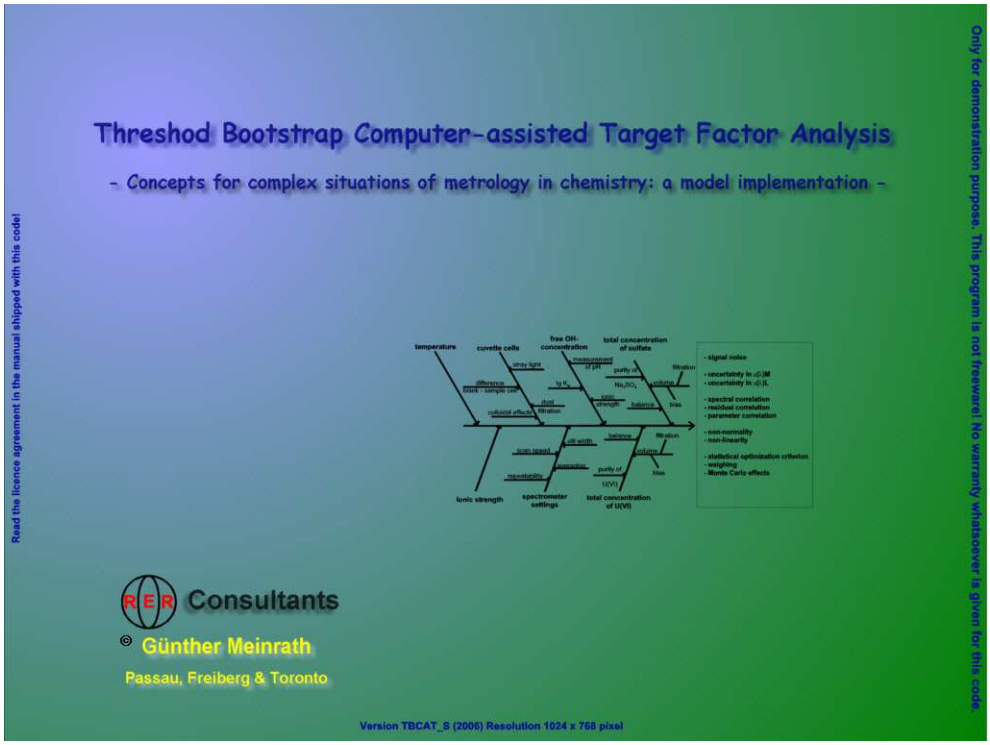
Threshold Bootstrap Computer-assisted Target Factor Analysis (TBCAT)

version S

A self-modelling tool for resolution of convoluted spectral information

by
Günther Meinrath



last addition: 10 March 2007

Table of Contents

License

Please read the terms of this agreement and any provided supplemental license terms (collectively 'agreement') carefully before installing TBCAT_S. By installing the software, the user agrees with the terms of this agreement. If the user is accessing the software electronically, he complies with this agreement by installing the software. If the user does not agree to all these terms, he must promptly deinstall this software and return the unused software or, if the software is accessed electronically, remove the software from the computer(s).

1. The copyright owner grants the user a non-exclusive and non-transferable right for the internal use only of the accompanying software and documentation. It is understood that the accompanying software is experimental – no warranty is included that the software may work on a certain computer and is fit-for-purpose (whatever the purpose might be).
2. The software is confidential and copyrighted. It is distributed free of charge but is not in the public domain. Title to software and all associated intellectual property rights is retained by the copyright owner. Any modification, especially by reversed engineering and decompilation, is forbidden except if explicitely authorized by the copyright owner. The user may not make copies except for personal archival purposes. No right, title or interest in or to any trademark, service mark, logo or trade name what so ever is granted under this agreement.
3. This software may be used, explicitely, for education purposes in an academic environment if it deems suitable. All necessary operations (except those requiring copyright infringement) may be performed with the software (e.g. installation of the software in a computer pool). There is any implicite or explicite claim that the software may be suitable for such a purpose. The sole responsibility remains with the user.
4. This agreement is effective until terminated. If any provision of this agreement is held to be unenforceable, this agreement will remain in effect with the provisions omitted, unless omission would frustrate the intent of the parties, in which this agreement will terminate immediately.
5. Source code is only available in the framework of a collaboration.

Citation

All reports of results obtained by application of the software, including illustrations, must be accompanied by the following citations:

G. Meinrath, S. Lis "Quantitative resolution of spectroscopic systems using computer-assisted target factor analysis (CAT)", Fresenius Journal of Analytical Chemistry 369 (2001) 124 - 133

G. Meinrath, P.Schneider "Quality Assurance in Chemistry and Environmental Science: Metrology from pH Measurement to Nuclear Waste Disposal. Springer Verlag Heidelberg (2007)

Disclaimer of Warranty

Unless specified in this agreement, all expressed or implied conditions, representations and warranties, including any implied warranty of merchantability, fitness for a particular purpose or non-infringenent are disclaimed, except to the extend that these disclaimers are held to be legally invalid.

To the extend not prohibited by the law, in no event will RER Consultants or the author or the programmer of the code be liable for any lost revenue, profit or data, or for special, indirect, consequential, incidential or punitive damages, however caused regardless of the theory of liability, arising out of or related to the use of or inability to use this software, even if RER Consultants has been advised of the possibility of such damages. In no event will RER Consultants' liability to the user, whether in contract, tort (including negligence), or otherwise exceed the amount paid by the user for Software under this agreement. The foregoing limitations will apply even if the above stated warranty fails of its essential purpose. This agreement is effective until terminated. The user may terminate this agreement by dstroying all copies of software. This agreement will terminate immediately without notice from RER Consultants if the user fails to comply with any provision of this agreement. Upon termination, the user must destroy all copies of software.

Introduction

The computer code '**C**omputer-assisted **T**arget Factor **A**nalysis' (CAT) has been implemented with the intention to provide a model-free tool for extraction of information from spectral data. The theoretical basis of CAT is factor analysis and its algorithmic backbone is the singular value decomposition algorithm.

There are ample programs described in literature capable to decompose spectral matrices into the single components and their respective concentrations. Nevertheless, CAT is unique in several aspects:

- CAT resolves the spectra on basis of a few informations provided by the user. The resolving procedure is otherwise automatic. Hence, the user input decides upon CAT's performance. Therefore, the attribute '-assisted' has been chosen as an indication that CAT isn't a completely automatic procedure. The attribute '-assisted' emphasizes that CAT is not a black box. The user needs to understand profundly what CAT is doing, how CAT is doing its work and which options can be manipulated in what manner.

- CAT allows to repeat the resolution procedure unassisted if the user has made her/his choices. Hence, as long as the input data doesn't vary too strongly the same set of user's parameters can be used to repeat spectral resolution without further need for interference. To the best knowledge of the author there is no other code available with this feature.

- Due to the capability to rerun slighly modified data sets without further need for user guidance, CAT is ideally suited to run in a bootstrap environment. Bootstrapping is a computer-intensive resampling technique to assess statistical properties of complex data sets without the need for complicated and often unavailable parametric statistical analysis. Hence, bootstrapping allows to extract location and dispersion estimates from data sets. Bootstrap methods can be applied to all kind of data sets provided the enormous amount of data generated from these methods. Fast modern computers and cheap digital storage render bootstrap methods to the ideal work horses for chemical data.

These three features are the key reason for the existance of CAT. But there is a fourth one, which at the time of writing these phrases is probably not aware to most chemists: CAT can be implemented into a code evaluating the complete measurement uncertainty of a spectroscopic measurement process. Thus, it is a very first implementation of a metrologically sound treatment of measurement uncertainty in complex chemical analysis. Metrology in chemistry, as metrology in general, is based on international treaties and agreements. Metrology is becoming the language of the international market place. The need to prove the quality of forwarded information will not exempt chemistry, least analytical chemistry. The need for comparable data of stated quality will change the way of analytical chemistry. CAT is an example how to realize certain aspects of the traceability chain in chemical analysis.

CAT is not a professional program. It is probably also not correct to say that CAT has been developped. CAT has been created by an evolutionary process. Users who expect the versatility and convenience of a modern commercial computer code will certainly not cheer the handling of this CAT implementation. The algorithms, however, are sound and tested. The algorithms are taken from available literature including Golub and Reinsch' Singular Value Decomposition (SVD), Box and Muller's algorithm to create normally distributed random deviates, some algorithms from the notable 'Numerical Recipes' and Nash's 'Compact Numerical Algorithms'.

The major motivation to start the implementation of CAT has been the search for a stable and easy-to-use method for deconvolution of overlapping spectra in rare-earth systems. The first tests have been made with an implementation in interpreted QuickBasic (QB). Because QB is limited to a memory of 64 KB an extended version was created in Visual Basic 5. As already mentioned before, there exist other proposals to deconvolute spectral information. But none of these proposals is simple enough to allow novices or even spectroscopists without any experience in chemometrics to handle the necessary input. It is believed that the algorithm implemented here has the required conceptional simplicity. However, since it is definitively not a black box program, the user must understand some mathematical and statistical concepts like matrices, correlation, and probability distributions. Experience with scientific programming is also recommended.

The code has been tested by applying it to problems where the approximate solution has been evaluated before. These tast cases have been UV-Vis spectra of U(VI) carbonate and hydrolysis systems in aqueous solutions. U(VI) shows an extremely weak absorption in the UV-Vis range that increases continuously towards the UV region. Hence, these spectra do not provide a baseline towards the UV. Such spectra are especially challenging because baseline corrections depend solely on the spectral baseline information collected towards the IR side of the spectra. The code has been applied to evaluate a larger number of rare earth systems in aqueous and non-aqueous solutions. Some data have been published in the reviewed literature.

CAT Capabilities

CAT performs the following tasks:

- linear background correction using either data from right and left side of a spectral band or a single side selected by the user. The range of spectral information considered as background is also specified by the user. The spectra where background correction should be performed can be selected individually by the user.
- estimation of the number of species contributing to a spectral system. The selection criterion is a figure representing the sum of least squares residuals of some properties of the resolved matrices, i.e. negative absorption or species concentration values.
- automatic generation of default files with input data for the rank of matrices (= number of species) and starting values for the iterative optimisation process.
- estimation of singular values, abstract singular vectors of spectral and concentration matrices.
- estimation of single species spectra and concentrations.
- estimation of formation constants for the coordinated species based on user-supplied information about the likely stoichiometry. The concentration quotients are weighted and a weighted estimate of the formation constant is provided together with a detailed table of the concentration estimates.
- evaluation of molar absorption estimates and absorption maxima.
- user-specified weighing of residual components in the target iteration step.
- least-squares fit of experimental spectra using evaluated mean value single component spectra.
- threshold bootstrap computer-assisted target factor analysis (TB CAT) procedure with user-specified number of resumpling cycles.
- generation of cumulative distributions for formation constants including a differentiation routine on basis of LOESS non-parametric regression algorithm.
- evaluation of empirical confidence bands for single component spectra for 68%, 90%, 95% and 99% confidence limits.

These features allow a complete analysis of spectral informations in a chemical system under study. In fact, in case of a TB CAT analysis the user obtains full information including confidence limits about a previously unknown chemical system studied by spectroscopy.

1 Theoretical Basics

Factor analysis decomposes a matrix of absorptions, say X, into two matrices E and C. Matrix E consists of n columns, where each column is the single component spectrum of one of the n relevant species. Matrix C consists of n rows, where each column holds n concentrations of the respective solution species. By matrix multiplication, elements x_{ik} of data matrix X can be calculated from the elements of the matrices E and C, resp. within the validity of Beer's Law

$$x_{ik} = \sum_{j=1}^{n} \varepsilon_{ij} \cdot c_{jk} \qquad (1)$$

x_{ik} : absorption observed at wavelength i in the k-th solution
ε_{ij} : molar absorption of the j-th species at wavelength i
c_{jk} : concentration of the j-th species in the k-th solution

or, expressed in matrix formulation

$$X = E\,C \qquad (2)$$

First step in the analysis is to determine the number n of factors significantly contributing to the observed spectral variance by Abstract Factor Analysis (AFA). A variety of equivalent techniques are available for AFA, e.g. Jacobi Rotation, nonlinear iterative partial least squares algorithm (NIPALS) and singular value decomposition (SVD). By these techniques, observed variances are interpreted by a set of mutually orthogonal vectors (the eigenvectors of matrix X), where each vector is chosen to extract successively as much of the data variance as possible. In this work, SVD algorithm is used for determination of eigenvectors and their eigenvalues λ. A real data matrix X_{rc}, of r rows and c columns with $r \geq c$, is decomposed according to Eq. 3:

$$X_{rc} = U_{rc}\, S_{cc}\, V^{T}_{cc} \qquad (3)$$

into a unitary matrix U of the column eigenvectors of X, a unitary transposed matrix V of the row eigenvectors of X and a diagonal matrix S, composed of the roots of the eigenvalues of X and elements $s_{ij} = 0$ ($i \neq j$). SVD extracts the roots of eigenvectors in decreasing relevance. The associated diagonal values s_{ii} with $s_{ii}^2 = \lambda_i$ (λ_i = i-th eigenvalue) are ordered with decreasing magnitude. It is straightforward to identify

$$U_{rc}\, S_{cc} = E^{\ddagger} \qquad (4)$$

$$V_{cc} = C^{\ddagger} \qquad (5)$$

Data matrices $E^{\ddagger}$ and $C^{\ddagger}$ contain the requested information, however in a mathematical, abstract form and associated with random error and bias.
If experimental data could be obtained unaffected by random errors and bias, AFA would result in a limited number of non-zero eigenvalues in S_{cc} corresponding to the dimensionality of the data matrix, that is the number of factors contributing to the experimental data under investigation. However, experimental data can hardly be obtained without random errors and bias. Therefore, all eigenvalues λ_i are non-zero, albeit usually quickly approaching very small

values with increasing i. Decision on the dimensionality of the data space therefore has to be based on statistical tests. Only the n largest eigenvalues λ and the associated column and row eigenvectors are contributing significantly to the experimental variance.

The remaining c-n eigenvalues and the associated eigenvectors $\lambda°$ form the so-called null space or error space (indicated by (°)) and are excluded from the further analysis. Forming matrices E* and C* from the first n row and column eigenvectors only allows calculation of a matrix X', where random errors and bias are reduced by omitting summation over the null-space. Therefore, X' differs from X by the amount of removed random error and bias. The informations thus filtered from the original data has to be transformed from the abstract orthogonal eigenvectors into physically meaningful vectors by rotating the n eigenvectors. This transformation is called target rotation by rotation matrix T:

$$X' = E\ C = E^*\ T\ T^{-1}\ C^*. \qquad (6)$$

The main task in this step is to identify a suitable rotation matrix T, that yields the required information.

In order to test a suspected single component spectrum s_i, a least square transformation vector t_i is calculated according to Eq. 7

$$t_i = E^{*+}\ s_i \qquad (7)$$

where E^{*+} is known as the pseudo inverse of E*. By Eq. 8, a vector x_i is predicted, being the best projection of s_i into the factor space of matrix E*:

$$x_i = E^*\ t_i \qquad (8)$$

The difference between s_i and x_i can be used as a measure for the acceptability of a suspected component vector s_i.

2 Statistical Basics

Least-squares, computer-intensive resampling and robust principal component analysis

The statistical treatment of factor analysis is a complex issue. Some references are given in the Appendix. It is fundamental to keep in mind that factor analysis is a method of linear regression. Linear regression is a common tool in interpreting data sets. Its availability on pocket calculators further contributed to its current position as the most often applied tool for interpreting univariate data sets with apparent linear relationship. The coefficient of correlation is commonly (and inadequately) considered as a measure for the quality of fit.

