Multivariate outlier detection with compositional data

P. Filzmoser[1], K. Hron[2]

[1] *Dept. of Statistics and Probability Theory, Vienna University of Technology, Vienna, AUSTRIA*
e-mail: P.Filzmoser@tuwien.ac.at
[2] *Dept. of Mathematical Analysis and Applications of Mathematics, Palacký University, Olomouc, CZECH REPUBLIC*

Abstract

Multivariate outlier detection is usually based on Mahalanobis distances, by plugging in robust estimates of location and covariance. For compositional data, carrying only relative information, a special transformation needs to be consulted in order to be able to work in the appropriate geometry. The effect of the transformation is discussed in this contribution. Furthermore, different possibilities for the interpretation of the identified multivariate outliers are presented.

1 Multivariate outliers

Multivariate data consist of n observations that are simultaneously measured on D variables. The observations are collected in the rows $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD})^t$ (for $i = 1, \ldots, n$) of the data matrix $\mathbf{X}$. It is often of interest to compare the n observations with respect to their multivariate information. It will be informative whether the observations form certain groups, or whether there are strongly deviating data points. Here we focus on the latter issue, and thus we are interested in identifying potential outliers. In the univariate case this is very simple, because a ranking of the data already allows to identify the largest and smallest observations, which are potential candidates of outliers. Comparing their robust Z-scores (robustly standardized observations) to the standard normal distribution allows to draw more detailed conclusions about univariate outlyingness, see, e.g., [10]. In the bivariate case the task is already more complicated. Bivariate outliers are not necessarily the most extreme observations along one coordinate. These could also be observations that are not extreme in any of the coordinates, but they deviate from the structure formed by the data majority. Fortunately, bivariate data can still be plotted, and the main data structure can thus be visualized. However, in the higher-dimensional case a visual representation of the data is usually impossible, and thus the identification of deviating data points is not reliable. In the best case it is possible to identify outliers that are visible in one or two dimensions, but real multivariate outliers can hardly be discovered. This calls for the necessity of an automatic procedure that highlighted potential multivariate outliers.

Similar to the univariate case, also for multivariate outlier detection we need to state an underlying model. Outliers would then severely deviate from this hypothetical

model. It is common to consider the mutlivariate normal distribution as the "model", and outliers are considered as points that are generated by a different (unspecified) distribution. Generally, it is difficult for any outlier detection method to find out, whether multivariate outliers indeed originate from a distribution other than the multivariate normal distribution, or whether they are just extreme values of the multivariate normal distribution ("extreme" not in terms of a single coordinate, but in the multivariate sense). Therefore, the term "potential multivariate outliers" is usually more appropriate, and the analyst may then investigate the reasons why the observation has been flagged. More discussion on "extremes" versus "outliers" can be found in [4].

Multivariate outlier detection is usually based on computing the Mahalanobis distance, which is defined for an observation $\mathbf{x}_i$ as

$$\mathrm{MD}(\mathbf{x}_i) = [(\mathbf{x}_i - T)^t C^{-1} (\mathbf{x}_i - T)]^{1/2}, \ i = 1, \ldots, n,$$

where T and C are estimators of location and covariance [8]. Clearly, for reliable outlier detection in the sense of robust statistics, both T and C have to be estimated in a robust way, and not in the traditional way by arithmetic mean vector and sample covariance matrix. Figure 1 shows the difference of classical and robust estimation for a two-dimensional data set, leading to classical and robust Mahalanobis distances, respectively. The ellipses shown correspond to certain fixed values of the Mahalanobis distance. Points lying at such an ellipse would thus correspond to this fixed distance. The way of estimation is important: while in the left-hand picture classical estimates were plugged in, for the right-hand picture robust estimates were used. It is obvious that the robust estimates lead to Mahalanobis distances that much better capture the inherent data structure. Using the classical estimates does not only incorrectly assign the data center, but the deviating data points also cause that the ellipses are inflated. Under the assumption of multivariate normal distribution, the (classical) squared Ma-

