



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



Christophe Leys<sup>a,\*</sup>, Olivier Klein<sup>a</sup>, Yves Dominicy<sup>b,1</sup>, Christophe Ley<sup>c</sup>

<sup>a</sup> Université libre de Bruxelles, Centre de Recherche en Psychologie Sociale et Interculturelle, Belgium

<sup>b</sup> Université libre de Bruxelles, Solvay Brussels School of Economics and Management, ECARES, Belgium

<sup>c</sup> Ghent University, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium

### A B S T R A C T

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 ( $n_s = 20, 100, 500$ ) and population correlation coefficient ( $\rho = .10, .30, .50$ ), 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](mailto:cleys@ulb.ac.be) (C. Leys).

<sup>1</sup> 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.

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 \(2017: study 2\)](#)<sup>2</sup> 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<sup>3</sup> ([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 decide 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”

<sup>2</sup> 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.

<sup>3</sup> 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.

Download English Version:

<https://daneshyari.com/en/article/5045595>

Download Persian Version:

<https://daneshyari.com/article/5045595>

[Daneshyari.com](https://daneshyari.com)