A PRINCIPAL COMPONENT ANALYSIS OF PATIENTS, DISEASE AND TREATMENT VARIABLES: A NEW PROGNOSTIC TOOL IN BREAST CANCER AFTER MASTECTOMY.

Buciński Adam[1], Załuski Jerzy[3], Krysiński Jerzy[2], Kaliszan Roman[2]

[1]The Institute of Animal Reproduction and Food Research of the Polish Academy of Sciences, Division of Food Science, Tuwima 10, 10-747 Olsztyn, Poland
[2]Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gen. Hallera 107, 80-416 Gdańsk, Poland
[3]Greatpoland Cancer Centre, Garbary Street 15, 61-866 Poznań, Poland

Received 28 March 2000; revised version received 10 December 2000; accepted 23 February 2001

Key words: principal component analysis (PCA), statistical analysis prognostic tool, breast cancer

ABSTRACT

Purpose: To demonstrate unique information potential of a powerful multivariate data processing method, principal component analysis (PCA), in detecting complex interrelationships between diverse patient, disease and treatment variables and in prognostication of therapy's outcome and response of patients after mastectomy.
Patients and Methods: One hundred-forty-two patients with breast cancer were retrospectively evaluated. The patients were selected from a group of 201 patients who had been treated and observed in the same oncology ward. The selection was based on availability of complete set of information describing each patient. The set consisted of 60 specific data. A matrix of 142 x 60 data points was subjected to PCA using a professional, statistical software (commercially available) and a personal computer.

Results: Two principal components, PC1 and PC2, were extracted. They accounted for 26% of total data variance. Projections of 60 variables and 142 patients were made on a plane determined by PC1 and PC2. A clear clustering of the variables and of the patients was observed. It was discussed in terms of similarity (dissimilarity) of the variables and the patients, respectively. A strikingly clear separation was demonstrated to exist between the group of patients living over 7 years after mastectomy and the group of deceased patients.

Conclusion: PCA offers a new promising alternative of statistical analysis of multivariable data on cancer patients. Using the PCA, potentially useful information on both the factors affecting treatment outcome and general prognosis, may be extracted from large data sets.

INTRODUCTION

Breast cancer has become the most common malignant disease causing death of women in the European Community. Its increasing incidence in all Western countries has been observed [1-3]. Early detection and optimal treatment are decisive for overall duration of patients' survival. A therapeutic strategy should naturally depend on the prediction of outcome and response to therapy. Such a reliable prediction is extremely difficult because of the lack of single prognostic parameters or identified combinations of thereof [1,4].

Several factors have been recommended to help to prognosticate both the overall patients' survival and recurrency-free interval. However, these factors usually seem controversial when considered separately. The best known example is the patient's age. Recent reports [5,6] demonstrated (in statistical terms) that the age of under 35 is an independent prognostic factor of unfavourable outcome. On the other hand, several studies have proved that age of over 50 years (post-menopausal women) can provide a better prognosis [7]. It would be doubtful, however, to linearly relate survival to age or to prognosticate outcome and response to therapy of an individual patient on the basis of her age alone.

In various types of cancer various prognostic indexes have been proposed. They are derived by the multiple regression analysis of the number of patients', disease and treatment parameters [8,9]. For instance, significant prognostic factors indicating survival in the case of small-cell lung cancer, when used simultaneously in a seven-variable regression equation are: the level of bicarbonate, alanine transaminase, alkaline phosphatase, sodium, potassium, urea and uric acid, together with erithrocyte sedimendation rate and the patient's age [9].

The fundamental problem of the multiple regression analysis is that parameters (independent variables) considered simultaneously cannot be mutually related, i.e., they should be orthogonal [10,11]. It is difficult, if at all possible, to find a representative (and sufficiently large for statistical purposes) set of readily available and informative patients', disease and treatment parameters which would be mutually orthogonal. Therefore, prognostic indexes derived by means of multiple regression analysis are of a rather limited reliability.

Problems of intercorrelation and multicolinearity among independent variables have often been encountered in science. They can be eliminated by using the so-called factorial methods of data analysis, in particular principal component analysis (PCA). These methods have recently been popularized especially in chemistry [12,13]. There is no need to discuss here the higher algebra which is the basis of the approach. Advanced software for personal computers, which are commonly available, can actually be used in a „black box" manner.

