Robust lower-rank approximation of data matrices with element-wise contamination

Ricardo A. Maronna[(a)]*and Víctor J. Yohai[(b)]
[(a)]University of La Plata and C.I.C.P.B.A.
[(b)]University of Buenos Aires and CONICET

Abstract

In this paper we propose a robust method to approximate an $n \times p$ data matrix with one of given rank q. The method is based on Yohai's (1987) regression MM estimates. It is intended to be resistant against the existence of both atypical rows and of scattered atypical cells, and to be able to cope with missing values. We propose an algorithm based on alternating M-regressions, and a starting estimate based on successive rank-one fits, which involves $O(npq)$ operations. Simulations show our estimate to outperform competing estimates, both in efficiency and resistance. Three high dimensional real datasets are analyzed. The running time of our estimate for large datasets is shown to be less than that of its competitors.

AMS subject clasification: Primary 62F35, Secondary 62J05

Key words and phrases: MM estimate. RAR estimate. Principal components. Multivariate outliers. Alternating regressions.

1 Introduction

Principal Components Analysis (PCA) is a widely used multivariate data-analytic tool which aims at finding a small number q of linear combinations of the p observed variables, which explain most of the variability in the data. PCA proceeds by finding directions of maximum or minimum variability in the data space. The classical approach measures variability through the variance, and the corresponding directions are the eigenvectors of the sample covariance matrix, which is well-known to be very sensitive to outliers.

The simplest and probably oldest approach to robust PCA consists of replacing the covariance matrix by a robust scatter matrix, e.g., Devlin et al. (1981),

*Ricardo Maronna (E-mail: rmaronna@mail.retina.ar) is professor, Faculty of Exact Sciences, University of La Plata, C.C. 172, La Plata 1900, Argentina; and researcher at C.I.C.P.B.A. Víctor Yohai (E-mail: vyohai@uolsinectis.com.ar) is professor, Faculty of Exact Sciences, University of Buenos Aires, and researcher at CONICET. This research was partially supported by grants PIP 5505 from CONICET and PICT 21407 and PAV 120 from ANPCyT, Argentina.

Campbell (1980), Naga and Antille (1990) and Croux and Haesbroeck (2000). Locantore et al. (1999) proposed a very simple and fast procedure based on an orthogonal-equivariant matrix.

Another approach consists of finding directions that maximize or minimize a measure of variability which is not the variance but a robust dispersion estimate. Such approach is generically called "projection pursuit"; see Li and Chen (1985), Croux and Ruiz-Gazen (1996 and 2005) and Maronna (2005). Hubert et al. (2005) propose a method for high dimensional data which combines projection pursuit and covariance matrix estimation.

All these approaches are resistant when the proportion of "atypical rows" is sufficiently small (in any event, smaller enough than 0.5). There are however situations where it is conceivable that several of the np data values may be contaminated. Situations of this type, with very large p, are computer vision, where each row represents an image and each column a pixel, and the analysis of microarray data where each row represents a case and each column a gene. Assume that each element is altered at random with probability ε. Then the probability π that a row contains at least a "bad" element is 1-$(1-\varepsilon)^p$, and this can be very large for large p even if ε is small. For example, with $p = 100$ and $\varepsilon = 0.01$ we have ≈ 0.63, so that the majority of rows would contain some atypical values. A model for this type of contamination has been discussed by Van Aelst *et. al.* (2006)

To cope with this situation, recall that computing the first q principal components is equivalent to finding a matrix $\widehat{\mathbf{X}}$ of rank q that minimizes $\left\|\mathbf{X}-\widehat{\mathbf{X}}\right\|_2$, where for a matrix $\mathbf{R} = [r_{ij}]$ we define the L_k norm as

$$\|\mathbf{R}\|_k = \left(\sum_{i=1}^{n}\sum_{j=1}^{p}|r_{ij}|^k\right)^{1/k}.$$

One known way to perform this minimization is the Eckart-Young algorithm (Gabriel and Zamir, 1979) which performs alternate least squares regressions.

A more robust estimate can be obtained by using the L_1 norm instead of the L_2 norm. Liu et al. (2003) apply this approach to microarray data.. It will be seen that this approach, although resistant to moderate element-wise contamination, may fail when some rows are completely atypical ("row-wise contamination"). This feature may be attributed to the lack of robustness of the L_1 regression estimate towards "bad" leverage points.

Verboon and Heiser (1994) and De la Torre and Black (2001) deal with element-wise M estimates. In particular, the later propose estimates of the form

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\rho\left(\frac{r_{ij}}{\widehat{\sigma}_j}\right) = \min, \tag{1}$$

where $\widehat{\sigma}_j$ is a column scale estimate. It is easy to verify that these estimates can also be computed by means of an alternating regressions algorithm, where each step now consists of an M estimate. The approach (1) presents the same

difficulties as regression M estimates. If ρ is convex, the estimate suffers the same lack of robustness as the L_1 estimate. If ρ is bounded, a robust estimate is needed to start the iterative process and to compute the scales $\widehat{\sigma}_j$. De la Torre and Black use a bounded ρ and propose a gradient algorithm rather than alternating regressions for the numerical solution of (1) combined with a "deterministic annealing method" to avoid bad local minima.

Recently Rey (2007) has addressed the problem of element-wise contamination and proposed a procedure that can be described as alternating weighted ridge regressions. He uses the classical solution as starting point, and the weights are obtained from a *convex* ρ-function. Both features make the procedure non-robust.

Croux et al. (2003) proposed a method which they call RAR, which may be considered as a weighted L_1 estimate:

$$\sum_{i=1}^{n}\sum_{j=1}^{p} w_{1i}w_{2j}\left|r_{ij}\right| = \min, \tag{2}$$

where the weights w_{1i} and w_{2j} depend on the data and are aimed at down-weighting outlying rows or columns. The algorithm is again based on alternating regressions, but using *weighted* L_1 regressions to avoid the effects of "bad" leverage points.

In this article we aim at defining estimates which are robust towards *both* element- and row-wise contamination, are fast to compute when p is large, and can cope with missing values. We shall deal with M estimates similar to (1) with a bounded ρ, following the approach of MM estimates (Yohai, 1987). A fast and robust initial estimate is first defined, which is needed to estimate the scales $\widehat{\sigma}_j$ and to start the iterative alternating regressions procedure. This initial estimate is composed of q steps of rank one, and hence its computing time is $O\left(npq\right)$. We define a criterion to choose q.

The main procedure is described in Section 2. In Section 3 we discuss a drawback of the proposed estimates, which consists of wrongly fitting a small proportion of cells. This drawback is corrected by a modified version of the MM estimate based on slight data perturbations.

Section 4 reports the results of a simulation study. In Section 5 the performances of different estimates are compared on three real data sets. In Section 6 we compare the computing times of estimates. Finally Section 7 contains the general conclusions of the article.

2 MM estimates for PCA

We shall consider approximating an $n \times p$ data matrix $\mathbf{X}$ by a matrix of given rank q plus a location vector, that is by $\widehat{\mathbf{X}} = \mathbf{T} + \mathbf{1}_n\mu'$, where $\mathbf{T}$ is a matrix of rank q, $\mathbf{1}_n$ is a column vector of n ones and $\mu = \left(\mu_1, ..., \mu_p\right)' \in R^p$ where in general $\mathbf{B}'$ denotes the transpose of $\mathbf{B}$. It will be useful to write $\mathbf{T}$ as

$$\mathbf{T} = \mathbf{A}\mathbf{B}' \tag{3}$$

where

$$\mathbf{A} = \left[\mathbf{a}^{(1)}...\mathbf{a}^{(q)}\right] = \begin{bmatrix} \mathbf{a}_1' \\ ... \\ \mathbf{a}_n' \end{bmatrix} \text{ and } \mathbf{B} = \left[\mathbf{b}^{(1)}...\mathbf{b}^{(q)}\right] = \begin{bmatrix} \mathbf{b}_1' \\ ... \\ \mathbf{b}_p' \end{bmatrix} \tag{4}$$

are full rank $n \times q-$ and $p \times q-$matrices. Here $\mathbf{a}^{(j)}$ $(j = 1, ..., q)$ and $\mathbf{a}_i$ $(i = 1, ..., n)$ denote respectively the columns and the rows of $\mathbf{A}$. Since our target is $\mathbf{T}$, we are not concerned with the identification of $\mathbf{A}$ and $\mathbf{B}$. It should be noted that if $\mathbf{B}$ is constrained to be orthonormal, then $\mathbf{a}_i$, $i = 1, .., n$ is a representation of the data in R^q.

