Description of the OPIUM Computer Program

The main feature of the OPIUM program[1] (besides extensive analysis of both the estimates and the residuals) is a strict separation of the dependence of a response variable y (electromotive force of a cell in potentiometry, absorbance in UV/VIS spectrophotometry, chemical shift in "fast exchange" NMR spectroscopy, integral intensity of a peak in "slow exchange" NMR spectroscopy, reaction heat in enthalpimetry, difference of resistances in VPO, effective mobility in capillary zone electrophoresis of a labile system, conductivity in conductimetry etc.) on the calibration and the chemical parameters $y = f(\boldsymbol{\vartheta}, \mathbf{c}, \mathbf{n}, \mathrm{V}, \mathbf{c}', \mathbf{q}, \boldsymbol{\beta}, x)$ into two independent parts. These are a *calibration function* (the dependence of the response variable y on the calibration parameters $\boldsymbol{\vartheta}$ and the free concentrations $\boldsymbol{c}$ of all components) $y = \phi(\boldsymbol{\vartheta}, \boldsymbol{c})$ and the system of nonlinear equations (mass balances)

$$c_k + \sum_{i=1}^{m} q_{ik}\beta_i \prod_{j=1}^{n} c_j^{q_{ij}} = \frac{\mathrm{c}_k\mathrm{V} + \mathrm{c}'_k x}{\mathrm{V} + x} = \frac{\mathrm{n}_k + \mathrm{c}'_k x}{\mathrm{V} + x} \quad , \qquad k = 1, \ldots, n \,,$$

n is the number of reactants, m is the number of complexes, c_j is the free concentration of the j-th reactant, q_{ij} is the element of the matrix of stoichiometric coefficients $\mathbf{q} \sim (m, n)$, β_i is the stability constant of the i-th complex, c_j (n_j) is the analytical concentration (mass amount) of the j-th reactant in the solution, V is the volume of the solution, c'_j is the analytical concentration of the j-th reactant in the burette and x is the added volume (only if a titration is applied). This separation (together with a correct definition of the variance-covariance matrix $\mathbf{V}$ of the random vector of the dependent variable $\mathbf{y}$) is a key to the treatment of many different types of data in any combination.

Formally, any subset of the parameters $\boldsymbol{\vartheta}, \mathbf{c}, \mathbf{n}, \mathrm{V}, \mathbf{c}', \mathbf{q}, \boldsymbol{\beta}$ can be "marked" unknown and calculated. Moreover, simple boundary conditions (linear constraints) for selected unknown parameters like for instance $a_1 = a_2 + a_3$¶ or $\mathrm{c}_1 = 0.0736 + 4\mathrm{c}_2 \;\wedge\; \mathrm{c}_3 = \mathrm{c}_2$ can be employed to suppress the correlation of estimates of the unknown parameters. The OPIUM program also analyzes data containing sections with *zero* free concentrations of some reactants (necessary, for example, when calculating dissociation constants of a weak acid and stability constants of complexes of the acid with several metals together, i.e. from both types of titrations).

The OPIUM program is robust enough to usually find a sole global minimum or at least one of local minima even from roughly approximate initial estimates of the unknown stability constants† (i.e. without any human assistance). This feature allows us to test many thousands of chemical model hypotheses by the "trial and error" method in several hours§ (the program is written in the standard ANSI C language and can be run on virtually any computer under any operating system). The best model(s) can be, then, verified using the built-in ESI (estimation of stoichiometric indices) approach[2], i.e. a joint calculation of all the unknown stability constants

¶ a_1, a_2, a_3 could be e.g. molar absorption coefficients

† If initial estimates of the unknown stability constants are not given, the OPIUM program generates them according to the simple formula $\log\beta_i = A + B\sum_{j=1}^{n} q_{ij}$, A and B are empirical constants. If all the other chemical parameters $\mathbf{c}$, $\mathbf{n}$, V, $\mathbf{c}'$, $\mathbf{q}$ are known, there is frequently just one local minimum on the hypersurface of the sum of squares of residuals and the method used for optimization of the unknown parameters usually ensures rapid convergence to the only solution. If they are **not**, some other local minima often exist and better initial estimates of $\boldsymbol{\beta}$ might be required to find the "correct" one. Fortunately, the initial estimates do not need to be exceedingly precise. The program can mostly get along any lower initial estimates of the unknown stability constants (owing to the problems of underflow we do not recommend to set the initial estimates of β_i close to the minimum positive real number representable by the computer being used, i.e. estimates of $\log\beta_i$ lesser than about -200 on computers using standard 8-byte real numbers). Initial estimates higher than the optimum values by several tens of orders of magnitude raise more serious problems. However, the program can usually find the correct solution even from those.

§ while being somewhere else and/or doing something more exciting and selected elements $q_{ij} \in \mathsf{R}$ of the matrix of stoichiometric coefficients $\mathbf{q}$. If the chemical model is sufficiently simple and the data are sufficiently precise, it is possible to attempt a direct computation of all the unknown stability constants and all the unknown stoichiometric coefficients without prior testing of hundreds or even thousands of hypotheses.

In the whole program, the first derivatives are calculated analytically, while the second derivatives (required in calculation of the maximum parametric and internal curvature of the model and Box bias of the estimate and residuals) are calculated numerically.

The percentage points of the selected distributions are calculated analytically directly from their definitions (where necessary, using the value of the *error* function $\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-p^2} dp$). The corresponding critical values are calculated by the Newton method.

Chemical model (mass balances)

To be a little more accurate, we have to define the number of reactants with unknown free concentration $r \le n$ (a lot of spectroscopic data are collected using buffered solutions, i.e. the free concentration instead of the analytical concentration of H^+ is known). Then the mass balances are¶

$$f_k = c_k + \sum_{i=1}^{m} q_{ik} c_{n+i} - \mathrm{C}_k = 0 \quad , \quad k = 1, \ldots, r \ ,$$

$$c_{n+i} = \beta_i \prod_{j=1}^{n} c_j^{q_{ij}} = \beta_i e^{\sum_{j=1}^{n} q_{ij} \ln c_j} = \beta_i 10^{\sum_{j=1}^{n} q_{ij} \log c_j} = 10^{\log \beta_i + \sum_{j=1}^{n} q_{ij} \log c_j} \quad , \quad i = 1, \ldots, m \ .$$

Partial derivatives of the components of the function $\mathbf{f}$ with respect to the logarithms of the unknown free concentrations of reactants (elements of the symmetric Jacobi matrix) are

$$\frac{\partial f_k}{\partial \ln c_k} = 1 + \sum_{i=1}^{m} q_{ik}^2 c_{n+i} \quad , \quad k = 1, \ldots, r \ ,$$

$$\frac{\partial f_k}{\partial \ln c_l} = \frac{\partial f_l}{\partial \ln c_k} = \sum_{i=1}^{m} q_{ik} q_{il} c_{n+i} \quad , \quad k \ne l = 1, \ldots, r \ .$$

