Journal of Experimental Social Psychology 74 (2018) 150–156

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance

Christophe Leys[a,*], Olivier Klein[a], Yves Dominicy[b,1], Christophe Ley[c]

[a] Université libre de Bruxelles, Centre de Recherche en Psychologie Sociale et Interculturelle, Belgium
[b] Université libre de Bruxelles, Solvay Brussels School of Economics and Management, ECARES, Belgium
[c] Ghent University, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium

ABSTRACT

A look at the psychology literature reveals that researchers still seem to encounter difficulties in coping with multivariate outliers. Multivariate outliers can severely distort the estimation of population parameters. Detecting multivariate outliers is mainly disregarded or done by using the basic Mahalanobis distance. However, that indicator uses the multivariate sample mean and covariance matrix that are particularly sensitive to outliers. Hence, this method is problematic. We highlight the disadvantages of the basic Mahalanobis distance and argue instead in favor of a robust Mahalanobis distance. In particular, we present a variant based on the Minimum Covariance Determinant, a more robust procedure that is easy to implement. Using Monte Carlo simulations of bivariate sample distributions varying in size (ns = 20, 100, 500) and population correlation coefficient (ρ = .10, .30, .50), we demonstrate the detrimental impact of outliers on parameter estimation and show the superiority of the MCD over the Mahalanobis distance. We also make recommendations for deciding whether to include vs. exclude outliers. Finally, we provide the procedures for calculating this indicator in R and SPSS software.

1. Introduction

Detecting outliers is a growing concern in psychology (Leys, Ley, Klein, Bernard, & Licata, 2013; Meade & Craig, 2012; Simmons, Nelson, & Simonsohn, 2011). Indeed, Simmons et al. (2011) showed how significant results could easily turn out to be false positives if outliers are dealt with only flexibly and post-hoc. Leys et al. (2013) showed that researchers took insufficient care to detect outliers, using either inappropriate methods or failing to report crucial information about the detection process. They provide a robust method to analyze univariate outliers. However, we argue that this problem is equally relevant for multivariate outliers. The aim of this paper is to underline the importance of such outliers and to propose a robust method of detection.

Quoting Barnett and Lewis (1994): "The study of outliers is as important for multivariate data as it is for univariate samples" (p. 25). In some respect, one can say that a correct approach is even more important for multivariate data sets (Meade & Craig, 2012), as (i) nowadays more and more observations are multi-dimensional (e.g., when several measurements are made on each individual) and (ii) the detection of multivariate outliers is a much more difficult task. This is due to the fact that in multiple dimensions there are several directions in which a point can be outlying. Multivariate outliers are particularly relevant in the context of designs involving more than two variables as is typically the case when relying on mediational models (Hayes, 2013; Muller, Judd, & Yzerbyt, 2005), which are commonly used in experimental social psychology. Moreover, in structural equation modeling, detecting multivariate outliers is of particular interest given the influence of these outliers on fit indices and is therefore a standard practice (Kline, 2015).

In this context, four issues should be addressed. Firstly, while it is obvious that outliers may appear in measured continuous variables where all values are theoretically possible, it is not as obvious how outliers apply to experimental designs: When one of the variables is manipulated, it should be contrast-coded (cf. Judd, McClelland, & Ryan, 2017) and naturally, there won't be univariate outliers on such IV (besides coding error). It is still possible to witness multivariate outliers in combinations of values of the IV and the DV but given the limited range of the IV, it may be more efficient to detect univariate outliers in each condition separately. However, detecting multivariate outliers can

* Corresponding author at: Centre de Recherche en Psychologie Sociale et Interculturelle, Université libre de Bruxelles, Avenue Franklin Roosevelt, 50 - CP191, Office: DC10.130, 1050 Brussels, Belgium.
E-mail address: cleys@ulb.ac.be (C. Leys).
[1] Yves Dominicy thanks the Fonds National de la Recherche Scientifique, Communauté Française de Belgique, for financial support via a Mandat de Chargé de Recherche FNRS.

http://dx.doi.org/10.1016/j.jesp.2017.09.011
Received 27 March 2017; Received in revised form 25 September 2017; Accepted 26 September 2017

be valuable in experimental research when the researcher is interested in the association between two or more measured variables (e.g., a continuous moderator and a DV, see an example below) as a function of a manipulated factor. Let us consider an actual example: Burrow and Rainone (2017: study 2)[2] manipulated the number of "likes" participants received on their profile picture (three levels IV: below average, average, above average) on a social networking website after having measured their sense of "purpose in Life", which was used as a continuous moderator. The authors hypothesized that receiving more "likes" will improve self-esteem (continuous DV) for people with a low "purpose in Life". In this design, although one variable is manipulated, the moderator is not. In such a design there may be multivariate outliers involving the relation between the moderator and the DV worth detecting. Assume an outlier high in "purpose in Life" and low on "Self-Esteem". Such a value can either create a false significant result if it is in the "below average" level of likes condition or obscure a true effect if it is in the "above average" level of likes. Such a situation invites researchers to carefully scrutinize the responses of these participants in the hope of understanding the reason of this observation (e. g. coding error, systematic answers, idiosyncrasies of the participant, etc.) and to decide whether to keep or remove the outlier following our recommendations (see below). In the present case, given that the study was not preregistered, it would have been best to provide the results with and without the potentially detected outliers.

