A Brief Overview of Robust Statistics

Olfa Nasraoui
Department of Computer Engineering & Computer Science
University of Louisville,
olfa.nasraoui_AT_louisville.edu

Robust Statistical Estimators

Robust statistics have recently emerged as a family of theories and techniques for estimating the parameters of a parametric model while dealing with deviations from idealized assumptions [Goo83,Hub81,HRRS86,RL87]. Examples of deviations include the contamination of data by gross errors, rounding and grouping errors, and departure from an assumed sample distribution. Gross errors or outliers are data severely deviating from the pattern set by the majority of the data. This type of error usually occurs due to mistakes in copying or computation. They can also be due to part of the data not fitting the same model, as in the case of data with multiple clusters. Gross errors are often the most dangerous type of errors. In fact, a single outlier can completely spoil the Least Squares estimate, causing it to break down. Rounding and grouping errors result from the inherent inaccuracy in the collection and recording of data which is usually rounded, grouped, or even coarsely classified. The departure from an assumed model means that real data can deviate from the often assumed normal distribution. The departure from the normal distribution can manifest itself in many ways such as in the form of skewed (asymmetric) or longer tailed distributions.

M-estimators

The ordinary Least Squares (LS) method to estimate parameters is not robust because its objective function, $\sum_{j=1}^{N} d_j^2$, increases indefinitely with the residuals d_j between the j^{th} data point and the estimated fit, with N being the total number of data points in a data set. Hence, extreme outliers with arbitrarily large residuals can have an infinitely large influence on the resulting estimate. M-estimators [Hub81] attempt to limit the influence of outliers by replacing the square of the residuals with a less rapidly increasing loss function of the data value, x, and parameter estimate, t, $\rho(x;t)$. The M-estimate, $T(x_1, \cdots, x_N)$ for the function ρ and the sample $x_1, \cdots, x_N$, is the value that minimizes the following objective

$$\min_t \quad \left\{ J = \sum_{j=1}^{N} \rho(x_j; t) \right\}. \tag{1}$$

The optimal parameter, T, is determined by solving

$$\frac{\partial J}{\partial t} = \sum_{j=1}^{N} \psi(x_j; t) = \mathbf{0}. \tag{2}$$

where, except for a multiplicative constant,

$$\psi(x_j; t) = \frac{\partial \rho(x_j; t)}{\partial t}. \tag{3}$$

When the M-estimator is equivariant, i. e., $T(x_1 + a, \cdots, x_N + a) = T(x_1, \cdots, x_N) + a$ for any real constant a, we can write ψ and ρ in terms of the residuals $x - t$. Also, in general, an auxiliary scale estimate, S is used to obtain the scaled residuals $r = \frac{x-t}{S}$. Hence, we can write

$$\psi(r) = \psi\left(\frac{x-t}{S}\right) = \psi(x; t),$$

and

$$\rho(r) = \rho\left(\frac{x-t}{S}\right) = \rho(x; t).$$

The ρ— functions for some familiar M-estimators are listed in Table 1. Note that LS can be considered an M-estimator, even though it is not a *robust* M-estimator. As seen in this table, M-estimators rely on both an accurate estimate of scale and a fixed tuning constant, c. Most M-estimators use a multiple of the Median of Absolute Deviations (MAD) as a scale estimate which implicitely assumes that the noise contamination rate is 50%. MAD is defined as follows

$$MAD(x_i) = med_i\left\{|x_i - med_j(x_j)|\right\}$$

The most common scale estimate used is $1.483 \times MAD$ where the 1.483 factor adjusts the scale for maximum efficiency when the data samples come from a Gaussian distribution.