However, the concept of linear regression is based on several assumptions. The figures created via linear least-squares regression are meaningful only if these assumptions are valid:

1. the expectation function is correct
2. the response is expectation function plus disturbance
3. the disturbance is independent of the expectation function
4. each disturbance has normal distribution

5. each disturbance has zero mean
6. the disturbances have equal variances
7. the disturbances are independently distributed

It has been shown that some of the requirements are not fulfilled for spectroscopic data. The least-squares criterion has some optimality properties only in case the seven requirements hold. If the requirements do not hold, the optimality properties of the linear least-squares estimates are likewise not valid.

Multivariate analysis depends on the same critera. A value obtained from a factor analysis is doubtful if the disturbances (often also termed residuals or noise) are not, for instance, identically and independently distributed (i.i.d.). Criterion 2, for instance, can be expressed in matrix formulation

$$A = X + N \tag{9}$$

where A is the matrix of spectral observations, X is the matrix of the true absorptions and N is the matrix of disturbances. After a factor analysis is performed, the matrix A' is formed being the estimate of X. Then, the matrix A-X = N', with N' being the estimate of the disturbances. An autocorrelation analysis of the disturbances will show immediately that the spectral residuals are not i.i.d. Consequently, the data in A' has no optimality criterion - it is in a certain degree arbitrary.

Here, certain statistical techniques can be introduced to estimate the range in which the optimal estimate can be found. These techniques do not depend on complicated mathematical analysis but on brute computing force. These techniques have been termed bootstrap techniques. In fact, the father of bootstrap techniques, B. Efron, intended to term the technique 'shotgun' because it "can blow the head off any problem if the statistician can stand the resulting mess". The resulting mess probably direted to the enormous amount of data generated by a bootstrap procedure. But in the times of GHz CPUs and Gbyte digital storage capacities, the mess can be handled easily.

Bootstrap methods are computer-intensive Monte Carlo methods. Bootstrap methods replace sophisticated and often unavailable statistical theory by the brute computing power. These techniques allow to concentrate on the answers that are asked for instead of traceing the questions to a small catalogue of mathematically solvable cases that itself are mostly only valid for large sample approximations, e.g. normal distributions..
A data set of experimental observations (x_i,y_i) is used as a basis of discussion. The data are sampled from an unknown probability distribution F conditional on x. The data are interpreted by a parametric model function eq. 10.

$$F(x)=F(x; X) \tag{10}$$

Due to unavoidable experimental errors, the total process is expressed actually by

$$y_i = F(x; X) + \xi_i \quad (11)$$

where ξ_i represents the random process while F(x, X) represents the deterministic part. Hence, the information on the parameters X is enclosed in the data (x_i,y_i) and extracted via the model function F(x, X). The presence of random noise ξ transfers to the parameters X (X=K, $ß_{101}$, $ß_{102}$, $ß_{103}$). The job is to obtain the probability distribution of a statistics θ, say $\theta = t(X)$. Having evaluated an estimator $\hat{\theta} = t(X)$ conditional on the predictors x, the next task is to assess the accuracy of $\hat{\theta}$ as a estimator for the true value of θ. The standard error of $\hat{\theta}$, the square root of its variance,

$$\sigma(\hat{\theta}; F) = [var_F(t(X))]^{½} \quad (12)$$

is the most common measure of accuracy for estimators $\hat{\theta}$ that are unbiased, or nearly so. A formal expresion exists if the statistic θ is the mean:

$$\sigma(\hat{\theta};F) = [\sigma^2(F)/n)]^{½} \quad (13)$$

An unbiased estimate for $\sigma^2(F)$ is available

$$\hat{\sigma}^2(F) = \frac{\sum_{i=1}^{n}(X_i - \hat{X})^2}{(n-1)}$$

Thus, the well-known estimate of a standard error of a mean is obtained.

$$\sigma(\hat{\theta}) = [\frac{\sum_{i=1}^{n}(X_i - \hat{X})^2}{(n-1)\,n}]^{½} \quad (14)$$

Note that the estimate of the standard error is valid independent of F. However, such a closed formula is available only for the sample mean. Other statistics, e.g., median, correlation coefficient, skewness or ratio of means, such formula are not available.
In order to apply eq. 12 an estimate for F, say $\hat{F}$, is required. It is common to circumvene this requirement (or better: to ignore the question completely in many cases related to chemistry) by using the normal distribution $N(\sigma, \hat{\theta})$ as an estimate for F. This plugging-in of $N(\sigma, \hat{\theta})$ for F will be termed as 'classical statistics' in the following.
Efron showed that a simple computer algorithm exists for estimating a distribution $\hat{F}$ of the observed data (x_i,y_i) that gives the standard error estimate

$$\sigma(\hat{\theta}; \hat{F}) = ((\overline{x - \bar{x}})^2 / n)^{½}, \quad (15)$$

almost the same as the classical estimate eq. 14. This algorithm gives

$$\sigma_{boot}(\hat{\theta}) = \sigma(\hat{\theta}; \hat{F}) = (\mathrm{var}\,\hat{F}(t(x^*)))^{1/2}, \quad (16)$$

where x* is a hypothetical data vector generated by indendently and identically distributed sampling from the distribution $\hat{F}$. Hence, the empirical distribution $\hat{F}$ is created by resampling from the original data (x_i,y_i) via Monte Carlo methods since the data represents a realization of the original but unknown distribution F. This procedure is called 'bootstrap' and the standard deviation σ_{boot} is termed 'bootstrap standard deviation'. $\hat{F}$ is the maximum likelihood approximation of F. The necessary number of resampings is large, e.g., 1000 resamplings. Therefore, bootstrap methods are computer-intensive Monte Carlo procedures that became feasible only by the increase of cheap computing power in the past twenty years.

It is straightforward to show that the linear least-squares approach gives confidence regions different from the bootstrap estimates.

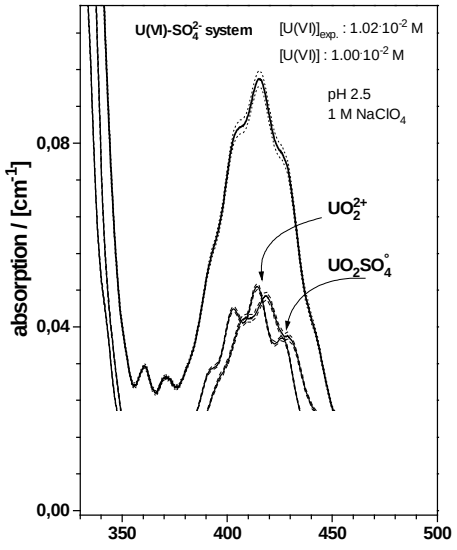


Figure a: Result of a MBB spectral analysis of an mixed component U(VI)-SO_4^{2-} UV-Vis absorption spectrum

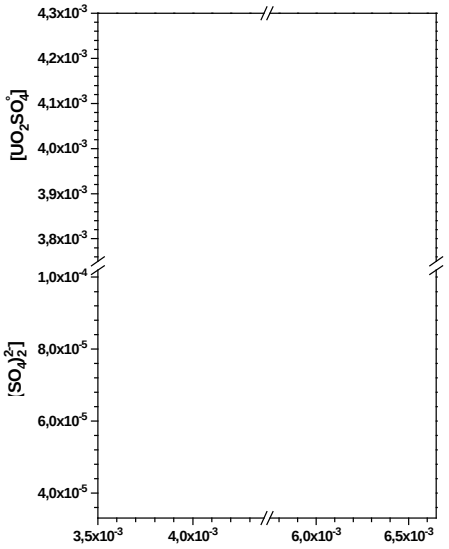


Figure b: Comparison of LSR 95% confidence ellipses with MBB results from 1000 resampling cycles for spectrum fig. a. The bias of LSR is obvious. Coverage of LSR ellipses by MBB results is almost negligible.

Figure a shows the interpreted three-component spectrum of a sulfate-containing U(VI) solution. Fig. b givs a comparison of a least-squares confidence region (LSR; dashed ellipses) and the corresponding distribution of 1000 bootstrap resamplings using a special bootstrap method called Moving Block Bootstrap (MBB). It is inteesting to note that the MBB estimates are found almost consistently outside the range given by the 99% confidence ellipses. The LSR mean values are found outside the MBB cloud of data points.

To check the independence of the residuals, an autoregression analysis is performed. Autoregression (AR) assumes that there is a dependence of neighboring residuals over a distance of r neighbors, where r is called the 'lag'. A simple AR scheme is the a second order autoregressive scheme, AR(2), where the lag r=2. An AR(2) scheme estimates correlation by interpreting a residual z_t from spectral analysis due to a model

$$z_t = \beta\, z_{t-1} + \gamma\, z_{t-2} + \varepsilon_t, \qquad (17)$$

where the t-th observation depends on the two previous observations z_{t-2} and z_{t-1} by the two parameters β and γ, resp. The disturbances ε_t are random contributions. Correlation is indicated by autoregression parameters β, γ significantly different from zero. By eq. 17, the seemingly random and uncorrelated residuals are interpreted by a model that exlains the magnitude of the t-th residual by the magnitudes of the both closest neighbour residuals to the left. This is a kind of time-series analysis. A non-zero value of β and γ doesn't automatically indicate correlation. he magnitude of β and γ must be significantly diferent from zero. Again we are using bootstrapping to evaluate approximate confidence regions for the two parameters by sampling from the residuals ε_t. A bootstrap analysis gives AR(2) coefficients $\beta = 0.2867 \pm 0.03293$ and $\gamma = 0.66924 \pm 0.0329$. Both AR(2) coefficient estimates are significantly different from zero, as indicated by the standard deviations.

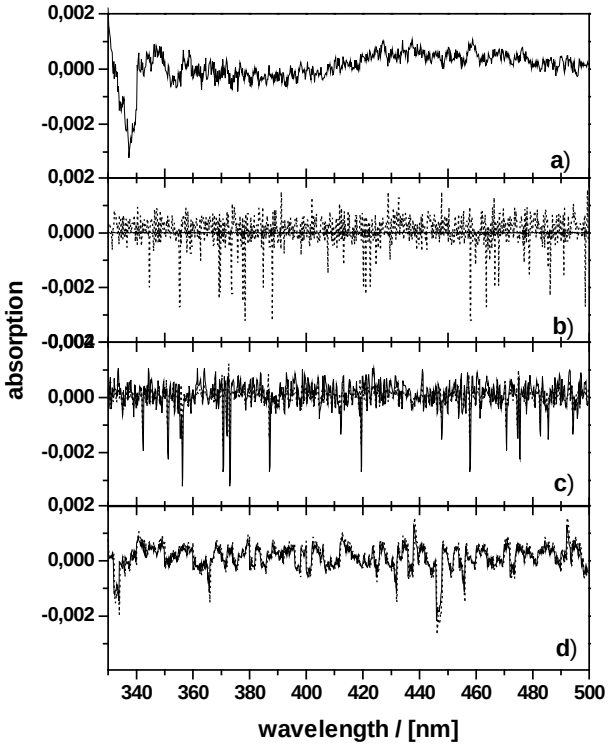


Figure c: Residuals (a) and results of AR(2) analysis for a Moving Block Bootstrap realization from the residuals with lag size r=1 (b), r=2 (c) and r=10 (d). MBB generated residuals are given an dotted lines, while AR(2) fits are given as straight lines.

Figure c shows the influence of the lag size on the reproduction of noise patterns generated by random sampling from the original residuals. It is obvious that the noise gets the more similar to the original noise the larger the lag size. It is very difficult to determine an optimal lag size. Hence, arbitraily selecting a lag size introduces considerable subjectivity. If the lag size is inadequate, additional noise is introduced into the regenerated residuals. As a matter of fact, it has been shown that both a fixed lag size isn't approriate in most cases and an inappropriate lag size may cause large breaks in the time series. A remedy is the threshold bootstrap.

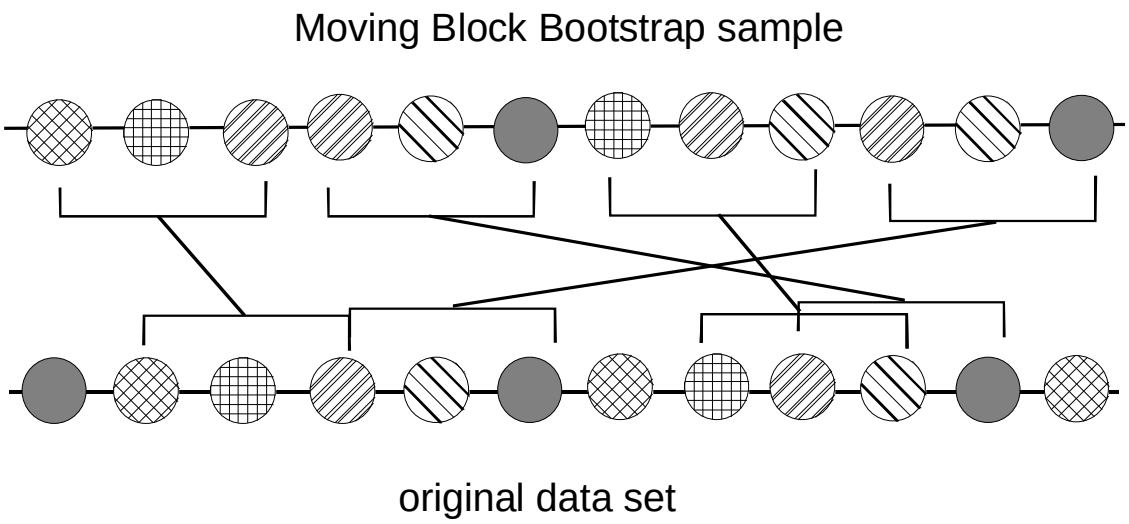


Figure d: Schematic representation of the Moving Block Bootstrap scheme. Twelve symbols representing the original data set are randomly redistributed in a Moving Block Bootstrap sample by choosing blocks of size r (here r=3).

Threshold bootstrapping residuals is a self-adaptive procedure. Hence, the noise itself defines the lag size. To perform a threshold bootstrap, the noise is mean-centered. Hence, the time series will run above and below the zero axis. A lag is now selected between points of a lag crossing the zero lines twice. Hence, the time series of noise is separated in individual regions of different length. When the different lags are assembled to from a new noise time series, no larger jumps at the connecting positions are introduced because in the original noise series, a transition above or below the mean value of the total series would have occured, too. By using a threshold bootstrap strategy, we can take into account the correlation in the spectral residuals. In eq. 9, the matrix N is a random matrix. The randomness may influence the interpretation of the spectral data. By threshold bootstrapping, an optimal scheme is applied to investigate the influence of N on the possible interpretation of the spectral information.

