

Chemometrics in f-element speciation: a metrological challenge?

Stefan Lis · Günther Meinrath

Received: 22 July 2010 / Accepted: 18 February 2011 / Published online: 9 March 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract The UV–Vis spectra of the f-elements, especially the lanthanides, are known for their rather sharp absorption bands showing systematic and reproducible changes upon coordination with suitable ligands. These absorption bands have been used as indicators allowing interpretation of the processes in solution. Modern absorption spectrometers in combination with digital recording of the spectra allow a precise collection of a large number of absorptions for many samples. The number of absorption thus collected easily amounts to several thousands of individual data. Chemometric techniques, especially factor analysis, are occasionally used to extract information from these data sets on basis of Bouguer–Lambert–Beer law. Recent studies indicate the sensitivity of parameter values obtained from uncritical use of chemometric techniques to various influence and nuisance factors. On the basis of selected examples, the effects of parameter correlations, residual correlations, and measurement uncertainty introduced by volume operations are demonstrated. Using the ISO Guide to the Expression of Uncertainty as a convention for assessing measurement

uncertainty, formation constants $\lg K$ derived from UV–Vis spectra of f-elements should be associated with a measurement uncertainty u of at least $u = 0.15$ ($k = 1$).

Keywords f-elements · UV–Vis spectroscopy · Chemometrics · Factor analysis · Guide to the Expression of Uncertainty

Introduction

The development of human civilization is closely related to the presence of water. Water is the most intensively monitored food item in the European community. All human activity affecting the environment is also affecting the water quality. Hence, a persistent interest exists in the analysis, understanding, quantification, and prediction of chemical processes in water. These processes can be rather elusive. Ligand exchange reactions with metal ions in aqueous solutions proceed to the nanosecond time scale and even below [1]. Nevertheless, these processes can be quantified as time averages, characterized by a variety of experimental methods and expressed in terms of thermodynamic quantities. To a limited extent, prediction of chemical reactions is possible within the limits of accuracy and precision of the experimental methods for quantification of reaction parameters, the simulation model, and the computational simulation tools [2].

UV–Vis spectroscopy is a classical technique for analyzing solution equilibria. Its versatility is enhanced by chemometric methods. Chemometrics is the field for application of numerical and statistical methods to data collected for chemical systems. This definition is rather general and encompasses a wide range of modern data treatment methods. The still increasing power of modern

This article is part of the Topical Issue “Quality Assurance of Thermodynamic Data”.

S. Lis (✉)
Department of Rare Earths, Faculty of Chemistry,
Adam Mickiewicz University, Grunwaldzka 6,
60-780 Poznań, Poland
e-mail: blis@amu.edu.pl

G. Meinrath
RER Consultants Passau, Fuchsbauerweg 50,
94036 Passau, Germany

G. Meinrath
Technische Universität Bergakademie Freiberg,
09596 Freiberg, Germany

computing machines allows investing a considerable amount of processor time for numerical treatment (“number crunching”) of chemical data. The introduction of chemometric techniques was largely motivated by the need to extract information from large data sets collected by electronic instruments in digital form. Of special interest are model-free numerical techniques. These methods do not require a priori information on factors giving rise to the measurement data. Latent variable methods, for instance principal component analysis (PCA) and partial least squares (PLS), distinguish between the intrinsic (latent) variables and the interpretation of these variables, the physical parameters. PCA and PLS have originated in other fields, e.g. econometrics and biometrics, but meanwhile are well accepted in analytical chemistry [3]. Their success in chemistry is supported by the often low dimensionality of chemical systems. Furthermore, the transformation of the latent variables into physically meaningful parameters is guided by experimentally available information about the system under study (e.g., non-negativity of absorptions and species concentrations). A variety of techniques and ready-to-use computer programs are available in literature [4–12]. The programs are often used as black boxes—a phenomenon not limited to latent variable methods [13–16].

