Robust Fuzzy Principal Component Analysis (FPCA). A Comparative Study Concerning Interaction of Carbon−Hydrogen Bonds with Molybdenum−Oxo Bonds

Thomas R. Cundari,*,† Costel Sârbu,‡ and Horia F. Pop‡

Department of Chemistry, Computational Research on Materials Institute (CROMIUM), The University of Memphis, Memphis, Tennessee 38152-6060, and Departments of Chemistry and Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania

Received April 2, 2002

Principal component analysis (PCA) is a favorite tool in chemometrics 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 approaches to improve PCA appears to be the fuzzification of the matrix data, thus diminishing the influence of outliers. In this paper we discuss a robust fuzzy PCA algorithm (FPCA). The new algorithm is illustrated on a data set concerning interaction of carbon−hydrogen bonds with transition metal−oxo bonds in molybdenum complexes. Considering, for example, a two component model, FPCA accounts for 97.20% of the total variance and PCA accounts only for 69.75%.

INTRODUCTION

Several statistical methods for the analysis of large quantities of data have been applied to chemical problems during the past decades.[1−6] 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 chemical situations.[7−17]

PCA is designed to reduce the number of variables that need to be considered to a small number of indices (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.

In the case of a *p*-dimensional problem, often the number of components needed to describe, say 90% of the sample variance, is considerably less than *p*, so that PCA essentially affords a technique whereby the dimensionality of the variable space can be reduced. It may well turn out that usually two or three principal components provide a good summary of all the original variables. Moreover, PCA offers a second important tool for multidimensional analysis that derives, in fact, from its original application in the social sciences and from which it took its name. In other words, PCA can also reveal those underlying factors or combinations of the original variables that principally determine the structure of the data distribution and that not infrequently are related to some real influencing factors in the sample population. An important issue in PCA is the interpretation of components, to help determine after the reduction of the observation space, which initial variables have the greatest shares in the variance of particular principal components. This information can be obtained using coefficients of determination (loadings) established between the components and the initial variables.

THEORETICAL CONSIDERATIONS

Classical Principal Component Analysis. Essentially the mathematical basis of PCA rests on eigenanalysis of the covariance or correlation matrix. Eigenanalysis of a matrix **M** involves finding unique pairs of vectors $\mathbf{e}_\mathrm{i}$ and scalars λ_i, called eigenvectors and eigenvalues, respectively such that the following equation is satisfied

$$\mathbf{M}.\mathbf{e}_\mathrm{i} = \lambda_\mathrm{i}.\mathbf{I}.\mathbf{e}_\mathrm{i} \tag{1}$$

where **I** is the identity matrix.

The principal components appear as linear combinations of the original variables in the form

$$PC = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p \tag{2}$$

where X_i represents the original variables and a_{ij} the elements of the eigenvectors $\mathbf{e}_\mathrm{i}$. A constraint that $a_{i1}^2 + a_{i2}^2 + ... + a_{\mathrm{ip}}^2 = 1$ is imposed on all components. The constraint is introduced because if this is not done then var(PC_i) can be increased by simply increasing any one of the a_{ij} values.

Since the vectors $\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_\mathrm{n}$ are orthonormal, that is

$$\mathbf{e}_i^\mathrm{T}.\mathbf{e}_i = 1, \quad \mathbf{e}_i.\mathbf{e}_j = 0 \quad \text{for } i \neq j \tag{3}$$

it is easy to observe also that we have

$$\mathbf{e}_i^\mathrm{T}\mathbf{M}\mathbf{e}_i = \lambda_i, \quad \mathbf{e}_i\mathbf{M}\mathbf{e}_j = 0 \quad \text{for } i \neq j \tag{4}$$

and

* Corresponding author phone: (940)369-7753; fax: (940)565-4318; e-mail: tomc@unt.edu.

† The University of Memphis.

‡ Babes-Bolyai University.

10.1021/ci025524s CCC: $22.00
Published on Web 09/19/2002

$$\mathbf{M} = \lambda_1 \mathbf{e}_1{}^{\mathrm{T}}\mathbf{e}_1 + \lambda_2 \mathbf{e}_2{}^{\mathrm{T}}\mathbf{e}_2 + ... + \lambda_n \mathbf{e}_p{}^{\mathrm{T}}\mathbf{e}_p \qquad (5)$$

The expression (5) is named a spectral decomposition of a matrix **M**. The basic property of the new variables is the lack of correlation among them (in the contrast to the initial variables). The variance of the *i*th component is λ_i or

$$\mathrm{Var}(\mathbf{e}_i X) = \lambda_i \qquad (6)$$

where

$$\mathrm{Cov}(\mathbf{e}_i X, \mathbf{e}_j X\,) = 0 \quad \text{for } i \neq j \qquad (7)$$

The first principal component *PC1* is that linear combination of sample values for which the "scores" have maximum variation. The second component *PC2* has scores that are uncorrelated with the scores for *PC1*. Among the many linear combinations with this property we select that which has maximum variation among its scores. The third component *PC3* is defined to be that linear combination which has the maximum variation among all those combinations whose scores are uncorrelated with the scores of the first two components. Subsequent components are defined analogously.

