# AN ADAPTIVE ROBUST ESTIMATOR FOR SCALE IN CONTAMINATED DISTRIBUTIONS

*Ramon Brcich, Christopher L. Brown and Abdelhak M. Zoubir*

Signal Processing Group, Institute of Telecommunications
Darmstadt University of Technology, Merckstrasse 25, D-64283 Darmstadt, Germany.
Email : {r.brcich,chris.brown,zoubir}@ieee.org

**ABSTRACT**

We consider the problem of scale estimation when a nominal distribution is contaminated. Knowledge of the scale is necessary in many signal detection and estimation problems and poor estimates of the scale can have deleterious effects on subsequent processing. The approach considered here is based on the M-estimation concept of Huber, but employs a score function which is a linear combination of basis functions whose weights are adaptively estimated from the observations. Results suggest that this adaptivity increases robustness over static M-estimators.

## 1. INTRODUCTION

Inspection of recent signal processing literature reveals renewed interest in robust methods. It is becoming widely accepted that robust performance can be achieved for only slight performance decrease in the nominal scenario. Emphasis has been on location estimation in the signal plus additive noise model and covariance estimation in array signal processing, such as [1, 2]. Here we concentrate on the scale estimation problem which is often concomitant with location estimation.

Consider the general signal in additive noise model,

$$y_n = s_n(\boldsymbol{\theta}) + \sigma x_n, \qquad n = 1, \ldots, N, \tag{1}$$

where $x_n$ is (standardised) iid noise, $\sigma$ is the noise scale and the signal $s_n$ is parameterised by $\boldsymbol{\theta}$. As noted in [3], the scale is a nuisance parameter, but is generally important in achieving good estimates of $\boldsymbol{\theta}$, although exceptions exist [4]. Here we only consider the problem of scale estimation as this can be incorporated into existing estimators for $\boldsymbol{\theta}$. Consequently, the scheme here will assume that an estimate of $\boldsymbol{\theta}$, and hence $s_n(\boldsymbol{\theta})$, is available.

The observations are assumed to follow a nominal distribution subject to some contamination or deviation. We consider robustness to imply insensitivity to such deviations. The noise will be modelled as follows,

$$f_X(x) = (1-\varepsilon) f_N(x) + \varepsilon f_C(x), \tag{2}$$

where $f_N$ is the symmetric nominal probability density function (pdf) with zero mean and $f_C$ is the symmetric, zero mean, contaminating distribution which occurs with probability $\varepsilon$. The noise pdf, $f_X(x)$, is standardised to have a scale of unity, although it is usually assumed that the contaminating distribution has a much larger scale than the nominal as it describes the presence of outliers in the observations.

It is well known that the sample standard deviation is not a robust estimator for the scale. Robust estimates of scale have been developed, such as Huber's minimax M-estimator [5]. A similar approach is used here except that the score function used in the M-estimator is able to adapt to changes in the contaminating distribution.

The outline of this paper is as follows. In Section 2 we briefly review the maximum likelihood estimator (MLE) and M-estimator for scale, before developing the adaptive robust scale M-estimator in Section 3. Simulations and a discussion follow in Section 4, before concluding in Section 5.

## 2. MAXIMUM LIKELIHOOD AND M-ESTIMATORS OF SCALE

For the reasons described in the introduction, we assume that the observations consist of noise only, the signal component having been removed. Given the (standardised) noise density, $f_X(x)$, the MLE for $\sigma$ is

$$\hat{\sigma}_{\mathrm{ML}} = \underset{\sigma}{\operatorname{argmin}} \sum\nolimits_{n=1}^{N} \rho\left(\frac{y_n}{\sigma}\right) \tag{3}$$

where $\rho(x) = -\log f_X(x)$ is the log-likelihood function of $f_X(x)$. Equivalently, the ML solution may be found by solving the log-likelihood equation,

$$\sum_{n=1}^{N} \psi\left(\frac{y_n}{\sigma}\right) = 0 \tag{4}$$

for $\sigma$ where

$$\psi(x) = -1 + x\dot{\rho}(x) = -1 - x\frac{\dot{f}_X(x)}{f_X(x)} \tag{5}$$

is the scale score function of $f_X(x)$ and $\dot{f}_X(x)$ denotes the derivative of $f_X(x)$ with respect to $x$. Note that, although related, scale score functions are not to be confused with location score functions.

