Available online at www.sciencedirect.com

SCIENCE DIRECT®

Talanta 65 (2005) 1215–1220

Talanta

www.elsevier.com/locate/talanta

Principal component analysis versus fuzzy principal component analysis A case study: the quality of danube water (1985–1996)

C. Sârbu[a,*], H.F. Pop[b]

[a] *Department of Analytical Chemsitry, Faculty of Chemistry and Chemical Engineering, "Babeş-Bolyai" University, Arany Janos Str. 11, RO-400028 Cluj-Napoca, Romania*

[b] *Department of Informatics, "Babeş-Bolyai" University, RO-400028 Cluj-Napoca, Romania*

Received 22 April 2004; received in revised form 4 August 2004; accepted 25 August 2004
Available online 29 September 2004

Abstract

Principal component analysis (PCA) is a favorite tool in environmetrics for data compression and information extraction. PCA finds linear combinations of the original measurement variables that describe the significant variations in the data. However, it is well-known that PCA, as with any other multivariate statistical method, is sensitive to outliers, missing data, and poor linear correlation between variables due to poorly distributed variables. As a result data transformations have a large impact upon PCA. In this regard one of the most powerful approach to improve PCA appears to be the fuzzification of the matrix data, thus diminishing the influence of the outliers. In this paper we discuss and apply a robust fuzzy PCA algorithm (FPCA). The efficiency of the new algorithm is illustrated on a data set concerning the water quality of the Danube River for a period of 11 consecutive years. Considering, for example, a two component model, FPCA accounts for 91.7% of the total variance and PCA accounts only for 39.8%. Much more, PCA showed only a partial separation of the variables and no separation of scores (samples) onto the plane described by the first two principal components, whereas a much sharper differentiation of the variables and scores is observed when FPCA is applied.

Keywords: Water quality; Principal component analysis; Fuzzy principal component analysis

1. Introduction

Multivariate statistical methods for the analysis of large quantities of data have been applied to chemical and environmental systems during the last decades [1–4]. One of these methods, principal component analysis (PCA) showed special promise for furnishing new and unique insights into the interactions in a wide range of pollution and ecotoxicological situations [5–11].

PCA is designed to transform the original variables into new, uncorrelated variables (axes) called the principal components, that are linear combinations of the original variables. The new axes lie along the directions of maximum variance. PCA provides an objective way of finding indices of this type so that the variation in the data can be accounted for as concisely as possible.

Principal component analysis as with any other multivariate statistical method is sensitive to outliers, missing data, and poor linear correlation between variables due to poorly distributed variables. As a result, the classical principal components may describe the shape of the majority of data incorrectly. It is therefore necessary to apply robust methods that are resistant to possible outliers [12]. In this order, during the last decades, two robust approaches have been developed. The first is based on the eigenvectors of a robust covariance matrix such as the MCD-estimator [13] or S-estimators of location and shape [14,15], and is limited to relatively low-dimensional data. The second approach is based on projection pursuit and can handle high-dimensional data [16]. A robust PCA approach which combines projection pursuit ideas with robust estimation of low-dimensional data has been de-

[*] Corresponding author. Tel.: +40 264 533898; fax: +40 264 590818.
E-mail address: csarbu@chem.ubbcluj.ro (C. Sârbu).

0039-9140/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2004.08.047

veloped [17] and applied to several bio-chemical datasets [18].

However, one of the most promising approaches to “robustify” PCA has been appearing to be the fuzzification of the matrix data to diminish the influence of outliers [19–26].

In this paper we discuss and apply a robust fuzzy PCA algorithm (FPCA) [27]. The efficiency of the new algorithm is illustrated on a data set concerning the water quality of the Danube River for a period of 11 consecutive years.

2. Theoretical considerations

2.1. Classical principal component analysis

Principal component analysis is also known as eigenvector analysis, eigenvector decomposition or Karhunen–Loéve expansion. As we have already mentioned above, the main purpose of PCA is to represent in an economic way the location of the samples in a reduced coordinate system where instead of *m*-axes (corresponding to *m* characteristics) only *p* ($p < m$) can usually be used to describe the data set with maximum possible information.

