

Robustness properties of a robust PLS regression method

K. Vanden Branden* and M. Hubert[†]

December 10, 2003

Abstract

The presence of multicollinearity in regression data is no exception in real life examples. Instead of applying ordinary regression methods, biased regression techniques such as Principal Component Regression and Ridge Regression have been developed to cope with such data sets. In this paper we consider Partial Least Squares (PLS) regression by means of the SIMPLS algorithm. Because the SIMPLS algorithm is based on the empirical variance-covariance matrix of the data and on least squares regression, outliers have a damaging effect on the estimates. To reduce this pernicious effect of outliers, we propose to replace the empirical variance-covariance matrix in SIMPLS by a robust covariance estimator. We derive the influence function of the resulting PLS weight vectors and the regression estimates, and conclude that they will be bounded if the robust covariance estimator has a bounded influence function. Also the breakdown value is inherited from the robust estimator. We illustrate the results using the MCD estimator and the reweighted MCD estimator (RMCD) for low-dimensional data sets. Also some empirical properties are provided for a high-dimensional data set.

Key words: Partial Least Squares Regression, SIMPLS, Influence Function, Minimum Covariance Determinant, Robustness.

*Assistant, Department of Mathematics, Katholieke Universiteit Leuven, W. de Croylaan 54, B-3001 Leuven, Belgium, Karlien.VandenBranden@wis.kuleuven.ac.be.

[†]Assistant Professor, Department of Mathematics, Katholieke Universiteit Leuven, W. de Croylaan 54, B-3001 Leuven, Belgium, Mia.Hubert@wis.kuleuven.ac.be

1 Introduction

Partial Least Squares Regression (PLSR) is the most commonly used regression technique in the field of chemometrics. It is used to link two sets of variables to each other by means of a linear model. The first set of variables contains the p regressors or predictor variables, which are usually denoted by X_1, X_2, \dots, X_p . The second set consists of q response variables Y_1, Y_2, \dots, Y_q . We assume that the regressors are more numerous than the number of response variables, thus $p \geq q$, as is generally the case in practice. The theory developed in this paper would also be valid if $p < q$ but then we would need to adapt a change in the notations. The PLSR model assumes that we have sampled n independent observations $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$ with $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ and $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iq})'$ from the linear model

$$\mathbf{y}_i = \boldsymbol{\beta}_0 + \mathbf{B}'\mathbf{x}_i + \mathbf{e}_i \quad (1)$$

with i.i.d. errors satisfying $E(\mathbf{e}_i) = \mathbf{0}$ and $\text{Cov}(\mathbf{e}_i) = \Sigma_e$. The intercept vector is denoted by $\boldsymbol{\beta}_0$ whereas the $(p \times q)$ slope matrix is represented by \mathbf{B} . Note that we print column vectors in bold and matrices are denoted by capital letters. The transpose of a vector or a matrix is indicated by $'$.

Estimates for the unknown parameters $\boldsymbol{\beta}_0$, \mathbf{B} and Σ_e in model (1) can be obtained via classical multiple linear regression (MLR) which minimizes the sum of the squared residuals. However, when the regressors are highly correlated, it is well-known that MLR gives rise to high variances. This problem is for example very prominent at high-dimensional data sets, where the number of regressors p largely exceeds the number of observations n . This typically occurs in multivariate calibration where the x -variables represent the absorbance or reflectance of light intensities of several samples at hundreds of wavelengths. Therefore biased regression techniques such as PLSR have been proposed.

PLSR can be seen as an extension of MLR. Instead of applying MLR to the full set of regressors, $k < p$ latent uncorrelated variables T_1, T_2, \dots, T_k are iteratively extracted from the data. These components then serve as the regressors in the reduced linear regression model with as response variables the original Y_i . Finally, backtransformation of the resulting estimates yields parameter estimates for the full model (1). Note that the optimal number of components $k = k_{\text{opt}}$ is typically determined by minimizing the Root Mean Squared Error of Prediction (RMSEP) or its cross-validated variant.

Several algorithms have been proposed to perform PLSR [1]. We consider here the SIM-

PLS algorithm [2] because it is computationally attractive, it applies to $q \geq 1$ and it coincides with the famous PLS1 for $q = 1$. In SIMPLS the components are derived by maximizing a covariance criterion which depends on some SIMPLS weight vectors. Because the construction of these SIMPLS weight vectors is based on the empirical variance-covariance matrix of the joint regressors and responses, outlying samples will have a pernicious influence on these vectors. We illustrate this effect by showing that the influence functions of these SIMPLS weight vectors and the regression estimates are unbounded. The influence function is a tool to measure the damage a single observation can cause to the estimates [3].

In this paper, we propose to replace the empirical variance-covariance matrix in the SIMPLS algorithm by a robust covariance matrix with bounded influence function. The influence functions of the resulting PLS weight vectors and the regression estimates then appear to be bounded as well. In particular, we can use the MCD estimator [4], the reweighted MCD estimator [5], S-estimators [6], ... in low-dimensional settings, i.e. when $n > 2(p + q)$. We also show that this robust method inherits the breakdown value [7], which is a more global measure of robustness, of the robust covariance estimator.

For high-dimensional data sets, we can apply another robust PLS regression method (RSIMPLS) which has recently been developed [8]. It relies on a robust covariance estimator for situations where the number of observations may be much smaller than the number of regressors p [9]. We show some empirical results to demonstrate the robustness of this approach.

In Section 2 we describe in detail how the weight vectors and the regression estimates $\hat{\mathcal{B}}$ and $\hat{\beta}_0$ are obtained in the SIMPLS algorithm [2] and we also introduce the robustification of this method. Section 3 presents the results of the influence functions of the PLS weight vectors and the regression estimates for low-dimensional data. Several graphical displays illustrate the non-robustness of the classical approach and the bounded influence functions obtained with the robust methods. We also derive the breakdown point of the proposed robust SIMPLS method. In Section 4 we describe some empirical properties of the robust PLS method developed in [8] for high-dimensional data. Section 5 contains some concluding remarks. Finally, most of the theoretical results and proofs are provided in the Appendix.

2 The algorithms

2.1 The SIMPLS algorithm

In the SIMPLS algorithm [2] k latent variables T_1, T_2, \dots, T_k are constructed in an iterative way by maximizing a covariance criterion. These k latent variables depend on the SIMPLS weight vectors \mathbf{r}_a and \mathbf{q}_a (for $a = 1, \dots, k$) that are obtained as the vectors that maximize the covariance between x - and y -components

$$\max_{\|\mathbf{r}_a\|=1, \|\mathbf{q}_a\|=1} \text{cov}(\tilde{X}\mathbf{r}_a, \tilde{Y}\mathbf{q}_a) = \max_{\|\mathbf{r}_a\|=1, \|\mathbf{q}_a\|=1} \mathbf{r}'_a S_{xy} \mathbf{q}_a \quad (2)$$

under the additional restrictions that the components $T_a = \tilde{X}\mathbf{r}_a$ be uncorrelated, i.e.

$$T'_a T_j = 0 \quad \forall j \neq a. \quad (3)$$

In (2) \tilde{X} and \tilde{Y} represent the mean-centered data matrices and $S_{xy} = \tilde{X}'\tilde{Y}/(n-1)$ is the empirical cross-covariance matrix between the x - and y -variables. The constraint (3) is imposed to generate a sequence of different solutions of (2) and in addition it avoids multicollinearity between the regressors in the second stage of the algorithm. The first pair of SIMPLS weight vectors $(\mathbf{r}_1, \mathbf{q}_1)$ is thus obtained as the first left and right singular vector of S_{xy} . This implies that \mathbf{r}_1 is the dominant eigenvector of $S_{xy}S_{yx}$ (with $S_{yx} = S'_{yx}$) and \mathbf{q}_1 is the dominant eigenvector of $S_{yx}S_{xy}$. With the restriction in (3), the following SIMPLS weight vectors \mathbf{r}_a and \mathbf{q}_a ($2 \leq a \leq k$) are obtained as the dominant eigenvectors of respectively $S^a_{xy}S^a_{yx}$ and $S^a_{yx}S^a_{xy}$ with S^a_{xy} the deflated cross-covariance matrix:

$$S^a_{xy} = (I_p - \mathbf{v}_{a-1}\mathbf{v}'_{a-1})S^{a-1}_{xy}$$

and $S^1_{xy} = S_{xy}$. Here $\{\mathbf{v}_1, \dots, \mathbf{v}_{a-1}\}$ represents an orthonormal basis of the x -loadings $\{\mathbf{p}_1, \dots, \mathbf{p}_{a-1}\}$ with $\mathbf{p}_j = \tilde{X}'T_j/(T'_jT_j)$, the least squares regression coefficient in the regression of \tilde{X} on T_j . Finally, note that \mathbf{p}_j can be written as $\mathbf{p}_j = \tilde{X}'\tilde{X}\mathbf{r}_j/(\mathbf{r}'_j\tilde{X}'\tilde{X}\mathbf{r}_j) = S_x\mathbf{r}_j/(\mathbf{r}'_jS_x\mathbf{r}_j)$ with S_x the empirical variance-covariance matrix of the regressors.

In the final stage of the algorithm, MLR is performed of the original y -variables on the extracted components T_1, T_2, \dots, T_k . Backtransforming these estimates to the full model (1) yields the slope estimate

$$\hat{\mathbf{B}} = R_k(R'_k S_x R_k)^{-1} R'_k S_{xy} \quad (4)$$

and the intercept

$$\hat{\boldsymbol{\beta}}_0 = \bar{\mathbf{y}} - \hat{\mathbf{B}}'\bar{\mathbf{x}}. \quad (5)$$

Tobacco data

To show the effect of one outlier on the SIMPLS weight vectors, we compute \mathbf{r}_1 and \mathbf{q}_1 explicitly for the *tobacco* data set [10], [11]. For $n = 25$ samples of tobacco leaf, $p = 6$ independent variables were measured: the percentage of total nitrogen, of chlorine, of potassium, of phosphorus, of calcium and of magnesium. A linear relationship is expected between these x -variables and $q = 3$ dependent variables (rate of cigarette burn in inches per 1000 seconds, per cent sugar in the leaf, and per cent nicotine). In [11] it was shown that $k = 1$ latent variable is satisfactory to explain this linear model.

We first estimate $(\mathbf{r}_1, \mathbf{q}_1)$ based on the raw data. Next we add contamination in the ninth observation $\mathbf{x}_9 = (1.93, 2.26, 2.15, 0.56, 3.57, 0.92)'$ by changing the third, fourth and sixth x -variable to 10 and also by adapting the response $\mathbf{y}_9 = (1.6, 17.84, 1.65)'$ into $(5, 17.84, 5)'$. We thus contaminate in the directions that were not important for the first pair of PLS weight vectors. The estimated PLS weight vectors for the SIMPLS algorithm are summarized in the first two columns of Table 1.

