Computational Statistics and Data Analysis 99 (2016) 1–11

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Robust regression estimation and inference in the presence of cellwise and casewise contamination

Andy Leung [*], Hongyang Zhang, Ruben Zamar

Department of Statistics, University of British Columbia, 3182-2207 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada

ARTICLE INFO

Article history:
Received 1 October 2015
Received in revised form 30 December 2015
Accepted 7 January 2016
Available online 19 January 2016

Keywords:
Cellwise outliers
Componentwise contamination
Regression S-estimation
Robust estimation
Filtering

ABSTRACT

Cellwise outliers are likely to occur together with casewise outliers in modern datasets of relatively large dimension. Recent work has shown that traditional robust regression methods may fail when applied to such datasets. We propose a new robust regression procedure to deal with casewise and cellwise outliers. The proposed method, called three-step regression, proceeds as follows: first, it uses a consistent univariate filter, that is, a procedure that flags and eliminates extreme cellwise outliers; second, it applies a robust estimator of multivariate location and scatter to the filtered data to down-weight casewise outliers; third, it computes robust regression coefficients from the estimates obtained in the second step. The three-step estimator is consistent and asymptotically normal at the central model under some assumptions on the tails of the distributions of the continuous covariates. The estimator is extended to handle both continuous and dummy covariates using an iterative algorithm. Extensive simulation results show that the three-step estimator is resilient to cellwise outliers. It also performs well under casewise contamination when compared to traditional high breakdown point estimators.

1. Introduction

The vast majority of procedures for robust linear regression are based on the classical Tukey–Huber contamination model (THCM) in which a relatively small fraction of cases may be contaminated. High breakdown point affine equivariant estimators such as least trimmed squares (Rousseeuw, 1984), S-regression (Rousseeuw and Yohai, 1984) and MM-regression (Yohai, 1985) proceed by down-weighting outlying cases, which makes sense and works well in practice, under THCM. However, in some applications, the contamination mechanism may be different in that random cells in a data table (with rows as cases and columns as variables) are independently contaminated. In this paradigm, a small fraction of random cellwise outliers could propagate to a relatively large fraction of cases, breaking down classical high breakdown point affine equivariant estimators (see Alqallaf et al., 2009). Since cellwise and casewise outliers may co-exist in some applications, our goal in this paper is to develop a method for robust regression estimation and inference that can deal with both cellwise and casewise outliers.

There is a vast literature on robust regression for casewise outliers, but only a scant literature for cellwise outliers and none for both types of outliers in the regression context. Recently, Öllerer et al. (2015) combined the ideas of coordinate descent algorithm (called the shooting algorithm in Fu, 1998) and simple S-regression (Rousseeuw and Yohai, 1984) to propose an estimator called the shooting S. The shooting S-estimator assigns individual weight to each cell in the data table

[*] Corresponding author.
E-mail addresses: andy.leung@stat.ubc.ca (A. Leung), hongyang.z@stat.ubc.ca (H. Zhang), ruben@stat.ubc.ca (R. Zamar).

http://dx.doi.org/10.1016/j.csda.2016.01.004
0167-9473/© 2016 Elsevier B.V. All rights reserved.

to handle cellwise outliers in the regression context. The shooting S-estimator is robust against cellwise outliers and vertical response outliers.

In this paper, we propose a three-step regression estimator which combines the ideas of filtering cellwise outliers and robust regression via covariance matrix estimate (Maronna and Morgenthaler, 1986; Croux et al., 2003), namely 3S-regression estimator. By filtering, here we mean detecting outliers and replacing them by missing values as in Agostinelli et al. (2015). Our estimator proceeds as follows: first, it uses a univariate filter to detect and eliminate extreme cellwise outliers in order to control the effect of outliers propagation; second, it applies a robust estimator of multivariate location and scatter to the filtered data to down-weight casewise outliers; third, it computes robust regression coefficients from the estimates obtained in the second step. With the choice of a filter that has simultaneous good sensitivity (is capable of filtering outliers) and good specificity (can preserve all or most of the clean data), the resulting estimator can be resilient to both cellwise and casewise outliers; furthermore, it attains consistency and asymptotic normality for clean data. In this regards, we propose a filter that is consistent under some assumptions on the tails of the covariates distributions. By consistent filter, we mean a filter that asymptotically can preserve all the data when they are clean.

