AN INVERSE PROBLEM APPROACH TO ROBUST REGRESSION

Jean-Jacques FUCHS

IRISA/Université de Rennes I
Campus de Beaulieu - 35042 Rennes Cedex - France
fuchs@irisa.fr

ABSTRACT

When recording data, large errors may occur occasionally. The corresponding abnormal data points, called outliers, can have drastic effects on the estimates. There are several ways to cope with outliers • detect and delete or adjust the erroneous data, • use a modified cost function. We propose a new approach that allows, by introducing additional variables, to model the outliers and to detect their presence. In the standard linear regression model this leads to a linear inverse problem that, associated with a criterion that ensures sparseness, is solved by a quadratic programming algorithm. The new approach (model + criterion) allows for extensions that cannot be handled by the usual robust regression methods.

1. INTRODUCTION

In linear regression, when fitting a linear model to noisy data, the presence of outliers i.e. observations which depart from the basic assumptions, can have dramatic effects on the quality of the estimates. The classical least-squares procedures, for instance, becomes unreliable in the presence of outliers in the data [2].

Mainly two differents approaches have been proposed to solve this problem [1]-[4]. The most well-known relies on identifying the abnormal observations in order to remove or correct them. This is generally achieved in an iterative way by removing one observation at a time but is computationally intensive and hazardous when several outliers have to be diagnosed. The other approach is robust estimation, it amounts to develop schemes that are less vulnerable to outliers. In the first approach one tries to detect and delete the abnormal observations in order to use a standard estimation scheme on the remaining data, in the second, one fits a robust estimate first and later decides upon the outliers by looking for instance at the residues.

The approach we present achieves both tasks (obtaining a robust estimate and detecting the outliers) simultaneously. As far as the basic parameters are concerned, it turns out to be analytically equivalent to the most well-known robust estimation scheme (the M-estimator with Huber's influence function [1]-[4]). It is nevertheless of interest because it sheds new light on this scheme and allows for extension that are not possible in the initial formulation.

2. ROBUST LINEAR REGRESSION

The objective is to fit a linear model to noisy observations :

$$Y = AX + N \tag{1}$$

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 which we assume to be of full column rank. Let a_i^T designate the i-th row of A and r_i the i-th residual : $r_i = y_i - a_i^T X$. The standard least squares approach minimizes $\sum r_i^2 = \|Y - AX\|_2^2$ and the minimum is attained at $X_{ls} = (A^T A)^{-1} A^T Y$. If the noise or model error N has a normal density with covariance matrix Σ proportional to identity i.e. $N \in N(0, \sigma^2 I)$, this solution corresponds to the maximum likelihood estimate.

If even a very limited number of observations do not follow the assumed gaussian density, the least squares estimate can be extremely far away from the true value [2]. This led Huber (see [3] for details) to seek the "worst" possible density among those of the form $p = (1-\epsilon)p_o + \epsilon p_1$ where $\epsilon \in (0,1)$, $p_o = N(0,1)$ and p_1 is an arbitrary density, i.e. the so-called ϵ-contaminated Gaussian densities. He then proposed to use the associated maximum likelihood estimator as a robust estimator. "Worst" meaning that one seeks the density p that yields the least possible information in each individual residual $y_i - a_i^T X$. One looks for the density p within the class that minimizes the Fisher information matrix :

$$\min_{p_1} \int_{-\infty}^{\infty} \frac{p'(z)^2}{p(z)} dz$$

0-7803-5041-3/99 $10.00 © 1999 IEEE

The optimum p^* is given by :

$$\begin{aligned} p^*(z) &= \frac{1-\epsilon}{\sqrt{2\pi}} e^{-z^2/2} && |z| \leq h \\ &= \frac{1-\epsilon}{\sqrt{2\pi}} e^{h^2/2-h|z|} && |z| > h \end{aligned} \quad (2)$$

where the constant h depends upon ϵ (see [3][4] for details) and varies from 0 to ∞ as ϵ goes from 1 to 0. This is a quite rational approach to robust estimation and one expects that the maximum likelihood (ML) method applied to this density leads to a robust regression estimate. We detail it below.