The OPIUM program combines several methods of calculation of the unknown free concentrations of reactants. Besides the Newton-Raphson method with the restricted† shift vector $\boldsymbol{\delta} = -\mathbf{J}^{-1}\mathbf{f}$ (using the Cholesky decomposition of the Jacobi matrix) and the Marquardt non-linear least squares method with the restricted shift vector $\boldsymbol{\delta} = -(\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda \mathbf{1})^{-1} \mathbf{J}^{\mathrm{T}} \mathbf{f}$ (using the decomposition of the Jacobi matrix to its eigenvalues and eigenvectors)§ a robust (but slow) non-derivative method has been adopted[3,4] (and slightly modified) with the restricted shift

$$\mathrm{C}_k \ge 0 \ : \ \delta_k = (p_k + \varrho) \left[\ln (\mathrm{C}_k - B_k) - \ln A_k \right] \ ,$$

$$\mathrm{C}_k \le 0 \ : \ \delta_k = (p_k + \varrho) \left(\ln |B_k| - \ln |\mathrm{C}_k - A_k| \right) \ ,$$

$$A_k = c_k + \sum_{\substack{i=1 \\ q_{ik} > 0}}^{m} q_{ik} c_{n+i} \quad , \quad B_k = \sum_{\substack{i=1 \\ q_{ik} < 0}}^{m} q_{ik} c_{n+i} \quad , \quad A_k + B_k = \mathrm{C}_k \quad , \quad k \in \langle 1, r \rangle \ ,$$

$$p_k + \varrho \in \left\{ 3, 2, 1, \tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{4}, \tfrac{1}{5}, \tfrac{1}{6}, \ldots \right\} \ .$$

¶ let the $n+m$ components are ordered such that the first r of them are reactants with unknown free (known analytical) concentrations, the next $n-r$ are reactants with known free (unknown analytical) concentrations and the last m are complexes

† the elements of the shift vector $\boldsymbol{\delta}$ must lie in the interval $\langle -1, 1 \rangle$

§ the λ modifier of step length and direction is initialized with the value $0.02 \operatorname{tr}(\mathbf{J}^{\mathrm{T}}\mathbf{J})/r$

In each iteration, only one of the free concentrations is corrected (mostly the one with the highest absolute value of the shift). While ϱ changes adaptively in the course of the computation, p_k is a characteristic constant of the k-th reactant. We choose it as the reciprocal value of one half of the weighted average of one (representing the reactant) and m absolute values of the elements of the k-th column of the matrix of stoichiometric coefficients $\mathbf{q}$ with the weights

$$(1+\omega)^{\log c_k} \quad , \quad (1+\omega)^{\log|q_{ik}|+\log\beta_i+q_{ik}\log c_k} \quad , \quad i=1,\ldots,m \ ,$$

ω is a conveniently selected empirical constant $\omega \sim 0.001-0.01$. Using mathematical symbols,

$$\frac{2}{p_k}=\frac{(1+\omega)^{\log c_k}+\sum_{i=1}^{m}|q_{ik}|(1+\omega)^{\log|q_{ik}|+\log\beta_i+q_{ik}\log c_k}}{(1+\omega)^{\log c_k}+\sum_{i=1}^{m}(1+\omega)^{\log|q_{ik}|+\log\beta_i+q_{ik}\log c_k}}$$

(the free concentration of the k-th reactant is replaced with its initial estimate).

Initial estimates of the unknown free concentrations of reactants are derived (at the first point of each titration section) using the known/fixed analytical concentrations of the reactants and the known/fixed ionic product of water (if it is defined). The logarithms of the unknown free concentrations are calculated with an accuracy to about one significant digit by the non--derivative method similar to the algorithms used in the programs COMICS[3] and COMPLEX[4]. To ensure a smooth convergence of the method of optimization of the unknown parameters, we need to calculate the logarithms of the free concentrations with an accuracy to approximately 13 or 14 significant digits (on computers using standard 8-byte real numbers). The Newton-Raphson method usually takes care of this providing (close to the solution) a quadratic convergence rate. If it fails, the non-derivative method is applied again. If it is possible, we increase the accuracy of this computation as long as the ratio of the biggest and the smallest eigenvalue of the Jacobi matrix is greater than 10^6. However, only rarely the logarithms of the unknown free concentrations are calculated with an accuracy to more than 2 significant digits. From these estimates the Marquardt method usually finds the solution. If it does not, everything is repeated (with greater demands on the accuracy of the computation by the non--derivative method and, if need be, more than once). This algorithm seems to be successful under all reasonable circumstances and mostly also very fast.

Thanks to a deliberate selection of the natural number k and of the rational number $p_k+\varrho$ the non-derivative method is virtually always convergent and especially in the initial phase of the computation also relatively fast. The Newton-Raphson method ensures a quadratic convergence rate close to the solution. The Marquardt method is of use when the Jacobi matrix is almost singular and cannot safely be inverted even near the solution of the mass-balance equations.

The described algorithm obviously best resembles the algorithms used in the programs CHEMEQUIL-2[5] and ES4EC[6,7].

At the second point of each titration section, the values computed at the first point are taken for initial estimates of the unknown free concentrations of reactants. The restricted linear extrapolation is applied at all the other points.

Starting with the second iteration, the values computed in the previous iteration are taken for initial estimates of the unknown free concentrations of reactants. If some of the unknown stability constants become negative, the simple non-derivative method cannot be applied even if both the derivative methods fail. Then it is necessary to multiply the negative stability constants with -1 and perform a restart of the method of optimization of the unknown parameters.

Partial derivatives of the function $y=f(\boldsymbol{\vartheta},\mathbf{c},\mathbf{n},\mathrm{V},\mathbf{c}',\mathbf{q},\boldsymbol{\beta},x)$ with respect to the unknown chemical parameters $\mathbf{p}\equiv\mathbf{c},\mathbf{n},\mathrm{V},\mathbf{c}',\mathbf{q},\boldsymbol{\beta}$ are‡

$$\frac{\partial y}{\partial p}=\sum_{i=1}^{n+m}\frac{\partial y}{\partial c_i}\frac{\partial c_i}{\partial p} \ .$$

‡ excluding two exceptions (see the description of calibration functions of types 1 and -5)

Derivatives $\frac{\partial c_i}{\partial p}$, $i=1,\ldots,r$, are the solution of the system of linear equations

$$\frac{\partial c_k}{\partial p}+\sum_{l=1}^{n}\frac{\partial c_l}{\partial p}\frac{1}{c_l}\sum_{\substack{i=1\\ q_{il}\neq 0}}^{m} q_{ik}q_{il}c_{n+i}=A_k \quad , \quad k=1,\ldots,r \ ,$$

for $p\equiv \mathrm{c}_v, \mathrm{n}_v, \mathrm{V}, \mathrm{c}'_v$

$$A_k=\frac{\partial \mathrm{C}_k}{\partial p}\ ¶,$$

for $p\equiv\beta_u$

$$A_k=-q_{uk}\prod_{j=1}^{n}c_j^{q_{uj}} \ ,$$

for $p\equiv q_{uv}\wedge k\neq v$

$$A_k=-q_{uk}\ln c_v\, c_{n+u} \ ,$$

for $p\equiv q_{uk}$

$$A_k=-(1+q_{uk}\ln c_k)c_{n+u} \ .$$

Derivatives $\frac{\partial c_i}{\partial p}$, $i=r+1,\ldots,n$, are zero†, derivatives $\frac{\partial c_i}{\partial p}$, $i=n+1,\ldots,n+m$ (i.e. $\frac{\partial c_{n+i}}{\partial p}$, $i=1,\ldots,m$), are

$$\frac{\partial c_{n+i}}{\partial p}=B_i+\sum_{\substack{k=1\\ q_{ik}\neq 0}}^{n}\frac{\partial c_k}{\partial p}\frac{1}{c_k}q_{ik}c_{n+i} \quad , \quad i=1,\ldots,m \ ,$$

for $p\equiv\beta_i$

$$B_i=\prod_{j=1}^{n}c_j^{q_{ij}} \ ,$$

for $p\equiv q_{ik}$

$$B_i=\ln c_k\, c_{n+i} \ ,$$

$B_i=0$ for any other p.

Measurements (calibration functions)

Overview of the supported calibration functions follows§.

