Listy Biometryczne
Vol. XXIV (1987) No.1
Biometrical Letters

TWO METHODS OF DIMENSIONALITY REDUCTION - A COMPARISON FOR MEDICAL DATA AND SIMULATION STUDY

Ewa Krusińska, Jerzy Liebhart

Institute of Computer Science, University of Wrocław,
Przesmyckiego 20, 51-151 Wrocław

Department of Internal Diseases, Medical Academy of Wrocław,
Traugutta 57, 50-417 Wrocław

Krusińska E., Liebhart J., 1987. Two methods of dimensionality reduction - a comparison for medical data and simulation .study. Listy Biometryczne XXIV, z.1, Wydawnictwo Naukowe Universytetu im. Adama Mickiewicza w Poznaniu (Adam Mickiewicz University Press) pp. 3-16, 4 tabl., PL ISSN 0458-0036.

The aim of this paper is to discuss the reduction of dimensionality in three statistical problems. These are discarding of variables in multiple linear regression, search for representative variables by regression analysis methods and search for variables with the greatest discriminatory power. The search for the subset from among a great number of predictor variables is practically impossible to be done in an optimal way because of high costs of computations. Therefore, a stepwise selection is commonly used in such cases. In this paper the Monte Carlo method consisting in generating a declared number of subsets and in choosing the best one from them is discussed in comparison with a stepwise selection. The methods were compared in two ways. The first comparison was performed for real data comprising results of examination of patients suffering from bronchial asthma and chronic bronchitis. In linear regression the Monte Carlo method gave better results than a stepwise procedure especially when the number of subsets generated was greater than the number of subsets analysed in the stepwise selection. For the second comparison a simulation study was performed on the basis of pseudo-random data with multivariate normal distribution. The comparison was done for an equal number of the analysed subsets.

1. ABOUT REDUCTION OF DIMENSIONALITY GENERALLY

The necessity of dimensionality reduction is often found in various statistical problems, e.g. in regression analysis, discriminant analysis or cluster analysis.

Key words: multiple linear regression, interdependence analysis, search of variables with the greatest discriminative power, reduction of dimensionality, stepwise selection, Monte Carlo method of selection

Let us consider p predictor variables. The subset consisting of r variables (the most important ones) should be chosen out of the complete set of p variables. To find the best subset or the optimal one all $\binom{p}{r}$ subsets should to be analysed with reference to a criterion used in the problem under consideration. In practice, when the number of predictor variables is large, the optimal choice often cannot be performed, even with a computer, because of the cost of analysing all $\binom{p}{r}$ subsets. In such situations the reduction of the number of variables can be done by the stepwise selection method, i.e. by the method in which the variables are added one by one. At each step the best variable is chosen out of the remaining features. Of course, the subset chosen in such a way cannot be the best one. The problem of the reduction of variables can also be approached by the Monte Carlo method, which consists in generating a declared number of subsets and in choosing the best subset among them. Two methods for generating subsets (with equal and unequal variable weights) are possible. The question is whether the subset found by the Monte Carlo method is better (with reference to a particular criterion used in the problem under consideration) than the subset established by a stepwise selection. In the paper, a comparison of the stepwise and Monte Carlo methods is made on the basis of real medical data concerning 105 individuals each with 47 variables, and on the basis of simulated data generated from the multivariate normal distribution.

2. CHOICE OF VARIABLES IN MULTIPLE LINEAR REGRESSION

Let us consider p predictor variables $x_1, x_2, \ldots, x_p$ and a dependent variable y. Further, let us assume that y is related to the predictor variables by the linear regression equation

$$y = b_o + b_1x_1 + \ldots + b_px_p + e \quad , \tag{1}$$

where $b_o, b_1, \ldots, b_p$ are regression coefficients and e is a random error term.

