DetMCD in a Calibration Framework

3 authors:

Tim Verdonck
University of Leuven
27 PUBLICATIONS   397 CITATIONS

Mia Hubert
University of Leuven
109 PUBLICATIONS   4,844 CITATIONS

Peter Rousseeuw
University of Leuven
209 PUBLICATIONS   32,638 CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   Functional Outlier Detection View project

Project   Deterministic estimation of location and scatter View project

DetMCD in a Calibration Framework

Tim Verdonck[1], Mia Hubert[2], and Peter J. Rousseeuw[3]

[1] Department of Mathematics and Computer Science, University of Antwerp
   Middelheimlaan 1, Antwerp, Belgium, *Tim.Verdonck@ua.ac.be*
[2] Department of Mathematics, Katholieke Universiteit Leuven
   Celestijnenlaan 200b, Leuven, Belgium, *Mia.Hubert@wis.kuleuven.be*
[3] Department of Mathematics, Katholieke Universiteit Leuven
   Celestijnenlaan 200b, Leuven, Belgium, *peter@rousseeuw.net*

Abstract. The minimum covariance determinant (MCD) method is a robust estimator of multivariate location and scatter (Rousseeuw (1984)). Computing the exact MCD is very hard, so in practice one resorts to approximate algorithms. Most often the FASTMCD algorithm of Rousseeuw and Van Driessen (1999) is used. The FASTMCD algorithm is affine equivariant but not permutation invariant. Recently a deterministic algorithm, denoted as DetMCD, is developed which does not use random subsets and which is much faster (Hubert et al. (2010)). In this paper DetMCD is illustrated in a calibration framework. We focus on robust principal component regression and partial least squares regression, two very popular regression techniques for collinear data. We also apply DetMCD on data with missing elements after plugging it into the M-RPCR technique of Serneels and Verdonck (2009).

Keywords: deterministic algorithm, outliers, robustness, RPCR, RSIMPLS

1 Introduction

The Minimum Covariance Determinant (MCD) method of Rousseeuw (1984) is a highly robust estimator of multivariate location and scatter. Given an $n \times p$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$, its objective is to find h observations (with $n/2 \le h \le n$) whose covariance matrix has the lowest determinant. The MCD estimate of location is then the average of these h points, and the scatter estimate is a multiple of their covariance matrix. The MCD has a bounded influence function and can attain the highest possible breakdown value (i.e. 50%) when $h = \lfloor (n+p+1)/2 \rfloor$. In addition to being highly resistant to outliers, the MCD is affine equivariant, i.e. the estimates behave properly under affine transformations of the data.

Although the MCD was already introduced in 1984, its practical use only became feasible since the introduction of the computationally efficient FASTMCD algorithm of Rousseeuw and Van Driessen (1999). The FASTMCD algorithm starts by drawing random subsets of size $p + 1$. It needs to draw many in order to obtain at least one that is outlier-free.

Recently, Hubert et al. (2010) have developed a deterministic algorithm for the MCD, denoted as DetMCD, which does not use random subsets and runs even faster than FASTMCD. Unlike the latter it is permutation invariant, i.e. the result does not depend on the order of the observations in the data set. It starts from only a few well-chosen initial estimates. By an extensive simulation study Hubert et al. (2010) have shown that DetMCD is as robust as FASTMCD and that the lack of affine equivariance is small. In Hubert et al. (2010) the performance of the DetMCD algorithm is illustrated in the context of principal component analysis, discriminant analysis, and MCD regression (Rousseeuw et al. (2004)). The latter method is a robust multivariate regression technique for low-dimensional predictors $\boldsymbol{x}_i$ and vector-valued response variables $\boldsymbol{y}_i$. The MCD regression estimates are obtained by matrix operations on the MCD location and scatter estimates of the joint $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ data. In this paper we investigate the use of the DetMCD algorithm in robust principal component regression (RPCR) and robust partial least squares regression (RSIMPLS). These two regression techniques fit a linear relationship between two sets of variables and are mostly used when the number of independent variables $\boldsymbol{x}_i$ is very large or when the regressors are highly correlated (also known as multicollinearity).

