Computational Statistics and Data Analysis 99 (2016) 148–170

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A practical approximation algorithm for the LTS estimator

David M. Mount [a,*], Nathan S. Netanyahu [b,c], Christine D. Piatko [d], Angela Y. Wu [e], Ruth Silverman [c]

[a] Department of Computer Science, University of Maryland, College Park, MD, USA
[b] Department of Computer Science, Bar-Ilan University, Ramat-Gan 52900, Israel
[c] Center for Automation Research, University of Maryland, College Park, MD, USA
[d] The Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA
[e] Department of Computer Science, American University, Washington, DC, USA

ARTICLE INFO

Article history:
Received 12 August 2013
Received in revised form 26 January 2016
Accepted 27 January 2016
Available online 3 February 2016

Dedicated to the memory of our dear friend and longtime colleague, Ruth Silverman

Keywords:
Robust estimation
Linear estimation
Least trimmed squares
Approximation algorithms
Computational geometry

ABSTRACT

The linear least trimmed squares (LTS) estimator is a statistical technique for fitting a linear model to a set of points. It was proposed by Rousseeuw as a robust alternative to the classical least squares estimator. Given a set of n points in $\mathbb{R}^d$, the objective is to minimize the sum of the smallest 50% squared residuals (or more generally any given fraction). There exist practical heuristics for computing the linear LTS estimator, but they provide no guarantees on the accuracy of the final result. Two results are presented. First, a measure of the numerical condition of a set of points is introduced. Based on this measure, a probabilistic analysis of the accuracy of the best LTS fit resulting from a set of random elemental fits is presented. This analysis shows that as the condition of the point set improves, the accuracy of the resulting fit also increases. Second, a new approximation algorithm for LTS, called Adaptive-LTS, is described. Given bounds on the minimum and maximum slope coefficients, this algorithm returns an approximation to the optimal LTS fit whose slope coefficients lie within the given bounds. Empirical evidence of this algorithm's efficiency and effectiveness is provided for a variety of data sets.

1. Introduction

Consider an n-element point set $P = \{p_1, \ldots, p_n\}$, where $p_i = (x_{i,1}, \ldots, x_{i,d-1}, y_i) \in \mathbb{R}^d$. In standard linear regression with intercept, we assume the model

$$y_i = \sum_{j=1}^{d-1} \beta_j x_{i,j} + \beta_d + e_i, \quad \text{for } i = 1, \ldots, n,$$

where $(x_{i,1}, \ldots, x_{i,d-1})$ are the given independent variables, the y_i's are the given dependent variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ is the unknown coefficient vector, and the e_i's are the errors. We refer to $(\beta_1, \ldots, \beta_{d-1})$ as the *slope coefficients* and β_d as the *intercept*. Given an estimator $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$, define the *ith residual* to be $r_i(\hat{\boldsymbol{\beta}}, P) = y_i - \left(\sum_{j=1}^{d-1} \hat{\beta}_j x_{i,j} + \hat{\beta}_d\right)$. Let $r_{[i]}(\hat{\boldsymbol{\beta}}, P)$ denote the ith smallest residual in terms of absolute value.

* Corresponding author.
E-mail addresses: mount@cs.umd.edu (D.M. Mount), nathan@cs.biu.ac.il, nathan@cfar.umd.edu (N.S. Netanyahu), christine.piatko@jhuapl.edu (C.D. Piatko), awu@american.edu (A.Y. Wu).

http://dx.doi.org/10.1016/j.csda.2016.01.016
0167-9473/

Robust estimators (see, e.g., Rousseeuw and Leroy, 1987) have been introduced in order to eliminate sensitivity to outliers, that is, points that fail to follow the linear pattern of the majority of the points. The basic measure of the robustness of an estimator is its *breakdown point*, that is, the fraction (up to 50%) of outlying data points that can corrupt the estimator arbitrarily. One of the most widely studied robust linear estimators is Rousseeuw's *least median of squares estimator* (LMS) (Rousseeuw, 1984), which is defined to be the estimator that minimizes the median squared residual. More generally, given an integer *coverage value* h, the objective is to find the hyperplane that minimizes the hth smallest squared residual. A number of papers, both practical and theoretical, have been devoted to computing this estimator in the plane and in higher dimensions (see, e.g., Souvaine and Steele, 1987; Edelsbrunner and Souvaine, 1990; Mount et al., 2000, 2004, 2007; Bernholt, 2005; Erickson et al., 2006).

