Robust calibration based on a robust covariance estimator in high dimensions

M. Hubert[1], S. Verboven[2], and K. Vanden Branden[1]

[1] Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium
[2] Universitaire Instelling Antwerpen, Universiteitsplein 1, B-2610 Antwerpen, Belgium

Keywords: Robust covariance, Principal Component Regression, Partial Least Squares Regression, Robust multivariate regression.

1 Introduction

Calibration is a very important statistical method in chemometrics. The measurements typically consist of spectra of several specimen (the X variables) and the concentration of some of their constituents (the Y variables). The number of channels or energy intervals in the spectra is usually very large, whereas the number of observations is often limited to at most 100 and the number of response variables to at most five.

We are thus faced with a linear model

$$Y_{n,q} = \beta_0 + X_{n,p}\beta_{p,q} + \varepsilon_{n,q}$$

with $p \gg n$ and $q \ll n$. Since X^tX is singular if $p > n$, the least squares estimator cannot be computed at this model. Therefore, biased estimators are frequently used, such as Principal Component Regression (PCR) and Partial Least Squares Regression (PLS). Both methods first obtain $k \ll p$ scores T as linear combinations of the original regressors, and then regress Y on T.

Although PCR and PLS are very popular in chemometrics, they are not resistant to outlying observations. Here, we propose two robust counterparts.

2 Robust PCR

To robustify PCR, we first need a robust PCA method which has to be applied on the X-variables. For this, we use the ROBPCA method (Hubert and Rousseeuw, 2002) since it is fast, transparent and highly robust. Assume that we select k principal components, then the scores matrix $T_{n,k}$ contains the projections of the observations on these components.

In the second stage of the algorithm, we apply the MCD-regression method (Rousseeuw et al., 2000) on the reduced regression model:

$$Y_{n,q} = \alpha_0 + T_{n,k}\alpha_{k,q} + \varepsilon'.$$

MCD-regression is a multivariate regression method whose breakdown value is as high as the MCD estimator (Rousseeuw, 1984). The regression coefficients $(\hat{\alpha_0}, \hat{\alpha})$ obtained in this stage are then backtransformed to the original p-dimensional parameter space, yielding estimates for β_0 and β.

3 Robust PLS

In PLS regression the computation of the scores takes into account the covariance between X and Y. We propose a robust version of the SIMPLS algorithm (de Jong, 1993). Leaving out the details of SIMPLS, it starts by computing the sample covariance matrix between X and Y. In our method we replace this classical covariance matrix by a robust covariance matrix. It is obtained by applying ROBPCA on X and Y simultaneously. This yields

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{XX} & \hat{\Sigma}_{XY} \\ \hat{\Sigma}_{YX} & \hat{\Sigma}_{YY} \end{pmatrix}$$

from which $\hat{\Sigma}_{XY}$ can be extracted. In the regression step, we again apply the MCD-regression method.

4 Selection of number of components

To select the number of components, both in PCR and PLS, we propose to make a plot of the robust R_j^2 versus j for $j = 1, \ldots, k_{\max}$:

$$R_j^2 = 1 - \frac{\det(\hat{\Sigma}_{\varepsilon,j})}{\det(\hat{\Sigma}_Y)}$$

with $\hat{\Sigma}_{\varepsilon,j}$ and $\hat{\Sigma}_Y$ the estimated covariance matrices of the errors and the response variables when j scores are retained in the model.

We also introduce several two and threedimensional diagnostic plots which are helpful to visualize and classify the outliers in the data. We show the performance and the robustness of the algorithms through simulations and applications on real data sets.

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

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

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

My abstract fits best to topic number: 20 and 18

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)