In Section 2 we describe the DetMCD algorithm in detail, whereas in Section 3 we briefly summarize RPCR and RSIMPLS. Section 4 presents the results of a simulation study in which we investigate the effect of replacing the FASTMCD algorithm with the DetMCD algorithm. We compare the robustness of the algorithms by adding different percentages of contamination in the simulated data sets, and we also compare their computation times. Moreover, both algorithms for MCD are compared on data with missing elements after plugging them into the M-RPCR method of Serneels and Verdonck (2009).

2   The DetMCD algorithm

In this section we describe the deterministic algorithm to compute the MCD, developed in Hubert et al. (2010). Given the data matrix $\boldsymbol{X}$ with rows $\boldsymbol{x}_i^T$, we denote the columns as X_j $(j = 1, \ldots, p)$. For a data set $\boldsymbol{X}$ with estimated center $\hat{\boldsymbol{\mu}}$ and scatter matrix $\hat{\boldsymbol{\Sigma}}$, the statistical distance of the i-th observation $\boldsymbol{x}_i$ is written as

$$D(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})}.$$

2.1   General procedure

First, each variable X_j is standardized by subtracting its median and dividing by the Q_n scale estimator of Rousseeuw and Croux (1993). This standardization makes the algorithm location and scale equivariant. The standardized data set is denoted by $\boldsymbol{Z}$ with rows $\boldsymbol{z}_i^T$ $(i = 1, \ldots, n)$ and columns Z_j $(j = 1, \ldots, p)$.

Next, seven initial estimates $\hat{\boldsymbol{\mu}}_l(\boldsymbol{Z})$ and $\hat{\boldsymbol{\Sigma}}_l(\boldsymbol{Z})$ $(l = 1, \ldots, 7)$ are constructed for the center and scatter of $\boldsymbol{Z}$. Apart from the last one, each computes a preliminary estimate $\boldsymbol{S}_l$ of the covariance or correlation matrix of $\boldsymbol{Z}$. They will be described in Section 2.2. As these $\boldsymbol{S}_l$ may have very inaccurate eigenvalues, the following steps are applied to each. The first two steps are performed to make the robust scatter matrix positive definite and more affine equivariant. They are similar to steps in the orthogonalized Gnanadesikan-Kettenring (OGK) algorithm (Maronna and Zamar (2002)).

1. Compute the matrix $\boldsymbol{E}$ of eigenvectors of $\boldsymbol{S}_l$ and put $\boldsymbol{B} = \boldsymbol{Z}\boldsymbol{E}$.
2. Estimate the covariance of $\boldsymbol{Z}$ by $\hat{\boldsymbol{\Sigma}}_l(\boldsymbol{Z}) = \boldsymbol{E}\boldsymbol{L}\boldsymbol{E}^T$ where
   $\boldsymbol{L} = \mathrm{diag}\left(Q_n^2(B_1), \ldots, Q_n^2(B_p)\right)$.
   Here $Q_n(B_1)$ is the Q_n scale estimator applied to the first column of $\mathbf{B}$.
3. To estimate the center of $\boldsymbol{Z}$ sphere the data, apply the coordinatewise median, and transform it back, i.e. $\hat{\boldsymbol{\mu}}_l(\boldsymbol{Z}) = \hat{\boldsymbol{\Sigma}}_l^{1/2}(\mathrm{med}(\boldsymbol{Z}\hat{\boldsymbol{\Sigma}}_l^{-1/2}))$.

For all estimates $(\hat{\boldsymbol{\mu}}_l(\boldsymbol{Z}), \hat{\boldsymbol{\Sigma}}_l(\boldsymbol{Z}))$ we compute the statistical distances

$$d_{i,l} = D(\boldsymbol{z}_i, \hat{\boldsymbol{\mu}}_l(\boldsymbol{Z}), \hat{\boldsymbol{\Sigma}}_l(\boldsymbol{Z})).$$

For each initial estimate l the h observations with smallest $d_{i,l}$ are taken and *concentration steps* (C-steps) are applied until convergence. (A C-step reduces the MCD objective function and is a major component of the FASTMCD algorithm.) The solution with smallest determinant is called the raw DetMCD. As in the FASTMCD algorithm, we then compute reweighted estimates to increase statistical efficiency while retaining high robustness.

