chemist may want to apply the principles of LETAGROP VRID and MIKO for special purposes, and might be helped by having the equations handy. It does not seem impossible that a similar treatment has already been applied to some other problem, although I am not aware of it.

NOTATION

A square matrix (in general $N \times N$) is denoted by a bold-face capital such as $\mathbf{S}$, $\mathbf{H}$, etc. The square matrices we shall consider will mostly be symmetrical (or even diagonal) or triangular (with zeros in lower left part). A vector will be denoted by a bold-face lower case symbol such as $\mathbf{k}$. If it is provided with a perpendicular arrow, or with no arrow, it is a column matrix (dimension in general $N \times 1$), if it carries a horizontal arrow, it is a row matrix (in general $1 \times N$).

Symbols (numbers refer to equations): $\mathbf{A}$ symmetrical matrix in (26, 29 and 35); a_i = known quantity in (2); $\mathbf{b} = \mathbf{k}_0$ in (33) etc; $\mathbf{b}_i$ = vector with one nonzero element in (22); $\mathbf{C}$ = symmetrical matrix in (21); $\mathbf{c} = \mathbf{k}_c$ = central set of "constants" (4); $\mathbf{H}$ = diagonal step matrix, elements = h_i (5); h_i = step in variation of k_i (5); i, j = integer indices; $\mathbf{k}$ = set of unknown "constants" k_i (3, 4); $\mathbf{k}_c$ = central set $\mathbf{k}$ (4); $\mathbf{k}_0$ = set $\mathbf{k}$ at calculated minimum of second-degree surface $U(\mathbf{k})$ (13); $\mathbf{M}$ = matrix in (36, 40); N = number of k_i to be varied; $\mathbf{p}$ = vector of linear coefficients, p_i, in $U(\mathbf{v})$, (7, 9); $\mathbf{p}'$ = $\mathbf{p}$ in reduced pit (38, 39); $\mathbf{R}$ = symmetrical matrix with second-degree coefficients r_{ij} in $U(\mathbf{v})$, (7,10,11,25); $\mathbf{R}'$ = $\mathbf{R}$ in reduced pit (38,39); $\mathbf{R}_n$ = submatrix of $\mathbf{R}$ (19a); $\mathbf{r}_m$ = part of column in $\mathbf{R}$ (19a); $\mathbf{S}$ = triangular twist matrix (6); s_{ij} = element in $\mathbf{S}$; $\mathbf{S}'$ = corrected twist matrix (17,18); ^{T} = transposed matrix; U = error square sum, or other function of $\mathbf{k}$ to be minimized (1,3); U_c = value for U with central set $\mathbf{k}_c$ (7); U_0 = value for U at calculated minimum (11,12b); $U'_c = U_c$ in reduced pit (38,39); U_i, U_{-i}, U_{ij} = value for U when all elements in $\mathbf{v}$ are = 0 except for $v_i = 1$ (U_i), or $v_i = -1$ (U_{-i}), or $v_i = v_j = 1$ (U_{ij}) (8); $\mathbf{v}$ = variation vector (4); $\mathbf{v}'$ = $\mathbf{v}$ corrected with $\mathbf{S}'$ (15,17); $\mathbf{v}_0$ = value for $\mathbf{v}$ at calculated minimum (12a); $\mathbf{w}$ = variation vector in reduced pit (36); w = weight (1); $\mathbf{W}$ = triangular correcting matrix (14,19a); w_{ij} = element in $\mathbf{W}$;

$\mathbf{w}_m$ = part of column in **W** (19a); **x** = vector in (21) and (31); **y** = vector in (31b); $y = y_{calc}$ = quantity defined by (2); y_{exp} = measured quantity; σ = standard deviation.

The superscripts, $^{-1}$ or ^{inv} refer to matrix inversion, the subscripts $_+$ and $_-$ are explained in the text before eqn. (31).

VARIATION AND SHOT

In each shot the preceding systematic variation of **k** around the central value $\mathbf{k}_c$ (or shorter, **c**) is conveniently described by the equation

$$\mathbf{k} = \mathbf{c} + \mathbf{S}\,\mathbf{H}\,\mathbf{v} \tag{4}$$

Here, **k** and **c** are column matrices (vectors) with N elements, $|k_1 k_2 ... k_N|$ and $|c_1 c_2 ... c_N|$. The step matrix, **H**, is diagonal (5) and its elements are the steps assigned to each constant, since the elements in **v** are either 0 or ± 1 (see below). The twist matrix, **S** (6) defines the directions in which to vary the vector **k**. **S** is triangular, and the diagonal terms are always chosen as 1. In the very first shot, all other terms (even those to the upper right) are set equal to zero, but the calculations then usually give nonzero values to the s_{ij}, which are improved by each shot, like the k_i.

$$\mathbf{H} = \begin{vmatrix} h_1 & 0 & \dots & 0 \\ 0 & h_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & h_N \end{vmatrix} \quad (5); \qquad \mathbf{S} = \begin{vmatrix} 1 & s_{12} & \dots & s_{1N} \\ 0 & 1 & \dots & s_{2N} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{vmatrix} \tag{6}$$

