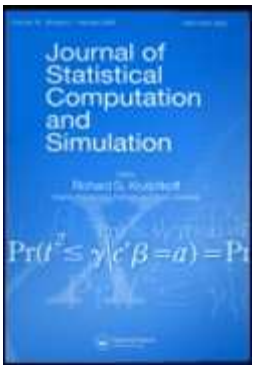
The heuristic approach in finding initial values for minimum density power divergence estimators



The Heuristic Approach in Finding Initial Values for Minimum Density Power Divergence Estimators

Journal:	*Journal of Statistical Computation and Simulation*
Manuscript ID:	GSCS-2008-0191.R1
Manuscript Type:	Original Paper
Date Submitted by the Author:	
Complete List of Authors:	Isaia, Ennio; University of Turin, Department of Statistics & Applied Mathematics Durio, Alessandra; University of Turin, Statistics & Applied Mathematics
Areas of Interest:	COMPUTATIONAL ALGORITHMS, ROBUST

Journal of Statistical Computation & Simulation
Vol. 00, No. 00, January 2008, 1–11

RESEARCH ARTICLE

The Heuristic Approach in Finding Initial Values for Minimum Density Power Divergence Estimators

A. Durio[a] and E.D. Isaia[a*]

[a] *Department of Statistics and Applied Mathematics, University of Turin, Italy*

(*25 July 2008*)

It is well known that in presence of outliers the maximum likelihood estimates are very unstable. In these situations an alternative is resorting to the estimators based on the minimum density power divergence criterion for which feasible computationally closed-form expressions can be derived, so that solutions can be achieved by any standard non linear optimization code. But since the function to be minimized is often ill-behaved, the convergence of the algorithm to optimal solutions strongly depends on the choice of the configuration of the initial values. A new procedure based on an heuristic local search approach is introduced in order to survey the parameters space and hence obtaining an accurate set of starting guesses for the gradient-method minimization routine.

Keywords: heuristic optimization; minimum density power divergence estimators; robust estimators; threshold accepting

AMS Subject Classification: 65C60; 90C59; 62G35.

1. Introduction

There is no doubt that a central role in parametric and Bayesian estimation is played by the likelihood function, although in presence of outliers its estimates are very unstable; for this reason alternative robust estimators have been proposed in literature [see for instance 1–3, and references therein]. In the following, as an alternative to the Maximum Likelihood criterion, we resort to the Minimum Density Power Divergence Estimation method, originally proposed by Basu, Harris, Hjort and Jones [4].

In many cases, for these estimators feasible computationally closed-form expressions are available, so that the solutions can be found applying any standard non linear optimization code, even if its convergence to optimal solutions strongly depends upon its initial values.

In this paper we focus the attention on the problem of the choice of starting values for the optimizing routine investigating the possibility to resort to an heuristic approach to find a subset of the parameters space yielding good initial values. More precisely, we suggest to explore the parameters surface applying an heuristic Threshold Accepting method [see 5, 6].

We illustrate the procedure examining the problem of the estimate of the parameters of Gaussian densities in presence of one cluster of outliers. Theory and

*Corresponding author. Email: isaia@econ.unito.it

ISSN: 0094-9655 print/ISSN 1563-5163 online

DOI: 10.1080/0094965YYxxxxxxxx
http://www.informaworld.com

algorithms are outlined, numerical examples are given and main results of a simulation study featuring several experimental scenarios are provided.

2. The generating model and the estimators

Before introducing our procedure, it is worthwhile to define the scenario we refer to. We consider the problem of estimating the parameters of a Gaussian density of dimension $d \geq 1$ when sample data, of size n, actually come from a mixture of two normal densities, i.e.

$$h(\mathbf{x}) \stackrel{def}{=} w_1 \, \phi_d(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + w_2 \, \phi_d(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \tag{1}$$

where the weights w_1 and w_2 are positive and sum to unit and $\phi_d(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of a d dimensional normal random variable with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

We remark that for our purposes data points drawn from the second component of equation (1) are considered belonging to a cluster of outliers of dimension $n\, w_2$, with $w_2 < w_1$. For this reason from now on we refer to equation (1) as the generating model, while we are interested in estimating the parameters of its first component, i.e. we assume that the sample data come from the Gaussian density $\phi_d(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$.

Clearly in this situation the estimates based on the Maximum Likelihood criterion are strongly influenced by the presence of outliers. As an alternative we consider the family of the Minimum Density Power Divergence Estimators.