2.2 Initial estimates

1) The first initial scatter estimate is obtained by computing the hyperbolic tangent (sigmoid) of each column of $\boldsymbol{Z}$, i.e. $Y_j = \tanh(Z_j)$ for $j = 1, \ldots, p$. Computing the classical correlation matrix of $\boldsymbol{Y}$ yields $\boldsymbol{S}_1 = \mathrm{corr}(\boldsymbol{Y})$.
2) Let R_j be the ranks of the column Z_j, and put $\boldsymbol{S}_2 = \mathrm{corr}(\boldsymbol{R})$. This is the Spearman correlation matrix of $\boldsymbol{Z}$.
3) For $\boldsymbol{S}_3$ normal scores are computed from these ranks, namely $T_j = \Phi^{-1}((R_j - 1/3)/(n + 1/3))$ where $\Phi(.)$ is the normal cumulative distribution function, and then we set $\boldsymbol{S}_3 = \mathrm{corr}(\boldsymbol{T})$.
4) The fourth scatter estimate is based on the spatial sign covariance matrix of Visuri et al. (2000): define $\boldsymbol{a}_i = \boldsymbol{z}_i/\|\boldsymbol{z}_i\|$ for all i and let $\boldsymbol{S}_4 = \mathrm{cov}(\boldsymbol{A})$.
5) For $\boldsymbol{S}_5$ we take the covariance matrix of the $\lceil n/2 \rceil$ standardized observations $\boldsymbol{z}_i$ with smallest norm.
6) The sixth scatter estimate is the raw OGK estimator.
7) Finally the classical mean $\hat{\boldsymbol{\mu}}_7(\boldsymbol{Z})$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_7(\boldsymbol{Z})$ of the full data set are used.

3   Robust calibration methods

In practice one often needs to estimate a linear relation between an $n \times p$ predictor data matrix $\boldsymbol{X}$ and an $n \times q$ predictand matrix $\boldsymbol{Y}$. When the errors are normally distributed, the optimal solution to this problem is to use the least squares estimator. However, when the number of predictors exceeds the number of cases the least squares regression estimator cannot be computed, and when the predictor data matrix $\boldsymbol{X}$ contains highly correlated columns the method is numerically unstable. Two popular regression techniques that tackle these problems are principal component regression (PCR) and partial least squares regression (PLSR).

The idea behind PCR is to replace the original regressors by their principal component scores ($\boldsymbol{T}$). Hubert and Verboven (2003) have proposed a robust PCR method (RPCR) by robustifying both steps of PCR. First a robust principal component analysis (PCA) method is applied to the regressors. For low-dimensional data the MCD estimator is used for this, whereas for high-dimensional data ROBPCA (Hubert et al. (2005)) is applied. The latter is a hybrid method that combines projection pursuit with the MCD. Next, a robust regression is performed with the robust scores as predictor variables. For a univariate response variable ($q = 1$) this is done by means of LTS regression (Rousseeuw (1984)), and for $q > 1$ by means of MCD regression.

In PLSR the scores are computed by maximizing a covariance criterion between the $\boldsymbol{x}$- and $\boldsymbol{y}$-variables. Unlike PCR, this technique uses the responses already from the start. A well-known PLSR method is the SIMPLS algorithm (de Jong (1993)). A robust SIMPLS method, RSIMPLS (Hubert and Vanden Branden (2003)), starts by applying ROBPCA to the $\boldsymbol{x}$- and $\boldsymbol{y}$-variables and then proceeds analogously to the SIMPLS algorithm. In the second stage of the algorithm again a robust regression is applied.