If $U(\mathbf{k})$ is a second-degree function, then $U(\mathbf{v})$ must also be so. If U_c is the value at the central point ($\mathbf{k} = \mathbf{c}$, $\mathbf{v} = 0$) then we may express $U(\mathbf{v})$ by means of

$$U = U_c - 2\mathbf{p}\,\mathbf{v} + \mathbf{v}\,\mathbf{R}\,\mathbf{v} \tag{7}$$

The variation vector **v** $|v_1 v_2 ... v_N|$ is, for simplicity, chosen so that each element is either 0 or ± 1. In addition to U_c, the value for the central point, we calculate U values for **v** vectors where all elements are equal to zero except for

$$\left.\begin{array}{ll} v_i = 1 \; ; & U_i = U_c - 2p_i + r_{ii} \\ v_i = -1 \; ; & U_{-i} = U_c + 2p_i + r_{ii} \\ v_i = v_j = 1 \; ; & U_{ij} = U_c - 2p_i - 2p_j + r_{ii} + r_{jj} + 2r_{ij} \end{array}\right\} \tag{8}$$

We thus calculate U_i and U_{-i} for $i = 1$ through N, and U_{ij} for $i = 1$ through N and $j = (i + 1)$ through N. For reasons of symmetry, $U_{ij} = U_{ji}$ and $r_{ij} = r_{ji}$. (In the actual calculation, we may reverse the sign for some coordinate to get closer to the minimum, but this does not change our equations).

We may now calculate the terms in $\mathbf{p}$ and $\mathbf{R}$:

$$p_i = 0.25(U_{-i} - U_i) \tag{9}$$

$$\left.\begin{aligned} r_{ii} &= 0.5(U_i + U_{-i}) - U_c \\ r_{ij} &= 0.5(U_{ij} - U_c) + (p_i + p_j) - 0.5\,(r_{ii} + r_{jj}) \end{aligned}\right\} \tag{10}$$

To find the vector $\mathbf{v}_0$ corresponding to the minimum U_0, we may express the equation $U(\mathbf{v})$ as follows:

$$U = U_0 + \overrightarrow{(\mathbf{v}-\mathbf{v}_0)}\,\mathbf{R}\,\downarrow(\mathbf{v} - \mathbf{v}_0) = U_0 + \overrightarrow{\mathbf{v}_0}\mathbf{R}\downarrow\mathbf{v}_0 - 2\overrightarrow{\mathbf{v}_0}\mathbf{R}\downarrow\mathbf{v} + \overrightarrow{\mathbf{v}}\mathbf{R}\downarrow\mathbf{v} \tag{11}$$

A comparison of eqns. (7) and (11) gives

$$U_c = U_0 + \overrightarrow{\mathbf{v}_0}\mathbf{R}\downarrow\mathbf{v}_0; \quad \overrightarrow{\mathbf{p}} = \overrightarrow{\mathbf{v}_0}\,\mathbf{R};$$

and hence

$$\overrightarrow{\mathbf{v}_0} = \overrightarrow{\mathbf{p}}\mathbf{R}^{-1}; \quad U_0 = U_c - \overrightarrow{\mathbf{p}}\downarrow\mathbf{v}_0 \tag{12a; 12b}$$

When $\mathbf{v}_0$ is known, the vector $\mathbf{k}_0$ at the calculated minimum point follows from (4)

$$\downarrow\mathbf{k}_0 = \downarrow\mathbf{c} + \mathbf{S}\downarrow\mathbf{H}\downarrow\mathbf{v}_0 \tag{13}$$

IMPROVING THE TWIST MATRIX

We want to choose the twist matrix $\mathbf{S}$ so that $\mathbf{R}$ becomes a diagonal matrix ($r_{ij} = 0$ for $i \neq j$). In general this is not the case after a shot so we have to improve $\mathbf{S}$ by the aid of information contained in $\mathbf{R}$. We have chosen to introduce a correcting matrix $\mathbf{W}$, and to make $\mathbf{W}$ triangular, with unit diagonal:

$$\mathbf{W} = \begin{vmatrix} 1 & w_{12} & w_{13} & \ldots & w_{1N} \\ 0 & 1 & w_{23} & \ldots & w_{2N} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & w_{N-1,N} \\ 0 & 0 & 0 & \ldots & 1 \end{vmatrix} \tag{14}$$

Let us write

$$\downarrow\mathbf{v} = \mathbf{W}\downarrow\mathbf{v}'; \quad \overrightarrow{\mathbf{v}} = \overrightarrow{\mathbf{v}'}\,\mathbf{W}^T \tag{15}$$

