

Robust PCR and Robust PLSR: a comparative study

S. Engelen^a, M. Hubert^a, K. Vanden Branden^a, and S. Verboven^b

Abstract. Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) are the two most popular regression techniques in chemometrics. They both fit a linear relationship between two sets of variables. The responses are usually low-dimensional whereas the regressors are very numerous compared to the number of observations. In this paper we compare two recent robust PCR and PLSR methods and their classical versions in terms of efficiency, goodness-of-fit, predictive power and robustness.

1. Introduction

Principal Component Regression and Partial Least Squares Regression are the two most popular regression techniques in chemometrics [11]. Both methods can handle multicollinearity in the data and can be applied if there are more regressors than objects. Hence, it are the standard tools for multivariate calibration where the concentrations of certain constituents in samples are modelled and predicted from their spectra.

PCR and PLSR are based on a bilinear model that explains the existence of a relation between a set of p -dimensional regressors and a set of q -dimensional response variables through k -dimensional scores \mathbf{t}_i with $k \gg p$. More precisely, for $i = 1, \dots, n$ with n the number of observations $(\mathbf{x}_i, \mathbf{y}_i)$, we assume that

$$(1.1) \quad \mathbf{x}_i = \bar{\mathbf{x}} + P_{p,k} \mathbf{t}_i + \mathbf{f}_i$$

$$(1.2) \quad \mathbf{y}_i = \bar{\mathbf{y}} + \mathcal{A}'_{q,k} \mathbf{t}_i + \mathbf{g}_i.$$

Here $\bar{\mathbf{x}}$ is the mean of the x -variables, $\bar{\mathbf{y}}$ the mean of the y -variables, $P_{p,k}$ the matrix of x -loadings and $\mathcal{A}_{k,q}$ represents the slope matrix in the regression of \mathbf{y}_i on \mathbf{t}_i . The superscript $'$ is used for the transpose of a vector or matrix, and subscripts as in $P_{p,k}$ indicate that the matrix P has p rows and k columns. The

Received by the editors September 20, 2003.

1991 *Mathematics Subject Classification.* 62F35; 62H20; 92E99.

Key words and phrases. Principal Component Analysis, Principal Component Regression, Partial Least Squares, Robustness.

error terms are denoted by \mathbf{f}_i and \mathbf{g}_i . In terms of the original predictor variables, this bilinear model can be written as

$$\mathbf{y}_i = \boldsymbol{\beta}_0 + \mathcal{B}'_{q,p} \mathbf{x}_i + \mathbf{e}_i$$

with

$$(1.3) \quad \mathcal{B}_{p,q} = P_{p,k} \mathcal{A}_{k,q}$$

$$(1.4) \quad \boldsymbol{\beta}_0 = \bar{\mathbf{y}} - \mathcal{B}'_{q,p} \bar{\mathbf{x}}.$$

According to the bilinear model, both PCR and PLSR proceed in two major stages. In the first step, following (1.1), they summarize the high-dimensional observations \mathbf{x}_i in scores \mathbf{t}_i of dimension $k \ll p$. The selection of the number of components k can be done by means of various different criteria, which will be discussed in Section 3. These k latent variables then become the regressors in the second step of the algorithm (1.2). Finally, estimates for \mathcal{B} and $\boldsymbol{\beta}_0$ are obtained via (1.3) and (1.4).

The main difference between PCR and PLSR lies in the construction of the scores \mathbf{t}_i . In PCR the scores are obtained by extracting the most relevant information present in the x -variables by performing a principal component analysis on the predictor variables and thus using a variance criterion. No information concerning the response variables is yet taken into account. In contrast, the PLSR scores are calculated by maximizing a covariance criterion between the x - and y -variables. Hence also information present in the responses is used in the first stage of the algorithm. By construction we thus expect that PLSR requires less components than PCR. This has been confirmed by several studies [3], [6], [11].

In this paper we want to investigate how recent robust methods of PCR and SIMPLS behave with respect to each other in terms of efficiency, goodness-of-fit, predictive ability and robustness. In Section 2 we shortly describe the four methods involved in this paper: classical PCR (CPCR), robust PCR (RPCR), SIMPLS and robust SIMPLS (RSIMPLS). In Section 3 a comparison between the four methods is given on the basis of a simulation study. All four methods will be illustrated on a real example in Section 4 whereas Section 5 summarizes our conclusions.

