The principal components analysis – method to reduce the collinearity in multiple linear regression model; application in medical studies

CRISTINA GENA DASCĂLU¹, CORINA DIMA COZMA²
¹Medical Informatics and Biostatistics Dept., Dental Medicine Faculty,
²The 6th Medical Clinic
University of Medicine and Pharmacy „Gr.T.Popa”
Str. Universitatii no. 16, Iași, 700115
ROMÂNIA
cristina_dascalu72@yahoo.com

Abstract: During the statistical analysis of medical data, in many situations it is necessary to identify the multiple correlations established between the studied parameters. In this purpose, one of the most useful methods is to build a model of multiple regression, which allows the modeling of a dependant variable values having at least the ordinal type, based on its linear relation with more than one independent variables satisfying the same restriction, called predictors. The main problem which affects the efficiency of such a model is the collinearity between predictors – a clue that they are not selected well because they overlap each other. There are different methods to eliminate this difficulty, but a very efficient one is the principal components analysis, applied as a method to reduce the predictors number, in a step preliminary to the regression model building. We will present the results obtained using these procedures on a set of 212 patients from three categories: healthy patients (65 cases), patients with liver cirrhosis (65 cases) and patients with liver hepatitis (82 cases), included into a study about the connections between the liver diseases and the heart’s health - measured using detailed electrocardiograms. The study’s purpose was to find if we can extract, using the heart’s activity analysis, certain conclusions about the liver’s state of health. The identified predictors are useful for further data processing.

Keywords: multivariate analysis, multiple linear regression, principal components analysis, collinearity, liver disease

1 Introduction
In many cases, in scientific researches, it is necessary to analyze the correlations between more than two parameters, in order to detect the internal influences between data. In such cases it is absolutely necessary to choose the right method to study the assumed correlations – because the parameters nature defines in fact the path we are going to follow. There are a few main possibilities to analyze the multiple correlations between data: the regressional models, the principal components analysis, the discriminant analysis, the clustering. The regressional analysis is the easiest available method between these, which tries to find a very clear pattern for data variation: one parameter is interpreted as being “dependant”, so it will be influenced in its variation by all the other parameters.

When all the studied parameters are quantitative and we intend to study in which way one of these parameters, defined as “dependant”, is influenced by all the other, the easiest way to quantify this influence is to build a model of linear multiple regression and to check how well it fits with the real studied phenomenon. The main difficulty which appears during this procedure is to find that the parameters identified as significant predictors are collinear between them, which means that their effects on the main model are overlapped. In order to eliminate this problem it is necessary to use other statistical methods for parameters processing, a very efficient one being the principal components analysis - a technique useful to select the most relevant parameters from a given list, with a minimal loss of information.

2. Problem Formulation

2.1 The linear regressions method
This method for data analysis is used when we need to model the values of a dependant variable according with the values of at least two independent variables, also called “predictors”, using the equation of a straight line. The main requirement that must be fulfilled by all the variables involved in the model is that these variables must have at least the scale type – but the model behaves best when all the variables are quantitative [1].

The linear regression model assumes that there is a linear relationship between the dependent variable and each predictor, described in the following formula:
\[ y_i = b_0 + b_1 x_{i1} + ... + b_p x_{ip} + e_i , \]
where:
- \( y_i \) is the value of the \( i \)-th case of the dependent scale variable
- \( p \) is the number of predictors
- \( b_j \) is the value of the \( j \)-th coefficient, \( j \in \{0, 1, ..., p\} \)
- \( x_{ij} \) is the value of the \( i \)-th case of the \( j \)-th predictor
- \( e_i \) is the error in the observed value for the \( i \)-th case.

We can notice that we are dealing here with an equation of 1st degree, with \( p \) variables; \( b_0 \) is the intercept or the model-predicted value of the dependent variable when the value of every predictor is equal to 0 (the point where the line intersects the Oy axes, in a representation using a Cartesian coordinates system). The error term \( e_i \) must fulfill also the following conditions [1]:
- Its distribution is normal, with a mean of 0;
- Its variance is constant across cases and independent of the variables in the model;
- Its value for a given case is independent of the values of the variables in the model and of its values term for other cases.

When we build the multiple linear regression model, we must follow a few steps.

First, it is necessary to check the model fit [2]. In this purpose, the ANOVA test is used, and the F statistic is calculated; if the F value is statistically significant (\( p \leq 0.05 \)), it follows that the model fits well with the analyzed data, and using it is better than guessing the mean.

