Journal of Chromatography, 556 (1991) 169–180
Elsevier Science Publishers B.V., Amsterdam

CHROM. 23 306

Factor analysis and experimental design in high-performance liquid chromatography

XI[a]. Factor analysis maps and chromatographic information

MICHEL RIGHEZZA
Laboratoire de Chimiométrie, Université d'Orléans, B.P. 6759, 45067 Orléans Cedex 2 (France)
and
JACQUES R. CHRÉTIEN*
Laboratoire de Chimiométrie, Université d'Orléans, B.P. 6759, 45067 Orléans Cedex 2 (France) and Institut de Topologie et de Dynamique des Systèmes, Associé au CNRS UA 34, 1 Rue Guy de la Brosse, 75005 Paris (France)

ABSTRACT

A progressive approach to the exploitation of chromatographic data, based on factor analysis, is presented. This approach is applied to the retention data k' of a large series of compounds in high-performance liquid chromatography. The chromatographic information, *i.e.*, affinity and selectivity, is extracted with help of principal component analysis (PCA) and correspondence factor analysis (CFA). The factor analysis gives rise to three factorial maps to present the chromatographic information: PCA affinity map, CFA trend analysis map and CFA distance analysis map. Examples of extraction of chromatographic information are given for the simultaneous exploitation of these maps. Possibilities and limitations of this approach are discussed.

INTRODUCTION

It is essential for the chromatographer to have a representation of the main information nested in huge series of data, and factor analysis is a good tool to extract and to represent this information [1,2]. The chromatographic properties, affinity, polarity and selectivity, can be revealed by factor analysis maps.

Up to now factor analysis has most often been used as a clustering technique [3]. Proximities between representative points, solutes or chromatographic systems, are considered to suggest similarities of the basic chromatographic phenomenon. In fact, the extracted factors most often remain abstract ones. In some instances it has

[a] For Part X, see ref. 12.

0021-9673/91/$03.50

been possible to give a real physicochemical meaning to these factors. For example, it has been shown that the first factorial axis can be assimilated to the contribution of the partition coefficient [4] or to the electronic factors [5].

Chromatographers can use factor analysis for the rapid and efficient selection of packings and/or eluents. Any data processing techniques or their representations generate a loss of minor information. These limitations need to be circumvented with the help of complementary factor analysis methods. The aim of this paper is to consider the complementarity evolved from the successive use of principal component analysis (PCA) and correspondence factor analysis (CFA). It will be shown how the chromatographer can retrieve required information in a series of factor analysis maps. Emphasis will be placed on affinity and/or selectivity with three maps: (1) PCA "affinity analysis" map, (2) CFA "trends analysis" map and (3) CFA "distances analysis" map.

DATA

Corresponding data have been published previously in a series of papers devoted to "factor analysis and experimental design in high-performance liquid chromatography (HPLC)" [4–12]. The original data set corresponds to 63 compounds studied on 43 systems. An overall view of these data has been presented in a previous paper in this series [12]. Therefore, the description of the solutes, a series of chalcones, $XC_6H_4CH=CHCOC_6H_4Y$, has been omitted in order to simplify the paper. The topic is focused on factor analysis maps, the art of the exploitation of chemometric techniques rather than on the physico-chemical exploitation of the data. The characteristics of the 43 chromatographic systems are presented Fig. 1: the packings, the eluents and the corresponding identifiers of the systems are given. References to the original papers are indicated at the top.

From the original incomplete data matrix of capacity factors (k'), two subsets have been considered. The first matrix is a set of 1548 k' relative to 36 compounds

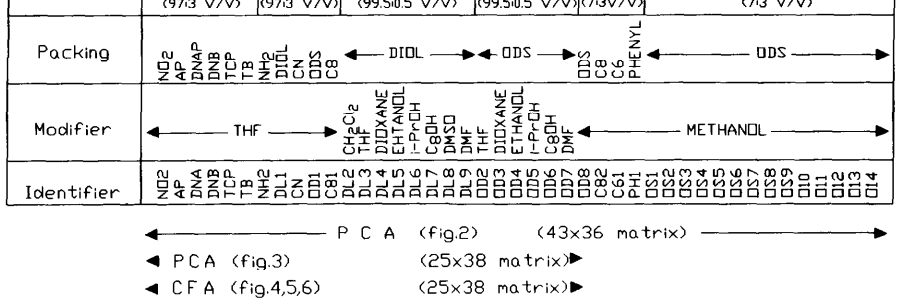

Fig. 1. Characteristics of the 43 studied chromatographic systems. These systems are the variables of an incomplete 43 × 38 retention data matrix of a series of 38 compounds. Factor analysis (PCA and CFA) was realized on the complete matrix and the reduced set of normal-phase data. MeOH = Methanol; i-ProH = isopropanol; C_8OH = Octanol-1.

eluted from 43 normal (NP) and reversed-phase (RP) chromatographic systems. The second matrix is a set of 950 k' relative to 38 compounds eluted from 25 normal-phase chromatographic systems.

