

Comparison of two principal component analysis methods to evaluate reversed-phase retention data

TIBOR CSERHÁTI*† and ZOLTÁN ILLÉS‡

† *Central Research Institute for Chemistry, Hungarian Academy of Sciences, H-1525 Budapest, POB 17, Hungary*

‡ *Institute for Managing Natural Resources and Protecting the Environment, Budapest, Hungary*

Abstract: The retention of twelve 2-nitro-4-cyanophenyl esters showing marked herbicidal activity was determined in 23 reversed-phase thin-layer chromatographic systems. The retention data set was evaluated by principal component analysis (PCA). To assess the effect of the information loss caused by normalization, PCA was separately carried out on the covariance (method A) and on the correlation matrix (method B). The ratio of the variances explained was very similar for both methods, however, the PC loadings and the coordinates of the two-dimensional nonlinear maps showed poor correlation. The distribution of the 2-nitro-4-cyanophenyl esters and that of chromatographic systems showed differences on the two-dimensional nonlinear maps of PC loadings and PC variables, however, the general trend was similar independently of the application of method A or B. The findings indicate that the application of the correlation matrix as basis for the PCA calculations may lead to slightly distorted results that strongly advocates the use of covariance matrix in PCA.

Keywords: *Principal component analysis; correlation matrix; covariance matrix; two-dimensional nonlinear mapping; stepwise regression analysis.*

Introduction

The development of new methods to interpret large data sets has been one of the major advances of chromatography during past decades. The major features of this evolving field are high-speed computers and a variety of mathematical–statistical methodologies. The modern multivariate mathematical–statistical methods make possible the simultaneous evaluation of a practically unlimited number of variables (chromatographic parameters) which highly facilitates the solution of theoretical and practical problems. In particular, multivariate methods have been applied in chromatography to identify basic factors which influence solute–solvent interactions and to classify solutes and solvents into groups having similar characteristics. The application of various multilinear regression methods requires the presence or selection of a dependent variable. However, in many cases the scientist is not interested in the dependence of one parameter (dependent variable) on the other parameters (independent variables), but rather wishes to find the relationship between all parameters without one being the dependent variable. Principal component analysis (PCA) complies with these

requirements [1]. The main advantages of PCA in chromatography are:

- (a) clustering of the variables according to their relationship (clustering chromatographic systems or solutes according to their retention behaviour);
- (b) the possibility of the extraction of one or more background variables having concrete physicochemical meaning for the theory and practice of chromatography;
- (c) easy visualization of the clusters by a nonlinear mapping technique [2];
- (d) a decrease in the number of variables (a decrease in the number of chromatographic systems or solutes to the minimum necessary for the solution of a problem).

PCA can be carried out both on the correlation and covariance matrices of the original data set. The application of a correlation matrix is common in PCA calculations. However, the normalization may cause information loss that can distort the results. As far as we are aware, the effect of normalization on the information content of PCA has never been studied in detail. In recent years, Quantitative Structure–Activity Relationship (QSAR) studies have promoted not only the realistic

* Author to whom correspondence should be addressed.

design of new bioactive compounds [3, 4] but also helped the better understanding of the various biochemical and/or biophysical processes accounting for biological efficiency [5, 6]. Lipophilicity as an important physico-chemical property of bioactive molecules has been frequently used in QSAR studies [7]. Reversed-phase thin-layer chromatography (RPTLC) has been extensively applied to determine lipophilicity [8, 9]. The lipophilicity determination by RPTLC offers some advantages: it is rapid and simple, does not need pure compounds and uses only a few micrograms of material. However, the support may partially retain its original adsorptive character even after impregnation [10], and the R_M value characterizing the molecular lipophilicity in RPTLC may depend on adsorption behaviour [11] and on the surface pH value of the support [12, 13]. In RPTLC, the supports (generally silica) were impregnated with paraffin or silicone oil [14]. According to our knowledge the suitability of the various impregnating agents for lipophilicity determinations in RPTLC has never been studied in detail.

