Journal of Multivariate Analysis 166 (2018) 32–49

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

On masking and swamping robustness of leading nonparametric outlier identifiers for multivariate data

Shanshan Wang [a,1], Robert Serfling [b,*,2]

[a] Operations Research and Advanced Data Analytics Group, Burlington Northern and Santa Fe (BNSF) Railway, Fort Worth, TX 76131, USA
[b] Serfling & Thompson Statistical Consulting and Tutoring, 1921 Sparrows Point Drive, Plano, TX 75023, USA

ARTICLE INFO

Article history:
Received 6 March 2017
Available online 17 February 2018

AMS 2010 subject classifications:
primary 62G35
secondary 62H99

Keywords:
Breakdown points
Masking robustness
Multivariate analysis
Nonparametric methods
Outlier detection
Swamping robustness

ABSTRACT

For any outlier detection procedure, a key concern is robustness with respect to possible misclassification errors, masking (Type I) and swamping (Type II). Although parametric model-based simulation results are informative, one also desires nonparametric masking and robustness measures that are more broadly applicable. To this effect, notions of finite-sample masking and swamping breakdown points formulated abstractly for outlyingness functions in arbitrary data settings (Serfling and Wang, 2014) are introduced in the present paper into the multivariate data setting. Formulas for the measures are derived for three important affine invariant nonparametric multivariate outlyingness functions: Mahalanobis distance, Mahalanobis spatial, and projection. Using the formulas, favorable masking and swamping breakdown points, balanced equally, are seen for the Mahalanobis distance outlyingness using minimum covariance determinant (MCD) location and scatter estimators, and likewise for the projection outlyingness with median and MAD for univariate location and scale. Also, Mahalanobis spatial outlyingness with MCD standardization is competitive when swamping robustness is given higher priority than masking robustness. A small simulation study with bivariate contaminated standard normal and contaminated exponential models yields results consistent with the theoretical formulas. Some practical recommendations are discussed.

1. Introduction

Outlier and anomaly identification is a critical first step in exploratory data analysis. Such data points may be nuisances to be eliminated or ignored, or they might present targets of special interest. Here we consider the setting of data in $\mathbb{R}^d$. While visualization can help identify "outliers" in low dimension, for dimension $d \geq 3$ one must rely on *algorithmic approaches* that examine the geometric structure of the data and give the degree of outlyingness for any point. In this regard, a particular outlier identification method is given as an *outlyingness function* defined on $\mathbb{R}^d$.

For using and comparing such outlier identifiers, it is important to know their tendencies toward the two kinds of misclassification error that can occur in the presence of outliers in the data: "masking" of an outlier as a nonoutlier (Type I error), and "swamping" of a nonoutlier as an outlier (Type II error). One thus wishes to characterize for an outlier identifier its

* Corresponding author.
E-mail addresses: Shanshan.Wang@BNSF.com (S. Wang), rjserfling@gmail.com (R. Serfling).
URL: http://www.utdallas.edu/~serfling (R. Serfling).

1 Part of doctoral dissertation research carried out at the University of Texas at Dallas under the direction of Robert Serfling.

2 Robert Serfling retired from the University of Texas at Dallas on September 1, 2017 but remains affiliated informally.

https://doi.org/10.1016/j.jmva.2018.02.003
0047-259X/

masking robustness and its *swamping robustness*. This can be explored by simulation studies based on particular hypothetical parametric models for the data, but for generality of application it is important to have objective, quantitative measures that are defined in a nonparametric fashion and thus are applicable very broadly. Such measures are provided by special notions of finite-sample breakdown point: *masking breakdown point (MBP)* and *swamping breakdown point (SBP)*, the minimum fractions of points in the data set which, if replaced suitably, cause the given procedure to mask outliers or to swamp nonoutliers, respectively. High MBP and SBP are desired. These robustness measures may be found theoretically, bypassing extensive computation for simulation studies oriented to particular parametric models, although nevertheless some such studies can provide added perspective. As interrelated criteria which trade off against each other, MBP and SBP should be considered in concert for any procedure. Note that they provide specialized robustness criteria distinct from the robustness properties of related location and scatter estimators that are possibly involved in the formulation of the given outlier identification procedure.

Here MBP and SBP are derived for four leading outlyingness functions for multivariate data. Three are affine invariant: *Mahalanobis distance outlyingness, Mahalanobis spatial outlyingness*, and *projection outlyingness*. The fourth, *spatial outlyingness*, is only orthogonally invariant but nevertheless is used in practice and also in defining Mahalanobis spatial outlyingness. In terms of their associated MBP and SBP results, we compare these procedures and make practical recommendations. Also, a simulation study is carried out and yields empirical results consistent with the theoretical nonparametric MBPs and SBPs.

