COMPUTATIONAL METHODS FOR ROBUST PROCEDURES

R. Lenth
The University of Iowa

ABSTRACT

In this paper, we discuss some computational techniques for implementing robust procedures. In particular, techniques for M-estimation are considered. Section 1 gives some methods for estimating the center of a population, with some attention given to defining what is meant by "location." In section 2, the question of scale estimation in conjunction with location estimates is discussed, along with a numerical example. Section 3 gives a generalization of the techniques for use in linear regression models.

1. ESTIMATING LOCATION

Suppose we have a set of data $x_1, x_2, \cdots, x_n$ which are independent observations from some population. Our objective is to estimate the "center," or location parameter, of the population. The estimator used should have good properties (such as low bias and high efficiency) over a variety of types of populations, including normal, nonnormal, and contaminated populations.

We should first define what we mean by location. There are some distributions for which the mean is undefined or infinite. This is often the case where the observations are ratios of random variables. For example, the ratio of two independent normal random variables with zero means has a Cauchy distribution. Thus the mean is not always a reasonable measure of location. However, all distributions have a median. If the distribution is symmetric, then the median is equal to the mean if it exists, and there is no ambiguity at all in defining the median as the "center" of the population. Moreover, when the population is asymmetric, the median is often considered more appropriate than the mean as a measure of location, an example being the distribution of incomes.

In view of the above, it may seem reasonable to use the sample median as an estimate of location, especially in view of the fact that it is relatively insensitive to the presence of

outliers in the data. However, the sample median is not as "robust" as the M-estimators we consider in this paper, in that it is often inefficient. The M-estimator corresponding to Huber's psi function, which will be described below, can be viewed as a compromise between the mean and the median. (However, it must be noted that the Huber estimate for a set of data is not necessarily between the sample mean and the sample median!) Thus, for symmetric populations it is quite clear that the Huber procedure yields estimates of the median. Unfortunately, the price we pay for the superior efficiency of M-estimators over the sample median is some degree of ambiguity in what is being estimated if the population is skewed. (We show later, however, that certain M-estimators corresponding to "re-descending" psi functions are related to estimates of the mode.)

We now proceed to define an M-estimator of location. Supposing that we have data values $x_1, x_2, \cdots, x_n$, the M-estimator $\hat{\theta}$ of location θ is a solution to the equation

$$g(\hat{\theta},\hat{\sigma}) = \sum_{i=1}^{n} \Psi\left(\frac{x_i-\hat{\theta}}{\hat{\sigma}}\right) = 0, \qquad (1.1)$$

where Ψ is some function (usually odd) chosen so as to provide desirable robustness properties, and $\hat{\sigma}$ is some (robust) estimate of spread. Commonly used Ψ functions include the following:

$$\Psi(t) = \begin{cases} t & \text{if } |t| \le c \\ c \text{ sign } t & \text{if } |t| > c \end{cases} \tag{1.2}$$

and

$$\Psi(t) = \begin{cases} \sin(t/c)/c & \text{if } |t| \le c\pi \\ 0 & \text{if } |t| > c\pi \end{cases} \tag{1.3}$$

The Ψ function (1.2) is due to Huber (1964), and is minimax for a class of symmetrically contaminated normal distributions while (1.3) is an example of a "redescending" Ψ function that performs well for very heavy-tailed distributions. See Andrews (1974). In practice, the "tuning constant" c in (1.2) and (1.3) is usually chosen to be around 1.5, for which the efficiency (relative to the sample mean) is greater than 95% in the normal case with known σ. There are a number of other Ψ functions in common use.

Note that in solving (1.1) we require some estimate $\hat{\sigma}$ of "spread," but we defer discussion of this to a later point and momentarily assume that we have a known scale value σ. In general, finding the solution $\hat{\theta}$ is a nonlinear problem, so that an iterative procedure is necessary. We mention three methods here, all of which require a starting value $\hat{\theta}_0$ (e.g., the sample mean or median):

1. (Newton's method)

$$\hat{\theta}_{j+1} = \hat{\theta}_j - g(\hat{\theta}_j, \sigma)/g'(\hat{\theta}_j, \sigma)$$

$$= \hat{\theta}_j + \sigma \sum_{i=1}^{n} \Psi\left(\frac{x_i - \hat{\theta}_j}{\sigma}\right) \Big/ \sum_{i=1}^{n} \Psi'\left(\frac{x_i - \hat{\theta}_j}{\sigma}\right).$$

2. (Iterative reweighting)

Define $w(t) = \Psi(t)/t$, $w_{ij} = w\left(\frac{x_i - \hat{\theta}_j}{\sigma}\right)$.

