PLISKA
STUDIA MATHEMATICA BULGARICA
ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **14** (2003), 59–70

PLISKA
STUDIA MATHEMATICA
BULGARICA

DIMENSION REDUCTION OF THE EXPLANATORY VARIABLES IN MULTIPLE LINEAR REGRESSION

P. Filzmoser and C. Croux

In classical multiple linear regression analysis problems will occur if the regressors are either multicollinear or if the number of regressors is larger than the number of observations. In this note a new method is introduced which constructs orthogonal predictor variables in a way to have a maximal correlation with the dependent variable. The predictor variables are linear combinations of the original regressors. This method allows a major reduction of the number of predictors in the model, compared to other standard methods like principal component regression. Its computation is simple and quite fast. Moreover, it can easily be robustified using a robust regression technique and a robust measure of correlation.

1. Introduction

Let us consider a problem where a certain phenomenon is depending on a series of other influential factors. For example, lung cancer could be influenced by different other health factors, but also on factors like smoking or even on environmental or psychological factors. In order to explain the phenomenon "lung cancer", we may consider a linear relationship to the other factors and ask for the contribution of these factors to the explanation of lung cancer. In practice we have to collect data in order to find an answer using a statistical analysis. The response variable (e.g. the percentage of lung cancer in different regions) is expressed by predictor

2002 *Mathematics Subject Classification:* 62J05, 62G35

Key words: Multiple linear regression, Principal component regression, Dimension reduction, Robustness, Robust regression

or explanatory variables (the other factors measured in the same regions) by a multiple linear regression model

$$y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i \qquad i = 1, \ldots, n. \tag{1}$$

n is the number of observations (regions), $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ are collected as rows in a matrix $\boldsymbol{X}$ containing the predictor variables, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is the response variable, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ are the regression coefficients which are to be estimated, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is the error term. The differences $y_i - \beta_0 - x_{i1}\beta_1 - \ldots - x_{ip}\beta_p$ express the deviation of the fit to the observed values and are called residuals. Traditionally, the regression coefficients are estimated by minimizing the sum of squared residuals

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_q x_{iq})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

This criterion is called *least squares* (LS) criterion, and the coefficient minimizing the criterion turns out to be

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \tag{2}$$

(see, e.g., Johnson and Wichern, 1998[8]).

Since the inverse of $\boldsymbol{X}^\top\boldsymbol{X}$ is needed in Equation (2), problems will occur if the rank of $\boldsymbol{X}$ is lower than $p+1$. This happens if the predictor variables are highly correlated or if there are linear relationships among the variables. This situation is called *multicollinearity*, and often a generalized inverse is then taken for estimating the regression coefficients. The inverse of $\boldsymbol{X}^\top\boldsymbol{X}$ also appears when computing the standard errors and the correlation matrix of the regression coefficients coefficients estimator $\hat{\boldsymbol{\beta}}_{LS}$. In a near-singular case the standard errors can be inflated considerably and cause doubt on the interpretability of these coefficients. Also note that the rank of $\boldsymbol{X}$ is always lower than $p+1$ if the number of observations is less or equal than the number of variables ($n \leq p$). This is a frequent problem which occurs in many applications e.g. in social sciences or in technical applications like in chemometrics, where the number of observations is much lower than the number of predictor variables.

The idea is to construct a limited set of k components $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$ which are linear combinations of the original variables. So there are existing vectors $\boldsymbol{b}_j$ such that $\boldsymbol{z}_j = \boldsymbol{X}\boldsymbol{b}_j$ for $1 \leq j \leq k$. Let $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)$ be the $n \times k$ matrix having the components in its columns. For ease of notation, we ask these components to be centered, so $\mathbf{1}_n^\top\boldsymbol{Z}$, with $\mathbf{1}_n$ a column vector with all n components equal to

1. Moreover, we also ask these components to be uncorrelated and to have unit variances:

$$\boldsymbol{Z}^\top \boldsymbol{Z} = \frac{1}{n-1}\boldsymbol{I}_k, \tag{3}$$

where $\boldsymbol{I}_k$ stands for an identity matrix of rank k. These components will then serve as new predictor variables in i the regression model. Note that, due to (3) the multicollinearity problem has completely vanished when using a regression model with $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$ as predictor variables. Moreover, when k is small relative to p, one has significantly reduced the number of predictor variables, leading to a more parsimoneous regression model.