DATA PROCESSING

Data were analysed with the use of PCA and CFA. Both approaches have been explained in previous papers [2,4,8].

The CFA method is particularly useful for a simultaneous comparison of the different characteristics of chromatographic systems and solutes. The usual maps obtained with CFA permit trends analysis of the chromatographic systems and solutes by means of their relative proximities. Hence, relative polarity, affinity and selectivity are accessible.

CFA can be extended to increase the analysis of the selectivity. The original maps of CFA are transformed by a translation of chromatographic systems (i) along the main axes (k). The translation factors are given by the eigenvalues (e_i) extracted from the data matrix. For a chromatographic system (i), the translation factor along the axis (k) is equal to $1/\sqrt{e_k}$. Because the chromatographic systems are translated far from the origin, they are drawn on the limit of the graph. The directions given by the chromatographic systems are then more useful for obtaining the relative selectivity of systems. This relative selectivity is related to the distance between the perpendicular projections of two solutes onto the directions defined by the systems. Hence, for a pair of solutes, it is possible to compare the selectivities of chromatographic systems by distance analysis.

RESULTS AND DISCUSSION

The factor analysis will be presented progressively by using PCA of the 43 × 36 complete matrix and of the 25 × 38 submatrix (Fig. 1) and CFA of the 25 × 38 submatrix.

The PCA of the matrix of normal and reversed-phases is presented Fig. 2. The projection of the 30 compounds is given on the plane defined by the first and second best factorial axes of inertia. These axes correspond to 60% and 34%, respectively, of the information content. Fig. 2 represents simultaneously the correlation circle of the chromatographic systems. Two main groups of systems defined two perpendicular directions which show clearly their independence. These groups correspond to the normal- and reversed-phase systems.

For the two chromatographic modes, a strong correlation of all the chromatographic systems was observed with the first two extracted factors. This separation into two groups can be related to the CFA of the same data matrix given elsewhere [12] (Fig. 4b). This CFA map has shown two clouds of the projected chromatographic systems with a large scattering of representative points for the NP systems due to the large and specific interactions involved. The representative points of the RP mode were closed, certainly owing to the simpler partition mechanism and the lack of various solvents. This type of information is not obvious in the above PCA map (Fig. 2). The latter gives some trends of the affinity of compounds for the system belonging to the two chromatographic modes. Each group of systems gives an average direction

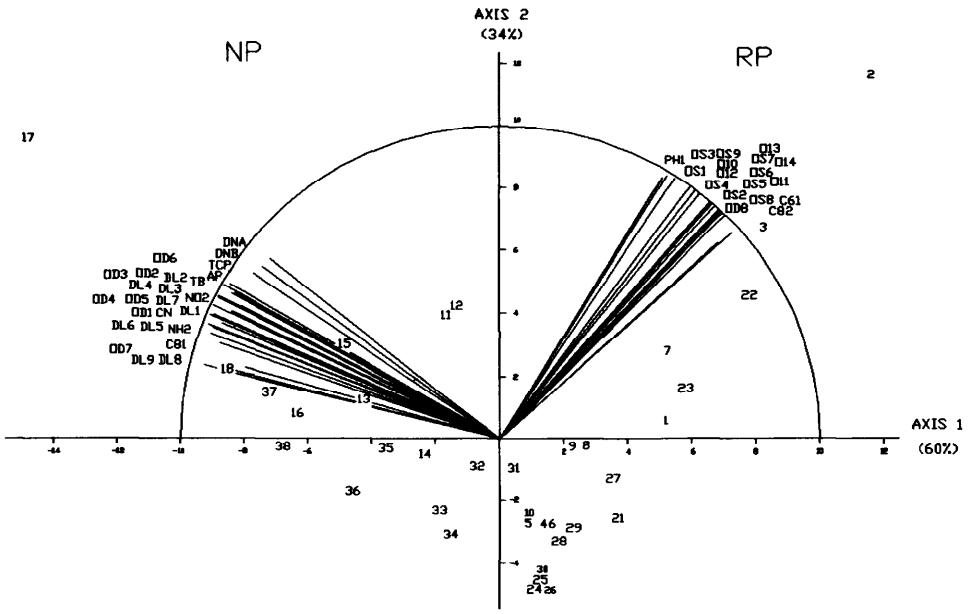

Fig. 2. Principal component analysis of a matrix of 1548 capacity factors in normal and reversed phases. The compounds are projected in the plane defined by the first and second best axes of inertia which represent 94% of the information content. On the correlation circle the independence of the two chromatographic modes appears clearly.

which presents a better dispersion for sets of compounds. For example, solutes 2, 3 and 7 have a better affinity for RP than for NP systems. In the same way, compounds 17, 18, 37 and 16 have a better affinity for NP systems. Compounds 5, 46, 29 and 28, which are located in the centre of the graph, present a similar affinity for RP and NP systems. The capacity factors increase according to the main directions defined by the NP and RP systems.