Secondly, the correlation coefficients \( R \) and \( R^2 \) are calculated; these coefficients show also how well the model works: \( R^2 \) squared, for example, shows the percentage from the dependant variable’s variation which is explained by the model.

Then, for each predictor involved in the model, we calculate its unstandardized and standardized coefficients and its significance level (expressed using the \( t \) statistic); in this way we can separate, from all the predictors involved in the model, only the significant ones – in order to eliminate further from the model the non-significant predictors (variables which do not contribute too much to the model). At this step we can also find the relative importance of each significant predictor, which varies proportionally with its standardized coefficient Beta.

At this step it is also good to calculate the part and partial correlation coefficients [3]; these coefficients help us to detect the possible multicollinearity problems. Such problems appear when the part and the partial correlations are very different by the zero-order correlation – which means that a large amount in the variance of the dependant variable that is explained by the analyzed predictor is also explained by the other predictors – so the predictors are “collinear” and their effects are overlapped. Another coefficient calculated at this step is the tolerance – or the percentage of the variance in a given predictor that cannot be explained by the other predictors; small tolerances show that large amounts in the variance in a given predictor are explained also by the other predictors, so again the multicollinearity is present, and large tolerances show that the multicollinearity is absent. Finally the multicollinearity is also measured using a Variance Inflation Factor (VIF) – that regards the standard error of the regression coefficients; a VIF factor greater than 2 is usually considered problematic, being a clue for predictors multicollinearity.

There are also a few diagnostics tests especially designed for collinearity [4]:
- in the predictors matrix, the eigenvalues are calculated; if these values are close to 0, it means that the predictors are highly intercorrelated, and therefore, small changes in the data values may lead to large changes in the estimates of the coefficients;
- in the same matrix, the condition indices are also computed, as the squared roots of the ratios of the largest eigenvalue to each successive eigenvalue; values greater than 15 indicate a possible problem with collinearity; greater than 30, a serious problem.

The last step of the analysis regards the collinearity removing, as long as this is possible. In order to do this, the easiest way is to rerun the model using the z scores for all the variables involved (predictors, as well as the dependant variable) instead of their direct values. In this way, only the most useful predictors will be included in the model.

### 2.2 The principal components analysis

A very useful method to deal with the predictors multicollinearity is also the principal components analysis.

This method of data analysis, described by Pearson (1901) and Hotelling (1933), concerns the finding of the best way to represent \( n \) samples by using vectors with \( p \) variables (predictors), in such a manner so the similar samples are represented by points as close as possible. In order to find the principal components from a set of predictors, the method used is the analysis of eigenvalues and eigenvectors, which starts from a data representation using a symmetrical matrix and transforms it.

Let \( X = \{x_1, x_2, ..., x_p\} \) be a set of points that makes a cloud in the \( R^p \) space [5]. We try to find the directions \( u_1, u_2, ... \) where the cloud’s dispersion is maximal. To do this, we find the cloud’s centroid and the lines \( L_1, L_2, ... \)
around which the cloud’s points are closely grouped and pass through the centroid, so the directions \( u^1, u^2, \ldots \) will be those lines directions. Denoting by \( d_j \) the distance from the \( x^j \) point to a line \( L \), the problem is to find the line where the quantity \( J = \sum_{j=1}^{p} d_j^2 \) is minimal. It is demonstrated that this line pass through the cloud’s centroid [6]. Making a data normalization by the translation \( x^* = x - m, x \in X \) (\( m = \) the average value of the points \( x \in X \)), the cloud \( X \) becomes \( X^* \), with average 0 and the centroid situated in the coordinate system’s origin; denoting by \( u \) the direction vector of the searched line, \( \| u \|=1 \), the criteria function becomes

\[
J(u) = \sum_{j=1}^{p} d_j^2 = \sum_{j=1}^{p} \left\| x^j - (x^j, u)^T \right\|^2
\]

The minimization of the function \( J \) is then equivalent with the maximization of the function

\[
I(u) = \sum_{j=1}^{p} \left\| x^j \right\|^2 - \sum_{j=1}^{p} (x^j, u)^2
\]

Because \( \| u \|=1 \) we determine for the I function the maximal value for the unit sphere vectors.