The classical way to obtain $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$ is to use the first k principal principal components (PCs) of the predictors $\boldsymbol{X}$. This approach is called principal component regression (PCR) (see, e.g., Basilevsky, 1994[2]). If all PCs are used in the regression model, the response variable will be predicted with the same precision as with the LS approach. However, one goal is to simplify the regression model by taking a reduced number of variables in the prediction set. While PCR can deal with multicollinearity, it is not a method which directly maximizes the correlation between the original predictors and the response variable. In fact, when constructing the components using PCR, the values of the response variables do not play a role at all. It was noted by Hadi and Ling (1998)[6] that in some situations PCR can give quite low values for the squared multiple correlation (SMC) between the response variable and the predictors. The SMC is also called the R^2-*coefficient*, and the SMC between $\boldsymbol{y}$ and $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)$ can be defined as

$$\mathrm{SMC}(\boldsymbol{y}, \boldsymbol{Z}) = \max_{b} |\mathrm{Corr}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{b})|, \tag{4}$$

where Corr stands for the usual bivariate correlation coefficient.

In Section 2 we prove two propositions which are suggesting a sequential approach in order to maximize the measure of determination of the response variable. The algorithms for constructing the components $\boldsymbol{z_1}, \ldots, \boldsymbol{z}_k$ is outlined in Section 3. The advantage of the proposed method will be that it can allow for a larger reduction of the number of predictor variables in the regression model. This is illustrated by means of a simulation study in Section 4. Section 5 concludes.

2. Sequential construction of the components

PCR is a two-step procedure: in the first step one computes PCs which are linear combinations of the predictor variables, and in the second step the response variable is regressed on the (selected) PCs. For maximizing the relation to the

response variable we could combine both steps in a single method. This method will find $k < p$ new predictor variables $\boldsymbol{z}_j$ $(j = 1, \ldots, k)$ which are linear combinations of the original variables and have high values of the SMC with the response variable. As before, we require that these predictor variables are centered, have unit variance, and are uncorrelated to each other. These components will be determined sequentially, as motivated by the next proposition.

Proposition 1. *Suppose that matrix* $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)$ *consists of the first* k *components. The* $(k+1)$*th component needs to be of the form* $\boldsymbol{z}_{k+1} = \boldsymbol{X}\boldsymbol{b}$ *for a certain coefficient vector* $\boldsymbol{b}$*. The problem of finding* $\boldsymbol{b}$ *such that the squared multiple correlation between* $\boldsymbol{y}$ *and* $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k, \boldsymbol{z}_{k+1})$*,*

$$\boldsymbol{b} \longrightarrow SMC(\boldsymbol{b}) := SMC(\boldsymbol{y}, (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k, \boldsymbol{z}_{k+1})) = SMC(\boldsymbol{y}, (\boldsymbol{Z}, \boldsymbol{X}\boldsymbol{b})), \tag{5}$$

is maximized is equivalent to finding $\boldsymbol{b}$ *such that*

$$\boldsymbol{b} \longrightarrow |\mathrm{Corr}(\boldsymbol{y}^k, \boldsymbol{X}^k\boldsymbol{b})| \tag{6}$$

is maximized. In (6), $\boldsymbol{X}^k$ *stands for the residual matrix of regressing* $\boldsymbol{X}$ *on* $\boldsymbol{Z}$*, and* $\boldsymbol{y}^k$ *is the residual vector of regressing* $\boldsymbol{y}$ *on* $\boldsymbol{Z}$*.*