[Table 1 about here.]

We see that in the contaminated case the third, fourth and sixth x -variable, as well as the first and third y -variable, now become the most important directions. This is completely the opposite of what we learned from the original data. We can furthermore observe that the angle between the uncontaminated \mathbf{r}_1 and the contaminated \mathbf{r}_1 equals 89.72 degrees, whereas for \mathbf{q}_1 we find an angle of 85.39 degrees. For the next pairs of SIMPLS weight vectors a similar effect can be observed. This implies that the uncontaminated SIMPLS weight vectors are almost orthogonal to the contaminated weight vectors.

Also the SIMPLS slope matrix $\hat{\mathcal{B}}$ is highly influenced by this artificial outlier. To illustrate this effect on the 6×3 dimensional slope matrix, we split $\hat{\mathcal{B}}$ in $[\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]$ where $\hat{\beta}_i$ represents the slope vector for the i th response variable Y_i . For each response variable Y_i , we calculate the norm of the difference of the slope estimate $\hat{\beta}_i$ obtained for the regular *tobacco* data with the estimated slope $\hat{\beta}_i^c$ for the slightly contaminated data set: $n(\hat{\beta}_i) = \|\hat{\beta}_i - \hat{\beta}_i^c\|$. We obtain the following values: $n(\hat{\beta}_1) = 0.27$, $n(\hat{\beta}_2) = 3.1$ and $n(\hat{\beta}_3) = 0.78$. The coefficients of the SIMPLS slope matrix obtained with contamination are thus highly different from those obtained for the original data. We see that mainly the SIMPLS estimate of β_2 has changed a lot. We furthermore note that the angle between the uncontaminated and contaminated

estimates is almost 90 degrees for the three slope vectors. This implies that the direction of the original slope matrix has clearly changed by adding an outlier.

This example illustrates that the SIMPLS method already breaks down when only one outlier is present in the data.

2.2 The robust approach

The construction of the SIMPLS weight vectors and the parameter estimates $\hat{\mathbf{B}}$ and $\hat{\beta}_0$ completely depends on the estimation of the variance-covariance matrix Σ of the data $Z = [X, Y]$ with

$$\Sigma = E(Z - \boldsymbol{\mu})(Z - \boldsymbol{\mu})' = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} \quad (6)$$

where $\boldsymbol{\mu}$ denotes the mean of the data:

$$\boldsymbol{\mu} = \begin{bmatrix} E(X) \\ E(Y) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}.$$

The SIMPLS algorithm is obtained by estimating Σ by the empirical variance-covariance matrix S with

$$S = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})', \quad \bar{\mathbf{z}} = \left(\sum_{i=1}^n \mathbf{z}_i \right) / n$$

and $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$. However, one ‘bad’ sample will already drastically influence these estimates as illustrated with the *tobacco* data. Therefore we will robustly estimate Σ . If the number of observations is sufficiently large with respect to the number of variables ($n > 2(p+q)$), we can use the MCD [4], the reweighted MCD [5], S-estimators [6], Here we will mainly concentrate on the MCD and the RMCD estimator as they can be quickly computed by means of the FAST-MCD algorithm [12]. To deal with high-dimensional data, we refer to the method developed in [8]. Its robustness properties will be studied in Section 4. Note that if we do not apply a dimension reduction, i.e. if we take $k = p$, we obtain the robust multivariate regression as introduced in [13].

MCD method

The objective of the MCD method is to find $h > \frac{n}{2}$ observations out of n whose empirical variance-covariance matrix has the lowest determinant. The default value for h is roughly

$\lceil \alpha n \rceil$ where α reflects a lower bound for the fraction of regular observations. Usually α is taken equal to 0.5 or 0.75. The value $\alpha = 0.5$ ensures a high resistance towards outliers, but $\alpha = 0.75$ raises the finite-sample efficiency of the estimator. Note that we need that $h > p + q = m$ since else the empirical variance-covariance matrix of every h -subset is singular.

Let H denote the h -subset with minimal determinant. The MCD estimates of location and covariance are then defined as the classical mean and variance-covariance matrix of this subset:

$$\hat{\boldsymbol{\mu}}_{\text{MCD}} = \frac{1}{h} \sum_{i \in H} \mathbf{z}_i \quad \hat{\boldsymbol{\Sigma}}_{\text{MCD}} = c_1 \frac{1}{h} \sum_{i \in H} (\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{\text{MCD}})(\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{\text{MCD}})'. \quad (7)$$

The factor c_1 is a consistency factor which for normally distributed data equals $c_1 = \alpha / F_{\chi_{m+2}^2}(q_\alpha)$, and $q_\alpha = \chi_{m,\alpha}^2$, the α -quantile of the χ^2 -distribution with m degrees of freedom [15]. The function $F_{\chi_{m+2}^2}$ then denotes the distribution function of a χ_{m+2}^2 -distributed random variable.

To increase the finite-sample efficiency, a reweighted MCD estimator was proposed by [5], [12]. After obtaining the raw MCD estimators in (7), each observation receives a weight w_i :

$$w_i = \begin{cases} 1 & \text{if } (\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{\text{MCD}})' \hat{\boldsymbol{\Sigma}}_{\text{MCD}}^{-1} (\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{\text{MCD}}) \leq \chi_{m,0.975}^2, \\ 0 & \text{otherwise.} \end{cases}$$

The reweighted MCD (RMCD) estimator is then defined as the classical mean and variance-covariance matrix of those observations with non-zero weight w_i , or equivalently:

$$\hat{\boldsymbol{\mu}}_{\text{RMCD}} = \frac{\sum_{i=1}^n w_i \mathbf{z}_i}{\sum_{i=1}^n w_i} \quad \hat{\boldsymbol{\Sigma}}_{\text{RMCD}} = c_2 \frac{\sum_{i=1}^n w_i (\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{\text{RMCD}})(\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{\text{RMCD}})'}{\sum_{i=1}^n w_i}$$

where $c_2 = 0.975 / F_{\chi_{m+2}^2}(\chi_{m,0.975}^2)$ is a consistency factor at the normal distribution.

Robust PLS

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 $\boldsymbol{\Sigma}$ in (6). From this we deduce $\hat{\boldsymbol{\Sigma}}_{xy}$ and $\hat{\boldsymbol{\Sigma}}_x$ which are robust estimates of $\boldsymbol{\Sigma}_{xy}$ and $\boldsymbol{\Sigma}_x$. Analogously to the SIMPLS algorithm, the robust PLS weight vectors are the dominant eigenvectors of the robust estimate $\hat{\boldsymbol{\Sigma}}_{yx}^a \hat{\boldsymbol{\Sigma}}_{xy}^a$ and $\mathbf{r}_a = \hat{\boldsymbol{\Sigma}}_{xy}^a \mathbf{q}_a / \|\hat{\boldsymbol{\Sigma}}_{xy}^a \mathbf{q}_a\|$ with

$$\hat{\boldsymbol{\Sigma}}_{xy}^a = \begin{cases} \hat{\boldsymbol{\Sigma}}_{xy} & \text{for } a = 1, \\ (I_p - \mathbf{v}_{a-1} \mathbf{v}_{a-1}') \hat{\boldsymbol{\Sigma}}_{xy}^{a-1} & \text{for } a > 1. \end{cases} \quad (8)$$

Here $\{\mathbf{v}_1, \dots, \mathbf{v}_{a-1}\}$ is an orthonormal basis of the set of x -loadings $\{\hat{\Sigma}_x \mathbf{r}_1, \dots, \hat{\Sigma}_x \mathbf{r}_{a-1}\}$.

To obtain a robust slope, we define analogously to (4)

$$\hat{\mathbf{B}} = R_k (R_k' \hat{\Sigma}_x R_k)^{-1} R_k' \hat{\Sigma}_{xy}. \quad (9)$$

This estimator $\hat{\mathbf{B}}$ is regression, x -affine and y -affine equivariant [7], [13] if all $k = p$ components are considered in the regression stage. According to (5) the robust intercept is defined as

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\mu}}_y - \hat{\mathbf{B}}' \hat{\boldsymbol{\mu}}_x. \quad (10)$$

Tobacco data

We resume the example of the *tobacco* data and estimate the first pair of PLS weight vectors of the raw and contaminated data with this robust PLS regression method using the reweighted MCD. We then obtain the PLS weight vectors listed in the last two columns of Table 1.

First of all, we notice that the estimates at the raw data hardly differ from the SIMPLS outcomes. Also adding contamination in one sample does not change the robust PLS weight vectors very much. In this case the PLS weight vectors for the uncontaminated and contaminated data are not orthogonal but almost coincide. The angle between \mathbf{r}_1 without and \mathbf{r}_1 with contamination only equals 4.84 degrees, and for \mathbf{q}_1 this angle is 2.18 degrees. For the raw MCD estimates, the same results apply.

Also the estimate of the slope matrix is just slightly influenced by the introduced outlier. The norms $n(\hat{\boldsymbol{\beta}}_i) = \|\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_i^c\|$ are $n(\hat{\boldsymbol{\beta}}_1) = 0.13$, $n(\hat{\boldsymbol{\beta}}_2) = 0.33$ and $n(\hat{\boldsymbol{\beta}}_3) = 0.09$. Moreover the angle between the contaminated and uncontaminated estimates is only 4.84 degrees.

3 Robustness properties in low dimensions

3.1 The influence functions

To measure the influence that one observation $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ exerts on an estimator, we calculate its influence function [3]. For this, we need to consider the functional version of the estimator which defines the estimator at a certain class of distributions. E.g. for the sample average \bar{x} , the corresponding functional is the first moment $E(X) = \int x dF(x)$ that is defined at all distribution functions with a finite first moment.

To obtain the influence function of \mathbf{r}_a , we consider the contaminated distribution $F_\epsilon = (1 - \epsilon)F + \epsilon\Delta_z$ with Δ_z the distribution putting all its mass in an observation \mathbf{z} and with $F \in \mathcal{F}$, the class of all elliptical symmetric unimodal distributions (see Appendix A for more details). Then we measure how the added observation \mathbf{z} effects the outcome of \mathbf{r}_a , by computing the difference between the estimator at the contaminated and the uncontaminated distribution. Finally, the influence function looks at the infinitesimal effect by dividing this difference through the amount of contamination ϵ , and then letting ϵ go to zero. The influence function of \mathbf{r}_a is thus defined as

$$\text{IF}(\mathbf{z}, \mathbf{r}_a; F) = \lim_{\epsilon \rightarrow 0} \frac{\mathbf{r}_a((1 - \epsilon)F + \epsilon\Delta_z) - \mathbf{r}_a(F)}{\epsilon} = \left[\frac{\partial \mathbf{r}_a(F_\epsilon)}{\partial \epsilon} \right]_{|\epsilon=0}.$$

For \mathbf{q}_a and $\hat{\mathcal{B}}$ a similar definition holds.

