Multivariate outlier detection in Stata

Vincenzo Verardi
University of Namur
(Centre for Research in the Economics of Development)
Namur, Belgium
and Université Libre de Bruxelles
(European Center for Advanced Research in Economics and Statistics
and Center for Knowledge Economics)
Brussels, Belgium
vverardi@ulb.ac.be

Catherine Dehon
Université libre de Bruxelles
(European Center for Advanced Research in Economics and Statistics
and Center for Knowledge Economics)
Brussels, Belgium
cdehon@ulb.ac.be

Abstract. Before implementing any multivariate statistical analysis based on empirical covariance matrices, it is important to check whether outliers are present because their existence could induce significant biases. In this article, we present the minimum covariance determinant estimator, which is commonly used in robust statistics to estimate location parameters and multivariate scales. These estimators can be used to robustify Mahalanobis distances and to identify outliers. Verardi and Croux (1999, Stata Journal 9: 439–453; 2010, Stata Journal 10: 313) programmed this estimator in Stata and made it available with the mcd command. The implemented algorithm is relatively fast and, as we show in the simulation example section, outperforms the methods already available in Stata, such as the Hadi method.

Keywords: st0192, mcd, detection, multivariate outliers, robustness, minimum covariance determinant

1 Multivariate detection

One of the principal objectives of statistics is to provide tools to describe multidimensional relations. However, if these tools are based on classical estimation methods and outliers are present in the dataset, the results can be biased. To deal with this drawback, it could be envisaged to identify the outliers and either downweight these observations or remove them from the dataset before calling on the statistical tool. Unfortunately, the identification of outliers is quite challenging because a visual inspection is troublesome when considering more than two dimensions.
Traditionally, the (multivariate) characterization of an outlier is measured by Mahalanobis distances, defined as
\[
d_i = \sqrt{(X_i - \mu)\Sigma^{-1}(X_i - \mu)}
\]
where \(X\) is the \(n \times p\) matrix containing the realizations of \(p\) random variables for \(n\) individuals, \(X_i\) is the \(i\)th row vector of matrix \(X\), \(\mu\) is the \(1 \times p\) multivariate location vector, and \(\Sigma\) is the \(p \times p\) covariance matrix. These distances are well known to be distributed as the square root of a chi-squared distribution with \(p\) degrees of freedom (\(\sqrt{\chi^2_p}\)) when \(X\) comes from a multivariate normal distribution. It is then rational to set a critical value (for example, the 97.5th percentile of a \(\sqrt{\chi^2_p}\)) above which an individual observation can be considered an outlier. Unfortunately, if outliers are present, the classical estimation of \(\mu\) (by the sample mean) is affected, and the estimation of \(\Sigma\) (by the sample covariance matrix) is inflated. Mahalanobis distances will hence be distorted. However, the identification of multivariate outliers using Mahalanobis distances is still possible if \(\mu\) and \(\Sigma\) are robustly estimated (that is, estimated using a method that is not excessively affected by outliers).

In Stata, an estimator aimed at robustly estimating the multivariate outlyingness (see Hadi [1992, 1994]) is available with the `hadimvo` command. Unfortunately, as will be shown in the simulation example section, the results of Hadi’s estimator may be adversely affected by outliers under some contamination scenarios. The weakness of Hadi’s method is the first step of the proposed iterative algorithm, which is based on ranking individuals by a measure that is not guaranteed to be robust to outliers.

In this article, we present the minimum covariance determinant (MCD) estimator of location and scatter, introduced by Rousseeuw (1985, 877), that has been proven to be particularly well-suited in this context and has become standard in robust statistics. Furthermore, we also present the intuition behind the algorithm Verardi and Croux (2009, 2010) used to implement this method (the `mcd` command) in Stata.

The structure of the article is as follows. In section 2, we briefly describe the MCD estimator, and in section 3, we present the computational algorithm implemented in Stata. In section 4, we use simulations to illustrate how the MCD estimator outperforms other available estimators. In section 5, we present the proposed Stata command and, in section 6, we provide a short summary discussion.

