A MODIFIED FORWARD SEARCH APPROACH APPLIED TO TIME SERIES ANALYSIS

M. Pesenti[a], M. Piras[a]

[a] Politecnico di Torino, DITAG, Land, 10129 Turin, Italy - (manuele.pesenti, marco.piras)@polito.it

KEY WORDS: Forward search, Time series analysis, Outlier detection, Discontinuities detection, Robust statistics, LMS

ABSTRACT

Studying the behavior of some slow phenomena, such as crustal or continental deformations, it is necessary to consider time series data, with a sufficient numerousness that depends on their characteristics. Time series analysis is a delicate step; in order to obtain a correct significance and validity it is necessary to use an internal coherent outlier free dataset. It is necessary to analyze the data in order to find any possible outliers and to verify inner coherence. Some interesting features about outlier removal, research of zero degree unknown discontinuities and their evaluations can be provided applying the forward search method. In this case, a modified version has been developed, starting from the more general FS technique. The method has been applied to artificial data which are created in order to know the exact entity of the introduced discontinuities with the aim of simulating a time series solution that originates from GPS permanent stations networks.

Simulated data were analyzed considering a period of 100 epochs and with a repeatability of 100. Experiments with different values of the ratio between the imposed jump in the series and its noise level were performed with the aim of firstly defining a percentage of positive results based on where the jumps are found and how good their entity is evaluated, and secondly in order to calculate the repeatability. The algorithms have been implemented in a automatic procedure, developed in R language. This work shows some of the obtained results and gives a statistical interpretation.

1. INTRODUCTION

1.1 Time series analysis in geodesy

Time series analysis is a useful procedure that can be used in any spatial geodesy technique and it can be applied both before and after raw data treatment with different aims. Datum definition is a first application in the geodetic field. Nowadays, it is considered 4-dimensional because the temporal coordinate is included in its definition. This is the direct consequence of the use of points external to the terrestrial surface and therefore are not located on the Earth and thanks to the high accuracy of the results. This condition is not compatible with the initial hypothesis where the fitting considers a static reference system. Nowadays, this hypothesis is a little truthful, because the deformations accumulated by the vertexes of the datum over these last few years have to be considered.

The main geodetic application of time series analysis is devoted to the analysis of deformations, where the aim is to measure the movements in the time of some points of interest on the Earth (landslide or crustal deformations). Another purpose of time series analysis is to estimate the accuracy of the results. For example, in the case of GPS measurements, the variance matrix of solution, which is estimated using the propagation law, is under-estimated. This derives from the initial hypothesis about the baseline being correct. It is possible to obtain a more credible variance value, considering a direct estimation of the residuals, using the repeatability concept with an appropriate time series analysis.

A complete time series analysis in the geodetic field is composed of three consecutive steps:

1. search for and detection of the discontinuities;

2. defining an adapted linear model devoted to modeling the movements of the vertexes, to remove the linear trend;

3. frequency analysis of the cyclic component.

The first and the second steps are important to define the third, because the object of the frequency analysis has to respect a fundamental and restrictive hypothesis: stationarity. In short, some statistical proprieties of the series (i.e. average and variance) must not change in time. In this way, the estimation is consistent and it has the same statistical proprieties as a single temporal sample. Steps 1 and 2 are fundamental because they permit the discontinuities and the linear trend to be identify and remove

We consider “discontinuity of null degree” or level shift, any behavior which is the direct consequence of an immediate change of the measurement condition. This change is identified with a constant value in the time series. For example, GPS antenna substitution in a permanent station site; this causes a different localization of the phase center with respect the point of interest.

Nowadays the diffusion process of GPS permanent stations has not been followed by an adequate structure and management. There are few cases, in the world, where an appropriate structure, such as EUREF and IGS service, is available. In these cases, periodical solutions and significant transformations are published. The problem is that these publications could be incomplete due to human error. This is the main reason why research and detection of the jumps in the time series of GPS coordinates are important. The general tendency, at present, is to have an automatic and efficient method which allows the network adjustment to be made, in a short time.

2. FORWARD SEARCH METHOD

2.1 Basic concept

If the number of outliers is greater than 20% of the observations, a traditional robust estimator (i.e Huber, BIBER) cannot be used. Another possible solution must be found. A particular method, which is often used in the economic field, but for some years, has been used in the geomatic field is here described. Most outlier detection methods attempt to divide the data into two parts, a larger "cleaned" part and the outlier. The cleaned data are then used for parameter estimation. Forward Search (FS) can be considered as a hybrid method which permits to pass from an LMS solution (robust approach) to an LS solution (classical approach).

