RR 90-07

MAJORIZATION WITH ITERATIVELY REWEIGHTED LEAST SQUARES: A GENERAL APPROACH TO OPTIMIZE A CLASS OF RESISTANT LOSS FUNCTIONS

Peter Verboon

Department of Data Theory

University of Leiden

Financial support by the Netherlands organization for scientific research (NWO) is gratefully acknowledged

MAJORIZATION WITH ITERATIVELY REWEIGHTED LEAST SQUARES: A GENERAL APPROACH TO OPTIMIZE A CLASS OF RESISTANT LOSS FUNCTIONS

Peter Verboon

In this paper a general algorithm is given for the optimization of a class of resistant loss functions. The theory on majorization is used to prove that this algorithm converges and that it is actually based on iteratively reweighted least squares. For the most important resistant loss functions, a majorization function is given. Finally some relations are shown between the majorization approach and the EM algorithm.

1. Introduction

It is already known for a very long time that the use of the least sum of squares criterion is very vulnerable when there are outliers in the data. This means that a technique which is based on a least squares loss function may come up with results which are very much influenced by the effect of one or more outliers. From the wide bulk of literature on this topic we may safely state that loss functions based on least squares (LS) are definitely not resistant to outliers.

The reason for this phenomenon is clear: due to the square large residuals will have a relatively large contribution to the loss. Since the objective is to minimize the loss these very large residuals will not be allowed, therefore the model will be attracted to the relative neighbourhood of the outliers and consequently outliers will be fitted at least moderately well. It follows that outliers cannot be neglected by a LS criterion, for this would cause too large residuals and consequently a high loss.

The obvious solution for this problem is to replace the LS criterion by a resistant one. However, the gain of resistance will obviously cost some extra theoretical and computational efforts. In this paper we will discuss three alternative loss functions that are supposed to possess resistant properties. These functions are: the *Huber* function (Huber, 1964), Tukey's *biweight* function (e.g. Andrews et al., 1972,) and a *three-part redescending* function (Hampel, 1968). The biweight and the three-part redescending functions are more radical than the Huber function, for they place a bound on the effect of the residuals upon the loss. In other words when a residual is larger than a particular value, it will cause no further increase in the loss. These functions have proved to be valuable in the context of robust regression and robust estimation of location.

From a computational point of view the optimization of these functions often consist of two steps: (i) the conditional optimization of the loss criterion for some set of weights, (ii) the computation of the weights as a function of the residuals derived from step (i).
It follows that weights play a very important role in these resistant loss functions. The use of weights to create resistant procedures can be looked upon from two different points of view, an intuitive and a technical one. From the intuitive point of view one could argue that the weights are used to increase or decrease the importance of some observations. An in some way important observation will obtain a large weight, while for instance an outlier will be down-weighted. In this way weights are a natural tool to deal with outliers and hence to create resistant procedures. However these weights are not known in advance and are found as a function of the residuals. The residuals then are computed on the basis of weighted observations. This brings us in a natural way to an iterative procedure in which weights and residuals are alternatingly computed.
From a technical point of view a particular loss function is considered and the weights are theoretically derived to create a convergent algorithm that minimizes the objective function.

To minimize the three functions mentioned above we are going to use a *majorization* approach from which an iteratively reweighted least squares algorithm is derived. For each objective function the majorization is based on a family of weighted LS functions, in which the weights are defined as functions of the residuals. The theory on majorization will provide us with the proof of convergence of the algorithm.

In the next section the majorization approach is briefly explained. In the remainder of this paper it is shown for each loss function how the majorization functions should be chosen, and thus how the weights are derived from the theory. We will also examine the derivative of each function, because this clearly shows the influence of the residuals upon the solution. Furthermore the derivative has a direct connection with the choice of the weights.

2. Majorization

The resistant loss functions that we are going to use in this study are complicated functions which cannot be minimized analytically. To minimize these functions we will use the majorization approach. A thorough discussion of this approach in the context of multidimensional scaling is given in De Leeuw (1988) and in De Leeuw and Heiser (1980). Some interesting applications are Meulman (1986), De Leeuw and Bijleveld (1988), Verboon and Heiser (1989).
The general idea of majorization is to define a family of simple, mostly quadratic functions (so-called majorization functions), and repeatedly minimize these instead of the (complicated) objective function. The majorization function should always be larger than or equal to the objective function and both functions should have at least one point in common. If these simple functions meet the requirements of a proper majorization function it can be proved that the algorithm converges to a minimum. This minimum is always the global minimum if the objective function is convex.

Consider some convex function $\phi(\mathbf{r})$, which is a function that cannot be minimized easily. The set $\mathbf{r}$ is the set of residuals, defined as the difference between observed values and a specified model. Thus $\mathbf{r} = \mathbf{z}$(= observed values) - $\mathbf{z}^*$(= model). To minimize $\phi(\mathbf{r})$ we define a family of majorization functions $\mu(\mathbf{r}; \mathbf{w})$. In each step of the algorithm we minimize $\mu(\mathbf{r}; \mathbf{w})$ as a function of $\mathbf{r}$ for fixed $\mathbf{w}$. The parameter set $\mathbf{w}$ defines the exact shape of the majorization function, this set will be different in each step. We will refer to the set $\mathbf{w}$ as *variable weights*: i.e. weights that are variable. The variable weights are computed as a function of the residuals that have been found in the previous step of the algorithm, thus: $\mathbf{w} = w(\mathbb{r})$, where $\mathbb{r}$ denotes the "old" set of residuals. The shape of $w(\mathbb{r})$ depends on $\phi(\mathbf{r})$ and $\mu(\mathbf{r}; \mathbf{w})$. Suppose we have some initial set of estimates together with their corresponding residuals, $\mathbb{r}$. The majorization algorithm can now be seen as consisting of three consecutive steps:

1) computation of variable weights $\mathbf{w} = w(\mathbb{r})$.
2) minimization of majorization function $\mu(\mathbf{r}; \mathbf{w})$ to find new model parameters.
3) computation of new residuals $\mathbf{r} = \mathbf{z} - \mathbf{z}^*$.

After these steps the function $\phi(\mathbf{r})$ is evaluated and if the change in loss compared to the previous step is smaller than some convergence criterion we stop, otherwise we return to step 1) and cycle through these steps until convergence.

In the present study the function $\mu(\mathbf{r}; \mathbf{w})$ is always chosen as a weighted least squares function. Because the weights may change each time we cycle through the algorithm, we can call this optimization procedure an iteratively reweighted least squares procedure (cf. Holland & Welsch, 1977).

