Uniwersytet
Ekonomiczny
w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 247 · 2015

Informatyka i Ekonometria 4

Justyna Majewska

University of Economics in Katowice
Faculty of Informatics and Communication
Department of Demography and Economic Statistics
justyna.majewska@ue.katowice.pl

IDENTIFICATION OF MULTIVARIATE OUTLIERS – PROBLEMS AND CHALLENGES OF VISUALIZATION METHODS

Summary: The identification of outliers is often thought of as a means to eliminate observations from a data set to avoid disturbance in further analyses. But outliers may as well be the interesting observations in themselves, because they can give us hints about certain structures in the data or about special events during the sampling period. Therefore, appropriate methods for the detection of outliers are needed. Literature is abundant with procedures for detection and testing of single outliers in sample data. The difficulty of detection increases with the number of outliers and the dimension of the data because the outliers can be extreme in any growing number of directions. An overview of multivariate outlier detection methods that are provided in this study because of its growing importance in a wide variety of practical situations. We focus on methods that can be visually presented.

Keywords: outlier, Mahalanobis distance, masking, swamping effect.

Introduction

An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the applied detection method [Ben-Gal, 2005]. In the literature many authors have proposed many definitions for an outlier with seemingly no universally accepted definition. The basic definition of an outlying observation is a data point or points that do not fit the model of the rest of the data. Hawkins [1980] defines an outlier "as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". Barnet and Lewis [1994] indicate that "an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs". Rousseeuw and von Zomeren [1990] stated that outliers are "an empirical reality but their exact definition is as elusive as the exact definition of a cluster". They argue that outliers "are observations that deviate from the model suggested by the majority of the point cloud, where the central model is a multivariate normal" [Rousseeuw and van Zomeren, 1990]. Booth et al. [1989] pointed out the difficulty of defining a multivariate outlier when they referred to a statistical outlier as a nonrepresentative observation whose "position may not be extreme enough on the basis of a single variable to demonstrate its outlying characteristics. However, the combined effects of several variables could be substantial enough to justify categorizing" it as an outlier. However, such words as *appear to deviate, deviates so much* imply some kind of subjectivity.

In univariate data, the identification of outlier seems relatively simple to carry out. A simple plot of the data, such as scatter plot, stem-and-leaf plot, QQ-plot etc., can often reveal which points are outliers. Identification of multivariate outliers is definitely more complex than in the univariate case. Practically, identification of outliers are hard to detect when dimension of p exceeds two [Rousseeuw and van Zomeren, 1990]. Some of the procedures for identifying multivariate outliers have been adapted from the univariate methods. And unfortunately, "many of the standard multivariate methods are derived under the assumption of normality and the presence of outliers will strongly affect inferences made from normal-based procedures" [Schwager and Margolin, 1982]. Various concepts for multivariate outlier detection methods exist in the literature [e.g. Barnett and Lewis, 1994; Rocke and Woodruff, 1996; Peña and Prieto, 2001].

1. Multivariate outliers identification

Multivariate outliers pose bigger challenges than univariate data as simple visual detection of multivariate outliers is virtually impossible. In most cases multivariable observations cannot be detected as outliers when each variable is considered independently. A simple example can be seen in Figure 1, which presents data points having two measures on a two-dimensional space and impossibility of using classical boxplot method to detect outliers in two-dimension space. The lower right observations (seen in the 2D space) are clearly a multivariate outliers but not a univariate. Thus, the test for outliers must take into account the relationships between the two variables, which in this case appear abnormal.

Outlier detection is possible only when multivariate analysis is performed, and the interactions among different variables are compared within the class of data.

Data sets with multiple outliers or clusters of outliers are subject to *masking* and *swamping effects*. Although not mathematically rigorous, the following definitions from Acuna and Rodriguez [2004] give an intuitive understanding for these effects:

Masking effect: it is said that one outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier the second instance is emerged as an outlier. Masking occurs when a cluster of outlying observations skews the mean and the covariance estimates toward it, and the resulting distance of the outlying point from the mean is small.

Swamping effect: it is said that one outlier swamps a second observation, if the latter can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier the second observation becomes a non-outlying observation. Swamping occurs when a group of outlying instances skews the mean and the covariance estimates toward it and away from other non-outlying instances, and the resulting distance from these instances to the mean is large, making them look like outliers.