The MLE requires $f_X(x)$ to be known and so without a priori knowledge of $f_X(x)$, estimation of $\sigma$ cannot be optimal in the ML sense. Therefore, performance is uncertain with respect to deviations from the nominal distribution.

In an M-estimator the log-likelihood function $\rho(x)$ of the MLE is replaced with a similarly behaved penalty function, $\varrho(x)$. The penalty function is chosen to confer robustness on the estimator under deviations from the assumed density. It follows that similarly to ML, $\sigma$ can be estimated from (4) with $\psi(x)$ replaced by $\varphi(x) = -1 + x\dot{\varrho}(x)$,

$$\sum_{n=1}^{N} \varphi\left(\frac{y_n}{\sigma}\right) = 0. \tag{6}$$

When $f_X(x)$ is unknown, the distance between the penalty and log-likelihood functions is uncertain. Selection of the penalty function is then of prime importance in ensuring the performance of the estimator is not highly sensitive to $f_X(x)$ but is robust over a wide class of noise models.

### 2.1. Huber's minimax estimator for scale

With this in mind, Huber proposed that a clipped quadratic score function be used in the M-estimator for scale,

$$\varphi_H(x;k) = \min(x^2, k^2) - \delta = \begin{cases} x^2 - \delta, & |x| \leq k \\ k^2 - \delta, & |x| > k, \end{cases} \tag{7}$$

which minimises the maximum relative asymptotic variance of the scale estimate. $\delta$ is determined such that the estimator is unbiased for a nominal Gaussian distribution,

$$\delta = 1 - 2k\phi(k) + 2(1-k^2)(\Phi(k) - 1) \tag{8}$$

where $\phi(x)$ and $\Phi(x)$ are the standard Gaussian pdf and cdf respectively. The parameter $k$ controls the sensitivity of the estimator to the contaminating distribution and increases as $\varepsilon$, the proportion of outliers, decreases. Hence, as $\varepsilon \to 0$, $k \to \infty$ and the score function used in the M-estimator becomes a quadratic function, reducing the estimator to the sample standard deviation. The optimum value of $k$ is determined from $\varepsilon$ as detailed in [5].

## 3. ADAPTIVE ROBUST SCALE ESTIMATION

In [6, 3] an adaptive M-estimator for location was developed by modelling the location score function as a linear combination of basis functions. To ensure robust behaviour against outliers, the bases were chosen to be the location score functions of several heavy tailed distributions.

The approach taken here is similar where the scale score function is parametrically modelled as a linear combination of basis functions,

$$\varphi(x) = \sum_{q=1}^{Q} a_q g_q(x) = \boldsymbol{a}^{\mathsf{T}} \boldsymbol{g}(x), \tag{9}$$

where the weights are $\boldsymbol{a} = (a_1, \ldots, a_Q)^{\mathsf{T}}$ and the bases are $\boldsymbol{g}(x) = (g_1(x), \ldots, g_Q(x))^{\mathsf{T}}$.

The bases are chosen for their ability to approximate $\psi(x)$. For instance, the bases can simply be a set of score functions obtained from distributions known to be close, in some sense, to $f_X(x)$. The weights can then be chosen to minimise some measure of distance between $\varphi(x)$ and $\psi(x)$ or to maximise the performance of the estimator.