Let us assume that we have a sample of n, n>p+1 individuals. Each of them is characterized by a vector $(x_{i1}, x_{i2}, \ldots, x_{ip})$ of p predictor variables and a dependent variable y_i (i=1,2,...,p). The usual least squares estimates $\hat{b}_o, \hat{b}_1, \ldots, \hat{b}_p$ of the regression coefficients can be found by solving normal equations. A linear dependence between the variable y and the predictor variables $x_1, x_2, \ldots, x_p$ is measured by the multiple correlation coefficient

$$R_{y(1,2,\ldots,p)} = \sqrt{\frac{SST-SSE}{SST}} \quad , \tag{2}$$

where SST = $\sum_{i=1}^{n}(y_i-\bar{y}_.)^2$ is the total sum of squares ($\bar{y}_.$ being the sample mean) and SSE = $\sum_{i=1}^{n}(y_i-\hat{b}_o-\hat{b}_1x_i-\ldots-\hat{b}_px_{ip})^2$ is the residual sum of squares. The squared multiple correlation coefficient $R^2_{y(1,2,...,p)}$ also called determination coefficient is commonly used as a criterion for the choice of the best subset of predictor variables in linear regression (e.g. Mardia, Kent and Bibby (1979)). The optimal subset consisting the r predictor variables gives the highest determination coefficient $R^2_{y(1,2,...,r)}$ between the predictor variables and the variable y. The optimal choice of variables requires investigation of $\binom{p}{r}$ subsets. Such a method was programmed and analysed e.g. by Bartkowiak (1978). Bartkowiak (1981) compared it with the method by leaps and bounds (Furnival and Wilson (1974)) and proved that the latter one needed more processor time.

The stepwise procedure of Bartkowiak (1978) used in this paper as a basis for a comparison with the Monte Carlo method, consists in a jerking search of variables in the regression set. The stepwise selection of variables is performed using a double criterion. The total number of chosen variables should not be greater than r and the partial correlation coefficient $\rho_{y(k+1)(1,2,...,k)}$ should be significant at the significance level equal to α (where the (k+1)th variable is added to the subset).

The significance is tested by:

$$F = \frac{R^2_{y(k+1)} - R^2_{y(1,...,k)}}{1-R^2_{y(k+1)}} \cdot \frac{n-k-2}{1} \sim F_{1,n-k-2} \qquad (3)$$

(e.g. Bartkowiak, 1982a).

Having selected a declared number of variables we can change the significance level by diminishing α. Then the variables which do not agree with the new level α are eliminated (backward elimination). If the number of variables in the subset is less than r we can add to the regression further variables significant at the new level α.

3. INTERDEPENDENCE ANALYSIS BY ANALYSIS OF REGRESSION

Let us consider the set $x_1,x_2,...,x_p$ of p variables. We would like to find a subset consisting of r variables so that the remaining p-r variables are explainable by variables in the chosen subset as fully as possible.

Let us label r variables in the subset by indices $\{i_1,i_2,...,i_r\}$. The variables of the subset $\{i_1,i_2,...,i_r\}$ are treated as independent (predictor) variables, the remaining p-r variables are treated as dependent ones. Then, we consider the residual sums of squares $SSE^{i_{r+k}}_{\{i_1,i_2,...,i_r\}}$ (k=1,2,...,p-r) of the (i_{r+k})th variable with respect to i_1, $i_2,...,i_r$ variables and we find such variable k_o for which the residual sum of squares is maximum, i.e.,

$$\mathrm{SSE}^{k_o}_{\{i_1,i_2,\ldots,i_r\}} = \max_{1\le k\le p-r} \mathrm{SSE}^{i_{r+k}}_{\{i_1,i_2,\ldots,i_r\}} \quad . \tag{4}$$

Then, we find the subset $\left\{i_1^o,i_2^o,\ldots,i_r^o\right\}$ for which the maximum residual sum of squares is minimum

$$\mathrm{SSE}^{k_o}_{\left\{i_1^o,i_2^o,\ldots,i_r^o\right\}} = \min_{\substack{\text{all subsets}\\ \{i_1,i_2,\ldots,i_r\}}} \max_{1\le k\le p-r} \mathrm{SSE}^{i_{k+r}}_{\{i_1,i_2,\ldots,i_r\}} \quad . \tag{5}$$

So the criterion for choosing the best subset is the maximum residual sum of squares. The smallest one indicates the best subset.

The stepwise choice of the subset can be performed upwards or downwards (programmed by Bartkowiak (1982c)). In the first method, the variables are added one by one to the subset. In the second one variables are eliminated one by one out of the complete set of variables.

