Bias in Robust Estimation Caused by Discontinuities and Multiple Structures

Charles V. Stewart*

Department of Computer Science

Rensselaer Polytechnic Institute

Troy, New York 12180–3590

stewart@cs.rpi.edu

November 27, 1996

Abstract

When fitting models to data containing multiple structures, such as when fitting surface patches to data taken from a neighborhood that includes a range discontinuity, robust estimators must tolerate both gross outliers and pseudo outliers. Pseudo outliers are outliers to the structure of interest, but inliers to a different structure. They differ from gross outliers because of their coherence. Such data occurs frequently in computer vision problems, including motion estimation, model fitting and range data analysis. The focus in this paper is the problem of fitting surfaces near discontinuities in range data.

To characterize the performance of least median of squares, least trimmed squares, M-estimators, Hough transforms, RANSAC, and MINPRAN on this type of data, the "pseudo outlier bias" metric is developed using techniques from the robust statistics literature, and it is used to study the error in robust fits caused by distributions modeling various types of discontinuities. The results show each robust estimator to be biased at small but substantial discontinuities. They also show the circumstances under which different estimators are most effective. Most importantly, the results imply present estimators should be used with care and new estimators should be developed.

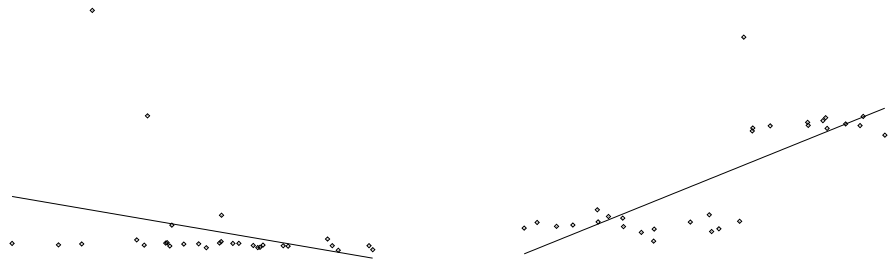
*This paper presents a substantial reformulation and improvement of an earlier version of this work, which was described in [25].

1 Introduction

Robust estimation techniques have been used with increasing frequency in computer vision applications because they have proven effective in tolerating the gross errors (outliers) characteristic of both sensors and low-level vision algorithms. Most often, robust estimators are used when fitting model parameters — e.g. the coefficients of either a polynomial surface, an affine motion model, a pose estimate, or a fundamental matrix — to a data set. For these applications, robust estimators work reliably when the data contain measurements from a single structure, such as a single surface, plus gross errors.

Sometimes, however, the data are more complicated than this, presenting a challenge to robust estimators not anticipated in the robust statistics literature. This complication occurs when the data are measurements from multiple structures while still being corrupted by gross outliers. These structures may be different surfaces in depth measurements or multiple moving objects in motion estimation. Here, the difficulty arises because robust estimators are designed to extract a single fit. Thus, to estimate accurate parameters modeling one of the structures — which one is not important — they must treat the points from all other structures as outliers. After successfully estimating the fit parameters of one structure, the robust estimator may be re-applied, if desired, to estimate subsequent fits after removing the first fit's inliers from the data.

An example using synthetic range data illustrates the potential problems caused by multiple structures. Figure 1 shows (non-robust) linear least-squares fits to data from a single surface and to data from a pair of surfaces forming a step discontinuity. In the single surface example, the least-squares fit is skewed slightly by the gross outliers, but the points from the surface are still generally closer to the fit than the outliers. Thus, the fit estimated by a robust version of least-squares will not be significantly corrupted by these outliers. In the multiple surface example, the least-squares fit is skewed so much that it crosses (or "bridges") the point sets from both surfaces, placing the fit in close proximity to both point sets. Since robust estimators use fit proximity to distinguish inliers and outliers and downgrade the influence of outliers, this raises two concerns about the accuracy of robust fits. First, an estimator that iteratively refines an initial least squares fit will have a local, and potentially



a b

Figure 1: Examples demonstrating the effects of (a) gross outliers and (b) both gross outliers and data from multiple structures on linear least-squares fits.

global, minimum fit that is not far from the initial, skewed fit. This is because points from both surfaces will have both small and large residuals, making it difficult for the estimator to "pull away" from one of the surfaces. Second, and more important, for the robust estimate be the correct fit, thereby treating the points from one surface as inliers and points from the other as outliers, the estimator's objective function must be lower for the smaller inlier set of the correct fit than the larger inlier set of the bridging fit. By varying both the proximity of the two surfaces and the relative sizes of their point sets, all robust estimators studied here can be made to "fail" on this data, producing fits that are heavily skewed.

Motivated by the foregoing discussion, the goal of this paper is to study how effectively robust estimators can estimate fit parameters given a mixture of data from multiple structures. Stating this "pseudo outliers problem" abstractly, to obtain an accurate fit a robust technique must tolerate two different types of outliers: gross outliers and pseudo outliers. Gross outliers are bad measurements, which may arise from specularities, boundary effects, physical imperfections in sensors, or errors in low-level vision computations such as edge detection or matching algorithms. Pseudo outliers are measurements from one or more additional structures. (Without losing generality, inliers and pseudo outliers are distinguished by assuming the inliers are points from the structure contributing the most points and pseudo outliers are points from the other structures.) The coherence of pseudo outliers distinguishes them from gross outliers. Because data from multiple structures are common in vision ap-

plications, robust estimators' performance on this type of data must be understood to use them effectively. Where they prove ineffective, new and perhaps more complicated robust techniques will be needed.[1]

To study the pseudo outliers problem, this paper develops a measure of "pseudo outlier bias" using tools from the robust statistics literature [10, pages 81-95] [12, page 11]. Pseudo outlier bias will measure the distance between a robust estimator's fit to a "target" distribution and its fit to an outlier corrupted distribution. The target distribution will model the distribution of points drawn from a single structure without outliers, and the outlier corrupted mixture distribution [27] will combine distributions modeling the different structures and a gross outlier distribution. The optimal fit is found by applying the functional form of an estimator to these distributions, rather than by applying the estimator's standard form to particular sets of points generated from these distributions. This gives a theoretical measure, avoids the need for extensive simulations, and, most importantly, shows the *inherent* limitations of robust estimators by studying their objective functions independent of their search techniques. The bias of a number of estimators — M-estimators [12, Chapter 7], least median of squares (LMS) [16, 21], least trimmed squares (LTS) [21], Hough transforms [13], RANSAC [7], and MINPRAN [26] — will be studied as the target and mixture distributions vary.

The application for studying the pseudo outliers problem is fitting surfaces to range data taken from the neighborhood of a surface discontinuity. While this is a simple application for studying the pseudo outliers problem, the problem certainly arises in other applications as well — essentially any application where the data could contain multiple structures — and the results obtained here should be used as qualitative predictions of potential difficulties in these applications. In the context of the range data application, three idealized discontinuity models are used to develop mixture distributions: step edges, crease edges and parallel surfaces. Step edges model depth discontinuities, where points from the upper surface of the step are pseudo outliers to the lower surface. Crease edges model surface orientation discontinuities, where points from one side of the crease are pseudo outliers to the other.

[1]Several versions of these techniques actually exist for fitting surfaces to range data. Their effectiveness, however, depends in part on the accuracy of an initial set of robust fits.

Finally, parallel surfaces model transparent or semi-transparent surfaces, where a background surface appears through breaks in the foreground surface, and data from the background are pseudo outliers to the foreground.

A final introductory comment is important to assist in reading this paper. The paper defines the notion of "pseudo outlier bias" using techniques common in mathematical statistics but not in computer vision, most importantly, the "functional form" of a robust estimator. The intuitive meaning of functional forms and their use in pseudo outlier bias are discussed at the start of Section 4, which then proceeds with the main derivations. Readers uninterested in the mathematical details should be able to skip Sections 4.2 through 4.6 and still follow the analysis results.

2 Robust Estimators

This section defines the robust estimators studied. These definitions are converted to functional forms suitable for analysis in Section 4. Because the goal of the paper is to expose inherent limitations of robust estimators, the focus in defining the estimators is their objective functions rather than their optimization techniques. Special cases of iterative optimization techniques where local minima are potentially problematic will be discussed where appropriate.

The data are $(\vec{x}_i, z_i)$, where $\vec{x}_i$ is an image coordinate vector — the independent variable(s) — and z_i is a range value — the dependent variable. Each fit is a function $z = \theta(\vec{x})$, often restricted to the class of linear or quadratic polynomials. The notation $\hat{\theta}(\vec{x})$ indicates the fit that minimizes an estimator's objective function, with $\hat{\theta}$ called the "estimate". Each estimator's objective function evaluates hypothesized fits, $\theta(\vec{x})$, via the residuals, $r_{i,\theta} = z_i - \theta(\vec{x}_i)$.

2.1 M-Estimators

A regression M-estimate [12, Chapter 7] is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_i \rho(r_{i,\theta}/\hat{\sigma}), \tag{1}$$

where $\hat{\sigma}$ is an estimate of the true scale (noise) term, σ, and $\rho(u)$ is a robust "loss" function which grows subquadratically for large $|u|$ to reduce the effect of outliers. (Often, as discussed below, $\hat{\theta}$ and $\hat{\sigma}$ are estimated jointly.) M-estimators are categorized into three types [11] by the behavior of $\psi(u) = \rho'(u)$; one estimator of each type is studied. *Monotone* M-estimators (Figure 2a), such as Huber's [12, Chapter 7], have non-decreasing, bounded $\psi(u)$ functions. *Hard redescenders* (Figure 2b), such as Hampel's [9] [10, page 150], force $\psi(u) = 0$ for $|u| > c$; hence c is a rejection point, beyond which a residual has no influence. *Soft redescenders* (Figure 2c), such as the maximum likelihood estimator of Student's t-distribution [5], do not have a finite rejection point, but force $\psi(u) \to 0$ as $|u| \to \infty$. The three robust loss functions are shown in Figure 2 and in order they are

$$\rho_m(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le c \\ \frac{1}{2}c(2|u| - c), & c < |u| \end{cases} \tag{2}$$

$$\rho_h(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le a \\ \frac{1}{2}a(2|u| - a), & a < |u| \le b \\ \frac{1}{2}a[(|u| - c)^2/(b - c) + (b + c - a)], & b < |u| \le c \\ \frac{1}{2}a(b + c - a), & c < |u| \end{cases} \tag{3}$$

and

$$\rho_s(u) = \frac{1}{2}(1 + f)\log(1 + u^2/f). \tag{4}$$

The ρ functions' constants are usually set to optimize asymptotic efficiency relative to a given target distribution [11] (e.g. Gaussian residuals).

M-estimators typically minimize $\sum \rho(r_{i,\theta}/\hat{\sigma})$ using iterative techniques [11] [12, Chapter 7]. The objective functions of hard and soft redescending M-estimators are non-convex and may have multiple local minima.