A submatrix of the main matrix is considered. It includes 950 k' data corresponding to the behaviour of the model series of chalcones on the normal phases only. This submatrix is studied by PCA. A projection of the compounds in the first factorial plane, defined by the first and second axes of inertia, which represent 89% and 6% of the information content, respectively, is shown in Fig. 3a. A simultaneous representation of the correlation circle confirms a strong correlation of each NP system principaly with axis 1 and secondarily with axis 2 and a similarity between all these systems. The average direction defined by the position of the systems on the correlation circle is superposed with axis 1. This graph gives a better in-depth analysis of similarities and differences of the systems. For example, the diol systems DL8 and DL9, with their strong modifiers (DMSO and DMF), are similar and relatively separated from the others. The group with ODS, C_8, diol and TB systems (OD1, $C_8$1, DL2–DL4, DL7, etc.) are not correlated with axis 2 and present the same affinity for the tested compounds. The non-commercially available phases TCP, DNAP (DNA) and DNB have the same specific affinity. The three groups reflect the three main trends of all the set of NP systems.

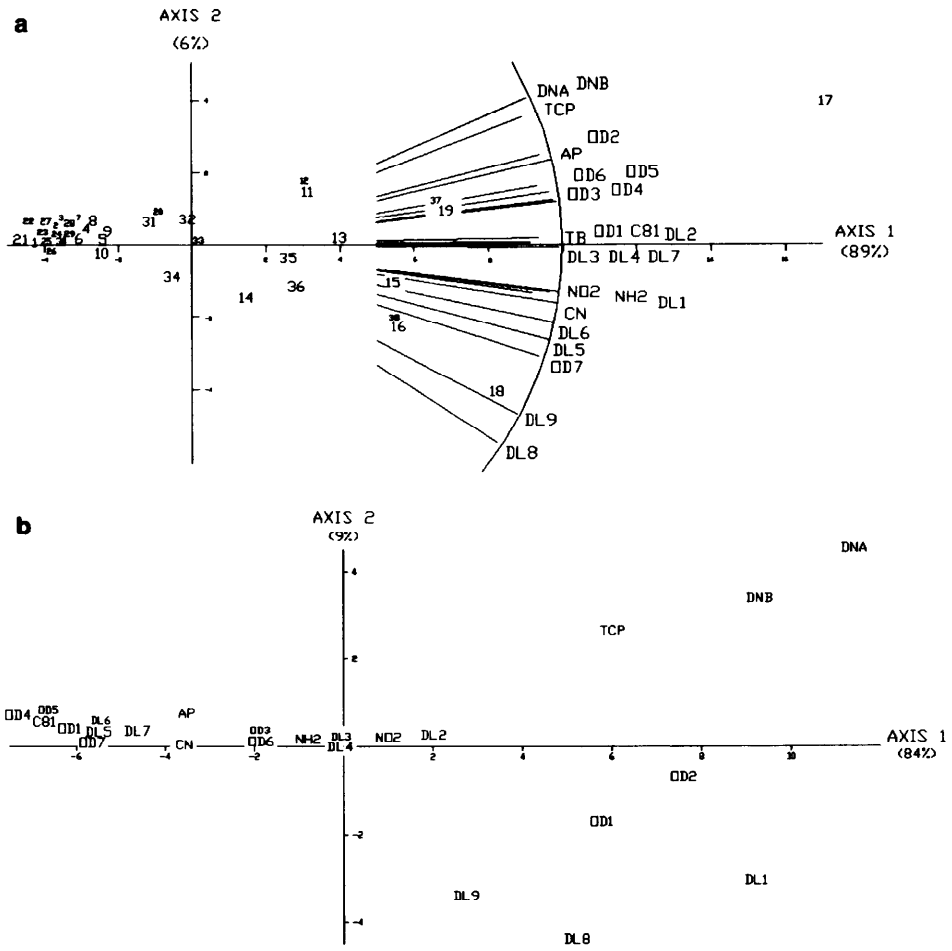

Fig. 3. (a) Principal component analysis of the submatrix of the 950 normal-phase capacity factor data. Projection of the compounds in the first plane (95% of the information content). The dispersion of the chromatographic systems on the correlation circle is related to their correlation with the extracted factors 1 and 2 and to their specific interactions with the compounds. (b) Principal component analysis of the normal-phase submatrix. Projection of the chromatographic systems in the first factorial plane (93% of the information content). This map gives the relative polarity of the chromatographic systems.