Type 1

$$y=E_0+S\log c_1+j_1c_1+j_2\frac{\beta_1}{c_1}$$

¶ it is clear that

$$\mathrm{C}_k=\frac{\mathrm{c}_k\mathrm{V}+\mathrm{c}'_k x}{\mathrm{V}+x}=\frac{\mathrm{n}_k+\mathrm{c}'_k x}{\mathrm{V}+x} \ ,$$

$$\frac{\partial \mathrm{C}_k}{\partial \mathrm{c}_v}=\frac{\partial \mathrm{C}_k}{\partial \mathrm{n}_v}=\frac{\partial \mathrm{C}_k}{\partial \mathrm{c}'_v}=0 \quad , \quad v\neq k=1,\ldots,r \ ,$$

$$\frac{\partial \mathrm{C}_k}{\partial \mathrm{c}_k}=\frac{\mathrm{V}}{\mathrm{V}+x} \quad , \quad \frac{\partial \mathrm{C}_k}{\partial \mathrm{n}_k}=\frac{1}{\mathrm{V}+x} \quad , \quad \frac{\partial \mathrm{C}_k}{\partial \mathrm{c}'_k}=\frac{x}{\mathrm{V}+x} \ ,$$

$$\left(\frac{\partial \mathrm{C}_k}{\partial \mathrm{V}}\right)_{d\mathrm{c}_k=0}=\frac{(\mathrm{c}_k-\mathrm{c}'_k)x}{(\mathrm{V}+x)^2} \quad , \quad \left(\frac{\partial \mathrm{C}_k}{\partial \mathrm{V}}\right)_{d\mathrm{n}_k=d\mathrm{c}_k\mathrm{V}+\mathrm{c}_k d\mathrm{V}=0}=-\frac{\mathrm{n}_k+\mathrm{c}'_k x}{(\mathrm{V}+x)^2}$$

† This is **not** true, if the "known" free concentration of the i-th reactant is considered an unknown parameter. The OPIUM program provides this option, if an alternative definition of a solution is used (solutions are **not** prepared by a titration).

§ including some "non-trivial" derivatives

$\boldsymbol{\vartheta} \equiv E_0,\ S,\ j_1,\ j_2$

$$\frac{\partial y}{\partial c_1} = \frac{S}{\ln 10 \cdot c_1} + j_1 - j_2 \frac{\beta_1}{c_1^2} \quad , \quad \frac{\partial y}{\partial \beta_1} = \frac{\partial y}{\partial c_1} \frac{\partial c_1}{\partial \beta_1} + \frac{j_2}{c_1}$$

Note: Function of type 1 describes the dependence of the difference y of potentials of a "hydrogen" (glass) electrode and a reference electrode in a cell with non-negligible liquid-junction potential on the free concentration c_1 ($\beta_1 \equiv K$ is the ionic product of water). Thus, H^+ must be the 1st reactant (1st column of the matrix $\mathbf{q}$), OH^- must be the 1st complex (1st row of the matrix $\mathbf{q}$).

Type 2

$$y = E_0 + S \log c_1$$

$\boldsymbol{\vartheta} \equiv E_0,\ S$

Type 3

$$y = E_0 + S \log c_k + j_1 c_1 + j_2 \frac{\beta_1}{c_1}$$

$\boldsymbol{\vartheta} \equiv E_0,\ S,\ j_1,\ j_2;\ k \in \langle 1, n \rangle$

Type 4

$$y = E_0 + S \log c_k$$

$\boldsymbol{\vartheta} \equiv E_0,\ S;\ k \in \langle 1, n \rangle$

Type 5

$$y = E_0 + S \log \frac{c_k}{c_l} + j_1 c_1 + j_2 \frac{\beta_1}{c_1}$$

$\boldsymbol{\vartheta} \equiv E_0,\ S,\ j_1,\ j_2;\ k \neq l \in \langle 1, n \rangle$

Type 6

$$y = E_0 + S \log \frac{c_k}{c_l}$$

$\boldsymbol{\vartheta} \equiv E_0,\ S;\ k \neq l \in \langle 1, n \rangle$

$$\frac{\partial y}{\partial c_k} = \frac{S}{\ln 10 \cdot c_k} \quad , \quad \frac{\partial y}{\partial c_l} = -\frac{S}{\ln 10 \cdot c_l}$$

Note: Functions of types 5 and 6 describe the dependence of the difference y of potentials of a redox (platinum) electrode and a reference electrode on the ratio of free concentrations c_k, c_l of two reactants (i.e. the simplest case of a redox system, e.g. Fe^{3+}/Fe^{2+}, Sn^{4+}/Sn^{2+}, Ce^{4+}/Ce^{3+} etc.).

Type 7

$$y = \sum_{i=1}^{n+m} a_i c_i = \mathbf{a}^{\mathrm{T}} \boldsymbol{c}$$

$\boldsymbol{\vartheta} \equiv a_i,\ i = 1, \ldots, n+m$

$$\frac{\partial y}{\partial a_i} = c_i \quad , \quad \frac{\partial y}{\partial c_i} = a_i \quad , \quad i = 1, \ldots, n+m$$

Type 8

$$y = d \sum_{i=1}^{n+m} a_i c_i = d\, \mathbf{a}^{\mathrm{T}} \boldsymbol{c}$$

$\boldsymbol{\vartheta} \equiv a_i,\ i = 1, \ldots, n+m;\ d \in \mathsf{R}$

Note: Functions of types 7 and 8 describe the dependence of an extensive *linearly additive* quantity y (absorbance in some "slow exchange" spectrometry (in the case of inseparable peaks), usually UV/VIS spectrophotometry, reaction heat in enthalpimetry, conductivity in conductimetry etc.) on the free concentrations $\boldsymbol{c}$ of all components.

$\boxed{\textbf{Type } 9}$

$$y = \frac{c_i}{c_j}$$

$i \neq j \in \langle 1, n+m \rangle$

$$\frac{\partial y}{\partial c_i} = \frac{1}{c_j} \quad , \quad \frac{\partial y}{\partial c_j} = -\frac{c_i}{c_j^2}$$

"Subtype" $9ln$

$$y = -\ln \frac{c_i}{c_j}$$

$i \neq j \in \langle 1, n+m \rangle$

$$\frac{\partial y}{\partial c_i} = -\frac{1}{c_i} \quad , \quad \frac{\partial y}{\partial c_j} = \frac{1}{c_j}$$

$\boxed{\textbf{Type } -1}$

$$y = -\log c_1$$

$\boxed{\textbf{Type } -2}$

$$y = -\log c_k$$

$k \in \langle 1, n \rangle$

$\boxed{\textbf{Type } -3}$

$$y = \alpha c_i$$

$\vartheta \equiv \alpha;\ i \in \langle 1, n+m \rangle$

$\boxed{\textbf{Type } -4}$

$$y = d\alpha c_i$$

$\vartheta \equiv \alpha;\ d \in \mathsf{R},\ i \in \langle 1, n+m \rangle$

$\boxed{\textbf{Type } -5}$

$$y = \sum_{i=1}^{n+m} a_i p_i c_i / \mathcal{C} = \mathbf{a}^{\mathrm{T}} \mathbf{P} \boldsymbol{c} / \mathcal{C} \ ¶†$$

$$\mathbf{P} = \mathrm{diag}(p_1, \ldots, p_{n+m})$$

$\boldsymbol{\vartheta} \equiv a_i,\ i = 1, \ldots, n+m$

$$\frac{\partial y}{\partial a_i} = p_i c_i / \mathcal{C} \quad , \quad \frac{\partial y}{\partial c_i} = p_i \left(a_i \mathcal{C} - \sum_{j=1}^{n+m} a_j p_j c_j \right) / \mathcal{C}^2 = p_i \left(a_i - y \right) / \mathcal{C}$$

$$i = 1, \ldots, n+m$$

¶ $p_i \in \mathsf{N} + \{0\}$, $i = 1, \ldots, n$, $p_{n+i} = \sum_{j=1}^{n} p_j q_{ij}$, $i = 1, \ldots, m$, $\mathcal{C} = \sum_{i=1}^{n+m} p_i c_i$, $\frac{\partial \mathcal{C}}{\partial q_{uv}} = p_v c_{n+u}$