In general, $\hat{\sigma}$ must be estimated from the data. Hard-redescending M-estimators often use the median absolute deviation (MAD) [11] computed from the residuals to an initial fit, $\hat{\theta}_0$:

$$\hat{\sigma} = k \ \underset{i}{\text{median}}\{|r_{i,\hat{\theta}_0} - \underset{j}{\text{median}}\{r_{j,\hat{\theta}_0}\}|\}, \tag{5}$$

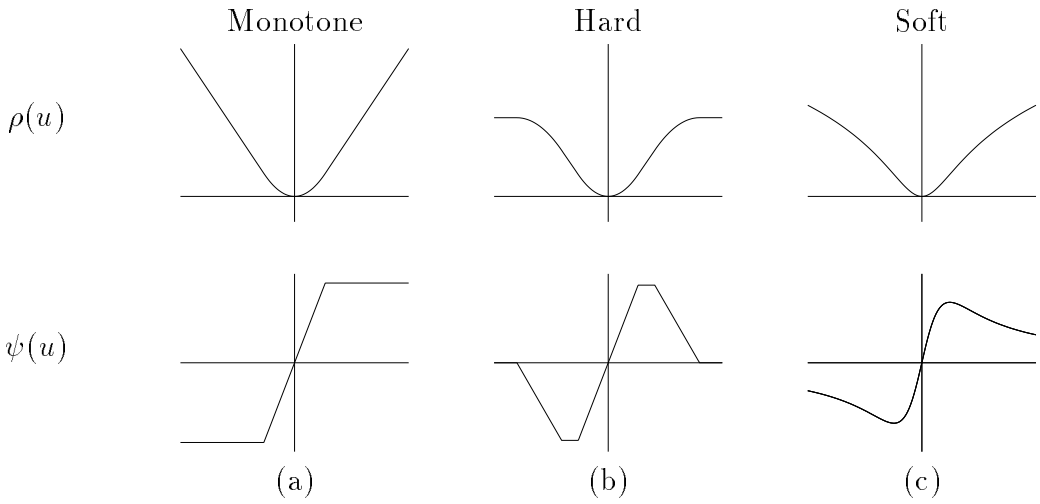


Figure 2: $\rho(u)$ and $\psi(u)$ functions for three M-estimators.

where $k = 1.4826$ for consistency at the normal distribution and $k = 1.14601$ for consistency at Student's t-distribution (when $f = 1.5$). Other M-estimators jointly estimate $\hat{\sigma}$ and $\hat{\theta}$ as

$$(\hat{\theta}, \hat{\sigma}) = \underset{\theta,\sigma}{\operatorname{argmin}} \sum_i \rho(r_{i,\theta}, \sigma). \tag{6}$$

In particular, Huber [12, Chapter 7] uses

$$\rho_m(r_{i,\theta}, \sigma) = [\rho_m(r_{i,\theta}/\sigma) + a]\sigma, \tag{7}$$

where $\rho_m(r_{i,\theta}/\sigma)$ is from equation 2 and a is a tuning parameter; Mirza and Boyer [5] use

$$\rho_s(r_{i,\theta}, \sigma) = \ln \sigma + \rho_s(r_{i,\theta}/\sigma), \tag{8}$$

where $\rho_s(r_{i,\theta}/\sigma)$ is from equation 4.

When fitting surfaces to range data, a different option for obtaining $\hat{\sigma}$ is often used [3]. If σ depends only on the properties of the sensor then $\hat{\sigma}$ may be estimated once and fixed for all data sets. Theoretically, when $\hat{\sigma}$ is fixed, the M-estimators described by equation 1 are no longer true M-estimators since they are not *scale equivariant* [10, page 259]. To reflect this, when $\hat{\sigma}$ is fixed *a priori*, they are called "fixed-scale M-estimators." Both standard M-estimators and fixed-scale M-estimators are studied here.

2.2 Fixed-Band Techniques: Hough Transforms and RANSAC

Hough transforms [13], RANSAC [4, 7], and Roth's primitive extractor [20] are examples of "fixed-band" techniques [20]. For these techniques, $\hat{\theta}$ is the fit maximizing the number of points within $\theta \pm r_b$, where r_b is an inlier bound which generally depends on $\hat{\sigma}$ (i.e. $r_b = c\hat{\sigma}$ for some constant c). Equivalently, viewing fixed-band techniques as minimizing the number of outliers, they become a special case of fixed-scale M-estimators with a simple, discontinuous loss function

$$\rho_f(u) = \begin{cases} 0, & |u| \leq c \\ 1, & |u| > c. \end{cases} \tag{9}$$

Fixed-band techniques search for $\hat{\theta}$ using either random sampling or voting techniques.

2.3 LMS and LTS

Least median of squares (LMS), introduced by Rousseeuw [21], finds the fit minimizing the median of squared residuals. (See [16] for a review.) Specifically, the LMS estimate is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\{\underset{i}{\operatorname{median}}\{(r_{i,\theta})^2\}\}. \tag{10}$$

Most implementations of LMS use random sampling techniques to find an approximate minimum.

Related to LMS and also introduced by Rousseeuw [21] is the least trimmed squares estimator (LTS). The LTS estimate is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{j=1}^{h} (r_\theta^2)_{j:N}. \tag{11}$$

where the $(r_\theta^2)_{j:N}$ are the (non-decreasing) ordered squared residuals of fit θ. Usually $h = \lfloor (N+1)/2 \rfloor$. LTS implementations also use random sampling.

2.4 MINPRAN

MINPRAN searches for the fit minimizing the probability that a fit and a collection of inliers to the fit could be due to gross outliers [24, 26]. It is derived by assuming that relative to

any hypothesized fit $\theta(x)$ the residuals of gross outliers are uniformly distributed[2] in the range $\pm Z_0$. Based on this assumption, the probability that a particular gross outlier could be within $\theta(\vec{x}_i) \pm r$ for $0 \leq r \leq Z_0$ is r/Z_0. Furthermore, if all n points are gross outliers, the probability k or more of them could be within $\theta(\vec{x}) \pm r$ is

$$\mathcal{F}(r,k,n) = \sum_{j=k}^{n} \binom{n}{j} (r/Z_0)^j (1 - r/Z_0)^{n-j}. \tag{12}$$

Given n data points containing an unknown number of gross outliers, MINPRAN evaluates hypothesized fits $\theta(\vec{x})$ by finding the inlier bound, r, and the associated number of points (inliers), $k_{r,\theta}$, within $\pm r$ of $\theta(\vec{x})$, minimizing the probability that the inliers could actually be gross outliers. Thus MINPRAN's objective function in evaluating a particular fit is

$$\min_r \mathcal{F}(r, k_{r,\theta}, n)$$

and MINPRAN's estimate is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}[\min_r \mathcal{F}(r, k_{r,\theta}, n)]. \tag{13}$$

MINPRAN is implemented using random sampling techniques (see [26]).

3 Modeling Discontinuities

The important first step in developing the pseudo outlier bias analysis technique is to model the data taken from near a discontinuity as a probability distribution. Attention here is restricted to discontinuities in one-dimensional structures, since this will be sufficient to demonstrate the limitations of robust estimators.

3.1 Outlier Distributions

To set the context for developing the distributions modeling discontinuities, consider the one-dimensional, outlier corrupted distributions used in the statistics literature to study robust location estimators [10, page 97] [12, page 11]:

$$F = (1 - \varepsilon)F_1 + \varepsilon G$$

[2] MINPRAN has been generalized to any known outlier distribution [26].

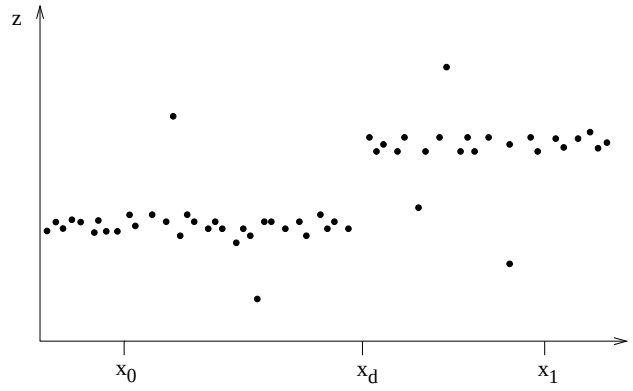


Figure 3: Example data set for points near a step discontinuity.

Here, F_1 is an inlier distribution (also called a "target distribution"), such as a unit variance Gaussian, and G is an outlier distribution, such as a large variance Gaussian or an uniform distribution over a large interval. The parameter ε is the outlier proportion. A set A of N points sampled from this distribution will contain on average εN outliers. Robust location estimators are analyzed using distribution F rather than using a series of point sets sampled from F.

3.2 Mixture Distributions Modeling Discontinuities

The present paper analyzes robust regression estimators by examining their behavior on distributions modeling discontinuities. These mixture distributions [27] will be of the form

$$H = (1 - \varepsilon_o)[\varepsilon_s H_1 + (1 - \varepsilon_s)H_2] + \varepsilon_o H_o. \tag{14}$$

H_1, H_2 and H_o will be inlier, pseudo outlier and gross outlier distributions, respectively, and ε_s and ε_o will control the proportion of points drawn from the three distributions.

To formulate H_1, H_2 and H_o, and to set ε_s and ε_o, consider a set, S, of data points taken from the vicinity of a discontinuity. For example, S might be the points in Figure 3 whose x coordinate falls in the interval $[x_0, x_1]$. H_1 is modeled as a two-dimensional distribution of points (x, z) with x values in an interval $[x_0, x_d]$ — assuming, without losing generality, more points are from the left side of the discontinuity location than the right. (Using a two-dimensional distribution could be counterintuitive since the x values, which may be thought of as image positions at which depth measurements are made, are usually fixed.) Here, x is

treated as uniform in the interval $[x_0, x_d]$, modeling the uniform spacing of image positions.[3] The depth measurement for an inlier is $z = \beta_1(x) + e$, where e is independent noise controlled by the Gaussian density $g(e; \sigma^2)$ with mean 0 and variance σ^2. $\beta_1(x)$ models the ideal curve from which the inliers are measured. The pseudo outlier distribution, H_2, can be defined similarly, with x values uniform in $[x_d, x_1]$ and depth measurements $z = \beta_2(x) + e$. Thus, for both distributions H_1 and H_2, the densities of x and z can be combined to give the joint density

$$h_i(x,z) = \begin{cases} \dfrac{g(z - \beta_i(x); \sigma^2)}{x_{i,1} - x_{i,0}}, & x_{i,0} \le x \le x_{i,1} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

where $i \in \{1, 2\}$ and $x_{i,0}$ and $x_{i,1}$ bound the uniform distribution on the x interval.

For H_o, the distribution of gross outliers in S, again x values are uniformly distributed, but this time over the entire interval $[x_0, x_1]$, and z values are governed by density $g_o(z)$, which will be uniform over a large range. This gives the joint density for a gross outlier:

$$h_o(x,z) = \begin{cases} \dfrac{g_o(z)}{x_1 - x_0}, & x_0 \le x \le x_1 \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

The mixture proportions ε_s and ε_o in (14) are easily specified. ε_o is just the fraction of gross outliers. ε_s is the "relative fraction" of inliers, i.e. the fraction points that are not gross outliers and that are from the inlier side of the discontinuity. Assuming the density of x values does not change across the discontinuity, ε_s is determined by x_d:

$$\varepsilon_s = \frac{x_d - x_0}{x_1 - x_0}. \tag{17}$$

Equivalently, given ε_s, $x_d = x_0 + \varepsilon_s(x_1 - x_0)$. (To distinguish inliers and pseudo outliers, assume $\varepsilon_s > 0.5$.) Notice that the "actual fraction" of inliers is $\varepsilon_1 = (1 - \varepsilon_o)\varepsilon_s$. Depending on which estimator is being analyzed, either the relative or the actual fraction or both will be important.

[3] For any point set sampled from this distribution, the x values will not be uniformly spaced, in general, but their expected values are. This expected behavior is captured when using the distribution itself in the analysis rather than points sets sampled from the distribution.

Using these mixture proportions, the above densities can be combined into a single, mixed, two-dimensional density:

$$h(x,z) = (1-\varepsilon_o)[\varepsilon_s h_1(x,z) + (1-\varepsilon_s)h_2(x,z)] + \varepsilon_o h_o(x,z) \tag{18}$$

Observe that the "target density" is just $h_1(x,z)$ and the "target distribution" is $H_1(x,z)$. The mixture distribution $H(x,z)$ and the target distribution $H_1(x,z)$ can be calculated from $h(x,z)$ and $h_1(x,z)$ respectively.

Using mixture density $h(x,z)$, data can be generated to form step edges and crease edges. The appropriate model is determined by the two curve functions β_1 and β_2. For example, a step edge of height Δz is modeled by setting $\beta_1(x) = c$ and $\beta_2(x) = c + \Delta z$, for some constant c. A crease edge is modeled when β_1 and β_2 are linear functions and $\beta_1(x_d) = \beta_2(x_d)$. Parallel lines with overlapping x domains can be created by using β_1 and β_2 from step edges, but setting $x_{1,0} = x_{2,0} = x_0$ and $x_{1,1} = x_{2,1} = x_1$, and letting ε_s represent the proportion of points from the lower line. In this case, the mixture proportions are divorced from the location of the discontinuity, which has no meaning. Thus, all three desired discontinuities can be modeled.

4 Functional Forms and Mixture Models

To analyze estimators on distributions H, each estimator must be rewritten as a functional, T — a mapping from the space of probability distributions to the space of possible estimates.