The above average direction defined by the chromatographic systems is related to the average value of the capacity factors; 89% of the information content of axis 1 reflects the weight of the k' when the main factor of inertia is extracted. In other words, axis 1 reflects k' variations. The dispersion of the previous three groups is due to the contribution of axis 2, which is the second extracted factorial axis. The differences in the chromatographic affinities of the compounds for all these NP systems can be linked to their contribution to the information content of the second axis.

The cloud of compounds located on the left-hand side of Fig. 3a contains those which have a low affinity for NP systems. These solutes have a better affinity for RP

systems, as was shown previously on the right-hand side of Fig. 2. The compounds that have the greatest k' values for NP systems are located on the right-hand side of Fig. 3a. On axis 2, the dispersion of solutes can be related to the dispersion of NP-systems. The trend analysis of the system directions and the solute projections can be linked qualitatively. For example, compound 17 has a great affinity for DNAP and DNB. Compound 18 has a great affinity for DL8.

The PCA of the NP matrix can also be used to analyse the behaviour of the chromatographic systems in the solute space. Fig. 3b gives the projection of these systems on the first factorial plane defined by axes 1 and 2, which take 84% and 9%, respectively, of the information content. Axes 1 and 2 also represent the variation of chromatographic polarity. Fig. 3b gives a more in-depth analysis of the dispersion of the systems than can be done with the correlation circle only (Fig. 3a). For any direction issued from the origin of the axes, the systems are classified according to their average chromatographic polarity. The systems which are located on the right-hand side of the graph contribute to the dispersion represented by the second factorial axis.

The complete PCA study must take into account projections of compounds (Fig. 3a) and systems (Fig. 3b). For example, compound 18 has a good affinity for DL9 and DL8 (Fig. 3a) but Fig. 3b shows that in the average direction of both systems, DL8 exhibits a greater polarity than DL9. Effectively, k' of compound 18 on DL8 is twice the k' value on DL9. In the same way, compound 17 has a relatively good affinity for DNA, DNB and RCP systems (Fig. 3a). Fig. 3b shows that this affinity increases in the order TCP < DNB < DNA. In order to optimize a separation of compounds, it is necessary to put the stress first on the affinity and second on the selectivity. Trends in affinity and selectivity can be elucidated simultaneously by CFA.

Trends in affinity are deduced from the relative proximity of compound and system projections. This exploitation must be conducted progressively. The different planes of projection are examined successively. The apparent trends must be weighted by the individual contributions of solutes and/or systems to the factorial axes. The usual CFA of the NP submatrix is given in Fig. 4a and b. Fig. 4a is the simultaneous projection of compounds and systems on the two main factorial axes of inertia. Axis 1 represents 55% of the information content and axis 2 24%. Fig. 4b shows the projections on the plane defined by the factorial axes 2 and 3. Axis 3 corresponds to 10% of the information content. Hence these factorial axes integrate 89% of the information content.

Compound 18 is projected near DL8 and DL9, on the first factorial plane (Fig. 4a). It exhibits a strong affinity for these two systems. The difference between the two systems is emphasized in Fig. 4b, where compound 18 is closer to DL8 than DL9. Hence this compound has a greater affinity for the DL8 system. Compound 17 has no particular proximity with any systems, but it has a strong contribution to the first and second axes. These axes are defined mainly by the DNB, TCP and DNA systems for axis 1 and by the OD6 and OD2 systems for axis 2. Hence, in this particular case compound 17 has a strong affinity for all these systems. The affinity increases with the proximity of the above systems. The DNA, TCP and DNB systems are closed to compound 19 in the first factorial plane. Owing to the lower contribution to the first and second axes of this compound, the exploitation of this map cannot determine the