The main concept is to employ a subsample m, extracted from an original dataset. This subsample is considered to be without outliers. This subsample is employed by the LMS to estimate the unknown parameters. The coefficients estimated in this way are applied to all the observations, through the Least Squares method, to evaluate the residual value of each observation.

The absolute values of the residuals are analyzed and m+1 observations with smallest residuals are identified. This group of observations is employed to calculate the new parameters, by means of the LMS. The loop continues until all the observations have been involved; at the end of the process, a LS solution is obtained.

This method can be represented as in the following flowchart:

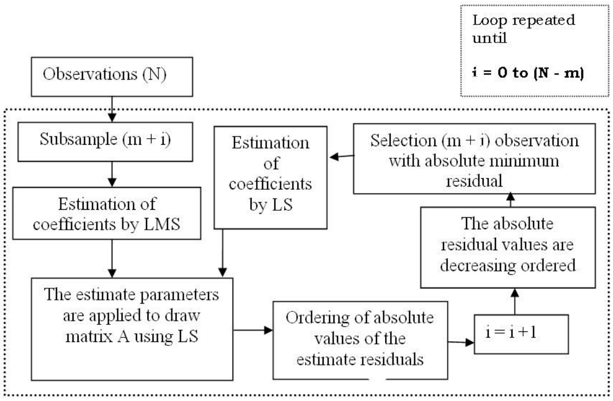
Loop repeated until
i = 0 to (N - m)
Observations (N)
Subsample (m + i)
Estimation of coefficients by LMS
Estimation of coefficients by LS
Selection (m + i) observation with absolute minimum residual
The absolute residual values are decreasing ordered
The estimate parameters are applied to draw matrix A using LS
Ordering of absolute values of the estimate residuals
i = i + 1

Figure 1. Sketch of the forward search

The new contribution of this method concerns variability because it allows one to pass from LMS to LS, and permits a continuous monitoring, epoch by epoch, of some parameters (residual, Cook's distances, coefficient estimates, t-statistic, etc). The purpose of this method is to divide the possible "clusters" into the dataset. It is possible to divide the data into "cleaned", which are employed to valuate the unknown coefficients and "outliers."

In the case where the values of the model parameters are known, it is not difficult to identify the possible outliers, because their relative observations have large residuals. The main difficulty arises when the outliers are included in the dataset used to estimate the parameters, which can contain some errors.

Many methods devoted to outlier detection try to divide the dataset into two parts: the main part is "cleaned" and the rest is the "outlier".

The method starts using the LMS method, which is based on the estimation of parameters x_p^*, and the median of the squared residuals, that is $\hat{v}_i^2(\hat{x})$, is minimized.

x_p^* minimizes the scaled estimated:

$$\sigma^2(\hat{x}) = e^2[med](\hat{x}) \quad (1)$$

where $e^2_{[k]}(\hat{x})$ is the kth ordered square of the residual.

Rousseuw defines a technique devoted to creating a subsample of p-dimension (APPENDIX A). The Forward Search (FS) uses this approach to define a subsample of parameters. A number of subsample elements equal to *m*, with a rank of A is usually defined and the *n-m* element is left to be tested. Some methods, such as LMS, use an augmented subsample to do the search, for example ns = m +1 or ns = m + 2.

The disadvantage of this is that increasing the number of subsample elements also increases the probability that it will contain outliers. In the case of FS, we start from ns = m, and we continue unit by unit, adding one observation at a time, until we obtain the condition ns = n.

During these n – m step, we can control the variation of some statistical parameter as values of residuals, Cook's distance, t-statistic. In particular, we can detect which new observation creates an immediate variation in the monitored parameters.

If we are using observations without any outliers, the observed statistical parameters should be constant.

In short, the Forward Search method is composed of three steps:

- Choice of the initial subsample;
- Adding observations at each epoch;
- Monitoring the statistical parameters.

2.1.1 Step 1: Choice of the initial subsample: If the model contains m parameters, the forward search algorithm starts with the selection of a subset of m units. Observations in this subset should be outlier free. Let Z=(X, y), so that Z is n x (m+1). If n is moderate and *m << n*, the choice of the initial subset can be performed by exhaustive enumeration of all $\binom{n}{m}$ distinct *m*-tuples $S^{(m)}_{i1,......,im} = \{z_{i1},...., z_{im}\}$ where $z^T_{1_i}$ is the i_1^{th} row of Z, for $1 \leq i_1,...., i_m \leq n$, and $i_j \neq i_{j'}$.