The rest of the paper is organized as follows. In Section 2, we introduce a family of consistent filters. In Section 3, we introduce 3S-regression. In Section 4, we show some asymptotic properties of 3S-regression. In Section 5, we evaluate the performance of 3S-regression in an extensive simulation study. In Section 6, we analyze a real dataset with cellwise and casewise outliers. In Section 7, we conclude with some remarks. We also provide a document referred to as "supplementary material", which contains all the proofs, additional simulation results, and other related material (see Appendix A).

2. Consistent filter

Filtering is a method for pre-processing data in order to control the effect of potential cellwise outliers. In this paper, we pre-process the data by flagging outliers and replacing them by missing values, NAs. This method of filtering has recently been used for robust estimation of multivariate location and scatter (Danilov, 2010; Agostinelli et al., 2015) and for clustering (Farcomeni, 2014a,b). Also, Farcomeni (2015) proposed a procedure to determine a data-driven choice for the number of filtered cells to increase the efficiency of the estimator.

Consistent filters are ones that do not filter good data points asymptotically. Gervini and Yohai (2002) introduced a consistent filter for normal residuals in regression estimation to achieve a fully-efficient robust regression estimator. Consistent filters are desirable because their good asymptotic properties are shared by the following-up estimation procedure. In this paper, we introduce a new family of consistent filters for univariate data.

Consider a random variable X with a continuous distribution function $G(x)$. We define the scaled upper and lower tail distributions of $G(x)$ as follows:

$$\begin{aligned} F^u(t) &= P_G\left(\frac{X-\eta^u}{\text{med}(X-\eta^u|X>\eta^u)} \le t|X>\eta^u\right) \quad \text{and} \\ F^l(t) &= P_G\left(\frac{\eta^l - X}{\text{med}(\eta^l - X|X<\eta^l)} \le t|X<\eta^l\right). \end{aligned} \tag{1}$$

Here, med stands for median, $\eta^u = G^{-1}(1-\alpha)$, $\eta^l = G^{-1}(\alpha)$, and $0 < \alpha < 0.5$. We use $\alpha = 0.20$, but other choices could be considered. To simplify the notation, we set $s^u = \text{med}(X-\eta^u|X>\eta^u)$ and $s^l = \text{med}(\eta^l - X|X<\eta^l)$. Alternatively, a combined tails approach could be used for symmetric distributions as in Gervini and Yohai (2002).

Let $\{X_1, \dots, X_n\}$ be a random sample from G, and let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the corresponding order statistics. Consistent estimators for $(\eta^u, s^u, \eta^l, s^l)$ are given by

$$\begin{aligned} \hat\eta^u_n &= \hat G^{-1}_n(1-\alpha), \qquad \hat s^u_n = \text{med}(\{X_i - \hat\eta^u_n | X_i > \hat\eta^u_n\}), \\ \hat\eta^l_n &= \hat G^{-1}_n(\alpha), \qquad \hat s^l_n = \text{med}(\{\hat\eta^l_n - X_i | X_i < \hat\eta^l_n\}), \end{aligned}$$

where $\hat G^{-1}_n(a) = X_{(\lceil na \rceil)}$, $0 < a < 1$, is the empirical quantile and $\text{med}(\{Y_1, \dots, Y_m\}) = Y_{(\lceil m/2 \rceil)}$ is the sample median (see Lemma 1.1 in the supplementary material (see Appendix A) for a proof of the consistency for $\hat s^u_n$ and $\hat s^l_n$). The empirical distribution functions for the scaled upper and lower tails in (1) are now given by

$$\begin{aligned} \hat F^u_n(t) &= \frac{\sum_{i=1}^n I(0 < (X_i - \hat\eta^u_n)/\hat s^u_n \le t)}{\sum_{i=1}^n I(X_i > \hat\eta^u_n)} \quad \text{and} \\ \hat F^l_n(t) &= \frac{\sum_{i=1}^n I(0 < (\hat\eta^l_n - X_i)/\hat s^l_n \le t)}{\sum_{i=1}^n I(X_i < \hat\eta^l_n)}. \end{aligned}$$