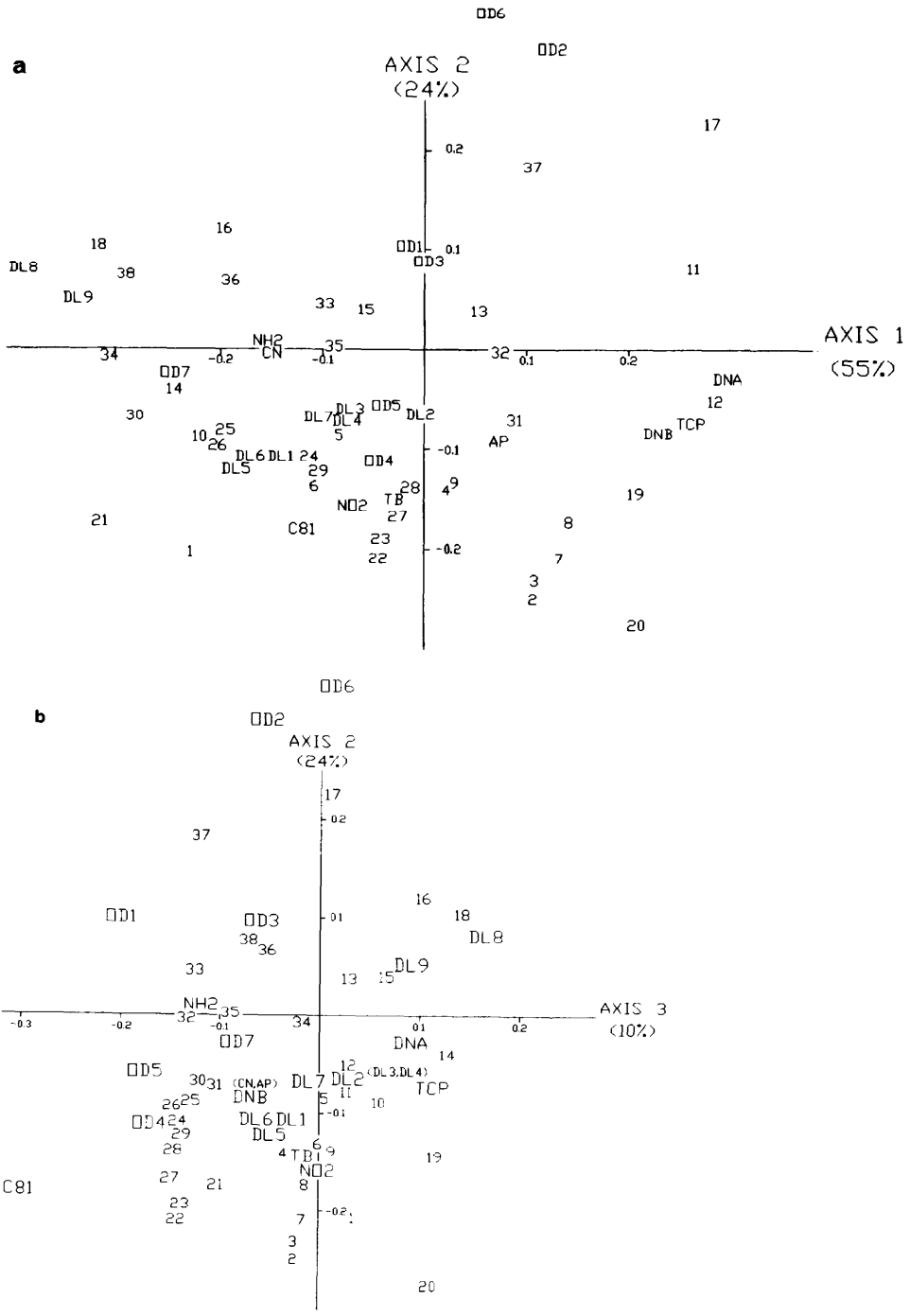

Fig. 4. (a) CFA trend analysis map. Simultaneous projection of the normal-phase chromatographic systems and compounds on the plane defined by the first and second best axes of inertia which represent 55% and 24%, respectively, of the information content. (b) CFA trend analysis map. Simultaneous projection of the normal-phase chromatographic systems and compounds on the plane defined by the second and third best axes of inertia which represent 24% and 10%, respectively, of the information content.

