A Projection Algorithm for Regression with Collinearity

Peter Filzmoser[1] and Christophe Croux[2]

[1] Department of Statistics and Probability Theory,
Vienna University of Technology,
Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria
[2] Department of Applied Economics, K.U.Leuven
Naamsestraat 69, B-3000 Leuven

Abstract. Principal component regression (PCR) is often used in regression with multicollinearity. Although this method avoids the problems which can arise in the least squares (LS) approach, it is not optimized with respect to the ability to predict the response variable. We propose a method which combines the two steps in the PCR procedure, namely finding the principal components (PCs) and regression of the response variable on the PCs. The resulting method aims at maximizing the coefficient of determination for a selected number of predictor variables, and therefore the number of predictor variables can be reduced compared to PCR. An important feature of the proposed method is that it can easily be robustified using robust measures of correlation.

1 Introduction

We consider the standard multiple linear regression model with intercept,

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

where n is the sample size, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ 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.

Problems can occur when the predictor variables are highly correlated, this situation is called *multicollinearity*. The inverse of $\boldsymbol{X}^\top\boldsymbol{X}$ which is needed to compute the least squares (LS) estimator $\hat{\boldsymbol{\beta}}_{LS}$ of $\boldsymbol{\beta}$, becomes ill-conditioned and is numerically unstable. This matrix is also used for computing the standard errors of the LS regression coefficients and for the correlation matrix of the regression coefficients. In a near-singular case they can be inflated considerably and cause doubt on the interpretability of the regression coefficients.

A number of techniques have been proposed when collinearity exists among the predictors. One possibility is principal component regression (PCR) (see, e.g., Basilevsky, 1994) where principal components (PCs) obtained from the predictors $\boldsymbol{X}$ are used within the regression model. Most of the problems

mentioned above are then being avoided. 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 of PCR is to simplify the regression model by taking a reduced number of PCs in the prediction set. We could simply take those first $k < p$ PCs in the regression model having the largest variances (*sequential selection*) but often PCs with smaller variances are higher correlated with the response variable. Hence, it might be more advisable to set up the PCR model by a *stepwise selection* of PCs due to an appropriate measure of association with the response variable. In more detail, in the first step we could search for that PC having the largest *Squared Multiple Correlation* (SMC) with the response variable, in the second step search for an additional PC resulting in the largest SMC, and so on. In fact, due to the uncorrelatedness of the principal components, this comes down to selecting the k components having the largest squared (bivariate) correlations with the dependent variable.

PCR can deal with multicollinearity, but it is not a method which directly maximizes the correlation between the original predictors and the response variable. It was noted by Hadi and Ling (1998) that in some situations PCR can give quite low values for the SMC. PCR is a two-step procedure: in the first step one computes PCs which are linear combinations of the x-variables, and in the second step the response variable is regressed on the (selected) PCs in a linear regression model. For maximizing the relation to the response variable we could combine both steps in a single method. This method has to find $k < p$ predictor variables z_j ($j = 1, \ldots, k$) which are linear combinations of the x-variables and have high values of the SMC with the y-variable. Note that the linear combination of the predictor variables giving the theoretical maximal value of SMC with the dependent variable is determined by the coefficients of the LS-estimator. Of course, due to the multicollinearity problem mentioned before, we will not aim at a direct computation of this LS-estimator.

2 Algorithm

The idea behind the algorithm is to find k components $z_1, \ldots, z_k$ having the property that the Squared Multiple Correlation between y and the components is as high as possible, under the constraints that these components are mutually uncorrelated and have unit variance. Under these contraints, it is easy to check that

$$SMC = \mathrm{Corr}^2(y, z_1) + \mathrm{Corr}^2(y, z_2) + \ldots + \mathrm{Corr}^2(y, z_k).$$

We will try to optimize the above SMC in a sequential manner. First by selecting a z_1 having maximal squared correlation with the dependent variable, and then by sequentially finding the other components having maximal correlation with y while still verifying the imposed side restrictions. Below

we propose an easy heuristical algorithm yielding a good approximation to the solution of the stated maximization problem under contraints.