Phytotoxicity of 3,5-dihalogeno-4-hydroxybenzonitriles [15] of their esters [16] and of the esters of 3-nitro-5-bromo-benzo-nitrile [17] is well known. The biological activity of the 2-nitro-4-cyanophenyl esters depends on the structural characteristics and on the lipophilicity of the compound [18]. The objectives of this work were to compare the results calculated from the same data set using PCA on the

covariance and on the correlation matrices, to elucidate the role of various chromatographic conditions (eluent composition, the type of impregnating agent and support) on the lipophilicity determinations of some 2-nitro-4-cyanophenyl esters and to assess the impact of the individual chromatographic parameters on the retention.

Experimental

The chemical structure of the 2-nitro-4-cyanophenyl esters is given in Table 1. The compounds were synthesized at the CHINOIN Chemical and Pharmaceutical Works (Budapest, Hungary). Silicone oils of various molecular weights ($S_1 = 30$ kDa, $S_2 = 44$ kDa, $S_3 = 60.5$ kDa, $S_4 = 64$ kDa, $S_5 = 81$ kDa, $S_6 = 93$ kDa, $S_7 = 114$ kDa and $S_8 = 142$ kDa) were purchased from Wacker Chemie GmbH (München, Germany) and were used as purchased. Kieselgel 60, Aluminiumoxid 60 and Cellulose TLC supports were purchased from Merck (Darmstadt, Germany). The impregnation of supports were carried out as follows.

The silicone oils were dissolved in chloroform at the concentration of 10 mg ml^{-1} . The support was added to this solution and shaken for 2 h (for ratios see Table 2). The chloroform was evaporated under vacuum, and plates of 20×20 cm and 0.25 mm layer thickness were prepared from the impregnated supports. This

Table 1
Chemical structure of 2-nitro-4-cyanophenyl esters

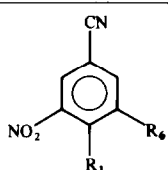
		
General structure		
Number of compound	R_1	R_6
1	O—CO—CH ₃	H
2	O—CO—CH ₃	Cl
3	O—CO—CH ₃	Br
4	O—CO—CH ₃	I
5	O—CO—O—CH ₃	H
6	O—CO—O—CH ₃	Cl
7	O—CO—O—CH ₃	Br
8	O—CO—O—CH ₃	I
9	O—CO—O—CH ₃	Br
10	O—CO—O—(CH ₂) ₂ —O—CH ₃	Br
11	O—CO—CH ₂ —CH=CH ₂	Br
12	O—CO—O—C ₆ H ₅	Br

Table 2
Reversed-phase chromatographic systems

Number of system	Support	Impregnating agent		Eluent composition (acetone:water, v/v)
		Type	(% w/w)	
I	Kieselgel	S ₁	5	1:1
II	Kieselgel	S ₂	5	1:1
III	Kieselgel	S ₃	5	1:1
IV	Kieselgel	S ₄	5	1:1
V	Kieselgel	S ₅	5	1:1
VI	Kieselgel	S ₆	5	1:1
VII	Kieselgel	S ₇	5	1:1
VIII	Kieselgel	S ₈	5	1:1
IX	Kieselgel	S ₁	5	1:3
X	Kieselgel	S ₂	5	1:3
XI	Kieselgel	S ₃	5	1:3
XII	Kieselgel	S ₄	5	1:3
XIII	Kieselgel	S ₅	5	1:3
XIV	Kieselgel	S ₆	5	1:3
XV	Kieselgel	S ₇	5	1:3
XVI	Kieselgel	S ₈	5	1:3
XVII	Kieselgel	S ₁	1.61	1:3
XVIII	Aluminium oxide	S ₁	5	1:3
XIX	Aluminium oxide	S ₁	1.61	1:3
XX	Aluminium oxide	S ₆	5	1:3
XXI	Cellulose	S ₁	5	1:4
XXII	Cellulose	S ₁	1.61	1:4
XXIII	Cellulose	S ₆	5	1:4

impregnation method was necessitated by the fact that:

- the immersion method produces different surface covering of the different supports, that makes comparison impossible;
- the silicone oils with the highest molecular mass do not move adequately in the predevelopment impregnation method.

The reversed-phase chromatographic systems are given in Table 2. The aim of our experimental design was to assess the impact of eluent composition and those of the support and impregnating agent on the reversed-phase retention of some 2-nitro-4-cyanophenyl esters. The inclusion of the lower impregnation rate (1.61% S₁ corresponds to 5% S₆ in the number of adsorbed molecules) was motivated by the hypothesis that in the case of one point adsorption (only one end of the silicone oil molecule is adsorbed on the support surface) the retention strength has to be similar on both impregnated supports.