Before we are going to use a particular function $\mu(\mathbf{r}; \mathbf{w})$, we should first verify that $\mu(\mathbf{r}; \mathbf{w})$ is indeed a proper majorization function. In order to do so, the following conditions formulated as (in)equalities should be true:

$$\mu(\mathbb{r}; \mathbf{w}) = \phi(\mathbb{r}), \tag{2.1}$$

$$\mu(\mathbf{r}; \mathbf{w}) \geq \phi(\mathbf{r}). \tag{2.2}$$

Condition (2.1) says that the majorization function should have exactly the same function value as the objective function in a so-called *supporting point*, which is defined as the set of parameter estimates in the previous step of the algorithm. By condition (2.2) it is required that the majorization function is always larger than the objective function.

In the following sections we will describe resistant loss functions, together with their majorization functions. For each of the latter functions, requirements (2.1) and (2.2) will be proved. Furthermore for each of the loss criteria the corresponding weight functions will be derived.

3. Huber's function

A straightforward extension of the LS criterion is Huber's loss function (Huber, 1964). The function is based on the idea of using absolute residuals instead of squared residuals when the residuals are large. It is defined as a summation over residual components:

$$\phi_H(\mathbf{r}) = \sum_{i=1}^{n} \phi_H(r_i), \tag{3.1}$$

with n as the number of elements in $\mathbf{r}$. The elements of the summation are defined as:

$$\phi_H(r_i) = \begin{cases} 1/2\, r_i^2 & \text{if} \quad |r_i| < c \\ c|r_i| - 1/2c^2 & \text{if} \quad |r_i| \geq c\,. \end{cases} \tag{3.2}$$

The objective is to find a set of parameter values that minimizes a combination of sums of squared residuals and sums of absolute residuals. In Huber's function c is some prechosen constant, the so-called *tuning constant*, which distinguishes small residuals from large ones. So the Huber function consists of two parts: for residuals smaller than the tuning constant the ordinary least sum of squared residuals is used and for large residuals the criterion is the least sum of the absolute residuals. The basic idea is that large residuals, due to outliers for instance, will have a less devastating effect upon the solution than in the least squares case. In other words the Huber function should be more resistant to outliers than least squares. Note that if c is chosen very large the Huber function becomes the ordinary least squares function, if c is chosen near zero Huber's function considers the least sum of absolute residuals. Huber's function therefore is also a generalisation of the so-called L_1 norm.

The derivative of Huber's function is:

$$\psi_H(r_i) = \begin{cases} r_i & \text{if} \quad |r_i| < c \\ c \,\mathrm{sgn}\,(r_i) & \text{if} \quad |r_i| \geq c. \end{cases} \tag{3.3}$$

So we see that the addition of an observation always affects the estimation of the function parameters. However, the influence of an observation which yields a residual larger than the tuning constant is less than in the LS case. Therefore this function will be more resistant than the LS criterion.

Since Huber's function cannot be minimized directly we use majorization. The following majorization function will be used to minimize Huber's function, which is also a summation over n elements:

$$\mu_H(\mathbf{r};\, \mathbf{w}) = \sum_{i=1}^{n} \begin{cases} 1/2 w_i r_i^2 & \text{if} \quad \mathbb{r}_i < c \\ 1/2 w_i r_i^2 + 1/2c\, \mathbb{r}_i - 1/2c^2 & \text{if} \quad \mathbb{r}_i \geq c. \end{cases} \tag{3.4}$$

This is a weighted quadratic functions of the residuals, where $\mathbb{r}_i$ represents the absolute value of the residual found in the previous step. We will prove that the function $\mu_H(\mathbf{r};\, \mathbf{w})$ is indeed a proper majorization function if we choose the weights as

$$w_i = \begin{cases} 1 & \text{if} \quad \mathbb{r}_i < c \\ \dfrac{c}{\mathbb{r}_i} & \text{if} \quad \mathbb{r}_i \geq c \end{cases} \tag{3.5}$$

Notice that because of this choice the value of the weights are always between 0 and 1.

If $\mu_H(\mathbf{r}; \mathbf{w})$ is a proper majorization function then the conditions in (2.1) and (2.2) should hold. For condition (2.2) there are four different cases to be examined (see also Heiser, 1987).

(i) $|r_i| < c$ and $\mathbb{r}_i < c$
Now $\mu_H(\mathbf{r}; \mathbf{w})$ and $\phi_H(\mathbf{r})$ are equal by definition.

(ii) $|r_i| \geq c$ and $\mathbb{r}_i < c$
Starting from the inequality $1/2[c - |r_i|]^2 \geq 0$ we find that $c|r_i| - 1/2c^2 \leq 1/2r_i^2$. Now using the second part of (3.2) and the first part of (3.4) we obtain $\phi_H(\mathbf{r}) \leq \mu_H(\mathbf{r}; \mathbf{w})$, if the weights are chosen as in (3.5).

(iii) $|r_i| \geq c$ and $\mathbb{r}_i \geq c$
We start from the inequality $w_i[r_i - \mathbb{r}_i]^2 \geq 0$, which gives

$$w_i\, \mathbb{r}_i^2 + w_i\, r_i^2 - 2w_i\mathbb{r}_i r_i \geq 0,$$

substituting the expression for the weights yields:

$$c\, \mathbb{r}_i + w_i\, r_i^2 - 2cr_i \geq 0 \quad <=> \quad 2cr_i \leq c\, \mathbb{r}_i + w_i\, r_i^2 \quad <=>$$

$$cr_i - 1/2c^2 \leq 1/2c\, \mathbb{r}_i + 1/2w_i\, r_i^2 - 1/2c^2 \quad <=> \quad \phi_H(\mathbf{r}) \leq \mu_H(\mathbf{r}; \mathbf{w}).$$

(iv) $|r_i| < c$ and $\mathbb{r}_i \geq c$
Since $0 \leq |r_i| < c \leq \mathbb{r}_i$,we also have $0 \leq |r_i|/c < 1 \leq \mathbb{r}_i/c$. The left part remains smaller than one if it is squared, which yields:

$$0 \leq r_i^2/ c^2 < 1 \leq \mathbb{r}_i/c,$$

and therefore $r_i^2 \leq c\, \mathbb{r}_i$.

