Journal of Multivariate Analysis 126 (2014) 167–183

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Strong consistency and robustness of the Forward Search estimator of multivariate location and scatter

Andrea Cerioli [a,*], Alessio Farcomeni [b], Marco Riani [a]

[a] University of Parma, Via Kennedy 6; 43100 Parma, Italy

[b] Sapienza - University of Rome, piazzale Aldo Moro, 5; 00186 Roma, Italy

ARTICLE INFO

Article history:
Received 16 May 2013
Available online 15 January 2014

AMS subject classifications:
62H10
62H12
62E20
62F12

Keywords:
Elliptical truncation
Forward Search
Generalized Mahalanobis distances
High-breakdown estimation
Multivariate trimming
Robust diagnostics

ABSTRACT

The Forward Search is a powerful general method for detecting anomalies in structured data, whose diagnostic power has been shown in many statistical contexts. However, despite the wealth of empirical evidence in favor of the method, only few theoretical properties have been established regarding the resulting estimators. We show that the Forward Search estimators are strongly consistent at the multivariate normal model. We also obtain their finite sample breakdown point. Our results put the Forward Search approach for multivariate data on a solid statistical ground, which formally motivates its use in robust applied statistics. Furthermore, they allow us to compare the Forward Search estimators with other well known multivariate high-breakdown techniques.

1. Introduction

The Forward Search (FS) is a powerful general method for detecting anomalies in structured data [2,6]. The idea behind the FS is simple and attractive. Given a sample of n observations and a generating model for them, the method starts from a subset of cardinality $m^* \ll n$ – often only few observations are required in practice, unless n is very large – which is robustly chosen to contain observations coming from the postulated model. This subset is used for fitting the model and the residuals, or other deviance measures, are computed. The subsequent fitting subset is then obtained by taking the $m^* + h$ observations with the smallest deviance measures. The algorithm iterates this fitting and updating scheme until all the observations are used in the fitting subset, thus yielding the classical statistical summary of the data. In practice h must be a finite number, possibly depending on n and on the postulated model. For instance, the typical choice with independent observations and moderate sample sizes is $h = 1$, while higher values are suitable with correlated data or very large samples. In the asymptotic framework of this work we have that $h \to \infty$ as $n \to \infty$, but we still assume a finite number of steps in the FS.

A major advantage of the FS is that it provides clear evidence of the impact that each unit, or block of units, exerts on the fitting process, with outliers and other peculiar observations entering in the last steps of the search. The presence of observations deviating from the null model can be displayed through pictures that monitor relevant quantities along the search, such as model residuals, distances, and their order statistics. For instance, if only $m < n$ units actually belong to the postulated population, we typically observe a peak in the monitoring plot of the minimum residual (distance) outside the fitting subset, when this subset only contains the m 'good' observations and the first outlier is about to enter. A further

[*] Corresponding author.
E-mail addresses: andrea.cerioli@unipr.it (A. Cerioli), alessio.farcomeni@uniroma1.it (A. Farcomeni), mriani@unipr.it (M. Riani).

0047-259X/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.12.010

bonus of the FS is that its main findings are usually insensitive to the specific choice of the initial subset, provided that it is outlier free, and virtually identical results are obtained through different criteria [7]. Typical methods for initializing the FS are Least Trimmed Squares in regression [21] and robust bivariate projections in multivariate analysis [36], but several alternative choices are also feasible.

The diagnostic power of the FS has been shown in many statistical contexts. For instance, in regression [3,4,8] the deletion residuals computed at each step of the FS can be monitored along the search, together with some of their relevant order statistics, for the detection of outliers and unsuspected structure in data, and so for building robust models. Such informative pictures are often called *forward plots*, because they are drawn by collecting several pieces of information, each of which is gathered from a different subset as the algorithm progresses. The full power of the FS thus stems from the combination of the different pieces, like in a "data movie" as opposed to a "data picture". Similar tools have also been developed for correlated observations [13], like in the case of spatial autoregressive models and in the kriging model of geostatistics. The FS for multivariate data replaces residuals with Mahalanobis distances, but keeps the general diagnostic approach unchanged. This leads to a (partial) ordering of multivariate data, and to robust and efficient diagnostic tools for the detection of multivariate outliers [5,30,27,18].