It has been observed by Rousseeuw and Leroy (1987) that LMS may not be the best estimator from the perspective of statistical properties. They argue in support of the *least trimmed squares* (or LTS) linear estimator (Rousseeuw, 1984). Given an n-element point set P and a coverage $h \leq n$, it is defined to be the estimator that minimizes the *sum* (as opposed to the maximum) of the h smallest squared residuals. For the sake of preserving scale, we convert this into a quantity that more closely resembles a standard deviation. More formally, given a non-vertical hyperplane $\hat{\boldsymbol{\beta}}$ define its *LTS cost* with respect to P and h to be

$$\Delta_{\hat{\boldsymbol{\beta}}}(P, h) = \left(\frac{1}{h-1} \sum_{i=1}^{h} r^2_{[i]}(\hat{\boldsymbol{\beta}}, P) \right)^{1/2}.$$

Note that this measure is scale equivariant, and minimizing it is equivalent to minimizing the sum of squared residuals. The *LTS estimator* is the coefficient vector of minimum LTS cost, which we denote throughout by $\boldsymbol{\beta}^*(P, h)$. Let $\Delta^*(P, h)$ denote the associated LTS cost. When P and h are clear from context, we refer to these simply as $\boldsymbol{\beta}^*$ and Δ^*, respectively. The *LTS problem* is that of computing $\boldsymbol{\beta}^*$ from P and h. We refer to the points having the h smallest squared residuals as *inliers* and the remaining points as *outliers*. This generalizes the ordinary least squares estimator (when $h = n$). It is customary to set $h = \lfloor (n + d + 1)/2 \rfloor$ for outlier detection (Rousseeuw and Leroy, 1987). In practice, h may be set to some constant fraction of n based on the expected number of outliers. Since h determines the percentage of squared residuals that will be trimmed from the sum to be minimized, it sometimes is referred to as the *trimming option*. The statistical properties of LTS are analyzed in Rousseeuw and Leroy (1987) and Rousseeuw (1984).

The computational complexity of LTS is less well understood than that of LMS. Various exact algorithms have been presented, which are based on variants of branch-and-bound search and efficient incremental updates (Agulló, 2001; Hofmann and Kontoghiorghes, 2010; Hofmann et al., 2010). Unfortunately, these algorithms are practical only for fairly small point sets. Hössjer (1995) presented an $O(n^2 \log n)$ algorithm for LTS in $\mathbb{R}^2$ based on plane sweep. In a companion paper (Mount et al., 2014) we presented an exact algorithm for LTS in $\mathbb{R}^2$, which runs in $O(n^2)$ time. We also presented an algorithm for $\mathbb{R}^d$, for any $d \geq 3$, which runs in $O(n^{d+1})$ time. (Throughout, we assume that d is a fixed constant.) For large n these running times may be unacceptably high, even in spaces of moderate dimension.

Given these relatively high running times, it is natural to consider whether this problem can be solved approximately. There are a few possible ways to formulate LTS as an approximation problem, either by approximating the residual, by approximating the quantile, or both. The following formulations were introduced in Mount et al. (2014). The approximation parameters ε_r and ε_q denote the allowed residual and quantile errors, respectively.

Residual Approximation: The requirement of minimizing the sum of squared residuals is relaxed. Given $0 < \varepsilon_r$, an *ε_r-residual approximation* is any hyperplane $\boldsymbol{\beta}$ such that

$$\Delta_{\boldsymbol{\beta}}(P, h) \leq (1 + \varepsilon_r)\, \Delta^*(P, h).$$

Quantile Approximation: Much of the complexity of LTS arises because of the requirement that *exactly* h points be covered. We can relax this requirement by introducing a parameter $0 < \varepsilon_q < h/n$ and requiring that the fraction of inliers used is smaller by ε_q. Let $h^- = h - \lfloor n\varepsilon_q \rfloor$. An *$\varepsilon_q$-quantile approximation* is any hyperplane $\boldsymbol{\beta}$ such that

$$\Delta_{\boldsymbol{\beta}}(P, h^-) \leq \Delta^*(P, h).$$

Hybrid Approximation: The above approximations can be merged into a single approximation. Given ε_r and ε_q as in the previous two approximations, let h^- be as defined above. An *$(\varepsilon_r, \varepsilon_q)$-hybrid approximation* is any hyperplane $\boldsymbol{\beta}$ such that

$$\Delta_{\boldsymbol{\beta}}(P, h^-) \leq (1 + \varepsilon_r)\Delta^*(P, h).$$

Note that approximating the LTS cost does not imply that the optimum slope coefficients themselves are well approximated. Computing an approximation to the slope coefficients seems to be difficult. In particular, for some pathological point sets there may be many solutions that have nearly the same LTS costs but very different slope coefficients. Consider, for example, fitting a plane to a set of points uniformly distributed within a sphere. In an earlier paper (Mount et al., 2014), we presented an approximation algorithm for LTS whose execution time is roughly $O(n^d/h)$. In the same paper, we presented asymptotic lower bounds for computing the LTS and the related LTA (least trimmed absolute value) estimators.