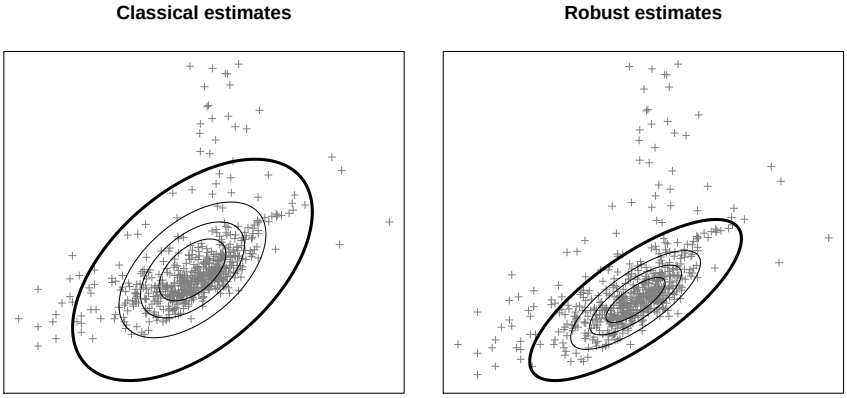

Figure 1: Visualization of Mahalanobis distances using classical (left) and robust (right) estimates of location and covariance.

halanobis distances follow a chi-square distribution with D degrees of freedom, see,

e.g., [9]. This distribution might also be considered for the robust case, and a quantile, e.g. 0.975, can be used as a cut-off value separating regular observations from outliers. This cut-off value is indicated in Figure 1 by the outer ellipse. Clearly, only the robust version allows for a reliable multivariate outlier detection.

Robust estimates of location and covariance can be obtained for instance from the MCD (Minimum Covariance Determinant) estimator, which is widely used because of its fast algorithm [11]. The MCD estimator is defined by those h observations that result in the smallest determinant of their sample covariance matrix. The robust location estimator is the arithmetic mean of these h observations, and the robust covariance is defined by the sample covariance matrix of the h observations, multiplied by a factor for consistency at normal distributions. Taking $h \approx n/2$ yields maximum robustness, but the estimator looses efficiency. Thus it is recommended to take $h \approx 3n/4$ as a compromise between efficiency and robustness. Figure 2 shows the results of the MCD estimator for the same data set as used in Figure 1. For the left-hand picture $h = n/2$ was used, and for the right-hand picture h was chosen as 3/4. The dark symbols refer to the solution of the MCD algorithm. The resulting MCD location estimator is indicated as grey dot, and the MCD covariance estimator (after multiplication with a constant for consistency) is indicated with an ellipse. Here, the solutions for both values of h are practically identical.

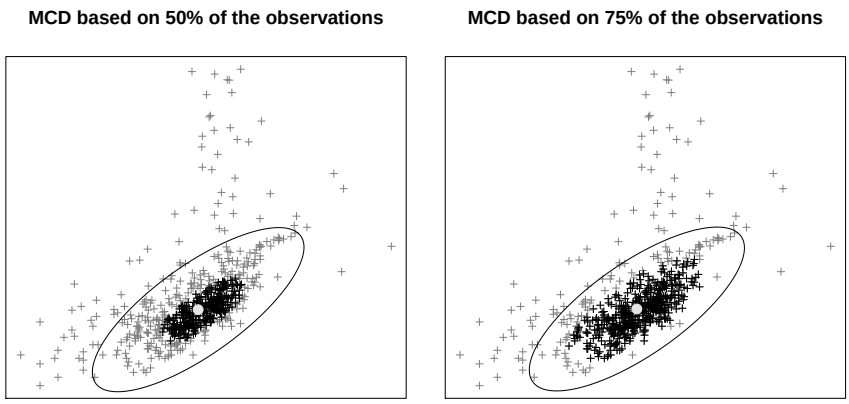

Figure 2: Functionality of the MCD estimator using different proportions of data.

2 Compositional data and outlier detection

Compositional data are multivariate data forming proportions in a whole. Thus this kind of data is containing only relative information, and only the ratios between the variables (called compositional parts) are relevant [1]. Compositional data occur in many areas, like in environmental sciences (e.g. concentrations of chemical compounds in a sample), or official statistics (e.g. household expenditures). D-part compositions are not represented in the Euclidean space but in the D-part simplex, a $(D-1)$-

dimensional subset of $\mathbf{R}^{D-1}$ that contains all D-part compositions that sum up to a prescribed constant sum. Therefore, whenever statistical methods that are designed for the Euclidean geometry (i.e. practically all standard methods) should be applied to compositional data, the data first have to be transformed from the simplex with its own geometry to the Euclidean space. Several possibilities were introduced by [1], but the transformation with the best properties is the ilr (isometric log-ratio) transformation introduced in [3]. The original compositions $\mathbf{x}_i$ with D parts are transformed by ilr to new observations $\mathbf{z}_i$ (for $i = 1, \ldots, n$) that can be viewed as coordinates in an orthogonal basis system on the simplex of dimensionality $D - 1$.

