Małgorzata Ćwiklińska-Jurkowska
Department of Theoretical Backgrounds of Biomedical Science
and Medical Informatics, Collegium Medicum in Bydgoszcz,
Nicolaus Copernicus University

EXPLORATORY DATA ANALYSIS FOR THE HEMATOLOGICAL FEATURES. PART II. APPLICATION

Abstract: Part II contains numerical and graphical exploratory analysis results and elements of their medical interpretation for the set of the hematological observations and hematological variables. On the basis of correlation matices, matrix of scatterplots with overlaid regression lines and two and three-dimensional biplots relationships between parameters of blood during hemoperfusion is examined. For comparison purposes also MDS and one-way and two-way cluster analysis are performed. Usefulness of applied methods of multivariate data ordination to inspect, e.g. variables' interdependencies was assessed. Applied methods gave very close results and the medical interpretation of the results confirm some physiological clotting ideas. The practical results confirm some hypothesis describing polymer-blood interactions. Additionally, the results of principal factor analysis and multidimensional metric scaling with cluster analysis are concordant. The variety of applied exploration data methods confirm results and give the possibility of looking at data from different point of views.

Introduction

The aim of the study was an investigation of the structure of the hematological data by examining some multidimensional exploratory data analysis methods such as scatterplot matrices and biplots (Gabriel 1971, 1981, 1990, Krzanowski 1988, 1995) and exploratory data analysis like factor analysis (mainly by PCA), MDS and cluster analysis.

A scatterplot matrix enables examination of relations between pairs of variables by graphical representation of a data matrix, while classical linear biplots are based on principal component analysis. Biplots are simply the scatter plots of multidimensional data into two or three dimensions with the superimposition of the variables. Thus, the relationship between the hematological data and variables can be investigated. The methodology of biplots and ither multivariate ordination methods is presented and the application on hematological data set is performed in Part I. "Methodolgy".

ISBN 978–83–7431–266–0 ISSN 0860-150X

Possible multivariate associations of numerical variables are examined in the work.

Aim

Exploring multivariate relationships between hematological data set is examined in the paper containing an interpretation from the medical point of view.

For drawing the plots and obtaining numerical results the program Statistica for Windows, SPPS and package R were used.

Illustration of multidimensional dependencies for hematological variables

The description of the examined hematological features is the following: "ADHEVIN" – Level of platelets' adhesion; "TIME" – time of perfusion ("CZAS"); "CZFIVIN" – Time of fibrinolise; "ERYTVIN" – Number of erythrocytes; "FIBGVIN" – Fibrinogen concentration; "KAOLVIN" – Kaolin-kefalin time; "KEFAVIN" – Kefalin time; "LEUKVIN" – Leukocytes number; "PROTVIN" – Prothrombin time; "STYPVIN" – Stypven-kefalin time and "TROMVIN" – Thrombocytes number.

Exploratory data analysis methods were applied (Bartkowiak et. al, 1996, Bartkowiak, 1995, Krzanowski W. J., 1988, 1995, Krishnaiah 1977, Larose D. T. 2005). Visualizing of interdependencies between variables describing parameters of human blood in experiments in vitro are presented in this section. The diagnostic tools for model of two-way tables are applied. Two kinds of exploratory data analysis are performed: firstly, scatterplot matrices and then biplots. Examining the data using scatterplot matrices is not in fact multidimensional analysis (only combing two-dimensional analysis), but can provide the insight into multidimensional data, if the dimension is reasonable.

The full matrix (Tab. 1) of correlations between the considered variables was calculated and the corresponding scatterplot matrix was obtained (Fig. 1). From Person correlations coefficients computed for all possible pairs of 11 variables we confirm the scatterplot matrix presented on Fig. 1. Looking at the obtained scatter matrices for all possible pairs of 11 variables we can verify that there is a positive significant ($p < 0.05$) Pearson correlation between the variables ADHEVIN and STYPVIN (0.41), TIME and PROTVIN (0.6), ERYTVIN and TROMVIN (0.77), FIBGVIN and TROMVIN (0.56), KAOLVIN and KEFAVIN (0.90), KAOLVIN and

Table 1

Linear correlation coefficients between pairs of variables

	ADHEVIN	TIME	CZFIVIN	ERYTVIN	FIBGVIN	KAOLVIN	KEFAVIN	LEUKVIN	PROTVIN	STYPVIN	TROMVIN
ADHEVIN		−0.0494	−0.0640	0.0056	0.2123	−0.3046	−0.0698	**−0.4893**	0.1382	**0.4099**	0.2254
TIME	−0.0494		0.0541	**−0.6061**	**−0.5317**	−0.1603	−0.2084	**−0.6659**	**0.5994**	0.0662	**−0.7435**
CZFIVIN	−0.0640	0.0541		−0.1619	0.0724	0.1764	0.2772	−0.0493	−0.2072	0.0082	−0.1210
ERYTVIN	0.0056	**−0.6061**	−0.1619		0.1501	**−0.4541**	**−0.4322**	0.1696	−0.2368	0.3114	**0.7685**
FIBGVIN	0.2123	**−0.5317**	0.0724	0.1501		0.1882	0.3240	0.2867	**−0.3966**	−0.1388	**0.5610**
KAOLVIN	−0.3046	−0.1603	0.1764	**−0.4541**	0.1882		**0.8992**	**0.4616**	**−0.4732**	−0.1933	**−0.4380**
KEFAVIN	−0.0698	−0.2084	0.2772	**−0.4322**	0.3240	**0.8992**		0.293	**−0.5567**	−0.1996	−0.3174
LEUKVIN	**−0.4893**	**−0.6659**	−0.0493	0.1696	0.2867	0.4616	0.2930		−0.3160	**−0.5511**	0.2289
PROTVIN	0.1382	0.5994	−0.2072	−0.2368	**−0.3966**	**−0.4732**	**−0.5567**	−0.316		−0.1274	−0.3574
STYPVIN	**0.4099**	0.0662	0.0082	0.3114	−0.1388	−0.1933	−0.1996	**−0.5511**	−0.1274		0.1342
TROMVIN	0.2254	**−0.7435**	−0.1210	**0.7685**	**0.5610**	**−0.4380**	−0.3174	0.2289	−0.3574	0.1342	

