AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 3&4, 301–317

Redescending M-estimators and Deterministic Annealing, with Applications to Robust Regression and Tail Index Estimation

Rudolf Frühwirth and Wolfgang Waltenberger
Austrian Academy of Sciences, Vienna, Austria

Abstract: A new type of redescending M-estimators is constructed, based on data augmentation with an unspecified outlier model. Necessary and sufficient conditions for the convergence of the resulting estimators to the Huber-type skipped mean are derived. By introducing a temperature parameter the concept of deterministic annealing can be applied, making the estimator insensitive to the starting point of the iteration. The properties of the annealing M-estimator as a function of the temperature are explored. Finally, two applications are presented. The first one is the robust estimation of interaction vertices in experimental particle physics, including outlier detection. The second one is the estimation of the tail index of a distribution from a sample using robust regression diagnostics.

Zusammenfassung: Ein neuer Typ von wiederabsteigenden M-Schätzern wird konstruiert, ausgehend von Datenerweiterung mit einem unspezifizierten Ausreißermodell. Notwendige und hinreichende Bedingungen für die Konvergenz zu Hubers "Skipped-mean"-Schätzer werden angegeben. Durch Einführung einer Temperatur kann die Methode des "Deterministic Annealing" angewendet werden. Der Schätzer wird dadurch unempfindlich gegen die Wahl des Anfangspunkts der Iteration. Die Eigenschaften des Schätzers als Funktion der Temperatur werden untersucht. Schließlich werden zwei Anwendungen vorgestellt. Die erste ist die robuste Schätzung von Wechselwirkungspunkten in der experimentellen Teilchenphysik, einschließlich der Erkennung von Ausreißern. Die zweite ist die Schätzung des "Tail index" einer Verteilung aus einer Stichprobe mittels robuster Regressionsdiagnostik.

Keywords: Regression Diagnostics.

1 Introduction

M-estimators were first introduced by Huber (2004) as robust estimators of location and scale. Their study in terms of the influence function was undertaken by Hampel and co-workers (Hampel et al., 1986). Redescending M-estimators are a special class of M-estimators. They are widely used for robust regression and regression clustering, see e.g. Müller (2004) and the references therein. According to the definition in Hampel et al. (1986), the ψ-function of a redescending M-estimators has to disappear outside a certain central interval. Here, we merely demand that the ψ-function tends to zero for $|x| \to \infty$. If ψ tends to zero sufficiently fast, observations lying farther away than a certain bound are effectively discarded. Redescending M-estimators are thus particularly resistant to

extreme outliers, but their computation is afflicted with the problem of local minima and a resulting dependence on the starting point of the iteration.

The problem of convergence to a local minimum can be cured by combining the iterative computation of the M-estimate with a global optimization technique, namely deterministic annealing. For a review of deterministic annealing and its applications to clustering, classification, regression and related problems see Rose (1998) and the references therein. To the best of our knowledge, the combination of M-estimators with deterministic annealing has been proposed only by Li (1996). It will be shown below, however, that his annealing M-estimators have infinite asymptotic variance at low temperature, a feature that we deem to be undesirable in certain applications.

The purpose of this note is to construct a new type of redescending M-estimators with annealing that converge to the Huber-type skipped mean Hampel et al. (1986) if the temperature T approaches zero. The starting point is a mixture model of data and outliers. Data augmentation is used to formulate an EM algorithm for the estimation of the unknown location of the data. The EM algorithm is then interpreted as a redescending M-estimator that can be combined with deterministic annealing in a natural way (Subsections 2.1 and 2.2). The most important case is a normal model for the data, but other models are possible. In Subsection 2.3 conditions are derived under which the corresponding M-estimator converges to the skipped mean. Section 3 explores the properties of the annealing M-estimator with a normal data model and illustrates the effect of deterministic annealing on a simple example with synthetic data. Section 4 presents two applications of the annealing M-estimator: first, robust regression applied to the problem of estimating an interaction vertex in experimental particle physics; and second, regression diagnostics applied to the problem of estimating the tail index of a distribution from a sample.

2 Redescending M-estimators with Deterministic Annealing

This section shows how a new type of redescending M-estimators can be constructed via data augmentation. The estimators are then generalized by introducing a temperature parameter so that deterministic annealing can be applied. The case of data models other than the normal one is discussed, and conditions for convergence to the skipped mean are derived.

2.1 Construction via Data Augmentation

The starting point is a simple mixture model of data with outliers, with the p.d.f.

$$h(x) = p \cdot f(x; \mu, \sigma) + (1 - p) \cdot g(x)\,. \tag{1}$$

By assumption, $f(x; \mu, \sigma)$ is the p.d.f. of the normal distribution with location μ and scale σ and can be written as

$$f(x; \mu, \sigma) = \varphi(r)\,, \qquad \text{with} \qquad r = (x - \mu)/\sigma\,,$$

$\varphi(\cdot)$ being the standard normal density. The distribution of the outliers, characterized by the density $g(x)$, is left unspecified.

Now let $(x_1, \ldots, x_n)$ be a sample of size n from the model in (1). The sample is augmented by a set of indicator variables I_j, $j = 1, \ldots, n$, where $I_j = 0\,(1)$ indicates that x_j is an inlier (outlier). If the scale σ is known, the location can be estimated by the EM algorithm (Dempster, Laird, and Rubin, 1977). In this particular case, the EM algorithm is an iterated re-weighted least-squares estimator, the weight of the observation x_j being equal to the posterior probability that it is an inlier. The latter is given by Bayes' theorem:

$$P(I_j = 0|x_j) = \frac{P(x_j|I_j = 0) \cdot P(I_j = 0)}{P(x_j|I_j = 0) \cdot P(I_j = 0) + P(x_j|I_j = 1) \cdot P(I_j = 1)}\,. \tag{2}$$

As we do not wish to specify the outlier distribution, we resort to a worst case scenario and set $P(I_j = 0) = P(I_j = 1) = 1/2$. In addition we require that in the vicinity of μ, an observation should be an inlier rather than an outlier, so $P(I_j = 1|x) \leq P(I_j = 0|x)$ for $|x - \mu|/\sigma \leq c$, $c > 0$, where c is a cutoff parameter. This can be achieved by setting the prior probability $P(x_j|I_j = 1)$ that x_j is an outlier to $P(\mu + c\sigma|I_j = 0) = \varphi(c)$. The posterior probability $P(I_j = 0|x_j)$ then reads:

$$P(I_j = 0|x_j) = \frac{f(x_j; \mu, \sigma)}{f(x_j; \mu, \sigma) + \varphi(c)} = \frac{\varphi(r_j)}{\varphi(r_j) + \varphi(c)}\,, \tag{3}$$

where $r_j = (x_j - \mu)/\sigma$. If $r_j = c$, the posterior probabilities of x_j being an inlier or an outlier, respectively, are the same.