A sensible measure of distance between $\varphi(x)$ and $\psi(x)$ is the mean squared error (MSE), from which the weights are defined as

$$\boldsymbol{a} = \underset{\boldsymbol{a}}{\operatorname{argmin}}\, \mathsf{E}\left[(\varphi(x) - \psi(x))^2\right]. \tag{10}$$

$\boldsymbol{a}$ is then obtained as the solution to the normal equations,

$$\mathsf{E}\left[\boldsymbol{g}(x)\boldsymbol{g}^{\mathsf{T}}(x)\right] \boldsymbol{a} = \mathsf{E}[\boldsymbol{g}(x)\psi(x)], \tag{11}$$

so that

$$\boldsymbol{a} = \mathsf{E}\left[\boldsymbol{g}(x)\boldsymbol{g}^{\mathsf{T}}(x)\right]^{-1} \mathsf{E}[\boldsymbol{g}(x)\psi(x)]. \tag{12}$$

This estimate exists given that the following conditions hold

C1. $\mathsf{E}[\boldsymbol{g}(x)\boldsymbol{g}^{\mathsf{T}}(x)]$ is finite and nonsingular.

C2. $\mathsf{E}[\boldsymbol{g}(x)\psi(x)]$ is finite.

Further implications of these conditions on the behaviour of the bases are discussed in the Appendix. The presence of the true scale score function $\psi$ in (12) can be avoided if

$$\lim_{x \to \pm\infty} x g_q(x) f_X(x) = 0, \tag{13}$$

in which case it can be shown that

$$\mathsf{E}[x\dot{\boldsymbol{g}}(x)] = \mathsf{E}[\boldsymbol{g}(x)\psi(x)] \tag{14}$$

and the estimator for $\boldsymbol{a}$ becomes

$$\boldsymbol{a} = \mathsf{E}\left[\boldsymbol{g}(x)\boldsymbol{g}^{\mathsf{T}}(x)\right]^{-1} \mathsf{E}[x\dot{\boldsymbol{g}}(x)], \tag{15}$$

which is independent of the true scale score function. In practice the expectation is replaced with an empirical mean. Note that this equation is slightly different from the case of estimating the weights of location score functions, as described in [3].

For similar theoretical and practical considerations as those articulated in [3], we impose the following constraints on $a_q$

$$0 \leq a_q \leq 1, \qquad \sum_{q=1}^{Q} a_q = 1. \tag{16}$$

The final algorithm comprises of the two alternating steps: estimate the scale score function and then find the M-estimate of the scale based upon the estimated scale score function. The algorithm is summarised in Table 1.

**Table 1**. Iterative algorithm for the adaptive robust scale estimator.

**Step 1.** Initialisation: Set $i = 0$. Obtain an initial estimate of $\sigma$, $\hat{\sigma}_0$.

**Step 2.** Scale the observations: $\hat{x}_n = y_n / \hat{\sigma}_i$.

**Step 3.** Estimate the scale score function: From $\hat{x}_n$, estimate the weights,

$$\hat{\boldsymbol{a}} = \left( \sum_{n=1}^{N} \boldsymbol{g}(\hat{x}_n)\boldsymbol{g}^{\mathsf{T}}(\hat{x}_n) \right)^{-1} \sum_{n=1}^{N} \hat{x}_n \dot{\boldsymbol{g}}(\hat{x}_n),$$

subject to (16). The scale score function estimate is $\varphi(x) = \hat{\boldsymbol{a}}^{\mathsf{T}} \boldsymbol{g}(x)$.

**Step 4.** Update the estimate of $\sigma_i$ to $\sigma_{i+1}$: Solve (6).

**Step 5.** Check for convergence: If $|\hat{\sigma}_{i+1} - \hat{\sigma}_i| < \epsilon |\hat{\sigma}_i|$ stop, otherwise set $i \to i+1$ and go to step 2.