The idea of a PCA is to reduce the dimensionality of an original multivariable data set by finding linear combinations of those parameters (variables) that explain most of the variability. Most of the systematic information, initially dispersed over a large matrix of input variables (often intercorrelated), is extracted and condensed in a few calculated abstract variables by using the PCA. Normally, two factors (principal components a PCs) are used to determine an abstract variable plane. Projections of data points ascribed to individual objects (patients) and individual input variables on the plane reflect, in a comprehensive graphical manner, similarites and disimilarities among them. In this way, a basic part of systematic information on the objects and the variables can be exploited by our mind, which naturally visualises relationships in up to three dimensions [14].

It is assumed in this study that the PCA may appear useful in the search for reliable prognostic factors in various kinds of cancer. Using the PCA all types of information on patients, disease, treatment, etc., can be exploited by means of a single analysis with variables ranging from sociological to genetic. This approach will be presented using information available for 142 breast cancer patients after mastectomy, who had been treated and observed in the Chemotherapy Ward, of the Wielkopolskie Oncology Center in Poznań. Mastectomy was done in 1990/1991 in the Surgery Ward of the same institution. The observation was caried on till 1997.

MATERIALS AND METHODS

Data on 201 patients with breast cancer were retrospectively collected and analysed. Considering individual data one has to realize that drugs, treatment strategy and diagnostic procedures were those usually applied in 1990-1991. A set of 60 variables was identified. Those variables were common for all the patients subjected to PCA. For reasons of statistical significance, a given variable was included in the analysis if it recurred at least 6 times. Hence, for instance, the age below 30 years was not an independent variable, and such patients were excluded from PCA, because they represented only few cases in our group of 201. As result, the final matrix of data subjected to PCA was 142 patients times 60 variables. The variables were defined and written in 0-1 manner. Exemplary data for patients Nos. 40, 52 and 99 (typically long survival) and Nos. 43 and 66 (unexpected short survival) are given in Table 1.

The principal component analysis of the 142 x 60 data matrix was performed by means of *Statistics* software (StatSoft Inc., Tulsa, OK, USA), run on a personal computer. It was found that the two first principal components, PC1 and PC2, accounted cumulatively for 26% of the variance of data within all 60 original variables. The programme calculated contributions to PC1 and PC2 of each original variable, numbered and listed in Table 1. Using those quantities (principal component „loadings") the variables are placed on the plane by means of coordinates PC1 and PC2 (Figure 1).

Table 1. Variables considered in principal component analysis (PCA) and its values for five selected patients.

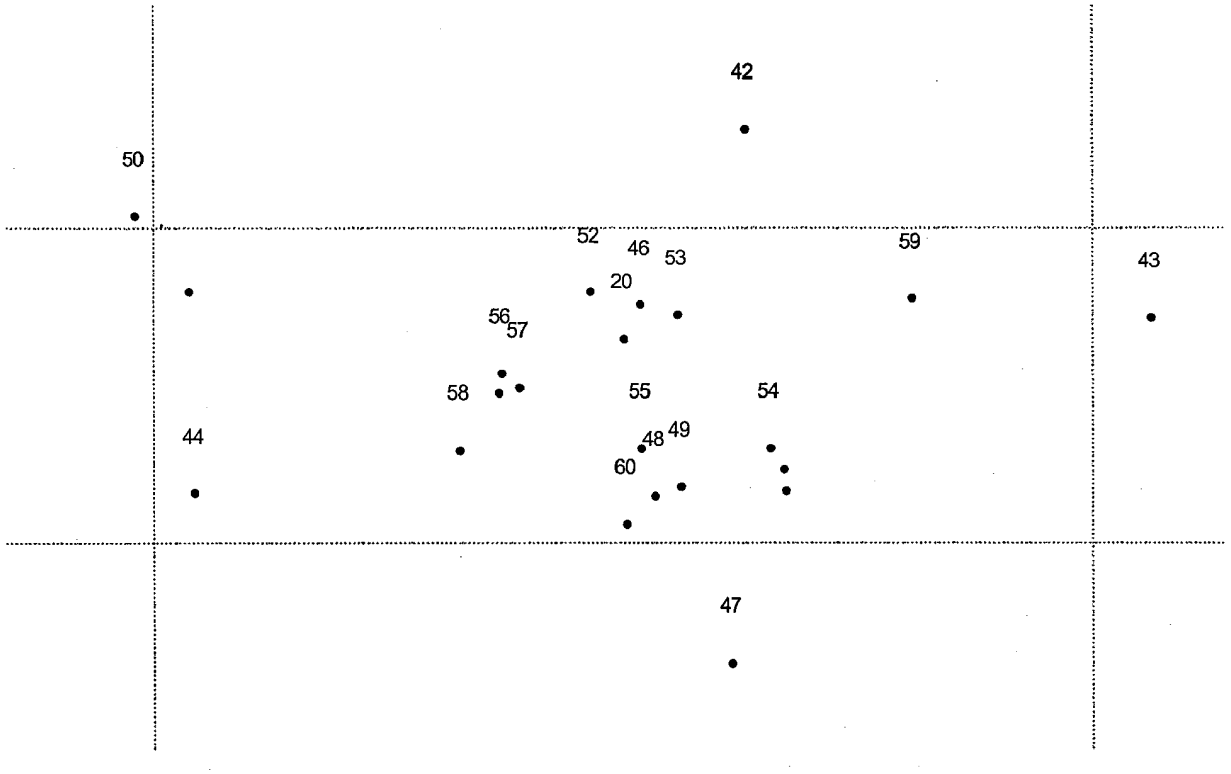
Variable No.	Variable Name	Variable Value for Patient				
		No. 40	No. 52	No. 99	No. 43	No. 66
1	Age: 31-50 years	0	0	0	1	0
2	Age: 51-60 years	0	1	1	0	1
3	Age: >60 years	1	0	0	0	0
4	Menopause: before	0	0	0	1	0
5	Menopause: during	0	0	0	0	0