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 data transformations have a large impact upon PCA.[18,19]

One of the most promising approaches to "robustify" PCA appears to be the fuzzification of the matrix data to diminish the influence of outliers.

Fuzzy Principal Component Analysis. Fuzzy clustering is an important tool to identify the structure in data.[20−23] In general, a fuzzy clustering algorithm with objective function can be formulated as follows: let $X = \{x^1, ..., x^n\} \subset \mathbf{R}^p$ be a finite set of feature vectors, where *n* is the number of objects (measurements) and *p* is the number of the original variables, $x_k{}^j = [x_1{}^j, x_2{}^j, ..., x_p{}^j]^{\mathrm{T}}$ and $L = (L^1, L^2, ..., L^s)$ 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[24−29]

$$J(P,L) = \sum_{i=1}^{s}\sum_{j=1}^{n}(A_i(x^j))^2 d^2(x^j,L^i) \qquad (8)$$

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

$$d(x^j,L^i) = ||x^j - L^i|| = [\sum_{k=1}^{p}(x_k{}^j - L^i{}_k)^2]^{1/2} \qquad (9)$$

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 $J(\bullet,L)$ is obtained for

$$A_i(x^j) = \frac{1}{\displaystyle\sum_{k=1}^{s}\frac{d^2(x^j,L^i)}{d^2(x^j,L^k)}}, i = 1,...,s \qquad (10)$$

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

$$L^i = \frac{\displaystyle\sum_{j=1}^{n}[A_i(x^j)]^2 x^j}{\displaystyle\sum_{j=1}^{n}[A_i(x^j)]^2}, i = 1,...,k \qquad (11)$$

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,[24−29] 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 (12), which is a slight generalization for fuzzy sets of the classical covariance matrix:

$$C_{kl} = \frac{\displaystyle\sum_{j=1}^{n}[A_i(x^j)]^2(x_{jk} - \bar{x}_k)(x_{jl} - \bar{x}_l)}{\displaystyle\sum_{j=1}^{n}[A_i(x^j)]^2} \qquad (12)$$

The algorithm proposed 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.[30−33]

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). [For the reason the values of α equal to 0 and 1 are excluded,

Table 1. Descriptive Statistics of Structural Features Considered[a]

variable	mean	median	minimum	maximum	range	SD	skewness	kurtosis
α	59	61	6	116	110	24	0.029	−0.470
β	50	53	3	85	82	20	−0.591	−0.335
γ	119	115	82	176	94	22	0.742	−0.118
MC	4.57	4.53	4.14	5.26	1.12	0.276	0.801	0.442
MCH	52	53	7	100	93	22	0.127	−0.609
MHC	116	115	66	171	105	25	0.061	−0.701
MO	1.68	1.68	1.66	1.70	0.040	0.010	0.164	−0.320
MOCH	2	−12	−178	174	352	100	−0.061	−1.08
MOH	117	118	80	174	94	24	0.567	−0.444
OH	3.07	3.11	2.37	3.98	1.61	0.394	0.183	−0.654
OHC	112	113	47	172	125	27	0.147	−0.387
OMC	42	44	5	75	70	16	−0.542	−0.282
R_{bp}	3.73	3.75	3.29	3.98	0.69	0.161	−0.381	−0.450
OC	3.56	3.54	3.11	4.41	1.30	0.269	0.592	0.424

[a] Bond lengths (MC, MO, OH, OC) are reported in angstrom units; bond angles (MCH, MHC, MOH, OHC, OMC) and dihedral angles (MOHC) are reported in degrees.

