ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
17 (1998) 1001–1013

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

The star plot: an alternative display method for multivariate data in the analysis of food and drugs

W. Wu, Q. Guo, P.F. de Aguiar, D.L. Massart *

ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan, 103, B-1090 Brussels, Belgium

Received 13 August 1997

Abstract

The star plot (SP) is a method of displaying multivariate data. It can be used to display data with more than two variables. Combined with principal component analysis (PCA), more than two PCs can be displayed in one plane. Different variants of this method are applied to an atomic absorption spectrometry (AAS) data set and to three near infra-red (NIR) spectral data sets. The results show that SP offers an easy way of visualising the multivariate data for food and drugs in a plane, and it is able to help the analyst to identify and to detect different qualities of food and drug composition. Moreover, when an object is added or removed, the PC's must be computed all over again, which is not the case for the SP-plot. The application of SP to the examples presented in the text suggests that the SP approach can be applied as an alternative method for displaying multivariate chemometric data in place of PCA or, to improve visualisation of the results already obtained with PCA.

Keywords: Multivariate data display; Star plot; PCA

1. Introduction

When objects can be shown in a plane, one can almost immediately recognise, amongst others, clustering tendencies, and extreme or outlying objects [1–4]. However, analytical data often have more than 2 variables, so that it is impossible to achieve direct graphical representation. To reduce the data dimensions, to allow the display of the multivariate data in a plane, one usually applies principal component analysis (PCA) [5,6]. This is also the method used as a reference throughout this paper.

To explore the data structure with the PCA approach, one will have to inspect two PCs at a time, thereby not using the rest of the information contained in the other PCs. When one needs more than 2 PCs to obtain the necessary information from the data, one will not be able to display this information completely in one plane. This can be seen, if not as a drawback, at least as a limitation on the use of PCA.

Another important limitation of the use of PCA is that this technique is data set dependent, that is, if one or more objects have to be added or removed from the data, one will have to recalculate the PCs following the whole procedure of selection and interpretation of the PCs once more.

* Corresponding author. Tel.: +32 2 4774737; fax: +32 2 4774735; e-mail: fabi@vub.vub.ac.be

0731-7085/98/$19.00
PII S0731-7085(98)00066-1

In this paper we review a method, called star plot (SP) [7–9], that has not been used frequently but that, in our opinion, could be very useful to the analyst due to its simplicity. It is based on the projection of objects on a plane, similar to what is done when plotting one PC against another. The two main differences between the PCA and the SP approaches are: (i) in the latter all the information is plotted in two dimensions, decreasing significantly the limitations encountered when applying PCA and facilitating the interpretation of the data structure, and (ii) when applied to the raw data, the SP is data set independent and each object is projected separately.

The aim of the study presented here is to evaluate the possibility of applying this method to pharmaceutical and food analysis and to compare it to the PCA approach. It must be stressed however, that SP is not a clustering method and aims only to guide the user on the interpretation of the data structure, similar to what PCA does.

In extreme cases, where neither PCA nor SP approaches are able to achieve good results (e.g. data set 4), a combination of both approaches is proposed. It should be clarified that, due to the combination of both PCA and SP, the advantage of SP of being data set independent is lost.

2. Theory

Using this method, an n-dimensional data set is mapped into a two-dimensional plane by two non-linear functions (sine and cosine). The data are transformed for each object separately. Each n-dimensional object is projected as an n-broken-line (star line) in the plane. After mapping, the end points of the star lines (star points) can be used to display the structure of the data.

2.1. Notation

In this paper capital letters A and B represent multi-dimensional data points. Each point is described by a vector $(x_1, x_2, x_3,\ldots, x_n)$, and capital letters X and Y are co-ordinates of a plane.

2.2. Procedure

Let us first show how the star plot displays a three-dimensional data point in a plane. Suppose A is a point in three-dimensional space with co-ordinates (x_1, x_2, x_3). In Fig. 1, a 3-segment-broken-line OCDA is used to display the three-dimensional data point, where the lengths of OC, CD, DA are equal to x_1, x_2, x_3, respectively, and the angles $(\theta_1, \theta_2, \theta_3)$ of these lines are functions of x_1, x_2, x_3 as follows:

$$\theta_1 = x_1\pi/(x_1 + x_2 + x_3) \quad (1)$$

$$\theta_2 = (x_1 + x_2)\pi/(x_1 + x_2 + x_3) \quad (2)$$

$$\theta_3 = (x_1 + x_2 + x_3)\pi/(x_1 + x_2 + x_3) \quad (3)$$

where π radians is equal to 180°.

In the same way, n-dimensional data can be expressed as n-segment-broken-lines in a plane. Because the shape of the n-segment-broken-line looks somewhat like a locus of a star, even like the orbit of a comet when n extends to infinity, this plot is called a star plot, the n-segment-broken-line the star line, the end point the star point and the X–Y plane the star plane.

Suppose A is an n-dimensional object defined by:

$$A = (x_1, x_2, x_3,\ldots, x_n) \quad (4)$$

The following equations are used to define the points on the n-segment-broken-line:

$$X_i = \sum_{k=1}^{i} x_k \cos\theta_k \quad (5)$$

$$Y_i = \sum_{k=1}^{i} x_k \sin\theta_k \quad (6)$$

where

$$\theta_i = \frac{\sum_{k=1}^{i} x_k\pi}{\sum_{j=1}^{n} x_j} \quad i = 1, 2, 3,\ldots, n \quad (7)$$

and i is the ith point of the n-segment-broken-line, X_i and Y_i are the co-ordinates of the ith point and θ_i is the angle of the ith segment.