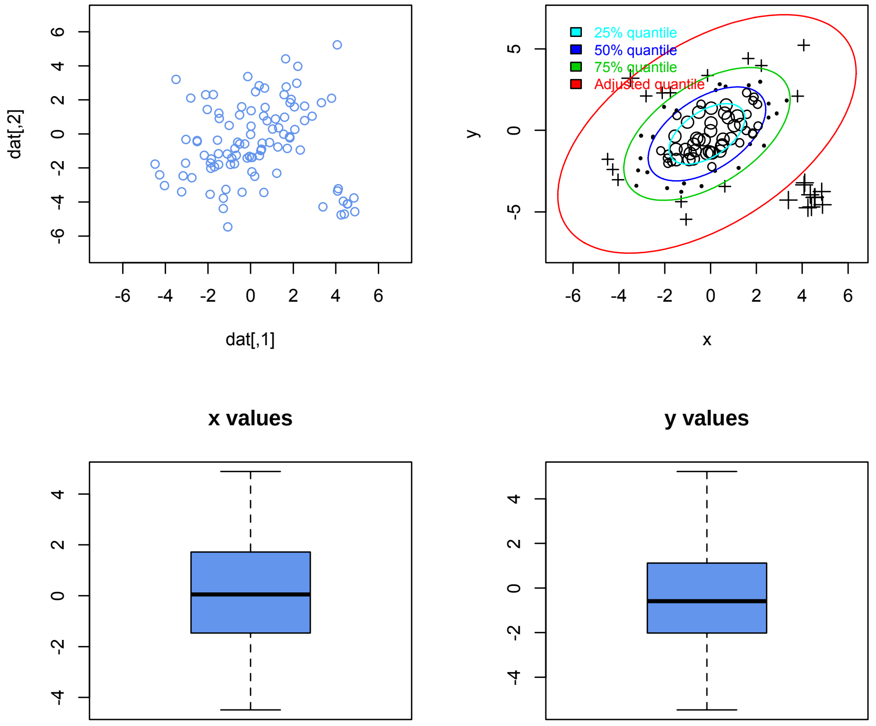

Fig. 1. An attempt to identify outliers from the set of simulated 100 observations (from N(100,5) distribution) in 2D with boxplot method and scatterplots (one of them with four ellipsoids where Mahalanobis distances are constant; these constant values correspond to the 0.25, 0.50, 0.75 and adjusted (see section 2.1) quantiles of the chi-square distribution)

Source: Own calculations in *R*.

A single step procedure with low masking and swamping is given in Iglewics and Martinez [1982].

The phenomenon of outlier masking and swamping also argues for the use of outlier resistant identification methods for detecting multivariate outliers. The degree of masking is measured in terms of an increase in Type II error, or false negatives, since observations that are truly outlying are classified as part of the uncontaminated population of data. Swamping refers to the increase in Type I error caused by outliers.

Becker and Gather [1999] developed the masking breakdown point[1] of outlier identification method that specifies the smallest fraction of outliers in a sample that can induce the masking affect. Becker and Gather proved that the masking breakdown point for an outlier detection method that uses a mean and covariance estimator is bounded by the breakdown points of these two estimators. Further, if the two estimators have the same breakdown point, then the masking breakdown point of the detector is equal to the estimator breakdown point.

2. Visualization of robust distance based methods

Distance-based methods are usually based on local distance measures and are capable of handling large databases [among others, Breunig et al., 2000].

2.1. The Mahalanobis robust distance

The Mahalanobis distance is a well-known criterion which depends on estimated parameters of the multivariate distribution. Given n observations from a p-dimensional dataset, denote the sample mean vector by $\boldsymbol{\mu}$ and the sample covariance matrix by $\mathbf{V}$. The Mahalanobis distance (MD) for each multivariate data point i, $i = 1,...,n$, is denoted by M_i and given by:

[1] Breakdown point is an important measure that is used to describe the resistance of robust estimators in the presence of outliers. Following Hodges [1967] and Hampel [1968, 1971], breakdown point of an estimator is the fraction of arbitrary contaminating observations that can be presented in a sample before the value of the estimator can become arbitrarily large. Lopuhaä and Rousseeuw [1991] have presented more formal definitions of the breakdown point for location and covariance estimators.

$$\mathrm{M}_i = \left(\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)^{1/2}$$

Accordingly, those observations with a large MD can be indicated as outliers [Aguinis et al., 2013]. For normally distributed data the Mahalanobis distance is approximately chi-square distributed with p degrees of freedom. Potential multivariate outliers *xi* will typically have large values *Mi*, and in this situation a comparison with the χ_p^2 distribution can be made.

Masking and swamping effects play an important rule in the adequacy of the MD as a criterion for outlier detection. Namely, masking effects might decrease the MD of an outlier. This might happen, for example, when a small cluster of outliers attracts $\boldsymbol{\mu}$ and inflate $\mathbf{V}$ towards its direction. On the other hand, swamping effects might increase the Mahalanobis distance of non-outlying observations. For example, when a small cluster of outliers attracts $\boldsymbol{\mu}$ and inflate $\mathbf{V}$ away from the pattern of the majority of the observations [see Penny and Jolliffe, 2001].

