Journal of Global Optimization **21:** 223–237, 2001.

Finding GM-estimators with global optimization techniques

RAFAEL BLANQUERO, EMILIO CARRIZOSA and EDUARDO CONDE
Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, Sevilla, Spain
(e-mail: rblanque@cica.es; ecarriz@cica.es; educon@cica.es)

Abstract. In this note we address the problem of finding the GM-estimator for the location parameter of a univariate random variable. When this problem is non-convex but d.c. one can use a standard covering method, which, in the one-dimensional case has a simple form. In this paper we exploit the structure of the problem in order to obtain d.c. decompositions with certain optimality properties in the application of the algorithm. Numerical results show that this general-purpose algorithm outperforms previous ad-hoc methods for this problem.

Key words: GM-estimators, Robust estimation, D.C. optimization, Covering methods

1. The model

Given a sample of n observations $y_1, y_2, \ldots, y_n$ the determination of a parameter that, in some sense, *represents* the data is a classical problem in Statistics. In the last decades the traditional least-squares method has been more and more frequently replaced by other approaches with better properties of robustness [19].

In particular, an M-estimator [14, 15], is an optimal solution of an optimization program of the form

$$\inf_{\theta\in\mathbb{R}} \sum_{1\leqslant j\leqslant n} \rho(r_j(\theta)), \tag{1}$$

where $(r_1(\theta), \ldots, r_n(\theta))$ is the vector of residuals,

$$r_j(\theta) = y_j - \theta \qquad j = 1, 2, \ldots, n$$

and $\rho : \mathbb{R} \longrightarrow \mathbb{R}$ is some continuous, even and nondecreasing function in $\mathbb{R}^+$.

The class of M-estimators has been further enlarged to the class of so-called *GM-estimators* (generalized M-estimators), in which the influence of each residual is made dependent on the observation y_j. In other words, θ^* is said to be a GM-estimator if it solves an optimization problem of the form

$$\inf_{\theta\in\mathbb{R}} \sigma(\theta), \tag{2}$$

with

$$\sigma(\theta) = \sum_{1\leqslant j\leqslant n} \rho_j(\theta),$$

Table 1. Examples of functions ρ

	Name	$\rho(t)$
1	Andrews	$\begin{cases} 1-\cos t & \text{if } \lvert t\rvert \leqslant \pi \\ 2 & \text{else} \end{cases}$
2	Biweight	$\begin{cases} 1-\left(1-t^2\right)^3 & \text{if } \lvert t\rvert \leqslant 1 \\ \frac{1}{2} & \text{else} \end{cases}$
3	Cauchy	$\frac{1}{2}\log\left(1+t^2\right)$
4	Fair	$\lvert t\rvert - \log(1+\lvert t\rvert)$
5	Huber	$\begin{cases} \frac{t^2}{2} & \text{if } \lvert t\rvert \leqslant 1 \\ \lvert t\rvert - \frac{1}{2} & \text{else} \end{cases}$
6	Logistic	$\log(\cosh t)$
7	$\ell_p \quad (p \geqslant 1)$	$\lvert t\rvert^p$
8	Talwar	$\begin{cases} \frac{t^2}{2} & \text{if } \lvert t\rvert \leqslant 1 \\ \frac{1}{2} & \text{else} \end{cases}$
9	Welsch	$\frac{1}{2}\left(1-\mathrm{e}^{-t^2}\right)^2$

each ρ_j is typically of the form

$$\rho_j(t) = \omega_j \rho\left(\frac{y_j - t}{C_j}\right)$$

for predetermined constants ω_j and C_j, and ρ a function as above. A list of typical examples of funtions ρ found in the literature [8, 15, 19], is presented in Table 1.

Finding a GM-estimator leads to solve Problem (2) in the case in which the scale parameter is assumed to be known. A much more realistic situation should considerate the estimation of the scale parameter simultaneously with the location coefficient [16]. Usually the scale estimation is computed by means of Eq. (3)

$$\sum_{1 \leqslant j \leqslant n} \chi\left(\frac{y_j - t}{s}\right) = 0 \tag{3}$$

where $\chi : \mathbb{R} \longrightarrow \mathbb{R}$. Nevertheless, in some models the estimation process alternates between the computation of the location parameter and the computation of the scale coefficient. For example, in MM-estimation [24] a three-stage procedure

is used. In the first stage a tentative approximation to the location parameter is computed assuming a fixed scale coefficient. In the second stage a M-estimate s^* of the errors scale is obtained using, in equation (3), residuals based on the initial estimate. Finally, in the third stage an M-estimate of the location parameter is computed by solving Problem (2) with $C_j = s^*$ for all $j = 1 \ldots n$.

In this paper, we restrict ourselves to this third stage of the process which is, in general, a difficult task, mainly due to the fact that the functions ρ_j used in Problem (2) may be non-convex, as is the case of the functions 1, 2, 3, 8, 9 of Table 1.

The usual strategy (e.g., [1, 8, 10, 15, 20, 21, 24]) for coping with (2) consists of deriving the normal equation

$$\sum_{1 \leqslant j \leqslant n} \rho_j'(\theta) = 0 \tag{4}$$

a zero of which is sought by means of an iterative procedure (e.g., Newton method).