For finding the first predictor variable z_1, we look for a vector $\boldsymbol{b}$ resulting in a high value of the function

$$\boldsymbol{b} \to |\mathrm{Corr}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{b})|. \tag{2}$$

The correlation in (2) is the usual sample correlation coefficient between two column vectors. Since the value of the objective function in (2) is invariant with respect to scalar multiplication of $\boldsymbol{b}$, we add the constraint that $\mathrm{Var}(\boldsymbol{X}\boldsymbol{b}) = 1$. The maximum of (2) would be obtained by choosing $\boldsymbol{b}$ as the LS estimator $\hat{\boldsymbol{\beta}}_{LS}$ due to model (1). However, to avoid the multicollinearity problem, we are not looking at the global maximum of this function, but we restrict ourselves at evaluating (2) at the discrete set

$$B_{n,1} = \left\{ \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}; i = 1, \ldots, n \right\}.$$

(Similar as in the algorithm of Croux and Ruiz-Gazen (1996) for principal component analysis.) The scores of the first component are then simply given by the vector $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{b}_1$, where $\boldsymbol{b}_1$ is the value maximizing the function (2) over the set $B_{n,1}$. Afterwards, $\boldsymbol{b}_1$ is rescaled in order to verify the side restriction of having unit sample variance for the scores of the first component. 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. Note that finding $\boldsymbol{b}_1$ can be done without any numerical difficulty, and in $O(n^2)$ computation time. Even in the case of more variables than observations ($p > n$), which is of interest for example in spectroscopy, the variable $\boldsymbol{z}_1$ can easily be computed. For very large values of n, one could pass to a subset of $B_{n,1}$.

For finding the scores of the second component $\boldsymbol{z}_2$, we need to restrict to the space of all vectors having sample correlation zero with $\boldsymbol{z}_1$. Denote $\boldsymbol{X}_j$ the j-th column of the data matrix $\boldsymbol{X}$, containing the realizations of the variable x_j with $1 \leq j \leq p$. Herefore we regress all data vectors $\boldsymbol{y}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ on the already obtained first component $\boldsymbol{z}_1$, just by means of a sequence of $p+1$ simple bivariate regressions. Since all these regressions are bivariate, they cannot imply any multicollinearity problems. We will continue then to work with the residual vectors obtained from these regressions, which we denote by $\boldsymbol{y}^1, \boldsymbol{X}_1^1, \ldots, \boldsymbol{X}_p^1$. Note that all these vectors are uncorrelated with $\boldsymbol{z}_1$. With $\boldsymbol{X}^1 = (\boldsymbol{X}_1^1, \ldots, \boldsymbol{X}_p^1)$, the second predictor variable is found by maximizing the function

$$\boldsymbol{b} \to |\mathrm{Corr}(\boldsymbol{y}^1, \boldsymbol{X}^1\boldsymbol{b})|. \tag{3}$$

The maximum in (3) can in principle be achieved by taking the LS estimator for $\boldsymbol{b}$. Using the statistical properties of LS estimators, it can be seen that this would yield a SMC value between y and z_1, z_2 equal to the SMC between y and the x-variables. But, again, since we are concerned with the collinearity

problem, we will approximate the solution of (3) by searching only in the set $B_{n,2} = \left\{ \frac{\boldsymbol{x}_i^1}{\|\boldsymbol{x}_i^1\|}; i = 1, \ldots, n \right\}$ where $\boldsymbol{x}_i^1$ denotes the i-th row vector of $\boldsymbol{X}^1$. The vector $\boldsymbol{b}_2$ maximizing the function (3) over the set $B_{n,2}$ defines, after rescaling to get unit variance, the scores on the second component $\boldsymbol{z}_2 = \boldsymbol{X}^1\boldsymbol{b}_2$. Since we passed to the space of residuals, the first two components will be uncorrelated, as required.

Note that if we would have worked with the theoretical maximum in (2) of the first step, then the objective function (3) would be equal to zero, since LS residuals are orthogonal to the explicative variables. In the latter case, there is no correlation left to explain after the first step. But since we are only approximating the solution, and not computing the LS-estimator directly, we will still have a non-degenerate solution to (3). A comparison of the numerical values of the maxima in (2) and (3) tells us how much explicative power is gained by adding z_2 to the model.

The other components $\boldsymbol{z}_3, \ldots, \boldsymbol{z}_k$ are now obtained in an analogous way as $\boldsymbol{z}_2$. Component $\boldsymbol{z}_l$ $(l = 3, \ldots, k)$ is found by maximizing

$$\boldsymbol{b} \rightarrow |\mathrm{Corr}(\boldsymbol{y}^{l-1}, \boldsymbol{X}^{l-1}\boldsymbol{b})| \tag{4}$$

where $y^{l-1}, \boldsymbol{X}^{l-1} = (\boldsymbol{X}_1^{l-1}, \ldots, \boldsymbol{X}_p^{l-1})$ are obtained by regressing the previously obtained residual series $\boldsymbol{y}^{l-2}, \boldsymbol{X}_1^{l-2}, \ldots, \boldsymbol{X}_p^{l-2}$ on component $\boldsymbol{z}_{l-1}$. We approximate the solution of (4) by considering only the n candidate vectors of the set $B_{n,l} = \left\{ \frac{\boldsymbol{x}_i^{l-1}}{\|\boldsymbol{x}_i^{l-1}\|}; i = 1, \ldots, n \right\}$.