This equality remains valid when multiplied by the nonnegative quantity $(\mathbb{r}_i - c)$:

$$(\mathbb{r}_i - c)\, r_i^2 \leq (\mathbb{r}_i - c)\, c\, \mathbb{r}_i,$$

after some rewriting we obtain:

$$r_i^2 \leq (c/\mathbb{r}_i)r_i^2 + c\, \mathbb{r}_i - c^2 \quad <=> \quad \phi_H(\mathbf{r}) \leq \mu_H(\mathbf{r}; \mathbf{w}).$$

Steps (i) to (iv) prove (2.2). To prove (2.1) we have:

(v) $r_i = \mathbb{r}_i$, leading to: $\phi_H(\mathbb{r}) = \mu_H(\mathbb{r}; \mathbf{w})$
The equality is immediately verified by substitution of $\mathbb{r}_i$ in (3.2) and (3.4) and by substituting the expression for the weights in (3.5).

We have shown that in each situation $\phi_H(\mathbf{r}) \leq \mu_H(\mathbf{r}; \mathbf{w})$ and that $\phi_H(\mathbb{r}) = \mu_H(\mathbb{r}; \mathbf{w})$. It follows that $\mu_H(\mathbf{r}; \mathbf{w})$ is a proper majorization function for $\phi_H(\mathbf{r})$.

So the minimization of Huber's function requires a two-step algorithm. To show what the Huber procedure is actually doing, a graphical presentation is given in Figure 3.1 of one cycle in the algorithm for the linear regression problem.

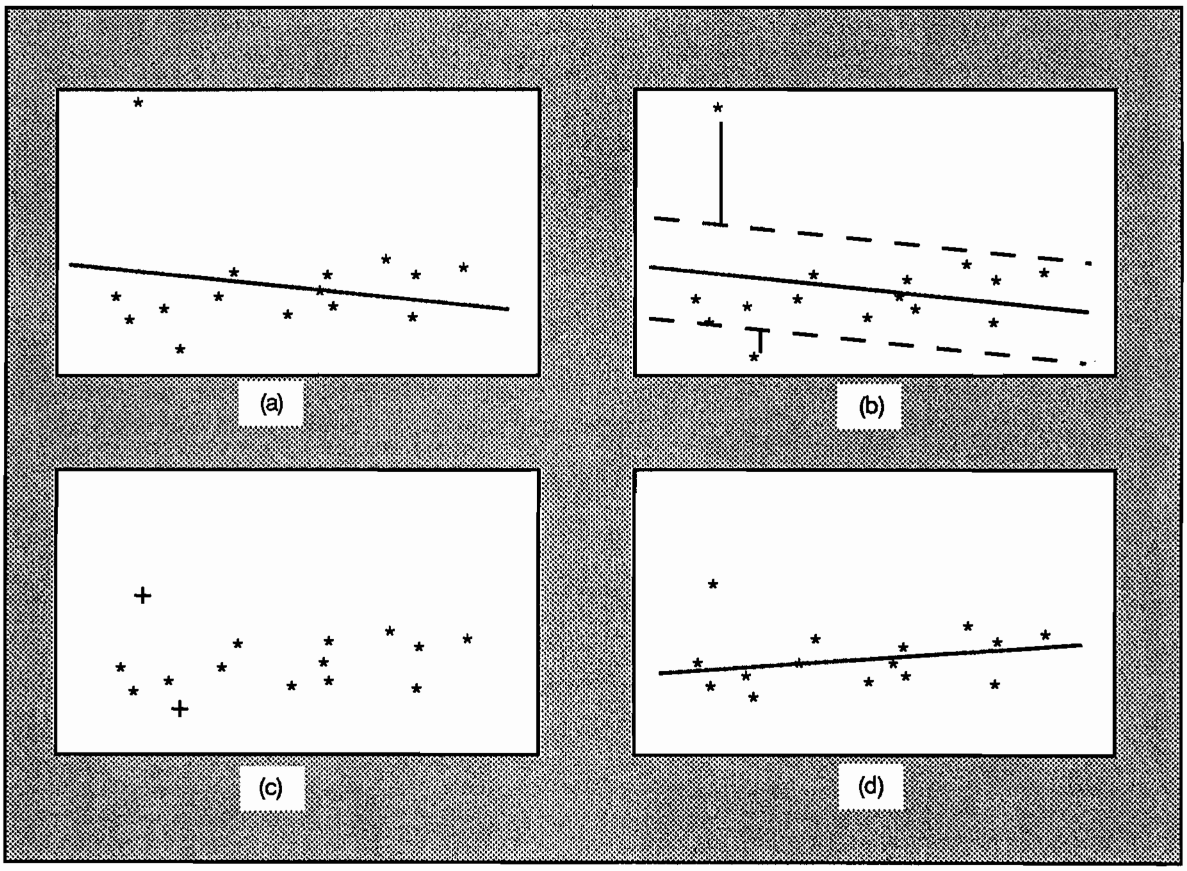

Figure 3.1 **Illustration of reweighted least squares algorithm for the Huber function.**

In (a) there are 15 points with the ordinary least squares regression line fitted through the points. This line clearly does not fit the data satisfactory. In (b) a band centered at the regression line is indicated by two parallel lines. The band width depends on the tuning constant. Points outside the band are translated , parallel to the y-axis, to the border of the band. In this way, new pseudo-observations, see (c), are defined: they coincide with the original points inside the band and with the translated points on the border. On these pseudo-data we again fit a least squares regression line (d). We may proceed by defining a new band with the same width around this line, find new pseudo-values and so on, until the steps converge to a situation where the regression line has become stable.

In Figure 3.2 the Huber function with its derivative are shown. The derivative clearly shows that the influence of outliers is bounded. After a certain value there is no further increase in influence upon the solution. This property was also illustrated in Figure 3.1 where points with large residuals were projected to the border of a band in order to diminish their influence in the next step of the algorithm.

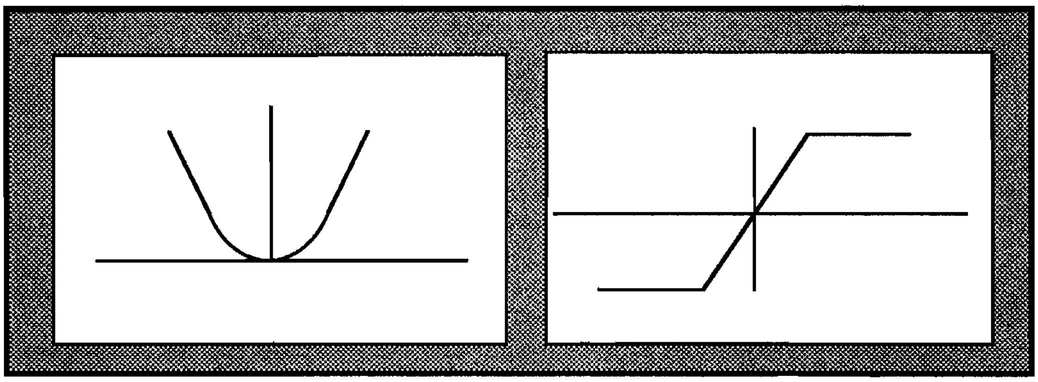

Figure 3.2 The Huber function and its derivative.