The quadratic matrix with \( d \) order,

\[
S = \sum_{j=1}^{p} x^j x^j^T
\]

is the spreading matrix of the normalized points cloud. The function’s \( I(u) = u^T S u \) extreme values on the unit sphere vectors are identical with the proper vectors of the \( S \) matrix, and the proper vector corresponding to the biggest proper value of \( S \) is identical with the direction where \( I(u) \) is maximal.

The proper vectors of the \( S \) matrix, denoted by \( u^1, u^2, \ldots, u^d \), and taken in descendant order of their corresponding proper values, are the principal directions, or the principal components of the points cloud and shows the cloud’s orthogonal directions. The cloud is maximally extended in the \( u^1 \) direction (main proper vector).

Practically, when we make the principal components analysis in order to reduce a set of factors we have to follow the next steps [3]:

First we calculate the so-called “communalities”: these values show the quantity of variance of each predictor; when communalities are high enough (over .750) it means that the extracted predictors represent well the whole data set; if we find communalities with low values, this is a sign that the corresponding predictors are not well selected and we have to replace them.

Then we calculate the variance corresponding to the initial set of predictors. The total eigenvalues show the quantity of variance in the original variables added by each extracted component; this quantity is also expressed in percentages (as a ratio between the variance added by each component and the total variance) and in cumulative percentages. Usually in this moment it is possible already to identify the principal components of the set of analyzed predictors – these components are those with eigenvalues > 1.

For the extracted components we calculate also the extraction sums of squared loadings, which show, in terms of cumulative percentages, the amount of variability of the initial set of predictors covered by the extracted components. In this way we calculate in fact the percentage of information loss which appears when we replace the whole set of predictors with the set of extracted predictors and, if this percentage is small enough, it shows that the extracted predictors can be indeed used with good results instead the whole data set.

A predictors transformation usually applied in this stage is their rotation, in order to spread uniform the global variance corresponding to the set of principal predictors between them – because, without this transformation, some predictors tends to bring much more variance into the model then the others.

The predictors graphical representation is “the scree plot” of the eigenvalues corresponding to them; this graphic allows to identify visually quickly the principal predictors, because they appear on the steep part of the slope, while the others appear on the shallow part of the slope. The principal components are all those before the last big drop of the line.

In the last stage we find the real signification of the principal predictors using the rotated component matrix. This matrix contains a column for each principal component, filled with the correlation coefficients between all the predictors in the initial set and that component; the predictor with the biggest correlation coefficient is in fact the principal corresponding component; the predictor with the biggest correlation coefficient is in fact the principal corresponding component. When we have two predictors with very close correlation coefficient, the principal component is the predictor with smaller correlation coefficients with the other principal components from the matrix.

Another way to establish this correspondence is using the component score coefficient matrix, which leads to the same results as the rotated component matrix but uses for this a component score, calculated by multiplying the case’s standardized variable values with the score coefficients.
3 Practical Application

In order to check the practical efficiency of these methods, we took under analysis a set of 212 patients from three categories: healthy patients (65 cases), patients with liver cirrhosis (65 cases) and patients with liver hepatitis (82 cases). These patients were included into a study about the connections between the liver diseases and the heart’s health - measured using detailed electrocardiograms. The study’s purpose was to find if we can extract, using the heart’s activity analysis, certain conclusions about the liver’s state of health.

The heart’s activity was recorded by ECG, and 38 numeric parameters were analyzed, as it follows: the diastolic blood pressure; the systolic blood pressure; the cardiac frequency; the diameter of the aorta at the ring; the diameter of the ascendant aorta; the inter-ventricular septum width; the left ventricle posterior wall width; the left ventricle mass; the right atrium diameter; the right ventricle diameter; the left ventricle mass index; the diastolic left ventricle diameter; the systolic left ventricle diameter; the shortening ratio; the diastolic volume; the systolic volume; the ejection ratio; the cardiac flow; the left atrium diameter; the E wave velocity; the A wave velocity; the E / A ratio; the iso-volume relaxation time; the E wave deceleration time; the systolic pressure into the pulmonary artery; the pulmonary artery at the ring diameter; the average pulmonary arterial pressure; TAPSE; the inferior cave vein diameter; the average arterial blood pressure; the peripheral vascular resistance; the pre-ejection time; the ejection time; the pre-ejection time / ejection time ratio; the nervous conductivity speed; QTc and the O₂ arterial satiation (CLINO and ORTO).

The data analysis was performed in SPSS 16.0, on a dataset made using Microsoft Visual FoxPro.