Secondly, it is also important to note that outliers on the IV and on the DV axis are not equivalent. An outlier on the DV will mainly impact the intercept whereas as an outlier on the IV will mainly affect the slope. Indeed, the slopes of the model are mainly determined by the respective leverage of each observation that is a function of the IVs only[3] (Cohen, Cohen, West, & Aiken, 2003). This implies that outlier detection is particularly crucial to perform on the IVs, as soon as there is more than one continuous IV. In the present paper, our examples use two continuous, measured, variables as IV and DV, but they could just as well use two continuous IVs.

Thirdly, outliers can be viewed as a source of bias, but they can also be considered as diagnostic tools allowing researchers to gain insights regarding the processes under study (McGuire, 1997). Consider a person who exhibits a very high level of in-group identification but a very low level of prejudice towards a specific out-group. This would count as an outlier under the theory that group identification leads to prejudice towards relevant out-groups. Detecting this person and seeking to determine why this is the case may help uncover possible moderators of the somewhat simplistic assumption that identification leads to prejudice. For example, this person might have inclusive representations of his/her in-group. One's social representation of the values of the in-group may thereby be found to be an important mediator (or moderator) of this relation. Merely disregarding this outlier or "excluding" it would have missed out the possibility of such a theoretical insight.

Lastly, and importantly, once outliers have been detected, it behooves the researcher to decides whether to include them or not in the subsequent analyses. It is now well known (Simmons et al., 2011) that such degrees of freedom can adversely impact the conclusions of subsequent statistical tests. It is therefore necessary to define a principled approach to excluding versus including outliers before data collection. We suggest to define a priori (i.e., in the context of a preregistration) an outlier management policy. There are two types of detected outliers: those that are part of the original population (false positives) and those that come from a different population (true negatives). There is no mathematical solution to discriminating these two categories. Both types of errors (keeping true negatives or removing false positives) have a cost in terms of type I and type II errors as well in the estimation of the parameters. Therefore, any general course of action (i.e., keeping vs. removing all outliers) is potentially costly. We invite researchers to first commit to a general policy of either keeping or removing outliers and to preregister this decision to the best of their abilities (cf. van't Veer & Giner-Sorolla, 2016). This decision can be informed by various factors: previous research in this area or statistical arguments. Once these outliers have been detected, and regardless of the policy being chosen, it is important to inspect them. Even if one wishes to keep them in principle, there may be cases in which removal is recommended. Here is a (not necessarily exhaustive) list of possible exclusion criteria (see also, Cohen et al., 2003):

- Values on two or more variables are logically, or physically, incompatible (e.g., weighting lbs. 100 and being 6′ 5 tall or expressing support for a positively worded proposition and for the same, negatively worded, proposition).
- Responses on control questions aimed at verifying participants' attention should also be inspected. If the respondent is detected as a multivariate outlier and has also failed such a question, it may raise suspicion as to the validity of his/her responses.
- In online surveys especially, participants may respond mechanically, not paying attention to the questions. As an alternative or supplement to control questions, the presence of systematic patterns (e.g., answering systematically at the extremes) should be checked and, if confirmed, can justify excluding outliers.
- If outliers are associated with a specific condition or stimulus, rather than being randomly distributed among conditions, this suggests that an unknown factor was confounded with the manipulation and the problem may be greater than just the outliers. In such a situation, excluding them may not be appropriate, because it would violate random allocation.

Each of these criteria should be specified in quantitative terms (e.g., starting from which discrepancy between height and weight shall a participant be considered out of range?). However, we are convinced that some reasons for excluding outliers may not be predicted a priori but still be perfectly valid. To deal with such instances, we invite researchers to address them by asking judges blind to the research hypotheses to make a decision on whether outliers that do not correspond to the a priori decision criteria should be included or not. Regardless, the most important aspect of this whole procedure is that it be specified before data collection. Given that our main scope is about detection of outliers, we refer readers further interested in the topic of coping with outliers to the papers of McClelland (2000), Cousineau and Chartier (2010) and Bakker and Wicherts (2014).

So far, we have not addressed the crucial question of how to detect outliers. Leys et al. (2013) have described a robust method for doing so in a sample of univariate observations. They have provided evidence that the commonly used rule, namely considering as outlier an observation which lies outside the interval formed by the mean plus or minus a coefficient (2, 2.5 or 3) times the standard deviation, should in fact be avoided, due to the fact that both the mean and the standard deviation themselves are heavily affected by outlying values. Instead, they proposed to use intervals formed by the median plus or minus a coefficient times the Median Absolute Deviation (MAD), as both the median and the MAD are very robust to aberrant observations, making this criterion much more sensitive. For more information, see Leys et al. (2013).