A convenient property of the Huber function is its monotonic ψ-function. By this property we know that the Huber function is convex and therefore we may conclude that after convergence of the majorization algorithm the global minimum is attained. The next sections are about redescending functions, which are not convex and therefore do not necessarily attain the global minimum after convergence.

4. The biweight function

One of the earliest proposals for a resistant loss function is Tukey's biweight function, an abbreviation of bi-square weight. As the name already suggest this procedure attributes weights to the observations. The function consists of two parts, like Huber's function. The first part defines the function $\phi_B(r)$ for small residuals and the second part is constant for large residuals. The function is defined as:

$$\phi_B(r_i) = \sum_{i=1}^{n} \phi_B(r_i), \tag{4.1}$$

where the elements of summation are:

$$\phi_B(r_i) = \begin{cases} c^2/6(1 - (1 - (r_i/c)^2)^3) & \text{if} \quad |r_i| \leq c \\ 1/6 & \text{if} \quad |r_i| > c. \end{cases} \tag{4.2}$$

So for residuals with absolute values larger than c there is no further increase of the loss: it is therefore said that the influence of the residuals is bounded. The biweight is actually a redescender. The derivative of the biweight function is:

$$\psi_B(r_i) = \begin{cases} r_i(1 - (r_i/c)^2)^2 & \text{if} \quad |r_i| \leq c \\ 0 & \text{if} \quad |r_i| > c \end{cases} \tag{4.3}$$

The derivative or influence curve shows that the increase in influence is zero for residuals larger than c. Because of this property the biweight belongs to the class of *hard redescending* functions. Both function are drawn in Figure 4.1.

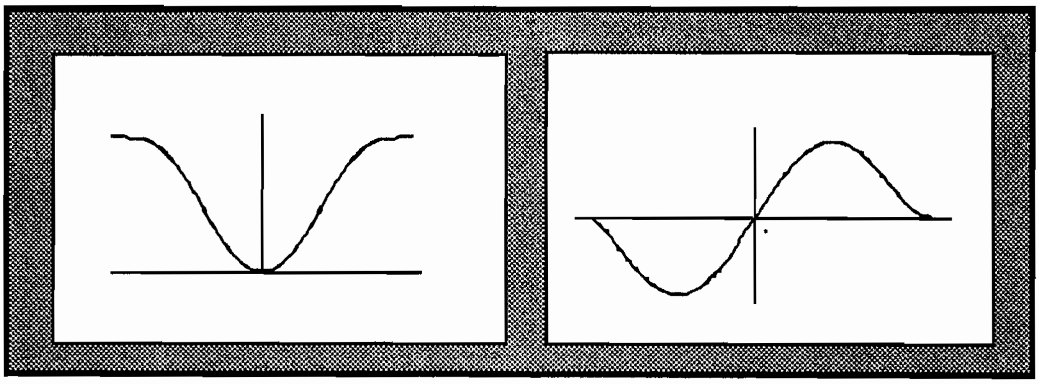

Figure 4.1 The Biweight function and its derivative.

In order to minimize the biweight function the following majorization function is used:

$$\mu_B(r; w) = \sum_{i=1}^{n} c^2/6(1\text{-}3w_i(1\text{-}(r_i/c)^2)+2w_i^{3/2}). \tag{4.4}$$

The r_i are again defined as the absolute residuals from the previous step. This function is also a weighted quadratic function of the residuals. The weights w_i in this function are defined as:

$$\mathrm{w_i} = \begin{cases} (1 - (\mathbb{r}_i/c)^2)^2 & \text{if} \quad \mathbb{r}_i \leq c \\ 0 & \text{if} \quad \mathbb{r}_i > c. \end{cases} \tag{4.5}$$

It is easy to minimize (4.4), this gives us new residuals which define new weights according to (4.5).

To prove that the proposed algorithm indeed converges, we must prove that $\mu_B(\mathbf{r}; \mathbf{w})$ is a proper majorization function. To do so we will examine the conditions (2.1) and (2.2) which must be satisfied for a majorization function. For ease of notation we assume without loss of generality a fixed tuning constant $c = 1$.

(i) It must be true that $\phi_B(\mathbb{r}) = \mu_B(\mathbb{r}; \mathbf{w})$. Substitution of (4.5) in (4.4) yields for the first part of the function:

$$\mu_B(\mathbb{r}; \mathbf{w}) = 1/6\ \Sigma(1- 3(1 - \mathbb{r}_i^2)^3 + 2(1 - \mathbb{r}_i^2)^3) = 1/6\ \Sigma(1- (1 - \mathbb{r}_i^2)^3) = \phi_B(\mathbb{r}).$$

The second part of both functions are equal by definition.

This is the first property (2.2) of a majorization function. Next it must also be true that the value of the majorization function is never smaller than the value of the objective function. Again, as with the Huber function, all possible situations must be considered. However in with this function there are only two situations. If the inequalities can be proved for any element, they will consequently be proved for the summation over the elements too.

(ii) $|r_i| > 1$.

We must prove: $\phi_B(r_i) \leq \mu_B(r_i; w_i)$ <=>

$$1/6 \leq 1/6\ (1- 3w_i\,(1 - r_i^2) + 2w_i^{3/2}) \tag{4.6}$$

This is equation is true since $w_i\,(1 - r_i^2) \leq 0$ and $w_i^{3/2} > 0$, thus the term $(1 - 3w_i\,(1 - r_i^2) + 2w_i^{3/2}) > 1$, which proves (4.6).

(iii) $|r_i| \leq 1$.

We must prove that $\phi_B(r_i) \leq \mu_B(r_i; w_i)$ <=>

$$1/6\ (1- (1 - r_i^2)^3) \leq 1/6\ (1- 3w_i\,(1 - r_i^2) + 2w_i^{3/2}). \tag{4.7}$$

We start from the general inequality $(a-b)^2 \geq 0$, which gives $a^2 \geq 2ab - b^2$. Using this inequality we may also write:

$$(1 - r_i^2)^2 \geq 2(1 - r_i^2)(1 - r_i^2) - w_i. \tag{4.8}$$

Next both sides are multiplied by the positive quantity $(1 - r_i^2)$, yielding:

$$(1 - r_i^2)^3 \geq 2(1 - r_i^2)^2(1 - r_i^2) - w_i(1 - r_i^2).$$

Substituting the second part of (4.8) for the term $(1 - r_i^2)^2$ does not change the inequality:

$$(1 - r_i^2)^3 \geq 2[2(1 - r_i^2)(1 - r_i^2) - w_i](1 - r_i^2) - w_i(1 - r_i^2).$$

Working out this expression yields

$$(1 - r_i^2)^3 \geq 3(1 - r_i^2)w_i - 2w_i^{3/2}.$$

Substracting both terms from one and multiplying by 1/6 gives

$$1/6(1 - (1 - r_i^2)^3) \leq 1/6(1 - 3(1 - r_i^2)(1 - r_i^2)^2 + 2(1 - r_i^2)^3),$$

which proves (4.7).

