CROATICA CHEMICA ACTA
CCACAA **77** (1–2) 117–125 (2004)

ISSN-0011-1643
CCA-2908
Original Scientific Paper

Procrustes Rotation and Pair-wise Correlation: a Parametric and a Non-parametric Method for Variable Selection*

Károly Héberger[a,] and José M. Andrade[b]**

[a]*Institute of Chemistry, Chemical Research Center, Hungarian Academy of Sciences,
P.O. Box 17, H-1525, Budapest, Hungary*

[b]*Department of Analytical Chemistry, University of A Corunna,
Campus da Zapateira s/n, E-15071, A Corunna, Spain*

RECEIVED FEBFUARY 5, 2003; REVISED JULY 1, 2003; ACCEPTED JULY 1, 2003

Key words
Procrustes analysis
robust method
feature selection
classification
pair-wise correlation
similarity measures

In this work attention is focused on selecting a small set of original (independent) variables, which take into account the most important information present in the data matrix. The performance of two already implemented methods is compared from a practical point of view: the Procrustes Rotation algorithm, which is parametric, and the Pair-wise Correlation (PCM), which is nonparametric, because the tests used to discriminate between the variables are nonparametric. Using a well-documented data set (aphid data), both methods gave comparable results. Procrustes Rotation selected four variables, whereas four or five variables were retained using the pair-wise correlation method (depending on the test used). Three variables were common to both approaches and the main structure in the original data set was retained in both cases. Therefore, both methods are appropriate for variable selection. Selection criterion for ranking the variables using PCM was further developed to retain the highest »dissimilarity« among similar variables.

INTRODUCTION

Most studies in analytical chemistry apply data sets in which several variables (parameters, responses, *etc.*) are measured for each sample. Typical examples extend from industrial quality control to environmental studies or spectroscopy-related measurements. However, some concern arises in the chemist's mind about the usefulness and adequacy of such complex data sets. Interesting questions should be addressed like: Are all those variables really needed to describe the problem? or Should all these variables be measured time and again to monitor the system under study?

Many times it is recognized that some sort of variable reduction can be made. Nevertheless, the problem remains how to perform this in an »objective«, reproducible way. In other words, how can variable selection be made without losing essential information? This topic is not new in the chemometric literature (see, *e.g.*, Ref. 1 for a nice introduction and several techniques). The variable selection process has been more or less solved for

[*] Dedicated to Professor Nenad Trinajstić on the occasion of his 65th birthday.
[**] Author to whom correspondence should be addressed. (E-mail: heberger@chemres.hu)

linear relationships,[2] although different approaches have been proposed for more complex situations (a complete review is out of the scope of this work).

Relevant variables can be selected by using the principal component analysis (PCA)[3] and partial least squares (PLS) regression.[4-9] The variable selection by PCA (PLS) is frequently made according to highest loadings (regression coefficients), but in general it selects a set of variables. As there is no criterion that could attribute probability to the selected variables, the significance of variables selected by PCA or PLS is not known. The definition of an index was proposed as a good alternative to search for the best linear fit throughout different spectral ranges (but it does not allow selection of individual variables, which would be of most interest in some applications).[10]

A variant of the evolving factor analysis was employed to define »resolvability indices«[11] and they were used to resolve chromatographic peaks by the window factor analysis. Variable selection can also be made in graphical ways based on the confidence region of the estimated concentrations.[12] Several methods are suggested for variable selection using artificial neural networks (ANN)[13-16] as well as genetic algorithms (GA),[17-19] since they can easily be modified for such purposes.

In this paper, attention will be focused on comparing two variable selection methods from a practical point of view. They have already been proved useful in some previous works for selecting original (individual) variables. One of them, Procrustes Rotation (PR), is a parametric approach that uses a selection algorithm[1,20-21] developed to select the minimum set of original variables that comprise most of the information of the initial data set; the other, Pair-wise Correlation Method (PCM), is a nonparametric alternative that relies upon ranking variables in the order of their »superiority«.[22-25] As their mathematical basis is quite different, it seemed interesting to compare how they perform on a common data set.

Therefore, they will be applied to a well-known data set from the field of biometrics. Thus, 40 winged aphid individuals (*alate adelges*) were captured in a light-trap and, then, 19 morphological parameters were measured on each of them.[26] The input matrix is given in Refs. 26–27 or is available from the authors upon request.

THEORETICAL BACKGROUND

Despite giving some essential details of the two multivariate techniques, this paper is not intended to address all the mathematical details nor to make a theoretical comparison between the two methods; more specific details can be found elsewhere (*e.g.*, Refs. 1, 20, 21, 27).

Procrustes Rotation

Procrustes rotation, PR, (sometimes called Procrustes analysis) is a multivariate technique, which rotates, translates or stretches two (or several) multivariate configurations whose points can be matched. The name stems from the Greek mythology: Procrustes stretched the visitors or chopped off their legs to match exactly the size of his bed.

