Statistical Methodology 20 (2014) 79–90

Contents lists available at ScienceDirect

Statistical Methodology

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

General foundations for studying masking and swamping robustness of outlier identifiers

Robert Serfling *, Shanshan Wang

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

ARTICLE INFO

Article history:
Received 15 November 2012
Received in revised form
29 June 2013
Accepted 9 August 2013

This paper is dedicated to the memory of Kesar Singh, an outstanding contributor to statistical science

Keywords:
Nonparametric
Outlier detection
Masking robustness
Swamping robustness

ABSTRACT

With greatly advanced computational resources, the scope of statistical data analysis and modeling has widened to accommodate pressing new arenas of application. In all such data settings, an important and challenging task is the identification of outliers. Especially, an outlier identification procedure must be robust against the possibilities of masking (an outlier is undetected as such) and swamping (a nonoutlier is classified as an outlier). Here we provide general foundations and criteria for quantifying the robustness of outlier detection procedures against masking and swamping. This unifies a scattering of existing results confined to univariate or multivariate data, and extends to a completely general framework allowing any type of data. For any space $\mathcal{X}$ of objects and probability model F on $\mathcal{X}$, we consider a real-valued outlyingness function $O(x, F)$ defined over x in $\mathcal{X}$ and a sample version $O(x, \mathbb{X}_n)$ based on a sample $\mathbb{X}_n$ from $\mathcal{X}$. In this setting, and within a coherent framework, we formulate general definitions of masking breakdown point and swamping breakdown point and develop lemmas for evaluating these robustness measures in practical applications. A brief illustration of the technique of application of the lemmas is provided for univariate scaled deviation outlyingness.

1. Introduction

With greatly advanced computational resources, the scope of statistical data analysis and modeling has widened to accommodate pressing new arenas of application. Now data is invariably *multivariate*,

* Corresponding author. Tel.: +1 972 883 2361.
E-mail address: serfling@utdallas.edu (R. Serfling).
URL: http://www.utdallas.edu/~serfling (R. Serfling).

1572-3127/$ – see front matter
http://dx.doi.org/10.1016/j.stamet.2013.08.004

typically with very high dimension and/or heavy tails and/or huge sample sizes, or *complex*, involving curves, images, sets, and other types of object, often with stream or network structure. In all such data settings, an increasingly important and challenging task is the identification of *outliers*. New contexts involving outliers and "anomaly detection" include fraud detection, intrusion detection, and network robustness analysis, to name a few. The outliers themselves are sometimes the cases of primary interest. *Nonparametric* notions and methods are especially important, since tractable parametric modeling is rather limited in the case of multivariate data (e.g. Olkin [7]) and even more so with complex data. Further, since visualization is feasible only in the case of numerical or vector data in low dimension, outlier detection methods become of necessity *algorithmic* in nature and the determination of their performance properties is complicated.

The key concern about performance of an outlier detection procedure is its *robustness*. It must be resistant to adverse performance effects due to the very presence of the outliers to be identified, or even due to the presence of a concentration of inliers. In particular, one must assess the proclivities of the procedure toward either of two kinds of misclassification error: *masking* (an outlier is undetected) and *swamping* (a nonoutlier is classified as an outlier). Robustness against both masking and swamping is clearly essential.

In handling complex modern data structures, ingenious outlier detection approaches have been crafted *ad hoc* in diverse settings. Typically their robustness performance is explored only through limited simulation studies, with results that lack both generality and precise interpretation. Highly needed are theoretical underpinnings. Here we develop *general foundations and criteria for quantifying robustness of outlier detection procedures against masking and swamping*.

We employ a leading type of robustness measure, the (finite sample) *breakdown point* (BP) of Donoho and Huber [5], i.e., the *minimum fraction of replacements of the sample data* (by outliers or inliers) sufficient to "break down" the statistical procedure, i.e., to render it drastically ineffective. This provides a distinctive quantitative approach toward measuring robustness.

In dealing with an *outlier identification procedure*, the BP approach to robustness involves two such measures: the *masking breakdown point* (MBP) and the *swamping breakdown point* (SBP). These are the minimum fractions of points in a data set which if arbitrarily placed as "outliers" or "inliers" suffice to cause the outlier detection procedure to *mask arbitrarily extreme outliers*, or *swamp arbitrarily central nonoutliers*, respectively. The higher the MBP and SBP values, the better the robustness performance of an outlier detection procedure.

Although the idea of BP for estimators such as the sample mean or variance is well established and quite simple and straightforward to define, the corresponding formulations of MBP and SWP are considerably more problematic and have received limited treatment. In the *parametric* setting of *univariate* data within the *contaminated normal model*, Davies and Gather [4] formulate versions of MBP and SBP using *addition* contamination. Becker and Gather [2] extend that MBP to the *multivariate* contaminated normal model, but extension of the SBP is not considered, although it is treated in Becker [1]. Dang and Serfling [3] introduce a version of MBP in the setting of *fully nonparametric* multivariate outlier identification based on the use of *depth functions* and apply this notion to compare several different depth-based outlier identifiers. Again, however, the SBP is left untreated.

The MBP and SBP are conceptually interrelated and are formulated in parallel ways, although the technical treatments to evaluate them differ in details. One cannot simply transform one problem into the other. Further, it turns out that for each of MBP and SBP there are two relevant versions representing complementary perspectives, making in all four robustness measures. Here we introduce a general framework for study of these together, establish key lemmas for their application, and carry out their application in the setting of univariate data.

Section 2 develops a completely general formulation of nonparametric outlier identification in terms of a real-valued outlyingness function $O(x, F)$, defined over x in any space $\mathcal{X}$ of objects and based on a probability distribution F on $\mathcal{X}$, and a sample version $O(x, \mathbb{X}_n)$ based on a sample $\mathbb{X}_n$ from $\mathcal{X}$. General definitions of MBP and SBP are provided within a unified conceptual framework for studying these robustness measures. Section 3 provides key technical lemmas for evaluating MBP and SBP in practical applications. Section 4 provides a brief illustration of the technique of application of the lemmas, using a leading outlier identifier in the case of *univariate* (real-valued) data, *scaled deviation outlyingness*. Complete treatment of univariate scaled deviation outlyingness as well as of *centered*