## 2 MCD

Before presenting the MCD estimator, it is helpful to recall the notion of generalized variance. This measure, originally introduced by Wilks (1932), is a one-dimensional assessment of multidimensional spread. For the sake of clarity, we explain this concept calling on a \(2 \times 2\) covariance matrix. The generalization to higher dimensions is straightforward.

---

1. These robust estimators are already implemented in S-plus, R, and SAS.
Let’s define the covariance matrix

\[ \Sigma = \begin{pmatrix} \sigma^2_{x_1} & \sigma_{x_1,x_2} \\ \sigma_{x_1,x_2} & \sigma^2_{x_2} \end{pmatrix} \]

where \( \sigma^2_{x_1} \), \( \sigma^2_{x_2} \), and \( \sigma_{x_1,x_2} \) are respectively the variance of variable \( x_1 \), the variance of variable \( x_2 \), and the covariance between variable \( x_1 \) and \( x_2 \). The generalized variance is defined as the determinant of \( \Sigma \) (i.e., \( \sigma^2_{x_1} \sigma^2_{x_2} - \sigma^2_{x_1,x_2} \)). To understand why this measure can be seen as a generalization of the variance, it is helpful to look more closely at the expression of the determinant. This expression is composed of two elements: the product of \( \sigma^2_{x_1} \) and \( \sigma^2_{x_2} \), and (minus) the squared covariance \( \sigma^2_{x_1,x_2} \). The first term \( \sigma^2_{x_1} \sigma^2_{x_2} \) represents the raw bi-dimensional spread of the observations. However, if \( x_1 \) and \( x_2 \) are not independent, each of the variables conveys some information on the other one. This redundant information should be accounted for. What remains is the bi-dimensional spread once the covariation has been accounted for.

Having defined the generalized variance, it is now easy to present the underlying principle of MCD. Consider that we want an estimator of the covariance matrix that withstands a contamination of up to 50% of sample points\(^2\) (it is then said to have a breakdown point of 50%). The basic idea of MCD is to identify the subsample containing 50% of the observations that is associated with the smallest generalized variance. This is equivalent to finding the subsample with the smallest multivariate spread. The MCD robust covariance matrix, labeled \( \Sigma_{MCD} \), and the MCD location vector, labeled \( \mu_{MCD} \), are then defined as the sample covariance matrix\(^3\) and location vector computed over this subsample.

To identify such a subsample, the idea is to consider all possible subsets containing 50% of the observations and flag the one with the smallest covariance matrix determinant. This is, of course, a cumbersome task. Imagine, for example, that we have a dataset of 100 observations. If we want to consider all possible subsamples containing 50 observations, the total number to check is \( \binom{100}{50} \approx 1.01 \times 10^{29} \). Fortunately, as shown by Rousseeuw and van Driessen (1999) and briefly explained in the following section, only a limited number of subsets needs to be considered in practice.

### 3 Fast-MCD algorithm

A feasible and stable algorithm for computing the MCD location and covariance matrix has been the pitfall of the method for a long time. This problem was solved by Rousseeuw and van Driessen (1999), who proposed the fast-MCD algorithm, which is particularly efficient. Verardi and Croux (2009, 2010) programmed this algorithm in Stata and created the \texttt{mcd} command. For the sake of brevity, we present the stylized algorithm here and refer the interested reader to the original article for further details.

---

1. A similar reasoning can be adopted for any \( h \% \) contamination where \( h \leq 50\% \).
2. Multiplied by a constant related to the assumed underlying distribution.

The algorithm is based on three steps. In the first step, \( N \) subsamples of size \( p + 1 \) (called the \( p \)-subsets) are randomly drawn from the dataset. For each \( j \) of the \( N \) \( p \)-subsets, the covariance matrix \( \hat{\Sigma}^p_j \) and the vector of location parameters \( \hat{\mu}^p_j \) are computed. Then for each \( p \)-subset, the determinant of \( \hat{\Sigma}^p_j \) is calculated. If the determinant is null, the number of sample points of the subset is increased until the determinant becomes nonnull. In the second step, for each \( p \)-subset, Mahalanobis distances are calculated for all points as in (1), using \( \hat{\mu}^p_j \) and \( \hat{\Sigma}^p_j \). The observations are then sorted in ascending order of the estimated Mahalanobis distances. After that, \( \hat{\mu}^p_j \) and \( \hat{\Sigma}^p_j \) are updated to the sample average and covariance matrix estimated on the 50% first-ranked points. The procedure is repeated until convergence. In the third and final step, all the determinants of the final \( N \) covariance matrices are compared. The subset that is associated with the smallest determinant is kept to estimate the final MCD covariance matrix and location vector.

