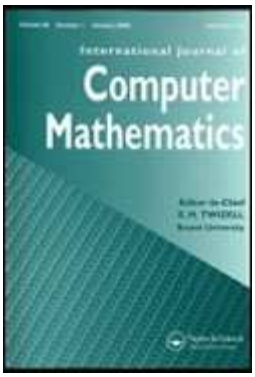
# Spectra-Based Multivalued Fingerprints as Predictive Vectors for Partial Least Squares Regressions Processes (SI-CMMSE-2006)

Miguel Ángel Gómez-Nieto, Irene Luque Ruiz, Manuel Urbano Cuadrado




# Spectra-Based Multivalued Fingerprints as Predictive Vectors for Partial Least Squares Regressions Processes (SI-CMMSE-2006)

| | |
|---|---|
| Journal: | *International Journal of Computer Mathematics* |
| Manuscript ID: | GCOM-2006-0166.R1 |
| Manuscript Type: | Original Article |
| Date Submitted by the Author: | 24-Nov-2006 |
| Complete List of Authors: | Gómez-Nieto, Miguel Ángel; University of Córdoba, Computing and Numerical Analysis<br>Luque Ruiz, Irene; University of Córdoba, Computing and NUmerical Analysis<br>Urbano Cuadrado, Manuel; Institute of Chemical Research of Catalonia ICIQ |
| Keywords: | Multivalued Fingerprints, PLS regression, FT-MIR spectroscopy, Data normalization, Data reduction |
| | |
| Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online. | |
| GCOM-2006-0166-r1.tex | |

*International Journal of Computer Mathematics*
Vol. 00, No. 00, Month 200x, 1–11

# *Spectra-Based Multivalued Fingerprints as Predictive Vectors for Partial Least Squares Regressions Processes (SI-CMMSE-2006)*

MANUEL URBANO-CUADRADO†, IRENE LUQUE RUIZ‡ and MIGUEL ÁNGEL GÓMEZ-NIETO*‡

†Institute of Chemical Research of Catalonia ICIQ. Avinguda Països Catalans, 16. E-43007 Tarragona (SPAIN)
‡ Department of Computing and Numerical Analysis. University of Córdoba. Campus Universitario de Rabanales, Albert Einstein Building. E-14071 Córdoba (SPAIN)

(*October 2006*)

A new method for transforming spectra into multivalued fingerprints is here presented and applied to multivariate regression. The method, aimed at enlarging differences between long-dimensional vectors showing a high degree of similarity, is based on the following stages: (1) spectral outliers removal; (2) data normalization aimed at transforming the spectral matrix into a new data set within the [0,1] range; and (3) selection of threshold values for assigning significance values to each variable according to both normalised and threshold values. A study case is described: the processing of mid infrared spectra in partial least squares regression processes for predicting total acidity and content of reducing sugars in wines. The original spectra matrix consisted of 156 objects (samples) and 1142 columns (predictors) —a wavelenght range of $3000 - 800\,cm^{-1}$ with a spectral resolution slightly greater than $2\,cm^{-1}$—. The fact of using the here proposed method yielded better predictions than those obtained by means of both classical treatments and spectral data without any processing.

*Keywords*: Multivalued Fingerprints, PLS regression, FT-MIR spectroscopy, Data normalization

*AMS Subject Classifications*: 62H12, 68P05, 92E99

## 1. Introduction

Either obtaining or generating real time data is a key aspect for achieving one of the most important characteristics of the information concept: in real time information [1]. For this reason, researchers in several chemical fields have recently devoted a lot of efforts to diminish the time necessary for the analysis tasks. The process between the material under study and the output of the target information about this material is still too long for many analytical systems. Several kinds of approaches have been developed with the aim of shortening this process: automated methods [2,3], screening methodologies [4], multiparametric determination [5,6], etc. These advances have been basically supported by two tools: better instrumentation and the speed offered by computers for both automated control and data analysis.

Regarding instrumentation, one of the major advances has been the development of techniques based on multichannel detectors. These enable collection of hundreds and even thousands of measurements in a short time, thus enhancing possibilities of extracting information from these very long data sets. Techniques like Near Infrared Spectroscopy (NIRS) [7,8], Fourier Transform Mid Infrared Spectroscopy (FT-MIRS) [9], Mass Spectrometry (MS) [10], Resonance Magnetic Nuclear (RMN) [11], etc., have provided the chemist with powerful techniques for the development of fast analytical methods.

In addition to the software for both controlling the instruments and acquiring data, computers have made possible the implementation of algorithms for multivariate analysis [12,13]. The analysis of large data sets requires different methods from those used for univariante approaches [14]. Multivariate analysis —with either quantitative or qualitative aims— tries to overcome disadvantages of working with long data arrays corresponding to spectra of materials that have not been previously treated chemically (main factor

*Corresponding author. Email: mangel@uco.es, Phone: +34 95 721 2082, Fax: +34 95 721 8630

*International Journal of Computer Mathematics*
ISSN 0020-7160 print/ISSN 1029-0265 online © 200x Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/0020716YYxxxxxxxx

for reducing the time of the analytical process) [15]. Scatter effects, segments of the spectrum that model the behaviour of interferences, the high number of variables, etc., are problems intended to be removed with the use of multivariate analysis.

The development of multivariate equations involves the training and testing stages using samples with properties measured by the reference method. The accuracy and precision achieved for multivariate models have as maximum limits those obtained with the reference method [16]. Usually these accuracy and precision levels are not achieved for several reasons: overcomplex systems under study (major components mask the properties of minor ones), very similar spectra, non representative training sets, low signal/noise ratios, etc.

Therefore, numerous efforts in chemometrics have focussed on improving the values of accuracy and precision. With this aim, methods for signal preprocessing —as derivatives treatment, scatter reduction correction, detrending, etc. [17-19]— have been developed. Regarding multivariate calibration methods, new approaches have also tried to improve prediction statisticals. Firstly, as a consequence of the long dimension of spectra, methods have been developed based on different statistical principles for the selection of variables showing the highest correlation with the property of interest [20,21]. Secondly, calibration methods based on the space transformation of the original variable space have been widely used and modified in order to build reliable prediction models [22,23]. Non-parametric techniques such as Artificial Neural Networks (ANN) have also been used in multivariate analysis for systems in which data do not show a linear structure [24].

In this paper we propose a new attempt to build predictive models by means of multivariate regression techniques applied to multivalued fingerprints. This method is based on transformation of spectra into multivalued fingerprints, using the latter as inputs of multivariate calibration and validation processes. Although fingerprints have been mainly employed in computational chemistry for calculating structural similarity [25,26], construction of fingerprint-based similarity matrices from spectra has improved discrimination power in pattern recognition approaches [27]. Here, we propose to consider fingerprint-based spaces as X-matrices for Partial Least Squares Regression (PLSR) [28,29].

## 2. A method for building multivalued fingerprints

Be an spectra matrix $S$ consisting of $N$ rows and $M$ columns representing samples and wavelengths, respectively, a fingerprints matrix $F$ (also with $N \times M$ dimensions) is obtained as follows:

### 2.1. *Removal of spectral outliers*

In order to avoid negative influences of some spectra on fingerprints building, a study of spectral outliers, based on Mahalanobis distance ($H$) computation, is carried out. Thus, spectra showing different behaviour from the population pattern are removed. This study of spectral outliers is one of the steps that compose correct calibration processes in chemical analysis by multichanel spectroscopy [30]. Usually, spectra with $H > 3.0$ (computed in a reduced space built often by principal components analysis) are considered outliers. After outliers removal, the original $S_{N\times M}$ matrix is reduced into a $S_{N'\times M}$ matrix, where $N' < N$. Then, the fingerprints matrix will also show $N' \times M$ dimensions.

### 2.2. *Normalization of the spectra matrix*

Each row of $S$ matrix represents the spectrum of a different sample $i$. So, each element $S(i, j)$ corresponds to an spectral datum (e.g., absorbance, reflectance, emission intensity, etc.) collected at condition $j$ (e.g., wavelength, excitation intensity, etc.). Due to the different spectral data per condition for all the samples, by column normalization has to be performed. Thus, the matrix $S_{N'\times M}$ is transformed into a normalised matrix $\bar{S}_{N'\times M}$, consisting of data within the range [0,1]. Several normalization methods can be used. In

this work, three of them were used, namely:

$$\forall i, \forall j, \bar{S}(i,j) = \frac{S(i,j) - min(S(n,j))}{max(S(n,j)) - min(S(n,j))}, \ \ (n = 1...N) \ \ Standard \tag{1}$$

$$\forall i, \forall j, \bar{S}(i,j) = \frac{S(i,j)}{max(S(n,j))}, \ \ (n = 1...N) \ \ Maximum \tag{2}$$

$$\forall i, \forall j, \bar{S}(i,j) = \frac{log(S(i,j)+1) - min(log(S(n,j)+1))}{max(log(S(n,j)+1)) - min(log(S(n,j)+1))}, \ \ (n = 1...N) \ \ Logarithmic \tag{3}$$

As can be observed in equations (1), (2) and (3), the three normalization techniques employed take into account the data distribution in the set of samples considered. This fact also justifies the spectral outliers removal to be carried out as a preliminary step.

### 2.3. *Building the fingerprints*

After building the matrix $\bar{S}_{N' \times M}$, this is employed for generating a new matrix $F_{N' \times M}$ of fingerprints. The matrix transformation process is as follows:

(i) Firstly, the fingerprints dimension $k$ is selected; being $k$ an integer representing the number of cases that can be assigned to each variable. For example, if the fingerprint dimension is $k = 3$ (ternary fingerprints), each variable is set to 0, 1 or 2.

(ii) Then, $k-1$ threshold values $U$ within the range [0,1] are chosen. The selection of these factors depends on the normalised data distribution. Following the above example, a possible set of threshold values is $U = (0.33, 0.66)$ for ternary fingerprints.

(iii) Finally, matrix $F_{N' \times M}$ is built, as follows:

$$F(i,j) = 0 \,|...|\, F(i,j) = k-1; if \ \bar{S}(i,j) \leq 0 \,|...|\, \bar{S}(i,j) \leq U(k-1) \tag{4}$$

Thus, a discrete level of significance is given to each measurement condition as a function between the normalised value of each variable and threshold values. Although the threshold factors $U$ are empirical, a statistical meaning can be extracted from the data structures built with them. Thus, the higher the threshold selected, the lower the number of significant variables.

Finally, the $F_{N' \times M}$ matrix is employed as the $X$ matrix in PLSR.

## 3. Partial Least Squares Regression

After building fingerprints, PLSR was employed due to several reasons. First, the multivalued fingerprint matrix is reduced into a latent space in order to visualize data trends. Thus, the study of the number and characteristics of PLS factors provides chemical information in a more intuitive than that obtained by using all the original variables. Second, since multivalued fingerprint matrices often have more variables than objects, we need regression techniques different of Multivariate Linear Regression (MLR). And third, PLSR considers variance of fingerprints and properties for building the latent space. Other techniques also based on data reduction only takes into account the fingerprints variance. For example, Principal Components Regression (PCR) retains relevant factors which only explain the fingerprint matrix.

### 3.1. *The PLSR algorithm*

Although several algorithms have been developed for computing partial squares components (the Non-linear Iterative Partial Least Squares NIPALS, the SIMPLS method, the PLS2 approach, etc., and their robust versions), the work methodology, described below, is common.

Be $F$ and $Y$ the matrices which describe $p$ observations and $m$ properties, respectively, for $n$ objects. A regression using factor extraction from data computes the factor score matrix $T = FW$ for an appropriate weight matrix $W$, and then considers the linear regression model $Y = TQ + E$, where $Q$ is a matrix of regression coefficients (loadings) for $T$, and $E$ is an error (noise) term.

Aimed at specifying $T$, two sets of weights $w$ and $q$ have to be found to create a linear combination of columns of $F$ and $Y$ such that their covariance is maximal. The goal is to obtain a first pair of vectors $t = Fw$ and $q = Yq$ with the constraints that $w^T w = 1$, $t^T t = 1$ (assures the orthonormality of the latent variables) and $t^T q$ be maximal (reflects the maximal covariance structure between the fingerprint and property spaces). When the first latent vector is found, their contributions are substracted from $F$ and $Y$ and this procedure is re-iterated until $F$ becomes a null matrix. In this case, the number of latent variables is equal to the rank of $F$, thus obtaining an exact decomposition of $F$ and $Y$.

Only a few latent variables are then considered for predicting the properties of new objects because of the overfitting phenomenon. Although the fact of using a high number of components involves accurate fittings between fingerprints and properties, this fact can also imply modelling of noise. Thus, the number of PLSR factor must be optimised taking into account the prediction error value [31,32].

## 4. An application case: wine analysis by FT-MIR

Recently, a number of PLSR equations for the prediction of wine parameters have been developed. Urbano et al. [33] compared the spectral mid and near infrared (MIR and NIR) regions in quantitative analysis of wines. In order to study the efficiency of the method here proposed, a PLSR equations system for determining total acidity and reducing sugars using MIR fingerprints was developed. Results were compared with those obtained from spectra without the proposed processing and with classical treatment (derivatives processing).