We shall use the approach of MM estimates which was proposed by Yohai (1987) to obtain linear regression estimates with prescribed efficiency and high breakdown point. In this approach we have a robust but possibly inefficient initial estimate; the residuals thereof are used to compute a robust error scale; and the final estimate is an M estimate computed iteratively using the initial one as starting point. We shall describe the implementation of this idea for PCA, dealing first with the iterative algorithm, then with the scale, and finally with the initial estimate.

2.1 The "alternating" algorithm

Let ρ be a "bounded ρ-function" in the sense of (Maronna et al., 2006); that is, $\rho(r)$ is a nondecreasing function of $|r|$ with $\rho(0) = 0$ and $\rho(\infty) = 1$, and $\rho(r)$ is increasing for $r > 0$ such that $\rho(r) < 1$. Let $\widehat{\sigma}_j$ $(j = 1, .., p)$ be estimates of column variability. Put

$$L(\mathbf{A}, \mathbf{B}, \mu) = \sum_{j=1}^{p} \widehat{\sigma}_j^2 \sum_{i=1}^{n} \rho\left(\frac{r_{ij}}{\widehat{\sigma}_j}\right), \tag{5}$$

where r_{ij} are the residuals:

$$r_{ij} = r_{ij}(\mathbf{A}, \mathbf{B}, \mu) = x_{ij} - \widehat{x}_{ij} \quad \text{with} \quad \widehat{x}_{ij} = \mathbf{a}_i'\mathbf{b}_j + \mu_j,$$

where $\mathbf{a}_i$ denotes the i-th row of $\mathbf{A}$ and $\mathbf{b}_j$ denotes the j-th row of $\mathbf{B}$. Our estimates are defined as

$$\left(\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mu}\right) = \arg\min_{\mathbf{A},\mathbf{B},\mu} L(\mathbf{A}, \mathbf{B}, \mu). \tag{6}$$

Note that if $\rho(r) = r^2$ then we have standard PCA. This does not happen with (1) unless the $\widehat{\sigma}_j$s are equal.

It is not difficult to verify that local minima of L verify the M-estimating equations

$$\sum_{i=1}^{n} \psi\left(\frac{r_{ij}}{\widehat{\sigma}_j}\right) \mathbf{a}_i = \mathbf{0}, \;\; j = 1, ..., p \tag{7}$$

$$\sum_{j=1}^{p} \widehat{\sigma}_j \psi\left(\frac{r_{ij}}{\widehat{\sigma}_j}\right) \mathbf{b}_j = \mathbf{0}, \; i = 1, ..., n \tag{8}$$

$$\sum_{j=1}^{p} \widehat{\sigma}_j \sum_{i=1}^{n} \psi\left(\frac{r_{ij}}{\widehat{\sigma}_j}\right) = 0. \tag{9}$$

with $\psi = \rho'$, and hence a solution can be computed through alternating M-regressions.

It is well known that these regressions can be expressed as weighted least squares problems. Put $W(t) = \psi(t)/t$ and $w_{ij} = W(r_{ij}/\widehat{\sigma}_j)$. Then (7)-(8)-(9) may be written as

$$\sum_{i=1}^{n} w_{ij} \left(x_{ij} - \mathbf{a}_i' \mathbf{b}_j - \mu_j\right) \mathbf{a}_i = \mathbf{0}, \; j = 1, .., p \tag{10}$$

$$\sum_{j=1}^{p} w_{ij} \left(x_{ij} - \mathbf{a}_i' \mathbf{b}_j - \mu_j\right) \mathbf{b}_j = \mathbf{0}, \; i = 1, ..., n \tag{11}$$

$$\sum_{j=1}^{p} \sum_{i=1}^{n} w_{ij} \left(x_{ij} - \mathbf{a}_i' \mathbf{b}_j - \mu_j\right) = 0. \tag{12}$$

This representation naturally suggests an iterative algorithm. Put

$$w_{ij}(\mathbf{A}, \mathbf{B}, \mu) = W\left(\frac{r_{ij}}{\widehat{\sigma}_j}\right).$$

Given initial $(\mathbf{A}_0, \mathbf{B}_0, \mu_0)$ at iteration k define $\mathbf{A}_k$ as the matrix with rows $\mathbf{a}_{k,1}, ..., \mathbf{a}_{k,n}$ and $\mathbf{B}_k$ as the matrix with rows $\mathbf{b}_{k,1}, ..., \mathbf{b}_{k,p}$ where $\mathbf{a}_{k,i}$ is the solution $\mathbf{a}$ of

$$\sum_{j=1}^{p} w_{ij} \left(\mathbf{A}_{k-1}, \mathbf{B}_{k-1}, \mu_{k-1}\right) \left(x_{ij} - \mathbf{a}' \mathbf{b}_{k-1,j} - \mu_{k-1}\right) \mathbf{b}_{k-1,j} = \mathbf{0}, \tag{13}$$

$\mathbf{b}_{k,j}$ is the solution $\mathbf{b}$ of

$$\sum_{i=1}^{n} w_{ij} \left(\mathbf{A}_k, \mathbf{B}_{k-1}, \mu_{k-1}\right) \left(x_{ij} - \mathbf{a}_{k,i}' \mathbf{b} - \mu_{k-1}\right) \mathbf{a}_{k,i} = \mathbf{0} \tag{14}$$

and $\mu_{k,j}$ is the solution μ of

$$\sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij} \left(\mathbf{A}_k, \mathbf{B}_k, \mu_{k-1}\right) \left(x_{ij} - \mathbf{a}_{k,i}' \mathbf{b}_{k,j} - \mu\right) = 0. \tag{15}$$

It follows from Section 9.1 of (Maronna et al., 2006) that if $W(x)$ is a nonincreasing function of $|x|$ then the loss function L in (5) decreases at each iteration. We thus get a simple procedure which converges to a local minimum. Actually the procedure stops when either the relative decrease in the loss function (5) is less than a prescribed tolerance value or the number of iterations exceeds a given limit. In our implementation the tolerance is 0.001 and the maximum number of iterations is 20.

We shall throughout use the bisquare ρ defined as $\rho(r) = \rho_1(r/c)$ where

$$\rho_1(r) = \min\left\{1, 1 - \left(1 - r^2\right)^3\right\}$$

and c is chosen so as to attain a given efficiency at the normal. We choose $c = 3.44$ which corresponds to 85% efficiency. The respective weight function is

$$W(r) = \left(1 - \left(\frac{r}{c}\right)^2\right)^2 \mathrm{I}(|r| \leq c), \tag{16}$$

where $\mathrm{I}(.)$ denotes the indicator function.

It is worth noting that, unlike in ordinary regression, the objective function (5) has at least $\min(n, p)$ local minima even in the classical case when $\rho(r) = r^2$.

2.2 The scales

Given the initial estimate, the scales are computed as M-scales of the initial residuals $r_{ij}(\mathbf{A}_0, \mathbf{B}_0, \mu_0)$. Recall that an M estimator of scale of a sample $\mathbf{r} = (r_1, .., r_n)$ (an M-scale for short) is the solution $\sigma = \widetilde{\sigma}(\mathbf{r})$ of

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{\sigma}\right) = \delta \tag{17}$$

with $\delta \in (0, 1)$. We choose

$$\delta = \frac{np - [q(n+p) + p]}{2np}. \tag{18}$$

Call c_0 the solution of $\mathrm{E}\rho(z/c) = 0.5$, with $z \sim \mathrm{N}(0, 1)$, which is $c_0 = 1.56$. Then $\widehat{\sigma}(\mathbf{r}) = \widetilde{\sigma}(\mathbf{r})/c_0$ is the M-scale normalized to be consistent at the normal. We define $\widehat{\sigma}_j = \widehat{\sigma}(r_{1j}, ..., r_{nj})$. Note that $q(n+p) + p$ is the total number of parameters. The reason for defining δ as in(18) is to take into account the "residual degrees of freedom". Using $\delta = 0.5$ instead of (18) would lead to the underestimation of σ.

The resulting estimate will be called an *MM estimate for PCA*.

2.3 The initial values

To compute our estimates we now need the initial $(\mathbf{A}_0, \mathbf{B}_0, \mu_0)$. Since the objective function (5) is not convex, the initial values have to be robust, in order to

avoid falling into "bad" local minima. De la Torre and Black (2001, Appendix) employ a procedure based on *classical* principal components to derive starting values, which may lead to wrong results if there are large outliers.