This section derives functional forms of the robust estimators defined in Section 2. It starts, in Section 4.1 by giving intuitive insight. Then, Section 4.2 introduces functional forms and empirical distributions on a technical level, using univariate least-squares location estimates as an example. Next, Section 4.3 derives several important distributions needed in the functionals. The remaining sections derive the required functionals. Readers uninterested in the technical details should read only Section 4.1 and then skip ahead to Section 5.

4.1 Intuition

To illustrate what it means for a functional T to be applied to a distribution H, consider least-squares regression. When applied to a set containing points (x_i, z_i), the least-squares objective function is $\sum_i [z_i - \theta(x_i)]^2 = \sum_i r_{i,\theta}^2$, which is proportional to the second moment of the residuals conditioned on θ, and the least squares estimate is the fit $\hat{\theta}$ minimizing this conditional second moment. A similar second moment, conditioned on θ, may be calculated for distribution $H(x, z)$, and the fit $\hat{\theta}$ minimizing this conditional second moment may be found. This is the least-squares regression functional. The functional form of an M-estimator, by analogy, returns the fit minimizing a robust version of the second moment of the conditional residual distribution calculated from H. Intuitions about the functional forms of other estimators are similar.

The estimate $T(H)$ can be used to represent or characterize the estimator's performance on point sets sampled from H. Although the robust fit to any particular point set may differ from $T(H)$, if $T(H)$ is skewed by the pseudo and gross outliers, then the fit to the point set will likely be skewed as well. Indeed, when an estimator's minimization technique is an iterative search, the skew may be worse than that of $T(H)$ because it may stop at a local minimum.

4.2 One-Dimensional Location Estimators

To introduce functional forms on a more technical level, this section examines the least-squares location estimate for univariate data. For a finite sample $\{x_1, \ldots, x_n\}$, the location estimate is

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_i (x_i - \theta)^2 = \frac{1}{n} \sum_i x_i, \tag{19}$$

which is the sample mean or expected value. The functional form of this is the location estimate of the distribution F from which the x_i's are drawn:

$$T_{loc}(F) = \hat{\theta} = \operatorname*{argmin}_{\theta} \int (x - \theta)^2 \, dF = \operatorname*{argmin}_{\theta} \int (x - \theta)^2 f(x) \, dx = \int x f(x) \, dx, \tag{20}$$

the population mean or expected value.

The functional form of the location estimate is derived from the sample location estimate by writing the latter in terms of the "empirical distribution" of the data, denoted by F_n, and then replacing F_n with F, the actual distribution. The *empirical density* of $\{x_1, \ldots, x_n\}$ is

$$f_n(x) = \frac{1}{n} \sum_i \delta(x - x_i).$$

where $\delta(\cdot)$ is the Dirac delta function, and the *empirical distribution* is

$$F_n(x) = \frac{1}{n} \sum_i u(x - x_i),$$

where $u(\cdot)$ is the unit step function. When the x_i's are independent and identically distributed, F_n converges to F as $n \to \infty$. The least squares location estimate is written in terms of the empirical density by using the sifting property of the delta function [8, page 56]:

$$\begin{aligned}
\operatorname*{argmin}_\theta \frac{1}{n} \sum_i (x_i - \theta)^2 &= \operatorname*{argmin}_\theta \frac{1}{n} \sum_i \int (x - \theta)^2 \delta(x - x_i)\, dx \\
&= \operatorname*{argmin}_\theta \int (x - \theta)^2 \frac{1}{n} \sum_i \delta(x - x_i)\, dx \\
&= \operatorname*{argmin}_\theta \int (x - \theta)^2 f_n(x)\, dx
\end{aligned}$$

Replacing f_n with the population density $f(x) = dF/dx$ yields the functional form of the location estimate as desired (20).

4.3 Residual Distributions and Empirical Distributions

Before deriving functional forms for the robust regression estimators, the mixture distribution $H(x, z)$ must be rewritten in terms of the distribution of residuals relative to a hypothesized fit, θ. This is because the estimators' objective functions depend directly on residuals r and only indirectly on points (x, z). In addition, several empirical versions of this "residual distribution" are needed.

Two different residual distributions are required: one for signed residuals and one for their absolute values. Let the distribution and density of signed residuals be $F^s(r|\theta, H)$ and $f^s(r|\theta, H)$ (including H in the notation to make explicit the dependence on the mixture distribution). These are easily seen to be (Figure 4a)

$$F^s(r|\theta, H) = \int_{x_0}^{x_1} \int_{-\infty}^{\theta(x)+r} h(x, z)\, dz\, dx, \tag{21}$$

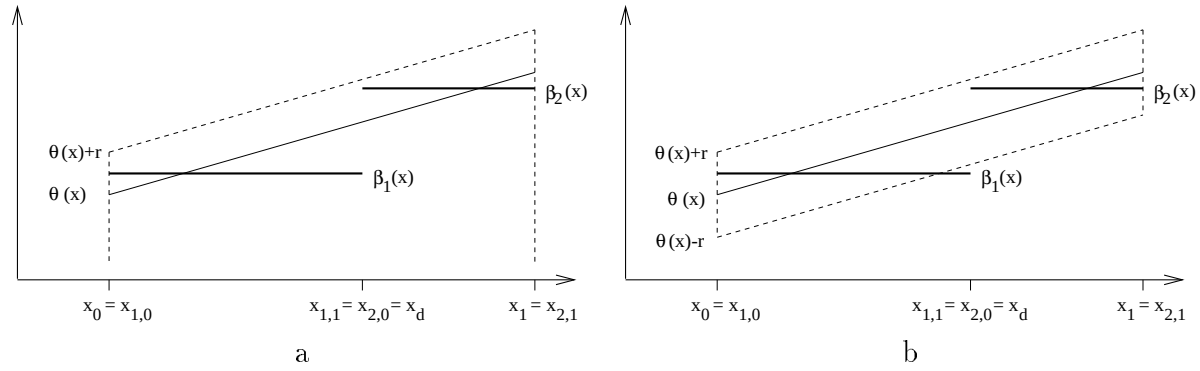


Figure 4: The cumulative distribution of residual r relative to fit $\theta(x)$ is the integral of the point densities, h_1 and h_2, from the curves and from the gross outlier density, h_o, over the region bounded above by $\theta(x) + r$, bounded on the sides by $x = x_0$ and $x = x_1$, and either (a) unbounded below for signed residuals or (b) bounded below by $\theta(x) - r$ for absolute residuals. Both figures show the region of integration for functions β_i and x boundaries modeling a step edge.

and

$$f^s(r|\theta, H) = dF^s(r|\theta, H)/dr = \int_{x_0}^{x_1} h(x, \theta(x)+r)\, dx. \tag{22}$$

Let the distribution and density of absolute residuals be $F^a(r|\theta, H)$ and $f^a(r|\theta, H)$, where $r \geq 0$. These are (Figure 4b)

$$F^a(r|\theta, H) = \int_{x_0}^{x_1} \int_{\theta(x)-r}^{\theta(x)+r} h(x, z) dz\, dx, \tag{23}$$

and

$$f^a(r|\theta, H) = dF^a(r|\theta, H)/dr = \int_{x_0}^{x_1} [h(x, \theta(x)+r) + h(x, \theta(x)-r)]\, dx. \tag{24}$$

Appendix A evaluates these integrals. Replacing h with h_1 in the above equations yields the residual distributions and densities for the target (inlier) distribution.

Now, several empirical distributions are needed below. First, given n points (x_i, z_i) sampled from $h(x, z)$, the empirical density of the data is simply

$$h_n(x, z) = \frac{1}{n} \sum_i \delta(x - x_i, z - z_i).$$

(h_n should not be confused with h_i from equation 15). Next, the empirical density of the signed residuals follows from $h_n(x,z)$ using the sifting property of the δ function [8, page 56]:

$$\begin{aligned} f_n^s(r|\theta,H_n) &= \int_{-\infty}^{\infty} h_n(x,\theta(x)+r)\,dx \\ &= \int_{-\infty}^{\infty} \frac{1}{n}\sum_i \delta(x-x_i,\theta(x)+r-z_i)\,dx \\ &= \frac{1}{n}\sum_i \delta(\theta(x_i)+r-z_i) \end{aligned} \tag{25}$$

Finally, the empirical distribution of the absolute residuals is

$$F_n^a(r|\theta,H_n) = \int_{-r}^{r} f_n^s(y|\theta)\,dy. \tag{26}$$

4.4 M-Estimators and Fixed-Band Techniques

The functionals for the robust regression estimators can now be derived, starting with that of fixed-scale M-estimators. The first step is to write equation 1 in a slightly modified form, which does not change the estimate:

$$\hat{\theta} = \operatorname*{argmin}_\theta \frac{1}{n}\sum_i \rho(r_{i,\theta}/\hat{\sigma}),$$

Next, writing this in terms of the empirical distribution produces

$$\begin{aligned} \operatorname*{argmin}_\theta \frac{1}{n}\sum_i \rho(r_{i,\theta}/\hat{\sigma}) &= \operatorname*{argmin}_\theta \frac{1}{n}\sum_i \rho((z_i-\theta(x_i))/\hat{\sigma}) \\ &= \operatorname*{argmin}_\theta \frac{1}{n}\sum_i \iint \rho((z-\theta(x))/\hat{\sigma})\ \delta(x-x_i,z-z_i)\,dx\,dz \\ &= \operatorname*{argmin}_\theta \iint \rho((z-\theta(x))/\hat{\sigma})\ h_n(x,z)\,dx\,dz \end{aligned}$$

Replacing the empirical density $h_n(x,z)$ with the mixture density $h(x,z)$ yields

$$T_\rho(H) = \operatorname*{argmin}_\theta \iint \rho((z-\theta(x))/\hat{\sigma})\ h(x,z)\,dx\,dz.$$

The change of variables $r = z - \theta(x)$ simplifies things further,

$$\begin{aligned} T_\rho(H) &= \operatorname*{argmin}_\theta \int \rho(r/\hat{\sigma}) \int h(x,\theta(x)+r)\,dx\,dr \\ &= \operatorname*{argmin}_\theta \int \rho(r/\hat{\sigma})\ f^s(r|\theta,H)\,dr. \end{aligned} \tag{27}$$

This is the fixed-scale M-estimator functional. Substituting equations 2, 3 and 4 gives functionals T_{ρ_m}, T_{ρ_h}, and T_{ρ_s} respectively for the M-estimators studied here.

For the M-estimators that jointly estimate $\hat{\theta}$ and $\hat{\sigma}$ (see equations 7 and 8), the functional is obtained by replacing $\rho(r/\hat{\sigma})$ with $\rho(r,\sigma)$ in equation 27, producing

$$T_{\rho,s}(H) = \underset{\theta}{\operatorname{argmin}} \int \rho(r,\sigma)\, f^s(r|\theta,H)\, dr. \tag{28}$$

Finally, recalling that fixed-band techniques are special cases of fixed-scale M-estimators, their functional is obtained by substituting equation 9 into equation 27, yielding

$$\begin{aligned} T_b(H) &= \underset{\theta}{\operatorname{argmin}} \left[\int_{-\infty}^{-r_b} f^s(r|\theta,H))\, dr + \int_{r_b}^{\infty} f^s(r|\theta,H))\, dr \right] \\ &= \underset{\theta}{\operatorname{argmin}} \left[1 - F^a(r_b|\theta,H)\right] \end{aligned} \tag{29}$$

Observe that $[1 - F^a(r_b|\theta,H)]$ is the expected fraction of outliers.