Having proved that in both possible conditions $\mu_B(r_i;\, w_i) \geq \phi_B(r_i)$, we may conclude that $\mu_B(r_i;\, w_i)$ is a proper majorization function for $\phi_B(r_i)$.

5. The three-part redescending function

Another redescending loss function is the *three-part redescender*, introduced by Hampel (1968). In the shape of the function we actually distinguish four parts: an ordinary quadratic part with a positive slope, a linear part, a quadratic part with a negative slope and finally a constant part. The function smoothly changes from one part to another. It is also defined as a summation over residual components:

$$\phi_R(\mathbf{r}) = \sum_{i=1}^{n} \phi_H(r_i). \tag{5.1}$$

To keep the formulas relatively simple we choose the function with equally spaced intervals. An additional advantage is that the function now depends on only one tuning constant instead of three. However, from a computational point of view it is better to choose the decreasing part of the derivative larger than the other parts. The components of Hampel's function (multiplied by a constant 2) are defined as:

$$\phi_R(r_i) = \begin{cases} r_i^2 & \text{if} \quad |r_i| \le c \\ 2c|r_i| - c^2 & \text{if} \quad c < |r_i| \le 2c \\ 6c|r_i| - r_i^2 - 5c^2 & \text{if} \quad 2c < |r_i| \le 3c \\ 4c^2 & \text{if} \quad |r_i| > 3c. \end{cases} \tag{5.2}$$

The first two components are identical to Huber's function. However, for larger residuals the function diminishes the increase in influence until even no increase is allowed for residuals larger than $3c$. This can also be seen if we write down the derivative:

$$\psi_R(r_i) = \begin{cases} r_i & \text{if} \quad |r_i| \le c \\ c \text{ sign}(r_i) & \text{if} \quad c < |r_i| \le 2c \\ 3c \text{ sign}(r_i) - r_i & \text{if} \quad 2c < |r_i| \le 3c \\ 0 & \text{if} \quad |r_i| > 3c \end{cases} \tag{5.3}$$

In Figure 5.1 an illustration is given of the three-part redescender and its derivative.

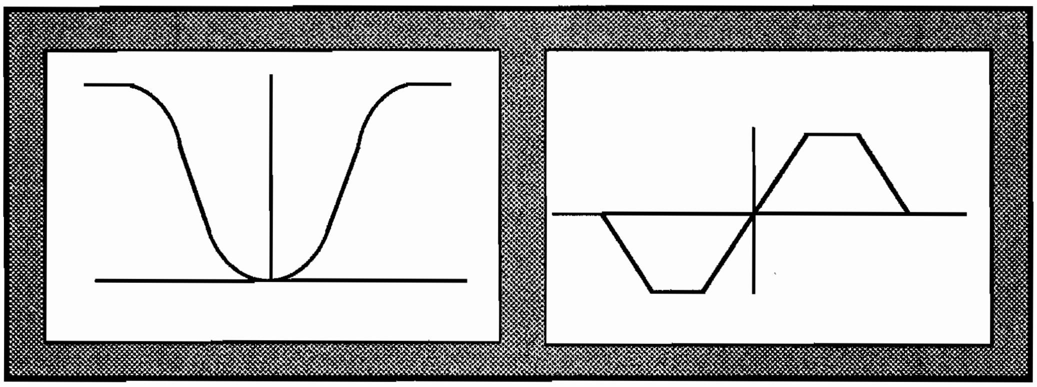

Figure 5.1 The three-part redescending function and its derivative.

The majorization function is again a weighted quadratic function, which is also defined in four pieces. The components of this function are:

$$\mu_R(r_i;\, w_i\,) = \begin{cases} w_i r_i^2 & \text{if} \quad |r_i| \le c \\ w_i r_i^2 + c|r_i| - c^2 & \text{if} \quad c < |r_i| \le 2c \\ w_i r_i^2 + 3c|r_i| - 5c^2 & \text{if} \quad 2c < |r_i| \le 3c \\ w_i r_i^2 + 4c^2 & \text{if} \quad |r_i| > 3c \end{cases} \tag{5.4}$$

The $\mathbb{r}_i$ are the absolute values of the residuals from the previous step. This function is a proper majorization function for the three-part redescender if we choose the weights as:

$$w_i = \begin{cases} 1 & \text{if} \quad \mathbb{r}_i \leq c \\ c/\mathbb{r}_i & \text{if} \quad c < \mathbb{r}_i \leq 2c \\ (3c/\mathbb{r}_i) - 1 & \text{if} \quad 2c < \mathbb{r}_i \leq 3c \\ 0 & \text{if} \quad \mathbb{r}_i > 3c \end{cases} \qquad (5.5)$$

Like in the case of Huber's function we will prove that $\mu_R(r_i;\, w_i)$ is a proper majorization function of $\phi_R(\mathbf{r})$. First we examine condition (2.1). This condition is easily verified by substituting $r_i = \mathbb{r}_i$, which yields for each of the four parts: $\mu_R(\mathbb{r}_i;\, w_i) = \phi_R(\mathbb{r}_i)$.

Condition (2.2) is more complicated to verify, for we have to check all 16 combinations, that is, for each part of the majorization function we must prove that it is above each part of the objective function. However, most of these situations are straightforward. Table 1 summarizes the situation.

