Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 2461 – 2486

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

# A practical approximation algorithm for the LMS line estimator[☆]

David M. Mount[a,*,1], Nathan S. Netanyahu[b,c], Kathleen Romanik[d,2], Ruth Silverman[c], Angela Y. Wu[e]

[a] Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA
[b] Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel
[c] Center for Automation Research, University of Maryland, College Park, MD, USA
[d] White Oak Technologies, Inc., Silver Spring, MD, USA
[e] Department of Computer Science, American University, Washington, DC, USA

Received 1 March 2005; accepted 25 August 2006
Available online 2 October 2006

## Abstract

The problem of fitting a straight line to a finite collection of points in the plane is an important problem in statistical estimation. Robust estimators are widely used because of their lack of sensitivity to outlying data points. The least median-of-squares (LMS) regression line estimator is among the best known robust estimators. Given a set of $n$ points in the plane, it is defined to be the line that minimizes the median squared residual or, more generally, the line that minimizes the residual of any given quantile $q$, where $0 < q \leqslant 1$. This problem is equivalent to finding the strip defined by two parallel lines of minimum vertical separation that encloses at least half of the points.

The best known exact algorithm for this problem runs in $O(n^2)$ time. We consider two types of approximations, a *residual approximation*, which approximates the vertical height of the strip to within a given error bound $\varepsilon_r \geqslant 0$, and a *quantile approximation*, which approximates the fraction of points that lie within the strip to within a given error bound $\varepsilon_q \geqslant 0$. We present two randomized approximation algorithms for the LMS line estimator. The first is a conceptually simple quantile approximation algorithm, which given fixed $q$ and $\varepsilon_q > 0$ runs in $O(n \log n)$ time. The second is a practical algorithm, which can solve both types of approximation problems or be used as an exact algorithm. We prove that when used as a quantile approximation, this algorithm's expected running time is $O(n \log^2 n)$. We present empirical evidence that the latter algorithm is quite efficient for a wide variety of input distributions, even when used as an exact algorithm.

*Keywords:* Least median-of-squares regression; Robust estimation; Line fitting; Approximation algorithms; Randomized algorithms; Line arrangements

[☆] A preliminary version of this paper appeared in Proceedings of the Eighth Annual ACM–SIAM Symposium on Discrete Algorithms, 1997, pp. 473–482.

[*] Corresponding author. Tel.: +1 301 4052704; fax: +1 301 4056707.

*E-mail addresses:* mount@cs.umd.edu (D.M. Mount), nathan@cs.biu.ac.il, nathan@cfar.umd.edu (N.S. Netanyahu), kromanik@comcast.net (K. Romanik), ruthsilverman@verizon.net (R. Silverman), awu@american.edu (A.Y. Wu).

[1] This material is based upon work supported by the National Science Foundation under Grant no. 0098151.

[2] This research was carried out in part while this author was at the Center for Automation Research, University of Maryland, College Park, MD, USA.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.08.033

## 1. Introduction

The problem of fitting a straight line to a finite collection of points in the plane is an important problem in statistical estimation. Robust estimators are of particular interest because of their lack of sensitivity to outlying data 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. Rousseeuw's least median-of-squares (LMS) regression (line) estimator (Rousseeuw, 1984) is among the best known 50% breakdown-point estimators.

The *LMS line estimator* (with intercept) is defined formally as follows. Consider a set $S$ of $n$ points $(x_i, y_i)$ in $\Re^2$. The problem is to estimate the parameter vector $\theta^* = (\theta_1^*, \theta_2^*)$ which best fits the data by the linear model

$$y_i = x_i\theta_1^* + \theta_2^* + e_i, \quad i = 1, \ldots, n,$$

where $(e_1, \ldots, e_n)$ are the (unknown) errors. Given an arbitrary parameter vector $(\theta_1, \theta_2)$, let $r_i = y_i - (x_i\theta_1 + \theta_2)$ denote the $i$th residual. The LMS estimator is defined to be the parameter vector that minimizes the median of the squared residuals. This should be contrasted with ordinary least squares (OLS), which minimizes the sum of the squared residuals. Intuitively, if less than half of the points are outliers, then these points cannot adversely affect the median squared residual.