Since (4) is not a sufficient condition for optimality in the non-convex case, such iterative procedures, strongly dependent of the initial guess, may be trapped in the neighborhood of a zero of (4) which does not represent a global minimum of (2).

When the problem is non-convex one has to choose between seeking a local minimum (close to the initial value, thus pressumably retaining some of its properties) or a global minimum. For the latter case, one can use well-known numerical procedures such as the covering method, first proposed by Piyavskii, [17], see also [3, 4, 5]. An outline of the algorithm is given below.

ALGORITHM 1.

Initialization: *Find a compact interval R^* known to contain an optimal solution of (2). Set* BEST.THETA *equal to an arbitrary $\theta_1 \in R^*$ and set* BEST.VALUE $:= \sigma$(BEST.THETA).

Construct $E^{(1)}$ and set $k = 1$.

Step k: *Set*

$$\text{LBOUND} = \min_{\theta \in R^*} E^{(k)}(\theta) \tag{5}$$

and let θ_{k+1} be its optimal solution. If BEST.VALUE $-$ LBOUND $\leqslant \varepsilon$ *then GoTo* **Output**. *Else, do:*

1. *Construct $E^{(k+1)}$ from $E^{(k)}$.*
2. *Update, if required,* BEST.VALUE *and* BEST.THETA
3. *GoTo step $k + 1$.*

Output: *STOP:* BEST.THETA *is an ε-optimal solution.*

The three critical issues of the algorithm are

1. Building R^*.
2. Defining the lower envelopes E^k of σ in order to guarantee convergence.
3. Solving the auxiliary problems (5).

Since ρ is assumed to be even and non-decreasing in $\mathbb{R}^+$, we can take as R^* the interval given by the extreme observations,

$$R^* = \left[\min_{j=1,\dots,n} y_j, \max_{j=1,\dots,n} y_j \right]. \tag{6}$$

In what follows we customize Algorithm 1 for the one-dimensional problem (2) under the extra assumption that each ρ_j is the difference of two convex functions, i.e., it is d.c., [11, 12, 22].

Hence, by the algebra of d.c. functions, σ is d.c. Let $\sigma = \sigma_1 - \sigma_2$ be a d.c. decomposition of σ, and let $\sigma_1'(\theta_i; \theta - \theta_i)$ be the directional derivative of σ_1 at θ_i in the direction $\theta - \theta_i$, i.e.,

$$\sigma_1'(\theta_i; \theta - \theta_i) = \lim_{t \downarrow 0} \frac{\sigma_1(\theta_i + t(\theta - \theta_i)) - \sigma_1(\theta_i)}{t}$$

Let h_i be given as

$$h_i(\theta) = \sigma_1(\theta_i) + \sigma_1'(\theta_i; \theta - \theta_i) - \sigma_2(\theta), \tag{7}$$

and define $E^{(k)}$ as the pointwise maximum of the functions h_i,

$$E^{(k)} = \max_{1 \leqslant i \leqslant k} h_i$$

Observe that this slightly differs from the standard choice $\widetilde{h}_i$,

$$\widetilde{h}_i(\theta) = \sigma_1(\theta_i) + \eta_i(\theta - \theta_i) - \sigma_2(\theta), \tag{8}$$

with $\eta_i \in \partial\sigma_1(\theta_i)$, the subdifferential of σ_1 at θ_i.

Since

$$\sigma'(\theta_i; \theta - \theta_i) = \max\{\eta(\theta - \theta_i) : \eta \in \partial\sigma_1(\theta_i)\},$$

[18], we see that $h_i \geqslant \widetilde{h}_i$. Moreover, such a choice will yield certain optimality properties, see Section 2.

We will show in this section that, with this choice, Problem (5) is solved after inspecting at most two easily located points, and the updating of $E^{(k)}$ is straightforward.

Moreover, the finiteness of the algorithm directly follows from [4, 17], as stated in Proposition 1

PROPOSITION 1. *For all* $k = 1, \dots,$

$$E^{(k)} \leqslant E^{(k+1)} \leqslant \sigma.$$

Moreover for $i = 1, \ldots, k$,

$$E^{(k)}(\theta_i) = \sigma(\theta_i).$$

If BEST.VALUE *and* LBOUND *are as above, then:*

(a) $\lim_{k\uparrow\infty} \text{BEST.VALUE}(k) = \min_{\theta\in R^*} \sigma(\theta)$
(b) $\lim_{k\uparrow\infty} (\text{BEST.VALUE}(k) - \text{LBOUND}(k)) = 0$

We discuss now how to solve (5), showing that finding an optimal solution of (5) and updating $E^{(k)}$ can be done in a straightforward manner.

PROPOSITION 2. *Let* $k \geqslant i$. *The sets*

$$I^k_{i+} := \{\theta \in R^* | \, h_i(\theta) \geqslant h_s(\theta), \, \theta \geqslant \theta_i, \, \forall s = 1, \ldots, k\}$$
$$I^k_{i-} := \{\theta \in R^* | \, h_i(\theta) \geqslant h_s(\theta), \, \theta \leqslant \theta_i, \, \forall s = 1, \ldots, k\}$$

are intervals with $\theta_i \in I^k_{i+} \cap I^k_{i-}$.