Table 1. Continued.

6	Menopause: after	1	1	1	0	1
7	Hormonal activity: 11-20 years	0	0	0	0	0
8	Hormonal activity: 21-30 years	0	1	0	1	1
9	Hormonal activity: 31-40 years	1	0	1	0	0
10	Number of births: 0	0	0	0	0	0
11	Number of births: 1	0	0	1	0	0
12	Number of births: 2	0	1	0	1	1
13	Number of births: 3	0	0	0	0	0
14	Number of births: >3	1	0	0	0	0
15	No record of breast cancer in I and II generation	1	1	1	1	1
16	Tumor size: <40 mm	0	1	0	0	1
17	Tumor size: >40 mm	1	0	1	1	0
18	Positive lymph nodes: none	0	1	0	0	0
19	Positive lymph nodes: 1-3	0	0	0	1	0
20	Positive lymph nodes: 4-8	1	0	1	0	1
21	Positive lymph nodes: >8	0	0	0	0	0
22	No infiltration of node capsule	0	1	1	0	0
23	No arrest in microvessels	1	1	1	1	1
24	Malignancy: Bloom's degree I	0	0	0	0	0
25	Malignancy: Bloom's degree II	1	0	0	0	0
26	Malignancy: Bloom's degree III	0	1	1	1	1
27	Surgery: Halsted's mastectomy	0	1	0	0	0
28	Surgery: Patey's mastectomy	1	0	1	0	1
29	Surgery: mastectomy	0	0	0	1	0
30	Adjuvant therapy: radiotherapy	1	0	1	0	1
31	Adjuvant therapy: chemotherapy	1	0	1	1	0
32	Adjuvant therapy: hormonotherapy	1	1	1	0	1
33	Adjuvant radiotherapy: none	0	1	0	1	0
34	Adjuvant radiotherapy: scar and lymph nodes	1	0	1	0	1
35	Adjuvant radiotherapy: peripheral lymph nodes	0	0	0	0	0
36	No neoadjuvant chemotherapy	1	1	0	1	0
37	Type of adjuvant chemotherapy: none	0	1	0	0	1
38	Type of adjuvant chemotherapy: CMF	1	0	1	1	0
39	Type of adjuvant hormonotherapy: none	0	0	0	1	0
40	Type of adjuvant hormonotherapy: tamoxifen	1	1	1	0	1
41	First line treatment: surgery	0	0	0	0	0
42	First line treatment: hormonotherapy	1	1	1	0	0
43	First line treatment: chemotherapy	0	1	1	0	0
44	First line treatment: radiotherapy	0	1	0	0	0
45	Type of first line chemotherapy: None	1	0	0	0	0
46	Type of first line chemotherapy: CMF	0	1	1	0	0
47	Type of first line chemotherapy: anthracyclines	0	0	0	0	0

Table 1. Continued.