Let us consider two examples, where $A = (1, 2, 3)$ and $B = (3, 2, 1)$. For the first example

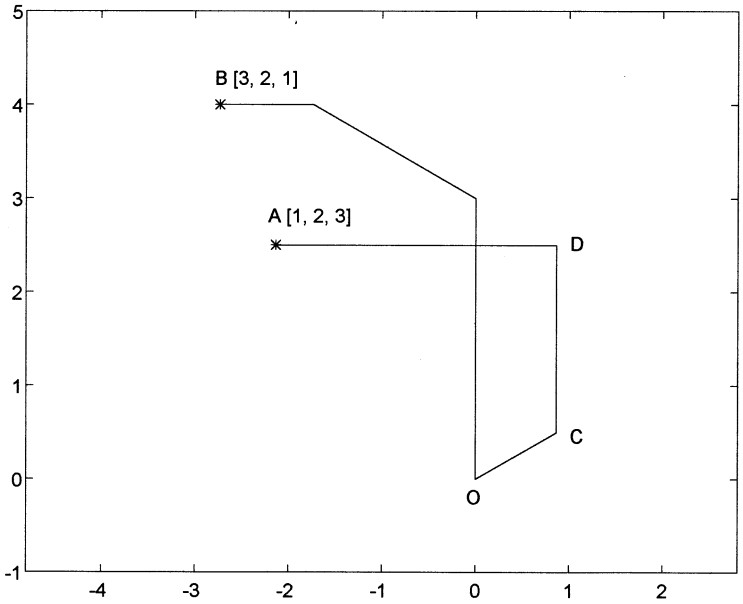

Fig. 1. Star plot of two three-dimensional data points A and B with values (1, 2, 3) and (3, 2, 1).

Table 1
Concentrations of trace metal contents in breads

No.	Bread name	Al	Fe	Zn	Ca	K	Mg	Cu
1	White bread	0.55	1.02	0.66	16.62	128.22	22.14	0.12
2	Bread roll	0.92	1.25	0.76	25.29	120.91	20.52	0.12
3	French stick	1.15	1.07	0.76	26.58	144.83	23.40	0.13
4	French roll	0.96	1.24	0.89	22.78	148.54	26.88	0.14
5	Wheat whole-meal bread	0.66	2.44	2.01	25.24	286.08	69.03	0.19
6	Seven-cereals bread	0.91	1.82	1.16	22.41	204.68	48.25	0.22
7	Croissant	2.11	0.82	0.56	26.17	113.97	17.79	0.12
8	Currant-bread	4.17	1.55	0.59	44.06	383.15	24.94	0.22
9	Sugar bread	1.09	0.86	0.67	31.58	124.36	17.95	0.10
10	Milk bread	0.92	1.20	0.72	55.97	148.32	20.57	0.12
11	Whole-meal bread	2.08	2.27	1.55	41.35	247.43	50.31	0.19
12	Brown bread	0.81	1.67	0.85	25.78	186.24	40.55	0.16
13	Rye bread	1.17	2.94	1.63	40.24	277.55	60.63	0.19

Values in mg 100 g^{-1}; data from [10].

(Fig. 1), the co-ordinates of broken-line points are calculated as follows:

$$\theta_1 = x_1\pi/(x_1 + x_2 + x_3) = \pi/6 \quad (8)$$

$$\theta_2 = (x_1 + x_2)\pi/(x_1 + x_2 + x_3) = \pi/2 \quad (9)$$

$$\theta_3 = (x_1 + x_2 + x_3)\pi/(x_1 + x_2 + x_3) = \pi \quad (10)$$

$$X_1 = x_1 \cos\theta_1 = 0.866 \quad (11)$$

$$Y_1 = x_1 \sin\theta_1 = 0.5 \quad (12)$$

$$X_2 = x_1 \cos\theta_1 + x_2 \cos\theta_2 = 0.866 \quad (13)$$

$$Y_2 = x_1 \sin\theta_1 + x_2 \sin\theta_2 = 2.5 \quad (14)$$

$$X_3 = x_1 \cos\theta_1 + x_2 \cos\theta_2 + x_3 \cos\theta_3 = 2.134 \quad (15)$$

$$Y_3 = x_1 \sin\theta_1 + x_2 \sin\theta_2 + x_3 \sin\theta_3 = 2.5 \quad (16)$$

