Ann. Inst. Statist. Math.
Vol. 43, No. 1, 131–146 (1991)

AN INFORMATION-THEORETIC FRAMEWORK FOR ROBUSTNESS

STEPHAN MORGENTHALER[1] AND CLIFFORD HURVICH[2]

[1] *Swiss Federal Institute of Technology, EPFL-DMA, 1015 Lausanne, Switzerland*
[2] *New York University, 735 Tisch Hall, Washington Sq., New York, NY 10003, U.S.A.*

(Received December 27, 1988; revised July 13, 1989)

Abstract. This is a paper about the foundation of robust inference. As a specific example, we consider semiparametric location models that involve a shape parameter. We argue that robust methods result via the selection of a representative shape from a set of allowable shapes. To perform this selection, we need a measure of disparity between the true shape and the shape to be used in the inference. Given such a disparity, we propose to solve a certain minimax problem. The paper discusses in detail the use of the Kullback-Leibler divergence for the selection of shapes. The resulting estimators are shown to have redescending influence functions when the set of allowable shapes contains heavy-tailed members. The paper closes with a brief discussion of the next logical step, namely the representation of a set of shapes by a pair of selected shapes.

Key words and phrases: Robustness, distributional shapes, Kullback-Leibler divergence.

1. Introduction

Huber (1964) argued that a location estimator ought to behave well not just at one particular distributional shape. Consequently, his robust estimator retains a low asymptotic variance in a class $\mathcal{F}$ of shapes. Robust estimators are good candidates for procedures to be used in computer packages, i.e. methods that are applied routinely. Hampel (1968, 1971, 1974) added a new interpretation. He argued that the local behavior of the estimator near a particular shape is of interest, and that this can be used to derive optimal robust estimators (Hampel *et al.* (1986)). Hampel's approach involves the statistical functional T which defines the parameter to be estimated. The center of a symmetric distribution, for example, can be described by many different functionals. Of these, Hampel proposes a specific choice, given approximate knowledge about the underlying distribution. Today, robustness is often taken to be equivalent to the stability of the inference procedure, an idea which follows in J. W. Tukey's footsteps.

Our paper is about the foundation of the robust approach. Following Huber and Hampel, we believe that robustness has to do with details of assumptions that underlie inferential procedures. Suppose our statistical model takes the form $\{P_\eta : \eta = (\theta, \lambda),\ \theta \in \mathbb{R}^p\}$, where θ describes the parameter we really wish to estimate and λ contains all the additional parameters that we must consider in order to be realistic. Note that such a parameter λ often takes values in an abstract space.

A typical non-robust method chooses a convenient λ_o and proceeds on that basis. On the other extreme lies the nonparametric approach which proceeds by estimating θ and λ together. Between these two lies the robust approach, where one selects an appropriate λ and proceeds on that basis. To formalize that selection we use a measure of disparity between models for θ having different λ-values. In order to find models that result in robust inferences about θ, one must, therefore, have two basic ingredients, namely

i) a class of allowable values for the parameter λ, and

ii) a measure of disparity between a contemplated value of λ and the true value of λ.

Given these two ingredients, we propose the following problem which generalizes Huber's approach (1964). One can think of the selection of a λ-value as finding a representer for the class of allowable λ-values. This representative element ought to be at the center of the class, meaning that the maximal disparity from the selected model to any of the others is made as small as possible. This leads to the minimax problem

$$\min_{\lambda \text{ a contemplated value}} \ \max_{\eta \text{ a true value}} \ \text{disparity}(\eta;\ \lambda).$$

These above two components of the robust approach are "necessary". There is no meaning in the term robustness unless we are sure what attribute of the inference we want to be robust about and under which circumstances.

In the next section we will explore an unrealistically simple situation where the robust approach is applicable. We then revisit the most widely discussed instance of robustness, namely the case of a vaguely known shape parameter. In this connection we study a natural measure of disparity that has never been used in robustness, namely the Kullback-Leibler divergence.

2. An example

Let X_1, X_2 be independent random variables with a normal distribution having unknown mean μ and unknown variance σ^2. We want to study interval estimation of μ. This means that μ is our θ and σ^2 is our λ. For this problem, Student's t-interval corresponds to the "nonparametric" solution, because it estimates σ^2. This interval is given by $[(X_1+X_2)/2 \pm 6.35|X_1-X_2|]$. It has a coverage probability equal to 95% and an expected length of 14.34σ.

We can also look at the described experiment as being composed of simpler models of the form $\{N(\mu, \sigma) : \sigma \text{ known}, \mu \in \mathbb{R}\}$. In this case the 95% confidence interval for μ is $i_\sigma = [(X_1 + X_2)/2 \pm 1.39\sigma]$. This solution corresponds to the "narrowly parametric" solution in our analog. The interval i_σ has, if the assumed model holds, an expected length of 2.77σ and a 95% coverage probability.

As the class of allowable λ-values, we take an interval $\sigma \in [\sigma_L, \sigma_U]$, for some $0 < \sigma_L < \sigma_U$. The disparity function is an asymmetric function of the contemplated scale λ and the true scale η. Its value $0 \leq d(\eta;\ \lambda)$ indicates how well the inference based on model λ behaves, if the data actually follows model η. The confidence interval i_λ has an expected length of 2.77λ and coverage probability $(2\Phi(1.96\lambda/\eta) - 1)$, assuming η describes the true model (Φ denotes the standard normal distribution). An ad hoc choice for d, balancing the increased expected length and the decreased coverage probability, is

$$d(\eta;\ \lambda) = 2.77(\lambda - \eta)_+ + \left(\log\left(\frac{0.95(1-x)}{0.05x}\right)\right)_+, \tag{2.1}$$

where $x = 2\Phi(1.96\lambda/\eta) - 1$, and $(z)_+ = z$, if $0 \leq z$ and $= 0$, otherwise.

Example. Suppose $\sigma \in [1, 9]$. The above disparity function between models leads to the minimax model $N(\mu, 7.45)$. However, a direct comparison of the "robust" interval with the "nonparametric" one seems unfair, because one would surely take into account that $\sigma \in [\sigma_L, \sigma_U]$ even in the adaptive approach. For that reason we include in the evaluation a modified version of Student's t which uses the estimate $s_r = \min(\sigma_U, \max(\sigma_L, s))$ of σ instead of the usual estimate s. The ratio $2^{1/2}(X_1 + X_2)/(2s_r)$ is no longer invariant under scale changes and the critical value depends on the unknown σ. To facilitate the comparison, we used the correct critical value for $\sigma = 7.45$. Expected lengths and coverage probabilities of the three intervals are shown in Figs. 1 and 2.

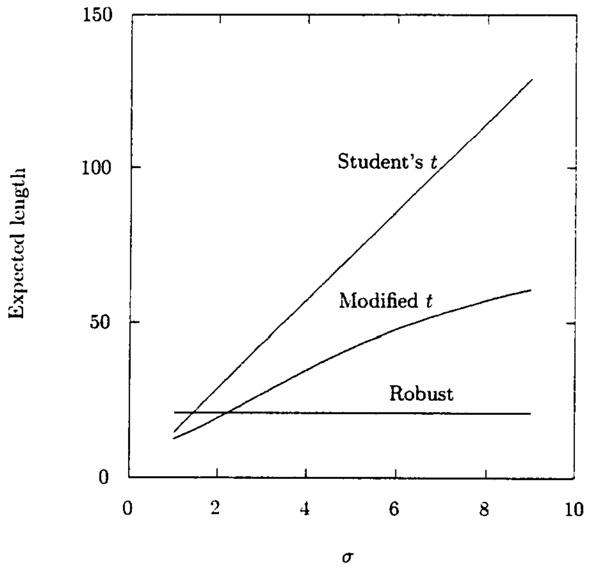

Fig. 1. Expected lengths of three confidence intervals.

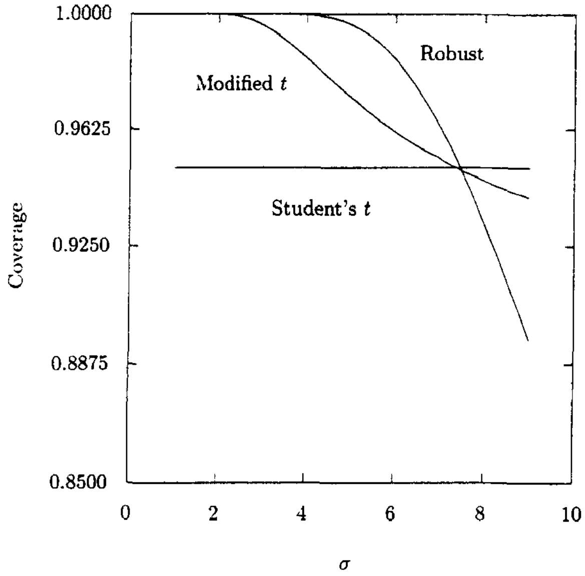

Fig. 2. Coverage probabilities of three confidence intervals.

Apparently, Student's t-interval pays a heavy penalty for its "adaptiveness", as is shown in the dramatic increase of the expected length. The modified version undoes some of this, but not all. In Fig. 2, we can evaluate the robust procedure with regard to coverage. The worst coverage is reached at $\sigma = 9$ with about 89.5%. Notice how in the region $\sigma \in [4, 6]$ the "robust" interval has a higher coverage than that of the "modified adaptive" one, even though the expected length is shorter.

This trivial example shows what we mean by robustness. Robust techniques come into play whenever there are "nuisance" parameters that are difficult to estimate, but may have a substantial bearing on the inference. In such situations one typically pays too heavy a price for estimating these parameters. Robustness is an alternative. One selects a value for the nuisance parameter, and then uses the selected model to make inferences.

In the simplest situations this approach works quite well. In more complicated cases, or if one wants an even better method, one can generalize this idea. As a first step, instead of selecting a single representer for the class of allowable λ-values, one can select two representative models. Both can then be fit to the data and a compromise between the two inferences can be found. In general this will improve over the simpler robust method, while still not attempting to estimate the nuisance parameter.

Example. Consider once more the example involving location and scale and suppose $\sigma \in [1, 9]$. In order to select a pair of models, the disparity function must be modified to measure the discrepancy between a pair of contemplated models and a true model. If one thinks again of the disparity as a kind of distance in the set $\mathcal{F}$, then the problem

$$\min_{(\lambda_1,\lambda_2)\in[\sigma_L,\sigma_U]^2} \max_{\eta\in[\sigma_L,\sigma_U]} (d(\eta;\ \lambda_1) \wedge d(\eta;\ \lambda_2))$$

makes sense ($x \wedge y = \min(x,\ y)$). The disparity (2.1) will now lead to the simultaneous model $\{N(\mu,\ 2.5),\ N(\mu,\ 8.1)\}$. Sophisticated compromise confidence intervals for this model are described in Morgenthaler (1986*b*). It is no longer feasible to analyze their distributional properties without simulation.

In the next section we will discuss Huber's robustness problem (1964) involving distributional shapes and we will consider two alternative choices for the disparity function.

3. Shape parameters

3.1 *Huber's disparity and entropy*

Huber (1964) considers point estimation for a location model $\{F(x-\mu): \mu \in \mathbb{R}\}$. This model has an infinite dimensional "nuisance parameter" λ, namely the distribution function $F(x)$, which is assumed to be symmetric around 0. We call such a parameter a shape parameter. This is an example which falls under the category of Section 2. Based on a smallish number of observations, it is difficult to estimate F or its density f. It is especially hard to infer much useful knowledge about the most crucial aspect of F, its tail behavior.

It is well known that there exist adaptive location estimators that are efficient if our class $\mathcal{F}$ of allowable shapes is $\mathcal{F} = \{$absolutely continuous, symmetric distributions with center of symmetry at 0$\}$ (Stone (1975) and Beran (1978)). However, efficiency is defined asymptotically and the performance of these efficient nonparametric estimators in smallish sample sizes ($n \leq 20$) is unknown.

In robustness one works with smaller classes $\mathcal{F}$, the leading example being the gross-error model $GE(\epsilon,\ \Phi) = \{(1-\epsilon)\Phi + \epsilon H: H$ symmetric$\}$, where Φ is the standard normal distribution and $\epsilon > 0$ is fixed (see Huber (1981)). Huber (1964) considers the disparity function

$V(F;\ G) =$ asymptotic variance of T_G ($=$ MLE based on G) when the data follows the true model F.

In order for this function to be well defined on $GE(\epsilon,\ \Phi) \times GE(\epsilon,\ \Phi)$, one must choose a subset of $GE(\epsilon,\ \Phi)$ that contains sufficiently nice models.

Example. (Huber (1964)) Let $SU(\epsilon,\ \Phi) = \{F \in GE(\epsilon,\ \Phi): F$ absolutely continuous, $-\log f$ convex, with support $\mathbb{R}\}$. (A distribution with the property "$-\log$(density) convex" is called strongly unimodal.) In this case one can show that there exists a unique $F_o \in SU(\epsilon,\ \Phi)$ such that

$$V(F;\ F_o) \leq V(F_o;\ F_o) \leq V(F_o;\ G)$$

for any choice of F and G from $SU(\epsilon, \Phi)$. The shape F_o is, therefore, at the center of $SU(\epsilon, \Phi)$ in the sense of the disparity function $V(F;\ G)$. (It should be noted that Huber's minimax estimator (1964) satisfies a more general proposition. One may choose F from a bigger class than $SU(\epsilon, \Phi)$.)

A divergence function between distributions that is of statistical importance, especially in connection with hypothesis testing, is the Kullback-Leibler divergence. When F and G are absolutely continuous with F's support contained in G's support, this function is equal to

$$I(F;\ G) = \int_{-\infty}^{\infty} (\log f - \log g) f dx \qquad \text{(Kullback (1968))}.$$

In all other cases, $I(F;\ G) = \infty$.

If our set $\mathcal{F}$ contains only smooth distributions then the Kullback-Leibler divergence is a reasonable measure for the problem of estimating μ. One connection is immediate. Note that $(-\log f)$ and $(-\log g)$ are the contrast functions corresponding to the M-estimators T_F and T_G, so that the function $I(F;\ G)$ has an intuitive interpretation in location estimation. A different connection has been revealed in Morgenthaler (1986*a*). When we attempt to adapt our estimation procedure to the underlying shape, we face the task of distinguishing between shapes based on the observed data. This is a testing problem and the Kullback-Leibler divergence enters quite naturally. In fact, an optimal procedure that compromises between two fixed shapes proceeds on the basis of the Kullback-Leibler divergence, not on the basis of Huber's distance measure.

It is important to realize that the K-L divergence only makes sense for smooth shapes. From now on we assume our shapes to be absolutely continuous with support equal to $\mathbb{R}$. In fact, our set $\mathcal{F}$ can always be represented as a subset of $L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$, because all of our distributions have bounded densities.

For the analysis of $I(F;\ G)$ it is helpful to consider first the simpler shape disparity

$$i(F;\ G) = -\int_{-\infty}^{\infty} \log g f dx.$$

To understand the behavior of this function, consider

$$En(G) = -\int_{-\infty}^{\infty} \log g g dx,$$

the entropy of G. The Gâteaux derivative of En at G in the direction F is, if it exists, equal to

$$\partial En(F;\ G) = \frac{d}{dt} En(G_t)\mid_{t=0}\ = i(F;\ G) - En(G),$$

where $G_t = (1-t)G + tF$. In other words, the shape $F^* \in \mathcal{F}$ ($\mathcal{F}$ convex) that maximizes En must be a saddle point of $i(F;\ G)$.

PROPOSITION 3.1. *Suppose $\mathcal{F}$ is a compact, convex set of absolutely continuous distributions that have support equal to $\mathbb{R}$, as well as bounded densities. Assume that En is bounded on $\mathcal{F}$. Then it follows that En has a unique maximizer $F^* \in \mathcal{F}$ and furthermore*

$$i(F;\ F^*) \le i(F^*;\ F^*) \le i(F^*;\ G), \qquad \forall F, G \in \mathcal{F}.$$

PROOF. The functional $-En$ is lower semi-continuous (Rockafellar (1971)). Let F, $G \in \mathcal{F}$ with $F \neq G$, and define $G_t = (1-t)G+tF$. The function $h\colon (0,\ 1) \to \mathbb{R}$ defined by $h(t) = -En(G_t)$ is in $C^2(0,\ 1)$ and satisfies

$$h''(t) = \int_{-\infty}^{\infty} (f-g)^2/g_t dx > 0 \qquad (g_t = (1-t)g+tf).$$

It follows that $h(t)$ is strictly convex. This implies the existence and uniqueness of the point F^* where En achieves its maximum. Since the Gâteaux derivative at F^* in direction F exists, it must satisfy

$$i(F;\ F^*) - En(F^*) = \partial En(F;\ F^*) \le 0, \qquad \forall F \in \mathcal{F}.$$

We therefore have one of the claimed inequalities. The second inequality simply follows from the well-known fact that

$$I(F^*;\ G) \ge 0 \qquad \forall G. \qquad \square$$

As an application of this proposition, let us reconsider the example of nice gross-error shapes, $SU(\epsilon,\ \Phi)$. Denote by $\mathrm{co}(SU(\epsilon,\ \Phi))$ the convex hull of $SU(\epsilon,\ \Phi)$. The assumptions of the previous proposition are satisfied by $\mathrm{co}(SU(\epsilon,\ \Phi))$.

PROPOSITION 3.2. *Huber's least favorable distribution F_o* (1964) *with density*

$$f_o(y) = \begin{cases} (1-\epsilon)\varphi(y), & \text{if } |y| \le k_\epsilon, \\ (1-\epsilon)\varphi(k_\epsilon)\exp(-k_\epsilon(|y|-k_\epsilon)), & \text{if } |y| > k_\epsilon, \end{cases}$$

maximizes entropy in the class $\mathrm{co}(SU(\epsilon,\ \Phi))$. (*The value of k_ϵ satisfies $2\varphi(k_\epsilon)/k_\epsilon - 2\Phi(-k_\epsilon) = \epsilon/(1-\epsilon)$; and φ denotes the standard normal density.*)

PROOF. The shape $F_o \in \mathrm{co}(SU(\epsilon,\ \Phi))$ maximizes $En(F)$, if and only if,

$$\partial En(G;\ F_o) = i(G,\ F_o) - En(F_o) \le 0, \qquad \forall G.$$

We will first show the existence of a constant $0 < c(G) < \infty$ for any $G \in SU(\epsilon,\ \Phi)$ such that

$$\begin{aligned} &g(x) - f_o(x) \le 0 \quad \text{for almost all } x \ge c(G) \quad \text{and} \\ &g(x) - f_o(x) \ge 0 \quad \text{for almost all } 0 \le x \le c(G). \end{aligned} \tag{3.1}$$

The first assertion of (3.1) is trivial. For a G with lighter than exponential tails, it follows from strong unimodality. For a G with exponential tails, it is a consequence of the construction of F_o, whose splice point k_ϵ is as large as possible.

The second assertion in (3.1) is a consequence of strong unimodality. Because $f_o(x) = (1-\epsilon)\varphi(x) \leq g(x)$ for all $0 \leq x \leq k_\epsilon$ and for all g in $SU(\epsilon,\ \Phi)$, any density g that does not satisfy (3.1) would have to cross over the density f_o (at least) twice in the interval $[k_\epsilon,\ \infty)$, say at x_1 and $x_1 < x_2$. Beyond x_2, g would be smaller than f_o; between x_1 and x_2 it would be bigger; and at some point between k_ϵ and x_2 it would be smaller. Consider the function $-\log g$ at the points k_ϵ, x_2, and a point $\tilde{x}$ between k_ϵ and x_1 with $g(\tilde{x}) < f_o(\tilde{x})$. Since $g(k_\epsilon) \geq f_o(k_\epsilon)$, the straight line connecting $(k_\epsilon,\ -\log g(k_\epsilon))$ and $(x_2,\ -\log g(x_2))$ lies below $-\log f_o(x)$ $(x > k_\epsilon)$. The inequality $-\log g(\tilde{x}) > -\log f_o(\tilde{x})$ shows that $-\log g$ cannot be convex.