Table 2. Correlation Matrix for Metric Variables

	α	β	γ	MC	MCH	MHC	MO	MOCH	MOH	OH	OHC	OMC	R_{bp}	OC
α	1.00	0.18	−0.19	−0.55	0.96	−0.96	−0.32	0.05	−0.18	0.67	−0.95	0.14	0.24	−0.39
β		1.00	−1.00	−0.74	0.11	−0.12	−0.17	−0.05	−0.97	0.68	−0.26	0.97	−0.06	0.59
γ			1.00	0.75	−0.11	0.12	0.17	0.04	0.98	−0.66	0.26	−0.97	0.09	−0.57
MC				1.00	−0.50	0.52	0.35	0.05	0.72	−0.52	0.57	−0.72	0.40	0.03
MCH					1.00	−1.00	−0.30	0.05	−0.06	0.58	−0.84	0.12	0.28	−0.37
MHC						1.00	0.32	−0.04	0.07	−0.58	0.84	−0.13	−0.25	0.37
MO							1.00	−0.02	0.15	−0.23	0.28	−0.19	0.05	0.04
MOCH								1.00	0.05	0.04	−0.03	−0.04	0.10	−0.02
MOH									1.00	−0.67	0.32	−0.90	0.11	−0.50
OH										1.00	−0.71	0.62	0.52	0.34
OHC											1.00	−0.16	−0.16	0.39
OMC												1.00	−0.03	0.66
R_{bp}													1.00	0.40
OC														1.00

see ref 30.] As a consequence the criterion function becomes

$$J(A,L;\alpha) = \sum_{j=1}^{n}[A(x^j)]^2 d^2(x^j,L) + \sum_{j=1}^{n}[\bar{A}(x^j)]^2 \frac{\alpha}{1-\alpha} \qquad (13)$$

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

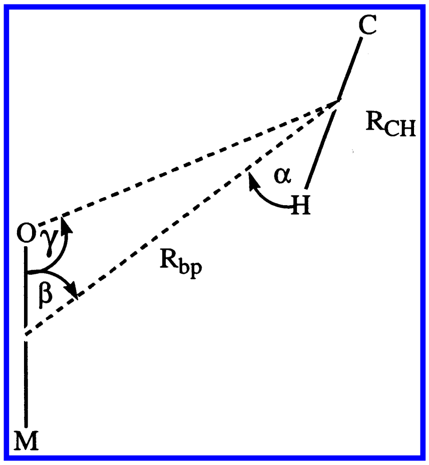
$$v = \sum_{j=1}^{n}[A(x)^j]^2 x^j / \sum_{j=1}^{n}[A(x^j)]^2 \qquad (14)$$

where

$$A(x^j) = \frac{\alpha/(1-\alpha)}{[\alpha/(1-\alpha)] + d^2(x^j,L)} \qquad (15)$$

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.

Scheme 1



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 (13) and, with the optimal membership

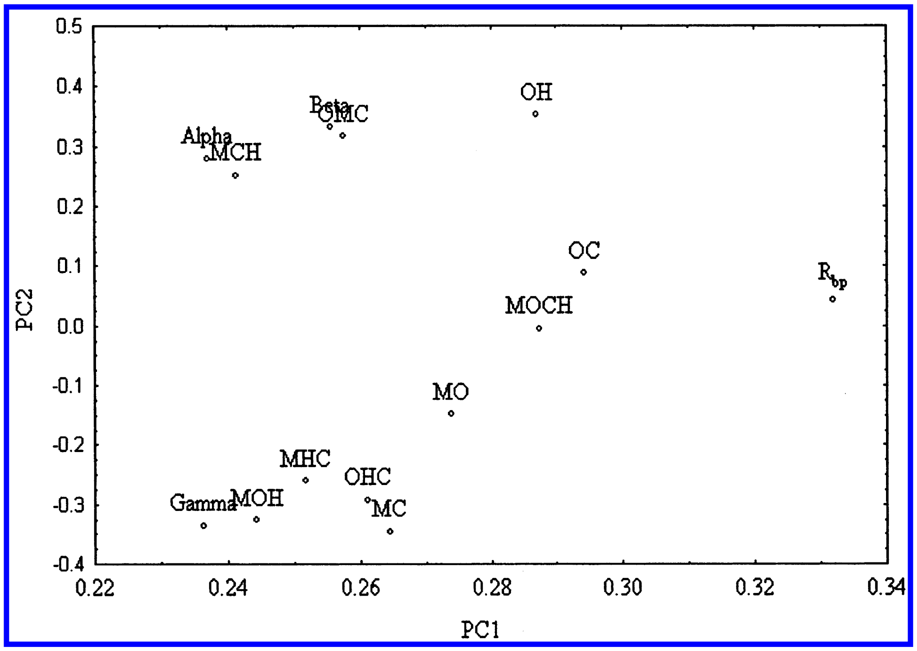


Figure 1. Loadings of the second PCA component plotted versus the loadings of the first component.