In the present paper, we propose such a robust and easy indicator for multivariate data sets, that is, observations of higher dimensions. Indeed, a survey made in the same journals as those used by Leys et al. (2013), namely the Journal of Personality and Social Psychology (JPSP) and Psychological Science (PS), revealed that few researchers seem to mind about multivariate outliers. We introduced "multivariate outlier"

[2] Note that we did not seek access to author data and that this example is only for didactic purpose and does not suggest any kind of suspicion about the results of the study.

[3] An implication of this is that in multivariate analysis of variance (MANOVA) multivariate outliers limited to the DVs are more informative with respect to the reliability of the measures than to the accuracy of prediction.

(without "s") as keywords and chose a period of 16 years (between 2000 and 2015). We found 8 hits for JPSP and 16 hits for PS. From these 24 papers, nine used the basic Mahalanobis distance (see below), five used another criterion (leverage using Student-t residuals or Cook's distance), and ten did not provide any information about the detection strategy. We then searched on PS only, with the keywords "multiple regression" for the same period and found 651 hits. This means that for over 97.5% of this type of multivariate analyses, either researchers did not search for multivariate outliers or they did not report any information about it. The 16 other teams looked for multivariate outliers, but either with a questionable method or without providing information about the method. There is, thus, a clear need for more awareness in our field about detecting outliers.

2. Multivariate outliers and the Mahalanobis distance

In a mathematical way of thinking, to detect outliers one has to take into consideration the shape/structure of the data set. Indeed, imagine a cloud of data points in $\mathbb{R}^2$ having an elliptical form, sampled for example from a bivariate normal distribution with mean $\mu = 0$ and covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

In an ellipse, some points are closer to the center than others (see Fig. 1), yet we cannot conclude that the more distant points (in terms of the classical Euclidean distance $\|x\| = \sqrt{x^T x}$, where x is a vector of variables) belong less to the sample than the closer points, as this is part of the underlying pattern of the normal distribution. Therefore, instead of the classical distance, it is recommended to use a distance taking into account the shape of the observations under scrutiny, and such a distance is the Mahalanobis distance (Mahalanobis, 1930) denoted here by d:

$$d = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)},$$

where x is a vector of variables $x = (x_1, x_2, \ldots, x_k)$, $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ is a vector of dimension k and Σ is a k x k symmetric matrix. It measures the distance from a point x to the center μ in the metric Σ, meaning that the distance depends on the shape Σ. Naturally, the values μ and Σ are unknown in practice, and hence need to be estimated. The usual estimators, obtained from a sample $X_1, \ldots, X_n$ are the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and sample covariance matrix $S = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T$.

With this basic Mahalanobis distance in hand, a criterion for outlier detection can be formulated as follows: An observation X_i is considered as outlying whenever

$$\sqrt{(X_i - \overline{X})^T S^{-1} (X_i - \overline{X})} > c_k$$

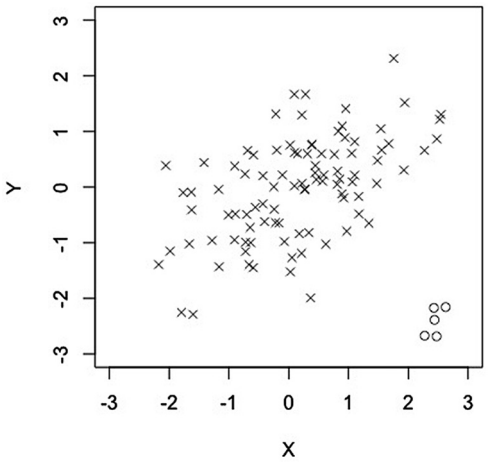

Fig. 1. Scatter plot of two variables X, Y sampled from a normal distribution Z(0,1) with a correlation $\rho = .5$ and 5 outliers (circles).

for a certain coefficient c_k depending on the dimension k of our observations. Note that, in dimension $k = 1$, this criterion boils down to $\sqrt{\frac{(X_i - \overline{X})^2}{s^2}} > c_1$ or $\frac{|X_i - \overline{X}|}{s} > c_1$, which is exactly the well-known criterion mean $\pm$ a coefficient c_1 times the standard deviation s (recall that the covariance matrix S corresponds to s^2). Thus, the basic Mahalanobis criterion is a multivariate extension of the univariate method of the mean $\pm$ a coefficient times the standard deviation. Obviously it suffers from the same criticism, namely a severe lack of robustness. Multivariate outliers may consequently not necessarily have large basic Mahalanobis distances, which is called masking effect. To overcome this problem, we advocate to use distances based on robust estimators of multivariate location and covariance matrix (see, e.g., Daszykowski, Kaczmarek, Vander Heyden, & Walczak, 2007, for a review).