The mathematics behind the technique is based on singular value decomposition. Let us suppose that the coordinates of n points in two configurations are given in the rows of two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, with each row of one matrix being matched to the corresponding row of the other. It is assumed that the dimensionalities of the two configurations are the same, so each matrix has the same number p of columns. If this is not the case, sufficient columns of zeros can be appended to the »smaller« matrix to make up for the deficit. Let us also suppose that one of the configurations, $\boldsymbol{X}$, is fixed and we wish to match the other one to it. Geometrically this is done by translating, rotating and/or reflecting, and then stretching or shrinking $\boldsymbol{Y}$ in such a way that the sum of squared distances, M^2, between the points of $\boldsymbol{Y}$ and the corresponding ones of $\boldsymbol{X}$ is a minimum. The »discrepancy« between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is then given by M^2. The smaller this value, the more similar are the two configurations, with a perfect match given by the value of zero. The steps of the process are carried out sequentially. For mathematical justification, see Ref. 1; here we just give the final result.

The first step, translation, is simply the mean centering of both $\boldsymbol{X}$ and $\boldsymbol{Y}$ at the outset. To determine the second and third steps, a singular value decomposition of the matrix product $\boldsymbol{X}^{\mathbf{T}}\boldsymbol{Y}$ can be performed, whose result can be written as $\boldsymbol{UDV}^{\mathbf{T}}$. Any physical rotation or reflection can be mathematically formulated using an orthogonal matrix, and in the present case the optimal rotation/reflection is given by the matrix product $\boldsymbol{VU}^{\mathbf{T}}$.

The key point here is to define the optimal number of principal components (q) to describe the data set. Krzanowski proposed a very effective test, Krzanowski's W_{m}, which is very similar to an F-test because it compares the increase in the predictive information obtained by introducing a given component into the model to the average information contained in the remaining components;[1] it leads to excellent results and is of general use.

As the number of variables increases, the number of possible combinations becomes so large that it is worth limiting the efforts to the »best« variables. The minimal number of original variables that can preserve the overall structure of the data set (described by q principal components) can be reasonably elucidated using the following approach:

(i) delete in turn each variable from the data set;

(ii) compute the first q principal components from the reduced data matrix (let this matrix be $\boldsymbol{R}$, reduced sample score matrix);

(iii) compare $\boldsymbol{R}$ and $\boldsymbol{T}$ (the »true« and target configuration, which is obtained considering the first q scores extracted when all the variables have been considered) in order to decide whether the loss of information is large; the M^2 statistics is employed here;

(iv) repeat this process for each variable;

(v) select the variable that led to the lowest disturbance of the data configuration and delete it from the original data set;

(vi) close the loop returning to step (i).

The process is repeated until only q variables remain (q equals the optimal number of principal components). These will therefore be the »best« q variables to use, in the sense that they are the ones that best capture the structure of all original p variables.

Generalized Pair-wise Correlation Method

The pair-wise correlation method (PCM) is a nonparametric alternative to variable selection. PCM, in its original form, can discriminate only between two variables.[22,23] It uses a piece of information present in the data that has been neglected till now. Let us assume that a dependent variable (Y) contains the information on the samples (type, origin, class, concentration, biological activity, *etc.*) and that two independent features (parameters or variables, $X1$ and $X2$) have been measured for each of them. To make a distinction, if any, a correct selection criterion is needed.[23] Four basic events can be distinguished: both variables enhance the correlation; one enhances and the other diminishes, and *vice versa*; both diminish the correlation between dependent variable Y and independent variables $X1$ or $X2$. Arranging properly the frequencies of the four basic events in a contingency table, a significant difference can be determined using nonparametric tests, *e.g.*, Conditional exact Fisher test, McNemar test, Chi-square test. Even a parametric test, the Williams t statistics is at our disposal.[23] Here, correlation was established among each X variable and a Y variable created by giving a code to each sample belonging to a given group of samples (class of aphid). In general, the assignment should be performed employing previous information or after a preliminary PCA study.

PCM generalization is necessary if the variable selection is to be carried out with more than two variables. The generalization can be done in several ways. First, simple PCM analysis should be done for all possible and different variable-pairs. Every pair-wise comparison can mark a variable as superior (»winner«), inferior (»loser«) or no decision can be made. If a given statistical test indicates a significant difference between the variables, the terms superior - inferior or winner - loser are used. Then, the variables are ranked according to the number of their superiority (wins). This is called simple ranking (SR). Ranking can be also done using the differences between superiority and inferiority (RD). Finally, ranking can be carried out according to the probability weighted differences between superiority and inferiority (RP). All the details of the generalized pair-wise correlation procedure (GPCM) can be found in Refs. 25, 26.