Table 3. Eigenvalues and the Ratios of the Variance Explained by Each Component

	PCA			FPCA		
component	eigenvalue	proportion %	cumulative %	eigenvalue	proportion %	cumulative %
1	718.95	**43.43**	**43.43**	1787.57	**57.42**	**57.42**
2	435.62	**26.32**	**69.75**	1238.30	**39.78**	**97.20**
3	274.62	**16.59**	**86.34**	33.26	**1.07**	**98.27**
4	82.33	4.98	91.32	20.68	0.67	98.94
5	65.89	3.97	95.29	12.03	0.38	99.32
6	52.76	3.20	98.49	11.18	0.36	99.68
7	21.18	1.28	99.77	8.69	0.28	99.96
8	1.87	0.11	99.88	0.872	0.03	99.99
9	1.30	0.08	99.96	0.202	0.01	100.00
10	0.429	0.02	99.98	0.063	0.00	100.00
11	0.165	0.01	99.99	0.025	0.00	100.00
12	0.051	0.01	100.00	0.017	0.00	100.00
13	0.024	0.00	100.00	0.006	0.00	100.00
14	0.021	0.00	100.00	0.005	0.00	100.00

Table 4. Loadings Corresponding to First Three Principal Components for PCA and FPCA

	PCA			FPCA		
variable	PC1	PC2	PC3	PC1	PC2	PC3
α	0.237	0.279	0.381	0.031	0.341	0.394
β	0.256	0.332	0.264	−0.048	0.346	−0.209
γ	0.236	−0.336	−0.300	0.443	−0.041	0.257
MC	0.265	−0.345	−0.046	0.369	−0.050	−0.101
MCH	0.241	0.252	−0.396	0.106	0.352	0.437
MHC	0.252	−0.260	0.353	0.315	−0.057	−0.387
MO	0.274	−0.149	0.062	0.172	0.106	−0.003
MOCH	0.287	−0.005	−0.076	0.273	0.018	0.020
MOH	0.244	−0.326	−0.295	0.478	−0.001	0.266
OH	0.287	0.353	−0.060	−0.123	0.484	0.004
OHC	0.261	−0.292	0.292	0.438	−0.004	−0.267
OMC	0.258	0.317	0.271	0.004	0.367	−0.205
R_{bp}	0.332	0.041	−0.141	0.094	0.376	−0.033
OC	0.294	0.088	0.365	0.076	0.333	−0.445

degrees $A(x^j)$, compute the largest eigenvalue of the matrix C given by (12). Select the optimal value of α according to the maximal eigenvalue.

2. Coding the variables $X_1, X_2, ..., X_p$ to have zero means and unit variances. This is usual but is omitted in some cases.

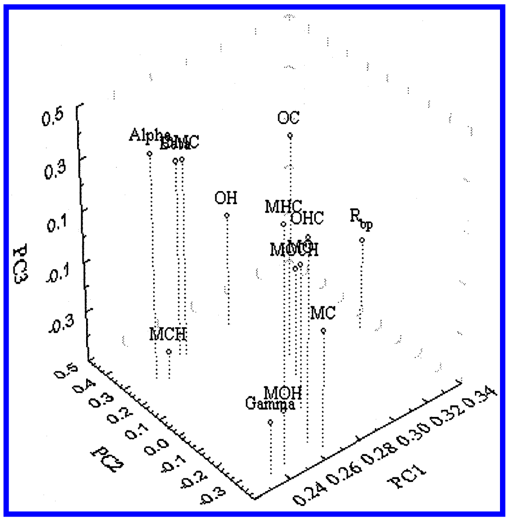


Figure 2. 3D plotting of the loadings corresponding to the first three PCA components.

3. Calculate the covariance matrix **C**, as given by relation (12). This is a correlation matrix if step 2 has been done.

4. Find the eigenvalue λ_1 and the corresponding eigenvector $\mathbf{e}_1$.

5. Determine the new fuzzy set A^{l+1} using eq 15.

6. If the fuzzy sets $A^{(l+1)}$ and $A^{(l)}$ are close enough, i.e., if

$$||A^{(l+1)} - A^{(l)}|| < \epsilon$$

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 (12) and determine its eigenvalues and eigenvectors as usually; these are the fuzzy principal components and the corresponding scatter values.

DATABASE GENERATION

A search of the Cambridge Structural Database[34] is performed for complexes with appreciable, intermolecular