It is necessary to give some comments about the difference between computer-intensive resampling methods and robust multivariate regression techniques like robust principal component analysis (RPCA). The robust regression techniques attempt to eliminate the effects of influential observations on the regression procedure. The term 'influential observations' refers to more than just 'outliers'. An 'outlier' can be easily spotted visually in case of

univariate data. In certain situations, simple and efficient tests (e.g. Dixon's test) are helpful in making a decision about a suspected outlier more objective. In univariate data (xy data), the definition of an outlying observation is much more difficult. Of course, there are data points whichobviously deviate crudely from the remainder of the data set. However, the characterisation of a data point as an 'outlier' is more tricky than an untrained scientist may imagine. Just consider a situation where a large data set holds multiple 'outliers'. Due to the variety of situations where the structure of a data set may give rise to the assumption that one or even several members of the set are 'outlying', the term 'influential observation' has been coined. Set members with high leverage are an example of such influential data points. Robust regression techniques try to identify influential data points and to reduce their influence on the data analysis procedure. As an example of robust regression techniques, the 'least median of squares' method is mentioned here. Instead of minimising the sum of squares of residuals, as in common linear regression, the median of quared residuals is idenified. This happens by collecting the residuals, odering them according to size and identifying the residual at position N/2+1 of an uneven number N of data points, or the average value between N/2 and N/2 + 1 for an even number N of data points. Optimisation criterion is the minimisation of this median value. It is shown that almost 50% of data may be outlying without changing the position of this median value - and consequently the position of the regression line. In case of classical least sum of squared residuals regression, one single outlier changes the position of the regression line. Robust regression does have its disadvantages. For classical regression, a one-step algorithm exist to identify the optimum regression line. In case of, say, least median of squares regression, a complicated search algorithm must be implemented. There is no algorithm which warrants that the optimum robust regression line has been found.

What is true for univariate data sets is the more true for multivariate data sets. Recently, robust median of squared residual methods for principal component analysis have been made available. These techniques exclusively treat influential observations. Other effects, especially correlation effects, deviations from normality and non-linearity are not treated. Because no algorithm exists warranting optimal parameter extraction, RPCA does not have advantages in the present context. It is clear, however, that preparing samples and collecting spectra must be performed as careful as possible. Any incorrect procedure may spoil the results produced from CAT.

3 Computer-assisted Target Factor Analysis

There are a larger number of algorithms available in literature that make use of the fundamental mathematical relationships summarized in eq. 6. Examples are the SIMPLISMA approach, the convexity approach, orthogonal projection method and the Iterative Key Set Factor Analysis. Each of these methods does have its strengths. However, for the present situation they all have a fundamental disadvantage: They need user interaction. If, however, 1000 to 2000 repetitions of a data analysis are to be repeated, user-interaction in each step is impossible.

CAT takes another way. A matrix A' holds a series of informations about the chemical system under study, however affected by noise and error. The error is a general nuisance factor and must be minimised as far as possible. Factor analysis cannot remedy the inadequacies from experimentation. The noise shall be termed 'measurement uncertainty' from now on. Measurement uncertainty is an unavoidable part of all experimental data. Minimising measurement uncertainty is one of the major tasks of a scientist. However, all quantitative information concerning the composition of samples, the concentrations of components and the spectral absorptions carries measurement uncertainty. And uncertainty means doubt. Because neither humans nor computers are gods, we cannot know truth. Hence, we have to evaluate our information taking the doubt into account. Doubt needs not to stop us making conclusions. But it helps ourselves and others to assess the reliability of a conclusion and to direct us to possible other interpretations. Hence, the task of a numerical interpretation is not to take the burden of interpretation from a researcher's shoulders but to allow him to adequately balance the weights...

Estimating rank

While the treshold bootstrap algorithm allows to account for residual correlation, an estimate of the number of spectral contributions is required, that is the true rank of the experimental data matrix A. There are a larger number of techniques to estimate the rank of A. But all these techniques (scree test, Bartlett's χ^2 criterion, Durbin-Watson statistics, canonical correlation etc.) make statements about random properties and give only probabilistic comments about the likely rank of A. Even if the probability for a certain value of the rank should be high, it doesn't exclude the other ranks with low but non-zero probability to be correct anyway. Eventually, all possible ranks must be tested and the resulting interpretations of a system must be balanced. It occurs that factor analysis doesn't suggest a unique interpretation of a system but points out different possible models. It is the scientist's task to develop a strategy to come to a final conclusion.

For UV-Vis spectraoscopic system, because the Bouguer-Lambert-Beer law allows to do so, a data matrix of different spectra of a chemical system can be interpreted with clear boundaries. The first boundary is the non-negativity of oth concentrations and absorptions. A furtherboundary can be selected if a component spectrum is known, for instance the uncoordinated, hydrated metal ion. Furthermore, the total concentrations of metal ion(s) and ligand(s) in the system are known, at least within the limits of measurement uncertainty. These informations are sufficient to resolve a chemical system. Hence, if a fast algorithm is available to resolve the mathematical equations at least semi-automatically, there is no need to rely on statistical tests in searching the true rank of an experimental data matrix A. Then, the total numerical process can be taken for any selected rank and the results can be compared. Often, one interpretation suggests itself while other ranks lead to numerical results with larger negative values in either the single component spectra and/or the species or ligand concentrations.

Performing the Target Rotation

The decisive step in target factor analysis is the target rotation step. If a rank is specified by the user, the mathematical-numerical steps to eigenvectors and eigenvalues are straightforwardly accomplished by efficient algorithms. Hence

$$X' = E\ C \qquad (18)$$

is available without requiring further attention.

The essential step is the target rotation matrix T. T must fulfill several requirements, especially leading to non-negative values in E* and C*. There is no easy way to generate a matrix with these properties. However, target transformation matrix T has a helpful property: T is comparatively small. If the system under study is a two-component system, dimensions of T are 2 x 2. For an n-component system, the dimensions of T are n^2.

CAT estimates the elements t_{ij} of T via a SIMPLEX algorithm, using the sum of neative elements in the estimates of E* and C* as a criterion. Using the non-negativity as an optimisation criterion has, however, a special case: If the diagonal elements t_{jj} in T are zero, then all elements in E* and C* are zero, too and hence, non-negative. It is necessary to avoid this trivial solution.

Avoiding Trivial Solutions in Target Rotation Matrix T

It is not sufficient to fix the values of the diagonal elements in T to, say, +1 each. While the matrices A, E* and C* must be positive semi-definite, T needs not. All values between -∞ and +∞ are principially allowed. Hence, the diagonal values of T must be set to either +1 or -1. And here, a problem arises. Which combination of +1 and -1 values is adequate?

Hence, all combinations of +1 and -1 values must be investigated. But the computational burden can be lessened because the first eigenvector is always either completely positive (t_{11} = 1) or negative (t_{11} = -1). Hence, the correct value for t_{11} can be taken directly from the properties of the first eigenvector in E. For all other diagonal elements in T, the combinations must be tested. Experience tells that there occur rare cases where two combinations of diagonal values give comparable but different results. CAT has a built-routine which allows to search for the adequate combination of diagonal elements.

The user will probably be surprised to be required to enter "0" instead of "-1". The diagonal elements are given as combinations of 1's and 0's and refered to as 'keys'. A 'key' 011 refers to diagonal elements (-1,1,1), while a key 0100 refers to (-1,1,-1,-1).

SIMPLEX optimisation of the off-diagonal elements in T

The SIMPLEX routine is an iterative optimisation algorithm that does not need analytical or numerical derivatives. Hence, the SIMPLEX can be applied with optimisation criteria different from the squared residuals (termed L_2 criterion instatistics). The SIMPLEX also finds minima in the error space if the optimisation criterion is, e.g., the smallest sum of negative values. In case of implementation in CAT, the SIMPLEX algorithm starts with randomly chosen values in the n(n-1) off-diagonal elements of T. Some precautions are necessary because there exist local minima in the error-space of the iteration procedure. Repeating an analysis with different starting values is often helpful. CAT has this feature built-in. In fact, CAT allows the resolution of a spectroscopic system by optimizing n(n-1) values of a transformation matrix T. After exploring certain properties of the data space like rank and combination of diagonal values, CAT provides a solution without interference by the user. And it repeats this analysis with slightly modified values independently. Hence, CAT can advantageously used as a core element of a threshold bootstrap process.

This is the philosophy of Threshold Bootstrap Computer Assisted Target Factor Analysis.

4 Performing a CAT Analysis

A typical CAT analysis follows the following procedure:

1. collection of a set of UV-Vis spectra under multivariate conditions
2. modification of the spectra according to the CAT input conventions
3. background correction of the spectra and, if necessary, generation of a uniform wavelength range
4. analysis of the spectral information to estimate the likely number of species and the corresponding key of diagonal elements
5. CAT analysis to test possible species combinations and stoichiometries
6. Input of the Type B measurement uncertainties
7. evaluation of the complete measurement uncertainty budget via Threshold Bootstrap Computer Assisted Target Factor Analysis (TB CAT)
8. evaluation of the distributions of formation constants and spectral information

5 Installation

Requirements

Screen resolution of at least 1024 x 768 pixels
Windows OS 98SE or XP with at least 500 MHz processor

Installation

TB CAT is available as a compressed ZIP file. The ZIP file should be unpacked into a temporary directory. Upon clicking the SETUP.exe the installation routine starts. Follow the guidance of the installation routine in the way that is common for the installation of new software on Windows OS. The routine creates a desk top icon but does not make modifications in the registry.

Known bugs

This code is a work-in-progress. The code is not a commercial program. It has a number of incomplete sub routines and dead-end menues. However, the code is currently running stable on a variety of Win 98SE and XP machines.

Input Files

For its performance, TB CAT needs a few data files which must be prepared by the user. These files are

a) spectral data files with component and concentration information in the header section
b) input files with starting information for the target rotation procedure
c) estimates of the ISO Type B uncertainties

Spectral data files are obtained from the experiments.
Header sections hold the information about the composition of experimental solutions.
Input files for the target rotation procedure are generated during the ANALYZE step of the CAT procedure. The user may request the system to generate several files in order to test different hypotheses.
The ISO Type B uncertainties are educated guesses or may be obtained from repeated experimentation

6 Running CAT

Data File Format Convention

The format of input data files is as follows (statement in *italics* are comments):

first line: beginn
beginn is a key word

second line: component 1, component 2,....
example: Co2+, NH3, Cl-:
1. Note: the component, whose spectral information is investigated, must be given in first place!
2. The component pH requires a special treatment, e.g. statement of ionic strength

third line: a,b,c
a,b,c are concentrations of the components in the given sample solution.
example: 2.1E-3, 0.0045, F2E-4
the key word F before a concentration indicates a free concentration. Otherwise, the free concentrations are obtained arithmetrically as the difference between the total concentration of a component (e.g. a ligand) and the concentration found by CAT analysis to be bound to the metal ion. Values of pH obtained from glass electrode potentiometry must always be given with an F.

fourth line: ende
ende is a key word indicating end of header section

fifth line: d,e,f
d: number of lines (generally the number of (wavelengths, absorption) pairs
e: the shortest wavelength
f: the longest wavelength

all following lines: g,h
g: wavelength value
h: absorption value

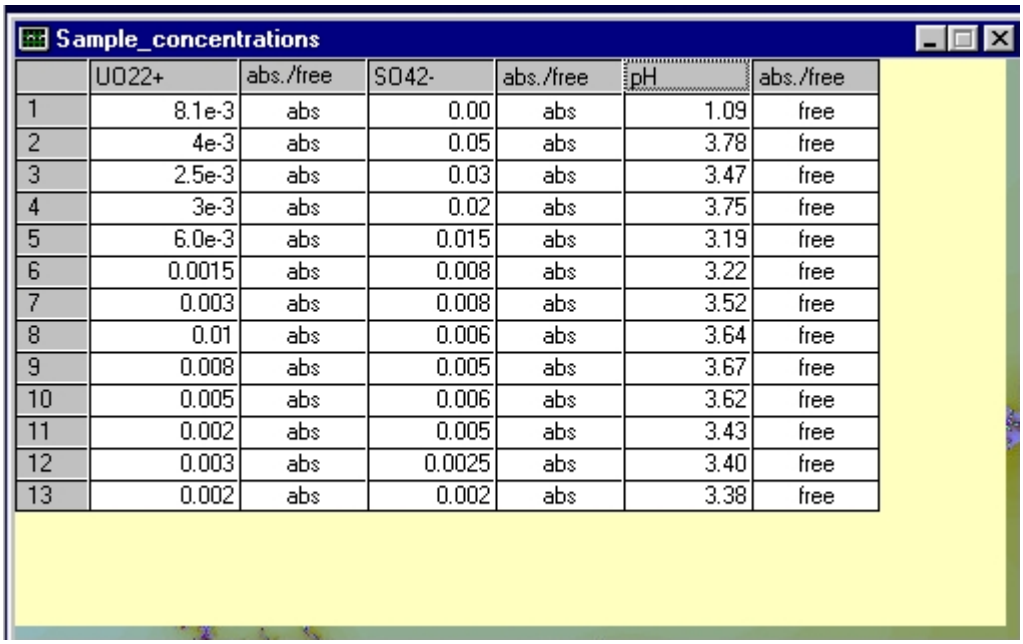
```
beginn
UO22+,SO42-,pH, .1
0.01,0.0015,F3.25
ende
2051,325,530
325,1.99791648825377
325.1,1.98821712932768
325.2,1.97811777040159
325.3,1.9681184114755
325.4,1.9576190525494
325.5,1.94731969362331
325.6,1.93622033469722
325.7,1.92642097577113
325.8,1.91562161684504
325.9,1.90362225791894
326,1.89322289899285
326.1,1.88122354006676
326.2,1.86932418114067
326.3,1.85732482221458
326.4,1.84492546328848
326.5,1.83202610436239
326.6,1.8187267454363
326.7,1.80692738651021
326.8,1.79422802758411
326.9,1.78092866865802
327,1.76852930973193
327.1,1.75482995080584
327.2,1.74043059187975
327.3,1.72773123295365
327.4,1.71403187402756
327.5,1.69973251510147
327.6,1.68493315617538
327.7,1.67033379724929
```

Figure e: First lines of a CAT input file of one UV-Vis spectrum

Sequence of spectra in the input procedure

With a single exception, the input sequence of the files can be selected arbitrarily. The exception is the spectrum of the component with the known single component spectrum. This spectrum MUST be loaded in first position. It is, therefore, helpful to name the data files appropriately to allow convenient sequential loading of the files.

Spectra are loaded into the program via the File menu. If the respective files are loaded, CAT displayes the concentration information as follows:



Sample_concentrations

	UO22+	abs./free	SO42-	abs./free	pH	abs./free
1	8.1e-3	abs	0.00	abs	1.09	free
2	4e-3	abs	0.05	abs	3.78	free
3	2.5e-3	abs	0.03	abs	3.47	free
4	3e-3	abs	0.02	abs	3.75	free
5	6.0e-3	abs	0.015	abs	3.19	free
6	0.0015	abs	0.008	abs	3.22	free
7	0.003	abs	0.008	abs	3.52	free
8	0.01	abs	0.006	abs	3.64	free
9	0.008	abs	0.005	abs	3.67	free
10	0.005	abs	0.006	abs	3.62	free
11	0.002	abs	0.005	abs	3.43	free
12	0.003	abs	0.0025	abs	3.40	free
13	0.002	abs	0.002	abs	3.38	free

Figure f: The sample concentrations window. CAT creates this window from the headers of the loaded spectrum files. The files are listed according to their input sequence. In the given case, the spectroscopic signal is generated by the metal ion UO_2^{2+}. Hence, the first spectrum holds only uncoordinated UO_2^{2+}(aq). The UV-Vis absorption spectrum of this species is therefore known in the subsequent analysis.