48	Type of first line hormonotherapy: none	0	0	0	0	0
49	Type of first line hormonotherapy: tamoxifen	0	1	0	0	0
50	Type of first line hormonotherapy: aminogluthetimide	1	0	1	0	0
51	Response to first line treatment: no response	0	0	0	0	0
52	Response to first line treatment: 4-8 months	1	1	0	0	0
53	Second line treatment: hormonotherapy	1	1	0	0	0
54	Second line treatment: chemotherapy	0	0	0	0	0
55	Type of second line hormonotherapy: none	0	0	0	0	0
56	Type of second line hormonotherapy: aminogluthemide	0	1	0	0	0
57	Type of second line chemotherapy: none	1	1	0	0	0
58	Type of second line chemotherapy: anthracyclines	0	0	0	0	0
59	Response to second line treatment: no response	1	0	0	0	0
60	Third line treatment: chemotherapy	1	0	0	0	0



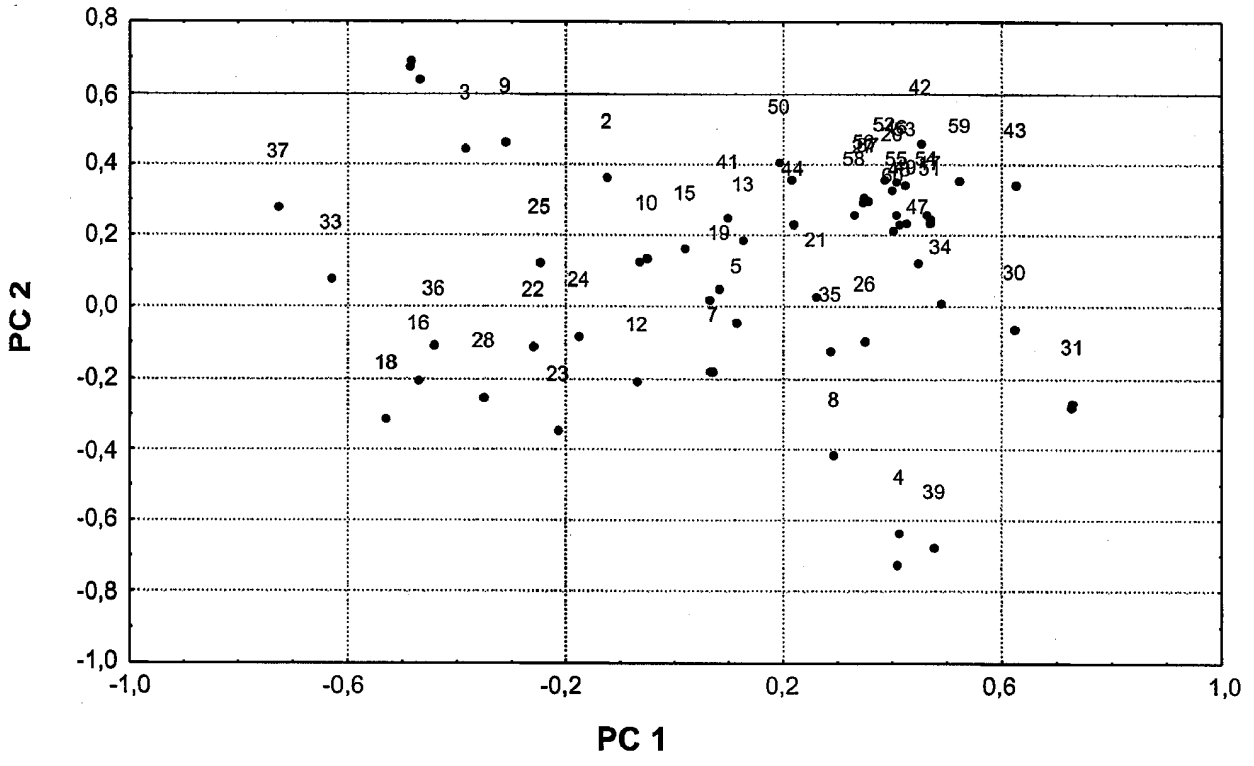


Figure 1. Projection of 60 variables listed in Table 1 on the plane with coordinates of the first two principal components, PC1 and PC2, distinguished in the principal component analysis (PCA) of data collected from 142 patients. The right top section of the chart is enlarged at the bottom of the Figure.