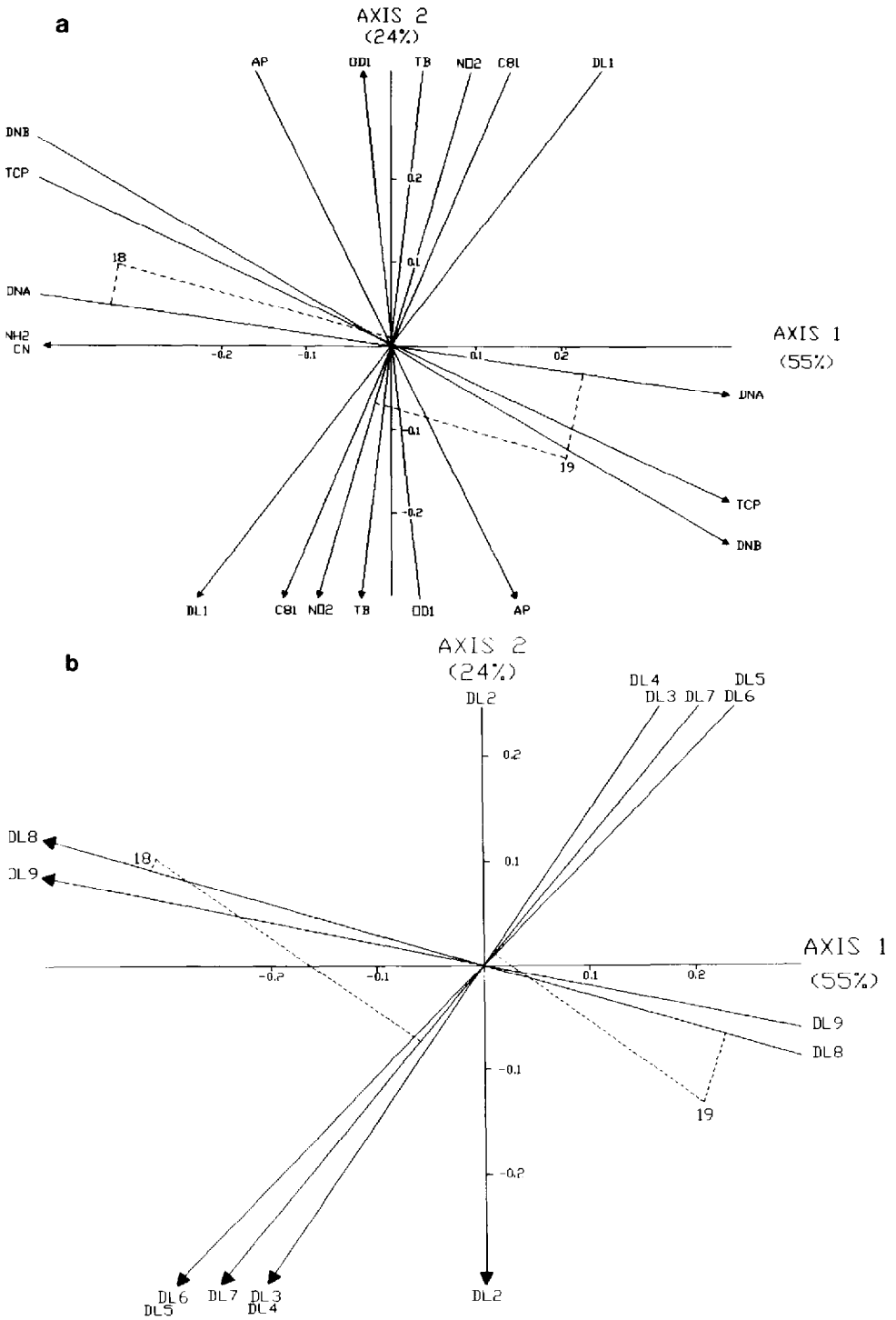

Fig. 5. (a) CFA distance analysis map. The chromatographic systems and compounds are selected from the previous CFA of the chromatographic normal-phase data matrix. The eleven presented chromatographic systems have the same eluent [heptane–THF (97:3, v/v)]. (b) CFA distance analysis map. The chromatographic systems and compounds are selected from the previous CFA of the complete normal-phase data matrix. The eight presented chromatographic systems have the same diol packing used with heptane plus different modifiers.

system which presents the greatest affinity. This is due to the determination of the principal factors, which are the average factors describing the behaviour of a large population of solutes and systems. Only "trends in affinity" can be reached here with different levels of reliability depending on the contribution to the extracted factors of solutes or systems analysed.

This common CFA map can be transformed to obtain the relative selectivity of the systems. Fig. 5a and b represent the transformed CFA maps from the previous ones limited to the example of solutes 18 and 19. Here, the eigenvalues of the first and second axes are 0.0428 and 0.0185, respectively. The selectivity is measured by the distance between the projections of two solutes on the axis defined by a transformed system. The relative selectivity of a chromatographic system, compared with the others, can be established. Two examples are proposed to illustrate the exploitation of the transformed CFA. In the first example (Fig. 5a), systems using the same eluent are extracted from the original CFA map. In the first factorial map, the eleven transformed systems define eleven directions drawn with solid lines. Arrows indicate that the true transformed projection of the systems are off the graph. Only two solutes are represented. The perpendicular projections of solutes on selected system axes are drawn with dashed lines. This map shows two main directions. The first is defined by the OD1, DL1, C81, NO2, TB and AP systems and the second by NH2, CN, DNB, TCP and DNA systems. These two directions are roughly perpendicular; the second main direction is parallel to the direction of solutes 18 19. The projection in dashed lines of these solutes are given on the directions of the DNA and NO2 systems. The observed distances between the projections of solutes 18–19 on the direction of a considered system are: (a) 12.3 arbitrary units for the DNA system and (b) 2.3 arbitrary units for the NO2 system. The corresponding selectivities of these two compounds, *i.e.*, their k' ratio, are 2.24 and 1.18 for the DNA and NO2 systems, respectively. The measure of the distances of the projections of solutes on particular system directions reflects the chromatographic selectivity.