The pH values are transformed into OH- concentrations inside the program using the information about ionic strength and the Davies equation. Therefore, in the spectrum headers, the user has to provide the ionic strength following the component name "pH"; e.g. '...,pH, 0.25'. If the equilibria under study depend on pH, the component pH should always given last. Figure e provides an example. The flag F informs CAT about the character of the data. The Sample Concentration window shows that the pH is understood as free concentrations.

7 Basic Procedures

TB CAT consists of a main window with seven principal menu items:

a) File

b) Do

c) Uncertainty

d) Data

e) Evaluate

f) About

g) Windows

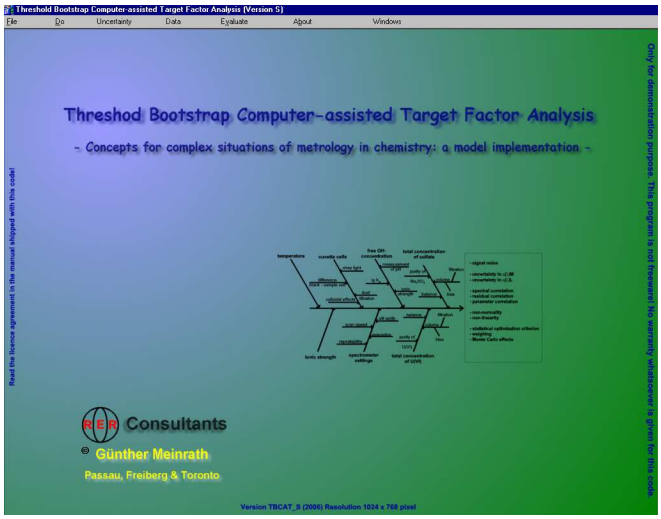


Figure g: The CAT main window.

a) File

The menu item FILE has the sub items

(1)open
(2)save
(3)save as (inactive)
(4)close
(5)exit

(1) Open : load the data files composing the matrix of experimental data

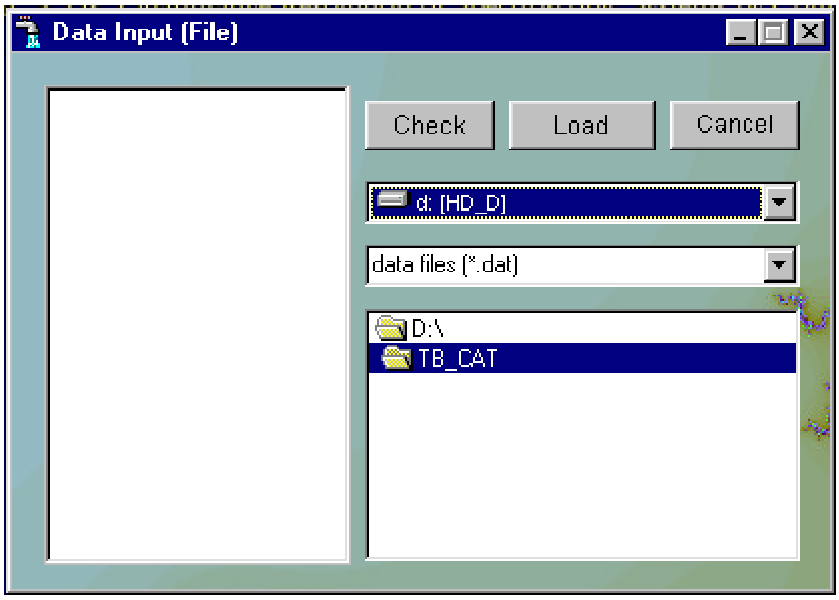


Figure h: Data Input screen of the **File/Open** menu item

In the left field, the files with extension specified by the user appear. The user may select the files which should have been adequately named before starting TB CAT. For raw data files, the extensions *.dat or *.asc are recommended.

Each file must follow the data file format convention explained in the previous §.

After selecting and before loading the files, the Check button allows to read the first line of each file. Thus the user can assure that the files do have the same wavelength range.

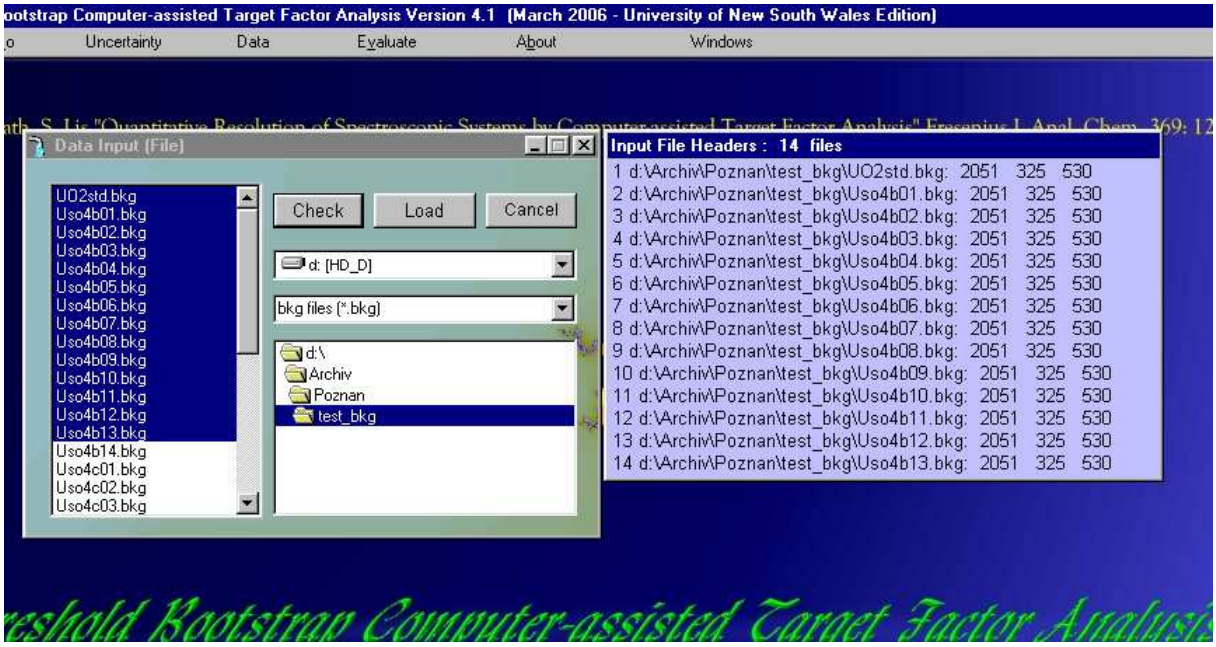


Figure i: Data Input Window with the File Header Checking Window

Upon loading, CAT performs a series of tests ensuring that the dataUpon loading, CAT performs a series of tests ensuring that the data format is appropriate for the subsequent operations. The spectral information is then displayed. This operations may take some time depending on the total amount of data.

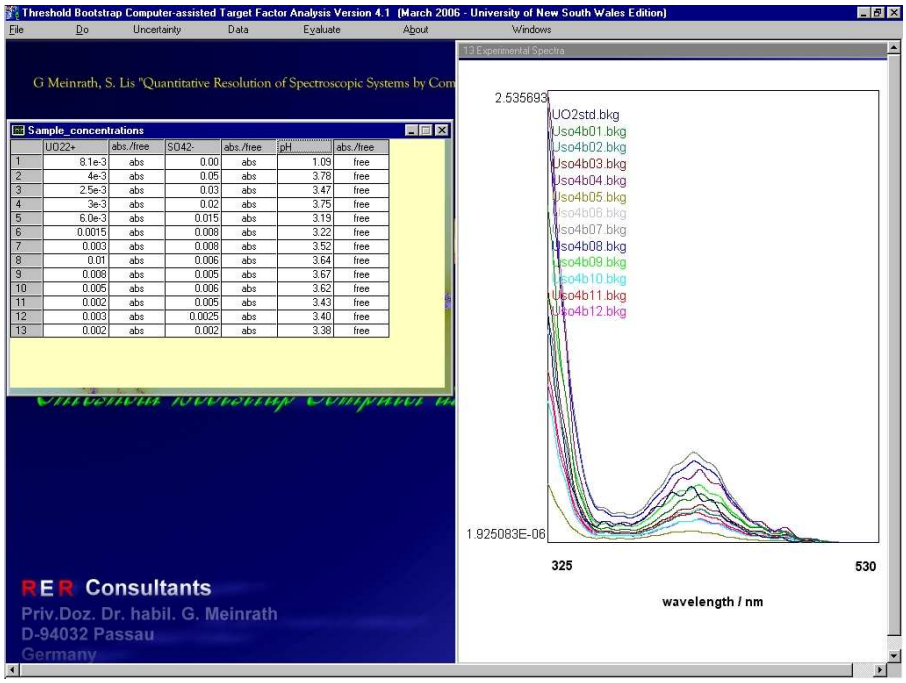


Figure j: Diagram with loaded spectral information. The position of the cursor inside the diagram is given interactively below the x-axis caption.

(2) Save

Information in the active window can be saved if the relevant information is available. Upon selecting **Save** a dialog window opens. Data in graphic windows is saved as ASCII data of the information displayed in the window. There is no way to save graphics data as a graphics format (e.g. JPG). The default extension is *.lis. The user may select own extensions.

(3) Save as

not implemented

(4) Close

not implemented

(5) Exit

Immediately shuts down TB CAT without further notice. No information will be saved.

b) Do

The DO menu is the working center of the TB CAT code. It offers the following items:

(1)Background correction
(2)analyze
(3)CAT
(4)molar absorption
(5)*TB CAT*

(1) Baseline correction

Raw spectra from a UV-Vis spectrometer usually require to be normalised to a common background. Occasionally, it is also required to limit the spectral information to a more narrow wavelength range. For this procedure, the user selects the range to the left and right of the absorption band representing the baseline. TB CAT uses a linear least squares regression fit of a straight line through the specified intervals to determine the baseline for each spectral file. With this information the files are corrected to a common baseline. The result will be displayed in the graphics window immediately. This process can be applied to a subsection of the loaded spectra. The user can select individual files in the *Select Files* window.

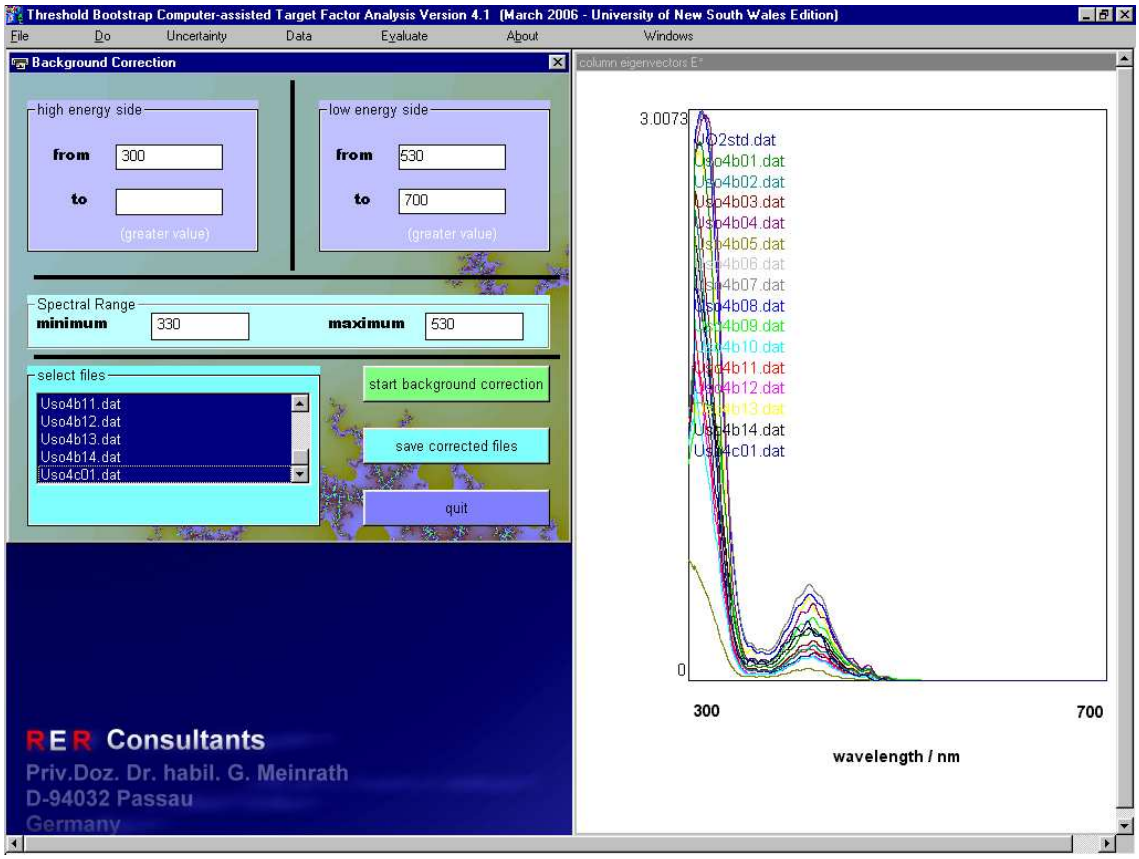


Figure k: Background correction window

Upon loading CAT sorts the imported data in ascending order of the dependent variable. It is hence possible to import data with disordered wavelength values or to import data ordered in descending order.

Specifying the baseline for least-squares correction

The both top windows allow specification of a baseline range at the high energy and low energy side. The left side is the low energy side, where the high wavelength appear. The right side is the high energy side where the low wavelength appear. It is for the sake of this procedure that the user should keep with the data input file convention. If this convention is reversed, the settings in this window must be reversed, too.

The shortest wavelength appear in the bottom field of the right top frame while the longest wavelength appear in the top field of the left top frame. The user is required to modify this position if the side should be used for baseline correction. If no modification is made and the empty field in the frame remains empty, CAT will not use this side in the regression leading to the definition of a baseline. This gives the user the possibility to exclude one side from the baseline calculation, e.g. if such a baseline is not available due to spectroscopic properties (continuous absorption to the UV) or if a baseline hasn't been recorded. The process of baseline correction can be repeated unlimited.

It is recommended to use the cursor position indicator fields of the graphical display for locating the baseline range to be applied in the baseline correction procedure.

Starting baseline correction and saving corrected file

The green button "start baseline correction" atarts the procedure. The button "save corrected files" opens a dialog window. The files will be saved by default with the same filename as the input files but an extention *.bkg. The user can also create a different directory clicking the "new dir" button. If a directory already holds files with the same file name, the user is asked to confirm the deletion of the old files.

(2) Analyze

The **analyze** item opens the Analyze Mode Input Form. It controls the determination of both the diagonal values of the target transformation matrix and the best starting values for the CAT routine.

