Implementations for the minimum covariance determinant estimator

J. Duintjer Tebbens[1,2] and J. Kalina[1]

Institute of Computer Science, Czech Academy of Sciences[1] and Faculty of Pharmacy in Hradec Králové, Charles University in Prague[2]

Robust estimation of location and scatter

In statistics, the term *robustness* is mostly used to indicate robustness with regards to outliers in the observed data. More precisely, a descriptive value is said to be robust of it is not significantly influenced by possible outliers in the data. The detection of outliers in p-dimensional data (i.e. observations with p recorded properties) is difficult if $p > 3$ because one can not rely on visual inspection. In the univariate case, a single outlier might still be relatively easily detected by measuring with a norm called Mahalanobis distance. This distance is in fact the energy norm for the inverse of the symmetric positive definite covariance matrix S and scales the p-dimensional space such that the variabilities of the individual properties are normalized. In a multivariate situation, with multiple outliers, the Mahalanobis distance itself is too strongly influenced by the outliers to give a reliable tool for their detection, a phenomenon called the masking effect.

If the aim is to estimate the location and scatter by robust estimators (i.e. to compute a robust mean vector and robust covariance matrix), one can compute the location and scatter for a subset of the observations which hopefully does not contain outliers. Assume we

- have n observations $x_i \in \mathbb{R}^p$ of p variables, given by the data matrix $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n\times p}$,
- and look for a subset of size h of the indices $\{1, 2, \ldots, n\}$, where $[(n + p + 1)/2] \le h \le n$, such that no index in the subset corresponds to an outlier.

A criterion to base the search of the subset on and that has been proved to lead to highly robust estimators of location and scatter is to *minimize the determinant* of the covariance matrix [2]. For a given subset H of size h of the indices $\{1, 2, \ldots, n\}$, let

- the corresponding mean $\bar{x}_H$ be

$$\bar{x}_H = \frac{\sum_{i\in H} x_i}{h} \in \mathbb{R}^p$$

- and the corresponding covariance matrix S_H be

$$S_H = \frac{1}{h-1}\sum_{i\in H}(x_i - \bar{x}_H)(x_i - \bar{x}_H)^T = \frac{1}{h-1}(X_H^T - \bar{x}_H\mathbf{1}_h^T)(X_H - \mathbf{1}_h\bar{x}_H^T) = \frac{1}{h-1}(X_H^c)^T X_H^c \in \mathbb{R}^{p\times p},$$

where $\mathbf{1}_h \in \mathbb{R}^h$ is the vector of ones, X_H the data matrix for the indices in H and $X_H^c \in \mathbb{R}^{h\times p}$ is the corresponding centered data matrix. The *Minimum Covariance Determinant Estimator* [3] defines the optimal subset H_0 of size h of $\{1, 2, \ldots, n\}$ as

$$H_0 = \arg\min_H \det(S_H) = \arg\min_H \det\Big(\sum_{i\in H}(x_i - \bar{x}_H)(x_i - \bar{x}_H)^T\Big)$$

and defines the corresponding estimates of location and scatter as $\bar{x}_{H_0}$ and S_{H_0}, respectively.

The fast MCD algorithm [4]

The computation of the Minimum Covariance Determinant Estimator requires minimization over all $\binom{n}{h}$ h-subsets of $\{1, 2, \ldots, n\}$, thus has combinatorial complexity and becomes infeasible for very moderate numbers of observations n. In the widely used *fast MCD [4]* algorithm:

- one attempts to approximate the minimum determinant
- with several determinant minimizing steps for a large ($\pm$ 500) number of trial h-subsets
- and selects the h-subset leading after the minimizing steps to the smallest determinant

The determinant minimizing steps are called C-steps (concentration steps) and rely on the following theorem:

Theorem (C-step [4])

Let H_1 be an h-subset with corresponding location $\bar{x}_{H_1}$ and scatter S_{H_1}. If $\det(H_1) \neq 0$ compute the Mahalanobis distances

$$d(i) = \sqrt{(x_i - \bar{x}_{H_1})^T S_{H_1}^{-1}(x_i - \bar{x}_{H_1})}, \qquad i = 1, \ldots, n$$

and find a re-ordering $j_1, \ldots, j_n$ of $\{1, 2, \ldots, n\}$ such that

$$d(j_1) \le d(j_2) \le \cdots \le d(j_n).$$

Then if H_2 is the h-subset consisting of the indices $\{j_1, \ldots, j_h\}$ and S_{H_2} is the corresponding covariance matrix,

$$\det(S_{H_2}) \le \det(S_{H_1})$$

with equality if and only if $\bar{x}_{H_1} = \bar{x}_{H_2}$ and $S_{H_1} = S_{H_2}$.

