

A NEW APPROACH TO ROBUST LINEAR REGRESSION

Jean Jacques Fuchs

*Irisa-Universit de Rennes, Campus de Beaulieu,
35042 Rennes Cedex, France.
e-mail : fuchs@irisa.fr*

Abstract: The least-squares estimation technique which minimizes the sum of the squared residuals is highly sensitive to outliers. One standard remedy is to minimize other functions of the residuals that downweight large residuals. A new approach to this technique is proposed. Additional unknowns that allow to detect and model the outliers are introduced. The robust estimates are obtained using a standard quadratic programming routine. Extensions to the total least-squares model in the presence of outliers are proposed.

Keywords: robust estimation, least-squares, quadratic programming

1. INTRODUCTION

The estimation of the linear regression model parameters is in general achieved by the method of least-squares. This approach is optimal if the additive noise is normal. It is however extremely sensitive to outliers i.e. to observations in which the response is abnormally large. This may happen in case of a failure of the sensor or in presence of spiky noise not accounted for in the model. To circumvent these difficulties one can model the outliers and develop corresponding optimal estimates but these are again in general highly sensitive to even small deviations of the real model from the assumed one.

The objective of robust procedures is to reject or downweight observations that are suspect relative to the model. Huber (Huber, 1973) was probably the first to propose a rational approach to robust estimation. Another basic reference is Poljak (Poljak, 1980). In this paper, a new approach to the problem is presented. Additional variables that allow to model the outliers, are introduced in the linear regression model and a new convex criterion is proposed. Its minimum can be obtained using a standard quadratic programming routine.

At the optimum, the basic variables are robust estimates of the regression parameters while the additional variables locate and estimate the amplitudes of the outliers. The so-obtained basic variables are shown to be identical to those known as Huber M-estimates. This new formulation is attractive because it allows for extensions not readily handled by other techniques. Extension to the colored standard noise case is considered in (Fuchs, 1999). Here an application to the total least-squares model is proposed. One seeks a solution to a over-determined system of linear equations with all the data matrices perturbed by noise and outliers.

In section 2, the robust linear regression model is presented together with the Iteratively Reweighted Least Squares approach which is currently the most used robust regression approach. The new approach is presented in section 3 where its equivalence to the M-estimator with Huber's function is established. Two optimization procedures are proposed. In section 4, the robust total least-squares model is considered and its extension in the presence of outliers and an optimization algorithm introduced. An example illustrates this section.

2. ROBUST LINEAR REGRESSION

2.1 The model

Consider the following linear regression model :

$$Y = AX + N \quad (1)$$

where Y is an n -dimensional vector of observations, X the p -dimensional vector of parameters to be estimated and N the additive noise vector. A is the (n, p) dimensional data matrix of full column rank. One should think of N as being composed of typical noise following the assumed model, e.g. $N(0, \sigma^2 I)$, and unexpected impulsive noise. If a_i^T designates the i -th row of A and $r_i = y_i - a_i^T X$ the i -th residual, then the standard least-squares approach minimizes $\sum r_i^2 = \|Y - AX\|_2^2$ and the optimum is given by $X = A^+ Y$ where $A^+ = (A^T A)^{-1} A^T$ is the pseudo-inverse of A .

For W an order n diagonal matrix of positive weights w_i , the weighted least-squares solution that minimizes $\sum w_i r_i^2 = (Y - AX)^T W (Y - AX)$ is $X_W = (A^T W A)^{-1} A^T W Y$. Note that X_W is optimal when N is a normal random vector with covariance matrix $\sigma^2 W^{-1}$.

The problem with these techniques is that even the presence of a small number of (large) outliers can lead to completely wrong estimates. A simple mean to detect wrong data is to start with a least-squares fit, to compute the residuals, to reject all observations whose residuals exceed a fixed threshold and to iterate this procedure until all residuals are below the threshold. More subtle techniques are easy to imagine along these lines. Of course the results always depend strongly on the value of the threshold.