For the chromatographer, the remaining problem is how to select a chromatographic system to separate this pair of solutes, *i.e.*, how to choose the best packing among others. It has been shown, in Fig. 5a, that the best selectivity is obtained with packings having the closest direction to the solute direction. For example, for the pair 18 19, the set of best packings is CN, NH2, DNA, TCP and DNB. The CFA map gives the relative affinity of solutes for these packings. To obtain an acceptable selectivity and a satisfactory affinity, the relative proximity of solutes and systems must be considered. For the pair 18–19, in Fig. 4a, the set of best packings is divided into two subsets. The first includes NH2 and CN systems and the second DNA, TCP and DNB systems. The first subset is between the two considered solutes whereas the second is close to solute 19. In the latter subset the best packing must be chosen. The selection of the retention time can be approached Fig. 3b.

The same type of map can be obtained to give a better understanding of the role of different eluents with the same diol packing (Fig. 5b). In this example the different contributions to factorial axes are not taken into account. Eight transformed chromatographic systems are presented. The projections of the model solutes 18–19 on DL7 and DL8 systems are drawn as dashed lines. Two main directions appear, the first for DL8 and DL9 systems and the second for DL2 up to DL7. The distances between the projections of compounds 18 and 19 on the two main directions are (a) 13 arbitrary

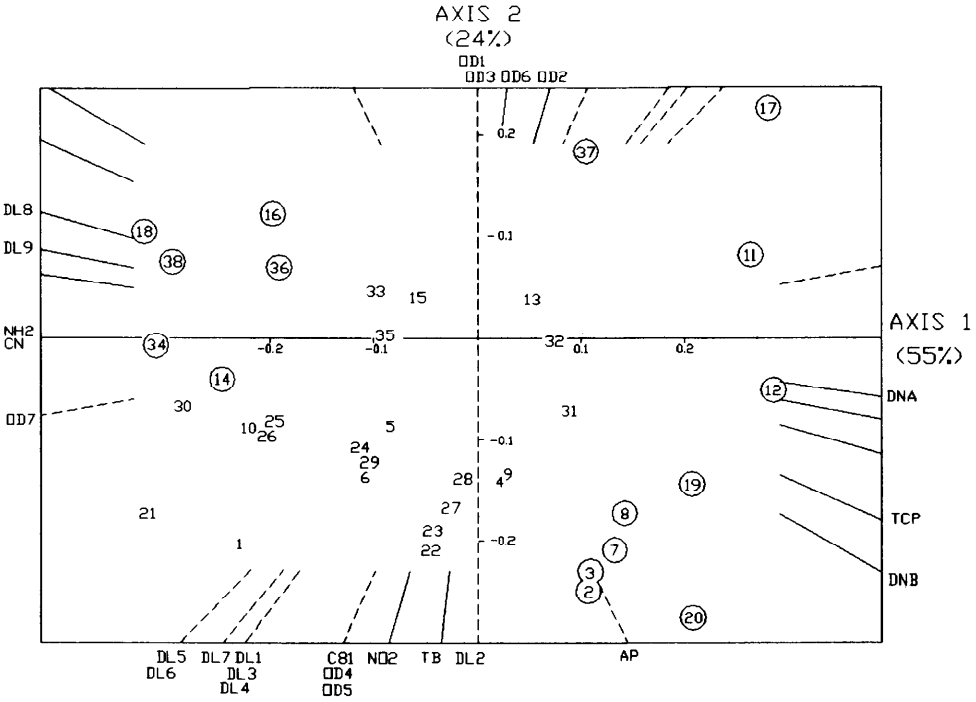

Fig. 6. CFA distance analysis map. The normal-phase data matrix is studied. Axes of the chromatographic systems are indicated. Axes of the chromatographic systems which have a lower contribution to inertia are drawn as dashed lines. Solutes which have a significant contribution to inertia of axes 1 and 2 are circled.

units for the DL8 system and (b) 2.5 arbitrary units for the DL7 system. The corresponding selectivities of these two compounds, *i.e.*, their k' ratio, are 2.7 and 1.2 for the DL8 and DL7 systems respectively. The best selectivity is obtained with solvents that have a parallel direction to the direction of the solute pair 18–19. DL8 and DL9 are selected, as they offer the greatest distance between the solutes projection. Systems have been selected regarding the selectivity by distances analysis. Now the affinity, *i.e.*, the time of chromatographic analysis, can be studied by "trends analysis" with CFA maps and by "affinity analysis" with PCA maps.

"Distance analysis" must be done carefully. The solute projections on the transformed systems are significant, at the first level of analysis, only if the chromatographic systems have a non-negligible contribution to the factorial axes. In the presented CFA, the most significant systems on axis 1 are DNA, DNB, NH2, TB, DL8 and DL9 and on axis 2 they are NO2, DNB, TCP, TB, CN, DL8, OD2 and OD6. With the other systems, their contribution to inertia must be considered to weight the distance analysis. The same restriction exists with solutes. On axis 1 solutes 11, 12, 14, 16–20, 34, 36 and 38 have a non-negligible contribution to inertia and on axis 2 solutes 2, 3, 7, 8, 16–20 and 37 have the same level of contribution to inertia.