† A little more complicated definition (using almost general diagonal matrix $\mathbf{P}$) enables correct description of the dependent variable no matter what linear transformation of the matrix $\mathbf{q}$ is actually in use. Moreover, dynamic definition of the analytical concentration $\mathcal{C}$ of the "active" component (**not** exclusively reactant or complex) is necessary for the calculation of unknown chemical parameters $\mathbf{c}$, $\mathbf{n}$, V, $\mathbf{c}$', $\mathbf{q}$ (the value of $\mathcal{C}$ changes).

$$\frac{\partial y}{\partial q_{uv}} = \sum_{i=1}^{n+m} \frac{\partial y}{\partial c_i}\frac{\partial c_i}{\partial q_{uv}} + p_v c_{n+u}\left(a_{n+u} - y\right)/\mathcal{C} \quad , \quad u = 1,\ldots,m \quad , \quad v = 1,\ldots,n$$

<u>Type −5</u>

$$y = a_i p_i c_i / \mathcal{C}\ \P$$

$$\frac{\partial y}{\partial a_i} = p_i c_i/\mathcal{C} \quad , \quad \frac{\partial y}{\partial c_i} = a_i p_i(\mathcal{C} - p_i c_i)/\mathcal{C}^2 = p_i(a_i - y)/\mathcal{C}$$

$$\frac{\partial y}{\partial c_j} = -a_i p_j p_i c_i/\mathcal{C}^2 = -p_j y/\mathcal{C} \quad , \quad j = 1,\ldots,n+m \ \wedge\ j \neq i$$

$$\frac{\partial y}{\partial q_{uv}} = \sum_{j=1}^{n+m} \frac{\partial y}{\partial c_j}\frac{\partial c_j}{\partial q_{uv}} + A_{uv}$$

$$A_{uv} = p_v c_i\left(a_i - y\right)/\mathcal{C} \quad , \quad n+u = i \quad , \quad v = 1,\ldots,n$$

$$A_{uv} = -p_v c_{n+u} y/\mathcal{C} \quad , \quad u = 1,\ldots,m \ \wedge\ n+u \neq i \quad , \quad v = 1,\ldots,n$$

Note: Function of type −5 describes the dependence of an <u>intensive</u> *linearly additive* quantity y (chemical shift or relaxation rate, i.e. reciprocal relaxation time, in "fast exchange" NMR spectroscopy, effective mobility in capillary zone electrophoresis of a labile system etc.) on the free concentrations $\boldsymbol{c}$ of all components.

<u>"Subtype" $-5ln$</u>

$$y = -\ln a_i - \ln p_i - \ln c_i + \ln \mathcal{C}\ \P$$

$$\frac{\partial y}{\partial \ln a_i} = -1 \quad , \quad \frac{\partial y}{\partial c_i} = p_i\left(1 - \frac{\mathcal{C}}{p_i c_i}\right)/\mathcal{C} = p_i\left(1 - e^{y+\ln a_i}\right)/\mathcal{C}$$

$$\frac{\partial y}{\partial c_j} = p_j/\mathcal{C} \quad , \quad j = 1,\ldots,n+m \ \wedge\ j \neq i$$

$$\frac{\partial y}{\partial q_{uv}} = \sum_{j=1}^{n+m} \frac{\partial y}{\partial c_j}\frac{\partial c_j}{\partial q_{uv}} + A_{uv}$$

$$A_{uv} = p_v c_i\left(1 - \frac{\mathcal{C}}{p_i c_i}\right)/\mathcal{C} = p_v c_i\left(1 - e^{y+\ln a_i}\right)/\mathcal{C} \quad , \quad n+u = i \quad , \quad v = 1,\ldots,n$$

$$A_{uv} = p_v c_{n+u}/\mathcal{C} \quad , \quad u = 1,\ldots,m \ \wedge\ n+u \neq i \quad , \quad v = 1,\ldots,n$$

Type −6

$$y = p_i c_i/\mathcal{C}$$

$i \in \langle 1, n+m\rangle$, $p_i \neq 0$

$$\frac{\partial y}{\partial c_i} = p_i(\mathcal{C} - p_i c_i)/\mathcal{C}^2 = p_i(1 - y)/\mathcal{C}$$

$$\frac{\partial y}{\partial c_j} = -p_j p_i c_i/\mathcal{C}^2 = -p_j y/\mathcal{C} \quad , \quad j = 1,\ldots,n+m \ \wedge\ j \neq i$$

$$\frac{\partial y}{\partial q_{uv}} = \sum_{j=1}^{n+m} \frac{\partial y}{\partial c_j}\frac{\partial c_j}{\partial q_{uv}} + A_{uv}$$

$$A_{uv} = p_v c_i\left(1 - y\right)/\mathcal{C} \quad , \quad n+u = i \quad , \quad v = 1,\ldots,n$$

¶ $p_i \neq 0, \forall\ j \neq i : a_j = 0$

$$A_{uv} = -p_v c_{n+u} y / \mathcal{C} \quad , \quad u = 1, \ldots, m \ \wedge \ n+u \neq i \quad , \quad v = 1, \ldots, n$$

Note: Function of type -6 describes the dependence of an intensive quantity y (intensity of an isolated (separated) peak in some "slow exchange" spectrometry, usually NMR spectroscopy, intensity of an isolated peak in capillary zone electrophoresis of an inert system etc., normalized to the sum of intensities of all the peaks embracing all components containing the active component) on the free concentrations $\boldsymbol{c}$ of all components (but mainly on c_i).

"Subtype" $-6ln$

$$y = -\ln p_i - \ln c_i + \ln \mathcal{C}$$

$$\frac{\partial y}{\partial c_i} = p_i \left(1 - \frac{\mathcal{C}}{p_i c_i}\right) / \mathcal{C} = p_i \left(1 - e^y\right) / \mathcal{C}$$

$$\frac{\partial y}{\partial c_j} = p_j / \mathcal{C} \quad , \quad j = 1, \ldots, n+m \ \wedge \ j \neq i$$

$$\frac{\partial y}{\partial q_{uv}} = \sum_{j=1}^{n+m} \frac{\partial y}{\partial c_j} \frac{\partial c_j}{\partial q_{uv}} + A_{uv}$$

$$A_{uv} = p_v c_i \left(1 - \frac{\mathcal{C}}{p_i c_i}\right) / \mathcal{C} = p_v c_i \left(1 - e^y\right) / \mathcal{C} \quad , \quad n+u = i \quad , \quad v = 1, \ldots, n$$

$$A_{uv} = p_v c_{n+u} / \mathcal{C} \quad , \quad u = 1, \ldots, m \ \wedge \ n+u \neq i \quad , \quad v = 1, \ldots, n$$

Type -7

$$y = \alpha \sum_{i=1}^{n+m} c_i$$

$\vartheta \equiv \alpha$

$$\frac{\partial y}{\partial \alpha} = \sum_{i=1}^{n+m} c_i \quad , \quad \frac{\partial y}{\partial c_i} = \alpha \quad , \quad i = 1, \ldots, n+m$$

Note: Function of type -7 describes the dependence of an extensive *linearly additive* quantity y, usually a change of some colligative property (difference of resistances in VPO, osmotic pressure in membrane osmometry etc.) on the free concentrations $\boldsymbol{c}$ of all components.

