Statistics and Probability Letters 94 (2014) 214–220

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

The finite sample breakdown point of PCS

Eric Schmitt [a,*], Viktoria Öllerer [b], Kaveh Vakili [a]

[a] Afdeling Statistiek, Celestijnenlaan 200b - bus 2400, 3001 Leuven, Belgium

[b] Faculty of Business and Economics, ORSTAT, KU Leuven, Belgium

ARTICLE INFO

Article history:
Received 1 July 2014
Received in revised form 18 July 2014
Accepted 20 July 2014
Available online 1 August 2014

Keywords:
Breakdown point
Robust estimation
Multivariate statistics

ABSTRACT

The Projection Congruent Subset (PCS) is a new method for finding multivariate outliers. PCS returns an outlyingness index which can be used to construct affine equivariant estimates of multivariate location and scatter. In this note, we derive the finite sample breakdown point of these estimators.

1. Introduction

Outliers are observations that depart from the pattern of the majority of the data. Identifying outliers is a major concern in data analysis because a few outliers, if left unchecked, can exert a disproportionate pull on the fitted parameters of any statistical model, preventing the analyst from uncovering the main structure in the data.

To measure the robustness of an estimator to the presence of outliers in the data, Donoho (1982) introduced the notion of finite sample breakdown point. Given a sample and an estimator, this is the smallest number of observations that need to be replaced by outliers to cause the fit to be arbitrarily far from the values it would have had on the original sample. Remarkably, the finite sample breakdown point of an estimator can be derived without recourse to concepts of chance or randomness using geometrical features of a sample alone (Donoho, 1982). Recently, Vakili and Schmitt (2014) introduced the Projection Congruent Subset (PCS) method. PCS computes an outlyingness index, as well as estimates of location and scatter derived from it. The objective of this paper is to establish the finite sample breakdown of these estimators and show that they are maximal.

Formally, we begin from the situation whereby the data matrix $\boldsymbol{X}$, is a collection of n so called *genuine* observations drawn from a p-variate model F with $p > 1$. However, we do not observe $\boldsymbol{X}$ but an $n \times p$ (potentially) corrupted data set $\boldsymbol{X}^{\varepsilon}$ that consists of $g < n$ observations from $\boldsymbol{X}$ and $c = n - g$ arbitrary values, with $\varepsilon = c/n$, denoting the (unknown) rate of contamination.

Historically, the goal of many robust estimators has been to achieve high breakdown while obtaining reasonable efficiency. PCS belongs to a small group of robust estimators that have been designed to also have low bias (see Maronna et al., 1992, Adrover and Yohai, 2002 and Adrover and Yohai, 2010). In the context of robust estimation, a low bias estimator reliably finds a fit close to the one it would have found without the outliers, when $c \leqslant n-h$ with $h = \lceil (n+p+1)/2 \rceil$. To the best of our knowledge, PCS is the first member of this group of estimators to be supported by a fast and affine equivariant algorithm (FastPCS) enabling its use by practitioners.

* Corresponding author. Tel.: +32 16 37 23 40.
E-mail addresses: eric.schmitt@wis.kuleuven.be (E. Schmitt), viktoria.oellerer@kuleuven.be (V. Öllerer), kaveh.vakili@wis.kuleuven.be (K. Vakili).

http://dx.doi.org/10.1016/j.spl.2014.07.026
0167-7152/

The rest of this paper unfolds as follows. In Section 2, we detail the PCS estimator. In Section 3, we formally detail the concept of finite sample breakdown point of an estimator and establish the notational conventions we will use throughout. Finally, in Section 4, we prove the finite sample breakdown point of PCS.

2. The PCS criterion

Consider a potentially contaminated data set $\boldsymbol{X}$ of n vectors $\boldsymbol{x}_i \in \mathbb{R}^p$, with $n > p + 1 > 2$. Given all $M = \binom{n}{h}$ possible h-subsets $\{H^m\}_{m=1}^M$, PCS looks for the one that is most *congruent* along many univariate projections. Formally, given an h-subset H^m, we denote $B(H^m)$ the set of all vectors normal to hyperplanes spanning a p-subset of H^m. More precisely, all directions $\boldsymbol{a} \in B(H^m)$ define hyperplanes $\{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{x}'\boldsymbol{a} = 1\}$ that contain p observations of H^m. For $\boldsymbol{a} \in B(H^m)$ and $\boldsymbol{x}_i \in \boldsymbol{X}$, we can compute the squared orthogonal distance, d_i^2, of $\boldsymbol{x}_i$ to the hyperplane defined by $\boldsymbol{a}$ as

$$d_i^2(\boldsymbol{a}) = \frac{(\boldsymbol{a}'\boldsymbol{x}_i - 1)^2}{\|\boldsymbol{a}\|^2}. \tag{2.1}$$