3. Cook's distance and leverage methods

Among outlier detection methods, Cook's distance and leverage are less common than the basic Mahalanobis distance, but still used. Cook's distance estimates the variations in regression coefficients after removing each observation, one by one (Cook, 1977). Therefore, as soon as there is more than one outlying value, the remaining outliers influence the estimators. As for the leverage method, it provides the same information as the Mahalanobis distance (Cohen et al., 2003): It is based on the study of residuals and their distance from the mean vector (e.g. Thode, 2002), which are computed using mean and variance, still polluted by outliers. This is why we recommend to use robust procedures to estimate the position μ and the scatter matrix Σ. The aim of robust methods is to estimate the location μ and the scatter matrix Σ even though the data has been contaminated. We introduce in the remaining of the text the robust method called Minimum Covariance Determinant.

4. The Minimum Covariance Determinant estimators

The Minimum Covariance Determinant approach was proposed by Rousseeuw (1984, 1985). The idea is quite simple: to find a fraction h of "good observations" which are not considered to be outliers and to compute the sample mean and covariance from this sub-sample. In other words, for a sample of size n, a number h of observations, where h lies between $n/2$ and n, is selected on which the empirical mean and empirical covariance matrix are calculated. This procedure is repeated for all possible sub-samples of size h and at the end the sub-sample which has the minimum determinant is selected. This deletes the effect of the most extreme observations, hence also of the outliers, and results in a very robust procedure. The goal is to find the "most central" subsample as that one will correspond to the one having least variability among the observations, meaning whose covariance matrix has minimal determinant, hence the name Minimum Covariance Determinant (MCD). The MCD estimators of location and scatter, denoted $\widehat{\mu}_{MCD}$ and $\widehat{\Sigma}_{MCD}$, correspond to the sample mean and covariance matrix of this most central sub-sample.

The MCD approach has nice statistical properties as its estimators are affine equivariant[4] and asymptotically normal (Butler, Davies, & Jhun, 1993). Moreover, the "breakdown point" (see, e.g., Donoho & Huber, 1983), which is an indicator of the insensitivity to outliers, of the MCD corresponds approximatively to $(n\text{-}h)/n$ (see, e.g., Hubert, Rousseeuw, & Van Aelst, 2008). Thus, for h close to $n/2$, it can reach a maximal breakdown point of 50%. An estimator's breakdown

[4] A quantity is affine-equivariant if its value changes linearly with the choice of coordinate system, which is important if for instance two people are measuring the same object in two distinct measurement units, say meters and centimeters: their results should be the same up to a factor 100. Affine-invariance means that the result does not depend at all on the coordinate system, hence expressed in meters or centimeters this quantity will remain the same.