Table 1. Overview of all conditions to be proved

absolute residuals	previous absolute residuals			
	$< c$	c-$2c$	$2c$-$3c$	$>3c$
$< c$	section 3	section 3	*	$4c^2 > \max \phi_R(\mathbf{r})$
c-$2c$	section 3	section 3	*	$4c^2 > \max \phi_R(\mathbf{r})$
$2c$-$3c$	$\min \mu_R(\mathbf{r};\mathbf{w}) = \max \phi_R(\mathbf{r})$	*	*	$4c^2 = \max \phi_R(\mathbf{r})$
$> 3c$	$\min \mu_R(\mathbf{r};\mathbf{w}) > \max \phi_R(\mathbf{r})$	*	*	$4c^2 = \phi_R(\mathbf{r})$

For residuals and previous residuals smaller than $2c$ we are dealing with the Huber function, for which the conditions are proved in section 3. The inequalities in the last column of the table can also directly be verified. Since the majorization function in this situation is constant ($4c^2$), we only have to compare it with the maximum values of the objective function. These maximum values of the objective function can easily be found by substituting the border values of the corresponding interval; the maximum values of $\phi_R(\mathbf{r})$ are respectively c^2, $3c^2$, $4c^2$ and $4c^2$.

The same reasoning holds for the last two elements of the first column, for we can easily find the minimum values of the majorization function and the maximum values of the objective function, which immediately prove the inequalities. The minimum values of

$\mu_R(\mathbf{r};\, \mathbf{w})$ are respectively $4c^2$ and $9c^2$, the maximum values of $\phi_R(\mathbf{r})$ are respectively $4c^2$ and $4c^2$.

There are still six situations left for which condition (2.2) has to be proved. These are marked with a star in Table 1.

<u>cell 3.2</u>: $2c < |r_i| \leq 3c;\ c < \mathbb{r}_i \leq 2c$

Starting from $(r_i - 2c)^2 \geq 0$, gives

$$r_i^2 - 4cr_i + 4c^2 \geq 0.$$

Furthermore $w_i(r_i - \mathbb{r}_i)^2 \geq 0$, with $w_i = c/\mathbb{r}_i$ gives $c\mathbb{r}_i - 2cr_i + w_ir_i^2 \geq 0$. Summing both inequalities yields:

$$w_ir_i^2 + r_i^2 - 6cr_i + c\mathbb{r}_i + 4c^2 \geq 0 \quad <=>$$

$$w_ir_i^2 + c\mathbb{r}_i - c^2 \geq 6cr_i - r_i^2 - 5c^2 \quad <=> \quad \mu_R(r_i;\, w_i) \geq \phi_R(r_i).$$

<u>cell 4.2</u>: $3c < |r_i|\,;\ c < \mathbb{r}_i \leq 2c$

Since $r_i > 3c$, $2cr_i - 5c^2 > 0$. Again we have $w_i(r_i - \mathbb{r}_i)^2 \geq 0$, with $w_i = c/\mathbb{r}_i$ giving:

$$c\mathbb{r}_i - 2cr_i + w_ir_i^2 \geq 0.$$

Summing both inequalities yields:

$$w_ir_i^2 + c\mathbb{r}_i - 5c^2 \geq 0 \quad <=> \quad w_ir_i^2 + c\mathbb{r}_i - c^2 \geq 4c^2 \quad <=>$$

$$\mu_R(r_i;\, w_i) \geq \phi_R(r_i\,).$$

<u>cell 1.3</u>: $|r_i| \leq c;\ 2c < \mathbb{r}_i \leq 3c$

We must prove: $\mu_R(r_i;\, w_i) \geq \phi_R(r_i)$, thus $w_ir_i^2 + 3c\mathbb{r}_i - 5c^2 \geq r_i^2$, with $w_i = 3c/\mathbb{r}_i - 1$

(i) since $\mathbb{r}_i \leq 3c$, $w_ir_i^2 \geq 0$,

(ii) since $\mathbb{r}_i > 2c$, $3c\mathbb{r}_i - 5c^2 > c^2$,

(iii) since $r_i \leq c$, $r_i^2 \leq c^2$,

(ii) and (iii) yield (iv): $3c\mathbb{r}_i - 5c^2 > r_i^2$, finally (i) and (iv) prove $\mu_R(r_i;\, w_i) \geq \phi_R(r_i)$.

cell 2.3: $c < |r_i| \leq 2c$; $2c < \mathbb{r}_i \leq 3c$
We must prove: $\mu_R(r_i;\, w_i) \geq \phi_R(r_i)$, thus $w_i r_i^2 + 3c\mathbb{r}_i - 5c^2 \geq 2cr_i - c^2$ with $w_i = 3c/\mathbb{r}_i - 1$. Consider this equation as a function of r_i and $\mathbb{r}_i$ which must be positive on the given interval:

$$f(r_i;\, \mathbb{r}_i) = w_i r_i^2 + 3c\mathbb{r}_i - 4c^2 - 2cr_i \geq 0.$$

We will eliminate r_i by taking the partial derivarive to r_i and setting it equal to zero:

$$\frac{\partial f(r_i;\, \mathbb{r}_i)}{\partial(r_i)} = 2w_i r_i - 2c = 0,$$

thus, $r_i = c/\mathrm{w}$. Independent of $\mathbb{r}_i$ the minimum of $f(r_i;\, \mathbb{r}_i)$ on the given interval is attained for $r_i = 2c$, since $0 < w_i < 1/2$. Now substitute $r_i = 2c$ and $w_i = 3c/\mathbb{r}_i - 1$, which gives

$$12c^3/\mathbb{r}_i + 3c\mathbb{r}_i - 12c2 \geq 0,$$

multiplying with the term $\mathbb{r}_i/3c$ gives

$$4c^2 + \mathbb{r}_i^2 - 4c\mathbb{r}_i \geq 0,$$

which is always true since $(\mathbb{r}_i - 2c)^2 \geq 0$.

cell 3.3: $2c < |r_i| \leq 3c$; $2c < \mathbb{r}_i \leq 3c$
From the inequality $3c/\, \mathbb{r}_i(\mathbb{r}_i - r_i)^2 \geq 0$, we obtain $3c/\, \mathbb{r}_i\, r_i^2 - 6cr_i + 3c\mathbb{r}_i \geq 0$. Adding the term $(- r_i^2 - 5c^2)$ and rewriting yields:

$$3c/\, \mathbb{r}_i\, r_i^2 - r_i^2 + 3c\mathbb{r}_i - 5c^2 \geq 6cr_i - r_i^2 + 5c^2 \quad <=>$$

$$\mu_R(r_i;\, w_i) \geq \phi_R(r_i\,).$$

cell 4.3: $3c < |r_i|$; $2c < \mathbb{r}_i \leq 3c$
We must prove: $\mu_R(r_i;\, w_i) \geq \phi_R(r_i\,)$, thus $w_i r_i^2 + 3c\mathbb{r}_i - 5c^2 \geq 4c^2$
The minimum for the left side is attained when r_i^2 is minimal, thus when $r_i = 3c$. Substituting this value and dividing by $3c$ yields: $w_i 3c + \mathbb{r}_i - 3c \geq 0$. If we substitute for the weights and multiply the result by $\mathbb{r}_i$, we obtain: $9c^2 + \mathbb{r}_i^2 - 6c\mathbb{r}_i \geq 0$. This equality is always true since $(\mathbb{r}_i - 3c)^2 \geq 0$.