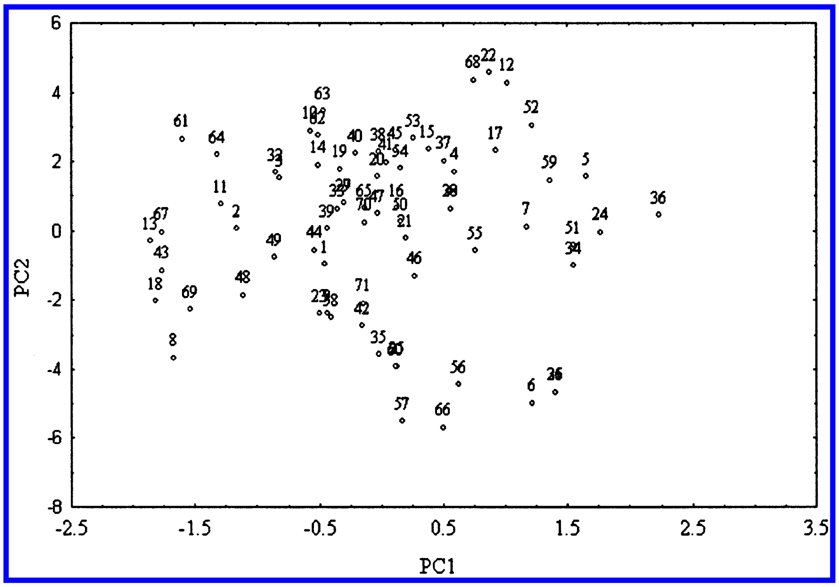


Figure 3. Plot of the scores of the second component versus the scores of the first component (PCA).

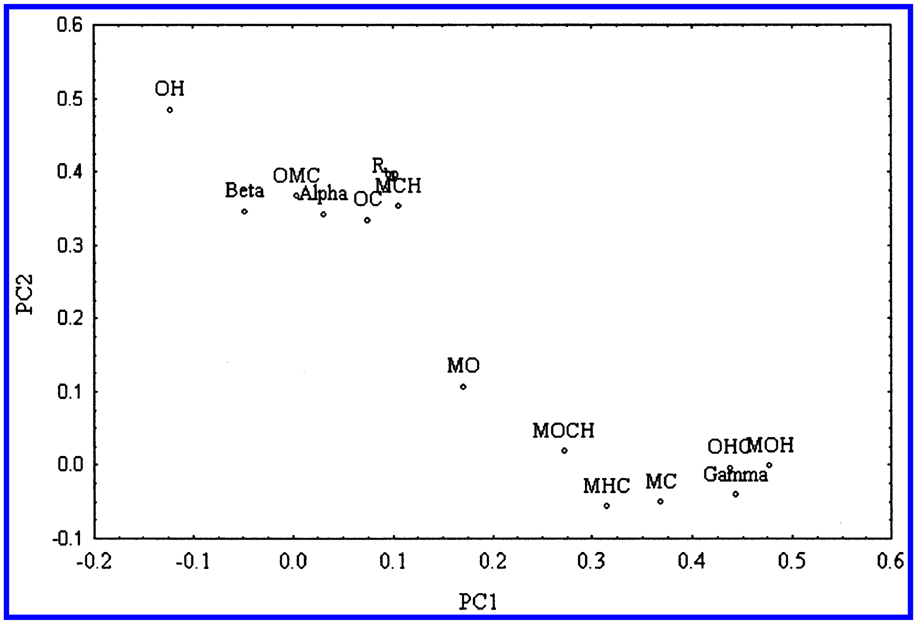


Figure 4. Loadings of the second FPCA component plotted versus the loadings of the first component.

Mo═O- - -H—C contacts. The working definition of "appreciable" is $R_{bp} < 3.75$ Å. The metric R_{bp} is taken as a measure of the strength of interaction between the MoO and CH moieties, i.e., as R_{bp} descreases, the strength of the MoO—HC interaction increases and *vice versa*. Interactions are considered intermolecular if they involve the C—H bond of one molecular with the Mo═O functionality in another. Refinements are employed to ensure the search focused on the most reliable crystal structures: INSIST-ON-COORDS (to search only systems in which fractional atomic coordinates are deposited), INSIST-NO-DISORDER (to search systems with no crystallographic disorder), INSIST-PERFECT-MATCH (to search entries with completely matched chemical and crystallographic connectivities), INSIST-ERROR-FREE (to include only entries whose published bond lengths agree with the recalculated values to within 0.05 Å), and INSIST-NO-POLYMERS (to exclude entries with polymeric bonds in the crystal connectivity). To counteract difficulties in locating H atoms in X-ray diffraction, C—H distances are set to the neutron diffraction average, 1.083 Å.

RESULTS AND DISCUSSION

Principal component analysis was performed on the structural features concerning interaction of carbon—hydrogen bonds with molybdenum—oxo bonds; in addition to all possible atom—atom distances and the angles subtended thereof, several additional metrics were defined and tabulated, Scheme 1. These include the distance R_{bp} and the angles α, β, and γ as well as the dihedral MOCH Scheme 1. Interactions between metal—oxo and carbon—hydrogen bonds are of importance with respect to microbial and industrial oxidation,[35,36] and for these reasons molybdenum was the focus of this research.

