3.6. A robust method for PLS regression based on SSVD
3.7. Model Calibration and Validation
3.8. Simulation study
3.9.
3.10. Robust properties of robust PLS

Chapter 4

An application in environmental field.

4.1 A look at data(4.1)Tj ET EMC /P <</MC7 >>BDC BT /TT1 1 Tf 0 Tc 0 Tw 12 0 0 1

4.5.1. SSVD and FAST-MCD on data(4.1)TjETEMC /P <</M47 >>BDC BT/TT1 1 Tf0 Tw 1

4.5.3. PLS regression on the subsample without multivariate outliers
4.5.4. Squared robust residual distance
4.5.5. PLS regression on the subsample without regression outliers

4.6. Conclusions

Conclusions and Perspectives

In the second chapter we describe the most important robust methods for estimating the regression

variables to the response variable. The model that is used most extensively is Multiple Linear Regression Model. It ca

$$\mathbf{X} + \mathbf{u}$$

wher is a stochastic erro

of regression coefficients when we get off the id

$$\hat{}_{RR} = \left(\mathbf{X}^T\mathbf{X} - \quad \mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

our scepticism about the value of jointness for Joint Continuum Regression and related techniques, such as PLS is unjustified". We are agree with Garthwaite (1994). He states there are problems where univariate PLS perform better than multivariate PLS and cases where it might be advantageous to seek components that are good predictors of all **y** variables.

1.3.26 TT3 1 Tf0 Tc 13.98 0 0 13.98 172.20024 519.32031 Tm(

Table 1. 1 NIPALS-PLS algorithm

1. $\mathbf{X}_0 = \mathbf{X}$; $\mathbf{Y}_0 = \mathbf{Y}$
2. For $h = 1, 2, \ldots, A$:
   2.1. Choose for n of

1.3.1.2. PLS regression model

We saw the PLS algorithm is an iterative process; i.e. after extraction of one component the algorithm starts again using the deflated matrices and computed in step 2.4 and 2.5. Thus we can achieve the sequence of the com ls up to the point when the rank of is reached. However,e

ode

$$\mathbf{R} = \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1} \qquad \textbf{1. 22}$$

Due to the propriety c., the mahe m is upper triangular and thus invertible.he m
column space and thhe $\mathbf{P}^{\mathrm{T}}$ should be equal to the identity mahe m

1.3.2.

de Jong,2 993) approtac.

matrix

If the number of **y** is smaller than the number of **x** variables, it will be efficient to calculate as the dominant eigenvector of small sym

1.3.3. Comparison of SIMPLS with NIPAL-PLS

We can now compare the NIPALS-PLS and the SIMPLS algorithm. Both the algorithm ma

tors . At first sight, these conditions would seem to guarantee equal re-

sults obtainedm(e)Tj12.0136 0 0 12.01.8H.4749639 583.82013 Tm with the two algorith gorithm to various data sets reveals that

A test set should be independent from the training set which is used to estimate the regression parameters in the model, but should still be representative of the population. Let denote the total numb

CHAPTER 2

OUTLIERS AND ROBUST STATISTICS

2.1.

This point is called an outlier in the y-direction and it has a rather large influence on the OLS line, which is quite different from the OLS line in

sarily have to be a regression outlier. When lies close to regr
sionline der

Therefore 2.9 is really a system of equations, the solution of which is not always easy to find. In practise, one uses iter tion sch

5

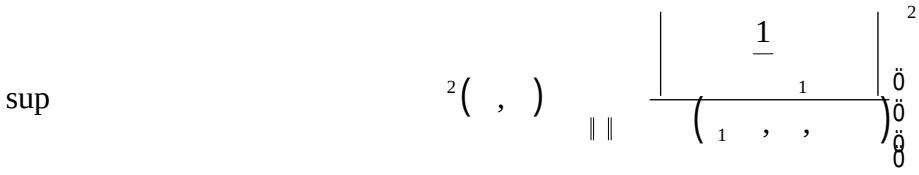
too harshly nor too lightly. The ideal estimator would penalize outliers without penalizing non-outliers. If it fails to penalize all the outliers, the estimator is performi

and delete one observation at a time. The size of the subset of observations used in fitting decreases as the method proceeds.
Another alternative is forward procedure (Atkinson, 1987) in which the basic idea is to order the observati