Given the r.v. $\mathbf{X}$ of dimension $d \geq 1$ with density $\varphi(\mathbf{x}|\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0 \in \mathcal{S} \subseteq \mathrm{R}^p$ and $p \geq 1$, for which we introduce the model $f(\mathbf{x}|\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \mathcal{S}$, the density power divergence between f and φ is defined, for $\alpha > 0$, as

$$d_\alpha(f, \varphi) = \int_{\mathrm{R}^d} \left\{ f^{1+\alpha}(\mathbf{x}|\boldsymbol{\theta}) - \left(1 + \frac{1}{\alpha}\right) \varphi(\mathbf{x}|\boldsymbol{\theta}_0) f^\alpha(\mathbf{x}|\boldsymbol{\theta}) + \frac{1}{\alpha} \varphi^{1+\alpha}(\mathbf{x}|\boldsymbol{\theta}_0) \right\} d\mathbf{x},$$

while for $\alpha = 0$ it is defined as the Kullback-Leibler divergence.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from $\mathbf{X}$, the Minimum Density Power Divergence Estimator ($MDPDE$) for $\boldsymbol{\theta}_0$ is the vector $\hat{\boldsymbol{\theta}}_\alpha$ that minimize the divergence $d_\alpha(f, \varphi)$ between the probability mass function $\hat{\varphi}_n$ associated with the empirical distribution of the sample and f, i.e. for $\alpha > 0$

$$\hat{\boldsymbol{\theta}}_\alpha = \arg\min_{\boldsymbol{\theta} \in \mathcal{S}} \left[\int_{\mathrm{R}^d} f^{1+\alpha}(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} f^\alpha(\mathbf{X}_i|\boldsymbol{\theta}) \right]. \tag{2}$$

In general, as α increases, the robustness of the $MDPDE$ increases while its efficiency decreases [e.g. 4]. For $\alpha = 0$ the $MDPDE$ becomes the maximum likelihood estimator, while for $\alpha = 1$ the divergence $d_1(f, \varphi)$ is the L_2 metric and the estimator minimizes the L_2 distance between the densities [see for instance 7–9].

Since for our purposes we stated that $\mathbf{X}$ is a d dimensional Gaussian random variable with density $\phi_d(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, recalling equation (2) the estimates for $\boldsymbol{\mu}_1$ and

$\boldsymbol{\Sigma}_1$ are given by

$$\arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \left[\int_{\mathrm{R}^d} \phi_d^{1+\alpha}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\, d\mathbf{x} - \left(1+\frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} \phi_d^{\alpha}(\boldsymbol{X}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) \right] =$$

$$= \arg\min_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \left[\frac{1}{(1+\alpha)^{d/2}\,(2\pi)^{\alpha d/2}|\boldsymbol{\Sigma}|^{\alpha/2}} - \left(1+\frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} \phi_d^{\alpha}(\boldsymbol{X}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) \right], \qquad (3)$$

since $\int_{\mathrm{R}^d} \phi_d^{1+\alpha}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\, d\mathbf{x} = (1+\alpha)^{-d/2}\, \phi_d(\mathbf{0}|\mathbf{0},\boldsymbol{\Sigma})$.

In other words our problem is to minimize the function

$$g_\alpha(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(1+\alpha)^{d/2}\,(2\pi)^{\alpha d/2}|\boldsymbol{\Sigma}|^{\alpha/2}} - \left(1+\frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} \phi_d^{\alpha}(\boldsymbol{X}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}). \qquad (4)$$

For brevity we refer to function (4) simply as $g_\alpha(\boldsymbol{\theta})$ with $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\sigma}]$, where $\boldsymbol{\mu}$ is the vector of the means and $\boldsymbol{\sigma}$ is the vector containing the elements of the columns in the lower triangular part of $diag\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)\mathbf{I}_{d\times d} + \mathbf{R}$, where $\mathbf{R}$ is the correlation matrix.

Thus, according to equation (3), we have to find, for a given value of α, the absolute minimum of function $g_\alpha(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ in the parameters set $\mathcal{S}$ of dimension $p = d\,(d+3)/2$.

Since we consider situations where sample data are contaminated by a substantial number of outliers, we choose α so that the absolute minimum of $g_\alpha(\boldsymbol{\theta})$ is given by a vector close to the vector of the true parameters of the first component of the generating model.

Function (4) appears to be a feasible computationally closed-form expression and hence the estimate of the parameters of the d dimensional Gaussian density can be performed by any standard non linear optimization code. From the computational point of view, we resort to the Newton-type `nlm` minimizing routine of the `R` software. Since the surfaces described by $g_\alpha(\boldsymbol{\theta})$ are often ill-behaved, the solutions of the routine may become stuck in a local minimum, preventing the proper survey of the entire surface, hence the importance of an accurate choice of the initial guesses.

3. An heuristic approach for the choice of initial guesses

In simulation field the convergence of the algorithm to optimal solutions, i.e. the ones minimizing a given function $g(\boldsymbol{\theta})$, can be ensured (almost always) setting the vector of initial guesses equal to the vector of the true parameters. Dealing with real data, it is intuitive to resort to some data-driven mechanism to obtain a vector $\tilde{\boldsymbol{\theta}}$ of good initial values to pass to the gradient-method minimization routine. In literature many solutions to this problem, facing different situations, have been proposed [see for instance 10–12, and references therein]. In this section we introduce an alternative method based on an heuristic approach for starting guesses generation.

