International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:7, 2012

Visualization and Indexing of Spectral Databases

Tibor Kulcsar, Gabor Sarossy, Gabor Bereznai, Robert Auer, Janos Abonyi

Abstract—On-line (near infrared) spectroscopy is widely used to support the operation of complex process systems. Information extracted from spectral database can be used to estimate unmeasured product properties and monitor the operation of the process. These techniques are based on looking for similar spectra by nearest neighborhood algorithms and distance based searching methods. Search for nearest neighbors in the spectral space is an NP-hard problem, the computational complexity increases by the number of points in the discrete spectrum and the number of samples in the database. To reduce the calculation time some kind of indexing could be used. The main idea presented in this paper is to combine indexing and visualization techniques to reduce the computational requirement of estimation algorithms by providing a two dimensional indexing that can also be used to visualize the structure of the spectral database. This 2D visualization of spectral database does not only support application of distance and similarity based techniques but enables the utilization of advanced clustering and prediction algorithms based on the Delaunay tessellation of the mapped spectral space. This means the prediction has not to use the high dimension space but can be based on the mapped space too. The results illustrate that the proposed method is able to segment (cluster) spectral databases and detect outliers that are not suitable for instance based learning algorithms.

Keywords—indexing high dimensional databases, dimensional reduction, clustering, similarity, k-nn algorithm.

I. Introduction

NEAR Infrared spectroscopy with Topological Mapping (TOPNIR) is widely used in oil industry to estimate product properties (e.g. aromatic components, cloud point, flash point, density etc.) of products and process streams [11]. TOPNIR performs a two dimensional mapping of the spectral space to visualize the operation regimes of the process. The key idea of this paper similar to low dimensional mappings can also be utilized to index the spectral database by giving small number f primary key variables, and sophisticated prediction and clustering algorithms [12], [14], [15] can be developed based on this indexing.

The authors (T.Kulcsar, J.Abonyi) are with the University of Pannonia Department of Process Engineering, Veszprm, H-8200, Hungary (phone: +36-70-944-8910, e-mail: janos@abonyilab.com)

The authors (R.Auer, G.Bereznai, G.Sarossy) are with the MOL Ltd. Department of DS Development Analytics MOL Hungarian Oil and Gas Plc. R&M Division, DS Development H-2443 Szzhalombatta, POB. 1.

The performance of dimensional reduction based indexing algorithms can be measured by the distance and neighborhood preserving properties of the mappings. In this paper the most important dimensional reduction techniques (aggregates, PCA - $Principal\ Component\ Analysis$) are applied and $Metric\ error$ and $trustworthy$; properties of the mappings are calculated.

The performance of instance based learning algorithms highly depends on the quality of the database used for estimation. Hence, data-driven modeling algorithms needs carefully designed and maintained training data. The coverage of the operating regimes and the structure of the indexed database should be consistent to support the fast searching for the nearest neighbors. The studied indexing techniques can also be used to visualize the structure the database and detect dense and compact operating regimes. For this propose a Delaunay triangulation based measure has been developed where the area of the triangles are indirectly proportional to the data coverage. This information about the coverage of the dataset is important because in spare areas the estimation models based on nearest neighbors (like TOPNIR's k-nn regression technique) show usually bad modeling performance that necessitate the insertion of new datapoints or removal of outliers.

The results illustrate that the proposed method is able to segment (cluster) spectral databases and detect outliers that are not suitable for instance based learning algorithms.

II. Topological Mapping for Visualization of Spectral Database

The TOPNIR algorithm utilizes spectral databases for the prediction of product properties based on on-line measured infrared spectra by utilizing the well known k-nn algorithm [16]. The performance of the prediction is based on the structure of the dataset of the reference spectra. TOPNIR uses an additional technique to identify and separate the operating ranges of the technology. For example fuels produced for summer and winter usage would have very different technological parameters that require two separate models for the prediction of the product properties. This additional visualization and indexing method used by TOPNIR is referred as Topological Mapping using Aggregates. [11]

The aggregates are equations that combine absorbances measured at significant wavelengths. In ideal case aggregates reflect product properties. Since these properties can be dependent on different rages of the spectra each aggregate built up to six wavelength to contain enough information related to a certain chemical property.