In [14] a general expression for the influence function of the eigenvalues and eigenvectors of a symmetrical positive-definite functional G is derived. Their lemma also applies to symmetrical semi-positive definite matrices such as $\Sigma_{xy}\Sigma_{yx}$. The formulation of this lemma is given in the Appendix A (see Lemma 2).

Because $\Sigma_{yx}\Sigma_{xy}$ and $\Sigma_{xy}\Sigma_{yx}$ are at least rank 1 matrices we can apply this lemma to obtain the influence function of \mathbf{r}_a and \mathbf{q}_a . Lemma 2 furthermore implies that once $\text{IF}(\mathbf{z}, \hat{\Sigma}_{xy}; F)$ is known, the influence functions of all the PLS weight vectors can be obtained. This influence function can be calculated by applying the next lemma [14] where the influence function of any affine equivariant scatter matrix is characterized by a special form.

Lemma 1 *For any affine equivariant scatter matrix functional Σ possessing an influence function, there exists two real-valued functions α_C and β_C on $[0, \infty)$, such that*

$$\text{IF}(\mathbf{z}, \hat{\Sigma}; F) = \alpha_C(d(\mathbf{z}))(\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})' - \beta_C(d(\mathbf{z}))\Sigma$$

with $d^2(\mathbf{z})$ the squared mahalanobis distance, i.e. $d^2(\mathbf{z}) = (\mathbf{z} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{z} - \boldsymbol{\mu})$.

Whether or not the influence function of $\hat{\Sigma}$ is bounded thus depends on the behavior of the function α_C . For the empirical variance-covariance matrix $\alpha_C = \beta_C = 1$ which implies that the influence function of $\hat{\Sigma}$ is unbounded. Table 2 contains the explicit formulation of the functions α_C and β_C for MCD and RMCD at $F \sim N(\boldsymbol{\mu}, \Sigma)$. These are derived from the influence functions of the MCD and RMCD estimator [15]. For $m = p + q = 3$ and $\alpha = 0.75$, both functions are shown in Figure 1. The function α_C is redescending to zero whereas β_C attains a constant different from zero.

[Table 2 about here.]

[Figure 1 about here.]

Since the PLS weight vectors are computed in an iterative way also the influence functions are obtained step by step. Therefore we first calculate the influence function of the first pair of PLS weight vectors and generalize this result to all the pairs of PLS weight vectors. The mathematical details are given in Appendix A by Theorem 3 and Theorem 4.

The expression for the first PLS weight vector $IF(\mathbf{z}, \mathbf{r}_1; F)$ in (14) (see Appendix A, Theorem 3) simplifies when $q = 1$. This is a situation which is often encountered in chemometrical applications where the response variable represents e.g. the concentration of a certain constituent in different samples.

Corollary 1 *If $q = 1$ the influence function of the PLS weight vector \mathbf{r}_1 equals*

$$IF(\mathbf{z}, \mathbf{r}_1; F) = \alpha_C(d(\mathbf{z})) \frac{(\mathbf{y} - \boldsymbol{\mu}_y)' \mathbf{q}_1}{\lambda_{11}} (I_p - \mathbf{r}_1 \mathbf{r}_1') (\mathbf{x} - \boldsymbol{\mu}_x) \quad (11)$$

whereas $IF(\mathbf{z}, \mathbf{q}_1; F) = 0$ with λ_{11} the square root of the dominant eigenvalue of $\Sigma_{xy} \Sigma_{yx}$.

It is important to note that the influence functions of the PLS weight vectors (see (11) and (14)–(18) in Appendix A), do not depend on β_C . This implies that an observation which has an unusual large robust distance $d(\mathbf{z})$ will exert no influence on robust estimators with vanishing α_C . However, for the SIMPLS algorithm this does not hold. The classical function $\alpha_C = 1$ and the presence of ξ - and φ -terms in (17) are responsible for the unboundedness of the influence function of the SIMPLS weight vectors. However, for the MCD and the RMCD method, the influence function is bounded and even equal to zero when α_C becomes zero.

We illustrate this effect by plotting the norm of the influence function. Because of dimension limitations we take $q = 1$ and $p = 2$. For $F = N((0, 0, 0)', \Sigma)$ with

$$\Sigma = \begin{pmatrix} 5 & 1/2 & 3 \\ 1/2 & 2 & 1/3 \\ 3 & 1/3 & 2 \end{pmatrix} \quad (12)$$

and a fraction $\alpha = 0.75$ we compute the norm of $IF(\mathbf{z}, \mathbf{r}_1; F)$ as derived in (11) for every observation $\mathbf{z} = (\mathbf{x}, y)$ in $[(-8, \dots, 8), 0, (-8, \dots, 8)]$. So for each value $x_1 \in [-8, 8]$ and each value y in $[-8, 8]$ we compute the theoretical influence function (11) with Σ as in (12) which can be decomposed as in (6).

The shape of the norm of the influence function is shown in Figure 2. Figure 2(a) illustrates that the influence function of the SIMPLS weight vector is unbounded whereas the influence function of the PLS weight vectors obtained with the MCD and RMCD method in Figure 2(b) and Figure 2(c) are bounded and redescending to zero. Note that the gross-error sensitivity γ^* , defined as the supremum of $\|IF(\mathbf{z}, \mathbf{r}_1; F)\|$, for MCD $\gamma_{\text{MCD}}^* = 1.84$ is larger than $\gamma_{\text{RMCD}}^* = 1.12$.

[Figure 2 about here.]

Finally, we present the result of the influence function of the slope matrix $\hat{\mathcal{B}}$ as defined in (4) and (9). Its influence function can be immediately derived from (17) which yields $IF(\mathbf{z}, R_k; F)$ with $R_k = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_k]$.

Theorem 1 *The influence function of the slope $\hat{\mathcal{B}} = R_k(R'_k \hat{\Sigma}_x R_k)^{-1} R'_k \hat{\Sigma}_{xy}$ based on k components at $F \in \mathcal{F}$ equals:*

$$\begin{aligned} IF(\mathbf{z}, \hat{\mathcal{B}}; F) &= IF(\mathbf{z}, R_k; F)\mathcal{A} - R_k(R'_k \Sigma_x R_k)^{-1} \left\{ IF(\mathbf{z}, R'_k; F)\Sigma_x \mathcal{B} \right. \\ &\quad \left. + R'_k \Sigma_x IF(\mathbf{z}, R_k; F)\mathcal{A} - IF(\mathbf{z}, R'_k; F)\Sigma_{xy} \right\} \\ &\quad + \alpha_C(d(\mathbf{z}))R_k(R'_k \Sigma_x R_k)^{-1} R'_k (\mathbf{x} - \boldsymbol{\mu}_x) \left\{ (\mathbf{y} - \boldsymbol{\mu}_y)' - (\mathbf{x} - \boldsymbol{\mu}_x)' \mathcal{B} \right\} \end{aligned}$$

with $\mathcal{A} = (R'_k \Sigma_x R_k)^{-1} R'_k \Sigma_{xy}$.

The influence function of the SIMPLS slope matrix is unbounded. This is illustrated in Figure 3(a) where we plot the norm of the slope matrix for the covariance matrix (12) for every $\mathbf{z} = (\mathbf{x}, y)$ with $x_1 \in [-8, 8]$, $x_2 = 0$ and $y \in [-8, 8]$ when $k = 1$ component is retained. The influence function of the robust slope matrix is clearly bounded and becomes zero when a strong outlier is present (i.e. when its robust Mahalanobis distance is very large), again because of the absence of the function β_C in (17) and (18). This can be seen in Figure 3(b) for MCD and Figure 3(c) for RMCD. If the regression is based on two components, we obtain similar figures. Again, the maximal norm of the influence function of the slope obtained with MCD ($\gamma_{\text{MCD}}^* = 1.79$) is slightly larger than the maximal norm obtained with RMCD ($\gamma_{\text{RMCD}}^* = 1.08$).

[Figure 3 about here.]

Remark 1 The influence function of the intercept vector (10) $\hat{\beta}_0 = \hat{\mu}_y - \hat{\mathcal{B}}' \hat{\mu}_x$ follows immediately from Theorem 1: $\text{IF}(\mathbf{z}, \hat{\beta}_0; F) = \text{IF}(\mathbf{z}, \hat{\mu}_y; F) - \text{IF}(\mathbf{z}, \hat{\mathcal{B}}'; F) \hat{\mu}_x - \hat{\mathcal{B}}' \text{IF}(\mathbf{z}, \hat{\mu}_x; F)$. For the SIMPLS algorithm it holds that $\text{IF}(\mathbf{z}, \hat{\mu}_x; F) = \mathbf{x} - \boldsymbol{\mu}_x$ and $\text{IF}(\mathbf{z}, \hat{\mu}_y; F) = \mathbf{y} - \boldsymbol{\mu}_y$ because the sample average is used. This implies that the influence function of the SIMPLS intercept is unbounded. The bounded influence function of the MCD location estimator can be found in [15] and for the reweighted MCD we refer to [16].

3.2 An example: tobacco data

We show the behavior of the influence function on the *tobacco* data. Therefore we introduce one outlier by taking $\mathbf{z} = (\mathbf{x}_9, \mathbf{y}_9)$. We vary the third x -variable, which has the value 2.15 in the original data set, and the third y -variable, which has the value 1.65 in the original data, into a value between -30 and 30. We then measure the influence that this observation \mathbf{z} exerts on the PLS weight vectors \mathbf{r}_1 and \mathbf{q}_1 by plugging the estimates in equations (14)–(15). The effect on both PLS weight vectors $(\mathbf{r}_a, \mathbf{q}_a)$ for different values of a is similar and therefore we only show the norm of the influence function of \mathbf{r}_1 in Figure 4(a).