Proof. Regression of each column of $\boldsymbol{X}$ on $\boldsymbol{Z}$ leads to a matrix of fitted values

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{X},$$

and the residuals we obtain are gathered in the matrix

$$\boldsymbol{X}^k := \boldsymbol{X} - \hat{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{X}. \tag{7}$$

Since the components were assumed to verify (3), Equation (7) can be written as

$$\boldsymbol{X}^k = \left(\boldsymbol{I}_n - \frac{1}{n-1}\boldsymbol{Z}\boldsymbol{Z}^\top\right)\boldsymbol{X}. \tag{8}$$

Now we want to find $\boldsymbol{b}$ such that $\boldsymbol{z}_{k+1} = \boldsymbol{X}\boldsymbol{b}$ maximizes (5).

A property of the squared multiple correlation coefficient, (see, e.g., Ryan, 1997[9], p. 177) gives

$$\mathrm{SMC}(\boldsymbol{b}) = \mathrm{SMC}(\boldsymbol{y}, \boldsymbol{Z}) + (1 - \mathrm{SMC}(\boldsymbol{y}, \boldsymbol{Z}))\; r^2_{\boldsymbol{y},\boldsymbol{X}\boldsymbol{b}|\boldsymbol{Z}}, \tag{9}$$

where $r_{\boldsymbol{y},\boldsymbol{X}\boldsymbol{b}|\boldsymbol{Z}}$ is the partial correlation coefficient between $\boldsymbol{y}$ and $\boldsymbol{X}\boldsymbol{b}$, controlling for $\boldsymbol{Z}$. This partial correlation coefficient is nothing else but the usual

correlation coefficient between the residual vectors resulting from regressions of $\boldsymbol{y}$ on $\boldsymbol{Z}$ and from $\boldsymbol{Xb}$ on $\boldsymbol{Z}$, respectively. Similar to Equation (8), the residual vector of regressing $\boldsymbol{y}$ on $\boldsymbol{Z}$ is given by

$$\boldsymbol{y}^k = \left(\boldsymbol{I}_n - \frac{1}{n-1}\boldsymbol{Z}\boldsymbol{Z}^\top\right)\boldsymbol{y},$$

and the residual vector of regressing $\boldsymbol{z}_{k+1} = \boldsymbol{Xb}$ on $\boldsymbol{Z}$ equals

$$\left(\boldsymbol{I}_n - \frac{1}{n-1}\boldsymbol{Z}\boldsymbol{Z}^\top\right)\boldsymbol{Xb} = \boldsymbol{X}^k\boldsymbol{b},$$

using Equation (8)). Therefore the partial correlation coefficient used in (9) can be rewritten as

$$r^2_{\boldsymbol{y},\boldsymbol{Xb}|\boldsymbol{Z}} = \mathrm{Corr}^2(\boldsymbol{y}^k, \boldsymbol{X}^k\boldsymbol{b}).$$

Since $\mathrm{SMC}(\boldsymbol{y}, \boldsymbol{Z})$ in (9) does not depend on $\boldsymbol{b}$, maximizing $\mathrm{SMC}(\boldsymbol{b})$ is equivalent to maximizing

$$|\mathrm{Corr}(\boldsymbol{y}^k, \boldsymbol{X}^k\boldsymbol{b})|$$

which proofs the proposition. □

Proposition 1 has an important practical aspect. Suppose that the first k components have already been obtained. Finding the $(k+1)$th component to maximize a squared multiple correlation coefficient can be reduced to a problem of maximizing a bivariate correlation coefficient. This advantage will be used in Section 3 where we introduce an algorithm for finding components having a high value for the SMC with the response variable. Now this $(k+1)$th component also needs to be uncorrelated with the previous ones and to have unit variance. So we require $\boldsymbol{b}$ to satisfy

$$\boldsymbol{Z}^\top\boldsymbol{z}_{k+1} = \boldsymbol{Z}^\top\boldsymbol{Xb} = \boldsymbol{0} \tag{10}$$

and

$$\boldsymbol{z}_{k+1}^\top\boldsymbol{z}_{k+1} = \boldsymbol{b}^\top\boldsymbol{X}^\top\boldsymbol{Xb} = n-1. \tag{11}$$

The next proposition is helpful here.