4. SEARCH OF VARIABLES WITH THE GREATEST DISCRIMINATORY POWER

Let us assume that we have a sample of n individuals. Each one is characterized by an observational vector $(x_1,x_2,\ldots,x_p)$ of predictor variables. The discriminatory power of the variables $x_1,x_2,\ldots,x_p$ is in keeping with the possibility of discrimination between the considered populations. The discriminatory power can be measured by various criteria e.g. the probability of misclassification, Wilks Λ statistic, trace criterion. In this paper Wilks Λ statistic is used to measure the discriminative power of variables. It is defined as the ratio of two determinants (e.g. Rao (1965)):

$$\Lambda = \frac{|W|}{|T|} \ , \tag{6}$$

where W is the within-group adjusted squares and products matrix and T are the total adjusted squares and products matrix. The criterion Λ takes the values of $0\le\Lambda\le1$. When $\Lambda = 0$, there is a complete discrimination, when $\Lambda = 1$, there is no discriminatory power at all.

The stepwise selection of variables with the greatest discriminative power was programmed by Bartkowiak (1982b) and can be performed upwards and downwards.

5. THE MONTE CARLO METHOD OF SUBSET CHOICE

An extensive theoretical description of the Monte Carlo methods can be found for instance in Zieliński (1972). The aim of the present authors was to evaluate the performance of one of such methods (Bartkowiak (1985)) in biometrical problems.

In the problem of subset selection, the Monte Carlo method can be used for generating declared number of subsets with r variables out of the complete set of p variables. Then only these generated subsets are taken into consideration and the values of the criterion function considered are compared only for them. These criteria are:

A. in multiple regression analysis - $R^2_{y(1,\dots,r)}$, i.e., the determination coefficient (formula (2)),

B. in the search of the most representative variables by regression analysis methods - $SSE^{k_\delta}_{\{i_1,\dots,i_r\}}$, i.e., the maximal residual sum of squares (formula (5)),

C. in the search of variables with the greatest discriminatory power - Wilks Λ statistic (formula (6)).

The simulation of the subset may be performed in two ways:

(a) numbers of the predictor variables taken to the subset are generated out of the set $\{1,2,\dots,p\}$ as pseudo-random numbers in accordance with uniform distribution,

(b) numbers of the predictor variables taken to the subset are generated in accordance with the distribution of the weights of p predictor variables (Bartkowiak (1985)).

Variant (b) is realized in the following way:

1° At the beginning of the simulation, the weight w_i $(i=1,2,\dots,p)$ of each variable equals 1/p as in the variant (a). The starting value of the criterion K equals $K = K_o$.

2° The subset is generated.

3° When the computed value of the considered criterion K, i.e., K_{comp} is better than K we set

$$w_{i_j} := 0.5\,(w_{i_j} + 1/r) \qquad (j=1,2,\dots,r)\ ,$$
$$w_{i_{r+k}} := 0.5\,w_{i_{r+k}} \qquad (k=1,2,\dots,p-r)\ ,$$
$$K := K_{comp}\ .$$

This means that the selected variables are "rewarded" and the remaining variables are "punished".

Then a new subset is generated once more (return to 2°).

4° When K_{comp} is worse than K, we set

$$w_{i_j} := 0.5\,w_{i_j} \qquad (j=1,2,\dots,r)\ ,$$
$$w_{i_{r+k}} := 0.5\,\{w_{i_{r+k}} + 1/(p-r)\} \qquad (k=1,2,\dots,p-r)\ .$$

This means that the selected variables are "punished", the remaining ones are "rewarded".

Then a new subset is generated once more (return to 2°).

For their analysis the authors used the statistical package A-STAT in the ALGOL 1900 language (Bartkowiak (1985)) with procedures realizing both variants (a) and (b).

6. MEDICAL EXAMPLE

The data used to compare the stepwise selection method to the Monte Carlo one were collected in the Department of Internal Diseases, Medical Academy of Wrocław and comprised results of examination of 105 patients who were divided into 4 groups:

I. patients with obstructive type of ventilatory defect (n=48),
II. patients with restrictive type of ventilatory defect (n=21),
III. patients with mixed ventilatory function loss (n=19),
IV. a control group without ventilatory defects (n=17).

Each patient was spirometrically examined to obtain the following basic spirometric parameters:

FVC (dm^3)	- forced vital capacity,
FEV_1 (dm^3)	- forced expiratory volume in one second,
FMF (dm^3/s)	- forced midexpiratory flow,
$FEF_{0.2-1.2}$ (dm^3/s)	- forced expiratory flow at 0.2 - 1.2 dm^3 of FVC,
MMFT (s)	- time of FMF.