Proof. First, $\theta_i \in I^k_{i+} \cap I^k_{i-}$ since $h_i(\theta_i) = \sigma(\theta_i) \geqslant E^{(k)}(\theta_i) \geqslant h_s(\theta_i)$ for all s.
Now, observe that $\sigma_1'(\theta_s; \theta - \theta_s)$ is convex in θ, with

$$\sigma_1'(\theta_s; \theta - \theta_s) = \begin{cases} \sigma_{1+}'(\theta_s)(\theta - \theta_s), & if \quad \theta \geqslant \theta_s \\ \sigma_{1-}'(\theta_s)(\theta - \theta_s), & if \quad \theta \leqslant \theta_s \end{cases},$$

σ_{1+}' and σ_{1-}' denoting the right and left derivatives of σ_1.
We have

$$\begin{aligned} &\{\theta \geqslant \theta_i | \, h_i(\theta) \geqslant E^{(k)}(\theta)\} = \\ = \quad &\{\theta \geqslant \theta_i | \, \sigma_1(\theta_i) + \sigma_1'(\theta_i; \theta - \theta_i) \geqslant \max_{1\leqslant s\leqslant k} \sigma_1(\theta_s) + \sigma_1'(\theta_s; \theta - \theta_s)\} \\ = \quad &\{\theta \geqslant \theta_i | \, \sigma_1(\theta_i) + \sigma_{1+}'(\theta_i)(\theta - \theta_i) \geqslant \max_{1\leqslant s\leqslant k} \sigma_1(\theta_s) + \sigma_1'(\theta_s; \theta - \theta_s)\} \\ = \quad &\{\theta \geqslant \theta_i | \, 0 \geqslant \max_{1\leqslant s\leqslant k} \sigma_1(\theta_s) + \sigma_1'(\theta_s; \theta - \theta_s) - \sigma_1(\theta_i) - \sigma_{1+}'(\theta_i)(\theta - \theta_i)\} \end{aligned}$$

which is an interval since it is the lower level set of a convex function. Hence $I^k_{i+} = R^* \cap \{\theta \geqslant \theta_i | \, h_i(\theta) \geqslant E^{(k)}(\theta)\}$ is an interval.
In the same manner I^k_{i-} is another interval, with θ_i as endpoint. □

From this we have

PROPOSITION 3. $E^{(k)}$ *is a piecewise concave function, having the intervals* $I^k_{1+}, I^k_{1-}, \ldots I^k_{k+}, I^k_{k-}$ *as domains of concavity.*

REMARK 1. *Proposition 2 enables us to easily update $E^{(k+1)}$ from $E^{(k)}$, since they only differ in one interval, namely $I_{k+1}^{k+1} = I_{k+1,-}^{k+1} \cup I_{k+1,+}^{k+1}$:*

$$E^{(k+1)}(\theta) = \begin{cases} E^{(k)}(\theta), & if \quad \theta \notin I_{k+1}^{k+1} \\ \sigma_1(\theta_{k+1}) + \sigma'_{1+}(\theta_{k+1})(\theta - \theta_{k+1}) - \sigma_2(\theta), & if \quad \theta \in I_{k+1,+}^{k+1} \\ \sigma_1(\theta_{k+1}) + \sigma'_{1-}(\theta_{k+1})(\theta - \theta_{k+1}) - \sigma_2(\theta), & if \quad \theta \in I_{k+1,-}^{k+1} \end{cases}$$

Hence, in order to solve (5) we have

$$\min_{\theta \in R^*} E^{(k)}(\theta) = \min_{1 \leqslant s \leqslant k} \{ \min_{\theta \in I_{s-}^k} h_s(\theta), \min_{\theta \in I_{s+}^k} h_s(\theta) \}.$$

This joint with Proposition 1 imply the following

PROPOSITION 4. *At Step k of Algorithm 1 the optimal solution θ_{k+1} of Problem (5) either coincides with θ_k (in which case it is optimal for Problem (2)), or it is an endpoint of an interval $I_i^k := I_{i-}^k \cup I_{i+}^k$.*

REMARK 2. *By Proposition 4, in order to minimize the new envelope function $E^{(k+1)}$ we only need to inspect two new points, namely the endpoints of the interval I_{k+1}^{k+1}.*

2. Constructing the d.c. decomposition

Solving Problem (2) by using Algorithm 1 requires the knowledge of a d.c. representation for the objective function. Although obtaining explicitly a decomposition for an arbitrary d.c. function may be far from trivial, this task turns out to be very simple in our case. Indeed, using the standard techniques of [12, 22], d.c. decompositions for all the non-convex functions ρ of Table 1 can be directly derived.