Then $\hat{\theta}_{j+1} = \sum_{i=1}^{n} w_{ij} x_i \Big/ \sum_{i=1}^{n} w_{ij}$

$$= \hat{\theta}_j + \sigma \sum_{i=1}^{n} \Psi\left(\frac{x_i - \hat{\theta}_j}{\sigma}\right) \Big/ \sum_{i=1}^{n} w_{ij}.$$

3. (H Algorithm)

$$\theta_{j+1} = \theta_j + k\sigma \sum_{i=1}^{n} \Psi\left(\frac{x_i - \hat{\theta}_j}{\sigma}\right) \Big/ n$$

where k is a "fudge factor" usually chosen between 1 and 2.

In all cases, iterations are discontinued when $\hat{\theta}_{j+1}$ is nearly equal to $\hat{\theta}_j$ according to some prescribed criterion. All three of these algorithms perform quite well when Ψ is monotone. Newton's method usually converges more quickly than the others, but can fail badly when Ψ is redescending, especially if σ is too small. The iterative reweighting scheme

makes $\hat{\theta}$ look like a weighted average, where the weights are chosen by the data. The H Algorithm converges at about the same rate as iterative reweighting. The rate is slower than Newton's Method but convergence is more stable when a redescending Ψ function is used. The H Algorithm is not of great importance in location estimation but its adaptation to regression models (given later) is quite useful.

Note that, when suitably expressed, the only differences among these algorithms are the denominators of the terms used to modify the previous estimate. In the case of Huber's Ψ function (1.2), it is easily seen that the denominator for Newton's Method is simply the number of "inliers" (residuals less than or equal to $c\sigma$ in absolute value), as compared to Σw_i and n/k for iterative reweighting and the H Algorithm, respectively. Since the inliers each receive unit weight, the denominator used in iterative reweighting is somewhat larger than in Newton's Method.

We strongly recommend that, if a redescending Ψ function is used, the starting value for the iterative process be an M-estimate corresponding to a monotone Ψ, such as Huber's. Furthermore, this starting value should use a severe "tuning constant" (i.e., a low value of c in (1.2)). This is an effort to avoid converging to a local, rather than global, solution.

There is an interesting connection between M-estimates based on certain redescending Ψ functions and estimates of

the mode. Consider a "kernel" density estimate:

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right), \tag{1.4}$$

where K is some density function. The mode of f_n can then be used as an estimate of the mode of the true distribution f. Now consider, for example, the kernel:

$$K(t) = \begin{cases} [1+\cos(t/c)]/(2c\pi) & \text{if } |t| \le c\pi \\ 0 & \text{if } |t| > c\pi. \end{cases} \tag{1.5}$$

Then the solution to $f_n(\theta)$ = maximum is the same as the M-estimator $\hat{\theta}$ corresponding to Andrews' Ψ function (1.3), if the scale parameters h in (1.4) and σ in (1.1) are equal. In practice, a mode estimator would more likely be based on a normal kernel (which is not unlike (1.5)), and a value of h somewhat smaller than $\hat{\sigma}$. Nonetheless, this comparison is helpful in understanding what sort of "location" is being estimated using (1.3) (or other redescending Ψ), and in anticipating some of the problems that may arise (such as local solutions) in computation of the M-estimate.

2. SCALE ESTIMATION

We mentioned earlier that, in general, some estimate $\hat{\sigma}$ of scale, or spread, is necessary in order to compute an M-estimate of location. This scale estimate should be fairly

insensitive to gross outliers in the data. With regard to computing M-estimates, we have two choices: to keep $\hat{\sigma}$ fixed throughout the iterative procedure, or to modify the value of $\hat{\sigma}$ after each iteration.