If the inlier density is normal the EM algorithm is tantamount to an iterated weighted mean of the observations:

$$\begin{aligned} \mu^{(k+1)} &= \sum_{j=1}^{n} w_j^{(k)} x_j \Big/ \sum_{j=1}^{n} w_j^{(k)}\,, \qquad \text{with} \\ w_j^{(k)} &= \frac{\varphi(r_j^{(k)})}{\varphi(r_j^{(k)}) + \varphi(c)}\,, \qquad \text{and} \\ r_j^{(k)} &= (x_j - \mu^{(k)})/\sigma\,. \end{aligned}$$

The iterated weighted mean can also be interpreted as an M-estimator of location (Huber, 2004), with

$$\psi(r; c) = \frac{r\varphi(r)}{\varphi(r) + \varphi(c)}\,, \qquad \text{and} \qquad \rho(r; c) = \int \psi(r; c)\,\mathrm{d}r\,.$$

This interpretation allows us to analyze the estimator in terms of its influence function and associated concepts such as gross-error sensitivity and rejection point.

2.2 Introducing a Temperature

The shape of the score function $\rho(r; c)$ can be modified by introducing a temperature parameter T into the weights. This allows to improve global convergence by using the

technique of deterministic annealing (Rose, 1998; Li, 1996). The modified weights are defined by

$$w(r;c,T) = \frac{\varphi(r/\sqrt{T})}{\varphi(r/\sqrt{T}) + \varphi(c/\sqrt{T})} = \frac{\exp(-r^2/2T)}{\exp(-r^2/2T) + \exp(-c^2/2T)}\,. \tag{4}$$

The redescending M-estimator with this weight function is called a normal-type or N-type M-estimator. Its ψ-function is given by

$$\psi(r;c,T) = \frac{r\exp(-r^2/2T)}{\exp(-r^2/2T) + \exp(-c^2/2T)}\,,$$

and its ρ-function by

$$\rho(r;c,T) = \frac{r^2}{2} - T\log\left(\exp(r^2/2T) + \exp(c^2/2T)\right) + T\log\left(1+\exp(c^2/2T)\right)\,. \tag{5}$$

Figure 1 shows the weight function, the ψ-function and the ρ-function of the N-type M-estimator for three different temperatures ($T = 10, 1, 0.01$). Note that (5) is not suitable for the numerical computation of $\rho(r;c,T)$ if T is very small. A numerically stable version of (5) is given by

$$\rho(r;c,T) = \begin{cases} \dfrac{r^2}{2} + T\log\dfrac{1+\exp(-c^2/2T)}{1+\exp((r^2-c^2)/2T)}\,, & \text{if } |r| < c, \\ \dfrac{c^2}{2} + T\log\dfrac{1+\exp(-c^2/2T)}{1+\exp((c^2-r^2)/2T)}\,, & \text{if } |r| > c. \end{cases}$$

If the temperature increases, the weight drops more slowly as a function of r. In the limit of infinite temperature we have

$$\lim_{T\to\infty} w(r;c,T) = \frac{1}{2}$$

for all c, and the M-estimator degenerates into a least-squares estimator. If the temperature drops to zero, the weight function converges to a step function.

Proposition 1 *Let $H(\cdot)$ be the unit step function (Heaviside function) with the additional convention $H(0) = 1/2$. Then*

$$\lim_{T\to 0} w(r;c,T) = H(c-r)\,.$$

Proposition 1 follows from the more general Proposition 2 below.

2.3 Non-normal Data models

The density φ used in defining the weight function in (4) need not be the standard normal density. In fact, every unimodal continuous symmetric density $f(x)$ with location 0, scale 1 and infinite range generates a type of redescending M-estimators. The behaviour of the weight function $w_f(r;c,T)$ at low temperature is determined by the tail behaviour of

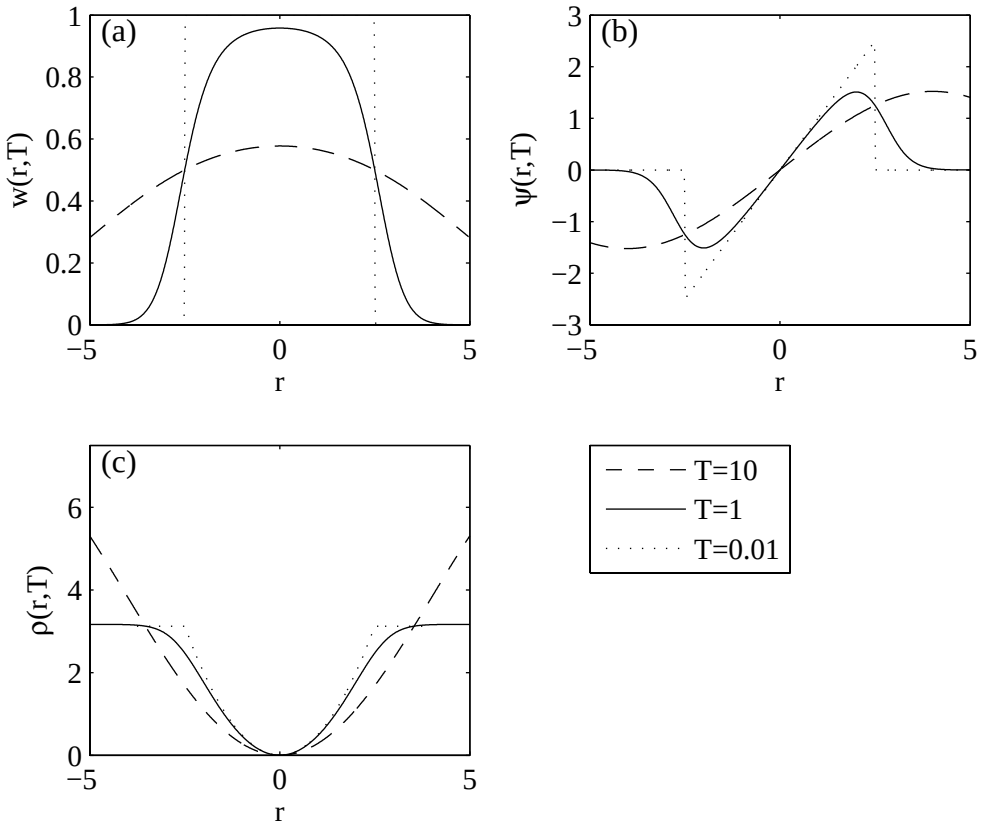

Figure 1: N-type M-estimator: a) $w(r;c,T)$, (b) $\psi(r;c,T)$ and (c) $\rho(r;c,T)$ for $c = 2.5$ and $T = 10, 1, 0.01$.