Heuristics optimization methods can be applied in all those situations where the objective function is not always well behaved enough to guarantee a solution to global optimality using standard gradient methods. Whereas the gradient method moves from a given point in the direction of the steepest descent of the gradient of the objective function in this point (assuming that this function is differentiable) and stops if a descent is no longer possible, iterative improvement chooses a solution from the neighborhood of a given solution which improves this solution best and

stops if the neighborhood does not contain an improving solution.

In order to find a minimum of an objective function $g : \mathcal{S} \in \mathrm{R}^p \to \mathrm{R}$, an heuristic local search method consists in moving iteratively through the solutions set $\mathcal{S}$ and choosing a new solution comparing the current solution with one somehow close to it [see 13]. Depending on the method of choosing solutions from the neighborhood of the current solution and on the way the stopping criteria are defined, different local search methods can be applied. For our purposes we resort to the Threshold Accepting method (TA) introduced by Dueck and Scheuer [14] as a variant of the simulated annealing proposed in the 50's by Metropolis *et alia* [15].

According to the Threshold Accepting method if the difference between the objective value of the chosen and the current solution is smaller than a threshold t, then TA moves to the chosen solution. The threshold t is a positive control parameter which decreases with increasing number of iterations and converges to 0. Thus, at each iteration we allow moves which do not deteriorate the current solution more than the current threshold t, and finally we only allow improving moves. In other words, we accept solutions which are worse than the previous ones in order to be able to escape local minima. Algorithm 1 resumes the main steps of the TA method for optimization problems.

Algorithm 1 (Threshold Accepting method)**:**

1. Choose a stopping criterion
2. Generate an initial solution $\mathbf{s}^c \in \mathcal{S}$
3. Choose the sequence of thresholds t_r $(r = 1, \ldots, n_R)$
4. **for** $r = 1$ **to** n_R
5. **repeat**
6. Generate $\mathbf{s}^n \in \mathcal{N}(\mathbf{s}^c)$ # *neighborhood to current solution*
7. **if** $g(\mathbf{s}^n) < g(\mathbf{s}^c) + t_r$
8. $\mathbf{s}^c = \mathbf{s}^n$ # *move to a better solution*
9. **endif**
10. **until** a stopping criterion is met
11. **endfor**

Applying the TA method, as described by Algorithm 1, requires the setting of several parameters. The stopping criterion states how many times we explore the local structure of the objective function for each threshold value; this means fixing the number (n_S) of iterations of the `repeat...until` loop. Since, in our context, the solution set $\mathcal{S} = \mathcal{S}_1 \times \ldots \times \mathcal{S}_p$ corresponds to the subset of R^p of all acceptable vectors of parameters values, we propose to choose $\mathcal{S}$ in a data-driven way, e.g. based on the sample moments of $\boldsymbol{X}$. For any given vector of current solutions $\mathbf{s}^* \in \mathcal{S}$, we define the neighborhood $\mathcal{N}(\mathbf{s}^*) = \{\mathbf{s} : |s_i - s_i^*| < \varepsilon_i, i = 1, \ldots, p\}$ with $\varepsilon_i = \gamma \cdot \mathrm{range}(\mathcal{S}_i)$, where $\gamma \in \,]\,0; 1\,[$. Finally the thresholds sequence $\{t_r\}$, for $r = 1, \ldots, n_R$, can be computed from the empirical distribution of $\Delta = |g(\mathbf{s}^1) - g(\mathbf{s}^2)|$ resorting to Algorithm 2.

Algorithm 2 (Thresholds sequence)**:**

1. **for** i $= 1$ **to** n_T
2. Generate $\mathbf{s}_i^1, \mathbf{s}_i^2 \in \mathcal{S}$
3. Compute $\Delta_i = |g(\mathbf{s}_i^1) - g(\mathbf{s}_i^2)|$
4. **endfor**
5. Compute the $n_R - 1$ quantile Q_r of order $r/(n_R - 1)$ with $r = 1, \ldots, n_R - 1$ of the λ right trimmed empirical distribution of Δ
6. Compute the thresholds sequence $\{t_r\}$, for $r = 1, \ldots, n_R - 1$, corresponding to the decreasing sorted Q_r adding $t_{n_R} = 0$

This choice of the thresholds sequence allows Algorithm 1 to accept solutions which are worse than the previous ones with a given decreasing probability.

Moreover, even if Althöfer and Koschnick [16] show that Algorithm 1 converges to a solution close to an absolute minimum with probability approaching to unit as the number of iterations grows to infinity, from a computational point of view the choice of the values of n_S, γ, n_R, n_T and λ must be made as a trade-off between the characteristics of the objective function and computing issues.