For instance, for Function 8 in Table 1, we have

$$\begin{aligned} \rho(t) &= \min\{\frac{t^2}{2}, \frac{1}{2}\} \\ &= \frac{t^2}{2} + \min\{0, \frac{1}{2} - \frac{t^2}{2}\} \\ &= \frac{t^2}{2} - \max\{0, \frac{t^2}{2} - \frac{1}{2}\} \end{aligned}$$

On the other hand, for Function 3 in Table 1, we have that

$$\rho''(t) = \frac{1 - t^2}{(1 + t^2)^2},$$

thus

$$\rho''(t) \geqslant \min_{0 \leqslant s} \frac{1-s}{(1+s)^2} = \frac{-1}{8},$$

yielding the d.c. decomposition of ρ

$$\rho(t) = \left(\rho(t) + \frac{1}{2}\frac{t^2}{8}\right) - \frac{1}{2}\frac{t^2}{8} \tag{9}$$

A similar analysis can be carried out for the remaining non-convex functions ρ of Table 1.

Once a d.c. decomposition of ρ,

$$\rho(t) = \alpha(t) - \beta(t)$$

is given, one immediately obtains a d.c. decomposition for ρ_j,

$$\rho_j(\theta) = \alpha\left(\frac{y_j - \theta}{C_j}\right) - \beta\left(\frac{y_j - \theta}{C_j}\right),$$

and then a d.c. decomposition for σ is given by

$$\sigma(\theta) = \sum_{1 \leqslant j \leqslant n} \alpha\left(\frac{y_j - \theta}{C_j}\right) - \sum_{1 \leqslant j \leqslant n} \beta\left(\frac{y_j - \theta}{C_j}\right)$$

For a given function ρ, more than one d.c. decomposition may be found. In spite of the fact that the choice of the decomposition influences the speed of convergence of the algorithm, see [3, 4] and Section 3, this is usually ignored in the literature. Here, good d.c. decompositions are sought, and we will prove that, under certain conditions, an optimal (in a sense described later) decomposition for this particularization of Algorithm 1 can be obtained.

This result will be used for finding an optimal d.c. decomposition for every term in (2) and, from here, a heuristic good representation for σ.

Let us consider the set $\mathcal{D}$ of real d.c. finite functions defined over an open interval of $\mathbb{R}$ that contains the closed bounded interval I where the optimization will be carried out. We recall the reader that any function $f \in \mathcal{D}$ has side derivatives and these are of bounded variation on I, [7].

The following result is given in [9]:

LEMMA 1. *Let $f \in \mathcal{D}$, $\hat{\theta} \in I$, and let $N_{\hat{\theta}}(t)$ denote the negative variation of f'_-, the left derivative of f, in the interval with endpoints $\hat{\theta}$ and t.*

Then the functions

$$\psi[\hat{\theta}](\theta) := f(\theta) + \int_{\hat{\theta}}^{\theta} N_{\hat{\theta}}(t)dt$$

$$\gamma[\hat{\theta}](\theta) := \int_{\hat{\theta}}^{\theta} N_{\hat{\theta}}(t)dt$$

are convex. In particular, $\psi[\hat{\theta}] - \gamma[\hat{\theta}]$ is a d.c. decomposition of f.

Our aim is to use the decomposition given in Lemma 1 for the functions ρ_j in Problem (2). The suitability of such a choice in Algorithm 1 is motivated in the following results.

First we show that the choice of $\hat{\theta}$ is not relevant, since we always obtain the same lower envelope.

LEMMA 2. *Let $f \in \mathcal{D}$ and $\theta_0, \hat{\theta} \in I$. Let $\psi[\hat{\theta}]$ and $\gamma[\hat{\theta}]$ be defined as in Lemma 1. Then, the function*

$$\theta \in I \longmapsto \psi[\hat{\theta}](\theta_0) + \psi[\hat{\theta}]'_-(\theta_0)(\theta - \theta_0) - \gamma[\hat{\theta}](\theta)$$

does not depend on $\hat{\theta}$.

Proof. First of all, note that ([18], Theorem 24.2):

$$\psi[\hat{\theta}]'(\theta_0, \theta - \theta_0) = \begin{cases} f'_-(\theta_0)\ (\theta - \theta_0) + N^-_{\hat{\theta}}(\theta_0)\ (\theta - \theta_0) & \text{if } \theta < \theta_0 \\ f'_+(\theta_0)\ (\theta - \theta_0) + N^+_{\hat{\theta}}(\theta_0)\ (\theta - \theta_0) & \text{if } \theta \geqslant \theta_0 \end{cases}$$

where f'_+ and f'_- are respectively the right and left derivatives of f and

$$N^-_{\hat{\theta}}(\theta_0) = \lim_{\theta \uparrow \theta_0} N_{\hat{\theta}}(\theta) \qquad N^+_{\hat{\theta}}(\theta_0) = \lim_{\theta \downarrow \theta_0} N_{\hat{\theta}}(\theta)$$