Due to these problems, robust estimators been used and substituted in the distance formula which yield robust distance. The use of robust estimates of the multidimensional distribution parameters can often improve the performance of the detection procedures in presence of outliers. Hadi [1992] addresses this problem and proposes to replace the mean vector by a vector of variable medians and to compute the covariance matrix for the subset of those observations with the smallest MD. A modified version of Hadi’s procedure was presented in Penny and Jolliffe [2001]. Caussinus and Roiz [1990] proposed a robust estimate for the covariance matrix, which is based on weighted observations according to their distance from the center. The authors also propose a method for a low dimensional projections of the dataset. They use the Generalized Principle Component Analysis to reveal those dimensions which display outliers. Other robust estimators such as M-estimator, S-estimator, MM-estimator, MVE, MCD and Fast-MCD (FMCD) estimator have been proven to identify outliers better than classical estimator. Among the robust estimators, FMCD has been shown to be the best estimator compare to other robust estimators [Rousseeuw, 1985; Rousseeuw and Leroy, 1987; Acuna and Rodriguez, 2004].

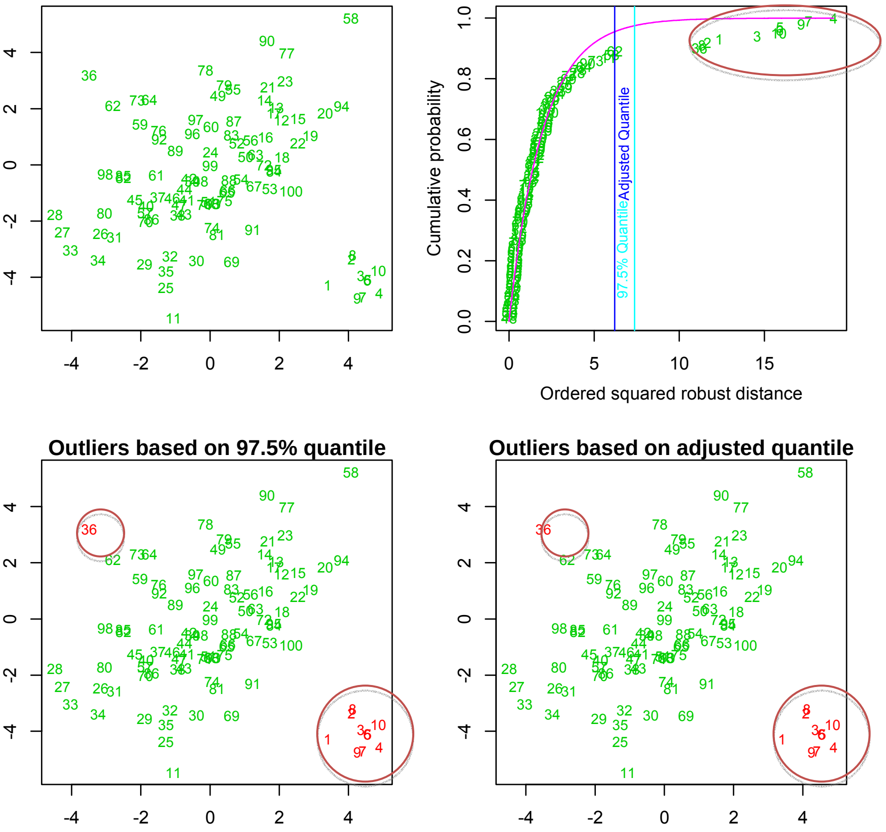

Fig. 2. The ordered squared robust Mahalanobis distances of the observations against the empirical distribution function of the squared the Mahalanobis distance. In addition the distribution function of *chisq_p^2* is plotted as well as two vertical lines corresponding specified in the argument list (default is 0.975) and the so-called adjusted quantile. Three additional graphics are created (the first showing the data, the second showing the outliers detected by the specified quantile of the *chisq_p^2* distribution and the third showing these detected outliers by the adjusted quantile)

Source: Figures made with mvoutlier package in *R*.

The Figure 2 presents the ordered squared robust Mahalanobis distances of the observations against the empirical distribution function of the squared the Mahalanobis distance. The outliers are detected by the specified quantile of the χ_p^2 distribution and by the adjusted quantile.

2.2. MVE and MCD methods

Rousseeuw [1985] studied whether it is at all possible to combine a high breakdown point with affine equivariance for multivariate estimation. It is found that Minimum Volume Ellipsoid estimator (MVE) and Minimum Covariance Determinant (MCD) estimator both are affine equivariant estimators[2] with a high breakdown. The mean of MVE was defined as center of the minimal volume ellipsoid covering at least h points of X. While the mean of MCD was defined as mean of the h points of X for which the determinant of the covariance matrix is minimal. In addition, Rousseeuw [1985] also found that 50% breakdown estimators MVE and MCD have low asymptotic efficiencies.