When keeping scale fixed, the most popular choice is the median absolute deviation (MAD) from the median, defined by

$$\text{MAD} = \underset{1\le i\le n}{\text{median}}\{\,|x_i - \underset{1\le j\le n}{\text{median}}\{x_j\}|\,\}.$$

Then we could use $\hat{\sigma} = \text{MAD}/.6745$ to make it an unbiased estimate of the standard deviation if the data comes from a normal population. While this is insensitive to outliers, this estimate is occasionally so small that convergence problems arise in the algorithms for $\hat{\theta}$, especially Newton's Method. Various other procedures based on order statistics, such as the interquartile range, could also be used.

Unfortunately, computing medians and other order statistics requires sorting of the data, an expensive procedure for large data sets. Thus an alternative is to compute something like a standard deviation and modify this scale estimate at each iteration. For instance, if scale is known then the M-estimate of location can be viewed as the maximum likelihood estimate of θ for a population having density proportional to $\exp\{-\rho((x-\theta)/\sigma)\}$, where ρ is an antiderivative of Ψ. If we use maximum likelihood techniques to estimate σ as well, we obtain for the $j^{\underline{th}}$ iteration

$$\hat{\sigma}_j = \frac{1}{n} \sum_{i=1}^{n} w_{ij}(x_i - \hat{\theta}_j)^2$$

(where Ψ is scaled so that, if $x_i = \hat{\theta}_j$, then $w_{ij} = 1$). The division by n in this expression could cause some problems, as the value of $\hat{\sigma}$ is likely to be small, leading to a large number of small weights, which in turn reduces $\hat{\sigma}$ substantially on the next iteration. A more desirable approach would be to use a truly weighted standard deviation, namely:

$$\hat{\sigma}_j^2 = \sum_{i=1}^{n} w_{ij}(x_i - \hat{\theta}_j)^2 / (\sum_{i=1}^{n} w_{ij} - 1). \tag{2.1}$$

Huber (1977) considered a scale invariant minimization problem of the form:

$$\hat{\sigma} \sum_{i=1}^{n} \rho\left(\frac{x_i - \hat{\theta}}{\hat{\sigma}}\right) + a\hat{\sigma} = \text{minimum},$$

which leads to the (iterated) scale estimate

$$\hat{\sigma}_{j+1}^2 = \frac{\hat{\sigma}_j^2}{a} \sum_{i=1}^{n} X((x_i - \hat{\theta}_j)/\hat{\sigma}_j), \tag{2.2}$$

where $X(t) = t\Psi(t) - \rho(t)$. For this to be asymptotically unbiased in the normal case, one should use $a = (n-1)E[X(U)]$ where U is a standard normal random variable. Note that for the Ψ function (1.2), we have $X = \frac{1}{2}\Psi^2$ so that (2.2) is proportional to $\sum_{i=1}^{n} w_i^2(x_i - \hat{\theta})^2$. This suggests that a possible compromise between (2.1) and (2.2) may be

$$\hat{\sigma}_j^2 = \sum_{i=1}^{n} w_{ij}^2 (x_i - \hat{\theta}_j)^2 / (\sum_{i=1}^{n} w_{ij}^2 - 1). \qquad (2.3)$$

Many other possibilities exist. It seems that (2.1) and (2.3) are more easily interpretable since they are equal to the usual standard deviation if all the weights are equal to 1.

An example of the iterative reweighting procedure in conjunction with the iterative scale estimator (2.1) is given in Table 1. Here we have 20 values generated from a "slash" distribution (the distribution of the ratio of independent normal and uniform random variables). The true median is 0 and the standard deviation is undefined. The data values are given in the first column and the remaining columns show the weights in each iteration (a blank entry indicates a weight of 1.0). The weighting function used corresponds to Huber's Ψ (1.2) with tuning constant $c = 1.5$.