Principal component analysis practically transforms the original data matrix ($\boldsymbol{X}_{\text{nxm}}$) into a product of two matrices, one of which contains the information about the samples ($\boldsymbol{S}_{\text{nxm}}$) and the other about the variables ($\boldsymbol{V}_{\text{mxm}}$). The $\boldsymbol{S}$ matrix contains the scores of the *n* objects on *m* principal components (the scores are the projection of the samples on principal components). The $\boldsymbol{V}$ matrix is a square matrix and contains the loadings of the original variables on the principal components (the loadings are the weights of the original variables in each principal component).

2.2. Fuzzy principal component analysis

Fuzzy clustering is an important tool to identify the structure in data [19–29]. In general, a fuzzy clustering algorithm with objective function can be formulated as follows: let $X = \{x^1, \ldots, x^n\} \subset \boldsymbol{R}^p$ be a finite set of feature vectors, where *n* is the number of objects (measurements) and *p* is the number of the original variables, $x_k^j = [x_1^j, x_2^j, \ldots, x_p^j]^T$ and $L = (L^1, L^2, \ldots, L^s)$ be a *s*-tuple of prototypes (supports) each of which characterizes one of the *s* clusters composing the cluster substructure of the data set; a partition of *X* into *s* fuzzy clusters will be performed by minimizing the objective function

$$J(P, L) = \sum_{i=1}^{s} \sum_{j=1}^{n} (A_i(x^j))^2 d^2(x^j, L^i) \tag{1}$$

where $P = \{A_1, \ldots, A_s\}$ is the fuzzy partition, $A_i(x^j) \in [0, 1]$ represents the membership degree of feature point x^j to cluster A_i, $d(x^j, L^i)$ is the distance from a feature point x^j to the prototype of cluster A_i, defined by the Euclidean distance norm

$$d(x^j, L^i) = ||x^j - L^i|| = \left[\sum_{k=1}^{p} (x_k^j - L_k^i)^2\right]^{1/2} \tag{2}$$

The optimal fuzzy set will be determined by using an iterative method where *J* is successively minimized with respect to *A* and *L*.

Supposing that *L* is given, the minimum of the function $J(\cdot, L)$ is obtained for:

$$A_i(x^j) = 1 / \sum_{k=1}^{s} \frac{d^2(x^j, L^i)}{d^2(x^j, L^k)}, \quad i = 1, \ldots, s \tag{3}$$

For a given *P*, the minimum of the function $J(P, \cdot)$ is obtained for:

$$L^i = \sum_{j=1}^{n} [A_i(x^j)]^2 x^j / \sum_{j=1}^{n} [A_i(x^j)]^2, \quad i = 1, \ldots, k \tag{4}$$

The above formula allows one to compute each of the *p* components of L^i (the center of the cluster *i*). Elements with a high degree of membership in cluster *i* (i.e., close to cluster *i*’s center) will contribute significantly to this weighted average, while elements with a low degree of membership (far from the center) will contribute almost nothing.

A cluster can have different shapes, depending on the choice of prototypes. The calculation of the membership values is dependent on the definition of the distance measure.

According to the choice of prototypes and the definition of the distance measure, different fuzzy clustering algorithms are obtained. If the prototype of a cluster is a point – the cluster center – it will give spherical clusters, if the prototype is a line it will give tubular clusters, and so on. In view of the linear form of the consequence part in linear fuzzy models, an obvious choice of fuzzy clustering was the Generalized fuzzy *n*-means algorithm [21–29], in which linear or planar clusters are allowed as prototypes to be sought.

The fuzzy set in this case may be characterized by a linear prototype, denoted $L(u, v)$, where v is the center of the class and u, with $||u|| = 1$, is the main direction. This line is named the first principal component for the set, and its direction is given by the unit eigenvector *u* associated with the largest eigenvalue $\lambda_{\max}$ of, for example, the covariance matrix given in relation (5), which is a slight generalization for fuzzy sets of the classical covariance matrix:

$$C_{kl} = \frac{\sum_{j=1}^{n} [A_i(x^j)]^2 (x_{jk} - \bar{x}_k)(x_{jl} - \bar{x}_l)}{\sum_{j=1}^{n} [A_i(x^j)]^2} \tag{5}$$

The algorithm applied in this paper permits the determination of the $A(x^j)$ values that best describe the fuzzy set *A* and the relation with its linear prototype (the first principal component). This algorithm is a natural extension of the fuzzy 1-lines algorithm [28–29].