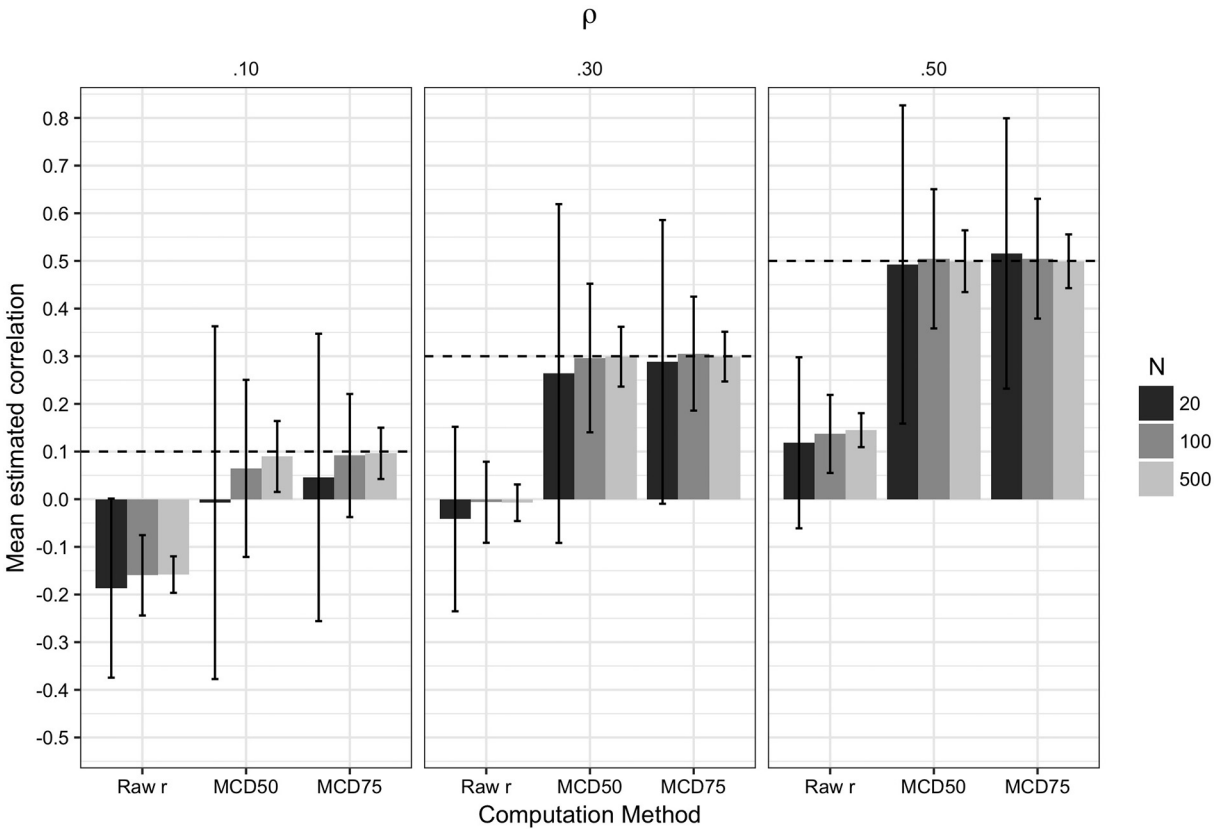

Fig. 2. Estimation of correlation between x and y with 5% outliers included.
Note: In each sample, 5% outliers are included at values varying randomly between 1.97 and 2.99 for X and − 1.97 and − 2.99 for Y. Error bars are standard deviations. Statistics are computed on 1000 simulations. Estimators are computed using the whole sample (Raw r, without any correction; which correspond to the estimators used with the basic Mahalanobis distance), the MCD50 centroïd and the MCD75 respectively.

point is the maximum proportion of observations that can be set to infinity without the estimator being influenced by it and providing a "false estimate" of the quantity. For instance, when a single observation has a very large value compared to the other observations in the sample, the mean of all observations becomes very large. However, this does not really reflect the average of that sample. The mean's breakdown point is 0. A robust approach with a high breakdown point would not be affected by the very large value, as it will consider this value as an outlier. The median is an example of such a robust estimator for the location of a data set.

From a practical point of view, the MCD is computationally demanding. However there exists an algorithm called FAST-MCD (Rousseeuw & Van Driessen, 1999), which renders the computation of the MCD faster. The MCD is implemented in R (see Fauconnier & Haesbroeck, 2009, for practical details). Rousseeuw and Van Driessen (1999) proposed the FAST-MCD command on R. One can compute it as well in SPSS via an R interface, because there does not exist a straightforward function for MCD in SPSS. See Appendix A for a step-by-step description of the algorithm.

Although we chose to focus here on the MCD which is one robust multivariate estimation method, there exist other alternatives. A brief list includes M-estimators (Maronna, 1976), the Stahel-Donoho estimator (Donoho, 1982; Stahel, 1981), S-estimators (Lopuhaä, 1989; Rousseeuw & Leroy, 1987) and MM-estimators (Tatsuoka & Tyler, 2000). Note that the above cited estimators are all affine equivariant. M-estimators are the computationally most attractive option, however they have a rather low breakdown point compared to the others which all can withstand a high fraction, up to 50%, of contaminated data. S-estimators are best computed using the FAST-S algorithm which uses techniques similar to the FAST-MCD.

5. The Mahalanobis-MCD distance

In view of what precedes, the robust criterion for multivariate outlier detection we shall propose corresponds to

$$\sqrt{(X_i - \widehat{\mu}_{MCD})^T (\widehat{\Sigma}_{MCD})^{-1} (X_i - \widehat{\mu}_{MCD})} > c_k,$$

where c_k remains to be determined. Note that as the MCD estimator is affine equivariant, the robust Mahalanobis distances are affine invariant. Theoretically, the squared Mahalanobis-MCD distance (in abbreviation MMCD distance) can be approximated by a χ_k^2 distribution (Rousseeuw & Van Zomeren, 1990), hence we can use $c_k = \sqrt{\chi^2_{k;1-\alpha}}$, which is the square-root of the upper-α quantile of the chi-square distribution with k degrees of freedom. Natural choices for 1-α are 90%, 95%, 97.5%, 99% and 99.9%, the latter being the most conservative choice. This criterion is a natural extension of the median plus or minus a coefficient times the MAD method (Leys et al., 2013).

6. Monte Carlo simulation

In order to show the superiority of the Mahalanobis-MCD distance over the basic Mahalanobis distance in terms of outlier detection capacities, we ran a Monte Carlo simulation using the following settings:

(a) We sampled two random variables X and Y from a normal distribution Z(0,1).
(b) We set a population correlation of ρ = .1, .3 and .5, related to Cohen's effect size standards (Cohen, 1992) between the variables.
(c) We use three sample sizes: 20, 100, and 500.
(d) We introduce a constant 5% of outlying values, respectively 1, 5 and 25 for each sample size. These values were set such that they

had to be detected as multivariate outliers but not univariate ones. That is, each value ranged from 1.96 to 2.99 on X and − 1.96 to − 2.99 on Y. Considering the example of Height and Weight relation, this would be as if the observation belonged to the 95–99.86% tallest individuals and to the 95–99.86% lightest individuals (which would be weird).