No experimental measurement can obtain information with arbitrary accuracy and precision [17]. In the academic world, results of chemical measurements are often reported without a systematic assessment of limits of measurability. In all fields where the results of chemical measurement will enter, public decision-making processes, e.g. environmental assessment, the results of chemical analysis, are coming under public scrutiny. Measurement values not being in agreement with appropriate requirements will increasingly be rejected for formal reasons. Two common situations where these requirements are enforced are breath alcohol determination [18] and food quality control [19]. Process and language of assessing measurement uncertainty is already quite formalized [20, 21]. Thermodynamic data of relevant geochemical reactions serve as a basis for environmental prediction. Thermodynamic data are a result of complex experimentation and data evaluation. Reference to concepts of measurement uncertainty assessment is almost completely missing. Integration of metrological concepts into chemometric evaluation procedures of thermodynamic data has been a major interest during the past decade. This is understood as work in progress.

Results and discussion

A complete assessment of measurement uncertainty in chemical analysis is statistically and computationally demanding. UV–Vis spectroscopic analysis of lanthanide

solution chemistry was selected as a model technique, because most chemists understand its underlying theory, and a well-developed computational toolbox is available based on factor analysis. The UV–Vis spectra of lanthanide ions in solution are known for their comparatively narrow but weak absorption bands showing distinct shifts due to changes in the electronic environment by complexation with ligands. The numerical analysis of experimental spectra by factor analytical methods is quite elaborate. In order to evaluate a complete measurement uncertainty budget under the aspects outlined in EURACHEM/CITAC’s guide, “Quantifying Uncertainty in Chemical Measurement” [20] requires further numerical operations. The analysis of uncertainty in complex chemical measurement, e.g. UV–Vis spectroscopic assessment of sample solutions, is influenced by a larger number of factors, some that the experimenter cannot completely control [9, 22].

Factor analysis—some basic elements

Bouguer–Lambert–Beer law is the fundamental relationship in UV–Vis spectroscopy.

$$a_{\lambda} = \varepsilon_{\lambda} c d \quad (1)$$

where a is the absorbance at wavelength λ , ε is the molar absorption coefficient, c is the concentration of the absorbing species, and d is the length of the light path through the sample. If several absorbing species are present in solution and in the linear absorption range of the sample, Eq. 1 is replaced by

$$a_{\lambda} = d \sum_{i=1}^n \varepsilon_{\lambda,i} c_i + \delta_i \quad (2)$$

where δ_i represents the residuals. Residuals are estimators for the disturbances and obtained as the differences between the optimal interpretation of the system and the experimental observations. The residuals result from measurement uncertainty, noise, and bias. If the absorption is measured at several different wavelengths and several samples with different solution compositions, Eq. 2 can be written in matrix form as

$$\mathbf{A} = \mathbf{E}\mathbf{C} + \Delta \quad (3)$$

In Eq. 3, \mathbf{A} gives the nm matrix of absorbances measured at n wavelengths in m samples. \mathbf{E} gives the nk matrix of molar extinction coefficients of k different species at n wavelengths, and \mathbf{C} is the km matrix of concentrations of k species in m different solutions. The absorbances in matrix \mathbf{A} are normalized for a given path length d . For linear equations, as Eq. 2, principal component analysis allows to derive the matrices \mathbf{E} and \mathbf{C} from a given matrix \mathbf{A} of experimentally measured spectra (the columns of \mathbf{A}) of the same chemical system.