These parameters are easily obtainable by a direct measurement of volume and flow rate of air during one forced expiration. Additionally, the total lung capacity-TLC(dm^3) and the residual volume-RV(dm^3) were measured by a helium method. Height and simple measurements of the chest were also collected for each patient. The total number of variables was 47. The hypotheses on univariate normality of these variables and on the homoscedasticity of covariance matrices were rejected at the significance level $\alpha = 0.05$. Some previous results obtained for this set of data were given by Krusińska and Liebhart (1985).

The statistical analysis was performed in 3 variants.

A. Multiple linear regression analysis

The total lung capacity TLC was treated as the dependent variable y. TLC is the volume of air contained in the lungs after maximum inspiration. TLC and the residual volume RV (the volume of air which always remains in chest when a deep expiration is finished) cannot be obtained by a simple spirometric examination during forced expiration. Special expensive techniques such as plethysmography, helium method or rentgenography are necessary. These methods are complicated and time consuming. Therefore, the linear regression equation for TLC was found stepwisely (Liebhart et al. (1985)) for the whole sample and for each group separately. The variables were chosen for the regression equation out of a set of 41 features (independent of TLC and RV). Now the results of the stepwise selection by the jerking method are compared with the Monte Carlo choice results. The number of generated subsets was equal to the number of subsets analysed by a stepwise method or was two and three times larger. Subsets from 2 to 6 variables were taken into consideration. The comparison of results for the whole sample of 105 patients is presented,in Table 1 (the Monte Carlo subsets better than those of the stepwise selection are marked with *). Additionally the best subset obtained by the Monte Carlo method (consisted of 2 up to 6 variables) and the stepwise selection subset are listed in the table, too. The results are very promising. All subsets obtained by the Monte Carlo method are better. From

Table 1. The comparison of results in linear regression

variant	number of variables in subset	$R^2_{y(1,2,\ldots,r)}$ number of subsets analysed by the Monte Carlo method			variables in the best subsets
		as in stepwise	2x stepwise	3x stepwise	
(a)	2	0.1845	0.7693*	0.7693*	FVC % of predicted value, one measurement of the chest
(b)		0.7693*	0.7670*	0.7693*	
stepwise			0.6795		height,FVC
(a)	3	0.9059*	0.9059*	0.9259*	age, two measurements of the chest
(b)		0.7649*	0.9059*	0.9258*	
stepwise			0.7237		height, FVC, one measurement of the chest
(a)	4	0.7798*	0.8963*	0.8963*	age, three measurements of the chest
(b)		0.8963*	0.9257*	0.9257*	
stepwise			0.7328		height, FVC, two measurements of the chest
(a)	5	0.9233*	0.9233*	0.9263*	age, height,, FEV_1, two measurements of chest
(b)		0.9280*	0.9263*	0.9278*	
stepwise			0.7424		FVC, FEV_1 predicted/ FVC predicted x 100, three measurements of the chest
(a)	6	0.9272*	0.9272*	0.9272*	age, height, FEV_1 predicted/ FVC predicted x 100, two measurements of the chest
(b)		0.9174*	0.9281*	0.9288*	
stepwise			0.7495		FVC, five measurements of the chest

* indicates better result obtained by the Monte Carlo method.

the medical point of view it is interesting that almost in all the subsets obtained by the stepwise or Monte Carlo method geometrical measurements of the chest have appeared. E.g. the Monte Carlo subset consisting of three variables contains the features: age and two measurements of the chest. The determination coefficient equals 0.9259 and is considerably larger than that for the stepwise selection subset (equal to 0.7237). So the regression equation with those three variables chosen by the Monte Carlo method gives good results of TLC prediction and may be used in practice. Such a prediction is easy and can be made simultaneously with the standard spirometric examination during one forced expiration.

B. Interdependence analysis by the regression analysis method.

The choice of the most representative variables was made by the regression analysis method. The results obtained for the whole sample of patients are summarized in Table 2. The stepwise selection was performed upwards and downwards. A criterion for the choice of the subset was the maximum residual sum of squares. Only a few results obtained by the Monte Carlo method are better than the results of the stepwise selection even when the number of generated subsets is three times larger than the number of subsets analysed by the stepwise method. Additionally, the stepwise selection subsets and the best Monte Carlo subsets are listed in the table. The stepwise selection subsets contain FVC % (of the predicted value) and measurements of the chest. It has appeared that the Monte Carlo selection subsets are enriched by other spirometric parameters, such as FEV_1, FMF/MMFT, RV % and age. So it can be said that the information provided by a simple spirometric examination cannot be omitted and the basic spirometric parameters besides measurements of the chest should be treated as representative of the whole set.