In bold: typed significant correlations on level 0. 05

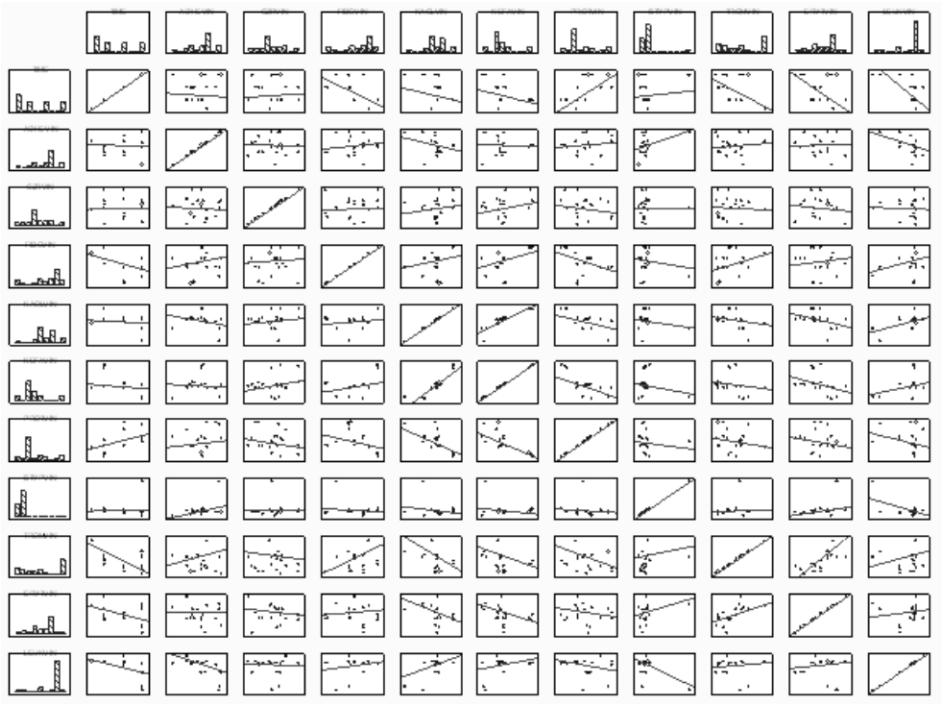

Fig. 1. Scatterplot matrix for all of variables' pairs with overlaid regression lines

LEUKVIN (0.46). Additionally there is a negative significant ($p < 0.05$) correlation between the pairs of variables ADHEVIN and LEUKVIN (– 0.49), CZAS and ERYTVIN (– 0.61), CZAS and FIBGVIN (– 0.53), TIME and LEUKVIN (– 0.67), CZAS and TROMVIN (– 0.74), ERYTVIN I KAOLVIN (– 0.45), ERYTVIN and KEFAVIN (– 0.4322), FIBGVIN and PROTVIN (– 0.40), KAOLVIN and PROTVIN (– 0.47), KAOLVIN and TROMVIN (– 0.44), KEFALVIN and PROTVIN (– 0.56), LEUKVIN and STYPVIN (– 0.55), LEUKVIN and PROTVIN (– 0.32) and finally KEFAVIN and PROTVIN (– 0.56).

The problems of in-vitro blood procedures are generally caused by changes of thrombocytes in time (TIME significantly negatively correlated with TROMVIN, $r = -0.74$), so the row connected with variable TROMVIN (number of thrombocytes) is most interesting. The highest correlated pair is visible in scatterplot matrix (TROMVIN and ERYTVIN, $r = 0.77$).

For pairs of variables correlated significantly positively (ADHEVIN and STYPVIN, TIME and PROTVIN, ERYTVIN and TROMVIN, FIBGVIN and TROMVIN, KAOLVIN and KEFAVIN and KAOLVIN with LEUKVIN) some of these relationships are self-evident. However, the other can find explanation from a point of view of clotting physiology. For example, they suggest an activation of clotting factors during the experiment (variables KAOLVIN and KEFAVIN), activation of clotting factors (variables KAOLVIN and KEFAVIN), consumption of fibrinogen (FIBGVIN) and segmentation of blood cells on polymer (variables TROMVIN and ERYTVIN). Significant positive correlation of a durations of experiment (TIME) and prothrombin time (PROTVIN) perhaps shows on catching on sorbent surface some clotting factors, active in extrinsic clotting pathway. Universally, one thinks that the prothrombin time changes insignificantly during a contact of blood with polymers, but by reason of considerable development of sorbent surface these interactions can be more clear. It can point to the advisability of taking into account the prothrombin time in estimation of hemocompatibility of polymers (foreign surfaces), which are planned to apply in clinical practice.

Next, for the following different pairs correlated significantly negatively (ADHEVIN and LEUKVIN, TIME and ERYTVIN, TIME and FIBGVIN, TIME And LEUKVIN, TIME and TROMVIN, ERYTVIN and KAOLVIN, ERYTVIN and KEFAVIN, FIBGVIN and PROTVIN, KAOLVIN and PROTVIN, KAOLVIN and TROMVIN, KEFALVIN and PROTVIN, LEUKVIN and STYPVIN and KEFAVIN with PROTVIN) some explanation from the physiological clotting point of view can be given. Those results for the times assessing intrinsic system of clotting (KAOLVIN and KEFAVIN) may point

to the activation of clotting factors, taking participation in the beginning of the activation. Simultaneous with the activation decreases also the number of plates of blood (thrombocytes)-TROMVIN correlated significantly negative with KAOLVIN).

Obtained results and the medical interpretation confirm some hypothesizes about interactions of polymeric sorbent with blood (Kao W. J et. al 1996, Lane et. al 1994, Lim et. al 1994).