### 4.1. *Material and chemometrics*

**4.1.1. Samples and sample preparation.** Different wines —including red and white wines; young and aged wines; wines from different procedence ("La Mancha", "Valdepeñas", "Jumilla", "Navarra", "Alicante" and "Madrid") and grape varieties ("Cencibel", "Cabernet Sauvignon", "Cencibel-Cabernet Sauvignon", "Merlot" and "Syrah")— were used in this study. Samples employed in the calibration and validation steps were 130 and 26, respectively. The samples were used as such, because filtering, dilution, preconcentration, interferents removal, etc., were not required. Thus, the analytical process was shortened considerably.

**4.1.2. Instrumentation.** The instrument employed for MIR spectra collection was an FT-MIR Nicolet Magna-IR550 Serie II (Nicolet Instrument Corp., Madison, Wisconsin, USA), capable of making measurements at 4 $cm^{-1}$ resolution in the spectral range covering 4000-400 $cm^{-1}$. The instrument was furnished with an infrared attenuated total reflection (ATR) solid, liquid and mellow sample cell with a zinc selenide crystal (Spectra Tech., Stamford, CT, USA) for Nicolet Spectrometers. The reflectance spectra (log1/R) were collected in duplicate. The region of the MIR spectra used was the 800-3000 $cm^{-1}$ wavelength range, shown in Figure 1. High degree of similarity between spectra can be observed [34].

The samples were also analysed in duplicate by the reference methods (titration with NaOH up to $pH = 7.0$ for total acidity and reduction of $Cu^{+2}$ in boiling alkaline medium for reducing sugars content), and standard error laboratory (SEL) was estimated from the duplicates.

**4.1.3. Chemometrics.** The software employed for outliers removal, normalising spectra and building fingerprints was developed by the authors in C programming language. The Unscrambler 7.8 (Camo Process AS, Oslo, Norway) was used for PLSR equations development.

PLSR equations were developed using both the training set and cross-validation strategy, based the latter on splitting training samples into fitting and internal test subsets. For this purpose, a group of samples are removed from the set of objects, and the training model is carried out with the remaining samples. Then, the model is internally tested with the samples not used for fitting. This is repeated cyclically in such way that all the samples are used once at least for internal tests. The statistical parameters obtained —$R^2$, Standard Error in Cross Validation (SECV), slope and bias— are the average of the values corresponding to all the cycles. In the study undertaken, 105 and 25 samples were used for the training and test segments, respectively, for each cycle. Thus, 6 cycles per training were completed.

A study of chemical outliers in training of equations by cross validation was carried out using the student test for each sample $i$. This parameter was calculated as follows:

$$T_i = \frac{y_i - \widehat{y}_i}{SECV} \tag{5}$$

where $y_i$ represents the value determined by the PLSR equation in cross-validation and $\widehat{y}_i$ represents the reference value. A $T_{cut-off} = 2.5$ was considered for detecting chemical outliers.

Validation of equations was carried out using the external data set (26 samples) and statistical parameters —$r^2$ and Standard Error in Prediction ($SEP$)— were also obtained.

### 4.2. *Study of spectral outliers*

Mahalanobis distance ($H$) was computed for clear wine groups —white and rosé wines were considered jointly—, and red wines separately. Two spectra of red wines and one spectrum of white wine behaved as outliers. Eight and 10 principal components were used for $H$ distance calculation. The criterion to fix the number of components was to obtain an increment of explained variance lower than 0.25%. On the other hand, the sum of explained variance for each model was close to 100%.

### 4.3. *Study of the normalization method and threshold values employed*

The standard and logarithmic methods led to similar normalised data matrices (Figs. 2 (A) and (B), respectively), which, in addition, showed the highest differences between spectra. However, the correlation between variables was lower than that obtained with the maximum method (Figure 2 (C)). Due to the similarity between the standard and logarithmic treatments, spectra normalised only by the standard and maximum methods were considered for subsequent studies.

As previously commented, data distributions differ considerably between them according to the method employed for data scaling. The maximum normalised variables show higher values than standard normalised ones (see Figs. 2 (A) and (C)). Selection of threshold values $U$, and in turn, quality of fingerprints, depends on data distribution. Figure 3 (A) shows the binary fingerprints for a given sample built from its standard normalised spectrum. As can be observed in this figure, medium values $U$ (0.4 and 0.6) led to medium-density fingerprints —yielded the best results from a qualitative point of view [27]—. Nevertheless, these threshold values produced high-density fingerprints when maximum normalization was employed (see Figure 3 (B)).

If the quantitative behaviour of fingerprints is similar to that shown in qualitative approaches, non-acceptable results will derive from the use of low-density and high-density fingerprints. This fact can be observed in Figs. 4 (A) and (B), which show the $R^2$ values as a function of the threshold value employed for building binary fingerprints using the standard and maximum methods, respectively. Thresholds that give medium-density fingerprints showed the best results.

Normalization methods were also compared with respect to the number of threshold. Better statistic parameters were achieved using the maximum method. Binary fingerprints, Figure 4 (B), referred to the maximum method, show $R^2$ values higher than those achieved with the standard normalization (Figure 4 (A)). Table 1 compares the statistic parameters $R^2$ and $SECV$ for other fingerprint dimensions and similar results were obtained. This can be explained by the correlation degree maintained compared to

original spectra when the maximum method is employed, as can be observed in Figure 2.

### 4.4. *Study of the multivalued fingerprint dimension*

The fingerprint dimensions studied were binary, ternary, quaternary and quintary. Figure 5 shows the values of statistic parameters —namely, (A) $R^2$, (B) relative standard error in prediction (RSD), (C) variance explained by the first 10 principal components and (D) the optimal number of PLSR factors— as a function of fingerprint dimension. $R^2$ and $RSD$ are referred to the predicted values by cross-validation strategy (6 segments were employed, ensuring that all the samples have been employed once at lest for validation). Maximum $R^2$ and minimum $RSD$ values (the best results) were obtained with ternary and quaternary fingerprints for total acidity and reducing sugars, respectively.

Observing Figure 5 (C), fingerprints matrices that yielded the best statistic values involved percentages of explained data variance lower than those obtained with binary and quintary fingerprints. This fact means that high percentages of explained variance involve, in turn, high levels of modelled noise. Data variance explained by PCA applied to MIR spectra without the processing proposed was 97% for the first 10 principal components. This value is higher than those explained by PCA applied to fingerprints. Thus, the fingerprints construction can be considered as a data reduction technique.

Figure 5 (D) shows the number of PLSR factors against the fingerprint dimension. As can be observed, this number increases as far as the quaternary fingerprints for the two wine parameters studied. This tendency is explained by the higher structural complexity derived from fingerprints matrix with higher dimensions, which requires a high number of factors in order to explain the relationship between predictors and properties. The decrease in the numbers of factors for quintary fingerprints can be due to their high noise level.

### 4.5. *Comparison with other chemometric approaches*

$R^2$, SECV, bias and slope values were compared with those obtained using both spectra without processing and derivative spectra (Norris method [12]). This comparison is shown in Figure 6. All the statistic parameters were improved using the method proposed in this work with the exception of $R^2$ for total acidity. The use of fingerprints significantly increases correlation between the reference and calculated values: fingerprints gave slope and bias values close to 1 and 0, respectively (optimal correlation).

According to the results shown in Figure 6, the data reduction achieved with the fingerprints matrix involved a refinement of the PLSR equations building. This is based, partly, on the capacity for outliers detection observed for fingerprints, as can also be observed in Figure 6 (the number of chemical outliers was lower than the limit accepted by the chemometric community —15% of the size of the samples set— for the two parameters). As Figure 6 shows, almost all the chemical outliers detected showed extreme values. Thus, these results point out to the fact of avoiding extrapolation problems when fingerprints were considered.

### 4.6. *Equations robustness and external validation*

Shenk and Westerhaus have proposed some criteria to evaluate the statistical results of the training and validation stages [30]: $R^2$ values higher than 0.90 indicate excellent precision, as well as $SEP$ values lower than $1.5 * SEL$; $R^2$ values between $0.70 - 0.90$ mean good precision, as do the $SEP$ values between $2 - 3 * SEL$; on the other hand, $R^2$ values lower than 0.70 indicate that the equation can only be used for screening purposes, which enable distinction between low, medium and high values for the measured parameter; and finally, if the $R^2$ value is lower than 0.50, the equation only discriminates high and low values.

The use of fingerprints is particularly useful for the reducing sugars prediction as it provides a quantitative method and not a screening method, which occurs when the method proposed is not employed according to $R^2$ values. External validation was carried and $R^2$ and $SEP$ obtained were 0.69 and $0.42\,meq/L$ for the total acidity, and 0.72 and $0.29\,g/L$ for the reducing sugars, respectively. Standard error laboratory

($SEL$) values were $0.35\,meq/L$ and $0.15\,g/L$ for total acidity and reducing sugars, respectively. Thus, equations show good precision.

A criterion employed for considering the equations as robust tools is that referred to the $SEP$ value: $SEP < 1.5 * SECV$. Then, fingerprints equations for determining both parameters were robust.

## 5. Conclusions

In this paper, we present a new method for developing multivariate equations (PLSR was employed) using multivalued fingerprints built form spectra as predictive vectors. An application case has been successfully studied: the prediction of total acidity and reducing sugars content in wines.

Fingerprints building can be considered as a data reduction technique, and the improvements on the statistic results are derived from the lower noise modelled as a consequence of the lower level of data variance explained. Thus, the prediction capacity of the equations was increased regarding both the original spectral data (without any processing) and the derived spectra.

The normalization step, involved in this method, was studied by means of the comparison between different normalization methods: standard, logarithmic and maximum methods. The best results were obtained using the maximum method due to the data correlation maintained with respect to the original space.

In the optimization of the fingerprint dimension, ternary and quaternary fingerprints (medium dimensions) behaved as the best PLSR inputs. Fingerprints with low or high dimensions (binary and quintary in our study) gave worse results. Besides, the higher capacity for outliers detection was especially useful for the refinement of the PLSR equations. The fingerprints matrix avoided extrapolation problems to detect extreme values as outliers in cross-validation training.

## Acknowledgments

We thank the Comisión Interministerial de Ciencia y Tecnología (CICyT) and FEDER for their financial support (Project TIN2006-02071).

## References

[1] S. Green, Information Systems Design. Thomson Computer Press, London, 1996.
[2] M. Valcárcel, M.D. Luque de Castro. Automatic Methods of Analysis, Elsevier, Amsterdam, 1988.
[3] M. Urbano-Cuadrado, M.D. Luque de Castro, M.A. Gómez-Nieto. Trigger-based Concurrent Control System for Automating Analytical Processes, Trends Anal. Chem. 23 (2004) 370–384.
[4] E. Trullols, I. Ruisnchez, F.X. Rius. Validation of Qualitative Methods, Trends Anal. Chem. 23 (2004) 137–145.
[5] M. Bonoli, M. Montanucci, T.G. Toschi, G. Lercker. Fast Separation and Determination of Tyrosol, Hydroxytyrosol and other Phenolic Compounds in Extra-virgin Olive Oil by Capillary Zone Electrophoresis with Ultraviolet-diode Array Detection, J. Chromatogr. A. 1011 (2003) 163–172.
[6] C. Jiménez, L. Moreno, C. de Haro, F.X. Muñoz, A. Florido, P. Rivas, A.M. Fernández, P.L. Martín, A. Bratov and C. Domínguez. Development of a Multiparametric System Based on Solid-State Microsensors for Monitoring a Nuclear Waste Repository, Sensor Actuat. B. 91 (2003) 103–108.
[7] A.M.C. Davies, R.K. Cho (Eds.). Near Infrared Spectroscopy: Proceedings of the 10th International Conference, NIR Publications, Chichester, 2002.
[8] A.M.C. Davies, R. Giangiacomo (Eds.). Near Infrared Spectroscopy: Proceedings of the 9th International Conference, NIR Publications, Chichester, 2000.
[9] R.H. Wilson, H.S. Tapp. Mid-infrared Spectroscopy for Food Analysis: Recent New Applications and Relevant Developments in Sample Presentation Methods, Trends Anal. Chem. 18 (1999) 85–93.
[10] E. de Hoffmann, V. Stroobant. Mass Spectrometry: Principles and Applications (Second Edition), Wiley, New York, 2001.
[11] C. Schorn. NMR Spectroscopy: Data Acquisition, Wiley-VCH, Weinheim, 2001.
[12] K.H. Esbensen. Multivariate Data Analysis - in Practice, Camo Process AS, Oslo, 2002.
[13] D.L. Massart, B.G.M. Vandeginsten, S. Buydens, S. De Jong, P.J. Lewi and J. Smeyers-Verbeke. Handbook of Chemometrics and Qualimetrics: Parts A and B, Elsevier, Amsterdam, 1998.
[14] T. Naes, T. Isakson, T. Davies. A User-Friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, 2002.
[15] D.A. Burns, E.W. Ciurczak. Handbook of Near Infrared Analysis, Marcel Dekker, Nueva York, 1992.
[16] D.M. Hawkins. The Problem of Overfitting, J. Chem. Inf. Comput. Sci. 44 (2004) 1–12.
[17] M. Blanco, T. Canals, J. Coello, J. Gené, H. Iturriaga, S. Maspoch. Direct Determination of Leather Dyes by Visible Reflectance Spectroscopy using Partial Least-squares Regression, Anal. Chim. Acta 419 (2000) 209–214.