The main advantage of GPCM is that it is nonparametric and the selection criteria are statistically well based, in contrast to many empirical approaches. The nonparametric character allows us to compare and select variables measured in different units without any scaling. Because of the well-defined confidence limits for the selection of variables, GPCM has an objective criterion[23] compared to the empirical, specific-problem solutions. Hence, the selection of individual explanatory variables can be achieved, which is a unique property among the methods currently focused on developing prediction models.

This work aims to compare the performance of two methods: Procrustes analysis and GPCM, both having well defined selection criteria, but based on different principles. Although PR is an unsupervised and GPCM is a supervised method, the comparison can be made inspecting the sample groups that the selected variables allow to obtain.

RESULTS AND DISCUSSION

It is interesting to present the original data set projected onto the plane formed by the first two principal components (PC1-PC2, they account for *ca.* 85.3 % of the total variance) as a reference to be compared with any

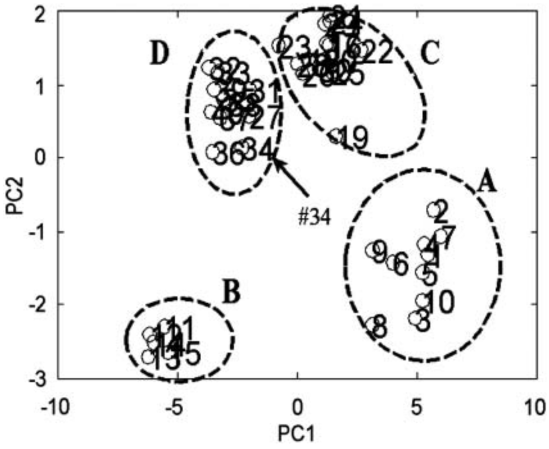

Figure 1: Reference view, PC1-PC2 considering the overall data set. Cluster A: aphids No. 1 to 10; cluster B: aphids No. 11 to 15; cluster C: aphids No. 16 to 26 and 28 to 30; cluster D: aphids No. 27 and 31 to 40. Sample No. 34 should not be included in any group.

other result obtained in the present study. In Figure 1, four more or less well-defined sample groups can be observed according to the four types of aphids defined by skilled biologists. Note that sample 19 is not a true »outlier«, but sample 34 is.

Procrustes Selection of Variables

The application of the Procrustes Rotation (PR) technique gave four essential variables (as pointed out above, we can retain the same number of variables as principal components; here four principal components are needed to describe the data set), namely, variables v5, v12, v14 and v18 (using autoscaled data). In order to visualize the goodness of the selection, Figure 2 presents the PC1-PC2 subspace calculated considering only the four selected variables. It can be seen that the results are really good and that PR seems to »maximize« the differences between the groups and »minimize« the differences within the groups; this is the general result obtained throughout the applications we have made using this technique. Sample 34 corresponds to a very special aphid where anal fold was not found, PR considered it as a »different« sample whilst the original PC1-PC2 did not reveal this point because this variable does not represent a large variance in the data set. On the contrary, when all the variables are considered, sample 19 seems to be a bit different from its neighboring group, which is not true, except for small divergences in some variables. Any other selection carried out containing a larger or a smaller number of principal components (*i.e.*, more or fewer variables) yielded poorer results. It can be argued that the PR technique led to a certain loss of information, as any variable reduction technique would do because the »true« information is only in the overall data set, but note that this technique retains the variables which contain most of the initial information (initial variance) and maintains most of the structure in the data. Many times this is of great importance, since PR avoids selecting variables associated to random error.

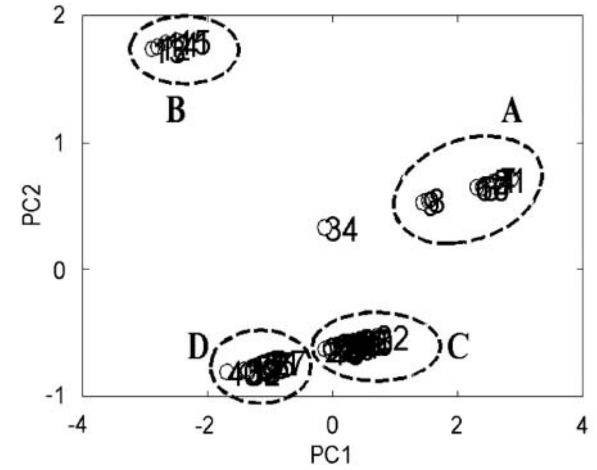

Figure 2: Clusters obtained using four variables after PR (variables 5, 12, 14 and 18 using autoscaled data). Cluster A: aphids No. 1 to 10; cluster B: aphids No. 11 to 15; cluster C: aphids No. 16 to 26 and 28 to 30; cluster D: aphids No. 27 and 31 to 33 and 35 to 40. Sample No. 34 stands alone correctly (see text).