Our approach is to perform successive fits of rank one. At step 1 put $\mathbf{R}_1 = \mathbf{X}$ and compute a rank-one fit $\left(\mathbf{a}^{(1)}, \mathbf{b}^{(1)}, \mu^{(1)}\right)$ to $\mathbf{R}_1$, i.e., $\widehat{\mathbf{R}}_1 = \mathbf{a}^{(1)}\mathbf{b}^{(1)\prime} + \mathbf{1}_n\mu^{(1)\prime}$. At step $m = 2, ..., q$ let $\mathbf{R}_m = \mathbf{R}_{m-1} - \widehat{\mathbf{R}}_{m-1}$ and fit $\left(\mathbf{a}^{(m)}, \mathbf{b}^{(m)}, \mu^{(m)}\right)$ to $\mathbf{R}_m$. Then $\mathbf{A}_0$ and $\mathbf{B}_0$ are obtained as in (4) and

$$\mu_0 = \sum_{m=1}^{q} \mu^{(m)}.$$

The problem is thus reduced to finding initial estimates for fits of *rank one*, which is a simpler although not trivial problem. The proposed procedure is described in the next Section.

2.4 Initial values for rank-one fits

In order to define an initial fit of rank one $\{\mathbf{a}, \mathbf{b}, \mu\}$ for $\mathbf{X}$, let $\mu_j = \text{median}_i \{x_{ij}\}$ and

$$y_{ij} = x_{ij} - \mu_j. \tag{19}$$

We now need a fit $\mathbf{Y} \approx \mathbf{a}\mathbf{b}'$. This will be performed through a sequence of alternate regressions using the simplest robust estimate of regression through the origin, namely the median of slopes. Let $\mathbf{a}_0$ be an initial $n-$vector. Given $\mathbf{a}_m$, define $\mathbf{a}_{m+1}$ and $\mathbf{b}_{m+1}$ by

$$b_{m+1,j} = \text{median}_i \left\{ \frac{y_{ij}}{a_{m,j}} \right\},\ j = 1, .., p$$

and

$$a_{m+1,i} = \text{median}_j \left\{ \frac{y_{ij}}{b_{m+1,j}} \right\},\ i = 1, ..., n.$$

The procedure does not seem to converge. Call N_{it} the number of iterations allowed. For each $m = 1, ..., N_{\text{it}}$ we have a fit $\widehat{\mathbf{Y}}_m = \mathbf{a}_m\mathbf{b}'_m$. We shall choose the "best" fit in the sense that $\mathbf{Y} - \widehat{\mathbf{Y}}_m$ is "smallest", as will be explained below.

We still have to choose the initial $\mathbf{a}_0$. We use each of the columns of $\mathbf{X}$ as $\mathbf{a}_0$. Since trying them all would make the computing cost $O\left(np^2\right)$, for a given N_{col} we use all columns if $p \leq N_{\text{col}}$ and take N_{col} columns at random otherwise.

To choose the "best" among the $N_{\text{it}} \times N_{\text{col}}$ fits we use a τ-scale. Let for a vector $\mathbf{r}$

$$S(\mathbf{r}) = s^2 \text{ave}_i \left\{ \rho\left(\frac{r_i}{s}\right) \right\}$$

where ρ is as in (5) and $s = \text{median}_i \{|r_i|\} / 0.675$. We finally choose the fit minimizing $S\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right)$.

The reason for choosing $S(\mathbf{r})$ rather than the simpler s was that exploratory simulations showed that a more efficient criterion yielded better results.

In our simulations and experiments we used $N_{\text{col}} = 20$ and $N_{\text{it}} = 1$. Increasing these values did not noticeably improve the estimate's performance.

It follows that for "large" p (say $p > 20$) the initial rank-one fits involve a random choice for reasons of time economy, which implies that some differences may be found in the results of different calls to the procedure.

An alternative approach for rank-one fits would be to use a matrix $\mathbf{\Sigma}$ of robust pairwise covariances, such as the Gnanadesikan-Kettenring covariances employed by Maronna and Zamar (2002). Compute the first principal direction $\mathbf{b}$ of $\mathbf{\Sigma}$. Compute $\mathbf{a}_i$ as the robust regression of $(y_{i1}, ..., y_{ip})$ on $\mathbf{b}$, with y_{ij} as in (19). This is a simple fit (e.g., median of slopes). Note that $\mathbf{\Sigma}$ may fail to be positive semidefinite; but hopefully the first principal direction is reliable, although not the higher order ones. Computing $\mathbf{\Sigma}$ is $O\left(np^2\right)$, which may be rather heavy for large p. Simulations showed the results of this approach to be no better than using the alternating regressions described above.

2.5 Choosing the rank

There remains the problem of choosing the rank q. The usual way to do this in classical PCA is by computing the proportion of unexplained variance. A robust analogue of the unexplained variance for a fit of rank q with residuals $r_{ij}^{(q)}$ is

$$v_q = \sum_{j=1}^{p} \widehat{\sigma}_j^2 \sum_{i=1}^{n} \rho\left(\frac{r_{ij}^{(q)}}{\widehat{\sigma}_j}\right). \tag{20}$$

Note that v_q is a robust residual scale (it is actually a "$\tau-$scale" in the sense of Yohai and Zamar (1988)). Define the residuals for the fit of rank 0 as $r_{ij}^{(0)} = x_{ij} - \widehat{\mu}_{0,j}$. Then the "proportion of unexplained variance" is defined as

$$u_q = \frac{v_q}{v_0}. \tag{21}$$

When $\rho(r) = \rho^2$ this coincides with the usual criterion.

The fact that the initial estimate is based on a sequence of rank-one fits avoids the need for the user to choose q in advance: rather, the sequence may be stopped when u_q is smaller than a prescribed threshold such as e.g. 0.05. Besides, once a fit of rank q is computed, the starting values for all fits of lower rank are available too. These two features may save much time if n and p are very large.

2.6 Some details

1. Location: The simplest approach to estimate location would be to first center the columns and then apply the former procedure to the centered data, omitting μ. But our simulations showed that including μ in the iterative procedure improves the performance with only a small increase in time.

2. Ill-conditioned rows: It may happen that at some iteration the regressions (13) cannot be carried out for some i, due to the ill conditioning of the predictor matrix. This happens especially when the w_{ij}s are very small, due to row i being completely atypical. In this case $\mathbf{a}_i$ is computed as an *ordinary* regression MM estimate, in the following sense. We use the current $\mathbf{a}_i$ —i.e., the i-th row $\mathbf{a}_{k-1,i}$— of $\mathbf{A}_{k-1}$ as starting estimate. We compute an M-scale of the respective residuals: $s_i = \widehat{\sigma}\left(r_{i1}, ..., r_{ip}\right)$ with

$$r_{ij} = x_{ij} - \mathbf{a}'_{k-1,i}\mathbf{b}_{k-1,j} - \mu_{k-1,j},$$

and define

$$\mathbf{a}_{k,i} = \arg\min_{\mathbf{a}} \sum_{j=1}^{p} \rho\left(\frac{x_{ij} - \mathbf{a}'\mathbf{b}_{k-1,j} - \mu_{k-1.j}}{s_i}\right). \tag{22}$$

3. Outlying rows: The former criterion is also adopted when "too many" elements of the row are potential outliers. More precisely, when more than half of $W\left(r_{ij}/\widehat{\sigma}_j\right)$ $(j = 1, ..., p)$ are less than 0.001 we use (22) even when the predictor matrix is well conditioned. This detail greatly improves on the resistance of the estimate towards row-wise contamination.

It may seem reasonable just to give zero weight to the whole row; but simulations have shown that this step does not necessarily improve the estimator's performance, and may worsen it the situations that will be considered in Section 3.

2.7 Missing values

Missing values appear frequently in large datasets, e.g. in microarray data, and are difficult to cope with for most robust procedures. Deleting all rows containing some missing values may result in too high a loss of data. But the proposed procedure can be adapted to the case of missing values, in case that these values are scattered over the data, i.e., they constitute only a small proportion in any row or column . Simply replace (5) with

$$L\left(\mathbf{A}, \mathbf{B}, \mu\right) = \sum_{j=1}^{p} \widehat{\sigma}_j^2 \sum_{i=1}^{n} M_{ij}\rho\left(\frac{r_{ij}}{\widehat{\sigma}_j}\right), \tag{23}$$

where M_{ij} is the indicator that x_{ij} is not missing. Each step of the procedure involves a linear fit to a single row or column, at which the missing cells are simply disregarded.

Recall that our procedure uses as starting values the outcome of successive rank-one fits. Here missing values produce no problems. But when running the full rank-q alternating regressions a problem may appear, namely that if a row contains more than $p-q$ missing values the regressions cannot be computed. In this case we replace the missing values with those from the initial fit.