Now we wish to choose $\mathbf{W}$ so that all mixed second-degree terms disappear if the variation is made with $\mathbf{v}'$ instead of with $\mathbf{v}$. This means that in

$$\overrightarrow{\mathbf{v}}\mathbf{R}\downarrow\mathbf{v} = \overrightarrow{\mathbf{v}'}\mathbf{W}^T\mathbf{R}\mathbf{W}\downarrow\mathbf{v}' \tag{16}$$

the matrix $\mathbf{W}^T\mathbf{R}\mathbf{W}$ should be strictly diagonal. The relationship with the new twist matrix $\mathbf{S}'$ is then, from (4) and (15):

$$\mathbf{k} - \mathbf{c} = \mathbf{S}\mathbf{H}\mathbf{W}\mathbf{v}' = \mathbf{S}'\mathbf{H}\mathbf{v}', \text{ or } \mathbf{S}\mathbf{H}\mathbf{W} = \mathbf{S}'\mathbf{H} \tag{17}$$

This gives

$$\mathbf{S}' = \mathbf{SHWH}^{-1} \tag{18}$$

To find $\mathbf{S}'$ we must thus calculate $\mathbf{W}$ from $\mathbf{R}$. The conditions for $\mathbf{W}^T\mathbf{RW}$ to be diagonal are the following equations

$$\begin{aligned}
&r_{11}w_{12} + r_{12} = 0\\
&r_{11}w_{13} + r_{12}w_{23} + r_{13} = 0\\
&r_{21}w_{13} + r_{22}w_{23} + r_{23} = 0\\
&r_{11}w_{14} + r_{12}w_{24} + r_{13}w_{34} + r_{14} = 0\\
&r_{21}w_{14} + r_{22}w_{24} + r_{23}w_{34} + r_{24} = 0\\
&r_{31}w_{14} + r_{32}w_{24} + r_{33}w_{34} + r_{34} = 0 \text{ etc.}
\end{aligned} \tag{19}$$

We let $\mathbf{R}_m$ be the $m \times m$ submatrix of $\mathbf{R}$ which contains rows no. 1 through m and columns no. 1 through m, and $\vec{\mathbf{r}}_m$ be the vector $|r_{1m}, r_{2m} \ldots r_{m-1,m}|$; $\mathbf{w}_m$ is defined analogously. Then (19) may be expressed as

$$\begin{aligned}
&w_{12} = -r_{12}/r_{11}\\
&\vec{\mathbf{w}}_3\mathbf{R}_2 + \vec{\mathbf{r}}_3 = 0;\ \vec{\mathbf{w}}_3 = -\vec{\mathbf{r}}_3\mathbf{R}_2^{-1}\\
&\vec{\mathbf{w}}_4\mathbf{R}_3 + \vec{\mathbf{r}}_4 = 0;\ \vec{\mathbf{w}}_4 = -\vec{\mathbf{r}}_4\mathbf{R}_3^{-1} \text{ etc.}
\end{aligned} \tag{19a}$$

By inserting (19) one may verify that the product $\mathbf{W}^T\mathbf{R}$ becomes a triangular matrix with zeros in the lower right half. On multiplication with $\mathbf{W}$, only the diagonal terms remain. Applying eqns. (19) we may thus find the various columns of $\mathbf{W}$, after which we can use (18) to calculate the new twist matrix.

Eqns. (19a) and (18) are those actually applied in the programs now in use, and were derived by us for the purpose. We have later been told that a specialist might have calculated $\mathbf{S}'$ by several alternative standard methods, *e.g.* Crout's or Choleski's, which might result in some economy of computer time. On the other hand the time required by the calculation of U is at any rate much longer than that spent with the matrices, so there has been no strong incentive to try any other approach.

THE STANDARD DEVIATIONS

If the equation of the supersurface $U(\mathbf{k})$ is known, one obtains the square of the standard deviation for y, $\sigma^2(y)$, by dividing the value U_0 at the minimum point by the number of degrees of freedom (part I,[1] eqn. 17); the latter is equal to the difference between the number of experiments, and the number of unknown constants to be determined. (The $\sigma(y)$ one obtains corresponds to the spread in measurements of weight = 1. If one has chosen weights that differ much from unity, the result may at first look surprising).

For defining the standard deviations, $\sigma(\mathbf{k})$, of the various constants, we have introduced (part I,[1] eqn. 47) what we called the "D boundary", which

is the curve or supercurve on which $U = U_0 + \sigma^2(y)$. We defined the standard deviation for each k_i as the maximum difference,

$$\sigma(k_i) = D_i = \max ((\mathbf{k}_D - \mathbf{k}_0)_i) \tag{20}$$

between the value for k_i at any point on the D boundary, and the value for k_i at the minimum. (Part I,[1] p. 168 and eqn. (51))

If $y(\mathbf{k})$ is a linear function, this definition for $\sigma(k_i)$ gives exactly the same result as the traditional one. For non-linear cases, we know of no generally recognized definition of the $\sigma(k_i)$.

To be strict, in our calculations we often use points closer to the minimum and assume that the second-degree approximation is valid out to $U = U_0 + \sigma^2$, which is permissible as a rule. If, in some special case, terms of higher order become important even at $U - U_0 = \sigma^2$, then it would seem preferable to use the idealized D boundary calculated from the second-degree surface coinciding with $U(\mathbf{k})$ close to the minimum. At any rate, in such a case our first wish would be to have better data rather than a better statistical treatment of those available.