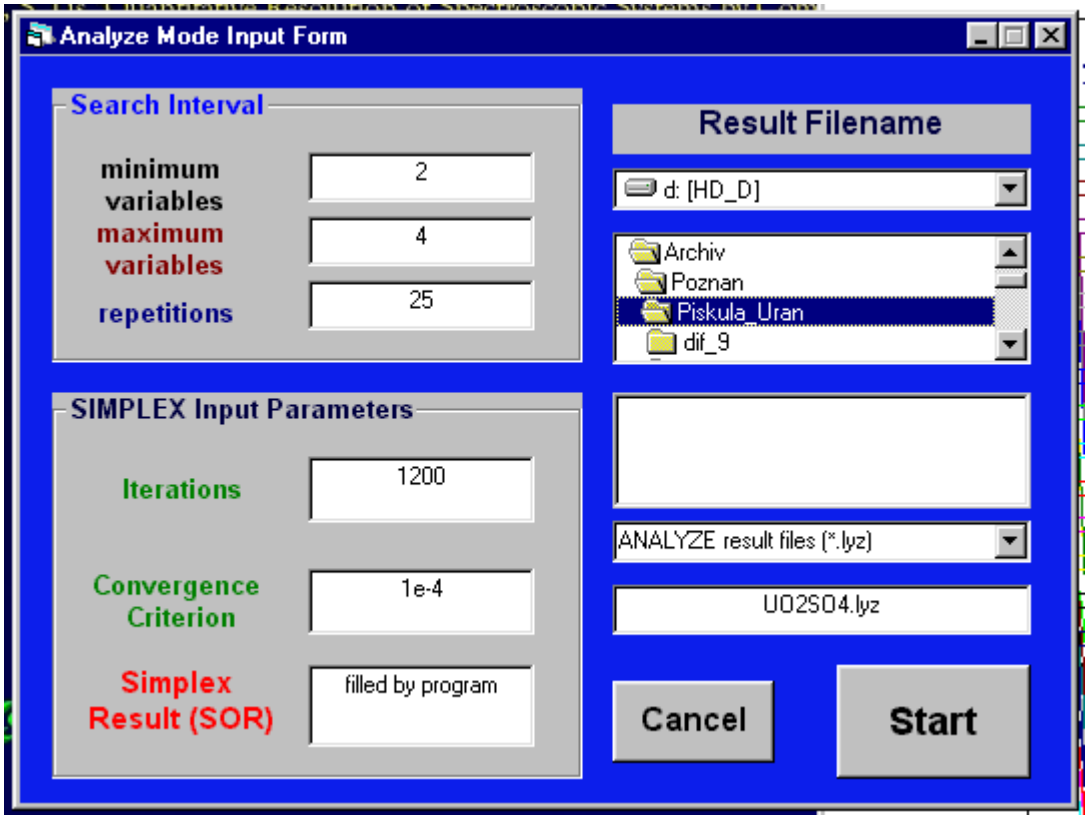


Figure l: Analyze Mode Input Form

Specifying the upper and lower limits of the rank and the repetition of the search

The left side frames *Search Interval* and *SIMPLEX Input Parameters* accept user input to explore the basic properties of the data matrix. The *minimum variables* field accepts an integer value for the lowest possible rank of the matrix, while the *maximum variables* field accepts an integer value for the upper limit of the true rank. The field *repetitions* specifies how often CAT creates a set of random starting values and repeats the interpretation of the matrix on assumption of a given rank and a given set of diagonal values +1 or -1. In fact, CAT repeats the interpretation for each possible permutation of diagonal elements. Each combination will be tested as often as specified under *repetitions*. Hence, if the minimum rank is 2 and the maximum rank is 4, CAT performs 100 tests if each permutation is tested 10 times.

SIMPLEX Input Parameters

The SIMPLEX algorithm is not as popular as it should be. The SIMPLEX algorithm is an efficient method to fit models to data which does not require derivatives. The concept of SIMPLEX is easy to grasp with a more detailed explanation. Such explanations are available in literature. There is no need to repeat such an explanation here. A SIMPLEX can be understood as a mesh of parameter values that moves over the parameter space according to detailed instructions. Its direction of movement is the minimum of the parameter surface where the surface is defined by the sum of negative values in E and C (L_1 criterion) or the sum of squared negative values in E and C (L_2 criterion). At present, only the L_2 criterion is implemented (L_2 is more commonly tremed 'least sum of squared residuals'). There are two conditions when the SIMPLEX algorithm stops. First conditions is fulfilled if the number of iterations specified in the field **Iterations** is satisfied. Second condition occurs if the convergence criterion is satisfied. The convergence criterion is satisfied if the difference in the mutual distances of the parameter values forming the SIMPLEX mesh is below a given critical value. The critical value is set in the **Convergence Criterion** field.

Rules of thumb for setting SIMPLEX parameters

In fact it is not necesssary to have a detailed understanding of the SIMPLEX to use TB CAT. Some experimentation is more teaching than theory. As a rule of thumb, 1000 iterations and a convergence criterion of 1 E-4 is adequate for most situations.

Result filename

The Analyze process creates a table of infomations which preferably should be saved for later inspection. The default extention for these files is *.lyz. A filename must be specified and confirmed by clicking the Return key. Then the **Start** button will become active and the Analyze procedure can be started.

CAT will calculate the possible permutations in the diagonal elements of T and the total number of runs given the minimum and maximum rank to be tested. The sum-of-residuals (SOR) indicate how well the single component spectra obtained from the experimental spectra under a certain hypothesis on the matrix rank (= number of species in solution) and the diagonal elements in T are able to explain the experimental spectra.

The Analyze procedure windows shows all central features of the TB CAT window. The yellow top window displays information about the SIMPLEX procedure. It shows the wideest and narrowest mesh of the SIMPLEX. Of central importance are the values listed under convergence criterion. The SIMPLEX has converged if all figures are equal or below 1.0. The second essential element is the lowest right-hand side figure. It gives the sum of squared negative values in E and C (as will be explained later, a weighing factor is included).

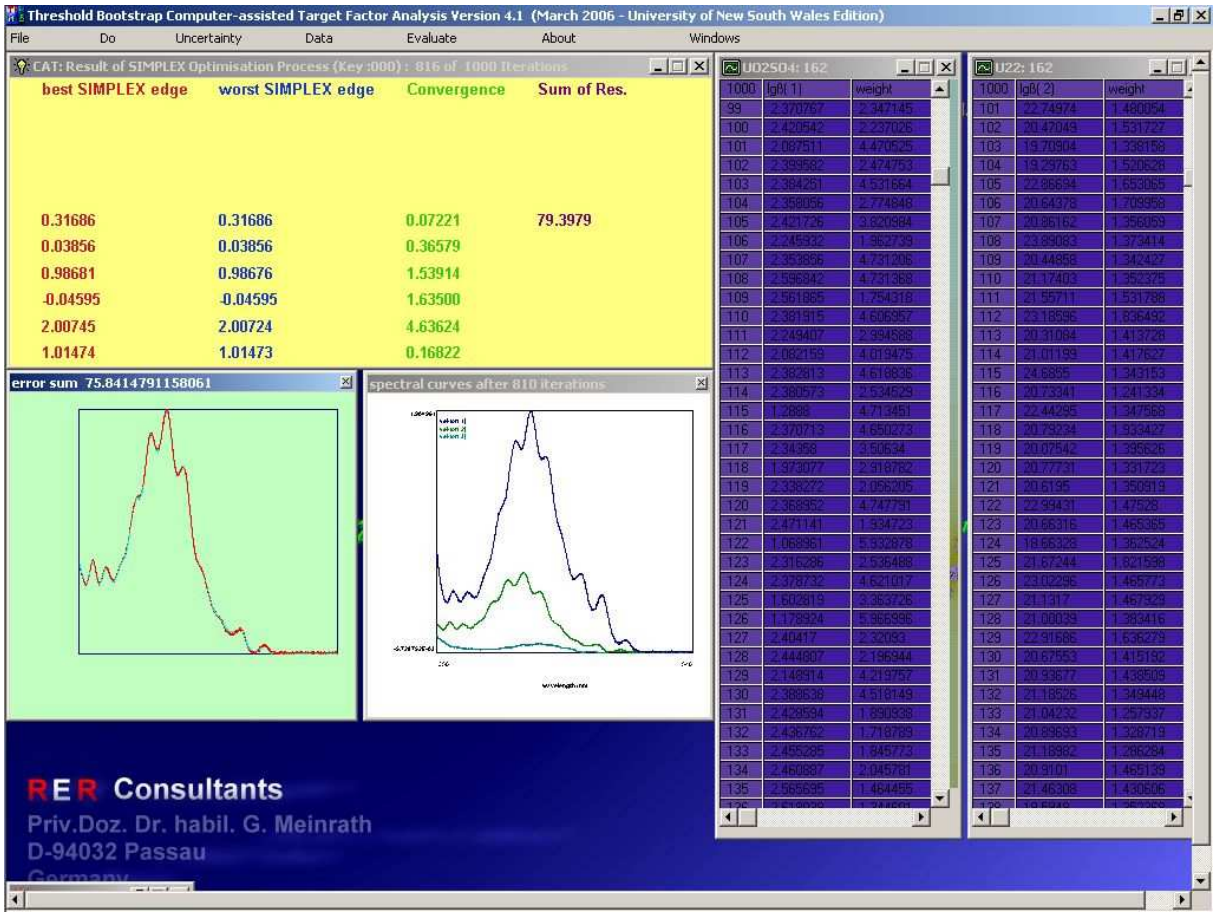


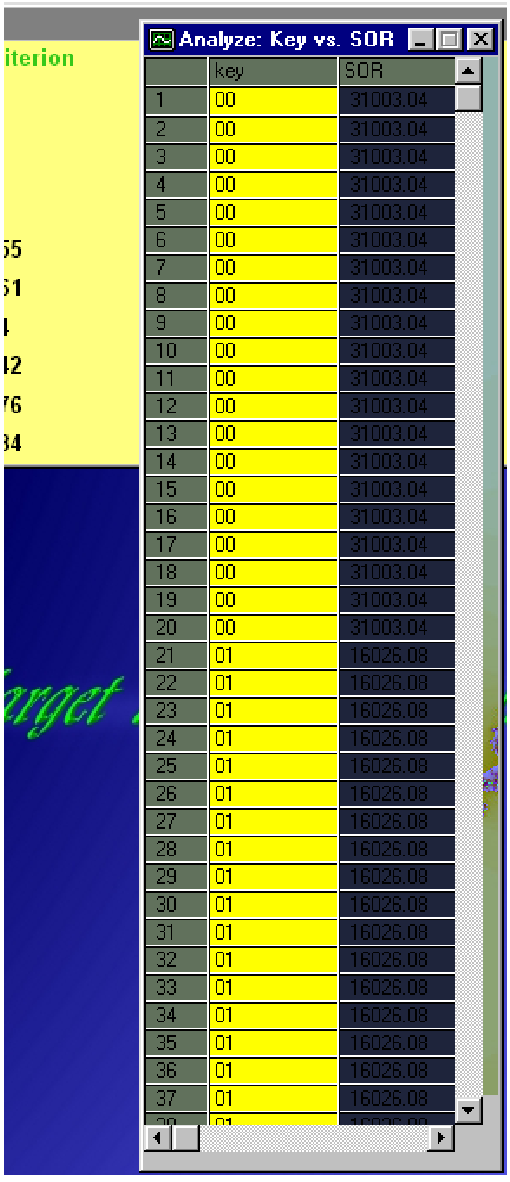
Figure m: The Analyze procedure windows

Graphics windows

Two graphics windows are on display during the ANALYZE procedure. These windows are also available during the CAT and the TB CAT step. The light green left-hand-side window compares the estimate of the first calculated spectral component with the experimental spectrum of the known component (note that CAT assumes this spectrum to be loaded in first position). Inside the algorithm, the spectrum of the known component and the first estimated component spectrum are normalized. Therefore, only the shape of the spectral information is of interest. The magnitude doesn't play a role at this step of the procedure. The experimental spectral curve is given in cyan, while the calculated spectrum is given in red colour. CAT tries to find a spectrum that optimally fits the cyan spectrum. The sum of squared differences between the both spectral curves are one of three components summing to the total sum of residuals (SOR). This value is specified in the header of the spectrum graphics window.

The white background right-hand-side graphics window shows the current best estimates for the single compnent spectral curves. Because the diagonal elements of T are fixed, the relative magnitudes of these spectra are irrelevant – only the spectral shape is of importance here. Arranging rapidly changing windows on screen can be tiresome. The both graphics windows

are coupled. The left window will always assume the same size as the right window. While the right window can be moved freely, the left window will always stay with the left border of the TB CAT main window and the lower border of the SIMPLEX window. If the right window is enlarged by dragging at the borders of the diagram, the left window will follow.



Analyze window (Key vs. SOR)

Central element of the Analyze process is the Analyze table. The left column gives the diagonal elements of the T matrix. A 1 represents a value +1 while a 0 represents a value -1. Because the value of the first element is given by the values of the first abstract factor, its value will not be permuted (this factor has either only positive or only negative values) and is always given as a zero in the keys. Internally, its value is taken into account by CAT. The sequence of 1's and 0's is termed a key. In the left uppermost field the total number of calculations is displayed. This figure allows a rough estimation of time consumption until completion of the total ANALYZE procedure.

During the Analyze procedure, the Analyze window will be filled with the key values and the respective SOR. Each key will be tested as often as specified by the user in the 'Repetitions' field of Analyze Mode Input Form. After termination of the Analyze procedure, the data list in the Analyze window will be sorted with increasing SOR. Hence, the smallest SOR representing the best fit will be displayed on top.

Figure n: The ANALYZE table

Creating default input files

Upon clicking a field in the SOR column of the Analyze window after termination of the procedure, a dialog box opens asking whether the user wants the program to generate a default input file with the relevant informations for subsequent program steps. The default input file will be saved as 'defaultxxx.elb'. (The extention *.elb honours the deceased former head of the Rare Earths Department at Faculty of Chemistry of Adam Mickiewicz University at Poznan/Poland, Prof. Marian Elbanowski). The 'xxx' stands for the key. Hence, it is possible to create an *.elb file with the best fitting input data sets of each possible key. Note

that the lenght of the key is not limited to three. Longer and shorter keys are also acceptable. The *.elb files are ASCII files.
An example of a default000.elb file is given below, which holds the following informations:

```
"key","000"
"niter",400
.269645854203445,2.69645854203445E-02,.0001
.213553310743779,2.13553310743779E-02,.0001
.987070096896006,9.87070096896006E-02,.0001
.335093279935166,3.35093279935166E-02,.0001
3.41121586767416,.341121586767416,.0001
7.88138696376839E-02,7.88138696376839E-03,.0001
```

The first line gives the key word "key", followed by the key in string format. The second line holds the key word "niter" which specifies the total number of iterations. The total number of iterations is fixed to 400. For the time being, modifications in total number of iterations must be made either by manipulating the ASCII file outside CAT or is handled within the CAT code.

The following lines define the parameters of the SIMPLEX. Because the key is currently 000, indicating diagonal values of the target rotation matrix to be -1, -1, -1, the dimension of the target matrix is 3 x 3. The 3 diagonal elements are set to -1, -1, -1. Hence, six variable matrix elements remain. The values of these matrix elements are given in the first column. SIMPLEX tries to optimise these elements further by creating a seven-dimensional polygon around the starting values. The polygon is built using the informations in the second column. It is easy to see that the values in the second column are just 10% of the first column. The last column specifies the stoppig condition of the SIMPLEX iterations. The stopping condition is reached if the seven-dimensional SIMPLEX net has contracted in that way, that the largest difference is 0.0001 times the smallest value in that dimension. For those unfamiliar with the SIMPLEX algorithm, the following readings by M.S. Caceci are recommended.

In fact, the user is not obliged to be familiar with the SIMPLEX because the 'Analyze' procedure generates the necessary files. The user has to specify the range of ranks he wants to investigate and to specify the number of repetitions. Because starting values are generated by a random procedure, each run has some different starting values. This random feature tries to overcome local minima, which would mislead the procedure. A minimum number of 10 repetitions is therefore recommended.