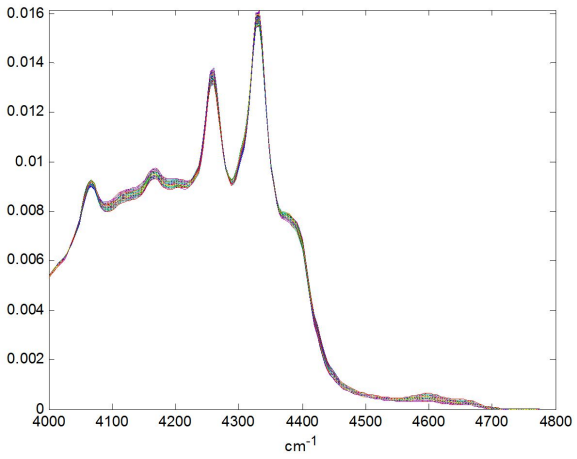

Fig. 1. Spectrums in sample database

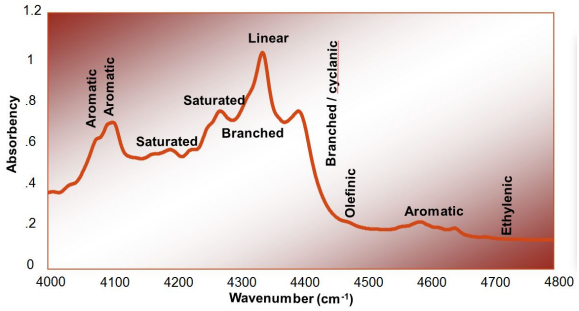

Fig. 2. Significant wavelengthes

The two main forms of the aggregates are shown by equation (1) and (2).

$$a_0 \frac{a_1 W_1 * a_2 W_2}{a_3 W_3 * a_4 W_4} \quad (1)$$

$$a_0 \frac{a_1 W_1 + a_2 W_2}{a_3 W_3 * a_4 W_4} \quad (2)$$

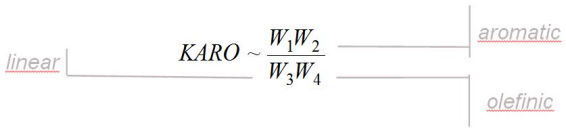

Fig. 3. Structure of KARO aggregate

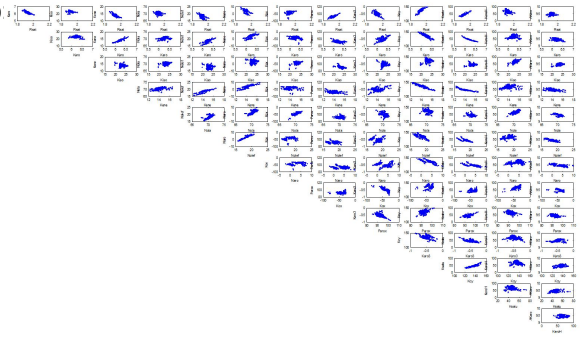

Fig. 4. Mappings using different aggregate pairs

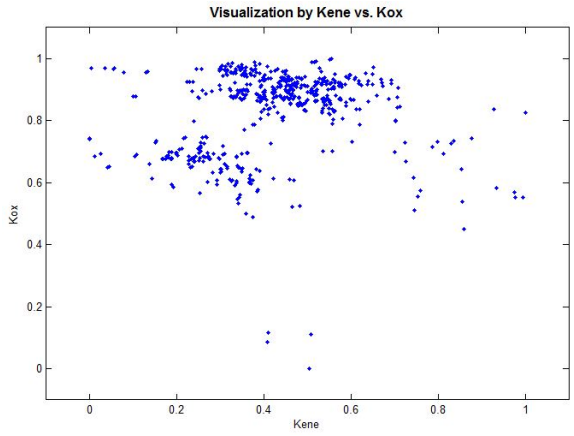

Fig. 5. Kene and Kox aggregates (Normalized values are shown)

For example the aromatic and the olefinic property have own ranges in the spectrum (see Fig. 2).

Two aggregates are used in the same time to give a two dimensional mapping of the spectral space.