If the pit is skew, the maximum deviation $\sigma(k_i)$ for a certain constant, k_i on the D boundary, may be considerably larger than $h_i\ \sigma(v_i)$, the shift in k_i from the minimum to the D boundary along the corresponding twisted axis (Fig. 2). The chemist has a choice either to define a new set of constants

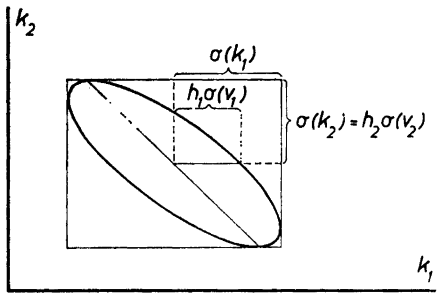

Fig. 2. Schematic comparison of two measures of standard deviation, $\sigma(k_i)$ and $h_i\sigma(v_i)$. Curve = D boundary, with minimum in center. For the last k_i, k_N, the two measures are identical. Because **S** is triangular, k_N is not changed by a shift in the other constants (see (6) and (4)).

(for instance to replace some of the earlier ones by their products, or ratios) so as to get smaller covariation, or to use the somewhat pessimistic maximum deviations $\sigma(k_i)$. At any rate, in our variation routine we need the quantities, $h_i\sigma(v_i)$, for adjusting our "steps" h_i. So, we have found it desirable to calculate both $\sigma(k_i)$ and $\sigma(v_i)$.

Let us shift the origin to the minimum point, and express $(U-U_0)$ as a second-degree function of some vector $\mathbf{x}$, which may be either $(\mathbf{k}-\mathbf{k}_0)$ or $(\mathbf{v}-\mathbf{v}_0)$:

$$U-U_0 = \overrightarrow{\mathbf{x}}\mathbf{C}\overrightarrow{\mathbf{x}} = \sigma^2(y) \tag{21}$$

C is a symmetrical $N \times N$ matrix. Differentiating (21), we get for a variation along the D boundary:

$$\overrightarrow{\mathrm{d}\mathbf{x}}\ \mathbf{C}\ \mathbf{x} = 0$$

The condition for the coordinate x_i to have a maximum or minimum is that $dx_i = 0$, independent of the other dx. The coefficients for the other dx must then be zero, whereas that for dx_i is a non-zero constant which we may call b_i. If now we let $\mathbf{b}_i$ be a vector for which all members are zero except for the i:th, which is b_i, then the condition for x_i to have an extreme value can be written as:

$$\overset{\downarrow}{\mathbf{C}}\overset{\downarrow}{\mathbf{x}} = \overset{\downarrow}{\mathbf{b}_i}; \quad \overset{\downarrow}{\mathbf{x}} = \mathbf{C}^{-1}\overset{\downarrow}{\mathbf{b}_i}; \quad \overset{\rightarrow}{\mathbf{x}} = \overset{\rightarrow}{\mathbf{b}_i}\,\mathbf{C}^{-1} \tag{22}$$

If we insert (22) into (21), the equation for the D boundary, we find

$$\sigma^2(y) = \overset{\rightarrow}{\mathbf{x}}\overset{\downarrow}{\mathbf{b}_i} = x_i b_i; \quad \sigma^2(y) = \overset{\rightarrow}{\mathbf{b}_i}\mathbf{C}^{-1}\overset{\downarrow}{\mathbf{b}_i} = c_{ii}^{\text{inv}}\, b_i^2 \tag{23}$$

Eliminating b_i we find for the extreme value (compare eqns. 20 and 21)

$$\sigma(x_i) = |\text{extreme } x_i| = \sigma(y)\sqrt{c_{ii}^{\text{inv}}} \tag{24}$$

Under the square root sign we find a diagonal element in the inverse matrix $\mathbf{C}^{-1}$. We may write the equation for the D boundary in two forms, using $\mathbf{v}$ or $\mathbf{k}$ ($\mathbf{v}_0$ and $\mathbf{k}_0$ are as usual the vectors at the minimum):

$$U - U_0 = \sigma^2(y) = \overset{\rightarrow}{(\mathbf{v} - \mathbf{v}_0)}\mathbf{R}\overset{\downarrow}{(\mathbf{v} - \mathbf{v}_0)} \tag{25}$$

$$U - U_0 = \sigma^2(y) = \overset{\rightarrow}{(\mathbf{k} - \mathbf{k}_0)}\mathbf{A}\overset{\downarrow}{(\mathbf{k} - \mathbf{k}_0)} \tag{26}$$

We may compare (21) and (25) and find that $\mathbf{x}$ corresponds to $\mathbf{v} - \mathbf{v}_0$ and $\mathbf{C}$ to $\mathbf{R}$. Our result (24) will then give

$$\sigma(v_i) = \sigma(y)\sqrt{r_{ii}^{\text{inv}}} \tag{27}$$

The expression to the right in (27) can be obtained immediately from our data and can be used, in conjunction with the ealier **SH**, to adjust the steps in the next variation.