Another class of robust estimators, the so-called M-estimators (Huber, 1981) propose to replace in the least-squares criterion the squares of the residuals by more general functions of the residuals, i.e., to replace $\sum r_i^2$ by $\sum f(r_i)$. Many functions have been proposed, the most frequently used is probably Huber's function :

$$\begin{aligned} f(r) &= r^2/2 & |r| \leq h \\ &= h|r| - h^2/2 & |r| > h \end{aligned} \quad (2)$$

where h is a threshold yet to be fixed. It is easy to check that this is a convex continuous function that is linear for r large and is a parabola for r small. The observations leading to large residuals are given less relative weight and thus less influence than the smaller residuals. For Huber's function (2), h allows to define the limit between small and large residuals. The performance of the M-estimator depends strongly on the way this threshold is fixed. It has clearly to be linked to the

assumed probability density function of the additive typical noise. For gaussian noise it is generally fixed at $h = 1.345 \sigma$. It remains to estimate σ and this estimate should itself be robust. A typical value for gaussian noise is $\hat{\sigma} = 1.483$ median $|r_i|$ (Holland, 1977), (Rousseeuw, 1987). These constants are fixed in order to achieve a given level of performance in case no outliers are present and the observations are only affected by additive gaussian noise.

2.2 Iteratively reweighted least-squares

Minimizing $\sum f(r_i)$ for a function like (2) has to be done iteratively and one possibility is to use the iteratively reweighted least-squares (IRLS) approach that we present for completeness. The minimum of $\sum f(r_i)$ is achieved at the point where its gradient vanishes. Let $\psi(r)$, known as the influence function, denote the derivative of $f(r)$ and ∇r the gradient of r with respect to X , the gradient of $\sum f(r_i)$ is then :

$$\begin{aligned} \sum \psi(r_i) \nabla r_i &= \sum \frac{\psi(r_i)}{r_i} r_i \nabla r_i \\ &= \frac{1}{2} \sum \frac{\psi(r_i)}{r_i} \nabla r_i^2 = 0 \end{aligned}$$

The last expression in this relation, can be seen as the gradient of $\frac{1}{2} \sum w_i r_i^2$ if one omits the dependance of the weights $w_i = \frac{\psi(r_i)}{r_i}$ upon r_i . An IRLS algorithm minimizing $\sum f(r_i)$ will, at step k , minimize $\sum w(r_i^{(k-1)}) r_i^2$. For $f(r)$ as in (2), one has $w(r_i) = h / \max(h, |r_i|)$. The IRLS algorithm starts with $W = I$ and iterates :

$$\begin{aligned} X &= (A^T W A)^{-1} A^T W Y \\ r_i &= y_i - a_i^T X \\ h &= 1.345 \times 1.483 \text{ median } |r_i| \\ W &= \text{diag}\left\{ \frac{h}{\max(h, |r_i|)} \right\} \end{aligned}$$

until some stopping criterion is satisfied.

The general idea is to start with the standard least-squares fit and to iteratively reweight the residuals, larger residuals receiving relatively less weight than observations with smaller residuals. For $f(r)$ as in (2) one can show that the algorithm converges.

3. THE NEW PROCEDURE

3.1 The criterion

Adding n new variables U in the standard regression model (1) one has :

$$Y = AX + U + N = BZ + N \quad (3)$$

where $B=[A \ I]$ and $Z^T = [X^T \ U^T]$. The aim is to use the unknowns U to model the outliers. To do so, we associate with the new model (3) the following convex optimization problem :

$$\min_{X,U} \|AX + U - Y\|_2^2 + \lambda \|U\|_1 \quad (4)$$

where $\lambda \in \mathbb{R}^+$ has yet to be fixed. It will play the same role as the threshold in the standard robust regression approach. $\|U\|_1 = \sum |u_i|$ stands for the ℓ_1 norm of the vector U . We will show that for a well chosen λ , the optimum is attained at a point $Z^* = (X^*, U^*)$ where X^* the robust estimate of interest is identical to the M-estimate with Huber's function and U^* is an auxilliary vector with very few non-zero components that *identify* the observations corrupted by outliers.