Examining of the structure of the data by factor analysis

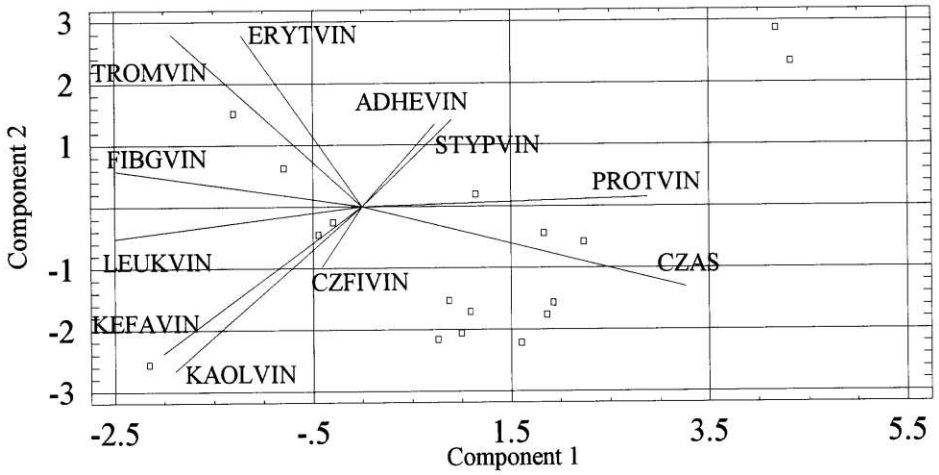

Fig. 2. Two-dimensional biplot for 11 variables

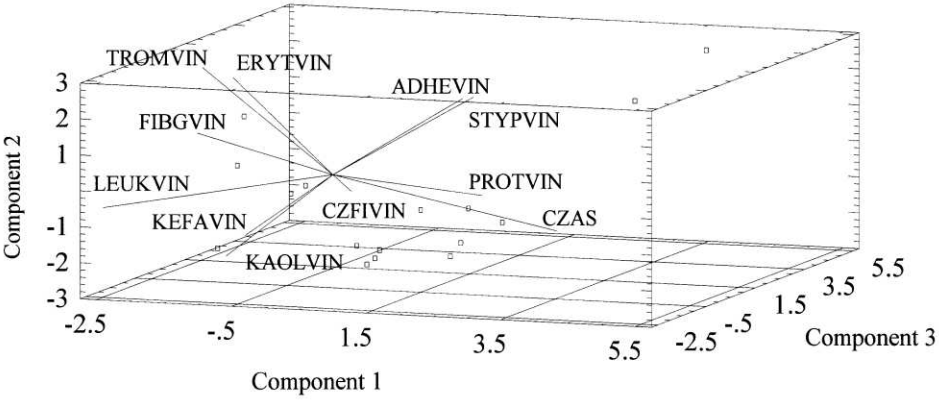

Fig. 3. Three-dimensional biplot for 11 variables, truly reproducing surface, on of which a two-dimensional biplot is presented

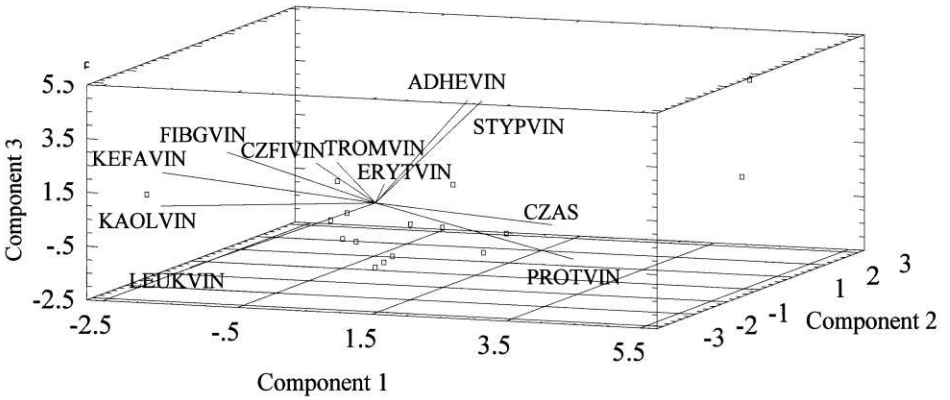

Fig. 4. Eigenectors-principal component coefficients for 11 original variables

The statistical graphics of biplots represent in the same plane both the variables and the cases. Eleven or six variables are represented by arrows (lines), while points represent cases. See the examples in Fig. 2–7. Graphs show visual inspection of main information from the data set. Presented at the biplots with principal components ((Fig. 2–7) vectors (visible as segments from the centroid) represent the original variables. The length of each vector is proportional to the contribution of the corresponding variable in the principal components. The angle between any two vectors is closely related to the correlation between presented variables. Cosine of this angle is a correlation in the case of total variance representation on biplot – then a high positive correlation is achieved for small angles (close to 0 degrees) and the high negative correlation, for angles between the vectors representing the variables, which are close to 180 degrees. Based on this angle, you can draw reliable conclusions about the correlation, if the original variables are well represented on the biplot (goodness of fit). If the vectors are short, they can not be applied to conclude about the correlations (Bartkowiak 1995).

In Fig. 2 and Fig. 5 the two-dimensional biplots constructed from the examined variables are presented. However, the biplot shown in Fig. 2 is based on all variables (11); the biplot shown in Fig. 5 was constructed using 6 chosen variables. Both groups of biplots were constructed from correlation matrices. Commonly, the points and the vectors in the biplot plane represent projections from the multivariate space onto the plane of the first two or three principal components. Points at biplots mean the individual observation boxes ($n = 32$), some of them overlap. From biplots for all 11 variables (Fig. 2–4 and Tab. 1) it can be noted that many subsets of variables

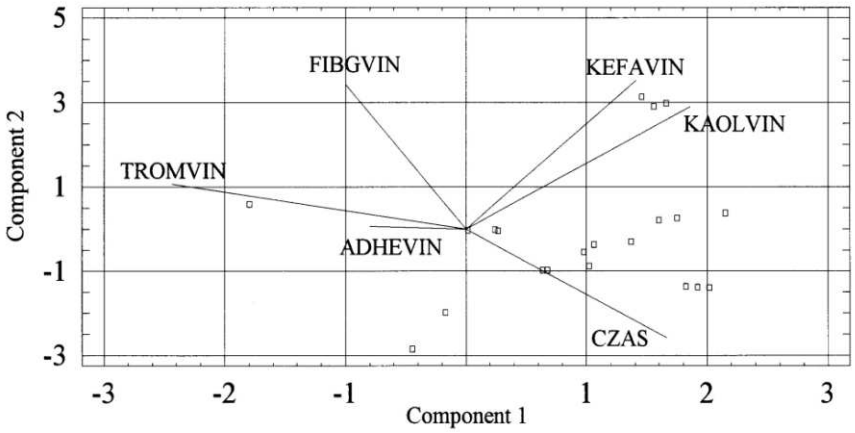

Fig. 5. Two-dimensional biplot for 6 selected variables

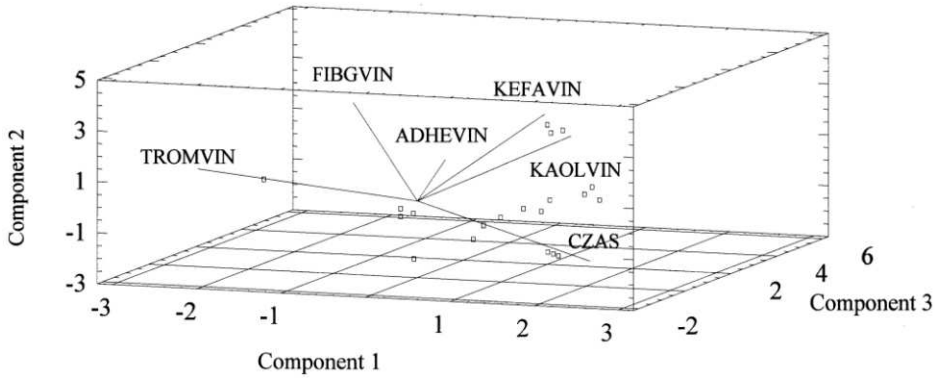

Fig. 6. Three-dimensional biplot for 6 selected variables, truly reproducing surface, on which two-dimensional biplot is presented