From (26) we find similarly

$$\sigma(k_i) = \sigma(y)\sqrt{a_{ii}^{\text{inv}}} \tag{28}$$

Under the square root sign is a diagonal term in $\mathbf{A}^{-1}$, the inverse of the square matrix in $U(\mathbf{k})$ (eqn. (26)). From the transformation (4) between $\mathbf{k}$ and $\mathbf{v}$ we find, comparing (25) and (26),

$$\mathbf{R} = (\mathbf{SH})^{\mathrm{T}}\mathbf{ASH} \tag{29}$$

By applying standard methods we derive from (29)

$$\mathbf{A}^{-1} = \mathbf{SH}\ \mathbf{R}^{-1}\ (\mathbf{SH})^{\mathrm{T}} \tag{30}$$

Using (30) we may thus calculate the diagonal terms in $\mathbf{A}^{-1}$ which we need for calculating $\sigma(k_i)$ with (28).

ELIMINATION OF "MINUS" CONSTANTS

When one searches the "best" values for a number of equilibrium constants, one or more of these may turn out to be negative at the minimum for U. For one thing, when the computer tries to solve equations containing one or more negative equilibrium constants, the calculations may go awry so that the computer stops or is caught in a loop. More important, however, negative equilibrium constants cannot have a physical meaning.

It is easy enough to add to the program some safeguard that will make it impossible for any of the "non-negative" unknown constants to become negative during a variation, or in the final check on U. What we really want is, however, the vector **k** that gives the lowest value for U and still has a physical meaning, thus containing no negative equilibrium constants.

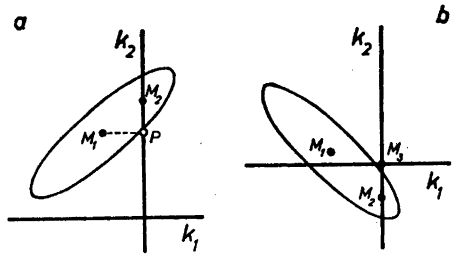

Fig. 3. Schematic. Two cases of "minus constants". Curve joins points of same U, such as the D boundary. a) First shot gives minimum M_1 with negative k_1. M_2 is minimum in "reduced pit" = section with $k_1 = 0$, P = projection of M_1 on reduced pit.
b) First shot gives minimum at M_1 with negative k_1 and positive k_2. In section with $k_1 = 0$, the minimum M_2 gives a negative k_2 so that k_2 must also be eliminated and M_3 is the "best" permissible minimum.

Fig. 3a gives a schematic two-dimensional picture for $U(k_1, k_2)$. M_1 is the calculated minimum, with a surrounding pit contour, (*e.g.* the D boundary) and k_1 is negative in M_1. The "best" value for k_2 would correspond to M_2, which gives a minimum in U in the "reduced pit", that is the section with $k_1 = 0$. This is not identical with the "projection" point P obtained by changing the negative value for k_1 at M_1 to zero. We shall now see, for an N-dimensional case, how we may find the minimum after eliminating the "minus" constants.

In the following discussion we shall write each vector, **x**, as a sum of two vectors of the same length as **x**. In one of them, $\mathbf{x}_+$, all those elements that correspond to the "minus" constants in the "shot" have been set equal to zero. In the other vector, $\mathbf{x}_-$ the only non-zero members are those that correspond to the "minus" constants. For instance, if there were five constants ($N = 5$), and k_2 and k_5 turned out to be negative in the shot, then the two vectors would be:

$$\mathbf{x}_+ = |x_1\ 0\ x_3\ x_4\ 0| \text{ and } \mathbf{x}_- = |0\ x_2\ 0\ 0\ x_5| \tag{31}$$

Similarly we may divide any square ($N \times N$) matrix **A** into four parts, where the first subscript refers to the rows and the second to the columns:

$$\mathbf{x} = \mathbf{x}_+ + \mathbf{x}_-; \quad \mathbf{A} = \mathbf{A}_{++} + \mathbf{A}_{+-} + \mathbf{A}_{-+} + \mathbf{A}_{--} \tag{31a}$$

In the example given, a_{41} would be in $\mathbf{A}_{++}$ and a_{35} in $\mathbf{A}_{+-}$. For the products of two vectors, a vector and a square matrix, and two square matrices, we would have, for instance:

$$\vec{\mathbf{x}}\downarrow\ \vec{\mathbf{y}}\downarrow = \mathbf{x}_+\mathbf{y}_+ + \mathbf{x}_-\mathbf{y}_-; \quad (\vec{\mathbf{x}}\mathbf{A})_+ = \mathbf{x}_+\mathbf{A}_{++} + \mathbf{x}_-\mathbf{A}_{-+};$$
$$(\mathbf{AB})_{++} = \mathbf{A}_{++}\mathbf{B}_{++} + \mathbf{A}_{+-}\mathbf{B}_{-+} \text{ etc.} \tag{31b}$$

Using the results of a "shot" we may express the equation for the second-degree U surface as follows

$$U - U_0 = (\vec{\mathbf{v}} - \vec{\mathbf{v}}_0)\,\mathbf{R}(\downarrow\mathbf{v} - \downarrow\mathbf{v}_0) \tag{32}$$