Since our goal is finding good initial guesses, i.e. in the neighborhood of the absolute minimum of $g(\boldsymbol{\theta})$, it is not so relevant for us to analyze the procedures yielding the best configuration of the TA parameters, but rather show how the heuristic approach can be useful in improving the classical methods for starting values generation. We suggest to use "raw" heuristic solutions as initial configurations for the gradient-method procedure.

In other words, for a given sample of data, we apply Algorithm 1 and use its solutions as starting guesses for the `nlm` routine. Algorithm 3 outlines this approach simply labelled as `TA + nlm`.

Algorithm 3 (`TA + nlm`)**:**

1. Set up the parameters for TA heuristic method
2. Perform the TA method according to Algorithms 1 and 2
3. Use its solutions as initial guesses for the `nlm` routine

4. Illustrative examples

In this section we provide two examples in order to show how the `TA + nlm` approach works in practice when sample data come from the generating model (1) and the $g_\alpha(\boldsymbol{\theta})$ function is given by equation (4).

We shall compare the behaviour of the `TA + nlm` procedure with a random approach, labelled `Rnd + nlm`, which consists in drawing N random sub-samples of size m from the data, computing the moments estimates and choosing as initial guesses for the `nlm` routine the vector of the sample moments which yields, among the N runs, the minimum value of the objective function $g_\alpha(\boldsymbol{\theta})$. We set the size m of the sub-samples equal to the number of the parameters to be estimated plus one, i.e. $m = p + 1$. This choice can be motivated by the intuitive idea that small sub-samples increase the randomization of the choice of starting values over the space of the parameters and $p + 1$ is the minimum number of observations needed for the moments estimates of $\boldsymbol{\theta}$ to exist.

Moreover, for a given sample of data, we estimate $B = 1000$ times the parameters of the density applying one of the $MDPDE$, and we count the percentage of times that respectively `Rnd + nlm` and `TA + nlm` approaches lead to optimal solutions, eventually dropping the degenerate ones, i.e in our simulation context we discard all the solutions for which the `nlm` routine yields a non positive definite variance-covariance matrix. In other words, for a given set of data we obtain 1000 random initial values and 1000 TA heuristic starting guesses and we count how many of these supply, through `nlm`, the optimal solutions. Obviously optimal solutions will be those corresponding to the absolute minimum of function $g_\alpha(\boldsymbol{\theta})$ which is close to the true value of the parameters of the first component of the generating model of equation (1).

Example I: we consider a situation where $n = 200$ sample data are randomly drawn from the univariate generating model (1) with $w_1 = 0.6$, $\mu_1 = 0$, $\sigma_1^2 = 1$, $w_2 = 0.4$, $\mu_2 = 8$ and $\sigma_2^2 = 1$. We perform our procedure considering the estimator defined

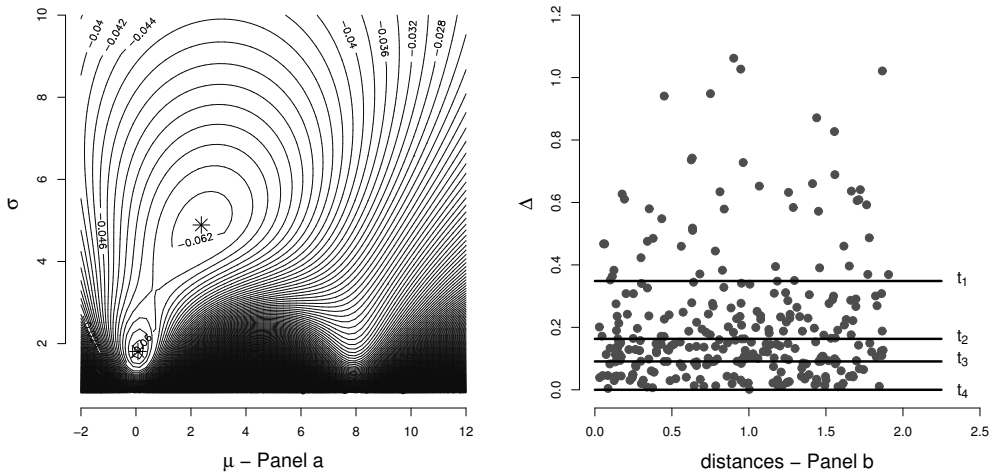


Figure 1. Panel a: contour plot of function $g_1(\boldsymbol{\theta})$ showing one absolute and one relative point of minimum. Panel b: scatter plot of points $(\delta_i; \Delta_i)$ and the thresholds t_r.

by equation (4) with $\alpha = 1$. In this situation the function $g_\alpha(\boldsymbol{\theta})$ shows one absolute point of minimum at $\boldsymbol{\theta} = [0.074, 1.808]$ and one relative point of minimum at $\boldsymbol{\theta} = [2.377, 4.890]$, as shown in Figure 1, panel (a).