Let ι' = $[i_1,......., i_m]$ and $e_{i,S_\iota^{(m)}}$ be the least squares residual for unit i given the observations in $S_\iota^{(m)}$. The m-tuple $S_*^{(m)}$ is taken as the initial subset and this satisfies:

$$e^2[med], S_*^{(m)} = \min_l \left\lfloor e^2[med], S_l^{(m)} \right\rfloor \tag{2}$$

where $e_k^2, S_l^{(m)}$ is the k^th ordered square residual among $e_i^2, S_l^{(m)}$, for i =1, ..., n, and med is the integer part of (n+m+1)/2.

If $\left(\frac{n}{m}\right)$ is too large, we use some larger numbers of samples, for example 1000.

2.1.2 Step 2: Adding observations at each epoch: Given a subset of dimension q ≥ m, the forward search moves to dimension q+1 by selecting the q+1 units with the smallest squared Least Squares residuals, the units being chosen by ordering all squared residuals. The forward search estimator $\hat{\beta}_{FS}$ is defined as a collection of Least Squares estimators in each step of the forward search; that is:

$$\hat{\beta}_{FS} = \left(\hat{\beta}_p^*, \ldots\ldots, \hat{\beta}_n^*\right) \tag{3}$$

In most moves from q to q+1 just one new unit joins the subset. It may also happen that two or more units join $S_*^{(q)}$ as one or more leave. At the next step the remaining outlier in the cluster seem less outlying and therefore several may be included at once. Obviously, several other elements therefore have to leave the subsample.

The method in not sensitive to the method used to select an initial subset, provided unmasked outliers are not included at the start. What is important in the FS procedure is that the initial subsample is either free of outliers or contains unmasked outliers which are immediately removed by the forward procedure.

2.1.3 Step 3: Monitoring of the statistical parameter: Step 2 is repeated until all units are included in the subsample. If just one observation enters $S_*^{(q)}$ at each move, the algorithm provides an ordering of the data according to the specified null model, with observations furthest from it joining the subset at the least stages of the procedure.

The estimate of σ^2 does not remain constant during the forward search as observations that have small residuals are sequentially selected. Thus, even in the absence of outliers, the residual mean square estimate $v^2_{S_*^{(m)}} < v^2_{S_*^{(n)}} = v^2$ for m < n.

3. PROGRESSIVE FORWARD SEARCH APPROACH

3.1 Choice of the initial subsample

Starting from the original forward search approach, we have implemented some specific parts, to characterize this algorithm when a time series has to be analyzed.

The choose of the initial subset is made with a numerical method which is devoted to identifying a small "cleaned" subset, in all the data. The significance value of the test should be low, in order to avoid an under-estimation of the model accuracy, which causes greater sensibility of the method with respect to the noise, in particular during the first step of the analysis. A satisfactory value of significance α equal to 0.3% from practical test, has been defined. Residual value (σ*), (Rousseeuw et al., 1987), defined with respect to the LMS model, through a robust approach which starts from a preliminary estimation of the residual S_0, is employed to define the weight matrix ω_I . This is possible if the following equations are used:

$$S_0 = 1.4826\left(1 + \frac{5}{n-p}\sqrt{med(r_i^2)}\right)$$

$$\omega_i = \begin{cases} 1 & if \quad \left|\dfrac{r_i}{S_0}\right| \le 2.5 \\ 0 & else \end{cases} \tag{4}$$

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n} \omega_i r_i^2}{\sum_{i=1}^{n} \omega_i - p}}$$

where *n* = number of elements used

r_i = residual of the element i of the series

p = number of estimated coefficient of the model

med = median operator

The selection of the cluster element can now be done, using the Neyman setting:

$$\left|\frac{r_i}{\sigma^*}\right| \le 3 \tag{5}$$

3.2 Monitoring of the statistical parameters

Using the ratio between the measured value (S_r) and the predicted value (S_p), it is possible to discriminate if an element belongs to the current cluster or not. Usually a multi-linear regression is used (see Figure 2)

The sensibility of the test can be calibrated in order to the adopt a confidence interval, for example 95%. In this way, the discontinuities are identified as repeated elements of the series with residuals greater than the forecasted values. The employment of a model derived from a restricted number of elements leads to having an under estimation of the accuracy. In the case where the accuracy is lower than the general noise of the series, the method is excessively sensible to the noise.