(e) Detection level was 9.21 and 13.82, which are the chi-square values (with 2 degrees of freedom) for quantiles 99 and 99.9 (most conservative).

Firstly, as we can see, the addition of 5% outlying values severely distorts the estimations of all estimators (see Fig. 2). Estimations of the mean correlation in the full sample (i.e., NA on Fig. 2) depart systematically more or less .30 units from the actual correlation in the population in the direction of the outlying values. This applies regardless of the correlation in the population or of the sample size. For example, detecting small effect sizes ($r = .1$, which is in the typical range of effect sizes in the social psychological research: Richard, Bond, & Stokes-Zoota, 2003), even with a large sample ($n = 500$), can yield an average correlation of −.16, which is in the opposite direction to the actual effect. Considering that these estimates are normally distributed, knowing the mean and the standard deviations easily allows to compute the likelihood of type I or type II errors. Obviously, given the level of inaccuracy of the estimates when outliers are included, the likelihood of at least one of these errors (depending on the location of the outliers) can be very large.

Secondly, Table 1 provides estimations of the correlations (and SD) using Mahalanobis distance, MCD50 (using a sub-sample of $h = n/2$, hence a breakdown point of 0.5), MCD75(using a sub-sample of $h = 3n/4$, hence a breakdown point of 0.25) methods to remove outliers as well as the true and false detection rates. It shows that MCD75 always yields the best estimations. It has the most efficient detection of outlying values as well as an acceptable false detection rate. Indeed, the MCD50 bases its estimates on a smaller sub-sample than MCD75 and, therefore, is less reliable. The basic Mahalanobis distance is seriously disturbed by outliers, hence not reliable at all. Note that the smaller the correlation, the harder it is to detect outliers, although MCD75 remains the best choice (a correlation of .1 implies that Mahalanobis method will not detect any outliers even in large samples). Of course, the superiority of MCD75 holds as long as there are less than 25% outliers (which is true of most social psychological research). We ran a simulation on 500 observations with 30% outliers. In this situation MCD50 becomes the best indicator (the basic Mahalanobis and MCD75 become totally unreliable).

Lastly, as shown by Table 1, the best detection level for MCD50 and MCD75 is the chi-square at $p = .001$ (cut-off = 13.82 for 2 dimensions), whereas the basic Mahalanobis should use a chi-square at $p = .01$ to get a barely reliable detection (cut-off = 9.21 for 2 dimensions) although we should remind ourselves that this holds true for outliers ranging between 1.96 and 2.99 on X and − 1.96 and − 2.99 on Y. Note that using the MAD with a recommended conservative cut-off of 3 (Leys et al., 2013) corresponds to a quantile of .999. We suggest using the same quantile with MCD. Table 2 and Table 3 show the scripts used to compute the basic Mahalanobis distance, MCD50 and MCD75 on R and SPSS statistical softwares, respectively.

Note that we have only reported the results for 2-dimensional

Table 2
Script for Mahalanobis distance, MMCD50 and MMCD75 calculation on R software.

```
library(MASS)
    #Creating covariance matrix for MCD («totalmatr» is the matrix containing your
    data
    #with x in column 1 and y in column 2)
    output50 < -cov.mcd(totalmatr,quantile.used = nrow(totalmatr)*.5)
    output75 < -cov.mcd(totalmatr,quantile.used = nrow(totalmatr)*.75)
    #Distances from centroid for each matrix
    md < -mahalanobis(totalmatr,colMeans(totalmatr),cov(totalmatr))
    mhmcd50 < -mahalanobis(totalmatr,output50$center,output50$cov)
    mhmcd75 < -mahalanobis(totalmatr,output75$center,output75$cov)
    #Detecting outliers for each method
    #The index of each detected outlier is recorded for each method for a
    alpha = .01
    #For more than two variables, df of cutoff variable (in bold) has to be adjusted
    alpha < -.01
    cutoff < -(qchisq(p = 1-alpha, df = 2))
    names_outliers_MH < -which(md > cutoff)
    names_outliers_MCD50 < -which(mhmcd50 > cutoff)
    names_outliers_MCD75 < -which(mhmcd75 > cutoff)
    #Excluding outliers in a new matrix (here based on MCD75) called totalmatr2
    excluded < -names_outliers_MCD75
totalmatr2 < -totalmatr[-excluded,]
```

Table 1
Comparison of detection performance and estimation of correlation as a function of outlier detection method used.