Regression model

Let m is the number of points, i.e. the number of ordered pairs of measured quantities on abscissa and ordinate $[x, y]$ ($[x_i, y_i]$, $i = 1, \ldots, m$) and n is the number of unknown parameters. If $m = n$, the values of the unknown parameters are the unique solution of a system of equations. If $m > n$, the system of equations is over-determined. Moreover, if the values of variables x, y and *known* parameters $\mathbf{p} \equiv \boldsymbol{\vartheta}, \mathbf{c}, \mathbf{n}, \mathrm{V}, \mathbf{c}', \mathbf{q}, \boldsymbol{\beta}$ are accompanied with measuring errors, the solution of the system of equations is **not** unique.

Let the measuring errors $\varepsilon_x, \varepsilon_y, \varepsilon_{p_k}$ of the variables x, y, p_k are random, independent, have normal distributions, zero means, variances $\sigma_x^2, \sigma_y^2, \sigma_{p_k}^2$ and act additively. Then the variables x, y, p_k are random, independent, have normal distributions, unknown means and variances $\sigma_x^2, \sigma_y^2, \sigma_{p_k}^2$

$$E(\varepsilon_x) = E(\varepsilon_y) = E(\varepsilon_{p_k}) = 0 \ ,$$

$$E(\varepsilon_x \varepsilon_y) = E(\varepsilon_x) E(\varepsilon_y) = 0 \ ,$$

$$E(\varepsilon_x \varepsilon_{p_k}) = E(\varepsilon_x) E(\varepsilon_{p_k}) = 0 \ ,$$

$$E(\varepsilon_y \varepsilon_{p_k}) = E(\varepsilon_y) E(\varepsilon_{p_k}) = 0 \ ,$$

$$E(\varepsilon_{p_k} \varepsilon_{p_l}) = E(\varepsilon_{p_k}) E(\varepsilon_{p_l}) = 0 \quad , \quad k \neq l \ ,$$

$$E(\varepsilon_x\varepsilon_y) = E(\varepsilon_x\varepsilon_{p_k}) = E(\varepsilon_y\varepsilon_{p_k}) = E(\varepsilon_{p_k}\varepsilon_{p_l}) = 0 \quad , \quad k \neq l \ ,$$
$$D(\varepsilon_x) = E(\varepsilon_x^2) = D(x) = \sigma_x^2 \ ,$$
$$D(\varepsilon_y) = E(\varepsilon_y^2) = D(y) = \sigma_y^2 \ ,$$
$$D(\varepsilon_{p_k}) = E(\varepsilon_{p_k}^2) = D(p_k) = \sigma_{p_k}^2 \ ,$$
$$\varepsilon_x \sim N(0,\sigma_x^2) \quad , \quad \varepsilon_y \sim N(0,\sigma_y^2) \quad , \quad \varepsilon_{p_k} \sim N(0,\sigma_{p_k}^2) \ .$$

However, it is usually assumed that only the values of the variables x and p_k are set (measured) independently in the course of a titration. Since it still holds that $y = f(\boldsymbol{\vartheta}, \mathbf{c}, \mathbf{n}, \mathrm{V}, \mathbf{c}', \mathbf{q}, \boldsymbol{\beta}, x) \equiv f(\mathbf{p}, x)$, an error of the functional relation $f(\mathbf{p}, x)$ should be added to the "independent" measuring error ε_y of the variable y to form the true regression error ε.

From the linear Taylor approximation of the function $f(\mathbf{p}, x)$ in a close neighborhood of the means of the variables x and p_k

$$f(\mathbf{p}, x) \doteq f(E(\mathbf{p}), E(x)) + \frac{\partial y}{\partial x}\varepsilon_x + \sum_k \frac{\partial y}{\partial p_k}\varepsilon_{p_k}$$

($x = E(x) + \varepsilon_x$, $p_k = E(p_k) + \varepsilon_{p_k}$) it follows that

$$\varepsilon \doteq \frac{\partial y}{\partial x}\,\varepsilon_x + \varepsilon_y + \sum_k \frac{\partial y}{\partial p_k}\,\varepsilon_{p_k}\,\P \ ,$$

$$E(\varepsilon) \doteq 0 \ ,$$
$$D(\varepsilon) \doteq E(\varepsilon^2) \doteq \left(\frac{\partial y}{\partial x}\right)^2 \sigma_x^2 + \sigma_y^2 + \sum_k \left(\frac{\partial y}{\partial p_k}\right)^2 \sigma_{p_k}^2 \ ,$$
$$D(\boldsymbol{\varepsilon}) \doteq \mathbf{V} \ .$$

The elements of the variance-covariance matrix $\mathbf{V}$ are†

$$v_{ii} \doteq D(\varepsilon_i) \doteq E(\varepsilon_i^2) \ ,$$
$$v_{ii} = \left(\frac{\partial y}{\partial x}\right)^2_{x=x_i, y=y_i} \sigma_{x_i}^2 + \sigma_{y_i}^2 + \sum_k \left(\frac{\partial y}{\partial p_k}\right)^2_{y=y_i, p_k=p_{k\,i}} \sigma_{p_{k\,i}}^2 \ ,$$
$$v_{ij} \doteq cov(\varepsilon_i, \varepsilon_j) \doteq E(\varepsilon_i\varepsilon_j) \ ,$$
$$E(\varepsilon_{x_i}\varepsilon_{x_j}) = E(\varepsilon_{x_i})E(\varepsilon_{x_j}) = 0 \quad , \quad x_i \not\equiv x_j \ ,$$
$$E(\varepsilon_{x_i}\varepsilon_{x_j}) = \sigma_{x_i}^2 \quad , \quad x_i \equiv x_j \ ,$$
$$E(\varepsilon_{y_i}\varepsilon_{y_j}) = E(\varepsilon_{y_i})E(\varepsilon_{y_j}) = 0 \quad , \quad i \neq j \ ,$$
$$E(\varepsilon_{p_{k\,i}}\varepsilon_{p_{k\,j}}) = E(\varepsilon_{p_{k\,i}})E(\varepsilon_{p_{k\,j}}) = 0 \quad , \quad p_{k\,i} \not\equiv p_{k\,j} \ ,$$

¶ all the necessary derivatives are given, except for

$$\frac{\partial \mathrm{C}_k}{\partial x} = \frac{(\mathrm{c}'_k - \mathrm{c}_k)\mathrm{V}}{(\mathrm{V}+x)^2} = \frac{\mathrm{c}'_k\mathrm{V} - \mathrm{n}_k}{(\mathrm{V}+x)^2}$$

† indexing of variables was taken into account, i.e.

$$\varepsilon_i = \left(\frac{\partial y}{\partial x}\right)_{x=x_i, y=y_i} \varepsilon_{x_i} + \varepsilon_{y_i} + \sum_k \left(\frac{\partial y}{\partial p_k}\right)_{y=y_i, p_k=p_{k\,i}} \varepsilon_{p_{k\,i}}$$

$$E(\varepsilon_{p_{ki}}\varepsilon_{p_{kj}}) = \sigma^2_{p_{ki}} \quad , \quad p_{ki} \equiv p_{kj} \ ,$$