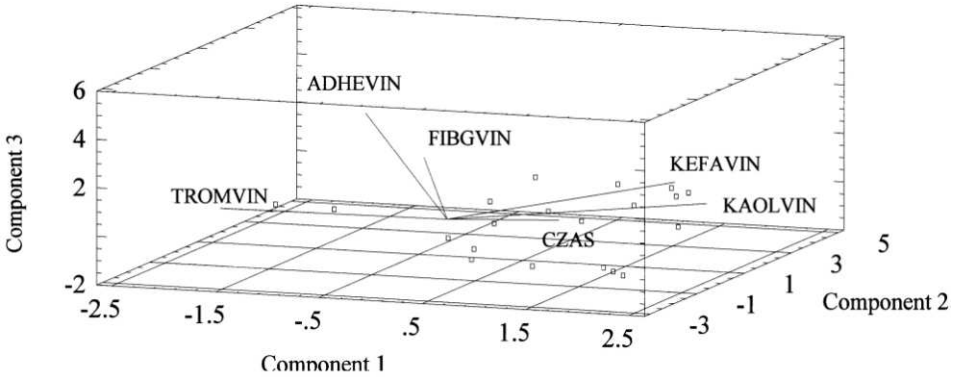

Fig. 7. Three-dimensional biplot for 6 variables, presenting deviation of vectors from plane of two first principal components

Table 2

Principal component analysis – set of 11 original variables

Number of principal component	Eigenvalue	Total variance percentage	Cumulated variance percentage
1	3.39581	30.871	30.871
2	3.06543	27.868	58.738
3	1.66496	15.136	**73.875**
4	1.04753	9.523	83.397
5	.88566	8.051	91.449
6	.42797	3.891	95.340
7	.29461	2.678	98.018
8	.15528	1.412	99.430
9	.04106	.373	99.803
10	.01675	.152	99.955
11	.00495	.45	100.000

Table 3

Eigenvectors-principal component coefficients for 11 original variables.

Variables	Eigenvectors			
	1	2	3	4
ADHEVIN	.106616	.22402	.50364	.49271
TIME	.474063*	–.22017	.03030	–.01168
CZFIVIN	–.0582577	–.16444	.28605	–.32623
ERYTVIN	–.176689	.46696*	–.11107	–.33431
FIBGVIN	–.361119	.09811	.20069	.48914
KAOLVIN	–.2748	–.44697*	.13184	–.05471
KEFAVIN	–.292155	–.39982*	.29908	.07581
LEUKVIN	–.408985*	–.09740	–.42429	.01627
PROTVIN	.417988*	.02436	–.27440	.33963
STYPVIN	.131751	.23607	.49885	–.41092
TROMVIN	–.280475	.47030*	–.00180	.08017

Variables with biggest contribution to eigenvectors are marked by asterisks

are highly correlated with each other. The question arises whether you can choose a smaller and more representative set of characteristics, perhaps dropping the features strongly correlated. In many publications as the most important promoter of hemocompatibility (compatibility with the blood) the effect on platelets is given. Therefore, as the first into a created subset

Table 4

Principal component analysis – set of 6 selected variables

Number of principal component	Eigenvalue	Total variance percentage	Cumulated variance percentage
1	2.37832	39.639	39.639
2	2.00856	33.476	**73.115**
3	1.02162	17.027	**90.142**
4	.438124	7.302	97.444
5	.107607	1.793	99.237
6	.045767	.763	100.00

Table 5

Eigenvectors-principal component coefficients for set of 6 selected variables

Variables	Eigenvectors		
	1	2	3
ADHEVIN	–.202265	.0105427	.915152
TIME	.418215	–.405896	.316225
FIBGVIN	–.251145	.538354*	.191984
KAOLVIN	.466875*	.457417*	–.023692
KEFAVIN	.354357	.554912*	.087172
TROMVIN	–.614468*	.167794	–.132212

Variables with biggest contribution to eigenvectors are marked by asterisks

of variables, the variable TROMVIN (thrombocytes) was selected. Next, the set has additionally been extended by five other features which found a relatively high representation at biplots (Table 4) – the cumulative percentage of variance (compared to the other six sets of original variables which contain variable TROMVIN). So the analysis of principal components and also for those biplots based on subset of six features was developed (Tab. 4, 5, Fig. 5–7). Biplots if Fig. 5–7 represent both all observations and variables subset containing 6 original features subset of a full multivariate data set on the same plot.

In the column "*Eigenvalue*" in Tables 2 and 4, the eigenvalues equal to the variances on the consecutive factors may be obtained. In the second column "Total variance percentage ", these values are expressed as a percent of the total variance (sum of eigenvectors). As we can notice from Tab. 2,

factor 1 accounts for 31% of the variability, factor 2 for 28%, factor 3 for 15% and so on. The third column consists of the cumulative variability extracted.

From Tab. 4 we can see that the first axis for 6 original variables explains 36% of variability of the whole multidimensional data set.

Analyzing the eigenvalues in Tab. 2 it can be stated that the presentation of data on a matrix of the first three principal components reproduces 73% of the total variation of the original six variables. Further, in Tab. 4 can see the better (in comparison to Tab. 2) result – 90% of the variability of representation after a three dimensional projection of the whole multidimensional data set into principal component subspace. It can be explained by a lower dimensionality of the reduced input data set (maybe some information is lost by dropping a number of original variables, from eleven to six-which almost half of them). From the tables of coefficients for principal components (Tab. 2 and 4) it can be easily determined which features have the most influence on another principal component. This is reflected in the illustration in the appropriate biplot (Fig. 2–7).

Scatterplot in the Fig. 5 is a projected multivariate scatter onto a plane but on the subset of 6 variables. The lines represent six original variables. Each case is represented by one point (individual data point) on the same principal variables axes.

For both 11 original variables and 6 original variables the scree Cattel criterion coming from showing subsequent eigenvalues indicate to bigger number of dimensionality than the Kaiser criterion, i.e. the number of eigenvalues bigger than 1 (which is equal to four in the case of 11 input variables and three in the case of 6 input variables).

According to the Kaiser criterion we an see that 4 variables have eigenvalues bigger than 1.

The approximation of the biplot variables is given by the biplot axes. The vectors labeled by names represent the considered variables. Table 4 shows the equations of the principal components. The most important variables are marked by characters "*". For example, the first principal component has the following equation

$$0.106616\ \text{ADHEVIN} + 0.474063\ \text{CZAS} - 0.0582577\ \text{CZFIVIN} - 0.176689\ \text{ERYTVIN} - 0.361119\ \text{FIBGVIN} - 0.2748\ \text{KAOLVIN} - 0.292155\ \text{KEFAVIN} - 0.408985\ \text{LEUKVIN} + 0.417988\ \text{PROTVIN} + 0.131751\ \text{STYPVIN} - 0.280475\ \text{TROMVIN}$$

where the values of the variables in the equation are standardized by subtracting their means and dividing by their standard deviations. For the principal components based on the set with all eleven variables the first

principal component can be defined as related to the activation of the extrinsic system. Important contribution to the value of this component also contends that the number of platelets, leukocytes and blood contact time with the sorbent. The second principal component is related to the activity and the number of thrombocytes and coagulation parameters intrinsic (Tab. 3).