So we proved that in all conditions $\mu_R(r_i;\, w_i) \geq \phi_R(r_i\,)$.
Since we know that the inequalities are satisfied for each loss component r_i, it is also proved that the equalities are true for the summation over these components. So by all the above proofs we have shown that $\mu_R(r_i;\, w_i)$ is a proper majorization function for the optimization of $\phi_R(r_i\,)$.

6. Relations with Maximum Likelihood theory

In this section we will show that an iteratively reweighted least squares (IRLS) algorithm based on the majorization approach has a strong relation with maximum likelihood (ML) theory. In fact we will show that the presented algorithm largely resembles the well-known EM-algorithm (Dempster, Laird and Rubin, 1977).
Consider the linear regression model as an illustration:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{r}, \tag{6.1}$$

where $\mathbf{y}$ is a n-vector, $\mathbf{X}$ a matrix (n x p) with predictor variables and $\mathbf{b}$ the parameter vector to be estimated. The n-vector $\mathbf{r}$ contains the error or the residuals. We will assume for simplicity that $\mathbf{r}$ has a distribution function $f(\mathbf{r})$ with a fixed variance of 1.

The objective is to find the parameter vector $\mathbf{b}$ that maximizes the likelihood function: $L^*(\mathbf{b}) = \prod_{i=1}^{n} f(r_i)$, or the more convenient form, the logarithm of it: $L(\mathbf{b}) = \sum_{i=1}^{n} \log f(r_i)$.
In order to maximize the likelihood function we set its derivative with respect to the parameter $\mathbf{b}$ equal to zero:

$$\frac{\partial L(\mathbf{b})}{\partial(\mathbf{b})} = \sum_{i=1}^{n} \frac{-f'(r_i)}{f(r_i)}\mathbf{x}_i = \mathbf{0}, \tag{6.2}$$

where $\mathbf{x}_i$ represents a row of $\mathbf{X}$. Now if we define weights for each of the n elements as:

$$w_i = \frac{-f'(r_i)}{f(r_i)r_i}\,, \tag{6.3}$$

we may write the stationary equation of the log likelihood as

$$\sum_{i=1}^{n} w_i r_i \mathbf{x}_i = \mathbf{0}. \tag{6.4}$$

This is exactly the stationary equation of the weighted least squares loss function:

$$\sigma(\mathbf{b}) = \frac{1}{2} \sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i'\mathbf{b})^2 . \tag{6.5}$$

If we assume that $f(\mathbf{r})$ is the standard normal distribution, i.e. $\mathbf{r} \sim N(0,1)$, and we substitute this distribution in (6.3) then the weights are completely identified and become

$$\mathbf{w} = \frac{-(2\pi)^{-1/2} e^{-1/2\mathbf{r}^2}(-\mathbf{r})}{(2\pi)^{-1/2} e^{-1/2\mathbf{r}^2}\,\mathbf{r}} = 1.$$

So in this context we find the well-known result that using unweighted least squares actually implies the assumption of standard normally distributed residuals.

When a different probability function is used we can see from (6.3) that the weights depend on the residuals. In that case we have a similar problem as described before in the majorization procedure, for the residuals are obtained for a set of fixed weights and the weights are computed from the residuals. Apparently there is a strong resemblance between IRLS and ML. When the assumption is made that the residuals are coming from a normal/independent (N/I) distribution then it has been shown by Dempster, Laird and Rubin (1980) that an IRLS algorithm with weights defined as in (6.3) is in fact an example of the more general EM algorithm.

Now let's examine our majorization problem in terms of EM. The EM algorithm consists of (not surprisingly) two basic steps : the *expectation* (E)step and the *maximization* (M)step.

In the M-step the *expected* log likelihood function, $M(\mathbf{r}; \mathbb{r})$, is maximized as a function of the parameters, thus in this step the equations in (6.2) are solved. So in an iterative procedure the M-step provides updates for the parameters. This step corresponds therefore with the minimization of the majorization function, which also provides updates for the parameters. The majorization function is not the objective function, but some intermediate auxiliary function based on the residuals derived in a previous step. Likewise the expected log likelihood function is not the objective likelihood but also some intermediate auxiliary function based on previous residuals.

In the E-step we must find this expected log likelihood function, which actually boils down to computing new weights. To see this we assume that the residual vector $\mathbf{r}$ is a *scaled normal* random variable with a N/I distribution, which means $\mathbf{r} = \mathbf{u}\boldsymbol{\xi}^{-1/2}$ where $\mathbf{u}$ is a standard normal random variable and $\boldsymbol{\xi}$ a positive random variable distributed independently of **u**. So we may say that the variable $\boldsymbol{\xi}$ symbolizes the departure of standard normality. Now in the E-step we must find the weights to define some function:

$$\mathrm{M}(\mathbf{r}; \mathbb{r}) = \mathbf{E}\,[\log f(\mathbf{r}) \mid \mathbb{r}\,]. \tag{6.6}$$

This function is the equivalent of the quadratic majorization function, $\mu(\mathbf{r}; \mathbf{w})$, which was used in the previous sections. After working out the term $\log f(\mathbf{r})$ (see Dempster *et al.*, 1980) and dropping all irrelevant terms, we find

$$\mathrm{M}(\mathbf{r}; \mathbb{r}) = \mathbf{E}\,[\,\sum_{i=1}^{n} \xi_i r_i^2 \mid \mathbb{r}\,]. \tag{6.7}$$

Since we know from Dempster *et al.* (1980, theorem 2) that

$$\mathbf{E}\,[\,\boldsymbol{\xi} \mid \mathbb{r}\,] = \frac{-f'(\mathbb{r})}{f(\mathbb{r})\;\mathbb{r}} = w(\mathbb{r}), \tag{6.8}$$

we find that the computation of the expectation step is actually applying the weights function (6.3) to find new weights. By substitution of (6.8) in (6.7) and multiplying by -1 we now find that the intermediate function has the following well-known form:

$$\mathrm{M}(\mathbf{r}; \mathbb{r}) = \frac{1}{2}\sum_{i=1}^{n} w_i r_i^2 . \tag{6.9}$$

The shape of the function $\mathrm{M}(\mathbf{r}; \mathbb{r})$ actually depends on the weights and therefore we may also write this function as $\mathrm{M}(\mathbf{r}; \mathbf{w})$. It is clear now that this weighted quadratic function corresponds with the majorization function $\mu(\mathbf{r}; \mathbf{w})$, which is also a weighted quadratic function depending on previous residuals through the weights **w**.