To increase efficiency, a one-step robust reweighted covariance matrix (called RMCD) is estimated, calculating the classic covariance matrix exclusively on the nonoutlying individuals (and multiplying it by a consistency factor).

To increase the speed of the algorithm, the second and third steps are performed only for the 10 \( p \)-subsets with the smallest preliminary determinants. The minimal number of \( p \)-subsets \( (N) \) to be considered can be determined according to the formula

\[
N = \left\lceil \frac{\log(1 - P_{\text{clean}})}{\log \left\{ 1 - \left(1 - \epsilon\right)^p \right\}} \right\rceil
\]

(2)

where \( P_{\text{clean}} \) is the (chosen) probability of having at least one noncorrupt \( p \)-subset (it is generally set to 0.99), \( \epsilon \) is the maximal proportion of outliers that we expect to contaminate the dataset (it is generally set to 20%), and \( p \) is the number of variables. The rationale behind this formula is presented in the original article of Rousseeuw and van Driessen (1999).

### 4 Simulated example

To illustrate why MCD is particularly well suited for multivariate outlier identification, we create a dataset of size \( n = 1,000 \) by randomly generating 10 independent continuous variables (labeled \( X_1, \ldots, X_{10} \)) from a Gaussian distribution with mean 0 and unit variance. This dataset is called the clean dataset. Subsequently, we contaminate the dataset by randomly replacing 10% of the observations of the first variable with random values drawn from a Gaussian distribution with mean 5 and standard deviation 0.1. This is called the contaminated dataset.

We then run the \texttt{mcd} command (on both the clean and the corrupt dataset) to estimate the covariance matrix and compute the robust distances. We obviously would like the estimated covariance matrix to be close to the identity matrix in both the clean and the contaminated dataset (because this is the covariance matrix that was used to generate the vast majority of the observations). In this example, we compare three
estimated covariance matrices: the classic one, the classic one with outliers identified by the Hadi method removed (defined as Hadi), and the RMCD covariance matrix. The test used to check whether the underlying population covariance matrix is the identity matrix, is a standard likelihood-ratio test. The calculated test statistic is $W = -2 \left[ np/2 + n/2 \log \left\{ \det(\hat{\Sigma}) \right\} - n/2 \text{trace}(\hat{\Sigma}) \right]$. As can be seen in table 1, before the contamination, all three estimation methods lead to an estimated covariance matrix that is not statistically different from the identity ($p$-value > 0.05). After the contamination, only the MCD estimated covariance matrix passes the test. As far as the MCD location vector is concerned, all elements are very close (and not statistically different) to 0.

<table>
<thead>
<tr>
<th>Table 1. Likelihood-ratio test on $\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>$W$</td>
</tr>
<tr>
<td>MCD</td>
</tr>
<tr>
<td>Hadi</td>
</tr>
<tr>
<td>Classic</td>
</tr>
</tbody>
</table>

To focus on the identification of outliers, we present in figure 1 two distance–distance plots comparing the Mahalanobis distances based on MCD estimations of location and scatter (on the y axis) with (on the x axis) the Hadi based distances (on the left) and the classic Mahalanobis distances (on the right). For the sake of clarity, the artificially generated outliers are identified with hollow circles. A line is drawn on both axes in correspondence to the critical value $\left( \sqrt{\chi^2_{10,0.975}} = 4.52 \right)$ to identify the value above which an individual can be considered an outlier.

(Continued on next page)
The result is striking. MCD robust distances allow the identification of all generated outliers, because all hollow circles are above the horizontal line. On the other hand, the methods based on Mahalanobis and Hadi distances behave (equivalently) poorly, because almost all hollow circles are on the left of the vertical line. Obviously, and as expected, about 2.5% of sample points generated by the (uncontaminated) normal distribution are above the benchmark set for robust distances.