In order to implement the TA method according to Algorithm 1, we roughly set $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 = [min(X); max(X)] \times [0.01; sd(X)]$, while for $\mathcal{N}(\mathbf{s}^*)$ we fix $\varepsilon_i = 0.2 \cdot \text{range}(\mathcal{S}_i)$, and furthermore we set $n_R = 4$, $n_T = 300$ and $\lambda = 0.2$.

Figure 2, panel (b), shows how the thresholds sequence is constructed. It displays the scatter plot of the n_T points $(\delta_i; \Delta_i)$, where δ_i are the euclidean distances between $\mathbf{s}_i^1$ and $\mathbf{s}_i^2$. The straight lines indicate the thresholds values t_r corresponding to the quantiles $\{Q_1, Q_{.67}, Q_{.34}, 0\}$ of the λ trimmed distribution of Δ.

Applying `TA + nlm` procedure with $n_S = 25$ we observe no degenerate solutions out of the $B = 1000$ runs and all of them, when passed as initial guesses to the `nlm` routine, supply the absolute minimum of function $g_1(\boldsymbol{\theta})$.

Recalling that each TA run consists in $n_S \times n_R = 25 \times 4 = 100$ candidates for initial guesses, to perform the `Rnd + nlm` procedure we set $N = 100$ and $m = 3$. This approach yields 995 non degenerate solutions and the 83.0% of them provide the absolute minimum. These results clearly testify the improvement we benefit resorting to `TA + nlm` procedure as an alternative to the classical `Rnd + nlm` approach. Our rough choice of the values for n_S, n_T, n_R, ε_i and λ has to be interpreted as a suitable compromise between the accuracy of `TA + nlm` solutions and computing issues.

To have an idea on how the TA algorithm performs with the tuning parameters we used, Figure 2, panel (b), shows the $B = 1000$ TA solutions that we used as initial guesses for `nlm`. Figure 2, panel (a) displays the $B = 1000$ random solutions that we passed to `nlm` as starting guesses. Only the points belonging to the left-lower cluster led us to the optimal solution. The means and the standard deviations of TA estimates for μ and σ are respectively $(-0.010, 1.852)$ and $(0.118, 0.169)$, while the corresponding means and standard deviations of the random solutions are respectively $(0.482, 2.312)$ and $(0.876, 1.171)$. If we consider only the random solutions leading to optimal solutions, i.e. the points of the left-lower cluster of Figure 2, panel (a), then their means and standard deviations become respectively $(0.093, 1.788)$ and $(0.162, 0.166)$. We furthermore remark that increasing the size

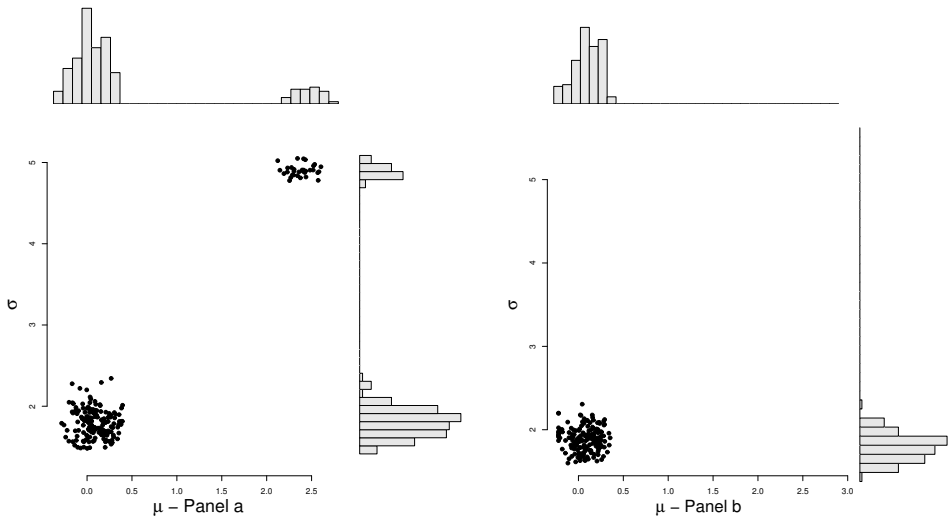


Figure 2. Scatter plots of the initial guesses for the `nlm` routine generated by the `Rnd` method (panel a) and by the heuristic `TA` approach (panel b).

m of the random sub-samples does not improve the goodness of the random initial guesses generation process. We found that setting for instance $m = 6$ the number of the points belonging to the right-upper cluster of Figure 2, panel (a), increases lowering thus the percentage of good solutions of `Rnd + nlm` procedure to an unsatisfactory 26%.