3 Example

We consider a data set from geochemistry which is available in form of a geochemical atlas (Reimann et al., 1998). An area of 188000 km^2 in the so-called Kola region at the boundary of Norway, Finland, and Russia was sampled. More than 50 chemical elements have been analyzed for all 606 samples. For some of the most interesting elements like gold (Au) it is not possible to obtain reliable estimations of the concentration because often the concentration is below the detection limit. It would thus be advantageous to estimate the contents of the "rare" elements by using the information of the other elements. Similar chemical structures of rocks and soil allow a dependency among the chemical elements which can sometimes be very strong. This leads to a regression problem with multicollinearity. Filzmoser (2001) used a robust PCR method to predict the contents of these "rare" elements. Here we apply our proposed algorithm to predict the concentration of cromium (Cr) using 54 other chemical elements as predictors. Cr belongs certainly not to the "rare" elements, but it is strongly related to many other elements, and hence it is suitable for testing our method and comparing it with conventional PCR. Figure 1 shows the comparison of (1) PCR with sequential selection

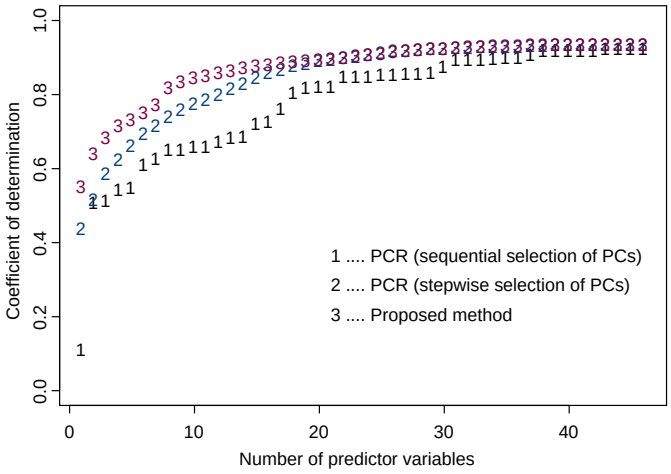

Fig. 1. Comparison of (1) PCR with sequential selection of PCs, (2) PCR with stepwise selection of PCs, (3) proposed method. The coefficient of determination is drawn against the number of predictor variables.

of PCs, (2) PCR with stepwise selection of PCs, and (3) the newly proposed method. The coefficient of determination is drawn against the number of predictor variables used in the linear model. Figure 1 shows that for each fixed number of predictor variables the coefficient of determination was highest for the new method. As already expected, the sequential selection of PCs according to their largest variances gives the lowest coefficient of determination. Our method would therefore allow a major reduction of the number of explanatory variables in the regression model.

4 Simulation study

We want to compare the proposed method with the classical multiple regression approach, and with PCR regression. Therefore, we generate a data set $\boldsymbol{X}_1$ with $n = 500$ samples in dimension $p_1 = 50$ from a specified $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ 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.

In the simulation we consider PCR of $\boldsymbol{y}$ on $\boldsymbol{X}$ by sequentially selecting the PCs according to the magnitude of their eigenvalues and by stepwise

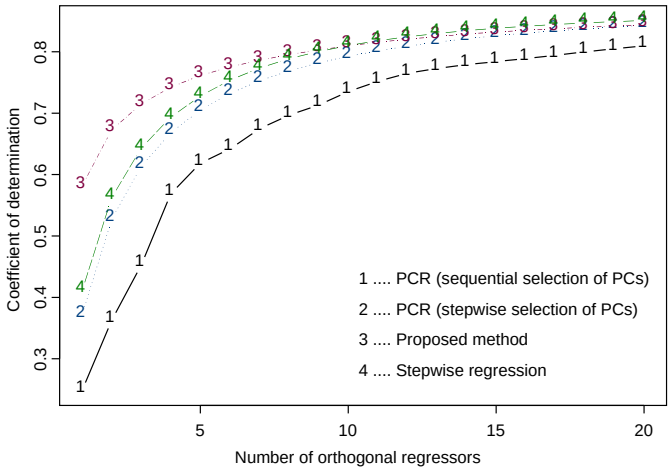

Fig. 2. Comparison of (1) PCR with sequential selection of PCs, (2) PCR with stepwise selection of PCs, (3) the proposed method, and (4) stepwise regression. The mean coefficient of determination is drawn against the number of predictor variables.