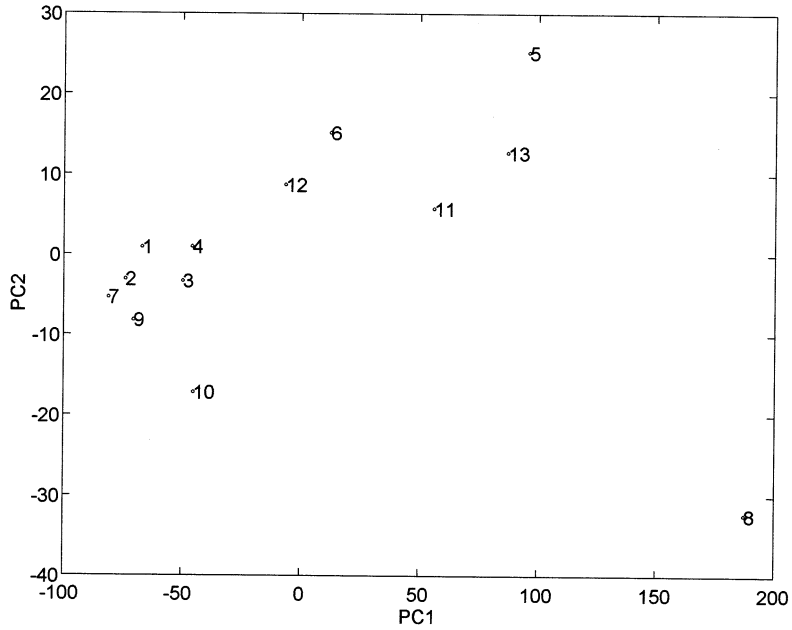

Fig. 2. PC1–PC2 score plot of minerals in 13 types of bread.

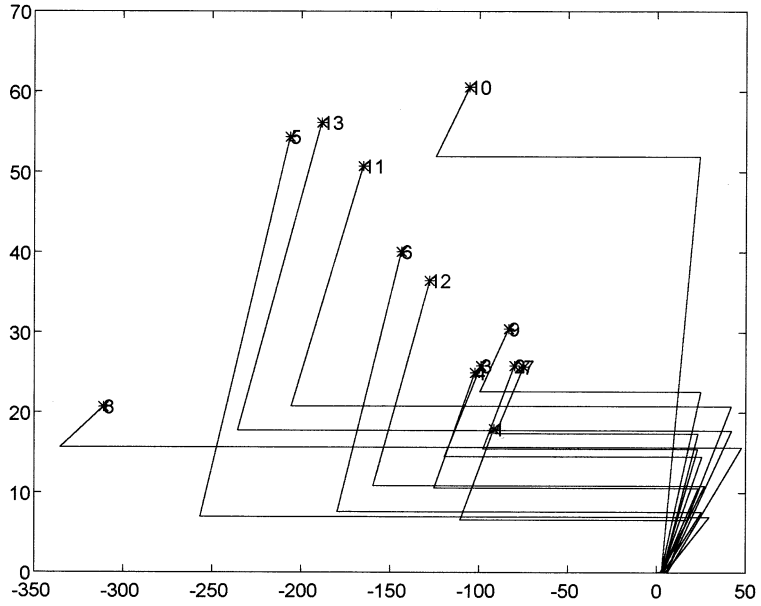

Fig. 3. Star plot of minerals in 13 types of bread for method 2.

For the second example (Fig. 1), we obtain: $\theta_1 = \pi/2$, $\theta_2 = 5\pi/6$, $\theta_3 = \pi$, $X_1 = 0$, $Y_1 = 3$, $X_2 = -1.732$, $Y_2 = 4$, $X_3 = -2.732$ and $Y_3 = 4$.

There are some different definitions of the angle in Eq. (7). In Ref. [7], it is also defined as a function of $x_1, x_2, \ldots, x_n$:

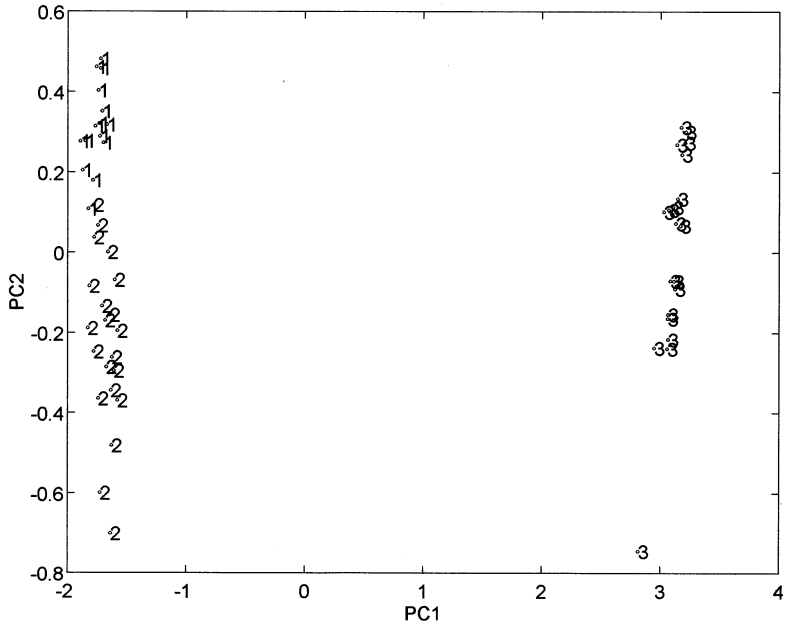

Fig. 4. PC1–PC2 score plot of data set 2, the SNV-transformed data.