Multivariate outlier detection, as described above, assumes the Euclidean geometry. Hence, this method cannot directly be applied to the original compositional parts, but only to appropriately transformed data. The method could not even be applied to the original compositions, because these are singular, and robust covariance estimation (e.g. using the MCD estimator) does not work with singular data. The ilr transformation is one possibility, but just for the purpose of outlier detection also other transformations could be used under certain conditions, see [5] for details. Here we will focus only on the ilr transformation.

An illustration showing the different geometry of original compositional parts and their transformation to ilr coordinates will be given in the following. We use data describing mean consumption expenditures of households from 2008 in the countries of the European Union. The data set is available at `http://epp.eurostat.ec.europa.eu` at `statistics_explained/index.php/Household_consumption_expenditure`. The expenditures on food, alcohol and tobacco, clothing, housing, furnishings, health, transport, communications, recreation, education, restaurants and hotels, and on other goods and services are reported for the countries Austria (A), Belgium (B), Bulgaria (BG), Cyprus (CY), Czech Republic (CZ), Denmark (DK), Estonia (EST), Finland (FIN), France (F), Germany (D), Greece (GR), Hungary (H), Ireland (IRL), Italy (I), Latvia (LV), Lithuania (LT), Luxembourg (L), Malta (M), Netherlands (NL), Poland (PL), Portugal (P), Romania (R), Slovakia (SK), Slovenia (SLO), Spain (ES), Sweden (S), and United Kingdom (GB). These are compositional data because the expenditures are parts of the overall household incomes. For example, if more money is devoted to one part, typically less money will be left for the other parts, and thus not the absolute numbers but only their ratios are informative. In the following example we will only use the parts food, alcohol and tobacco, and recreation, because then the data structure can be visualized easily. Figure 3 shows the original compositions in ternary diagrams (left column) and the ilr-transformed data (right column). The ternary diagram is a convenient graphical tool for compositions. The closer a point comes to the edge of the plot, the higher is the proportion on this compositional part [1]. The indicated ellipses on the right-hand side correspond to the 0.5 and 0.975 quantiles of the corresponding chi-square distribution, respectively, and thus the outer ellipses can be used as outlier cut-off. For the upper right plot, classical estimators were used to compute the Mahalanobis distances (and to construct the ellipses), while for the lower right plot robust estimators resulting from MCD were plugged in. While in the classical case only IRL would appear as outlier, the robust version additionally flags BG, GR, and

R as multivariate outliers. It is then interesting to see these ellipses in the ternary diagrams on the left-hand side. Due to the geometry on the simplex, the ellipses appear distorted, and the identified outliers are more difficult to see. It is clear that an outlier detection in this original space would work differently and thus lead to wrong conclusions. However the ternary diagram can be used for an interpretation of the outliers: IRL has higher relative expenditures on alcohol and tobacco, while those for food and recreation are comparable. GR, and even more extreme BG and R devote much more money to food than to recreation, but at the same time their proportional expenditures on alcohol and tobacco are higher compared to other countries. Of course, one should not conclude now that people in these countries drink and smoke much more than in other countries, without comparing the prices for alcohol and tobacco.