To obtain the criterion function, we have to determine a fuzzy partition $\{A, \bar{A}\}$; the set A is characterized by its linear prototype.

In relation to the complementary fuzzy set, $\bar{A}$, we will consider that the dissimilarity between its hypothetical prototype and the point x^j is constant and equal to $\alpha/(1-\alpha)$, where α is a real constant from the interval (0, 1). As a consequence the criterion function becomes

$$J(A, L; \alpha) = \sum_{j=1}^{n} [A(x^j)]^2 d^2(x^j, L) + \sum_{j=1}^{n} [\bar{A}(x^j)]^2 \frac{\alpha}{1-\alpha}. \quad (6)$$

The prototype $L(u, v)$ that minimizes the function $J(A, \cdot, \alpha)$ is given by

$$v = \sum_{j=1}^{n} [A(x^j)]^2 x^j \Big/ \sum_{j=1}^{n} [A(x^j)]^2 \quad (7)$$

where

$$A(x^j) = \frac{\alpha/(1-\alpha)}{[\alpha/(1-\alpha)] + d^2(x^j, L)}. \quad (8)$$

It follows from here that α represents the membership degree of the farthest point (the largest outlier) from the first principal component. Since this is an input parameter, we need a heuristics for determining the best suitable value of α. As opposed to the general case, we now do have such a mechanism. Of course, we are interested to find fuzzy membership degrees that contribute to producing a better fitted first principal component along the data set. But, since the eigenvalue associated to a principal component describes the scatter of data along that component, we are also interested in producing a first principal component characterized by an eigenvalue that is as large as possible. As a consequence, we will prefer that particular value of α that maximizes the eigenvalue associated to the first principal component.

Because of the fact that we are interested in real-world applications of this algorithm, an exact value of α is not required. Instead, we will simply work through a loop between 0 and 1, with a step to be chosen by the user, and select the value of α that maximizes our criterion.

The steps in a fuzzy principal component analysis can now be stated:

1. Determine the best value of α. For this, loop with α between 0 and 1. For each iterative value of α minimize the objective function (6), and, with the optimal membership degrees $A(x^j)$, compute the largest eigenvalue of the matrix C given by (5). Select the optimal value of α according to the maximal eigenvalue.
2. Coding the variables $X_1, X_2, \ldots, X_p$ to have zero means and unit variances. This is usual, but is omitted in some cases.
3. Calculate the covariance matrix $\boldsymbol{C}$, as given by relation (5). This is a correlation matrix if step 2 has been done.
4. Find the eigenvalue λ_1 and the corresponding eigenvector $\boldsymbol{e}_1$.
5. Determine the new fuzzy set A^{l+1} using Eq. (8).
6. If the fuzzy sets $A^{(l+1)}$ and $A^{(l)}$ are close enough, i.e., if $||A^{(l+1)} - A^{(l)}|| < \varepsilon$
   where ε has a predefined value (i.e. 10^{-5}), then stop, else increase l by 1 and go to the step 3; else, continue with step 7.
7. Using the fuzzy membership degrees determined above, recompute the covariance matrix C as in (5), and determine its eigenvalues and eigenvectors as usually; these are the fuzzy principal components and the corresponding scatter values.

3. Results and discussion

The data collection was performed at Galaţi site, Romania, according to standardized methods for sampling, sample preparation and analysis of Danube River water for a period of 11 consecutive years [30]. Galaţi site is selected as representative for the Danube estuary region.

Nineteen different water parameters were checked monthly (pH, chemical oxygen demand-COD, equivalent oxygen, calcium, magnesium, calcium/magnesium ratio, chloride, sulphate, hydrogen carbonate, nitrite, nitrate, phosphate, ammonia, ammonium, alkalinity, hardness, dry residue, suspension).

