Journal of Statistical Planning and Inference 162 (2015) 62–74

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

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

Shanshan Wang, Robert Serfling *,[1]

Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080-3021, USA

ARTICLE INFO

Article history:
Received 25 February 2013
Received in revised form 4 February 2015
Accepted 5 February 2015
Available online 14 February 2015

MSC:
primary 62G35
secondary 62-07

Keywords:
Nonparametric
Outlier detection
Boxplot
Masking robustness
Swamping robustness
Breakdown point

ABSTRACT

In the wide-ranging scope of modern statistical data analysis, a key task is identification of *outliers*. For any outlier identification procedure, one needs to know its robustness against *masking* (an "outlier" is undetected as such) and *swamping* (a "nonoutlier" is classified as an "outlier"). Masking and swamping robustness are interrelated aspects which must be studied together. For such purposes, Serfling and Wang (2014) provide a general framework applicable in any data space. Implementation, however, with particular outlier identifiers in particular types of data space, requires additional theoretical development specialized to the chosen setting. Even the case of univariate data presents nontrivial challenges. Here we apply the framework to study the masking and swamping robustness properties of two leading types of nonparametric outlier identifiers, *scaled deviation outlyingness* and *centered rank outlyingness*. The results shed new light on the choice between (Median, MAD) and (trimmed mean, trimmed standard deviation) in using scaled deviation outlyingness. Also, our findings explain how the *boxplot*, a leading descriptive tool, performs using a hybrid outlyingness function incorporating a quantile-based component to describe the middle half of a data set and a scaled deviation outlyingness component for outlier detection. For both goals, the boxplot greatly favors swamping robustness over masking robustness. We also formulate a variant boxplot offering a more favorable trade-off between these two criteria.

1. Introduction

Crucial to data analysis is the identification of outliers and anomalies, which may be nuisances to be eliminated or ignored, or possibly targets of special interest. Errors to be avoided in so far as possible are "masking" of an outlier as a nonoutlier and "swamping" of a nonoutlier as an outlier. It is important to evaluate, for any outlier identification procedure, both its *masking robustness* and its *swamping robustness*, which are interrelated and trade off against each other.

Quantitative measures for such purposes, the *masking breakdown point (MBP)* and the *swamping breakdown point (SBP)*, are studied here for two well-established types of univariate nonparametric outlier identifiers, *scaled deviation* and *centered rank*. Besides providing better understanding of these two important outlyingness functions, the results clarify how the design of the popular boxplot greatly favors swamping robustness over masking robustness. These insights provide a basis for designing a variant boxplot that balances somewhat more equally across these two performance characteristics.

* Corresponding author.
E-mail address: serfling@utdallas.edu (R. Serfling).
[1] http://www.utdallas.edu/~serfling.

http://dx.doi.org/10.1016/j.jspi.2015.02.002
0378-3758/

Notions of MBP and SBP have been developed in a series of papers by Davies and Gather (1993), Becker and Gather (1999), Dang and Serfling (2010), and Serfling and Wang (2014). The latter paper provides the first broad foundational framework for coherent study of MBP and SBP for any outlier identification procedure in any data space, and the treatment includes several key general lemmas aimed at facilitating practical application of the framework. However, application of this general framework is not immediate, but in fact requires innovative further development that is specialized to the particular data setting under consideration. Even the case of univariate data space is nontrivial, and, as the first application of the general framework, is the target of the present paper.

Our setting is *nonparametric* outlier identification, where the bulk of the data consists of "regular" observations from a distribution F that is unknown and not assumed to belong to a specified parametric family. The central goal is to characterize the outlyingness of points x relative to the distribution F, in terms of an *outlyingness function* $O(x, F)$. Such a function corresponds to a global view of x, in comparison with the density function $f(x)$ which quantifies local probability mass at x. It yields a "center" (the minimum outlyingness point), a "middle half" region of the 50% least outlying points, and thresholds for selected degrees of outlyingness. The sample version $O(x, \mathbb{X}_n)$ analogously structures a data set $\mathbb{X}_n$. Being based on a function and thus 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.*

Nonparametric outlier identification differs in orientation and style from *parametric* outlier identification, which is oriented to a specified model for the "regular" observations, typically the normal, and has goals such as parametric model-fitting after elimination of outliers, or robust regression modeling in the normal model setting. For example, the "forward search" method (Atkinson and Riani, 2000; Atkinson et al., 2010) utilizes explicitly the assumed parametric model and carries out diagnostics via graphical displays that are interpreted subjectively.

In our nonparametric treatment, sample "outliers" are points with $O(x, \mathbb{X}_n)$ above some specified threshold λ, and these may include both an unknown number of "regular" observations from F and some "contaminants" arising from other sources than sampling from F and typically 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, such contaminants can seriously disrupt the performance of $O(x, \mathbb{X}_n)$ as a surrogate for $O(x, F)$, so we need $O(x, \mathbb{X}_n)$ to be robust against both masking and swamping, on the basis of well-defined quantitative criteria.

Our objective measures of masking and swamping robustness, the MBP and SBP, are the minimum fractions of points in $\mathbb{X}_n$ which, if replaced in a suitable way, cause the given procedure to *mask outliers* or to *swamp nonoutliers*, respectively. Higher MBP and SBP are better.

More precisely, for each of MBP and SBP, there are two complementary versions, Type A and Type B, making four robustness measures in all. 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 are thus paired together, whereas the Type B measures involve a given choice of F threshold γ and thus are paired together.

Unfortunately, the masking and swamping robustness of outlier identifiers cannot be inferred directly from the "ordinary" robustness properties of the various estimators that may be involved in their formulation. Rather, the notions of MBP and SBP and these four specialized measures are needed. Also, since MBP and SBP trade off against each other, these must be considered in concert.

While the breakdown point (BP) for *estimators* is a well-established and widely applied concept, notions of MBP and SBP are more problematic and have received only limited treatment prior to the general framework of Serfling and Wang (2014). Davies and Gather (1993) treat certain notions of Type A MBP and Type B SBP in the *univariate parametric* setting of the *contaminated normal model*, Becker and Gather (1999) treat Type A MBP in the setting of the *multivariate contaminated normal model*, and Dang and Serfling (2010) treat Type A MBP in the general *nonparametric* multivariate setting.

Here we comprehensively treat MBP and SBP for the sample versions of two important univariate outlyingness functions: *scaled deviation outlyingness*

$$\widetilde{O}(x, F) = \left| \frac{x - \mu(F)}{\sigma(F)} \right|, \quad -\infty < x < \infty, \tag{1}$$

with $\mu(F)$ and $\sigma(F)$ location and spread measures, respectively, and *centered rank outlyingness*

$$O(x, F) = |2F(x) - 1|, \quad -\infty < x < \infty, \tag{2}$$

each increasing as x moves outward from $\mu(F)$ or Median(F), respectively. Scaled deviation outlyingness dates from Mosteller and Tukey (1977), while centered rank outlyingness is rooted in classical inference based on quantiles and ranks. These outlyingness measures are well suited to the nonparametric approach, since they neither require nor put to use an assumption of symmetry as is inherent in the contaminated normal parametric approach.

The results of this paper are as follows. In Section 2 we define the MBP and SBP measures and provide key lemmas instrumental in evaluating them. In Section 3 we develop and discuss MBP and SBP results for scaled deviation outlyingness and centered rank outlyingness. Our results shed new light on the choices (*Mean, SD*), (*Median, MAD*), and (*trimmed mean,*