[18] C.A. Andersson. Direct Orthogonalization, Intell. Lab. Syst. 47 (1999) 51–63.
[19] Y. Roggo, L. Duponchel, C. Ruckebusch, J.P. Huvenne. Statistical Tests for Comparison of Quantitative and Qualitative Models Developed with Near Infrared Spectral Data, J. Mol. Struct. 654 (2003) 253–262.
[20] R. Leardi. Genetic Algorithms in Chemometrics and Chemistry: a Review, J. Chemom. 15 (2001) 559–569.
[21] J. Ferré, F.X. Rius. A Graphical Criterion to Examine the Quality of Multicomponent Analysis: Implications for Wavelength Selection, Trends Anal. Chem. 16 (1997) 155–162.
[22] M. Forina, C. Casolino, E.M. Almansa. The Refinement of PLS Models by Iterative Weighting of Predictor Variables and Objects, Chemom. Intell. Lab. Syst. 68 (2003) 29–40.
[23] W.R. Windham, P.C. Flinn. Comparison of MLR and PLS Regression in NIR Analysis of Quality Components in Diverse Feedstuff Populations. Near Infrared Spectroscopy. Bridging the Gap between Data Analysis and NIR Applications, K.I. Hildrum, T. Isaksson, T. Ns and A. Tandberg (Eds.) Ellis Horwood, Chichester, 1992.
[24] F. Despagne, D.L. Massart. Neural Networks in Multivariate Calibration, Analyst 123 (1998) 157R–178R.
[25] P. Willet, J.M. Barnard, G. Downs. Chemical Similarity Searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
[26] D.H. Rouvray and A.T. Balaban. Chemical Applications of Graph Theory: Applications of Graph Theory, R.J. Wilson and L.W. Beineke (Eds.), Academic, New York, 1979.
[27] M. Urbano Cuadrado, G. Cerruela García, Irene Luque Ruiz,, M.A. Gómez-Nieto. A Method for Clustering and Screening of Long-dimensional Chemical Data Based on Fingerprints and Similarity Measurements, J. Math. Chem. 40 (2006) 15–27.
[28] K.H. Esbensen. Multivariate Data Analysis - in Practice, Camo Process AS, Oslo, 2002..
[29] P. Geladi, B. Kowlaski. Partial Least Square Regression: A Tutorial, Anal. Chim. Acta 35 (1985) 1–17.
[30] J.S. Shenk, M.O. Westerhaus. Calibration the ISI Way, Near Infrared Spectroscopy: the Future Waves, NIR Publications, Chichester, 1996.
[31] S. Wold, M. Sjostrom, L. Eriksson. PLS-Regression: A Basic Tool of Chemometrics, Chemom. Intell. Lab. Syst. 58 (2005) 109–130.
[32] S. Jong. SIMPLS: An Alternative Approach to Partial Least Squares Regression, Chemom. Intell. Lab. Syst. 18 (1993) 251–263.
[33] M. Urbano Cuadrado, M.D. Luque de Castro, P.M. Pérez Juan and M.A. Gómez-Nieto. Comparison and Joint Use of Near Infrared Spectroscopy and Fourier transform Mid Infrared Spectroscopy for the Determination of Wine Parameters, Talanta 66 (2005) 218–224.
[34] M. Urbano Cuadrado, M.D. Luque de Castro, P.M. Pérez Juan and M.A. Gómez-Nieto. Study of Spectral Analytical Data using Fingerprints and Scaled Similarity Measurements, Anal. Bioanal. Chem. 381 (2005) 953–963.

Table 1. $R^2$ and $SECV$ obtained with standard and maximum normalization methods

| Parameter | Fingerprint dimension | Standard | | Maximum | |
|---|---|---|---|---|---|
| | | $R^2$ | SECV | $R^2$ | SECV |
| Total Acidity | Binary | 0.53 | 0.43 meq/L | 0.71 | 0.41 meq/L |
| | Ternary | 0.72 | 0.35 meq/L | 0.74 | 0.31 meq/L |
| | Quaternary | 0.62 | 0.41 meq/L | 0.70 | 0.33 meq/L |
| | Quintary | 0.58 | 0.43 meq/L | 0.68 | 0.37 meq/L |
| Reducing Sugar | Binary | 0.56 | 0.35 meq/L | 0.66 | 0.29 meq/L |
| | Ternary | 0.58 | 0.37 meq/L | 0.62 | 0.32 meq/L |
| | Quaternary | 0.61 | 0.31 meq/L | 0.75 | 0.23 meq/L |
| | Quintary | 0.55 | 0.38 meq/L | 0.67 | 0.37 meq/L |

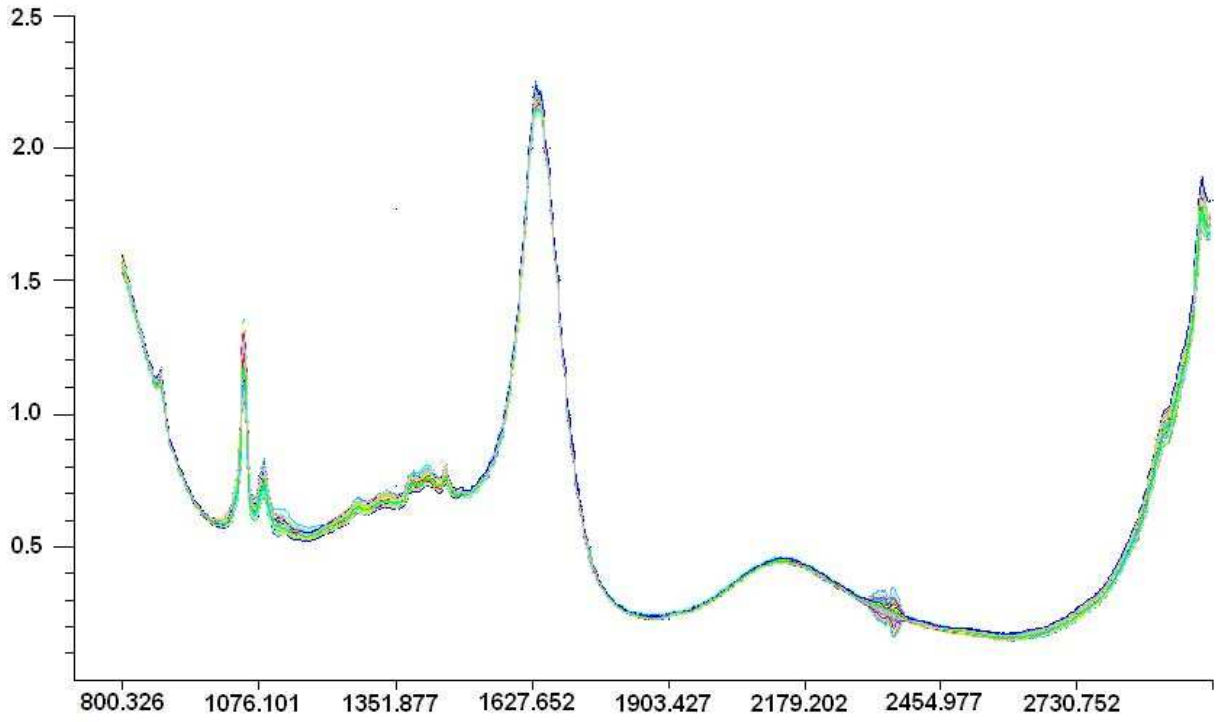



Figure 1. MIR spectra of the samples employed for models fitting and testing

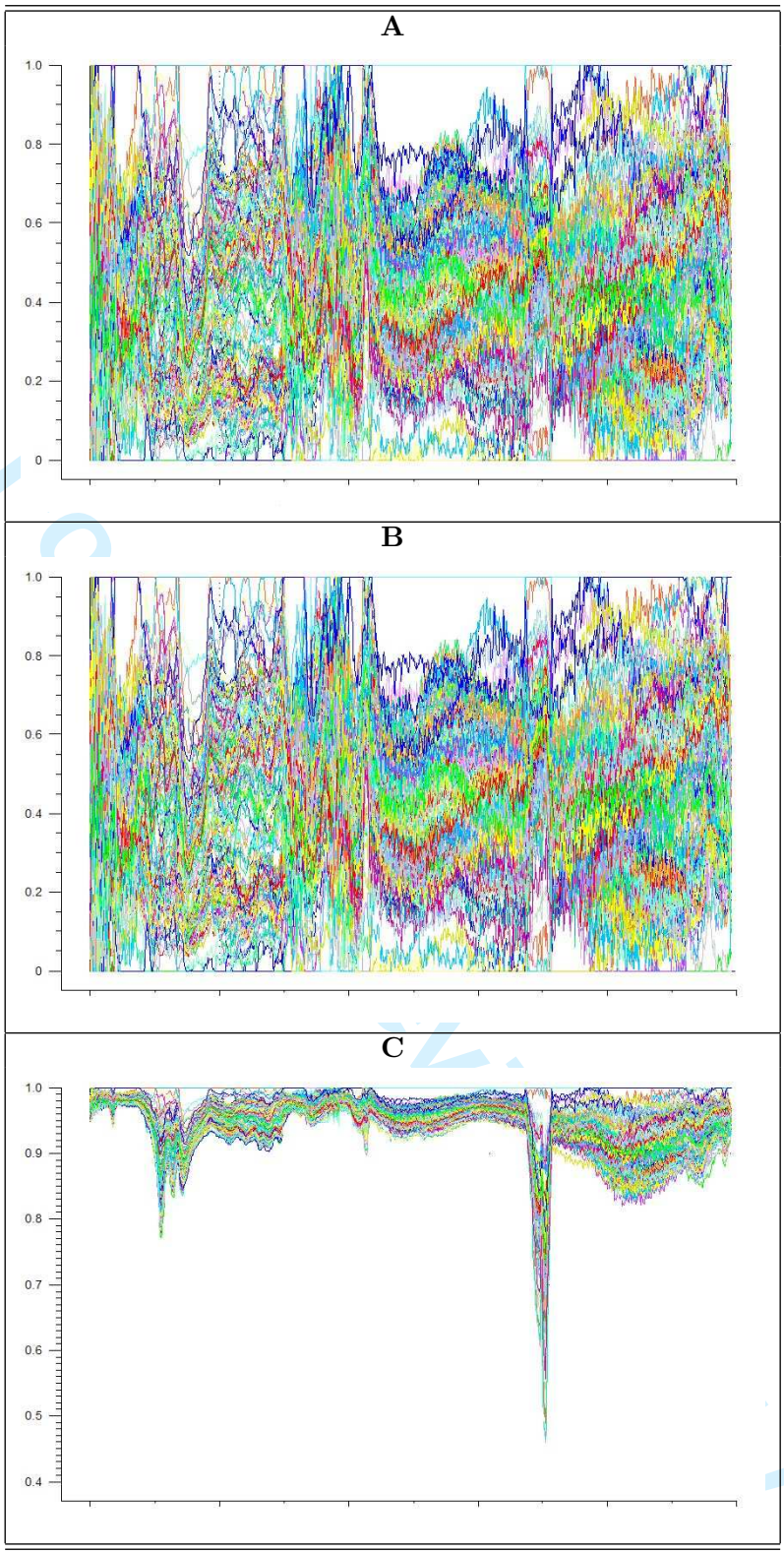



Figure 2. Spectra normalised within the range [0,1] by the (A) standard, (B) logarithmic and (C) maximum methods

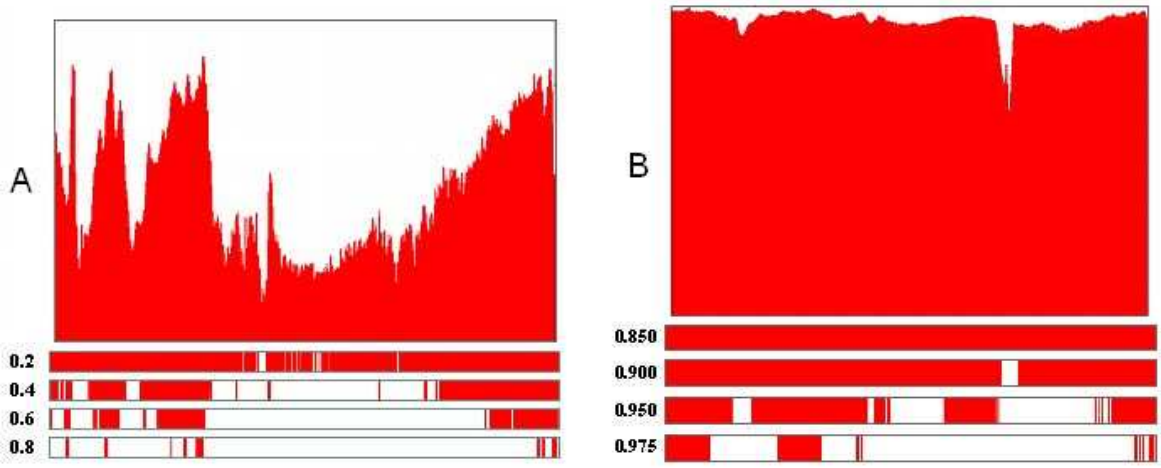



Figure 3. Normalised spectra of the sample T166 and their corresponding binary fingerprints for the (A) standard and (B) maximum methods