Proposition 2. *Let $\tilde{\boldsymbol{b}}$ be a non zero p-dimensional vector. Let $\boldsymbol{B}$ be the matrix containing the coefficients for the components in $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots \boldsymbol{z}_k)$, i.e. $\boldsymbol{Z} = \boldsymbol{XB}$. Consider now*

$$\hat{\boldsymbol{b}} = c\left(\tilde{\boldsymbol{b}} - \frac{1}{n-1}\boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{X}\tilde{\boldsymbol{b}}\right) \tag{12}$$

with the scalar

$$c = \sqrt{\frac{n-1}{\tilde{\boldsymbol{b}}^\top \boldsymbol{X}^\top \boldsymbol{X}^k \tilde{\boldsymbol{b}}}}, \tag{13}$$

and with $\boldsymbol{X}^k$ *as in Proposition 1. Then* $\hat{\boldsymbol{b}}$ *verifies the side restrictions (10) and (11). Furthermore, the two vectors* $\hat{\boldsymbol{b}}$ *and* $\tilde{\boldsymbol{b}}$ *yield the same value in the maximization problem (5):*

$$SMC(\hat{\boldsymbol{b}}) = SMC(\tilde{\boldsymbol{b}}).$$

Proof. First note that, using $\boldsymbol{XB} = \boldsymbol{Z}$ and $\boldsymbol{Z}^\top \boldsymbol{Z} = (n-1)\boldsymbol{I}_k$,

$$\boldsymbol{Z}^\top \boldsymbol{X}\hat{\boldsymbol{b}} = c\boldsymbol{Z}^\top \boldsymbol{X}\tilde{\boldsymbol{b}} - \frac{c}{n-1}\boldsymbol{Z}^\top \boldsymbol{XBZ}^\top \boldsymbol{X}\tilde{\boldsymbol{b}} = c\boldsymbol{Z}^\top \boldsymbol{X}\tilde{\boldsymbol{b}} - c\boldsymbol{Z}^\top \boldsymbol{X}\tilde{\boldsymbol{b}} = \boldsymbol{0}$$

For the second side restriction (11) we use some simple algebra to get

$$\begin{aligned} \hat{\boldsymbol{b}}^\top \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{b}} &= c^2 \tilde{\boldsymbol{b}}^\top \boldsymbol{X}^\top \boldsymbol{X}\tilde{\boldsymbol{b}} - \frac{c^2}{n-1}\tilde{\boldsymbol{b}}^\top \boldsymbol{X}^\top \boldsymbol{ZZ}^\top \boldsymbol{X}\tilde{\boldsymbol{b}} \\ &= c^2 \tilde{\boldsymbol{b}}^\top \boldsymbol{X}^\top \left(\boldsymbol{I} - \frac{1}{n-1}\boldsymbol{ZZ}^\top\right) \boldsymbol{X}\tilde{\boldsymbol{b}} = c^2 \tilde{\boldsymbol{b}}^\top \boldsymbol{X}^\top \boldsymbol{X}^k \tilde{\boldsymbol{b}} = n-1 \end{aligned}$$

using (13), which corresponds indeed to (11).