5 The MCD command

5.1 Syntax

The mcd command (Verardi and Croux 2009, 2010) estimates the robust (one-step, reweighted MCD) covariance matrix (and names it covRMCD), using the fast-MCD algorithm described above. mcd also calculates robust distances according to (1), where \( \mu \) and \( \Sigma \) are estimated with MCD estimators of location and scatter. Type \texttt{search mcd, sj} and follow the online instructions to download \texttt{mcd} and its affiliated commands. The syntax of the \texttt{mcd} command is

\[ \texttt{mcd varlist [if] [in] [, e(#) proba(#) trim(#) generate(newvar1 newvar2) bestsample(newvar) raw setseed(#)]} \]
5.2 Options

The `mcd` command has seven options. The first two options allow the change of the number of \(p\)-subsets analyzed in the algorithm. More specifically, the `e(#)` option requests that Stata change the maximum expected percentage of outliers in the dataset \([\varepsilon in (2)]\). The default is `e(0.2)`, but it can take any value belonging to the interval \((0, 0.5)\). The `proba(#)` option requests that Stata fix the probability of having at least one noncorrupt \(p\)-subset among the \(N\) considered by the algorithm \([P_{\text{clean}} in (2)]\). The default is `proba(0.99)`, but it can take any value belonging to the interval \((0, 1)\).

The third option, `trim(#)`, allows the user to specify the percentage of outliers that the estimator can withstand before breaking up. The default is `trim(0.5)`, but it can take any value belonging to the interval \((0, 0.5)\). The fourth option, `generate()`, asks the user to provide two variable names that return identified outliers and robust distances, respectively.

The fifth option, `bestsample()`, asks the user to provide a name for a variable that flags the points that are in the subset with the smallest covariance matrix determinant. The sixth option, `raw`, specifies that Stata return the genuine MCD location and scatter matrices rather than the one-step, reweighted MCD (the default).

The last option, `setseed(#)`, allows the user to specify the seed (to a value chosen by the user). By default, the seed is not set, which means that results may change over replications; see [8] `set seed` for information on setting the seed.

6 Conclusion

Several multivariate outlier detection tools are available in statistical software. Unfortunately, most of them use estimates of the covariance that are sensitive to the presence of the outliers to be detected, thus leading to masking and swamping effects. Several robust alternatives have been proposed, but the complexity behind their practical implementation is such that they did not manage to emerge as standard tools used by applied researchers. In this article, we used the `mcd` command (which allows a robust identification of multivariate outliers) that Verardi and Croux (2009, 2010) programmed in Stata to tackle the problem. To illustrate the usefulness of the method, we ran some simulations to show how it dramatically outperforms the previously available methods.

`mcd` also provides robust estimations of multivariate location and scatter. Relying on these, it is now easy to implement robust multivariate statistical techniques. Such an approach was taken for robust principal component analysis (Croux and Haesbroeck 2000), discriminant analysis (Hubert and Van Driessen 2004), or canonical correlations analysis (Croux and Dehon 2002) by simply replacing the classic estimators with the MCD ones. Moreover, as shown by Verardi and Croux (2009) in a companion article, in the context of regression analysis, the identification of multivariate outliers among explanatory variables allows for the detection of leverage points.

4. That is, points with a robust distance larger than \(\sqrt{\chi^2_{p,0.975}}\).
7 Acknowledgments

We would like to thank Christophe Croux for his precious comments. Vincenzo Verardi is an associate researcher of the National Science Foundation of Belgium and gratefully acknowledges their financial support.

8 References


About the authors

Vincenzo Verardi has a PhD in economics from the Université Libre de Bruxelles. He is an associate researcher of the National Science Foundation of Belgium and is a professor of economics and econometrics at the University of Namur and the Université Libre de Bruxelles. His research fields are applied econometrics, robust methods, political economy, and public economics.

Catherine Dehon has a PhD in statistics from the Université Libre de Bruxelles. She is an associate professor of statistics and econometrics at Université Libre de Bruxelles. Her current fields of research include robust regression, robust multivariate analysis, and recently, the robustification of econometric methods.