Given $\theta < \theta_0$, one has that:

$$\begin{aligned}
&\psi[\hat{\theta}](\theta_0) + \psi'_-[\hat{\theta}](\theta_0)(\theta - \theta_0) - \gamma[\hat{\theta}](\theta) \\
=& f(\theta_0) + \int_{\hat{\theta}}^{\theta_0} N_{\hat{\theta}}(\xi)d\xi + f'_-(\theta_0)\ (\theta - \theta_0) + N^-_{\hat{\theta}}(\theta_0)\ (\theta - \theta_0) \\
&- \int_{\hat{\theta}}^{\theta} N_{\hat{\theta}}(\xi)d\xi \\
=& f(\theta_0) + f'_-(\theta_0)\ (\theta - \theta_0) - \int_{\theta_0}^{\theta} N_{\hat{\theta}}(\xi)d\xi + N^-_{\hat{\theta}}(\theta_0)\ (\theta - \theta_0) \\
=& f(\theta_0) + f'_-(\theta_0)\ (\theta - \theta_0) - \int_{\theta_0}^{\theta} \{N_{\hat{\theta}}(\xi) - N^-_{\hat{\theta}}(\theta_0)\}d\xi \\
=& f(\theta_0) + f'_-(\theta_0)\ (\theta - \theta_0) - \int_{\theta_0}^{\theta} \{N_{\theta_0}(\xi) - N^-_{\theta_0}(\theta_0)\}d\xi \\
=& f(\theta_0) + f'_-(\theta_0)\ (\theta - \theta_0) + N^-_{\theta_0}(\theta_0)(\theta - \theta_0) - \int_{\theta_0}^{\theta} N_{\theta_0}(\xi)d\xi \\
=& \psi[\theta_0](\theta_0) + \psi'_-[\theta_0](\theta_0)(\theta - \theta_0) - \gamma[\theta_0](\theta)
\end{aligned}$$

since

$$\begin{aligned}
N_{\hat{\theta}}(\xi) - N^-_{\hat{\theta}}(\theta_0) &= N_{\hat{\theta}}(\xi) + N_{\hat{\theta}}(\theta_0) - N_{\hat{\theta}}(\theta_0) - N^-_{\hat{\theta}}(\theta_0) \\
&= N_{\theta_0}(\xi) + N_{\hat{\theta}}(\theta_0) - N^-_{\hat{\theta}}(\theta_0) \\
&= N_{\theta_0}(\xi) - N^-_{\theta_0}(\theta_0)
\end{aligned}$$

In the same manner, one can cope with $\theta \geqslant \theta_0$, so that the function

$$\theta \in I \longmapsto \psi[\hat{\theta}](\theta_0) + \psi[\hat{\theta}]'_-(\theta_0)(\theta - \theta_0) - \gamma[\hat{\theta}](\theta)$$

does not depend on $\hat{\theta}$. □

The following result establishes the optimal behaviour of the decomposition of Lemma 1 for applying Algorithm 1, in the following sense: given $\theta_0 \in I$, the bounding function that it provides is a pointwise majorant for the bounding function obtained from any other d.c. decomposition. In this way, the minimum value of the envelope $E^{(k)}$ provided for this d.c. decomposition will be greater or equal than the minimum of the envelope built from any other representation, so the algorithm should converge more quickly.

PROPOSITION 5. *Let $f \in \mathcal{D}$ and $\theta_0 \in I$. Let $\psi - \gamma$ be the d.c. decomposition of Lemma 1 and $\tau - \lambda$ any other d.c. decomposition of f. Then, one has for all $\theta \in I$*

$$\psi(\theta_0) + \psi'_-(\theta_0)(\theta - \theta_0) - \gamma(\theta) \geqslant \tau(\theta_0) + \tau'_-(\theta_0)(\theta - \theta_0) - \lambda(\theta).$$

Proof. Let $\theta < \theta_0$. Then, since $f = \tau - \lambda$, with τ and λ convex, f'_- is the difference of two non-decreasing functions, namely

$$f'_- = \tau'_- - \lambda'_-.$$

Making use of the properties of the bounded variation functions, [7], one has that:

$$N_{\theta_0}(\xi) \leqslant T_{\theta_0}(\xi) = \lambda'_-(\xi) - \lambda'_-(\theta_0)$$

where $N_{\theta_0}(\xi)$ denotes the negative variation of f'_- and $T_{\theta_0}(\xi)$ is the total variation of λ'_- in the interval $[\theta_0, \xi]$. Since λ is convex, integrating the previous expression between θ_0 and θ, it follows that ([18], Corollary 24.2.1):

$$\int_{\theta_0}^{\theta} N_{\theta_0}(\xi)d\xi \leqslant \lambda(\theta) - \lambda(\theta_0) - \lambda'_-(\theta_0)(\theta - \theta_0)$$

from where, since $\lambda = \tau - f$, we obtain:

$$\begin{aligned} & f(\theta_0) + f'_-(\theta_0)\,(\theta - \theta_0) - \int_{\theta_0}^{\theta} N_{\theta_0}(\xi)d\xi \\ & \geqslant \tau(\theta_0) + \tau'_-(\theta_0)(\theta - \theta_0) - \lambda(\theta) \end{aligned}$$

A similar analysis can be carried out for the case $\theta \geqslant \theta_0$, yielding the result. □

The following result establishes the optimality of a classical d.c. decomposition for certain types of functions. Indeed, the d.c. decomposition for a $\mathcal{C}^2$-function proposed in [2], is a particular case of the representation of Lemma 1 and, using Proposition 5, it is optimal for applying Algorithm 1.

