A Highly Robust Estimator for Computer Vision

Xinhua Zhuang

Electrical and Computer Engineering
University of Missouri-Columbia
Columbia, Missouri 65211
U.S.A.

Robert M. Haralick

Department of Electrical Engineering, FT-10
University of Washington
Seattle, Washington 98195
U.S.A.

Abstract

The paper presents a highly robust estimator called an MF-estimator for general regression.

In the paper, we argued that the kind of estimators needed by computer vision must be highly robust, and that the classical robust estimators do not render a high robustness. We also explain that the high robustness becomes possible only through partially but completely modeling the unknown log likelihood function.

Partial modeling explores a number of important heuristics implicit in the regression problem and takes place by taking them into consideration with the Bayes statistical decision rule, while maximizing the log likelihood function.

Experiments with the simplest location estimation show the superior performance of the MF-estimator over the classical M-estimator, as was expected.

The authors believe that the proposed MF-estimator will pave a solid road towards solving a lot of robust estimation problems which have arised in low level computer vision and many other scientific and engineering fields.

1. Introduction

According to Huber (1981), the technical term "robust" was coined in 1953 (by G.E.P. Box), and the subject matter acquired recognition as a legitimate topic only in the mid-sixties. For a long time, theoretical statisticians tended to shun the subject as being inexact and "dirty". Lately, it seems that the pendulum has swung to the other extreme, and that "robust" has now become a magic word which is evoked to add respectability.

A few years ago, Fischler and Bolles (1981) and Haralick (1986) strongly and convincingly argued that computer vision, one of the most interesting and challenging areas in artificial intelligence, requires all of its algorithms to be robust. Their arguments explained what the realistic assumption about errors caused by low level image processing should be and why most of the existing algorithms in computer vision cannot be practically useful. As well recognized, all machine vision feature extractors, recognizers, and matchers explicitly or implicitly needed for low level computation are unavoidably error prone and seem to make occasional errors, which indeed are blunders. The realistic assumption for errors should be contaminated Gaussian noise, which is a regular white Gaussian noise with probability 1 - ϵ plus an outlier process with probability ϵ (Huber 1981). The least-squares estimator is very sensitive to minor deviations from the Gaussian noise model assumption. As a matter of fact, the occurrence of outliers definitely makes typical estimators such as ordinary least-squares estimators, the estimators of least virtue. Similar conclusions are also experimentally observed by (Roach and Aggarwal 1980; Fang and Huang 1982, 1984; Jerian and Jain 1983, 1984), to name a few. In Haralick et. al. (1989), the classical robust M-estimator is successfully applied to solve a single pose or a single rigid motion estimation from corresponding point data. However, lots of experiments conducted in Haralick et. al. (1989) conclusively show that the M-estimator only allows a low proportion of outliers. In order to solve the multiple pose or multiple rigid motion estimation problem, the M-estimator with a low robustness is of no help. From our point of view, the kind of estimators for motion analysis and perhaps many other computer vision problems must be highly robust. That's because it is usually hard to control the proportion of outliers in contaminated data and because, for a multiple rigid motion case relative to one rigid motion, other rigid motions and outliers all should be taken as outliers.

The paper presents a highly robust estimator called an MF-estimator for general regression. In Section 2, we explain why the classical M-estimator can't render a high robustness, and why the high robustness becomes possible only through partially but completely modeling the unknown log likelihood function. Partial modeling takes place by taking a number of heuristics implicit in the regression problem and the Bayes decision rule into consideration, while maximizing the log likelihood function. The MF-estimator is also presented in the section. In Section 3, we show the experimental results of applying both the proposed MF-estimator and the classical M-estimator to the typical location estimation problem. The MF-estimator exhibits a much higher robustness than the M-estimator does. The final section is the conclusion.

2. How to Develop a Highly Robust Estimator for General Regression.

2.2 Classical Robust Estimators Can't Render a High Robustness.

As well known, the classical robust estimator such as the M-estimator, L-estimator, and R-estimator (Huber 1981) possesses the following properties:

A. They have a reasonably good (optimal or nearly optimal) efficiency at the assumed noise distribution.

B. They are robust in the sense that the degradation in performance caused by a small number of outliers is relatively small.