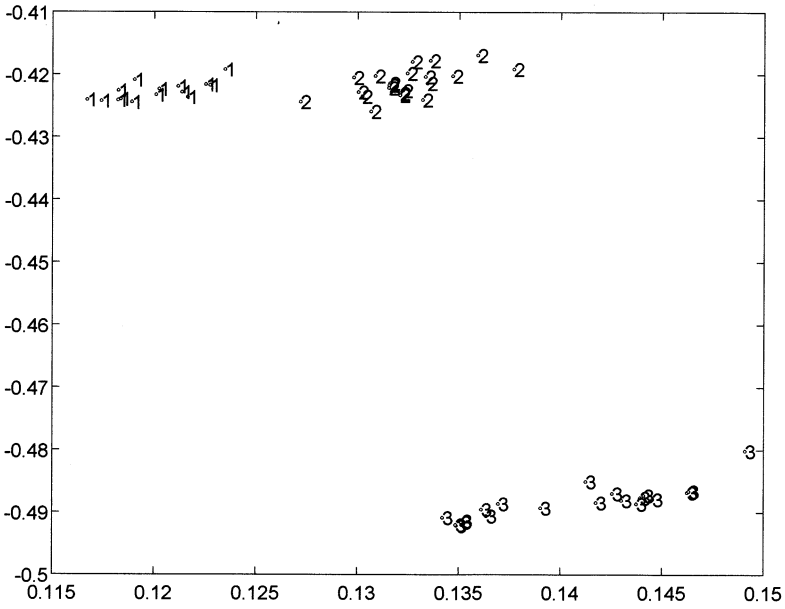

Fig. 5. Star plot of data set 2, method 3 for the SNV-transformed data.

$$\theta_i = \frac{(x_i - x_{\min})\pi}{x_{\max} - x_{\min}} \tag{17}$$

where $x_{\min}$ and $x_{\max}$ are the minimum and maximum value of $x_1, x_2, \ldots, x_n$, respectively.

Kaffka and coworkers [8,9] defined it as a function of the variable index (i):

$$\theta_i = \frac{2i\pi}{n} \tag{18}$$

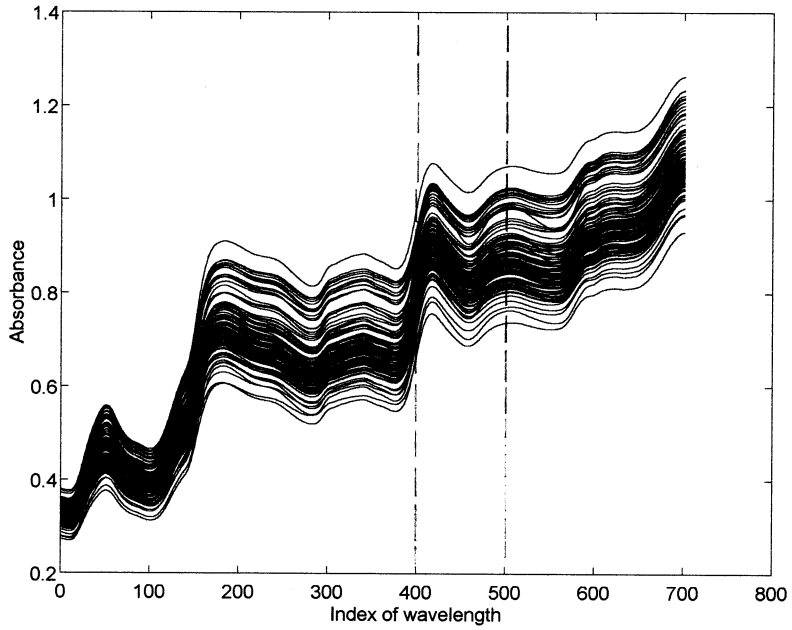

Fig. 6. The NIR spectra of data set 3.

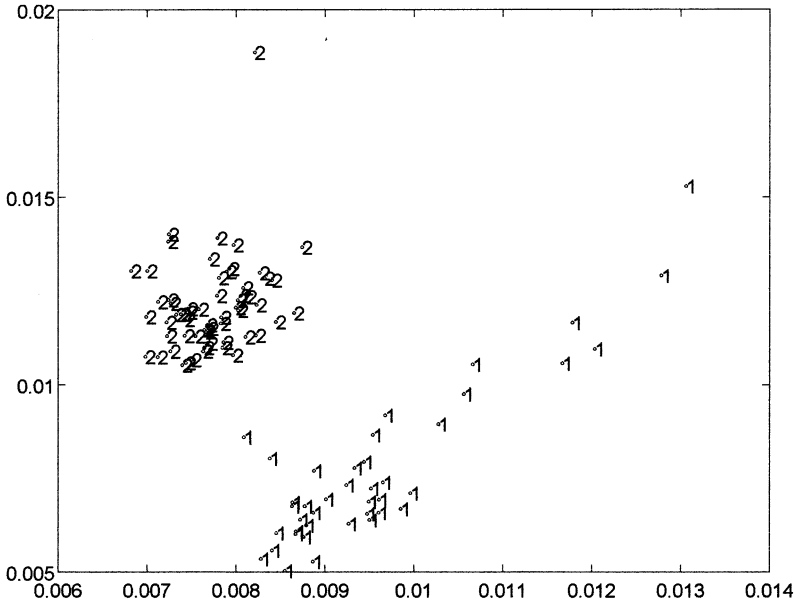

Fig. 7. Star plot of method 3 for data set 3 with the selected region of wavelengths.

In order to conveniently compare the SP procedures, methods 1, 2 and 3 are referred to the star plot approaches using the angle defined by Eqs. (7), (17) and (18), respectively.

Method 1 is a variant of the two others, and is proposed by us. This variant was developed due to the fact that in one of the examples studied, the other two methods were not able to provide better results than PCA.