We estimate Σ with the MCD method and the SIMPLS method and plug the estimates into (14) for every contaminated ninth observation \mathbf{z} . The influence function is then bounded and almost everywhere equal to zero. For the original SIMPLS method we see that the influence function is unbounded. When we zoom in on the region where the original sample is located in Figure 4(b), i.e. corresponding to the case when no outlier is added to the data, we see that $\text{IF}(\mathbf{z}, \mathbf{r}_1; F)$ is not exact equal to zero, but the norm of the influence function of \mathbf{r}_1 obtained with RMCD and MCD is still smaller than the norm of the influence function of the SIMPLS weight vector which is expanding quadratically. We thus see that a regular observation does possess some influence on \mathbf{r}_1 , as we prefer, but this influence is clearly bounded. The norm of the influence function of \mathbf{r}_1 based on the raw MCD estimator attains higher values, whereas the influence function based on RMCD differs from zero in a slightly larger region.

[Figure 4 about here.]

Finally, we made similar figures to investigate the robustness of the slope matrix $\hat{\mathcal{B}} = [\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]$, where $\hat{\beta}_i$ represents the slope vector corresponding to the i th response Y_i . Because the results were very similar to those in Figure 4, we did not include them here.

3.3 Breakdown property

A robust method should also be resistant to groups of outliers, which is usually measured by means of the breakdown value [7]. The finite-sample breakdown value $\epsilon^*(T, Z_n)$ of an estimator T at a sample Z_n is defined as the smallest fraction of outliers that can cause the estimator to take on arbitrary values far from $T(Z_n)$:

$$\epsilon^*(T, Z_n) = \min_{1 \leq l \leq n} \left\{ \frac{l}{n} : \sup_{\tilde{Z}_n} \|T(Z_n) - T(\tilde{Z}_n)\| = \infty \right\}$$

where \tilde{Z}_n is obtained by replacing $l \in \{1, 2, \dots, n\}$ samples of Z_n by arbitrary samples. Simulation studies in [17] and [8] already show that the robust approach can deal with a large fraction of different types of outliers. We now show that this method inherits the breakdown value of the robust estimator of location and scatter. For this we assume that the m -dimensional observations are in general position which means that no more than m observations lie on an $(m - 1)$ -dimensional subspace.

Theorem 2 *Let $Z_n = \{(\mathbf{x}_i, \mathbf{y}_i); i = 1, \dots, n\}$ denote a set of $m = (p + q)$ -dimensional observations with $n > 2m$ and let $\hat{\boldsymbol{\mu}}_n$ and $\hat{\Sigma}_n$ denote estimators of location and scatter for $\boldsymbol{\mu}$ and Σ . Denote $\min\{\epsilon^*(\hat{\boldsymbol{\mu}}_n, Z_n), \epsilon^*(\hat{\Sigma}_n, Z_n)\} = \frac{[\epsilon^*n]}{n}$. Then the breakdown value of the estimator $(\hat{\mathcal{B}}; \hat{\boldsymbol{\beta}}_0)$ also satisfies $\epsilon^*((\hat{\mathcal{B}}; \hat{\boldsymbol{\beta}}_0), Z_n) = \frac{[\epsilon^*n]}{n}$.*

Remark 2 Theorem 2 implies that the breakdown value of the SIMPLS slope matrix and intercept is $1/n$. This illustrates again the non-robustness of the SIMPLS algorithm. For the MCD method, the breakdown value is $[\epsilon^*n]/n$ with $\epsilon^* = (n - h)/n$. The highest possible breakdown value of approximately 50% is attained by taking $h = [(n + p + q + 1)/2] \approx n/2$ [7, 13].

3.4 Computation times

The Matlab implementation of this robust PLS regression method is available at our web site www.wis.kuleuven.ac.be/stat/robust.html. The program is part of the Matlab toolbox for Robust Calibration. Note that its computation time is relatively small due to the FAST-MCD algorithm [12] which allows to compute the MCD estimator at large sample sizes. For small data sets some computation times are provided in [8]. E.g. with $n = 100$ observations and $p = 5$ (for $k = 1, 2, \dots, 5$), it only takes about six seconds on a Pentium IV with 1.60 GHz. For $n = 1000$, $p = 20$ and $k = 3$ it only requires 38.9 seconds.

4 Robustness properties in high dimensions

For cases where $n < p$ the previous section does not suffice because the theoretical results and the MCD method on which the results are based can only be applied in practice when $n > 2(p + q)$. However, the classical results still hold, i.e. the influence function of the classical PLS weight vectors, slope matrix and intercept vector are unbounded.

In [8] a robust PLS regression method (RSIMPLS) is developed for such high-dimensional data sets based on a robust PCA method in high-dimensions [9]. It is also recommended to use this method when p is larger than, say 10, as the computation of the MCD estimator then becomes less precise. No theoretical results could yet be proven about this robust PLS method but many simulation results ([17], [8]) have already proven the robustness of this method. Therefore we will study the empirical influence function and the empirical breakdown value of a specific data set. We consider the *gas* data set [18] which consists of NIR spectra of 60 gasoline samples measured in 2 nm intervals from 900 nm to 1700 nm, so $p = 401$. For each sample the octane number Y is measured ($q = 1$). We first preprocessed the data by robustly centering and robustly scaling the octane number of each gasoline sample by means of the univariate MCD estimator. We then analyzed the data using our robust techniques and no damaging outliers were found to be present in the data. Moreover, it was found by looking at robust RMSECV values that $k = 2$ components were appropriate, hence we used $k = 2$ in the sequel.

4.1 An empirical influence function

The empirical influence function is defined as

$$\frac{\hat{\beta}(\tilde{Z}_{60}) - \hat{\beta}(Z_{60})}{1/60} \quad (13)$$

with $\hat{\beta}(Z_{60})$ the estimated slope matrix for the clean *gas* data $Z_{60} = \{(\mathbf{x}_i, y_i), i = 1, 2, \dots, 60\}$ and $\hat{\beta}(\tilde{Z}_{60})$ the estimated slope matrix for the contaminated data \tilde{Z}_{60} which is obtained from Z_{60} by varying one observation.

The contamination is added in the first observation (\mathbf{x}_1, y_1) by varying the first x -variable x_{11} , which originally is -0.05, and the preprocessed response variable y_1 which is -1.33 for the clean data. We let x_{11} and y_1 take values between -40 and 40 in steps of 2 in the contaminated data \tilde{Z}_{60} . For each data set \tilde{Z}_{60} , we obtain the norm of the empirical influence

function as defined in (13). We then obtain the three-dimensional figure in Figure 5 where for each value $x_{11} \in [-40, 40]$ and for each $y_1 \in [-40, 40]$ the corresponding norm is shown.

Figure 5(a) clearly illustrates the non-robustness of the SIMPLS slope. Changing the y -variable for one observation has a huge effect on the slope estimate. It might seem that changing only one x -variable does not damage the slope estimate since the figure seems to remain flat. However, notice the high values on the z -axis. In this example, changing x_{11} does influence the slope estimate, but the damage is still bounded. This is caused by the fact that $\hat{\beta}(\tilde{Z}_{60})$ tends to a very bad slope estimate such that adding a worse type of contamination does not significantly change this estimate. The norm of (13) thus becomes constant. Changing x_{11} and y_1 simultaneously has a similar effect.

Figure 5(b) again shows the boundedness of the robust PLS method in high dimensions. The peaks in this figure appear at the value for $y_1 = -2$ (close to the original value of -1.33). Also note that the height of these peaks remains similar for the large $|x_{11}|$ values.

[Figure 5 about here.]

4.2 An empirical breakdown property

We also investigate empirically the breakdown value of the robust PLS regression method in high dimensions. Therefore we add various amounts of contamination to the *gas* data set by varying the first i observations of the clean data with $i = 0, 1, 2, \dots, 30$. This is forced by changing the univariate response to 20 for these i observations. So we add 0 to 50% of contamination to the data. For each amount of contamination we calculate the contaminated slope vector $\hat{\beta}_i^c$ and compare it to the uncontaminated slope vector $\hat{\beta}$ by calculating the norm $\|\hat{\beta} - \hat{\beta}_i^c\|$. We then plot in Figure 6 this norm against the amount of contamination added.

We already know that the classical SIMPLS method can not deal with one outlier. This is also seen in Figure 6 (solid line). With only one outlier the estimated slope is already highly influenced. If we use the RSIMPLS method where we assume that the fraction of regular observations is at least $\alpha = 75\%$ (dashed line) the method breaks down when 23.33% of contamination is present in the data. If we however take $\alpha = 0.5$ (dotted line) the method copes with up to 48.33% of irregular observations. Thus the breakdown value ϵ^* is roughly $1 - \alpha$ as for the MCD estimator.

[Figure 6 about here.]

4.3 Computation times

Also this robust PLS method is available in the Matlab toolbox for Robust Calibration. For $n = 1000$, $p = 500$, $q = 1$ and $k = 3$ it only takes 75.5 seconds. For situations with $n < p$, e.g. $n = 50$, $p = 100$, $q = 1$ and $k = 5$ it takes 6.4 seconds whereas for $n = 1000$ and $p = 2000$, $q = 1$, $k = 3$ its computation time is 352 seconds.

5 Conclusions

The SIMPLS algorithm is highly non-robust towards outlying observations. We have illustrated that a single sample can change the direction of the SIMPLS weight vectors and the regression estimates arbitrarily. This also appears in their unbounded influence functions.

In low-dimensions a bounded influence covariance estimator provides an alternative for the SIMPLS algorithm. We have shown that the influence function of all pairs of PLS weight vectors and of the regression estimates is then bounded. This makes this method resistant towards point contamination. In particular we have graphically illustrated the bounded influence function obtained with the MCD and RMCD estimator.

A robust method should however also be resistant to groups of outliers, which is usually measured by means of the breakdown value. We have shown that the breakdown value of the robust PLS regression method corresponds to the breakdown value of the robust estimator of location and scatter.

For high-dimensional data, we recommend to use the RSIMPLS method. We have illustrated on a real data set that the empirical influence function remains bounded and that it can resist large fractions of contamination.

Appendix A. Influence properties of the PLS weight vectors

To derive the influence function of the SIMPLS and the robust PLS weight vectors and the slope matrix, we introduce the functional forms of these estimators. We assume that our m -dimensional observations $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$ are sampled from an elliptical symmetric unimodal distribution $F_{\mu, \Sigma}$, with $\boldsymbol{\mu} \in \mathbb{R}^m$ and Σ an $(m \times m)$ symmetrical positive definite matrix (SPD(m)). This means that the density of $F_{\mu, \Sigma}$ can be written as

$$f_{\mu, \Sigma}(\mathbf{z}) = \frac{g((\mathbf{z} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu}))}{\sqrt{\det(\Sigma)}}$$

with g a strictly monotone decreasing function.