Rousseeuw and van Zomeren [1990] proposed computation of distances based on very robust estimates of location and covariance. MVE estimator for mean and covariance are used to compute robust distance. They applied it to various data sets and found that robust distance can identify outliers more efficiently compared to MD and also found to be useful to identify outliers in multivariate data.

Butler et al. [1993] showed that the MCD has better statistical efficiency than the MVE since the MCD is asymptotically normal. Additionally, Davies, showed that the MVE has a lower convergence rate than the MCD. According to Rousseeuw and van Driessen [1999], theoretical findings combined with the need for accurate estimators for use in outlier detection schemes, the MCD began to gain favor over the MVE as the preferred robust estimator for outlier detection. The main drawback to using the MCD, however, is the high computational complexity involved with searching the space of half-samples of a dataset to find the covariance matrix with minimum determinant.

Fast-MCD (FMCD) was developed due to the existing algorithms that is limited to a few hundred objects in few dimensions [Rousseeuw and Katrien, 1999]. As a result, FMCD give accurate results for large datasets and exact MCD for small datasets [Rousseeuw and Katrien, 1999]. The main drawback of MCD strategy for robust distance detection is their large computational burden that limits their utility relative to large-scale problems. The result of identification method is presented in Figure 3.

[2] If an estimator is affine equivariant, stretching or rotating the data will not affect the estimator. Dropping this requirement greatly increases the number of available estimators, and in many cases, non-affine equivariant estimators have superior performance to affine equivariant estimators.

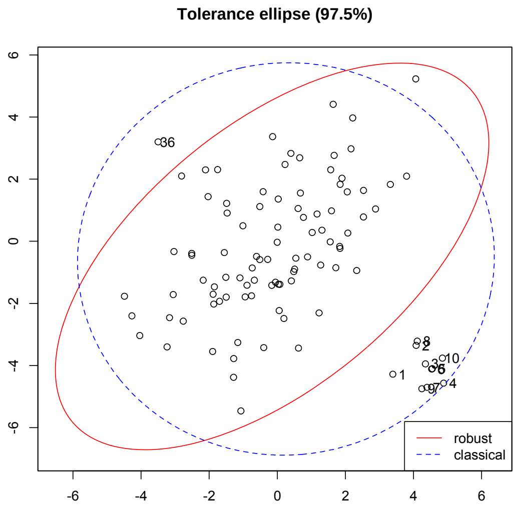

Fig. 3. Outliers identification by robust MCD with tolerance ellipsoid (0,975)

Source: Figure made with rrcov package in *R*.

3. Non-traditional methods based on robust PCA

A common limitation with all robust distance-based outlier detection methods is the requirement to find a subset of outlier-free data from which robust estimates of the mean vector and covariance matrix can be obtained. Unfortunately, there is no existing method that can find an outlier-free subset with 100% certainty. Researchers have proposed alternative non-traditional outlier detection methods that attempt to avoid robust Mahalanobis distances altogether. In the following paragraphs, the significant non-traditional and most interesting outlier detection methods found in the literature are outlined.

3.1. Method for outlier identification in high dimensions

In this subsection we use fast algorithm for identifying multivariate outliers in high-dimensional and/or large datasets [Filzmoser, Maronna, and Werner, 2007]. Based on the robustly sphered data, semi-robust principal components are computed which are needed for determining distances for each observation. Separate weights for location and scatter outliers are computed based on these distances. The combined weights are used for outlier identification. Figure 4 present: vector with final weights for each observation (weight 0 indicates potential multivariate outliers), vector with final weights for each observation (small values indicate potential multivariate outliers), vector with weights for each observation (small values indicate potential location outliers), vector with weights for each observation (small values indicate potential scatter outliers).