However, it is well known that $\bar{\mathbf{x}}$ is not robust, because even a single outlier in sample can move $\bar{\mathbf{x}}$ arbitrarily far away. To quantify such ef-

2.5.2. Affine Equivariant Estimators with High Breakdown Point.

The f



$$= \begin{cases} \quad (\quad - \quad) \, (\quad)^{-} \, (\quad - \quad) \, ¢ \, c \\ \text{otherwise} \end{cases}$$

of observations without replacement, as we did for the estimation of MVE. For each subgroups (indexed by) the arithmetic m

2.6. Robust Estimation of Singular Value Decomposition

A numerically stable way to perform many operations that are used in multivariate statistics is the Si ngular Value Decom

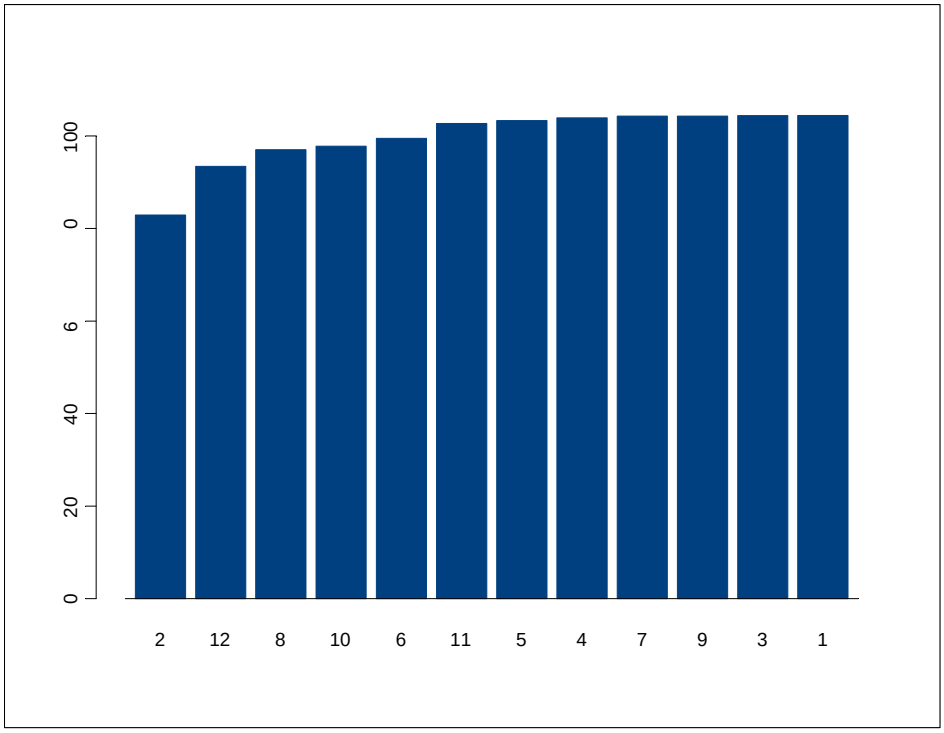
te the resu

thousands), so we cannot apply MCD as well as the other affine equivariant estimators with high breakdown point. A second problem is the computation of these robust estimators in high dimensions. Indeed, FAST-MCD algorithm, as implemented in S-PLUS, cannot handle more than 50 variables. Moreover, when the number of variables s la g r than, say 10 the computation of the MCD estimator becomes less p ecise.

method, which combines “leave-one-out” methods and Singular Value Decomposition (SVD).

The SVD is tivated by the following geom fact: e thi ge of the unit sphere under any $n \times p$ matrix is a hyperellipse. We may define a hyperellipse in $\mathbb{R}$ as the surface obtained by stretching he u

Figure 2. 7. Barplot of 12 singular values



Rousseeuw and Van Driessen ran FAST-
u efau 1 /2 12 2 1 /2 7

of the study suggest limits on the breakdown point of the procedure. A simulation study should be performed as follows:

1.

where the number of observations can be very large and the number of variables relatively small.