## 4. SIMULATIONS AND DISCUSSION

To test the adaptive robust scale estimator, consider the following scenario. We wish to estimate the scale $\sigma_N$ of a nominal Gaussian process, $X_N$. The observations include a component from a contaminating process $X_C$. Of course if $X_C$ has much larger scale than $X_N$ these contaminating observations may appear as outliers.

The traditional approach to scale estimation ignores the presence of outliers and uses the sample standard deviation while a robust approach could use Huber's minimax estimator. The problem with this is that the best point, $k$, at which to clip the quadratic function is dependent on the relative scale of the nominal and contaminating processes, as well as other properties of $X_C$. As noted previously, the sample standard deviation can be obtained by setting $k = \infty$, for no clipping, hence we denote this estimator $\varphi_H(x; k = \infty)$. The adaptive robust scale estimator will be compared to these estimators where the basis functions consist of a number of clipped quadratic functions with different $k$.

For the simulation results shown here we allow $0.1 \leq \sigma_N \leq 10$, $0 \leq \varepsilon \leq 0.1$ and set $X_C \sim \mathcal{N}(0, 100)$. Overall, the observed process can be modelled as

$$X \sim (1-\varepsilon)\mathcal{N}(0, \sigma_N^2) + \varepsilon\mathcal{N}(0, 100) \quad .$$

The number of observations is $N = 1000$. We consider 3 clipped quadratic bases, $\varphi_H(x; k), k = 1, 2, 3$, in the adaptive robust estimator. These functions are shown in Figure 1.

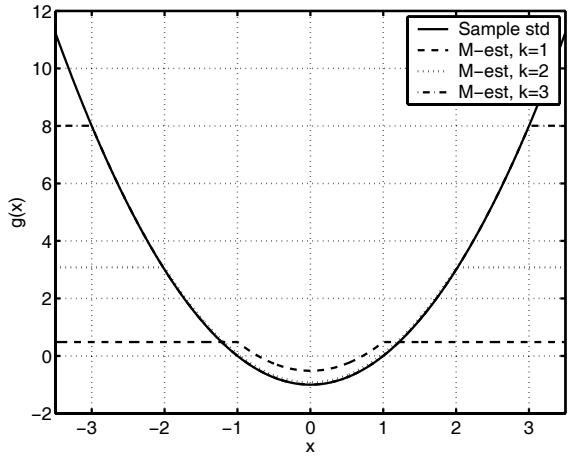



**Fig. 1**. Score functions used in simulations.

The MSEs of the proposed adaptive M-estimator and four M-estimators with static score functions, $\varphi_H(x; k), k = 1, 2, 3, \infty$, were evaluated over 500 Monte Carlo realisations. Each estimator was found to be best under *some* parameter settings. However, none was uniformly best, or even uniformly better than any other estimator.

Significantly, when compared to the static M-estimators, the relative performance of the proposed estimator was fairly constant over the parameter space $(\varepsilon, \sigma_N)$. This is shown in Figure 2 where the darker the square, the better the relative performance for that particular parameter setting – a black square indicates the best method (lowest MSE) while white indicates the worst performance (highest MSE). Of the 5 estimators, the proposed estimator's MSE was usually 2nd or 3rd lowest – it was occasionally best, but never 4th or 5th (last). By contrast, the others showed much more variable relative performance, see Figure 3. Therefore, it could be claimed that the proposed estimator is more robust in this case.

Since the full set of results is too large to show here, Table 2 shows the MSE of the estimators versus $\varepsilon$ with $\sigma_N = 1$. As expected, all methods perform well for un-contaminated observations ($\varepsilon = 0$), however as contamination increases, the sample