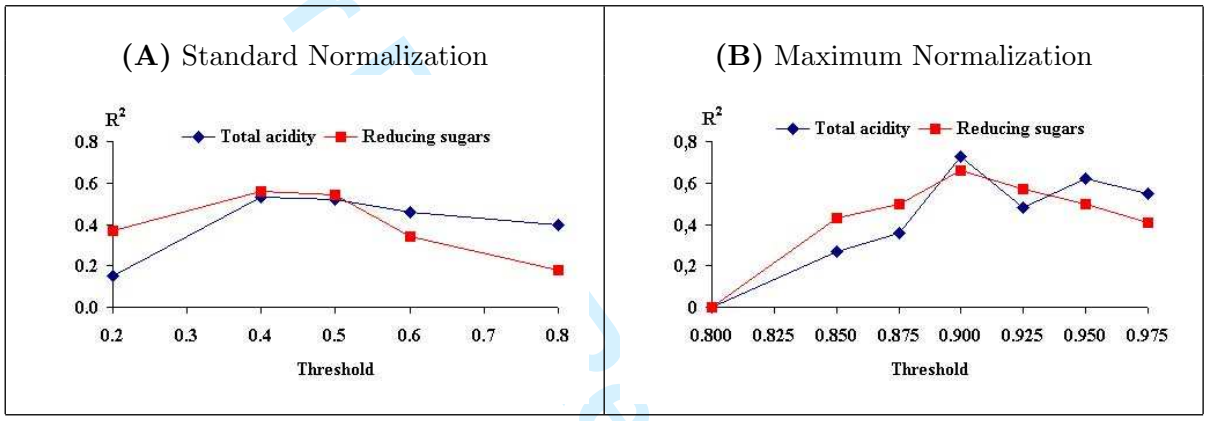



Figure 4. $R^2$ vs. threshold plots for the (A) standard and (B) maximum methods in building binary fingerprints

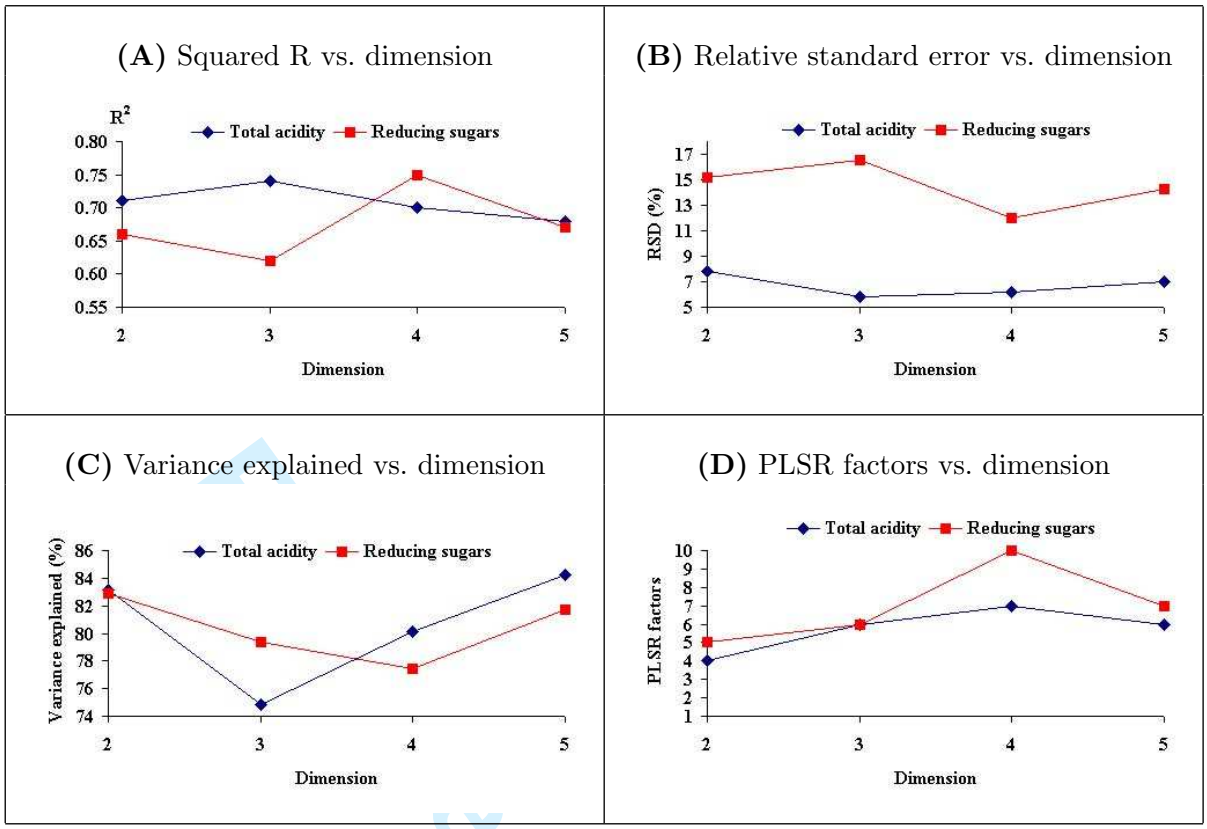



Figure 5. Statistic parameters as a function of fingerprint dimension: (A) $R^2$, (B) relative standard error in prediction, RSD, (C) variance explained by the fist 10 principal components and (D) optimal number of PLSR factors