We shall write for simplicity

$$\mathbf{b} = \mathbf{k}_0 \tag{33}$$

Using (4) we find

$$\downarrow\mathbf{k} - \downarrow\mathbf{b} = (\mathbf{SH})\,(\downarrow\mathbf{v} - \downarrow\mathbf{v}_0) \tag{34a}$$

$$\downarrow\mathbf{v} - \downarrow\mathbf{v}_0 = (\mathbf{SH})^{-1}\,(\downarrow\mathbf{k} - \downarrow\mathbf{b}) \tag{34b}$$

Introducing (34b) into (32), we find, using (29):

$$U - U_0 = (\vec{\mathbf{k}} - \vec{\mathbf{b}})\mathbf{A}(\downarrow\mathbf{k} - \downarrow\mathbf{b}) \tag{35}$$

$$\mathbf{A} = ((\mathbf{SH})^{\mathrm{T}})^{-1}\,\mathbf{R}(\mathbf{SH})^{-1} \tag{35a}$$

Eqn. (35) is the fundamental relationship $U(\mathbf{k})$; we shall remember that $\mathbf{v}$ is an auxiliary set of coordinates, the directions and origin of which are shifted during the calculation.

Now we shall consider only the "reduced pit", and thus the section in (U,$\mathbf{k}$) space in which all the "minus" constants are exactly zero, so that $\mathbf{k}$ is equal to the "plus" part $\mathbf{k}_+$. Moreover, we shall choose as the starting point the projection $\mathbf{b}_+$ on that section, of the calculated minimum $\mathbf{b}$, using a new variation vector and a new matrix $\mathbf{M}$, with only "plus" components. Then we have (compare eqn. 4)

$$\mathbf{k} = \mathbf{k}_+; \quad \mathbf{b} = \mathbf{b}_+ + \mathbf{b}_-; \quad \downarrow\mathbf{k} = \downarrow\mathbf{b}_+ + \mathbf{M}_{++}\downarrow\mathbf{w}_+ \tag{36}$$

hence

$$\downarrow\mathbf{k} - \downarrow\mathbf{b} = \mathbf{M}_{++}\downarrow\mathbf{w}_+ - \downarrow\mathbf{b}_-; \quad \vec{\mathbf{k}} - \vec{\mathbf{b}} = \vec{\mathbf{w}}_+\mathbf{M}^{\mathrm{T}}{}_{++} - \vec{\mathbf{b}}_- \tag{37}$$

Introducing (37) into (35) we find

$$U = U'_{\mathrm{c}} - 2\vec{\mathbf{p}}'\downarrow\mathbf{w} + \vec{\mathbf{w}}\mathbf{R}'\downarrow\mathbf{w} \tag{38}$$

where

$$U'_{\mathrm{c}} = U_0 + \vec{\mathbf{b}}_-\mathbf{A}_{--}\downarrow\mathbf{b}_-; \quad \vec{\mathbf{p}}' = \vec{\mathbf{b}}_-\mathbf{A}_{-+}\mathbf{M}_{++}; \quad \mathbf{R}' = \mathbf{M}^{\mathrm{T}}{}_{++}\mathbf{A}_{++}\mathbf{M}_{++} \tag{39}$$

The matrix **M** can be chosen arbitrarily. We shall use what is perhaps the simplest choice, namely the same transformation matrix **SH** as earlier, only after eliminating the rows and the columns that correspond to "minus" constants

$$\mathbf{M}_{++} = (\mathbf{SH})_{++} \tag{40}$$

In the program we have developed,[3] U'_c, $\mathbf{p}'$ and $\mathbf{R}'$ are calculated in a special block MIKO, after which the minimum and standard deviations of the "reduced pit" are again calculated in the block GROP, since it seems practical to use the operations already available for a normal pit. However, one could also have calculated the reduced minimum directly, from

$$\mathbf{k}_+ = \mathbf{b}_+ + \mathbf{b}_-\mathbf{A}_{-+}\mathbf{A}_{++}^{-1} \tag{41}$$

(This follows from solving (38) for $\mathbf{w}_0$, like in (12a), and inserting (39). Before inverting $\mathbf{A}_{++}$ in (41), one must reduce its size by eliminating all "minus" rows and columns).

After the "minus" constants have been made zero in this way it sometimes happens that in the new minimum other constants get "minus" values so that they also have to be eliminated. An illustration is given in Fig. 3b, which may represent a section from $(N + 1)$-dimensional space. The procedure in MIKO will then be repeated until all remaining constants are positive or zero.

The position for the minimum in the section with $\mathbf{k}_- = 0$, and the standard variations for the constants $\mathbf{k}_+$, will come out correctly from (38), provided the approximation of a second-degree surface is valid. We have several times compared the position for a minimum as calculated from (38), thus using U values for points outside of the section with $\mathbf{k} = 0$, with the values obtained directly from U values for points in this section. The agreement has been surprisingly good, and the differences have been too small to be of any practical importance; this indicates that in "normal" cases the second-degree approximation is a useful one.