The transition from experimentally measured spectra in the columns of matrix A to the matrix E of single component spectra and the matrix C of species concentrations in the m sample solutions requires some numerical and mathematical transformations that have to be guided by the information available to or derived by the experimenter (e.g., non-negativity of absorptions and concentrations). In the first step of the analysis, the singular values and singular vectors of matrix A are determined by a numerical algorithm. Today, singular value decomposition is applied routinely using the SVD algorithm [22]. As a result, the matrix of column singular vectors $E^\#$, the matrix of row singular vectors $C^\#$, and the vector of singular values V are obtained.

$$A = E^\#VC^\# \quad (4)$$

The singular vectors are orthogonal to each other. Hence, no singular vector interprets experimental variance already interpreted by another vector. The singular values in vector V are ordered in decreasing magnitude. Their values are an indicator for the relevance of a singular vector. Because, in theory, all experimental variance is a result of an absorbing species (cf. Eq. 2), the number of non-zero singular vectors should not be larger than the number of species in the system. However, because experimental information is affected by measurement uncertainty, all singular values in vector V are non-zero.

Types of measurement uncertainty

Factor analysis is not a foolproof technique even at the purely mathematical level. The reason is the underlying theory. The basic theory of factor analysis sets requirements to the data structure which are similar to those for ordinary linear regression: (a) the expectation function is correct; (b) the disturbances are independent of the signal; (c) each disturbance δ has a normal distribution; (d) the disturbances δ have equal variances; and (e) the disturbances δ are independently distributed. Failure to comply with these requirements will introduce bias. The effect, especially of requirement (e), can be visualized by using computer-intensive resampling methods, e.g. the moving block bootstrap (MBB).

MBB creates a large number of new data sets by generating new noise patterns from the residuals and adds it to the mean value (optimal) interpretation of the system. Correlation between neighboring residuals (the estimates of the disturbances) is maintained by composing the new noise pattern from chunks (lags) of residuals with a specified lag size. The MBB is a statistical approach to time series analysis and suitable for residual analysis of spectral information. Details have been given previously [23]. To illustrate the effect of correlated residuals, the species

concentrations estimated by least squares regression (LSR) and by MBB are compared for the complexation of U(VI) by sulfate in Fig. 1. Three species are assumed to be present in solution UO_2^{2+} , UO_2SO_4 , and $UO_2(SO_4)_2^{2-}$. The ellipses give the 95% confidence regions for the concentrations of the species; in other words, if the experiment would be repeated a large number of times, the measured species concentrations should be found within the ellipses in 95% of the cases. Figure 1 can give an approximate representation only, because the concentration domain should be an ellipsoid in 3D space. To allow representation on paper, the confidence ellipses for two species concentrations are given keeping the third species at its mean value concentration. Thus, the ellipses are conditional confidence regions. Note that the ellipses are very narrow and elongated as a result of high parameter correlation (which should not be confused with residual correlation). The MBB point clouds (obtained from 1000 resamplings) do only marginally overlap with the least square confidence ellipses. The simulation indicates that a difference in the expected species concentrations results from the assumption of correlation between residuals. Hence, assumption of residual independence introduces bias in the numerical evaluation resulting, for instance, in the evaluation of equilibrium information from spectroscopically determined species concentrations. A further aspect highlighted by Fig. 1 are the least square mean values that are not covered by the MBB point clouds. Hence, focusing exclusively on least squares mean values will usually contribute additional bias into the evaluation of data from UV–Vis spectra.

A number of other nuisance contributions affect the evaluation of chemical information from UV–Vis spectroscopy. In the ISO terminology [24], contributions to measurement uncertainty are obtained either by a detailed statistical evaluation (Type A evaluation) or by other means, e.g. separate experimentation (Type B evaluation). A typical Type B contribution are volume operations, e.g. by Eppendorf pipettes. Using a calibrated balance and thermostated distilled water, the variability can be assessed by repeated transfer of a given volume water to the balance. From the density of water, the volume and its variation can be assessed. A typical example is given in Fig. 2.

Ten samples of nominally 1 cm^{-3} are pipetted to a balance at $20 \text{ }^\circ\text{C}$. The density of water at $20 \text{ }^\circ\text{C}$ is 0.998 g cm^{-3} . The ten samples given in Fig. 2 scatter between 0.990 and 1.015 g cm^{-3} and enclose the theoretical value. An interpretation of the data set by a Kolmogorov–Smirnov test indicates that the null hypotheses (the sampled data are normally distributed and deviations from normality are random) cannot be rejected. Volume operations contribute about 2–5% of measurement uncertainty.