The set of the h observations with smallest $d_i^2(\boldsymbol{a})$ is then defined as

$$H^{\boldsymbol{a}} = \{i : d_i^2(\boldsymbol{a}) \leqslant d_{(h)}^2(\boldsymbol{a})\}, \tag{2.2}$$

where $d_{(h)}$ denotes the hth-order statistic of a vector $\boldsymbol{d}$.

We begin by considering the case in which $d_{(h)}^2(\boldsymbol{a}) > 0$. For a given subset H^m and direction $\boldsymbol{a}$ we define the *incongruence index* of H^m along $\boldsymbol{a}$ as

$$I(H^m, \boldsymbol{a}) := \log\left(\frac{\underset{i\in H^m}{\text{ave}}\, d_i^2(\boldsymbol{a})}{\underset{i\in H^{\boldsymbol{a}}}{\text{ave}}\, d_i^2(\boldsymbol{a})}\right) \tag{2.3}$$

with the conventions that $\log(0/0) := 0$. This index is always positive and will be smaller the more members of H^m correspond with, or are similar to, the members of $H^{\boldsymbol{a}}$. To remove the dependency of Eq. (2.3) on $\boldsymbol{a}$, we measure the incongruence of H^m by considering the average over many directions $\boldsymbol{a} \in B(H^m)$ as

$$I(H^m) := \underset{\boldsymbol{a}\in B(H^m)}{\text{ave}}\, I(H^m, \boldsymbol{a}). \tag{2.4}$$

The optimal h-subset, H^*, is the one satisfying the PCS criterion:

$$H^* = \underset{\{H^m\}_{m=1}^M}{\text{argmin}}\; I(H^m).$$

Then, the *PCS estimators of location and scatter* are the sample mean and covariance of the observations with indexes in H^*:

$$\left(\boldsymbol{t}^*(\boldsymbol{X}), \boldsymbol{S}^*(\boldsymbol{X})\right) = \left(\underset{i\in H^*}{\text{ave}}\, \boldsymbol{x}_i, \underset{i\in H^*}{\text{cov}}\, \boldsymbol{x}_i\right).$$

Finally, we have to account for the special case where $d_{(h)}^2(\boldsymbol{a}) = 0$. In this case, we enlarge H^* to be the subset of all observations lying on $\boldsymbol{a}$. More precisely, if $\exists\, \boldsymbol{a}^* \in B(H^*) : |\{i : d_i^2(\boldsymbol{a}^*) = 0\}| \geqslant h$, then $H^* = \{i : d_i^2(\boldsymbol{a}^*) = 0\}$.

2.1. Illustrative example

To give additional insight into PCS and the characterization of a cloud of point in terms of congruence, we provide the following example. Fig. 1 depicts a data set $\boldsymbol{X}^\varepsilon$ of 100 observations, 30 of which come from a cluster of outliers on the right. For this data set, we draw two h-subsets of 52 observations each.

Subset H^1 (dark blue diamonds) contains only genuine observations, while subset H^2 (light-orange circles) contains 27 outliers and 25 genuine observations. Finally, the 17 observations belonging to neither h-subset are depicted as black triangles. For illustration's sake, we selected the members of H^2 so that their covariance has smaller determinant than *any* h-subsets formed of genuine observations. Consequently, robust methods based on a characterization of h-subsets in terms of density alone will always prefer the contaminated subset H^2 over any uncontaminated h-subset (and in particular H^1).

The outlyingness index computed by PCS differs from that of other robust estimators in two important ways. First, in PCS, the data is projected onto directions given by p points drawn from the members of a given subset, H^m, rather than indiscriminately from the entire data set. This choice is motivated by the fact that when ε and/or p are high, the vast majority of random p-subsets of $\{1, \ldots, n\}$ will be contaminated. If the outliers are concentrated, this yields directions almost parallel to each other. In contrast, for an uncontaminated H^m, our sampling strategy always ensures a wider spread of directions and this yields better results.