Even if it is beyond the extent of this paper, we observe that optimal solutions can certainly be reached applying only Algorithms 1 and 2 carefully tuning their parameters. This implies an accurate study of the solutions set, however rather simple in the situation we are describing. For our data sample, if we set $\mathcal{S}_1 = [-2; 10]$, $\mathcal{S}_2 = [0.01; 3]$, $\varepsilon_1 = 0.15$, $\varepsilon_2 = 0.65$, $\lambda = 0.20$, $n_R = 4$, $n_T = 500$ and $n_S = 1200$, we obtain $B = 1000$ TA solutions that supply the absolute minimum, with means $(0.074, 1.808)$ equal to the absolute minimum of $g_1(\boldsymbol{\theta})$ and standard deviations $(0.007, 0.005)$ which are acceptably low.

Example II: we consider now a more complex situation where $n = 200$ sample data are randomly drawn from the generating model (1) with $w_1 = 0.75$, $w_2 = 0.25$ and

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 5 \\ 4 \\ 2 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.12 & 0.10 & 0.02 & 0.01 \\ 0.10 & 0.14 & 0.01 & 0.01 \\ 0.02 & 0.01 & 0.03 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 \end{bmatrix} \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 6 \\ 3 \\ 5 \\ 3 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.30 & 0.09 & 0.18 & 0.07 \\ 0.09 & 0.10 & 0.08 & 0.05 \\ 0.18 & 0.08 & 0.20 & 0.08 \\ 0.07 & 0.05 & 0.08 & 0.05 \end{bmatrix}$$

In this way we reproduce the subset of Fisher's iris data [17] for the two species of iris, namely Setosa and Versicolor, but modifying the weights of the original clusters.

We observe that, according to our notation, the $\boldsymbol{\theta}$ vector of the first component of model (1) corresponds to

$$\begin{aligned} \boldsymbol{\theta} &= [\mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ \sigma_1,\ \sigma_2,\ \sigma_3,\ \sigma_4,\ \rho_{21},\ \rho_{31},\ \rho_{41},\ \rho_{32},\ \rho_{42},\ \rho_{43}] \\ &= [5, 4, 2, 0, 0.35, 0.37, 0.17, 0.10, 0.77, 0.33, 0.29, 0.15, 0.27, 0.58]\,. \end{aligned}$$

We perform our procedure considering the estimator defined by equation (4) with $\alpha = 0.2$. Choosing values $\alpha < 0.2$ the absolute minimum of $g_\alpha(\boldsymbol{\theta})$ is given by a

vector which is far from the one of the first component of the generating model. On the other hand, as α increases from 0.2 the absolute point of minimum becomes more identifiable so that the choice of good initial guess is less critical. For our generated sample, the absolute minimum of $g_{0.2}(\boldsymbol{\theta})$ is -4.55 given by the vector

$$\boldsymbol{\theta}^* = [5.03, 4.04, 2.01, 0.01, 0.42, 0.44, 0.17, 0.10, 0.83, 0.39, 0.26, 0.29, 0.27, 0.58] .$$

For TA initial guesses generation, we consider the set $\mathcal{S} = \mathcal{S}_1 \times \ldots \times \mathcal{S}_{14}$, where $\mathcal{S}_{j=1,\ldots,4} = [min(X_i); max(X_i)]_{i=1,\ldots,4}$, $\mathcal{S}_{j=5,\ldots,8} = [0.01; sd(X_i)]_{i=1,\ldots,4}$, and $\mathcal{S}_{j=9,\ldots,14} = [\min(r_{i,l} \mp 0.8, \mp 1)]_{i<l}$ with $r_{i,l} = Cov(X_i, X_l)/sd(X_i)\, sd(X_l)$.
We remark that on the vectors $\mathbf{s} \in \mathcal{S}$ of the Generate step of Algorithms 1 and 2 we impose the condition that the $s_{j=5,\ldots,14}$ lead to a positive definite variance-covariance matrix.
We furthermore decide to keep unchanged (except for n_S) the remaining settings, i.e. we fix $\varepsilon_i = 0.2 \cdot \text{range}(\mathcal{S}_j)$, $n_R = 4$, $n_T = 300$ and $\lambda = 0.2$. Given the high dimension ($p = 14$) of the solution set, for all the $B = 1000$ solutions of Algorithm 3 provide the vectors corresponding to the absolute point of minimum, it is necessary to increase at $n_S = 300$ the number of times we explore the local structure of the objective function.
In this situation `TA + nlm` approach shows still an improvement on `Rnd + nlm` procedure (with $N = 1200$ and $m = 15$), which hits the absolute minimum only 817 times out of 960 non degenerate runs (i.e. 85.1%).
It is interesting to remark that clearly the setting we chose for the TA algorithm yields solutions to be variable over the $B = 1000$ runs. In fact, while the means of the TA estimates for $\boldsymbol{\theta}$ are

$$[5.07, 4.00, 2.31, 0.50, 0.44, 0.44, 0.58, 0.30, 0.24, 0.04, 0.04, 0.15, 0.09, 0.19] ,$$

their standard deviations are

$$(0.15, 0.15, 0.06, 0.25, 0.06, 0.06, 0.19, 0.18, 0.29, 0.36, 0.37, 0.37, 0.33, 0.33) .$$

5. Some results from a simulation study

In this section we provide some of the results of a simulation study we set up to check the goodness of the `TA + nlm` procedure described by Algorithm 3. To this aim, we consider some different experimental configurations given by a specified generating model of equation (1), with $d = 1, 2$.