Let \mathcal{F} denote the set of all such distributions, then we call a map $\hat{\Sigma} : \mathcal{F} \rightarrow \text{SPD}(m)$ a statistical functional for a covariance matrix Σ . The corresponding estimator of Σ is $\hat{\Sigma}_n = \hat{\Sigma}(F_n)$ with F_n the empirical distribution of the data $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$. We also assume that $\hat{\Sigma}$ is Fisher-consistent for Σ at $F_{\mu, \Sigma}$ meaning that $\hat{\Sigma}(F_{\mu, \Sigma}) = \Sigma$ and we require that $\hat{\Sigma}$ is affine equivariant, i.e. if $Z \sim F_{\mu, \Sigma}$ then $\hat{\Sigma}(AZ + \mathbf{b}) = A\hat{\Sigma}(F)A'$ for any vector $\mathbf{b} \in \mathbb{R}^m$ and any non-singular $(m \times m)$ matrix A . Note that the notation $\hat{\Sigma}(Z) = \hat{\Sigma}(F_{\mu, \Sigma})$, if $Z \sim F_{\mu, \Sigma}$.

Furthermore we assume that $\Sigma_{yx}^a \Sigma_{xy}^a$ has at least one positive eigenvalue λ_{1a}^2 with $\lambda_{1a}^2 > \lambda_{2a}^2 \geq \dots \geq \lambda_{qa}^2 \geq 0$ and $\mathbf{q}_a = \mathbf{b}_{1a}, \dots, \mathbf{b}_{qa}$ denote the corresponding eigenvectors for $1 \leq a \leq q$. We define the eigenvectors of $\Sigma_{xy}^a \Sigma_{yx}^a$ as $\mathbf{r}_a = \mathbf{a}_{1a}, \dots, \mathbf{a}_{pa}$ with $\mathbf{a}_{ja} = \Sigma_{xy}^a \mathbf{b}_{ja} / \lambda_{ja}$ for $j = 1, \dots, q$. For ease of notation, we will use the same notation for the functionals of the eigenvectors and eigenvalues of $\Sigma_{yx}^a \Sigma_{xy}^a$ and of $\Sigma_{xy}^a \Sigma_{yx}^a$.

The Fisher-consistency of Σ at F also implies that $\Sigma_{yx}^a \Sigma_{xy}^a$ is Fisher-consistent for $\Sigma_{yx}^a \Sigma_{xy}^a$. Moreover this also means that the functionals of the weight vectors and eigenvalues are Fisher-consistent and thus $\mathbf{a}_{1a}(F) = \mathbf{r}_a$, $\mathbf{b}_{1a}(F) = \mathbf{q}_a$ and $\lambda_{ja}^2(F) = \lambda_{ja}^2$.

Denote the set of all symmetrical semi-positive definite matrices of size $d \times d$ but of rank $l \leq d$ by SSPD(d, l). The next lemma then provides the influence functions of the positive eigenvalues of $\Lambda \in \text{SSPD}(d, l)$ and of the corresponding eigenvectors.

Lemma 2 *Let $G : \mathcal{F} \rightarrow \text{SSPD}(d, l)$ be a statistical functional and F an m -dimensional distribution from \mathcal{F} . Suppose that $IF(\mathbf{z}, G; F)$ exists and let $G(F) = \Lambda$. Denote with $\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_l$ the eigenvectors of Λ with corresponding non-zero eigenvalues $\rho_1 > \dots > \rho_l$. Then the influence functions of ρ_j and $\boldsymbol{\tau}_j$ are given by:*

$$IF(\mathbf{z}, \rho_j; F) = \boldsymbol{\tau}_j' IF(\mathbf{z}, G; F) \boldsymbol{\tau}_j$$

and

$$IF(\mathbf{z}, \boldsymbol{\tau}_j; F) = \sum_{\substack{i=1 \\ i \neq j}}^l \frac{1}{\rho_j - \rho_i} (\boldsymbol{\tau}_i' IF(\mathbf{z}, G; F) \boldsymbol{\tau}_j) \boldsymbol{\tau}_i + \sum_{i=l+1}^d \frac{1}{\rho_j} (\boldsymbol{\tau}_i' IF(\mathbf{z}, G; F) \boldsymbol{\tau}_j) \boldsymbol{\tau}_i.$$

If $l = d$ we obtain the same expression as in [14].

We can now obtain the influence function of the first pair of PLS weight vectors.

Theorem 3 *Let \mathbf{q}_1 be the dominant eigenvector of $\Sigma_{yx}\Sigma_{xy}$ with corresponding eigenvalues $\lambda_{11}^2 > \dots > \lambda_{q1}^2 > 0$ and let \mathbf{r}_1 be the dominant eigenvector of $\Sigma_{xy}\Sigma_{yx}$. The influence functions of \mathbf{r}_1 and \mathbf{q}_1 are given by*

$$IF(\mathbf{z}, \mathbf{r}_1; F) = \alpha_C(d(\mathbf{z})) \left\{ \sum_{i=2}^q \frac{(\lambda_{11}\xi_{i1}\varphi_{11} + \lambda_{i1}\xi_{11}\varphi_{i1})}{\lambda_{11}^2 - \lambda_{i1}^2} \mathbf{a}_{i1} + \frac{\varphi_{11}}{\lambda_{11}} \sum_{i=q+1}^p \xi_{i1} \mathbf{a}_{i1} \right\} \quad (14)$$

$$IF(\mathbf{z}, \mathbf{q}_1; F) = \alpha_C(d(\mathbf{z})) \sum_{i=2}^q \frac{(\lambda_{i1}\xi_{i1}\varphi_{11} + \lambda_{11}\xi_{11}\varphi_{i1})}{\lambda_{11}^2 - \lambda_{i1}^2} \mathbf{b}_{i1} \quad (15)$$

with α_C , β_C and $d(\mathbf{z})$ as defined in Lemma 1, $\mathbf{r}_1 = \mathbf{a}_{11}, \mathbf{a}_{21} \dots \mathbf{a}_{p1}$ the eigenvectors of $\Sigma_{xy}\Sigma_{yx}$, $\mathbf{q}_1 = \mathbf{b}_{11}, \mathbf{b}_{21}, \dots \mathbf{b}_{q1}$ the eigenvectors of $\Sigma_{yx}\Sigma_{xy}$ and

$$(\mathbf{x} - \boldsymbol{\mu}_x)' \mathbf{a}_{i1} = \xi_{i1}, \quad (\mathbf{y} - \boldsymbol{\mu}_y)' \mathbf{b}_{j1} = \varphi_{j1} \quad (16)$$

for $i = 1, \dots, p$ and $j = 1, \dots, q$.

Proof of Theorem 3:

To obtain the influence function of \mathbf{q}_1 we first apply Lemma 2:

$$IF(\mathbf{z}, \mathbf{q}_1; F) = \sum_{i=2}^q \frac{1}{\lambda_{11}^2 - \lambda_{i1}^2} (\mathbf{b}'_{i1} IF(\mathbf{z}, \hat{\Sigma}_{yx}\hat{\Sigma}_{xy}; F) \mathbf{q}_1) \mathbf{b}_{i1}.$$

Applying the product rule and Lemma 1 immediately yields the influence function of $\hat{\Sigma}_{yx}\hat{\Sigma}_{xy}$ and thus:

$$\begin{aligned} IF(\mathbf{z}, \mathbf{q}_1; F) &= \sum_{i=2}^q \frac{1}{\lambda_{11}^2 - \lambda_{i1}^2} \left[\alpha_C(d(\mathbf{z})) \{ \mathbf{b}'_{i1} (\mathbf{y} - \boldsymbol{\mu}_y) (\mathbf{x} - \boldsymbol{\mu}_x)' \Sigma_{xy} \mathbf{q}_1 \right. \\ &\quad \left. + \mathbf{b}'_{i1} \Sigma_{yx} (\mathbf{x} - \boldsymbol{\mu}_x) (\mathbf{y} - \boldsymbol{\mu}_y)' \mathbf{q}_1 \} - 2\beta_C(d(\mathbf{z})) (\mathbf{b}'_{i1} \Sigma_{yx} \Sigma_{xy} \mathbf{q}_1) \right] \mathbf{b}_{i1}. \end{aligned}$$

Because $\mathbf{b}'_{i1} \Sigma_{yx} \Sigma_{xy} \mathbf{q}_1 = 0$ for $i \neq 1$ and $\Sigma_{xy} \mathbf{b}_{i1} = \lambda_{i1} \mathbf{a}_{i1}$ the previous equality becomes

$$IF(\mathbf{z}, \mathbf{q}_1; F) = \sum_{i=2}^q \frac{\alpha_C(d(\mathbf{z}))}{\lambda_{11}^2 - \lambda_{i1}^2} \left[\lambda_{11} \mathbf{b}'_{i1} (\mathbf{y} - \boldsymbol{\mu}_y) (\mathbf{x} - \boldsymbol{\mu}_x)' \mathbf{r}_1 + \lambda_{i1} \mathbf{a}'_{i1} (\mathbf{x} - \boldsymbol{\mu}_x) (\mathbf{y} - \boldsymbol{\mu}_y)' \mathbf{q}_1 \right] \mathbf{b}_{i1}.$$

Applying the notations from (16) finally results in expression (15).

The influence function of the first PLS weight vector \mathbf{r}_1 is obtained in the same way as for \mathbf{q}_1 or by noting that

$$\lambda_{11}(F)\mathbf{r}_1(F) = \hat{\Sigma}_{xy}(F)\mathbf{q}_1(F).$$

■

Next, we derive the influence function of the PLS weight vectors $(\mathbf{r}_a, \mathbf{q}_a)$ in general.