Table 1: A few common M- and W-estimators

Type	$\rho(r)$	pt $\psi(r)$ pt	pt w(r) pt	Range	used
				of r	scale
L_2 (mean)	$\frac{1}{2}r^2$	r	1	$\mathbb{R}$	none
L_1 (median)	$\|r\|$	sgn(r)	$\frac{sgn(r)}{r}$	$\mathbb{R}$	none
Huber	$\frac{1}{2}r^2$	r	1	$\|r\| \leq k$	MAD
	$k\|r\| - \frac{1}{2}k^2$	k sgn(r)	$\frac{ksgn(r)}{r}$	$\|r\| > k$	
Cauchy	$\frac{c^2}{2}log\left[1 + (\frac{r}{c})^2\right]$	$\frac{r}{1+(\frac{r}{c})^2}$	$\frac{1}{1+(\frac{r}{c})^2}$	$\mathbb{R}$	MAD
Tukey's	$\frac{1}{6}\left[1 - (1 - r^2)^3\right]$	$r(1 - r^2)^2$	$(1 - r^2)^2$	$\|r\| \leq 1$	$c \times$ MAD
biweight	$\frac{1}{6}$	0	0	$\|r\| > 1$	
Andrews	$\frac{1}{\pi^2}(1 - cos\pi r)$	$\frac{1}{\pi}sin\pi r$	$\frac{1}{\pi r}sin\pi r$	$\|r\| \leq 1$	$c \times$ MAD
	$\frac{2}{\pi^2}$	0	0	$\|r\| > 1$	
Welsch	$\frac{c^2}{2}\left[1 - exp(-(\frac{r}{c})^2)\right]$	$r\, exp(-(\frac{r}{c})^2)$	$exp(-(\frac{r}{c})^2)$	$\mathbb{R}$	MAD

W-estimators

W-estimators [Goo83] represent an alternative form of M-estimators. Each W-estimator has a characteristic weight function, $w(.)$ that represents the importance of each sample in its contribution to the estimation of T, which is related to the corresponding M-estimator as follows

$$\psi(r) = w(r)\, r. \tag{4}$$

The optimal parameter is determined by solving

$$\sum_{j=1}^{N} w\left(r_j\right) r_j = 0, \tag{5}$$

which is similar to the equations for a ``weighted LS'' regression problem. W-estimators offer a convenient and simple iterative computational procedure for M-estimators, where the W-estimator equations in the current iteration are solved by fixing the weight values, $w\left(r_j\right)$, to those of the previous iteration. The resulting procedure is referred to as the Iterative Reweighted Least Squares (IRLS or RLS). As in the case of M- and W-estimators, the IRLS relies on an accurate and prefixed scale estimate for the definition of its weights. The most common scale estimate used is $1.483 \times MAD$.

The ρ, ψ, and w functions for some familiar M- and W-estimators are listed in Table 1.

L-estimators

Also known as trimmed means for the case of location estimation (t), L-estimators [KJ78] are based on a definition of quantiles as follows.

$$\psi_p\left(r_j\right) = \begin{cases} p-1 & \text{if } r_j < 0, \\ p & \text{otherwise.} \end{cases} \tag{6}$$

In (6), $r_j = x_j - t$ is the signed residual from the j^{th} data sample, x_j, to the location estimate, t. The loss function is defined as

$$\rho_p\left(r_j\right) = r_j \psi_p\left(r_j\right) = \begin{cases} r_j\left(p-1\right) & \text{if } r_j < 0, \\ r_j p & \text{otherwise.} \end{cases} \tag{7}$$

The p^{th} quantile of the sample $x_1, \cdots, x_N$, is the value of t that solves

$$\min_t \quad J = \sum_{j=1}^{N} \rho_p\left(r_j\right). \tag{8}$$

It is easy to check that for $p = \frac{1}{2}$, the half-quantile, $\psi_p(r_j) = \frac{1}{2} sgn(r_j)$ corresponds to the sample median. The major inconvenience in L-estimators is that they are not easy to optimize, and that they rely on a known value for the noise contamination rate, $1 - p$. They are also among the least efficient (accurate) estimators because they completely ignore part of the data.