$f(x)$, as described by the concept of regular variation at infinity (Seneta, 1976). We recall that a function $f(x) : [0,\infty) \to (0,\infty)$ is called regularly varying at infinity with index $\xi \in \mathbb{R}$ if it satisfies

$$\lim_{x\to\infty} \frac{f(\lambda x)}{f(x)} = \lambda^{\xi} \qquad \text{for any } \lambda > 0.$$

If $f(x)$ is a probability density function, ξ has to be in the interval $(-\infty, -1)$. The definition can be extended in the obvious sense to $\xi = -\infty$. If $\xi = -\infty$,

$$\lim_{x\to\infty} \frac{f(\lambda x)}{f(x)} = \begin{cases} 0 & \text{for } \lambda > 1, \\ \infty & \text{for } \lambda < 1. \end{cases}$$

In this case $f(x)$ is also called rapidly varying at infinity (Seneta, 1976). Normal densities are rapidly varying at infinity, as are all densities with exponential tails.

Proposition 2

(a) Let $H(\cdot)$ be as in Proposition 1. $f(r)$ is rapidly varying at infinity if and only if

$$\lim_{T\to 0} w_f(r;c,T) = H(c-r)$$

for all $c > 0$.

(b) $f(r)$ is regularly varying at infinity with index $\xi \in \mathbb{R}$ if and only if

$$\lim_{T\to 0} w_f(r; c, T) = \frac{r^\xi}{r^\xi + c^\xi} = \frac{c^{-\xi}}{r^{-\xi} + c^{-\xi}}$$

for all $c > 0$. □

The proof is omitted, but can be obtained from the authors on request.

Example 1 (The Hyperbolic Secant Distribution)
The hyperbolic secant distribution is a symmetric distribution with exponential tails. The standardized density is equal to

$$h(r) = \frac{1}{2\cosh(r\pi/2)}\,.$$

The ψ-function of the corresponding (HS-type) redescending M-estimator is shown in Figure 2(a), for three different temperatures ($T = 10, 1, 0.01$). It is easy to show that $h(r)$ is rapidly varying at infinity. According to Proposition 2, the weight function $w_h(r; c, T)$ converges to $H(c - r)$ for $T \to 0$, and the corresponding M-estimator approaches the skipped mean. □

Example 2 (Student's t-Distribution)
Student's t-distribution is a symmetric distribution with tails falling off according to a power law. The standardized density with $\nu > 2$ degrees of freedom is equal to

$$t_\nu(r) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi(\nu-2)}\,\Gamma(\nu/2)} \left(1 + \frac{r^2}{\nu-2}\right)^{-(\nu+1)/2}\,.$$

The ψ-function of the corresponding (t_ν-type) redescending M-estimator with $\nu = 3$ is shown in Figure 2(b), for three different temperatures ($T = 10, 1, 0.01$). The density $t_\nu(r)$ is regularly varying at infinity with index $\xi = -(\nu + 1)$. From Proposition 2 follows:

$$\lim_{T\to 0} w_{t_\nu}(r; c, T) = \frac{c^{\nu+1}}{c^{\nu+1} + r^{\nu+1}}\,.$$

For $\nu \to \infty$, this function approaches the step function $H(c - r)$. □

3 N-type M-estimators of Location

In this section the properties of the annealing M-estimator with a normal data model are explored. The effect of deterministic annealing on the objective function of the estimator is illustrated on a simple example with two clusters (data and outliers).

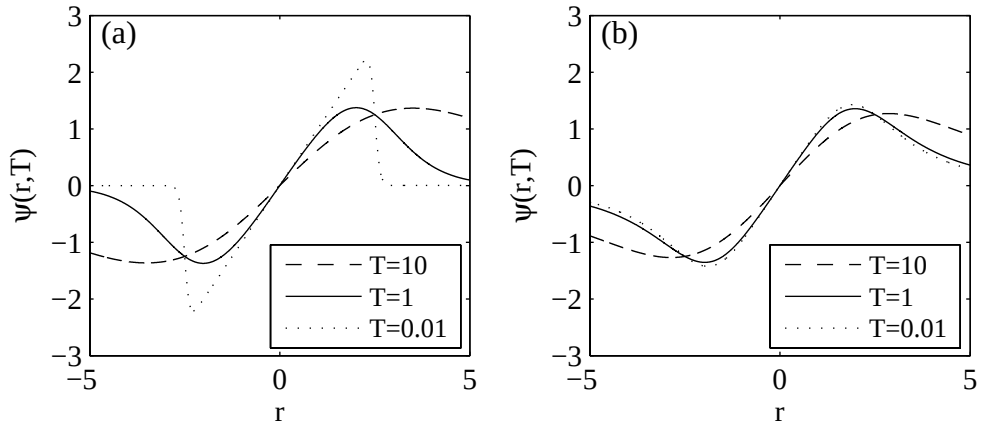

Figure 2: $\psi(r;c,T)$ of redescending M-estimators of (a) hyperbolic secant-type and (b) Student's t_3-type , for $c = 2.5$ and $T = 10, 1, 0.01$.