C. Somewhat larger deviations from the assumed distribution does not cause a catastrophe.

The MF-estimator to be presented in the paper will represent a new brand of robust estimators that possess the above properties A and B as well as the following property C ', which is much stronger than the above property C:

The authors gratefully acknowledge the support of K.C. Wong Education Foundation, Hong Kong.

C '. The degradation in performance caused by somewhat larger deviations from the assumed distribution is still relatively small.

To construct a highly robust estimator we first need to say a few words about the minimax approach, which was widely used in developing the classical robust estimators such as the M-estimator, L-estimator, and R-estimator.

Assume that the true underlying shape F lies in some neighborhood P_ϵ of the assumed standard normal distribution Φ, where $P_\epsilon(\Phi) = \{F \mid F=(1-\epsilon)\Phi + \epsilon H, H \in \$\}$ with $\$$ representing the set of unknown foreign distributions, that the observations are independent with common distribution $F(x - \theta)$, and that the location parameter θ is to be estimated.

The minimax approach to robustly estimating the location parameter is based on minimizing its maximum asymptotic variance for all possible distributions $F \in P_\epsilon$. Suppose that F_0 attains maximum asymptotic variance for location in the set P_ϵ, then the corresponding probability density function f_0 has the form as follows (Huber 1981),

$$f_0(x) \quad = \frac{1-\epsilon}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right), \quad \text{for } |x| \le a$$

$$= \frac{1-\epsilon}{\sqrt{2\pi}} \exp\left(\frac{a^2}{2} - a|x|\right), \quad \text{for } |x| > a.$$

with the robust control parameter a and the outlier proportion parameter ϵ connected through

$$\frac{2\phi(a)}{a} - 2\Phi(-a) = \frac{\epsilon}{1-\epsilon}$$

where $\phi = \Phi'$ is the standard normal density. Moreover, the asymptotically efficient maximum likelihood estimate of location for F_0 (called by the M-estimator) in fact has been proven to possess certain minimax properties in P_ϵ (Huber 1981).

As clearly seen, for the M-estimate of location the robust distribution function F_0, instead of the standard normal distribution Φ, is exclusively used to model each possible distribution function F in the neighborhood $P_\epsilon(\Phi)$. As well known, the maximum likelihood estimate of location by using the standard normal distribution Φ leads to the least-squares estimate.

The reason why the function f_0 (or F_0) has a reasonably good efficiency at the assumed standard normal distribution Φ is that, with a high probability, the residual $|x - \theta|$ will be less than or equal to a, in other words, the M-estimator will do mostly the same job as the least-squares estimator does.

The reason why the function f_0 (or F_0) is robust when there are only a small number of outliers is that, with a probability about or less than $1 - \epsilon$, the residual $|x - \theta|$ will be less than or equal to a, for this major part of residuals the M-estimator behaves much like the least-squares estimator as demanded, and with a probability about or perhaps less than ϵ, the residual $|x - \theta|$ will be larger than a. To gain robustness, this minor part of residuals are also taken care of by the M-estimator. To see it clearly, we should point out that the unknown outlier process is modeled as the standard normal process by the least-squares estimator, i.e., $h(x) = \phi(x)$, or as the process whose probability density $h(x)$ equals zero as $|x| \le a$ or

$$\frac{1-\epsilon}{\epsilon\sqrt{2\pi}}\left[\exp\left(\frac{a^2}{2} - a|x|\right) - \exp\left(-\frac{x^2}{2}\right)\right] \text{ as } |x| > a$$

by the M-estimator. As seen, the important fact that the magnitude of an outlier residual is more likely larger than the magnitude of a nonoutlier residual is fully neglected by the least squares estimator, which produces a nonrobust estimate. The same fact is more or less reflected in the construction of the M-estimator. In fact, the possibility of an outlier having a residual magnitude being less than or equal to a is completely excluded by the M-estimator. From the previous expression, for $h(x)$, it is easy to verify that the probability density function of outliers has only two modes, approximately peaking around $\pm \frac{a + \sqrt{a^2+8}}{2}$ and flattens as residual magnitudes go beyond $\pm \frac{a + \sqrt{a^2+8}}{2}$.