2. Robust Calibration Methods

2.1. Principal Component Regression

As explained in the introduction, CPCR starts by performing a Principal Component Analysis (PCA) on the x -variables. The PCA loading matrix $P_{p,k}$ then contains the first k dominant eigenvectors of the empirical covariance matrix of the \mathbf{x}_i , and the scores satisfy $\mathbf{t}_i = P'_{k,p} (\mathbf{x}_i - \bar{\mathbf{x}})$. In the second step of CPCR multiple linear least squares regression (MLR) is applied on the $(\mathbf{t}_i, \mathbf{y}_i)$ to obtain an estimate of the slope matrix $\mathcal{A}_{k,q}$ in (1.2).

In [9] a robust PCR method is proposed by robustifying both steps of CPCR. First a robust PCA method is applied on the regressors. For low-dimensional

data ($p < n/2$), the MCD estimator [14] is used as a robust estimator of the covariance matrix of the \mathbf{x}_i , and for high-dimensional data the ROBPCA method [8]. This estimator combines projection pursuit techniques with robust covariance estimation in low dimensions. Next a robust regression method is applied. If there is only one response variable the reweighted LTS regression [14] is preferred, else the MCD regression [16] is performed.

2.2. Partial Least Squares Regression

We consider the SIMPLS algorithm [2] being the leading PLSR method because of its speed and efficiency. Let $\tilde{X} = \{(\mathbf{x}_i - \bar{\mathbf{x}})'\}_{i=1}^n$ and $\tilde{Y} = \{(\mathbf{y}_i - \bar{\mathbf{y}})'\}_{i=1}^n$ be the centered data matrices. The first normalized PLSR weight vectors \mathbf{r}_1 and \mathbf{q}_1 are obtained as linear combinations of \tilde{X} and \tilde{Y} that maximize

$$\text{cov}(\tilde{X}\mathbf{r}_1, \tilde{Y}\mathbf{q}_1).$$

The solution of this maximization problem is found by taking \mathbf{r}_1 and \mathbf{q}_1 as the first left and right singular eigenvectors of $S_{xy} = \tilde{X}'\tilde{Y}/(n-1)$, the cross-covariance matrix of the x - and y -variables. For each observation the first coordinate of the score \mathbf{t}_i is computed as $t_{i1} = \tilde{\mathbf{x}}_i'\mathbf{r}_1$.

The other PLSR weight vectors \mathbf{r}_a and \mathbf{q}_a for $a = 2, \dots, k$ are obtained by imposing an orthogonality constraint to the elements of the scores. If we require that $\sum_{i=1}^n t_{ia}t_{ib} = 0$ for $a \neq b$, a deflation of the cross-covariance matrix S_{xy} provides the solutions for the other PLSR weight vectors. This deflation is carried out by first calculating the x -loading $\mathbf{p}_a = S_x\mathbf{r}_a/(\mathbf{r}_a'S_x\mathbf{r}_a)$ with S_x the empirical variance-covariance matrix of the x -variables. Next an orthonormal base $\{\mathbf{v}_1, \dots, \mathbf{v}_a\}$ of $\{\mathbf{p}_1, \dots, \mathbf{p}_a\}$ is constructed and S_{xy} is deflated as

$$S_{xy}^a = S_{xy}^{a-1} - \mathbf{v}_a(\mathbf{v}_a'S_{xy}^{a-1})$$

with $S_{xy}^1 = S_{xy}$. In general the PLSR weight vectors \mathbf{r}_a and \mathbf{q}_a are obtained as the left and right singular vector of S_{xy}^a . Finally, when the scores are k -dimensional, MLR is performed of the responses \mathbf{y}_i on these scores \mathbf{t}_i .

A robust method RSIMPLS has recently been developed in [10]. It starts by applying ROBPCA [8] on the x - and y -variables in order to replace S_{xy} and S_x by robust estimates and then proceeds analogously to the SIMPLS algorithm. Similar to RPCR a robust regression method (ROBPCA regression) is performed in the second stage. In [20] it is proved that for low-dimensional data the RSIMPLS approach yields bounded influence functions for the weight vectors \mathbf{r}_a and \mathbf{q}_a , and for the regression estimates. Also the breakdown value is inherited from the MCD estimator.