Table 5 shows the equations of the principal components. For example, the first principal component has the following equation

$$0.202265\ \text{ADHEVIN} + 0.418215\ \text{CZAS} - 0.251145\ \text{FIBGVIN} + 0.466875\ \text{KAOLVIN} + 0.354357\ \text{KEFAVIN} - 0.614468\ \text{TROMVIN}.$$

However, in the case of 6 selected features, the above first principal component is associated with the number of platelets, the other components of the plasma coagulation. In this case, for both first principals the important component of variability is the time of the experiment (Table 5). Principal component analysis confirms the suitability of the number of platelets as an important parameter to measure the impact of polymer on the blood. The different signs in the equation (signs of coefficients of loadings) mean distinct contribution into the principal component. It is worth noting a difference for biplots (either two or three-dimensional and) in a configuration relative to each other vectors representing eleven features and only six selected variables characteristics. It is of course the fact that both principal component analyses are taken into account other data matrix X: respectively *nx6* and *nx11* matrix ($n = 32$). Thus biplots from Fig. 5–7 contain only some columns of the matrix *nx11*. In addition to the configuration vectors (features), also the configuration of points (observations) on biplots for 11 variables (Fig. 2–4) and six variables (Fig. 5–7) are different, since they reproduce on biplots significantly different percentage of variability (Table 2 and 4).

Comparing pairs of correlated variables (Tab. 1) with biplots for all 11 variables, an adequacy of biplots for these variables to a certain extent can be observed, though only on a three-dimensional biplot the total variation is represented satisfactorily, in 74% (Table 3). The adequacy of the correlation matrix is more obvious for 6 variables, because in this case reconstruction of variability on biplots is 73% and 90% for two-dimensional and for three dimensional, respectively (Tab. 4). It is worth noting that in the case of a large percentage of the representing of the of variation (like fig. 5–6), a large absolute value of negative correlation is related to the angle between vectors denoting variables close to 180 degrees, and the large positive correlation value means the angle close to 0 degree.

Comparing the biplot 2 and 3-dimensional for the 11 variables it may be observed that the worst representation in the two-dimensional biplot has the CZFIVIN variable (Fig. 2 and 3), and for the six variables has ADHEVIN feature (Fig. 5 and 6). These vectors' characteristics significantly deviate from the plane of the first two principal components. It is seen by selecting in displaying the biplot the third major component parallel to the edge of the graph (Fig. 4 and 7). This is confirmed by the analysis of the third column in Tables 3 and 5. In these tables the high coefficients are found for the relevant variables in comparison with the coefficients in the same row for the primary and last components (first and second column compared with third and/or fourth).

Despite the small angle between variables and TROMVIN ADHEVIN on biplot for the original six variables (Fig. 5) and a small angle between variables and KAOLVIN CZFIVIN biplot for the original 11 variables (Fig. 2), correlations between these variables are not large ($r = 0.23$ and 0.18 respectively), because – as stated above – the representation of the CZFIVIN and ADHEVIN variables is weak. More generally, for biplot devised for standardized features, namely the correlation matrix, no correlations can be inferred if the vectors' features are shorter (Bartkowiak 1995).

Comparing of both cases (i.e. for 11 and 6 variables) on two and three-dimensional biplots (Fig. 2 and 3 and 5 and 6), the closer representation of actual data matrix X of a three-dimensional biplots by the two-dimensional biplots on the plane it can be held for six variables. The diminishing of variability representations from matrix for 11 and 6 variables is 15% and 17%, respectively.

In plot on Fig. 8 we can see which variables have smaller length and therefore are nor well represented in the plane (e.g. CZFVIN, ADHEVIN, STYPVIN).

It is interesting if inspecting the biplots and the interpretation of factor analysis results give comparable results. Other multivariate data ordination, namely principal factor analysis is also performed. The principal factor analysis for the whole input data set is presented graphically in Fig. 8. In the plot of factor loadings in Fig. 8 eleven variables were reduced to two specific factors. Fig. 9 shows similar results with the nonessential difference, coming from the arrangement of points. Next, Fig. 10 is obtained with Varimax rotation – difference with Fig. 10 is caused by the rotation. Additionally, the variables except the points are visualized in comparison with biplots. Corresponding Cattel scree plot is visible in Fig. 11. In graphical method for the *scree* Cattell test we can see the eigenvalues shown in a simple line plot, the values are numerically presented also in the first column of Tab. 2.