Theorem 4 For \mathbf{r}_a and \mathbf{q}_a the dominant eigenvectors of $\Sigma_{xy}^a \Sigma_{yx}^a$ and $\Sigma_{yx}^a \Sigma_{xy}^a$ with eigenvalues $\lambda_{1a}^2 > \lambda_{2a}^2 \geq \dots \geq \lambda_{qa}^2 \geq 0$ the influence functions are given by

$$\begin{aligned} IF(\mathbf{z}, \mathbf{r}_a; F) = & \alpha_C(d(\mathbf{z})) \sum_{i=2}^q \frac{1}{\lambda_{1a}^2 - \lambda_{ia}^2} \left(\lambda_{1a} \xi_{ia} \varphi_{1a} + \lambda_{ia} \xi_{1a} \varphi_{ia} \right) \mathbf{a}_{ia} + \frac{\alpha_C(d(\mathbf{z})) \varphi_{1a}}{\lambda_{1a}} \sum_{i=q+1}^p \xi_{ia} \mathbf{a}_{ia} \\ & - \sum_{i=2}^q \frac{1}{\lambda_{1a}^2 - \lambda_{ia}^2} \left(\lambda_{1a} \mathbf{a}'_{ia} IFv(a-1) \Sigma_{xy} \mathbf{q}_a + \lambda_{ia} \mathbf{r}'_a IFv(a-1) \Sigma_{xy} \mathbf{b}_{ia} \right) \mathbf{a}_{ia} \\ & - \frac{1}{\lambda_{1a}} \left[\sum_{i=q+1}^p \left(\mathbf{a}'_{ia} IFv(a-1) \Sigma_{xy} \mathbf{q}_a \right) \mathbf{a}_{ia} + \alpha_C(d(\mathbf{z})) \sum_{i=1}^{a-1} \mathbf{v}_i \mathbf{v}'_i (\mathbf{x} - \boldsymbol{\mu}_x) \varphi_{1a} \right] \end{aligned} \quad (17)$$

and

$$\begin{aligned} IF(\mathbf{z}, \mathbf{q}_a; F) = & \alpha_C(d(\mathbf{z})) \sum_{i=2}^q \frac{1}{\lambda_{1a}^2 - \lambda_{ia}^2} \left(\lambda_{ia} \xi_{ia} \varphi_{1a} + \lambda_{1a} \xi_{11} \varphi_{ia} \right) \mathbf{b}_{ia} \\ & - \sum_{i=2}^q \frac{1}{\lambda_{1a}^2 - \lambda_{ia}^2} \left(\mathbf{b}'_{ia} \Sigma_{yx} IFv(a-1) \Sigma_{xy} \mathbf{q}_a \right) \mathbf{b}_{ia}. \end{aligned} \quad (18)$$

for $a = 1, \dots, p$ with α_C and $d(\mathbf{z})$ as defined in Lemma 1, $\mathbf{r}_a = \mathbf{a}_{1a}, \mathbf{a}_{2a}, \dots, \mathbf{a}_{pa}$ the eigenvectors of $\Sigma_{xy}^a \Sigma_{yx}^a$, $\mathbf{q}_a = \mathbf{b}_{1a}, \mathbf{b}_{2a}, \dots, \mathbf{b}_{qa}$ the eigenvectors of $\Sigma_{yx}^a \Sigma_{xy}^a$, $IFv(a-1)$ as defined in (20) and with $(\mathbf{x} - \boldsymbol{\mu}_x)' \mathbf{a}_{ia} = \xi_{ia}$ and $(\mathbf{y} - \boldsymbol{\mu}_y)' \mathbf{b}_{ja} = \varphi_{ja}$ similar as in (16).

Proof of Theorem 4:

Before proving Theorem 4, we first obtain an expression for the influence function of the vectors \mathbf{v}_a . Therefore define $\mathbf{v}_{an} = (I_p - \sum_{j=1}^{a-1} \mathbf{v}_j \mathbf{v}'_j) \Sigma_x \mathbf{r}_a$, the unnormalized vector as a result of the Gram-Schmidt orthogonalization process of $\{\Sigma_x \mathbf{r}_1, \Sigma_x \mathbf{r}_2, \dots, \Sigma_x \mathbf{r}_a\}$ and let $\mathbf{v}_a = \mathbf{v}_{an} / \|\mathbf{v}_{an}\|$ be the normalized vector. By making use of the product rule we find that

$$\begin{aligned} IF(\mathbf{z}, \mathbf{v}_a; F) = & \frac{(I_p - \sum_{i=1}^a \mathbf{v}_i \mathbf{v}'_i) \{ \alpha_C(d(\mathbf{z})) (\mathbf{x} - \boldsymbol{\mu}_x) \xi_{1a} + \Sigma_x IF(\mathbf{z}, \mathbf{r}_a; F) \}}{\|\mathbf{v}_{an}\|} \\ & - \frac{(I_p - \mathbf{v}_a \mathbf{v}'_a) IFv(a-1) \Sigma_x \mathbf{r}_a}{\|\mathbf{v}_{an}\|} \end{aligned} \quad (19)$$

with

$$\text{IF } v(a-1) = \begin{cases} \text{IF}(\mathbf{z}, \sum_{i=1}^{a-1} \mathbf{v}_i \mathbf{v}_i'; F) & \text{if } a > 1, \\ 0 & \text{otherwise.} \end{cases} \quad (20)$$

The influence function of $\hat{\Sigma}_{yx}^a \hat{\Sigma}_{xy}^a$ can then be derived using the property that $\hat{\Sigma}_{yx}^a \hat{\Sigma}_{xy}^a = \hat{\Sigma}_{yx}(I_p - \sum_{i=1}^{a-1} \mathbf{v}_i \mathbf{v}_i') \hat{\Sigma}_{xy}$ and the product rule. Lemma 19 and Lemma 2 then lead to the expressions as stated in Theorem 4.

Corollary 2 *If $q = 1$ the influence function of \mathbf{r}_a and \mathbf{q}_a at $F \in \mathcal{F}$ satisfies*

$$\begin{aligned} \text{IF}(\mathbf{z}, \mathbf{r}_a; F) &= \frac{1}{\lambda_{1a}} \left[(I_p - \mathbf{r}_a \mathbf{r}_a') \{ \alpha_C(d(\mathbf{z})) \varphi_{1a}(\mathbf{x} - \boldsymbol{\mu}_x) - \text{IF } v(a-1) \Sigma_{xy} \mathbf{q}_a \} \right. \\ &\quad \left. - \alpha_C(d(\mathbf{z})) \sum_{i=1}^{a-1} \mathbf{v}_i \mathbf{v}_i' (\mathbf{x} - \boldsymbol{\mu}_x) \varphi_{1a} \right] \end{aligned}$$

and $\text{IF}(\mathbf{z}, \mathbf{q}_a; F) = 0$.

Remark 3 The influence function of \mathbf{r}_a in (17) has no component in the direction of \mathbf{r}_a . This follows from $\mathbf{r}_a' \text{IF}(\mathbf{z}, \mathbf{r}_a; F) = 0$ using the fact that $\|\mathbf{r}_a\| = 1$. A similar result holds for the influence function of \mathbf{q}_a in (18).

Appendix B. Robustness properties of the PLS slope matrix

Proof of Theorem 1:

Using the product rule, we obtain:

$$\begin{aligned} \text{IF}(\mathbf{z}, \hat{\mathcal{B}}; F) &= \text{IF}(\mathbf{z}, R_k; F)(R'_k \Sigma_x R_k)^{-1} R'_k \Sigma_{xy} + R_k \text{IF}(\mathbf{z}, (R'_k \hat{\Sigma}_x R_k)^{-1}; F) R'_k \Sigma_{xy} \\ &\quad + R_k (R'_k \Sigma_x R_k)^{-1} \text{IF}(\mathbf{z}, R'_k; F) \Sigma_{xy} + R_k (R'_k \Sigma_x R_k)^{-1} R'_k \text{IF}(\mathbf{z}, \hat{\Sigma}_{xy}; F). \end{aligned}$$

This expression is simplified by noting that

$$\text{IF}(\mathbf{z}, (R'_k \hat{\Sigma}_x R_k)^{-1}; F)(R'_k \Sigma_x R_k) + (R'_k \Sigma_x R_k)^{-1} \text{IF}(\mathbf{z}, (R'_k \hat{\Sigma}_x R_k); F) = 0$$

and by using Lemma 1. ■

Proof of Theorem 2:

Assume that $l < \lceil \epsilon^* n \rceil$ observations of the original set Z_n are replaced by arbitrary observations. Let \tilde{Z}_n denote the contaminated set. We obtain the breakdown value of the slope matrix $\hat{\mathcal{B}}$ by showing that the two-norm of the slope matrix $\hat{\mathcal{B}}$ is finite. Because $\hat{\mathcal{B}} = R_k (R'_k \hat{\Sigma}_x R_k)^{-1} R'_k \hat{\Sigma}_{xy}$ it holds that:

$$\begin{aligned} \|\hat{\mathcal{B}}(\tilde{Z}_n)\|_2 &= \|R_k(\tilde{Z}_n)(R_k(\tilde{Z}_n)' \hat{\Sigma}_x(\tilde{Z}_n) R_k(\tilde{Z}_n))^{-1} R_k(\tilde{Z}_n)' \hat{\Sigma}_{xy}(\tilde{Z}_n)\|_2 \\ &\leq \|R_k(\tilde{Z}_n)\|_2 \|(R_k(\tilde{Z}_n)' \hat{\Sigma}_x(\tilde{Z}_n) R_k(\tilde{Z}_n))^{-1}\|_2 \|R_k(\tilde{Z}_n)'\|_2 \|\hat{\Sigma}_{xy}(\tilde{Z}_n)\|_2. \end{aligned}$$

First, we use the property [19] that $\|R_k(\tilde{Z}_n)\|_2 \leq \sqrt{k} \|R_k(\tilde{Z}_n)\|_1 \leq \sqrt{pk}$ and $\|R_k(\tilde{Z}_n)'\|_2 \leq \sqrt{p} \|R_k(\tilde{Z}_n)'\|_\infty \leq p$ because $\|\mathbf{r}_a\| = 1$ for $a = 1, 2, \dots, k$. We also have that $\|\hat{\Sigma}_{xy}(\tilde{Z}_n)\|_2 \leq \|\hat{\Sigma}(\tilde{Z}_n)\|_2 \leq \lambda_{\max}(\hat{\Sigma}(\tilde{Z}_n))$ with $\lambda_{\max}(\hat{\Sigma}(\tilde{Z}_n))$ the maximal eigenvalue of $\hat{\Sigma}(\tilde{Z}_n)$. This eigenvalue is bounded because we only replaced $l < \lceil \epsilon^* n \rceil$ observations.

Finally, we look at $\|(R_k(\tilde{Z}_n)' \hat{\Sigma}_x(\tilde{Z}_n) R_k(\tilde{Z}_n))^{-1}\|_2 = 1/\lambda_{\min}(R_k(\tilde{Z}_n)' \hat{\Sigma}_x(\tilde{Z}_n) R_k(\tilde{Z}_n))$ with $\lambda_{\min}(R_k(\tilde{Z}_n)' \hat{\Sigma}_x(\tilde{Z}_n) R_k(\tilde{Z}_n))$ the minimal eigenvalue of $R_k(\tilde{Z}_n)' \hat{\Sigma}_x(\tilde{Z}_n) R_k(\tilde{Z}_n)$. Because we assume that this matrix $R_k(\tilde{Z}_n)' \hat{\Sigma}_x(\tilde{Z}_n) R_k(\tilde{Z}_n)$ is invertible, all its eigenvalues are different from zero and thus also $\|(R_k(\tilde{Z}_n)' \hat{\Sigma}_x(\tilde{Z}_n) R_k(\tilde{Z}_n))^{-1}\|_2 < \infty$. These results imply that $\|\hat{\mathcal{B}}(\tilde{Z}_n)\|_2 < \infty$.