The computational complexity of ROBPCA and RSIMPLS is discussed in detail in [8] and [10]. The computation time remains feasible due to the FAST-MCD algorithm [18]. To give an example, on a Pentium IV with 1.60 GHz, it requires in a full Matlab implementation approximately 7 seconds to perform RSIMPLS on a data set with $n = 100$, $p = 500$ and $k = 5$.

3. Experimental Study

We will compare the efficiency, the goodness-of-fit (GOF), the predictive power and the robustness of CPCR, RPCR, SIMPLS and RSIMPLS by performing a simulation study on uncontaminated and contaminated data. Note that the robustness of RPCR and RSIMPLS has also been shown through simulations in [9] and [10], but there the emphasis was put only on the parameter estimation and not on the predictive performance of the methods. The experiments described in this section consider univariate responses ($q = 1$) which are mostly used in practice. We thus consider the regression model

$$y_i = \beta_0 + \mathcal{B}'_{1,p} \mathbf{x}_i + e_i$$

with $\mathcal{B}_{p,1} = (\beta_1, \dots, \beta_p)'$. The regression vector including the intercept is denoted as $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$.

3.1. Simulation settings

We compare the algorithms on high-dimensional data sets $X_{50,100}$ of size $n = 50$ and $p = 100$ and low-dimensional data sets $X_{100,6}$ of size $n = 100$ and $p = 6$. They were constructed according to the bilinear model (1.1) and (1.2):

$$\begin{aligned} T &\sim N_2(\mathbf{0}_2, \Sigma_t) \\ X &= TI_{2,p} + N_p(\mathbf{0}_p, 0.1I_p) \\ Y &= T\mathcal{A}_{2,1} + N(0, 1) \end{aligned}$$

with $\mathbf{0}_2 = (0, 0)'$, $I_{k,p} = \delta_{ij}$ and $\mathcal{A}_{2,1} = (1, 1)'$. Furthermore we set $\Sigma_t = \begin{pmatrix} 8 & 0 \\ 0 & 2 \end{pmatrix}$ when $p = 100$ and $\Sigma_t = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$ for $p = 6$. Hence the optimal number of components $k_{\text{opt}} = 2$.

Next, contamination is added by replacing 10% of the observations by different types of outliers. Denote T_ϵ , X_ϵ , and Y_ϵ as the contaminated parts of the data.

1. *Bad leverage points* were constructed by substituting $T_\epsilon \sim N_2((15, 15)', \Sigma_t)$ in equation (1.1):

$$X_\epsilon = T_\epsilon I_{2,p} + N_p(\mathbf{0}_p, 0.1I_p).$$

Note that the corresponding y -values did not change.

2. *Vertical outliers* have uncontaminated x -values, but their y -values were changed by adjusting the error term in (1.2):

$$Y_\epsilon = T\mathcal{A}_{2,1} + N(15, 0.1).$$

For each situation, $m = 100$ data sets were generated and they were analyzed with $k = 1, 2$ and 3 components.

The efficiency of the considered methods is evaluated by means of the MSE of the estimated regression parameters $\hat{\beta}$. It is defined by

$$\text{MSE}_k(\hat{\beta}) = \frac{1}{m} \sum_{l=1}^m \|\hat{\beta}_k^{(l)} - \beta\|^2$$

where $\hat{\beta}_k^{(l)}$ denotes the estimated parameter based on k components in the l th simulation. The MSE indicates to what extent the slope and intercept are correctly estimated. So the goal is to obtain an MSE value close to zero.

Next we want to study how well the methods fit the regular data points. Because of the simulation settings, we know exactly their indices which we store in the set G_r . Then we define the goodness-of-fit criterion as

$$(3.1) \quad \text{GOF}_k = 1 - \frac{\text{var}_{i \in G_r}(r_{i,k})}{\text{var}_{i \in G_r}(y_i)}$$

with $r_{i,k}$ the residual of the i th observation when k components are computed. The objective is to obtain a GOF close to 1. Note that this GOF is an adaptation of the robust R^2 proposed in [9] where G_r contains the non-outlying data points detected by the estimation procedure itself (RPCR or RSIMPLS). The value of k where the R^2 -curve stabilizes is then selected as the optimal one.