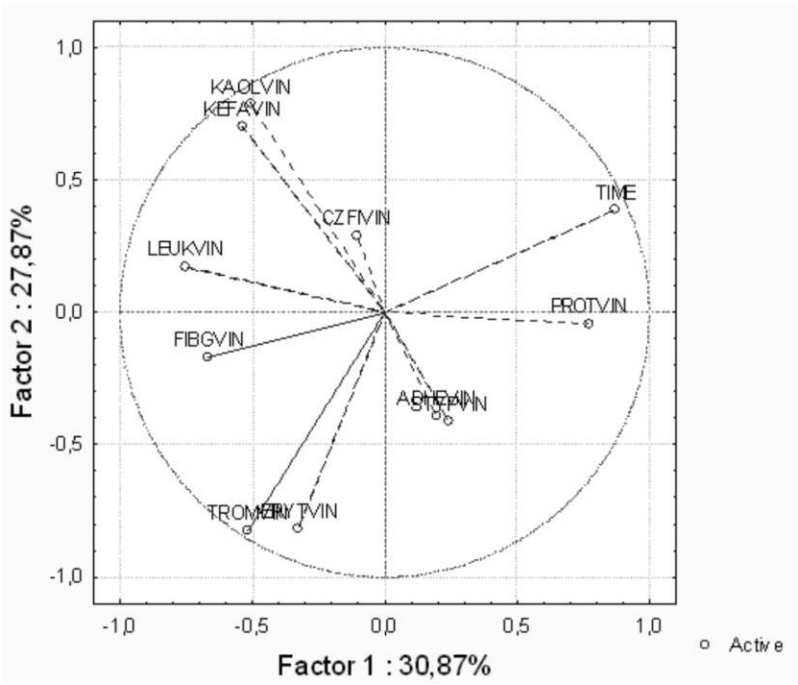

Fig. 8. Projection of variables on Factor1xFactor2 plane

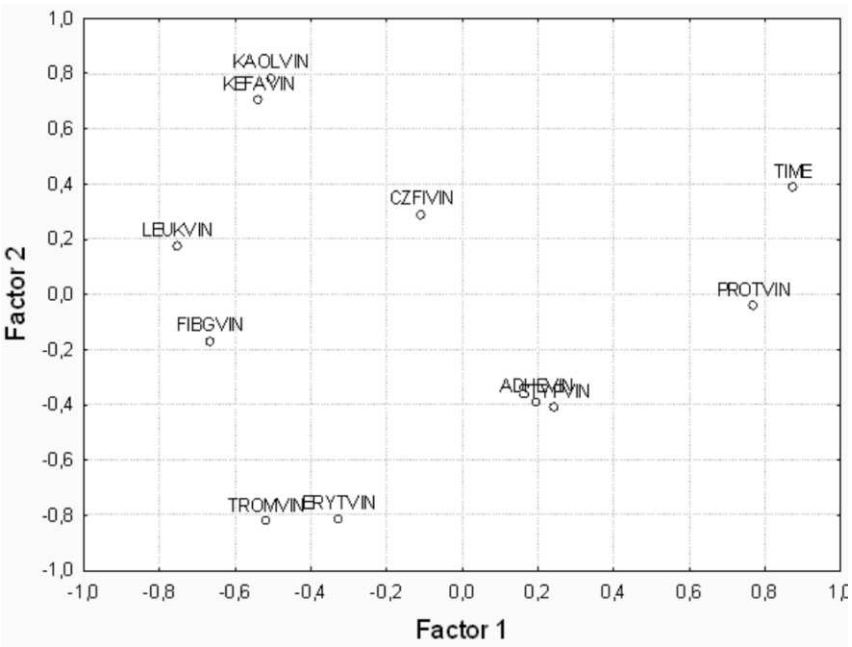

Fig. 9. Principal components without rotation

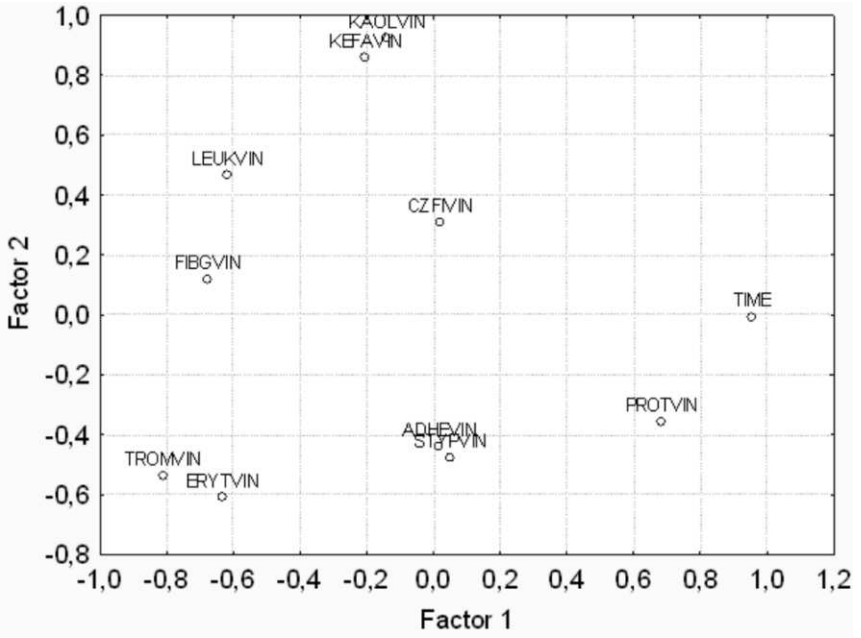

Fig. 10. Principal components with Varimax rotation

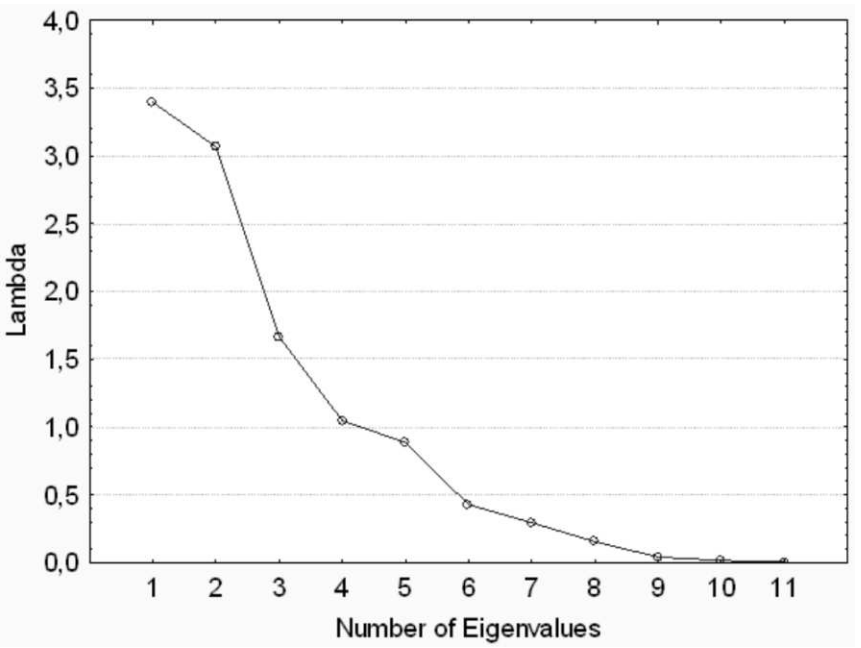

Fig. 11. Cattel scree plot

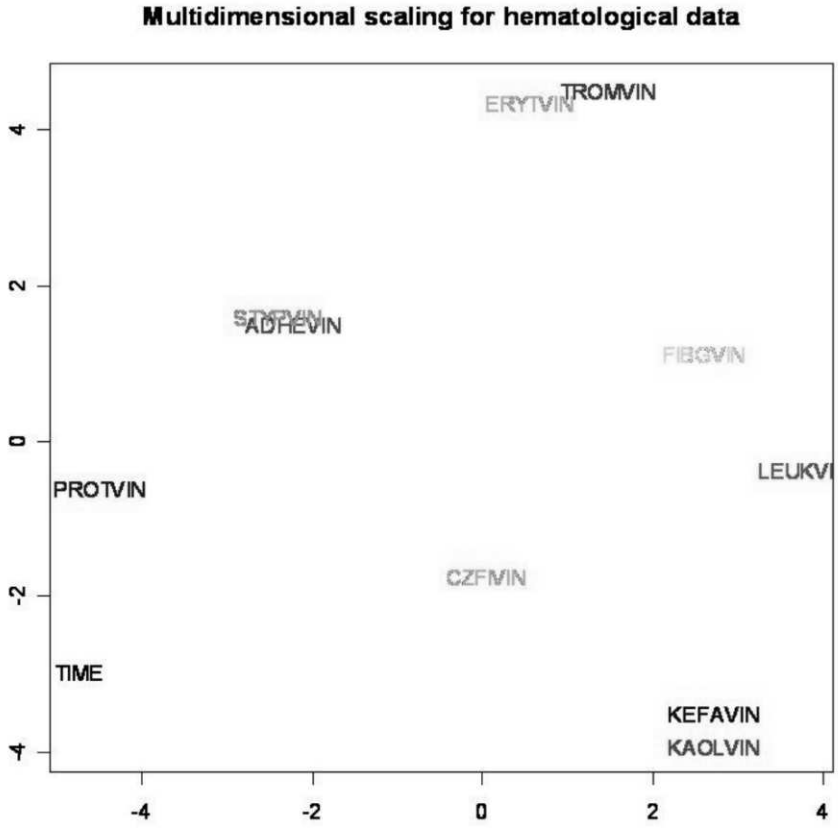

Fig. 12. MDS for all 11 variables

The Cattel scree plot showing subsequent eigenvalues indicate to bigger number of dimensionality representation (equal to 6) than the Kaiser criterion, i.e. the number of eigenvalues bigger than 1 (which is equal to 3).