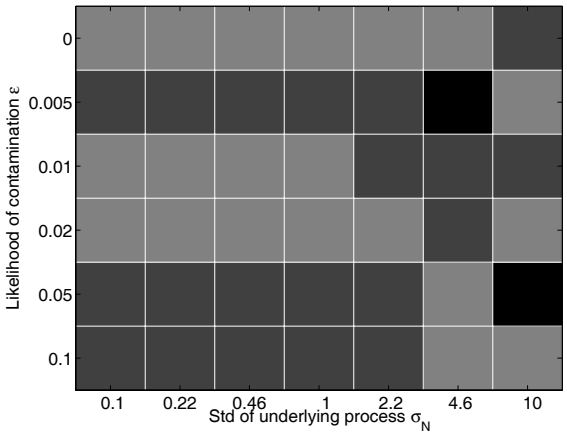



**Fig. 2**. Relative performance of the proposed adaptive M-estimator (darker squares indicate better performance).

standard deviation breaks down. The adaptive M-estimator has similar performance to the static M-estimator with $k = 2$. In the case shown here, $k = 1$ is sufficiently wide to capture enough observations from the $X_N$, while rejecting those from $X_C$, hence its good relative performance. Conversely, when $\sigma_N$ is large, the higher value for $k$ in the M-estimator using $\varphi_H(x; k = 3)$ is needed to avoid clipping too much data from $X_N$.

Finally we comment on the bias of the adaptive M-estimator. When each of the bases used in the adaptive M-estimator is considered individually in a static M-estimator, they yield unique, unbiased estimates of the scale when the nominal distribution is Gaussian and there is no contaminant. This is a consequence of the monotonicity of $\mathsf{E}[\varphi_H(x; k)]$ for the chosen $k = 1, 2, 3$, and that $\delta$ was set to yield unbiased M-estimates. As a result of this, it can be shown that adaptive M-estimator which uses a linear combination of these bases also produces unique unbiased estimates of scale for a nominal Gaussian distribution with no contamination subject to the constraints (16).

## 5. CONCLUSION

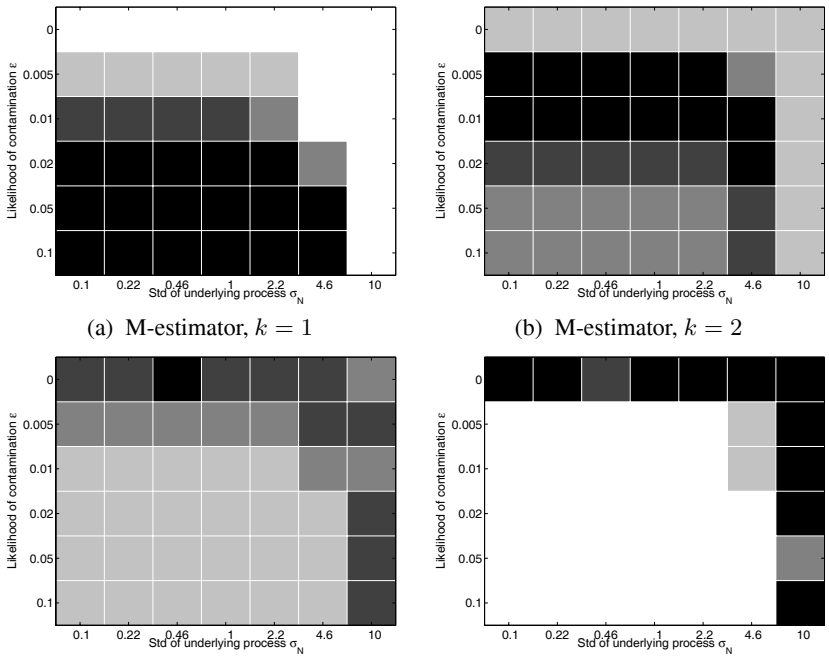
An M-estimator for scale was proposed which adaptively estimates the scale score function. This estimator can be included in robust parameter estimation problems, such as the signal in additive noise scenario, where the scale is a nuisance parameter.

A simulation study was carried out which compared the performance of the proposed adaptive M-estimator with that of static M-estimators with fixed score functions. In the study the adaptive scheme achieved good performance (lower MSE) across a wider range of the parameter space than the static M-estimator. This suggests that this estimator is more robust than the static M-estimator.