Finally, we measure the predictive ability of the methods by means of the Root Mean Squared Error (RMSE) as in [10]. First we generate a test set G_t of uncontaminated points with size $n_t = 50$, and then we compute

$$(3.2) \quad \text{RMSE}_k = \sqrt{\frac{1}{n_t} \sum_{i=1}^{n_t} (y_i - \hat{y}_{i,k})^2}$$

with $\hat{y}_{i,k}$ the predicted y -value of observation i from the test set when the regression parameter estimates are based on the training set (X, Y) of size n and k scores are retained. The optimal number of components is often selected as that k for which this RMSE value (or its cross-validated version) is minimal.

The results of the simulations are listed in Tables 1–3 for the high-dimensional data sets and in Tables 4–6 for the low-dimensional situation.

3.2. Simulation results

When no contamination is added (Tables 1 and 4), the classical methods perform somewhat better than their robust versions, as we would expect. At the optimal $k_{\text{opt}} = 2$, CPCR and SIMPLS can hardly be distinguished. At high-dimensional data, the MSE(CPCR) is minimal, but the GOF and RMSE values are slightly in favor of SIMPLS. At low-dimensional data, we see almost no differences. When only one component is selected ($k = 1$), SIMPLS outperforms CPCR noticeably. This confirms the findings of [6] that PLSR constructs its components more efficiently than PCR. However, when we choose more components than required, here $k = 3$, we notice that SIMPLS suffers much more from overfitting than CPCR.

		Method			
		CPCR	RPCR	SIMPLS	RSIMPLS
$k = 1$	MSE($\hat{\beta}$)	1.12	1.14	0.63	0.65
	GOF	0.69	0.69	0.81	0.80
	RMSE	1.81	1.82	1.50	1.51
$k = 2$	MSE($\hat{\beta}$)	0.16	0.22	0.23	0.29
	GOF	0.89	0.89	0.91	0.90
	RMSE	1.14	1.15	1.13	1.15
$k = 3$	MSE($\hat{\beta}$)	0.20	0.29	3.19	0.91
	GOF	0.89	0.89	0.97	0.91
	RMSE	1.14	1.16	1.27	1.19

TABLE 1. $n = 50, p = 100$, no contamination.

		Method			
		CPCR	RPCR	SIMPLS	RSIMPLS
$k = 1$	MSE($\hat{\beta}$)	1.93	1.14	1.92	0.65
	GOF	0.17	0.71	0.31	0.82
	RMSE	3.07	1.84	2.85	1.50
$k = 2$	MSE($\hat{\beta}$)	2.49	0.22	2.80	0.31
	GOF	0.39	0.89	0.45	0.91
	RMSE	2.68	1.15	2.68	1.16
$k = 3$	MSE($\hat{\beta}$)	2.78	0.28	20.24	1.02
	GOF	0.41	0.89	0.80	0.93
	RMSE	2.68	1.15	3.01	1.19

TABLE 2. $n = 50, p = 100$, 10% bad leverage points.

This is reflected in the large MSE values of SIMPLS. At high-dimensional data $\text{MSE}_3(\text{SIMPLS}) = 3.19$ is even considerably larger than $\text{MSE}_3(\text{RSIMPLS}) = 0.91$.

If we add contamination, CPCR and SIMPLS clearly break down. The MSE of the regression parameter estimates increase drastically and even attain their minimum at $k = 1$ (except for CPCR in Table 3). The GOF values are very low, especially when the data contain bad leverage points. This shows that the regular data points are badly fitted. The high RMSE values indicate the low predictive ability of the classical methods. SIMPLS is more sensitive to vertical outliers than

		Method			
		CPCR	RPCR	SIMPLS	RSIMPLS
$k = 1$	MSE($\hat{\beta}$)	3.39	1.15	3.06	0.62
	GOF	0.65	0.69	0.76	0.81
	RMSE	2.44	1.81	2.25	1.48
$k = 2$	MSE($\hat{\beta}$)	2.68	0.22	14.97	0.29
	GOF	0.80	0.89	0.66	0.90
	RMSE	2.11	1.15	2.55	1.15
$k = 3$	MSE($\hat{\beta}$)	3.37	0.28	60.56	1.01
	GOF	0.77	0.89	0.57	0.92
	RMSE	2.13	1.16	3.26	1.20

TABLE 3. $n = 50, p = 100, 10\%$ vertical outliers.