There are 14 aggregates defined in the TOPWIN software used as a framework of the TOPNIR algorithm. Fig. 4 shows the mappings defined by the possible combination of these aggregates. As can be seen, the database contains samples from two different operating modes (summer and winter diesel) and some of these mappings are able to separate these operating regimes. It is interesting to see that there are also paris of aggregates where correlation among them is to high to provide informative mapping. Since two aggregates are used for visualization, these two aggregates should contain enough information about different rages of the spectra representing all the studied properties. In case the prediction performance related to a given product property is in the focus of the indexing of the database it is important to select the optimal pair of aggregates.

III. Principal Component Analysis

One of the most widely applied dimensionality reduction method is the *Principal Component Analysis* (PCA) [17], [18]. PCA is also known as Hotteling or as Karhunen-Loéve transformation [17], [18]. PCA differs from metric and non-metric dimensionality reduction methods, because instead of the preservation of the distances or the global ordering relations of the objects (in this case spectra) it tries to preserve the variance of the data. PCA represents the data as linear combinations of a small number of basis vectors. This method finds the projection that stores the largest variance possible in the original data and rotates the set of the objects such that the maximum variability becomes visible. Geometrically, PCA transforms the data into a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on. If the data set ($\mathbf{X}$) is characterized with D dimensions and the aim of the PCA is to find the d-dimensional reduced representation of the data set, the PCA works as follows:

1) PCA subtracts the mean from each of the data dimensions,
2) then it calculates the $D \times D$ covariance matrix of the data set,
3) following this PCA calculates the eigenvectors and the eigenvalues of the covariance matrix,
4) then it chooses the d largest eigenvectors,
5) and finally it derives the new data set from the significant eigenvectors and from the original data matrix.

The corresponding d-dimensional output is found by linear transformation: $\mathbf{Y} = \mathbf{QX}$, where $\mathbf{Q}$ is the $d \times D$ matrix of linear transformation composed of the d largest eigenvectors of the covariance matrix, and $\mathbf{Y}$ is the $d \times D$ matrix of the projected data set. *Independent Component Analysis* (ICA) [20] is similar to PCA, except that it tries to find components that are independent.

The PCA is ideal for dimensional reduction but according to the original idea only the first two principal component should be used for indexing. This is ideal because these two principal component are the most close to orthogonality so applicable for visualization too.

IV. Mapping Quality

Instance based prediction algorithms used for property estimation are based on the assumption that similar spectra represent samples having similar product property. This concept is illustrated by Fig. 7, where samples close to each other in the spectral space are also neighbors in the space of the property variables.

There are some measures of the spectral space (e.g D_{max} and i_m). The D_{max} is the maximum distance in the spectral space, the i_m is the minimum distance under that the samples are not distinguished.

In this work is the classical MDS stress function, Sammon stress function and residual variance are used to measure the distance preservation of the mappings to be analyzed. The neighborhood preservation of the mappings and the local

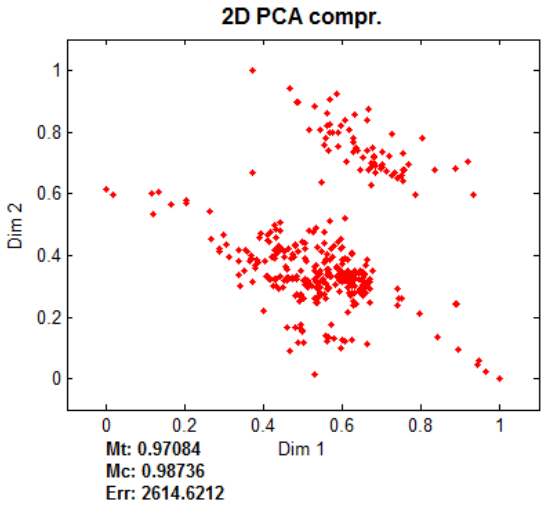

Fig. 6. Indexing using 2 principal component

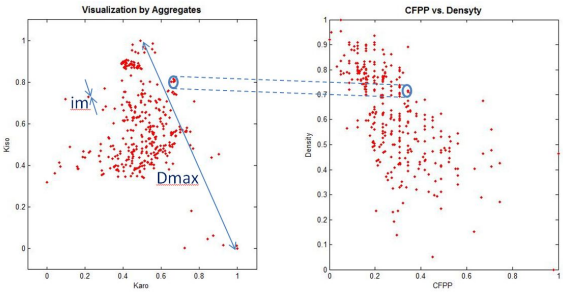

Fig. 7. Kene and Kox aggregates

and global mapping qualities are measured by functions of trustworthiness and continuity. Kaski and Venna pointed out that every visualization method has to make a tradeoff between gaining good trustworthiness and preserving the continuity of the mapping [19].