			Mahalanobis			MCD50						MCD75					
			α = .01			**α = .001**			**α = .01**			**α = .001**			**α = .01**		
N	ρ		r	HR	FAR	r	HR	FAR	r	HR	FAR	r	HR	FAR	r	HR	FA
20	.10	M	**−.10**	.24	.00	**.02**	.73	.08	**.03**	.81	.12	**.03**	.71	.02	**.05**	.85	.05
		SD	.28	.43	.01	.39	.44	.10	.43	.39	.11	.32	.45	.04	.33	.36	.05
	.30	M	**.15**	.50	.00	**.25**	.86	.08	**.25**	.90	.12	**.26**	.87	.02	**.27**	.94	.05
		SD	.30	.50	.01	.36	.35	.10	.40	.30	.11	.29	.34	.04	.30	.23	.05
	.50	M	**.44**	.80	.00	**.46**	.94	.07	**.46**	.96	.11	**.48**	.97	.02	**.48**	.99	.05
		SD	.24	.40	.01	.30	.25	.10	.33	.20	.10	.23	.17	.04	.24	.10	.05
100	.10	M	**−.12**	.19	.00	**.02**	.68	.01	**.08**	.90	.03	**.02**	.66	.00	**.09**	.95	.02
		SD	.11	.19	.01	.16	.37	.01	.16	.25	.02	.15	.33	.01	.13	.16	.02
	.30	M	**.12**	.44	.00	**.28**	.91	.01	**.30**	.98	.03	**.28**	.92	.00	**.30**	1.00	.02
		SD	.14	.23	.01	.13	.23	.01	.12	.10	.02	.12	.18	.01	.11	.03	.02
	.50	M	**.40**	.76	.00	**.50**	.99	.01	**.50**	1.00	.03	**.50**	1.00	.00	**.50**	1.00	.02
		SD	.12	.18	.01	.09	.06	.01	.10	.01	.02	.08	.02	.01	.09	.00	.02
500	.10	M	**−.13**	.17	.01	**.01**	.65	.00	**.09**	.98	.02	**.00**	.63	.00	**.10**	.99	.01
		SD	.05	.09	.00	.09	.22	.00	.06	.07	.01	.08	.19	.00	.06	.03	.01
	.30	M	**.11**	.44	.00	**.29**	.95	.00	**.30**	1.00	.02	**.28**	.95	.00	**.30**	1.00	.01
		SD	.06	.11	.00	.05	.07	.00	.05	.00	.01	.05	.07	.00	.05	.00	.01
	.50	M	**.40**	.75	.00	**.50**	1.00	.00	**.50**	1.00	.02	**.50**	1.00	.00	**.50**	1.00	.01
		SD	.05	.08	.00	.04	.00	.00	.04	.00	.01	.04	.00	.00	.04	.00	.01

Note: r = correlation when outliers are excluded, HR = hit rate (ratio between number of correctly detected outliers and the total number of added outliers in the sample), FAR = False alarm rate (ratio between the number of observations in the original sample detected as outliers and the N of the original sample). .001 and .01 refer to the two thresholds used for excluding outliers based on the MCD, the Mahalanobis distance always uses .01, using .001 always yields far worse estimations.

Table 3
Script for Mahalanobis distance, MMCD50 and MMCD75 calculation on SPSS software. You have to install the plug-in that enables you to run R syntax within SPSS, as there does not exist a pre-implemented function for MCD in SPSS. The plug-in can be downloaded from https://www.ibm.com/developerworks/library/ba-call-r-spss/index.html. This online source contains as well other useful information about calling R from SPSS.

```
BEGIN PROGRAM R.
    # Pull the data into a data frame
  totalmatr = spssdata.GetDataFromSPSS()
    # Pull the data dictionary into another data frame
  totalmatrDict = spssdictionary.GetDictionaryFromSPSS()
    # INSERT HERE the script given above in Table 2 for R
    # List the outliers
  print("outling data:")
  print(totalmatr[excluded,])
    # Set up a new SPSS database with the same dictionary
  spssdictionary.SetDictionaryToSPSS("Test2", totalmatrDict)
    # Copy the data to the new SPSS database
  spssdictionary.SetDataToSPSS("Test2", totalmatr2)
    # Tell SPSS you are done creating data
  spssdictionary.EndDataStep()
    END PROGRAM.
```

Note: You may need to install packages before the first run of your script.
Note: You can check for outliers via Mahalanobis distance in SPSS using the following path in the menu: Regression > Save > Mahalanobis.

tables. What about datasets involving more than 2 variables? The logic described above also holds for more complex dimensional spaces and, although we cannot report simulations for all possible dimensions, the same conclusion holds: The MCD75 allows detecting outliers without being contaminated by the outliers themselves.

7. Conclusion

Given the results of our survey of two journals, emphasizing a poor management of multivariate outliers, we showed that the methods conventionally used are problematic because they are polluted by the outlying value they aim at detecting. We argue in favor of robust estimators with a suitably high breakdown point, as these estimators are the least affected by outliers. Moreover, we would suggest the use of estimators that are affine invariant such as the MCD approach proposed in this paper.