A similar modification can be applied to any procedure based on alternating regressions, like L_1 or RAR. Recently Serneels and Verdonck (2007) proposed an approach to adapt any robust PCA estimate for missing values.

3 False outliers and perturbed estimates

To assess the performance of our MM estimate with "clean" data, we performed a fit of rank $q = 4$ to a simulated normal sample with $n = 50$ and $p = 10$, generated as $\mathbf{X} = \mathbf{AB}' + 0.5\mathbf{E}$ where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{E}$ have i.i.d. standard normal elements and dimensions $n \times q$, $p \times q$ and $n \times p$ respectively. We computed the residuals $r_{ij}^{(LS)}$ and $r_{ij}^{(MM)}$ of the classical estimate (which will henceforth be called "LS" for short) and of the MM estimate, respectively. Figure 1 shows the QQ-plot of the ordered absolute values of the latter vs. those of the former.

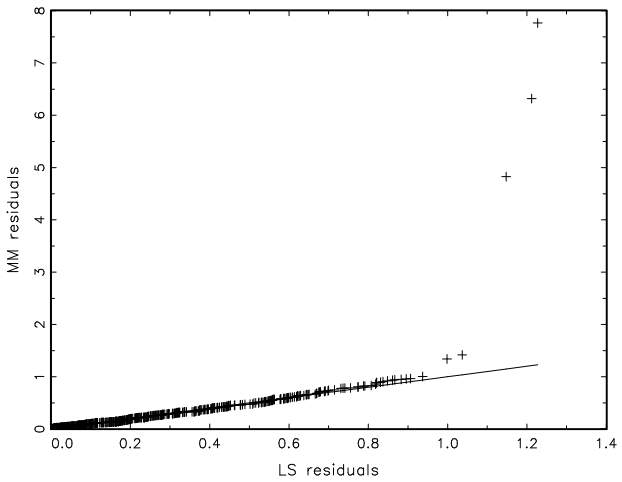

Figure 1: Ordered absolute residuals of the MM estimate vs. those of the LS estimate.

We see that most of the points are near the identity line, but the MM estimate has three very large residuals. That is, there are cells where MM fits clearly worse than LS, and which would appear as outliers. The same pattern occurs very frequently if we repeat the simulation.

It would seem that the problem is unimportant, since it involves only 3/500=0.6% of the cells. But consider now a real set of ionospheric data, which will be analyzed in more detail in Section 5. Each of the $n = 225$ rows of the dataset is a set of responses measured at $p = 31$ pulse numbers.

The upper left-hand panel of Figure 2 shows the LS fit of row 175 for rank $q = 4$; that is, x_{ij} and $\widehat{x}_{ij}$ for $i = 175$ vs. the pulse numbers $j = 1, ..., 31$. This is a typical row, and we see that the fit is almost perfect. The right-hand panel shows the MM fit. There are six clearly wrong values which make the fit unsatisfactory. The bottom row of the figure shows the same for row 102. This same failure pattern occurs in 31 rows affecting 177 cells.. The proportion of cells affected is very low: 2.5%, but affects 13% of the rows, making the MM fit quite unreliable.

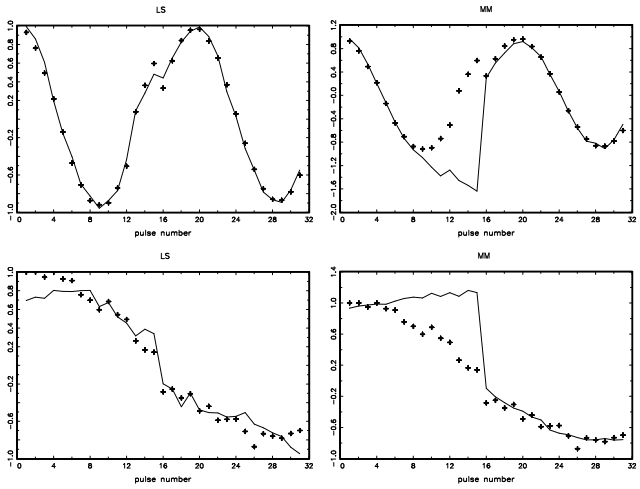

Figure 2: Ionosperic data: Observations (+++) and fits (—) by the LS and MM estimates (left-hand and right-hand columns) for data rows 175 (upper plot row) and 102 (lower row).

We have not been able to completely understand the reason for these misfits. The fact that it also occurs with the RAR estimate, as will be demonstrated in Section 5, makes the problem still more puzzling.

We are thus led to an approach that can avoid false outliers, without a high loss in robustness, and without a high increase in computational cost. After trying and discarding simpler approaches, we propose one based on perturbing the data. The intuitive idea behind this approach is that false outliers are just artifacts, and should hence be unstable if the data are perturbed, while real outliers should be stable under perturbations. Assume that we have already computed the MM estimate $\left\{\widehat{\mathbf{A}}_1,\widehat{\mathbf{B}}_1,\widehat{\mu}_1\right\}$ with residuals $r_{ij}^{(1)}$ and weights $w_{ij}^{(1)} = W\left(r_{ij}^{(1)}/\widehat{\sigma}_j\right)$. Let γ be a "small" constant and $m > 1$. Generate the perturbed data

$$x_{ij}^{(\mathrm{per})} = x_{ij} + \gamma\widehat{\sigma}_j z_{ij} \tag{24}$$

where z_{ij} are i.i.d. $\mathrm{N}(0,1)$. Let $\left\{\widehat{\mathbf{A}}_2,\widehat{\mathbf{B}}_2,\widehat{\mu}_2\right\}$ be the estimate computed on the perturbed data, yielding weights $w_{ij}^{(2)}$. In general we generate $m-1$ perturbed datasets with yield weights $w_{ij}^{(k)}$, $k = 2,..,m$. Let

$$w_{ij}^{(\mathrm{med})} = \mathrm{median}\left\{w_{ij}^{(1)},...,w_{ij}^{(m)}\right\}, \tag{25}$$

and define finally

$$w_{ij}^{(\mathrm{per})} = \mathrm{I}\left(w_{ij}^{(\mathrm{med})} > 0\right). \tag{26}$$

The "perturbed estimate" which will be called PertMM(m, γ) is defined as the weighted LS estimate with weights $w_{ij}^{(\mathrm{per})}$. Note that this amounts to using a simple rejection rule decided by a "majority vote". The choice of $w_{ij}^{(\mathrm{per})}$ over $w_{ij}^{(\mathrm{med})}$ was made after simulations showed that this choice improved the behavior of the estimate towards "false outliers" without a loss of robustness.

Note that a large γ increases the chances of getting rid of the false outliers, but also decreases the robustness. Also, a large m increases the method's stability but also the computing time. The current approach was chosen after trying and discarding others through a reduced simulation study, which led us to choose $\gamma = 0.5$ with $m = 5$. Applying this estimate to the data of Figure 1 yields no outliers, and the fit given by the estimate to the cases of 2 is similar to the LS fit. The merits of this approach will be assessed more thoroughly in Sections 4 and 5.

We remark that the perturbed estimates involve a random element which may cause differences in the results of different calls to the procedure.

The weighted LS fit required by the perturbed estimate is computed by a sequence of alternating weighted LS regressions. This procedure needs starting values, and if the data contain outliers its outcome may depend on these values. We use as starting values for the weighted LS estimate the values $\left\{\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1, \widehat{\mu}_1\right\}$ given by the original MM estimate. Overlooking this detail may lead to a loss of robustness.

Our experiments show that the phenomenon of false outliers appears only with "good" data. The difference between MM and PertMM would be very difficult to measure through a simulation, because the proportion of misfitted cells is very low and the corresponding residuals, although large for the eye, are not so large that less than 1% of them can alter a performance measure like the mean squared error.

4 A simulation study

A simulation study was run to compare the different estimates. The underlying idea is that we want to reconstruct the data matrix subject to contamination. The "clean" observations were generated as in model (3)-(4), i.e.

$$\mathbf{X} = \mathbf{T} + \mathbf{E} \tag{27}$$

where $\mathbf{T} = \mathbf{A}\mathbf{B}' = [t_{ij}]$ is a matrix of rank q, and $\mathbf{E}$ is "pure noise" with i.i.d. elements $e_{ij} \sim \mathrm{N}\left(0, \sigma_{\mathrm{noise}}^2\right)$. The elements a_{ij} of $\mathbf{A}$ are i.i.d. $\mathrm{N}(0,1)$, and $\mathbf{B}$ is fixed. We chose $\mathbf{B}$ to induce "high leverage rows". Let first $b_{jk} = j^k$ for $j = 1, ..., p$ and $k = 1, ..., q$. The columns of $\mathbf{B}$ are then normalized to have zero means and unit Euclidean norms. The reason for the unit norms is to ease the interpretation of $\sigma_{\mathrm{noise}}^2$, which becomes the "noise to signal ratio": $\sigma_{\mathrm{noise}}^2 = \mathrm{var}(e_{ij}) / \mathrm{ave}_{ij}\{\mathrm{var}(x_{ij})\}$. The reason for the centering will be explained below. We used $q = 1, 2, 3$. We show the results for $q = 2$, which is representative of the other cases. We used $p = 10$ and $n = 50$ throughout.