The criterion (Fuchs, 1997) is a combination of the standard least-squares criterion and a regularization term on the ℓ_1 -norm of the U -variables. It is convex but not continuously differentiable. It can be transformed into a quadratic program and its global minimum obtained using standard routines available in any scientific program library.

3.2 Optimality conditions

Let us first show that the X^* minimizing (4) and the optimum of :

$$\min_X \sum f(r_i) \quad \text{with} \quad r_i = y_i - a_i X \quad (5)$$

with the function f as in (2), are identical for $\lambda = 2h$.

Indeed the minimization with respect to (w.r.t.) U in (4) can be done independtly of the one w.r.t. X . Let us rewritte (4) as :

$$\sum_i (u_i - r_i)^2 + \lambda |u_i| \quad (6)$$

The minimum w.r.t. u_i is obtained for $u_i^* = r_i - (\lambda/2) \text{ sign } r_i^*$ if $|r_i| > \lambda/2$ and for $u_i^* = 0$ otherwise. With obvious notations, this can be summarized as :

$$U = \max(0, r - \frac{\lambda}{2}) + \min(0, r + \frac{\lambda}{2})$$

Substituting for these values in (6) leads to

$$\min_X \sum_i \{r_i^2\} \mathbb{I}_{|r_i| \leq \lambda/2} + \{\lambda |r_i| - (\lambda/2)^2\} \mathbb{I}_{|r_i| > \lambda/2} \quad (7)$$

where \mathbb{I} designates the indicator function. For $\lambda = 2h$, this is precisely $\sum 2f(r_i)$ establishing the equivalence.

As we shall see, this is a quite fruitful point of view since it allows for new extensions and already indicates that to obtain these specific M-estimates

the IRLS algorithm can be replaced by a quadratic programing routine.

Before we present some potential extensions let us develop the optimality conditions (Fuchs, 1998). Since the criterion (4) is convex a necessary and sufficient condition for $Z^* = (X^*, U^*)$ to be the optimum is that the vector 0 be a subgradient of the criterion at Z^* . Since (4) is non-smooth at zero only, we distinguish the non-zero components of Z^* , denoted \bar{Z}^* , from the remaining components. For the components in \bar{Z}^* the subgradient reduces to the gradient and has to vanish.

\bar{Z}^* denotes the union of the components of X^* and the non-zero components in U^* , themself denoted \bar{U}^* . In a similar way, we denote \bar{B} the matrix formed with the columns in B associated with the non-zero components in Z^* so that $BZ^* = \bar{B}\bar{Z}^*$. Equating the gradient at \bar{Z}^* with zero, one gets :

$$\bar{Z}^* = \bar{B}^+ Y - \frac{\lambda}{2} (\bar{B}^T \bar{B})^{-1} \underline{\text{sign}} (\bar{Z}^*) \quad (8)$$

where \bar{B}^+ is the pseudo-inverse of \bar{B} and $\underline{\text{sign}} (\bar{Z}^*) = [0^T \text{sign} (\bar{U}^{*T})]^T$, with $\text{sign} (u_i) = -1, +1$ for u_i respectively $< 0, \geq 0$. This is not an explicit expression of the optimum \bar{Z}^* of the criterion since \bar{Z}^* appears on both sides.

For the other components in Z^* , i.e., the zero components in U^* , the vector 0 must be a subgradient of the criterion (4), this condition becomes :

$$|y_j - b_j^T Z^*| < \frac{\lambda}{2} \quad \forall j \ni u_j^* = 0 \quad (9)$$

These two relations (8, 9) fully define the optimum Z^* and though they are not explicit (the optimum can only be obtained in an iterative way) they can be helpful in checking whether or not the point one would like to be the solution is indeed the optimum. Taking a close look at (8) it appears that it is sufficient to know (or guess) the indices and the signs of the non-zero components in U^* to completely define the optimum. Since these non-zero components actually designate the outliers this information is of course unknown *a priori*.