Finally, we have to show that $\hat{\boldsymbol{b}}$ and $\tilde{\boldsymbol{b}}$ yield the same value in (6). We see that

$$\boldsymbol{X}^k \hat{\boldsymbol{b}} = c\ \boldsymbol{X}^k \tilde{\boldsymbol{b}} - \frac{c}{n-1}\boldsymbol{X}^k \boldsymbol{B}\boldsymbol{Z}^\top \boldsymbol{X}\tilde{\boldsymbol{b}}.$$

Using (8) we have

$$\boldsymbol{X}^k \boldsymbol{B} = \left(\boldsymbol{I}_n - \frac{1}{n-1}\boldsymbol{ZZ}^\top\right) \boldsymbol{XB} = \boldsymbol{Z} - \boldsymbol{Z}\frac{1}{n-1}\boldsymbol{Z}^\top \boldsymbol{Z} = 0,$$

and thus

$$\boldsymbol{X}^k \hat{\boldsymbol{b}} = c\ \boldsymbol{X}^k \tilde{\boldsymbol{b}}.$$

The proof is complete with

$$\mathrm{Corr}(\boldsymbol{y}^k, \boldsymbol{X}^k \tilde{\boldsymbol{b}}) = \mathrm{Corr}(\boldsymbol{y}^k, c\ \boldsymbol{X}^k \tilde{\boldsymbol{b}}) = \mathrm{Corr}(\boldsymbol{y}^k, \boldsymbol{X}^k \hat{\boldsymbol{b}}),$$

and using the result of Proposition 1. □

Due to the fact that the predictor variables are orthogonal to each other, it is not difficult to check that (7) can be rewritten, for $k > 1$, as

$$\boldsymbol{X}^k = (I_n - \frac{\boldsymbol{z}_{k-1}\boldsymbol{z}_{k-1}^\top}{n-1})\boldsymbol{X}^{k-1}.$$

The residuals in $\boldsymbol{X}^k$ can thus be obtained as simple bivariate regression of the columns of $\boldsymbol{X}^{k-1}$ on $\boldsymbol{z}_{k-1}$. Similarly, $\boldsymbol{y}^k$ is given by the residuals of regressing $\boldsymbol{y}^{k-1}$ on $\boldsymbol{z}_{k-1}$. Again, no multiple regressions are needed here.

In the next section we will outline a sequential algorithm for finding predictor variables having a high SMC with the response variable. We stress again that the linear combination of the original explicative variables giving the theoretical maximal value of correlation with the dependent variable would be determined by the coefficients of the LS estimator. But due to the multicollinearity problem or the case that $p > n$ mentioned before, we will not aim at a direct computation of this LS estimator.

3. Algorithm

For finding the first predictor variable $\boldsymbol{z}_1$ we have to find a vector $\boldsymbol{b}$ such that

$$\boldsymbol{b} \longrightarrow |\mathrm{Corr}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{b})| \tag{14}$$

is maximized under the condition $\boldsymbol{z}_1^\top \boldsymbol{z}_1 = n - 1$. Rather than looking for the global maximum we approximate the solution by restricting to the set of candidate solutions to

$$B_{n,1} = \left\{ \tilde{\boldsymbol{b}}_{j,1} = \boldsymbol{x}_j \,|\, j = 1, \ldots, n \right\}. \tag{15}$$

The set $B_{n,1}$ is the collection of vectors pointing at the data, and can be thought of as a collection of potentially interesting directions. This kind of approximate algorithm was used by Croux and Ruiz-Gazen (1996)[4] for principal component analysis, and it requires $O(n^2)$ computation time. For very large values of n, one could pass to a subset of $B_{n,1}$, whereas for very small n additional directions can be simulated.

We will transform every candidate solutions in order to fulfill the restriction of a unit variance for the first component. In analogy to (12) we just need to set

$$\hat{\boldsymbol{b}}_{j,1} = c\ \tilde{\boldsymbol{b}}_{j,1}$$

with

$$c = \sqrt{\frac{n-1}{\tilde{\boldsymbol{b}}_{j,1}^\top \boldsymbol{X}^\top \boldsymbol{X} \tilde{\boldsymbol{b}}_{j,1}}} = \sqrt{\frac{n-1}{\boldsymbol{x}_j^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{x}_j}},$$

to have $\boldsymbol{z}_1^\top \boldsymbol{z}_1 = (\boldsymbol{X}\hat{\boldsymbol{b}}_{j,1})^\top(\boldsymbol{X}\hat{\boldsymbol{b}}_{j,1}) = n-1$. Then we simply take the maximal value of (14) over the candidates $\hat{\boldsymbol{b}}_{j,1}$, i.e.