3.1 Basic Properties

The influence function is always proportional to ψ:

$$\mathrm{IF}(r;\psi(r;c,T),F) = \psi(r;c,T)/K(c,T)\,.$$

If the model distribution is the standard normal distribution, $K(c,T)$ is given by (Hampel et al., 1986):

$$K(c,T) = \int_{\mathbb{R}} r\,\psi(r;c,T)\,\varphi(r)\,\mathrm{d}r = 2\int_0^\infty r\,\psi(r;c,T)\,\varphi(r)\,\mathrm{d}r\,.$$

Unfortunately, the integral cannot be written in closed form. Figure 3(a) shows $K(c,T)$ as a function of T, for $c = 1.5 : 0.5 : 3$, computed by numerical integration. Complications at very small values of T can be avoided by splitting the interval of integration $[0,\infty)$ at c. The low- and high-temperature limits can be computed explicitly:

$$\lim_{T\to 0} K(c,T) = 2\Phi(c) - 1 - 2c\varphi(c)\,, \qquad \lim_{T\to\infty} K(c,T) = \frac{1}{2}\,.$$

The point of maximum influence can be computed by means of the Lambert W-function (Coreless et al., 1996):

$$r_{\max}(c,T) = \sqrt{2T\omega(c,T)+T}\,, \qquad \text{with} \qquad \omega(c,T) = W\left(\tfrac{1}{2}\exp(c^2/2T - 1/2)\right)\,.$$

$r_{\max}$ is shown in Figure 3(b). The maximum value of the influence function is the gross-error sensitivity γ^*:

$$\gamma^*(c,T) = \max_r \mathrm{IF}(r;c,T) = \frac{\psi(r_{\max}(c,T);c,T)}{K(c,T)} = \frac{1}{K(c,T)}\frac{2\sqrt{T}\omega(c,T)}{\sqrt{2}\omega(c,T)+1}.$$

Figure 3(c) shows the gross-error sensitivity as a function of T, for $c = 1.5 : 0.5 : 3$.

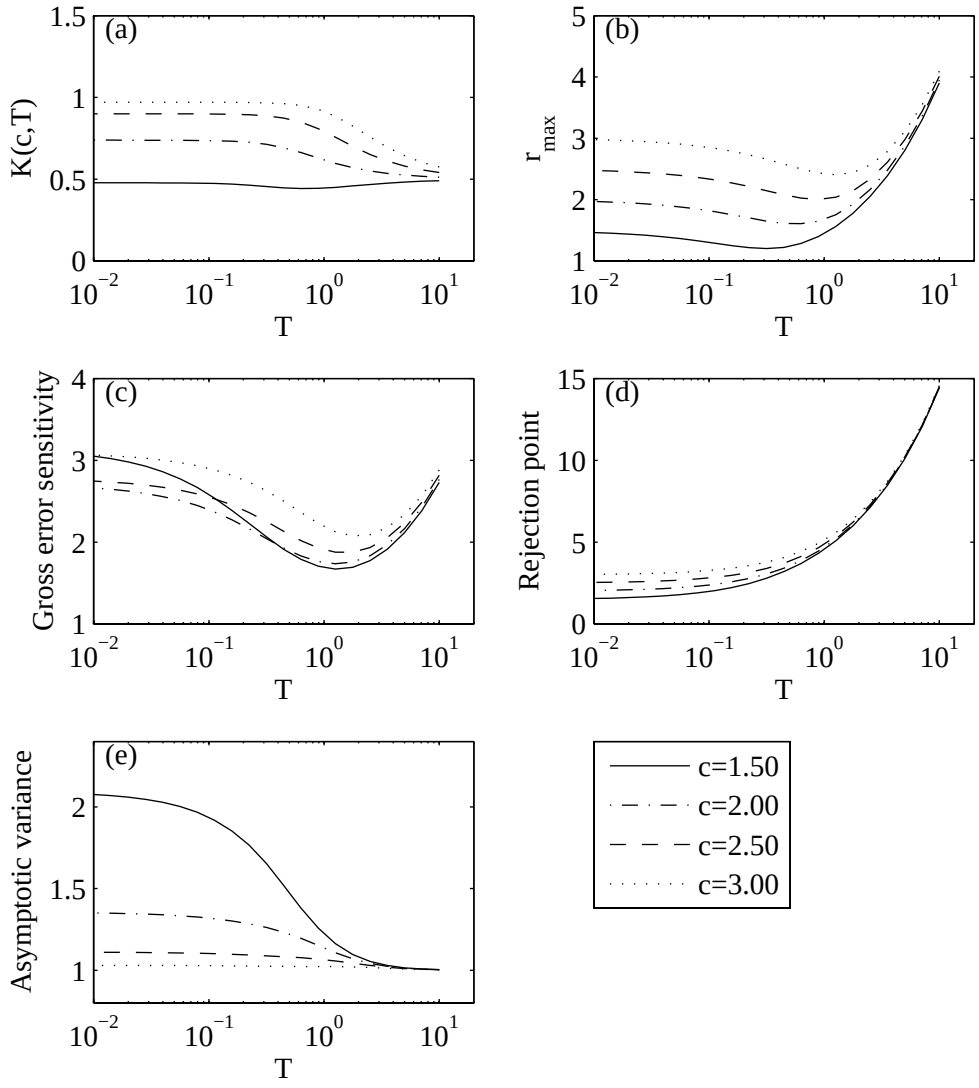

Figure 3: N-type M-estimator: (a) $K(c,T)$, (b) $r_{\max}$, (c) gross error sensitivity, (d) effective rejection point for $\varepsilon = 10^{-3}$ and (e) asymptotic variance, as a function of the temperature T, for $c = 1.5 : 0.5 : 3$.

The minimum value lies in the range $1 < T < 2$, so if one aims to minimize γ^*, the final temperature should be chosen in that range. In the low-temperature limit we have

$$\lim_{T\to 0} \gamma^*(c,T) = \frac{c}{2\Phi(c) - 1 - 2c\varphi(c)}\,.$$

At $T = 0$, γ^* is minimal for $c \approx 2.14$. The weight function $w(r;c,T)$ is always positive, so the M-estimator does not have a finite rejection point. However, an effective rejection point can be computed for a threshold ε:

$$\rho^*_{\text{eff}}(c,T,\varepsilon) = \sup\{r : \text{IF}(r;c,T) > \varepsilon\}\,.$$

Figure 3(d) shows the effective rejection point for $\varepsilon = 10^{-3}$. In the limit $T \to 0$ the effective rejection point approaches the cutoff value c. Finally, the asymptotic variance at the standard normal distribution, given by

$$V(c,T) = \frac{\int_{\mathbb{R}} \psi(r;c,T)\varphi(r)\,\mathrm{d}r}{K(c,T)^2}\,,$$

is shown in Figure 3(e). The low- and high-temperature limits are given by:

$$\lim_{T\to 0} V(c,T) = \frac{1}{2\Phi(c) - 1 - 2c\varphi(c)}\,, \qquad \lim_{T\to\infty} V(c,T) = 1\,.$$

Figure 3 shows that, for a given cutoff value c, it is not possible to minimize the gross-error sensitivity and the rejection point at the same time. The choice of the stopping temperature therefore depends on the problem at hand. If the asymptotic efficiency is important the cutoff value c should be between 2.5 and 3, at the cost of a somewhat higher gross-error sensitivity and a larger rejection point. Cutoff values larger than 3 are not recommended.