4.5 LMS and LTS

Deriving the functional equivalent to LMS requires first deriving the cumulative distribution of the squared residuals and then writing the median in terms of the inverse of this distribution. Defining $y = r^2$, the empirical distribution of squared residuals is

$$F_{n,y}(y|\theta,H) = F_n^a(\sqrt[+]{y}|\theta,H),$$

since it is simply the percentage of points whose absolute residuals relative to fit θ are less than $\sqrt[+]{y}$. Now,

$$\text{median}\{(r_{i,\theta})^2\} = F_{n,y}^{-1}(1/2|\theta,h), \tag{30}$$

In other words, the median is the inverse of the cumulative, evaluated at $1/2$.[4] This is the standard functional form of the median [10, page 89]. Substituting equation 30 in 10 and

[4]When LMS is implemented using random sampling where p points are chosen to instantiate a fit, the median residual is taken from among the remaining $n-p$ points. To reflect this, the $1/2$ in equation 30 could be replaced by $(n-p)/2+p$.

replacing the empirical distribution $F_{n,y}$ with $F_y(y|\theta, H) = F^a(\sqrt[+]{y}\,|\theta, H)$ produces the LMS functional:

$$T_L(H) = \underset{\theta}{\operatorname{argmin}}\, F_y^{-1}(1/2|\theta, H) \tag{31}$$

Turning now to LTS, normalizing its objective function and writing it in terms of the empirical density of residuals yields

$$\frac{1}{n}\sum_{j=1}^{\lfloor (N+1)/2 \rfloor} (r_\theta^2)_{j:N} = \int_{-r_m}^{r_m} r^2 f_n^s(r|\theta, H_n)\, dr$$

where $r_m^2 = F_{n,y}^{-1}(1/2|\theta, H_n)$ is the empirical median square residual. The functional form of LTS then is easily written as

$$T_T(H) = \underset{\theta}{\operatorname{argmin}} \int_{-F_y^{-1}(1/2|\theta,H)}^{F_y^{-1}(1/2|\theta,H)} r^2 f^s(r|\theta, H)\, dr \tag{32}$$

4.6 MINPRAN

MINPRAN's functional is derived by first re-writing MINPRAN's objective function, replacing the binomial distribution with the incomplete beta function [19, page 229]:

$$\min_r \mathcal{F}(r, k_{r,\theta}, n) = \min_r I(k_{r,\theta},\ n - k_{r,\theta} + 1,\ r/Z_0)$$

where

$$I(v, w, p) = \frac{\Gamma(v+w)}{\Gamma(v)\Gamma(w)} \int_0^p t^{v-1}(1-t)^{w-1}\, dt.$$

and $\Gamma(\cdot)$ is the gamma function. This is done because $I(v, w, p)$ only requires $v, w \in \Re^+$, whereas the binomial distribution requires integer values for $k_{r,\theta}$ and n. Now, since $F_n^a(r|\theta, H_n)$ is the empirical distribution of the absolute residuals (see equation 26), $k_{r,\theta} = n \cdot F_n^a(r|\theta, H_n)$ for all $r > 0$. Thus, MINPRAN's objective function can be re-written equivalently as

$$\min_r\ I(n \cdot F_n^a(r|\theta, H_n),\ n(1 - F_n^a(r|\theta, H_n)) + 1,\ r/Z_0),$$

Replacing F_n^a by F^a and substituting equation 13 gives the functional

$$T_M(H) = \underset{\theta}{\operatorname{argmin}} \left\{ \min_r I(n \cdot F^a(r|\theta, H),\ n(1 - F^a(r|\theta, H)) + 1,\ r/Z_0) \right\}. \tag{33}$$

Observe that n, the number of points, is still required here, but $T_M(H)$ is considered a functional [10, page 40].

5 Pseudo Outlier Bias

Now that the functional forms of the robust estimators have been derived, the pseudo outlier bias metric can be defined. Given a particular mixture distribution $H(x,z)$, target distribution $H_1(x,z)$, and a functional T, let

$$\hat{\theta} = T(H) \quad \text{and} \quad \hat{\theta}_1 = T(H_1).$$

These fits are assumed to minimize the estimator's objective functional globally. Then, pseudo outlier bias is defined as the normalized L^2 distance between the fits:

$$\|\hat{\theta} - \hat{\theta}_1\|_2 = \frac{1}{(x_1 - x_0)} \left\{ \int_{x_0}^{x_1} [(\hat{\theta}(x) - \hat{\theta}_1(x))/\sigma]^2 \, dx \right\}^{1/2}. \tag{34}$$

As is easily shown, this metric is invariant to translation and independent scaling of both x and z. (For fixed-scale M-estimators, $\hat{\sigma}$, which is provided *a priori*, must be scaled as well. For MINPRAN, the outlier distribution must be scaled appropriately.)

When the set of the possible curves $\theta(x)$ includes $\beta_1(x)$, it can be shown that for each of the functionals derived in Section 4, $T(H_1) = \hat{\theta}_1 = \beta_1$. In other words, the estimator's objective function is minimized by β_1.[5] When $T(H_1) = \beta_1$, the pseudo outlier bias metric becomes

$$\|\hat{\theta} - \beta_1\|_2 = \frac{1}{(x_1 - x_0)} \left\{ \int_{x_0}^{x_1} [(\hat{\theta}(x) - \beta_1(x))/\sigma]^2 \, dx \right\}^{1/2}. \tag{35}$$

Intuitively, pseudo outlier bias measures the L^2 norm distance between the two estimates, $T(H)$ and $T(H_1)$, normalized by the length of the x interval over which $H(x,z)$ is non-zero and by the standard deviation of the noise in the z values. Since $T(H_1) = \beta_1$ for the cases studied here, a metric value of 0 implies that T is not at all corrupted by the presence of either gross or pseudo outliers, and a metric value of 1 implies that on average over the x domain $T(H)$ is one standard deviation away from β_1.

[5] In the analysis results given in Section 6, the set of curves will be linear functions of the form $\theta(x) = mx + b$. $\beta_1(x)$ and $\beta_2(x)$ will also be linear. These curves are continuous and have infinite extent in x, unlike the densities modeling data drawn from them.

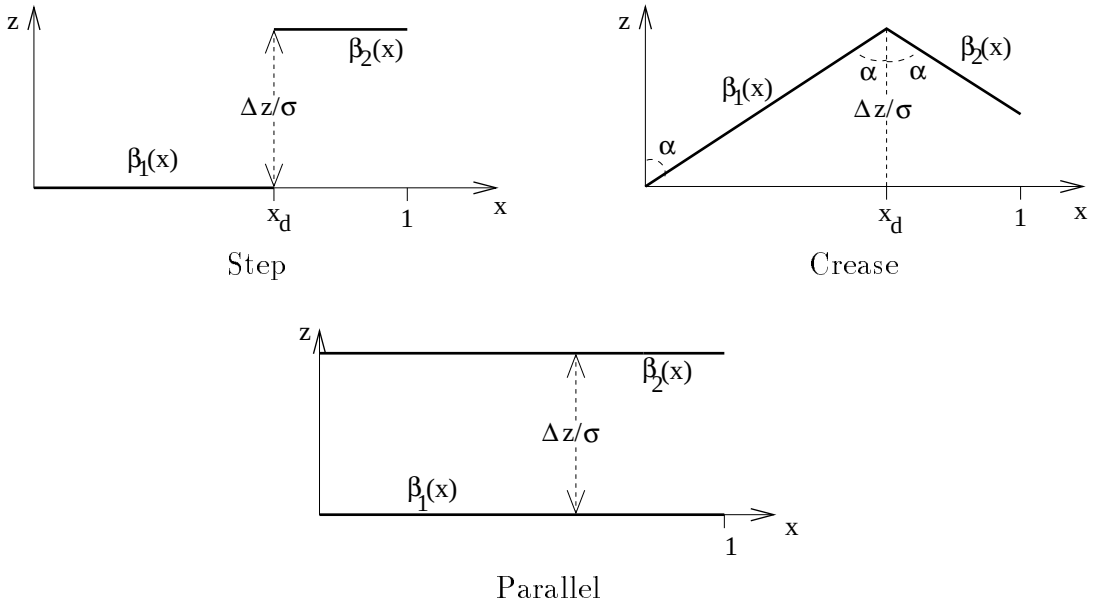


Figure 5: Parameters controlling the curve models for step edges, crease edges, and parallel lines. In each case, $\beta_1(x)$ is the desired correct fit and points from $\beta_2(x)$ are pseudo outliers. $\Delta z/\sigma$ is the scaled discontinuity magnitude, and ε_s controls the percentage of points from $\beta_1(x)$.

6 Bias Caused by Surface Discontinuities

Pseudo outlier bias (or "bias" for short) can now be used to analyze robust estimators' accuracy in fitting surfaces to data from three different types of discontinuities: step edges, crease edges, and parallel lines with overlapping x domains. To do this, Section 6.1 parameterizes the mixture density, outlines the technique to find $T(H)$, and discusses the relationship between results presented here and results for higher dimensions. Then, analysis results for specific estimators are presented: fixed-scale M-estimators and fixed-band techniques (Section 6.2) which require a prior estimate of $\hat{\sigma}$, standard M-estimators (Section 6.3) which estimate $\hat{\sigma}$, and LMS, LTS and MINPRAN (Section 6.4) which are independent of $\hat{\sigma}$. In each case, the bias is examined as both the discontinuity magnitude and mixture of inliers, pseudo outliers and gross outliers vary.

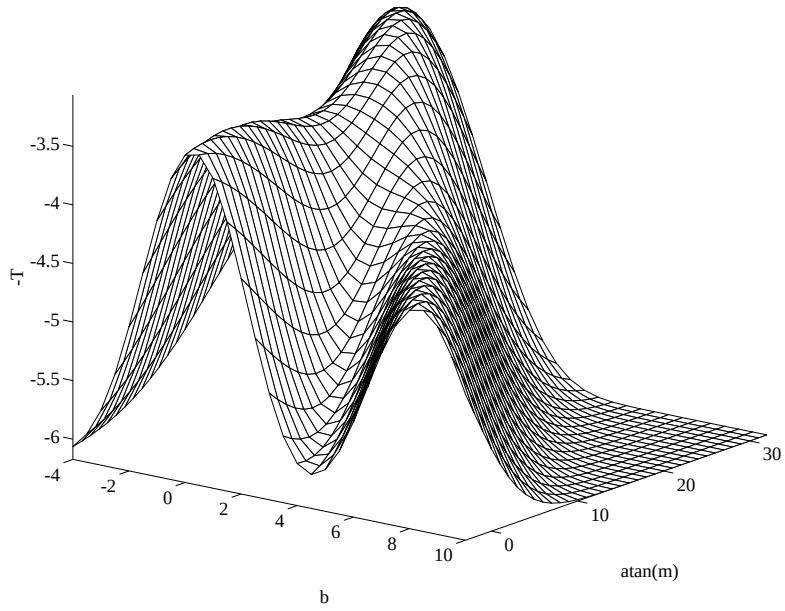


Figure 6: Surface plot of the objective functional of $T_{\rho_h}(H)$, i.e. $\int \rho(r/\hat{\sigma})\ f^s(r|\theta, H)\, dr$, for the hard redescending, fixed-scale M-estimator on a step edge with $\varepsilon_s = 0.6$ and $\Delta z/\sigma = 7.5$ when fits have the form $\theta(x) = mx + b$. (The plot shows the negation of the objective functional, so local minima in the functional appear as local maxima in the plot.) There are three local optimal: one at $\theta(x) = \beta_1(x)$, the second at $\theta(x) = \beta_2(x)$, and the third at a heavily biased fit, $\theta(x) = \tan(27.92°) - 1.91$. The biased fit is the global optimum.

6.1 Discontinuity Models and Search

Figure 5 shows the models of step edges, crease edges, and parallel lines. The translation and scale invariance of both the estimators and pseudo outlier bias, along with several realistic assumptions, allow these discontinuities to be described with just a few parameters. (Refer back to Section 3 for the exact parameter definitions.) For all models, $\sigma = 1.0$ and the x interval is $[0, 1]$. For step edges, $\beta_1(x) = 0$ and $\beta_2(x) = \Delta z/\sigma$ — retaining the σ parameter to make clear the scale invariance — and $x_{1,0} = 0$, $x_{2,1} = 1$, and $x_{1,1} = x_{2,0} = x_d$. With these values, $\varepsilon_s = x_d$. To move from step to crease edges, only the curves $\beta_1(x)$ and $\beta_2(x)$ must be changed. Referring to Figure 5b, these functions are $\beta_1(x) = (\Delta z/\sigma)(x/x_d)$ and $\beta_2(x) = (\Delta z/\sigma)(-x/x_d + 2)$ — α plays no explicit role because it is not scale invariant. For parallel lines (Figure 5c), $\beta_1(x)$ and $\beta_2(x)$ are the same as for step edges, $x_{1,0} = x_{2,0} = 0$ and $x_{1,1} = x_{2,1} = 1$, and the parameter x_d plays no role. Finally, the outlier distribution $g_o(z)$ is uniform for z within $\pm z_0/2$ of $\Delta z/2$ and 0 otherwise.