For the intercept vector $\hat{\beta}_0$ it holds that $\|\hat{\beta}_0(\tilde{Z}_n)\|_2 \leq \|\hat{\boldsymbol{\mu}}_y(\tilde{Z}_n)\|_2 + \|\hat{\mathcal{B}}'(\tilde{Z}_n)\|_2 \|\hat{\boldsymbol{\mu}}_x(\tilde{Z}_n)\|_2 < \infty$ because only $l < \lceil \epsilon^* n \rceil$ observations are replaced. ■

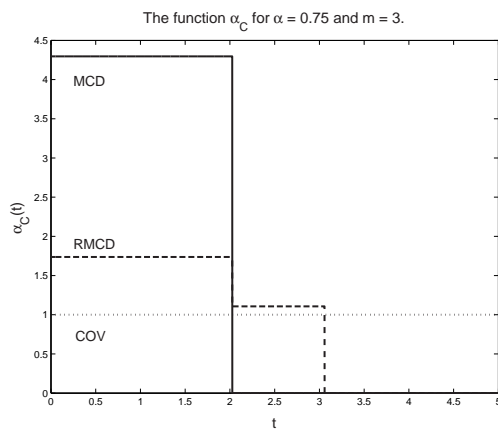
References

- [1] M. Tenenhaus, *La Regression PLS: theorie et pratique*, Editions Technip, Paris, 1998.
- [2] S. de Jong, *Chemom. and Int. Lab. Syst.*, 18 (1993) 251–263.
- [3] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York, 1986.
- [4] P.J. Rousseeuw, *J. Am. Math. Soc.*, 79 (1984) 871–880.
- [5] P.J. Rousseeuw, B.C. van Zomeren, *J. Am. Math. Soc.*, 85 (1990) 633–651.
- [6] P.J. Rousseeuw, V.J. Yohai. in J. Franke, W. Härdle and R.D. Martin, *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics No. 26, Springer-Verlag, New York 1984, 256–272.
- [7] P.J. Rousseeuw, A.M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.
- [8] M. Hubert, K. Vanden Branden, *J. Chemom.*, 17 (2003) 537–549.
- [9] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, under revision (*Technometrics*), available at www.wis.kuleuven.ac.be/stat/robust.html.
- [10] R.L. Anderson, T.A. Bancroft, *Statistical Theory in Research*, McGraw-Hill, New York, 1952.
- [11] A. Lazraq, R. Cléroux, *J. Chemom.*, 15 (2001) 523–536.
- [12] P.J. Rousseeuw, K. Van Driessen, *Technometrics*, 41 (1999) 212–223.
- [13] P.J. Rousseeuw, S. Van Aelst, K. Van Driessen, J. Agulló, Under revision (2002).
- [14] C. Croux, G. Haesbroeck, *Biometrika*, 87 (2000) 603–618.
- [15] C. Croux, G. Haesbroeck, *J. Multivariate Anal.*, 71 (1999) 161–190.
- [16] H. Lopuhaä, *Ann. Statist.*, 27 (1999) 1638–1665.

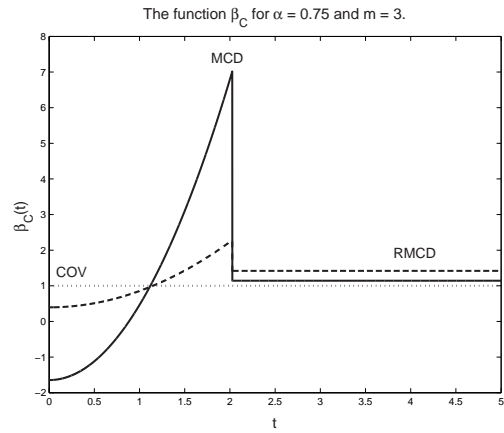
- [17] S. Engelen, M. Hubert, K. Vanden Branden, S. Verboven, to appear in Theory and Applications of Recent Robust Methods, edited by M. Hubert, G. Pison, A. Struyf and S. Van Aelst, Series: Statistics for Industry and Technology, Birkhauser, Basel, 2004.
- [18] J.H. Kalivas, Chemometr. Intel. Lab. Syst., 37 (1997) 255–259.
- [19] G.H. Golub, C.F. Van Loan, Matrix Computations, Johns Hopkins university press, Baltimore, 1989.

Figures

Figure 1



(a)



(b)

Figure 2

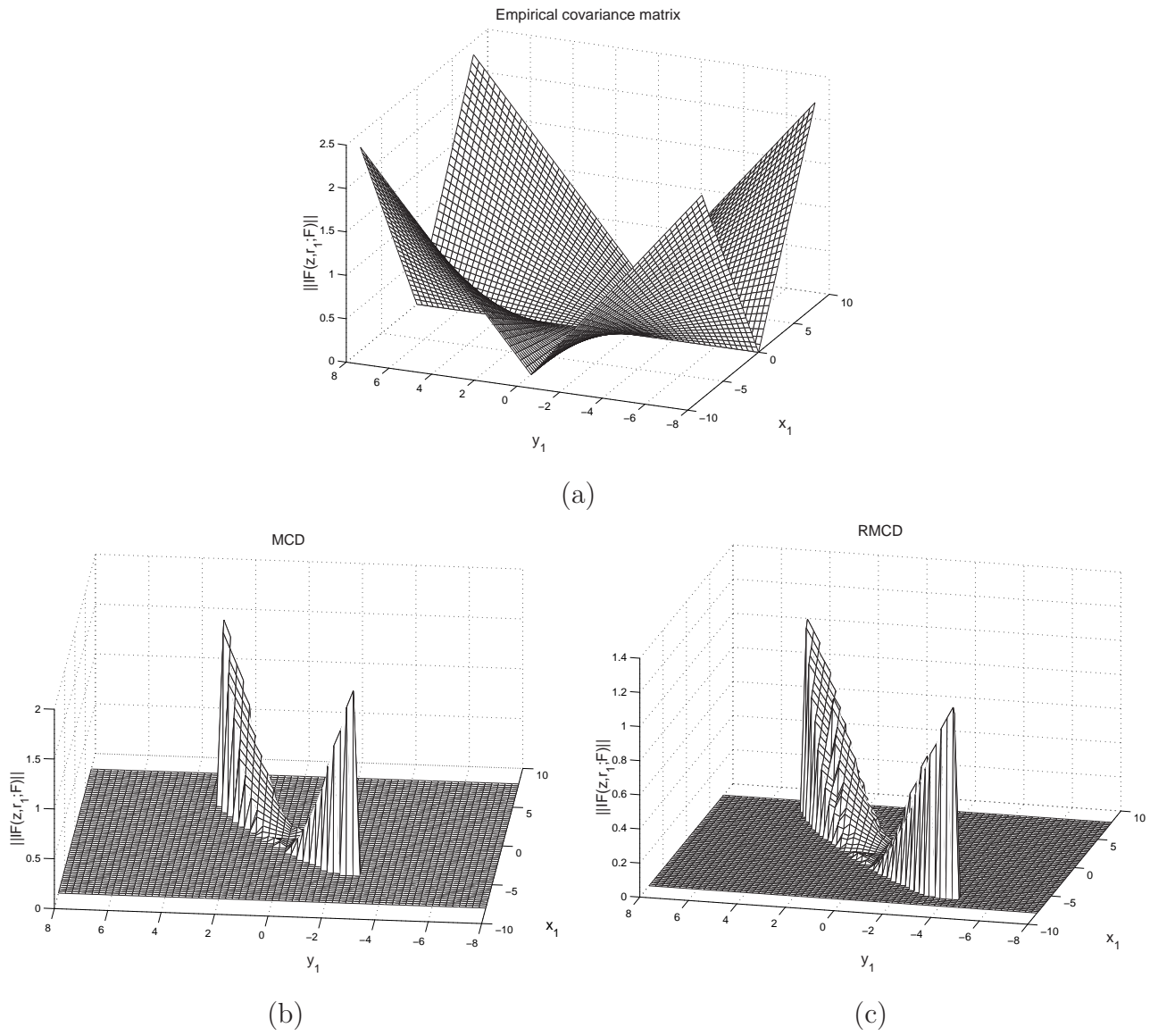
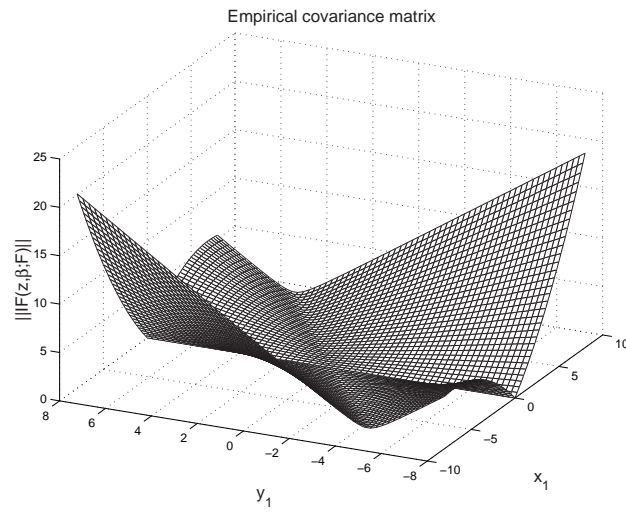
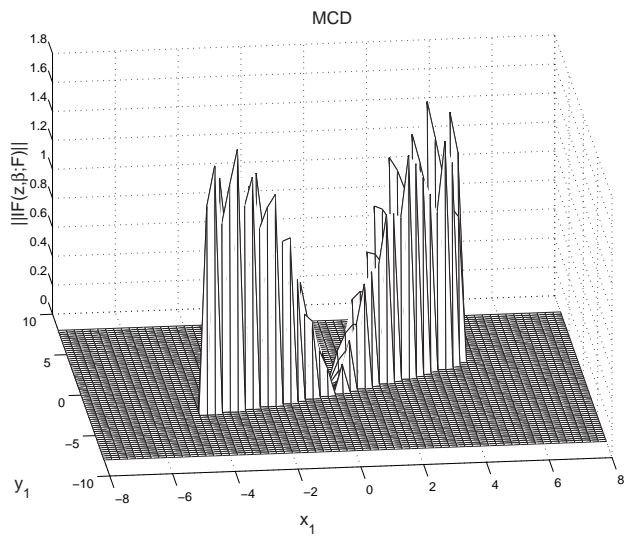