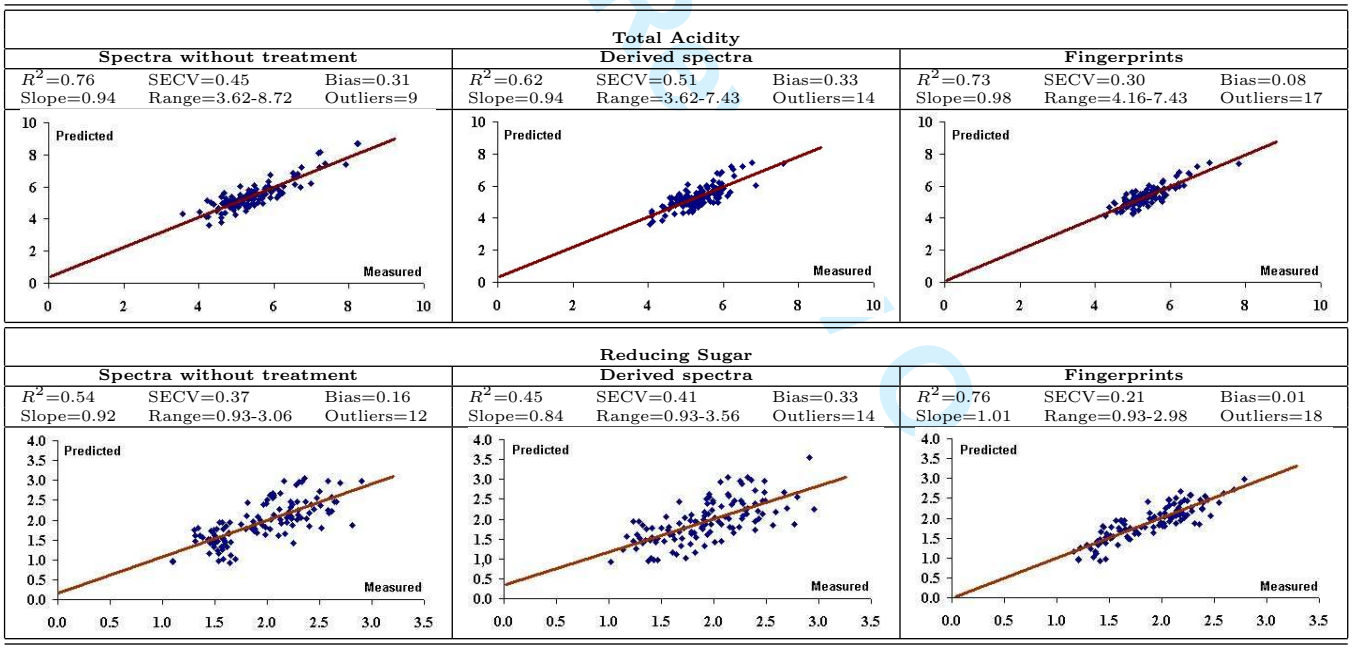
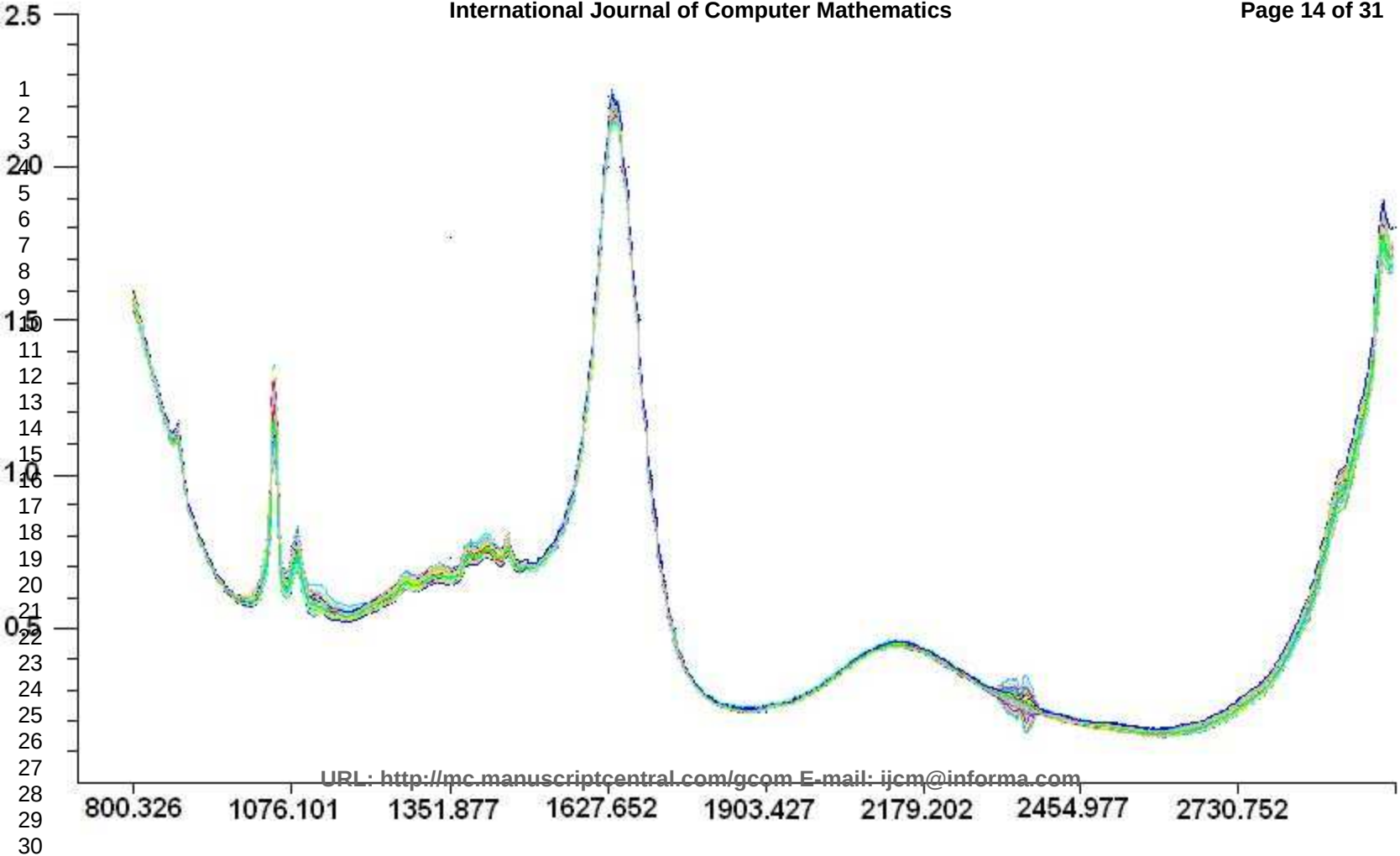
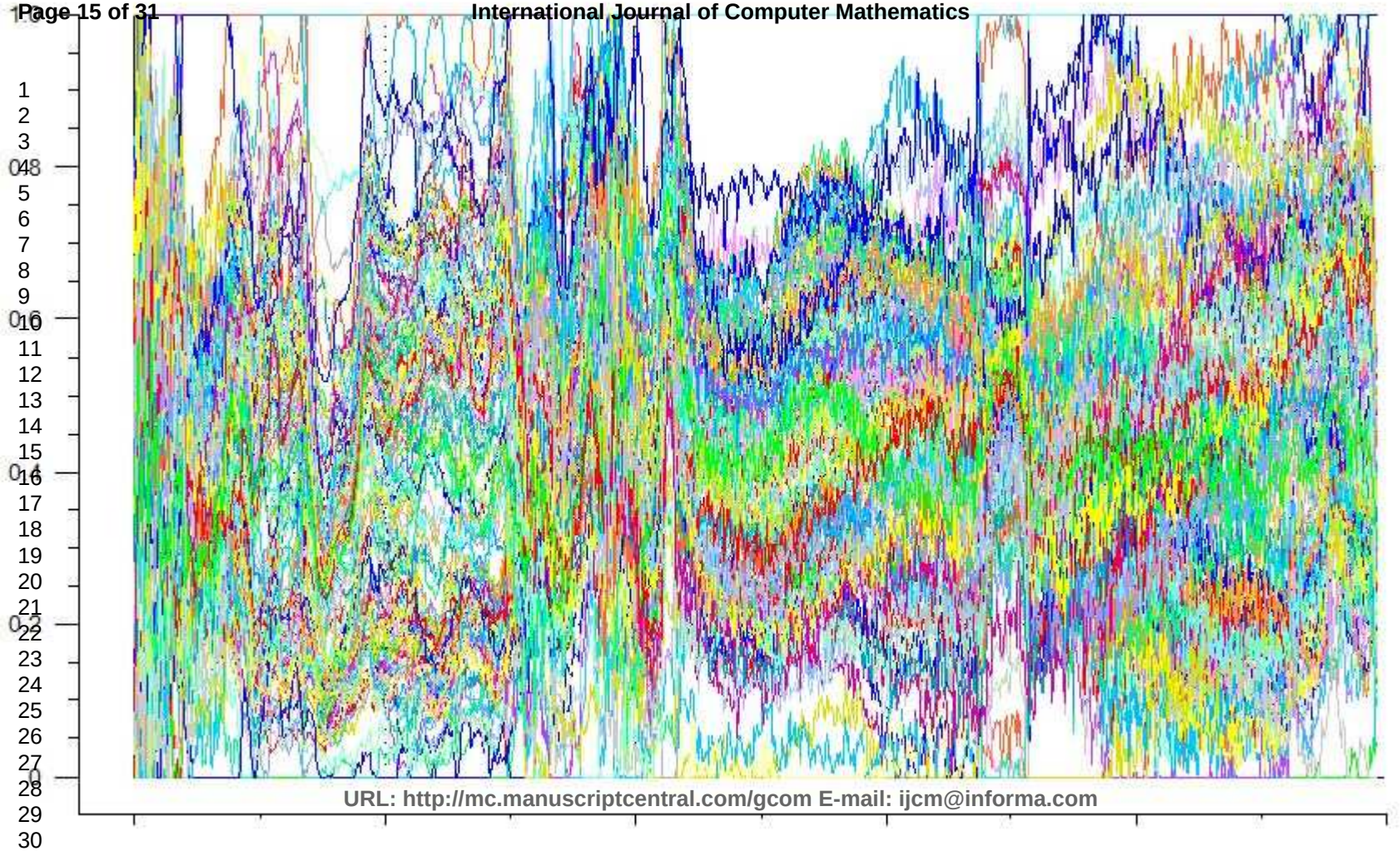
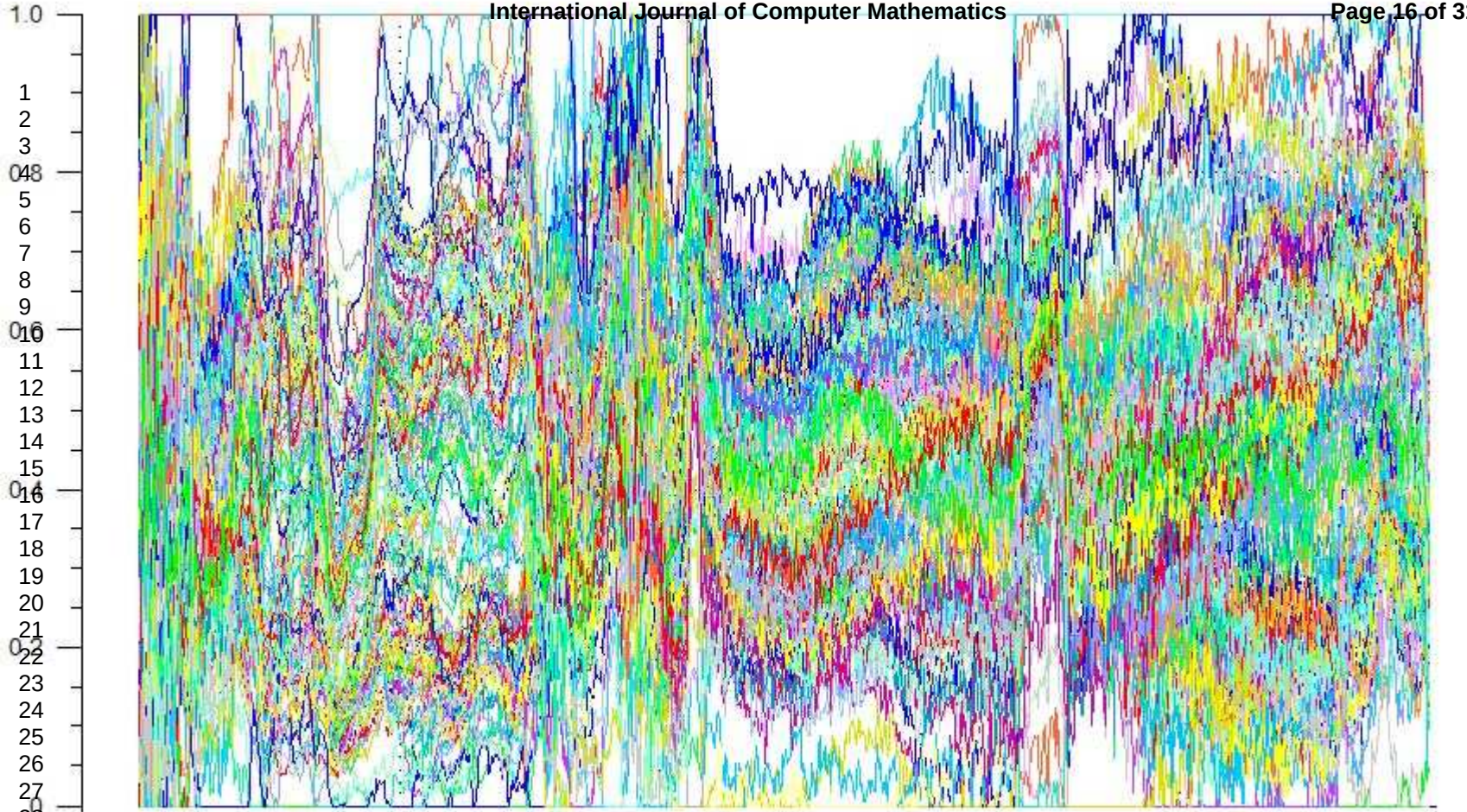
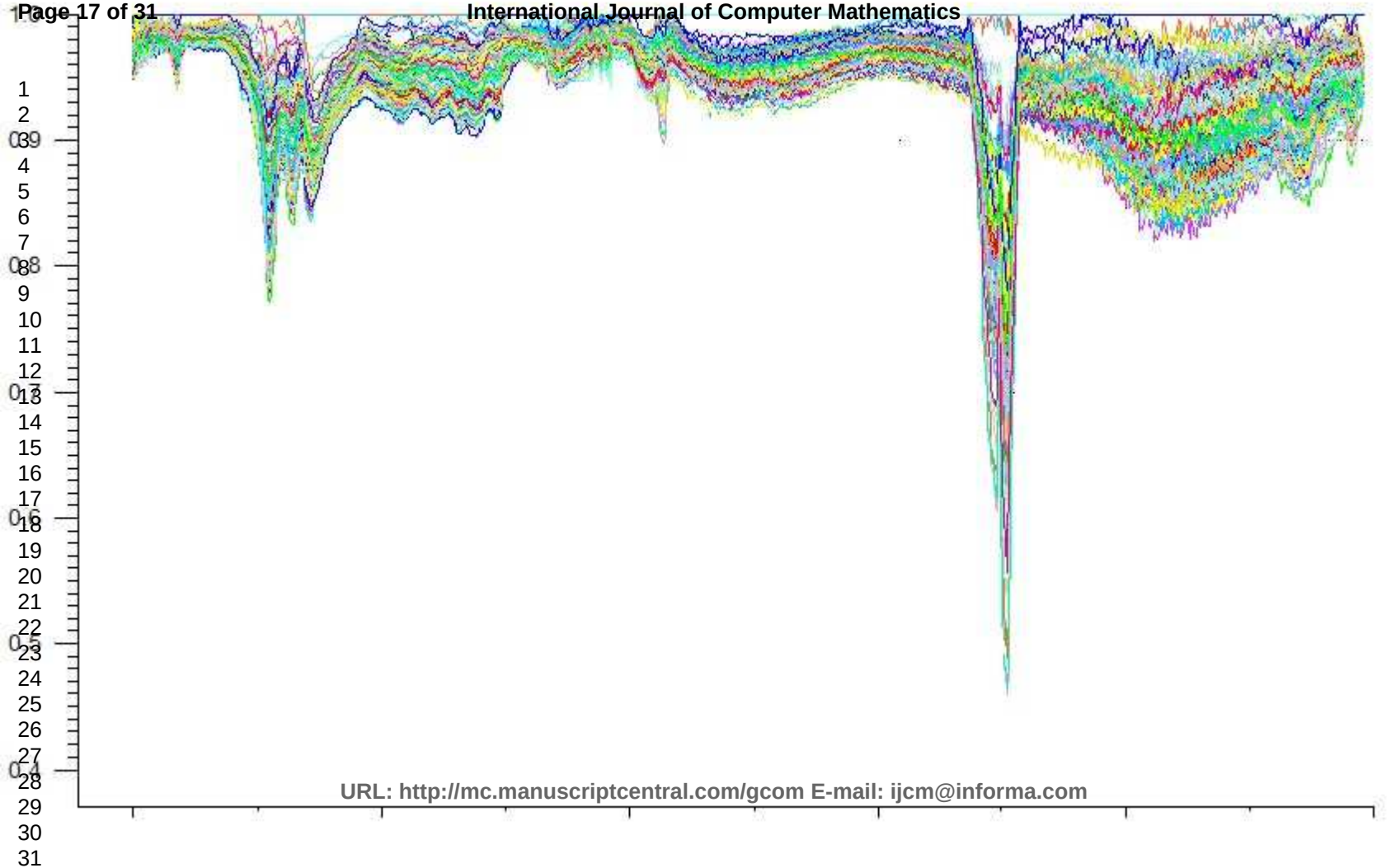


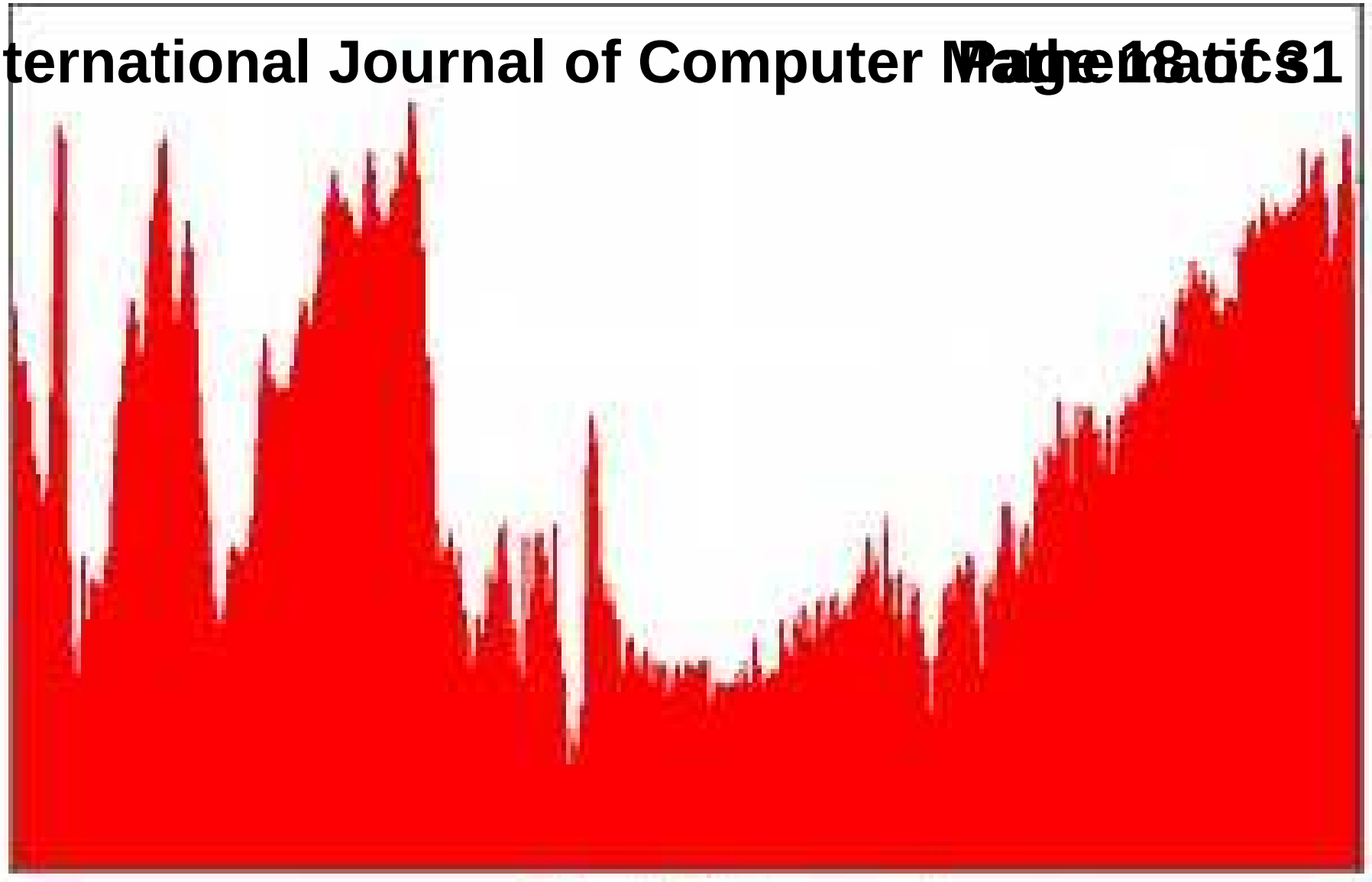

Figure 6. Comparison of graphical and statistics results using multivalued fingerprints regard to both spectra without the processing proposed and derivative spectra with the Norris method