3.2 Effect of Deterministic Annealing

The effect of deterministic annealing on the minimization of the objective function of the N-type estimator is illustrated on a simple problem of location estimation with synthetic data. The data are generated from the following mixture model with mean-shift outliers (see Equation (1)):

$$h(x) = p\varphi(x) + (1-p)\varphi((x-m)/\sigma)\,.$$

We have chosen the following mixture parameters:

$$p = 0.7, \quad m = 6, \quad \sigma = 1\,,$$

which results in two barely separated standard normal components. An example data set with 500 observations is shown in Figure 4. There are 364 inliers and 136 outliers. The scale estimate s is computed by taking the median of the absolute deviations from the half-sample mode, which in this situation is a better measure of the inlier location than the sample median (Bickel and Frühwirth, 2006). Its normal-consistent value for the example data is 1.31, whereas the normal-consistent MAD is equal to 1.56. The cutoff has been set to $c = 2.5$.

It is instructive to observe the evolution of the objective function

$$M(\mu;c,T) = \sum_{i=1}^{n} \rho((x_i-\mu)/s;c,T)$$

with falling temperature T (see Figure 5). At large T, the weights are nearly independent of the residuals, and the objective function is almost quadratic. If the temperature is decreased, the objective function starts to reflect the structure of the data, eventually showing two clear local minima. These minima could be used to detect clusters in the data (Garlipp and Müller, 2005). As the objective function is minimized at each temperature, the final estimate is now totally independent of the starting value. As long as the high-temperature

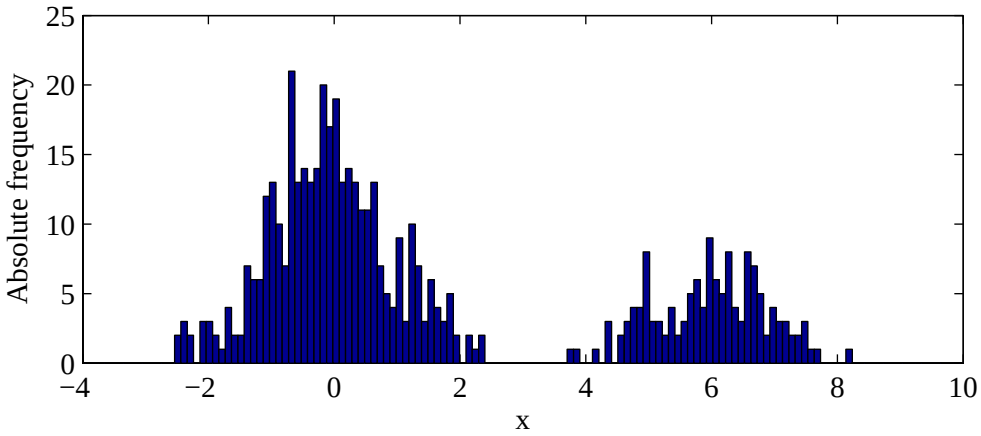

Figure 4: Example data set with 500 observations from a mixture of two standard normal distributions. The difference of the means is equal to six. There are 364 inliers and 136 outliers.

minimum is closer to the deeper low-temperature minimum convergence to the latter is virtually guaranteed.

If the separation m between inliers and outliers is decreased, the final objective function eventually has a single minimum. Figure 6 shows the final objective function at $T = 0.1$ for $m = 6, 5, 4, 3$. At $m = 5$ the second local minimum has disappeared, but the objective function still has a point of inflection close to the outlier location, and the estimate is unbiased. At $m = 4$ the point of inflection is barely visible, and the estimate shows a small bias. At $m = 3$ the point of inflection has disappeared, and the estimate shows a clear bias. In contrast, the median and the half-sample mode are totally unaffected by the change in separation.

Deterministic annealing in combination with redescending M-estimators has already been proposed by Li (1996). One of the weights function used there is a modified Welsch estimator with the weight function

$$w(r;T) = \exp(-r^2/2T)\,,$$

which is equal to the numerator of the N-type weight function in (4). It is easy to show that the asymptotic variance of this estimator at the standard normal distribution is equal to

$$V(T) = \frac{(1+T)^3}{(2+T)^{3/2}T^{3/2}}\,,$$

and consequently

$$\lim_{T\to 0} V(T) = \infty\,.$$

The same holds for the other two weight functions proposed by Li (1996).

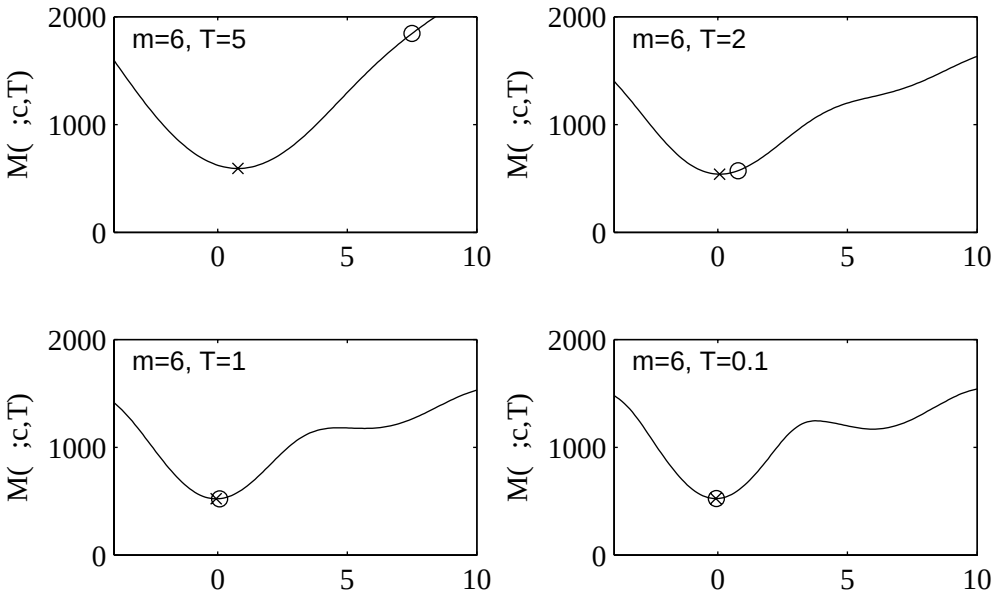

Figure 5: Evolution of the objective function $M(\mu; c, T)$ with falling temperature. The open circle (○) is the starting point of the iteration at the respective temperature, the x-mark ($\times$) is the final estimate at the respective temperature.