The esters were dissolved in acetone at a concentration of 2 mg ml⁻¹, 5 µl of each solution were spotted onto the plates. After development the spots were detected by their visible spectra. Each determination was run in quadruplicate.

The R_M values characterizing the molecular lipophilicity in RPTLC have been calculated:

$$R_M = \log (1/R_f - 1). \quad (1)$$

PCA was carried out in two different manners. In both cases the compounds were taken as variables, the R_M values served as observations. The calculation was separately carried out on the covariance (Method A) and on the correlation matrix (Method B) of the data set. The two-dimensional nonlinear map of PC loadings and variables was also calculated. To assess the similarities and dissimilarities between the calculations, the PC loadings of the methods have been compared. The first five PC loadings (containing the overwhelming majority of the total variance) of method A was taken as dependent and the first five PC loadings of the method B as independent variables. The calculations were carried out with stepwise regression analysis [19], the acceptance level for the independent variables was set to 95% significance level. Stepwise regression analysis was carried out five times the first five PC loadings of method A being the dependent variables.

To compare the nonlinear maps, linear regressions were calculated between the corresponding coordinates of the two-dimensional nonlinear maps.

To find physical or physicochemical meanings of the PC loadings and of the coordinates of the nonlinear maps stepwise regression analysis was applied again. The structural characteristics of the 2-nitro-4-cyanophenyl esters were taken as independent variables. The calculations were carried out seven times, the first five PC loadings and the two coordinates of the nonlinear map being the dependent variables. The other conditions were the same as before.

Results and Discussion

Retention behaviour of solutes and chromatographic systems

The molecular mass of the impregnating silicone oil did not have any appreciable effect on the strength of retention, the retention order was similar for each silicone oil (Fig. 1). This finding indicates that the molecular mass has a negligible effect on the lipophilicity determination of these esters. The retention decreased with decreasing quantity of impregnating agent, that is the mass of the silicone oil and not the number of the adsorbed silicone oil molecules account for the retention. The type

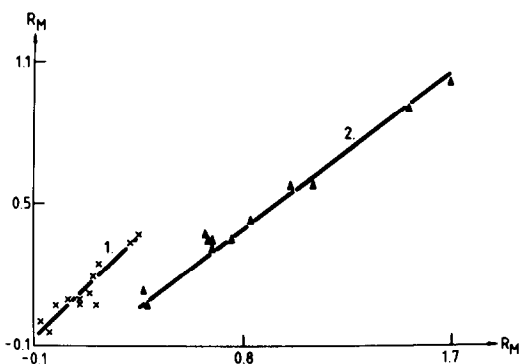


Figure 1
Relationship between the R_M values of some 2-nitro-4-cyanophenyl esters determined in various RPTLC systems. (1) Systems II-VIII; (2) Systems IX-XVII. For symbols see Table 2.

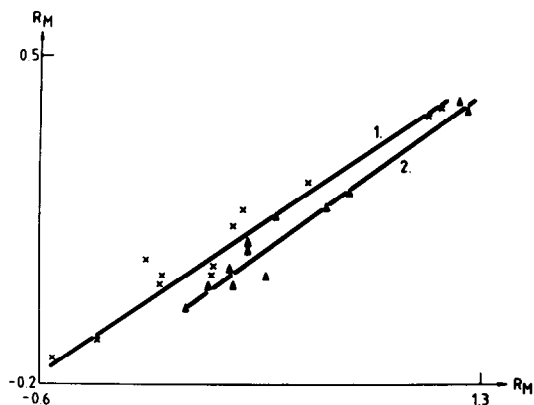


Figure 2
Relationship between the R_M values of some 2-nitro-4-cyanophenyl esters determined in various RPTLC systems. (1) Systems I-XXII; (2) Systems VI-XX. For symbols see Table 2.

of support influenced the strength of retention but not the retention order (Fig. 2), that is for the lipophilicity determination of these compounds each support can be successfully applied.