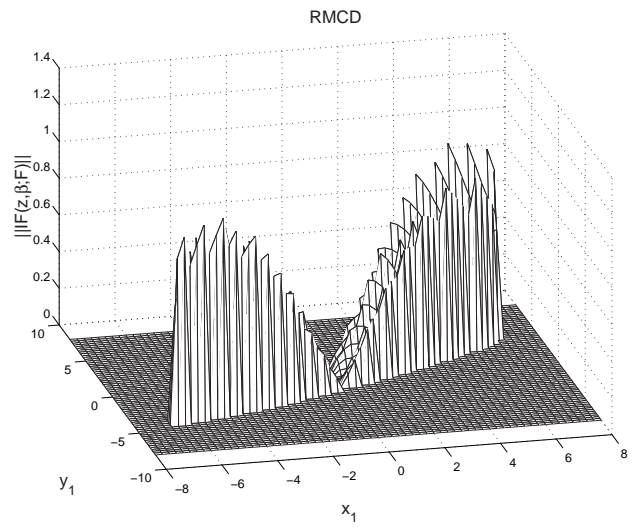
Figure 3



(a)

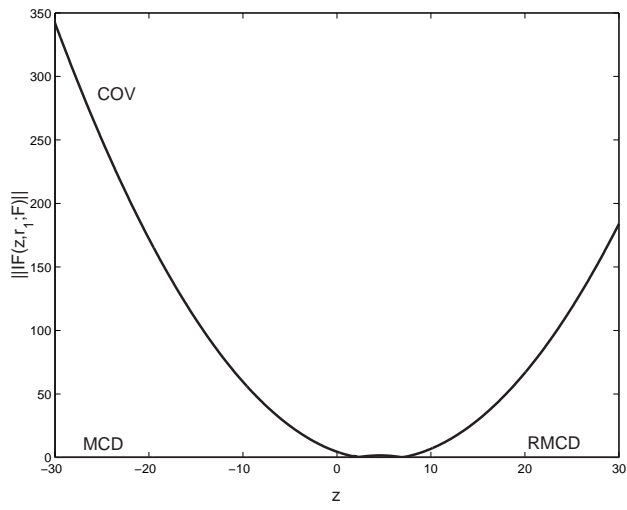


(b)

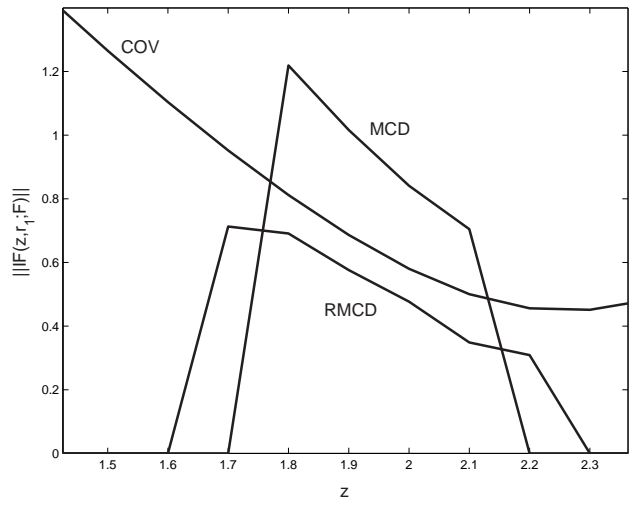


(c)

Figure 4

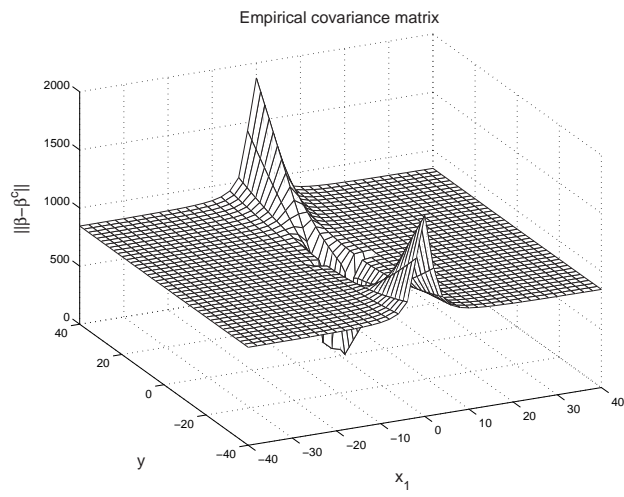


(a)

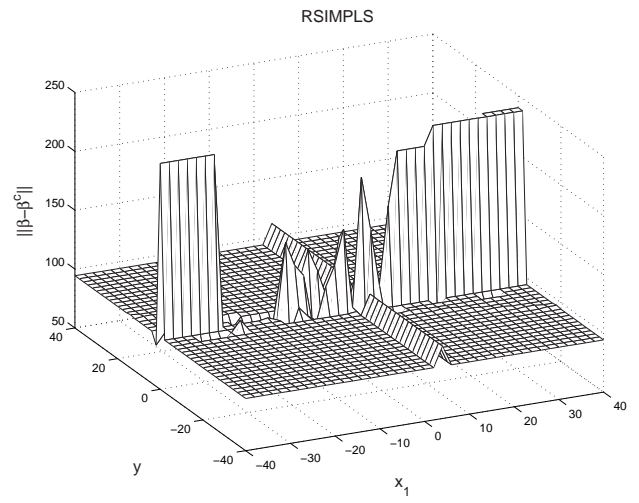


(b)

Figure 5



(a)



(b)

Figure 6

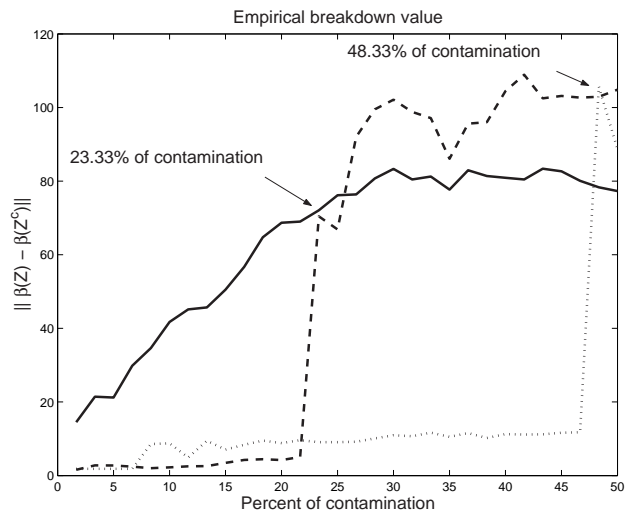


Figure Captions

Figure 1: The function α_C (a) and β_C (b) for $m = 3$ and $\alpha = 0.75$.

Figure 2: The norm of the influence function of the first SIMPLS weight vector \mathbf{r}_1 based on (a) the empirical variance-covariance estimator; (b) the MCD estimator; (c) the RMCD estimator.

Figure 3: The norm of the influence function of the slope vector $\boldsymbol{\beta}$ based on (a) the empirical variance-covariance estimator; (b) the MCD estimator; (c) the RMCD estimator.

Figure 4: The norm of the influence function of \mathbf{r}_1 for the *tobacco* data when one observation is contaminated: (a) in a large range for \mathbf{z} ; (b) in a small range for \mathbf{z} .

Figure 5: The empirical influence function of the slope vector of the PLS regression method in high dimensions for (a) SIMPLS (b) robust SIMPLS.

Figure 6: The empirical breakdown value of the SIMPLS (solid line) and robust PLS regression method in high dimensions for $\alpha = 0.75$ (dashed line) and $\alpha = 0.5$ (dotted line).

Tables

Table 1

| | SIMPLS | | Robust PLS | |
|----------------|----------|-------------------|------------|-------------------|
| | Raw data | Contaminated data | Raw data | Contaminated data |
| \mathbf{r}_1 | 0.51 | 0.03 | 0.51 | 0.54 |
| | -0.67 | 0.00 | -0.68 | -0.67 |
| | -0.12 | -0.51 | -0.12 | -0.05 |
| | -0.02 | -0.62 | -0.02 | -0.02 |
| | 0.50 | 0.01 | 0.49 | 0.48 |
| | 0.17 | -0.59 | 0.17 | 0.16 |
| \mathbf{q}_1 | 0.05 | -0.73 | 0.04 | 0.07 |
| | -0.97 | -0.27 | -0.98 | -0.98 |
| | 0.24 | -0.62 | 0.20 | 0.18 |

Table 1: PLS weight vectors of the *tobacco* data obtained with SIMPLS and a robust PLS regression method based on the RMCD estimator.

Table 2

| | |
|---------------------------|---|
| $\alpha_{\text{MCD}}(t)$ | $\frac{I(t \leq \sqrt{q_\alpha})}{f_1(\alpha, m+4)}$ |
| $\alpha_{\text{RMCD}}(t)$ | with $f_1(x, y) = F_{\chi_y^2}(q_x)$ and $q_x = \chi_{m,x}^2$ $\alpha_{\text{MCD}}(t)(1 - f_2(0.975, m)) + \frac{I(t \leq \sqrt{q_{0.975}})}{f_1(0.975, m+2)}$ with $f_2(x, y) = \frac{f_1(x, y+4)}{f_1(x, y+2)}$ |
| $\beta_{\text{MCD}}(t)$ | $\frac{1}{f_2(\alpha, m) - mf_3} \left\{ 1 + I(t \leq \sqrt{q_\alpha}) \left(\frac{q_\alpha}{f_1(\alpha, m+2)m} - \frac{f_3 t^2}{f_1(\alpha, m+4)} \right) - \frac{\alpha q_\alpha}{f_1(\alpha, m+2)m} \right\}$ with $f_3 = \frac{1}{2} (1 - f_2(\alpha, m)) + \frac{q_\alpha}{m} \left(1 - \frac{\alpha}{2f_1(\alpha, m+2)} \right)$ |
| $\beta_{\text{RMCD}}(t)$ | $1 + (1 - f_2(0.975, m)) \left(\beta_{\text{MCD}}(t) \left(\frac{m}{2} + 1 \right) - \frac{t^2 \alpha_{\text{MCD}}(t)}{2} \right)$ |

Table 2: The expressions for α_C and β_C at $F \sim N(\boldsymbol{\mu}, \Sigma)$ for MCD and RMCD.