The reason why the function f_0 cannot render high robustness is that the robust control parameter a is tied to the outlier proportion parameter ϵ and approaches zero as ϵ tends to 1. That means when the outlier proportion ϵ becomes larger, all nonoutlier residuals ought to be smaller to comply with the M-estimator. If the nonoutlier residuals do not become smaller correspondingly, some or even all of them will be treated as outliers by the M-estimator and the contained information which is useful for location estimation thus lost as a result. This often leading to bad estimates. Things become even worse when a limited sample size is used in the estimation process since the M-estimator cannot recognize enough nonoutliers necessary for a reasonably good estimate.

It seems, if we model each possible distribution function F in the neighborhood $P_\epsilon(\Phi)$ by using a single fixed robust distribution function, a high robustness will never be possible. In the paper, we attempt to fit the most likely values to each unknown foreign probability density function at the observed data individually instead of completely modeling each unknown foreign probability density function. Specifically, we will use more heuristic reasoning rather than purely mathematical reasoning. In fact, we will combine the Bayes statistical decision rule with a number of deeply explored heuristic considerations and turn the general robust regression problem into a model fitting problem, which is not only more flexible but also more tractable.

2.2 Partially Modeling Log Likelihood Function Using Heuristics

Assume that p unknown parameters $\theta_1, \ldots \theta_p$, shortened as the vector θ, are to be estimated from N observations $y_1, \ldots y_N$, each of which is a m-dimensional vector. Assume $f_k : R_p \to R_m$, $k = 1, \ldots, N$. Let r_k be the residual between y_k and $f_k(\theta)$, i.e., $r_k = y_k - f_k(\theta)$. Furthermore, assume each single observation y_k with probability $1 - \epsilon$ is a "good" one, i.e., not an outlier, and with probability ϵ, is a "bad" one, i.e., an outlier, where $0 \le \epsilon \le 1$. In the former case, the residual r_k is Gaussian distributed with zero mean and an unknown covariance matrix $\sigma^2 I_m$, in the latter obeys an unknown foreign distribution. All r_k's are independent, identically distributed with the common probability density function f, namely

$$f(r_k) = \frac{1-\epsilon}{(\sqrt{(2\pi}\,\sigma)^m} \exp\left(-\frac{\|r_k\|^p}{2\sigma^2}\right) + \epsilon h(r_k) \qquad (1)$$

where h is unknown. Thus, the log likelihood function of observing $y_1, \ldots y_N$ conditioned on $\theta_1, \ldots \theta_p, \sigma, \epsilon, h(r_1), \ldots, h(r_N)$ are expressed by Q as follows,

$$\begin{aligned} Q &= \log P(y_1, \cdots, y_N \mid \theta_1, \cdots, \theta_p, \sigma, \epsilon, h(r_1), \cdots, h(r_N)) \\ &= \log \prod_k P(y_k \mid \theta_1, \cdots, \theta_p, \sigma, \epsilon, h(r_k)) \\ &= \sum_k \log P(y_k \mid \theta_1, \cdots, \theta_p, \sigma, \epsilon, h(r_k)) \qquad (2) \\ &= \sum_k \log f(r_k) \end{aligned}$$

which, when combined with (1), comprises the first model assumption. To be successful in gaining high robustness, we need to further explore possible Heuristics implicit in the regression problem instead of hurrying in maximizing the log likelihood function Q.

