Journal of Multivariate Analysis 100 (2009) 876–887

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Multivariate generalized S-estimators

E. Roelant [a,*], S. Van Aelst [a], C. Croux [b]

[a] Department of Applied Mathematics and Computer Science, Ghent University - UGent, Krijgslaan 281-S9, B-9000 Gent, Belgium

[b] Katholieke Universiteit Leuven, Faculty of Economics and Business and Leuven Statistical Research Centre, Naamsestraat 69, B-3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 14 February 2008
Available online 7 September 2008

AMS subject classifications:
62F40
62F35
62J05

Keywords:
Bootstrap
Efficiency
Multivariate regression
Robustness

ABSTRACT

In this paper we introduce generalized S-estimators for the multivariate regression model. This class of estimators combines high robustness and high efficiency. They are defined by minimizing the determinant of a robust estimator of the scatter matrix of differences of residuals. In the special case of a multivariate location model, the generalized S-estimator has the important independence property, and can be used for high breakdown estimation in independent component analysis. Robustness properties of the estimators are investigated by deriving their breakdown point and the influence function. We also study the efficiency of the estimators, both asymptotically and at finite samples. To obtain inference for the regression parameters, we discuss the fast and robust bootstrap for multivariate generalized S-estimators. The method is illustrated on a real data example.

1. Introduction

In this paper we introduce a new class of estimators for the multivariate regression model, called Generalized S-estimators (GS). Generalized S-estimators are defined by minimizing the determinant of a robust estimator of the scatter matrix of differences of residuals. Using differences instead of the residuals themselves has the advantage that at most models this will lead to an increase in statistical efficiency, while the robustness of the estimators, as measured by their breakdown point, remains the same. The breakdown point of an estimator is the highest possible percentage of outliers than an estimator can withstand. It turns out that it is possible to achieve the highest possible value for the breakdown point, 50%, even when working with differences of residuals. GS-estimators estimate the slope and the scatter matrix of the error terms of the multivariate regression model without needing to estimate the intercept first. The intercept is treated as a nuisance parameter, and the resulting estimator is therefore considered to be "intercept free".

The multivariate regression model encompasses both the multivariate location-scale model, as a multivariate regression model with only an intercept, and the univariate regression model. While GS-estimators were already considered for univariate regression [1], they were not studied yet for the multivariate location-scale model. In the latter model, the "intercept free" property of the GS estimator translates into "location free" estimation. Hence, GS-estimators allow for estimation of scatter while not needing to estimate the location. More important, since the GS-estimator is based on differences, it has the independence property, meaning that when the components of a random vector are independent, the scatter matrix estimate is diagonal [2]. This is not true for S-estimators of scatter in general. The independence property is highly important in independent component analysis (ICA). Briefly, the ICA problem consists of finding an original random vector with independent components when only an unknown linear mixture is observed [3]. Oja et al. [4] proposed a method for ICA that is based on the use of two different scatter matrices that are required to have the independence property; see

* Corresponding author.
E-mail address: Ella.Roelant@ugent.be (E. Roelant).

0047-259X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.09.002

also Tyler et al. [2]. By using the GS-estimator, a high breakdown approach to robust ICA is obtained. Other scatter matrix estimators, based on differences of observations were proposed by [5,6]. They are of the M-type and their breakdown point decreases with the dimension [7], and thus do not have a high degree of robustness. The GS-estimators are estimators that can have a 50% breakdown point and at the same time the independence property.

Consider the multivariate linear regression model given by

$$\mathbf{y} = \alpha + \mathcal{B}^{\mathrm{T}}\mathbf{u} + \epsilon \tag{1}$$

where $\mathbf{u}$ is the p-variate predictor, $\mathbf{y}$ the q-variate response and ϵ the q-variate error term which has center zero and a positive definite scatter matrix Σ. The unknown parameters $\theta = (\alpha, \mathcal{B}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{(p+1)\times q}$ and $\Sigma \in \mathbb{R}^{q\times q}$ are to be estimated from the observations $\mathcal{Z}_n = \{\mathbf{z}_i := (\mathbf{x}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}} = (1, \mathbf{u}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}, i = 1, \ldots, n\} \subset \mathbb{R}^{p+q+1}$. The classical estimator for this model is the least squares estimator, but it is well-known that this estimator can be highly influenced by outliers.