Conditional Exact Test with Ranking According to the Differences between the Number of Wins and of Losses (RD)

Autoscaling for GPCM was not needed, since the method is nonparametric and scaling does not influence the portion of information used. This was verified by conducting several assays with original and autoscaled data. No differences were obtained, as expected.

Table I summarizes the selected variables according to the selection criterion and the ordering method given in the preceding section (see p. 3.). The sum of differences (between the number of wins and of losses) is 75. The selection criterion was defined earlier as 95 %, which means that the variables are selected until the number of 71.25 (=0.95*75) is achieved. The last three variables have equal differences, namely four. These are »uncertain« variables as one of them might be useful (as it will be demonstrated later), but it is unknown which one.

TABLE I. Selection with conditional Fisher's exact test and ranking according to the differences between the number of wins and losses

	Number of wins	Number of losses	Difference	No decision	Ranked by
v8	17	1	16	0	**1**
v14	10	2	8	6	**2**
v12	8	1	7	9	**3**
v1	7	1	6	10	**4**
v2	7	1	6	10	**5**
v13	7	1	6	10	**6**
v3	7	1	6	10	**7**
v9	7	2	5	9	**8**
v7	6	2	4	10	*9*
v4	6	2	4	10	*10*
v5	4	0	4	14	*11*
v10	6	3	3	9	12

In summary, PCM selected 11 variables from among 19 using conditional Fisher's exact test and ranking according to the differences. However, the differences do not decrease uniformly. No doubt that v8 is the best variable, and v14 and v12 are the second and third best ones. Variables v1, v2, v3, and v13 carry the same amount of information, similarly to v7, v4, and v5. Hence, a second PCM has to be performed on the data set of the selected 11 variables. The results are summarized in Table II where v5, v14 and v12 were selected besides v8. Some versatility is inherent to the method as a whole. The smallest subset contains v8 and v5 (Conditional exact Fisher's test and ranking by differences, CEpW). No

TABLE II. Second PCM ranking of the eleven variables selected by the first PCM

	Number of wins	Number of losses	Difference	No decision	Ranked by
v8	10	1	9	0	1
v5	3	0	3	8	2
v14	3	2	1	6	3
v12	1	1	0	9	4
v2	0	1	–1	10	*5*
v13	0	1	–1	10	*6*
v3	0	1	–1	10	*7*
v1	0	1	–1	10	*8*
v7	0	2	–2	9	**9**
v4	0	2	–2	9	**10**
v9	0	2	–2	9	**11**
v10	0	3	–3	8	**12**

doubt that the best variable is v8, when compared to the other eleven variables (v8 won 17 times, so it became superior to the rest of variables 17 times, Table I). But v5 is superior to v8, as shown by the contingency table (Table III) which compares only v5 and v8 (there is only one unique time when v8 loses!). Therefore, v8 can override many unimportant or not especially good variables; see, for instance, the example of variable 13 in Table IV. Variable v8 is clearly superior to v13, but there is no information in the data set to decide between v5 and v13 (Table V). There are other »not useful« variables (*e.g.*, v2, v3, v1, v4), which makes v8 superior, whereas v5 cannot override them (simply because the information portion used does not allow this).

TABLE III. Contingency table for choosing between v5 and v8[(a)]

	Δ(v8)<0	Δ(v8)>0	
Δ(v5)<0	D: 40	B: 0	Ignored: 618
Δ(v5)>0	C: 51	A: 71	
Crit. value		20 v5 won	0<20
α (user)		0.05 α (theor)	6.3E-21

[(a)] A, B, C, and D denote the events – A: both v8 and v5 enhance the correlation; B: v8 enhances, whereas v5 diminishes; C: v5 enhances, whereas v8 diminishes; D: both v8 and v5 diminish the correlation. Δ is the difference according to Refs. 22 and 23.

TABLE IV. Contingency table for choosing between v13 and v8[(a)]

	Δ(v8)<0	Δ(v8)>0	
Δ(v13)<0	D: 116	B: 43	Ignored: 214
Δ(v13)>0	C: 5	A: 402	
Crit. value		18 v8 won	5<18
Δ (user)		0.05 α (theor)	1.9E-09

[(a)] A, B, C, and D denote the events – A: both v8 and v13 enhance the correlation; B: v8 enhances, whereas v5 diminishes; C: v5 enhances, whereas v8 diminishes; D both v8 and v5 diminish the correlation. Δ is the difference according to Refs. 22 and 23.

Table V. Contingency table for choosing between v5 and v13 [(a)]

	Δ(v13)<0	Δ(v13)>0	
Δ(v5)<0	D: 40	B: 0	Ignored: 615
Δ(v5)>0	C: 0	A: 125	
Crit. value		0 Neither won	0>= 0
A (user)		0.05 α (theor)	1

[(a)] A, B, C, and D denote the events – A: both v5 and v13 enhance the correlation; B: v13 enhances, whereas v5 diminishes; C: v5 enhances, whereas v13 diminishes; D: both v13 and v5 diminish the correlation. Δ is the difference according to Refs. 22 and 23.