According to (1), the probability of the observation y_k being a nonoutlier conditioned on θ, σ, ϵ, $h(r_k)$ is given by λ_k:

$$\lambda_k = \frac{1-\epsilon}{(\sqrt{2\pi}\sigma)^m} \exp\left(-\frac{\|r_k\|^2}{2\sigma^2}\right) / f(r_k) \tag{3}$$

Using the Bayes Statistical decision rule, we can classify the observation y_k as a nonoutlier if $\lambda_k > 0.5$ or an outlier otherwise. Let G denote the indices of "good" observations and B the indices of "bad" observations, where

$$\begin{aligned} G &= \{k : \lambda_k > 0.5\} \\ B &= \{k : \lambda_k \le 0.5\} \end{aligned} \tag{4}$$

The second model assumption consists of the following heuristic condition:

$$\begin{aligned} \frac{\#G}{N} &= 1 - \epsilon \\ \frac{\#B}{N} &= \epsilon \end{aligned} \tag{5}$$

where # represents "the number of".

To obtain a reliable estimation, a minimum of "good" observations are demanded. The minimum number, denoted as L, is very problem-dependent and can be experimentally or theoretically determined. The results will no longer be reliable when $\#G$ drops below L. Thus, for the third model assumption we use the following heuristic condition

$$\#G \ge L \tag{6}$$

The fourth model assumption states that all $h(r_k)$ could be taken as equal, namely

$$h(r_k) = \delta, \; k = 1, \cdots, N \tag{7}$$

It comes from the heuristic consideration that for two different observations the one with a smaller residual magnitude should be more likely a nonoutlier than the other one. It means the partition $\{G,B\}$ ought to have the following property:

$$\max \{ \|r_k\| : k \in G \} < \min \{ \|r_k\| : k \in B \} \tag{8}$$

Let

$$g = \min \left\{ \frac{1-\epsilon}{(\sqrt{2\pi}\sigma)^m} \exp\left(-\frac{\|r_k\|^2}{2\sigma^2}\right) : k \in G \right\} \tag{9}$$

$$b = \max \left\{ \frac{1-\epsilon}{(\sqrt{2\pi}\sigma)^m} \exp\left(-\frac{\|r_k\|^2}{2\sigma^2}\right) : k \in B \right\}$$

Then, it holds that

$$g > b \tag{10}$$

which indicates that the partition $\{G, B\}$ can be equally well defined by

$$\begin{aligned} G &= \left\{ k : \frac{1-\epsilon}{(\sqrt{2\pi}\sigma)^m} \exp\left(-\frac{\|r_k\|^2}{2\sigma^2}\right) > \epsilon\delta \right\} \\ B &= \left\{ k : \frac{1-\epsilon}{(\sqrt{2\pi}\sigma)^m} \exp\left(-\frac{\|r_k\|^2}{2\sigma^2}\right) \le \epsilon\delta \right\} \end{aligned} \tag{11}$$

where δ can be an arbitrary number in the interval $\frac{1}{\epsilon}[b,g]$. This means that the choice of δ can be quite broad. This observation will play an important role in our algorithmic development.

Let the fitting error of detected nonoutliers be defined by $\max\{\|r_k\| : k \in G\}$ and the distance of detected outliers from detected nonoutliers can be defined by $\min\{\|r_k\| : k \in B\} - \max\{\|r_k\| : k \in G\}$. For each set of parameters $\{\theta_1, \ldots \theta_p, \sigma, \epsilon, \delta\}$, a partition $\{G,B\}$ of $\{1, \ldots, N\}$ can be generated by (4). It is reasonable to search for a parameter set having the least possible fitting error and the largest possible distance. This requirement will amount to minimizing the following cost function:

$$C(G,B) = \frac{\#G}{N} \max \|r_k\| + \frac{\#B}{N} \frac{1}{\min_B \|r_k\| - \max_G \|r_k\|} \tag{12}$$