It is possible to create default input files with different keys. For example, it is recommended to select the best result for each key within the first 5% to 10% of total results generated in the 'key vs. SOR' list. If for one key several default files should be created, the already existing default file with this key must be renamed. Otherwise, the existing key will be overwritten witout any further notice. Note that default files should be stored in the same directory as the spectra files.

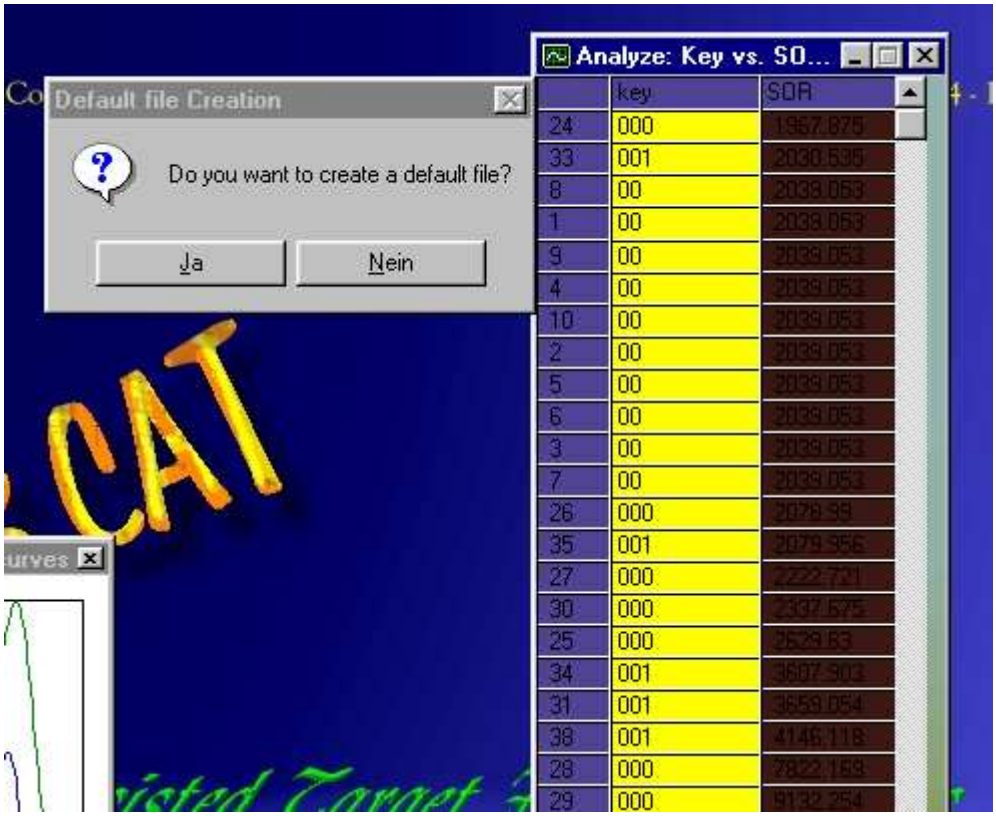


Figure o: After finishing Analyze, the data list is sorted with the optimal combination of starting values and diagonal elements on top of the list. Upon clicking in a field, a saving dialog opens allowing to create a default file with parameters of the selected run.

ANALYZE procedure returns informations about the optimum settings of the CAT and TB CAT procedures. It ensures that all possible interpretations of the chemical system under study are considered. The user gets a first impression of the system. After finishing the ANAYZE procedure, the user holds the most relevant interpretations in convenient default files allowing systematic and detailed quantitative evaluation of the system.

(3) CAT

CAT is a mean square principal component analysis procedure. It returns best fit values for sample composition, single component spectra and formation quotients.

CAT proceeds in two steps. In the first step, the abstract factor analysis is performed. For this purpose, CAT needs the informations which have previously been collected into the default input files, defaultxxx.elb. By filling the fields 'Starting Values', 'Step Width' and 'Convergence', the user is free to enter values by hand – a tiresome procedure.

Figure p: CAT input window. The default file default000.elb has already been loaded and the number of iterations is set to 1000. In the next step, the concentration information file example.dat will be loaded. The values from the default000_D.elb file have been copied to the lower left hand side fields with caption 'Starting Values', 'Step Width' and 'Concergence'. These fields can be manipulated by the user.

The available default files are specified in the directory list box bottom right of the data input window. Loading a default file and specifying the number of iterations is sufficient to start the optimization procedure.

By clicking the respective "run CAT '!'" button, CAT is searching for the rotation matrix values t_{ij} and displays the results. The result may look as in fig. q.

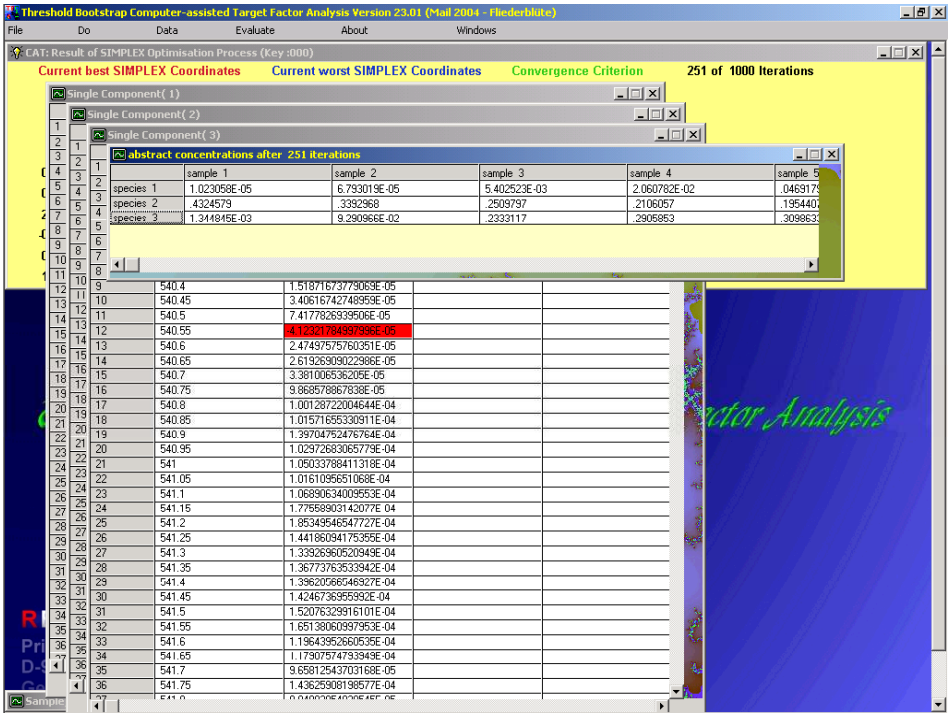


Figure q: Result of the CAT procedure. The abstract concentrations and the normalised single component spectra are given.

The estimated single component spectra are given numerically together with the concentration estimates. The user has the possibility to study these results which already represent the final spectral curves. But at this stage, there is no chemical interpretation because the rotation matrix T is normalised by the diagonal elements -1 or 1. It is necessary to transfer the normalised curves and concentrations into concentrations and molar absorptions consistent with the chemical composition of the samples. This is done by the 'Molar Absorption' menu item. In most cases the user will therefore will proceed with the next menu item - the 'Molar Absorption' step.

(4) Molar Absorption

The Molar Absorption menu item opens with a window collecting basic chemical informations about the system under study in the upper section and returning calculated information in the lower section.

CAT needs informations on the chemical system under study, e.g. the probable composition of the species formed, their names (even though 'species 1' and 'metal ion' are also acceptable).

The component names must already be given in the headers of the spectrum files. CAT automatically puts these names under 'components names'. 'Species identification' allows the user to change the default names 'species 1', 'metal ion', 'species 2', ... into meaningful names which allow an easier identification for the given system. The 'stoichiometry' section is the most essential input.

In the 'stoichiometry' section the user must give the chemical composition of the species. If component 1 is the metal ion M and component 2 is a ligand, say X, then a species MX is represented by the stoichiometric coefficient '11'. The metal ion alone is represented by the coefficient '10'. Note that the left hand field under 'stoichiometry' identifies the species, while the position of a stoichiometry coefficient represents a component. The sequence of stoichiometric coefficients must comply with the sequence of components given in the spectrum file headers – to be found under 'Chemical Informations - a) component names'.

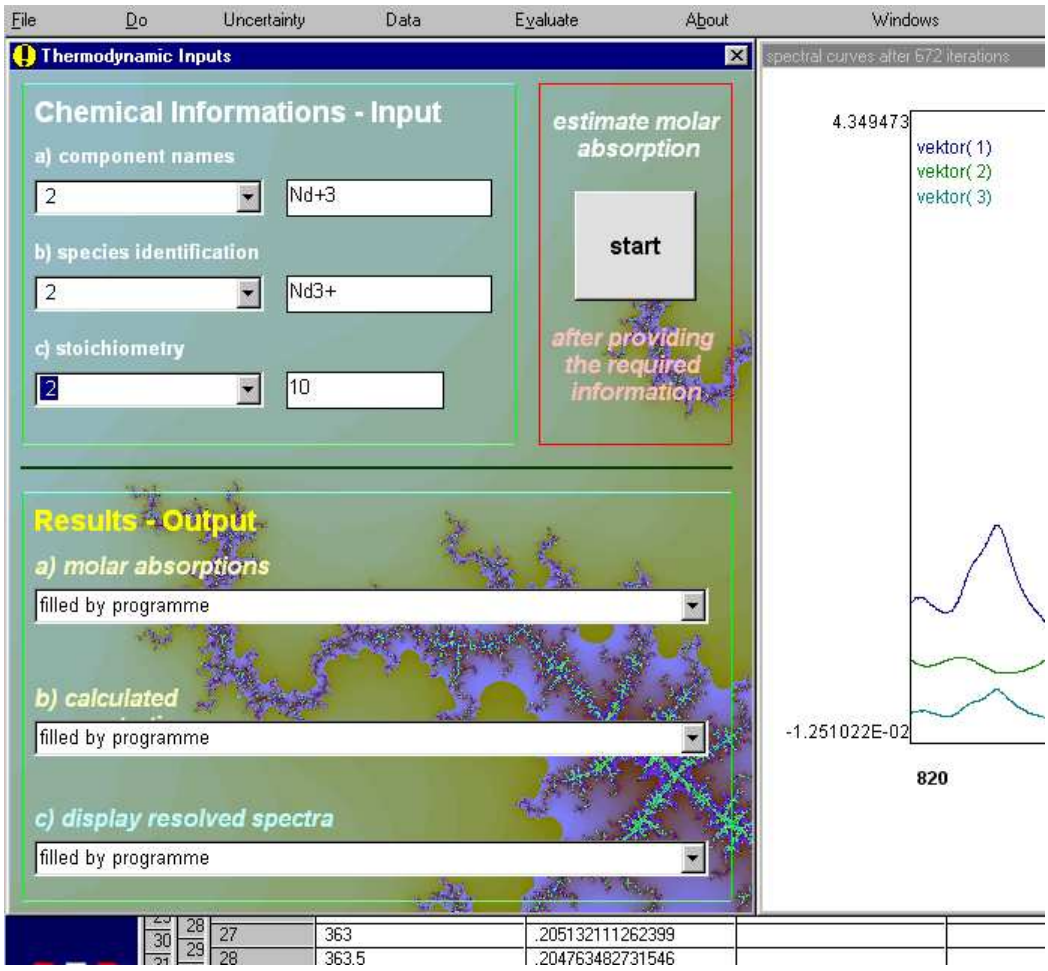
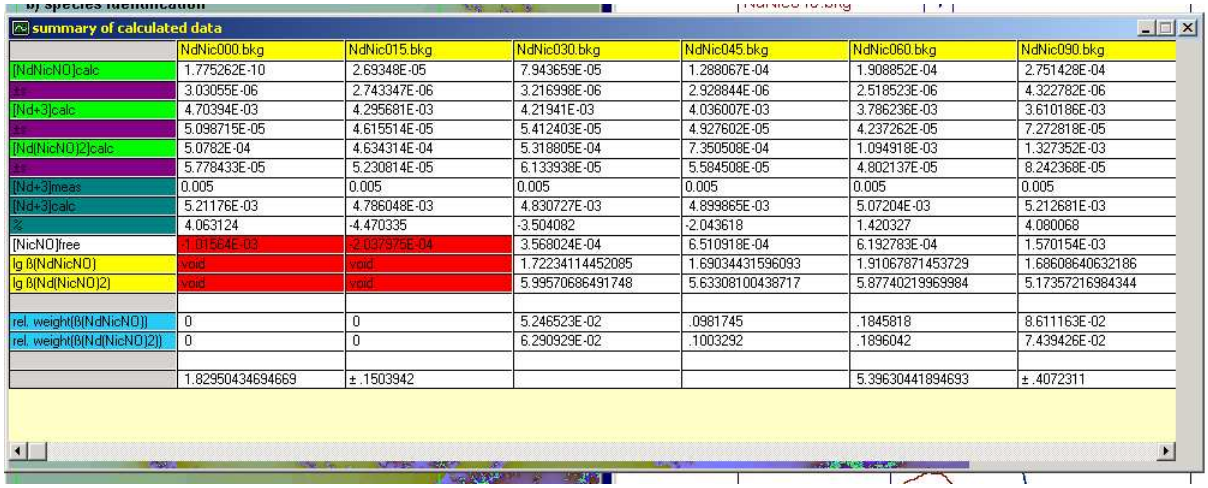


Figure r: The Basic Thermodynamic Information window of the Molar Absorption menu item. Note that the species where the single component spectrum is known (here Nd^{3+}) (cf. the 'species identification' field), always appears in second place!

It is important to give stoichiometric coefficients for all components, even if the species is not holding a component. Having a system with components Fe3+, Cl- and pH (for OH-), a species $Fe(OH)_2^+$ is characterised by the stiochiometric coefficients '102', while $FeCl^{2+}$ is characterised by '110'. If only '11' is given for $FeCl^{2+}$, CAT will crash quickly.