The starting estimates were obtained by setting all $w_{i0} = 1$, which corresponds to the usual sample mean and standard deviation for $\hat{\theta}_0$ and $\hat{\sigma}_0$, respectively. Subsequent iterations, as indicated earlier, are weighted means and standard deviations using $w_{ij} = \Psi((x_i - \hat{\theta}_j)/\hat{\sigma}_j)/((x_i - \hat{\theta}_j)/\hat{\sigma}_j)$ $= \min[1, 1.5\hat{\sigma}_j/|x_i - \hat{\theta}_j|]$ in this case. Note that the weights settle down rather quickly, and by the tenth iteration the changes are very small. The value of $\hat{\sigma}$ cannot be viewed as an estimate of population standard deviation (which is infinite in this example). It is merely an outlier-insensitive measure of spread.

Table 1. Iterative reweighting procedure: an example.

	WEIGHTS					
DATA	INITIAL	ITER #1	ITER #2	ITER #3	ITER #4	ITER #10
-1.21	1	1				
.25	1	1				
-.24	1	1				
-.66	1	1				
.75	1	1				
.04	1	1				
2.28	1	1				
.50	1	1				
.60	1	1				
-4.21	1	1				
.53	1	1				
43.75	1	.414	.297	.253	.234	.219
1.47	1	1				
.21	1	1				
.44	1	1				
-2.33	1	1				
-1.02	1	1				
-1.36	1	1				
25.08	1	.768	.535	.451	.416	.387
1.31	1	1				
Σw_i	20	19.182	18.832	18.704	18.650	18.606
$\hat{\theta}$ = weighted $\bar{x}$	3.309	1.810	1.262	1.055	.966	.894
$\hat{\sigma}$ = weighted s	11.152	8.296	7.159	6.663	6.435	6.245

3. REGRESSION METHODS.

The techniques for location and scale estimation can be adapted easily to the estimation of coefficients in a linear regression model. In particular consider a model of the form

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \text{error}.$$

(If an intercept is desired, take $x_{i1} \equiv 1$.) Robust regression estimates are obtained by solving

$$\sum_{i=1}^{n} \rho([y_i - \sum_{j=1}^{p} \beta_j x_{ij}]/\hat{\sigma}) = \text{minimum} \qquad (3.1)$$

in place of the usual "least-squares" criterion. Here, ρ is an antiderivative of the Ψ function used earlier. If ρ is smooth then (3.1) is equivalent to finding a solution to

$$\sum_{i=1}^{n} x_{im} \Psi(\Delta_i/\hat{\sigma}) = 0 \quad (m = 1,2,\cdots,p), \qquad (3.2)$$

where $\Delta_i = y_i - \sum_{j=1}^{p} \beta_j x_{ij}$. As in the location case, $\hat{\sigma}$ is a robust scale estimate, and the solution is (usually) unique if ρ is convex. (There are pathological cases in which there is no unique solution even though ρ is convex.) For nonconvex ρ (redescending Ψ) we again must be careful that we don't arrive at a local rather than global solution to (3.1).

For purposes of describing the algorithms, it is convenient

to rewrite the model in matrix notation:

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon},$$

where $\underset{\sim}{y}$ is the vector of observations $y_1, y_2, \cdots, y_n$, $\underset{\sim}{\beta}$ is the vector of coefficients, X is the $n \times p$ matrix of x_{ij} values and $\underset{\sim}{\epsilon}$ is the vector of errors. Further, let $\underset{\sim}{\Delta} = \underset{\sim}{y} - X\underset{\sim}{\beta}$ be the vector of residuals and $W = \text{diag}(w_1, w_2, \cdots, w_n)$ be the matrix of weights assigned by the robust procedure.

Newton's Method modifies the estimate $\hat{\underset{\sim}{\beta}}$ of $\underset{\sim}{\beta}$ in each according to

$$\hat{\underset{\sim}{\beta}} \longleftarrow \hat{\underset{\sim}{\beta}} + A^{-1}X'W\underset{\sim}{\Delta} \tag{3.3}$$

where A is the $p \times p$ matrix consisting of the values $a_{\ell m} = \sum_{i=1}^{n} x_{i\ell} x_{im} \psi'(\Delta_i/\hat{\sigma})$.