The v5, v8, v12, and v14 subset is satisfactory, as evidenced by the PC1-PC2 score plot defined when only these variables are considered (see Figure 3). The most remarkable difference from the »true configuration« is that the relative ordering of the groups is lost to some extent and that samples 8 and 9 lie too close to a group to which they do not belong. As for the overall data set, sample 34 is not perceived as different (in opposition to PR).

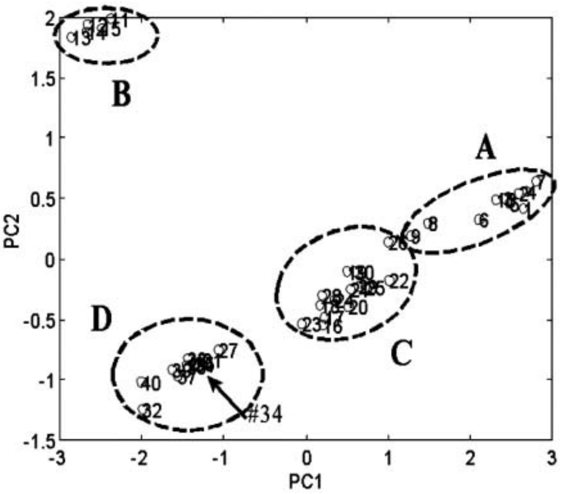

Figure 3: The best four variables achieved using PCM, pre-selection of variables (conditional exact Fisher's test ranking according to differences) and further selection of the pre-selected subset (second PCM): variables 5, 8, 12 and 14. Cluster A: aphids No. 1 to10; cluster B: aphids No. 11 to 15; cluster C: aphids No. 16 to 26 and 28 to 30; cluster D: aphids No. 27 and 31 to 40.

Although a statistically correct criterion is defined for the GPCM variable selection, we do not know in advance how many variables will be retained and it could happen that the amount of selected variables might not necessarily be applicable for a given purpose. For instance, a reduction from 19 to 11 variables was not too large in the aphid data set or, as another example, a drastic reduction in the number of parameters to be measured in an environmental routine control might be needed. This is understandable because the ranking of the variables follows a decreasing similarity. Some diversity of variables is valuable for prediction. If we have enough variables, then a quasi-continuous ranking of similarity can be achieved. Degeneracy can be observed in many practical cases: some of the variables are indistinguishable from each other. In such cases, the farthest variable (from top to bottom, *e.g.*, in Tables I and II) should be retained to preserve the largest portion of diversity between similar variables. Considering the results in Table I, it is easy to find the best diverse subset: v8, v14, v12, v9, and v5, three of them (v5, v12 and v14) selected also by PR.

In fact, the best subset of variables found by GPCM is the one involving the above five variables (Table I). Clearly, this subset gives a satisfactory result when the PC1-PC2 space is considered (see Figure 4). Again, the main criterion was to reproduce the original sample distribution as much as possible; every mathematical distance was defined. Note that although Figure 4 does not reveal sample 34 as anomalous, the »within-group« distribution reveals it closer to the original one than the PR selection.

A comparative summary of the findings observed for these two techniques is given in Table VI.

TABLE VI. Comparison of Procrustes rotation and generalized pair-wise correlation[(a)]

Procrustes rotation	Generalized pair-wise correlation
No previous information is needed about sample grouping.	A dependent (grouping) variable is needed before starting the procedure.
Parametric algorithms, data scaling modifies results (recommended: autoscaling, mean centering).	Non-parametric algorithms, scaling does not affect selections.
The procedure is applied once.	The procedure should be applied, at least twice, successively.
Not rooted to regression techniques.	Originally developed for regression problems, but it is better to rank the variables.
»Within-group« differences are minimized.	»Within-group« differences can be retained.
One well defined selection criterion is used.	Several selection criteria (Conditional Fishers' exact-, McNemar-, Chi square test) ensure some versatility to the method.

[(a)] Both methods have well-defined criteria to decide the number of variables to be retained; both can select individual variables.