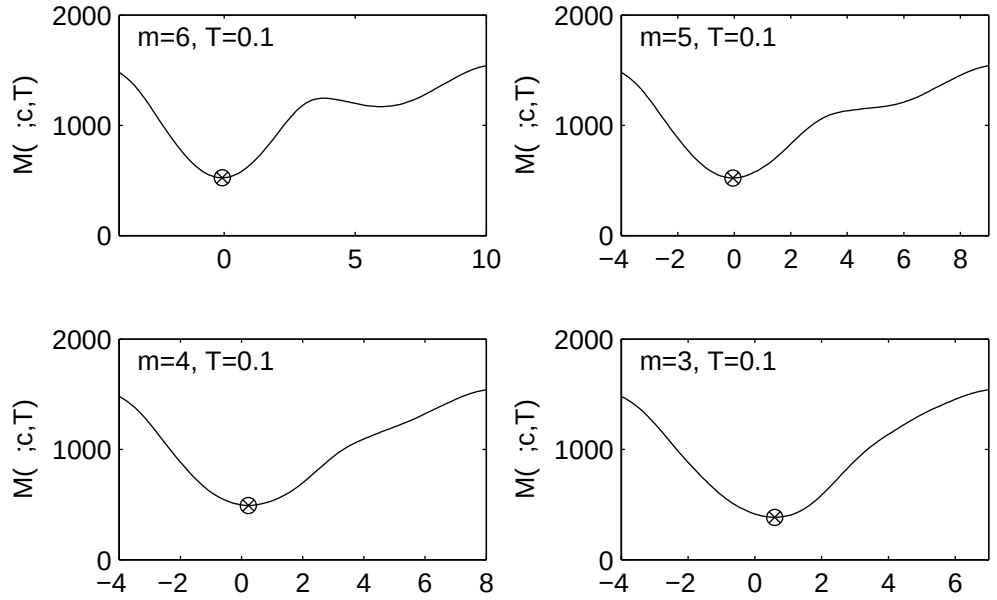

Figure 6: The objective function $M(\mu; c, T)$ at the final temperature $T = 0.1$, for different values of the mean shift m between inliers and outliers. The open circle (○) is the starting point of the iteration, the x-mark ($\times$) is the final estimate.

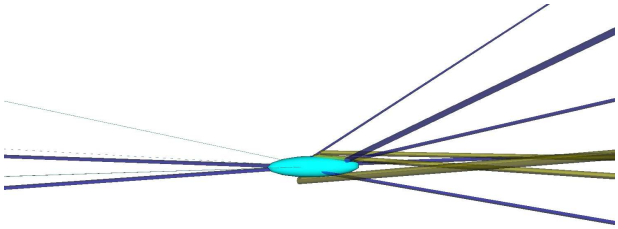

Figure 7: A primary vertex with several outgoing tracks.

4 Applications

In this section we present two application of the annealing M-estimator. In the first one the estimator is applied to the problem of estimating robustly the interaction vertex of a particle collision or a particle decay. The results show that annealing is instrumental in identifying and suppressing the outliers. In the second application the annealing M-estimator is used for regression diagnostics in the context of the estimation of the tail index of a distribution from a sample.

4.1 Robust Regression and Outlier Detection

The N-type estimator can be applied to robust regression with minimal modifications. The procedure is illustrated with the following problem from experimental particle physics.

An interaction vertex or briefly vertex is the point where particles are created by a collision of two other particles, or where an unstable particle decays and produces two or more daughter particles. The position of the vertex has to be estimated from the parameters of the outgoing particles, the so-called track parameters. The track parameters consist of location, direction and curvature. They have to be estimated before the vertex can be estimated. As an illustration, Figure 7 shows a primary vertex, the interaction point of two beam particles in the accelerator, and several outgoing tracks. The precision of the estimated track parameters is indicated by the width of the tracks.

The least-squares (LS-)estimator of the vertex position $\boldsymbol{v}$ minimizes the sum of the squared standardized distances of all tracks from the vertex position $\boldsymbol{v}$:

$$\hat{\boldsymbol{v}}_{\mathrm{LS}} = \operatorname*{argmin}_{\boldsymbol{v}} L(\boldsymbol{v}) \quad \text{with} \quad L(\boldsymbol{v}) = \frac{1}{2}\sum_{i=1}^{n} r_i^2(\boldsymbol{v}) = \frac{1}{2}\sum_{i=1}^{n} d_i^2(\boldsymbol{v})/\sigma_i^2 \,.$$

The distance d_i is approximated by an affine function of $\boldsymbol{v}$, obtained by a first-order Taylor expansion of the track model, which is the solution of the equation of motion of the particle:

$$d_i(\boldsymbol{v}) \approx c_i + \boldsymbol{a}_i^T \boldsymbol{v} \,.$$

The σ_i^2 are known from the estimation procedure of the track parameters.

With the redescending N-type M-estimator each track gets a weight w_i:

$$w_i = \frac{\exp(-r_i^2/2T)}{\exp(-r_i^2/2T) + \exp(-c^2/2T)} \,.$$

Table 1: Results of vertex estimation with the N-type M-estimator, using simulated data. For details see the text.

	Inliers		Outliers		Vertices
Annealing Schema	$w < 0.5$	$w > 0.5$	$w < 0.5$	$w > 0.5$	n_{rec}
No annealing, $T = 1$	0.312	0.688	0.859	0.141	1422
No annealing, $T = 0.01$	0.512	0.488	0.899	0.101	1004
Annealing, $T_{\text{end}} = 1$	0.101	0.899	0.828	0.172	1913
Annealing, $T_{\text{end}} = 0.01$	0.092	0.908	0.829	0.171	1939

As a consequence, outlying or mis-measured tracks are downweighted by a factor w_i. As the factor w_i depends on the current vertex position $\boldsymbol{v}$, the M-estimator is computed as an iterated reweighted least-squares estimator. The dependence on the starting point is cured by annealing. The final weights can be used for a posterior classification of the tracks as inliers ($w_i > 0.5$) or outliers ($w_i < 0.5$).