The results of both PCAs are compiled in Table 3. The retention behaviour of the 23 chromatographic systems can be described by five background variables. In other words, five chromatographic systems are sufficient to obtain the overwhelming majority of information content of the 23 systems. Unfortunately, PCA does not define these five systems, only indicates their mathematical possibility.

The conclusions drawn from the two-dimensional nonlinear maps of PC loadings calculated with different methods (Figs 3 and 4) are the same; namely that the type of halogen substitution accounts for the similar or dissimilar retention behaviour of the esters (group A: esters without halogen substitution; group B: esters with chloro substituent; group C: esters with iodo substituent; and group D: esters with bromo substituent).

Table 3

Results of principal component analysis carried out on the covariance (A) and on the correlation matrix (B) of the original data set

Number of principal component	Per cent variance explained		Per cent total variance explained	
	A	B	A	B
1	33.12	34.28	33.12	34.28
2	22.86	17.89	55.98	52.18
3	17.83	16.15	73.81	68.32
4	9.55	12.98	83.36	81.30
5	9.41	10.12	92.76	91.42

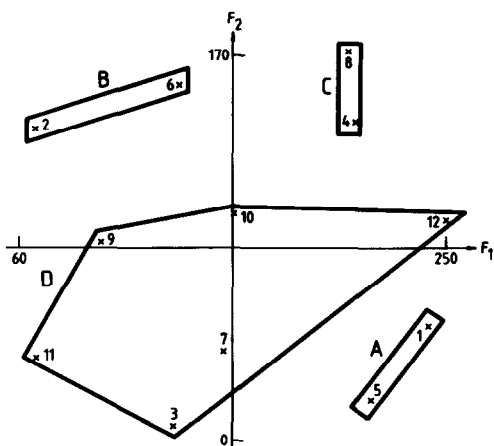


Figure 3
Two-dimensional nonlinear map of PC loadings. Covariance matrix. Number of iterations: 134. Maximum error: 4.86×10^{-2} . Numbers refer to 2-nitro-4-cyanophenyl esters in Table 1.

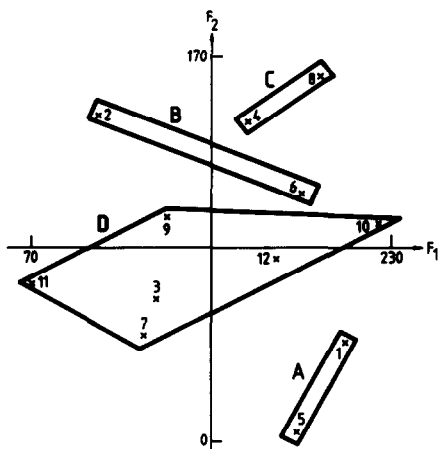


Figure 4
Two-dimensional nonlinear map of PC loadings. Correlation matrix. Number of iterations: 53. Maximum error: 6.42×10^{-2} . Numbers refer to 2-nitro-4-cyanophenyl esters in Table 1.

RPTLC Systems I–VIII form a loose cluster, whereas the others are quasi-randomly distributed on both the two-dimensional nonlinear maps of PC variables (Figs 5 and 6). This finding can be explained by the supposition that the effect of higher acetone concentration governs the retention and overshadows the influence of other chromatographic parameters (type of support, quality and quantity of impregnating agent) on the retention. At lower acetone concentrations each variable has a similar impact on the retention order of 2-nitro-4-cyanophenyl esters resulting in the random-like distribution of chromatographic systems.

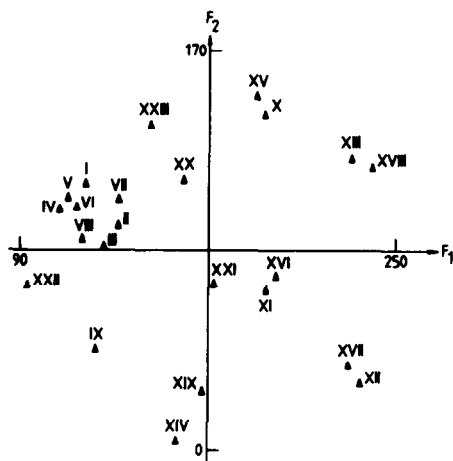


Figure 5
Two-dimensional nonlinear map of PC variables. Covariance matrix. Number of iterations: 117. Maximum error: 4.54×10^{-2} . Numbers refer to chromatographic systems in Table 2. F_1 and F_2 are the dimensionless coordinates of the two-dimensional nonlinear map without any concrete physicochemical meaning.