The standardized methods for water quality analysis were used: potentiometry with glass electrode (pH, alkalinity), titrimetry (COD, EO_2, Ca^{2+}, Cl^-, HCO_3^-, hardness), atomic absorption spectrometry (Mg^{2+}), turbidimetry (SO_4^{2-}), colorimetry (NO_2^-, PO_4^{3-}, NH_3, NH_4^+, Fe^{2+}), UV-spectroscopy (NO_3^-), filtration and drying (dry residue, suspension).

The results obtained from the initial dataset (130 samples × 19 characteristics) are presented in two tables. Table 1 shows the data statistics. These results are very informative and confirm that the chemical and physical features concerning the water quality of the Danube River are related to each other and so could be reduced. Table 2 is the table of components. It lists the eigenvalues of the correlation matrix considering only the first five principal components for PCA and FPCA, ordered from largest to smallest. This table also shows the proportion for each component.

3.1. Classical PCA

In the case of classical PCA considering all the original data the first component explains only 24.1% of the total variance and the second one 15.7%; a two component model, for example, thus accounts only for 39.8% of the total variance, Table 2.

Fig. 1 shows the plot of loadings and scores corresponding to the first two principal components of the original water samples. Sample 57 appears to be a very strong outlier in the

Table 1
Descriptive statistics of chemical and physical features concerning the water quality of the Danube River for a period of 11 consecutive years (all concentrations units are mg/L, dry residue and suspension in mg)

Variable	Mean	Median	Minimum	Maximum	Range	S.D.
PH	7.67	7.68	6.84	8.40	1.56	0.297
COD	33.37	30.54	15.76	66.21	50.45	9.41
EO_2	8.35	7.78	3.94	16.55	12.61	2.35
Ca^{2+}	57.55	56.00	16.03	104.00	87.97	12.67
Mg^{2+}	26.81	24.00	7.20	81.60	74.40	13.01
Ca^{2+}/Mg^{2+}	2.66	2.33	0.410	8.88	8.47	1.49
Fe^{2+}	0.269	0.250	0.001	0.925	0.924	0.177
Cl^-	57.96	54.93	25.90	138.81	112.91	19.66
$SO_4{}^{2-}$	93.86	80.00	25.00	400.00	375.00	61.66
$HCO_3{}^-$	215.21	206.73	87.13	426.40	339.27	56.29
$NO_2{}^-$	0.147	0.132	0.003	0.660	0.657	0.117
$NO_3{}^-$	3.62	3.19	0.420	16.96	16.54	2.43
$PO_4{}^{3-}$	0.479	0.337	0.001	1.63	1.63	0.393
NH_3	0.022	0.011	0.01	0.200	0.200	0.028
$NH_4{}^+$	0.619	0.388	0.01	3.39	3.39	0.678
Alkalinity	3.53	3.39	1.42	7.60	6.18	0.938
Hardness	13.89	13.18	8.96	25.80	16.84	3.23
Dry residue	389.45	372.5	152.50	670.00	517.50	90.72
Suspension	51.32	45.00	2.50	180.00	177.50	26.73

Table 2
Eigenvalues and proportion considering only the first five principal components for PCA and FPCA (initial data and without outlier)

Component	PCA				FPCA			
	Raw data (130 × 19)		Without outlier (129 × 19)		Raw data (130 × 19)		Without outlier (129 × 19)	
	Eig[a]	Prop[b]	Eig[a]	Prop[b]	Eig[a]	Prop[b]	Eig[a]	Prop[b]
1	4.59	24.1	3.46	18.2	10.23	89.5	10.76	91.7
2	2.98	15.7	2.50	13.1	0.25	2.2	0.19	1.6
3	1.85	9.8	1.96	10.3	0.20	1.7	0.16	1.4
4	1.72	9.0	1.65	8.7	0.15	1.3	0.12	1.0
5	1.30	6.9	1.52	8.0	0.12	1.0	0.10	0.8

[a] Eigenvalue.

[b] Proportion (%).

plot of scores as it is very distant from the other water samples. As a general rule, outliers should be deleted because of the least-squares property of principal components. In other words, a sample that is distant from the other points in the measurement space can pull the principal components towards it and away from the direction of maximum variance.

Fig. 2 shows the results of the first two principal components mapping experiment with sample 57 removed from the data. Again, graphing scores onto the plane described by PC1 and PC2 illustrates a random scatter of samples without well delimited classes and the scatterplot of loadings is more or less similar to the original representation.