R-estimators

In this approach [Jae72] , each residual is weighted by a score based on its rank as in the following objective.

$$\min_{\theta} \quad \left\{ J = \sum_{j=1}^{N} a_N(R_j) r_j \right\}. \tag{9}$$

where $R_j = R(d_j)$ is the rank of the j^{th} residual in $\{r_1, \cdots, r_N\}$ and $a_N(.)$ is a nondecreasing score function satisfying $\sum_k a_N(k) = 0$. For example, the Wilcoxon scores are given by $a_N(k) = \frac{k}{(N+1)} - \frac{1}{2}, k = 1, \cdots, N$. Like L-estimators, the major inconvenience in R-estimators is that they are not easy to optimize, and that the definition of the score function implicitly necessitates prior information about the noise contamination rate.

The Least Median of Squares Estimator (LMedS)

Instead of minimizing the sum of squared residuals, r_j, as in LS to estimate the parameter vector Θ, Rousseuw [RL87] proposed minimizing their median as follows

$$\min_{\Theta} \quad \text{med} \quad << 390 >> \quad r_j^2. \tag{10}$$

This estimator effectively trims the $\lfloor \frac{n}{2} \rfloor$ observations having the largest residuals, and uses the maximal residual value in the remaining set as the criterion to be minimized. Hence it is equivalent to *assuming* that the noise proportion is 50%, and its breakdown point asymptotically approaches 50% for 2-dimensional data sets, and $\left(\frac{\lfloor n/2 \rfloor - p + 2}{n} \right)$ for p-dimensional data sets. It can also be seen that (10) is unwieldy from an optimization point

of view, because of its non-differentiable form. This means that a quasi-exhaustive search on all possible parameter values needs to be done to find the global minimum. As a variation, random sampling/searching of some kind has been suggested to find the best fit [RL87], leading to a reduced complexity of $\mathcal{O}(N \log N)$ instead of the high $\mathcal{O}(N^2)$ complexity of the exhaustive option. A major drawback of LMedS is its low efficiency, since it only uses the middle residual value and hence assumes that the data set contains a 50% fraction of noise. When the data set contains less than 50% noise, the LMedS estimates suffer in terms of accuracy, since not all the good points are used in the estimation; and when the data set contains more than 50% noise, the LMedS estimates can suffer considerably in terms of robustness, as will be illustrated in Chapter 3.

The Least Trimmed of Squares Estimator (LTS)

LTS [RL87] offers a more efficient way to find robust estimates by minimizing the objective function given by

$$\min_{\theta} \quad \sum_{j=1}^{h} \left(r^2\right)_{j:N}, \qquad (2)$$

where $\left(r^2\right)_{j:n}$ is the j^{th} smallest residual or distance when the residuals are ordered in ascending order, i.e.,

$$\left(r^2\right)_{1:N} \leq \left(r^2\right)_{2:N} \leq \cdots \leq \left(r^2\right)_{N:N}.$$

Since h is the number of data points whose residuals are included in the sum, this estimator basically finds a robust estimate by identifying the $(n-h)$ points having the largest residuals as outliers, and discarding (trimming) them from the data set. The resulting estimates are essentially LS estimates of the trimmed data set. It can be seen that h should be as close as possible to the number of good points in the data set, because the higher the number of good points used in the estimates, the more accurate the estimates are. In this case, LTS will yield the best possible estimate. One problem with LTS is that its objective function does not lend itself to mathematical optimization. Besides, the estimation of h itself is difficult in practice. As will be illustrated in Chapter 3, when faced with more noise than assumed, LTS will lack robustness. And when the amount of noise is less than the assumed level, it will lack efficiency, i.e., the parameter estimates suffer in terms of accuracy, since not all the good data points are taken into account. This reliance on a known or assumed amount of noise present in the data set (contamination rate) means that an exhaustive search over all possible contamination rates needs to be done; and that the optimal estimates have to be chosen based on some kind of validity test because

the LTS criterion is monotonically nondecreasing with h when h is less than the actual number of noise points. In addition, the LTS objective function is based on hard rejection. That is, a given data point is either totally included in the estimation process or totally excluded from it. This is not a good strategy if there are points in the region of doubt. As in the case of L-estimators, LTS suffers from a low efficiency, because it completely ignores part of the data.