If the solution is considered to belong to the previous cluster, a new model is estimated, considering this element and adopting LS method.

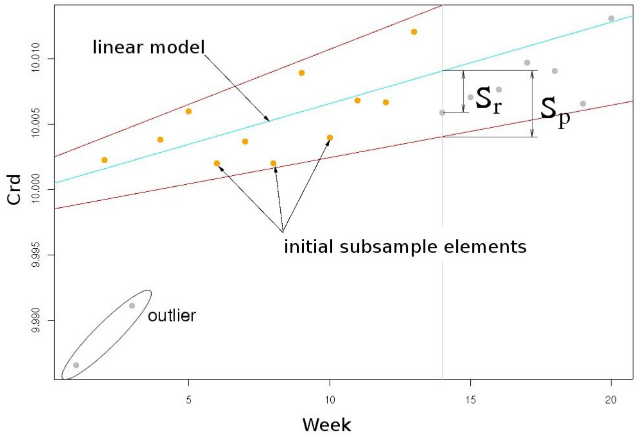

Figure 2. Monitoring of the statistical parameters

This procedure allows a model to be estimated, increasing the redundancy. A checked dataset is used to define this new model. At the end of cluster identification, the estimation jump procedure can start. A single model of regression is estimated for each identified cluster. When the different clusters are detected, the first and the last value of the regression model are saved. In this way the jumps are calculated as the difference between the values assumed by the model in the two different consecutive clusters. This is shown in figure 3.

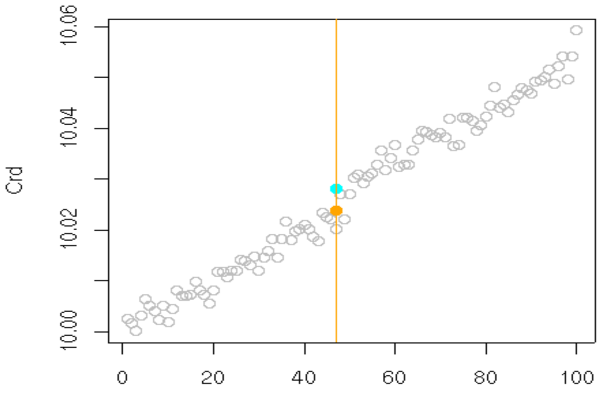

Figure 3. Definition of the discontinuities

4. ANALYSIS OF THE RESULTS

4.1 Test settings

The statistical approach involved in our analysis is necessary because of the kind of time series and the noise; the test was carried out considering several artificial time series with the analogue behavior of the real GPS permanent station data.

In this way, a known jump entity has been considered in the time series as a constant value in a known epoch. This is a great advantage because it allows to be sure about the time and the entity of the discontinuity.

Different jump values have been considered in order to evaluate the relationship between the sensibility of the method and the noise level. Different jump values were considered for each level of noise. These values were increased by 10% per time, until reaching at 4σ. Each test has been repeated 100 times, with same jump/noise ratio. The results obtained using the "*progressive forward search*" are described in the follow parts.

4.2 First test: α = 99.7%

The following parameters were considered in each test: jump value, epoch and estimated value.

All correct identified jumps are flagged as "correct", considering a tolerance equal to +/- 4 epochs with respect to where the jump was really located. Positive results are obtained (> 50% of tentative) considering a ratio jump/noise equal to a 2.5σ, as shown in figure 4.

The quality of the jump estimation is described in figure 5, where the ratio between the real value and the estimated value is identified, but only in the "correct" cases.