After transformation of the multi-dimensional

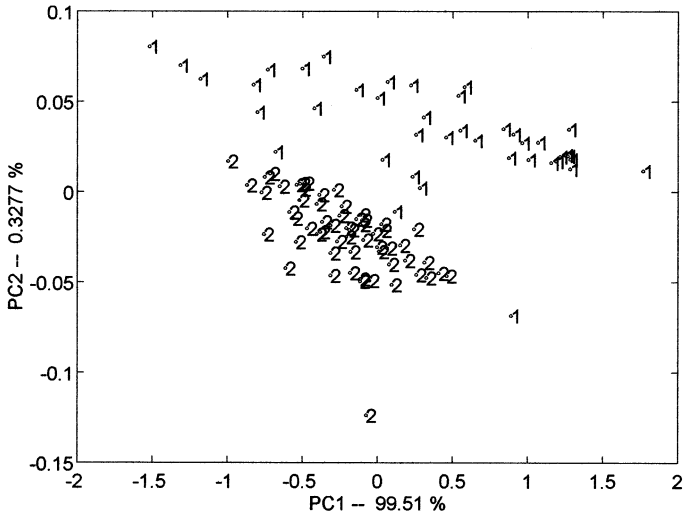

Fig. 8. PC1–PC2 score plot for data set 3 with the selected region of wavelengths.

data into a two-dimensional star plane, the structures of the data can be visually observed. This approach can be considered as projection on nonlinear base lines. Since the star plane can display multi-dimensional data, it can be used as a tool to deal with multivariate data by extracting useful information. The location of the last points (star points) can be used to cluster the data.

3. Experimental

3.1. Data

The star projection method is applied to the following data sets:

3.1.1. Data set 1: trace metal contents of breads

Seven minerals (Al, Zn, Fe, Cu, Mg, Ca, and K) in 13 different types of Belgian bread and bread products, have been determined with atomic absorption spectrometry by Yang [10]. The thirteen types are white bread, bread roll, French stick, French roll, wheat whole-meal bread, seven cereals, croissant, currant bread, sugar-loaf, milk bread, whole-meal bread, brown bread and rye bread. For each type, at least three samples are collected and pooled. Two representative sub-samples are obtained from each pooled sample, and digested in a closed-vessel microwave digestion system. Graphite furnace atomic absorption spectrometry is used to determine the concentration of Al and Cu. The other minerals are measured by flame atomic absorption spectrometry. Recovery experiments and standard reference materials are used to validate the analysis methods. The measurement results are listed in Table 1.

3.1.2. Data set 2: NIR spectra of active drugs and comparator

56 NIR spectra (1100–2500 nm; 700 wavelengths) of tablets containing different dosages (100 and 200 mg) of an experimental active ingredient and a comparator (another drug with similar pharmacological properties) were studied. They were measured with an NIRS systems Model 6500 near-infrared reflectance spectrometer configured with the aperture slit facing upward, and presented as $\log(1/R')$ absorption values. The spectra were measured through the blister package, which contributed to the spectrum at around 1700 nm. To eliminate end effects of the measurements, 15 wavelengths were discarded at, respectively, the beginning and the end of the spectra. The data set (56×670) is marked by three groups. Group 1 (100 mg) contains 15 spectra, group 2 (200 mg) 21 spectra and group 3 (comparator) 20 spectra.