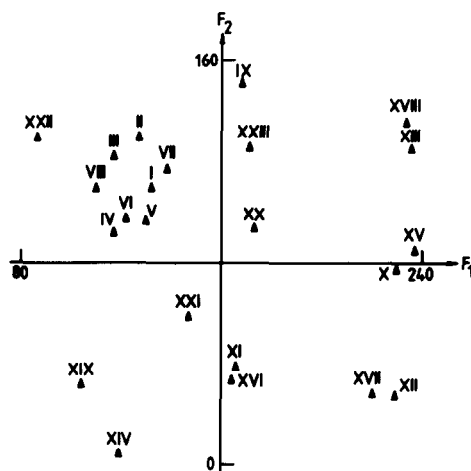


Figure 6
Two-dimensional nonlinear map of PC variables. Correlation matrix. Number of iterations: 107. Maximum error: 5.37×10^{-2} . Numbers refer to chromatographic systems in Table 2. F_1 and F_2 are the dimensionless coordinates of the two-dimensional nonlinear map without any concrete physicochemical meaning.

The parameters of correlations between the substituents of the 2-nitro-4-cyanophenyl esters and the PC loadings and coordinates of the two-dimensional nonlinear map are compiled in Table 4. The first and fifth PC loadings were not related to any substituents. The first PC loading containing the highest ratio of variance is only correlated with the presence of halogen substituents in the molecule. As the halogen substitution accounts for the overwhelming majority of variance, these sub-

Table 4

Parameters of the linear correlations between the principal component loadings, coordinates of the two-dimensional nonlinear map and the substituents of 2-nitro-4-cyanophenyl esters

Results of stepwise regression analysis, $n = 12$						
I	$L_{A2} = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3$					
II	$L_{A3} = a + b_4 \cdot x_4$					
III	$L_{A4} = a + b_1 \cdot x_1 + b_5 \cdot x_5$					
IV	$F_1 = a + b_4 \cdot x_4$					
V	$F_2 = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3$					
Number of equation	F	r^2	Per cent path coefficients			
			1	2	3	5
I	18.59	0.8745	33.53	24.96	41.51	—
II	7.11	0.4158	—	—	—	—
III	8.55	0.6552	44.37	—	—	55.63
IV	5.66	0.3614	—	—	—	—
V	15.09	0.8499	34.19	23.10	42.71	—
x_1	= presence of Cl substituent in the molecule.					
x_2	= presence of Br substituent in the molecule.					
x_3	= presence of I substituent in the molecule.					
x_4	= presence of O—CO—CH ₂ —CH=CH ₂ substituent in the molecule.					
x_5	= presence of O—CO—O—(CH ₂) ₂ —O—CH ₃ substituent in the molecule.					
F_1 and F_2	= coordinates of the two-dimensional nonlinear map.					
L	= principal component loading; subscripts indicate methods A or B and the number of PC loadings.					

Table 5

Parameters of the linear correlations between the principal component loadings of the methods A and B. Results of stepwise regression analysis ($n = 12$; for symbols see Table 4)

I	$L_{A2} = a + b_1 \cdot L_{B2} + b_2 \cdot L_{B3} + b_3 \cdot L_{B5}$				
II	$L_{A3} = a + b_1 \cdot L_{B2} + b_2 \cdot L_{B3} + b_3 \cdot L_{B4}$				
III	$L_{A4} = a + b_1 \cdot L_{B3} + b_2 \cdot L_{B4} + b_3 \cdot L_{B5}$				
IV	$L_{A5} = a + b_1 \cdot L_{B4} + b_2 \cdot L_{B5}$				
Number of equation	F	r^2	Per cent path coefficients		
			1	2	3
I	62.39	0.9590	59.34	28.01	12.65
II	32.52	0.9242	27.40	48.31	24.29
III	46.23	0.9455	23.77	37.90	38.33
IV	37.41	0.8926	47.49	52.51	—

stituents govern the reversed-phase retention of 2-nitro-4-cyanophenyl esters. This result entirely supports our previous qualitative conclusions drawn from the maps of PC loadings.