C. The choice of the most discriminative variables

The choice of the most discriminative variables was performed by differentiating between 4 considered groups - 3 types of ventilatory defects (obstruction, restriction, mixed type) and a norm. From the medical point of view such a choice is very important because it permits to find the most diagnostic features in differentiating between various diseases or, as in our example, between various ventilatory defects and a norm, Then the automatic assistance in diagnosis by the use of discriminant functions may be performed for the most diagnostic features (chosen statistically). As it can be seen from Table 3 presenting the comparison of the stepwise and Monte Carlo selection an improvement of the results in choosing of the most discriminative features by the Monte Carlo method was obtained only in one case. The Monte Carlo selection subset consisting of two variables is better than that obtained by the stepwise method. It has appeared that FVC % and RV % differentiate better between the various ventilatory defects and the norm than FEV_1 % and RV %. Further results of the stepwise selection confirm that the

Table 2. The comparison of results in interdependence analysis

variant	number of variables in subset	$\mathrm{SSE}^{k_o}_{\left\{i_1^o, i_2^o, \dots, i_r^o\right\}}$ number of subsets analysed by the Monte Carlo method			variables in the best subsets
		as in stepwise	2x stepwise	3x stepwise	
(a)	2	0.9587	0.9587	0.9587	FMF/MMFT, one measurement of the chest
(b)		0.9730	0.9587	0.9514*	
stepwise (upwards)			0.9587		two measurements of the chest
stepwise (downwards)			0.9804		
(a)	3	0.9386	0.9386	0.9852	FEV_1, two measurements of the chest
(b)		0.9462	0.9569	0.9132*	
stepwise (upwards)			0.9349		FVC % (of predicted value), two measuremens of the chest
stepwise (downwards)			0.9615		
(a)	4	0.8947*	0.8947*	0.8947*	age, FEV_1, two measurements of the chest
(b)		0.8922*	0.9078	0.8968	
stepwise (upwards)			0.8967		FVC %, three measurements of the chest
stepwise (downwards)			0.9227		
(a)	5	0.8645	0.8645	0.8296*	FEV_1, FEV_1 %, RV %, two measurements of the chest
(b)		0.8683	0.8600	0.8600	
stepwise (upwards)			0.8597		FVC %, four measurements of the chest
stepwise (downwards)			0.8940		
(a)	6	0.8261*	0.8261*	0.8261*	age, FVC %, RV %, FMT/MMFT, two measurements of the chest
(b)		0.8424	0.8352*	0.8419*	
stepwise (upwards)			0.8421		FVC %, five measurements of the chest
stepwise (downwards)			0.8535		

* indicates better result obtained by the Monte Carlo method

forced vital capacity (FVC) is a very important parameter in the recognition of ventilatory defect (it has occurred as the third parameter in the stepwise selection). Generally, it can be stated that the basic spirometric parameters (obtained during one forced expiration) and RV % are the most essential in differentiating between ventilatory defects.

Table 3. The comparison of results in discriminant analysis

variant	number of variables in subset	value of Wilks Λ statistic number of subsets analysed by the Monte Carlo method			variables in the best subsets
		as in stepwise	2x stepwise	3x stepwise	
(a)	2	0.5191	0.3888	0.4082	FVC % (of predicted value), RV %
(b)		0.4059	0.2629*	0.3815	
stepwise (upwards)			0.3072		FEV$_1$ %, RV %
stepwise (downwards)			0.5113		
(a)	3	0.4687	0.2536	0.2974	FVC %, FMF, RV %
(b)		0.3376	0.2966	0.2239	
stepwise (upwards)			0.1730		FVC %, FEV$_1$ %, RV %
stepwise (downwards)			0.4925		
(a)	4	0.2332	0.1924	0.2180	FVC, FVC %, FEV$_1$, RV %
(b)		0.1842	0.1752	0.2128	
stepwise (upwards)			0.1523		FVC %, FEV$_1$ %, FEV$_1$/FVC x 100,RV %
stepwise (downwards)			0.2293		
(a)	5	0.2191	0.1790	0.1904	FEV$_1$ %, TLC, three measurements of the chest
(b)		0.1833	0.1645	0.1606	
stepwise (upwards)			0.1343		FVC %, FEV$_1$ %, FEV$_1$/FVC x 100, FEV$_1$ predicted/FVC predited x 100, RV %
stepwise (downwards)			0.1660		
(a)	6	0.1653	0.1912	0.1648	FVC %, FEV$_1$/FVC x100 RV %, FVC/MMFT, one measurement of the chest
(b)		0.1560	0.1612	0.1731	
stepwise (upwards)			0.1263		FVC %, FEV$_1$ %, FEV$_1$/FVC x 100, FEV$_1$ predicted/FVC predicted x 100, RV %, one measurement of the chest
stepwise (downwards)			0.1454		