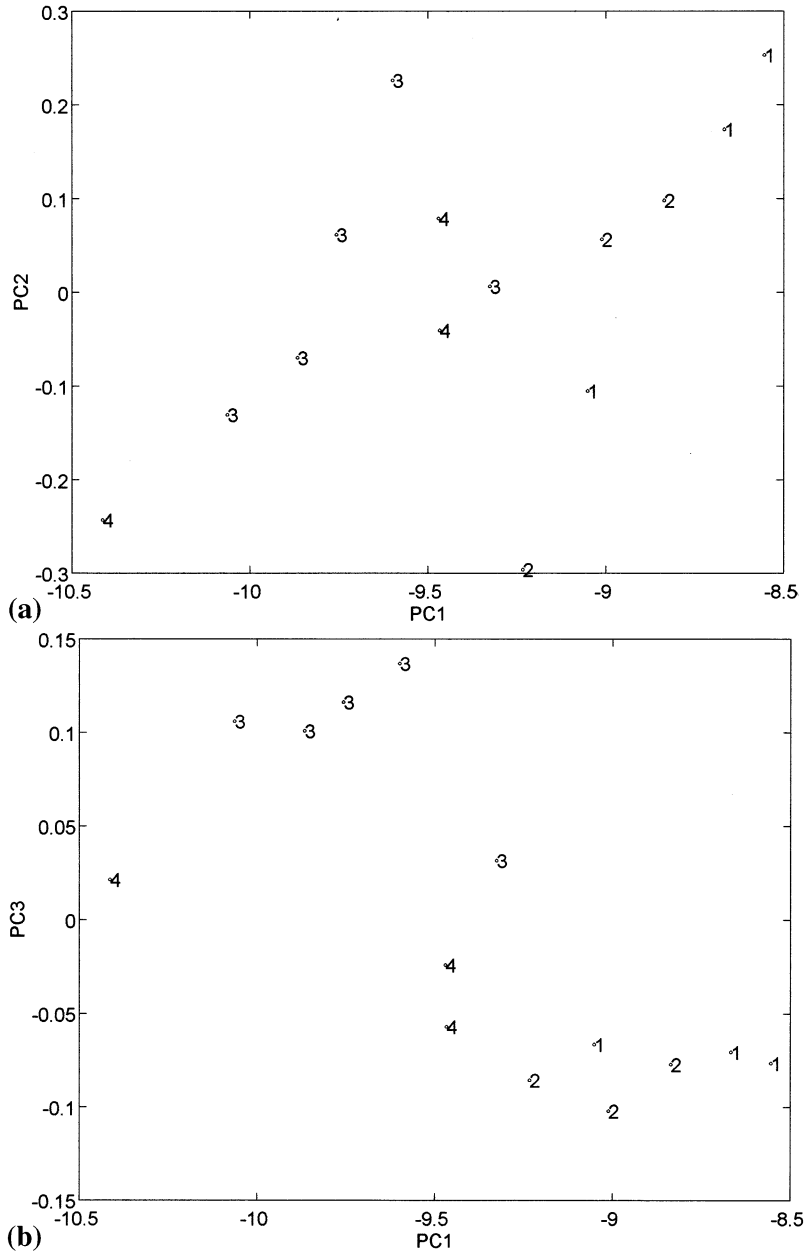

Fig. 9. Score plot of data set 4: (a) PC1 PC2; (b) PC1 PC3; (c) PC1 PC7.

The effects of scatter light, particle size distribution, packing density and instrument noise are always present in NIR spectra. These effects can influence the results of PCA and SP, making the visualisation of possible clusters a little more difficult. To avoid this, a standard normal variant (SNV) transform [11] was applied to the data set. For convenience, the wavelength is expressed by its index in the resulting data matrix.

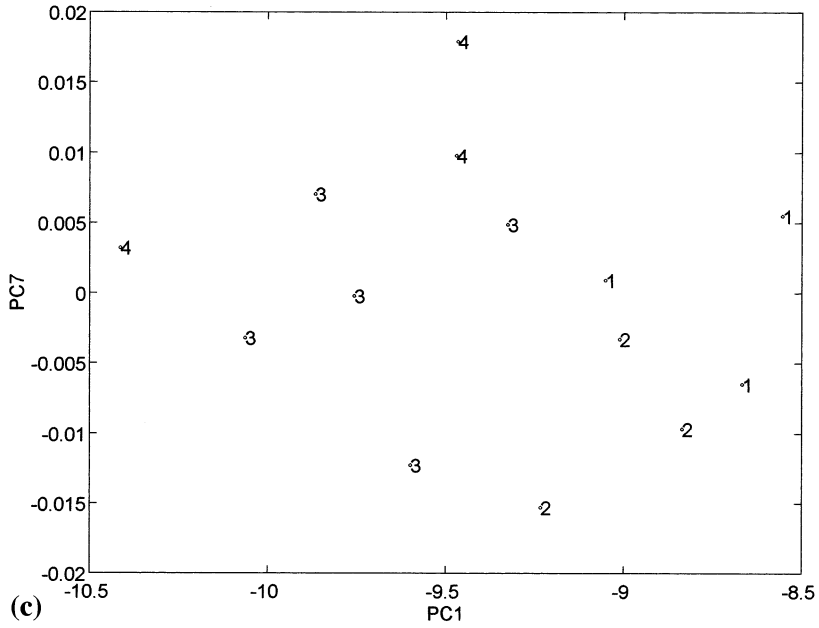

Fig. 9. (*Continued*)

3.1.3. Data set 3: NIR spectra of wheat

100 NIR spectra (1100–2500 nm) with a resolution of 2 nm were recorded using an Infraprover FT from Bran + Luebbe, in the diffuse reflectance mode [12]. The spectra are presented as $\log(1/R')$ absorption values. The data set consists of 41 samples which contain low moisture (12.45–13.99%) and 59 samples which contain high moisture (15.08–17.36%). From here on these two groups are referred to as class 1 and class 2.

3.1.4. Data sets 4: NIR spectra of active drugs and placebos

NIR spectra of drugs were treated. Spectra of 14 tablets containing placebos or drugs in different dosages (3 and 12 mg) were obtained in the 8000–400 cm^{-1} wavelength range with a resolution of 6 cm^{-1} using an Infraprover FT from Bran + Luebbe. The spectra are recorded in the diffuse reflectance mode (10 scans per spectrum). They are presented as $\log(1/R')$ absorption values, where R' is the reflectance of the sample versus that of a white ceramic reflectance. The aim is to see if it is possible to identify the tablets by using simple visualisation.

4. Results and discussion

The SP approaches described in the theory section, as well as the PCA approach, were applied to all data sets.