The main computational costs of one C-step can be summarized as follows:

- construction of the current covariance matrix S_{H_1} : $\mathcal{O}(np^2)$ flops.
- Cholesky- or eigendecomposition of S_{H_1} (this also yields $\det(S_{H_1})$) : $\mathcal{O}(p^3)$ flops.
- computation of the distances $d(i)$: $\mathcal{O}(np^2)$ flops.

We will consider C-steps based on eigendecomposition, that is, they compute

$$S_{H_1} = Z_1 D_1 Z_1^T$$

with D_1, Z_1 the eigenvalue and eigenvector matrix, respectively, and find the Mahalanobis distances $d(i)$ using

$$d(i) = \sqrt{(x_i - \bar{x}_{H_1})^T Z_1 D_1^{-1} Z_1^T (x_i - \bar{x}_{H_1})}, \qquad i = 1, \ldots, n.$$

Our contribution consists of two cheap, $\mathcal{O}(np)$ permutations that can be added to the C-step to improve its power with regards to minimizing the determinant.

An *a-posteriori* permutation

Suppose after a C-step, we have selected a new h-subset based on the ordered distances $d(j_1) \le d(j_2) \le \cdots \le d(j_n)$ as described in the previous theorem. In other words, the new h-subset H_2 consists of the indices $\{j_1, \ldots, j_h\}$. A natural question is whether among the discarded indices $\{j_{h+1}, \ldots, j_n\}$ there may be indices that, if included in H_2, would yield a covariance matrix with smaller determinant. This can be checked in a computationally inexpensive way as follows.

If instead of H_2 we use the h-subset $\{j_1, \ldots, j_{h-1}, j_r\} \equiv H_r$ for some index $j_r \in \{j_{h+1}, \ldots, j_n\}$, then the data matrix for H_r differs from the data matrix for H_2 in one column only. Therefore, the corresponding covariance matrices are small rank updates from eachother.

Theorem (Low rank update of a covariance matrix ([1], Theorem 3.2.2))

Let $d_r = x_{j_h} - x_{j_r} \in \mathbb{R}^p$, let S_r denote the covariance matrix for H_r and let $f = e_h - \mathbf{1}_h/h \in \mathbb{R}^h$, where e_h denotes the hth unit vector. Then there holds

$$S_r = S_{H_2} - d_r f^T X_2^c - (X_2^c)^T f d_r^T + \|f\|^2 d_r d_r^T.$$

All vectors involved in the low-rank update can be computed with $\mathcal{O}(p)$ flops. Moreover, information on the determinant of S_r can be obtained from the determinant of S_{H_2} in $\mathcal{O}(p)$ flops as well: Using the eigendecomposition $S_{H_2} = Z_2 D_2 Z_2^T$, the eigendecomposition of S_r for the modified h-subset H_r can be written as

$$\begin{aligned} S_r &= S_{H_2} - d_r f^T X_2^c - (X_2^c)^T f d_r^T + \|f\|^2 d_r d_r^T \\ &= Z_2\left(D_2 - Z_2^T d_r f^T X_2^c Z_2 - (X_2^c Z_2)^T f d_r^T Z_2 + \|f\|^2 Z_2^T d_r d_r^T Z_2\right) Z_2^T. \end{aligned}$$

Thus the eigenvalues of S_r are the eigenvalues of a symmetric rank-three update of the diagonal matrix D_2 and each eigenvalue can be obtained, using (inverse) power iteration, in $\mathcal{O}(p)$ flops. To keep the flop count at $\mathcal{O}(p)$, we propose to compute only the s, $s \le 5$, largest eigenvalues of each covariance matrix S_r. After testing for all $j_r \in \{j_{h+1}, \ldots, j_n\}$, we select the index j_r for which the product of the s largest eigenvalues of S_r is minimal. The total flop count for this *a posteriori* permutation is of order $(n-h)sp$.

A look-ahead permutation

The weakness of the *a posteriori* permutation is that it tends to find, in numerical tests, an index j_r to exchange the index j_h of H_2 with, which would have been found anyway in the next C-step, i.e. the index j_r often becomes a member of H_3 anyway. The proposed *a posteriori* permutation is therefore mainly useful to add to the very last C-step to be performed.

To overcome this weakness, we propose a second permutation which looks ahead at the indices of H_3 and attempts to add an index to H_2 that will not be in H_3. Assume that with a candidate h-subset H_2 we compute the Mahalanobis distances

$$d(i) = \sqrt{(x_i - \bar{x}_{H_2})^T S_{H_2}^{-1}(x_i - \bar{x}_{H_2})}, \qquad i = 1, \ldots, n \qquad (1)$$

and find a re-ordering $k_1, \ldots, k_n$ of $\{1, 2, \ldots, n\}$ such that

$$d(k_1) \le d(k_2) \le \cdots \le d(k_n).$$

Then H_3 would be defined as the indices $\{k_1, \ldots, k_h\}$. We can test whether indices in $\{k_{h+1}, \ldots, k_n\} \setminus H_2$ yield a lower determinant of S_{H_2} when interchanged with i_h. This can be done in $\mathcal{O}((n-h)sp)$ flops as before. When the index for which the product of the s largest eigenvalues of S_r is minimal is found, we replace H_2 with H_r and have to recompute the Mahalanobis distances

$$d(i) = \sqrt{(x_i - \bar{x}_{H_r})^T S_{H_r}^{-1}(x_i - \bar{x}_{H_r})}, \qquad i = 1, \ldots, n$$

to perform the next C-step. Fortunately, this does not require the full $\mathcal{O}(np^2)$ flops for a regular C-step. Thanks to the fact that H_r is a small-rank update of H_2, it can be done in $\mathcal{O}(np)$ flops using (1) and the Sherman-Morrison formula.