3. THE M-ESTIMATORS

Maximizing the log-likelihood of p^* in (2) leads to the minimization of :

$$\min_X \sum_1^n f(r_i) \quad (3)$$

with the function f -known as Huber's function- defined by :

$$\begin{aligned} f(r) &= r^2/2 && |r| \leq h \\ &= h|r| - h^2/2 && |r| > h \end{aligned} \quad (4)$$

It is a parabola in the vicinity of zero and increases linearly for $|r| > h$. The interpretation is that observations with large residuals are given less weight and hence less influence than those with small residuals, h being the threshold.

Since the development has been done under the assumption that the standard noise has unit variance, to adapt it to real data, one has either to divide the residuals or multiply the threshold h by a *robust* estimate $\hat{\sigma}$ of the standard deviation of the residuals.

Estimates obtained minimizing $\sum_1^n f(r_i/\hat{\sigma})$ for this function (4) or other even functions are known as M-estimates in the statistical litterature. Many other functions have been proposed. The minimization (3) is generally achieved using a so-called iteratively reweighted least squares scheme (IRLS) [5], [9]. Let us now present and justify such a scheme. Since f in (4) is convex, the minimum in (3) is found by making the gradient vanish. Denoting ∇r_i the gradient of the i-th residual r_i with respect to X, this leads to :

$$\sum_i \psi(r_i)\ \nabla r_i = 0$$

where $\psi(r)$ known as the influence function, is the derivative of $f(r)$. This set of p non-linear equations can be rewritten as :

$$\sum \frac{\psi(r_i)}{r_i} r_i\ \nabla r_i = (1/2)\ \sum \frac{\psi(r_i)}{r_i}\ \nabla r_i^2 = 0$$

where the last expression can be seen as the gradient of the weighted residuals : $\sum_I w_i r_i^2 = (Y - AX)^T D (Y - AX)$ with diagonal weighting matrix $D = \text{diag}(w_i) = \text{diag}\ (\frac{\psi(r_i)}{r_i})$. An IRLS algorithm consists then in initializing $D_0 = I$ and iterating :

$$\begin{aligned} X_1 &= (A^T D_0 A)^{-1} A^T D_0 Y \\ R_1 &= Y - AX_1 \\ h_1 &= 1.345 \times 1.483\ \text{med}_i\ |r_i - \text{med}_j\ r_j| \\ D_1 &= \text{diag}\{\frac{h_1}{\max(h_1, |r_i|)}\} \end{aligned} \quad (5)$$

until some stopping criterion is satisfied. Note that for f as in (4), $w_i = \frac{h}{\max(h,|r_i|)}$.

The idea is to start with the standard least squares fit and to iteratively reweight the residuals. The weights are functions of the residuals from the previous iteration such that observations with larger residuals receive relatively less weight than observations with small residuals. Abnormal points tend to receive less weight than typical points.

For completness we have included the *robust* reevaluation of the modified threshold h [5]. The factor 1.345 is the usual value of h for unit-variance gaussian noise, it is chosen to guaranty a given level of efficiency in the absence of outliers [1]. It is multiplied here by an estimate of the scale $\hat{\sigma} = 1.483$ med $|r_i - \text{med}\ r_j|$ where med denotes the median. This is an approximately unbiased estimate of the standard deviation of the residuals when the error model is gaussian [1].

For $f(r)$ as in (4) one can show that the algorithm converges though the convergence can be relatively slow. One can show that when M-estimates are ML estimates the associated IRLS algorithm is also an expectation-maximize (EM) algorithm and convergence results of these algorithms thus apply.

4. THE PROPOSED APPROACH

Let us rewritte the standard regression model (1) by adding n new variables U, one in each observation equation :

$$Y = AX + IU + N = BZ + N \quad (6)$$

where $B = [A\ I]$ and $Z^T = [X^T\ U^T]$. The objective is to use the unknowns U to model the outliers. Remember that these are occasional large measurement errors that can be caused by disturbances, conversion failures, etc. In any case they are extremely sparse which means that only very few components in U should be non-zero. Adding these new variables U, transforms the initially over-determined set of linear equations into an under-determined one. While (1) has (generically) no

solution, (6) has an infinite number of solutions and one now needs a criterion that selects among them a solution in which U is sparse. We will show that :

$$\min_{X,U} \|Y - AX - U\|_2^2 + \lambda \|U\|_1 \qquad (7)$$

is such a criterion. It is a combination of the standard least squares criterion and a regularization term on the ℓ_1-norm of the U-variables. This criterion is convex but not continuously differentiable.