We considered the following types of contamination:

Element-wise contamination (centered) Given $\varepsilon \in (0,1)$ and K, cells are chosen independently at random with probability ε and their contents replaced with $\mathrm{N}\left(0, K^2\right)$.

Element-wise contamination (shifted) Same as above, but replacing the contents with K

Row-wise contamination The contents of the first $n\varepsilon$ rows are replaced with $pn\varepsilon$ i.i.d. values with distribution $\mathrm{N}\left(K, \alpha^2\right)$ where $\alpha = 0.1$. Since the columns of $\mathbf{B}$ have zero means, these rows become almost orthogonal to $\mathbf{B}$, which makes this contamination most difficult to deal with for the estimates.

In all cases, the values of K ranged between σ_{noise} and 20.

Since the results corresponding to shifted element-wise contamination yield the same conclusions as those from centered element-wise contamination, the former will henceforth be omitted.

To evaluate the results, let $\widehat{x}_{ij}$ be the fitted values for a given estimate. We want to compare $\widehat{\mathbf{X}}$ to $\mathbf{X}$. The MSE is computed only for the non-contaminated cells:

$$\mathrm{MSE} = \mathrm{ave}_{ij}\left\{\left(\widehat{x}_{ij} - x_{ij}\right)^2 : i \in I\right\} \tag{28}$$

where I is the set of non-contaminated cells.

The data are generated N_{rep} times with $N_{\text{rep}} = 500$. In a given situation we thus have for each estimate N_{rep} values of the MSE. The simplest way to summarize them is to take their average. This criterion however turned out to be unreliable, because in some situations, an estimate with a generally good behavior had a few very large failures (only 1%) which deformed the criterion. For this reason we preferred as a criterion a trimmed mean, namely the average of the 90% smallest MSEs. We report both trimmed and overall means of MSEs for the sake of completeness.

The estimates used were the following.

1. LS: the classical estimate
2. L_1: the L_1 estimate
3. RAR: the RAR estimate as described by Croux et al. (2003). Since the criterion (2) does not always decrease at each iteration, we chose to stop either after no noticeable changes occurred, or a pre-specified number of iterations (20) was attained, and then chose the solution with minimum (2).
4. SPC: the "spherical principal components" of Locantore et al. (1999).
5. S-M: the estimate proposed by Maronna (2005) that minimizes an M-scale of the distances. To make the estimate faster, SPC was used as a starting value rather than the random procedure used in Maronna's paper.

6. MM: the MM estimate (6) with bisquare ρ-function and efficiency 0.85.

7. PertMM: The perturbed MM estimate described in Section 3, PertMM(m, γ) with $\gamma = 0.5$ and $m = 5$.

Since SPC and S-M gave very similar results, only the latter will be henceforth reported.

4.1 Complete data

Table 1 displays the maximum (over outlier size K) trimmed means of MSEs of estimates.

		Contamination type						
		No	Element			Row		
σ_{noise}	ε	0	0.05	0.1	0.2	0.05	0.1	0.2
0.2	LS	0.030	1.690	4.047	10.310	0.412	0.413	0.402
	MM	0.032	0.036	0.039	0.165	0.033	0.044	0.374
	PertMM	0.031	0.036	0.039	0.146	0.032	0.044	0.461
	L_1	0.040	0.049	0.129	0.876	0.425	0.556	0.544
	RAR	0.045	0.053	0.091	2.687	0.049	0.094	0.546
	S-M	0.032	2.762	5.777	11.452	0.033	0.036	0.179
0.5	LS	0.187	1.869	4.221	10.480	0.573	0.567	0.557
	MM	0.202	0.216	0.212	0.378	0.205	0.232	0.762
	PertMM	0.195	0.216	0.225	0.483	0.201	0.241	0.700
	L_1	0.227	0.263	0.355	1.253	0.669	0.713	0.712
	RAR	0.248	0.279	0.371	1.993	0.260	0.346	0.765
	S-M	0.196	2.988	6.009	11.534	0.197	0.209	0.530

Table 1: Simulation: Maximum over K of 0.9–trimmed means of MSEs

We see that

- MM, PertMM and S-M are rather efficient as compared to LS for $\varepsilon = 0$
- For element-wise contamination, MM and PertMM are clearly the best; S-M fails completely (as expected), RAR performs clearly better than S-M better but much worse than MM and PertMM, and L_1 performs not too badly for $\varepsilon = 0.05$ but poorly for $\varepsilon = 0.1$ and 0.2. The performance of RAR deteriorates sharply for $\varepsilon = 0.2$.
- For row-wise contamination, S-M is the best (as could be expected); MM and PertMM are next to best; L_1 performs worse than LS, and RAR is better than L_1 but still inferior to MM and PertMM. When $\varepsilon = 0.2$ all estimates are at least as bad as LS, except for S-M when $\sigma_{\text{noise}} = 0.2$.

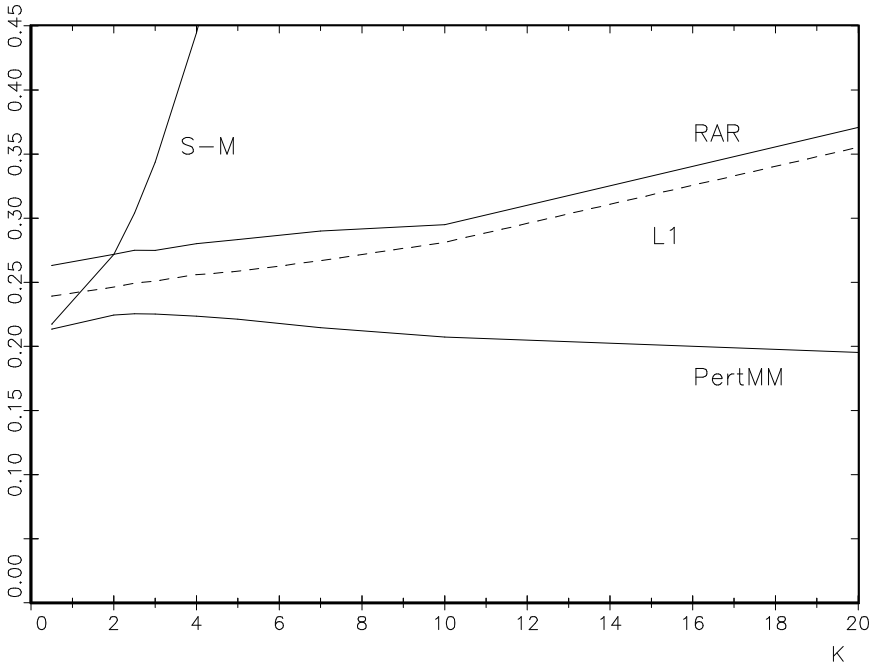

Figure 3: Simulation results: trimmed means of MSEs of estimates as a function of outlier size K for $\sigma_{\text{noise}} = 0.5$ and 10% element-wise contamination.

- The behavior of PertMM is similar to that of MM, showing that the perturbation procedure does not affect its robustness.

Figures 3 and 4 plot the trimmed MSEs of PertMM and RAR as a function of the outlier size K for 10% element- and row-wise contamination, respectively, and $\sigma_{\text{noise}} = 0.5$. The value at $K = 0$ corresponds to $\varepsilon = 0$.

It should be remarked that these results do not imply that the competing estimates are useless. The competing estimates may yield reliable results under smaller ε and K and choices of $\mathbf{B}$ with less leverage.

For completeness, we display in Table 2 the results corresponding to the overall (without trimming) averages of MSEs. The qualitative conclusions remain the same for $\varepsilon \leq 0.1$. For $\varepsilon = 0.2$ we see however that some values are almost twice the corresponding ones in the former table. Is worth noting that those differences are due to 1% of the simulation samples.