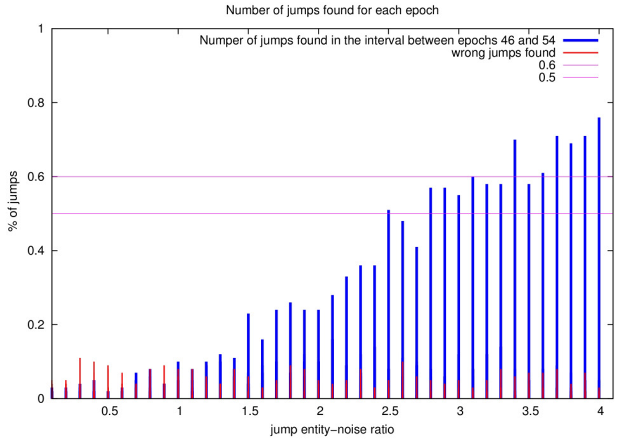

Figure 4. Percentage of success

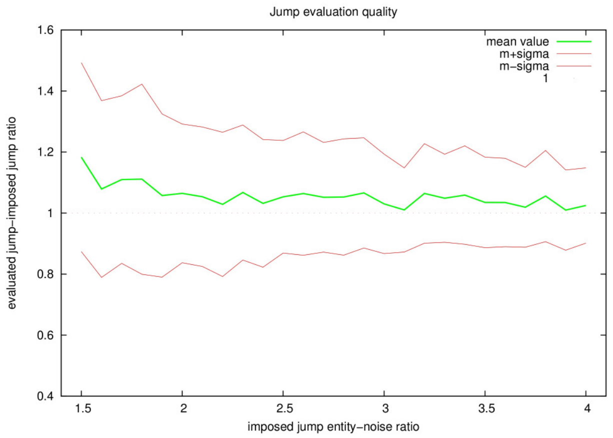

Figure 5. Ratio between the estimated value and the real jump values

A average trend is detectable in figure 5, with 10% of difference with respect to the real value. If the jump value increases, there are more correct identifications. This aspect permits the estimation accuracy to be improved about 20-35% of the real value of the jump. This improvement starts when the jump/noise ratio is equal to 2.5σ.

4.3 First test: α = 95%

After the first test, we considered a significance level equal to 4.6%, considering the right term in [5] equal to 2. The results of this test are described in figures 6 and 7.

In this case the total number of “successes” decrease with respect to the total number of tests carried out. On the other hand, the accuracy of the estimation is still the same.

Time series analysis is an important step because it allows undocumented anomalies to be detected eventually. When the jumps are documented, their values are unknown and it is not always easy to estimate them due to the nature of the jump.

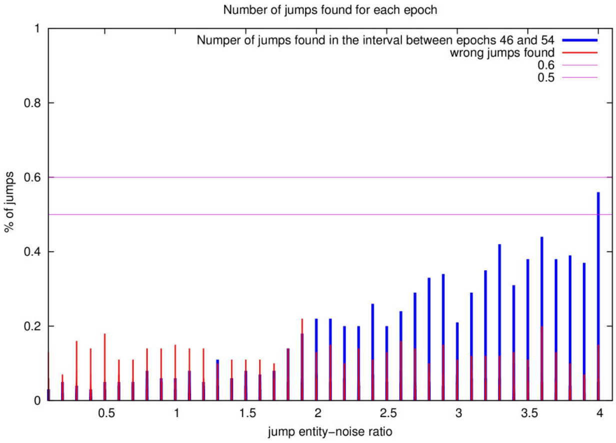

Figure 6. Percentage of success

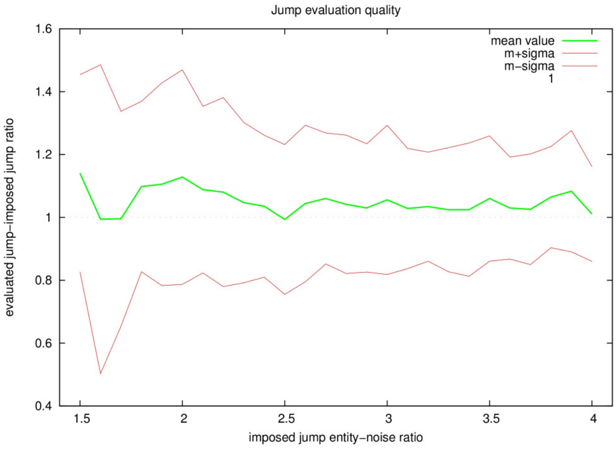

Figure 7. Ratio between the estimated value and the real jump values

5. CONCLUSION

Analysing the forward search method it is possible to denote that the choice of the initial subset is a critical step. It is very important to define a subset which significantly describes the population. This is one of the main characteristics of robust statistics.

The size definition of the subset has to respect some constrains, which are apparently contrasting. In fact the subset has to be sufficiently small but also significantly large, in order to be significant from the statistical point of view. The minimum temporal distance between two consecutive discontinuities is another important discriminating factor. The success of the progressive forward search depends on the quality of the model used to describe the behaviour of the time series.