The foregoing shows that the parameters ε_s, ε_o, $\Delta z/\sigma$, and z_0 completely specify a two surface discontinuity model, the resulting mixture density, $h(x,z)$, and therefore, the distribution, $H(x,z)$. Hence, after specifying the class of functions (linear, here) for hypothesized fits, a given robust estimator's pseudo outlier bias can be calculated as a function of these parameters. This calculation requires an iterative, numerical search to minimize $T(H)$, and may require several starting points to avoid local minima. (See Figure 6 for an example plot of T_{ρ_h}'s objective functional.) Thus, for a particular type of discontinuity and for a particular robust estimator, the parameters may be varied to study their effect on the estimator's pseudo outlier bias, thereby characterizing how accurately the estimator can fit surfaces near discontinuities.

As a final observation, although the results are presented for one-dimensional image domains, they have immediate extension to two dimensions. For example, a two-dimensional analog of the step edge presented here is $\beta_1(x,y) = 0$ for $x \in [0, x_d]$ and $y \in [0,1]$ and $\beta_2(x,y) = \Delta z/\sigma$ for $x \in [x_d, 1]$ and $y \in [0,1]$. It is straightforward to show that this model results in exactly the same pseudo outlier bias as a one-dimensional step model having the same mixture parameters and gross outlier distribution. Similar results are obtained for natural extensions of the crease edge and parallel lines models. Thus, one-dimensional discontinuities are sufficient to establish limitations in the effectiveness of robust estimators.

6.2 Fixed-Scale M-Estimators and Fixed-Band Techniques

The first analysis results are for fixed-band techniques and fixed-scale M-estimators. These techniques represent an ideal case where the noise parameter $\hat{\sigma} = \sigma$ is known and fixed in advance. Figure 7 shows the bias of fixed-band techniques (T_F) and three fixed-scale M-estimators (T_{ρ_m}, T_{ρ_s}, and T_{ρ_h}) as a function $\Delta z/\sigma$ when $\varepsilon_s = 0.6$ and when $\varepsilon_s = 0.8$. The bias of the least-squares estimator, calculated by substituting $\rho(u) = u^2$ into equation 27, is included for comparison. The ρ function tuning parameters values are directly from the literature ($c = 1.345$ for ρ_m [11], $a = 1.31$, $b = 2.04$, $c = 4.00$ for ρ_h [10, page 167], and $f = 1.5$ for ρ_s [18]), and $r_b = 2.5\hat{\sigma}$ for T_b. Interestingly, the proportion of gross outliers, ε_o, has no effect on the results. This is because the fraction of the outlier distribution within r of a fit is the same for all fits θ and for all r except when $\theta(x) \pm r$ is extreme enough to

cross outside the bounds of the gross outlier distribution.

The sharp drops in bias shown in Figure 7 (a) and (b) for fixed-band techniques and the hard redescending M-estimator (and to some extent for the soft redescending M-estimator in (b)) correspond to $\hat{\theta}(x) = T_\rho(H)$ shifting from the local minimum associated with a heavily biased fit to the local minimum near $\beta_1(x)$, the optimum fit to the target distribution. Plotting the step height at which this drop occurs as a function of ε_s gives a good summary of these estimators' bias on step edges. Figure 8 does this, referring to this height as the "small bias height" and quantifying it as the step height at which the bias drops below 1.0.

The plots in Figures 7 and 8(a) show that fixed-band techniques and fixed-scale M-estimators are biased nearly as much as least-squares for significant step edge and crease edge discontinuity magnitudes. The estimators fare much better on parallel lines (Figure 7(e) and (f)); apparently, asymmetric positioning of pseudo outliers causes the most bias. To give an intuitive feel for the significance of the bias, Figure 9 shows step edge data generated using $\varepsilon_s = 0.6$ and $\Delta z/\sigma = 7.5$, model parameters for which the robust estimators are strongly biased.

Overall, the hard redescending, fixed-scale M-estimator is the least biased of the techniques studied thus far. Compared to other fixed-scale M-estimators, its finite rejection point — the point at which outliers no longer influence the fit — makes it less biased by pseudo outliers than monotone and soft redescending fixed-scale M-estimators. On the other hand, it is less biased than fixed-band techniques because it retains the statistical efficiency of least-squares for small residuals.

The hard redescending, fixed-scale M-estimator can be made less biased by reducing the values of its tuning parameters, as shown in Figure 8(b), effectively narrowing ρ_h and reducing its finite rejection point. (The parameter set $a = 1.0$, $b = 1.0$, $c = 2.0$ comes from [2]; the set $a = 1.0$, $b = 2.0$, $c = 3.0$ was chosen as an intermediate set of values.) Using small parameter values has two disadvantages, however: the optimum statistical efficiency of the standard parameters is lost, giving less accurate fits to the target distribution, and some good data may be rejected as outliers. Despite these disadvantages, lower tuning parameters should be used since avoiding heavily biased fits is the most important objective.

Finally, in practice, the non-convex objective functions of hard and soft redescending

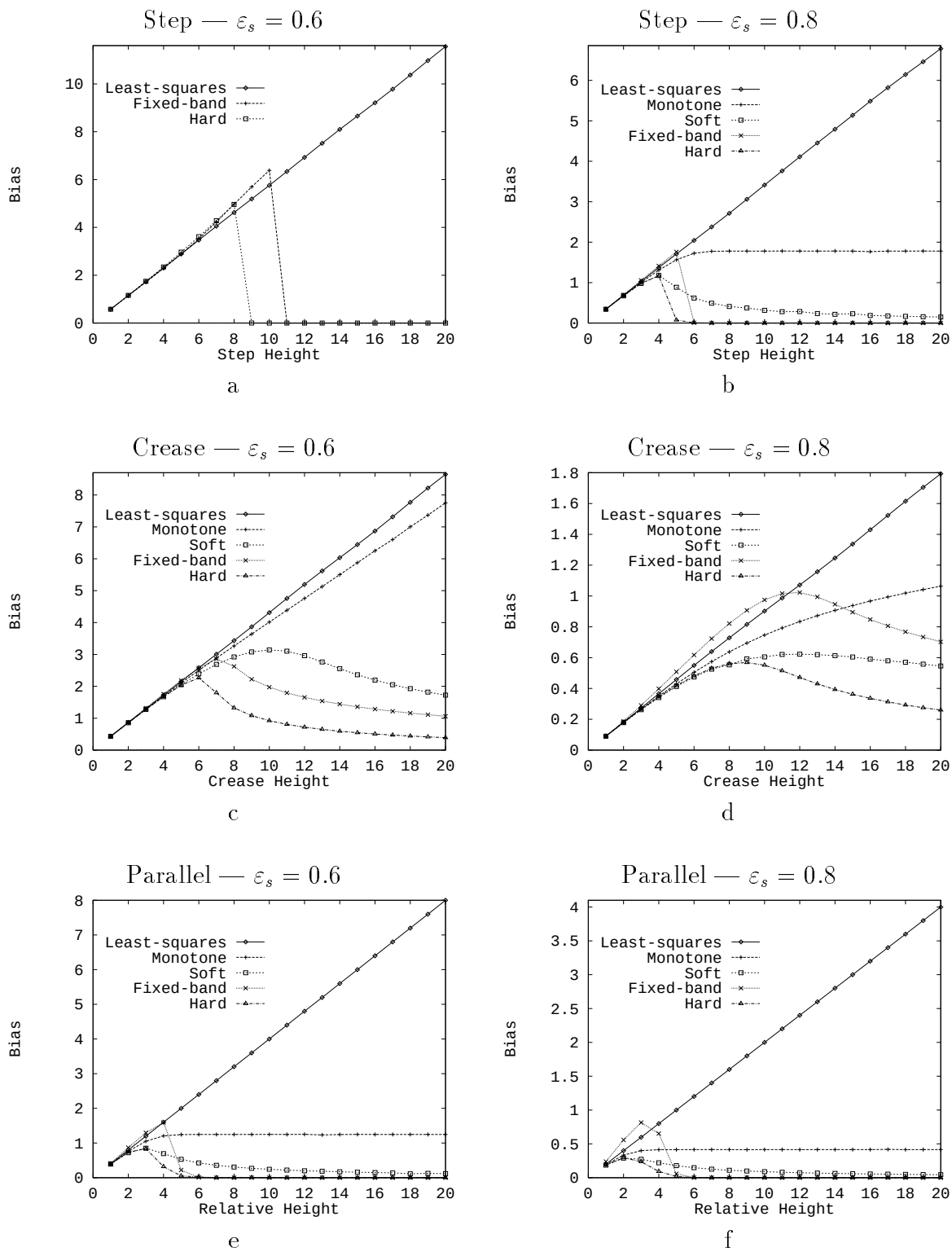


Figure 7: Bias of fixed-band techniques, fixed-scale M-estimators and least-squares on step edges, (a) and (b), crease edges, (c) and (d), and parallel lines, (e) and (f), as a function of height when $\varepsilon_s = 0.6$ and $\varepsilon_s = 0.8$. The horizontal axis is the relative discontinuity magnitude (height), $\Delta z/\sigma$, and the vertical axis is the bias (see equation 35). Plots not shown in (a) are essentially equivalent to the least-squares plots.

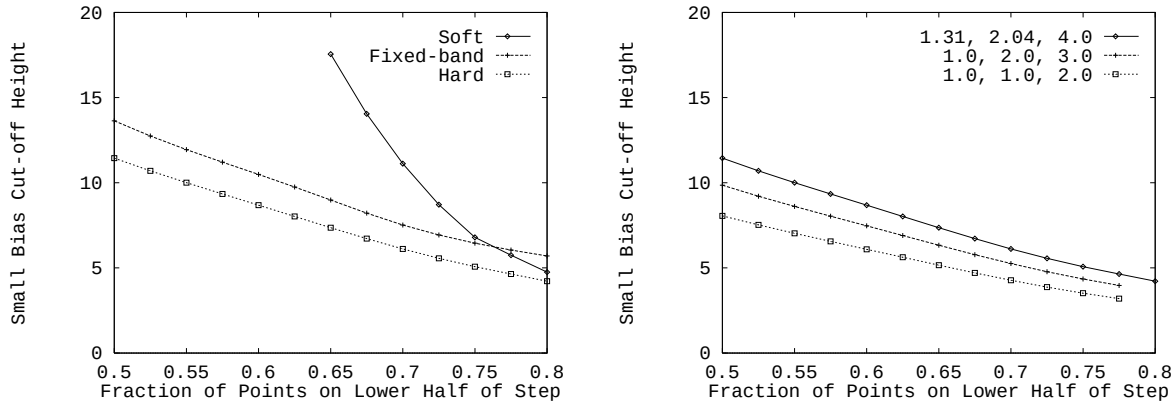


Figure 8: Small bias cut-off heights as a function of ε_s, the relative fraction of points on the lower half of the step. Plots in (a) show the heights for fixed-band techniques and two fixed-scale M-estimators. Plots in (b) show the heights for different tuning parameters of the hard redescending fixed-scale M-estimator. Heights not plotted for small ε_s are above $\Delta z/\sigma = 20$. When height is not plotted for large ε_s, bias is never greater than 1.0.

fixed-scale M-estimators can lead to more biased results than indicated here. Iterative search techniques, especially when started from a non-robust fit, may stop at a local minimum corresponding to a biased fit when the fit to the target distribution is the global minimum of the objective function. Therefore, to avoid local minima, fixed-scale M-estimators should use either a random sampling search technique or a Hough transform.