The minimization (7) can be transformed into a quadratic program. Indeed, if 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 $u_i^+ + u_i^-$, this unconstrained non-smooth optimization problem is converted into a quadratic program with the X variables unconstrained and these new variables u_i^+ and u_i^- constrained to be greater or equal to zero [6]. Its unique solution is easily and quickly obtained, even for large number of unknowns, using standard programs available in any scientific program library.

The criterion (7) is similar to the one used in [7],[8]. The ℓ_1-norm regularization term is known to lead to sparseness and the weight or hyper-parameter λ allows to tune the sparseness in U. The criterion can be given a Bayesian interpretation with a Laplace prior for the U variables.

5. THE EQUIVALENCE OF BOTH APPROACHES

Let us now show that the two approaches : • minimizing (3) with Huber's function (4) and • minimizing (7) lead to the same X estimates.

Criterion (7) is separable, it can be rewritten :

$$\min_{X,U} \sum_i (r_i - u_i)^2 + \lambda |u_i| \quad \text{with} \quad r_i = y_i - a_i^T X$$

The minimum with respect to u_i only concernes the i-th term and is obtained either for $u_i^* = 0$ if $|r_i| \leq \lambda/2$ or at $u_i* = r_i - (\lambda/2)$ sign r_i^* if $|r_i| > \lambda/2$, replacing u_i by these optimal values, the criterion becomes :

$$\min_X \sum_i \{r_i^2\} \mathbb{1}_{|r_i| \leq \lambda/2} + \{\lambda |r_i| - (\lambda/2)^2\} \mathbb{1}_{|r_i| > \lambda/2}$$

where $\mathbb{1}$ designates the indicator function. This new formulation is equivalent to the minimization of $\sum_i f(r_i)$ with f the Huber function (4) provided $\lambda = 2h$. Minimizing (3) with function (4) is thus equivalent to :

$$\min_{X,U} \frac{1}{2} \|Y - AX - U\|_2^2 + h\|U\|_1 \qquad (8)$$

This is an interesting point for both approaches. It allows for new interpretations of the standard approach and extends its domain of applicability as we will argue below. From an computational point of view it shows that, for this specific M-estimate, the IRLS procedure can be replaced by a quadratic programming routine.

6. OPTIMALITY CONDITIONS

The optimality conditions for (8) allow for new interpretations of the robust estimation procedure. A necessary and sufficient condition for $Z^* = (X^*, U^*)$ to be the global optimum of (8) is that the vector 0 be a subgradient of the criterion at Z^*. Since (8) is non-smooth at zero only, we distinguish the non-zero components of Z^*, denoted $\bar{Z}^*$, from the zero components. For the components in $\bar{Z}^*$ the subgradient reduces to the gradient and has to vanish.

Let $\bar{Z}^*$ denote the union of the components of X^* and the non-zero components in U^*, themself noted $\bar{U}^*$. 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 of (8) at $\bar{Z}^*$ with zero, one gets :

$$\bar{Z}^* = \bar{B}^+ Y - h(\bar{B}^T \bar{B})^{-1} \underline{\text{sign}}\ (\bar{Z}^*) \qquad (9)$$

where $\bar{B}^+$ denotes the pseudo-inverse of $\bar{B}$ and $\underline{\text{sign}}\ (\bar{Z}^*) = [0^T \text{sign}\ (\bar{U}^{*T})]^T$, with sign $(u_i) = -1, +1$ for u_i respectively $< 0, > 0$. Note that 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 (8), this condition becomes :

$$|(Y - BZ^*)_j| < h \qquad \forall j \ \ni u_j^* = 0 \qquad (10)$$

where $(.)_j$ denotes the j-th component.

These two relations (9, 10) fully define the optimum Z^* and though they are not explicit (the optimum can only be obtained in an iterative way) they clearly indicate how the presence of outliers affects the least squares solution and the bias that is due to the presence of the threshold h. Taking a close look at (9) 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.

The non-zero components in U actually designate the outliers.

Further insight can be gained by considering the dual form of the optimization problem (8) which is :

$$\begin{aligned} \min_{X,U} \quad & \|AX + U\|_2^2 \\ \text{subject to :} \quad & A^T(Y - AX - U) = 0 \\ & \|Y - AX - U\|_\infty \leq h \end{aligned} \qquad (11)$$

It is obtained by first transforming (8) into a quadratic program and then taking the dual in the usual way [6]. Introducing the augmented residuals $\epsilon_i = r_i - u_i = y_i - a_i^T X - u_i$, the constraints in (12) say that at the optimum these residuals are bounded by h in absolute value. In fact (see (10)) they are strictly smaller than h for observations that are considered to be outlier free and equal to h otherwise.