Note: The RAR estimate as described in (Croux et al., 2003) works with the variables standardized to unit MAD. Since in this simulation the estimates are evaluated through their MSEs computed on the raw variables, it might be argued that the criterion is unfair to RAR. For this reason a version of RAR without normalization was also included in the simulation; the results were not

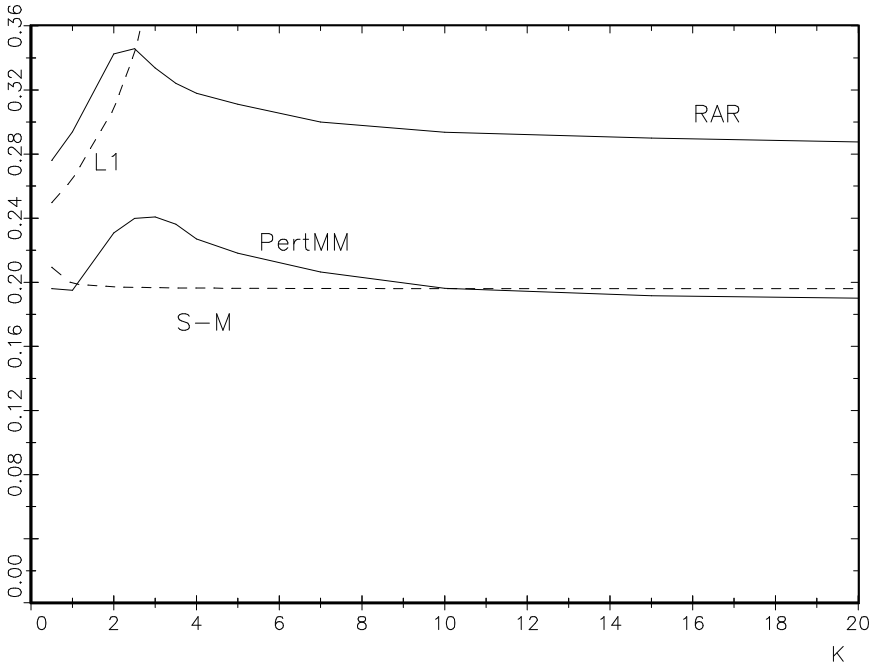

Figure 4: Simulation results: trimmed means of MSEs of estimates as a function of outlier size K for $\sigma_{\text{noise}} = 0.5$ and 10% row-wise contamination

very different from those reported here for the default version.

4.2 Missing data

The study was repeated with the same design but with 10% missing values located at random. Now I in (28) is the set of cells without contamination and without missing values. Here SPC and S-M cannot be used. As to L_1 and RAR, the alternating regressions procedure can also be applied, but not the starting estimates, since the procedure by Croux and Riz-Gazen uses linear combinations of cells. Therefore the starting values were obtained by means of alternating median of slopes regression.

Tables 3 and 4 give the trimmed and the overall averages of MSEs for $\sigma_{\text{noise}} = 0.5$. Comparison with Table 1 suggests that the missing values do not have a serious impact on the robust estimates, except for RAR with row-wise contamination.

		Contamination type						
		No	Element			Row		
σ_{noise}	ε	0	0.05	0.1	0.2	0.05	0.1	0.2
0.2	LS	0.030	1.913	4.510	10.980	0.425	0.429	0.418
	MM	0.033	0.036	0.046	0.300	0.034	0.048	0.709
	PertMM	0.031	0.034	0.041	0.266	0.032	0.047	0.834
	L_1	0.041	0.057	0.208	1.194	0.459	0.558	0.572
	RAR	0.048	0.069	0.383	6.433	0.051	0.101	0.573
	S-M	0.033	3.029	6.169	12.180	0.033	0.037	0.214
0.5	LS	0.190	2.091	4.687	11.153	0.582	0.588	0.574
	MM	0.206	0.209	0.220	0.771	0.210	0.241	1.362
	PertMM	0.198	0.211	0.233	0.817	0.204	0.249	1.149
	L_1	0.231	0.258	0.437	1.629	0.700	0.740	0.743
	RAR	0.253	0.284	0.618	2.807	0.266	0.357	0.798
	S-M	0.199	3.274	6.419	12.221	0.200	0.216	0.566

Table 2: Simulation: Maximum over K of overall means of MSEs

Contamination	No	Elem.	Row
ε	0	0.1	0.1
LS	0.181	3.120	0.606
MM	0.207	0.222	0.287
PertMM	0.193	0.210	0.286
L_1	0.257	0.382	0.746
RAR	0.284	0.394	0.511

Table 3: Simulation with 10% missing values: Maximum over K of 0.9–trimmed means of MSEs

5 Examples

In this Section the performances of the estimates on some real datasets are compared.

5.1 False outliers

We now compare the behavior of the estimates with respect to the phenomenon of "false outliers" discussed in Section 3 through some datasets where each row corresponds to a curve, which have the advantage that "true" and "false" outliers are easy to distinguish visually. We consider two datasets taken from the "Data Repository" in (Bay, 1999). The first one (called henceforth "Ionospheric") has been used by Sigillito *et al.* (1989) and also analyzed by Maronna and Zamar (2002) and Maronna (2005). It consists of $n = 225$ "good" radar measurements of the ionosphere on 34 continuous characteristics with values in $[-1, 1]$, each corresponding to a pulse number. Variables 1, 2 and 27 have MAD=0. Since this would prevent the use of RAR, they were omitted, so that here $p = 31$.The

Contamination	No	Elem.	Row
ε	0	0.1	0.1
LS	0.184	3.489	0.625
MM	0.214	0.279	0.374
PertMM	0.196	0.217	0.486
L_1	0.264	0.509	0.780
RAR	0.293	0.589	0.535

Table 4: Simulation with 10% missing values:: Maximum over K of overall means of MSEs

other dataset (henceforth "LRS") is part of the Low Resolution Spectrometer Database in the Infra-Red Astronomy Satellite Project, and contains $n = 531$ high quality spectra measured on $p = 93$ frequency bands. It has also been analyzed by Maronna and Zamar (2002). Table 5 gives the proportions of unexplained variabilities of the LS and MM estimates, as measured by the proportion of unexplained variance and by (21), respectively.

	Ionospheric		LRS	
q	LS	MM	LS	MM
1	0.463	0.422	0.489	0.342
2	0.243	0.154	0.092	0.093
3	0.131	0.100	0.054	0.061
4	0.081	0.073	0.029	0.031

Table 5: Proportions of unexplained variability for rank q for the Ionospheric and LRS datasets

We discuss first the Ionoshperic dataset. We choose rank $q = 4$ which according to Table 5 yields less than 10% unexplained variability for both estimates. As explained in Section 3, the MM estimate with $q = 4$ gives disastrous fits to 31 rows of this dataset, such as those of Figure 2. Table 6 shows the number of failures for all estimates. It is seen that also RAR gives an important number of wrong fits, while PertMM, L_1 and S-M show a reliable behavior.

Dataset	MM	PertMM	L_1	RAR	S-M
Ionospheric	31	0	0	34	2
LRS	35	4	0	22	0

Table 6: Number of failures of the estimates for two real datasets

We now consider the LRS data. We choose rank $q = 2$ which according to Table 5 yields less than 10% unexplained variability. Figure 5 shows the failure of MM with $q = 2$ in a typical row. It is seen that the LS fit is excellent, while that of MM goes completely astray. Table 6 gives the results for all estimates. Here MM and RAR show again their propensity to misfits, although the proportion of failed rows is smaller than with the Ionospheric dataset. PertMM, L_1 and S-M show a reliable behavior.

These results suggest that PertMM has a satisfactory behavior with respect to the phenomenon of false outliers.

In order to avoid visual examination of all rows to find potential failures of a given estimate, we limited the search to the rows in which the estimate's absolute residuals were much larger than those of LS for a large number of consecutive columns. More precisely, call r_{ij} the estimate's residuals; for row i call s_i the 0.75-quantile of $\left\{\left|r_{ij}^{LS}\right|, j = 1, ..., p\right\}$. Then the fit to row i was plotted if the number of consecutive js for which $|r_{ij}| > 2.5s_i$ was larger than $p/15$. This criterion included most situations like the ones of Figure 2. The final decision whether the row was actually considered a failure was based on the visual impression. Although the criterion is admittedly subjective, we think that the cases that we classified as failures leave no room for doubt.