$\forall\, i,j,k : i \neq j \ \wedge \ x_i \not\equiv x_j \ \wedge \ p_{ki} \not\equiv p_{kj}$

$$v_{ij} = 0 \ ,$$

$\forall\, i,j,k : i \neq j \ \wedge \ x_i \equiv x_j \ \wedge \ p_{ki} \equiv p_{kj}$

$$v_{ij} = \left(\frac{\partial y}{\partial x}\right)_{x=x_i, y=y_i} \left(\frac{\partial y}{\partial x}\right)_{x=x_j, y=y_j} \sigma^2_{x_i} + \sum_k \left(\frac{\partial y}{\partial p_k}\right)_{y=y_i, p_k=p_{ki}} \left(\frac{\partial y}{\partial p_k}\right)_{y=y_j, p_k=p_{kj}} \sigma^2_{p_{ki}} \ .$$

From the quadratic Taylor approximation of the function $f(\mathbf{p}, x)$ in a close neighborhood of the means of the variables x and p_k

$$f(\mathbf{p}, x) \doteq f(E(\mathbf{p}), E(x)) + \frac{\partial y}{\partial x}\varepsilon_x + \sum_k \frac{\partial y}{\partial p_k}\varepsilon_{p_k} + \frac{1}{2}\left(\frac{\partial^2 y}{\partial x^2}\varepsilon_x^2 + \sum_k \frac{\partial^2 y}{\partial p_k^2}\varepsilon^2_{p_k}\right) +$$

$$+ \sum_k \frac{\partial^2 y}{\partial x \partial p_k}\varepsilon_x \varepsilon_{p_k} + \sum_k \sum_{l>k} \frac{\partial^2 y}{\partial p_k \partial p_l}\varepsilon_{p_k}\varepsilon_{p_l}$$

$(x = E(x) + \varepsilon_x,\ p_k = E(p_k) + \varepsilon_{p_k})$ it follows that

$$\varepsilon \doteq \frac{\partial y}{\partial x}\varepsilon_x + \varepsilon_y + \sum_k \frac{\partial y}{\partial p_k}\varepsilon_{p_k} + \frac{1}{2}\left(\frac{\partial^2 y}{\partial x^2}\varepsilon_x^2 + \sum_k \frac{\partial^2 y}{\partial p_k^2}\varepsilon^2_{p_k}\right) + \sum_k \frac{\partial^2 y}{\partial x \partial p_k}\varepsilon_x\varepsilon_{p_k} + \sum_k\sum_{l>k} \frac{\partial^2 y}{\partial p_k \partial p_l}\varepsilon_{p_k}\varepsilon_{p_l} \ ,$$

$$E(\varepsilon) \doteq \frac{1}{2}\left(\frac{\partial^2 y}{\partial x^2}\sigma_x^2 + \sum_k \frac{\partial^2 y}{\partial p_k^2}\sigma^2_{p_k}\right) \ ,$$

$$E(\boldsymbol{\varepsilon}) \doteq \boldsymbol{\mu} \ .$$

$$D(\varepsilon) = E(\varepsilon^2) - [E(\varepsilon)]^2 \doteq \left(\frac{\partial y}{\partial x}\right)^2 \sigma_x^2 + \sigma_y^2 + \sum_k \left(\frac{\partial y}{\partial p_k}\right)^2 \sigma^2_{p_k} + \frac{1}{4}\left\{\left(\frac{\partial^2 y}{\partial x^2}\right)^2 \left[E(\varepsilon_x^4) - \sigma_x^4\right] +\right.$$

$$\left. + \sum_k \left(\frac{\partial^2 y}{\partial p_k^2}\right)^2 \left[E(\varepsilon^4_{p_k}) - \sigma^4_{p_k}\right]\right\} + \sum_k \left(\frac{\partial^2 y}{\partial x \partial p_k}\right)^2 \sigma_x^2\sigma^2_{p_k} + \sum_k\sum_{l>k}\left(\frac{\partial^2 y}{\partial p_k \partial p_l}\right)^2 \sigma^2_{p_k}\sigma^2_{p_l} +$$

$$+ \frac{\partial y}{\partial x}\frac{\partial^2 y}{\partial x^2}E(\varepsilon_x^3) + \sum_k \frac{\partial y}{\partial p_k}\frac{\partial^2 y}{\partial p_k^2}E(\varepsilon^3_{p_k}) \ .$$

The last relation can substantially be simplified using the approximate formulae $E(\varepsilon_x^4) \doteq \sigma_x^4$, $E(\varepsilon^4_{p_k}) \doteq \sigma^4_{p_k}$ and $E(\varepsilon_x^3) \doteq E(\varepsilon^3_{p_k}) \doteq 0$, i.e.

$$D(\varepsilon) \doteq \left(\frac{\partial y}{\partial x}\right)^2 \sigma_x^2 + \sigma_y^2 + \sum_k \left(\frac{\partial y}{\partial p_k}\right)^2 \sigma^2_{p_k} + \sum_k \left(\frac{\partial^2 y}{\partial x \partial p_k}\right)^2 \sigma_x^2 \sigma^2_{p_k} + \sum_k\sum_{l>k}\left(\frac{\partial^2 y}{\partial p_k \partial p_l}\right)^2 \sigma^2_{p_k}\sigma^2_{p_l} \ ,$$

$$D(\boldsymbol{\varepsilon}) \doteq \mathbf{V} \ .$$

Unfortunately, both the relation for $E(\varepsilon)$ and the relation for $D(\varepsilon)$ are too complicated for most practical purposes anyway. The elements of the vector $\boldsymbol{\mu}$ are

$$\mu_i \doteq E(\varepsilon_i) \ ,$$

$$\mu_i = \frac{1}{2}\left[\left(\frac{\partial^2 y}{\partial x^2}\right)_{x=x_i, y=y_i} \sigma^2_{x_i} + \sum_k \left(\frac{\partial^2 y}{\partial p_k^2}\right)_{y=y_i, p_k=p_{ki}} \sigma^2_{p_{ki}}\right] \ .$$

If the “independent” measuring error $\boldsymbol{\varepsilon}_y$ is a predominant part of the true regression error $\boldsymbol{\varepsilon}$, the distribution of the random vector $\boldsymbol{\varepsilon}$ is close to normal and the maximum likelihood estimate of the unknown parameters is close to the solution of the over-determined set of m equations (for n unknown variables) minimizing the generalized least squares method criterion¶. If the quadratic instead of the linear Taylor approximation was applied, an estimate of the vector $\boldsymbol{\mu}$ would have to be subtracted from the random vector $\mathbf{y}$ first.

All the elements of the vector $\boldsymbol{\vartheta}$ of all the calibration functions implemented so far are "linear" and thus the OPIUM program does not require their initial estimates. The unknown linear and nonlinear parameters (variables of the vector of residuals $\mathbf{f}$) are separated[8]. The unknown nonlinear parameters are computed using the approximate Jacobi matrix[9]

$$\mathbf{J}_2 = (\mathbf{1}_m - \mathbf{A}\mathbf{A}^+)\mathbf{J}_{n-k}$$

(instead of the full-size Jacobi matrix $\mathbf{J} \equiv (\mathbf{A} \quad \mathbf{J}_{n-k})$ of the function $\mathbf{f}$ with respect to all unknown parameters†), m is the number of points, n is the number of unknown parameters, k is the number of unknown linear parameters, $\mathbf{A} \sim (m, k)$ and $\mathbf{J}_{n-k} \sim (m, n-k)$ are the "linear" and "nonlinear" (left and right) submatrices of the matrix $\mathbf{J} \sim (m, n)$§.

When selected elements of the vector $\boldsymbol{\vartheta}$ are the only unknown parameters, then a calculation is made of their unbiased estimates by the iterative linear least squares method (the variance--covariance matrix $\mathbf{V}$ is in general a function of the unknown parameters), where biased estimates are also available, obtained by the ridge regression method.