First we tried to build directly the multiple linear regression model using as dependant variable the diagnosis type (with 3 possible categories: healthy, cirrhosis and hepatitis) and as independent variables all the 38 parameters mentioned before and extracted from the ECG analysis.

The ANOVA test shows that the multiple linear regression model fits significantly with the analyzed data (F = 9.972, p = .000). The correlation coefficients R (0.829) and R squared (0.687) come to strengthen the previous observation, having quite good values: 68.7% from the diagnosis variation is covered by the regresional model.

In order to identify which are the significant predictors of this model we checked the significance levels of all the predictors and we obtained 11 significant predictors for the model (tab. I).

According to the Beta coefficient values, it follows that the most important predictor is Peripheral vascular resistance ($\beta = -0.908$), followed by Average pulmonary arterial pressure ($\beta = -0.528$). The less important predictor is Inferior cave vein diameter ($\beta = -0.103$). Checking the tolerance values, we can see that most of them are different of 0 but, unfortunately, the VIF factor is smaller than 2 in only 2 cases from all 11 – so we have here important problems concerning the predictors collinearity.

Table I.
The predictors coefficients and significance levels (1st stage)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>Correlations</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
<td>Zero-order</td>
<td>Partial</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-----</td>
<td>------------</td>
<td>------</td>
<td>-------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>1 Left ventricle mass index</td>
<td>-0.11</td>
<td>0.05</td>
<td>-0.380</td>
<td>-2.147</td>
<td>.033</td>
<td>.270</td>
</tr>
<tr>
<td>2 Cardiac flow</td>
<td>-0.595</td>
<td>0.149</td>
<td>-0.482</td>
<td>-3.992</td>
<td>.000</td>
<td>.296</td>
</tr>
<tr>
<td>3 Iso-volume relaxation time</td>
<td>0.015</td>
<td>0.03</td>
<td>0.263</td>
<td>4.375</td>
<td>.000</td>
<td>.481</td>
</tr>
<tr>
<td>4 Systolic pressure - pulm.</td>
<td>0.018</td>
<td>0.05</td>
<td>0.369</td>
<td>3.921</td>
<td>.000</td>
<td>.115</td>
</tr>
<tr>
<td>artery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Average pulm. arterial</td>
<td>-0.037</td>
<td>0.08</td>
<td>-0.528</td>
<td>-4.785</td>
<td>.000</td>
<td>.030</td>
</tr>
<tr>
<td>pressure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Inferior cave vein</td>
<td>-0.025</td>
<td>0.012</td>
<td>-1.013</td>
<td>-2.051</td>
<td>.042</td>
<td>-0.058</td>
</tr>
<tr>
<td>diameter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Average arterial blood</td>
<td>0.019</td>
<td>0.05</td>
<td>0.307</td>
<td>3.979</td>
<td>.000</td>
<td>-0.259</td>
</tr>
<tr>
<td>pressure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Peripheral vascular</td>
<td>-0.001</td>
<td>0.00</td>
<td>-0.908</td>
<td>-5.894</td>
<td>.000</td>
<td>-0.492</td>
</tr>
<tr>
<td>resistance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 QTc</td>
<td>-0.012</td>
<td>0.003</td>
<td>-0.233</td>
<td>-3.788</td>
<td>.000</td>
<td>0.062</td>
</tr>
<tr>
<td>10 O₂ arterial satiation - CLINO</td>
<td>-0.078</td>
<td>0.026</td>
<td>-0.212</td>
<td>-2.985</td>
<td>.003</td>
<td>-0.245</td>
</tr>
<tr>
<td>11 O₂ arterial satiation - ORTO</td>
<td>0.052</td>
<td>0.020</td>
<td>0.226</td>
<td>2.575</td>
<td>.011</td>
<td>-0.169</td>
</tr>
</tbody>
</table>

Therefore we will try to solve this issue by applying to the initial set of 38 parameters the principal components analysis, in order to identify the most relevant between them.
From the variance analysis we found a number of 12 principal components (with initial eigenvalues > 1.00), according to the Scree Plot in Fig. 1. The cumulative initial eigenvalue of these components is 75.41 % - so these components correspond to an information (variability) loss of 24.59 %, compared with the whole set of parameters.