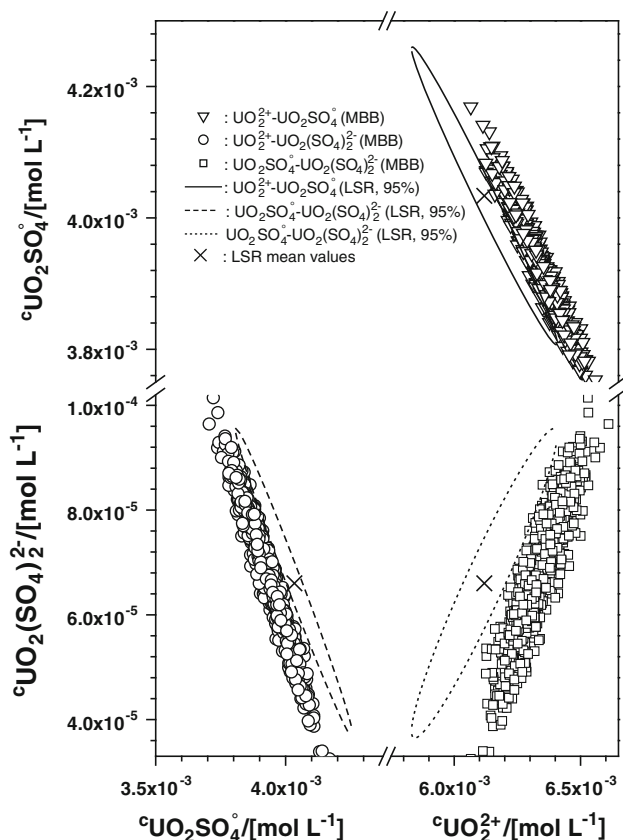


Fig. 1 Comparison of U(VI) species concentrations confidence regions obtained from mean value analysis (95% confidence ellipses) and Monte Carlo (MC) resampling (point clouds). The confidence ellipses are obtained from least squares regression (LSR). The MC point clouds take into account residual correlation and are obtained from a Moving Block Bootstrap (MBB) analysis

The both examples were selected to illustrate Type A and Type B of measurement uncertainty contributions. Table 1 summarizes other relevant influence factors. Analysis of influence factors is advantageously done by creating a cause-and-effect diagram [20].

Ambiguities

Matrices and vector given in Eq. 4 are mathematical structures without any physical meaning. The singular vectors, for instance, usually do have negative values. It is the task of the experimenter to transform these matrices into physically meaningful information. The first step is to decide on the dimensionality of the system. In most cases, the magnitude of the singular values is analyzed with the intention to derive clues on the threshold between meaningful information and noise. Deciding the dimensionality of the system is crucial because each dimension corresponds to chemical species. Figure 3 gives typical UV–Vis spectra of solutions holding a lanthanide metal ion and