Both RPCR and RSIMPLS thus apply the MCD estimator. For RPCR with $q > 1$ this is done in the PCA and in the regression step. RSIMPLS uses MCD in the first stage only, as part of ROBPCA.

When missing values occur in the data, RPCR and RSIMPLS cannot be applied anymore. However, in Serneels and Verdonck (2009) a method (M-RPCR) is presented to perform RPCR on data with missing elements according to the missing at random mechanism (MAR). As the algorithm is based on the expectation-maximization approach, it is iterative and consequently it applies MCD many times. Note that for RSIMPLS the same methodology could be applied, but this has not been worked out yet.

4   Simulation study

In this section we first compare RPCR and RSIMPLS using the FASTMCD and the DetMCD algorithms on several simulated data sets without missing values. Different types of outliers are added to the data. This allows to

study the efficiency at uncontaminated data, as well as the robustness at contaminated data. Next, we study the M-RPCR method on data with missing elements.

4.1 Simulation design

A straightforward way to set up a simulation for PCR and SIMPLS is to generate data according to the bilinear latent variable model (Burnham et al. (1999)) with given complexity k. We will consider

$$\begin{cases} \boldsymbol{X} = \boldsymbol{T}_k \boldsymbol{P}_k^T + N_p\left(\boldsymbol{0}_p, 0.1\boldsymbol{I}_p\right) \\ \boldsymbol{Y} = \boldsymbol{T}_k \boldsymbol{Q}_k^T + N_q\left(\boldsymbol{0}_q, 0.1\boldsymbol{I}_q\right) \end{cases} \tag{1}$$

where $q = 3$, $k = 2$, $\boldsymbol{T}_k \sim N_k\left(\boldsymbol{0}_k, \boldsymbol{\Sigma}\right)$, and $\boldsymbol{\Sigma} = \mathrm{diag}(6, 2)$. We will specify p later. For the matrix $\boldsymbol{Q}_k$ we took

$$\boldsymbol{Q}_k = \begin{pmatrix} -2 & 1 & 2 \\ 1 & -1 & -2 \end{pmatrix}.$$

The loadings $\boldsymbol{P}_k$ are defined as the eigenvectors of the covariance matrix of k independent uniform variables on [0,1]. The vector $\boldsymbol{0}_p$ denotes the vector of length p with all entries equal to zero, and $\boldsymbol{I}_p$ is the identity matrix of size p.

The contaminated parts of the data $\boldsymbol{T}_\varepsilon$, $\boldsymbol{X}_\varepsilon$, and $\boldsymbol{Y}_\varepsilon$ were constructed as follows. Bad leverage points were generated as $\boldsymbol{X}_\varepsilon = \boldsymbol{T}_\varepsilon \boldsymbol{P}_k^T + N_p(\boldsymbol{0}_p, 0.1\boldsymbol{I}_p)$ with $\boldsymbol{T}_\varepsilon \sim N_k\left(30\,\boldsymbol{1}_k, 0.1\boldsymbol{\Sigma}\right)$. Vertical outliers are obtained as $\boldsymbol{Y}_\varepsilon = \boldsymbol{T}_k \boldsymbol{Q}_k^T + N_p\left(30\,\boldsymbol{1}_p, 0.1\boldsymbol{I}_p\right)$.

Note that $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the latent variable model (1) satisfy the regression relation

$$\boldsymbol{Y} = \boldsymbol{X}\mathfrak{B} + \boldsymbol{E}$$

with $\mathfrak{B} = \boldsymbol{P}_k \boldsymbol{Q}_k^T$ and $\boldsymbol{E}$ normally distributed errors. In order to evaluate the different methods, the following criteria are used:

- The bias in the regression coefficients:

$$e_B = \frac{1}{pq} \left\| \mathfrak{B} - \hat{\mathfrak{B}} \right\|_F,$$

  where $\| \cdot \|_F$ denotes the Frobenius norm of the matrix.
- The predictive ability for a test set $(\boldsymbol{X}_t, \boldsymbol{Y}_t)$ of the same size as $(\boldsymbol{X}, \boldsymbol{Y})$:

$$e_P = \frac{1}{nq} \left\| \boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_t \right\|_F.$$

- The computation time t (in seconds).