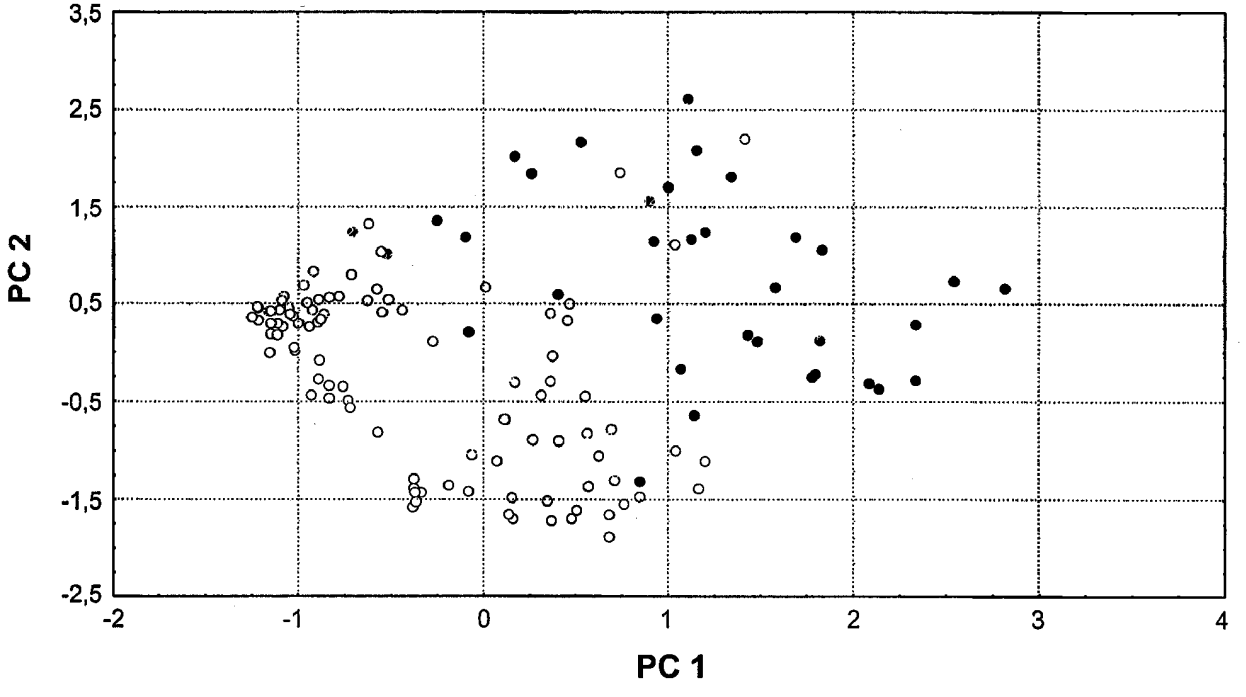


Figure 2. Projection of 142 patients (described by 60 variables listed in Table 1) on the plane with coordinates of the first two principal components, PC1 and PC2, distinguished in the principal component analysis (PCA) of the 142 x 60 data matrix. Blank circles denote patients living longer than 7 years after surgery; shaded circles denote deceased patients.

RESULTS AND DISCUSSION

The following most evident regularities can easily be sean in Figure 1. On the left hand side the chart includes variables with a better prognosis (especially in the top left section). On the right hand side (especially in the chart's bottom section) the variables with poorer prognosis are grouped. For example, majority of dots on the left represent variables 37 and 33, indicating that neither chemotherapy nor radiotherapy was applied. On the left, there is also variable 18 informing about the absence of positive lymph nodes, and variable 16 indicating tumor size smaller than 40 mm. The top left section includes variables 6, 3 and 9 informing, respectively, about the menopause (its termination), age (over 60 years) and hormonal activity duration (31-40 years).