In our example we have used simulated events from the CMS experiment (CMS collaboration, 1994; CMS Collaboration, 2007) at CERN, real data not yet being available. We have studied the estimation of the primary (beam-beam collision) vertex. For more details about the estimation problem, see Waltenberger, Frühwirth, and Vanlaer (2007). The primary particles produced in the beam-beam collision are the inliers, whereas short-lived secondary particles produced in decays of unstable particles are the outliers, along with mis-measured primary tracks. Primary and secondary tracks can be identified from the simulation truth. Estimation of the primary vertex was done by the N-type M-estimator, with the least-squares estimator as the starting value. The annealing schedule was $T_0 = 256$, $T_{i+1} = T_{\text{end}} + q(T_i - T_{\text{end}})$, with $q = 0.25$.

The results are summarized in Table 1. The first column shows the type of annealing used, the second and third columns show the classification of the primary tracks by their final weights, the fourth and fifth columns show the classification of the secondary tracks, and the last column shows how many vertex estimates were within $100\mu m$ of the true vertex position, known from the simulation. Without annealing, the N-type M-estimator performs better at $T = 1$ than at $T = 0.01$. However, the results show that annealing is essential for the correct classification of primary and secondary tracks. The natural stopping temperature of the annealing procedure is $T_{\text{end}} = 1$, but cooling to $T_{\text{end}} = 0.01$ gives a slight improvement.

A similar method can be employed for the estimation of the track parameters. In this case, several observations may compete for inclusion into the track, and the computation of the weights has to be modified accordingly (Frühwirth and Strandlie, 1999).

4.2 Tail Index Estimation

The tail index α of the distribution of a random variable X is defined by

$$\alpha = \sup\{\delta > 0 : \mathrm{E}(|X|^{\delta}) < \infty\} .$$

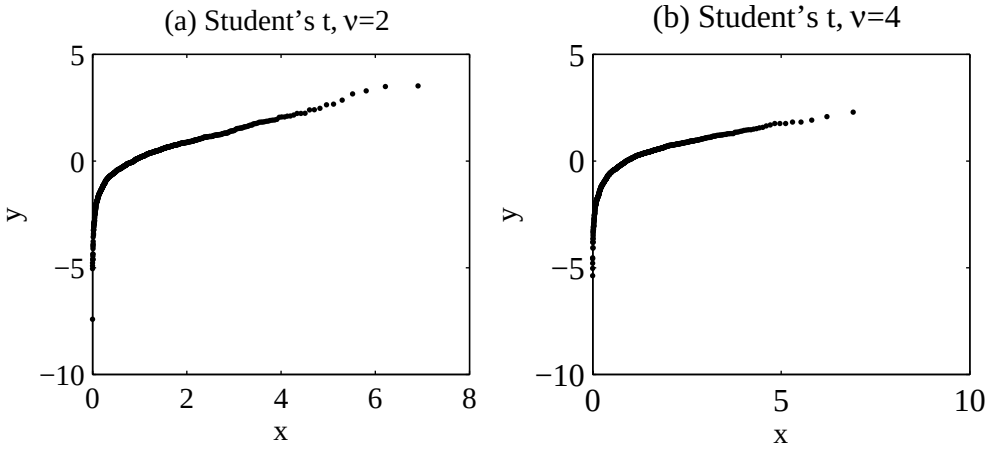

Figure 8: Pareto quantile plots of two samples of size $n = 1000$ from the t-distribution, with (a) $\nu = 2$ and (b) $\nu = 4$, respectively.

The tail index determines how many moments of X exist. A consistent estimator of α^{-1} from a sample $(x_1, \ldots, x_n)$ is the Hill estimator (Hill, 1975):

$$\hat{\alpha}_k^{-1} = \frac{1}{k} \sum_{j=1}^{k} \log X_{(n-j+1)} - \log X_{(n-k)} \,.$$

As pointed out by Beirlant, Vynckier, and Teugels (1996), the choice of k is a problem of regression diagnostics. This can be understood by looking at the Pareto quantile plot of the sample. The latter is a scatter plot (x_j, y_j), $j = 1, \ldots, n$, with

$$x_j = -\log\left(\frac{j}{n+1}\right) , \qquad y_j = \log X_{(n-j+1)} \,, \qquad j = 1, \ldots, n \,.$$

As an example, Figure 8 shows the Pareto quantile plot of two samples from the t-distribution, with $\nu = 2$ and $\nu = 4$, respectively. The sample size is $n = 1000$.

If a line is fitted to the linear part of the plot, its slope is an estimate of $1/\alpha$. The problem is therefore to find the linear part of the plot. It is worth noting that standard robust regression methods such as LMS or LTS (Rousseeuw and Leroy, 1987) will fail, as by definition the tail is not the majority of the data.

The N-type M-estimator can be used for regression diagnostics in order to find the linear part of the Pareto quantile plot. The algorithm is based on the idea of the forward search (Atkinson and Riani, 2000) and is composed of the following steps:

Algorithm A

A1. Compute the scale of y_i, using the asymptotic expression for quantiles and a kernel estimator for the probability density function. As the kernel estimator is unreliable in the very extreme part of the tail, the largest half percent of the sample is discarded.

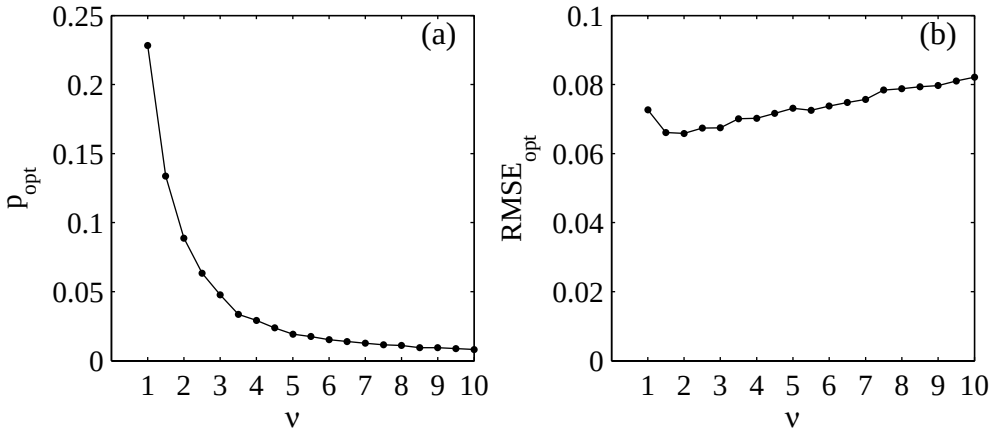

Figure 9: (a) Optimal proportion $p_{\rm opt}$ of the sample and (b) optimal RMSE of the Hill estimator, as a function of ν.