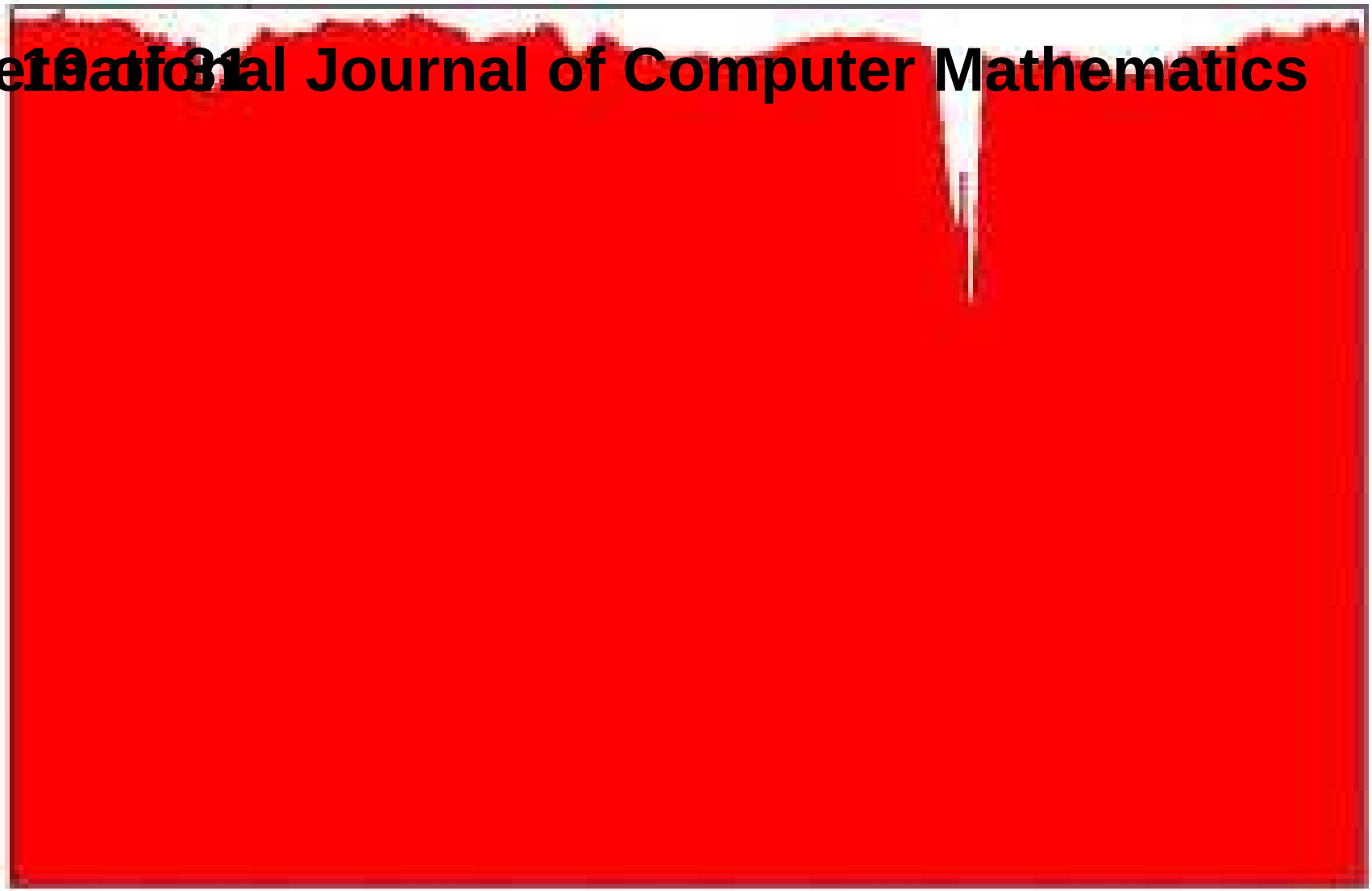

2.5
2
1.5
1
0.5
800.326
1076.101
1351.877
1627.652
1903.427
2179.202
2454.977
2730.752