4.1. Data set 1

Fig. 2 shows the results of the application of the PCA approach to data set 1. The score plot of PC1–PC2 suggests the presence of two clusters, one containing points number 1, 2, 3, 4, 7 and 9 and another composed by points number 5, 6, 11, 12 and 13. Points 8 and 10 are outliers corresponding to currant bread and milk bread. The outlier 8 is due to the high amount of K and Al characteristic of currants, while the outlier 10 is due to the high amount of Ca coming from the milk added to this kind of bread (Table 1). The cluster containing points 1, 2, 3,

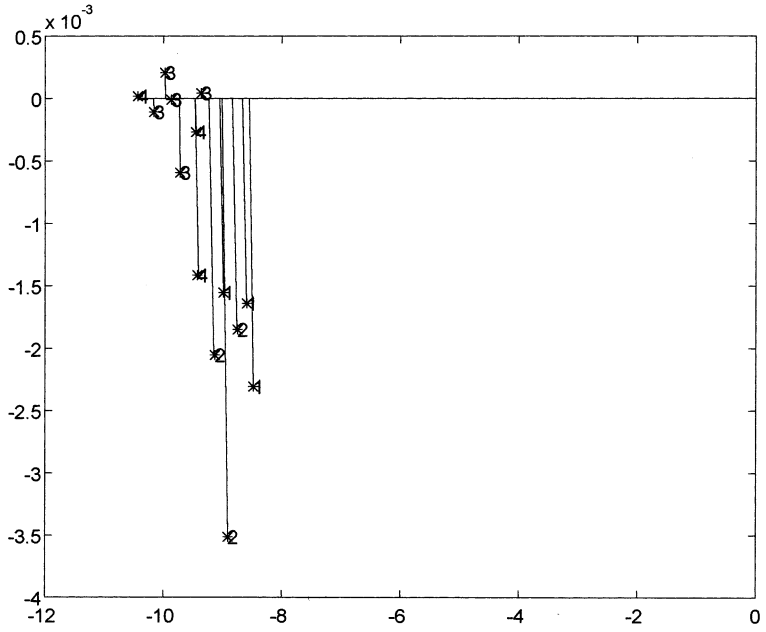

Fig. 10. Star plot of method 3 for data set 4, using the important PCs as input for the SP method.

4, 7 and 9 consists of white breads with a relatively low concentration of mineral content due to the lower extraction rate flour, while the cluster containing points 5, 6, 11, 12 and 13 is formed by brown breads with a high concentration of mineral content due to the high extraction rate flour.

The SP approach was applied to the raw data described in Table 1. The application of the three SP procedures to this data set gave similar clustering results and to avoid presenting a large number of figures, only that figure obtained with the SP method 2 is presented (Fig. 3). It can be seen from this relatively simple case that it is possible to obtain the same information with SP as that obtained by PCA.

4.2. Data set 2

In this data set, it is our intention to show that the method can also be applied to many variables.

The score plot of PC1 against PC2 obtained from this transformed data set (Fig. 4) shows that PC1 is able to separate clearly the two active drugs (groups 1 and 2) from the comparator (group 3), while groups 1 and 2 are reasonably separated by PC2.

The large number of variables and objects presented in this data set lead to many broken-lines so that the star plots would be difficult to interpret. Therefore, in Fig. 5 only the star points are displayed.

The three SP methods were applied to this transformed data set, using 670 wavelengths as the input variables. The results demonstrate that the third method gave better results than the first two methods (Fig. 5). From this figure one can see that not only all three groups are separated but also that the separation between groups 1 and 2 is better than the one obtained with PCA.

When wavelengths are used as the input variables, an important characteristic of the SP methods that should be noticed is that they are data set independent, that is, one can add or remove objects from the data set without the need to recalculate the values of the remaining objects. When applying PCA however, the addition or deletion of a single object obliges the user to: (i) recalculate the PCs, (ii) select once more the PCs

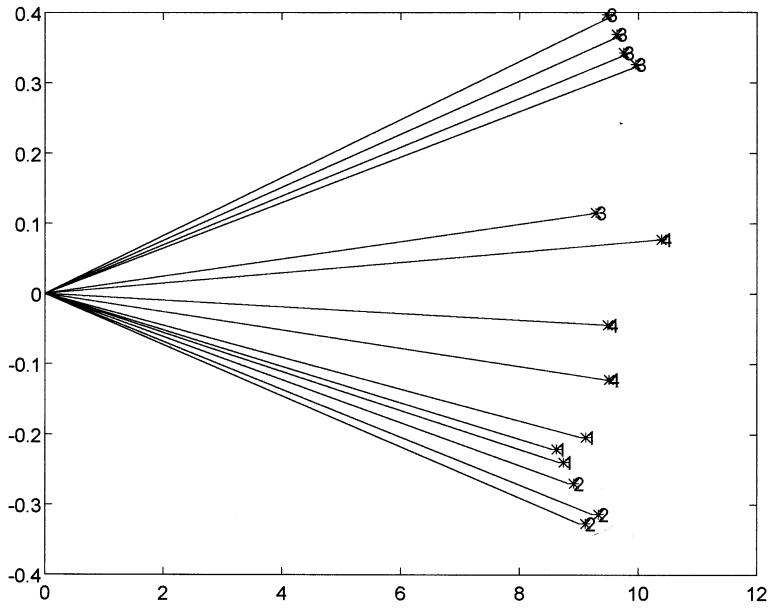

Fig. 11. Star plot of method 1 for data set 4, using the important PCs as input for the SP method.

of interest and (iii) produce the relevant score plot(s), before being able to evaluate the structure of the data.

4.3. Data set 3

This data set contains 701 variables (absorbances at 701 wavelengths) and is shown in Fig. 6. Using all these original variables (wavelengths) as the input, both SP and the PC1–PC2 plot were not able to separate class 2 from class 1. This is probably due to the large number of irrelevant variables which mask the difference in moisture between the two classes. It is known, however, that the original variables in the wavelength range 400–500 are the more relevant for the peak region of water. Therefore, the wavelengths in this region were selected as the input variables, and the SP and PCA methods were once more applied to the reduced data set.