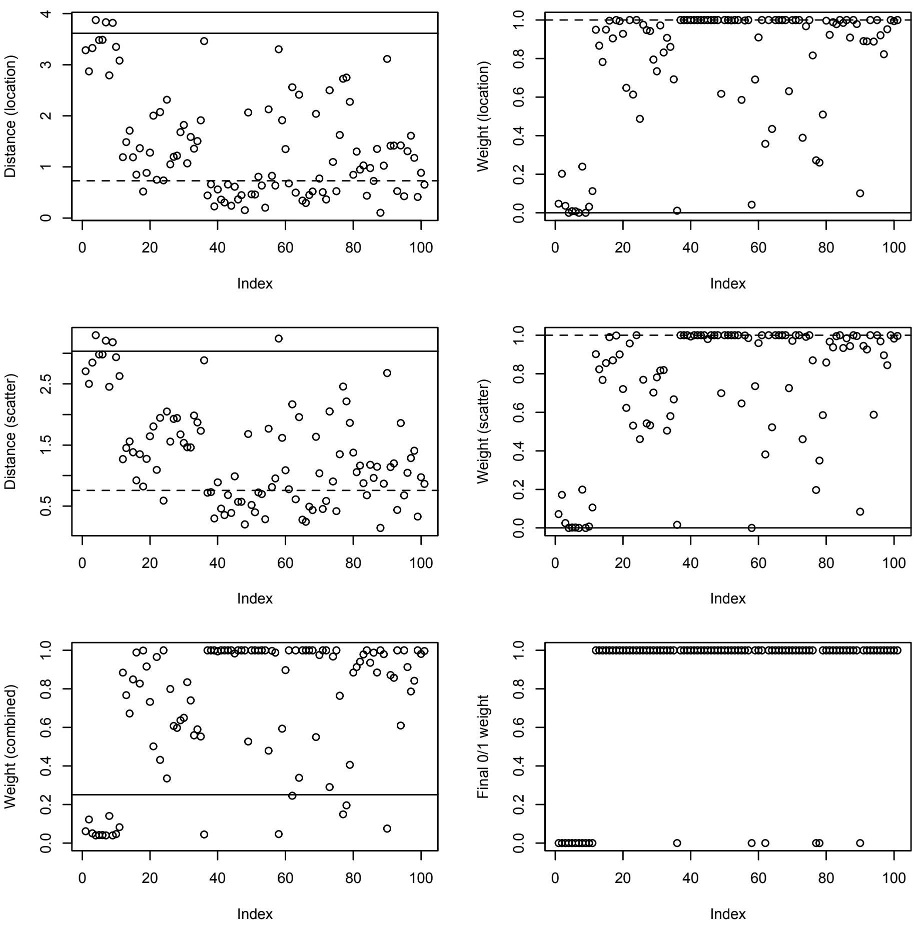

Fig. 4. Results of outliers' identification method of Filzmoser, Maronna, and Werner [2007]

Source: Figures made with mvoutlier package in *R*.

This method detected more outliers than MCD method, see Figure 5.

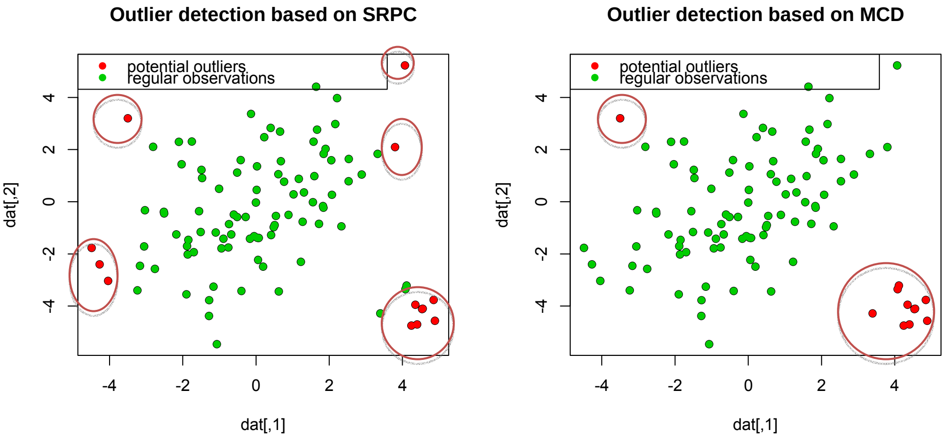

Fig. 5. Identified outliers using Filzmoser et al. [2008] method

Source: Figures made with mvoutlier package in *R*.

3.2. Outliers identification method based on functional approach

Some of data, for example mortality data, can be treat as set of curves, which are the realizations on the functional space. By visualizing these curves we can identify outliers in the observed curves using functional equivalents of boxplots and bagplots. Hyndman and Shang [2010] proposed the functional bagplot and a functional boxplot in order to visualize functional data and to detect any outliers present.