Another multidimensional data ordination method is visible in Fig. 12 and 13. It is the multidimensional metric scaling result for the Euclidean distance.

In Fig. 12 the "rearranged" hematological features in a proficient way are presented. The obtained configuration best approximates the observed Euclidean distances. For other applied distances or for correlation dissimilarity measure the MDS results (not presented in the paper) are nearly the same as for the Euclidean distance. The diminished set of 11 original variables into its subset of 6 variables, the same as in Tab. 4 and 5 and on Fig. 5–7, are arranged by MDS into the plot in Fig. 13.

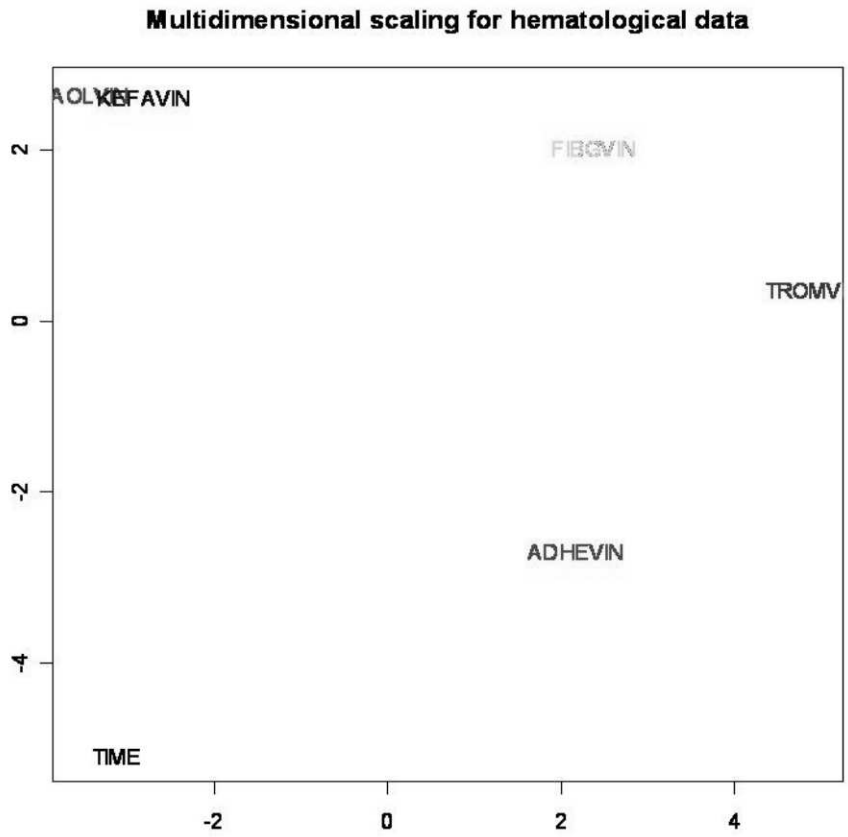

Fig. 13. MDS for 6 variables: "TIME" "ADHEVIN" "FIBGVIN" "KAOLVIN" "KEFAVIN" "TROMVIN"

The arrangement of variables in MDS for 11 and 6 variables is very close to those obtained earlier and described above. MDS is an alternative to factor analysis. However, MDS and factor analysis are basically different methods, though the type of research tasks to which these two techniques can be applied are similar. For example, MDS does not require normality and linearity such restrictions. Moreover, MDS can be applied to any kind of distances or similarities, while factor analysis is based on the covariance matrix.

To discover structures on the basis of distances between variables, grouping of records into groups of similar objects is performed by cluster analysis. In the obtained subsets (in clusters) the similarity of the records is maximized and the similarity of the records in other clusters is minimized.

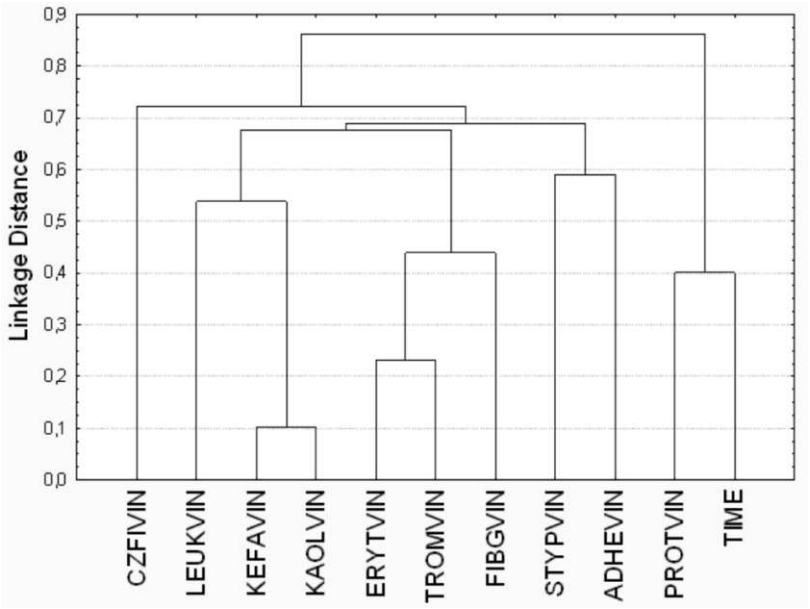

Fig. 14. Hierarchical single linkage clustering by Euclidean distance for whole data set