For each scenario, for which function $g_\alpha(\boldsymbol{\theta})$ of equation (4) shows, for a given value of α, one absolute and one relative point of minimum, we draw $H = 100$ random samples of size n and on each of them we compute $B = 100$ times the $MDPDE$ estimates of the parameters of the first component of the generating model (1) resorting to `Rnd + nlm` and `TA + nlm` algorithms.

For each of the 100 samples, we record the percentage of times that `nlm` routine initialized with random and heuristic initial guesses detects the absolute point of minimum (`AM`) and the number of non degenerate solutions out of the $B = 100$ runs (`ND`). The results of the simulation are recorded as the mean and the standard deviation of `AM` and we also indicate for completeness the mean percentage of `ND`.

Scenario I: we consider a simple situation where $n = 250$ sample data are randomly drawn from the univariate generating model with $w_1 = 0.60$, $\mu_1 = 0$, $\sigma_1^2 = 1$, $w_2 = 0.40$, $\mu_2 = \delta$ and $\sigma_2^2 = 0.25$, where the mean of the second component, i.e.

Table 1. Scenario I: mean percentage of non degenerate solutions (`mean(ND)`), mean (`mean(AM)`) and standard deviation (`sd(AM)`) of the percentage of times that `nlm` detects the absolute point of minimum.

		$\alpha = 0.8$		$\alpha = 0.9$		$\alpha = 1$	
		`Rnd+nlm`	`TA+nlm`	`Rnd+nlm`	`TA+nlm`	`Rnd+nlm`	`TA+nlm`
$\delta = 8$	`mean(ND)`	99.3	100.0	98.9	100.0	97.8	100.0
	`mean(AM)`	87.2	97.9	90.3	99.8	91.9	100.0
	`sd(ND)`	4.5	0.8	3.2	0.9	3.9	0
$\delta = 10$	`mean(ND)`	99.1	100.0	99.3	100.0	98.5	100.0
	`mean(AM)`	81.3	100.0	88.4	100.0	90.7	100.0
	`sd(ND)`	3.8	0	4.2	0	4.1	0

the one describing the distribution of the outliers, is set to $\delta = 8, 10$. We analyze this situation resorting to three different $MDPDE$ obtained setting $\alpha = 0.8, 0.9, 1$. The settings for heuristic and random initial guesses generation are the same as those outlined in example I.

Table 1 shows the results of the simulation applying `Rnd + nlm` and `TA + nlm` approaches. If we choose at random the starting guesses, then as δ moves from 8 to 10, i.e. the cluster of outliers goes far from the first component of the generating model, `Rnd + nlm` procedure gives somewhat worse results while `TA + nlm` remains substantially stable.

The results of Table 1 allow us to state that the heuristic method performs better than the random one for all the experimental configurations, this in terms of percentage of non degenerate solutions (`mean(ND)`) and in terms of percentage of times that `nlm` detects the absolute minimum (`mean(ND)`).

Scenario II: we inspect a situation where $n = 250$ sample data are randomly drawn from the bivariate generating model with

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 7 \\ 7 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and where the weight of the first component are respectively $w_1 = 0.55$ and $w_1 = 0.60$. For this scenario we consider the three different $MDPDE$ given by setting $\alpha = 0.8, 0.9, 1$.

The settings for heuristic and random initial guesses generation are the same as those outlined in example II.

Table 2 records the results of the simulation resorting to `Rnd + nlm` and `TA + nlm` approaches.

For each estimator and for all the experimental configurations, it's clear the advantage we benefit from `TA + nlm` procedure, in fact the corresponding `mean(AM)` values are systematically greater than those we observe resorting to `Rnd + nlm`. Moreover, the quite large values of `sd(AM)` denote in some measure a sample dependence of the `Rnd + nlm` procedure .

The behaviour of $MDPDE$ is well outlined if we consider the rows of Table 2 for a fixed value of w_1. As α increases the estimator becomes more robust. Moreover, if we examine the columns of Table 2, it is clear that, when the components of the generating model begin to be heavier, the performance of `Rnd + nlm` improves and this independently of the value of α.

Table 2. Scenario II: mean percentage of non degenerate solutions (`mean(ND)`), mean (`mean(AM)`) and standard deviation (`sd(AM)`) of the percentage of times that `nlm` detects the absolute point of minimum.