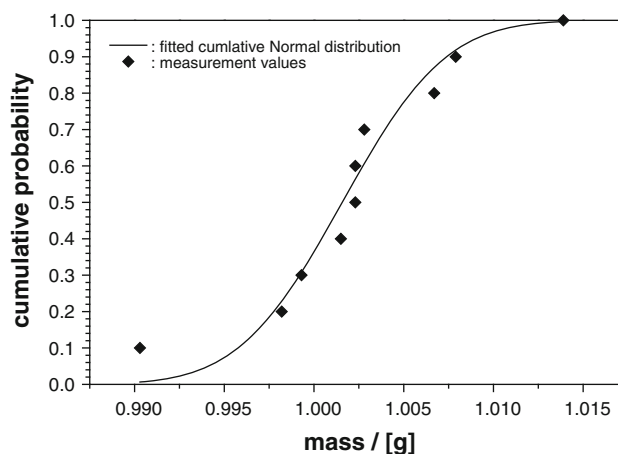


Fig. 2 Empirical cumulative distribution of water weight determined with 1-mL Eppendorf pipettes using a calibrated balance (Mettler). The ten samples obtained are interpreted by a Kolmogorov–Smirnov test with the closest fitting cumulative normal distribution with a mean of 1.00(2) mL and a standard deviation of $4.5 \cdot 10^{-3}$ mL. Obviously, the respective volume operation is affected by an uncertainty. It is impossible to perform volume operations with arbitrary accuracy. Volume operations are only one of several factors affecting the accuracy of experimental results derived from data obtained by UV–Vis spectroscopy (from [30])

Table 1 Influence factors affecting the result of analytical parameters obtained by UV–Vis spectroscopy

Type A	Type B
Signal noise	Lanthanide concentration
Spectral correlation	Ligand concentration
Residual correlation	Volume determination
Parameter correlation	Baseline correction
Non-normality	Repeatability
Non-linearity	
Statistical optimization criterion	
Monte Carlo effects	

The influence factors are grouped into Type A and Type B uncertainties

varying amounts of a ligand, pyridine 2,6-dicarboxylic acid N-oxide.

While the sharp absorption bands of the trivalent lanthanides are advantageous for the chemometric analysis of the chemical reaction in the sample solution, the small shifts and, consequently, strong overlap of the absorption bands of the various species are not. The strong overlap of the single component spectra of individual species implies high spectral correlation. Some consequences have already been shown in Fig. 1: the confidence ranges are not symmetrically distributed about a mean value but found in an elongated region where—depending on the correlation—high values of one parameter prefer low values of the other parameter and vice versa. This high correlation also affects

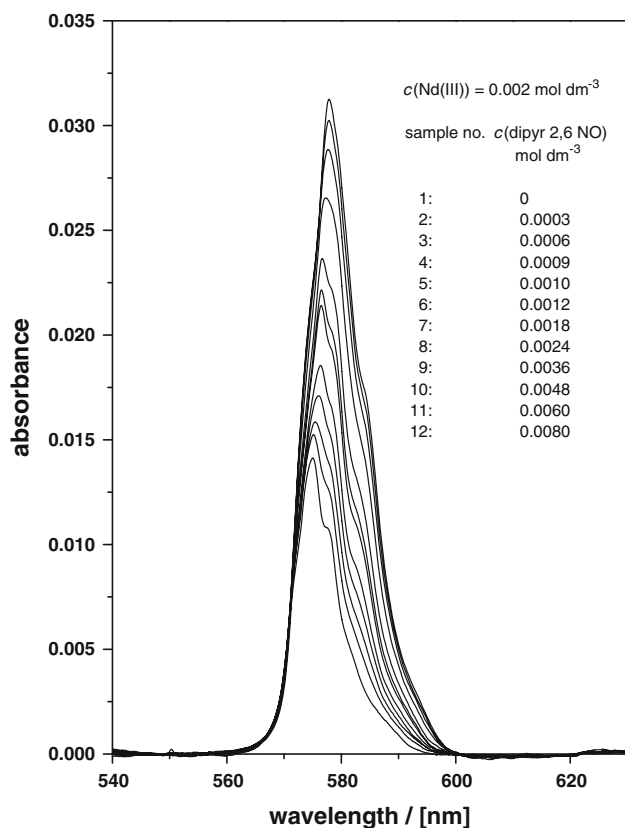


Fig. 3 A set of UV–Vis spectra in the region 540–620 nm obtained from solutions of Nd(III) holding varying amounts of nicotinic acid N-oxide. The spectral bands are rather *narrow*. The modifications observed in the spectra related to the presence of pyridine 2,6-dicarboxylic acid N-oxide ligand, however, are likewise small and consist mainly in an increase in the absorption band. A consequence of the rather small shifts is high parameter correlation (cf. Fig. 1)

the efficiency of obtaining the dimensionality of the system (the number of species in solution) from an analysis of the singular values. This is illustrated by Fig. 4 where the singular values (left axis) are given for the spectra shown in Fig. 3.