Upon clicking the right hand side 'GO' button, CAT searches for the best fitting agreement between the metal ion concentration in the species and the total metal ion concentration. Free ligand concentrations are calculated from difference (if absolute concentrations have been given) or takes the free concentrations to calculate formation quotients for each solution. If negative concentrations follow, the calcuation is stopped. Sample solutions with negative concentrations will be highlighted in red colour as shown below.



summary of calculated data

	NdNic000.bkg	NdNic015.bkg	NdNic030.bkg	NdNic045.bkg	NdNic060.bkg	NdNic090.bkg
[NdNicNO]calc	1.775262E-10	2.69348E-05	7.943659E-05	1.288067E-04	1.908852E-04	2.751428E-04
±	3.03055E-06	2.743347E-06	3.216998E-06	2.928844E-06	2.518523E-06	4.322782E-06
[Nd+3]calc	4.70394E-03	4.295681E-03	4.21941E-03	4.036007E-03	3.786236E-03	3.610186E-03
±	5.098715E-05	4.615514E-05	5.412403E-05	4.927602E-05	4.237262E-05	7.272818E-05
[Nd(NicNO)2]calc	5.0782E-04	4.634314E-04	5.318805E-04	7.350508E-04	1.094918E-03	1.327352E-03
±	5.778433E-05	5.230814E-05	6.133938E-05	5.584508E-05	4.802137E-05	8.242368E-05
[Nd+3]meas	0.005	0.005	0.005	0.005	0.005	0.005
[Nd+3]calc	5.21176E-03	4.786048E-03	4.830727E-03	4.899865E-03	5.07204E-03	5.212681E-03
%	4.063124	-4.470335	-3.504082	-2.043618	1.420327	4.080068
[NicNO]free	-1.01564E-03	-2.037975E-04	3.568024E-04	6.510918E-04	6.192783E-04	1.570154E-03
lg ß(NdNicNO)	void	void	1.72234114452085	1.69034431596093	1.91067871453729	1.68608640632186
lg ß(Nd(NicNO)2)	void	void	5.99570686491748	5.63308100438717	5.87740219969984	5.17357216984344
rel. weight(ß(NdNicNO))	0	0	5.246523E-02	.0981745	.1845818	8.611163E-02
rel. weight(ß(Nd(NicNO)2))	0	0	6.290929E-02	.1003292	.1896042	7.439426E-02
	1.82950434694669	± .1503942			5.39630441894693	± .4072311

Figure s: The 'summary of calculated data' window. The system has been interpreted by two cooridnated species, termed $NdNicNO^{2+}$ and $Nd(NicNO)_2^+$. The first solution holds only Nd^{3+}, hence no formation quotient has been calculated. For the second solution, a negative free ligand concentration, [NicNO]free, is obtained. Hence, the field is highlighted red and the formation quotient holds the word 'void'. The violet fields hold uncertainty estimates on basis of Clifford's method (see References). These uncertainties represent misfit and are solely used to assign a weight to each calculated quantity. The weights of all components of a formation quotient are multiplied. After all weights are calculated, the weights of all non-void samples are normalised and given in the row 'rel. weight' for each species. Given the weights and the formation quotients in each sample solution the formation quotients in the lowest row are calculated together with an uncertainty estimate.

The 'Summary of Calculated Data' window can be saved. If the window has the focus, it is the format to be saved if 'Save' is selected in the 'File' menue. The data in the file are put in an ASCII file which can be evaluated externally.

Further informations may be obtained from the 'Basic Thermodynamic Information' window. Here the 'Results' section is now filled with additional information. The molar absorptions are given in the top section. The middle section allows to inspect the agreement between

specified total metal ion concentration and the calculated sum of metal ion. The difference is given in per cent. In the bottom section, the user can select a spectrum, where experimental spectrum, fitted total spectrum and the contributions of the species as calculated from the given model are specified. This spectrum appears in a new graphics window.

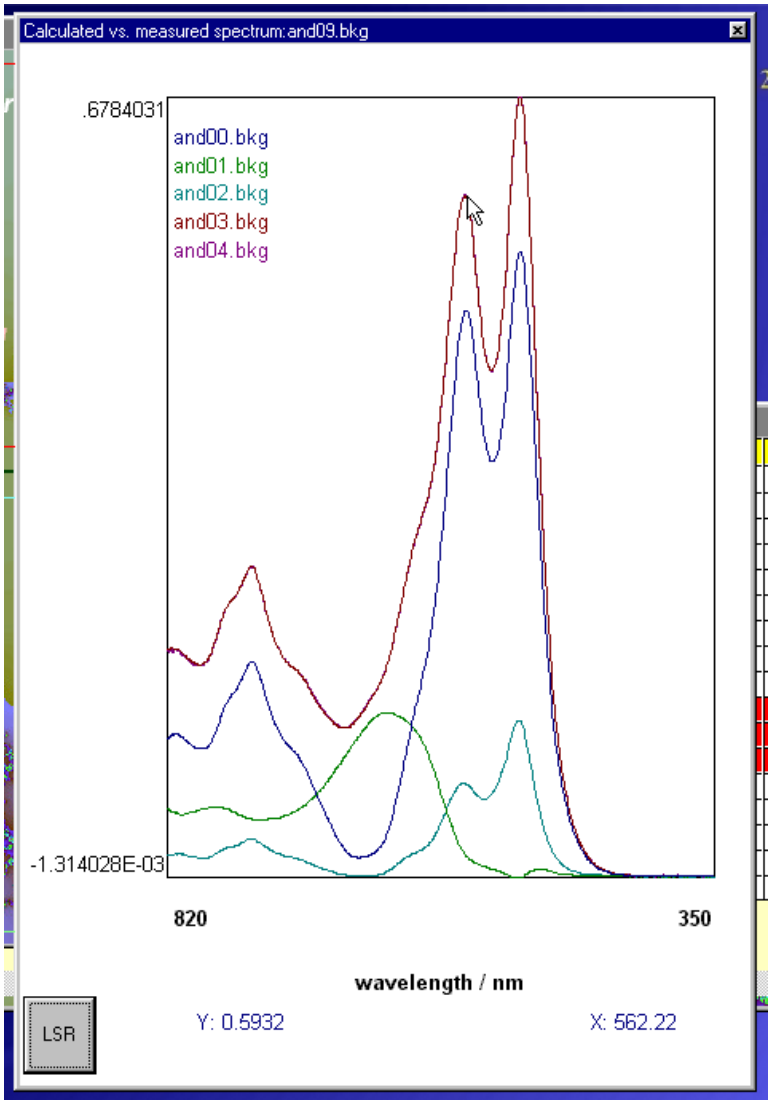


Figure t: Example of an interpreted spectrum and its single species contributions – available from the 'display resolved spectra' field (cf. figure r: under 'Results – Output'). Pointing with the mouse into the diagram allows to read the spectroscopic informations: here the absorption peak is at 562.2 nm with an absorption of 0.593. The button LSR at the lower left hand side allows to display a least squares interpretation. The underlying algorithm is QR decomposition. The calculations may become quite time-consuming but provides the variance-covariance matrix.

The displayed spectrum can be saved in an ASCII file and loaded into an appropriate graphics program. Thus, all spetra can be transfered into a presentation graphics.

(5) TB CAT

The simple CAT procedure returns formation quotient(s) and single component spectra for the system under study. The constraints are the key (the specific permutation of the diagonal elements in the target transformation matrix), the single component spectrum of the free metal ion and the values for the composition of the sample solutions. CAT gives values but without information concerning the stability of the values. As long as these values are considered as results per se, there is no need to inquire into their reliability and stability. Without an understanding of reliability and stability of a result, no further conclusions and comparisons, e.g. with similar values from other sources or separate experimentation, should be made.

The menue items CAT and Molar Absorption evaluate mean value based data interpretations on the assumption that the informations about component concentrations are perfectly true. Hence, if a concentration is given as '0.0002 mol dm^{-3}, the algorithm assumes 0.00020000000... mol dm^{-3} of that component. However, can the user be sure that the concentration isn't, say, 0.0001998 mol dm^{-3}? Or 0.00020314 mol dm^{-3} ? There is almost no certainty. Each volume operation introduces some uncertainty. Mixing samples from several standard (standardised by what? or to what?) solution is prepared with a small but non-negligible uncertainty. Uncertainties accumulate. Further uncertainties, e.g. in spectral residual correlation, add into the procedure.
Threshold bootstrap computer-assisted target factor analysis tries to approach the problem of limited certainty of the knowledge about a system by computer-intensive resampling methods. In fact, TB CAT creates a large number of new input files with input quantities varying within specified limits from run to run. The residuals may vary, the input concentrations may vary, the volumes may vary.

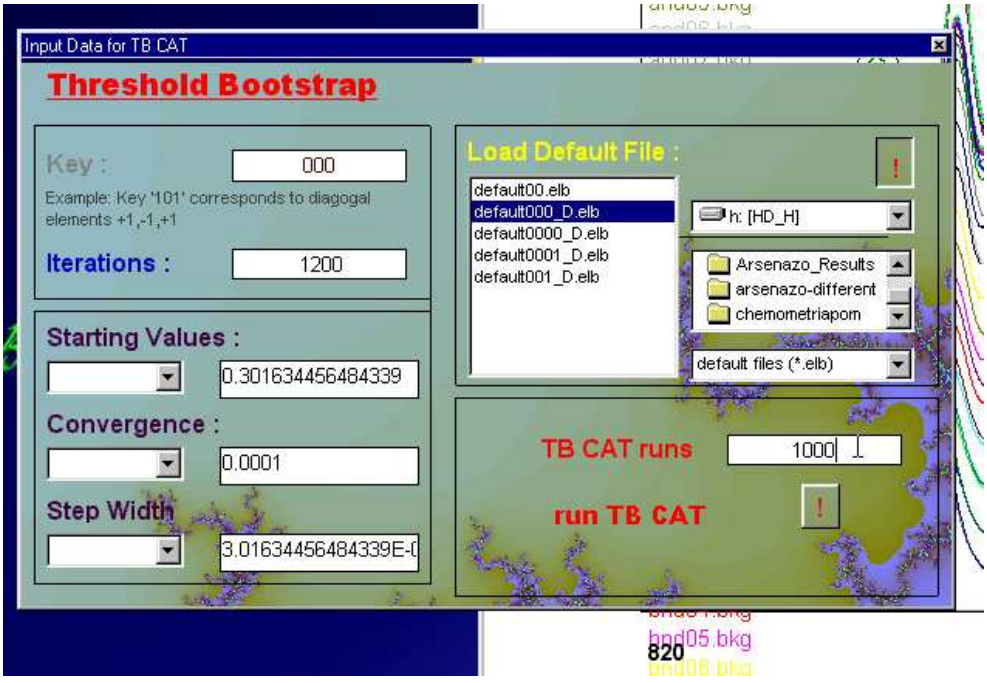


Figure u: The TB CAT input window. The window is the same as figure p, but now the 'TB CAT runs' field is enabled.

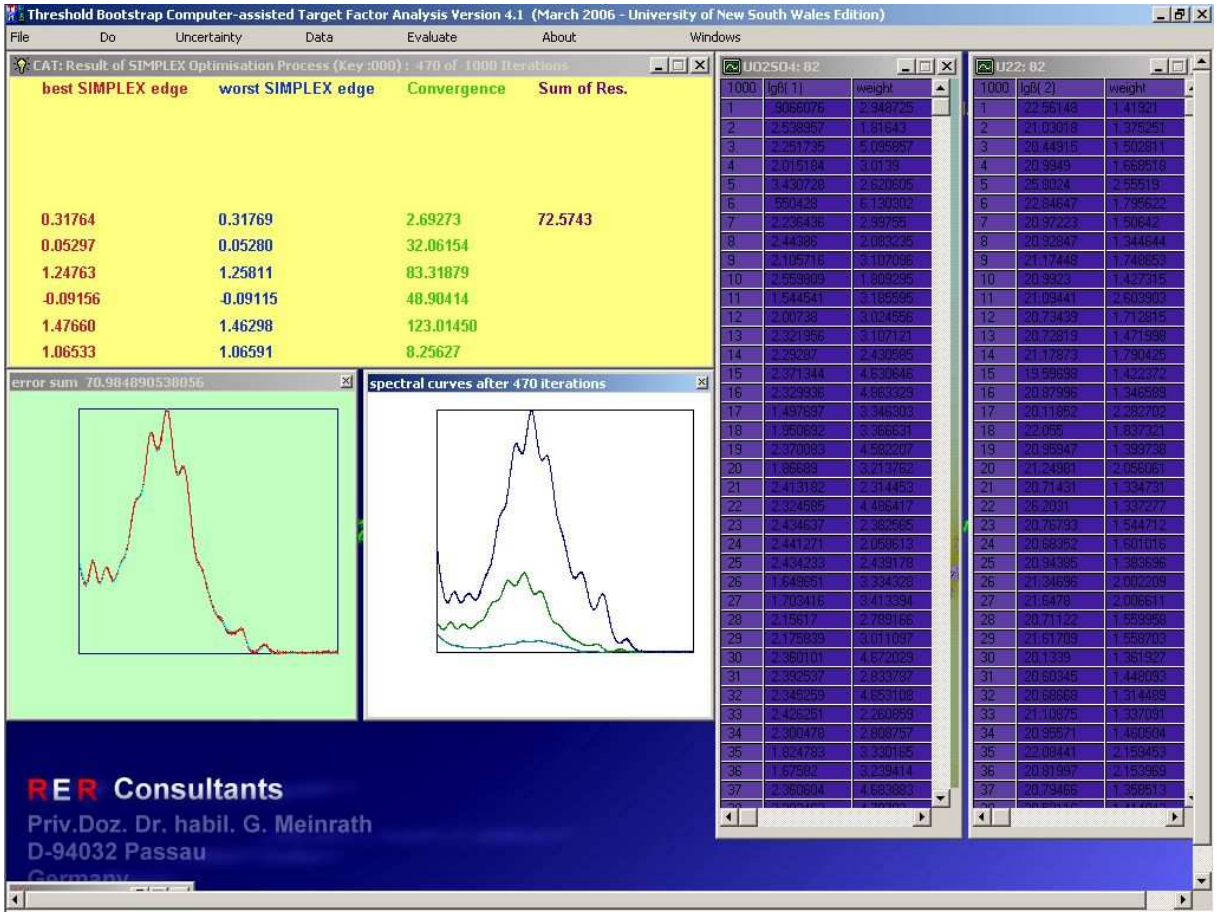


Figure v: TB CAT at work. The formation constants obtained in each TB CAT run are displayed in the two windows at right. CAT automatically generates the required number of result windows, with names taken from the species information given by the user. In the top left field, the total number of runs is given. The current status is shown in the header of each column. Upon termination, TB CAT sorts each column and writes the information to the disk using a naming convention specified below.

The user proceeds as if a CAT would be calculated. In the 'Input Data for CAT' window, the field 'TB CAT runs' is enabled. The user specifies the total replicate runs. All informations are requested by the program as has been shown in the previous section. TB CAT reruns the data the specified amounts of times (at least 1000 TB CAT runs should be performed). For each species, a table is created where the calculated formation constants and a weight factor are tabled. From these tables, probability densities and spectral uncertainties are derived.

TB CAT should procede without further need of user interaction. After having finished the calculations, the list of formation quotients will be sorted and the weights will be transformed into probability densities. At the same time, TB CAT creates a large number of spectral data files in the directory from which the experimental spectra have been loaded. The name convention is as follows:

AAAANNN_B.iii

AAAA: name of the species (from previous user input in the Thermodynamic Data Window)
NNN : key (may have more or less than three positions)
B : number of species (pure spectral component is always no. 2)
iii : no. of TB CAT run

example: "metal ion000_2.123"

Furthermore, a file 'cdf*.dat' is created holding the cumulative probability density for each formation quotient. Example: cdf_species(2)000.dat

From these files, the user may create various probability densities as described in the section 'Evaluate'.

c) Uncertainty

This menu item allows the user to communicate the uncertainties to be associated with relevant influence quantities to the TB CAT routine. This menu item has two fixed elements:

(1) Load/Save

For the convenience of the user, a ASCII file may be either created or loaded holding the relevant values for the uncertainties to be associated with the components.