It should be noticed that the axis directions of the vector $(\mathbf{SH})_{++}\mathbf{w}_+$, which can be considered as a projection of $(\mathbf{SH})\mathbf{v}_+$, in general do not correspond to the main axes of the reduced pit. As a matter of fact, the transformation $(\mathbf{SH})_{++}$ may make the reduced pit more skew than before, and so the quantity $h_i\sigma(w)_i$ comes out larger than $\sigma(k_i)$ as often as not. If one does not want to be confused by this result, or to have to explain it to everyone who sees it, one can just omit having it printed; it is of no use in the calculation.

Details of the new program are given in part IV.[3] Applications of the new LETAGROP, with VRID and MIKO, to various chemical problems will be shown in a number of future publications from this department.

THE SPECIES SELECTOR

In equilibrium analysis — and especially in studies of polynuclear complexes — the first question to ask is, which complexes $\mathbf{X}_i$ exist in appreciable amounts in the system. We may imagine that we have extensive data of the most common type, $Z(\log a, B)$, to be explained by a set of complexes A_pB_q, each

with a triplet (p,q,β_{pq}) (see for instance, part I,[1] p. 160); the following discussion will be valid, however, also for other types of data.

Suppose that we have tried to explain our data by a certain set of complexes X_i; $X_1 ... X_N$, each with its equilibrium formation constant k_i, and that we have found a minimum for U at M_1, where all the equilibrium constants, $k_1 - k_N$ are positive. In Fig. 4, the space ($k_1 ... k_N$) is represented by the vertical line. Now we add a new complex X′, with the equilibrium constant k', and let the computer calculate the minimum point M′ and the standard deviations for the combination ($k_1 ... k_N$, k').

Fig. 4 indicates schematically a few possible positions of M′ and the surrounding elliptical D boundary. In Fig. 4a, a "better" minimum is found at M′, with a positive k', so there are good reasons not to forget X′ in succeeding calculations. In Fig. 4b, a "better" minimum is found a M′, to be sure, but it requires a negative k', so that X′ will be "thrown out" by MIKO.

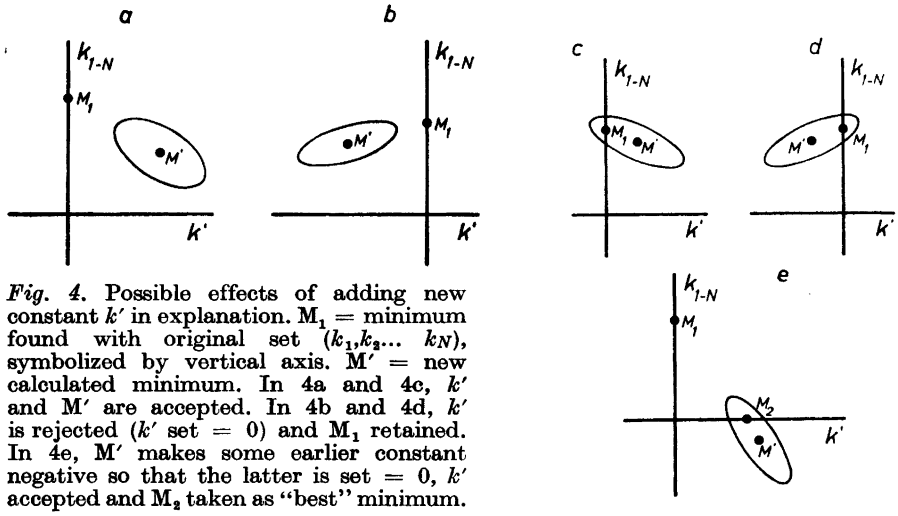

Fig. 4. Possible effects of adding new constant k' in explanation. M_1 = minimum found with original set (k_1,k_2... k_N), symbolized by vertical axis. M′ = new calculated minimum. In 4a and 4c, k' and M′ are accepted. In 4b and 4d, k' is rejected (k' set = 0) and M_1 retained. In 4e, M′ makes some earlier constant negative so that the latter is set = 0, k' accepted and M_2 taken as "best" minimum.

These two cases are clear-cut, but there are borderline cases such as 4c and 4d, where $\sigma(k')$ is less than $|k'|$ at M′. As the program is now written, X′ will be thrown out in case 4d but retained in case 4c. The distinction is really not very sharp. Adding the complex X′ to the "explanation" would give a slight improvement in U in case 4c, and no improvement in case 4d. However, from such data one would probably be content, in either case, to state a maximum value for k' (for instance $k' + 3\sigma(k')$, see discussion by Dunsmore, Hietanen and Sillén,[11] p. 2648), and to state that the data are insufficient either to prove or to disprove the existence of X′. A statistician might feel inclined to give a confidence level; we fould feel inclined to look into the systematic errors, and to search for better data, or evidence of other kinds.

It often happens that the addition of a new complex throws out an earlier one, as indicated in Fig. 4e. In the calculated minimum M′, the newcomer

has a positive k' whereas some old complex gets a negative k and hence is set equal to 0. The new "best" set is represented by M_2 in Fig. 4e.