Experiment

We generated 10 data sets $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{100\times 10}$ each with 100 observations and 10 variables. 80 observations were normally distributed with mean vector 0 and covariance matix $\Sigma = 0.6 \cdot I_{10} + 0.4 \cdot \mathbf{1}_p \cdot \mathbf{1}_p^T$ and 20 randomly placed outliers were normally distributed with mean vector $3 \cdot \mathbf{1}_p$ and covariance matix $2 \cdot \Sigma$. For 25 random initial choices of H_0 and each of the ten datasets, we performed 4 regular C-steps and compared with 4 C-steps including the *look-ahead* permutations (dashed curve) and with 4 C-steps including both proposed permutations (solid curve). The quality of the results is measured by the squared norm of $\bar{x}_{H_4}$ (left figure) and the Frobenius norm of $\Sigma - S_{H_4}$ (right figure). The curves give the ratio of these measures for the improved vs. regular C-steps, averaged over all 25 random initial h-subsets.

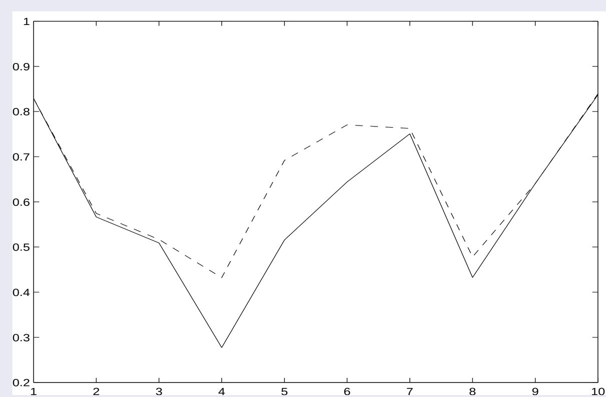

Figure 1 : Ratios (averaged over 25 random initial choices of H_0) of $\|\bar{x}_{H_4}\|^2$ for look-ahead improved C-steps (dashed) or C-steps improved with both proposed permutations (solid) versus regular C-steps; the x-axis gives the data set number.

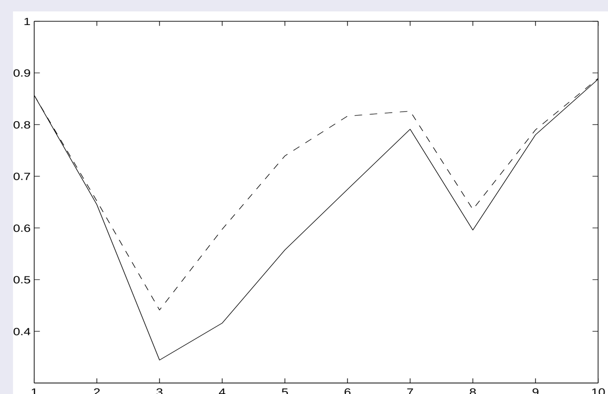

Figure 2 : Ratios (averaged over 25 random initial choices of H_0) of $\|\Sigma - S_{H_4}\|_F$ for look-ahead improved C-steps (dashed) or C-steps improved with both proposed permutations (solid) versus regular C-steps; the x-axis gives the data set number.

Acknowledgements

The work of J. Kalina was financially supported by the Neuron Fund for Support of Science. The work of J. Duintjer Tebbens was supported by the grant GA13-06684S of the Czech Science Foundation.

1. Athanasiadis, S. *The small sample size problem in gene expression tasks*, Diploma thesis, Faculty of Pharmacy, Charles University, 2015.
2. Grübel, R. *A minimal characterization of the covariance matrix* Metrika, vol. **35**, 49–52, 1988.
3. Hubert, M. and Debruyne, M. *Minimal covariance determinant* Metrika, Wiley Interdisciplinary Reviews: Computational Statistics, vol. **2**, 36–43, 2010.
4. Rousseeuw, P. and Van Driessen, K. *A fast algorithm for the minimum covariance determinant estimator* Technometrics, vol. **34**(3), 212–223, 1999.

J. Duintjer Tebbens
duintjertebbens@cs.cas.cz
Insitute of Computer Science
Czech Academy of Sciences