Suppose we have a set of curves $\{y_i(x)\}$, $i = 1,...,n$, which are realizations on the functional space *I*. After visualizing these curves for large *n* using functional equivalents of boxplots and bagplots we want to identify outliers in the observed curves. In this concept the notion of ordering a set of curves is crucial. This methods use approach to ordering obtained using a principal component decomposition of the set of observed curves. If we let:

$$y_i(x) = \mu(x) + \sum_{k=1}^{n-1} z_{i,k} \varphi_k(x)$$

where $\{\phi_k(x)\}$ represents the eigenfunctions, then we can use an ordering method from multivariate analysis based on the principal components scores $\{z_{i,k}\}$. The simplest procedure is to consider only the first two scores, $z_i = (z_{i,1}, z_{i,2})$. Then an ordering of the curves is defined using the ordering of $z_i = (z_{i,1}, z_{i,2})$. For example, bivariate depth can be used [Rousseeuw et al., 1999]. Alternatively, the value of the kernel bivariate density estimate at z_i can be used to define an ordering.

Age-specific mortality rates are very good example to illustrate this method. There are two major advantages in ordering via the principal component scores. The first, it leads to a natural method for defining visualization methods such as functional bagplots and functional boxplots. The second, it seems to be better able to identify outliers in real data. Outliers will usually be more visible in the principal component space than the original (functional) space [Filzmoser et al., 2008]. Thus finding outliers in the principal component scores does no worse than searching for them in the original space. Often, it is the case that the first two principal component scores[3] suffice to convey the main modes of variation.

Because principal component decomposition is itself non-resistant to outliers, Hyndman and Shang [2010] applied a functional version of Croux and Ruiz-Gazen's [2005] robust principal component analysis, which uses a projection pursuit technique. This method was described and used in Hyndman and Ullah [2007].

The *functional bagplot* is based on the bivariate bagplot of Rousseeuw et al. [1999] applied to the first two (robust) principal component scores. The bagplot is constructed on the basis of the halfspace location depth denoted by $d(\theta,\mathbf{z})$ of some point $\theta \in R^2$ relative to the bivariate data cloud $\{\mathbf{z}_i;\ i = 1,...,n\}$. The depth region D_k is the set of all θ with $d(\theta,\mathbf{z}) \geq k$. Since the depth measurements are convex polygons, we have $D_{k+1} \subset D_k$. For a fixed center, the regions grow as the radius increases. Thus, the data points are ranked according to their depth. The bivariate bagplot displays the median point (the deepest location), along with the selected percentages of convex hulls. Any point beyond the highest percentage of the convex hulls is considered as an outlier. Each point in the scores bagplot corresponds to a curve in the functional bagplot. The functional bagplot also displays the median curve which is the deepest location, the 95% confidence intervals for the median, and the 50% and 95% of surrounding curves ranking by depth. Any curve beyond the 95% convex hull is flagged as a functional outlier (see Figure 6).

The *functional highest density region (HDR) boxplot* is based on the bivariate HDR boxplot of Hyndman [1996] applied to the first two (robust) principal component scores. The HDR boxplot is constructed using the Parzen-Rosenblatt bivariate kernel density estimate $\hat{f}(\mathbf{w};a,b)$. For a bivariate random sample $\{\mathbf{z}_i;\ i = 1,...,n\}$ drawn from a density f, the product kernel density estimate is defined by Scott [1992]:

[3] Hyndamn and Shang [2008] found empirically that the first two principal component scores are adequate for outlier identification.

$$\hat{f}(\mathbf{w};a,b)=\frac{1}{nab}\sum_{i=1}^{n}K\left(\frac{w_1-z_{i,1}}{a}\right)K\left(\frac{w_2-z_{i,2}}{b}\right)$$

where $\mathbf{w} = (w_1, w_a)^T$, K is a symmetric univariate kernel function such that $\int \mathrm{K(u)du}=1$ and *(a, b)* is a bivariate bandwidth parameter such that $a > 0$, $b > 0$, $a \to 0$ and $b \to 0$ as $n \to \infty$. The contribution of data point $\mathbf{z}_i$ to the estimate at some point w depends on how distant $\mathbf{z}_i$ and $\boldsymbol{w}$ are.

A highest density region is defined as $R_\alpha = \left\{z : \hat{f}(z;a,b) \geq f_\alpha\right\}$ where f_α is such that $\int_{R_\alpha} \hat{f}(z;a,b)dz = 1-\alpha$. That is, it is the region with probability coverage $1 - \alpha$ where every point within the region has higher density estimate than every point outside the region.

The advantage of ranking by the HDR is its ability to show multimodality in the bivariate data. The HDR boxplot displays the mode, defined as $\sup_z f(\mathbf{z};a,b)$, along with the 50% HDR and the 95% HDR. All points not included in the 95% HDR are shown as outliers (see Figure 7). The functional HDR boxplot is a one-to-one mapping of the scores HDR bivariate boxplot.