$$\boldsymbol{b}_1 = \underset{j=1,\ldots,n}{\operatorname{argmax}} |\operatorname{Corr}(\boldsymbol{y}, \boldsymbol{X}\hat{\boldsymbol{b}}_{j,1})|, \tag{16}$$

and the first predictor variable is constructed as $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{b}_1$. Note that the multicollinearity problem is avoided, since no least squares estimator on the original predictor variables is computed here.

For finding the second component $\boldsymbol{z}_2$, we want to to maximize the SMC between $\boldsymbol{y}$ and $(\boldsymbol{z}_1, \boldsymbol{z}_2)$. Following Proposition 1, we first regress all variables $\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ on the first component $\boldsymbol{z}_1$, just by means of a sequence of $p+1$ simple bivariate regressions. We will continue then to work with the residual vectors $\boldsymbol{y}^1$ and $\boldsymbol{X}^1 = (\boldsymbol{x}_1^1, \ldots, \boldsymbol{x}_p^1)$ and try to find $\boldsymbol{b}$ such that

$$\boldsymbol{b} \longrightarrow |\operatorname{Corr}(\boldsymbol{y}^1, \boldsymbol{X}^1\boldsymbol{b})| \tag{17}$$

is maximized under $\boldsymbol{z}_1^\top \boldsymbol{z}_2 = 0$ and $\boldsymbol{z}_2^\top \boldsymbol{z}_2 = n-1$. Similar as before we consider the set of candidate directions

$$B_{n,2} = \left\{ \tilde{\boldsymbol{b}}_{j,2} = \boldsymbol{x}_j^1; j = 1, \ldots, n \right\} \tag{18}$$

for an approximative solution. $B_{n,2}$ is a collection of vectors pointing at the data in the residual space $\boldsymbol{X}^1$. To ensure that the new component will be uncorrelated with the previous one and have unit variance, we use Proposition 2 and transform the vectors of $B_{n,2}$ into

$$\hat{\boldsymbol{b}}_{j,2} = c \left(\boldsymbol{x}_j^1 - \frac{1}{n-1} \boldsymbol{b}_1 \boldsymbol{z}_1^\top \boldsymbol{X} \boldsymbol{x}_j^1 \right)$$

with

$$c = \sqrt{\frac{n-1}{\boldsymbol{x}_j^{1\top} \boldsymbol{X}^\top \boldsymbol{X}^1 \boldsymbol{x}_j^1}}.$$

Now the function (17) is evaluated over the candidates $\hat{\boldsymbol{b}}_{j,2}$, for $1 \le j \le n$ and with

$$\boldsymbol{b}_2 = \underset{j=1,\ldots,n}{\operatorname{argmax}} |\operatorname{Corr}(\boldsymbol{y}^1, \boldsymbol{X}^1\hat{\boldsymbol{b}}_{j,2})|, \tag{19}$$

we obtain the second predictor variable $\boldsymbol{z}_2 = \boldsymbol{X}\boldsymbol{b}_2$.

In a similar way, the other components $\boldsymbol{z}_3, \ldots, \boldsymbol{z}_k$ are sequentially obtained. Note that, according to (7) the computation of the residuals in $\boldsymbol{y}^k$ and $\boldsymbol{X}^k$ do not require any matrix inversion.