The results obtained from the initial data set (71 objects × 14 characteristics)—covariance matrix—are presented in four tables. Table 1 shows the data statistics. These results

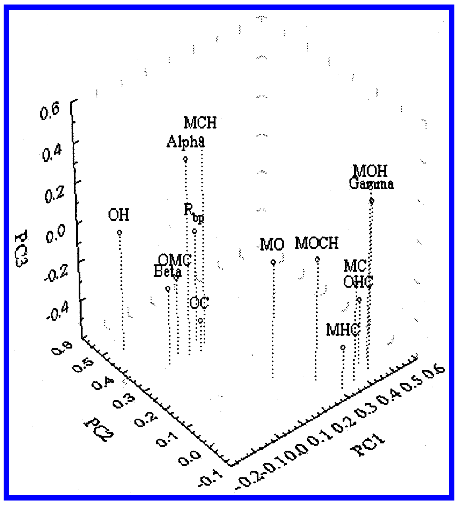


Figure 5. 3D plotting of the loadings corresponding to the first three FPCA components.

and the correlation data in Table 2 confirm that the structural features concerning interaction of C−H and Mo═O bonds are related to each other and so could be reduced. Table 3 is the table of components. It lists the eigenvalues of the covariance matrix, ordered from largest to smallest. This table also shows the proportion and cumulative proportion for each component. Table 4 displays the eigenvectors associated with each of the components in Table 3. The first eigenvector goes with the first eigenvalue, and so on.

Classical PCA. In the case of classical PCA the first component explains only 43.43% of the total variance and the second one 26.32%; a two component model, for example, thus accounts only for 69.75% of the total variance, Table 3.

The column corresponding to the first eigenvector obtained by PCA indicates that the contribution to the first component is very similar, Table 4. However, the greatest contribution is realized by the R_{bp} metric (0.332), the next highest is the OC distance (0.294), OH (0.287), and MOCH (0.287). It is interesting to observe that the contribution to the first principal component of OH and MOCH is practically the same and the smallest was realized by the angles α and γ.

The second component describes remaining variance after the first component is removed from the data. In this case the highest contribution to PC2 is realized by the OH (0.353), the second being β (0.332). It is very useful to remark that, for instance, γ (−0.336), MC (−0.345), MOH (−0.326) are very similar but the values are negative; this suggests that the values of these characteristics decrease when, for example, OH and β increase (negative correlation). The lowest contribution to PC2 is given by MOCH (−0.005), R_{bp} (0.041), and OC (0.088). For now, we observe that these results are more or less in agreement with the correlation data in Table 2 and also note the position of R_{bp} as an outlier. This statement is well supported by the 2D (*i.e.*, PC1 versus PC2) and 3D (*i.e.*, PC1 versus PC2 versus PC3) representations of loadings presented in Figures 1 and 2. In addition, graphing scores onto the plane described by PC1 and PC2 (Figure 3) illustrates a random scatter of compounds without well delimited classes.

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 57.42% of the total variance and the second one 39.78%: a two component model thus accounts for 97.20% of the total variance (as compared to 69.75% for PCA) and a three components model accounts for 98.27% (as compared to 86.34% for PCA), for the fuzzy PCA method, Table 3. Hence, the FPCA-derived components account for significantly more of the variance than their classical PCA counterparts.

The data in Table 4 corresponding to the first FPCA eigenvector illustrates that the greatest contribution to the first component is realized by the MOH (0.478), γ (0.443), OHC (0.438), MC (0.369), MHC (0.315), and MOCH (0.273), Table 4. A less significant contribution is obtained from MO (0.172) and MCH (0.106). With respect to the second FPCA component it is easy to observe that the highest contribution is realized by OH (0.484), R_{bp} (0.376), OMC (0.367), MCH (0.352), β (0.346), α (0.341), and OC (0.333), Table 4. One may conclude that R_{bp} depends mainly on these structural features. Considering these results, it clearly appears that the first component might be considered as “the factor of metal−oxo complex” and the second “the factor of R_{bp}”. The third component seems to be “the factor of reverse correlation” OC (−0.445), MCH (0.437), α (0.394),