## 6. APPENDIX

The constraints on the bases imposed by conditions C1 and C2 are more clearly interpreted by considering their asymptotic behaviour. First, assume that the tails of the noise density decay asymptotically at an algebraic rate,

$$\lim_{x \to \pm\infty} f_X(x) = c|x|^{-\alpha-1}, \tag{17}$$




(a) M-estimator, $k = 1$ (b) M-estimator, $k = 2$

(c) M-estimator, $k = 3$ (d) Sample std, $k = \infty$

**Fig. 3**. Relative performance of the existing static M-estimators.

| | $\varepsilon$ ($\times 10^{-2}$) | | | | | |
|---|---|---|---|---|---|---|
| Method | 0 | 0.5 | 1 | 2 | 5 | 10 |
| Adaptive M-estimator | 0.5 | 0.7 | 1.3 | 2.7 | 10.7 | 41.9 |
| M-estimator, $k = 1$ | 1.1 | 1.0 | 1.1 | 1.7 | 4.6 | 16.7 |
| M-estimator, $k = 2$ | 0.5 | 0.7 | 0.9 | 2.2 | 11.5 | 58.9 |
| M-estimator, $k = 3$ | 0.5 | 0.8 | 1.9 | 6.9 | 54.4 | 460.5 |
| Sample std, $k = \infty$ | 0.5 | 68.4 | 198.3 | 547.9 | 2019.7 | 5277.5 |

**Table 2**. MSE ($\times 10^{-3}$) for $\sigma_N = 1$.

where $c$ is a constant and $\alpha > 0$ is the rate of decay, in the case of symmetric alpha stable distributions, $0 < \alpha \leq 2$.

Second, assume that the basis function $g(x)$ decays asymptotically at an algebraic rate,

$$\lim_{x \to \pm\infty} g(x) = b|x|^{-\beta}, \tag{18}$$

where $b$ is some constant and $\beta$ is the rate of decay.

Given that $g(x)$ and $f_X(x)$ are bounded over $-\infty < x < \infty$, or at least that

$$\int_{-b_1}^{b_2} g(x) f_X(x)\, dx < \infty, \qquad \infty < b_1, b_2 < \infty \tag{19}$$

such that for $x \in \{\{x < -b_1\} \cap \{x > b_2\}\}$ the asymptotic algebraic decay is accurate, the following holds,

*Conditions C1 and C2 are satisfied if $\beta \geq 0$, that is, if $|g(x)|$ is asymptotically non-increasing.*

Note that $\beta \geq 0$ also satisfies (13). If the noise density or basis function decays at a non-algebraic but faster rate, such as exponentially, then this condition is again satisfied.

## 7. REFERENCES


[1] S. Visuri, H. Oja, and V. Koivunen, "Subspace-based direction-of-arrival estimation using nonparametric statistics," *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 2060–73, September 2001.

[2] R. Kozick and B. Sadler, "Maximum likelihood array processing in non-Gaussian noise with Gaussian mixtures," *IEEE Transactions on Signal Processing*, vol. 48, no. 12, pp. 3520–35, December 2000.

[3] R. Brcich and A. Zoubir, "Robust estimation with parametric score function estimation," in *ICASSP*, Orlando, USA, May 2002, vol. 2, pp. 1149–52.

[4] C. Brown, R. Brcich, and A. Taleb, "Suboptimal robust estimation using rank score functions," in *ICASSP*, Hong Kong, April 2003, vol. 4, pp. 753–6.

[5] P. Huber, *Robust Statistics*, John Wiley, 1981.

[6] A. Taleb, R. Brcich, and M. Green, "Suboptimal robust estimation for signal plus noise models," in *Asilomar*, Pacific Grove, USA, October 2000, vol. 2, pp. 837–41.