PROPOSITION 6. *Let $I \subset \mathbb{R}$ be a compact interval and let $f : I \longmapsto \mathbb{R}$ be a twice continuously differentiable function. Given $\theta_0 \in I$, the d.c. decomposition*

$$f(\theta) = \tau(\theta) - \lambda(\theta)$$

with

$$\tau(\theta) = f(\theta_0) + f'(\theta_0)(\theta - \theta_0) + \int_{\theta_0}^{\theta} (\theta - \xi)[f''(\xi)]^+ d\xi$$

$$\lambda(\theta) = \int_{\theta_0}^{\theta} (\theta - \xi)[f''(\xi)]^- d\xi$$

where $[A]^+ = \max(0, A)$ and $[A]^- = \max(0, -A)$, gives the representation of Lemma 1.

Proof. The function λ in (6) can be written as follows:

$$\int_{\theta_0}^{\theta} (\theta - \xi)[f''(\xi)]^- d\xi = \int_{\theta_0}^{\theta} \Big(\int_{\xi}^{\theta} [f''(\xi)]^- \omega \Big) d\xi = \int_{\theta_0}^{\theta} \Big(\int_{\theta_0}^{\omega} [f''(\xi)]^- d\xi \Big) d\omega$$

Taking into account that $N(\omega) = \int_{\theta_0}^{\omega} [f''(\xi)]^- d\xi$ provides the negative variation of f' on the interval $[\theta_0, \omega]$, we obtain the d.c. decomposition of Lemma 1. □

To sum up, Propositions 5 and 6 provide a heuristic procedure for constructing d.c. decompositions for σ in Algorithm 1, based on the construction of an optimal decomposition for every term ρ_j.

3. Numerical Examples

In order to test the sensitivity of the procedure with respect to the d.c. decomposition chosen, we address here the particular case in which ρ is Function 3 of Table 1.

This instance has attracted the interest of many researchers in Statistics since it is, essentially, the problem of the Maximum Likelihood Estimation for the single-parameter Cauchy distribution, whose probability density function is given by

$$f(y; \theta) = \frac{1}{\pi} \frac{1}{1 + (y - \theta)^2}$$

Indeed, the log-likelihood function for a sample of n independent observations $y_1, \ldots, y_n$ becomes

$$L(\theta) = -n \log \pi - \sum_{1 \leqslant j \leqslant n} \log \left(1 + (y_j - \theta)^2\right)$$

Table 2. Sample data for Cauchy likelihood

A	3	7	12	17											
B	2	5	7	8	11	15	17	21	23	26					
C	4.1	7.7	17.5	31.4	32.7	92.4	115.3	118.3	119.0	129.6	198.6	200.7	242.5	255.0	274.7
	274.7	303.8	334.1	430.0	489.1	703.4	978.0	1656.0	1697.8	2745.6					
D	952.1	987.0	992.7	994.1	994.5	995.1	996.0	997.9	998.0	998.6	998.7	998.7	998.8	998.8	998.9
	998.9	999.0	999.1	999.3	999.3	999.3	999.4	999.4	999.5	999.7	999.8	999.8	999.9	999.9	1000.0
	1000.0	1000.0	1000.1	1000.1	1000.2	1000.3	1000.3	1000.5	1000.5	1000.6	1000.7	1001.0	1001.2	1001.3	1002.0
	1002.1	1005.2	1007.3	1011.6	1047.9										
E	74.3	141.4	188.0	298.6	327.1	364.0	439.3	550.7	582.5	583.1	598.2	605.6	606.4	702.3	714.3
	755.2	759.7	765.2	835.8	858.1	860.4	861.4	1029.3	1148.8	1162.8	1243.9	1309.9	1312.2	1347.2	1359.6
	1383.0	1414.9	1518.1	1559.2	1562.3	1594.7	1618.0	1626.1	1640.2	1737.8	1762.3	1766.6	1773.3	1850.8	1857.8
	1922.6	2072.3	2080.6	2104.7	2163.6	2243.5	2306.3	2314.1	2374.1	2375.0	2530.4	2550.4	2571.8	2579.1	2683.5
	2837.0	2855.8	2893.3	2895.6	2904.4	2959.2	3009.0	3045.3	3115.6	3138.2	3145.2	3252.3	3351.2	3447.8	3539.9
	3639.6	3670.9	3687.2	3702.8	3796.5	3824.8	3864.0	3934.3	3982.5	4007.2	4013.4	4071.3	4146.8	4151.2	4225.7
	4261.1	4273.0	4497.1	4599.1	4606.8	4607.8	4747.2	4809.4	4877.8	4905.2					

Table 3. Computational results for Cauchy likelihood

Sample	Sample Size	Interval	Starting Point	Number of iterations		
				Wingo	Breiman-Cutler	Algorithm 1
A	4	[3,17]	9.5	71	16	12
B	10	[2,26]	13.0	107	21	12
C	25	[4.1,2745.6]	242.5	1048	391	23
D	50	[952.1,1047.9]	999.5	-	20	11
E	100	[74.3,4905.2]	2332.1	-	1919	63

whose maximum can be computed by solving (2) with ρ equal to the Cauchy function and taking $w_j = C_j = 1$ for $j = 1, \ldots, n$.