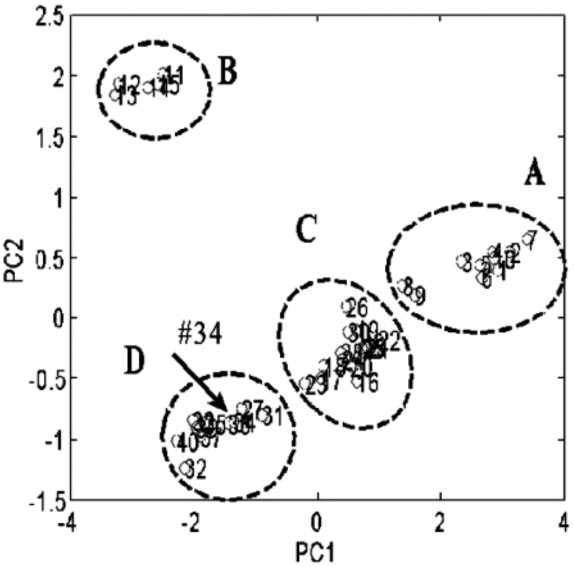

Figure 4: The best subset achieved using PCM, conditional exact Fisher's test preserving the largest portion of diversity, variables 5, 8, 9, 12 and 14. Cluster A: aphids No. 1 to 10; cluster B: aphids No. 11 to 16; cluster C: aphids No. 16 to 26 and 28 to 30; cluster D: aphids No. 27 and 31 to 40.

McNemar Test with Ranking According to the Number of Wins (SR)

Using the previous criteria, only the following variables are excluded in the pre-selection step: v6, v16, v11, and v17. However, an alternative selection criterion is preferred by some authors.[28] The McNemar test is a variant of the sign test; it is often used for decisions of 2 x 2 contingency tables. By applying it here, the same results were obtained as with the previous criteria, and the retained parameters were v8, v14, v12, and v5. From uncertain variables (v2, v13, v3, v1, for which the number of wins equals three and the number of losses is one), the farthest v1 might be selected to preserve the utmost diversity (see Figure 5). Note that this subset is exactly the same as the best one given by conditional Fisher's exact test as selection criterion and, further, up to three retained variables fully agree with those retained by the Procrustes algorithms.

When these results are considered altogether, two conclusions can be drawn: (i) PCM can lead to satisfactory selection of the most important variables to describe the problem at hand; and (ii) several assays need to be performed in order to select the »optimal« variables. These include a second PCM on the already pre-selected subset, and using and comparing various selection criteria and different ranking methods. Moreover, it is worth noting that all PCM subsets contained, at least, three of the variables selected by the Procrustes method. The main exception is that variable 18 (retained in PR) is replaced by variable 8 as this variable has the largest difference between the number of wins and losses.

Variable 5 needs some further consideration, for it turned out to be of the utmost importance for describing

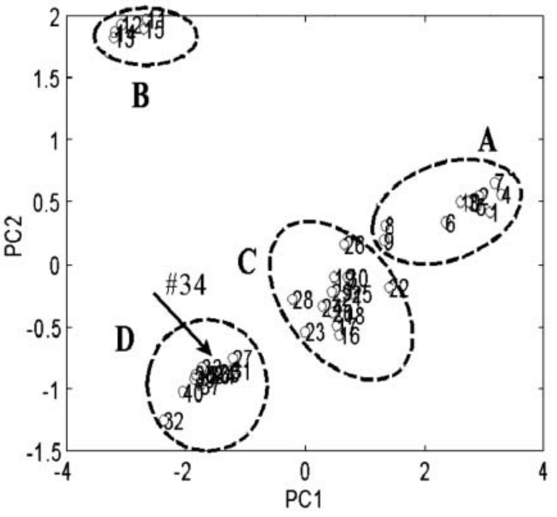

Figure 5: The best subset achieved using PCM, McNemar test with ranking according to the number of wins and preserving maximum diversity: variables 1, 5, 8, 12 and 14. Cluster A: aphids No. 1 to 10; cluster B: aphids No. 11 to 16; cluster C: aphids No. 16 to 26 and 28 to 30; cluster D: aphids No. 27 and 31 to 40.