The results of the application of the SP methods show that methods 2 and 3 lead to better results than method 1. Because methods 2 and 3 give equivalent results, only the results of method 3 are presented (Fig. 7). The broken-lines are not displayed in order to avoid too many lines in the figure. One can clearly see that both classes (1 and 2) are completely separated by the SP method.

PCA was applied to the same reduced data (Fig. 8). The analysis of this figure demonstrates that the first 2 PCs explain most of the variance between the two classes. However, PCA can only display the two clustering tendency, but the two classes are still overlapped, so that not all objects can be classified in one of the two clusters.

4.4. Data set 4

This data set is composed of 1267 variables (absorbances at 1267 wavelengths) and PCA was applied to analyse the data. It was verified that 4 PCs, namely PC1, PC2, PC3 and PC7 were needed to extract the necessary information that led to separation of the four groups (Fig. 9a–c). The score plot of PC1–PC2 (Fig. 9a) shows that PC1 separates the clusters of 3 mg dosage (group 1) and its placebo (group 2) from these of 12 mg dosage (group 3) and its placebo (group 4). PC2 mainly explains only irrelevant information such as interference of excipients,

particle size distribution, packing density and instrument noise existing in the NIR spectra. In the plane of PC1–PC3 (Fig. 9b), the cluster of groups 3 and 4 can be separated, while the cluster of groups 1 and 2 are difficult to separate. In the plane of PC1–PC7 (Fig. 9c), the cluster of groups 1 and 2 are classified well whilst the cluster of groups 3 and 4 are overlapped. To have an overview of the complete separation of the four groups, one would need to have a plot in four dimensions which is not possible.

In this example the application of SP to the raw data did not give better results than the ones obtained with PCA.

One would expect that the combination of PCA and SP approaches, that is the use of 4 PCs (PC1, PC2, PC3 and PC7) as the inputs for the SP methods, would improve the separation of the groups in only one plane.

Combined with PCA, the two SP methods described in the literature, i.e. methods 2 and 3, were first applied. The best results were obtained with PCA + method 3. However, the separation of the four groups was still not completely achieved (Fig. 10).

When applying the combination of PCA with the variant proposed in this manuscript (method 1), complete separation is obtained (Fig. 11). The advantage of using SP combined with PCA here is that the information can be summarised in one single plot.

5. Conclusions

The SP is a non-linear multivariate mapping method. This method was applied to an atomic absorption spectrometry (AAS) data set, as well as to three NIR spectral data sets. The results of its application has shown that the visualisation of objects with more than two variables in a plane is possible for all examples.

Combined with PCA, more than two PCs can be displayed in one plane which can be very useful in cases where several PCs are necessary to allow the interpretation of the data structure.

From the examples treated in this paper it is not possible to suggest one of the three methods as being the best. Instead of selecting one among them, it is more prudent to perform all three methods and verify which one is best for the problem studied. For small data sets the difference in time for the computations may not be significant but for large data sets SP methods are significantly faster than PCA. Moreover, SP is data set independent whilst PCA depends on the data set.

Summarising, SP methods can be applied as an alternative method to display multivariate chemometric data instead of PCA.

Acknowledgements

The authors thank the Nationaal Fonds van Wetenschappelijk Onderzoek, the DWTC and the Standards, Measurement and Testing program of the EU for financial assistance.

References

[1] D.L Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: a Textbook, Elsevier, Amsterdam, 1988.
[2] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, New York, 1983.
[3] M.J. Adams, Chemometrics in Analytical Spectroscopy, The Royal Society of Chemistry, Cambridge, 1995.
[4] D.H. Coomans, O.Y. de Vel, Pattern Analysis and Classification, the Handbook of Environmental Chemistry, Springer, Berlin, 1994.
[5] S. Wold, Principle component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.
[6] W. Wu, D.L. Massart, S. de Jong, The kernel PCA algorithms for wide data. Part I: theory and algorithms, Chemom. Intell. Lab. Syst. 36 (1997) 165–172.
[7] S.Y. Shong, F.L. Zhang, Multivariate Statistics: a Textbook. Department of Mathematics, Shenyang College of Pharmacy, Shenyang, 1991.
[8] C. van der Vlies, K.J. Kaffka, W. Plugge, Qualifying pharmaceutical substances by fingerprinting with NIR spectroscopy and PQS, Pharm. Technol. Eur. April (1995) 43–48.
[9] K.J. Kaffka, L.S. Gyarmati, Qualitative (comparative) analysis by near infrared spectroscopy, in: R. Biston, N. Bartiaux-Thill (Eds.), Proc. 3rd Int. Conf. on NIR

Spectroscopy, vol. 1, Agriculture Research Centre, Gembloux, Belgium, 1990, pp 135–139.

[10] Q. Yang, W. Penninckx, R. Van Cauwenbergh, J. Smeyers-Verbeke, H. Deelstra, D.L. Massart, Visualisation of the mineral distribution pattern in breads by principal component analysis, Analysis 21 (1993) 379–382.

[11] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and detrending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777.

[12] J. Kalivas, Two data sets of near infrared spectra, Chemom. Intell. Lab. Syst. 37 (1997) 255–259.