7. COLORED NOISE EXTENSION

So far we have assumed that the different observations were independent or at least uncorrelated. In practice this might not be the case and quite often one is in the situation where the covariance matrix Σ of the noise N is known up to a multiplicative constant $\Sigma = \sigma^2 \bar{\Sigma}$. In standard least squares this is of no consequence and one whitenes the observations by pre-multiplying them by $\bar{\Sigma}^{-1/2}$. Such a whitening step is however unfeasible in the presence of outliers since it will distribute the effect of the outliers on all the data. Let us recall (6) the augmented regression model we introduced earlier :

$$Y = AX + U + N$$

where N the standard noise model has now a normal density $N(0, \sigma^2 \bar{\Sigma})$. Clearly whitening the observations transforms this model into :

$$\bar{Y} = \bar{A}X + \bar{\Sigma}^{-1/2} U + \bar{N}$$

whith $\bar{Y} = \bar{\Sigma}^{-1/2} Y$, $\bar{A} = \bar{\Sigma}^{-1/2} A$ and where $\bar{N}$ is again $N(0, \sigma^2 I)$. Standard robust regression estimation methods no longer apply since the effect of even a single outlier (one non-zero component in U) is now distributed over all the components in $\bar{\Sigma}^{-1/2}\, U$. On the other hand, the criterion (8) we introduced simply becomes :

$$\min_{X,U} \frac{1}{2} \|\bar{Y} - \bar{A}X - \bar{\Sigma}^{-1/2} U\|_2^2 + h\|U\|_1 \qquad (12)$$

and no difficulty occurs since the regularization term continues to penalize just U and makes it sparse. Due to limited space we do not present simulations that highlight this point.

8. CONCLUSIONS

Adopting an inverse problem approach to robust regression estimation, we have developped a new model and criterion that happens to lead to an optimum that is strictly equivalent to the M-estimator with Huber's function also known as the minimax M-estimator which is probably the most used robust estimator. This new approach presents several advantages. It allows to replace the iteratively reweighted least squares approach that is used to obtain M-estimates, requires good initial estimates and encounters sometimes convergence difficulties by a standard quadratic programming routine available in any standard scientific program library.

The new approach also allows for extensions not readily handled with other techniques. We have indicated here its extension to the colored noise case that cannot be handled by the standard approaches [10]. Further extensions include the robust subset selection problem [11] where the p dimensional initial model may be over-parametrized and a lower order model (using a subset of columns of A) could be more informative than the full model. Work in this direction is under progress.

9. REFERENCES

[1] P.J. Rousseeuw and A.M. Leroy. Robust regression and outlier detection. *John Wiley and sons.*, New York, 1987.

[2] R.R. Hocking Linear regression. *Encyclopedia of Statistical Sciences, John Wiley and sons 5.*, New York, 1985, pp.59-64 .

[3] P.J. Huber. Robust Statistics. *John Wiley and sons.*, New York, 1981.

[4] B.T. Poljak and Y. Z. Tsypkin. Robust identification. *Automatica* vol. 16, 1, 53-63, 1980.

[5] P.W. Holland and R.E. Welsh. Robust regression using iteratively reweighted least squares. *Comm. Stat.* A6, 813-828, 1977.

[6] D. G. Luenberger. *Introduction to linear and nonlinear programming.* Addison Wesley, 1973.

[7] J.J. Fuchs. Multipath time-delay estimation. *IEEE Proc. ICASSP*, I, pp. 527–530, Munich. An extended version is submitted to *IEEE-T-SP* Mar 1997.

[8] J.J. Fuchs. Detection and estimation of superimposed signals. *IEEE Proc. ICASSP* , III, pp. 1649-1652, Seattle, May 1998.

[9] S.A. Ruzinsky and E.T. Olsen. ℓ_1 and ℓ_∞ minimzation via a variant of Karmarkar's algorithm. *IEEE-T-ASSP* 37, 245–253, Feb. 1989.

[10] S.L. Portnoy. Robust estimation in dependent situations. *Annals of Stat.* 5, 22–43, 1977.

[11] E. Ronchetti and R.G. Staudte. A robust version of Mallow's C_p. *J. Amer. Statis. Assoc.* 89, 550–559, 1994.