The iterative reweighting procedure is, of course, standard weighted least-squares,

$$\hat{\underset{\sim}{\beta}} = (X'WX)^{-1}X'W\underset{\sim}{y},$$

which can be rewritten as

$$\hat{\underset{\sim}{\beta}} \longleftarrow \hat{\underset{\sim}{\beta}} + (X'WX)^{-1}X'W\underset{\sim}{\Delta}. \tag{3.4}$$

The H Algorithm is given by

$$\hat{\underset{\sim}{\beta}} \longleftarrow \hat{\underset{\sim}{\beta}} + k(X'X)^{-1}X'W\underset{\sim}{\Delta}, \qquad (3.5)$$

where k is a "fudge factor."

Again, the only distinction among these algorithms is the "denominator" (in this case an inverted matrix) in the modification term. Newton's Method converges at the fastest rate if no problems arise, but there is some danger of A being singular or the solution oscillating among a few values. These possibilities must be anticipated in writing a computer program based on Newton's Method, requiring some rather elaborate algorithms. Iterative reweighting and the H Algorithm converge at the same rate, but the H Algorithm is clearly advantageous, since $X'X$ may be inverted (decomposed) once and for all and no further matrix inversions are necessary. A "fudge factor" of $k = n/\sum_{i=1}^{n} w_i$ has worked quite successfully in this author's experience.

If one chooses to use fixed scale, an appropriate scale estimate would be

$$\hat{\sigma} = \text{median}\{|y_i - \sum_{j=1}^{p} \tilde{\beta}_j x_{ij}|\}/.6745$$

where $\tilde{\underset{\sim}{\beta}}$ is the least absolute values (L_1) estimate of $\underset{\sim}{\beta}$. For iterated scale estimates, one may use a natural adaptation of one of those used in location estimates. For example, (2.1) would become:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} w_i (y_i - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2 / (\sum_{i=1}^{n} w_i - p), \qquad (3.6)$$

or we could use $a = (n-p)E[\psi(u)]$ in the regression equivalent of (2.2).

One convenient numerical method for the H Algorithm is the "Modified Gram-Schmidt" procedure, which involves an orthogonalization of the columns of X. Then the $\underset{\sim}{y}$ vector is swept out according to the modified columns of X, resulting in the vector of least-squares residuals. Each iteration is then simply a weighted sweep of the residual vector $\underset{\sim}{\Delta}$. The advantage of this method is that the residuals are always explicitly available for purposes of computing the weights. For a detailed description of this method, see Lenth (1977).

4. SOME COMMENTS

Techniques for computing robust M-estimators are fairly well-developed, and the results are generally much more appealing than least-squares estimates when outliers are present, and comparable to least-squares when the data is "clean." Variations of the techniques given here can be applied to a wide variety of problems, including graduation and smoothing. See Huber (1979) and Lenth (1979).

Hogg (1979) suggests that a good way to use robust procedures in practice is to perform the "classical" (e.g.,

least-squares) technique and then see if the results are much different if an M-estimate such as a Huber or Andrews procedure (possibly with just a few iterations) is performed. If the two results are comparable, then one can proceed with the classical method and its associated statistical properties; otherwise, take a hard look at the data with special consideration of the seeming outliers indicated by low weights.

If, however, we can find no reasonable explanation for the anomalies in the data, there is no real justification for discarding them. This is where some knowledge of the statistical properties of M-estimates is needed, and unfortunately there are few practical results known in this regard. While there are some asymptotic results available, many of these depend on knowledge of scale, and little is known in the case of small sets of data. Much more work is needed before we can construct meaningful confidence limits and tests based on robust estimators.

BIBLIOGRAPHY

Andrews, D.F. (1974). A robust method for multiple linear regression, Technometrics, 16, 523-31.

Hogg, R.V. (1979). Statistical robustness: one view of its use in applications today, American Statistician, 33, 108-15.

Huber, P.J. (1964). Robust estimation of a location parameter, Ann. Math. Statist., 35, 73-101.

_____ (1977). Robust Statistical Procedures, Philadelphia: Society of Industrial and Applied Mathematics.

_____ (1979). Robust smoothing, Robustness in Statistics, New York: Academic Press, 33-47.

Lenth, R.V. (1977). A computational procedure for robust multiple regression, Tech. Report 53, Dept. of Statistics, Univ. of Iowa.

Lenth, R.V. (1979). On robust smoothing splines and validation methods. In preparation.