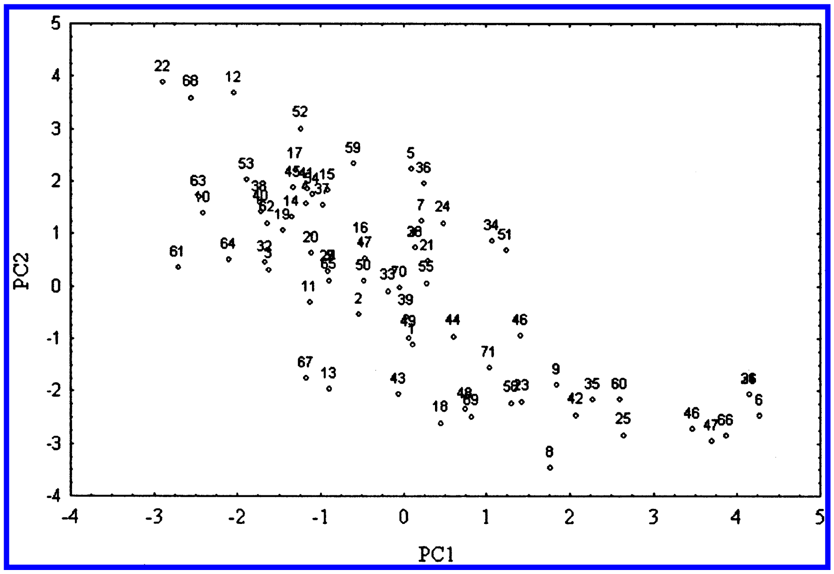


Figure 6. Plot of the scores of the second component versus the scores of the first component (FPCA).

MHC (−0.387), OHC (−0.267), MOH (0.266), γ (0.257), β (−0.209), and OMC (−0.205). These statements are very well confirmed by the 2D and 3D representation of loadings as is shown in Figures 4 and 5.

CONCLUSIONS

A fuzzy principal component analysis (FPCA) method for robust estimation of principal components has been described in this paper. The efficiency of the new algorithm was illustrated on a data set concerning interaction of carbon−hydrogen bonds with molybdenum−oxo bonds. The FPCA method achieved better results mainly because it is more compressible than classical PCA. Considering, for example, a two component model, FPCA accounts for 97.20% of the total variance, and the first two PCA components account only for 69.75%, Table 3. It would take five to six classical principal components, Table 3, to account for the same total variance as two fuzzy principal components. The latter is, of course, much more desirable makes for a significantly easier data set analysis.

Additionally, PCA showed only a partial separation of the molybdenum−oxo complexes onto the plane described by the first two principal components, Figure 3, whereas a much sharper differentiation of the metric variables along the diagonal is observed when FPCA is used, Figure 6. Some of them are well grouped in the lower right corner separated more or less from the majority situated upper left corner, Figure 6.

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 chemometric methods such as principle component regression (PCR) and partial least squares (PLS) techniques.

ACKNOWLEDGMENT

This research and the stay in Memphis of C.S. and H.F.P. are sponsored by the U.S. National Science Foundation, International Programs grant number INT-0077531.

REFERENCES AND NOTES

(1) Massart, D. L.; Kaufman, L. *The Interpretation of Analytical Chemical Data by the Use of* Cluster Analysis; Wiley: New York, 1983.

(2) Brereton, R. G. Chemometrics: *Applications of Mathematics and Statistics to the Laboratory*; Ellis Horwood; Chichester, 1990.

(3) Meloun, M.; Mlitky, J.; Forina, M. *Chemometrics for Analytical Chemistry, vol I: PC-aided statistical data analysis*; Ellis Horwood; Chichester, 1992.

(4) Einax, J.; Zwanziger, H.; Geiss, S. *Chemometrics in Environmental Analysis*; John Wiley & Sons Ltd.: Chichester, 1997.

(5) Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics, Part B*; Elsevier: Amsterdam, 1998.

(6) Mellinger, M. Multivariate Data Analysis: Its Methods. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 29−36.

(7) Wold, S. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.

(8) Mackiewicz, A.; Ratajczak, W. Principal Component Analysis (PCA). *Comput. Geosci*. **1993**, *19*, 303−342.

(9) Katritzky, A. R.; Tamm, T.; Wang Y.; Karelson, M. A Unified Treatment of Solvent Properties. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 692−698.

(10) Vendrame, R.; Braga, R. S.; Takahata, Y.; Galvao, D. S. Structure−Activity Relationship Studies of Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons Using Calculated Molecular Descriptors with Principal Component Analysis and Neural Network Methods. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1094−1104.

(11) Benigni, R.; Passerini, L.; Livingstone, D. J.; Johnson, M. A.; Giuliani, A. Infrared Spectra Information and Their Correlation with QSAR Descriptors. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 558−562.

(12) De Groot, P. J.; Postma, G. J.; Melssen, W. W.; Buydens, L. M. S.; Deckert, V.; Zenobi, R. Application of Principal Component Analysis to Detect Outliers and Spectral Deviations in Near-Field Surface-Enhanced Raman Spectra. *Anal. Chim. Acta* **2001**, *446*, 71−83.