The minimal cost requirement comprises the last heuristic condition, namely,

$$C(G,B) = \min \tag{13}$$

Now the general robust regression problem can be stated as finding the parameter set $\{\theta_1, \ldots \theta_p, \sigma, \epsilon, \delta\}$ through maximizing the log likelihood function of the model which is partially modeled by basic assumptions consisting of (1)-(2), (4)-(5), (6), (7), and (13). That is,

$$\max Q(\theta_1, \cdots, \theta_p, \sigma, \epsilon, \delta)$$

$$= \sum \log \left\{ \frac{1-\epsilon}{(\sqrt{2\pi}\sigma)^m} \exp\left(-\frac{\|r_k\|^2}{2\sigma^2}\right) + \epsilon\delta \right\}$$

subject to

$$\begin{aligned} &\#G \ge L \\ &\#G = (1-\epsilon)N \\ &C(G,B) = \min \end{aligned} \tag{14}$$

where conditions (1)-(2) and (7) have been included in the model log likelihood function Q, and L is a problem dependent number to be experimentally or theoretically determined.

As clearly seen, combining the Bayes statistical decision rule with a number of heuristic considerations has turned the general robust regression problem into a more appropriate model fitting problem. The algorithm developed thereof is called by the name MF-estimator.

2.3 Discussion

The M-estimator, the L-estimator and R-estimator are all residual based, the MF-estimator is no exception. Assume $m = 1$ (i.e., 1-dimensional case) and that the true parameters are $\theta_1, \ldots \theta_p$ forming N true residuals, $r_k = y_k - f_k(\theta_1, \ldots, \theta_p)$, $k = 1, \ldots, N$. Suppose those N residuals can be distinctively divided into two groups, the nonoutlier set G and the outlier set B, so that each residual in G is small in magnitude and much smaller than each residual in magnitude in B. To use the M-estimator effectively, the following three conditions must be satisfied:

$\#G \le (1-\epsilon)N$ or $\#B \le \epsilon N$

$|r_k| \le a$, r_k distributed around zero if $k \in G$

$|r_k| > a$, r_k distributed around $\pm \dfrac{a + \sqrt{a^2+8}}{2}$ or beyond if $k \in B$

Suppose we add more outliers into B but leave G alone so that the two sets, G and newly formed B, can still be distinctively divided as before. Then the second condition will not hold any more since the number of outliers becoming larger leads to a —> 0. Comparatively, adding more qualified outliers will not influence the effective use of the MF-estimator. The following two conditions are necessary for a good use of the M-estimator, the L-estimator, the R-estimator and are sufficient for good use of the MF-estimator:

1. A large number of qualified nonoutliers permitting a satisfactory nonlinear least-squares fitting;

2. Qualified outliers are far from nonoutliers based on the residual consideration.

Unlike the M-estimator, which uses single probability density function to model all possible unknown outlier probability densities, the MF-estimator only assumes that $h(r_k)$ are equally valued. It does not assume which value they should take, nor about the whole shape of $h(.)$. From residual based consideration, the assumption of all $h(r_k)$ being equally valued is reasonable, especially when only a limited sample size is allowable as is the case in many computer vision problems.

From a practical point of view, however, it is hard to guarantee that observed or processed data have a large enough distance between outliers and nonoutliers. Therefore, in order for the proposed MF-estimator to be useful, this problem must be solved. In other words, the second condition stated as above must be removed from a proposed MF-estimator. This and other issues will be talked over in Section 4.

2.4 MF-Estimator

In this subsection we will show how to solve the model fitting problem, i.e. (14). As a matter of fact, what we really need is to maximize Q w.r.t. θ, σ, ϵ, δ subject to $\#G = (1-\epsilon)N$. As will be seen, it's quite simple to include condition (6), i.e. $\#G \ge L$, in the MF-estimator. As for condition (13), i.e., $C(G,B) = \min$, it will be reached through only a few trials for the initial value of δ. That is because there exists a whole interval instead of a single value for a good workable choice of δ, as stated before.