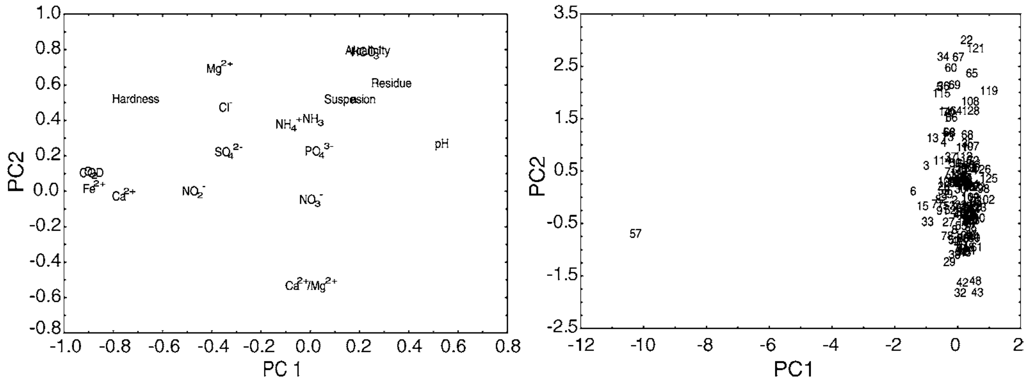

Fig. 1. Scatterplot of loadings and scores corresponding to the first two principal components (PCA, original data).

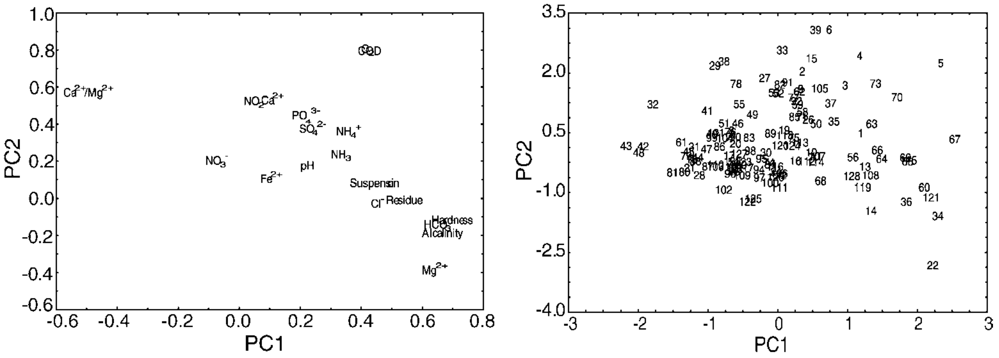

Fig. 2. Scatterplot of loadings and scores corresponding to the first two principal components (PCA, without outlier).

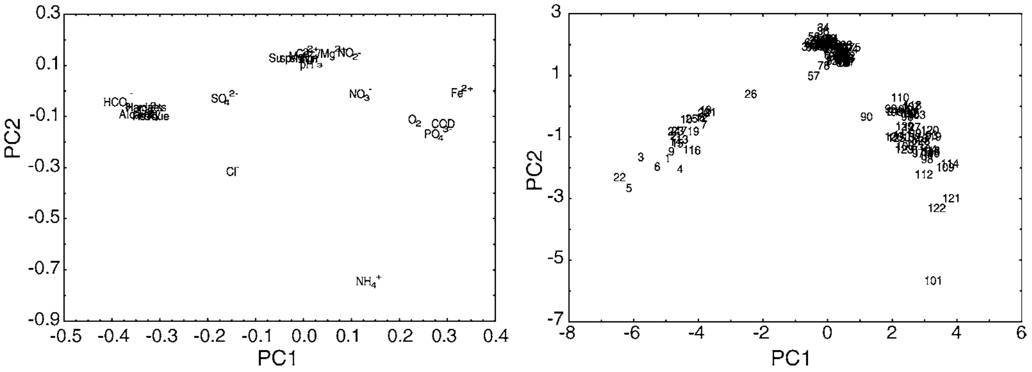

Fig. 3. Scatterplot of loadings and scores corresponding to the first two principal components (FPCA, initial data).