Table 2. 2

n	p	# $_{p}$ $_{p}$	# $_{p}$ $_{p}$	SSVD		FAST-MCD	
				Robust	Time	Robust	Time
10	2	6	4	yes	<1	yes	<1
10							

2.10. The deletion of Mult ivar iat e Out liers and

It is very imp ultivariate outliers. The classis
al app

from the linear pattern set by the majority of the data. Therefore, a combination of high-breakdown regression and robust MD_i gives more insight into regression data.

To

CHAPTER 3

ROBUST METHODS FOR PARTIAL LEAST SQUARES REGRESSION

4.1. Outliers in PLS Regression

cludes methods which use a robust cross-covariance matrix and a robust regression method.
In this chapter we describe the

j p

Form the loading vector $\mathbf{c}$ for the $\mathbf{Y}$ -block and normalize it

$$\mathbf{c} = \left(c_1, c_2, \mathbf{L}\, c_k, \mathbf{L}, c_q\right) \qquad \mathbf{3.\ 11}$$

$$\mathbf{c} = \frac{\mathbf{c}}{\|\mathbf{c}\|} \qquad \mathbf{3.\ 12}$$

Calculate the score vector $\mathbf{u}$

is maximised. The use of the $mad_h^2(\theta)$ (see footnote 5) instead of common $var(\theta)$ modifies the covariance estimates. It protects the derived components from abnormal observations. A full description of PLAD algorithm is given below.

1. Center or standardise both $\mathbf{X}$ and $\mathbf{y}$
2. For h ,2, , a

$$\mathbf{W}_h^M = \left(\mathbf{w}_1^M, \mathbf{K}, \mathbf{w}_h^M\right)$$

sion

Partial Reweighted Least Squares (PRLS) regression uses a weight vector in order to classify at each iteration the observations in outliers and non outliers. The latter is obtained while regressing the predictors and the

5%. Alternative rules could involve relative change in residuals or weights. They found that convergence occurs very quickly.
In step 3 of IRPLS the authors suggest to pass the predicted residuals rathe

of the m

PLS has the additional step of using t

optim by op
It would seem appropriate to use the

Let us assume that the data $\mathbf{Z} = [\mathbf{X}, \mathbf{y}]$ come from a joint distribution with mean equal to zero and population covariance matrix consisting of the elements

3.3.2.

necessary. If we replace these population values by the sample values, they can be influenced by the presence of outliers. A global robust version can come from using again the robust covariance matrix

$$= \begin{pmatrix} & \\ & \\ & \end{pmatrix}_r \qquad \textbf{3. 28}$$

depends on the estimation of the mean $\boldsymbol{\mu}_{\mathbf{Z}}$ and the variance–covariance matrix $_{\mathbf{Z}}$ of the data $_{n,m}$ $\mathbf{X}_{n,p}, \mathbf{Y}_{n,q}$

Applying the MCD or the RMCD method once to the $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$ yields an estimate of the variance–covariance matrix $_{\mathbf{z}}$ in 3.30. From this we deduce $_{\mathbf{XY}}$ which are robust estimates of .

208>Tj Tf 11.8231 M60 6f 0st estim

3.5. A robust method for PLS regression based on SSVD

In this paragraph we suggest a robust method for PLS regression. We proceed in the following way.

Stage 1. We apply SSVD on $_{n,m} = \lfloor _{n,p}, _{n,q} \rfloor$. This yields a "clean sample" (sample without multivariate outliers)

Stage 2. We apply PLS method (NIPALS or SIMPLS approach) on "clean sample" to obtain robust scores $\mathbf{t}_{h,}$.

Stage 3. Once the scores are derived, a linear regression is performed obtaining a robust estimate of ú. The regression model is the same as in 1.23, but now based on the robust scores. Finally a robust estimate of $ú_{PLSR}$ is obtained by rew estimates for the parameters in the original model.