The method used for optimization of the unknown parameters is based on an algorithm similar to the MDLS (modified damped least squares) method of Meyer and Roth[10] with the partially restricted shift vector $\boldsymbol{\delta} = -\alpha(\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{f}$ (using the QR-decomposition of the Jacobi matrix‡), $\boldsymbol{\Omega}$ is a diagonal matrix with the same diagonal elements as the matrix $\mathbf{J}^{\mathrm{T}}\mathbf{J}$¶. Deliberate selection of a method of calculation of the unknown free concentrations of reactants of appropriate (i.e. **not** always the highest available) robustness considerably increases the efficiency of the method of optimization of the unknown parameters. Failure of the selected method (starting with the second iteration, the values computed in the previous iteration are taken for initial estimates of the unknown free concentrations) is evaluated as a step to a higher value on the hypersurface of sum of squares of residuals.

Moreover, several empirical tricks have been employed which significantly improve the "sub-method" of optimization of the unknown stability constants. Probably the most effective of them is the change of all the negative unknown stability constants to positive ones (simply by their multiplication with -1) as soon as the λ modifier of step length and direction exceeds its upper limit. Not only that is why the OPIUM program rarely fails. Usually it finds the

¶ In addition to a point estimate of mean of estimate of the unknown parameter(s) the OPIUM program also offers classical analysis of this estimate in a linearization approximation (variance, t- and F-interval estimate of mean, significance test, correlation matrix, partial and multiple correlation coefficients etc.) as well as jack-knife and likelihood ratio analysis (interval estimate of mean, significance test). At each point, several types of residuals are calculated along with a number of scalar and vector quantities for indication of golden points, outliers and high leverage points. The OPIUM program also tests hypotheses of homogeneity, randomness, independence and normality of the standardized residuals.

† the efficiency of the separation of the unknown linear and nonlinear parameters compared with the solution of the full-size problem depends on the calibration function(s) used, initial estimates of the unknown nonlinear parameters and fitted data, but the computed results are always exactly the same!

§ Strictly speaking, the matrix $\mathbf{J}_{n-k}$ is a function (of the unknown nonlinear parameters) composed of the "nonlinear" (right) submatrix of the matrix $\mathbf{J}$ (which is a function of both the unknown linear and the unknown nonlinear parameters) and the closed-form dependence of the unknown linear parameters on the unknown nonlinear parameters. Hence, optimum values of all the unknown linear parameters have to be computed (by the linear least squares method) in each iteration of the nonlinear least squares method (for actual values of the unknown nonlinear parameters and variance-covariance matrix $\mathbf{V}$) before the matrix $\mathbf{J}_{n-k}$ is calculated and transformed to the matrix $\mathbf{J}_2$ by its left multiplication with the square matrix $\mathbf{1}_m - \mathbf{A}\mathbf{A}^+$.

‡ taking full advantage of the contingent block structure of its "linear" submatrix

¶ the λ modifier of step length and direction is initialized with the value 0.01

correct solution even from initial estimates of the unknown stability constants lying far beyond the limit of their physical sense.

Four-level termination criterion has been adopted into the non-linear least squares method (only the $1.a$ level is applied in the case of iterative linear least squares method, "b" levels are used instead of "a" levels for testing of unknown free concentrations of reactants and stability constants computed in logarithmic scale):

1. *a*) $|\delta_i|/|p_i| \le \varepsilon_1$, $i = 1, \dots, n$,
   b) $|\delta_i| \le -\log(1 - \varepsilon_1)$, $i = 1, \dots, n$,
2. *a*) $|g_i p_i| \le \varepsilon_2$, $i = 1, \dots, n$,
   b) $|g_i| \le \varepsilon_2 \ln 10$, $i = 1, \dots, n$,
3. $|\cos\alpha_i| = \frac{1}{2}|g_i|/\sqrt{\mathbf{j}_i^{\mathrm{T}}\mathbf{j}_i\mathbf{f}^{\mathrm{T}}\mathbf{f}} \le \varepsilon_3$, $i = 1, \dots, n$,
4. the Jacobi matrix $\mathbf{J}$ has full rank, $r(\mathbf{J}) = n$,

n is the number of unknown parameters, ε_1, ε_2, ε_3 are small real numbers, $\mathbf{p}$ is the vector of unknown parameters, $\boldsymbol{\delta}$ is the vector of parameter shifts, $\mathbf{f}$ is the vector of residuals, $\mathbf{j}_i$ is the i-th column of the Jacobi matrix $\mathbf{J}$ of the function $\mathbf{f}$ (with respect to the unknown parameters), $\mathbf{g} = 2\mathbf{J}^{\mathrm{T}}\mathbf{f}$ is the gradient of the function minimized $\|\mathbf{f}\|^2 = \mathbf{f}^{\mathrm{T}}\mathbf{f}$ and α_i is the angle between $\mathbf{j}_i$ and $\mathbf{f}$. This test ensures that the OPIUM program **never** terminates the nonlinear lest squares method computation at a false minimum or a saddle point.

Principal component analysis

The dependence of the measured quantity y on the free concentrations $\boldsymbol{c}$ of components in a common spectrophotometric or "fast exchange" NMR experiment¶ can be expressed in a matrix form

$$\mathbf{A} = \boldsymbol{\varepsilon}\boldsymbol{C} \quad , \quad \mathbf{A} = \boldsymbol{\delta}\boldsymbol{\Pi}\boldsymbol{C}/\mathcal{C} \ \dagger,$$

$$\mathbf{A} \sim (m,n) \quad , \quad \boldsymbol{\varepsilon} \sim (m,k) \quad , \quad \boldsymbol{\delta} \sim (m,k) \quad , \quad \boldsymbol{C} \sim (k,n) \ ,$$

m is the number of wavelengths (NMR active nuclei), n is the number of solutions, k is the number of independent ("active") components, $\mathbf{A}$ is the absorbance matrix (the matrix of chemical shifts) ($a_{ij} \equiv y_{(i-1)\times n+j}$), $\boldsymbol{\varepsilon}$ is the matrix of molar absorption coefficients, $\boldsymbol{\delta}$ is the matrix of the chemical shifts of pure components and $\boldsymbol{C}$ is the matrix of the free concentrations of components (a column per solution).

Obviously $r(\mathbf{A}) \le \min(m, n, k)$. Thus, if both $m > k$ and $n > k$, the true rank of the absorbance matrix (the matrix of chemical shifts) gives us a reasonable estimate of the number of light absorbing (δ influencing) components in a solution $k \ge r(\mathbf{A})$. Moreover, the number of observations can dramatically be reduced from $m \times n$ to $k \times n$ by means of the singular value decomposition of the matrix $\mathbf{A}$ (no more than k suitable linear combinations of the m original wavelengths are needed).

For $m \ge n$

$$\mathbf{A} = \mathbf{U}_n^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{V} \ ,$$

$$\mathbf{U}_n \sim (n,m) \quad , \quad \boldsymbol{\Sigma} \sim (n,n) \quad , \quad \mathbf{V} \sim (n,n) \ ,$$

$$\mathbf{U}_n\mathbf{U}_n^{\mathrm{T}} = \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{1}_n \ ,$$

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \dots, \sigma_n) \quad , \quad \sigma_1 \ge \dots \ge \sigma_n \ge 0 \ ,$$

$$\boldsymbol{\Sigma}\mathbf{V} = \mathbf{U}_n\boldsymbol{\varepsilon}\boldsymbol{C} \quad , \quad \boldsymbol{\Sigma}\mathbf{V} = \mathbf{U}_n\boldsymbol{\delta}\boldsymbol{\Pi}\boldsymbol{C}/\mathcal{C} \ .$$

¶ see the definitions of calibration functions of types 7 and -5

† Since only the "active" components are taken into account, matrix $\boldsymbol{\Pi}$ is diagonal containig only the non-zero diagonal elements of the matrix $\mathbf{P}$ (see the definition of a calibration function of type -5), i.e. $\boldsymbol{\Pi} \sim (k,k)$, $r(\boldsymbol{\Pi}) = r(\mathbf{P}) = k$.