A projection is said to be *trustworthy* when the nearest neighbors of a point in the reduced space are also close in the original vector space. Let n be the number of the objects to be mapped, $U_k(i)$ be the set of points that are in the k size neighborhood of the sample i in the visualization display but not in the original data space. The measure of trustworthiness of visualization can be calculated in the following way:

$$M_1(k) = 1 - \\ -\frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in U_k(i)} \left(r\left(i,j\right) - k\right) \qquad (3)$$

where $r(i,j)$ denotes the ranking of the objects in input space.

The projection onto a lower dimensional output space is said to be *continuous* [19] when points near to each other in the original space are also nearby in the output space. The measure of continuity of visualization is calculated by the following equation:

$$M_2(k) = 1 - \\ \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in V_k(i)} \left(s\left(i,j\right) - k\right), \qquad (4)$$

where $s(i,j)$ is the rank of the data sample i from j in the output space, and $V_i(k)$ denotes the set of those data points that belong to the k-neighbors of data sample i in the original space, but not in the mapped space used for visualization.

In our analysis when the mapping algorithms are based on geodesic distances, the ranking values of the objects in both cases (trustworthiness and continuity) are calculated based on the geodesic distances.

Both trustworthiness and continuity functions are function of the number of neighbors k. Usually, the qualitative measures of trustworthiness and continuity are calculated for $k = 1, 2, \ldots, k_{max}$, where k_{max} denotes the maximum number of the objects to be taken into account. At small values of parameter k the local reconstruction performance of the model can be tested, while at larger values of parameter k the global reconstruction is measured.

The non-metric stress can be formulated as follows[1]:

$$E_{nonmetric} = \sqrt{\sum_{i<j}^{N} (\widehat{d}_{i,j} - d_{i,j})^2 / \sum_{i<j}^{N} d_{i,j}^2}, \qquad (5)$$

where $\widehat{d}_{i,j}$ yields the disparity of $\mathbf{x}_i$ and $\mathbf{x}_j$, and $d_{i,j}$ denotes the distance between the vectors $\mathbf{y}_i$ and $\mathbf{y}_j$.

V. Delaunay Triangulation

In mathematics and computational geometry, a Delaunay triangulation for a set P of points in a plane is a triangulation $DT(P)$ such that no point in P is inside the circumcircle of any triangle in $DT(P)$ [13]. Delaunay triangulations maximize the minimum angle of all the angles of the triangles in the triangulation; they tend to avoid skinny triangles. The triangulation is named after Boris Delaunay for his work on this topic from 1934.

For a set of points on the same line there is no Delaunay triangulation (the notion of triangulation is degenerate for this case). For four or more points on the same circle (e.g., the vertices of a rectangle) the Delaunay triangulation is not unique: each of the two possible triangulations that split the quadrangle into two triangles satisfies the "Delaunay condition", i.e., the requirement that the circumcircles of all triangles have empty interiors. By considering circumscribed spheres, the notion of Delaunay triangulation extends to three and higher dimensions. Generalizations are possible to metrics other than Euclidean. However in these cases a Delaunay triangulation is not guaranteed to exist or be unique.

The Fig. 8 shows the triangulation of the mapped plain using the the PCA's first two component. As it can be seen the spectral database contains samples in two separable samples. The separation can be done where the triangular area is larger than in the center of the clusters.

The distribution of the triangles' helps to cluster the samples. The Fig. 9 show the distribution of the triangles' area. One can se that the expected value is really small in term of the maximal area. The red line shows the expected value. There are about 700 triangles but less than 150 is larger than this average area.

[1]Traditionally, the non-metric stress is often called Stress-1 due to Kruskal [19]

VI. Results

The indexing of the spectral database of MOL was made by four different approaches. In the first three cases three aggregate parings were studied. These aggregates are built in the TOPWIN software. These spectrums were taken in Duna refinery of MOL Ltd. (Szazhalombatta). The database has 651 samples from winter and summer operation.