		Method			
		CPCR	RPCR	SIMPLS	RSIMPLS
$k = 1$	MSE($\hat{\beta}$)	1.06	1.11	0.27	0.27
	GOF	0.55	0.54	0.77	0.77
	RMSE	1.79	1.81	1.28	1.27
$k = 2$	MSE($\hat{\beta}$)	0.02	0.03	0.03	0.05
	GOF	0.83	0.83	0.83	0.83
	RMSE	1.10	1.11	1.10	1.11
$k = 3$	MSE($\hat{\beta}$)	0.10	0.13	0.51	0.60
	GOF	0.83	0.83	0.84	0.84
	RMSE	1.11	1.11	1.12	1.13

TABLE 4. $n = 100, p = 6$, no contamination.

CPCR. This is probably due to the fact that the response variable is already used in the construction of the SIMPLS scores, contrary to the CPCR scores that only depend on the x -variables.

The results of the robust methods on the other hand are almost identical to the uncontaminated case. RSIMPLS is again superior to RPCR for $k = 1$ and $k = 2$, but shows more overfitting when $k = 3$. Note that the GOF values of RSIMPLS are always higher than those of RPCR. Both GOF and RMSE appear to be good criteria to select k . We see that the differences $\text{GOF}_3 - \text{GOF}_2$ are very

		Method			
		CPCR	RPCR	SIMPLS	RSIMPLS
$k = 1$	MSE($\hat{\beta}$)	1.84	1.06	1.83	0.24
	GOF	0.10	0.54	0.13	0.78
	RMSE	2.51	1.81	2.49	1.26
$k = 2$	MSE($\hat{\beta}$)	2.04	0.03	2.05	0.05
	GOF	0.18	0.83	0.18	0.83
	RMSE	2.42	1.11	2.42	1.11
$k = 3$	MSE($\hat{\beta}$)	2.56	0.17	4.52	0.65
	GOF	0.19	0.83	0.21	0.83
	RMSE	2.43	1.11	2.47	1.13

TABLE 5. $n = 100, p = 6, 10\%$ bad leverage points.

		Method			
		CPCR	RPCR	SIMPLS	RSIMPLS
$k = 1$	MSE($\hat{\beta}$)	3.31	1.10	2.55	0.24
	GOF	0.52	0.53	0.74	0.78
	RMSE	2.39	1.81	2.04	1.25
$k = 2$	MSE($\hat{\beta}$)	2.35	0.03	2.41	0.05
	GOF	0.78	0.83	0.78	0.83
	RMSE	1.97	1.10	1.96	1.11
$k = 3$	MSE($\hat{\beta}$)	4.46	0.13	10.79	0.62
	GOF	0.75	0.83	0.68	0.84
	RMSE	2.02	1.11	2.16	1.13

TABLE 6. $n = 100, p = 6, 10\%$ vertical outliers.

small compared to $\text{GOF}_2 - \text{GOF}_1$. But we can not conclude that we should choose k for which GOF_k is maximal. On the other hand, the minimal value of RMSE is always reached at the correct $k = 2$. This suggests to select k such that RMSE_k is minimal.

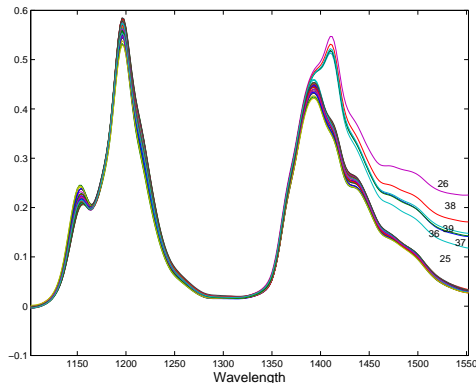


FIGURE 1. The spectra of the octane data.

4. An example

In this section, we compute the GOF and the RMSE values of the two classical and the two robust calibration methods on a real data set. The *octane* data [5] consist of NIR absorbance spectra over $p = 226$ wavelengths ranging from 1102 to 1552 nm with measurements every two nm. For each of the $n = 39$ production gasoline samples the octane number y was measured, so $q = 1$.

This data set has been studied before in [8]. It is well known that the octane data set contains six outliers to which alcohol was added. In Figure 1 we see that the spectra of those six samples clearly stand out. They were also detected by applying the robust PCA method ROBPCA to the 226 regressors [8]. Therefore we put the set of regular observations G_r equal to the full data of 39 observations minus these six outliers.

