A robustified version of the SIMPLS algorithm

K. Vanden Branden[1], and M. Hubert[1]

[1] Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium

Keywords: Partial Least Squares Regression, SIMPLS, Robust covariance matrix, Robust multivariate regression.

1 Introduction

A well-known problem in the field of chemometrics is to estimate a linear relationship between two sets of variables. The independent variables X $(n \times p)$ can be very numerous (some hundreds, thousands), while the number of observations is typically very small (some tens). Also the number of dependent variables Y $(n \times q)$ is in general limited to at most five. This problem leads to the multivariate regression model

$$Y = \alpha + X\beta + \epsilon$$

with $p > n$, intercept term α $(n \times q)$, slopes β $(p \times q)$ and error term ϵ $(n \times q)$. Because p is larger than n, the inverse of $X'X$ does not exist and hence we can not perform a least squares regression. Therefore Partial Least Squares (PLS) regression (Tenenhaus, 1998) has been developed to estimate the parameters of this model.

PLS regression mainly consists of two steps. In the first stage a matrix of scores $T = [t_1, t_2, \ldots, t_k]$ is obtained with k the number of components we want to retain in the final regression. The calculation of these scores t_h is essentially based on an empirical covariance matrix. In the second stage the multivariate least squares regression of Y on the scores matrix T is performed.

Since both stages are very sensitive to outliers in the data, we propose a robust PLS algorithm based on a robust covariance matrix in high dimensions and a robust multivariate regression method.

2 Partial Least Squares Regression

Two popular algorithms for PLS regression are PLS2 and SIMPLS (de Jong, 1993). The motivation is to maximize a covariance criterion under certain restrictions. In both algorithms the X and Y variables are first mean centered. In the first stage of PLS2 the data matrix X is deflated in each step. This results in scores t_h, $h = 1 \ldots, k$, that are linear combinations of this deflated matrix and not of the original mean centered data matrix X. The interpretation of the scores matrix T is therefore not straightforward.

The SIMPLS algorithm avoids this problem by deflating the sample covariance matrix $S_0 = X'Y$ in every step of the algorithm. In the first step of this algorithm normalized weights r_h and q_h are calculated such that the covariance between the scores $t_h = Xr_h$ and $u_h = Yq_h$ is maximized under the condition that $t_l't_k = 0$ for $l > k$. The solution of this problem is known, i.e. the weights r_1 and q_1 are the first left and right eigenvectors from the singular value decomposition (SVD) of S_0. The other weights r_h and q_h, $h = 2, \ldots, k$, are given by the first pair of singular values from the SVD of a deflated covariance matrix $S_{h-1} = P_{h-1}^{\perp} S_{h-2}$ with $P_{h-1}^{\perp}$ the orthogonal projector on the space spanned by $[X't_1, \ldots, X't_{h-1}]$.

Finally, in the second stage, the multivariate least squares regression of Y on T is performed. This leads to the estimators of the slopes β and the intercept α.

3 A robustified version of the SIMPLS algorithm

We introduce a robust version of the SIMPLS algorithm by robustifying the two main stages of the SIMPLS algorithm. In the first stage we propose to obtain the weights r_1 and q_1 as the first pair of singular values of the SVD of a robust covariance matrix. For this we apply the ROBPCA algorithm (Hubert and Rousseeuw, 2002) to the data (X, Y). Note that here X and Y are the original data matrices. This results in a robust covariance matrix $\hat{\Sigma}$ and a robust center $\hat{\mu}$:

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{XX} & \hat{\Sigma}_{XY} \\ \hat{\Sigma}_{YX} & \hat{\Sigma}_{YY} \end{pmatrix} \qquad \hat{\mu} = \begin{pmatrix} \hat{\mu}_X \\ \hat{\mu}_Y \end{pmatrix}.$$

The classical covariance matrix S_0 in the SIMPLS algorithm is then replaced by $\hat{\Sigma}_{XY}$. The scores t_h and u_h are now obtained as linear combinations of the robust centered matrices X en Y, namely $t_h = (X - 1_n\hat{\mu}_X')r_h$ and $u_h = (Y - 1_n\hat{\mu}_Y')q_h$.

A second robustification involves the multivariate least squares regression in the second stage. We replace this regression by a robust multivariate regression method based on the MCD estimator of location and scatter, the so called MCD-regression method (Rousseeuw et al., 2000).

Simulations and examples on real data sets demonstrate the robustness of this algorithm.

References

S. de Jong (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18, 251–263.

M. Hubert, and P.J. Rousseeuw (2002). ROBPCA: a new approach to robust principal component analysis. Submitted.

P.J. Rousseeuw, S. Van Aelst, K. Van Driessen, and A. Agulló (2000). Robust Multivariate Regression. Submitted.

M. Tenenhaus (1998). *La Régression PLS.* Éditions Technip, Paris.

Please fill in this form and mail it together with your abstract.

My abstract fits best to topic number 20

List of Topics:

1. Algorithms
2. Applications
3. Biostatistics
4. Computing and graphics
5. Data analysis
6. Data mining
7. Economics, finance
8. Efficiency and robustness
9. Functionals and bias
10. Fuzzy statistics
11. Geostatistics
12. Inference for robust methods, model testing
13. Location depth and regression depth
14. Multivariate methods
15. Neural networks
16. Rank-based methods
17. Regression quantiles, trimming
18. Robust covariance
19. Robust designs
20. Robust regression
21. Time series analysis
22. Wavelets
23. Other (please specify)