By calculating the Component Scores Coefficient Matrix and selecting the parameters with the highest values, we found that the 12 principal components correspond to the following parameters: the left ventricle mass (0.952); the average pulmonary arterial pressure (0.774); the E wave deceleration time(0.825); the right ventricle diameter (0.776); the pre-ejection time / ejection time ratio (0.935); the systolic left ventricle diameter (0.924); the E wave velocity (0.758); the O₂ arterial satiation (CLINO) (0.860); the systolic blood pressure (0.916); the ejection ratio (0.892); the E / A ratio (0.735) and the diameter of the aorta at the ring (0.848), in this order.

In the second step of our study, we rebuilt the multiple linear regression model using as dependant variable the diagnosis type (with 3 possible categories: healthy, cirrhosis and hepatitis) and as independent variables the 12 parameters identified as principal components.

The ANOVA test shows again that the multiple linear regression model fits significantly with the analyzed data (F = 11.859, p = .000), but the correlation coefficients are weaker than in the previous step: R = 0.646 and R squared = 0.417 – only 41.7% from the diagnosis variation is covered by the regressional model.

In order to identify which are the significant predictors of this model we checked the significance levels of all the predictors and we obtained 8 significant predictors for the model (tab. II).

The most important predictor is E wave deceleration time (β = 0.341), followed by Systolic blood pressure (β = 0.255). The less important predictor is E/A ratio (β = -0.157) but, as general observation, all the values of β are significantly smaller than in the 1st stage of the study. Checking the tolerance values, we notice that all of them are very different of 0 and the VIF factor is smaller than 2 in all the cases. Therefore we have no problems concerning the predictors collinearity and the regressional model fits acceptably the data of our study.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>Zero-order</th>
<th>Partial</th>
<th>Part</th>
<th>Tolerance</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Average pulmonary arterial pressure</td>
<td>-0.015</td>
<td>.005</td>
<td>-2.19</td>
<td>-3.344</td>
<td>.001</td>
<td>.030</td>
<td>-.231</td>
<td>-.181</td>
</tr>
<tr>
<td>2</td>
<td>E wave deceleration time</td>
<td>.006</td>
<td>.001</td>
<td>.341</td>
<td>5.685</td>
<td>.000</td>
<td>.414</td>
<td>.374</td>
<td>.308</td>
</tr>
<tr>
<td>3</td>
<td>Right ventricle diameter</td>
<td>.029</td>
<td>.010</td>
<td>.178</td>
<td>2.824</td>
<td>.005</td>
<td>.234</td>
<td>.196</td>
<td>.153</td>
</tr>
<tr>
<td>4</td>
<td>Systolic left ventricle diameter</td>
<td>.025</td>
<td>.010</td>
<td>.144</td>
<td>2.388</td>
<td>.018</td>
<td>.218</td>
<td>.167</td>
<td>.129</td>
</tr>
<tr>
<td>5</td>
<td>O₂ arterial satiation - CLINO</td>
<td>-.059</td>
<td>.022</td>
<td>-.161</td>
<td>-2.702</td>
<td>.007</td>
<td>-.245</td>
<td>-.188</td>
<td>-.146</td>
</tr>
<tr>
<td>6</td>
<td>Systolic blood pressure</td>
<td>.012</td>
<td>.003</td>
<td>.255</td>
<td>4.579</td>
<td>.000</td>
<td>.238</td>
<td>.309</td>
<td>.248</td>
</tr>
<tr>
<td>7</td>
<td>Ejection ratio</td>
<td>-.028</td>
<td>.009</td>
<td>-.186</td>
<td>-3.231</td>
<td>.001</td>
<td>-.289</td>
<td>-.223</td>
<td>-.175</td>
</tr>
<tr>
<td>8</td>
<td>E / A ratio</td>
<td>-.399</td>
<td>.171</td>
<td>-.157</td>
<td>-2.338</td>
<td>.020</td>
<td>-.311</td>
<td>-.164</td>
<td>-.127</td>
</tr>
</tbody>
</table>

4 Conclusion
The multiple linear regression model is useful to identify the correlations between more than two parameters; it allows to select from large set of predictors only the significant ones, and also to classify them according to their importance in the variance of the dependant variable. The principal components analysis solves successfully the collinearity problems which can appear
between the predictors into a regressional model, allowing building better models of this type. Even if, using the principal components analysis, we usually lose some significant predictors identified through the regressional analysis, the major advantage of this method is that the new predictors identified have a better behavior, a more homogenous structure (beta values within a smaller variation range) and better tolerance and VIF values.

So these methods can be successfully used in complex set of data analysis, because they give an image about the general parameters behavior, the correlations between them and their internal links.

References: