Redescending M-estimators

Georgy Shevlyakov[1], Stephan Morgenthaler[2], Alexander Shurygin [3]

Abstract

In finite sample studies redescending M-estimators outperform bounded M-estimators (see for example, Andrews *et al.,* 1972). Even though redescenders arise naturally out of the maximum likelihood approach if one uses very heavy-tailed models, the commonly used redescenders have been derived from purely heuristic considerations. Using a recent approach proposed by Shurygin, we studied the optimality of redescending M-estimators. We show that redescending M-estimator can be designed by applying a global minimax criterion to locally robust estimators, namely maximizing the *minimum variance sensitivity* of an estimator over a given class of densities. As a particular result, we proved that Smith's estimator, which is a compromise between Huber's skipped mean and Tukey's biweight, provides the guaranteed level of an estimator's variance sensitivity over the class of distribution densities with a bounded variance.

AMS Subject Classification: 62G35

Keywords: M-estimators; minimax robustness; change-of-variance function; redescending M-estimators

1 Introduction

Let $x_1, x_2, \ldots, x_n$ be i.i.d. observations from a distribution with density $f(x, \theta)$ depending on the scalar parameter θ to be estimated. Estimating equations of the form

$$\sum_{i=1}^{n} \psi(x_i, \widehat{\theta}_n) = 0, \tag{1}$$

were studied by Godambe (1960), who showed that under regularity conditions the maximum likelihood score $\psi_f(x, \theta) \propto -\partial/\partial\theta \log(f(x, \theta))$ is the best choice. Huber (1964) demonstrated that if the model $f(x, \theta)$ is only approximately true, the optimality is not even approximately true. He also derived optimal minimax ψ-functions and, because estimators different from the maximum likelihood choice appeared to be useful, called estimators of the form (1) M-estimators. In this paper, we restrict ourselves to the simple problem of estimating the location parameter θ of

[1]School of Information and Mechatronics, Gwangju Institute of Science and Technology, Korea
[2]EPFL SB IMA, Station 8, 1015 Lausanne, Switzerland
[3]Department of Mechanics and Mathematics, Moscow State University, Moscow, Russia

a symmetric density $f(x-\theta)$ in which case the ψ-functions take the form $\psi(x-\theta)$. Henceforth, without any loss of generality, we set $\theta = 0$. There are two principal methods for designing robust estimators in this situation, namely Huber's minimax method of quantitative robustness (Huber, 1964, 1981), and Hampel's method of qualitative robustness based on the influence function (Hampel *et al.,* 1986). We will now briefly comment on the two approaches.

Under rather general regularity conditions on ψ and f, M-estimators are consistent and asymptotically normally distributed with asymptotic variance equal to

$$V(\psi, f) = A(\psi, f)/B^2(\psi, f)\,, \tag{2}$$

where

$$A(\psi, f) = \int \psi^2 f\, dx, \qquad B(\psi, f) = \int \psi' f\, dx. \tag{3}$$

As pointed out by Godambe (1960), for a given $f \in C^1(\mathbb{R})$, the variance is minimized when the ψ-function is equal to $\psi_f = -f'/f$, in which case the minimal variance is simply equal to $1/A(\psi_f, f) = 1/B(\psi_f, f)$. The efficiency of any ψ-function is defined as $\mathrm{Eff}(\psi, f) = V(\psi_f, f)/V(\psi, f)$.

In cases where the ψ-function is not smooth or the density contains point masses, the integrals in (3) must be interpreted with care. Given a class Ψ of ψ-functions and a class $\mathcal{F}$ of densities f (various suggestions can be found in (Huber, 1964, 1967, 1981; Deniau *et al.,* 1977; Hampel *et al.,* 1986), it is often possible to identify a minimax estimator ψ^*. These minimax estimators satisfy the property of guaranteed accuracy for any density $f \in \mathcal{F}$ in the sense that there exists a worst-case density f^* such that

$$V(\psi^*, f) \le V(\psi^*, f^*) = \inf_{\psi\in\Psi} \sup_{f\in\mathcal{F}} V(\psi, f). \tag{4}$$

This least favorable density f^* minimizes the Fisher information for location over the class $\mathcal{F}$

$$f^* = \arg\min_{f\in\mathcal{F}} I(f), \qquad \text{where } I(f) = \int \left(f'/f\right)^2 f\, dx\,. \tag{5}$$

The minimax score function ψ^* is equal to the maximum likelihood choice for the least informative density f^*, i.e. $\psi^* = -f^{*\prime}/f^*$. The upper bound on the asymptotic variance of the optimal estimator in (4) depends strongly on the characteristics of the chosen class of distributions $\mathcal{F}$. The key point of Huber's approach is the solution of the variational problem (5). The survey of the least favorable distributions for different distribution classes can be found in (Shevlyakov and Vilchevski, 2002). Here we recall the well-known Huber solution for the class of ε-contaminated normal distributions

$$\mathcal{F}_H = \{f\colon\ f(x) = (1-\varepsilon)\varphi(x) + \varepsilon h(x), \quad 0 \le \varepsilon < 1\}, \tag{6}$$

where $\varphi(x) = (2\pi)^{-1/2}\exp\left(-x^2/2\right)$ is the standard normal density, $h(x)$ is an arbitrary density, and ε is a contamination parameter describing the uncertainty of our

knowledge about the true underlying distribution. The least informative density then has exponential tails

$$f_H^*(x) = \begin{cases} (1-\varepsilon)\varphi(x), & \text{for} \quad |x| \le k, \\ (1-\varepsilon)(2\pi)^{-1/2} \exp\left(-k|x| + k^2/2\right), & \text{for} \quad |x| > k, \end{cases}$$

where $k = k(\varepsilon) > 0$ satisfies

$$2\varphi(k)/k - 2\Phi(-k) = \varepsilon/(1-\varepsilon), \quad \Phi(x) = \int_{-\infty}^{x} \varphi(t)\, dt, \tag{7}$$

and is tabulated in (Huber, 1981; p. 87). The optimal score function is a truncated linear function $\psi_H^*(x) = \max\left[-k, \min(x, k)\right]$.

It is likely that one could construct classes of densities in which the minimax estimator would be a bad choice because it would attempt to ensure an adequate variance for some abstruse density. In the above example and in other cases discussed in the robustness literature this does, however, not happen. It is, for example, quite surprising to note that the least informative distribution has exponential rather than the more severe Pareto-like tails. Under Pareto-like tails, quite extreme outliers would be expected to be found among the data. Such gross errors are, however, not very difficult to spot and, as Huber's result shows, the estimator that protects against exponential tails does already a sufficiently good job of it. If f^* had Pareto tails, it would follow that $\lim_{|x|\to\infty} \psi^*(x) = 0$, that is ψ^* would be redescending. Large outliers have none or negligible effect on the estimate computed with a redescender. By Huber's result, such estimators pay the price of an increased asymptotic variance at intermediate tail indices.

Hampel's local method consists in constructing an estimator with a predetermined influence function, which in turn determines the qualitative robustness properties of an estimation procedure such as its sensitivity to large outliers, to rounding off, etc. Let F be a distribution corresponding to $f \in \mathcal{F}$, our class of densities, and let $T(F)$ be a functional defined on a subset of all distribution functions. The natural estimator defined by T is $T_n = T(F_n)$, i.e. the functional evaluated on the sample distribution function F_n. The influence function $\mathrm{IF}(x; T, f)$ of this functional at one of the model densities is defined as

$$\mathrm{IF}(x; T, f) = \lim_{t\to 0} \left[T((1-t)F + t\Delta_x) - T(F)\right]/t,$$

where Δ_x is the degenerate distribution taking mass 1 at the point x (Hampel, 1968; Hampel, 1974; Hampel *et al.*, 1986; pp. 83-87). The influence function measures the impact of an infinitesimal contamination at x on the value of an estimator, formally being the Gâteaux derivative of the functional $T(F)$. For an M-estimator with score function ψ, the influence function takes the form (Hampel, 1974; Hampel *et al.,* 1986)

$$\mathrm{IF}(x; \psi, f) = \frac{\psi(x)}{B(\psi, f)}\,.$$

From the influence function, several local measures of robustness can be defined, in particular, the supremum of its absolute value, $\gamma^*(T, f) = \sup_x |\mathrm{IF}(x; T, f)|$, called

gross-error-sensitivity of T at f (Hampel *et al.,* 1986). This sensitivity gives an upper bound upon the asymptotic bias of the estimator and measures the worst influence of an infinitesimal contamination on the value of the estimator. Minimizing the asymptotic variance under the condition of a finite gross-error-sensitivity leads formally to the same estimator as Huber's approach. The use of the gross-error-sensitivity as a leading indicator of robustness excludes redescenders, because redescendency has no effect on this indicator.

By Monte Carlo simulation of a variety of M-estimators in a variety of situations, Andrews *et al.,* (1972, p. 216) found that the redescending M-estimator they dubbed 25A overall performed better than the others. Redescending ψ-functions are conventionally designed to vanish outside some central region, in other words, the following class (Hampel *et al.,* 1986) is considered

$$\Psi_r = \{\psi(x): \quad \psi(x) = 0 \quad \text{for} \quad |x| \geq r\}, \tag{8}$$

where $0 < r < \infty$ is fixed. Some examples of redescending estimators are given by the aforementioned Hampel's 25A estimator from (Andrews *et al.,* 1972) with the redescending three-part score function

$$\psi_{a,b,r}(x) = \begin{cases} x, & \text{for} \quad 0 \leq |x| \leq a, \\ a\ \text{sign}(x), & \text{for} \quad a \leq |x| \leq b, \\ a\,\dfrac{r-|x|}{r-b}\,\text{sign}(x), & \text{for} \quad a \leq |x| \leq b, \\ 0, & \text{for} \quad r \leq |x| \end{cases}$$

with $0 < a \leq b < r < \infty$, the Huber skipped mean with ψ-function

$$\psi_{\text{sk(r)}}(x) = x\, 1_{[-r,r]}(x),$$

the biweight (Mosteller & Tukey, 1977, p. 205)

$$\psi_{\text{bi(r)}}(x) = x(r^2 - x^2)^2\, 1_{[-r,r]}(x), \tag{9}$$

and a compromise between the last two, called Smith's estimator, with ψ-function (Stigler, 1980)

$$\psi_{\text{Sm(r)}}(x) = x(r^2 - x^2)\, 1_{[-r,r]}(x), \tag{10}$$

where $1_{[-r,r]}(x)$ is the indicator function taking unit value in $[-r, r]$ and zero otherwise.

Optimality of redescending estimators can be derived by implementing either Huber's or Hampel's program with the "hard rejection" condition (8). But, given their success in finite sample simulation studies, the following two questions are of interest: first, is it possible to derive redescendency without simply imposing it via the rejection point r, and, second, is it possible to weaken the condition imposed by the rejection point by considering score functions tending to zero for $|x| \to \infty$ as in (Holland and Welsch, 1977)? The basis of our paper is a reversal of the conventional setting described above. We propose to minimize the maximum of some measure

of sensitivity under the guaranteed value of the estimator's variance or efficiency in a given class of distributions. The conventional point-wise measures of sensitivity such as the influence and change-of-variance functions are not appropriate for this purpose. We propose the use of a related global indicator. The corresponding results are briefly reviewed in Section 2.

An outline of the remainder of the paper is as follows. In Section 2, we present a brief survey of the global variational optimization approach to robust estimation, show that redescending M-estimators of location with the score functions vanishing at infinity are natural within this approach, and exhibit some examples. In Section 3, we pose a maximin problem of maximizing the minimum variance sensitivity of an M-estimator of location over a given nonparametric class of densities. We show that for the class of densities with bounded variance the optimal estimator is the redescending M-estimator with the score function (10). In Section 4, some conclusions are drawn.

2 Global Stability

2.1 Variance sensitivity

In (Shurygin, 1994a, 1994b, 2000a, 2000b) a new measure of an estimator's sensitivity, derived from the asymptotic variance $V(\psi, f)$ is introduced. Formally, he derived it as

$$\begin{aligned} \frac{\partial V(\psi,f)}{\partial f} &= \frac{\partial}{\partial f}\frac{\int \psi^2 f\,dx}{\left(\int \psi' f\,dx\right)^2} = \frac{\int \psi^2\,dx}{\left(\int \psi' f\,dx\right)^2} - 2\,\frac{\int \psi'\,dx \int \psi^2 f\,dx}{\left(\int \psi' f\,dx\right)^3} \\ &= \frac{\int \psi^2\,dx}{B(\psi,f)^2} - 2\,\frac{\int \psi'\,dx\, A(\psi,f)}{B(\psi,f)^3}. \end{aligned} \tag{11}$$

Of course, in order for this to make mathematical sense, the density f and the function ψ must be smooth. Equation (11) defines a global measure of the stability of an estimator under an improper model where the outliers occur with equal chance anywhere on the real line. Since the existence of the asymptotic variance (2) is guaranteed by the existence and positiveness of the integrals $A(\psi, f)$ and $B(\psi, f)$ in (3), finiteness of the Lagrange derivative holds under the condition of square integrability, i.e. $\psi \in L^2(\mathbb{R})$. Such ψ-functions are automatically redescending, because the integral only converges if $\psi(x) \to 0$ for $|x| \to \infty$. From this it follows that $\int \psi'\,dx = 0$, so that for estimators with a finite Lagrange functional derivative, the second term summand in (11) vanishes and the Lagrange derivative of the asymptotic variance can in fact be simplified.

Definition. The scalar

$$\mathrm{VS}(\psi, f) = \frac{\partial V(\psi,f)}{\partial f} = \frac{\int \psi^2\,dx}{\left(\int \psi' f\,dx\right)^2} \tag{12}$$

is called *the variance sensitivity* of the M-estimator defined by ψ.

The variance sensitivity is an extremely stringent indicator of stability of an estimator. Its finiteness is in fact equivalent with redescendency. It is well known that the square of the influence function averaged with respect to the model density is equal to the asymptotic variance. It is thus not surprising that

$$\mathrm{VS}(\psi, f) = \int \left(\mathrm{IF}(x, \psi, f)\right)^2 dx\,.$$

By analogy with the influence function, the change-of-variance function and its sensitivity are defined as

$$\mathrm{CVF}(x; T, f) = \lim_{t\to 0} \left[V\left(T, (1-t)F + \frac{t}{2}(\Delta_x + \Delta_{-x})\right) - V(T, f)\right] \Big/ t$$

and $\kappa^*(T, f) = \sup_x \mathrm{CVF}(x; T, f)/V(T, f)$ (Hampel *et al.,* 1981). For M-estimators that satisfy $\psi \in C^1(\mathbb{R})$, the change-of-variance function is (Hampel *et al.,* 1986)

$$\mathrm{CVF}(x; \psi, f) = \frac{A(\psi, f)}{B^2(\psi, f)} \left(1 + \frac{\psi^2(x)}{A(\psi, f)} - 2\frac{\psi'(x)}{B(\psi, f)}\right), \tag{13}$$

where $A(\psi, f)$ and $B(\psi, f)$ are given by (3). Thus, up to an additive constant, the variance sensitivity is also simply equal to the integral of the CVF, if that integral exists.

The variance sensitivity of any M-estimator with increasing ψ-function is infinite. The mean, the trimmed means and the Huber estimators all have infinite variance sensitivity. Writing the median as a limit of such estimators shows that even the median has infinite sensitivity, which comes as a surprise. In fact, from the point of view of the bias, this estimator is the most robust possible.

2.2 Stability of an estimator

In what follows, we restrict ourselves to estimation of a location parameter and consider the following problem: what is the least variance sensitive ψ-function for a given distribution F with density $f \in C^1(\mathbb{R})$? This is a sensible problem, because any estimator with low variance sensitivity has an asymptotic variance that is little affected if the assumed model is only approximately true. Robustness with regard to uncertainty in the model density has thus been taken care of. Under this point of view, exploring the range of the variance sensitivity and of the efficiency at a given density is all that remains to be done. As shown in (Shurygin, 1994a, 1994b, 2000a, 2000b), the solution of the problem is given by

$$\psi_{\mathrm{MVS}}(x) = \arg \min_{\psi \in C^1(\mathbb{R})} \mathrm{VS}(\psi, f) = -f'(x). \tag{14}$$

This and further similar results are obtained by the calculus of variations techniques through writing out the Euler-Lagrange equations for the appropriate functionals. In case of minimization of the variance sensitivity, the problem is reduced

to minimization with respect to ψ of the nominator of the fraction (12) subject to its bounded denominator, i.e., the functional $J(\psi) = \int (\psi^2 + \lambda\psi' f)\, dx$, where λ is the Lagrange multiplier corresponding to the aforementioned condition. The Euler-Lagrange equation has the form $2\psi - \lambda f' = 0$ giving the required result $\psi_{\mathrm{MVS}}(x) = -f'(x)$.

The estimator with the score function (14) is called as the estimator of *minimum variance sensitivity* with $\mathrm{VS}_{min} = \mathrm{VS}(\psi_{\mathrm{MVS}}, f)$. It is easy to check that the minimum sensitivity functional takes the form

$$\mathrm{VS}_{min}(f) = \mathrm{VS}(\psi_{\mathrm{MVS}}, f) = \left(\int \psi^2_{\mathrm{MVS}} dx \right)^{-1} = \left(\int (f'(x))^2 dx \right)^{-1}. \qquad (15)$$

By comparing an estimator's variance sensitivity with the above minimum, we define the *stability* of any ψ-estimator as

$$0 \le \mathrm{Stb}(\psi, f) = \frac{\mathrm{VS}_{min}(f)}{\mathrm{VS}(\psi, f)} \le 1\,.$$

Setting different weights for efficiency and stability, various optimality criteria can be constructed (Shurygin, 1994a, 1994b, 2000a, 2000b). In particular, the structure of redescenders is specified by the analog of Hampel's lemma (Hampel *et. al.,* 1986): the maximum efficiency under the guaranteed stability

$$\max_{\psi \in C^1(\mathbb{R})} \mathrm{Eff}(\psi, f), \qquad \mathrm{Stb}(\psi, f) \ge \underline{\mathrm{Stb}}, \qquad 0 \le \underline{\mathrm{Stb}} < 1$$

is provided by the redescending M-estimator called *conditionally optimal* with the following score function

$$\psi_{c.opt}(x) = \frac{\psi_f(x)}{1 + \lambda / f(x)}, \qquad (16)$$

where $\psi_f = -f'/f$ is the maximum likelihood score function and λ is the Lagrange multiplier corresponding to the restriction upon stability. From (16) it follows that if the restriction upon stability does not matter, that is, when $\underline{\mathrm{Stb}} = 0$ and therefore $\lambda = 0$, then $\psi_{c.opt}(x) = \psi_f(x)$; otherwise $\psi_{c.opt}(x)$ is redescending, i.e., $\psi_{c.opt}(x) \to 0$ as $|x| \to \infty$. Obviously, the conditionally optimal estimator (16) also maximizes stability under guaranteed efficiency.

In practice, the freedom in choosing the level of guaranteed stability or efficiency may be inconvenient. In this case, a reasonable choice can be made setting the equal weights for efficiency and stability, i.e., when $\mathrm{Eff}(\psi) = \mathrm{Stb}(\psi)$: this estimator is called *radical*, and the score function of the radical M-estimator is given by the maximum likelihood score function ψ_f multiplied by the weight function $\sqrt{f(x)}$

$$\psi_{rad}(x) = \psi_f(x)\sqrt{f(x)} = -f'(x) \Big/ \sqrt{f(x)}\,. \qquad (17)$$

For distributions from the exponential family, the estimators of minimum sensitivity and the radical estimators belong to the class of M-estimators with the exponentially weighted maximum likelihood score functions previously proposed on intuitive grounds by Meshalkin (1971).

Score function	Normal		Laplace		Cauchy		Slash	
	Eff	Stb	Eff	Stb	Eff	Stb	Eff	Stb
ψ_f	1.00	0.00	1.00	0.00	1.00	0.50	1.00	0.50
ψ_{MVS}	0.65	1.00	0.75	1.00	0.80	1.00	0.79	1.00
ψ_{rad}	0.84	0.84	0.89	0.89	0.92	0.92	0.92	0.92

Table 1: *The normal distribution has density* $\varphi(x) = (2\,\pi)^{-1/2} \exp(-x^2/2)$. *The Laplace density is* $f(x) \propto \exp(-|x|)$, *the Cauchy density is* $f(x) \propto (1 + x^2)^{-1}$, *and the Slash density is defined as* $f(x) = (\varphi(0) - \varphi(x))/x^2$ *for* $x \neq 0$ *with* $f(0) = \varphi(0)/2$. *The maximum likelihood estimators defined by* ψ_f *are the mean (normal), the median (Laplace) and the redescenders with* $\psi_f(x) = 2x/(1 + x^2)$ *(Cauchy) and* $\psi_f(x) = \left[2 - (x^2 + 1)\, \exp(-x^2/2)\right] / \left[x\,(1 - \exp(-x^2/2))\right]$ *(Slash).*

Now let us consider several examples of the application of the variational optimization approach to M-estimators of location and, using the characteristics of efficiency and stability and compare the performance of three estimators: the maximum likelihood estimator ψ_f, the estimator of minimum variance sensitivity ψ_{MVS}, and the radical estimator ψ_{rad}. Table 1 shows the results for four distributions, two of them, Cauchy and Slash, with very heavy tails. Since the corresponding score functions are not square integrable on the real line, the maximum likelihood estimators of location for the normal and Laplace distributions, namely, the mean and median, have zero stability. On the contrary, the redescending maximum likelihood score functions for the heavy-tailed Cauchy and Slash distributions provide a reasonably moderate level of stability equal to 0.5. The redescending minimum variance sensitivity and radical estimators perform well both in efficiency and stability, especially the radical estimator.

3 Robustness

3.1 Preliminaries

The above optimization problem is local, dealing with a given density f. What if the density f is unknown or it belongs to some class $\mathcal{F}$? As pointed out above, finite variance stability implies a high degree of robustness with regard to changes in the model density f. While it is of interest to study classes of densities, it is not necessary to look at full neighborhoods as in Huber's theory. Nor is it necessary to include heavy-tailed densities in the class.

The simplest and most direct way to examine the variability of the newly introduced characteristics, the variance sensitivity and estimator's stability, is to compare their values on some set of distributions. Partially, it was done in Section 2 for the normal, Laplace, Cauchy, and Slash distributions. Now we enlarge that set of examples.

Example 1. *Consider the class of exponential power densities*

$$f(x;q) \propto e^{-|x|^q}, \qquad 1 \le q < \infty\,,$$

and the corresponding radical estimator ψ_{rad} *(17). The efficiency and stability of these estimators are*

$$\mathrm{Eff}(\psi_{rad}) = \mathrm{Stb}(\psi_{rad}) = \frac{64}{81}\left(\frac{9}{8}\right)^{1/q},$$

which is decreasing in q*, attains a maximal value at the Laplace distribution* ($q = 1$)*:* $\mathrm{Eff}(\psi_1) = \mathrm{Stb}(\psi_1) = 8/9 \approx 0.89$ *and a minimal value at* $q = \infty$*:* $\mathrm{Eff}(\psi_\infty) = \mathrm{Stb}(\psi_\infty) = 64/81 \approx 0.79$*. For the normal distribution* ($q = 2$)*, the values of efficiency and stability are:* $\mathrm{Eff}(\psi_2) = \mathrm{Stb}(\psi_2) = 32/(27\sqrt{2}) \approx 0.84$.

So, under exponential power distributions, we observe a rather low variability of the examined characteristics.

Example 2. *Let now* $f(x)$ *be the Huber least favorable density (6) in the class of* ε*-contaminated normal distributions and consider a radical estimator of the center of symmetry* $\theta = 0$ *with the score function (17). The efficiency and stability of the radical estimator are*

$$\mathrm{Eff}\,\widehat{\theta}(\varepsilon) = \mathrm{Stb}\,\widehat{\theta}(\varepsilon) = \frac{32}{27\sqrt{2}}\,\frac{\Phi^2(\sqrt{3/2}\,k(\varepsilon))}{\Phi(k(\varepsilon))\,\Phi(\sqrt{2}\,k(\varepsilon))},$$

where $\Phi(x) = \int_0^x \varphi(t)\,dt$ *and* $k(\varepsilon)$ *satisfies (7). The range of the examined characteristics is defined by their minimum and maximum values attained at the normal distribution with* $\varepsilon = 0$*:* $\mathrm{Eff}\,\widehat{\theta}(0) = \mathrm{Stb}\,\widehat{\theta}(0) = 32/(27\sqrt{2}) \approx 0.84$*, and the Laplace distribution with* $\varepsilon \to 1$*:* $\lim\limits_{\varepsilon\to 1}\mathrm{Eff}\,\widehat{\theta}(\varepsilon) = \lim\limits_{\varepsilon\to 1}\mathrm{Stb}\,\widehat{\theta}(\varepsilon)) = 8/9 \approx 0.89$*. In this case also, we have a rather narrow range of the variability of estimator's efficiency and stability both being relatively high.*

Example 3. *Consider now the standard normal distribution density* $f(x) = \varphi(x)$ *and the optimal* V*-robust redescending hyperbolic tangent estimator minimizing asymptotic variance* $V(\psi, f)$ *subject to the bounded sensitivity of the change-of-variance function* $\kappa^*(\psi, f) = \sup_x \mathrm{CVF}(x;\psi,f)/V(\psi,f) \le k$ *(Hampel et al., 1986; pp. 160-167) with the score function*

$$\psi_{tanh}(x) = \begin{cases} x, & 0 \le |x| \le p, \\ (A(k-1))^{1/2}\tanh\left[\dfrac{(k-1)^{1/2}B}{2A^{1/2}}(r-|x|)\right]\mathrm{sign}(x), & p \le |x| \le r, \\ 0, & r \le |x|, \end{cases}$$

where the recommended choice $r = 4.0$ *and* $k = 4.5$ *(Hampel et al., 1986; p. 163) implies* $A = 0.804598$*,* $B = 0.877210$ *and* $p = 1.634416$*. The efficiency and stability of this estimator are*

$$\mathrm{Eff}(\psi_{tanh}) = 0.96, \quad \mathrm{Stb}(\psi_{tanh}) = 0.53.$$

Here we observe a highly efficient estimator with an acceptable stability.

3.2 Maximization of the minimum variance sensitivity

From these examples, one may expect that the optimal M-estimators of location that maximize the minimum variance sensitivity will provide a high guaranteed level of stability of estimation over wide nonparametric classes of distributions.

Now we briefly describe this minimax approach and the expected results, and justify them further. Consider the following maximin problem

$$(\psi^*, f^*) = \arg\max_{f\in\mathcal{F}} \min_{\psi\in\Psi} \mathrm{VS}(\psi, f), \tag{18}$$

where $\mathcal{F}$ and Ψ are some suitable classes of distribution densities and score functions, respectively. This setting is almost equivalent to Huber's setting of the problem of minimax estimation of location (Huber, 1964) up to the substitution of the asymptotic variance $V(\psi, f)$ by the variance sensitivity $\mathrm{VS}(\psi, f)$.

The solution of the inner minimization problem in (18) is given by the minimum variance sensitivity score function $\psi_{\mathrm{MVS}}(x) = -f'(x)$ (14), and since the minimum sensitivity takes the form (15), the solution of problem (18) is reduced to the solution of the variational problem of minimization of the following functional

$$f^* = \arg\min_{f\in\mathcal{F}} J(f), \qquad J(f) = \int (f'(x))^2\, dx \tag{19}$$

with the subsequent application of formula (14) to the least favorable density f^*

$$\psi^*(x) = -f^{*\prime}(x). \tag{20}$$

The optimal pair (ψ^*, f^*) provides the guaranteed stability of estimation over a chosen class of distribution densities: the variance sensitivity $\mathrm{VS}(\psi^*, f)$ of the maximin M-estimator with the score function ψ^* does not exceed the value of $\mathrm{VS}(\psi^*, f^*)$ for all $f \in \mathcal{F}$

$$\mathrm{VS}(\psi^*, f) \le \mathrm{VS}(\psi^*, f^*).$$

3.3 General conditions of regularity imposed on the classes $\mathcal{F}$ and Ψ

To formulate the aforementioned results precisely, we need sufficient conditions of regularity, which will guarantee the desired asymptotic properties of M-estimators of location, namely consistency, asymptotic normality, the existence of the Lagrange derivative and variance sensitivity put on densities f and score functions ψ. These conditions can be formulated in many ways, for example, strengthening the conditions imposed on score functions and weakening those put on densities, and vice versa; here we use a standard set of assumptions (see Hampel *et al.*, 1986, pp. 125-127) with some essential changes:

($\mathcal{F}1$) f is symmetric and unimodal.

($\mathcal{F}2$) f is continuously differentiable and it may have a finite support $\{x : f(x) > 0\}$ $=(-l, l)$ with $f(l) = f'(l) = 0$.

($\mathcal{F}3$) Fisher information for location $I(f) = \int (f'/f)^2 f\, dx$ satisfies $0 < I(f) < \infty$.

($\Psi 1$) ψ is well-defined and continuous on $\mathbb{R}^+ \setminus C(\psi)$, where $C(\psi)$ is finite. At each point of $C(\psi)$ there exist finite left and right limits of ψ.

($\Psi 2$) The set $D(\psi)$ of points at which ψ is continuous but in which ψ' is not defined or not continuous is finite.

($\Psi 3$) $\int \psi f\, dx = 0$.

($\Psi 4$) $\int \psi^2\, dx < \infty$.

($\Psi 5$) $0 < \int \psi' f\, dx = -\int \psi f'\, dx < \infty$.

By ($\mathcal{F}2$) we allow densities with finite support, provided the existence of nonzero Fisher information as well as of several integrals. Note that from ($\Psi 4$) the existence of $\int \psi^2 f\, dx$ follows.

3.4 Minimax property

Following Huber (1964, 1981), we consider the functional $M(\psi, f)$, the reciprocal of the variance sensitivity $\mathrm{VS}(\psi, f)$ (12)

$$M(\psi, f) = \frac{1}{\mathrm{VS}(\psi, f)} = \frac{\left(\int \psi' f\, dx\right)^2}{\int \psi^2\, dx} \tag{21}$$

and redefine the functional $J(f)$, an analogue of Fisher information in this case, as

$$J(f) = \sup_{\psi \in \Psi} M(\psi, f),$$

where Ψ is a class of score functions satisfying the conditions ($\Psi 1$)–($\Psi 5$).

Now we are in position to formulate precisely the minimax property.

Theorem 1. *Let $\mathcal{F}$ be a convex set of densities such that every $f \in \mathcal{F}$ satisfies the conditions ($\mathcal{F}1$)–($\mathcal{F}3$).*

1. If there is an $f^* \in \mathcal{F}$ such that $J(f^*) \le J(f)$ for all $f \in \mathcal{F}$, and if $\psi^* = -f^{*\prime}$ belongs to Ψ, then (ψ^*, f^*) is a saddle point of the game

$$M(\psi, f^*) \le M(\psi^*, f^*) = J(f^*) \le M(\psi^*, f)$$

for all $\psi \in \Psi$ and all $f \in \mathcal{F}$.

2. Conversely, if (ψ^*, f^*) is a saddle point, then $J(f^*) \le J(f)$ for all $f \in \mathcal{F}$ and f^* is uniquely determined.

3. A necessary and sufficient for f^* to minimize $J(f)$ is

$$\frac{d}{dt}J(f_t)|_{t=0} = \int \psi^{*\prime}(f_1 - f^*)\,dx \geq 0$$

for all distribution densities $f_1 \in \mathcal{F}$, where

$$f_t = (1-t)f^* + tf_1, \quad 0 \leq t \leq 1$$

is the variation of f^* in the form of a mixture of densities.

PROOF. Theorem 1 literally word for word repeats Theorem 2 of Huber (1964) with the evident substitutions, namely, $M(\psi, f) \to K(\psi, f)$, $J(f) \to I(f)$, $\psi^* \to \psi_0$, etc. Moreover, in this case the proof is technically simpler due to the fact that the corresponding functionals, namely, $M(\psi, f)$ and $J(f)$, are simpler in structure than their counterparts in Huber's theory. For instance, the analog of Lemma 6 of Huber (1964) that states the convexity of the functional $K(\psi, f)$ with respect to f, immediately follows from the definition of the functional $M(\psi, f)$. ∎

3.5 Optimal redescending M-estimators

Similarly to Huber's problem, the optimal maximin solution strongly depends on the characteristics of the chosen class of densities $\mathcal{F}$.

Consider, for example, the class of densities with a bounded variance

$$\mathcal{F} = \left\{ f\colon\ f(x) \geq 0, \quad \int f(x)dx = 1, \quad \sigma^2(f) = \int x^2 f(x)\,dx \leq \overline{\sigma}^2 \right\}. \tag{22}$$

Theorem 2. *Under the conditions ($\mathcal{F}1$)–($\mathcal{F}3$), the least favorable density minimizing the functional $J(f)$ in class (22) is given by*

$$f^*(x) = \begin{cases} \dfrac{15}{16\,r^5}(r^2 - x^2)^2, & \text{for} \quad |x| < r, \\ \\ 0 & \text{for} \quad |x| \geq r, \end{cases} \tag{23}$$

where $r = \sqrt{7}\,\overline{\sigma}$.

The corresponding optimal M-estimator providing the maximum of the minimum variance sensitivity is Smith's estimator with the score function $\psi^*(x)$ given by (10).

PROOF. The proof consists of two parts. First, using the methods of the calculus of variations, we obtain the candidate for the optimal solution and, second, we check its optimality applying Theorem 1.

Since the objective functional $J(f)$ monotonically decreases with respect to increasing distribution variance, it is sufficient to consider the constraint in the form of equality: $\sigma^2(f) = \overline{\sigma}^2$.

Using the method of the Lagrange multipliers, we rewrite the constrained variational problem

$$\text{minimize } J(f) \text{ subject to } \int f(x)\,dx = 1 \text{ and } \sigma^2(f) = \overline{\sigma}^2$$

as the unconstrained variational problem

$$\text{minimize } \int \left((f'(x))^2 + 2\lambda\, f(x) + 2\mu\, x^2 f(x) \right)\, dx,$$

where λ and μ are the Lagrange multipliers corresponding to the restrictions of normalization and upon a variance, respectively.

Then the Euler equation takes the form

$$f''(x) = \lambda + \mu\, x^2\,,$$

and the stationary functions are the polynomials of the fourth order. Taking into account the restrictions of symmetry, of unimodality, of nonnegativeness, and of a fixed variance, we arrive at formula (23).

Now we check the optimality of (23) using the necessary and sufficient condition given in Part 3 of Theorem 1. Since the corresponding optimal score function $\psi^*(x) = -f^{*\prime}(x)$ is Smith's, direct computation gives $\psi^{*\prime}(x) \propto x^2 - r^2/3$ and thus

$$\frac{d}{dt} J(f_t)|_{t=0} = \int \psi^{*\prime}(x)(f_1 - f^*)\,dx \propto \overline{\sigma}^2 - \int x^2 f_1(x)\,dx \geq 0,$$

which completes the proof. ∎

Example 4. *Let $\overline{\sigma} = 1$. In this case, the score function takes the form*

$$\psi^*(x) = x(7 - x^2)\, 1_{[-\sqrt{7},\, \sqrt{7}]}(x).$$

In order to compute its characteristics at the normal density $\varphi(x)$, we need the following results

$$\int_{-\sqrt{7}}^{\sqrt{7}} \psi_{Sm}^2(x)\varphi(x)\,dx = 0.018; \qquad \int_{-\sqrt{7}}^{\sqrt{7}} \psi'_{Sm}(x)\varphi(x)\,dx = 0.122;$$

$$\int_{-\sqrt{7}}^{\sqrt{7}} \psi_{Sm}^2(x)\,dx = \frac{15}{49\,\sqrt{7}} = 0.116; \quad \int_{-\infty}^{\infty} (\varphi'(x))^2\,dx = 0.141.$$

The asymptotic variance, variance sensitivity, efficiency, minimum variance sensitivity, and stability can now be computed as

$$V(\psi_{Sm}, \varphi) = \frac{0.018}{0.122^2} = 1.198, \quad \text{Eff} = \frac{1}{1.198} = 0.835;$$

$$\text{VS}(\psi_{Sm}, \varphi) = \frac{0.116}{0.122^2} = 7.775, \quad \text{VS}_{min} = \frac{1}{0.1410} = 7.092, \quad \text{Stb} = \frac{7.092}{7.775} = 0.912.$$

Note the high levels of the estimator's efficiency and stability, naturally the latter higher than the former since the objective functional relates to stability.

When considering a bounded fourth moment, it is evident that the above proof will go through, with the least favorable density proportional to a certain fourth degree polynomial, twice integrated. Let

$$\mathcal{F} = \left\{ f\colon\ f(x) \geq 0, \quad \int f(x)dx = 1, \quad \mu_4(f) = \int x^4 f(x)\, dx \leq \overline{\mu}_4 \right\}. \tag{24}$$

Theorem 3. *Under the conditions ($\mathcal{F}1$)–($\mathcal{F}3$), the least favorable distribution density minimizing the functional $J(f)$ in class (24) is given by*

$$f^*(x) = \begin{cases} \dfrac{7}{16\, r^7}(r^4 - x^4)(2\, r^2 - x^2), & \text{for} \quad |x| < r, \\ \\ 0 & \text{for} \quad |x| \geq r, \end{cases}$$

where $r = (55\, \overline{\mu}_4/3)^{1/4}$.

In the particular case, when a distribution variance and fourth moment are given

$$\sigma^2(f) = \overline{\sigma}^2, \quad \mu_4(f) = \overline{\mu}_4, \quad \text{and} \quad \overline{\mu}_4 = 27\, \overline{\sigma}^4/11,$$

the least favorable distribution density has the form

$$f^*(x) = \begin{cases} \dfrac{35}{32\, r^7}(r^2 - x^2)^3, & \text{for} \quad |x| < r, \\ \\ 0 & \text{for} \quad |x| \geq r, \end{cases}$$

where $r = 3\, \overline{\sigma}$.

The proof of Theorem 3 is easily adapted from the proof of Theorem 2.

Corollary. *From Theorem 3 (Part 2) it follows that the optimal M-estimator providing the maximum of the minimum variance sensitivity is Tukey's biweight estimator with the score function $\psi_{bi}(x)$ given by (9).*

4 Concluding Remarks

The influence function is a basic instrument for describing statistical functionals. Beside the smoothness and boundedness of the influence function, we introduce square integrability as another characteristic of interest. Square integrability implies stability of the asymptotic variance and it excludes estimators whose ψ-functions are not redescending. For such highly resistant estimators, it can be argued that optimality at an ideal model is all that is required.