This is a global optimization problem (the objective function can exhibit up to n local optima) which resolution using deterministic methods has been carried out in [23] by means of a derivative-free algorithm due to Brent (1973), and afterwards in [5], making use of a covering method.

Since ρ is a $\mathcal{C}^2$-function, we can use Proposition 6 to obtain an optimal d.c. decomposition for ρ_j, yielding $\rho_j(\theta) = \psi_j(\theta) - (\psi_j(\theta) - \rho_j(\theta))$ with

$$\psi_j(\theta) = \begin{cases} \frac{1}{2}\left(\log 2 - 1 + (y_j - \theta)\right), & \text{if } \theta < y_j - 1 \\ \frac{1}{2}\log\left(1 + (y_j - \theta)^2\right), & \text{if } y_j - 1 \leqslant \theta \leqslant y_j + 1 \\ \frac{1}{2}\left(\log 2 - 1 - (y_j - \theta)\right), & \text{if } \theta > y_j + 1 \end{cases}$$

and, using (2), a (good) d.c. decomposition for the objective function of (2) can be built from here.

Table 3 shows the number of iterations spent by Algorithm 1 and the other ones previously mentioned, for the three sets of data used by Wingo [23], and two additional sets of [4], see Table 2 and Figures 1 and 2.

The results for Wingo and Breiman-Cutler methods have been taken from [23] and [5]. Note that the function $L(\theta)$ is unimodal for data set D.

In this implementation of Algorithm 1, we have taken as initial set $R^* = [\min_{1 \leqslant j \leqslant n} y_j, \max_{1 \leqslant j \leqslant n} y_j]$, since it always contains the optimal solution, as we mentioned before.

The comparison of Algorithm 1 with Breiman-Cutler method is specially interesting, since they both have the same structure; in fact, it has been proved in [3, 4] that the latter is a particular case of the former using a specific d.c. decomposition, namely (9). This shows that the choice of the d.c. representation may strongly affect the speed of convergence of the algorithm, since each iteration in both algorithms takes approximately the same time, whereas the number of iterations in Breiman-Cutler method may be drastically higher.

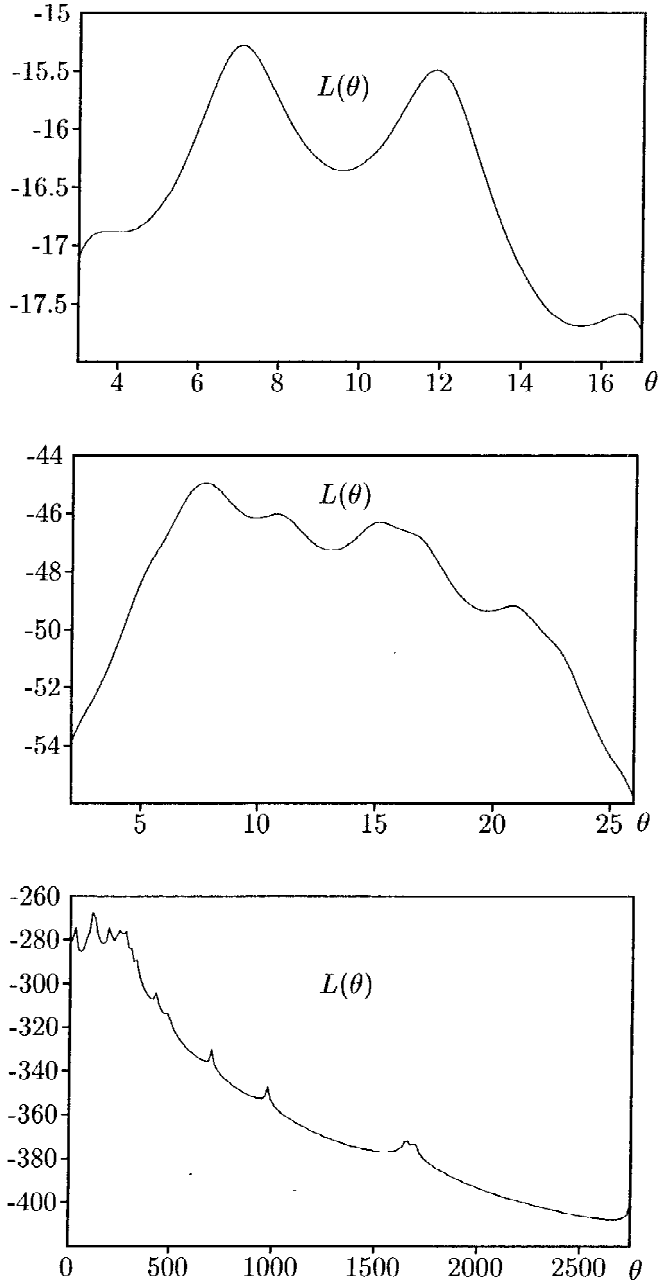

Figure 1. Log-likelihood functions for samples A, B and C.

4. Acknowledgements

This research has been supported by Grant PB96-1416-CO2-02 of DGES, Spain.