Now suppose that our first treatment of the data — with the MESAK method,[14] or other graphical methods — has provided us with, say 8 or 10 "possible" formulas for the complexes present. A promising strategy is the following. We start with, say, two or three that we suspect to be the main species and let LETAGROP calculate the "best" values for their equilibrium constants. Then the other "possible" complexes are added, one after another, and the "best" constants calculated for each combination. Some of the newcomers will be thrown out immediately, some will stay, and some "old" complexes may be thrown out in the process.

After trying all "possible" complexes we have a set of complexes and equilibrium constants which is in general different from the starting set. Then there is a chance that some of those complexes that were thrown out at an early stage would have been retained if combined with some complex that has been "caught" later on in the process. So, the whole set of rejected complexes should be passed through once more, and this should be repeated until no additional complex is "caught" during a cycle. The policy to be followed in borderline cases such as 4c and 4d will be a matter of judgment and experience.

With the development of faster and faster computers, it may become practical one day to pass through the computer all conceivable formulas from A_1B_1, A_2B_0 and A_0B_2, to, say $A_{24}B_{24}$, rather than try to limit the search to the most probable formulas. However, it seems that one must always start with a small group of complexes — one, two or three — that give something that looks like a pit in U, even if it need not be very deep. If the starting complexes, and constants, give calculated values for y (which may for instance be Z) that have no resemblance to the experimental ones, then U will be of the order of $n\bar{y}^2$ and we are on a high plateau in $U(\mathbf{k})$ where the second-degree approximation is of no help. To find the starting pit, preliminary graphical methods at present seem the best way.

SYSTEMATIC ERRORS

One of the greatest advantages of well-designed graphical methods as compared to purely numerical methods of treating data is that systematic errors can be made to stand out very clearly (Ref.[15], p. 191, 196). If experimental data are fed thoughtlessly into a computer, there is a risk that systematic errors may be overlooked, and erroneous conclusions may be drawn. This risk exists even if the first preliminary set of species and equilibrium constants are as usual derived by graphical methods, and the final check of experimental against calculated values is also made in a graph. The risk is especially great with those (*e.g.* young workers) who have not had extensive experience with graphical methods at their best.

To minimize this risk it seems necessary, as a matter of routine, to *treat also the systematic errors as unknown constants to be determined* (Part I[1], p. 171). If the systematic errors are "turned loose" in the same way as other constants,

some minor complex may be thrown out. This would mean that the data can be explained as well or better if that complex is left out and a certain systematic error assumed. If the error seems reasonable, there is no reason to maintain stubbornly the existence of that complex. If the error is larger than expected, one must try to find out whether the error or the complex is the more likely explanation.

Systematic errors, such as unavoidable small analytical errors, and errors in the emf constants E_0, are usually different in different groups of experimental data. This is one reason for making a distinction between common constants, valid for all data (such as equilibrium constants) and group constants (such as E_0), and to give the program a choice to vary either. This device was first applied to Sylvia Gobom's emf data on acetate-Pb^{2+} complexes.[10]

Acknowledgements. I wish to thank my friends at the department of inorganic chemistry, and especially Dr. Nils Ingri, for a very pleasant cooperation and many valuable discussions. I also wish to thank Dr. Germund Dahlquist, Professor of applied mathematics at KTH, for reading the manuscript, and for helpful comments. Dr. Roy Whiteker was kind enough to correct the English.

This work has been financially supported by *Statens Tekniska Forskningsråd*. The National Swedish Office for Administrative rationalisation and economy (earlier the Swedish board for computing machinery) has kindly provided free time at the computers Besk, Facit, Ferranti Mercury, Univac 1107, and IBM 7090.

REFERENCES

1. Part I. Sillén, L. G. *Acta Chem. Scand.* **16** (1962) 159.
2. Part II. Ingri, N. and Sillén, L. G. *Acta Chem. Scand.* **16** (1962) 173.
3. Part IV. Ingri, N. and Sillén, L. G. *Arkiv Kemi* (1964). *In print.*
4. Ingri, N. *Acta Chem. Scand.* **16** (1962) 439; **17** (1963) 573, 581, 597.
5. Ahlberg, I. *Acta Chem. Scand.* **16** (1962) 887.
6. Dyrssen, D. and Lumme, P. *Acta Chem. Scand.* **16** (1962) 1785.
7. Biedermann, G. and Ciavatta, L. *Acta Chem. Scand.* **16** (1962) 2221.
8. Grenthe, I. and Tobiasson, I. *Acta Chem. Scand.* **17** (1963) 2101.
9. Tobias, R. S. and Yasuda, M. *Inorg. Chem.* **2** (1963) 1307.
10. Gobom, S. *Acta Chem. Scand.* **17** (1963) 2181.
11. Dunsmore, H. S., Hietanen, S. and Sillén, L. G. *Acta Chem. Scand.* **17** (1963) 2644.
12. Dunsmore, H. S. and Sillén, L. G. *Acta Chem. Scand.* **17** (1963) 2657.
13. Hietanen, S., Row, B. R. L. and Sillén, L. G. *Acta Chem. Scand.* **17** (1963) 2735.
14. Sillén, L. G. *Acta Chem. Scand.* **15** (1961) 1981.
15. Sillén, L. G. *Acta Chem. Scand.* **10** (1956) 186.

Received March 16, 1964.