We propose two take-home messages:

1) It is important to preregister the ways outliers will be detected and handled, with an analysis plan detailed enough to anticipate most situations. Using procedures similar to those suggested for univariate outliers we recommend to report the method used for detection, the cut-off selected, the number and value of outliers removed, and possibly the results obtained with and without outliers (especially if the procedure has not been preregistered or if the preregistered decision has to be changed post hoc).
2) We suggest using the MCD75 rather than the basic Mahalanobis distance. Of course MCD50 has a higher breakdown point, but if there are no reasons to believe that more than 25% outliers contaminate the data, MCD75 is a better indicator. This is especially true for small samples.

Appendix A

The brute-force algorithm of MCD is as follows:

(1) Determine all sub-samples containing h observations.
(2) For each sub-sample of size h estimate the covariance matrix and compute the determinant of all those covariance matrices.
(3) Choose the sub-sample with the smallest determinant.
(4) Estimate the dispersion and the location with this sub-sample: $\widehat{\mu}_{MCD}$ and $\widehat{\Sigma}_{MCD}$.

This algorithm is inefficient and hardly computable for large dimensions. Rousseeuw and Van Driessen (1999) developed the FAST-MCD:

(1) Choose a random sub-sample containing h observations.
(2) Estimate the covariance matrix and the location vector.
(3) Calculate the Mahalanobis distance for the n observations using the covariance matrix and location vector of step (2).
(4) Choose the h smallest distances and create a new subset.
(5) Repeat steps (2)–(4) until the difference of the determinants of the covariance matrices for two sub-sequent sub-samples is smaller than a pre-specified threshold.
(6) Repeat steps (1)–(5) m times and choose the sub-sample that has the covariance matrix with the smallest determinant.

References

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t-tests: The power of alternatives and recommendations. *Psychological Methods, 19*(3), 409–427.
Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd edition). New York: John Wiley & Sons.
Burrow, A., & Rainone, N. (2017). How many likes did I get?: Purpose moderates links between positive social media feedback and self-esteem. *Journal of Experimental Social Psychology, 69*, 232–236. http://dx.doi.org/10.1016/j.jesp.2016.09.005.
Butler, R., Davies, P., & Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics, 21*(3), 1385–1400.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.
Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple correlation/regression analysis for the behavioral sciences.* UK: Taylor & Francis.
Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics, 19*(1), 15–18.
Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research, 3*(1), 58–67.
Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., & Walczak, B. (2007). Robust statistics in data analysis – a review: basic concepts. *Chemometrics and Intelligent Laboratory Systems, 85*(2), 203–219.
Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators.* Ph.D. dissertationHarvard University.
Donoho, D. L., & Huber, P. J. (1983). The notion of breakdown point. In P. J. Bickel, K. Doksum, & J. L. HodgesJr. (Eds.). *A Festschrift for Erich L. Lehmann* (pp. 157–184). California: Wadsworth.
Fauconnier, C., & Haesbroeck, G. (2009). Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology, 6*(4), 363–379.
Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* London: Guilford Press.
Hubert, M., Rousseeuw, P. J., & Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science, 23*(1), 92–119.
Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data Analysis: A Model Comparison Approach to Regression, ANOVA, and Beyond* (3rd ed). Abingdon, UK: Routledge.
Kline, R. B. (2015). *Principles and practice of structural equation modeling.* London: Guilford publications.
Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766.
Lopuhaä, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics, 17*(4), 1662–1683.
Mahalanobis, P. C. (1930). On tests and measures of groups divergence. *Journal of Asiatic Sociology of Bengal, 26*, 541–588.
Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics, 4*(1), 51–67.
McClelland, G. H. (2000). Nasty data: Unruly, ill-mannered observations can ruin your analysis. In H. T. Reis, & C. M. Judd (Eds.). *Handbook of research methods in social and personality psychology* (pp. 393–411). Cambridge: Cambridge University Press.
McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology, 48*, 1–30.
Meade, A. W., & Craig, B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455.
Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89*(6), 852–863.
Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331–363. http://dx.doi.org/10.1037/1089-2680.7.4.331.
Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871–880.
Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, & W. Wertz (Eds.). *Mathematical Statistics and Applications, Vol. B* (pp. 283–297). Netherlands: Reidel.
Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection.* New York: John Wiley & Sons, Inc.

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212–223.

Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points (with discussion). *Journal of the American Statistical Association, 85*(411), 633–651.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Stahel, W. A. (1981). Breakdown of covariance estimators. *Research Report 31, Fachgruppe für Statistik*. Switzerland: E.T.H. Zürich.

Tatsuoka, K. S., & Tyler, D. E. (2000). On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *The Annals of Statistics, 28*(4), 1219–1243.

Thode, H. C. (2002). *Testing for normality.* New York: Marcel Dekker.

van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2–12. http://dx.doi.org/10.1016/j.jesp.2016.03.004.