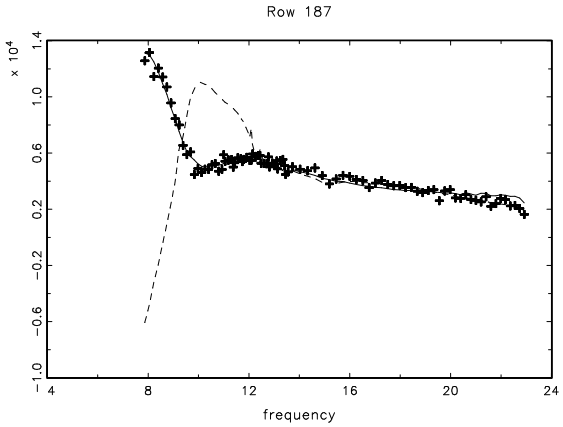

Figure 5: LRS dataset: data (++) and fits by LS (—) and MM (- - -) of row 187

5.2 Resistance to outliers

One difficulty that appears in the assessment of the behavior of estimates on real data is the lack of *a priori* knowledge of what a "good behavior" should be. For this reason we considered adding light element-wise contamination to the data and measure the quality of the fits to the original data. We first experimented

with the two former datasets. As could be expected, LS and S-M broke down under the contamination; but while the other estimates yielded different results, none of them could be said to outperform the others.

We then considered a dataset corresponding to electron-probe X ray microanalysis of archeological glass vessels (Janssens et al, 1988), where each of $n = 180$ vessels is represented by a spectrum on $p =$1920 frequencies. For our analysis we restricted ourselves to frequencies 15 to 500, since the responses corresponding to the others have either null or almost null MAD; therefore we actually have $p = 486$.

	Original data		Contaminated	
q	LS	MM	LS	MM
1	0.286	0.726	0.974	0.555
2	0.142	0.235	0.959	0.184
3	0.010	0.033	0.944	0.037

Table 7: Vessel dataset: Proportion of unexplained variability for rank q

The left-hand half ("Original data") of Table 7 gives the proportions of unexplained variability. We choose $q = 3$. Table 8 gives the α-quantiles of the absolute residuals corresponding to the different estimates. Since MM and PertMM give very similar results, the former are henceforth omitted.

α	LS	S-M	L_1	RAR	PertMM
0.2	1.5	1.5	1.4	1.3	1.4
0.5	4.9	5.2	4.9	4.3	4.5
0.7	9.4	10.1	9.3	8.5	8.5
0.9	25.0	29.0	24.0	33.4	25.1
0.95	39.8	49.1	39.2	78.5	47.1
0.97	54.8	69.9	55.2	136.6	71.3

Table 8: Vessel dataset: α-quantiles of absolute residuals

Table 8 indicates that LS and MM give similar fits to the original data. However, Table 7 suggests that LS gives a much better fit. The reason for this seeming contradiction is that each estimate uses a different measure for the fit. The upper panel of Figure 6 shows the data represented in the space of the first two (classical) principal components, and the lower panel shows the MM representation with rank $q = 2$, rotated to match the upper panel. Both figures are similar. The presence of clusters is due to the existence of four different types of vessels. The cluster on the right hand side of both plots contains 22%

of the data. MM uses a robust scale which reflects the dispersion of the main cluster on the left, while LS uses the variance which is inflated by the presence of the right-hand cluster.

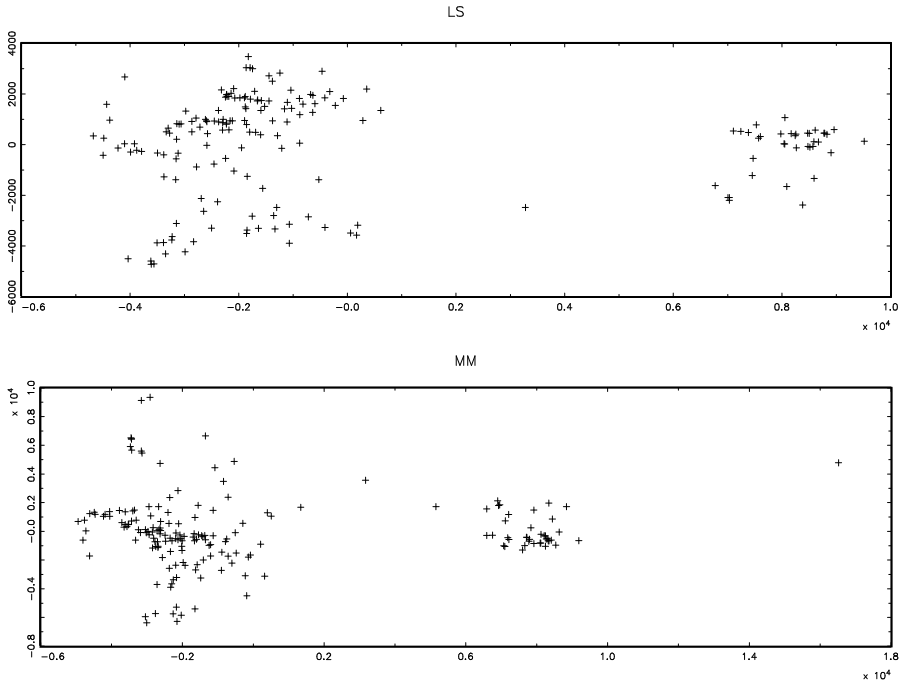

Figure 6: Vessel data: first two principal components from LS and MM estimates

In this dataset the phenomenon of false outliers can again be seen to affect the MM estimate in an important number of rows, while PertMM and RAR are only slightly affected. The details are omitted for brevity.

To assess resistance towards outliers, we now contaminate at random 3% of the cells by replacing their contents with $K = 8676 = \max_{i,j} x_{ij}$. The right-hand half ("Contaminated") of Table 7 shows the changes in the proportions of unexplained variability. It is seen that the criterion (21) is robust, while the classical one clearly breaks down. Table 9 gives the α-quantiles of absolute errors $|\widehat{x}_{ij} - x_{ij}|$ where x_{ij} are the *original* (i.e., not contaminated) data. It is seen that

- LS and S-M give similarly bad fits
- PertMM gives the best fit
- L_1 and RAR give similar fits, inferior to that of PertMM.

α	LS	S-M	L_1	RAR	PertMM
0.2	96.2	91.7	2.0	1.4	1.5
0.5	237.0	233.7	7.5	5.3	4.8
0.7	353.1	364.5	16.5	12.3	9.4
0.9	606.9	652.0	72.9	70.7	28.3
0.95	778.0	841.7	167.3	160.2	50.6
0.97	913.0	989.1	292.1	241.8	75.13

Table 9: Vessel dataset: α-quantiles of absolute errors of contaminated data

In some cases, L_1 completely breaks down, as shown in Figure 7 for row (spectrum) 91.

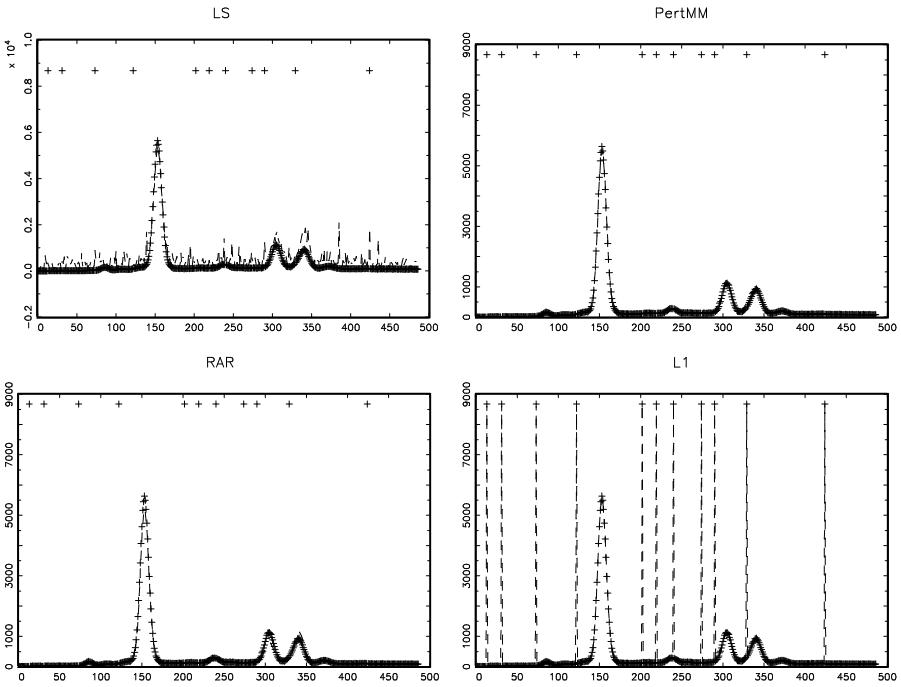

Figure 7: Contaminated vessel data: fits to row 91 as functions of frequency.