The Reweighted Least Squares Estimator (RLS)

Instead of the noise proportion, some algorithms explicitly cast their objective functions in terms of a set of weights that distinguish between inliers and outliers. However, these weights usually depend on a scale measure which is also difficult to estimate. For example, the RLS estimator [HW77] tries to minimize

$$\min_{\theta} \quad \sum_{j=1}^{N} w_j r_j^2. \tag{11}$$

where r_j^2 are robust residuals resulting from an approximate LMedS or LTS procedure. Here the weights w_j essentially trim outliers from the data used in LS minimization, and can be computed after a preliminary approximate phase of LMedS or LTS. The function w_j is usually continuous and has a maximum at 0 and is monotonicaly non-increasing with r_j^2. In addition, w_j depends on an error scale σ which is usually heuristically estimated from the results of LMedS or LTS. RLS can be considered to be equivalent to W-estimators if there exists a function $\psi()$ satisfying (4). A major advantage of RLS is its ease of computation using the IRLS procedure as for the case of W-estimators. However, RLS was intended to refine the estimates resulting from other robust but less efficient algorithms. Hence it is extremely dependent on a good initialization.

Resistance Properties of M-estimators

An estimator is resistant if a small number of gross errors or any number of small rounding and grouping errors have only a limited effect on the resulting estimate [Goo83]. As seen below, most of the resistance properties of an M-estimator can be inferred from the shape of its Influence Curve.

The Influence Curve

The Influence Curve (IC) tells us how an infinitesimal proportion of contamination affects the estimate in large samples. Formally [Goo83], the IC gives a quantitative expression of the change in the estimate that results from perturbing the samples underlying distribution, F, by a point mass at sample location x. For an estimator given by the functional $T(F)$ and defined by the ψ function $\psi(u)$, the IC at F_0 is

$$IC\left(x; F_0, T\right) = \lim_{\epsilon \to 0} \frac{T\left[(1-\epsilon)F_0 + \epsilon\delta_x\right] - T\left(F_0\right)}{\epsilon},$$

where δ_x is a point mass perturbation at x and F_0 is the underlying or empirical distribution. The IC can be shown to reduce to

$$IC\left(x; F_0, T\right) = \frac{S\left(F_0\right)\psi\left[\frac{x - T(F_0)}{S(F_0)}\right]}{\int \psi'\left[\frac{x - T(F_0)}{S(F_0)}\right] dF_0(x)}$$

where $S\left(F\right)$ is the auxiliary scale estimate. The IC can be further shown to be of the form

$$IC = (\text{const})\psi(u). \tag{12}$$

Hence, the shape of IC depends only on the shape of the ψ function, not the data distribution.

The Breakdown Bound

The Breakdown (BD) bound or point [Hub81] is the largest possible fraction of observations for which there is a bound on the change of the estimate when that fraction of the sample is altered without restrictions. M-estimators of location with an odd $\psi()$ function have a BD bound close to 50% provided that the auxiliary scale estimator has equal or better BD bound [RL87].

Rejection Point

The rejection point is defined as the point beyond which IC becomes zero [Goo83]. Except possibly through the auxiliary scale estimate, observations with residuals beyond the rejection point have zero influence. Hence they make no contribution to the final estimate. Estimators who have a finite rejection point are said to be redescending and are well protected against very large outliers. However, a finite rejection point usually results in the underestimation of scale. This is because when the samples near the tails of a distribution are ignored, too little of the samples may remain for the estimation process. This in turn adversely affects the efficiency of the estimator. An estimator is efficient if the variance of its estimate is as close as possible to the variance of the best estimator for a given distribution. For the Gaussian distribution the best estimator is the mean which also yields the minimum variance of the estimate. In general, it is best for a robust estimator to use as many of the good samples of a distribution as possible, in order to maintain a good efficiency. Another adverse effect of finite rejection is that if a large part of the sample is ignored, the objective function may have many local minima

[Goo83].

Gross Error Sensitivity

The Gross Error Sensitivity (g.e.s.) expresses asymptotically the maximum effect a contaminated observation can have on the estimator. It is the maximum absolute value of the IC. The asymptotic bias of an estimator, defined as the maximum effect of the contamination of a given distribution with a proportion ϵ from an outlying distribution, is given by $\epsilon.(\mathrm{g.e.s.})$. Unfortunately, it was reported in [Goo83] that in general, poor g.e.s. corresponds to higher Gaussian efficiency, and vice versa.