Table I shows the quality measures of the mapping techniques. The FNN rate is the rate of false Nearest Neighbors in case of two different properties. (The details of this measure are given in the appendix.)

As can be seen from the results given in Table I, the proposed mappings can be effectively applied when property based indexing is necessary. Among the presented mappings in most of the cases PCA produces the best indicators.

The outlier samples and the spare areas of the database can be identified easily by the proposed Delaunay triangulation based techniqe. Where the area of the triangles are large the coverage is low. So new samples should be inserted in the database (to increase the density of the data in the related region which will improve the prediction performance of the model).

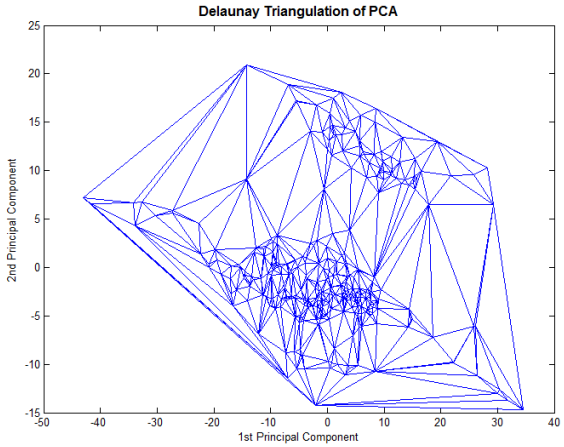

Fig. 8. Delaunay triangulation of mapped plain

The outlier samples are on the corners of the enclosing polygon. The prediction of these samples is problematic, because their neighbors are in one direction. If we predict these points using sample only from one direction (target angle is lass than 180) the prediction will not follow the plain determined the samples.

The distribution of the triangles' area shows that most of triangles are in the same range but there are some much larger triangles than the expected value. Triangles having larger area than a given threshold will be ignored. Fig. 9 shows the distribution of the triangles area and the horizontal line signs the threshold value which above the triangles should be ignored.

Fig. 10 shows a clustering of the PCA mapped plain. This clustering was made by removing the triangles which were larger than the upper quartile of the areas' distribution.

TABLE I
VISUALIZATION QUALITY OF DATA SET

Method	M_t	M_c	Metric Error	FNN rate (CFPP0)	FNN rate (Density)
Kene vs. Kox	0.87286	0.94273	11402.6385	0.15126	0.14286
Karo vs. Kiso	0.7513	0.82782	15246.5048	0.13445	0.17087
Nolef vs Naro	0.67636	0.74684	15608.9904	0.20168	0.070028
PCA 2D Reduction	0.97084	0.98736	2614.6212	0.15686	0.12325

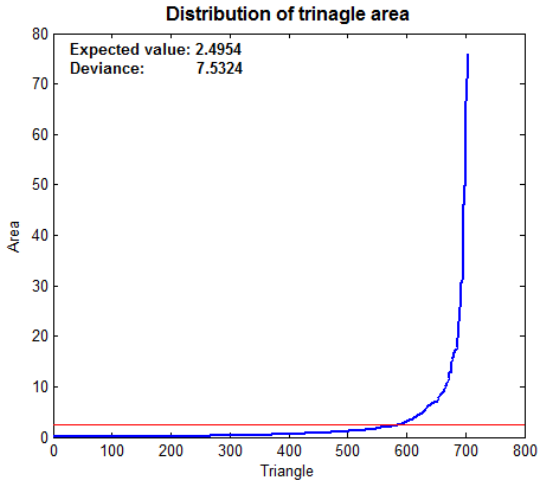

Fig. 9. Distribution of triangles' areas

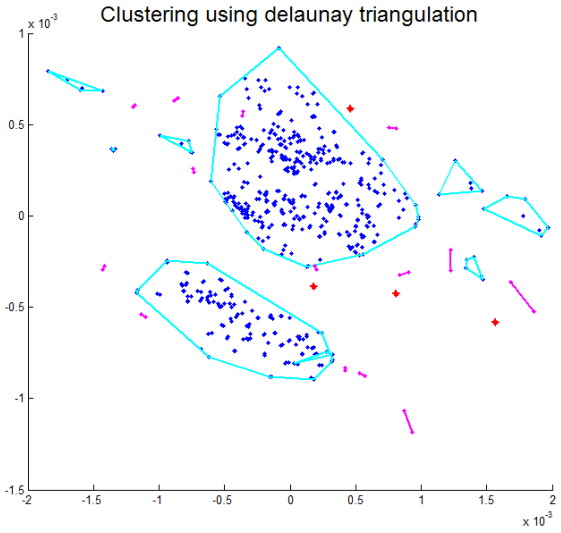

Fig. 10. Clustering based on Delaunay trinagulation

Using this clustering the search in the spectral database can become more effective since it can be used to segment the database and remove outliers that could worsen the prediction performance.