3.3 Optimization algorithms

In the sequel we rewritte (4) as

$$\min_{X,U} \frac{1}{2} \|AX + U - Y\|_2^2 + h \|U\|_1 \quad (10)$$

to make it strictly equivalent to the criterion leading to Huber M-estimates (5). This is an unconstrained non-smooth optimization problem that can be converted into a quadratic program (Luenberger, 1973). To do so one introduces new variables $u_i^+ = \max(u_i, 0)$, $u_i^- = \max(-u_i, 0)$ and replaces u_i by $u_i^+ - u_i^-$ and $|u_i|$ by $u_i^+ + u_i^-$ and further constrains these new variables to

be greater or equal to zero. The problem then becomes :

$$\min \frac{1}{2} \|AX - Y + U^+ - U^-\|_2^2 + h \mathbf{1}^T (U^+ + U^-)$$

subject to $U^+ \geq 0, U^- \geq 0$

where $\mathbf{1}^T$ is a row-vector of ones. This is now a quadratic program whose solution is easily and quickly obtained, even for large number of unknowns, using standard programs available in any scientific program library.

Yet another way to obtain the optimum of (10) that is easily derived from this new formulation consists in minimizing (10) alternatively w.r.t. X and U , starting from $U^0 = 0$.

◇ For U^{k-1} fixed, the minimum in X is attained at $X^k = A^+(Y - U^{k-1})$.

◇ For X^k fixed, defining the residuals $r^k = Y - AX^k$ the minimum of (10) in U is attained at $U^k = \max(r^k - h, 0) + \min(r^k + h, 0)$.

Using the global convergence theorem (Luenberger, 1973), one can establish that this algorithm converges to a point where the above given necessary and sufficient conditions for an optimum are satisfied. Indeed since ◇ the estimates remain bounded, ◇ the criterion decreases strictly as long as the necessary and sufficient conditions for an optimum are not satisfied and ◇ both mappings are one-to-one and continuous, this algorithm converges to a global minimum.

4. ROBUST TOTAL LEAST-SQUARES

4.1 The total least-squares model

When applying least-squares to the basic linear regression model (1) $Y = AX + N$ one implicitly assumes that errors are only present in the observations Y since one minimizes the ℓ_2 norm of the residuals $r = Y - AX$. If errors are present in A as well, one should resort to the total least-squares model (Golub, 1980), a very specific subclass of the more general static error-in-variables models (Anderson, 1996).

Reformulating the basic least-squares problem as :

$$\min \|r\|_2^2 \quad \text{s.t.} \quad AX = Y + r$$

leads quite naturally to the following total least-squares (TLS) formulation :

$$\min \|[R \ r]\|_F^2 \quad \text{s.t.} \quad (A + R)X = Y + r \quad (11)$$

where $\|C\|_F$ stands for the Frobenius norm $\|C\|_F^2 = \sum_{i,j} c_{i,j}^2$. As r allows for errors in Y , R allows for errors in A . This is a highly non linear optimization problem in the unknowns R, r and X . The solution of this problem is now well

known, it is linked to the singular value decomposition of $C = [A \ Y]$. There are many different ways to establish this link. We present three of them to highlight the different aspects of the problem. .

Probably the most powerful one, relies on matrix approximation results. Defining the augmented matrices $C = [A \ Y]$ and $\Gamma = [R \ r]$, one reformulates (11) as follows :

$$\min \|\Gamma\|_F^2 \quad \text{s.t.} \quad \text{rank}(C + \Gamma) = p \quad (12)$$

i.e one seeks a perturbation $\Gamma = ((\gamma_{i,j}))$ of lowest Frobenius norm that makes the rank of $C + \Gamma$ drop from $(p+1)$ to p . The optimum X of (11) is such that $(C + \Gamma)Z = 0$ with $Z = [X \ -1]^T$. For the Frobenius norm the solution Γ^* to this problem is generically unique and of rank one (Stewart, 1990). It is built from the singular decomposition triplet associated with the lowest singular value of C and the optimum X^* of the TLS problem is obtained by normalizing to -1 the last component of the associated right singular vector.