		$\alpha = 0.8$		$\alpha = 0.9$		$\alpha = 1$	
		`Rnd+nlm`	`TA+nlm`	`Rnd+nlm`	`TA+nlm`	`Rnd+nlm`	`TA+nlm`
$w_1 = 0.55$	`mean(ND)`	98.8	99.6	99.5	99.9	100.0	100.0
	`mean(AM)`	31.5	92.3	54.6	96.4	72.2	98.3
	`sd(AM)`	5.7	2.0	6.1	2.6	6.9	2.3
$w_1 = 0.60$	`mean(ND)`	99.2	100.0	100.0	100.0	100.0	100.0
	`mean(AM)`	63.4	98.7	69.9	100.0	82.6	100.0
	`sd(AM)`	3.9	1.3	4.4	0.0	4.7	0.0

6. Conclusions and final remarks

In presence of outliers robust estimates of the parameters of a multivariate Gaussian density can proficiently be obtained resorting to the estimators based on the Minimum Density Power Divergence criterion, for which computationally closed-form expressions are available so that solutions can be obtained applying any standard non linear optimization code. But Since the function to be minimized is in most cases ill-behaved, the convergence of the algorithm to optimal solutions strongly depends on the choice of the configuration of the initial guesses.

Usually starting values for gradient based routines are achieved applying (more or less sophisticated) data-driven random algorithms. In this paper we suggest to exploit the method of an heuristic local search to identify an optimal subset of the parameter space yielding good initial values for the optimization code. It seems to the authors that this approach in exploring the parameters surface behaves very well featuring a valuable alternative to a basic random technique and this primarily in terms of parsimony of computing time.

To be certain that the performance of the proposed method remains the same when considering less basic situations a lot of work must still be done. For instance we could introduce more components to the generating model (1) and then check our procedure in finding good initial guesses for `nlm` routine when considering $MDPD$ estimators for a mixture of densities. In this case the results of the heuristic approach could be compared with the one yielded by more opponent strategies, such as EM algorithm, Kmeans procedure or other similar techniques.

Acknowledgements

Authors are indebted to the coordinating editor and to the anonymous referees for carefully reading the manuscript and for their many important remarks and suggestions.

References

[1] P.J. Huber *Robust Statistics*, John Wiley, New York, 1981.
[2] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel *Robust Statistics : The Approach Based on Influence Functions*, John Wiley, 1986.
[3] M. Markatou and E. Ronchetti, *Robust Inference: the Approach Based on Influence Functions*, in *Robust Inference*, G.S. Maddala and C.R. Rao, eds., North-Holland, 1997, pp. 49–75.
[4] A. Basu, I.R. Harris, N. Hjort, and M. Jones, *Robust and Efficient Estimation by Minimizing a Density Power Divergence*, Biometrika 85 (1998), pp. 549–559.
[5] M. Gilli, E. Këllezi, and H. Hysi, *A Data-Driven Optimization Heuristic for Downside Risk Minimization*, Journal of Risks 8 (2006), pp. 1–19.
[6] P. Winker *Optimization Heuristics in Econometrics*, John Wiley, 2001.

[7] G.R. Terrell, *Linear Density Estimates*, Proceedings of the Statistical Computing Section, American Statistical Association (1990), pp. 297–302.
[8] D.W. Scott, *Parametric Statistical Modeling by Minimum Integrated Square Error*, Technometrics 43 (2001), pp. 274–285.
[9] A. Durio and E.D. Isaia, *On Robustness to Outliers of Parametric L_2 Estimate Criterion in the Case of Bivariate Normal Mixtures: a Simulation Study*, in *Theory and Applications of Recent Robust Methods*, A.S. M. Hubert G. Pison and S.V. Aelst, eds., Birkhäuser, 2004, pp. 93–104.
[10] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society 39 (1977), pp. 1–22.
[11] G. McLachlan, *On the choice of initial values for the EM algorithm in fitting mixture models*, The Statistician 37 (1988), pp. 417–425.
[12] C. Biernacki, G. Celeux, and G. Govaert, *Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models*, Computational Statistics & Data Analysis 41 (2003), pp. 561–575.
[13] C.H. Papadimitriou and K. Steiglitz *Combinatorial Optimization : Algorithms and Complexity*, Dover Publications, 1998.
[14] G. Dueck and T. Scheuer, *Threshold Accepting: a General Purpose Optimization Algorothm*, Journal of Computational Physics 90 (1990), pp. 161–175.
[15] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, *Equations of State Calculations by Fast Computing Machines*, Journal of Chemical Physics 21 (1953), pp. 1087–1092.
[16] I. Althöfer and K.U. Koschnick, *On the convergence of Threshold Accepting*, Applied Mathematics and Optimization 24 (1991), pp. 183–195.
[17] R.A. Fisher, *The use of multiple measurements in taxonomic problems*, Annals of Eugenics 7 (1936), pp. 179–188.