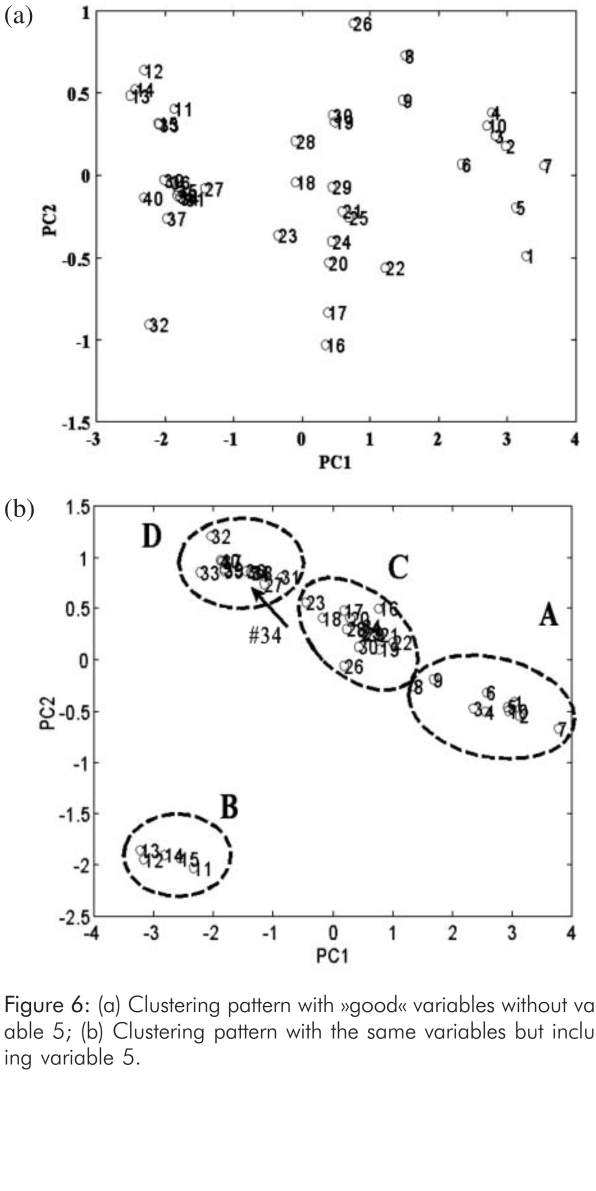

Figure 6: (a) Clustering pattern with »good« variables without variable 5; (b) Clustering pattern with the same variables but including variable 5.

TABLE VII. Comparison of several variable selection methods while selecting the first five »best« variables[(a)]

v5	v14	v4	v16	v15	MLR
v5	v12	v14	v18	v4	PR
v8	v14	v12	v9	v5	GPCM1
v8	v14	v12	v5	v1	GPCM2
v5	v14	v6	v7	v19	PLS1
v14	v3	v1	v15	v13	PLS2

[(a)] MLR: the first five variables by forward selection $p < 6.3$ %. Note that 12 aphids are classified wrongly.

PR: Procrustes rotation using scaled data and four principal components.

GPCM1: best »diverse« subset as described in the text and Table I using conditional Fisher's exact test.

GPCM2: best »diverse« subset as described in the text, (Figure 5) using the McNemars test.

PLS1: with 5 PLS components, regression coefficients are not scaled.

PLS2: with 5 PLS components, regression coefficients are scaled.

the sample group distribution. If we use the information gathered from the usage of PCM, we may select different sets of variables (as shown above). Variable 5 was always included after successive completion of PCM. Although in Table I it was ranked only 11th, it took second place after the second PCM procedure. Considering the results in Table I, it may seem at first glance that the equivalent variables 1, 2, 13, 3 and 9 would be more useful (their difference values are six, whereas v5 has only four). Nevertheless, this does not mean that variable 5 should not be selected; on the contrary, the importance of v5 was shown by the second PCM. It is true that v1, v2, v13, v3 and v9 are superior to v5 as they win over unimportant variables more times than v5. If we consider the second subset (Table II), v5 is superior to the best v8 variable as well (Table III)!

The non-negligible importance of variable 5 is demonstrated in Figure 6. There, an example of a variable subset, where this parameter was deleted, is shown. Note the very poor results obtained under this circumstance (other assays were conducted on similar plots but they are not shown here).

Williams t *Test with Probability Weighted Ordering*

It is interesting to compare the parametric alternative of PCM to various nonparametric tests (conditional exact Fisher's McNemar tests). When the Williams *t* test with probability weighted ordering (WtpW) was assayed, unsatisfactory results were obtained. Up to seven variable subsets were studied, but none of them gave useful PC1-PC2 plots, so it has to be concluded that the subsets derived from this variable selection mode were not good. One reason might be that the Williams *t*-test requires the assumption of normality for all variables, which cannot always be assured in real data sets.

Comparison of Variable Selection Methods

We compared some other »classical« options to select variables related to the highest regression coefficients (absolute value). In this case, for a PLS model with 5 latent variables, the largest regression coefficients were associated with variables v5, v6, v7, v14 and v19 (Table VII) but the sample groups they produced were worse than those displayed for the Procrustes rotation (Figure 2) and GPCM (Figure 4).

The last variable is not significant either in MLR or PR at the 5 % level. The important role of v5 demonstrated by successive usage of GPCM. In many cases, the same variables were selected: v5, v8, v14, v12, are among the most selections.

In is noteworthy that the worst classification was achieved by Multiple Linear Regression by simply predicting the grouping (independent) variable using the five predictor variables: 12 aphids were classified wrongly. It is interesting to note that in the case of PLS the scaling of regression coefficients deteriorated the selection of good (accepted) variables. This selection did not show any grouping.

Acknowledgments. – This work was supported by the Hungarian Research Foundation: OTKA T 037684.