If one is willing to think of n , the number of rows in A , as being variable and going to infinity one can adopt a statistical setting (Gleser, 1981). One combines the observations in a sequence of $(p+1)$ -dimensional vectors $V_i^T = [a_i^T \ y_i]$. These vectors are modeled as $V_i = W_i + E_i$ where the vectors E_i are independent and identically distributed random vectors with mean zero and covariance matrix $\sigma^2 I$ and the vectors W_i are modeled as $W_i^T = [\nu_i^T \ X^T \nu_i]$. If the random error vectors E_i are further assumed to be gaussian, maximizing the likelihood of the observations V_i with respect to X and the ν_i 's amounts to solve the TLS problem (11). After some manipulations, eliminating the ν_i vectors, the criterion can be rewritten as :

$$\min_X \frac{\|AX - Y\|_2^2}{1 + X^T X} = \min_Z \frac{\|CZ\|_2^2}{\|Z\|_2^2}$$

with C and Z as above. From the last formulation and the variational characterization of singular values it follows again that the solution is linked to the *smallest* right singular vector of C .

Yet another fruitful approach amounts to work with the covariance matrix estimates and to adopt a signal subspace approach. Using the notations of the previous paragraph, one assumes that the empirical *covariance* matrix of ν_i i.e. $\frac{1}{n} \sum \nu_i \nu_i^T$ is full rank, the empirical covariance matrix of W_i then has a single zero eigenvalue associated with eigenvector $[X^T \ -1]$. The covariance matrix of V_i being the sum of the covariance matrix of W_i and $\sigma^2 I$, is full rank with minimal eigenvalue an estimate of σ^2 and associated eigenvector an estimate of $[X^T \ -1]$. But the *smallest* eigenvector of $\frac{1}{n} \sum V_i V_i^T$ is also the *smallest* right singular vector of $C = [A \ Y]$.

Note that in the more general static errors-in-variables model the noise covariance matrix is assumed to be any positive definite diagonal matrix and not just $\sigma^2 I$. In the present situation, this amounts to say that the noises on the different columns of C have different variances. This leads to indeterminacy as follows easily from the signal subspace approach.

4.2 The robust total least-squares approach

If one now considers that beside the standard noise or perturbations, some outliers may be present, a straightforward way to proceed is to propose to replace in (11) or (12) the Frobenius norm of the errors i.e. $\sum \gamma_{i,j}^2$ by $\sum f(\gamma_{i,j})$ leading to the following problem :

$$\min \sum_{i,j} f(\gamma_{i,j}) \quad \text{s.t.} \quad \text{rank}(C + \Gamma) = p$$

Even if the function $f(\cdot)$ is chosen to be convex, as is the case in (2), this is no longer a convex optimization problem and nothing guarantees that it admits a unique optimum. The matrix approximation result (Stewart, 1990) no longer applies since $\sum f(\gamma_{i,j})$ is not a unitarily invariant matrix norm. It is no even trivial to get a local optimum. However proceeding as above when we transformed (5) into (6), we rewrite this problem as :

$$\begin{aligned} \min_{\Gamma, V} \frac{1}{2} \|\Gamma - V\|_F^2 + h |V|_1 \quad (13) \\ \text{s.t.} \quad \text{rank}(C + \Gamma) = p \end{aligned}$$

With V a matrix of variables intended to model the outliers and $|V|_1 = \sum_{i,j} |v_{i,j}|$. For $f(\cdot)$ the Huber function (2), these two formulations are equivalent as we have established it in (6, 7). In the new formulation, it is possible to develop an algorithm converging to a local optimum.

4.3 Optimization algorithm

In this section we develop an algorithm that converges to a minimum of (13). Let us first define $t_{\min}(A)$ that associates to a matrix A the rank-one matrix built upon its *smallest* singular triplet : $t_{\min}(A) = \sigma_{\min}(A) u_{\min}(A) v_{\min}^T(A)$.