To maximize Q subject to $\#G = (1-\epsilon)N$, we basically follow the gradient-ascent rule. The partial derivatives of Q w.r.t. $\theta_1, ..., \theta_p$, σ, ϵ, δ are as follows,

$$\frac{\partial Q}{\partial \theta_i} = \frac{1}{\sigma^2}\sum_k \lambda_k r_k \frac{\partial f_k}{\partial \theta_i}, \quad i=1, ..., p \tag{15}$$

$$\frac{\partial Q}{\partial \sigma} = \frac{1}{\sigma}\left\{\frac{1}{\sigma^2}\sum_k \|r_k\|^2 \lambda_k - m\sum_k \lambda_k\right\} \tag{16}$$

$$\frac{\partial Q}{\partial \epsilon} = \frac{N-\lambda}{\epsilon} - \frac{\lambda}{1-\epsilon}, \quad \text{where } \lambda = \sum_k \lambda_k \tag{17}$$

$$\frac{\partial Q}{\partial \delta} = \frac{N-\lambda}{\delta} \tag{18}$$

We then define the location step, the scale step, and the distribution step to determine $\theta_1, ..., \theta_p$, σ, ϵ, δ respectively, from the n-th to the (n+1)-th iterative step, as follows,

The Location Step. Because of (15), to assure

$$\sum_i \frac{\delta Q}{\delta \theta_i}\Delta\theta_i \ge 0$$

it is necessary to have

$$\sum_k \lambda_k r_k \left(\sum_i \frac{\partial f_k}{\partial \theta_i}\Delta\theta_i\right) \ge 0,$$

which is satisfied if there holds

$$\lambda_k r_k = \left(\frac{\partial f_k}{\partial \theta_1}, ..., \frac{\partial f_k}{\partial \theta_p}\right)\Delta\theta, \quad k=1, ..., N \tag{19}$$

Let

$$X = \begin{pmatrix} \frac{\partial f_1}{\partial \theta_1}, & ..., & \frac{\partial f_1}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_N}{\partial \theta_1}, & ..., & \frac{\partial f_N}{\partial \theta_p} \end{pmatrix} \tag{20}$$

If the singular value decomposition of X is

$$X = UWV' \tag{21}$$

(where $U'U = I_p$, $V'V = VV' = I_p$ and $W = \text{diag}\,[\sigma_1^2, ..., \sigma_p^2]$), then the least-squares solution to (19) is given by

$$\Delta\theta = VW^{-1}U'(\lambda_1 r_1, ..., \lambda_N r_N)' \tag{22}$$

The Scale Step. Because of (16), we simply define the value of σ at the (n+1)-th iterative step as follows,

$$\sigma^2 = \frac{1}{m}\sum_k \frac{\lambda_k}{\lambda}\|r_k\|^2 \tag{23}$$

The Distribution Step. First, we consider how to make a meaningful change of the outlier proportion, i.e., ϵ. This is quite simple. The change of ϵ from the n-th step to the (n+1)-th step should be made so that $\dfrac{\partial Q}{\partial \epsilon} \ge 0$ and $0 < \epsilon + \Delta\epsilon < 1$.

We need only to set

$$\Delta\epsilon = \begin{cases} -\alpha_1\epsilon, & \text{if } \dfrac{1}{N}\left(\dfrac{N-\lambda}{\epsilon} - \dfrac{\lambda}{1-\epsilon}\right) < -\alpha_2\ ; \\ \alpha_1(1-\epsilon), & \text{if } \dfrac{1}{N}\left(\dfrac{N-\lambda}{\epsilon} - \dfrac{\lambda}{1-\epsilon}\right) > \alpha_2\ ; \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

because of (17), where $0 < \sigma_1, \sigma_2 < 1$.

To make a meaningful change of δ, however, we need not only to take care of

$$\frac{\partial Q}{\partial \epsilon}\Delta\epsilon + \frac{\partial Q}{\partial \delta}\Delta\delta \ge 0$$

and

$$\delta + \Delta\delta > 0$$

but also to reinforce the heuristic condition $\#G = (1 - \epsilon)N$, (see (14)). As pointed out before, for each set of parameter trial values, a partition $\{G,B\}$ defined by (4) can be calculated. Suppose $\#G > (1 - \epsilon)N$ with the current trial value of ϵ and calculated G. That means the outliers are estimated to be lower than needed. To correct the unbalancing, we need only to increase $\epsilon\delta$. If the change of ϵ has already been determined as positive previously, we leave δ alone. Otherwise, we need to increase δ. Similarly, when $\#G < (1 - \epsilon)N$ and $\Delta\epsilon > 0$, we need to decrease δ. Combining these ideas with the requirements that