Local Shift Sensitivity

The local Shift Sensitivity (l.s.s.) measures the effect of the removal of a mass ϵ at y and its reintroduction at x. Therefore, it measures the effect of rounding and grouping errors on an estimator. For highest resistance, it is required that the l.s.s. be bounded. For a continuous and differentiable IC, l.s.s. is given by the maximum absolute value of the slope of IC at any point. In [Goo83], it was reported that in general, a lower (hence better) l.s.s. corresponds to higher Gaussian efficiency.

Winsor's Principle

Winsor's principle [Tuk60] states that all distibutions are normal in the middle. Hence, the $\psi-$ function of M-estimators should resemble the one that is optimal for Gaussian data in the middle. Since the Maximum Likelihood estimate for Gaussian data is the mean which has a linear $\psi-$ function, it is desired that $\psi(u) \approx ku$ for small $|u|$, where k is a nonzero constant. In general, a $\psi-$ function that is linear in the middle results in better efficiency at the Gaussian distribution.

Symmetry

As shown in [Goo83], it is necessary for the $\psi-$ function to be odd in order to have an unbiased estimator, when the samples' underlying distribution is symmetric. If the underlying distribution is symmetric with center T, then an estimator, $T_N(x_1, \cdots, x_N)$ is said to be unbiased if $E[T_N(x_1, \cdots, x_N)] = T$. This crucial property is satisfied by all known M-estimators.

Simultaneous M-estimators of Location and Scale

A simultaneous M-estimator of location and scale [Goo83] for the sample $x_1, \cdots, x_N$ is the combination of a location estimator T_N and a scale estimator w_N that satisfy the pair of equations

$$\sum_{j=1}^{N} \psi\left(\frac{x_j - T_N}{cw_N}\right) = 0 \tag{13}$$

and

$$\sum_{j=1}^{N} \chi\left(\frac{x_j - T_N}{cw_N}\right) = 0, \tag{14}$$

where c is a tuning constant, and for a symmetric underlying sample distribution, ψ is an odd function, and χ is an even function. The problem with this approach lies in how to choose an appropriate χ-function that will yield a scale estimate that is as accurate as possible, that is meaningful, and that is somehow related to the ψ-function so that the simultaneous optimization process, which usually alternates solving the two equations, makes global sense. For all these reasons, this approach has hardly been used in the past. As in the case of location M-estimators, χ has the same shape as the IC of the scale estimator.

M-estimators and W-estimators For Regression

The classical nonrobust multiple regression model relates a response vector $\mathbf{y} = (y_1, \cdots, y_N)^t$ to the explanatory variables in the matrix $\mathbf{X}$ in the form

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

where the rows of $\mathbf{X} = \begin{pmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_N} \end{pmatrix}$ represent the individual observation vectors $\mathbf{x_i} = (1, x_{i1}, \cdots, x_{ip})$,

which are augmented by 1 in the first dimension to allow for a constant term, β_0, in the linear regression model. The elements of $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^t$ are unknown parameters to be estimated, and ϵ is an error vector.

The Least Squares chooses the estimate $\hat{\beta}$ as the value of β that minimizes the sum of squared residuals,

$$\sum_{i=1}^{N} (y_i - \mathbf{x}_i\beta)^t (y_i - \mathbf{x}_i\beta) = (\mathbf{y} - \mathbf{X}\beta)^t (\mathbf{y} - \mathbf{X}\beta)$$

which results in the optimal closed form solution

$$\hat{\beta} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y},$$

and the fitted values $\hat{y}_i = \mathbf{x}_i\hat{\beta}$.

The M-estimator for β, based on the loss function $\rho(t)$ and the data $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$ is the value $\hat{\beta}$ which minimizes

$$\sum_{i=1}^{N} \rho(y_i - \mathbf{x}_i\beta).$$

$\hat{\beta}$ is determined by solving the set of p simultaneous equations

$$\sum_{i=1}^{N} \psi(y_i - \mathbf{x}_i\beta)\mathbf{x}_i^t = \mathbf{0}$$

where $\psi(t) = \frac{\partial\rho(t)}{\partial t}$.