Let us now make this discussion precise. Our setting is *nonparametric* outlier identification in $\mathbb{R}^d$, where the bulk of the data consists of "regular" observations from a distribution F on $\mathbb{R}^d$ that is unknown and not assumed to belong to a specified parametric family. The central goal is to characterize the outlyingness of points $\boldsymbol{x}$ in $\mathbb{R}^d$, relative to the distribution F, in terms of an *outlyingness function* $O(\boldsymbol{x}, F)$. Such a function yields a "center" or "median" (the minimum outlyingness point), a designation of the "middle half" region (the 50% least outlying points), and regions for selected other outlyingness thresholds. An outlyingness function corresponds to a global view of each point $\boldsymbol{x}$, in comparison with the density function which quantifies local probability mass at $\boldsymbol{x}$. A sample version $O(\boldsymbol{x}, \mathbb{X}_n)$ analogously structures a data set $\mathbb{X}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ assumed to include "regular" observations from F and possibly contaminants from other sources. *Nonparametric* outlier identification differs from *parametric* outlier identification, which is based on a specified model for the "regular" observations, for example the normal, and which pursues goals such as parametric model-fitting after elimination of outliers, or robust regression modeling in the normal model setting. Being based on a function and hence algorithmic in its formulation, a *nonparametric* outlier identification procedure does not depend critically on graphical views or other subjective criteria and can be used in online data analysis and statistical learning. Of course, outlyingness functions and hence also MBP and SBP can nevertheless also be used in the setting of a contaminated parametric model.

In our nonparametric treatment, the sample "outliers" are defined as those points with $O(\boldsymbol{x}, \mathbb{X}_n)$ above a specified threshold λ. These include unknown numbers of "regular" observations from F and "contaminants" from other sources, the latter typically lying in or toward the tail regions of the data. One goal is to detect the presence of contaminants and sort them out from the regular points. However, since contaminants can seriously disrupt the performance of $O(\boldsymbol{x}, \mathbb{X}_n)$ as a surrogate for $O(\boldsymbol{x}, F)$, *we need* $O(\boldsymbol{x}, \mathbb{X}_n)$ *to be robust in terms of MBP and SBP*. Actually, for each of MBP and SBP, there are two complementary versions, Type A and Type B, making altogether four relevant robustness measures to be considered.

Type A MBP measures the extent to which an extreme outlier of F can be masked in the sample as a nonoutlier at λ outlyingness level, while *Type B MBP* measures how deeply (centrally) in the sample a γ level outlier of F can be masked as a nonoutlier. On the other hand, *Type A SBP* measures how centrally a nonoutlier of F can be swamped as a λ level sample outlier, while *Type B SBP* measures the most extreme sample λ threshold at which a γ level nonoutlier of F can be swamped as a sample outlier. The Type A measures are based on a given choice of sample threshold λ and thus pair together, while the Type B measures are based on a given choice of F threshold γ and likewise pair together.

While the concept of replacement breakdown point (RBP) for *estimators* is well-established, straightforward, and widely applied, notions of MBP and SBP have proved more problematic to formulate. The seminal paper Davies and Gather [4] treats versions of Type A MBP and Type B SBP for some outlier detection procedures for the *univariate contaminated normal model*. In the setting of the *multivariate contaminated normal model*, Becker and Gather [1] treat Type A MBP for *Mahalanobis distance outlyingness*. Dang and Serfling [2] define Type A MBP in the *nonparametric* multivariate setting and evaluate it for several outlyingness functions. Serfling and Wang [16] extend that treatment to develop a broad foundational framework for coherent study of both Type A and Type B MBP and SBP for *any outlyingness function in any data space*. Application of that framework in any particular data setting requires, however, specialized development relevant to that data setting. Its application in the setting of *univariate data*, treating all four MBP and SBP measures for the scaled deviation and centered rank outlyingness functions, is carried out by Wang and Serfling [19]. That work is related to the present treatment of all four measures for selected *multivariate* outlyingness functions, as follows: scaled deviation outlyingness is a special case of both Mahalanobis distance outlyingness and projection outlyingness, while centered rank outlyingness is a special case of Mahalanobis spatial outlyingness.

The multivariate outlyingness functions under consideration are introduced in Section 2. Precise formulation of the four MBP and SBP measures, along with key representations of them in terms of RBPs and the RBPs for relevant location and spread or scatter estimators that will arise here, are provided in Section 3. Theoretical MBP and SBP results for the selected outlyingness functions are presented and discussed in Section 4. Versions of Mahalanobis distance outlyingness and projection outlyingness are seen to possess highly favorable MBP and SBP results, balancing equally well between masking and swamping robustness. When only moderate Type A masking robustness can be tolerated, Mahalanobis spatial