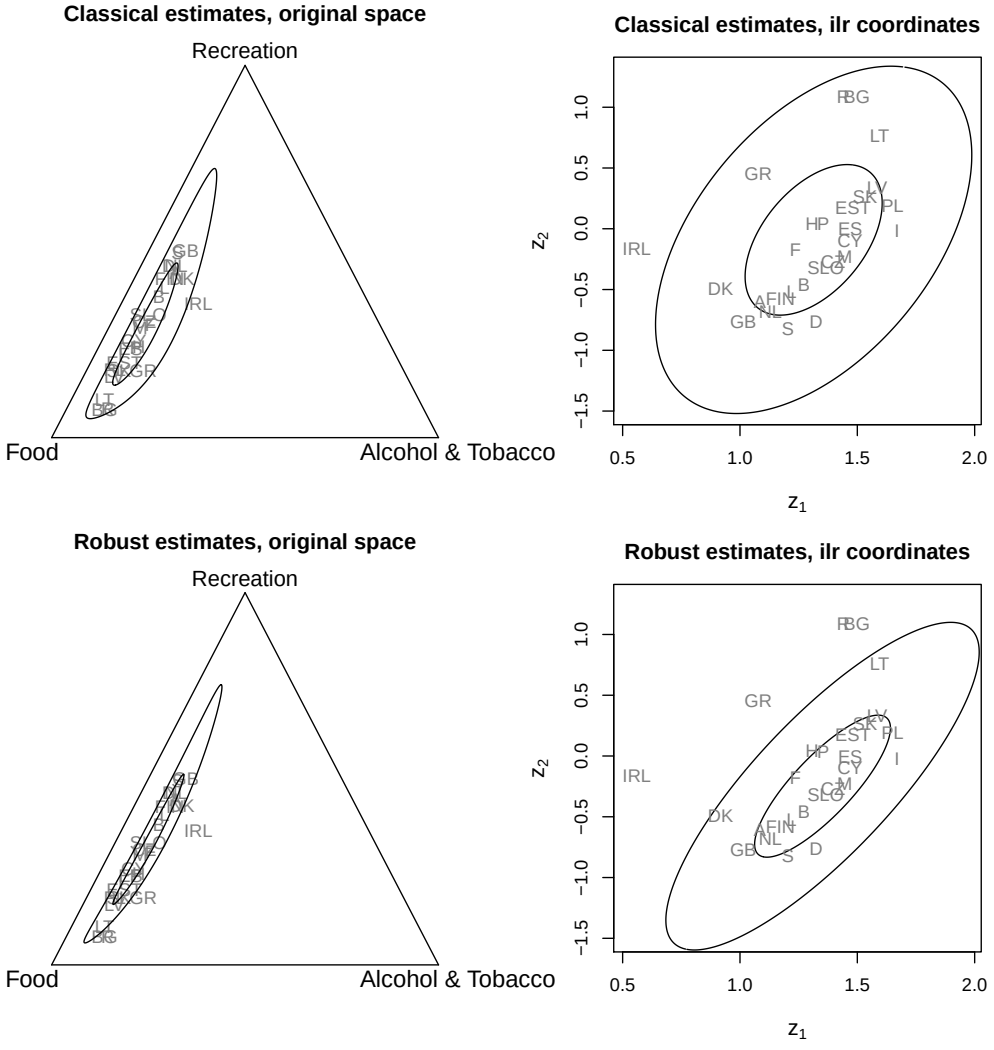

Figure 3: Visualization of data reflecting certain expenditures in the ternary diagram (left) and in ilr coordinates (right), using classical (top) and robust (bottom) estimators for outlier detection.

The distance-distance plot of [11], comparing the Mahalanobis distances based on classical and robust estimates, respectively, can be extended for compositional data. Here it can be of interest if the transformation of the data to the Euclidean space is relevant to outlier detection, or whether the same results would appear without any transformation. This is realized for the expenditures data example from Figure 3, and the resulting plots are shown in Figure 4. The left-hand plot is based on classical estimates, while for the right-hand plot robust estimates (MCD) were used. The horizontal axes represent the Mahalanobis distances using the untransformed original data, and the vertical axes are for the ilr-transformed data. The latter information was already shown in Figure 3. The lines indicate the cut-off values. The conclusions based on classical estimated would be the same, namely that IRL is an outlier. In the robust case one would come to quite different answers.