For each of the four methods we computed the goodness-of-fit index as defined in (3.1). Applying (3.2) is not possible here because a test data set is beyond our reach, and given the small number of observations, we do not want to split the data set into a training and a test data set. Therefore, we use the cross-validated R-RMSECV value as proposed in [9] and [10]. It is defined as

$$(4.1) \quad \text{R-RMSECV}_k = \sqrt{\frac{1}{33} \sum_{i \in G_r} (y_i - \hat{y}_{-i,k})^2}.$$

Here $\hat{y}_{-i,k}$ represents the predicted y -value for observation i based on k components when observation i was left out of the estimation of the regression parameters. Note that when the outliers are not known in advance, the set of regular observations G_r is determined by the calibration method itself. Hence different sets could be obtained for e.g. RPCR and RSIMPLS. This would make it difficult to compare R-RMSECV values of several methods.

We evaluate criteria (3.1) and (4.1) for $k = 1, \dots, 4$ components. The results are summarized in Table 7. For $k = 1$ the classical methods perform very

		Method			
		CPCR	RPCR	SIMPLS	RSIMPLS
$k = 1$	GOF	-0.04	0.81	0.13	0.82
	R-RMSECV	2.08	0.93	1.92	0.90
$k = 2$	GOF	0.80	0.98	0.85	0.98
	R-RMSECV	0.94	0.31	0.81	0.32
$k = 3$	GOF	0.98	0.98	0.98	0.99
	R-RMSECV	0.31	0.30	0.30	0.30
$k = 4$	GOF	0.98	0.98	0.98	0.99
	R-RMSECV	0.31	0.30	0.29	0.29

TABLE 7. Analysis of the octane data.

badly. For CPCR this even results in a negative $\text{GOF}_1(\text{RPCR}) = -0.04$. Also $\text{GOF}_1(\text{SIMPLS}) = 0.13$ is very small compared to the robust values. The R-RMSECV values of CPCR and SIMPLS are approximately twice as high than those of RPCR and RSIMPLS.

When we retain $k = 2$ components, the GOF values of the classical methods increase considerably, and their R-RMSECV values have been more than halved, yielding results that are comparable with the robust ones for $k = 1$. RPCR and RSIMPLS clearly improve their fit with a very high $\text{GOF}_2 = 0.98$ and R-RMSECV values of 0.31 resp. 0.32.

For $k = 3$ the classical values correspond with the robust ones for $k \geq 2$. This clearly shows that the first component of CPCR and SIMPLS is completely determined by the outliers and it confirms the conclusion in [19] to retain the second and the third component of SIMPLS. The results obtained with RPCR and RSIMPLS remain stable from $k = 2$ on, so these two methods suggest to retain $k_{\text{opt}} = 2$ components. Note that here the R-RMSECV value is not at its minimal value, but it does not hardly change when k is increased.

5. Conclusions

Both the simulation study and the analysis of a real data set show that CPCR and SIMPLS are very sensitive to outliers in the data, whereas RPCR and RSIMPLS can resist several types of contamination.

When the correct number of components is used in the calibration, RPCR and RSIMPLS are comparable in terms of efficiency, goodness-of-fit, predictive power and robustness. For smaller k , RSIMPLS is to be preferred, whereas RPCR is less sensitive to overfitting when a larger set of components is selected.

Finally, the proposed GOF and RMSE/R-RMSECV measures are shown to be good indicators to select the optimal number of components. To speed up the heavy computations involved in the cross-validated R-RMSECV, we are currently developing faster algorithms. This will allow to perform fast and robust model selection in multivariate calibration.

All the methods described in this paper can be downloaded from the web sites <http://www.wis.kuleuven.ac.be/stat/robust.html> and <http://win-www.ruca.ua.ac.be/u/statis> as part of the Matlab library for Robust Calibration [21].

Note that in this paper we have concentrated on calibration methods that are particularly useful for small data sets in high-dimensions, which are very common in chemometrics, food science and bioinformatics. Successful applications of robust PCA in bioinformatics are e.g. presented in [13] and [7]. One of the referees wondered whether they also could be applied to problems in computer vision where both the number of samples and the number of variables can be very large. Currently ROBPCA is being implemented by Darren Cosker (3D Vision and Geometry, Department of Computer Science, Cardiff University, UK) to build statistical models of shape and appearance, i.e. Active Shape Models (ASM) and Active Appearance Models (AAM) [1]. In particular the method is used for building a model of mouth tracking. Other robust methods in computer vision have e.g. been proposed in [17], [22], [4], and [15]. For an overview of the use of high-breakdown methods in computer vision, see e.g. [12].