6.3 M-Estimators

Next, consider standard M-estimators, which estimate $\hat{\sigma}$ from the data. To calculate $T(H)$ for the monotone and soft redescending M-estimators, simply calculate $\hat{\theta} = T_{\rho,s}(H)$ for any mixture distribution using equation 7 or 8 as the objective functional. For the hard redescending M-estimator, which estimates $\hat{\sigma}$ from an initial fit, the optimum fit to the mixture distribution is found in three stages: first find the optimum LMS fit, then calculate the median absolute deviation (MAD) [10, page 107] to this fit, scaling it to estimate $\hat{\sigma}$, and finally calculate $\hat{\theta} = T_{\rho_h}(H)$ with $\hat{\sigma}$ fixed. Two different scale factors for estimating $\hat{\sigma}$ are considered: the first, 1.4826, ensures consistency at the normal distribution; the second, 1.14601, ensures consistency at Student's t-distribution (with $f = 1.5$). Using the latter allows accurate comparison between the hard and soft redescending M-estimators since the

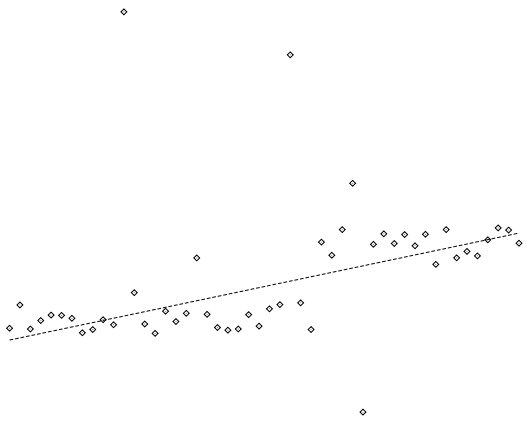


Figure 9: Example step edge data generated when $\varepsilon_s = 0.6$ and $\Delta z/\sigma = 7.5$, a discontinuity where each the objective function of each robust estimator (except LTS) is minimized by a biased fit. The example fit shown is $\hat{\theta}(x)$ for the hard redescending, fixed-scale M-estimator.

latter is the maximum likelihood estimate for Student's t distribution [5].

Figure 10 shows bias plots for the soft redescending M-estimator and for the hard redescending M-estimator using the two different scale factors (plot "Hard-N" for the normal distribution and plot "Hard-t" for the t-distribution). Results for the monotone M-estimator are not shown since its bias matches that of least-squares almost exactly. Overall, the results are substantially worse than for fixed-scale M-estimators, especially for $\varepsilon_s = 0.6$. This is a direct result of $\hat{\sigma}$ being a substantial over-estimate of σ: for example, when $\varepsilon_s = 0.6$ and $\Delta z/\sigma = 10$, $\hat{\sigma}/\sigma \geq 2.4$ for all estimates. (See [22] for analysis of bias in estimating $\hat{\sigma}$.) These over-estimates allow a large portion of the residual distribution to fall in the region where ρ is quadratic, causing the estimator to act more like least-squares. Because of this, M-estimators are heavily biased by discontinuities when they must estimate $\hat{\sigma}$ from the data.

6.4 LMS, LTS and MINPRAN

The last estimators examined are LMS, LTS, and MINPRAN, methods which neither require $\hat{\sigma}$ *a priori* nor need to estimate it while finding $\hat{\theta}(x)$. Figure 11 shows bias plots for these estimators on step edges, crease edges and parallel lines, using $\varepsilon_0 = 0.1$ and $z_0 = 100$. Figure 12 shows small bias cut-off heights on step edges for LMS, LTS and MINPRAN, and it demonstrates the effects of changes in the mixture proportions on LMS and LTS.

LMS and LTS work as well as any technique studied as long as the actual of fraction

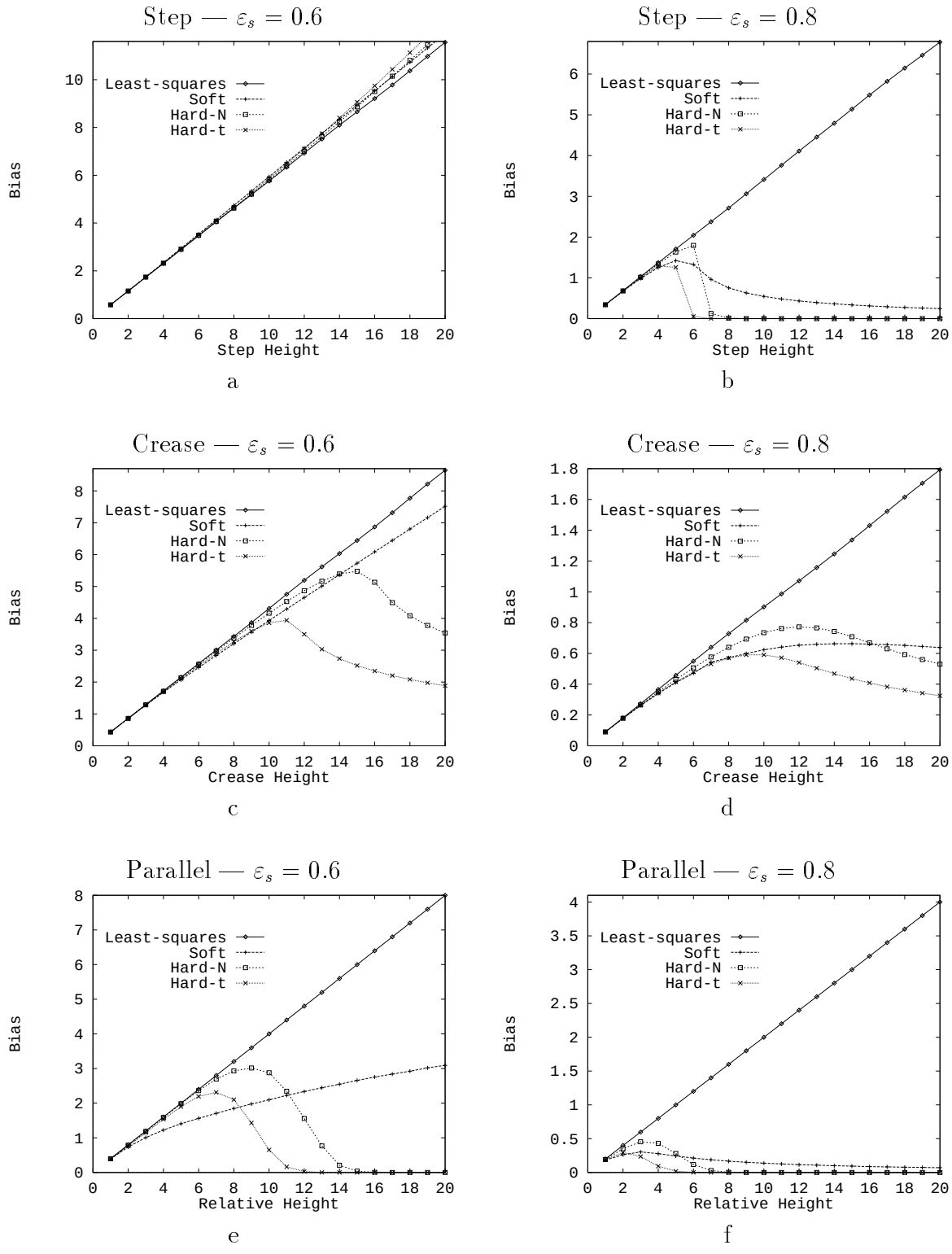


Figure 10: Bias of M-estimators and least-squares on step edges, (a) and (b), crease edges, (c) and (d), and parallel lines, (e) and (f) when $\varepsilon_s = 0.6$ and $\varepsilon_s = 0.8$. The horizontal axis is the relative discontinuity magnitude (height), $\Delta z/\sigma$, and the vertical axis is the bias (see equation 35).

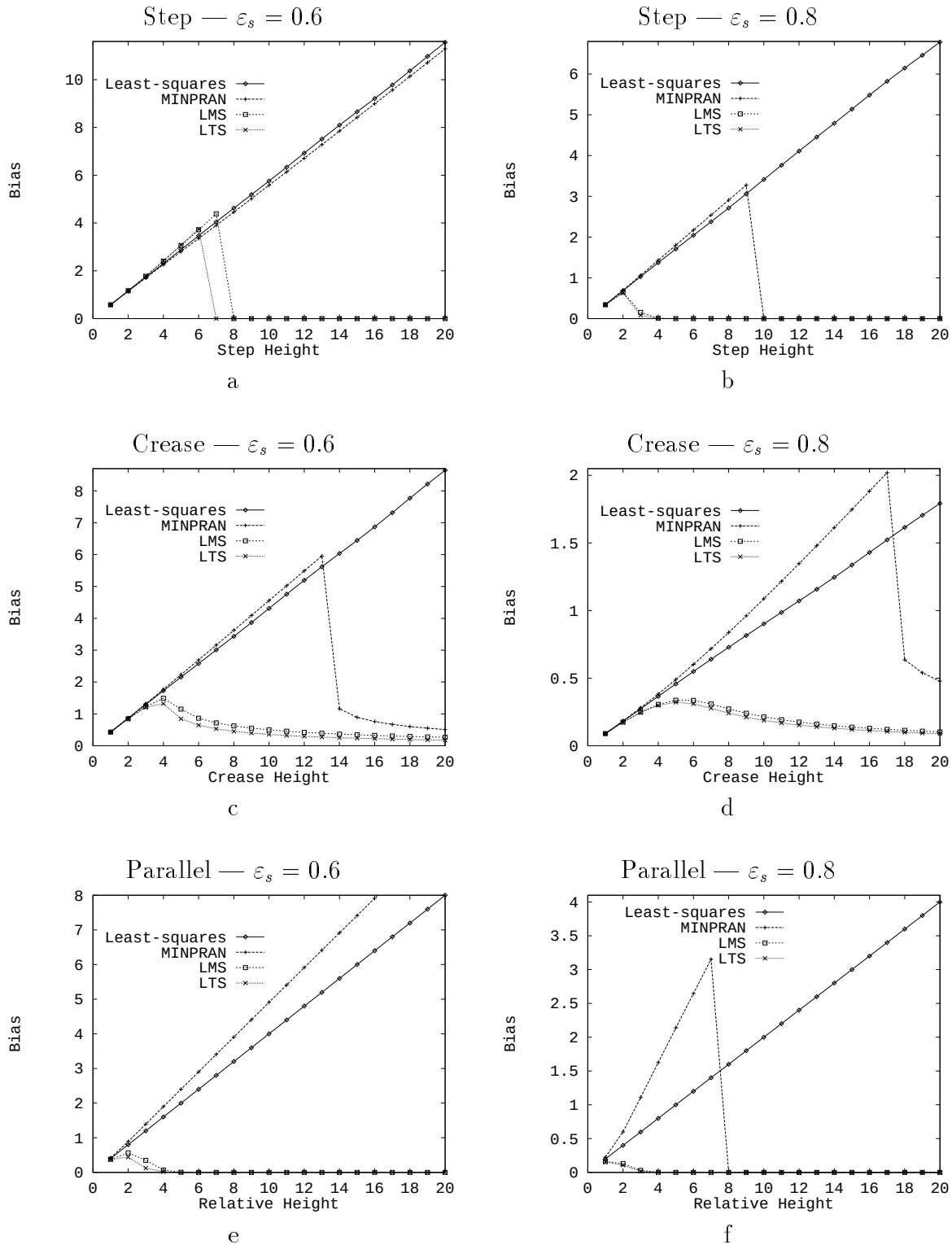


Figure 11: Bias of MINPRAN, LMS, LTS and least-squares on step edges, (a) and (b), crease edges, (c) and (d), and parallel lines, (e) and (f) when $\varepsilon_s = 0.6$ and $\varepsilon_s = 0.8$. The horizontal axis is the relative discontinuity magnitude (height), $\Delta z/\sigma$, and the vertical axis is the bias (see equation 35).