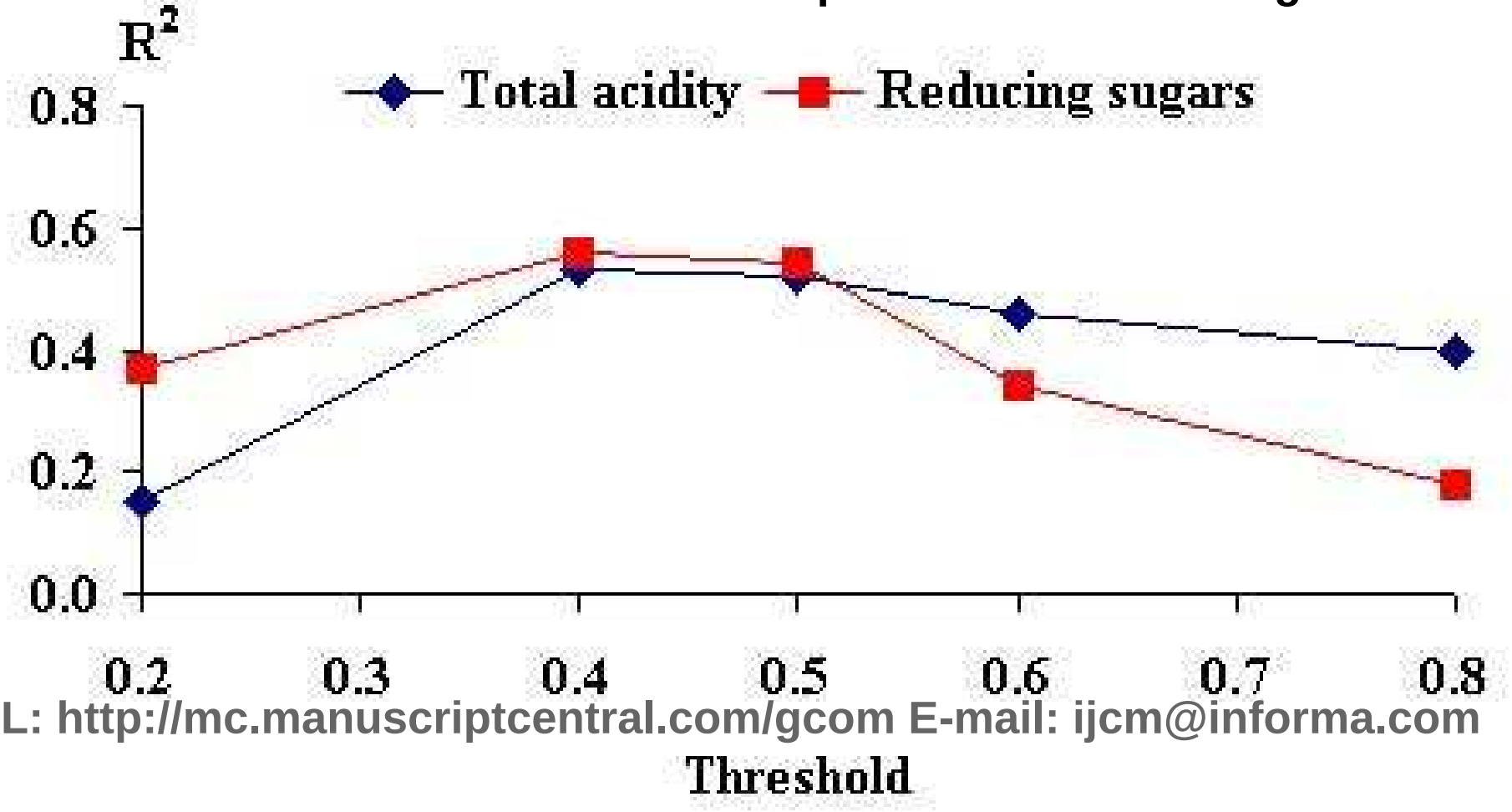


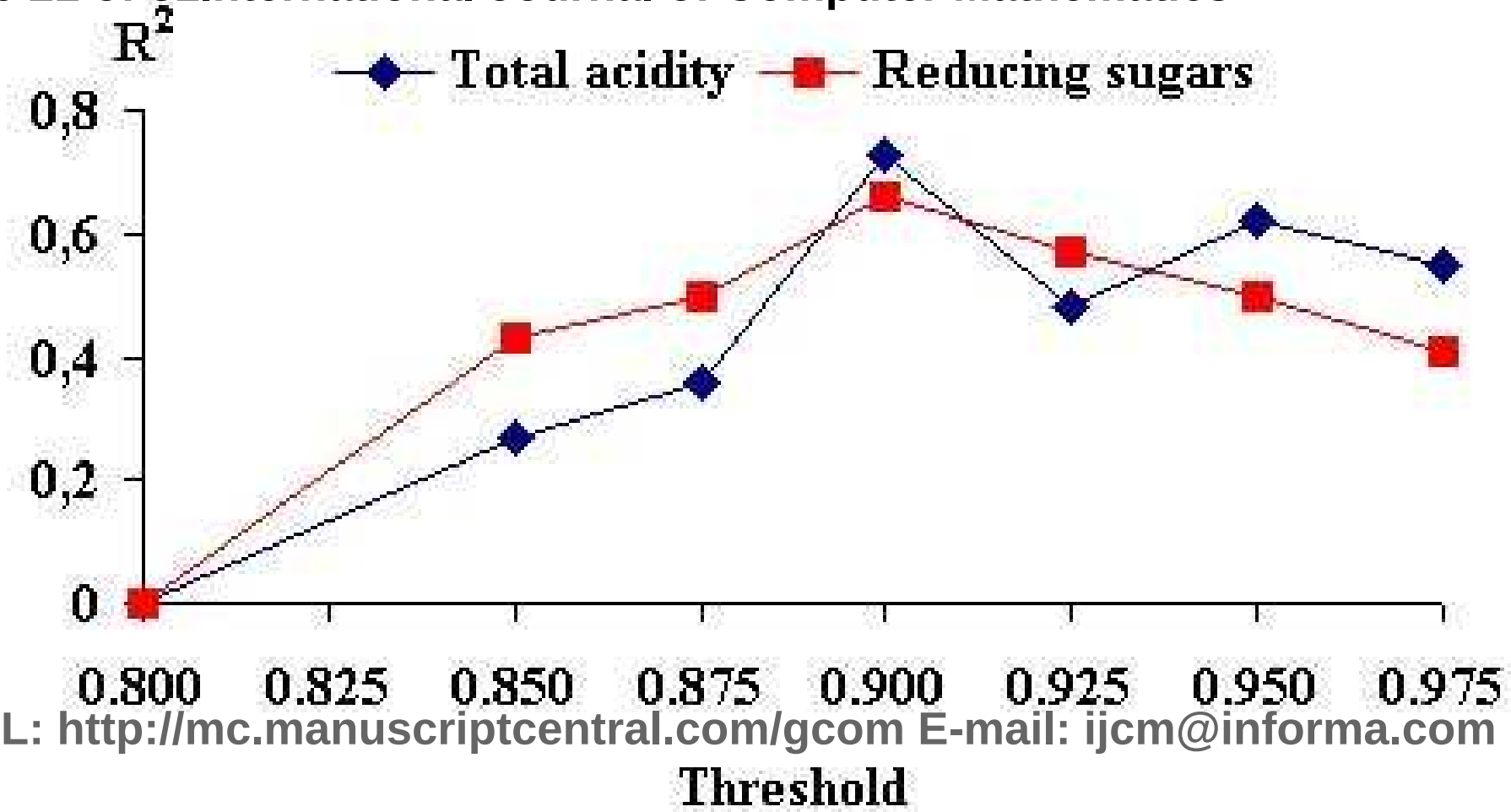


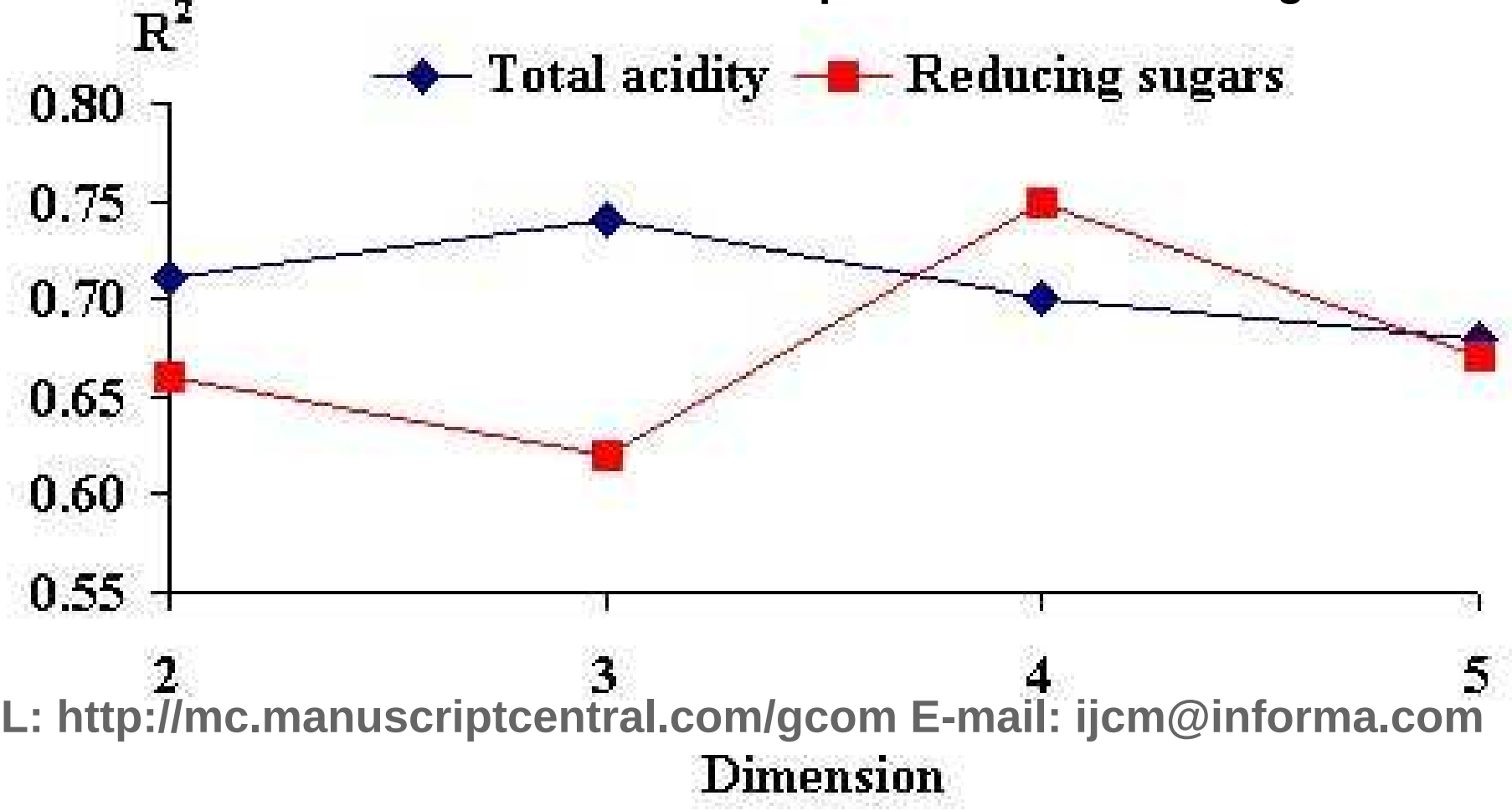


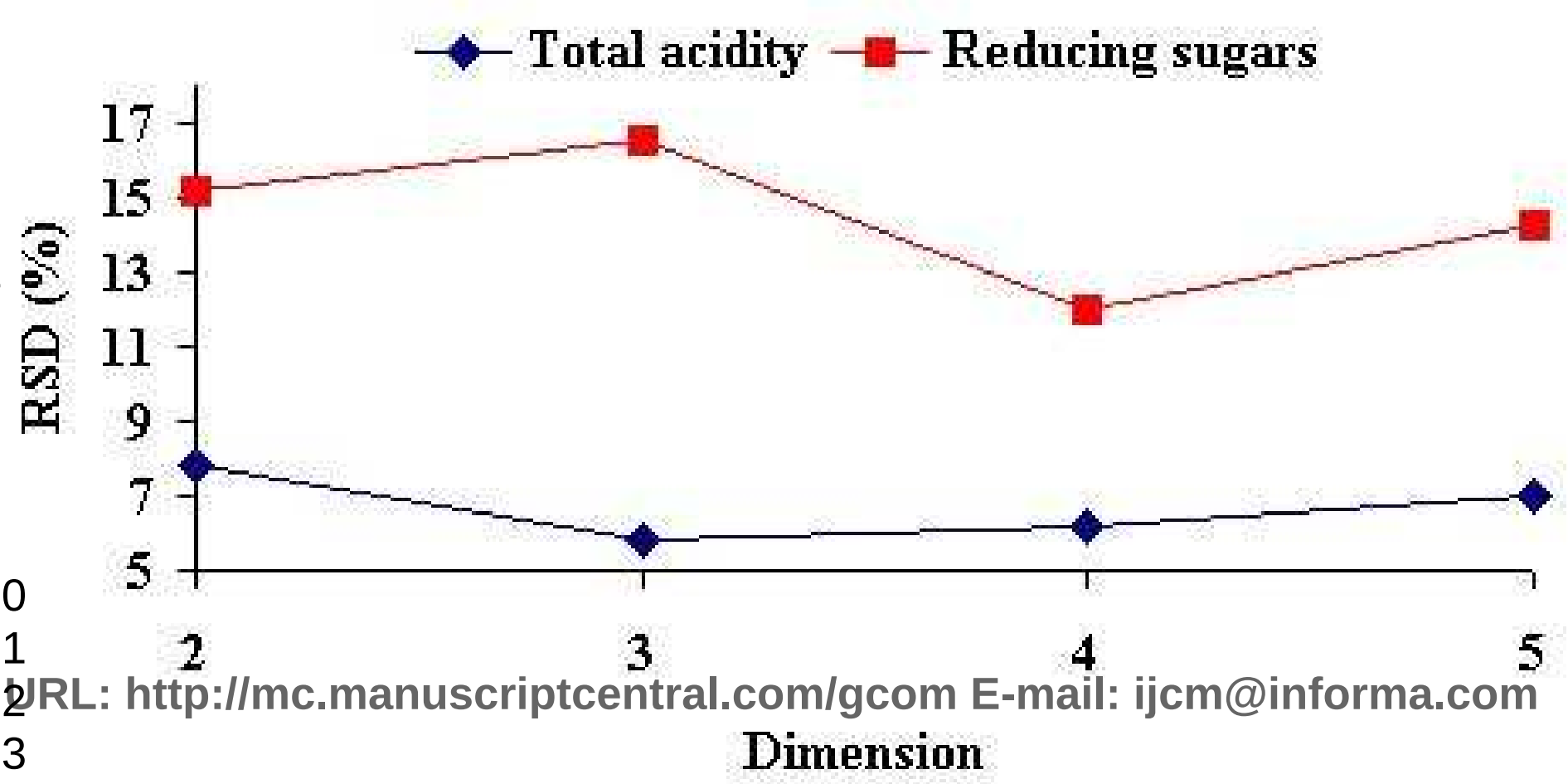


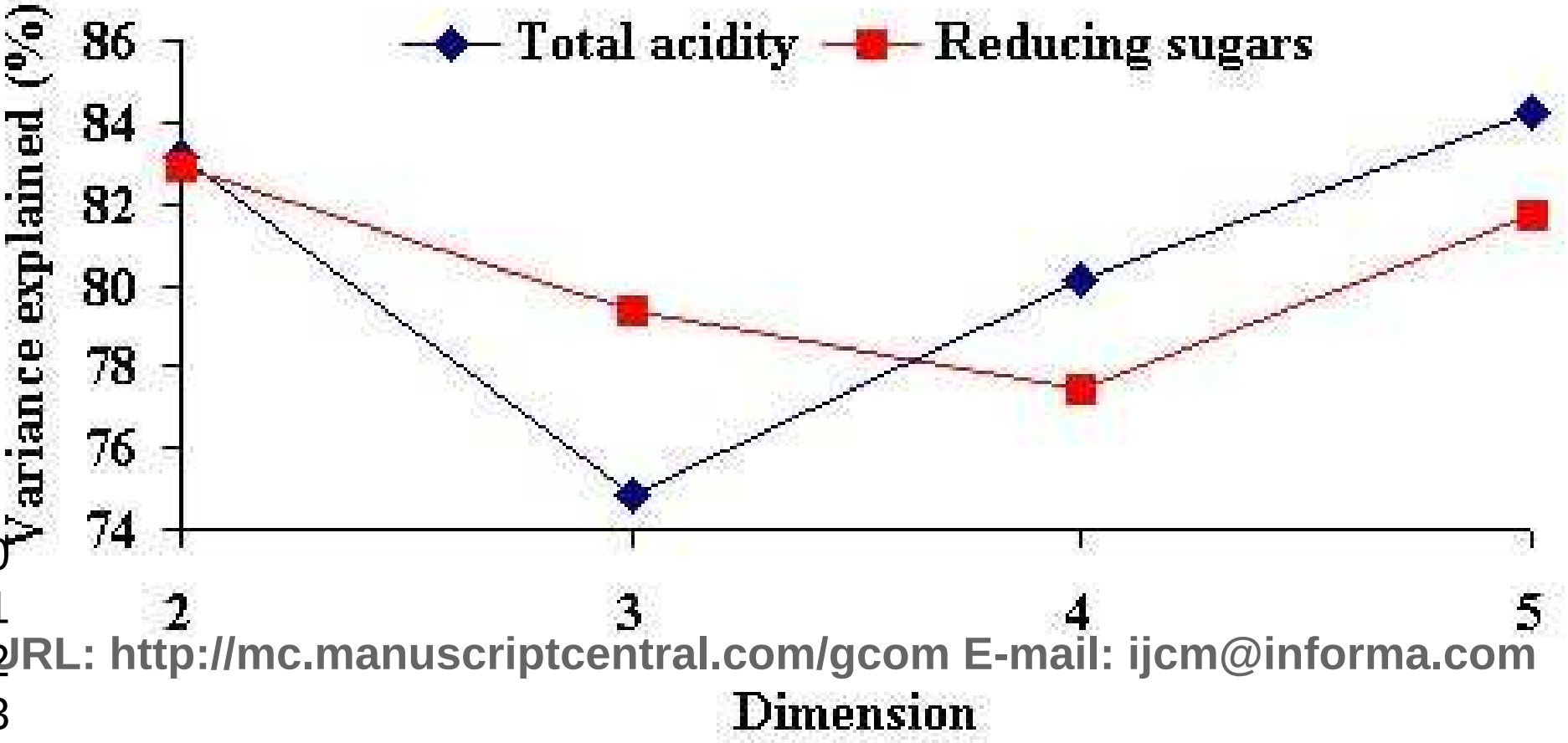


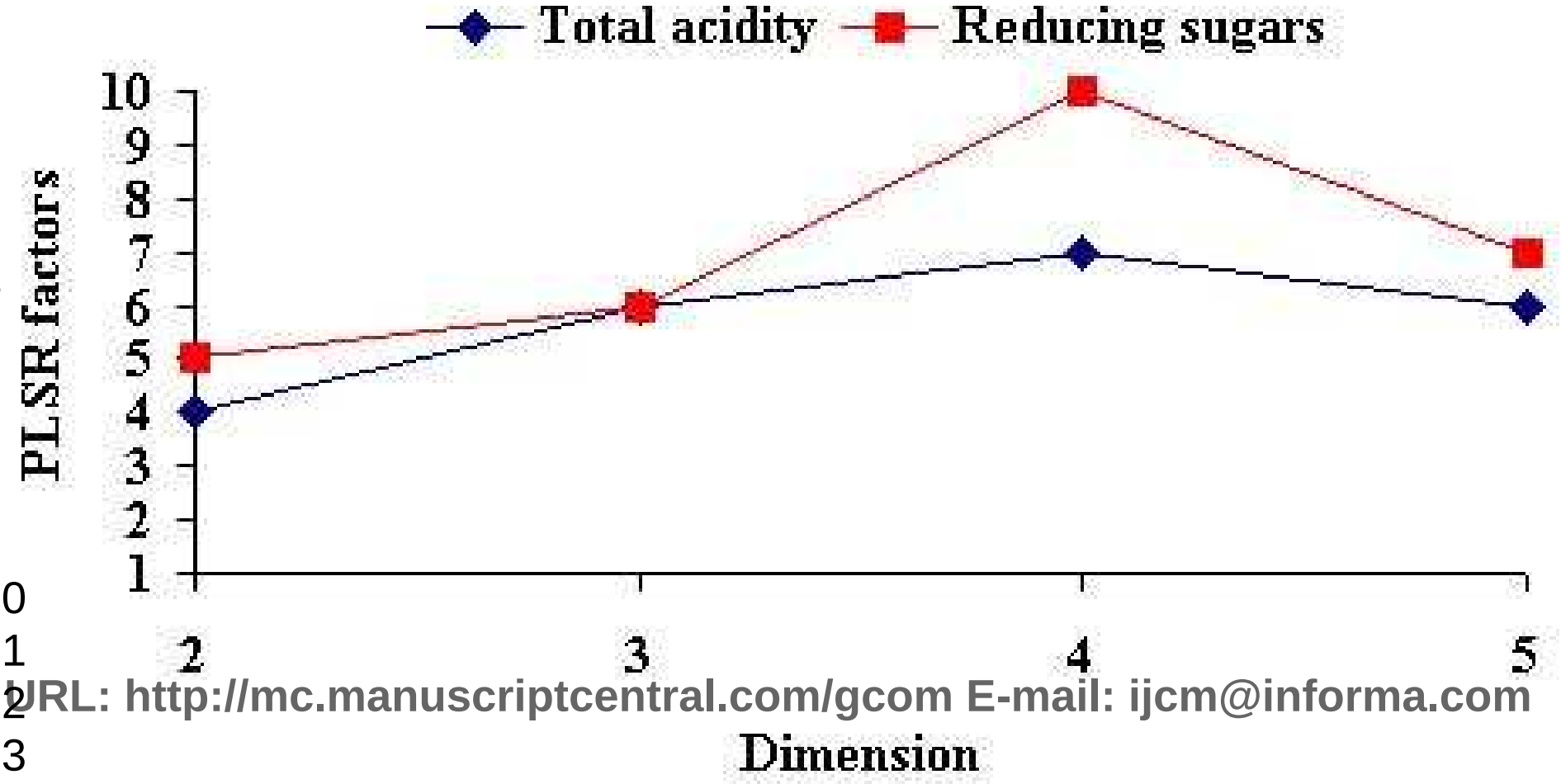

R²
Total acidity
Reducing sugars
0.8
0.6
0.4
0.2
0.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Threshold