Given an ideal model density f, redescending M-estimators naturally arise, either when minimizing the variance sensitivity, or when maximizing the efficiency under

a lower bound on the stability. Generally, the optimal score functions $\psi(x)$ are of the form of a product of a weight function $w(x)$ and the maximum likelihood score function $\psi_f(x) = -f'(x)/f(x)$, that is,

$$\psi(x) = w(x)\,\psi_f(x).$$

The weights are:

- $w(x) = f(x)$ for the estimator of minimum variance sensitivity (14),
- $w(x) = f(x)/[\lambda + f(x)]$ for the conditionally optimal estimator (16),
- $w(x) = \sqrt{f(x)}$ for the radical estimator (17).

Among these, we recommend the radical M-estimator, which is at the same time highly efficient and stable (see Table 1; Shurygin, 1994a, 1994b, 2000a, 2000b; Shevlyakov *et.al.,* 2005).

If no ideal model density is assumed and instead a class $\mathcal{F}$ of model densities is considered, we may maximize the minimum variance sensitivity. We show that such well-known redescending M-estimators as Smith's estimator and Tukey's biweight can be justified in this manner.

This paper deals solely with the so-called V-robustness not touching at all the aspects of B-robustness. However, we admit that redescending M-estimators of location optimal in the V-robustness sense perform quite well also with respect to bias (Hampel *et al.,* 1986).

Finally, it is noteworthy that the Shurygin's global stability approach used in this paper can be extended to the problems of robust estimation of scale, regression, multivariate location and scatter (Shurygin, 1994a, 1994b, 1995, 1996, 2000a, 2000b).

Acknowledgements

The work by the second author has been supported by a grant of the Swiss National Science Foundation. The authors would like to thank the referees for their constructive critiques and comments, which led to a marked improvement of the paper.

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972), *Robust Estimates of Location.* (Princeton Univ. Press, Princeton).

Deniau, C., Oppenheim, G., and Viano, C. (1977), M-estimateurs. *Austérisque.*, No 43-44, 31-40.

Godambe, V.P. (1960), Optimum property of regular maximum likelihood estimation, *Ann. Math. Statist.*, **31**, 1208-1212.

Hampel, F. R. (1974), The influence curve and its role in robust estimation, *J. Amer. Statist. Assoc.*, **69**, 383-393.

Hampel, F.R., Rousseeuw, P.J., and Ronchetti, E. (1981), The change-of-variance curve and optimal redescending M-estimators, *J. Amer. Statist. Assoc.*, **76**, 643-648.

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

Holland, P.W. and Welsch, R.E. (1977), Robust regression using iteratively reweighted least squares, *Commun. Statist.*, A **6**, 813-888.

Huber, P.J. (1964), Robust estimation of a location parameter, *Ann. Math. Statist.*, **35**, 1-72.

Huber, P.J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions. In: *Proc. 5th Berkeley Symp. on Math. Statist. Prob.* **1.** Berkeley Univ. California Press, 221-223.

Huber, P.J. (1981), *Robust Statistics*, (John Wiley, New York).

Jaeckel, L.A. (1971), Robust estimates of location: Symmetry and asymmetric contamination, *Ann. Math. Statist.* **42**, 1020-1034.

Meshalkin, L.D. (1971), Some mathematical methods for the study of non-communicable diseases, In *Proc. 6th Intern. Meeting of Uses of Epidemiol. in Planning Health Services,* Vol. 1, (Primosten, Yugoslavia), 250-256.

Mosteller, F. and Tukey, J. W. (1977), *Data Analysis and Regression*, Addison-Wesley, Reading, MA.

Rousseeuw, P.J. (1981), A new infinitesimal approach to robust estimation, *Z. Wahrsch. verw. Geb.*, **56**, 127-132.

Rousseeuw, P.J. (1982), Most robust M-estimators in the infinitesimal sense, *Z. Wahrsch. verw. Geb.*, **61**, 541-555.

Shevlyakov, G.L. and N.O. Vilchevski (2002), *Robustness in Data Analysis: criteria and methods*, (VSP, Utrecht).

Shevlyakov, G.L., Shin, V.I. and Kim, K. (2005), A radical stable M-estimator of a correlation coefficient for a bivariate normal distribution. In: *Abstracts Int. Conf. on Robust Statistics, ICORS2005,* (Univ. of Jyväskylä, Jyväskylä, Finland), 91-92.

Shurygin, A.M. (1994a), New approach to optimization of stable estimation, In *Proc. 1 US/Japan Conf. on Frontiers of Statist. Modeling* , (Kluwer Academic Publishers, Netherlands), 315-340.

Shurygin, A.M. (1994b), Variational optimization of the estimator stability, *Automation and Remote Control*, **55**, 1611-1622.

Shurygin, A.M. (1995), Dimensions of multivariate statistics, *Automation and Remote Control*, **56**, 1138-1154.

Shurygin, A.M. (1996), Regression: choice of a model and stable estimating, *Automation and Remote Control*, **57**, 104-115.

Shurygin, A.M. (2000a), Estimator stability in geological applications, *Mathematical Geology*, **32**, 19-29.

Shurygin, A.M. (2000b), *Applied Stochastics: robustness, estimation and forecasting*, (Finances and Statistics, Moscow) (in Russian).

Stigler, S.M. (1980), Studies in the history of probability and statistics XX-XVIII: R. H. Smith, a Victorian interested in robustness, *Biometrika*, **67**, 217-221.

Troutman, J.L. (1983), *Variational Calculus with Elementary Convexity*, (Springer-Verlag, New York).

Tukey, J.W. (1960), A survey of sampling from contaminated distributions. In: *Contributions to Probability and Statistics. (Olkin, I., Ed.).* (Stanford Univ. Press, Stanford,) 448-485.