Since $r(\mathbf{A}) \doteq k$, $\mathbf{\Sigma} \doteq \mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0_{k+1}, \ldots, 0_n)$¶. Then

$$\mathbf{\Sigma}_k \mathbf{V}_k \doteq \mathbf{U}_k \boldsymbol{\varepsilon} \boldsymbol{C} \quad , \quad \mathbf{\Sigma}_k \mathbf{V}_k \doteq \mathbf{U}_k \boldsymbol{\delta} \mathbf{\Pi} \boldsymbol{C}/\mathcal{C} \ ,$$

where the matrices $\mathbf{\Sigma}_k \sim (k,k)$, $\mathbf{V}_k \sim (k,n)$ and $\mathbf{U}_k \sim (k,m)$ are appropriate submatrices of the matrices $\mathbf{\Sigma}$, $\mathbf{V}$ and $\mathbf{U}_n$†.

For $m < n$

$$\mathbf{A} = \mathbf{U}^{\mathrm{T}} \mathbf{\Sigma} \mathbf{V}_m \ ,$$

$$\mathbf{U} \sim (m,m) \quad , \quad \mathbf{\Sigma} \sim (m,m) \quad , \quad \mathbf{V}_m \sim (m,n) \ ,$$

$$\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{V}_m \mathbf{V}_m^{\mathrm{T}} = \mathbf{1}_m \ ,$$

$$\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_m) \quad , \quad \sigma_1 \geq \ldots \geq \sigma_m \geq 0 \ ,$$

$$\mathbf{\Sigma} \mathbf{V}_m = \mathbf{U} \boldsymbol{\varepsilon} \boldsymbol{C} \quad , \quad \mathbf{\Sigma} \mathbf{V}_m = \mathbf{U} \boldsymbol{\delta} \mathbf{\Pi} \boldsymbol{C}/\mathcal{C} \ .$$

Since $r(\mathbf{A}) \doteq k$, $\mathbf{\Sigma} \doteq \mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0_{k+1}, \ldots, 0_m)$. Then

$$\mathbf{\Sigma}_k \mathbf{V}_k \doteq \mathbf{U}_k \boldsymbol{\varepsilon} \boldsymbol{C} \quad , \quad \mathbf{\Sigma}_k \mathbf{V}_k \doteq \mathbf{U}_k \boldsymbol{\delta} \mathbf{\Pi} \boldsymbol{C}/\mathcal{C} \ ,$$

where the matrices $\mathbf{\Sigma}_k \sim (k,k)$, $\mathbf{V}_k \sim (k,n)$ and $\mathbf{U}_k \sim (k,m)$ are appropriate submatrices of the matrices $\mathbf{\Sigma}$, $\mathbf{V}_m$ and $\mathbf{U}$.

"Trial and error" method

Let p, k, l are natural numbers and m, n are non-negative integers $(m+n < p \wedge k,l > m)$ with the following meaning:

p	...	number of "unknown" complexes,
m	...	number of "unknown" complexes with *higher* priority (all of them **must** definitely be in every generated combination),
n	...	number of "unknown" complexes with *lower* priority,
$p-m-n$	...	number of "unknown" complexes with *common* priority (at least one of them **must** be in each generated combination),
k, l	...	lower (upper) limit of order of combination.

Then the number of combinations (chemical model hypotheses) is

$$\begin{array}{rc} n < k-m : & \sum_{i=k-m}^{l-m} \binom{p-m}{i} \\ k-m \leq n \leq l-m : & \sum_{i=k-m}^{l-m} \binom{p-m}{i} - \sum_{i=k-m}^{n} \binom{n}{i} \\ n > l-m : & \sum_{i=k-m}^{l-m} \binom{p-m}{i} - \sum_{i=k-m}^{l-m} \binom{n}{i} \end{array} .$$

For each combination, the OPIUM program creates a shortened information about the goodness of fit (containing quantities such as sum of squares of residuals (U), Akaike information criterion (AIC), Schwarz[12]-Rissanen[13], Hannan-Quinn[14] and Anděl-Perez-Negrao[15] statistic) and stores it into a file§.

¶ To estimate the "effective" rank k of the matrix $\mathbf{A}$ (both the equations would hold exactly only if there were no measuring errors), sequence of residual standard deviations is calculated[11]

$$s_i = \sqrt{\frac{\mathrm{tr}\,(\mathbf{\Sigma}^2) - \sum_{j=1}^{i} \sigma_j^2}{p-i}} \quad , \quad i = 0, \ldots, p-1 \quad , \quad p = \min(m,n) \ .$$

Then the inequalities $s_0 > \ldots > s_{k-1} > \sigma_y \geq s_k \geq \ldots \geq s_p \geq 0$ should be valid, where σ_y is the standard deviation of the random variable y (absorbance, chemical shift).

† i.e. $\mathbf{A}' \doteq \boldsymbol{\varepsilon}' \boldsymbol{C}$, $\mathbf{A}' \doteq \boldsymbol{\delta}' \mathbf{\Pi} \boldsymbol{C}/\mathcal{C}$, where the matrices $\mathbf{A}' = \mathbf{\Sigma}_k \mathbf{V}_k^{\mathrm{T}} \sim (k,n)$, $\boldsymbol{\varepsilon}' = \mathbf{U}_k^{\mathrm{T}} \boldsymbol{\varepsilon} \sim (k,k)$ and $\boldsymbol{\delta}' = \mathbf{U}_k^{\mathrm{T}} \boldsymbol{\delta} \sim (k,k)$ are "reduced" equivalents of the matrices $\mathbf{A}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$

§ quite suitable for processing with standard Unix tools (`grep`, `sort`, `awk` etc.)

1. Kvala M., Luke I., CHEMOMETRICS '95, Pardubice, Czech Republic, 1995, Book of Abstracts p. 63
2. Havel J., Meloun M., *Talanta* **33** (1986) 525
3. Perrin D. D., Sayce I. G., *Talanta* **14** (1967) 833
4. Ginzburg G., *Talanta* **23** (1976) 149
5. Tripathi V. S., *Talanta* **33** (1986) 1015
6. De Robertis A., De Stefano C., Sammartano S., *Anal. Chim. Acta* **191** (1986) 1015
7. De Stefano C., Princi P., Rigano C., Sammartano S., *Computers Chem.* **13** (1989) 343
8. Golub G. H., Pereyra V., *SIAM J. Numer. Anal.* **10** (1973) 413
9. Maeder M., Zuberbühler A. D., *Anal. Chem.* **62** (1990) 2220
10. Meyer R. R., Roth D. M., *J. Inst. Math. Appl.* **9** (1972) 218
11. Kankare J. J., *Anal. Chem.* **42** (1970) 1322
12. Schwarz G., *Ann. Statist.* **6** (1978) 461
13. Rissanen J., *Automatica* **14** (1978) 465
14. Hannan E. J., Quinn B. G., *J. Roy. Statist. Soc.* **B41** (1979) 190
15. Andl J., Perez M. G., Negrao A. I., *Kybernetika* **17** (1981) 514