W-estimators offer an alternative form of M-estimators by writing $\psi(t) = w(t)t$. Hence the p simultaneous equations become

$$\sum_{i=1}^{N} (y_i - \mathbf{x}_i\beta) w_i\mathbf{x}_i^t = \mathbf{0}$$

where $w_i = w\left(y_i - \mathbf{x}_i\beta\right)$. The above equations can be combined into the following single matrix equation

$$\mathbf{X}^t\mathbf{W}\mathbf{X}\beta = \mathbf{X}^t\mathbf{W}\mathbf{y},$$

where $\mathbf{W} = \mathbf{diag}\left(w_1, \cdots, w_N\right)^t$ is a diagonal matrix with $W_{ii} = w_i$. This results in the optimal closed form solution

$$\hat{\beta} = \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{y},$$

and the fitted values $\hat{y}_i = \mathbf{x_i}\hat{\beta}$.

In a similar fashion to the simple location M-estimator, the IC for a regression M-estimator can be shown to take the form [HRRS86]

$$IC\left((\mathbf{x}, y); F_0, \hat{\beta}\right) = (\mathbf{const})\ \psi\left[y - \mathbf{x}\hat{\beta}\left(F_0\right)\right]\mathbf{B}^{-1}\mathbf{x}^t, \tag{15}$$

where $\mathbf{B} = \int \psi'\left[y - \mathbf{x}\hat{\beta}\left(F_0\right)\right]\mathbf{x}^t\mathbf{x}dF_0(\mathbf{x}, y)$. Though $\psi\left[y - \mathbf{x}\hat{\beta}\left(F_0\right)\right]$ is bounded for robust M-estimators, the term $\mathbf{B}^{-1}\mathbf{x}^t$ can grow infinitely large depending on the position of $\mathbf{x}$. Hence M- and W-estimators of regression have an unbounded IC. This means that leverage points for which the independent variables, $\mathbf{x}$, are outlying compared to the remainder of the data can have an unbounded influence on the parameter estimates.

The Minimum Volume Ellipsoid Estimator (MVE)

MVE [RL87] tries to generalize LMedS to the case of multivariate location estimation. The approach is based on seeking the ellipsoid with smallest volume, including at least h points of the sample data set. A subsample, K, consisting of $p+1$ observations is first drawn from the data set, $\mathcal{X}$, with dimensionality p. Then, the subsample mean and covariance matrix are computed as per maximum likelihood,

$$\mathbf{m}_K = \frac{\sum_{\mathbf{x}_i \in K}\mathbf{x}_i}{p+1}$$

and

$$\mathbf{\Sigma}_K = \frac{\sum_{\mathbf{x}_i \in K} (\mathbf{x}_i - \mathbf{m}_K)(\mathbf{x}_i - \mathbf{m}_K)^t}{p}$$

It can easily be shown that the factor, λ_K needed to inflate the ellipsod covering the subsample K to the ellipsoid $\lambda_K \mathbf{\Sigma}_K$, which includes $h = 50\%$ of the points, is given by $\lambda_K = \text{med}<< 390 >> (\mathbf{x}_i - \mathbf{m}_K)^t \mathbf{\Sigma}_K^{-1} (\mathbf{x}_i - \mathbf{m}_K)$. The volume of the resulting ellipsoid is proportional to $|\mathbf{\Sigma}_K|^{\frac{1}{2}} (\lambda_K)^p$. After repeating this sampling process many times, the sample resulting in the minimum volume generates the optimal MVE parameters, $\mathbf{m}$ and $\mathbf{\Sigma}$ given by $\mathbf{m} = \mathbf{m}_K$ and $\mathbf{\Sigma} = \frac{\lambda_K^2 \mathbf{\Sigma}_K}{\chi^2_{n,0.5}}$, where the denominator adjusts the final covariance estimate to include all the good data points for the case of Gaussian data. Instead of the otherwise exhaustive sampling needed, Rousseeuw suggests selecting only m subsamples to guarantee a probability, q, of selecting at least one good subsample, where m is given by the relation $q = 1 - (1 - (1 - \epsilon)^p)^m$, and ϵ is the noise contamination rate. However, it is clear that an accurate lower bound on the number of subsamples cannot be computed if the noise contamination rate is not known in advance. MVE is also limited by its assumption that the noise contamination rate is 50%.