* indicates better result obtained by the Monte Carlo method

7. SIMULATION STUDY

To make a more comprehensive comparison of the two methods discussed a simulation study was performed. The theory of multivariate regression analysis and discriminant analysis is classically developed for normal variables. Therefore, the pseudo-random data with multivariate normal distribution were generated in 10 variants to check the performance of the two methods of dimensionality reduction when the assumption on normality is fulfilled. The data used for the comparison were obtained by the GENMN program (Bartkowiak, Krusińska (1986)).

The multivariate random variable $X = (X_1, X_2, \ldots, X_s)$ is generated from the multivariate normal distribution with zero mean vector and the covariance matrix Σ using the method of Zieliński (1979). As the matrix Σ, the Lietzke matrix (Lietzke et al. (1964)) is taken.

To compare the methods of dimensionality reduction, two groups of data consisting of $n = 1000$ individuals were generated. The number of variables equalled 11. The first $p=10$ variables were treated as predictor variables, the last one was treated as the dependent variable y for linear regression. The first group of data had zero mean vector, the second one had the vector

$$(-3\sigma,\ -2\sigma,\ -\sigma,\ -\tfrac{1}{2}\sigma,\ 0,\ 0,\ \tfrac{1}{2}\sigma,\ \sigma,\ 2\sigma,\ 3\sigma)$$

as the mean vector (where σ is standard deviation). The different means in the second group were chosen for variables to impose on them different discriminatory power. As a common covariance matrix Σ for two groups of data the Lietzke matrix was taken. The data were generated 10 times. The regression analysis and the interdependence analysis were performed in each group separately (totally 20 trials). The choice of variables with the greatest discriminatory power was performed in differentiating between two generated groups (10 trials).

The essential results are summarized in Table 4. The numbers of results in the Monte Carlo selection that were better than those of the

Table 4. Number of results better in the Monte Carlo selection and number of results better for variant (b) with unequal weights

comparison	number of variables in subset	linear regression (for 20 trials)	interdependence analysis (for 20 trials)	discriminant analysis (for 10 trials)
number of results better in the Monte Carlo selection than in the stepwise one	2	20	20	0
	3	20	14	4
	4	20	14	3
	5	20	4	2
	6	20	2	0
number of results better for variant (b) with unequal weights than for variant (a)	2	0	3	4
	3	2	3	8
	4	2	4	0
	5	10	0	2
	6	5	1	9

stepwise selection are given in it. The number r of the variables in the subset increases from 2 to 6 variables. To study the effect of weights on the results of selection procedure the variants (a) and (b) (see Section 5) for subset generating were also compared. As seen in Table 4, all trials for the linear regression were successful. In the interdependence analysis the number of successful trials decreased with the increase in the number of variables in the subsets. In more than half of the trials the Monte Carlo method results were also better. The results obtained for the choice of variables with the greatest discriminatory power were not so good. Only several trials were successful.

Table 4 also contains a comparison of the results of the Monte Carlo method with (a) unequal and (b) equal variable weights. The question was whether the weights (different and changing during the run of the generating procedure) would improve results of choosing a subset in the considered problems. As we can see, it is impossible to find a rule concerning the effect of the weights on the results of the subset choice. Only in 53 trials (for the total number equal to 250) the results obtained with different weights are better than those with equal weights. It cannot be also stated on the basis of our simulation study whether the influence of weights on the selection results decreases or increases with the increase of the number of variables in a subset.

ACKNOWLEDGMENT

The authors wish to thank Dr Anna Bartkowiak of the Institute of Computer Science, University of Wrocław, for her encouragement in carrying out the present study and for her helpful remarks and discussion of this paper. We are also thankful to her for making available the statistical package A-STAT.

We are also indebted to the Editor of the Journal for useful comments on an earlier version of this paper.