REFERENCES

1. W. J. Krzanowski, *Principles of Multivariate Analysis; a User's Perspective,* (Revised edn.) Clarendon Press, Oxford, 2000.
2. N. R. Draper and H. Smith, *Applied Regression Analysis*, 2nd ed., Wiley, New York, 1981.
3. E. R. Malinowski, *Factor Analysis in Chemistry,* 2nd ed., Wiley, New York, 1991.
4. A. Garrido-Frenich, M. D. Gil-García, J. L. Martínez-Vidal, and M. Martínez-Galera, *Quím. Anal.* **18** (1999) 319–327.
5. F. Lingren, P. Geladi, S. Ränner, and S. Wold, *J. Chemometrics* **8** (1994) 349–363.
6. F. Lingren, P. Geladi, A. Berglund, M. Sjöström, and S. Wold, *J. Chemometrics* **9** (1995) 331–342.
7. U. Norinder, *J. Chemometrics* **10** (1996) 95–105.
8. N. J. Messick, J. H. Kalivas, and P. M. Lang, *Microchem. J.* **55** (1997) 200–207.
9. A. Hõskuldsson, *Chemometrics Intell. Lab. Syst*. **55** (2001) 23–38.
10. H. C. Goicoechea and A. C. Olivieri, *Talanta* **49** (1999) 793–800.
11. R. G. Brereton and A. Elbergali, *J. Chemometrics* **8** (1994) 423–437.
12. J. Ferré and F. X. Rius, *Trends Anal. Chem*. **16** (1997) 155–162.
13. R. Todeschini, D. Galvagni, J. L. Vílchez, and M. del Olmo, N. Navas, *Trends Anal. Chem*. **18** (1999) 93–98.
14. H. Wikel and E. R. Dow, *Bioorg. Med. Chem. Lett.* **3** (1993) 645–651.
15. F. Despagne and D. L. Massart, *Chemometrics Intell. Lab. Syst*. **40** (1998) 145–163.
16. V. V. Kovalysin, I. V. Tetko, A. I. Luik, V. V. Kholodovych, A. E. P. Villa, and D. J. Livingstone, *J. Chem. Inf. Comput. Sci*. **38** (1998) 651–659.
17. D. Jouan-Rimbaud, D. L. Massart, R. Leardi, and O. E. De Noord, *Anal. Chem.* **67** (1995) 4295–4301.
18. A. S. Bangalore, R. E. Shaffer, D. W. Small, and M. A. Arnold, *Anal. Chem*. **68** (1996) 4200–4212.
19. R. Leardi and A. L. Gonzalez, *Chemometrics Intell. Lab. Syst*. **41** (1998) 195–207.
20. A. Carlosena, J. M. Andrade, M. Kubista, and D. Prada, *Anal. Chem*. **67** (1995) 2373–2378.
21. J. M. Andrade, S. Muniategui, P. Lopez-Mahia, and D. Prada, *Fuel* **76** (1997) 51–59.
22. K. Héberger and R. Rajkó, *Discrimination of Statistically Equivalent Variables in Quantitative Structure – Activity Relationships*, in: Fei Chen and G. Schüürmann, (Eds.), *QSAR in Environmental Sciences VII*. SETAC Special Publication Series, SETAC Press, 1997, Pensacola, Florida, Chapter 29, pp. 425–433.
23. R. Rajkó and K. Héberger, *Chemometrics Intell. Lab. Syst.* **57** (2001) 1–14.
24. K. Héberger and R. Rajkó, *SAR QSAR Environ. Res.* **13** (2002) 541–554.
25. K. Héberger and R. Rajkó, *J. Chemometrics* **16** (2002) 436–443.
26. J. N. R. Jeffers, *Appl. Stat*. **16** (1967) 225–236.
27. W. J. Krzanowski, *Appl. Stat*. **36** (1987) 22–33.
28. W. J. Conover, *Practical Nonparametric Statistics*, 2nd ed., Wiley, New York, 1980, Chapter 3.5.

SAŽETAK

Prokrustova rotacija i korelacija po parovima: parametrijska i neparametrijska metoda selekcije varijabli

Károly Héberger i José M. Andrade

U ovom je članku pozornost posvećena odabiru maloga skupa izvornih (neovisnih) varijabli, koji uzima u obzir najvažnije informacije prisutne u matrici podataka. Uspoređena je s praktičnoga stajališta izvedba dvaju već implementiranih metoda: Prokrustov rotacijski algoritam, koji je parametrijski, i korelacija po parovima, koja je neparametrijska. Autori su upotrijebili dobro dokumentirani skup podataka. Obje su metode dale usporedljive rezultate. Prokrustova rotacija je odabrala četiri varijable, dok je korelacija po parovima zadržala četiri ili pet varijabli, ovisno o testu koji je upotrebljen. Obje su metode odabrale tri iste varijable. Stoga su autori zaključili da su obje metode upotrebljive za odabir varijabli. Kriterij odabira rangiranja varijabli pomoću korelacije po parovima je proširen tako da zadrži najveću moguću različitost među varijablama.