Each of these performance measures should be as close to zero as possible. We considered both low-dimensional ($n = 100$ and $p = 6$) and high-dimensional ($n = 40$ and $p = 200$) data, and 100 data sets were generated for each situation. The number of selected components in RPCR and RSIMPLS were fixed to the actual complexity $k = 2$.

For the contamination percentage ϵ we chose the values 0%, 10% and 25%. The parameter $\alpha = (n - h)/n$, denoting the fraction of outliers the method should be able to resist, was set to 25%, 25% and 50% respectively. To study M-RPCR with FASTMCD and DetMCD we randomly replaced 10% of the elements of $\boldsymbol{X}$ by missing values.

All simulations were carried out in MATLAB 7.4 (The MathWorks, Natick, MA). Many functions were taken from LIBRA, the Matlab library for Robust Analysis (Verboven and Hubert (2005)).

4.2   Results

In the following tables we report for each performance criterion the average over 100 runs. Tables 1 and 2 show the results for the low-dimensional and the high-dimensional data without missing elements. We can conclude that the algorithms perform similarly well for the first two performance criteria, irrespective of the data dimension. However, we see that the algorithms differ in computation time. As DetMCD is much faster than FASTMCD, its computational advantage carries over to RPCR and RSIMPLS. The speedup is most prominent for RPCR in low dimensions, because MCD is applied more often there than in RSIMPLS. In high dimensions this effect is reduced as RPCR then applies ROBPCA to the regressors, and ROBPCA includes MCD but also a time-consuming projection pursuit part that is not changed.

		Clean	Bad leverage		Vertical outliers	
	ϵ	0	0.10	0.30	0.10	0.30
RPCR (FASTMCD)	e_B	0.0147	0.0144	0.0153	0.0150	0.0175
	e_P	0.0446	0.0447	0.0450	0.0447	0.0450
	t	3.15	3.14	3.13	3.15	3.14
RPCR (DetMCD)	e_B	0.0143	0.0143	0.0152	0.0147	0.0159
	e_P	0.0446	0.0447	0.0449	0.0447	0.0449
	t	0.51	0.51	0.53	0.51	0.52
RSIMPLS (FASTMCD)	e_B	0.0147	0.0146	0.0154	0.0149	0.0163
	e_P	0.0446	0.0447	0.0450	0.0447	0.0449
	t	2.42	2.41	2.40	2.41	2.40
RSIMPLS (DetMCD)	e_B	0.0145	0.0145	0.0153	0.0148	0.0162
	e_P	0.0446	0.0447	0.0450	0.0447	0.0449
	t	1.76	1.75	1.76	1.76	1.76

Table 1. Simulation results for low-dimensional data.

		Clean	Bad leverage		Vertical outliers	
	ϵ	0	0.10	0.30	0.10	0.30
RPCR (FASTMCD)	e_B	0.0025	0.0026	0.0029	0.0025	0.0025
	e_P	0.0948	0.0974	0.1072	0.0946	0.0969
	t	3.79	3.78	3.78	3.79	3.78
RPCR (DetMCD)	e_B	0.0025	0.0026	0.0028	0.0025	0.0025
	e_P	0.0943	0.0976	0.1074	0.0946	0.0959
	t	1.90	1.89	1.89	1.90	1.90
RSIMPLS (FASTMCD)	e_B	0.0027	0.0027	0.0030	0.0027	0.0030
	e_P	0.0967	0.0979	0.1129	0.0992	0.1125
	t	2.38	2.37	2.36	2.38	2.37
RSIMPLS (DetMCD)	e_B	0.0026	0.0027	0.0030	0.0027	0.0030
	e_P	0.0962	0.0973	0.1118	0.0983	0.1100
	t	1.74	1.73	1.72	1.74	1.73

Table 2. Simulation results for high-dimensional data.