3.2. Fuzzy PCA

From the beginning we have to remark that the results obtained by applying FPCA are quite different from the PCA results. We can see that, for example, the first principal component explains 89.5% of the total variance and the second one 2.2%: a two component model thus accounts for 91.7% of the total variance (as compared to 39.8% for PCA) and a three components model accounts for 93.4% (as compared to 49.6% for PCA), for the fuzzy PCA method, Table 2. Hence, the FPCA-derived components account for significantly more of the variance than their classical PCA counterparts.

The first FPCA eigenvector illustrates that the greatest “negative” contribution to the first component is realized by the HCO_3^- (−0.387), alkalinity (−0.343), hardness (−0.331), dry residue (−0.322) and Ca^{2+} (−0.321); relative high positive contribution can be mentioned for Fe^{2+} (0.329), COD (0.289), PO_4^{3-} (0.280). The positive correlation between the last three quality parameters and also their negative correlation with, for example, HCO_3^-, alkalinity, might be explained considering Fe^{3+}/Fe^{2+} equilibrium, the simultaneous titration of Fe^{2+} with permanganate and, probably, the Fe^{3+}/PO_4^{3-} complex equilibrium. A less significant contribution is obtained from pH (0.005) and Mg^{2+} (−0.005). With

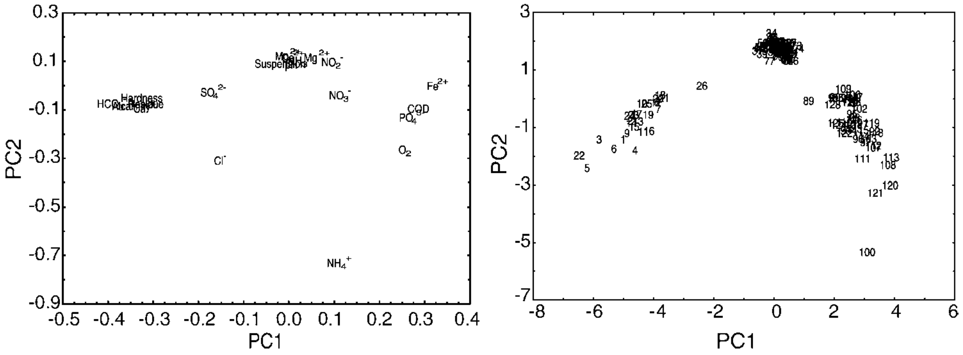

Fig. 4. Scatterplot of loadings and scores corresponding to the first two principal components (FPCA, without outlier).

respect to the second FPCA component the highest contribution was realized by NH_4^+ (−0.796), Cl^- (−0.360), PO_4^{3-} (−0.220).

Considering these results, it clearly appears that the first component might be considered as “the factor of alkalinity-hardness” and the second “the factor of ammonium salts”. The third component seems to be “the factor of the chemical oxygen demand” EO_2 (0.743), COD (0.316). These statements are very well confirmed by the 2D representations of loadings as is shown in Figs. 3 and 4. In addition, it is evident (see also Figs. 3 and 4) that the Danube water samples can be divided into three subgroups which would suggest that other conclusions should be drawn about the water quality history of the Galati region. As a consequence three distinct periods could be differentiated: the first one before 1990 (1985–1988), the second one around 1990 (1989–1993) and respectively after 1990 (1994–1996), in a good agreement to the economical activity changes in Romania. These subgroups are well cut along the first principal component mainly by the “alkalinity-hardness factor” and along the second principal component by “ammonium factor”.

4. Conclusions

A fuzzy principal component analysis method for robust estimation of principal components has been applied in this paper. The efficiency of the new algorithm was illustrated on a data set concerning the quality of the Danube River. The FPCA method achieved better results mainly because it is more compressible than classical PCA, i.e. the first fuzzy principal component accounts for significantly more of the variance than their classical counterparts. Considering, for example, a two component model in the case of all original data, FPCA accounts for 91.7% of the total variance, and the first two PCA components account only for 39.8%, Table 2. Additionally, PCA showed a relatively different separation of the variables and no separation of scores (samples) onto the plane described by the first two principal components, whereas a much sharper differentiation of the variables and scores is obtained when FPCA is used. These facts (greater accounting for total variance and shaper delineation of principal components) should encourage the application of fuzzy principal components analysis methodology to other database “mining” efforts as well as encourage “fuzzification” of other important environmetric methods like principle component regression (PCR) and partial least-squares (PLS) techniques. We appreciate, moreover, that using fuzzy principal component analysis it should be possible to explain some of the discrepancies, found in the literature, relating to multivariate analysis of data in terms of efficiency, goodness-of-fit, predictive power and robustness.