In the univariate regression case a lot of research has been done to construct more robust estimators. Classes of robust estimators in this setting include M-estimators [8], least median of squares and least trimmed squares estimators [9], S-estimators [10], MM-estimators [11], CM-estimators [12] and τ-estimators [13]. Croux et al. [1] introduced a class of regression estimators, called generalized S-estimators or GS-estimators. While an S-estimator of regression minimizes an M-estimator of scale of the residuals, a GS-estimator minimizes an M-estimator of scale applied on the pairwise differences of the residuals, instead of on the residuals themselves. It has been shown that for bounded loss functions these univariate GS-estimators have nice properties such as a high breakdown point and a higher efficiency than the original S-estimators. Moreover, they do not require the assumption of symmetric errors (see also [14–16]). In the univariate regression model we also mention the rank-based regression estimates of [17], which are also based on the differences of the residuals. In this paper, we extend the definition of GS-estimates to multivariate regression.

Recently, several robust estimators for multivariate regression have been introduced. Methods based on robust estimators for multivariate location and scatter applied to the joint distribution of responses and explanatory variables have been proposed by Ollila et al. [18] using sign covariance matrices, Ollila et al. [19] using rank covariance matrices and [20] using the minimum covariance determinant estimator. An alternative approach is to define a robust regression estimator by minimizing a robust estimate of the covariance matrix of the residuals. Agulló et al. [21] proposed the multivariate least trimmed squares estimator, Van Aelst and Willems [22] considered multivariate regression S-estimators, while [23] introduced τ-estimators for multivariate regression. Also note that the idea of univariate least absolute deviation estimation has been extended to the multivariate case by [24–26]. All these procedures, however, are not based on differences of residuals.

GS-estimators generalize S-estimators in the sense that they are identical to S-estimators, but computed from differences $\mathbf{z}_i - \mathbf{z}_j$, instead of using the original observations $\mathbf{z}_i$. This is the same idea as used in the definition of generalized L-, M-, and R-estimators [27] for the location model. It should be stressed that S-estimators are not included in the class of GS-estimators. Another suitable name for GS-estimators may be "Symmetrized S-estimators".

The remainder of the paper is organized as follows. In Section 2 we introduce the multivariate regression GS-estimators and determine their breakdown point. Section 3 describes the algorithm for computing the GS-estimators. In Section 4 we define the functional form of the estimator. We show that the GS-functional is Fisher-consistent if the differences of the errors have an elliptical distribution. We also derive the influence function of the GS-functional. Asymptotic variances and corresponding efficiencies are given in Section 5. Section 6 discusses the fast and robust bootstrap method for GS-estimators. Section 7 presents a real data example and Section 8 concludes. Proofs are omitted and can be found in the technical report [28].

2. Definition and breakdown point

We now define Generalized S-estimators for the multivariate regression model given in (1).

Definition 1. Let $\mathcal{Z}_n = \{\mathbf{z}_i := (\mathbf{x}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}} = (1, \mathbf{u}_i^{\mathrm{T}}, \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}, i = 1, \ldots, n\} \subset \mathbb{R}^{p+q+1}$. The GS-estimates of multivariate regression $(\widehat{\mathcal{B}}_n, \widehat{\Sigma}_n)$ minimizes among all $(B, C) \in \mathbb{R}^{p\times q} \times PDS(q)$, with $PDS(q)$ the set of positive definite symmetric $q \times q$ matrices, the determinant $|C|$, subject to the condition

$$\binom{n}{2}^{-1} \sum_{i<j} \rho([(\mathbf{r}_i - \mathbf{r}_j)^{\mathrm{T}} C^{-1} (\mathbf{r}_i - \mathbf{r}_j)]^{1/2}) = k \tag{2}$$

where $\mathbf{r}_i = \mathbf{y}_i - B^{\mathrm{T}}\mathbf{u}_i - \alpha$.

Note that the objective function does not depend on the intercept α. The constant k can be chosen as $k = E_{F\times F}[\rho(\|\epsilon_1 - \epsilon_2\|)]$, which ensures consistency at the model with error distribution F (see Section 4). The choice $\rho(u) = u^2$ yields the non-robust least squares (LS) estimator. To obtain robust estimates, we impose the following properties on the loss function ρ:

- ρ is twice continuously differentiable and $\rho(0) = 0$
- ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$ for some $c < \infty$.