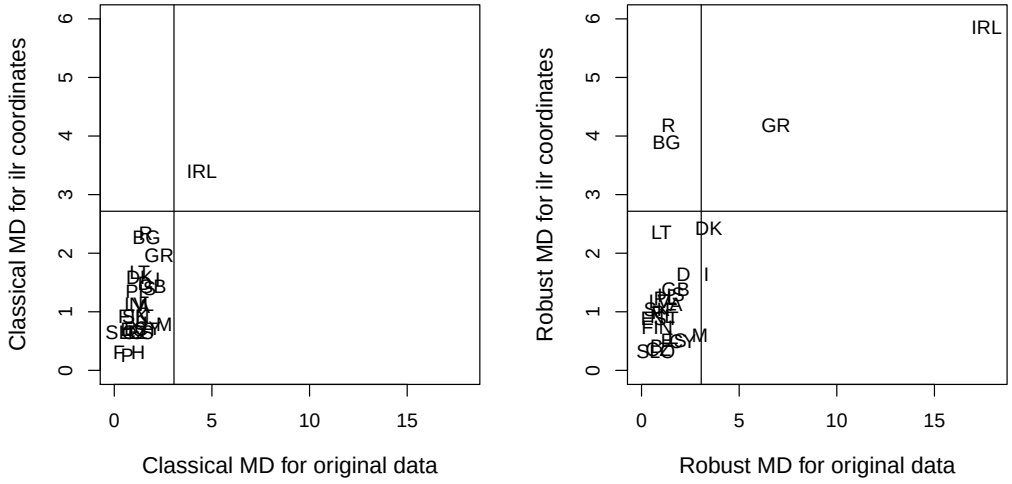

Figure 4: Distance-distance plots comparing outlier detection for the original and ilr-transformed data, based on classical (left) and robust (right) estimates.

3 Interpreting multivariate outliers

Multivariate outlier detection becomes really relevant as soon as the complete data information can no longer be visualized. For compositional data this refers to data with $D > 3$ compositional parts. The distance-distance plot can then still be presented because Mahalanobis distances prepare the information in a univariate way.

Figure 5 shows the distance-distance plot in the same way as Figure 4, but here the complete expenditures data set is used, with all twelve compositional parts. When using classical methods (left), no outliers are detected. In the robust case (right), now different countries than before are identified as outliers. Note that outlier detection based on the untransformed data would give a different answer. The outliers BG, D, P, and PL thus can only be identified following an appropriate data transformation.

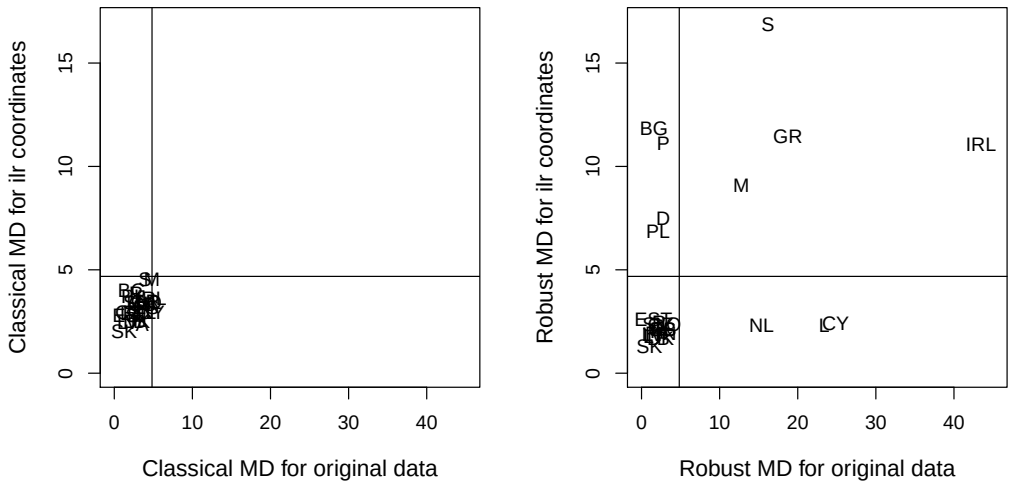

Figure 5: Distance-distance plots for the complete expenditures data set. Outlier detection is compared for the original and ilr-transformed data, based on classical (left) and robust (right) estimates.