References

[1] D.L. Massart, B.G.M. Vandenginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A textbook, Elsevier, Amsterdam, 1988.

[2] R.G. Brereton, Chemometrics: Applications of Mathematics and Statistics to the Laboratory, Ellis Horwood, Chichester, 1990.

[3] J. Einax, H. Zwanziger, S. Geiß, Chemometrics in Environmental Analysis, John Wiley & Sons Ltd., Chichester, 1997.

[4] M. Otto, Chemometrics: Statistics and Computer Application in Analytical Chemistry, Wiley–VCH, Weinheim, 1999.

[5] M. Mellinger, Chemom. Intell. Lab. Syst. 2 (1987) 29–36.

[6] S. Wold, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.

[7] A. Mackiewicz, W. Ratajczak, Comput. Geosci. 19 (1993) 303–342.

[8] R. Vendrame, R.S. Braga, Y. Takahata, D.S. Galvao, J. Chem. Inf. Comput. Sci. 39 (1999) 1094–1104.

[9] W. Krawczyk, A. Parczewski, Anal. Chim. Acta 446 (2001) 107–114.

[10] V. Simeonov, C. Sârbu, D.L. Massart, S. Tsakovski, Mikrochim. Acta 137 (2001) 243–248.

[11] I.M. Farnham, A.K. Singh, K.J. Stetzenbach, K.H. Johnnesson, Chemom. Intell. Lab. Syst. 60 (2002) 265–281.

[12] K. Kafadar, Chemom. Intell. Lab. Syst. 60 (2002) 127–134.

[13] P.J. Rousseeuw, J. Am. Stat. Assoc. 79 (1984) 871–880.

[14] P.J. Rousseeuw, A. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.

[15] L. Davies, Ann. Stat. 15 (1987) 1269–1292.

[16] G. Li, Z. Chen, J. Am. Stat. Assoc. 80 (1985) 759–766.

[17] M. Hubert, P.J. Rousseeuw, S. Verboven, Chemom. Intell. Lab. Syst. 60 (2002) 101–111.

[18] M. Hubert, S. Engelen, Bioinformatics 20 (2004) 1728–1736.

[19] L.A. Zadeh, Inf. Control 8 (1965) 338–353.

[20] J.C. Bezdek, R. Ehrlich, W. Full, Comput. Geosci. 10 (1984) 191–203.

[21] D. Dumitrescu, C. Sârbu, H.F. Pop, Anal. Lett. 24 (1994) 1031–1054.

[22] H.F. Pop, C. Sârbu, O. Horowitz, D. Dumitrescu, J. Chem. Inf. Comput. Sci. 36 (1996) 465–482.

[23] C. Sârbu, H.F. Pop, Chemosphere 40 (2000) 513–520.

[24] T.N. Yang, S.-D. Wang, Pattern Recognition Lett. 20 (1999) 927–933.

[25] C. Sârbu, H.F. Pop, Talanta 54 (2001) 125–130.

[26] C. Sârbu, H.F. Pop, Fuzzy soft-computing methods and their applications in chemistry, in: K.B. Lipkowitz, D.B. Boyd, T.R. Cundari (Eds.), Reviews in Computational Chemistry, Wiley–VCH, 2004, pp. 249–332. Chapter 5.

[27] T.R. Cundari, J. Deng, H.F. Pop, C. Sârbu, J. Chem. Inf. Comput. Sci. (2000) 40.

[28] H.F. Pop, C. Sârbu, Anal. Chem. 68 (1996) 771–778.

[29] H.F. Pop, C. Sârbu, Rev. Chim. (Bucharest) 48 (1997) 888–891.

[30] P. Popa, R. Mocanu, C. Sârbu, Rev. Chim. (Bucharest) 49 (1998) 846–854.