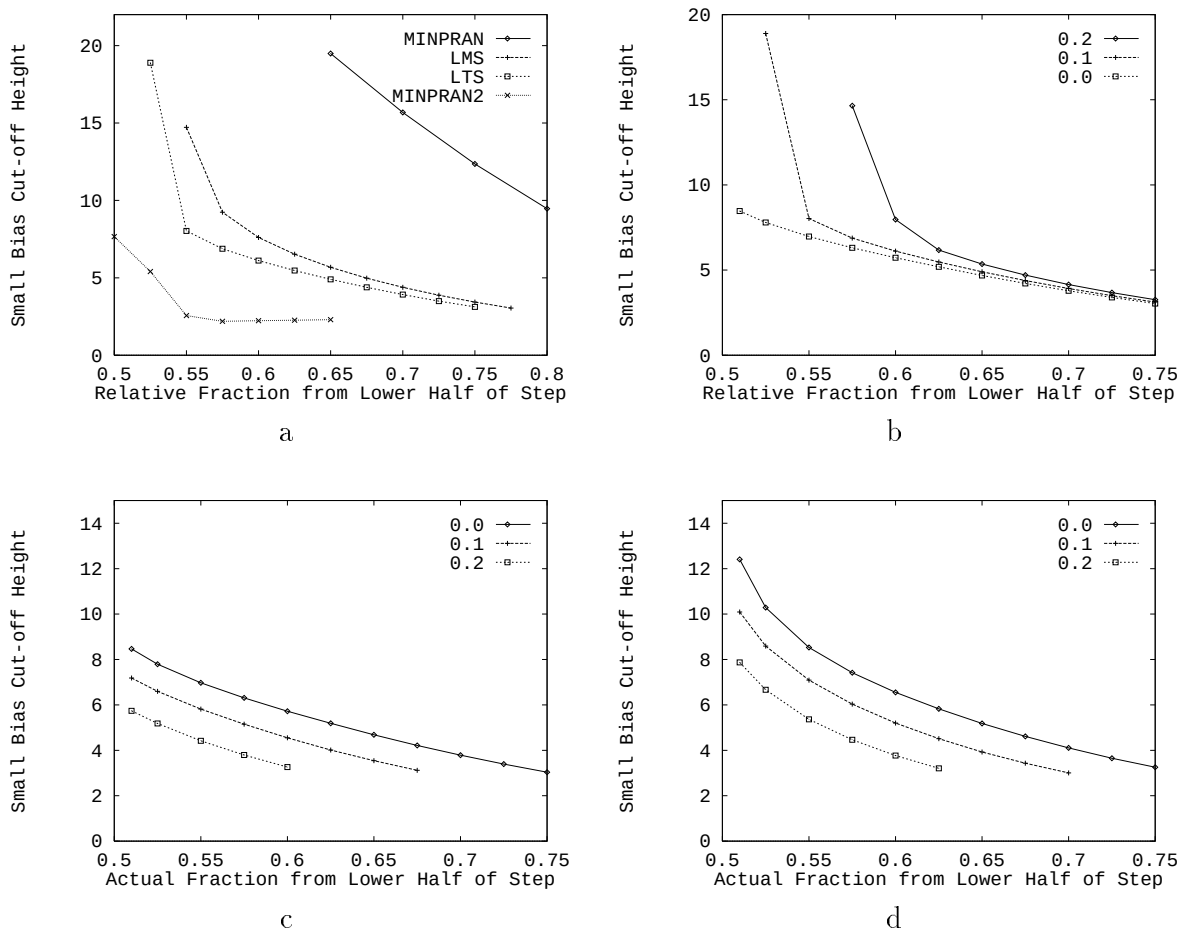


Figure 12: Small bias cut-off heights. Plot (a) shows these for LMS, LTS, MINPRAN, and the modified MINPRAN optimization criteria (MINPRAN2) as a function of ε_s, the relative fraction of inliers. Plot (b) shows these for LTS as a function of ε_s for different gross outlier percentages ε_o. Plots (c) and (d) show these for LTS and LMS respectively as a function of $(1-\varepsilon_o)\varepsilon_s$, the actual fraction of inliers. Heights not plotted for small ε_s or $(1-\varepsilon_o)\varepsilon_s$ are above $H/\sigma = 20$. When height is not plotted for large ε_s or $(1-\varepsilon_o)\varepsilon_s$, bias is never greater than σ.

inliers — data from $\beta_1(x)$ — is above 0.5. Since this fraction is $(1-\varepsilon_o)\varepsilon_s$, the bias of LMS and LTS, unlike that of M-estimators, depends heavily on both ε_o and ε_s. (For random sampling implementations of LMS and LTS, where p points instantiate a hypothesized fit and the objective function is evaluated on the remaining $n-p$ points, the bias curves in Figure 11 and the steep drop in cut-off heights in Figure 12 will shift to the right, but only marginally since usually $n \gg p$.) Figures 12b and c demonstrate this dependence in two ways for LTS. Figure 12b shows small bias cutoffs as a function of ε_s, the *relative* fraction of inliers — points on the lower half of the step. The bias cutoffs are lower for lower ε_o simply because fewer gross outliers imply more actual inliers when ε_s remains fixed. Figure 12c shows small bias cutoffs as a function of the *actual* fraction of inliers. In this context, varying ε_o while $(1-\varepsilon_o)\varepsilon_s$ is fixed changes the fraction of gross outliers versus pseudo outliers. As the plot shows, the coherent structure of the pseudo outliers causes more bias than the random structure of gross outliers. This same effect is shown for LMS in Figure 12d. Finally, the magnitude of z_0, which controls the gross outlier distribution, has little effect on the bias results, except in the unrealistic case where it approaches the discontinuity magnitude.

LTS is less biased than LMS, especially when the actual fraction of inliers is only slightly above 0.5. This can be seen most easily by comparing the low bias cutoff plots in Figure 12c and d. Like the advantage of hard redescending M-estimators over fixed-band techniques (Section 6.2), this occurs because LTS is more statistically efficient than LMS [21] — its objective function depends on the smallest 50% of the residuals rather than just on the median residual. It is important to note that although LMS's efficiency can be improved by application of a one-step M-estimator starting from the LMS estimate, this will not improve substantially a heavily biased fit, since a local minimum of the M-estimator objective function will be near this fit.

With a minor modification to its optimization criteria, MINPRAN can be made much less sensitive to pseudo outliers, improving dramatically on the poor performance shown in Figures 11 and 12. The idea is to find two disjoint fits (no shared inliers), $\hat{\theta}_a$ and $\hat{\theta}_b$, with inlier bounds $\hat{r}_a$ and $\hat{r}_b$ and inlier counts $k_{\hat{\theta}_a,\hat{r}_a}$ and $k_{\hat{\theta}_b,\hat{r}_b}$, minimizing $\mathcal{F}(\hat{r}_a+\hat{r}_b, k_{\hat{\theta}_1,\hat{r}_a}+k_{\hat{\theta}_b,\hat{r}_b},\ n)$ [23, 26]. If $\hat{\theta}$ is the single fit minimizing the criterion function, with inlier bound $\hat{r}$ and inlier

count $k_{\hat{\theta},\hat{r}}$, then the two fits $\hat{\theta}_a$ and $\hat{\theta}_b$ are chosen instead of the single fit $\hat{\theta}$ if

$$\mathcal{F}(\hat{r}_a + \hat{r}_b, k_{\hat{\theta}_a,\hat{r}_a} + k_{\hat{\theta}_b,\hat{r}_b}, n) < \mathcal{F}(\hat{r}, k_{\hat{\theta}}, n). \tag{36}$$

Thus, the modified optimization criteria tests whether one or two inlier distributions are more likely in the data [27]. Figure 12 shows the step edge small bias cut-off heights for this new objective function, denoted by MINPRAN2. These are substantially lower than those of the other techniques, including LTS. Further, these results, unlike those of MINPRAN, are only marginally affected by the parameters ε_o and z_0. Unfortunately, the search for $\hat{\theta}_a$ and $\hat{\theta}_b$ is computationally expensive, and so the present implementation of MINPRAN2 uses a simple search heuristic that yields [23, 26] more biased results than the optimum shown here. It is, however, as effective as the fixed-scale, hard redescending M-estimator and, unlike LMS and LTS, it does not fail dramatically when there are too few inliers.

6.5 Discussion and Recommendations

Overall, the results show that all the robust estimators studied estimate biased fits at small but substantial discontinuity magnitudes. This bias, which relative to the bias of least-squares is greater for crease and step edges and less for parallel lines, occurs even if $\hat{\sigma}$ or the distribution of gross outliers or both are known *a priori*. Further, it must be emphasized that *this bias is not an artifact of the search process*: the functional form of each estimator returns the fit corresponding to the global minimum of the estimator's objective function.

The reason for the bias can be seen by examining the cumulative distribution functions (cdfs) of absolute residuals. Figure 13 plots this cdf, $F^a(r|\theta, H)$, when θ is the target fit ($\theta = \beta_1$) and when θ is the least-squares fit to H, for H modeling crease and step discontinuities. For $\Delta z/\sigma = 6.0$, the cdf of the biased fit is almost always greater that of the target fit, meaning that in a discrete set of samples, the biased fit, which crosses through both point sets, will on average yield smaller magnitude residuals than the target fit, which is close to only the target point set. (The situation is somewhat better when $\Delta z/\sigma = 9.0$.) Therefore, robust estimators, such as the ones studied, whose objective functions are based solely on residuals, are unlikely to estimate unbiased fits at small magnitude discontinuities.

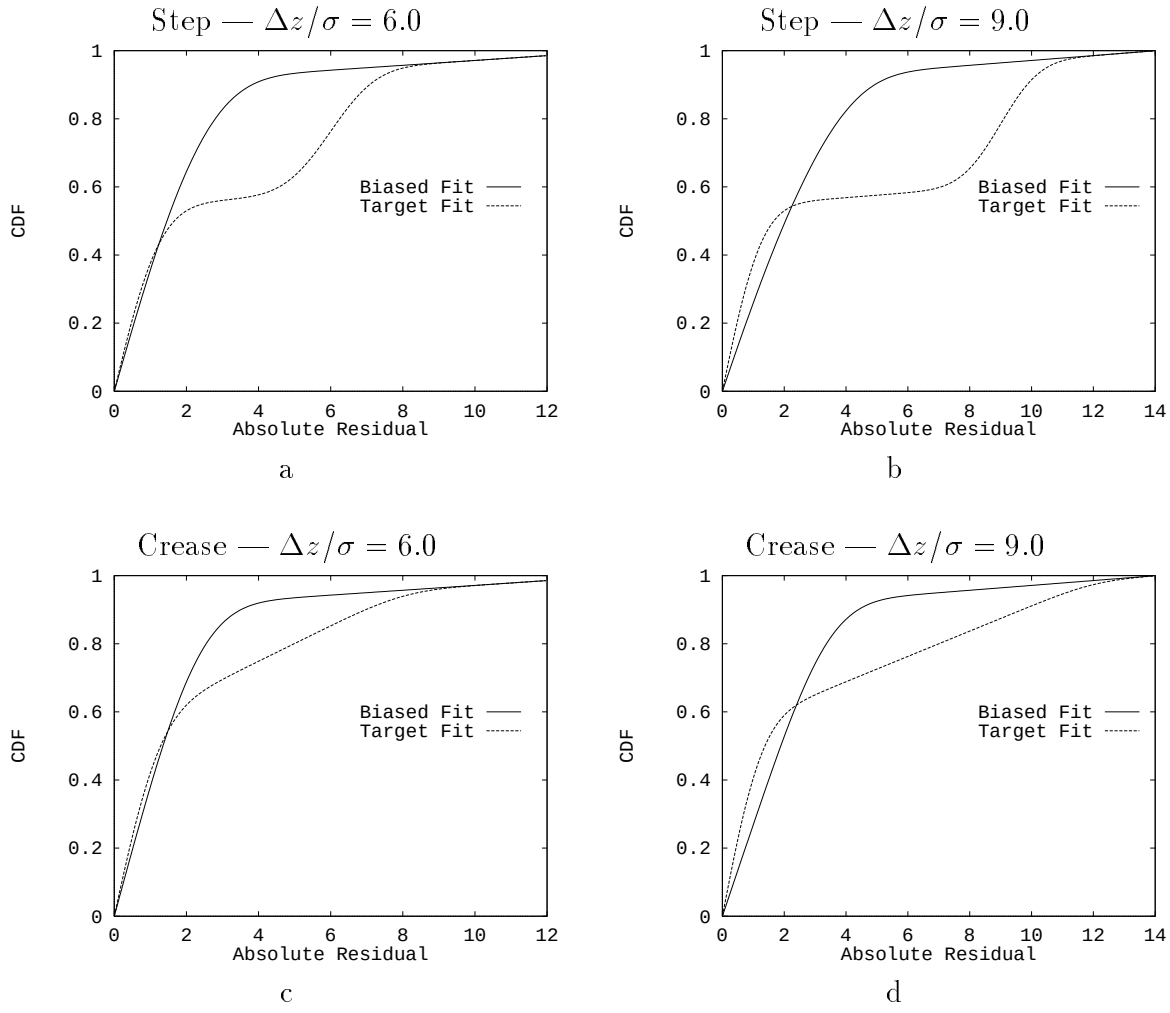


Figure 13: Each figure plots the cumulative distribution functions (cdf) of absolute residuals for the target fit and for a biased (least-squares) fit: (a) and (b) are relative to a step discontinuity, and (c) and (d) are relative to a crease discontinuity. For all plots, the mixture fractions are fixed at $\varepsilon_o = 0.1$ and $\varepsilon_s = 0.6$. All robust estimators are substantially biased at $\Delta z/\sigma = 6.0$ for both step and crease discontinuities.