A distances analysis map of the complete CFA is given Fig. 6. Systems that have a significant contribution to inertia are drawn as solid lines and other systems as dashed lines. Solutes that draw inertia of the solute cloud are circled.

Distances analysis of these solutes is possible with the eleven significant chromatographic systems. Specific treatment of chromatographic systems and solutes which have a weak contribution to the total inertia will be presented in a forthcoming paper. The interest in this map is that it offers an easy way to analyse the selectivity between two solutes. With this map, the selectivity is estimated by the distances between solute projections on chosen chromatographic systems. For example, let us consider the solute pair 2–3. Their own direction is approximatively perpendicular to axis 1 and parallel to axis 2. The best selectivity should be given with systems such as NO2, TB, OD2 or OD6. From the data matrix, OD6 is the most selective system. In the same manner, the best system, indicated by their identifier in parentheses, can be selected for some pairs of solutes such as 7 and 8 (OD6), 11 and 12 (NO2), 14 and 34 (TCP), 17 and 18 (DNA) and 20 and 37 (OD6).

CONCLUSIONS

To analyse a large set of homogeneous chromatographic data, it is necessary to use complementary chemometric methods. PCA and CFA do not extract the same factors.

With PCA the stress is put on the chromatographic affinity of solutes or on the chromatographic polarity of systems. The affinity of solutes and the polarity of systems are shown with two maps which are the projections of the solutes in their first factorial plane and the projection of the systems in their first factorial plane. The average direction(s) found on the correlation circle show the axis of variation of k'. When only NP or RP retention data are processed, the axis of k' variation is superimposed on the first factorial axis.

CFA puts the stress on relative chromatographic affinity and on selectivity. The data processing used in CFA hides the contribution of k' to the first factorial axis seen in PCA. CFA allows simultaneous solute and system projections on the same map. The relatives proximities of solutes and systems reflect the chromatographic affinity. The usual CFA maps give chromatographic trends of affinity and selectivity. Such projections could be called "trends analysis maps".

The selectivity analysis is improved with appropriate transformation of system projections. The distances between the solute projections on the system directions reflect the system selectivities. Selectivity analysis can be done easily with the transformed CFA maps, also called "distances analysis maps". Such a map can be exploited rapidly. The influent systems offering the best selectivity have the same direction as the two solutes considered.

Chromatographic analysis of a data set can be reduced to the study of factor analysis maps. The chromatographic properties, affinity and selectivity, are more or less nested in the three factor analysis maps: PCA affinity, CFA trends analysis map and CFA distances analysis map. An extensive study of chromatographic data requires the simultaneous grasping of the information content deduced from these three maps.

REFERENCES

1 E. R. Malinowski and D. G. Howery, *Factor Analysis in Chemistry*, Wiley, New York, 1980.

2 R. F. Hirsch, R. J. Gaydosh and J. R. Chrétien, *Anal. Chem.*, 52 (1980) 723.
3 D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte and L. Kaufman, *Chemometrics: a Textbook (Data Handling in Science and Technology*, Vol. 2), Elsevier, Amsterdam, 1988, p. 371.
4 B. Walczak, M. Dreux, J. R. Chrétien, K. Szymoniak, M. Lafosse, L. Morin-Allory and J. P. Doucet, *J. Chromatogr.*, 353 (1986) 109.
5 B. Walczak, L. Morin-Allory, J. R. Chrétien, M. Lafosse and M. Dreux, *Chemometr. Intell. Lab. Syst.*, 1 (1986) 79.
6 B. Walczak, J. R. Chrétien, M. Dreux, M. Lafosse, L. Morin-Allory, K. Szymoniak and F. Membrey, *J. Chromatogr.*, 353 (1986) 123.
7 B. Walczak, J. R. Chrétien, M. Dreux, L. Morin-Allory and M. Lafosse, *Chemometr. Intell. Lab. Syst.*, 1 (1987) 177.
8 B. Walczak, M. Lafosse, J. R. Chrétien, M. Dreux and L. Morin-Allory, *J. Chromatogr.*, 369 (1986) 27.
9 J. R. Chrétien, B. Walczak, L. Morrin-Allory, M. Dreux and M. Lafosse, *J. Chromatogr.*, 371 (1986) 253.
10 B. Walczak, L. Morin-Allory, M. Lafosse, M. Dreux and J. R. Chrétien, *J. Chromatogr.*, 395 (1987) 183.
11 B. Walczak, M. Dreux, J. R. Chrétien, L. Morin-Allory, M. Lafosse and G. Felix, *J. Chromatogr.*, 464 (1989) 237.
12 M. Righezza and J. R. Chrétien, *J. Chromatogr.*, 544 (1991) 393–411.