The interpretation of multivariate outliers is not straightforward for compositional data. It is not possible to go back to the single variables, because the values itself are not meaningful, only the ratios between the parts. One possibility to interpret the outliers is to construct the *compositional biplot* as proposed in [2]. It presents observations and variables in one plot, just in the style of the original biplot introduced by [7], but it requires a slightly different interpretation. In order to provide the consistency to robust outlier detection, the biplot needs to be based on robust principal component analysis, see [6]. However, here we want to find a way to present the basic underlying information, namely the log-ratios between the compositional parts. Figure 6 shows this information for all ratios of the variables "Alcohol & Tobacco" (left) and "Education" (right), respectively, to the remaining compositional parts (see horizontal axes). Presented are the Z-scores of the countries for the specific log-ratios, where the median was used for centering, and the MAD (median absolute deviation) for scaling. Z-scores below -2 or above $+2$ are candidates for causing multivariate outliers. For example, Figure 6 (right) shows that S is outlying with respect to the complete part "Education". Indeed, the relative expenses of people in Sweden on education are very small compared to other countries.

References

[1] Aitchison J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

[2] Aitchison J., Greenacre M. (2002). Biplots of Compositional Data. *Applied Statistics*. Vol. **51**, pp. 375-392.

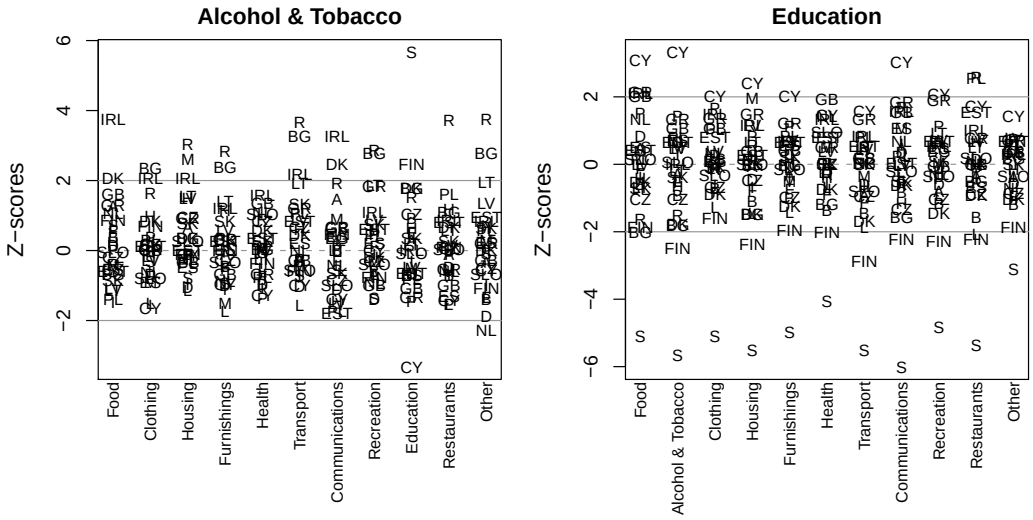

Figure 6: Robust Z-scores of the log-ratios "Alcohol & Tobacco" (left) and "Education" (right) to the remaining parts.

[3] Egozcue J.J., Pawlowsky-Glahn V., Mateu-Figueras G. (2002). Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*. Vol. **35**, pp. 279-300.

[4] Filzmoser P., Garrett R.G., Reimann C. (2005). Multivariate Outlier Detection in Exploration Geochemistry. *Computers and Geosciences*. Vol. **31**, pp. 579-587.

[5] Filzmoser P., Hron K. (2008). Outlier Detection for Compositional Data Using Robust Methods. *Mathematical Geosciences*. Vol. **40**, pp. 233-248.

[6] Filzmoser P., Hron K., Reimann C. (2009). Principal Component Analysis for Compositional Data with Outliers. *Environmetrics*. Vol. **20**, pp. 621-632.

[7] Gabriel K.R. (1971). The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika*. Vol. **58**, pp. 453-467.

[8] Mahalanobis P.C. (1936). On the Generalised Distance in Statistics. *Proceedings of the National Institute of Science of India*, A2, pp. 49-55.

[9] Maronna R., Martin D., Yohai V. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons Canada Ltd., Toronto, ON.

[10] Reimann C., Filzmoser P., Garrett R.G., Dutter R. (2008). *Statistical Data Analysis Explained. Applied Environmental Statistics with R*. John Wiley, Chichester.

[11] Rousseeuw P., Van Driessen K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*. Vol. **41**, pp. 212-223.