So the relation between the EM algorithm and our majorization (IRLS) algorithm can be summarized as follows: in the M-step we minimize a weighted least squares function, while in the E-step new weights are computed as the expected values of a random variable given the previous residuals, which define a new quadratic function.

We already showed that both procedures consisted of a step in which weights were computed. In addition we can also show that these weights are exactly the same. First notice that the weights which we defined in (3.5), (4.5) and (5.5) could also be defined in a more general way as:

$$w(\mathbb{r}) = \frac{\psi(\mathbb{r})}{\mathbb{r}}. \tag{6.10}$$

This definition is valid for all objective functions that we considered in previous sections. It is now quite easy to see that this definition is exactly the same as the definition in (6.8), provided that the distribution function, f(r), is an exponential function of the following form:

$$f(\mathbf{r}) = a\, e^{-\phi(\mathbf{r})}, \tag{6.11}$$

where a is a constant and $\phi(\mathbf{r})$ is one of the objective functions discussed before. Substituting (6.11) in (6.8) yields:

$$w(\mathbf{r}) = \frac{-a\, e^{-\phi(\mathbf{r})}\,(-\psi(\mathbf{r}))}{a\, e^{-\phi(\mathbf{r})}\,\mathbf{r}} = \frac{\psi(\mathbf{r})}{\mathbf{r}}, \tag{6.12}$$

which proves that the weights are exactly the same.

From the above relation between the maximum likelihood theory and the majorization approach we can now get a better understanding about the meaning of the choice of the resistant functions. The shape of these functions correspond with the assumed shape of the N/I distributions of the residuals. Thus if we take as our objective function a hard redescender like for example the biweight, we implicitly assume a heavy tailed distribution of the residuals. On the other hand minimizing the L_1 norm corresponds with much thinner tails.

The objective functions that we used were introduced as so-called W-estimators because we showed that in order to optimize them we could define convergent algorithms via iteratively reweighted least squares. It is now also explained why these estimators are sometimes called M-estimators, for they are based on generalizations of maximum likelihood estimators.

7. What's the use

The majorization approach for resistant procedures was explained by assuming a set of residuals **r**, defined as the difference between the data and some kind of model,

$$\mathbf{r} = \mathbf{z} - \mathbf{z}^*. \tag{7.1}$$

In all objective functions it was implicitly assumed that **r** is a n-vector and furthermore that the n observations or objects which constitute **z** are independent. Because of this

general approach it follows that we are able to use the described resistant procedures for any problem that can be written in the form of (7.1). In other words we can insert for **z*** a number of different linear models and may directly apply one of the resistant procedures.
The first obvious illustration which can be written in this form is the linear multiple regression problem:

$$\mathbf{r} = \mathbf{z} - \mathbf{X}\mathbf{b}. \tag{7.2}$$

Here the matrix **X** (n x p) is a set of predictor variables, which are combined by the weights **b** to predict the data **z** (criterion variable) as good as possible. A lot of research in the context of robustness has already been done concerning this problem. For instance Rousseeuw & Leroy (1987) and Hampel et al. (1987) are among the most prominent references.
The three resistant functions from this paper have very frequently been applied to the multiple regression problem, both in Monte Carlo studies as with emperical data. It is clear from the many literature that these functions are indeed very useful for the analysis of contaminated data sets.
A direct generalization of (7.2) is the *multivariate* multiple regression problem:

$$\mathbf{R} = \mathbf{Z} - \mathbf{X}\mathbf{B}. \tag{7.3}$$

It this problem we aim to predict m variables **z** instead of a single one. Now **B** is a matrix (p x m) with regression weights. Since we can split the problem from (7.3) in m independent problems of the form:

$$\mathbf{r}_j = \mathbf{z}_j - \mathbf{X}\mathbf{b}_j, \tag{7.4}$$

multivariate multiple linear regression can also be handled by the majorization approach.

Finally we mention the principal component (PCA) problem. The PCA problem can be written in exactly the same form as (7.3), except for a transpose sign for **B**, since we can never have more components than variables. However, the matrix **X** is in PCA a set of unobserved variables, also called principal components. Another difference with (7.3) is that p (here the number of principal components) can never be larger than m (the number of observed variables). Like in multivariate multiple regression we can also split the problem as in (7.4). Although a resistant procedure for PCA obviously differs from one for regression analysis the general majorization approach can still be used. In PCA there is only an additional step within the weighted least squares part to compute **X**. This, however, does not basically change the general problem formulated in (7.1).

These are only a few illustrations for which the general approach for optimizing resistant loss functions can be used. With the theory from this paper we now have a tool to study a whole family of linear models with respect to their behaviour in relation to outliers.

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W. (1972). *Robust estimates of location: survey and advances*. Princeton, NJ: Princeton University Press.

De Leeuw, J. (1988). Convergence of the majorization method for multidimensional scaling. *Journal of Classification, 5*, 163-180.

De Leeuw, J. and Bijleveld, C. (1988). *Fitting longitudinal reduced rank regression models by alternating least squares*. Research Report RR-88-03. Leiden: Department of Data Theory.

De Leeuw, J. and Heiser, W.J. (1980). Multidimensional scaling with restrictions on the configuration. In: P.R. Krishnaiah (Ed.), *Multivariate Analysis V.* , p. 501-522. Amsterdam: North-Holland.

Dempster, A. P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal Royal Statistical Society, Series B, 39*, 1-38.

Dempster, A. P., Laird, N.M. and Rubin, D.B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. In P.R. Krishnaiah (Ed.), *Multivariate Analysis V,* p. 35-57. Amsterdam: North-Holland.

Hampel, F.R. (1968). *Contributions to the theory of robust estimation*. Ph.D. thesis, University of California, Berkeley.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: the approach based on influence functions*. New York: Wiley.

Heiser, W.J. (1987). Correspondence analysis with least absolute residuals. *Computational Statistics and Data Analysis, 5*, 337-356.

Holland, P.W. and Welsch, R.E. (1977). Robust regression using iteratively reweighted least squares. *Communications in Statistics, A6*, 813-827.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*, 73-101.

Meulman, J.J. (1986). *A distance approach to multivariate analysis*. Leiden: DSWO Press.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust regression and outlier detection.* New York: Wiley.

Verboon, P. and Heiser, W.J. (1989). *Robust loss functions in the orthogonal Procrustes problem.* Research Report RR-89-03. Leiden: Department of Data Theory.