We propose to minimize (13) alternatively with respect to Γ and V . Remember that V will have very few non-zero components since it models the outliers. We start with $V^0 = 0$.

◇ For V^{k-1} fixed, we seek the minimum in Γ of :

$$\begin{aligned} \min_{\Gamma} \frac{1}{2} \|\Gamma - V^{k-1}\|_F^2 + h |V^{k-1}|_1 \\ \text{s.t.} \quad \text{rank}(C + \Gamma) = p \end{aligned}$$

defining $\Delta = \Gamma - V^{k-1}$, this becomes :

$$\begin{aligned} \min_{\Delta} \frac{1}{2} \|\Delta\|_F^2 + h |V^{k-1}|_1 \\ \text{s.t.} \quad \text{rank}(C + V^{k-1} + \Delta) = p \end{aligned}$$

This is the basic TLS problem (see (12)) whose optimum is attained at $\Delta^k = -t_{\min}(C + V^{k-1})$. We thus have $\Gamma^k = V^{k-1} - t_{\min}(C + V^{k-1})$.

◇ For Γ^k fixed, one seeks the optimum in V of :

$$\begin{aligned} \min_V \frac{1}{2} \|\Gamma^k - V\|_F^2 + h |V|_1 \\ \text{s.t.} \quad \text{rank}(C + \Gamma^k) = p \end{aligned}$$

the constraints is inactive and the optimum is attained at $V^k = \max(0, \Gamma^k - h) + \min(0, \Gamma^k + h)$. The criterion decreases at each step and the algorithm converges to a stationary point where $\|\Delta^*\|_{\infty} = h$ and the non-zero components of V^* and those in Δ^* with maximal absolute values h are at the same locations. $\Gamma^* = V^* + \Delta^*$ the global perturbation that makes the rank of $C + \Gamma^*$ drop to p is no longer of rank one but X^* , the solution to the robust TLS problem remains uniquely defined by its relation to the *smallest* right singular vector of $C + V^*$.

4.4 Simulation results

Let us present some results for the very simplest TLS model with $p = 1$: a line in a plane. The observed data consist of n pairs of real numbers (a_i, y_i) , $i = 1, \dots, n$. The appropriate model for these data is given by :

$$\begin{aligned} a_i &= \nu_i + e_i \\ y_i &= x\nu_i + n_i \end{aligned}$$

where e_i and n_i represent *noise* distorting the exact linear relationship. There is only one unknown the slope x of the line. The noise e_i is gaussian, zero mean with variance σ_e^2 denoted $N(0, \sigma_e^2)$, while the noise n_i is a mixture of two gaussians $(1 - \epsilon)N(0, \sigma_e^2) + \epsilon N(0, K \sigma_e^2)$. The rate $\epsilon \in (0, 1)$ defines the degree of contamination by outliers of the noise acting on the y -observations. For ϵ equal to zero there are no outliers but due to the presence of the noise e_i on the ν -variables the basic least-squares solution would be biased and one has to resort to a TLS solution. For non-zero ϵ the TLS solution also becomes biased because it assumes that both noises have the same variance. It is worthwhile to compare with the errors-in-variables model (Rissanen, 1988). One should indeed realize that if both noises e_i and n_i have the same ϵ -contaminated gaussian density then the TLS estimates is unbiased as follows easily from the subspace interpretation of the TLS solution.

We have performed the simulations with the following values : $n = 32$, $\sigma_e^2 = .01$, $\epsilon = .25$, $K = 25$. The n values of the abscissae ν_i are random uniformly distributed in $(0, 1)$ and the slope x is also drawn from this density. We implemented the algorithm presented in section 4.3. The result of just one simulation are presented in Figure 1.