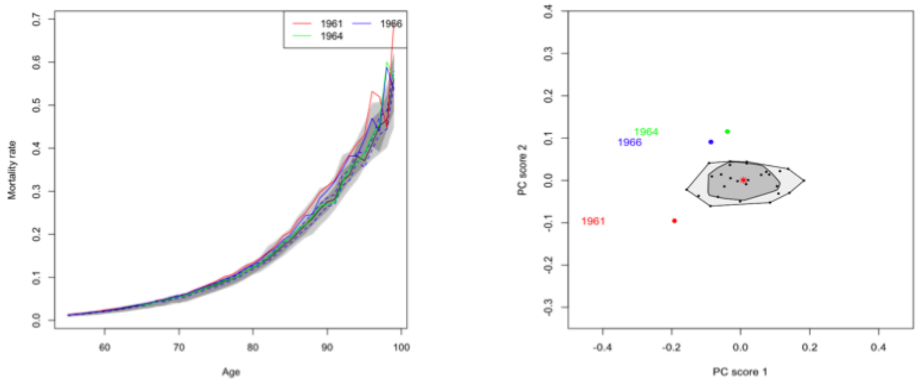

Fig. 6. The functional and bivariate bagplot [UK 1961-1990, male, age: 55-100]

Source: Figures made with the rainbow package in *R*.

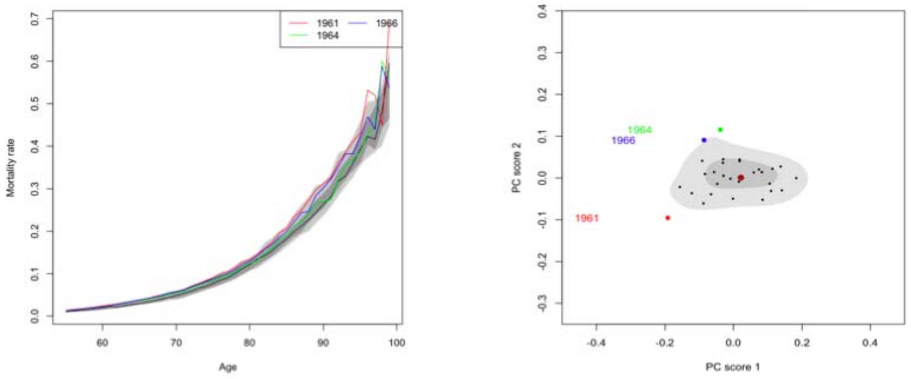

Fig. 7. The functional and bivariate HDR boxplot [UK 1961-1990, male, age: 55-100]

Source: Figures made with the rainbow package in *R*.

Both of methods identified the same outliers. Of the two new methods, Hyndman and Shang [2010] prefer the functional HDR boxplot as it also provides an additional advantage in that it can identify unusual “inliers” that full in sparse regions of the sample space.

Conclusions

The procedure of outlier identification would not be comprehensive without displaying the results graphically. In this paper we review most interesting approaches to outliers’ detection.

It is known that using robust (high-breakdown) estimators for location and covariance is also very effective in finding multivariate outliers. In particular, examining the structure of outliers found by high-breakdown estimators is a diagnostic effort that is often somewhat neglected. The distance-projection plot has the advantage of being quite easy to interpret, but here is always a chance that the “outlier-free” sample contains some outliers.

Any single bivariate plot can not reveal all multivariate structure, so different bivariate plots should be made providing complementary information. Therefore, we recommend using of different plots to know better structure, shape and dependencies between the data. Every described in this paper method is presented with artificial example/real data example.

References

Acuna E., Rodriguez C.A. (2004), *Meta Analysis Study of Outlier Detection Methods in Classification*, Technical paper, University of Puerto Rico at Mayaguez, Proceedings IPSI 2004, Venice.

Aguinis H., Gottfredson R.K., Joo H. (2013), *Best-Practice Recommendations for Defining, Identifying, and Handling Outliers*, "Organizational Research Methods", p. 270-301.

Barnett V., Lewis T. (1994), *Outliers in Statistical Data* (2nd Edition), John Wiley and Sons.

Becker C., Gather U. (1999), *The Masking Breakdown Point of Multivariate Outlier Identification Rules*, "Journal of the American Statistical Association" 94, p. 947-955.

Ben-Gal I. (2005), *Outlier Detection* [in:] O. Maimon, L. Rockach (eds.), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kluwer Academic Publishers.

Breunig M.M., Kriegel H.P., Ng R.T., Sander J. (2000), *Identifying Density-based Local Outliers*, Proceedings ACMSIGMOD 2000, p. 93-104.

Booth D.E., Alam P., Ahkam S.N., Osyk B. (1989), *A Robust Multivariate Procedure for the Identification of Problem Savings and Loan Institutions*, "Decision Sciences", 20, p. 320-333.