(13) Krawczyk, W.; Parczewski, A. Application of Chemometric Methods in Searching for Illicit Leuckart Amphetamine Sources. *Anal. Chim. Acta* **2001**, *446*, 107−114.

(14) Statheropoulas, M.; Mikedi, K. PCA − ContVarDia: An Improvement of the PCA − VarDia Technique for Curve Resolution in GC-MS and TG-MS Analysis. *Anal. Chim. Acta* **2001**, *446*, 353−370.

(15) Airian, C. Y.; Shen, H.; Brereton, R. G. PCA in Liquid Chromatography Proton Nuclear Magnetic Resonance: Differentiation of Three Regio-Isomers. *Anal. Chim. Acta* **2001**, *447*, 199−210.

(16) Simeonov, V.; Sârbu, C.; Massart, D. L.; Tsakovski, S. Danube River Data Modelling by Multivariate Data Analysis. *Mikrochim. Acta* **2001**, *137*, 243−248.

(17) Farnham, I. M.; Singh, A. K.; Stetzenbach K. J.; Johnnesson K. H. Treatement of Nondetects in Multivariate Analysis of Groundwater Geochemistry Data. *Chemom. Intell. Lab. Syst*. **2002**, *60*, 265−281.

(18) Kafadar, K. The Influence of John Tukey’s Work in Robust Methods for Chemometrics and Environmetrics. *Chemom. Intell. Lab. Syst*. **2002**, *60*, 127−134.

(19) Hubert, M.; Rousseeuw, P. J.; Verboven, S. A Fast Method for Robust PrincipalComponents with Applications to Chemometrics. *Chemom. Intell. Lab. Syst*. **2002**, *60*, 101−111.

(20) Zadeh, L. A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338−353.

(21) Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Functions Algorithms*; Plenum Press: New York, 1981.

(22) Kandel, A. *Fuzzy Techniques in Pattern Recognition*; Wiley: New York, 1982.

(23) Bezdek, J. C.; Ehrlich, R.; Full, W. FCM. The Fuzzy C Means Clustering Algorithm. *Comput. Geosci*. **1984**, *10*, 191−203.

(24) Dumitrescu, D.; Sârbu, C.; Pop, H. F. A Fuzzy Divisive Hierarchical Clustering Algorithmfor the Optimal Choice of Sets of Solvent Systems. *Anal. Lett.* **1994**, *24*, 1031−1054.

(25) Pop, H. F.; Dumitrescu, D.; Sârbu, C.; A Study of Roman Pottery (terra sigillata) UsingHierarchical Fuzzy Clustering. *Anal. Chim. Acta* **1995**, *310*, 269−279.

(26) Pop, H. F.; Sârbu, C.; Horowitz, O.; Dumitrescu, D. A Fuzzy Classification of the Chemical Elements. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 465−482.

(27) Sârbu, C.; Pop, H. F. Fuzzy Clustering Analysis of the First 10 MEIC Chemicals. *Chemosphere* **2000**, *40*, 513−520.

(28) Cundari, T. R.; Deng, J.; Pop, H. F.; Sârbu, C. Structural Analysis of Transition Metal β-X Substituent Interactions. Toward the Use of Soft Computing Methods for Catalyst Modeling. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1052−1061.

(29) Sârbu, C.; Pop, H. F. Fuzzy Classification and Comparison of Some Romanian andAmerican Coals. *Commun. Math. Comput. Chem*. (*Match*) **2001**, *44*, 387−400.

(30) Pop, H. F.; Sârbu, C. A New Fuzzy Regression Algorithm. *Anal. Chem.* **1996**, *68*, 771−778.

(31) Sârbu, C.; Pop, H. F. Fuzzy Regression. Heteroscedasticity. *Rev. Chim*. (Bucharest) **1997**, *48*, 732−737.

(32) Pop, H. F.; Sârbu, C. Fuzzy Regression. Outliers. *Rev. Chim*. (Bucharest) **1997**, *48*, 888−891.

(33) Sârbu, C.; Pop, H. F. Fuzzy Robust Estimation of Central Location. *Talanta*, **2001**, *54*, 125−130.

(34) Allen F. H.; Kennard O. 3D Search and Research using the Cambridge Structural Database. *Chem. Design Autom. News* **1993**, *8*, 31−37.

(35) Nugent, W. A.; Mayer, J. M. *Metal−Ligand Multiple Bonds*; Wiley: New York, 1988.

(36) Mayer, J. M Hydrogen Atom Abstraction by Metal-Oxo Complexes: Understanding the Analogy with Organic Radical Reactions. *Acc. Chem. Res.* **1998**, *31*, 441−450.

CI025524S