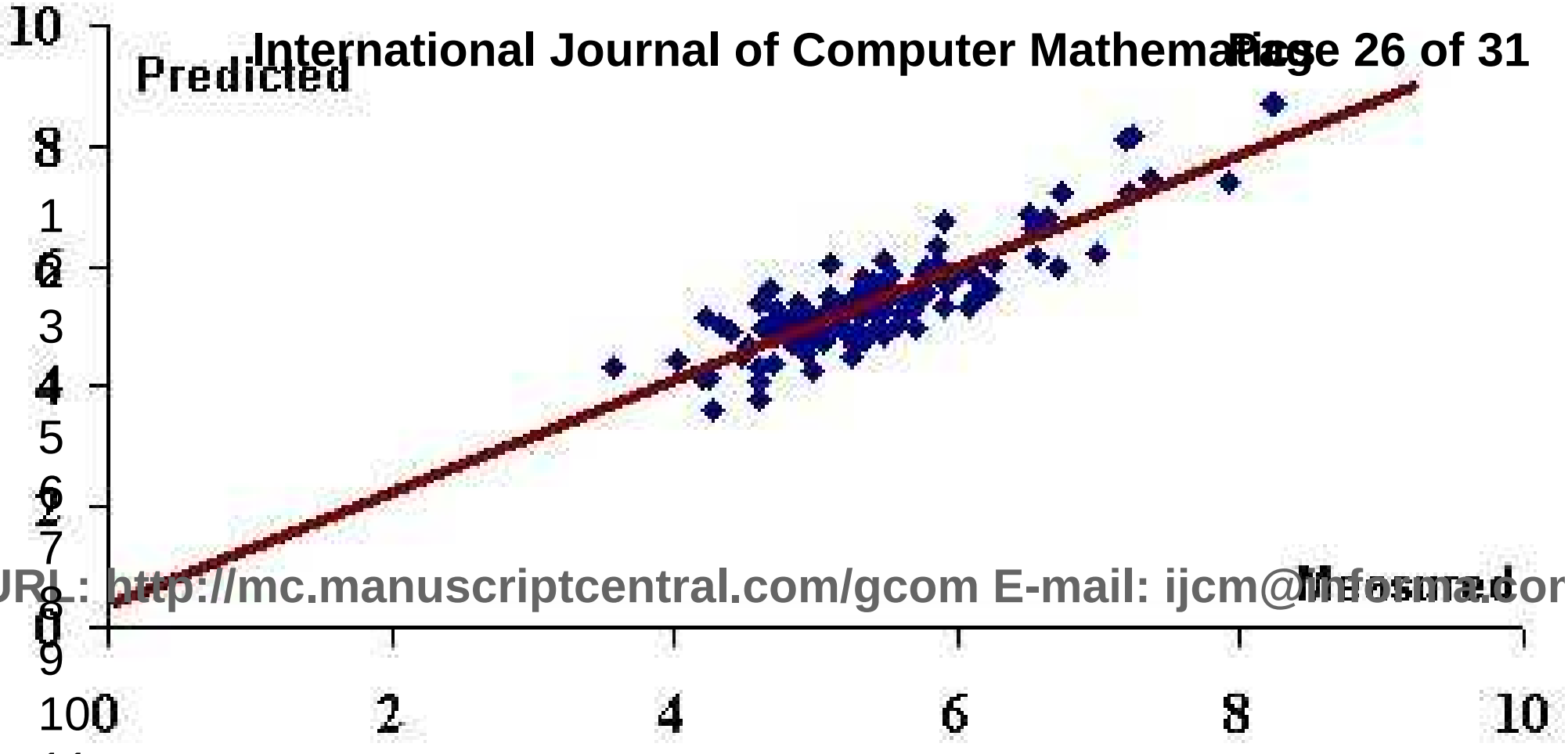

R²
Total acidity
Reducing sugars
0,8
0,6
0,4
0,2
0
0.800
0.825
0.850
0.875
0.900
0.925
0.950
0.975
Threshold

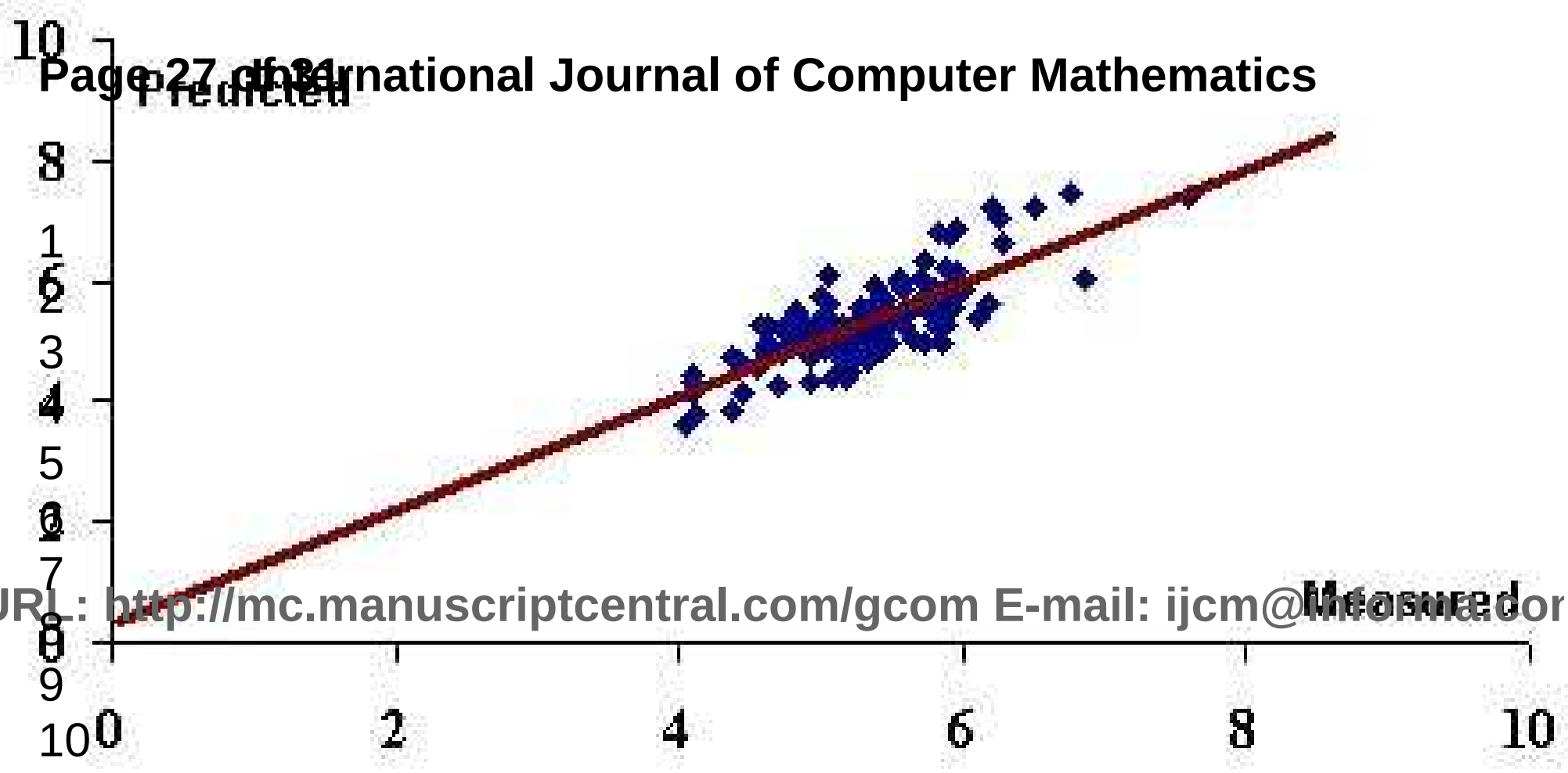

R²
Total acidity
Reducing sugars
0.80
0.75
0.70
0.65
0.60
0.55
2
3
4
5
Dimension

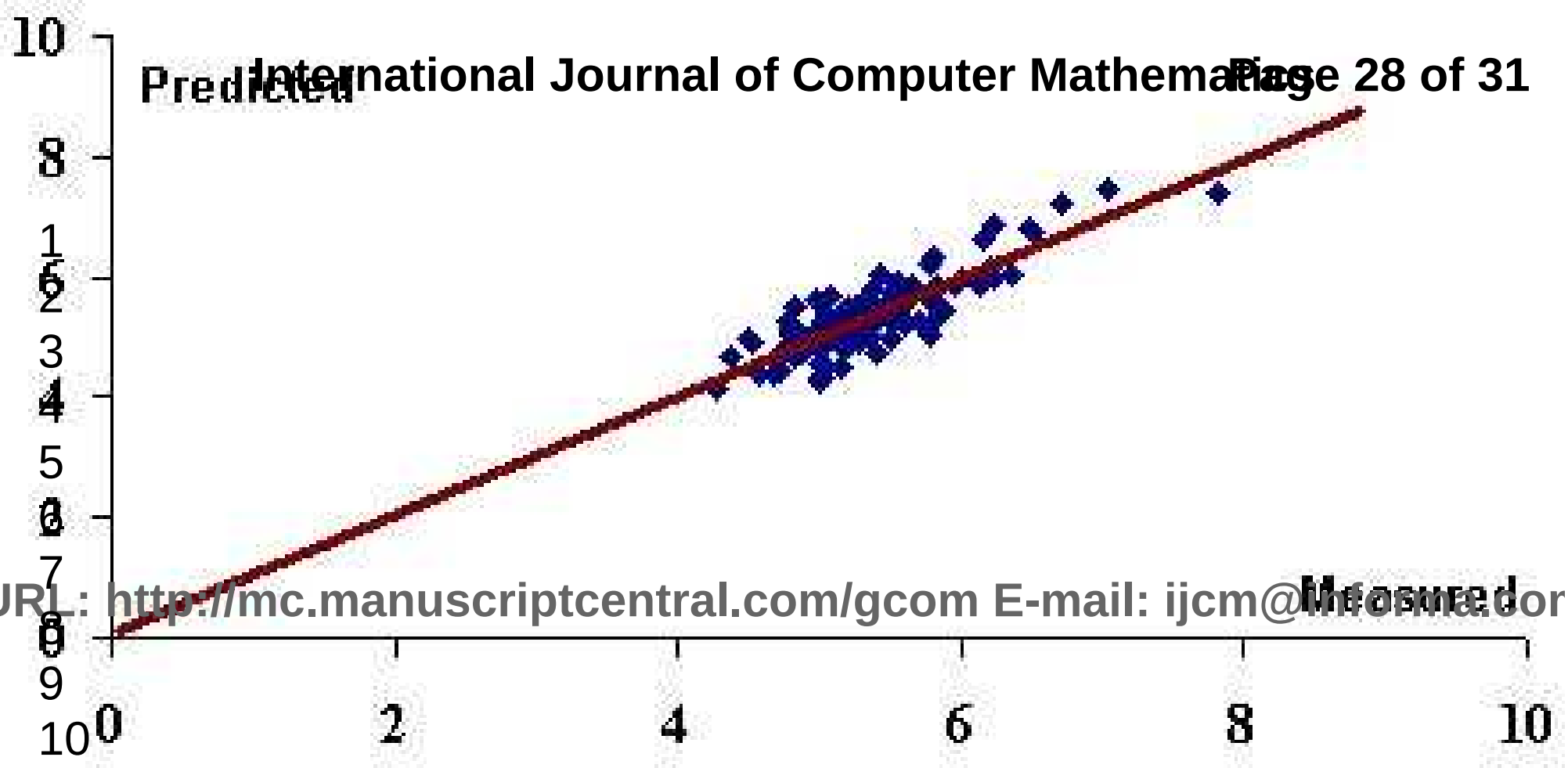

Total acidity
Reducing sugars
RSD (%)
17
15
13
11
9
7
5
2
3
4
5
Dimension


Total acidity
Reducing sugars
Variance explained (%)
86
84
82
80
78
76
74
2
3
4
5
Dimension


Total acidity
Reducing sugars
PLSR factors
10
9
8
7
6
5
4
3
2
1
2
3
4
5
Dimension


Predicted
Measured
0
2
4
6
8
10




Predicted
Measured
0
2
4
6
8
10