For these reasons, it is necessary to adopt a simplified priori hypothesis about the model employed, which in the geodetic case can be defined using polynomial and trigonometric terms. In a multi-linear model the cyclic component is considered to be known, therefore it is not included in the unknown terms.

The use of synthetic data allows to be analyzed more easily the algorithm of the progressive forward search, the advantages and disadvantages to be identified and some other techniques to be selected that, jointed together with the forward search, could create a good procedure devoted to outlier detection.

The statistical approaches are not unambiguous, because there are several parameters (objective function, discriminating factor, etc) which can have different effects on the final estimation. The analysis of a synthetic set assumes an important role because it permits the contribution of the single parameter in the algorithm to be defined. Using the progressive forward search it is possible to detect the discontinuities with a good level of success (>60%) , if the jump/noise ratio is greater than 1.5-2σ.

References

Atkinson A. C., Riani M., Cerioli A., 2004. *Exploring multivariate data with the forward search,* Springer, London

Kailath T., Sayed A. H., Hassibi B., 2000. *Linear estimation*, Prentice Hall, New Jersey.

Teunissen P., 2001. *Dynamic data processing*, Delft University Press.

Rousseeuw P. J., Leroy A.M., 1987. *Robust regression and outlier detection*, Wiley & sons, New York

Perfetti N., 2006. Detection of station coordinates discontinuities within the Italian GPS Fiducial Network, *Journal of Geodesy*, pp. 1432-1394.

Pesenti M., Piras M., Roggero M., 2007. Analisi di serie storiche di dati da SP GPS, *ASITA proceeding* XI conferenza Torino, Italy, Vol.1 pp.

APPENDIX A. Least Median of Squares approaches

The main idea is to define a robust estimator, using the median value and with a high breakdown point. Rousseuw proposes minimising the median of the squared residuals (*LMS*) using one of Hampel’s ideas.

The Least Median of Squares is equal to:

$$\min(med(v_i^2)) \qquad (6)$$

The LMS estimate is equivalent to linear transformations of the design matrix and it has a breakdown point equal to 50%. One disadvantage is its poor asymptotic efficiency, that is, the dispersion of the estimator around the expected value should be small.

There are two different approaches that can give us the solution of the LMS estimator: the *combinatorial method* and the *resampling algorithm.*

Let ($a_{i,j}$, l_i) be a given complete dataset for i = 1,…, n and j = 1,…, m. The aim is to estimate the solution x, while obtaining the minimum value of the objective function $med(v_i^2)$. The number of unknown parameters are described by m instead of q and (q ≥ m) is the number of data points used to define the design matrix A.

The total number of subsamples, without repletion, that can be created is:

$$C_q^n = \frac{n!}{(q!(n-q)!} \qquad (7)$$

In the first approach, for each subsample we can compute an LS estimate, defining the residuals and calculating the $med(v_i^2)$. The j^th^ subsample which provides the smallest median value is the solution x, where the objective function is minimum. This algorithm is not always applicable. The second one is, on the other hand, a reasonable method.

Using the resampling algorithm, it is possible to set the number of subsamples in advance, instead of calculating all the subsamples. Rousseuw chose the number of subsample (ns) so as to have the probability α of at least one of the subsamples of being "good". If we denote the fraction of bad observation in a sample with ε, a subsample is good if it consists of q "good" observations. Assuming that ns/q is large, the probability α is equal to:

$$\alpha = 1-(1-(1-\varepsilon)^q)^{ns} \qquad (8)$$

This number is almost equal to 99% if ns ≈ $4{,}6 \cdot 2^q$. It is quite interesting to see that the probability α is independent of the number of observations n, but depends only on the number of subsamples and the number of observations that belong to a subsample.

The choice of q is variable. It is in fact possible to have:

- q = m, which represents the case where the number of observations chosen for a subsample is equal to the number of unknown parameters
- q > m, which represents the case where the number of observations extracted to create the subsample is larger than the number of unknown parameters. Aktinson and Weisberg have shown that the results of the resampling algorithm using q = m are very "rough" if n is small compared to m. Adopting q > m seems to produce better results in these cases, but the number of systems of normal equations to be solved increases quickly. Other authors have demonstrated that using q = m+1, it is always possible to find the absolute minimum of the LMS estimator, using of the combinatorial approach.