We can conclude that PertMM outperforms its competitors on this dataset.

For the reasons given at the end of Section 4, a version of RAR using the raw instead of the normalized variables was also used with the dataset of this Section, but the results were not better than with the original version.

6 Computing times

The computing times of the estimates depend on the programming language, number of iterations, etc. However, an analysis of the number of operations involved in their computation may help predict their comparative behaviors for large n, p and q.

RAR uses the initial Croux and Ruiz-Gazen's (2005) PCA estimate for which the number of operations is $O\left(pqn^2\right)$, and each RAR iteration requires weighted q-variate L_1 regressions which require at least $O\left(npq^2\right)$ operations, and computing the MVE which is $\left[O\left(q^2\left(n+p\right)\right)+O\left(q^3\right)\right]$ times the number of subsamples. We may thus expect a slow behavior for very large n and q.

The MM estimate requires an initial fit of rank q based on stepwise rank-one fits, with a cost $O\left(npq\right)$. The subsequent rank-q iterative process requires $O\left(npq^2\right)$ operations. Experiments showed that the initial fit is already rather close to the solution, and therefore the rank-q process usually requires less iterations than the initial fit, which implies an important time economy.

An experiment was run to compare the running times of PertMM and RAR. For each situation, each estimate was computed 5 times, and the average times reported. The estimates were implemented in Gauss and run on a 1200 MHz PC. Table 10 gives the mean computing times (in minutes) required for a fit of rank q to the LRS dataset of Section 5 and Figure 8 plots the results. It is seen that PertMM is much faster than RAR for all q, and also faster than L_1 for $q > 10$.

It should be noted that fits with large q may be needed in situations with very large p such as computer vision; e.g., De la Torre and Black (2001) employ $q = 20$.

q	PertMM	RAR	L_1
2	2.7	3.9	0.9
4	4.6	6.1	2.0
6	5.6	11.2	2.9
10	7.9	18.1	7.4
15	10.9	34.6	14.9
20	15.3	45.4	23.5

Table 10: Average computing times (in min.) of Perturbed-MM and RAR estimates for fits of rank q to LRS data

It is curious that the running time of PertMM seems sublinear in q. It must be kept in mind that in a matrix language like Gauss, Matlab or R, the relationship between the number of operations and the computing time is in general not linear, and is difficult to predict.

Remark: The computing time of RAR could be drastically reduced by replacing the MVE with an estimate which admits of iterative improvements,

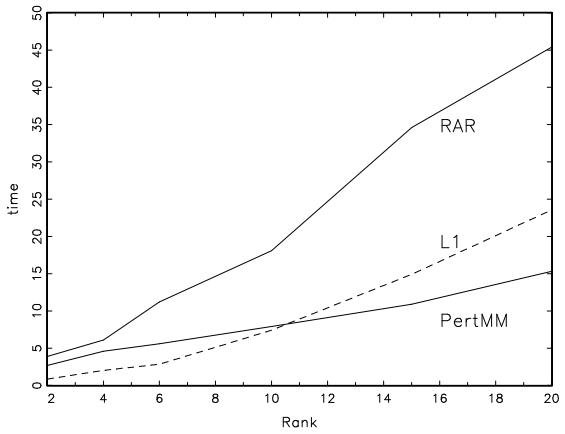

Figure 8: Average computing times (in min.) of PertMM, RAR and L_1 estimates for fits of rank q to LRS data

such as the MCD or an S-estimate. In this case the subsampling procedure would be needed only at the first iteration, and the matrix would be just updated at each iteration instead of computed from scratch.

7 Conclussions

In this article we have addressed the problem of fitting a low-rank matrix to a large dataset containing possible both element- and row-wise contamination and missing values. We have proposed a procedure based on the approach of MM estimation. We give a fast procedure for its numerical computation, based on iterative alternating regressions . Our procedure presents a tendency to produce a small proportion of false outliers; this drawback is overcome by introducing "small" perturbations in the data. Examples with real datasets indicate that the proposed procedure gives sensible results and in particular overcomes the problem of false outliers, while simulations show that the perturbed estimate is not less robust than the original one.

We have compared our estimate against two estimates designed for row-wise contamination: SPS and S-M; and two based on alternate regressions, L_1 and RAR, that can in principle withstand both row- and element-wise contamination. The simulations indicate that, as expected, the first two outperform the others for row-wise contamination, but fail even under light element-wise contamination; that L_1 fails under row-wise contamination, and that our perturbed MM estimate outperforms the last two under both types of contamination. Our procedure is also more "efficient" in the sense that it is comparable to classical PCA when there is no contamination. As respects computing times, our estimate is clearly faster than RAR, and also faster than L_1 for large rank.

As respects the problem of false outliers, our experiments with real data indicate that both our original MM estimate and RAR are affected by it, but that PertMM generally overcomes this drawback.

References

Bay, S.D. (1999), The UCI KDD Archive [http://kdd.ics.uci.edu], Irvine, CA: University of California, Department of Information and Computer Science.

Campbell, N.A.(1980), Robust procedures in multivariate analysis I: Robust covariance estimation, *Applied Statistics,* **29**, 231–237.

Croux, C., Filzmoser, P. , Pison, G. and Rousseeuw, P.J. (2003), Fitting multiplicative models by robust alternating regressions, *Statistics and Computing*, **13,** 23-36.

Croux, C. and Haesbroeck, G. (2000), Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, *Biometrika*, **87**, 603–618.

Croux, C. and Ruiz-Gazen, A. (1996), A fast algorithm for robust principal components based on projection pursuit. In: Prat, A. (Ed.), Compstat: Proceedings in Computational Statistics, Heidelberg: Physica-Verlag, 211–216.

Croux, C., Ruiz-Gazen, A. (2005), High breakdown estimators for principal components: the projection-pursuit approach revisited, *Journal of Multivariate Analysis*, **95**, 206-226.

De la Torre, F. and Black, M.J. (2001), Robust principal components analysis for computer vision, In *Proc. International Conference on Computer Vision*, 2001. http://citeseer.ist.psu.edu/torre01robust.html

Devlin, S.J., Gnanadesikan, R. and Kettenring, J.R. (1981) Robust estimation of dispersion matrices and principal components, *Journal of the American Statistical Association,* **76**, 354–362.

Gabriel, K.R. and Zamir, S. (1979), Lower-rank approximation of matrices by least squares with any choice of weights, *Technometrics,* 21, 489-498.

Hubert, M., Rousseeuw, P.J., Vanden Branden, K. (2005), ROBPCA: a new approach to robust principal component analysis, *Technometrics*, **47**, 64-79.

Janssens, K., Deraedt, I., Freddy, A., Veekman, J. (1998). Composition of 15-17$^{\text{th}}$ century archeological glass vessels excavated in Antwerp, Belgium. *Mikrochimica Acta*,**15** (Suppl.), 253-267.

Liu, L., Hawkins, D.M., Ghosh, S. and Young, S.S. (2003), Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences,* **100**, 13167-1317.

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T. and Cohen, K.L. (1999), Robust principal components for functional data, *Test,* **8**, 1–28.

Maronna, R.A. (2005), Principal components and orthogonal regression based on robust scales, *Technometrics*, **47**¸264-273.

Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*, New York, John Wiley and Sons.

Maronna, R.A. and Zamar, R.H. (2002), Robust estimation of location and dispersion for high-dimensional data sets, *Technometrics,* **44,** 307-317.

Naga, R. and Antille, G. (1990) Stability of robust and non-robust principal component analysis, *Computational Statistics & Data Analysis,* **10**, 169–174.

Rey, W. (2007). Total singular value decomposition. Robust SVD, regression and location-scale. http://arxiv.org/abs/0706.0096

Serneels, S. and Verdonck, T. (2007). Principal component analysis for data containing outliers and missing elements. *Computational Statistics and Data Analysis*, in press.

Van Aelst, S., Alquallaf, F., Yohai, V.J. and Zamar, R.H. (2006). A model for contamination in multivariate data. Unpublished manuscript.

Verboon, P. and Heiser, W.J. (1994), Resistant lower rank approximation of matrices by iterative majorization, *Computational Statistics and Data Analysis,* **18**, 457-467.

Yohai, V.J. (1987), High breakdown-point and high efficiency estimates for regression, *The Annals of Statistics,* **15,** 642-656.

Yohai, V.J. and Zamar, R.H. (1988). High breakdown estimates of regression by means of the minimization of an efficient scale, *Journal of the American Statistical Association,* **83,** 406-413.