VII. CONCLUSION

To reduce the computational requirement of searching and predicting in spectral databases dimensional reduction techniques were proposed. Two methods (Aggregates and PCA) were applied to index high dimensional spectral spaces onto into a two dimensional map to support prediction processes. These methods are able to visualize spectral space. The combination of indexing and visualization reduces the computational requirement because only one algorithm is needed for both functions.

Measures and indicators were defined to evaluate the quality of the mappings. The proposed metric error represents how the mapping preserves the distances during the transformation. As the results show the 2D PCA indexing provides the smallest error. The neighborhood preserving property was also evaluated (M_t and M_c). These indicators show the same performance. The PCA's neighborhood preserving is greater than 0.97 which means a really good result.

To visualize and explore the hidden structure of the spectral database a novel tool based on Delaunay triangulation was proposed. Using this tool the dense and compact operating regimes can be identified. Due to the the area of the triangles are indirectly proportional to the data coverage the spare areas of operating regimes can be found easily which information can be used to support model validation and development (e.g. experiment design).

The presented framework can support the development data driven models of on-line analyzers using Near-Infrared spectroscopy.

APPENDIX A
FALSE NEAREST NEIGHBOR (FNN) METHOD

The main idea of the FNN algorithm stems from the basic property of a function. If there is enough information in the regression vector to predict the future output, then any of two regression vectors which are close in the regression space will also have future outputs which are close in some sense. For all regression vectors embedded in the proper dimensions, for two regression vectors that are close in the regression space and their corresponding outputs are related in the following way:

$$y_i - y_j = df(\mathbf{x}_i)[\mathbf{x}_i - \mathbf{x}_j] + o([\mathbf{x}_i - \mathbf{x}_j])^2 \quad (6)$$

where $df(\mathbf{x})$ is the jacobian of the function $f(.)$ at $\mathbf{x}_i$.

Ignoring higher order terms, and using the Cauchy-Schwarz inequality the following inequality can be obtained:

$$|y_i - y_j| \leq \|df(\mathbf{x}_i)\|_2 \|\mathbf{x}_i - \mathbf{x}_j\|_2 \quad (7)$$

$$\frac{|y_i - y_j|}{\|\mathbf{x}_i - \mathbf{x}_j\|_2} \leq \|df(\mathbf{x}_i)\|_2 \quad (8)$$

If the above expression is true, then the neighbors are recorded as true neighbors. Otherwise, the neighbors are false neighbors.

Based on this theoretical background, the outline of the FNN algorithm is the following. [21]

1) Identify the nearest neighbor to a given point in the regressor space. For a given regressor: $\mathbf{x}_i$ find the nearest neighbor $\mathbf{x}_j = \mathbf{x}_{(i,1)}$.
2) Determine if the following expression is true or false

$$\frac{|y_i - y_j|}{||\mathbf{x}_i - \mathbf{x}_j||_2} \leq R$$

   where R is a previously chosen threshold value. If the above expression is true, then the neighbors are recorded as true neighbors. Otherwise, the neighbors are false neighbors.
3) Continue the algorithm for all times i in the data set.

The FNN algorithm is sensitive to the choice of the R threshold. In the threshold value was selected by trial and error method based on empirical rules of thumb, $10 \leq R \leq 50$. However, choosing a single threshold that will work well for all data sets is impossible task. In this case, it is advantageous to estimate R based on 8 using the the maximum of the Jacobian, $R = max_i \|df(\mathbf{x}_i)\|$, as it was suggested by Rhodes and Morari.

While this method uses data based models for the estimation of $\|df(\mathbf{x})\|$, the performance and the capabilities of this identified model can deteriorate the estimate of $max(df)$. When df is over estimated the model orders could be under estimated, and vice-versa. Hence, the modeler has to be careful at the construction of this model (e.g. the model can be over or under parameterized, etc.).