$$\frac{\partial Q}{\partial e}\Delta\epsilon + \frac{\partial Q}{\partial \delta}\Delta\delta \geq 0$$

and

$$\delta + \Delta\delta > 0$$

together, we set

$$\Delta\delta = \begin{cases} t\delta, & \text{if } \dfrac{\#G}{(1-\epsilon)N} > 1 + \alpha_3 \text{ and } \Delta\epsilon < 0; \\ -t\delta, & \text{if } \dfrac{\#G}{(1-\epsilon)N} < 1 - \alpha_3 \text{ and } \Delta\epsilon > 0. \end{cases} \tag{25}$$

with

$$t = \min\left\{\alpha_4, \frac{\alpha_5}{N-\lambda}\left(\frac{N-\lambda}{\epsilon} - \frac{\lambda}{1-\epsilon}\right)\Delta\epsilon\right\},$$

$$0 < \alpha_3, \alpha_4, \alpha_5 < 1.$$

All λ_k, λ, r_k, X, etc. in (20)-(25) are calculated using the trial values for θ, σ, ϵ, δ at the n-th iterative step.

Up to now, three of the five heuristic conditions have been taken care of in the algorithmic development. These are (1)-(2), (4)-(5), and (7). It is time to include condition (6) and condition (13) in our algorithmic development. We can simply summarize the MF-estimator as follows:

Step 1. Chose an initial approximation: θ^0, σ^0, ϵ^0, δ^0.

Step 2. Iterate. Given the estimation: θ^n, σ^n, ϵ^n, and δ^n at the n-th step, compute the change $\Delta\theta$, $\Delta\sigma$, $\Delta\epsilon$, and $\Delta\delta$ by using (22)-(25).

Step 3. After convergence, compute the corresponding partition and cost, i.e., $C(G,B)$.

Step 4. If $\#G \geq L$ and $C(G>B) < \zeta$, where ζ is a prespecified, small cost bound, then go to step 7.

Step 5. If $\#G < L$, then the appropriate initial value of δ, which leads to the most preferable partition $\{G,B\}$ with $\#G \geq L$ and $C(G,B) = \min$ cannot go beyond the interval $(0,\delta^0)$ and thus will be found by using the golden bisection technique to the interval $(0,\delta^0)$ while following steps 2-3. Go to step 7.

Step 6. $\delta^0 \leftarrow 2\delta^0$ and go to step 2.

Step 7. Reestimate the parameters $\theta_1, \ldots, \theta_p$ by iteratively applying the nonlinear least-squares method to the good observation data, i.e., G, to increase the estimation precision. Stop.

3. Simplest Location Estimation

To prove the MF-Estimator possesses a high robustness, we are to conduct experiments with the simplest location estimation.

Suppose there are N one-dimensional observations, $x_1, \ldots, x_N$, on a location parameter θ. The observations are independent with common density function $f(x - \theta)$ which is modeled by

$$f(x) = \frac{1-\epsilon}{\sqrt{2\pi}}\exp\left(-\frac{bd^2}{2}\right) + \epsilon h(x)$$

where $h(x)$ is unknown. To estimate the location θ, we use both the M-estimator and the MF-estimator.

The location parameter θ is randomly selected. The $N - M$ random numbers from $N(0,1)$ are generated and added to θ to form $N - M$ good observations about θ. The remaining M observations are supposed to form outliers. They are randomly generated, for instance, from a uniform distribution, under the condition that each of them keeps a certain distance from θ. In our experiments, the outlier residual magnitudes are managed as being larger than one. All nonoutlier residual magnitudes do not go beyond $\frac{\sqrt{2}}{2} \approx 0.707$ with the probability 95%. Thus the outlier residuals and nonoutlier residuals are "distinctively" divided.