References

- [1] T.F. Cootes, G.J. Edwards, C.J. Taylor, *Active Appearance Models* in Proc. European Conference on Computer Vision 1998 (H.Burkhardt & B. Neumann Ed.s). Vol. 2, 484–498, Springer, 1998.
- [2] S. de Jong, *SIMPLS: an alternative approach to partial least squares regression*. Chemometrics Intell. Lab. Syst. **18** (1993 (a)), 251–263.
- [3] S. de Jong, *PLS fits closer than PCR*. J. of Chemometrics **7** (1993 (b)), 551–557.
- [4] F. De la Torre, M.J. Black, *Robust principal component analysis for computer vision*. Int. Conf. on Computer Vision, ICCV-2001, Vancouver, (2001), 362–369.
- [5] K.H. Esbensen, S. Schönkopf, T. Midtgaard *Multivariate Analysis in Practice*. Camo, Trondheim, 1994.
- [6] I.E. Frank, J.H. Friedman *A statistical view of some chemometrics regression tools*. Technometrics **35** (1993), 109–135.
- [7] M. Hubert, S. Engelen, *Robust PCA and classification in biosciences*. Submitted (2003).

- [8] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, *ROBPCA: a new approach to robust principal component analysis*. tentatively accepted for *Technometrics*, (2003).
- [9] M. Hubert, S. Verboven, *A robust PCR method for high-dimensional regressors*. *J. of Chemometrics*, **17** (2003), 438–452.
- [10] M. Hubert, K. Vanden Branden, *Robust methods for Partial Least Squares regression*. *J. of Chemometrics*, to appear (2003).
- [11] H. Martens, T. Naes, *Multivariate Calibration*. Wiley, Chichester, 1998.
- [12] P. Meer, *Robust techniques for computer vision* in Emerging topics in Computer Vision (G. Medioni & S.B. Kang Ed.s), Prentice Hall, 2004.
- [13] F. Model, T. Knig, C. Piepenbrock, P. Ardojan, *Statistical process control for large scale microarray experiments* *Bioinformatics*, **18** Suppl 1 (2002), 55–63.
- [14] P.J. Rousseeuw, *Least median of squares regression*. *J. Am. Statist. Assoc.* **79** (1984), 871–880.
- [15] C.V. Stewart, *Robust parametric estimation in computer vision*. *Siam Reviews* **41** (1999), 513–537.
- [16] P.J. Rousseeuw, S. Van Aelst, K. Van Driessen, J. Agulló, *Robust multivariate regression*. Under revision (2002).
- [17] D. Skocaj, H. Bischof, A. Leonardis, *A robust PCA algorithm for building representations from panoramic images*. *Proc. European Conf. on Computer Vision*; Vol. 4; pp. 761–775; Copenhagen, Denmark; 2002.
- [18] P.J. Rousseeuw, K. Van Driessen, *A fast algorithm for the minimum covariance determinant estimator*. *Technometrics* **41** (1999), 212–223.
- [19] M. Tenenhaus, *La Régression PLS*. Édition Technip, Paris, 1998.
- [20] K. Vanden Branden, M. Hubert *Robustness properties of a robust SIMPLS method*. Submitted (2003).
- [21] S. Verboven, M. Hubert *A Matlab toolbox for robust calibration*. In preparation.
- [22] L. Xu, A. Yuille *Robust principal components analysis by self-organizing rules based on statistical physics approach*. *IEEE Transaction on Neural Networks*. **6** (1995), 131–143.

^(a)Katholieke Universiteit Leuven, W. de Croylaan 54, B-3001 Leuven, Belgium
E-mail address: `sanne.engelen@wis.kuleuven.ac.be`
E-mail address: `mia.hubert@wis.kuleuven.ac.be`
E-mail address: `karlien.vandenbranden@wis.kuleuven.ac.be`

^(b)University of Antwerp, Middelheimlaan 1, B-2020 Antwerpen, Belgium
E-mail address: `sabine.verboven@ua.ac.be`