4. Simulation

In this section we want to compare the proposed method with the regression method and with PCR. It should be demonstrated that for our method the number of selected predictor variables in the model is indeed lower than for the other classical methods. Therefore, we generate a data set $\boldsymbol{X}_1$ with $n = 500$ samples in dimension $p_1 = 50$ from a specified $N(\mathbf{0}, \boldsymbol{\Sigma})$ distribution. For obtaining collinearity we generate $\boldsymbol{X}_2 = \boldsymbol{X}_1 + \boldsymbol{\Delta}$, and the columns of the noise matrix $\boldsymbol{\Delta}$ are independently distributed according to $N(0, 0.001)$. Both matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are combined in the matrix of independent variables $\boldsymbol{X} = (\boldsymbol{X}_1|\boldsymbol{X}_2)$. Furthermore, we generate a dependent variable as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\delta}$. The first 25 elements of the vector $\boldsymbol{a}$ are generated from a uniform distribution in the interval $[-1, 1]$, and the remaining elements of $\boldsymbol{a}$ are 0. The variable $\boldsymbol{\delta}$ comes from the distribution $N(0, 0.8)$. So, $\boldsymbol{y}$ is a linear combination of the first 25 columns of $\boldsymbol{X}_1$ plus an error term.

As an output of the simulation study we choose the SMC coefficient between $\boldsymbol{y}$ and the predictor variables, and this coefficient can be compared over the different methods. First of all, we compare the proposed method with a stepwise variable procedure which selects the predictor variables out of the columns of $\boldsymbol{X}$. Furthermore, we compare with two different approaches to PCR: once the predictor variables (which are the PCs) are selected according to the magnitude of their variances (*sequential selection*), and once the predictors are selected due to the largest increase of the SMC coefficient (*stepwise selection*).

We computed $m = 1000$ replications and a maximum number $l = 20$ of predictor variables. The results are summarized by computing the average SMC over all m replications. Denote $R^2_{LS}(t, j)$ the resulting SMC coefficient of the t-th replication if j regressors are considered in the model. Then

$$R^2_{LS}(j) = \frac{1}{m} \sum_{t=1}^{m} R^2_{LS}(t, j), \tag{20}$$

for each number of regressors $j = 1, \ldots, l$. Figure 1 shows the mean SMC values $R^2_{LS}(j)$ for each considered number j of regressors. We find that the proposed method gives a higher mean coefficient of determination especially for a low number of predictor variables which is most desirable. PCR with sequential selection gives the worst results: for obtaining the same mean coefficient of determination, one would have to take considerably more predictor variables in the model than for our proposed method.

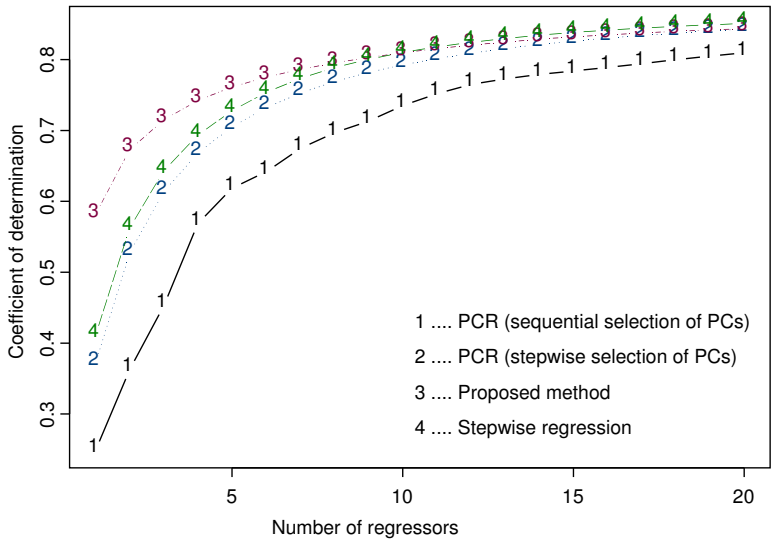

Figure 1: Comparison of (1) PCR with sequential selection of PCs, (2) PCR with stepwise selection of PCs, (3) the proposed method, and (4) stepwise variable selection. The mean coefficient of determination is drawn against the number of predictor variables.