In addition to having a high breakdown-point, the LMS estimator is regression-, scale-, and affine-equivariant, i.e., its estimate transforms "properly" under these types of transformations. (See Rousseeuw and Leroy, 1987, pp. 116–117, for exact definitions.) The LMS estimator may be used in its own right or as an initial step in a more complex estimation scheme (Yohai, 1987). It has been widely used in numerous applications of science and technology, and is considered a standard technique for robust data analysis. Our main motivation for studying the LMS estimator stems from its usage in computer vision. (See, e.g., Meer et al., 1991; Netanyahu et al., 1997; Stein and Werman, 1992; Stewart, 1996.) For example, Fig. 1 demonstrates the enhanced performance obtained by LMS versus OLS in detecting straight road segments in a noisy aerial image (Netanyahu et al., 1997).

In some applications the number of outliers may differ from 50%, so it is common to generalize the problem definition to minimize the squared residual of a specified quantile. This is also called the least-quantile squared (LQS) estimator (Rousseeuw and Leroy, 1987). In our formulation, we will assume that in addition to the point set $S$, the algorithm is given a *residual quantile* $q$, where $0 < q \leqslant 1$, and it returns the line that minimizes the $\lceil nq \rceil$th smallest squared residual.

Define a *strip* $\sigma = (\ell_1, \ell_2)$ to be the closed region of the plane lying between two parallel lines $\ell_1$ and $\ell_2$. The *vertical height* of a nonvertical strip, height$(\sigma)$, is the length of its intersection with any vertical line. For $0 < q \leqslant 1$, define LMS$(S, q)$ to be the strip of minimum vertical height that encloses at least $\lceil nq \rceil$ points from the set $S$. It is easy to see that the center line of LMS$(S, \frac{1}{2})$ is the LMS regression line, and the height of the strip is twice the median of the absolute residuals. We call LMS$(S, 0.5)$ the *LMS strip*. The algorithms discussed here can be generalized to compute the strip that minimizes the perpendicular width of the strip. This is done by scaling each vertical height by the cosine of the angle corresponding to the slope, as described by Edelsbrunner and Souvaine (1990).

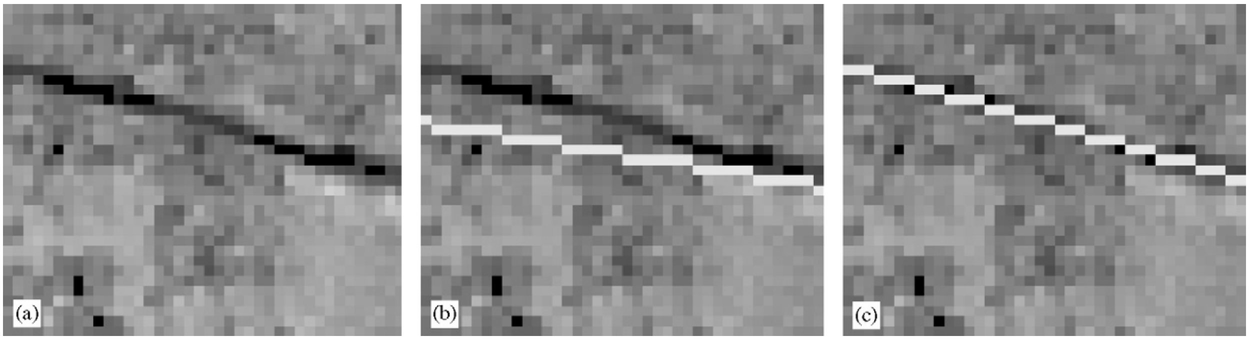

Fig. 1. (a) A road segment with outlying pixels; (b) its OLS line fit; (c) its LMS line fit.