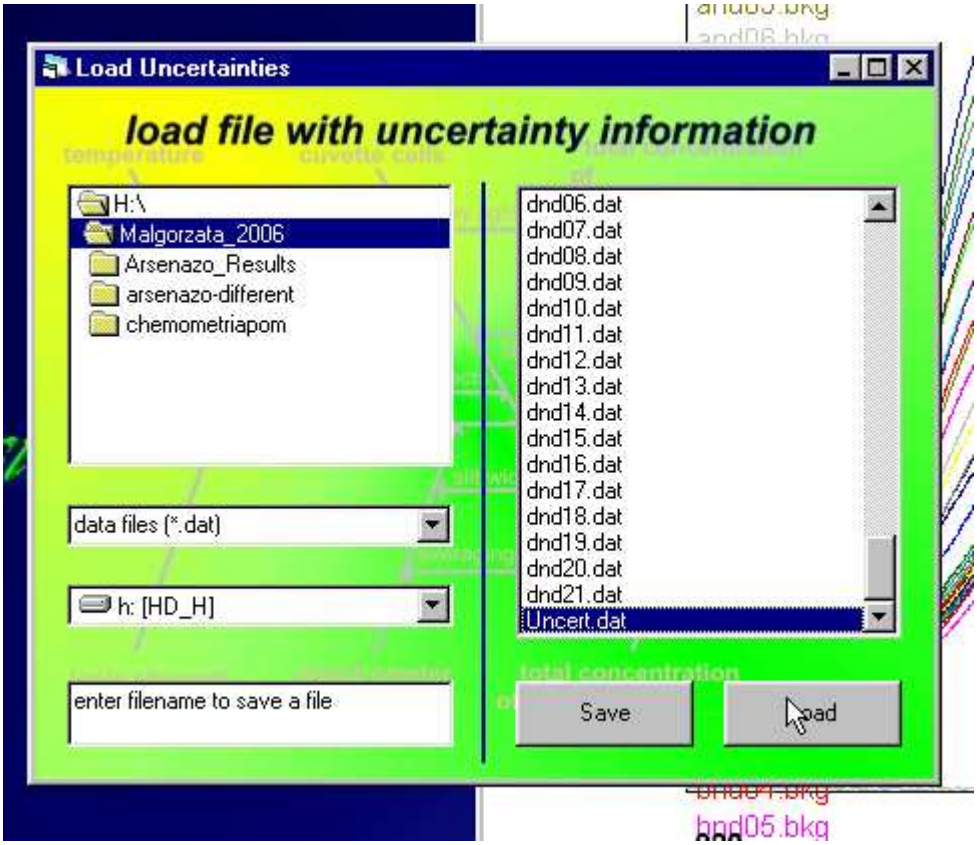


Figure w: Window for loading and saving information on measurement uncertainty.

(2) Repeatability

Upon clicking on this item, an input window will open allowing to specify the repeatability. Repeatability tries to account for random variations in recording spectral files. It is a common observation that the spectral curve of a sample is quite precisely reproducible, but the maximum of the spectrum may show some variability.

The other menu items are created during run-time on basis of the informations given in the file headers of the spectrum files. The component names will appear in the menu. Upon clicking a component name, a window opens allowing to specify uncertainties for volume operation (pipetting), balance (refering to the certainty that the weighted amount also arrives in the sample cuvette) and purity. Note that the uncertainties in these quantities, not the quantity itself must be specified. The purity of common laboratory chemical can be nominally 99.5%. After some time, it is reasoanble to assume an uncertainty of 0.5% to 1% in this value.

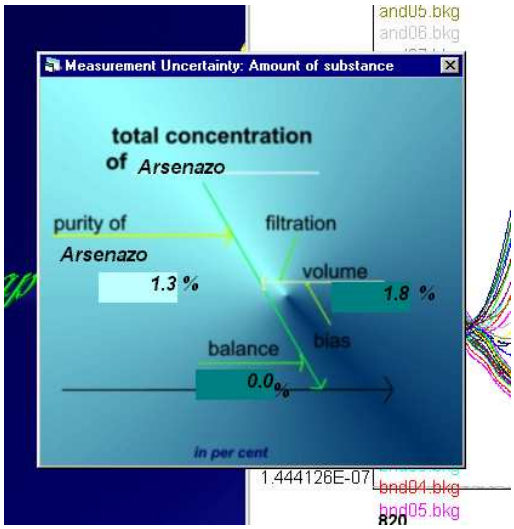


Figure x: An input window for the measurement uncertainty. All information are transfered to the code by upon closing the window.

If a component is pH, a special window opens, allowing to state the gross uncertainty in pH measurement. It is not possible to state the measurement uncertainty for each individual sample.

If all values are given, these may be saved into an ASCII file (see above). TB CAT has default values. Information given in the Uncertainty menu is only used during TB CAT procedures.

(d) Data

Under this menu item those data are listed that can be scrutinized by the user. This includes mainly the information about component concentrations in the different sample solutions and the spectral input data.

Opening the 'sample concentrations' wondow during a TB CAT analysis allows to follow the modifications made by the TB CAT algorithm on the sample concentrations on basis of the uncertainty information provided by the user under the 'Uncertainty' menu item.

(e) Evaluate

The menue item 'Evaluate' has possible selections 'Penalties', 'list Eigenvalues', 'plot Eigenvectors', 'Spectral Uncertainty' and 'Differentiate'

(1) Penalties

In searching the best target transformation matrix T, CAT relies on the common least-sum-of-squared-residuals criterion. However, there are several criteria which CAT must satisfy simultaneously. First, an optimum agreement between a known spectrum and a calculated spectrum is desirable. Second, the number of negative values in the estimated single component spectra and the species concentrations is strived for. Third, the difference between the measured spectra and those calculated from the numerical procedure should be a minimum. CAT and TB CAT balance these three components of the total sum-of-residuals value by three values, termed 'penalties'. Upon clicking the respective menu item, the currentr values can be modified. There never has been any need to do so.

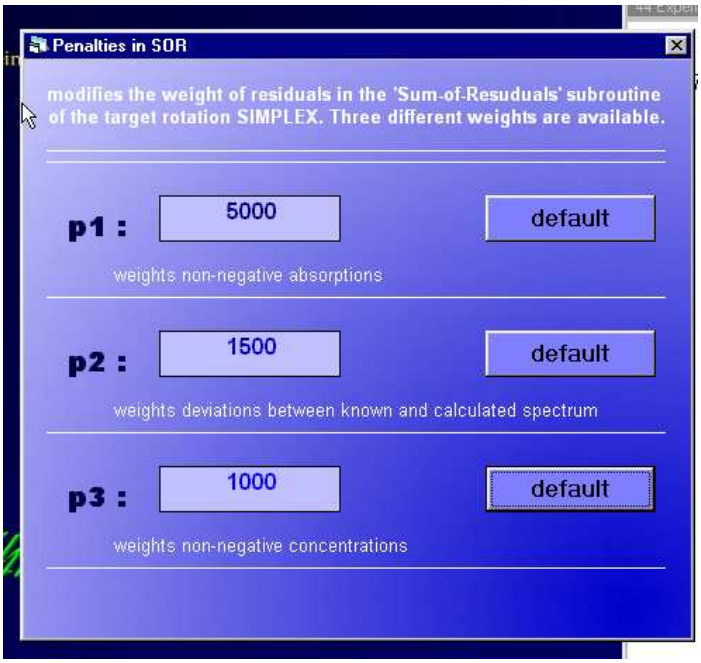


Figure y: The Penalty window. The default values have never been modified in practice.

(2) list Eigenvalues

Clicking this menu item provides a ordered list holding the singular values obtained from the abstract factor analysis step of CAT. Putting the focus to this window allows to save the list as an ASCII file.

(3) Plot Eigenvectors

Clicking this menu item plots the column eigenvectors into a diagram. Setting the focus to this diagram allows to save the graphical data into an ASCII file for subsequent analysis or processing by an external graphics program.

(4) Spectral Uncertainty

The spectral uncertainty window displays the first single component files (*.1) of the species available. These files are created during the TB CAT procedure. The user may choose four different confidence limits. If the respective selections have been made, TB CAT starts to generate a cumulative distribution function at each(!) wavelength. This process takes some time. If the process has finished, it is indicated in the title bar ('CDF done').

The generated ASCII file has the following name convention:

CDF_UVAAAANNN_B.dat

example: CDF_UVspecies(1)000_1.dat.

The letter code is the same as above. These files may be loaded into any graphics program and manipulated.

Figure z: Spectral Uncertainty window

(5) Differentiate

The menue item 'differentiate' transforms a cumulative distribution function of a formation constant into a probability density. A cumulative density function is by default saved as

cdf_AAAANNN.dat.

If a file is selected, the GO button starts to transfer the cumulative distribution data in file cdf_AAAANNN.dat into a probability density ASCII file dif_AAAANNN.dat. For this purpose a stepwise weighted linear regression algorithm (LOESS smoother) is used. Example: dif_species(2)000.dat.

A window opens showing the differentiated curve. However, the display may be misleading. To judge the probability density curve, it is necessary to open the data file in an external graphics program.

(f) About

opens a window giving some informations about the code. It mainly holds information on the person to which the respective code was personalized. It is understood that this code is NOT distributed without the explicite consent of the copyright owner.

The major practical purpose of this menu item however is to stop CAT temporarily. CAT is programmed to make full use of the computer resources. It doesn't like to share the CPU time. On the other hand, a TB CAT evaluation of, say 35, spectra with, say, three species may take several hours up to several days (large spectral data sets and slow CPU). To temporarily stop CAT, just open the 'About' window. All activity stops while the program waits to close this window. After closing it, CAT proceeds.

(g) Windows

The Window list is a special item of the VB programming environment. Under this menu item all currently accessible windows are listed and can be conveniently displayed even if a certain window should be covered completely by other windows.

A Summary of Recommendations

To get happy with TBCAT_S, the following suggestions may be helpful.

1) Do not expect to get a perfect code. TBCAT_S somehow runs – but it also crashes often. Often it is helpful just to restart the procedure.

2) Allow at least 24 hours before you engage into physical action against your computer.

3) After each procedure (Background correction, Analyze, CAT and TBCAT), save the results and restart the program. Some efforts have been made to allow just a modification in the 'basic thermodynmic data' window and restart the fitting from the same window. In this procedure the likely chemical composition of different species is commonly tested. It works often – but likes to fail in other cases. Be patient. At this stage of the analysis, where different species models must be tested to assess the likely meaning of the numerical data, the chemist's intuition and knowledge is required. No computer can replace that (fortunately). If the chemistry of the system is unclear, manual documentation of the already tested species models is required because TB CAT does not store or otherwise documents the user input.

4) An advantage of factor analysis is the large number of data to be handled simultaneously. Please remind that TBCAT_S, especially in the TBCAT subroutine, generates huge amount of data. Visual Basic commonly efficiently cares for the memory – but the memory handling of its operation system has several weak points. The "unexpected error" message has often been observed.

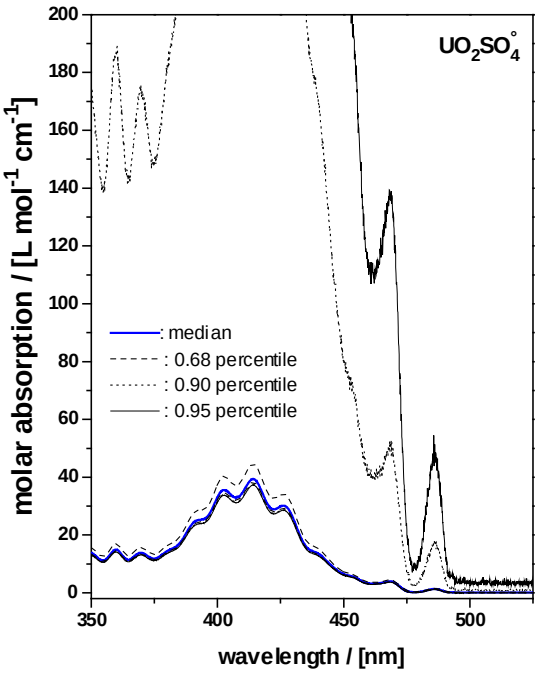
5) Familiarize yourself with the program on basis of synthetic data sets or simple spectroscopic systems. CAT has handled already quite complex systems with up to seven components. But experience shows that such systems should be split into smaller units to reduce correlation between the spectra.

6) Remind the input conventions: There must be always one component whose single component spectrum is known. This spectrum must be loaded in first place.

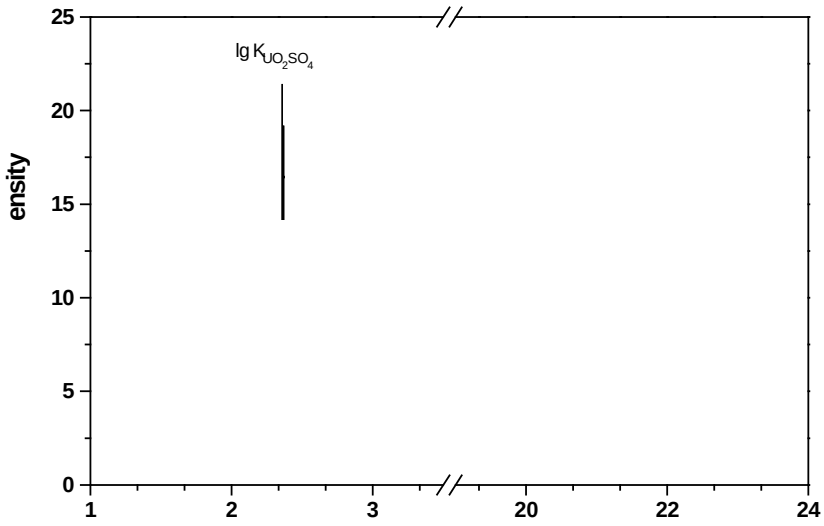
7) TBCAT_S is designed to resolve spectra of metal ions in solutions with ligands. Therefore, it is assumed that the free metal ion's spectrum is the known component. But TBCAT_S also works if the known spectrum belongs to a ligand, e.g. arsenazo III. Then, the absorption spectrum of arsenazo III takes the role of the known component and all input has to be modified accordingly. But there is no fundamental problem for CAT to evaluate such systems. A limitation is the situation where two components absorb. CAT cannot resolve this situation. A feasible way is to limit the analysis to wavelength ranges where only one component absorbs.

8) TBCAT_S is work-in-progress. Feel free to make suggestions.

Example results



Example 1: Single component spectrum of $UO_2SO_4°$. The blue curve gives the median. Note the extreme values for the upper 0.90 and 0.95 percentile spectral curves. The underlying distributions are highly non-Normal and skewed.



Example 2: Probability densities for formation constants of solution species.

References

There are a large number of publications having been influential in developing TBCAT_S. The following small selection is meant as a starter. The methods described in these references are of a general interest for the application of computers in chemistry.

M.S. Caceci "Estimating Error Limits in Parametric Curve Fitting". Anal. Chem. 61 (1989), 2324

M.S. Caceci, W.P. Cacheris; "Fitting Curves to Data: The SIMPLEX Algorithm is the Answer". Byte (1984) 340

A.A. Clifford, "Multivariate Error Analysis". Applied Science, London/UK (1973)

G.H. Golub, C. Reinsch, "Singular Value Decomposition and Least Squares Solutions". Numer. Math. 14 (1970) 403

J.A. Nelder, R. Mead, "A SIMPLEX Method for Function Minimization". Computer J. 7 (1965) 308

J.C. Nash, "Compact Numerical Methods for Computers". Adam Hilger Bristol/UK (1981)

G.E.P. Box, M.E. Muller, "A Note on the Generation of Random Normal Deviates". Ann. Math. Stat. 29 (1958) 610

B. Efron, "Computers and the Theory of Statistics: Thinking the Unthinkable" SIAM Review 21 (1979) 460

B. Efron, R. Tibshirani, "Statistical Analysis in the Computer Age" Science 253 (1991) 390

P.J. Rousseeuw, B.C. van Zomeren, "Unmasking Multivariate Outliers and Leverage Points" J. Am. Stat. Asssoc. 85 (1990) 633