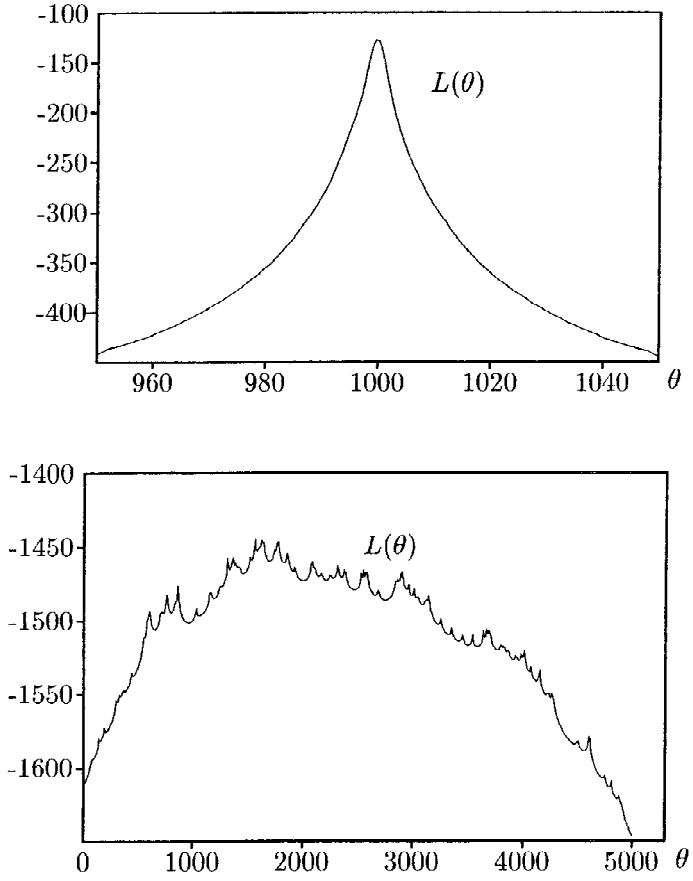

Figure 2. Log-likelihood functions for samples D and E.

References

1. Andrews, D.F. (1974), A robust method for multiple linear regression, *Technometrics* 16, 523–531.
2. Bittner, L. (1970), Some representation theorems for function and sets and their application to nonlinear programming, *Numerische Mathematik* 16, 32–51.
3. Blanquero, R. and Carrizosa, E. (2000), On covering methods for d.c. optimization, *Journal of Global Optimization* 18, 265–274.
4. Blanquero, R. (1999), *Localización de servicios en el plano mediante técnicas de optimización d.c.*, PhD. Thesis, Universidad de Sevilla.
5. Breiman, L. and Cutler, A. (1993), A deterministic algorithm for global optimization, *Mathematical Programming* 58, 179–199.
6. Brent, R.P. (1973), *Algorithms for Minimization without Derivatives*, Prentice-Hall, Englewood Cliffs, NJ.
7. De Barra, G. (1974), *Introduction to Measure Theory*, Van Nostrand Reinhold Company, New York.

8. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), *Robust statistics. The approach based on influence functions*, Wiley, New York.
9. Hartman, P. (1959), On functions representable as a difference of convex functions, *Pacific Journal of Mathematics* 9, 707–713.
10. Holland, P.W. and Welsch, R.E. (1977), Robust regression using iteratively weighted least squares, *Commun. Stat. (Theory and Methods)* 6, 813–828.
11. Horst, R., and Pardalos, P.M. (1995), *Handbook of Global Optimization*, Kluwer Academic Press, Dordrecht.
12. Horst, R., and Tuy, H. (1996), *Global Optimization. Deterministic Approaches*, Springer-Verlag, Berlin.
13. Horst, R. and Thoai, N.V. (1999), DC programming: overview, *Journal of Optimization Theory and Applications* 103, 1–43.
14. Huber, P.J. (1973), Robust regression: asymptotics, conjectures and Monte Carlo, *Annals of Statistics* 1, 799–821.
15. Huber, P.J. (1981), *Robust Statistics*, Wiley, New York.
16. Maronna, R.A. and Yohai, V.J. (1991), The breakdown point of simultaneous general M-estimates of regression and scale, *Journal of the American Statistical Association* 86, 699–703.
17. Piyavskii, S.A. (1972), An algorithm for finding the absolute extremum of a function, *USSR Computational Mathematics and Mathematical Physic* 12, 57–67.
18. Rockafellar, R.T. (1970), *Convex Analysis*, Princenton University Press, Princenton, NJ.
19. Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust regression and outlier detection*, Wiley, New York.
20. Späth, H. (1992), *Mathematical algorithms for linear regression*, Academic Press, San Diego, CA.
21. Staudte, R.G. and Sheather, S.J. (1990), *Robust estimation and testing*, Wiley, New York.
22. Tuy, H. (1998), *Convex Analysis and Global Optimization*, Kluwer Academic Press, Dordrecht.
23. Wingo, D.R. (1983), Estimating the location of the Cauchy distribution by numerical global optimization, *Communications in Statistics Part B: Simulation and Computation* 12(2), 201–212.
24. Yohai, V.J. (1987), High breakdown-point and high efficiency robust estimates for regression, *The Annals of Statistics* 15, 642–656.