The true slope was $x = .3582$. There are three lines displayed : one with a wrong slope close to one which is the result given by the basic TLS approach in the presence of the outliers. The two others lines are those obtained by the proposed algorithm (in the presence of the outliers) and the the basic TLS in the absence of outliers (same noise seed but $\epsilon = 0$). The 32 data pairs are represented by the "o"s, there are are 8 outliers that are further marked with a "+" sign in the "o"s. There are also 8 squares that represent the outliers free pairs (to each "+" corresponds a square with the same abscissae). Clearly for this noise realization the estimate of the slope obtained by the proposed method is close to the exact one and appears to be quite insensitive to the presence of the outliers.

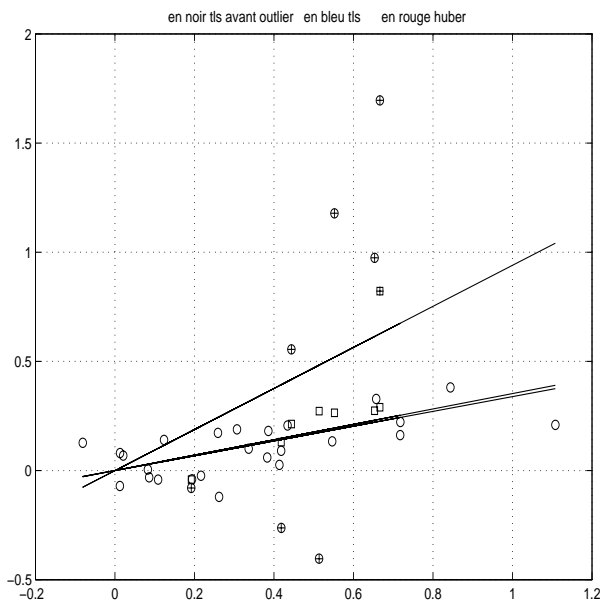


Fig. 1. Illustration of the results for a total least-squares model with outliers.

5. CONCLUSION

Introducing new variables in the standard regression model (that allow to handle the outliers) we have proposed a criterion that leads to robust regression estimates that are strictly identical to the M-estimates with Huber's function.

The criterion can be optimized using standard quadratic programming routines. This new approach allows for extensions not readily handled

by other techniques. We have considered here its application to the robust linear total least-squares model.

6. REFERENCES

- B.D.O. Anderson, M. Deistler and W. Scherrer. (1996) Solution set properties for static errors-in-variables problems. *Automatica* vol. 32, 7, 1031-1035.
- J.J. Fuchs. (1997) Multipath time-delay estimation. *IEEE Proc. ICASSP*, I, pp. 527-530, Munich. to appear in *IEEE-T-SP* Jan. 1999.
- J.J. Fuchs. (1998) Detection and estimation of superimposed signals. *IEEE Proc. ICASSP*, III, pp. 1649-1652, Seattle.
- J.J. Fuchs. (1999) An inverse problem approach to robust regression. *IEEE Proc. ICASSP* Phoenix.
- L.J. Gleser. (1981) Estimation in a multivariate Errors-in-variables regression model : large sample results. *Annals of Statistics.*, vol. 9, 1, 24-44.
- G.H. Golub and C.F. Van Loan. (1980) An analysis of the total least squares problem. *SIAM J. Numer. Anal.* vol. 17, 6, 883-893.
- P.W. Holland and R.E. Welsh. (1977) Robust regression using iteratively reweighted least squares. *Comm. Stat.* A6, 813-828.
- P.J. Huber. (1973) Robust regression. *Annals of Statistics.*, vol. 1, 799-821.
- P.J. Huber. (1981) Robust Statistics. *John Wiley and sons.*, New York.
- D. G. Luenberger. (1973) Introduction to linear and nonlinear programming. *Addison Wesley*
- B.T. Poljak and Y. Z. Tsyarkin. (1980) Robust identification. *Automatica* vol. 16, 1, 53-63.
- J. Rissanen. (1988) Estimation of errors-in-variables models. *IEEE Proc. 27th. CDC* pp. 1828-1830, Austin
- P.J. Rousseeuw and A.M. Leroy. (1987) Robust regression and outlier detection. *John Wiley and sons.*, New York.
- G.W. Stewart and J. Sun. (1990) Matrix perturbation theory. *Academic Press*