On the other hand, majority of dots on the left represent variables 31 and 38, indicating the need for adiuvant CMF chemotherapy. Next to them are variables 30 and 43, indicating adiuvant radiotherapy and chemotherapy as a first rate treatments, respectively. In the bottom right section, there is variable 4 informing that the patient was before menopause. As expected, in the right hand side of Figure 1 there are located variables: No. 17 indicating tumor size > 40 mm, No. 26 indicating the III degree of Bloom malignancy; No. 20 indicating 4-8 positive lymph nodes, etc.

The above regularities are not controversial and may be considered as a proof of the PCA information potential regarding the prognostic value of individual patient's disease and treatment variables. Some observations in Figure 1 may be disputable. For example, variable 27 is more to the right than variables 29 and 28. This would mean that Halsted mastectomy provides worse prognosis than simple mastectomy and the Patey mastectomy.

Using the PCA one obtains quantitative information on the similarity of information provided by given variables: the smaller is the distance between two variable points the more similar are the effects of the two variables on the system. Hence, charts developed for a series of well-known variables can be used to calibrate other unknown variables. Our study is basically methodological, but it demonstrates simultaneously that this approach can be applied to a number of now available factors of sociological, morphologic, molecular genetic and therapeutic nature.

The principal component analysis also extracts systematic information on objects described by sets numerous variables. The objects here are patients of the total number of 142. Distribution of patients on a plane with coordinates PC1 and PC2 can be obtained in the same way as the distribution of variables. Such a projection of points assigned to individual patients is presented in Figure 2. The chart shows clear cut clustering of patients surviving (blank circles) and not surviving (shaded circles) the 7-year period after mastectomy.

It also shows the exellent prognostic potency of the PCA performed on the factors listed in Table 1. Only two patients (Nos. 43 and 66) have been wrongly classified to the surviving group, and three patients (Nos. 40, 52 and 99) to the deceased group. The parameters of those five outliers are given in Table 1. It is difficult to draw straightforward conclusions from such a limited number of cases. It can be hypothesized, however, that the analysis of the outliers'specific features, especially that of the unexpectedly surviving patients, may help to identify a combination of factors providing a good prognosis. Moreover, the PCA offers a possibility of testing a practically unlimited number of either mutually related or unrelated factors. The presented report is methodological in nature, and is based on a limited set of readily available data collected retrospectively. It would be advisable to use the PCA to process other large sets of data, especially series of rationally designed and determined data.

REFERENCES

1. von Kleist S. Prognostic factors in breast cancer: theoretical and clinical aspects (review). Anticancer Research 1996; 16: 3907 - 12.

2. O'Higgins N. Aspects of breast cancers. Chirurgie 1992; 118: 342 - 27.

3. McGuire WL. Adjuvant therapy of node-negative breast cancer. N Engl J Med 1989; 320: 525 - 7.

4. Dhingra K, Hortobagyi GN. Critical evaluation of prognostic factors. Semin Oncol 1996; 23: 436 - 45.

5. Bonnier P, Romain S, Charpin C, et al. Age as a prognostic factor in breast cancer: relationship to pathologic and biological features. Int J Cancer 1995; 62: 138 - 44.

6. Chung M, Chang HR, Bland KI, Wanebo HJ. Younger women with breast carcinoma have a poorer prognosis than older women. Cancer 1996; 77: 97 - 103.

7. Aaltomaa S, Lipponen P, Eskelinen M, et al. Prognostic factors after 5 years follow-up in female breast cancer. Oncology 1992; 49: 93 - 8.

8. Beyer J, Kramer A, Mandanas R, et al. High-dose chemotherapy as salvage treatment in germ cell tumors: a multivariate analysis of prognostic variables. J Clin Onco 1996; 14: 2638 - 45.

8. Maestu I, Pastor M, Gómez-Codina J, et al. Pretreatment prognostic factors for survival in small-cell lung cancer: a new prognostic index and validation of three known prognostic indicies on 34 patients. Annals of Oncology 1997; 8: 547 - 53.

10. Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. Academic Press, London 1979.

11. Jolliffe IT. Principal Component Analysis. Springer, New York 1986.

12. Massart DL, Vandeginste BGM, Deming SN, et al. Chemometrics: a textbook. Elsevier, Amsterdam 1988.

13. Brereton RG, editor: Multivariate Pattern Recognition in Chemometrics. Elsevier, Amsterdam 1992.

14. Kaliszan R. Structure and Retention in Chromatography. A Chemometric Approach. Harwood Academic, Amsterdam 1997.