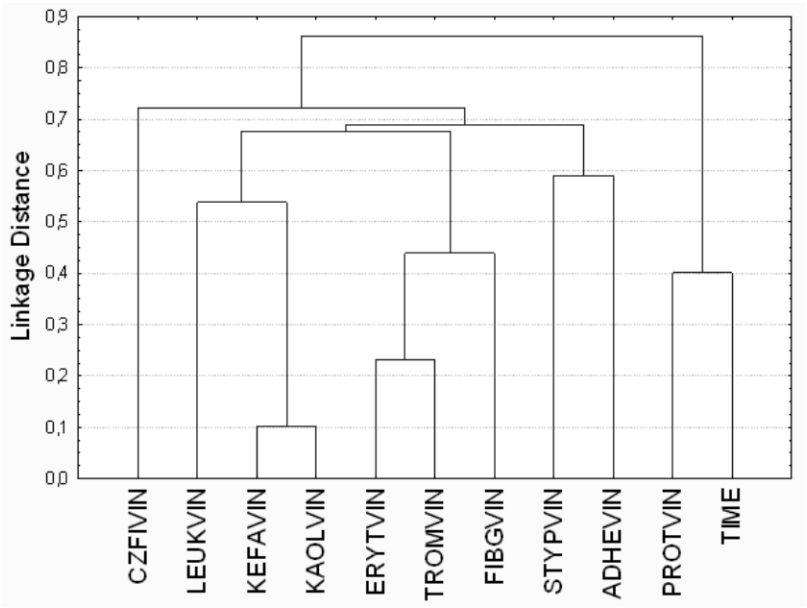

Fig. 15. Hierarchical single linkage clustering by Pearson coefficient for whole data set

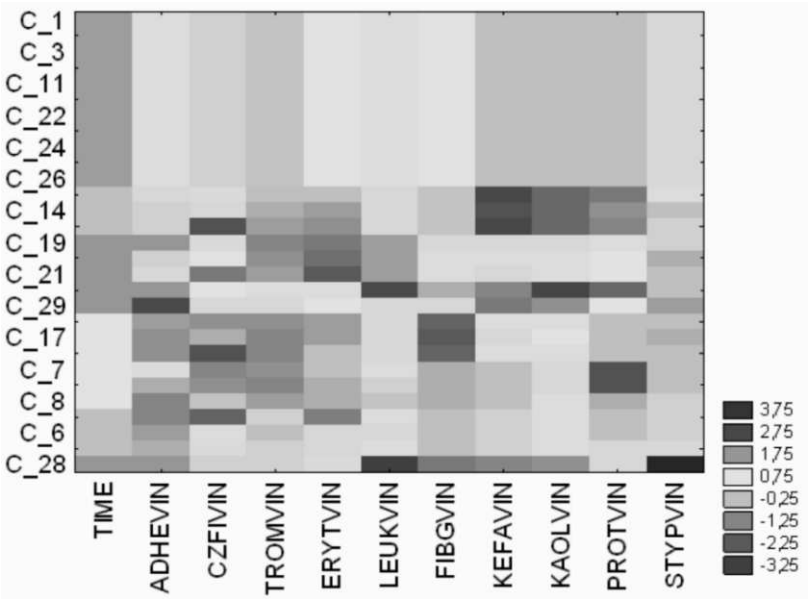

Fig. 16. Cluster two-way joining for hematological features and observations

The applied hierarchical cluster analysis for complete linkage and Euclidean distance is presented on Fig. 11. Other distances gave the same results (for example dissimilarity equal to Person correlation coefficient – Fig. 12). The clustering of variables shows the similarity to other multidimensional reduction methods results. Clusters for two-way joining are presented in a "heat map" (Fig. 13). The dissimilarity of variables is close – the columns in two-way joining plot are strictly related to the tree coming from one-way hierarchical clustering.

Different multivariate methods applied, like the factor analysis (PCA and principal factors), correlations, MDS and clusters gave similar or complementary results.

In all applied methods a low-dimensional graphic representation of the hematological data set is obtained. Relationships obtained by the different ordination methods confirm the dependencies obtained by each other with only subtle differences.

Final remarks

In biomedical problems the question arises how to categorize observed data into meaningful structures. Generally, medical problems are characte-

rized by high complexity. When one wants to describe medical phenomenon, large number of variables is needed. However, with high dimensional tasks the interpretation is difficult. Therefore, the reduction of the dimensionality is needed. Two groups of dimensionality exists (selection or extraction), here we applied the extraction reduction of dimensionality. Grouping of the variables is obtained by similar methods of PCA and factor components methods. If, after dimensionality reduction, one has obtained only two or three dimensions, a physician can interpret the problem using appropriate illustration.

In exploratory data analysis we can examine higher dimensional data sets, where the relationships or trends are difficult to see. The insight into multidimensional data is possible looking at the same time at many graphs using scatteplot technique. However, for bigger number of variables constituent plots are smaller. For p-dimensional data sets one can not simply visualize the whole information coming from data, so the need of representative ordination is needed. The possibility of analyzing relationships between many variables on only one plot gives the two or three-dimensional biplot technique. However, then the representative features in the reduction of dimensionality connected with the eigenvalues of the matrix in principal component is needed.

The research question of interest is usually expressed in terms of both cases (observations) and variables. For example, the biplots or two-way clusters may be applied. Applied explanatory data analyses methods can discover structures in data, however this does not automatically supply an explanation or interpretation.

Conclusion

The relationship between the hematological data and variables can be investigated by graphic representation: scatterplot matrices and biplots. Two-dimensional or three-dimensional biplots is a method of dimensionality reduction giving the possibility of observing simultaneously both variables and observations, so observations may be also visualized in the context of many variables. Useful maps are obtained by multidimensional metric scaling, which is also the ordination multivariate data method, which do not need the assumptions related to PCA. The results confirm some hypothesis describing polymer-blood interactions and may suggest new unknown facts.

The results of factor analysis, cluster analysis and two-way clusters and multidimensional scaling are all concordant.

REFERENCES

[1] Bartkowiak A. Liebhart J. Szustalewicz A. 1996. Visualizing the correlation structure by a biplot extended to 3 dimensions. 34th International Center of Biocybernetics. Seminar. Statistics and Clinical Practice. Warsaw, 24–28 June, 1996. pp. 50–52.

[2] Bartkowiak A. Lisp-Stat. Narzędzie eksploratywnej analizy danych. In Polish. Uniwersytet Wrocławski. 1995.

[3] Krzanowski W. J., (1988). Principles of Multivariate Analysis, A User's Perspective. Oxford Univ. Press: Clarendon.

[4] Krzanowski W. J. (1995). Recent advances in descriptive multivariate analysis. Oxford University Press, New York 1995.

[5] Krishnaiah P. R. (ed.) 1977. Multivariate Analysis II. North Holland Vol 2. p. 595.

[6] Kao W. J., Sapatnekar S., Hiltner A., Anderson J. M.: Complement mediated leukocyte adhesion on poly(etherurethane-ureas) under shear stress in vitro. J. Biomed. Mater. Res. 1996, 32 (1): 99–109.

[7] Larose D. T. 2005. Discovering Knowledge In Data. An Introduction to Data Mining. Wiley and Sons.

[8] Lane D. A., Bowry S. K. The scientific basis for selection of measures of thrombogenicity. Nephrol. Dial. Transplant. 1994, 9: 18–28.

[9] Lim F., Yang C. Z., Cooper S. L.: Synthesis, characterization and ex vivo evaluation of polydimethylsiloxane polyurea urethanes. Biomaterials 1994, 15 (6): 408–416.