The same conclusions can be drawn when missing values are added to the data, as seen in Tables 3 and 4. Because the M-RPCR method iterates RPCR several times, the RPCR speedup is very useful.

		Clean	Bad leverage		Vertical outliers	
	ϵ	0	0.10	0.30	0.10	0.30
M-RPCR (FASTMCD)	e_B	0.0170	0.0170	0.0179	0.0172	0.0199
	e_P	0.0450	0.0455	0.0457	0.0455	0.0459
	t	45.89	44.09	64.11	45.65	69.22
M-RPCR (DetMCD)	e_B	0.0169	0.0169	0.0180	0.0170	0.0189
	e_P	0.0450	0.0455	0.0456	0.0455	0.0460
	t	5.03	4.91	7.05	4.91	5.59

Table 3. Simulation results for low-dimensional data with missing values.

		Clean	Bad leverage		Vertical outliers	
	ϵ	0	0.10	0.30	0.10	0.30
M-RPCR (FASTMCD)	e_B	0.0027	0.0027	0.0030	0.0027	0.0027
	e_P	0.1001	0.1022	0.1149	0.0995	0.1005
	t	14.81	14.83	18.25	15.19	15.51
M-RPCR (DetMCD)	e_B	0.0027	0.0027	0.0030	0.0027	0.0027
	e_P	0.0995	0.1024	0.1147	0.0993	0.1000
	t	4.83	4.74	5.83	4.87	4.67

Table 4. Simulation results for high-dimensional data with missing values.

5   Summary and conclusion

In this paper we have illustrated our recently proposed deterministic algorithm for MCD in a calibration framework. Replacing FASTMCD by DetMCD in the robust regression techniques RPCR and RSIMPLS gives similar results concerning robustness and predictive ability, but with improved computation speed. This becomes even more important when the data also contain missing elements and the DetMCD algorithm is applied many times in an iterative way. We conclude that DetMCD is a fast and robust alternative to FASTMCD in this calibration framework.

References

BURNHAM, A.J., MACGREGOR, J.F. and VIVEROS, R. (1999): Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems 48(2), 167-180.*

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

HUBERT, M., ROUSSEEUW, P.J. and VANDEN BRANDEN, K. (2005): ROBPCA: a new approach to robust principal component analysis. *Technometrics 47, 64-79.*

HUBERT, M., ROUSSEEUW, P.J. and VERDONCK, T. (2010): A deterministic algorithm for the MCD. *Submitted.*

HUBERT, M. and VANDEN BRANDEN, K. (2003): Robust methods for partial least squares regression. *Journal of Chemometrics 17, 537-549.*

HUBERT, M. and VERBOVEN, S. (2003): A robust PCR method for high-dimensional regressors. *Journal of Chemometrics 17, 438-452.*

MARONNA, R.A. and ZAMAR, R.H. (2002): Robust estimates of location and dispersion for high-dimensional data sets. *Technometrics 44, 307-317.*

ROUSSEEUW, P.J. (1984): Least median of squares regression. *Journal of the American Statistical Association 79, 871-880.*

ROUSSEEUW, P.J. and CROUX, C. (1993): Alternatives to the median absolute deviation. *Journal of the American Statistical Association 88, 1273-1283.*

ROUSSEEUW, P.J., VAN AELST, S., VAN DRIESSEN, K. and AGULLO, J. (2004) Robust multivariate regression. *Technometrics 46, 293-305.*

ROUSSEEUW, P.J. and VAN DRIESSEN, K. (1999): A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41, 212-223.*

SERNEELS, S. and VERDONCK, T. (2009): Principal component regression for data containing outliers and missing elements. *Computational Statistics and Data Analysis 53(11), 3855-3863.*

VERBOVEN, S. and HUBERT, M. (2005): LIBRA: a Matlab library for robust analysis. *Chemometrics and Intelligent Laboratory Systems 75, 127-136.*

VISURI, S., KOIVUNEN, V. and OJA, H. (2000): Sign and rank covariance matrices. *Journal of Statistical Planning and Inference 91, 557-575.*