Comparison of the two methods of principal component analysis

The variances explained by the first five principal component loadings show very little deviation according to the method of calculation (Table 3). This finding suggests that the information loss caused by the normalization of data is negligible, and the application of the correlation matrix has no drawbacks. However, the distribution of the solutes on the two-

dimensional nonlinear maps of PC loadings shows some marked differences (Figs 3 and 4). From chromatographic point of view, these differences do not influence the evaluation of the maps but indicate the inherent uncertainty associated with the application of the correlation matrix for PCA calculations.

As in the case of the two-dimensional nonlinear maps of PC loadings, the distribution of the chromatographic systems also shows marked differences (Figs 5 and 6). In contrast to the highly similar quantity of explained variances this finding indicates again that the two calculation methods can lead to different results.

Some parameters of the correlations between the principal component loadings of the methods A and B are given in Table 5. The first PC loading of method A did not show any correlation with the PC loadings of method B. As the first loading generally explains the majority of variance, this result emphasizes again the difference between the two methods of calculation. The other correlations are generally good, and the ratio of the variance explained is high. However, this result is somewhat misleading, because the correlation between the corresponding PC loadings is poor. The inclusion of more than one significant independent variable in the equations indicates that the orthogonality of the two PC loading sets is different and that explains the differences between the two-dimensional non-linear maps.

References

- [1] K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, pp. 213–254. Academic Press, London (1979).
- [2] J. W. Sammon, Jr, *IEEE Transactions on Computers* **C18**, 401–407 (1969).
- [3] J. Andrews, W. E. Stuper, P. Brugger and S. Jurs, *Computer Assisted Studies of Chemical Structure and Biological Function*. J. Wiley and Sons, New York (1979).
- [4] C. Hansch, in *Structure–Activity Relationships* (C. J. Cavallito, Ed.), pp. 75–160. Pergamon Press, Oxford (1973).
- [5] A. Lopata, F. Darvas, K. Valkó, Gy. Mikite, E. Jakucs and A. Kiss-Tamás, *Pestic. Sci.* **14**, 513–520 (1983).
- [6] T. Cserhádi, J. Bojarski, É. Fenyvesi and J. Szejtli, *J. Chromatogr.* **351**, 356–362 (1986).
- [7] R. Franke, in *QSAR and Strategies in the Design of Bioactive Compounds* (J. K. Seydel, Ed.), pp. 59–67. VCH Verlagsgesellschaft, Weinheim, Germany (1985).
- [8] C. B. C. Boyce and B. B. Milborrow, *Nature* **208**, 537–538 (1965).
- [9] G. L. Biagi, A. M. Barbaro and M. C. Guerra, *J. Chromatogr.* **41**, 371–379 (1969).
- [10] W. V. van Giesen and L. H. M. Janssen, *J. Chromatogr.* **237**, 199–213 (1982).
- [11] M. C. Guerra, A. M. Barbaro, G. Cantelliforti, M. T. Foffani, G. L. Biagi and A. Borea, *J. Chromatogr.* **216**, 93–102 (1981).
- [12] Z. Illés and T. Cserhádi, *J. Plan. Chromatogr.* **1**, 231–234 (1988).
- [13] Z. Illés and T. Cserhádi, *J. Plan. Chromatogr.* **2**, 92–94 (1989).
- [14] E. Stahl, *Dünnschichtchromatographie*, pp. 479 and 202. Springer-Verlag, Berlin (1962).
- [15] K. Carpenter and B. J. Heywood, *Nature* **200**, 28–29 (1963).
- [16] B. J. Heywood, *Chem. Ind.* **47**, 1946–1952 (1966).
- [17] Z. Szigeti, *Acta Biochim. Biophys. Acad. Sci. Hung.* **19**, 133 (1984).
- [18] C. Bujtás, T. Cserhádi, E. Cseh, Z. Illés and Z. Szigeti, *Biochem. Physiol. Pflanzen* **182**, 465–471 (1987).
- [19] H. Mager, *Moderne Regressionsanalyse*, pp. 135–157. Salle, Sauerlander, Frankfurt am Main (1982).

[Received for review 7 July 1991]