However, despite the wealth of empirical and simulation evidence in favor of the method, only few theoretical properties are available for the resulting estimators. The key ingredient for deriving such properties is the distribution of the basic quantities, i.e. residuals or distances, which are monitored along the search. These quantities are computed after a sequence of data driven steps. Therefore, obtaining their distribution is far from trivial, even in an asymptotic framework. Some approximate results that are useful in practice are available in the regression setting, based on the combination of the distribution theory of order statistics for residuals and truncation arguments under the normal distribution. Similar results are also available in the multivariate framework, with Mahalanobis distances in place of model residuals, and provide sound statistical thresholds for outlier nomination in finite samples, even of small and moderate sizes [30]. A detailed asymptotic analysis for the FS estimators has been developed only recently in [23,24], but for the univariate regression context only. Their study involves theory for a new class of weighted and marked empirical processes, quantile process theory, and a fixed point argument to describe the iterative nature of the FS algorithm. The main analytic results are an asymptotic representation of the FS residuals, scaled by the estimated variance, and convergence of the corresponding empirical process.

In this paper we deal with the estimators obtained through the single-population multivariate version of the FS, for which no asymptotic result is available yet. We do agree that the ultimate goal should be to study the weak convergence of the empirical process defined through the FS as the algorithm progresses, e.g. by extending the complex analytic machinery of [24] to the multivariate case. However, this difficult task is outside the scope of this paper and is left for further research. Our goal in the present work is slightly less ambitious. The multivariate FS estimators are usually assumed to be consistent and robust, following intuition and empirical experience, but formal proofs of such properties are still lacking. Our purpose is to fill the gap and to provide justification for such statements. Our asymptotic results thus put the FS approach for multivariate data on a solid statistical ground, which formally motivates its use in robust applied statistics and provides justification for its very good diagnostic properties. Furthermore, our proofs of consistency and robustness are important because they allow us to compare the FS estimators (3) and (4) with other well known multivariate high-breakdown estimators, for which similar properties have been established in the past. These include the Minimum Covariance Determinant and its reweighted version [14,25,15,10,11], S-estimators [17,20,35], Projection estimators [28,37,34] and Trimmed Likelihood estimators [16]. Some preliminary comparisons are provided in the paper, while more extensive theoretical and empirical results will be given elsewhere.

A problem related to the one that motivates our work was also considered by García-Escudero and Gordaliza [19], who derived the asymptotic distribution of the so-called radius process. This process is defined by trimmed Mahalanobis distances similar to (5) below, when the trimming level $1-\gamma$ varies in $(0, 1]$. However, an important difference is that in the radius process the multivariate estimators of location and scatter are computed only once and for all; then, they are kept fixed when the trimming level changes. This makes the radius process a valuable tool for the purpose of multivariate outlier detection, when sufficiently "good" robust parameter estimators already exist. On the other hand, the adaptive nature of the FS implies that the fitting subset changes with trimming level. New location and scatter estimators are thus defined at each step of the FS. These estimators, as well as the corresponding ellipsoids in the multivariate normal model, are dependent and the degree of dependence is unknown. We conjecture that, even with this additional degree of dependence at subsequent steps of the FS, the corresponding radius process is weakly convergent, but a formal proof of this statement is lacking. Therefore, in this paper we limit ourselves to analyze the pointwise asymptotic properties of the FS estimators, when they are computed for a finite sequence of steps. The strong consistency and asymptotic equicontinuity properties that we derive are clearly positive results towards our conjecture, and point to weak convergence results similar to those obtained by García-Escudero and Gordaliza [19]. A careful asymptotic analysis of the relationship between the radius process and the empirical process defined through the FS requires techniques that go well beyond the scope of this paper. Nevertheless, we note that our requirement of the existence of "good" robust parameter estimators is less stringent than in the radius process. For the consistency results derived in this paper, it is sufficient to start the FS with consistent estimators (see Assumption 1 in Section 3), while García-Escudero and Gordaliza [19] assume convergence with a rate of $n^{-1/2}$.

Our basic model for the data generating mechanism is the multivariate normal distribution

$$M_0 : y_i \sim N_v(\mu, \Sigma), \quad i = 1, \dots, n, \tag{1}$$