5. Conclusions

Multiple linear regression is a standard method in statistics which is frequently used. However, for many applications the number of observation is even smaller than the number of variables in the x-part which makes the usage of this simple method impossible. Also, problems can occur if the regressor variables are highly correlated. A simple solution to these problems is to take PCR because this method constructs uses PCs as regressors. But they are not optimized with respect to the prediction of the dependent variable since the PCs do not take any information of $\boldsymbol{y}$ into account.

Like PCR, the proposed method constructs uncorrelated predictor variables which are, however, selected to have high values for the SMC with the response variable. We propose a simple and fast method for selecting the predictors which are linear combinations of $\boldsymbol{X}$. The simulation study has demonstrated that the proposed method allows a major reduction of the predictor variables compared to PCR and multiple regression. This is an important advantage because in most applications the regression model should be kept as simple as possible.

The proposed method can easily be robustified. Recall that the outlined algorithm only uses computations of bivariate correlation measures and bivariate regressions. These could be replaced by robust counterparts. A popular robust regression estimator is the Least Trimmed Squares estimator (Rousseeuw and Leroy, 1987[10]), becoming much faster to compute in the simple regression model. Robust correlation measures have been considered for example in Croux and Dehon (2002)[3]. This robust version of the proposed method is preferable to robust PCR (Filzmoser, 2001[5]) because in general a lower number of predictors can be used in the model.

Often it is important to find a simple interpretation of the regression model. Since PCR as well as our proposed method are searching for linear combinations of the x-variables, the resulting predictor variables will in general not be easy to interpret. In practice it is not always necessary to find an interpretation of the predictors, only the functional relation between $\boldsymbol{y}$ and $\boldsymbol{X}$ is important. However, if an interpretation is desired, one has to switch to other methods which can deal with collinear data, like ridge regression (Hoerl and Kennard, 1970[7]). There are also interesting developments of methods in the chemometrics literature. Araújo et al. (2001)[1] introduced a projection algorithm for sequential selection of x-variables in problems with collinearity and with very large numbers of x-variables.

REFERENCES

[1] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H.C. Chame. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems* **57** (2001), 65–73.

[2] A. Basilevsky. Statistical factor analysis and related methods: Theory and applications. Wiley & Sons, New York, 1994.

[3] C. Croux, C. Dehon. Estimators of the multiple correlation coefficient: Local robustness and confidence intervals. *Statistical Papers*, (2002), to appear, `http://www.econ.kuleuven.ac.be/christophe.croux`.

[4] C. Croux, A. Ruiz-Gazen. A fast algorithm for robust principal components based on projection pursuit. In: Computational Statistics (ed. A. Prat), Physica-Verlag, Heidelberg, 1996, 211–216.

[5] P. Filzmoser. Robust principal component regression. In: Computer data analysis and modeling. Robust and computer intensive methods. (eds. S. Aivazian, Yu. Kharin, and H. Rieder), Belarusian State University, Minsk, (2001), 132–137.

[6] A. S. Hadi, R. F. Ling. Some cautionary notes on the use of principal components regression. *The American Statistician* **1** (1998), 15–19.

[7] A. E. Hoerl, R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** (1970), 55–67.

[8] R. Johnson, D. Wichern. Applied multivariate statistical analysis. 4th edition, Prentice-Hall, London, 1998.

[9] T. P. Ryan. Modern regression methods. Wiley & Sons, New York, 1997.

[10] P. J. Rousseeuw, A. M. Leroy. Robust regression and outlier detection. Wiley & Sons, New York, 1987.

Peter Filzmoser
Department of Statistics and Probability Theory,
Vienna University of Technology,
Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria
e-mail: P.Filzmoser@tuwien.ac.at

C. Croux
Department of Applied Economics
K. U. Leuven
Naamsestraat 69, B-3000 Leuven