While none of the estimators works as well as desired, the following recommendations for choosing among them are based on the results presented above:

- When $\hat{\sigma}$ is known *a priori*, one should use a hard redescending M-estimator objective function such as Hampel's with reduced tuning parameter values and either a random-sampling search technique or a weighted Hough transform. To ensure all inliers are found and to obtain greater statistical efficiency, an one-step M-estimator with larger tuning parameters should be run from the initial optimum fit. This technique is preferable to LTS and LMS because it is less sensitive to the number of gross outliers.

- When $\hat{\sigma}$ is not known *a priori*, but the distribution of gross outliers is known, one should use the modified MINPRAN algorithm, MINPRAN2 [23, 26].

- When neither $\hat{\sigma}$ nor the distribution of gross outliers is known, LTS should be used, although its performance degrades quickly when there are too few inliers. LTS is preferable to LMS because of its statistical efficiency.

7 Summary and Conclusions

This paper has developed the pseudo outlier bias metric using techniques from mathematical statistics to study the fitting accuracy of robust estimators on data taken from multiple structures — surface discontinuities, in particular. Pseudo outlier bias measures the distance between a robust estimator's optimum fit to a target distribution and its optimum fit to an outlier corrupted mixture distribution. Here, the target distribution models the points from a single surface and the mixture distribution models points from multiple surfaces plus gross outliers. Each estimator's optimum fit is found by applying its functional form to one of these model distributions. Thus, like other analysis tools from the robust statistics literature, pseudo outlier bias depends on point distributions rather than on particular point sets drawn from these distributions. While this has some limitations — the actual fitting error for particular points sets may be more or less than the pseudo outlier bias and it ignores problems that may arise from multiple local minima in an objective function — it represents a simple, efficient, and elegant method of analyzing robust estimators.

Pseudo outlier bias was used to analyze the performance of M-estimators, fixed-band techniques (Hough transforms and RANSAC), least median of squares (LMS), least trimmed squares (LTS) and MINPRAN in fitting surfaces to three different discontinuity models: step edges, crease edges and parallel lines. For each of these discontinuities, two surfaces generate data, with the larger set of surface data forming the inliers and the smaller set forming the pseudo outliers. By characterizing these discontinuity models using a small number of parameters, formulating the models as mixture distributions, and studying the bias of the robust estimators as the parameters varied, it was shown that each robust estimator is biased for substantial discontinuity magnitudes. This effect, which relative to that of least-squares is strongest for step edges and crease edges, persists even when the noise in the data or the gross outlier distribution or both are known in advance. It is disappointing because in vision data — not just in range data — multiple structures (pseudo outliers) are more prevalent than gross outliers. In spite of the disappointment, however, specific recommendations, which depend on what is known about the data, were made for choosing between current techniques.[6]

These negative results indicate that care should be used when robustly estimating surface parameters in range data, either to obtain local low-order surface approximations or to initialize fits for surface growing algorithms [3, 5, 6, 15]. (Similar problems may occur for the "layers" techniques that have been applied to motion analysis [1, 6, 28].) Robust estimates will be accurate for large scale depth discontinuties and sharp corners, but will be skewed at small magnitude discontinuites, such as near the boundary of a slightly raised or depressed area of a surface. Obtaining accurate estimates near these discontinuities will require new and perhaps more sophisticated robust estimators.

Acknowledgements

The author would like to acknowledge the financial support of the National Science Foundation under grants IRI-9217195 and IRI-9408700, the assistance of James Miller in various aspects of this work, and the insight offered by the anonymous reviewers which led to sub-

[6] See [14, 17] for new, related techniques.

stantial improvements in the presentation.

Appendix A: Evaluating $F_s(r|\theta, H)$

This appendix shows how to evaluate the conditional cumulative distribution and conditional density of signed residuals, $F^s(r|\theta, H)$ (equation 21) and $f^s(r|\theta, H)$ (equation 22). The distribution and density of the absolute residuals are obtained easily from these.

Expanding the expression in equation 21 for $F^s(r|\theta, H)$, using equation 18 for h, gives

$$F^s(r|\theta, H) = \varepsilon_o F_o^s(r|\theta, H) + (1-\varepsilon_o)[\varepsilon_s F_1^s(r|\theta, H) + (1-\varepsilon_s)F_2^s(r|\theta, H)]. \tag{37}$$

Here,

$$F_o^s(r|\theta, H) = \int_{x_0}^{x_1} \int_{-\infty}^{\theta(x)+r} \frac{g_o(z)}{x_1 - x_0}\, dz\, dx = G_o(\theta(x) + r) \tag{38}$$

where $G_o(\cdot)$ is the cumulative distribution of the gross outliers, and for $i = 1, 2$

$$F_i^s(r|\theta, H) = \int_{x_{i,0}}^{x_{i,1}} \int_{-\infty}^{\theta(x)+r} \frac{g(z-\beta_i(x);\sigma^2)}{x_{i,1} - x_{i,0}} dz dx. \tag{39}$$

To simplify evaluating $F_i^s(r|\theta, H)$, change variables and then change the order of integration. Starting with the change of variables, make the substitutions $v = z - \beta_i(x)$ and $dv = dz$ (intuitively, v is the fit residual at x), define $\phi(x) = \theta(x) - \beta_i(x)$, and let $\lambda_i = 1/(x_{i,1} - x_{i,0})$. Then, the integral becomes

$$F_i^s(r|\theta, H) = \int_{x_{i,0}}^{x_{i,1}} \int_{-\infty}^{\phi(x)+r} \lambda_i g(v;\sigma^2)\, dv\, dx.$$

Since the integrand is now independent of x, rewriting the integral to integrate over strips parallel to the x axis will produce a single integral. Consider a strip bounded by v and $v+\Delta v$ (Figure 14). The integral over this strip is approximately $\lambda_i g(v) w(v) \Delta v$, where $w(v)$ is the width of the integration region at v. In the limit as $\Delta v \to 0$, this becomes exact and the integral over the entire region becomes

$$F_i^s(r|\theta, H) = \int_{-\infty}^{v_1} \lambda_i g(v;\sigma^2) w(v) dv, \tag{40}$$

where v_1 is the maximum of $\phi(x) + r$ over $[x_{i,0}, x_{i,1}]$.

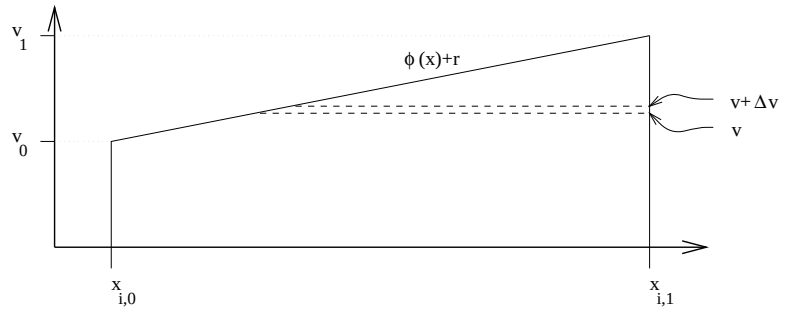


Figure 14: Calculating $F_i^s(r|\theta, H)$ for $i = 1, 2$ requires integrating the point density for curve i over strips of width Δv parallel to the x axis. The density $g(v; \sigma^2)$ is constant over these strips.

Evaluating $w(v)$ depends on $\phi(x)$. This paper studies linear fits and linear curve models, so $\phi(x)$ is linear. In this case, let $\phi(x) = mx + b$, assume $m > 0$, and let $v_0 = mx_{i,0} + b + r$ and $v_1 = mx_{i,1} + b + r$ (see Figure 14). Then, $w(v) = x_{i,1} - x_{i,0}$ for $v < v_0$ and $w(v) = x_{i,1} - (v - b - r)/m$ for $v_0 \leq v \leq v_1$. Thus, using G to denote the cdf of the gaussian,

$$\begin{aligned} F_i^s(r|\theta, H) &= \lambda \int_{-\infty}^{v_0} (x_{i,1} - x_{i,0}) g(v; \sigma^2)\, dv \;+\; \lambda \int_{v_0}^{v_1} \left(x_{i,1} - \frac{v - b - r}{m} \right) g(v; \sigma^2)\, dv \\ &= G(v_0; \sigma^2) + \lambda \frac{mx_{i,1} + b + r}{m} [G(v_1; \sigma^2) - G(v_0; \sigma^2)] + \lambda \frac{\sigma^2}{m} [g(v_1; \sigma^2) - g(v_0; \sigma^2)] \end{aligned} \tag{41}$$

A similar result is obtained when $m < 0$, and when $m = 0$, $v_0 = v_1$ and so $F_i^s(r|\theta, H) = G(v_0; \sigma^2)$.

To compute the density $f_s(r|\theta, H)$, start from the mixture density in equation 22 and integrate each component density separately. This is straightforward when the density g_o is uniform and, as above, $\theta(x)$ and $\beta_i(x)$ are linear.

References

[1] S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum likelihood estimation of mixture models and MDL encoding. In *Proceedings IEEE International Conference on Computer Vision*, pages 777–784, 1995.

[2] P. J. Besl, J. B. Birch, and L. T. Watson. Robust window operators. In *Proceedings IEEE International Conference on Computer Vision*, pages 591–600, 1988.

[3] P. J. Besl and R. C. Jain. Segmentation through variable-order surface fitting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:167–192, 1988.

[4] R. C. Bolles and M. A. Fischler. A Ransac-based approach to model fitting and its application to finding cylinders in range data. In *Proceedings Seventh International Joint Conference on Artificial Intelligence*, pages 637–643, 1981.

[5] K. L. Boyer, M. J. Mirza, and G. Ganguly. The Robust Sequential Estimator: A general approach and its application to surface organization in range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:987–1001, 1994.

[6] T. Darrell and A. Pentland. Cooperative robust estimation using layers of support. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:474–487, 1995.

[7] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24:381–395, 1981.

[8] J. D. Gaskill. *Linear Systems, Fourier Transforms, and Optics*. John Wiley and Sons, 1978.

[9] F. R. Hampel, P. J. Rousseeuw, and E. Ronchetti. The change-of-variance curve and optimal redescending M-estimators. *Journal of the American Statistical Association*, 76:643–648, 1981.

[10] F. R. Hampel, P. J. Rousseeuw, E. Ronchetti, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 1986.

[11] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Commun. Statist.-Theor. Meth.*, A6:813–827, 1977.

[12] P. J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[13] J. Illingworth and J. Kittler. A survey of the Hough transform. *CVGIP*, 44:87–116, 1988.

[14] K.-M. Lee, P. Meer, and R.-H. Park. Robust adaptive segmentation of range images. *(submitted to) IEEE Transactions on Pattern Analysis and Matchine Intelligence*, 1996.

[15] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation of range images as the search for geometric parametric models. *International Journal of Computer Vision*, 14:253–277, 1995.

[16] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim. Robust regression methods for computer vision: A review. *International Journal of Computer Vision*, 6:59–70, 1991.

[17] J. V. Miller and C. V. Stewart. MUSE: Robust surface fitting using unbiased scale estimates. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 300–306, 1996.

[18] M. J. Mirza and K. L. Boyer. Performance evaluation of a class of M-estimators for surface parameter estimation in noisy range data. *IEEE Transactions on Robotics and Automation*, 9:75–85, 1993.

[19] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992.

[20] G. Roth and M. D. Levine. Extracting geometric primitives. *CVGIP: Image Understanding*, 58:1–22, 1993.

[21] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.

[22] P. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283, 1993.

[23] C. V. Stewart. A new robust operator for computer vision: Application to range images. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 167–173, 1994.

[24] C. V. Stewart. A new robust operator for computer vision: Theoretical analysis. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 1994.

[25] C. V. Stewart. Expected performance of robust estimators near discontinuities. In *Proceedings IEEE International Conference on Computer Vision*, pages 969–974, 1995.

[26] C. V. Stewart. MINPRAN: A new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:925–938, 1995.

[27] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, New York, 1985.

[28] J. Y. A. Wang and E. H. Adelson. Layered representation for motion analysis. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 361–366, 1993.