The singular values decrease rapidly. The first singular value explains about 85% of the experimental variance (the residual variance is given at the right side axis). The first two singular vectors interpret almost 98% of the variance leaving a marginal difference between a possible third factor and the noise in the data. Not surprisingly, a larger number of methods have been developed to extract the likely number of factors from experimental data [e.g., 3, 25–27]. Given the narrow margin left for decision, it is always recommended to check whether the system can be reasonably interpreted with more or less species that suggested by whatever test criterion. In some cases, factor analysis may be able to provide two independent and equally satisfactory interpretations of the same system with different number of species [28].

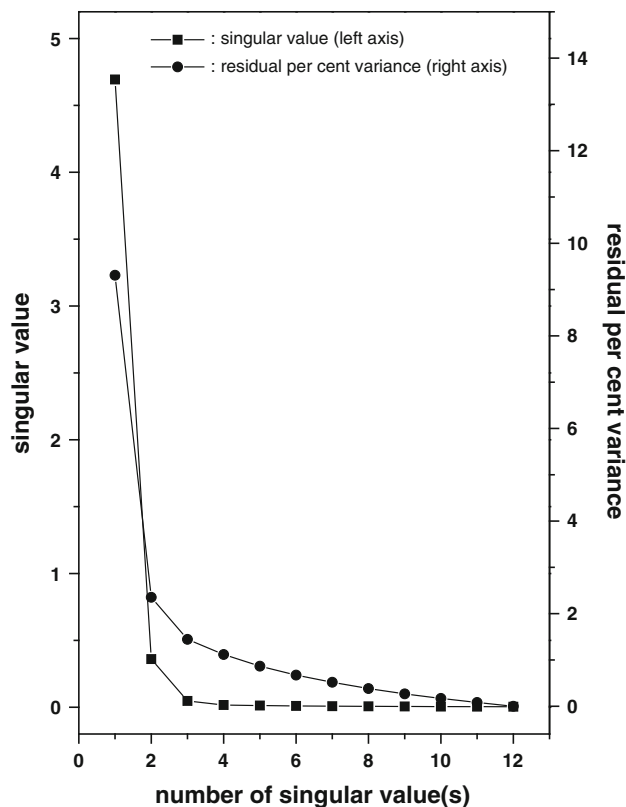


Fig. 4 Residual variance in the singular values as a function of singular values included into the target rotation analysis. The first singular value explains almost 85% of the experimental variance. It is almost impossible to give an unambiguous statement on the number of significant singular values. This situation is typical in factor analysis of experimental data. A variety of statistical approaches have been proposed. Different approaches, however, will also result in different conclusions on the number of relevant factors (=species) in a given set of UV–Vis spectra.

System analysis

Factor analysis is an effective method to analyze larger sets of spectra. The previous paragraphs, however, have shown that factor analysis is influenced by a variety of influence factors. Furthermore, factor analysis alone is not sufficient to arrive at single component spectra, species concentrations and derived information like molar extinction coefficients, and formation constants of the solution species. Some objective criteria can be used to evaluate possible solutions of the numerical equations, e.g., the non-negativity of single component spectra and species concentrations. Formally, a transformation matrix T with dimensions kk has to be found which transforms the abstract matrices E' (with $E' = E^{\#}V$) and $C^{\#}$ into the physically meaningful single component matrix E° and C° . Then,