A2. Fit a robust regression line with the N-type M-estimator to the m largest points in the Pareto quantile plot. The starting line is the LMS regression line. The temperature is set to $T = 1$.

A3. Freeze all weights and extend the fit successively to the lower portion of the plot, by adding m points at a time. The choice of m is a trade-off between speed and safety.

A4. Stop adding points when the new weights get too small, indicating failure of the linear model.

We have tested the algorithm on samples from the t-distribution with ν degrees of freedom, with $n = 1000$ and $\nu = 1 : 0.5 : 10$. The baseline is the Hill estimator using the optimal value of k. The latter was found for each value of ν by computing Hill estimators with different values of k and choosing the one that minimizes the root mean-square error (RMSE) of $\hat{\alpha}_k^{-1}$ with respect to the true value $\alpha^{-1} = 1/\nu$. Figure 9 shows the optimal proportion $p = k/n$ and the RMSE of the corresponding Hill estimators as a function of ν.

Algorithm A as described above was run with $m = 10$, i.e. one percent of the sample size. The cutoff parameter c was adjusted at the 99%-quantile of the χ^2_1 distribution, i.e. at $c = 2.576$. The fit was stopped as soon as at least half of the new weights were smaller than 99% of the maximum weight $w_{\max} = 1/[1 + \exp(-c^2/2)] = 0.965$. Figure 10 summarizes the results. The left hand panel (a) shows box plots of the proportion of the sample included in the regression, one for each value of ν. The right hand panel (b) shows the resulting RMSE of $\hat{\alpha}_k^{-1}$ with respect to the true value $\alpha^{-1} = 1/\nu$, for all ν. The figure clearly shows that there is a tendency to include more data points than required for the optimal estimate. As a consequence, the RMSE is somewhat larger than in the optimal case. On the other hand, in a real life situation no external information at all may be available about the optimal value of k. In this case the regression diagnostics approach, which is entirely driven by the data, is a viable alternative.

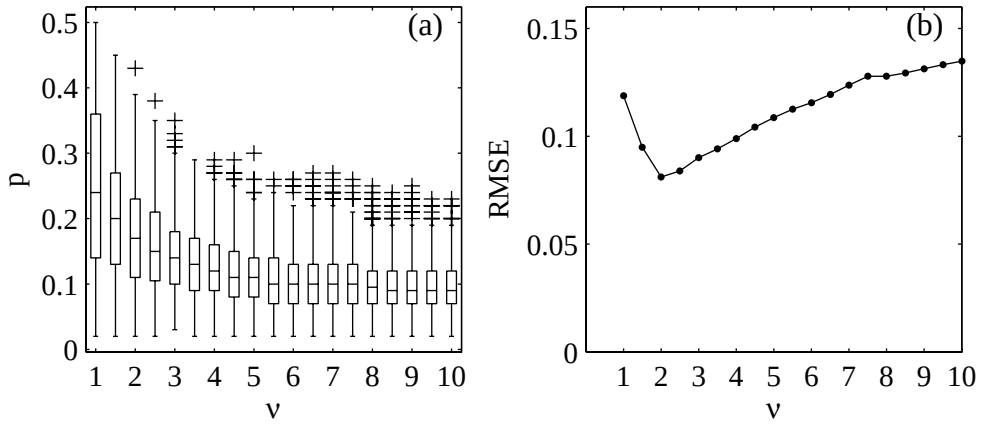

Figure 10: (a) Proportion p of the sample used by Algorithm A and (b) resulting RMSE of Algorithm A, as a function of ν.

5 Summary

A new type of redescending M-estimators has been introduced, suitable for combination with deterministic annealing. It has been shown that the annealing M-estimator converges to the skipped mean if and only if the inlier density is rapidly varying at infinity. Deterministic annealing helps to make the estimator insensitive to the starting point of iteration. Possible applications are location estimation, robust regression and regression diagnostics. The new type of estimators is particularly useful if the scale of the observations is known. In other cases the scale has to be estimated from the data, preferably in a robust way.

References

Atkinson, A., and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer.

Beirlant, J., Vynckier, P., and Teugels, J. L. (1996). Tail index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Asssociation*, *91*, 1659.

Bickel, D. R., and Frühwirth, R. (2006). On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications.

CMS collaboration. (1994). *CMS Technical Proposal* (Tech. Rep.). (Technical Report CERN/LHCC 94-38, CERN, Geneva)

CMS Collaboration. (2007). *CMS Detector Information.* (url: `http://cmsinfo.cern.ch/outreach/CMSdetectorInfo/CMSdetectorInfo.html`)

Coreless, R. M., Gonnet, G. H., Hare, D. E. G., Jerrey, D. J., and Knuth, D. E. (1996). On the Lambert W Function.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from

incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1.

Frühwirth, R., and Strandlie, A. (1999). Track fitting with ambiguities and noise: a study of elastic tracking and nonlinear filters. *Computer Physics Communications*, *120*, 197.

Garlipp, T., and Müller, C. (2005). Regression clustering with redescending M-estimators. In D. Baier and K. Wernecke (Eds.), *Innovations in Classification, Data Science, and Information Systems.* Berlin, Heidelberg, New York: Springer.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, *3*, 1163.

Huber, P. J. (2004). *Robust Statistics: Theory and Methods*. New York: John Wiley & Sons.

Li, S. Z. (1996). Robustizing robust M-estimation using deterministic annealing. *Pattern recognition*, *29*, 159.

Müller, C. (2004). Redescending M-estimators in regression analysis, cluster analysis and image analysis. *Discussiones Mathematicae — Probability and Statistics*, *24*, 59.

Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems.

Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*.

Seneta, E. (1976). *Regularly Varying Functions*. Berlin, Heidelberg, New York: Springer.

Waltenberger, W., Frühwirth, R., and Vanlaer, P. (2007). Adaptive vertex fitting. *Journal of Physics G: Nuclear and Particle Physics*, *34*, 343.

Authors' Address:

Rudolf Frühwirth and
Wolfgang Waltenberger

Institute of High Energy Physics
Austrian Academy of Sciences
Nikolsdorfergasse 18
1050 Wien
Austria

E-Mail: fru@qhepu3.oeaw.ac.at