selecting the PCs according to the largest increase of the R^2 measure, our proposed regression method, and stepwise regression (forward selection of the predictor variables). We computed $m = 1000$ replications and a maximum number $k = 20$ of predictor variables.

We can summarize the resulting coefficients of determination 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

$$R^2_{LS}(j) = \frac{1}{m} \sum_{t=1}^{m} R^2_{LS}(t, j), \tag{5}$$

for each number of regressors $j = 1, \ldots, k$. Figure 2 shows the mean coefficient of determination for each considered number j of regressors. We find that our 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.

5 Discussion

The two-step procedure of PCR, namely computing the PCs of the predictor variables and performing regression of the response variable on the PCs can

be reduced to a single step procedure. Like PCR, the proposed method is able to deal with the problem of multicollinearity, but the new predictor variables which are linear combinations of the original x-variables lead in general to a higher coefficient of determination compared to PCR. Since one usually tries to explain the main variability of the response variable by a possibly low number of predictor variables, the proposed method is preferable to PCR, and also to the stepwise regression technique as was shown by the simulation study.

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. Our example has shown that the interpretation of the predictor variables is not always necessary. 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). There are also interesting developments of methods in the chemometrics literature. Araújo et al. (2001) introduced a projection algorithm for sequential selection of x-variables in problems with collinearity and with very large numbers of x-variables.

An important advantage of our proposed method is that it can easily be robustified. It is well known that outliers in the y-variable and/or in the x-variables can have a severe influence on regression estimates, even for bivariate regressions. Hence, robust regression techniques like least median of squares (LMS) regression or least trimmed squares (LTS) regression (Rousseeuw, 1984) have been developed which can resist the effect of outliers. The classical correlations used in (2) and (3) can also be replaced by robust versions. Note that Croux and Dehon (2001) introduced robust measures of the multiple correlation.

References

ARAÚJO, M.C.U., SALDANHA, T.C.B., GALVÃO, R.K.H., YONEYAMA, T., and CHAME, H.C. (2001): The Successive Projections Algorithm for Variable Selection in Spectroscopic Multicomponent Analysis. *Chemometrics and Intelligent Laboratory Systems*, *57*, 65–73.

BASILEVSKY, A. (1994): *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley & Sons, New York.

CROUX, C. and DEHON, C. (2001): Estimators of the Multiple Correlation Coefficient: Local Robustness and Confidence Intervals, to appear in *Statistical Papers*, `http://www.econ.kuleuven.ac.be/christophe.croux`.

CROUX, C. and RUIZ-GAZEN, A. (1996): A Fast Algorithm for Robust Principal Components Based on Projection Pursuit. In: A. Prat (ed.): *Computational Statistics*. Physica-Verlag, Heidelberg, 211–216.

FILZMOSER, P. (2001): Robust Principal Compnent Regression. In: S. Aivazian, Yu. Kharin, and H. Rieder (Eds.): *Computer Data Analysis and Modeling.*

Robust and Computer Intensive Methods. Belarusian State University, Minsk, 132–137.

HADI, A.S. and LING, R.F. (1998): Some Cautionary Notes on the Use of Principal Components Regression. *The American Statistician*, *1*, 15–19.

HOERL, A.E. and KENNARD, R.W. (1970): Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*, 55–67.

REIMANN, C., ÄYRÄS, M., CHEKUSHIN, V., BOGATYREV, I., BOYD, R., DE CARITAT, P., DUTTER, R., FINNE, T.E., HALLERAKER, J.H., JÆGER, Ø., KASHULINA, G., LEHTO, O., NISKAVAARA, H., PAVLOV, V., RÄISÄNEN, M.L., STRAND, T., and VOLDEN, T. (1998): *Environmental Geochemical Atlas of the Central Barents Region*. NGU-GTK-CKE special publication. Geological Survey of Norway, Trondheim.

ROUSSEEUW, P.J. (1984): Least Median of Squares Regression. *Journal of the American Statistical Association*, *79*, 871–880.