ACKNOWLEDGMENT

The financial support of the TAMOP-4.2.1/B-09/1/KONV-2010-0003 and TAMOP-4.2.2/B-10/1-2010-0025 projects are gratefully acknowledged.

REFERENCES

[1] H. Yamamotoa, H. Yamaji, E. Fukusaki, H. Fukuda, *Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting*, Biochemical Engineerin Journal 40 (2008) 199-204.
[2] J. Parkkinen, A.C. Legrand, *Index and Search in a Spectral Imaging Database by PCA and NMF for Archiving Paintings Application*, Department of Computer Science and Statistics, University of Joensuu, 2007.
[3] J.F. MacGregor, T. Kourtl, *Statistical process control of multivariate processes*, Control Fag. Practice, VoL 3, No. 3, pp. 403-414, 1995.
[4] I. Kopanakis, B. Theodoulidis, *Visual data mining modeling techniques for the visualization of mining outcomes*, Journal of Visual Languages and Computing 14 (2003) 543-589, 2003.
[5] X. Blasco, J.M. Herrero, J. Sanchis, M. Martnez, *A new graphical visualization of n-dimensional Pareto front for decision-making in multiobjective optimization*, JInformation Sciences 178 (2008) 39083924.
[6] N.Krmer, A.L. Boulesteix, G. Tutz, *Penalized Partial Least Squares with applications to B-spline transformations and functional data*, Chemometrics and Intelligent Laboratory Systems 94 (2008) 6069.
[7] Rolf Ergon, *Informative PLS score-loading plots for process understanding and monitoring*, Journal of Process Control 14 (2004) 889897.
[8] M. Greenacrea, T. Hastieb, *Dynamic visualization of statistical learning in the context of high-dimensional textual data*, Web Semantics: Science, Services and Agents on the World Wide Web 8 (2010) 163168.
[9] W.R. Browett, M.J. Stillman, *DComputer-aided chemistryII. A spectral-database management program for use with microcomputers*, Computers and Chemistry 11 (1987) 7382.
[10] Ehud Gudes, *A uniform indexing scheme for object-oriented databases*, Information Systems 22 (1997) 199-221.
[11] B. Descales, D. Lambert, J.R. Llinas, A. Martens, S. Osta, M. Sanchez, S. Bages, *Method for determining properties using near infra-red (NIR) spectroscopy, Eutech Engineering Solutions* (2000) US6.070.128.
[12] Yaser R. Sonbul, *Topological near infrared analysis modeling of petroleum refinery products*, Saudi Arabien Oil Company (2005) US6.897.071 B2.
[13] L. Jin, J.A. Fernndez Pierna, Q. Xu, F. Wahl, O.E. de Noord, C.A. Saby, D.L. Massart, *Delaunay triangulation method for mutivariate calibration*, Analytica Chimica Acta 488 (2003) 114.
[14] I. Lee, J. Yang, *Common Clustering Algorithms*, Comprehensive Chemometrics (2009) 577-618.
[15] S. Mimaroglu, E. Erdil, *Combining multiple clusterings using similarity graph*, Pattern Recognition 44-3 (2011) 694-703.
[16] Y. Wu, K. Ianakiev, V. Govindaraju, *Improved k-nearest neighbor classification*, Pattern Recognition 35-1 (2002) 2311-2318
[17] K.H. Esbensen, P. Geladi, *Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice*, Comprehensive Chemometrics (2009) 211-226.
[18] Ian R. Greenshields, Joel A. Rosiene, *A fast wavelet-based Karhunen-Loeve transform* , Pattern Recognition 31-77 (1998) 839-845.
[19] Jarkko Venna, Samuel Kaski, *Local multidimensional scaling*, Neural Networks 19 67 (2006) 889-899.
[20] F. Westad, M. Kermit, *Independent Component Analysis*, Comprehensive Chemometrics (2009) 227-248.
[21] I. Marn Carrin, E. Arias Antnez, M.M. Artigao Castillo, J.J. guila Guerrero, J.J. Miralles Canals *Thread-based implementations of the false nearest neighbors method*, Parallel Computing, Volume 35 1011 (2009) 523-534.