The experimental results with hundreds of thousand trials, convincingly show that the MF-estimator performs much better than the M-estimator. For a sample size N = 10, 20, 30, the largest possible outlier proportion which can be reached, is ϵ = 0.5, 0.5, 0.5, respectively, by the M-estimator, or ϵ = 0.9, 0.9, 0.93, respectively, by the MF-estimator. For the MF-estimator, only two good points or observations are needed in order to get an accurate location estimate no matter how many outliers occur.

4. Conclusion

A highly robust estimator called the MF-estimator has been discussed. It's comparison with the classical M-estimator in the simplest location estimation case seems to be promising. The superior performance of the MF-estimator has also been proven in many other application topics such as automatic selection of multiple thresholds, multiple motions from a mixture point corresponding data, optic flow-multiple motions, multiple views-multiple motions, and so forth. We will present those results in separate papers in the near future. However, more theoretical and experimental work remains to be done. There are at least two major problems remaining to be solved. They are:

1. How to relieve the requirement for good initial approximation in order for the MF-estimator to work?

2. How to remove the distance condition which guarantees that outliers and nonoutliers can be distinctively divided in order for the MF-estimator to be practically useful?

To relieve the requirement for good initial approximation, we must realize that it is actually a problem related to global optimization. It is common to be trapped in local maxima (or local minima), while following the gradient-ascent (or gradient-descent) rule. To be initial-value independent or avoid being trapped in local extrema, we must not strictly observe the gradient-ascent (or gradient-descent) rule for maximization (or minimization), very much like the stochastic search rule used in Boltzmann machine to find the global minimum of an energy function. To directly satisfy the requirement for good initial approximate, we must use knowledge about parameter space to be searched in order to have an efficient search.

To solve the distance problem, we should divide the whole set of residuals into three sets: a good one, a bad one, and a fuzzy one. The good one will contribute to least-squares parameter estimation, the bad one will have a far enough distance from the good one (that is why it is called bad) and thus not prevent the good one from least-squares parameter estimation. The fuzzy one has neither a far enough distance from the good one nor small enough residuals. Thus, this one will cause the good one to do least-squares parameter estimation. This gives us a hint to shut down the participation of the fuzzy one in robust estimation process.

Those and others comprised are further researched.

References

[1] Fang, J.-Q. and Huang T.S., "A Corner Finding Algorithm for Image Analysis and Registration", *AAAI-82*, pp. 46-49.

[2] Fang J.-Q. and Huang T.S, "Solving Three-Dimensional Small-Rotation Motion Equations: Uniqueness, Algorithms and Numerical Results", *Computer Vision, Graphics, and Image Processing*, 26, 1984, pp. 183-206.

[3] Fang J.-Q and Huang T.S., "Some Experiments on Estimating the 3-D Motion Parameters of a Rigid Body from Two Consecutive Image Frames", *IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-6*, No. 5, 1984, pp. 545-554.

[4] Fischler M.A. and Bolles R.C., "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", *Communications of the ACM*, Vol. 24, No. 6, June, 1987.

[5] Haralick R.M., "Computer Vision Theory: The Lack Thereof", *Computer Vision, Graphics, and Image Processing*, 36, 1986, pp. 372-386.

[6] Haralick R.M., Joo H., Lee C.N., Zhuang X., Vaidya V.G., and Kim M.B., "Pose Estimation from Corresponding Point Data", *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 19, No. 6, November/December, 1989.

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

[8] Jerian C. and R. Jain, "Determining Motion Parameters for Scenes with Translation and Rotation", *Proc. ACM Siggraph/Sigart Interdisciplinary Workshop on Motion*, Toronto, April, 1983, pp. 71-77 and PAMI-6, No. 4, July 1984, pp 523-530.

[9] Roach J.W. and Aggarwal J.K., "Determining the Movement of Objects from a Sequence of Images", *IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-2*, No. 6, Nov., 1980, pp. 554-562.