Butler R.W., Davies P.L. and Jhun M. (1993), *Asymptotics for the Minimum Covariance Determinant Estimator*, "The Annals of Statistics", 21, p. 1385-1400.

Caussinus H., Roiz A. (1990), *Interesting Projections of Multidimensional Data by Means of Generalized Component Analysis*, COMPSTAT90, Physica-Verlag, Heidelberg, p. 121-126.

Croux C., Ruiz-Gazen A. (2005), *High Breakdown Estimators for Principal Components: The Projection-pursuit Approach Revisited*, "Journal of Multivariate Analysis", 95(1), p. 206-226.

Fawcett T., Provost F. (1997), *Adaptive Fraud Detection*, "Data-mining and Knowledge Discovery", 1(3), p. 291-316.

Filzmoser P., Maronna R., Werner M. (2008), *Outlier Identification in High Dimensions*, "Computational Statistics and Data Analysis", 52, p. 1694-1711.

Hadi A.S. (1992), *Identifying Multiple Outliers in Multivariate Data*, "Journal of the Royal Statistical Society", Series B, 54, p. 761-771.

Hawkins D.M. (1980), *Identification of Outliers*, Chapman and Hall, London.

Human Mortality Database (2015), University of California, Berkeley (USA), and Max Planck Institute for Demographical Research (Germany), viewed 15/09/07, available online at: www.mortality.org.

Hyndman R.J., Shang H.L. (2008), *Rainbow Plots, Bagplots, and Boxplots for Functional Data*, "Journal of Computational and Graphical Statistics" 19(1), p. 29-45.

Hyndman R.J., Ullah S. (2007), *Robust Forecasting of Mortality and Fertility Rates: A Functional Data Approach*, "Computational Statistics and Data Analysis", 51, p. 4942-4956.

Iglewics B., Martinez J. (1982), *Outlier Detection Using Robust Measures of Scale*, "Journal of Statistical Computation and Simulation", 15, p. 285-293.

Peña D., Prieto F.J. (2001), *Multivariate Outlier Detection and Robust Covariance Matrix Estimation*, "Technometrics", 43, p. 286-300.

Penny K.I., Jolliffe I.T. (2001), *A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Safety Data*, "The Statistician", 50(3), p. 295-308.

Rocke D.M., Woodruff D.L. (1996), *Identification of Outliers in Multivariate Data*, "Journal of the American Statistical Association" 91, p. 1047-1061.

Rousseeuw P. (1985), *Multivariate Estimation with High Breakdown Point* [in:] W. Grossmann et al. (eds.), "Mathematical Statistics and Applications", Vol. B, p. 283-297.

Rousseeuw P.J., Driessen K.A. van (1999), *Fast Algorithm for the Minimum Covariance Determinant Estimator*, "Technometrics", 41, p. 212-223.

Rousseeuw P.J., Katrien V.D. (1999), *A Fast Algorithm for the Minimum Covariance Determinant Estimator*, "Technometrics", 41(3), p. 212-223.

Rousseeuw P., Leroy A. (1987), *Robust Regression and Outlier Detection*, Wiley Series in Probability and Statistics.

Rousseeuw P., Ruts I., Tukey J. (1999), *The Bagplot: A Bivariate Boxplot*, "The American Statistician", 53(4), p. 382-387.

Rousseeuw P.J., Zomeren B.C. van (1990), *Unmasking Multivariate Outliers and Leverage Points*, "Journal of the American Statistical Association", 85(411), p. 633-651.

Schwager S.J., Margolin B.H. (1982), *Detection of Multivariate Normal Outliers*, "Annals of Statistics", 10, p. 943-95.

IDENTYFIKACJA WIELOWYMIAROWYCH OBSERWACJI ODSTAJĄCYCH – PROBLEMY I WYZWANIA METOD WIZUALIZACYJNYCH

Streszczenie: Proces identyfikacji obserwacji odstających jest często rozważany jako wstęp do eliminacji obserwacji nietypowych ze zbiorów danych w celu uniknięcia jakichkolwiek problemów w dalszej analizie danych. Tymczasem obserwacje nietypowe dostarczają niejednokrotnie istotnych informacji o strukturze danych lub wyjątkowych zdarzeniach podczas badanego okresu. Dlatego potrzebne są właściwe metody identyfikacji tychże obserwacji. Literatura jest bogata w metody wykrywania obserwacji nietypowych w jednowymiarowych przypadkach. W wielowymiarowej przestrzeni proces ten znacznie się komplikuje. W artykule prezentujemy wybrane metody wizualizacyjne wykrywania wielowymiarowych obserwacji nietypowych.

Słowa kluczowe: obserwacja odstająca, odległość Mahalanobisa, efekt maskowania, efekt zanurzania, wizualizacja.