REFERENCES

Bartkowiak, A. (1978). Multiple regression with stepwise selection of variables. *Zastosowania Matematyki* 16, 293-315.

Bartkowiak, A. (1981). Stosowalność zmodyfikowanego algorytmu Gaussa-Jordana w niektórych zagadnieniach statystycznych. *Raport* N-91, Instytut Informatyki U.Wr., Wrocław.

Bartkowiak, A. (1982a). *Opis merytoryczny programów statystycznych opracowanych w Instytucie Informatyki U.Wr.*, 2nd edition, Wydawnictwo Uniwersytetu Wrocławskiego, Wrocław.

Bartkowiak, A. (1982b). Algorithm 82. Stepwise selection of discriminative variables by the use of the Wilks criterion. *Zastosowania Matematyki* 17, 351-364.

Bartkowiak, A. (1982c). Algorithm 84. The choice of representative variables by stepwise regression. *Zastosowania Matematyki* **18**, 527-538.

Bartkowiak, A. (1985). *A-STAT. Zbiór procedur statystycznych w języku ALGOL 1900 na m.c. ODRA 1305.* Tom 2. Wydawnictwo Uniwersytetu Wrocławskiego, Wrocław.

Bartkowiak, A., Krusińska, E. (1986). *SABA. Opis programów statystycznych w języku ALGOL 1900.* Tom 2. Wydawnictwo Uniwersytetu Wrocławskiego, Wrocław.

Furnival, G.M., Wilson, R.W. (1974). Regression by leaps and bounds. *Technometrics* **16**, 499-512.

Krusińska, E., Liebhart, J. (1985). Heuristic and stepwise selection of variables. A comparison for some respiratory disease data. In: *Abstracts of contributed papers*, 1st European Biometric Conference, Budapest, 1-4 April 1985, 77-77.

Liebhart, J., Karkowski, Z., Krusińska, E., Liebhart, E., Małolepszy, J. (1985). Obliczanie wartości całkowitej pojemności płuc (TLC) na podstawie krzywej natężonego wydechu i prostych pomiarów klatki piersiowej. *Pneumonologia Polska* **53**, 520-526.

Lietzke, M.H., Stoughton, R.W., Lietzke, M.P. (1964). A comparison of several methods for inverting large symmetric positive define matrices. *Math. Comput.* **18**, 449-456.

Mardia, K.V., Kent, J.T., Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, London.

Rao, C.R. (1965). *Linear Statistical Inference and Its Applications.* Wiley, New York.

Seber, G.A.F. (1984). *Multivariate Observations.* Wiley, New York.

Zielinski, R. (1979) *Metody Monte Carlo.* WNT Warszawa.

Zieliński, R. (1979). *Generatory liczb losowych.* 2nd edition, WNT, Warszawa.

Received 12 September 1985: revised 9 December 1987

DWIE METODY REDUKCJI WYMIAROWOŚCI - PORÓWNANIE DLA DANYCH MEDYCZNYCH I BADANIA SYMULACYJNE

Streszczenie

W pracy przedyskutowano redukcję wymiarowości przestrzeni cech w trzech zagadnieniach statystycznych. Był to wybór zmiennych do równania regresji liniowej, poszukiwanie najlepszych reprezentantów metodami analizy regresji oraz wybór zmiennych o największej sile dyskryminacji. We wszystkich tych zagadnieniach z uwagi na wysoki koszt obliczeń nie można w praktyce stosować optymalnej metody wyboru podzbioru przy dużej liczbie

zmiennych objaśniających. Dlatego też w takim wypadku stosowana jest zwykle metoda krokowa. W tej pracy postępowanie krokowe porównano z metodą opartą na generowaniu pewnej deklarowanej liczby podzbiorów i wyborze najlepszego podzbioru spośród nich. Metody były porównane na dwa sposoby. Pierwsze zestawienie wykonano dla danych medycznych obejmujących chorych z astmą oskrzelową lub przewlekłym nieżytem oskrzeli. Metoda Monte Carlo dała lepsze wyniki szczególnie dla regresji liniowej, gdy liczba wygenerowanych podzbiorów przekraczała liczbę podzbiorów analizowanych w metodzie krokowej. Ponadto przeprowadzono badania symulacyjne na danych wygenerowanych z wielowymiarowego rozkładu normalnego. Liczby generowanych podzbiorów cech w metodzie Monte Carlo i podzbiorów analizowanych w metodzie krokowej były równe.