$$E^{\circ} = E'T \quad (5)$$

$$C^{\circ} = T^{-1}C^{\#} \quad (6)$$

The step of transforming the matrices from abstract singular vectors into physically meaningful matrices with matrix E° holding the estimates of the k single component spectra and matrix C° holding the k species concentrations in each of the m solutions is usually termed target factor analysis [29]. A transformation matrix T is usually estimated by a complex process, often of a trial-and-error nature. In order to account for the various Type A and Type B uncertainties, a Monte Carlo resampling procedure has to be implemented which requires at least 1000 repetitions of the target transformation procedure. Computer-intensive methods therefore must be based on robust and stable algorithms. For factor analysis, the singular value decomposition (SVD) satisfies this requirement [3]. Hence, an algorithm has been implemented which estimates target transformation matrix T without user-specified input where SVD provides the matrix diagonalization step. Analysis of experimental data comprises at least 1000 repetitions of the steps indicated by Eqs. 4–6. In each run, a new matrix A is generated by redistributing correlated noise, and the values of the Type B influence factors are modified randomly within the given distributions. The resulting parameters are evaluated and stored in a file. Finally, the empirical distribution functions are evaluated. The complete procedure was termed threshold bootstrap computer-assisted target factor analysis (TB CAT). A more detailed description is found in [9].

A larger number of different lanthanide systems were systematically analyzed. Table 2 gives a selection together with some references. In all situations (including the somewhat unusual polyoxometallate ligands), the relative simplicity of the lanthanide systems became evident. Numerical analysis in all cases indicated not more than two coordinated metal species. The evaluation of the species concentration information revealed the delicate equilibria. Small numerical variations may infer considerable

modifications in the concentration domain. If a unique and stable numerical solution was available, the evaluated formation constants were found to be rather narrow with complete measurement uncertainties u (u in logarithmic scale) for $\lg K$ in the order of $u = 0.15$ ($k = 1$). If spectral correlation ρ was above $|\rho| > 0.8$, the evaluated formation constants can easily be found distributed over several orders of magnitude.

Conclusions

A chemist is interested in information on the number of species in a given chemical system, their formation constants, and their single component spectra. Factor analysis is, in principle, capable of extracting this information from a limited number of UV–Vis spectra. The procedure is numerically stable and straightforward. Thus, factor analysis is an almost unique tool to study chemical systems with unusual ligands, e.g. polyoxometallates, where little a priori information is available. Aqueous systems involving lanthanide ions are especially suitable for a treatment by factor analysis because of their narrow absorption bands and marked shifts with coordination.

However, factor analysis is capable to provide a numerical solution for almost all data satisfying the mathematical requirements. Combining target factor analysis with an efficient search algorithm (e.g., simplex algorithm) will almost certainly retrieve a solution if there is one. This solution, however, need not to be unique. This finding is a major reason to view any interpretation based solely on factor analysis with skepticism. A further reason is the sensitivity of the numerical solution to some minor changes in the input quantities. The correlation in residuals is just one such contribution. Together with the small uncertainties in auxiliary variables (e.g., ligand concentrations, pH, and volume operations), the resulting variabilities may range over orders of magnitude.

Therefore, some provisions are recommended by which the sensitivity of a numerical solution to modifications in the input data can be assessed. The studies performed with TB CAT systematically identified, quantified, and evaluated those influence factors with marked effect on the modeling output. Thus, the complete analyses are complying with the requirements of ISO's Guide to the Expression of Uncertainty (GUM) [24]. The results indicated that some chemical systems are better suited for an analysis by UV–Vis spectroscopy than others. Some systems are more amenable to factor analysis treatment than others. Probably, chemical systems less suitable for factor analysis may advantageously be analyzed by other experimental techniques. Thus, application of the rules of metrology in chemistry carry does not only the hope of

Table 2 f-Element systems analyzed including an assessment of measurement uncertainty

Ligand	Reference
Na ₂₇ [NaAs ₄ WO ₄₀ O ₁₄₀]	[9]
K ₇ PW ₁₁ O ₃₉ ·17 H ₂ O	[31]
Picolinic acid N-oxide	[22]
Nicotinic acid N-oxide	[32]
Pyridine 2,4 dicarboxylic acid N-oxide	[28]
Pyridine 2,6 dicarboxylic acid N-oxide	[33]
Diethyl(2-oxopropyl)phosphonate	[34]
Diethyl(2-oxo-2-phenylethyl) phosphonate	[34]
Arsenazo (III)	[35]
Sulfate ^a , hydroxide ^a	[36]