We obtain a robust estimate of

Stage 5. We calculate the robust scores and final regression estimates by applying PLS regression on these observation with weight c_{ih} equal to 1. This stage has the advantage that it might again include observations which are "good" leverage points.
We call this method SPLS.

servations n. To avoid such small calibration sets, the authors alternatively define

$$h \qquad \textbf{3. 56}$$

approach is fast because it avoids cross-validation but m erely m easures the variance of the residuals instead of the prediction error.

3.7. Simulation study

To test effectiveness of a robust m ethod, a simulation study has to be designed. The study want to answer the question: can the robust m tect multiple outliers in various conf igurations? The goal w as to give the robust procedure a variety of outlier situations, not only in size and number of outliers but also in the placem ent of them. In addition, the results of the study suggest lim its on the breakdown point of the procedure. A simulation study should be performfollows: 62 559.94031 Tm( )TjET7EMC /P <</MCID 6 >>BI

CHAPTER 4

A ROBUST MODEL FOR THE EVALUATION OF VEHICLE EMISSIONS.

4.1. A look at data

Several epidemiological studies demonstrated short-term associations be-
veen high levels of pollution and increased acute mortality and morbid-

The analysis is based on a research developed by the Italian National Re-
acil (CNR),. Istituto Motori (IM). It concerns the relationship ions be-
ematics pa-
driving cy-

Before the IM research the evaluation of emissions of a vehicle in real use was generally obtained by utilizing methods which were based on the
n the definition of vehi-
ons (urban congested.

Vehicles emissions are an important source of environmental pollu-

2) was considered as an individual response. X

n. so it's necessary to estimate pollution emissions caused by classes
raffic. road. etc) in order to reduce en-

However. the IM showed the way the performance of vehicle affects exhaust pollutant emissions is complex and requires a larger number of parameters than the simple mean speed to be effectively described. Moreover. the different species of emissions are related each other as well as the different aspects characterizing vehicle performa

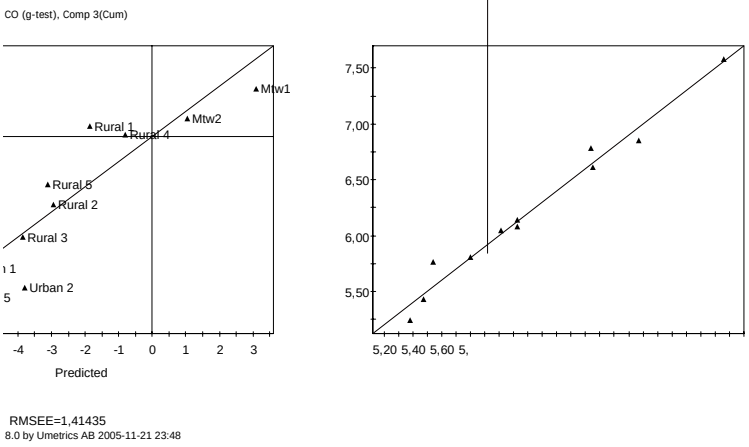
Table 4. 1

Dipendent variables

When the linear combination of the ’s ( )and the linear combination of the ’s ( ) are plotted against each other for the first few PLS dime>>BDC BT/TT0 1 Tf-0.02.20325 624.56007 Tm(nsions, one obtains a view of the relation bet

als are large or small. For this reason it is interesting to look at the standardized residual plot.

Figure 4. 7



4.4.1.

4.4.2. Multicollinearity diagnostics on the subsample without multivariate outliers

Multicollinearity diagnostics are not immune to the presence of contamination. Therefore, the identification of linear dependencies in a factor space, combined with the detection of outlie portant problem of regression analysis. For this reason we calcu54944 573.68019 Tm(late ag)Tj12.045 0 0 12.045

4.4.3. PLS regression on the subsample without multivariate outliers

Let us apply

t onent p

l e

Figure 4. 13

We can note several differ

APPENDIX

Robust methods for Partial Least Squares Regression in environmental field

Robust P