Implementation for the minimum covariance determinant estimator
J. Duintjer Tebbens1,2 and J. Kalina3

Institute of Computer Science, Czech Academy of Sciences1 and Faculty of Pharmacy in Hradec Králové, Charles University in Prague2

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 if 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 cannot rely only 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 variances 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 \( X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{p \times n} \) and look for a subset of size \( h \) of the indices \{1, 2, …, n\}, where \( \{n + p + 1)/2\} \leq h \leq n \), 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 \( \Sigma \). For a given subset \( A \) of \( \{1, 2, \ldots, n\} \), then the data matrix for \( A \) differs from the data matrix for \( H \) in one column only. Therefore, the corresponding covariance matrices are small rank updates from each other.

**Theorem (Low rank update of a covariance matrix)** [3, Theorem 3.2.2]

Let \( d_i = x_i - \bar{x}_H \in \mathbb{R}^p \), let \( S \) denote the covariance matrix for \( H \), and let \( f = x_i - \bar{x}_H \in \mathbb{R}^p \), where \( \bar{x}_H \) denotes the hth unit vector. Then there holds

\[
S = S_H - d_f S_H d_f^T + \|d_f\|^2 d_f d_f^T.
\]

All vectors involved in the low-rank update can be computed with \( O(p) \) flops. Moreover, information on the determinant of \( S \), which can be obtained from the determinant of \( S_H \) in \( O(p) \) flops as well. Using the eigendecomposition \( S_H = Z D Z^T \), the eigendecomposition of \( S \) for the modified h-subset \( H \) can be written as

\[
S = S_H - d_f d_f^T Z (Z^T d_f)^2 Z^T + \|d_f\|^2 Z^T d_f d_f^T Z^T.
\]

Thus the eigenvalues of \( S \) 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 \( O(p) \) flops. To keep the flop count at \( O(p) \), we propose to compute only the \( s \) largest eigenvalues of each covariance matrix \( S_h \). After testing for \( s \) indices \( j \in \{1, 2, \ldots, n\} \), we select the index \( j \) for which the product of the \( s \) largest eigenvalues of \( S_H \) is minimal. The total flop count for this a posteriori permutation is of order \( (n - h) s p \).

A look-ahead permutation
The weakness of the a posteriori permutation is that it tends to find, in numerical tests, an index \( j \) to exchange the index \( j \) of \( H \) with, which would have been found anyway in the next C-step, i.e. the index \( j \) often becomes a member of \( H \) anyway. Therefore, 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 \( H \) and attempts to add an index to \( H \) that will not be in \( H \). Assume that with a candidate h-subset \( H \) we compute the Mahalanobis distances

\[
d(i) = \sqrt{\langle x_i - \bar{x}_H \rangle^T S_H^{-1} \langle x_i - \bar{x}_H \rangle}, \quad i = 1, \ldots, n
\]

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

\[
d(k_1) \leq d(k_2) \leq \cdots \leq d(k_n).
\]

Then \( H \) would be defined as the indices \{\( k_1, \ldots, k_n \)\}. We can test whether indices in \{\( k_1, \ldots, k_n \)\} \( H \) yield a lower determinant of \( S_H \) when interchanged with \( t \). This can be done in \( O((n - h) s p) \) flops as before. When the index for which the product of the \( s \) largest eigenvalues of \( S_H \) is minimal is found, we replace \( H \) with \( H \) and have to recalculate the Mahalanobis distances

\[
d(i) = \sqrt{\langle x_i - \bar{x}_H \rangle^T S_H^{-1} \langle x_i - \bar{x}_H \rangle}, \quad i = 1, \ldots, n
\]

to perform the next C-step. Fortunately, this does not require the full \( O(n s p) \) flops for a regular C-step. Thanks to the fact that \( H \) is a small-rank update of \( H \), it can be done in \( O(s p) \) flops using \( \Sigma \) and the Sherman-Morrison formula.

**Experiment**
We generated 10 data sets \( X = [x_1, \ldots, x_n] \in \mathbb{R}^{10 \times 100} \) each with 100 observations and 10 variables. 80 observations were normally distributed with mean vector \( 0 \) and covariance matrix \( \Sigma = 0.6 \cdot I_1 + 0.4 \cdot I_2 \). \( I_1 \) and \( I_2 \) are 100 randomly placed outliers were normally distributed with mean vector \( 3 \) \( I_1 \) and covariance matrix \( 2 \Sigma \). For 25 random initial choices of \( H \) 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 \( \bar{x}_H \) (left figure) and the Frobenius norm of \( \Sigma - \bar{S}_H \) (right figure). The curves give the ratio of these measures for the improved versus regular C-steps, averaged over all 25 random initial h-subsets.

**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 GA33-06884S of the Czech Science Foundation.