^a With U(VI) as metal ion

traceable and internationally comparable measurement values, but also criteria by which the most suitable experimental method for the analysis of a chemical system can be selected.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Ohtaki H, Radnai T (1993) *Chem Rev* 93:1157–1204
- Ödegaard-Jensen A, Ekberg Ch, Meinrath G (2004) *Talanta* 63:907
- Malinowski ER (2002) *Factor analysis in chemistry*, 3rd edn. Wiley-Interscience, New York
- Manne R, Grande BV (2000) *Chemom Intell Lab Syst* 50:35
- Tauler R (1995) *Chemom Intell Lab Syst* 30:133–146
- De Braekeler K, De Juan A, Massart DL (1999) *J Chromatog A* 832:67–86
- Windig W, Guilment J (1991) *Anal Chem* 63:1425–1432
- Windig W (1992) *Chemom Intell Lab Syst* 16:1
- Meinrath G, Lis S (2001) *Fresenius J Anal Chem* 369:124–133
- Cazallas R, Citroes MJ, Etxebarria N, Fernandez LA, Madariaga JA (1994) *Talanta* 41:1637–1644
- Cartwright H, Farley HM (1988) *J Chemometrics* 2:111–119
- Maeder M (1987) *Anal Chem* 59:527–530
- Filella M, May PM (2005) *Talanta* 65:1221–1225
- Ekberg Ch, Meinrath G, Strömberg B (2003) *J Chem Thermodynamics* 35:55–66
- Chalmers RA (1993) *Talanta* 40:121–126
- Thompson M (1994) *Analyst* 119:127N
- Menditto A, Patriarc M, Magnusson B (2007) *Accred Qual Assur* 12:45–47
- Gullberg RG (2006) *Accred Qual Assur* 11:562–568
- Haouet MN, Altissimi MS, Framboas M, Galarini R (2006) *Accred Qual Assur* 11:23–28
- <http://www.measurementuncertainty.org>
- <http://www.ntmdt.ru/download/vim.pdf>
- Meinrath G, Lis S (2002) *Anal Bioanal Chem* 372:333–340
- Meinrath G (2000) *Anal Chim Acta* 415:105–115
- ISO/IEC (2008) *Guide 98–3:2008: uncertainty of measurement—part 3: guide to the expression of uncertainty in measurement*. ISO, Geneva/CH
- Malinowski ER (1977) *Anal Chem* 49:606–612
- Rutledge DN, Barros AS (2002) *Anal Chim Acta* 454:277–295
- Tu XM, Burdick DB, Millican DW, McGown LB (1989) *Anal Chem* 61:2219–2224
- Meinrath G, Hnatejko Z, Lis S (2004) *Talanta* 63:287–296
- Hopke PK (1989) *Chemom Intell Lab Syst* 6:7–19
- Meinrath G, Schneider P (2007) *Quality assurance in chemistry and environmental science*. Springer, Heidelberg/FRG
- Meinrath G, Lis S, But S, Elbanowski M (2001) *Talanta* 55:371–386 (29)
- Meinrath G, Lis S, Elbanowski M (2004) *J Alloy Comp* 380:413–417
- Meinrath G, Lis S, Böhme U (2006) *J Alloy Comp* 408–412:962–969
- Hnatejko Z, Lis S, Pawlicki G, Meinrath G (2008) *J Alloys Comp* 451:395–399
- Kaczmarek M, Meinrath G, Lis S, Kufelnicki A (2008) *J Sol Chem* 37:933–946
- Meinrath G, Lis S, Piskula Z, Glatty Z (2006) *J Chem Thermodynamics* 38:1274–1284