Random Sample Consensus (RANSAC)

This approach [FB81] relies on random sampling selection to search for the best fit. The model parameters are computed for each randomly selected subset of points. Then the points within some error tolerance are called the consensus set of the model, and if the cardinality of this set exceeds a prespecified threshold, the model is accepted and its parameters are recomputed based on the whole consensus set. Otherwise, the random sampling and validation is repeated as in the above. Hence, RANSAC can be considered to seek the best model that maximizes the number of inliers. The problem with this pioneering approach is that it requires the prior specification of a tolerance threshold limit which is actually related to the inlier bound.

Minimum Probability of Randomness (MINPRAN)

MINPRAN [Ste95] relies on the assumption that the noise comes from a well known distribution. As in RANSAC, this approach uses random sampling to search for the fit and the inliers to this fit that are least likely

to come from the known noise distribution. Even with random sampling, MINPRAN's computational complexity is $\mathcal{O}\left(N^2 + SN\log N\right)$, which is higher than that of LMedS, where N is the size of the data set and S is the number of samples.

All the above estimators are either obliged to perform an exhaustive search or assume a known value for the amount of noise present in the data set (contamination rate), or equivalently an estimated scale value or inlier bound. When faced with more noise than assumed, all these estimators will lack robustness. And when the amount of noise is less than the assumed level, they will lack efficiency, i.e., the parameter estimates suffer in terms of accuracy, since not all the good data points are taken into account. They are also limited to estimating a single component in a data set.

Acknowledgment

This work is supported by the National Science Foundation (CAREER Award IIS-0133948 to O. Nasraoui). Partial support of earlier stages of this work by the Office of Naval Research grant (N000014-96-1-0439) and by the National Science Foundation Grant IIS 9800899 is also gratefully acknowledged.

Bibliography

FB81

M. A. Fischler and R. C. Bolles.
Random sample consensus for model fitting with applications to image analysis and automated cartography.
Comm. of the ACM., 24:381-395, 1981.

Goo83

C. Goodall.
M-estimators of location: An outline of the theory.
In D. Hoaglin, F. Mosteller, and J. W. Tukey, editors, *Understanding Robust and Exploratory Data Analysis*, pages 339-403. New York, 1983.

HRRS86

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel.
Robust Statistics the Approach Based on Influence Functions.
John Wiley & Sons, New York, 1986.

Hub81

P. J. Huber.
Robust Statistics.
John Wiley & Sons, New York, 1981.

HW77

P. W. Holland and R. E. Welsch.
Robust regression using iteratively reweighted least-squares.
Commun. Statist. Theory Meth, A6(9):813-827, 1977.

Jae72
L. A. Jaeckel.
Estimating regression coefficients by minimizing the dispersion of the residuals.
Annals of Mathematical Statistics, 43:1449-1458, 1972.

KJ78
R. Koenker and G. Basset Jr.
Regression quantiles.
Econometrica, 36:33-50, 1978.

RL87
P. J. Rousseeuw and A. M. Leroy.
Robust Regression and Outlier Detection.
John Wiley & Sons, New York, 1987.

Ste95
C. V. Stewart.
Minpran: A new robust estimator for computer vision.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(10):925-938, Oct. 1995.

Tuk60
J. W. Tukey.
A survey of sampling from contaminated distributions.
In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, , and H. B. Mann, editors, *Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling*, pages 448-485. Stanford, CA: Stanford University Press, 1960.

About this document ...

A Brief Overview of Robust Statistics

This document was generated using the **LaTeX2HTML** translator Version 2002 (1.62)

The command line arguments were:
latex2html `-split 0 -image_type gif RobustStatistics.tex`