The inequalities (3.1) imply that F_o maximizes entropy within $SU(\epsilon,\ \Phi)$. Take any $G \in SU(\epsilon,\ \Phi)$. Both f_o and g are symmetric probability densities, implying

$$\int_{c(G)}^{\infty} (g(x) - f_o(x))dx = -\int_0^{c(G)} (g(x) - f_o(x))dx \leq 0.$$

Furthermore, since $-\log(f_o(x))$ is increasing on $\mathbb{R}_+$, we find

$$\begin{aligned}\int_{c(G)}^{\infty} -\log f_o(x)(g(x) - f_o(x))dx &\leq -\log f_o(c(G)) \int_{c(G)}^{\infty} (g(x) - f_o(x))dx \\ &= \log f_o(c(G)) \int_0^{c(G)} (g(x) - f_o(x))dx \\ &\leq \int_0^{c(G)} \log f_o(x)(g(x) - f_o(x))dx.\end{aligned}$$

This implies

$$\int_0^{\infty} -\log f_o(x)(g(x) - f_o(x))dx \leq 0.$$

With this, Proposition 3.2 is proved, because the convexity of $En(F)$ implies that its maximum over the set $\text{co}(SU(\epsilon,\ \Phi))$ is achieved in $SU(\epsilon,\ \Phi)$. □

The disparity functions $V(F;\ G)$ and $i(F;\ G)$ behave similarly and both lead to tractable optimization problems. Both of them possess a saddle point. Because of this fact, however, we cannot generalize to the selection of two shapes. Suppose we consider the minimax problem involving pairs of shapes,

$$\inf_{(G_1,G_2)\in\mathcal{F}\times\mathcal{F}} \sup_{F\in\mathcal{F}} i(F;\ G_1) \wedge i(F;\ G_2).$$

Any pair $(F^*,\ G)$, with F^* the entropy maximizer and G arbitrary, solves this problem. A similar comment applies to Huber's disparity $V(F;\ G)$.

3.2 *Kullback-Leibler divergence*

It is clear how one must modify the two divergence measures from the last section. In Huber's case (1964) we ought to consider asymptotic efficiency rather than asymptotic variance. In the case of $i(F;\ G)$, we ought to switch to $I(F;\ G)$. We have previously derived the formula $I(F;\ G) = i(F;\ G) - En(F) = \partial En(F;\ G) + En(G) - En(F)$. The function $I(F;\ G)$ has a geometrical interpretation. It is equal to the difference between the concave "surface" $En(F)$ and its "tangent" $\partial En(F;\ G) + En(G)$ at the point G. This representation makes the analysis of

$$\min_{G\in\mathcal{F}} \max_{F\in\mathcal{F}} I(F;\ G) \tag{3.2}$$

feasible.

PROPOSITION 3.3. *Let $\mathcal{F}$ be a compact, convex set of absolutely continuous distributions having support equal to $\mathbb{R}$, as well as bounded densities. And suppose that En is bounded on $\mathcal{F}$. Then there exists a unique point $G^* \in \mathcal{F}$ that solves* (3.2).

PROOF. $\mathcal{F}$ is the convex hull of its extremal points. Denote by L the linear functional which is equal to En at all these extremal points. The functional $(En - L)$ is then a concave, bounded functional on $\mathcal{F}$. It has therefore a unique maximizer $G^* \in \mathcal{F}$. Furthermore, it follows that $\partial(En - L)(F;\ G^*) = \partial En(F;\ G^*) - L(F) + L(G^*) \leq 0$, $\forall F \in \mathcal{F}$. Let $F_e \in \mathcal{F}$ be an extremal point. We then find, using the previous inequality, that

$$I(F_e;\ G^*) = \partial En(F_e;\ G^*) + En(G^*) - En(F_e) \leq En(G^*) - L(G^*).$$

Let $G \in \mathcal{F}$ be any shape such that the inequality

$$I(F_e;\ G) \leq En(G^*) - L(G^*)$$

holds for all extremal points F_e. From this it follows that

$$En(G) \leq En(G^*) - L(G^*) - \partial En(F_e;\ G) + En(F_e), \qquad \forall F_e.$$

This implies

$$En(G) - L(G) \leq En(G^*) - L(G^*) + \partial L(F_e;\ G) - \partial En(F_e;\ G), \qquad \forall F_e.$$

Since both $\partial En(\cdot\ ;\ G)$ and $\partial L(\cdot\ ;\ G)$ are continuous linear functionals, there must exist an extremal point F_e where $\partial L(F_e;\ G) - \partial En(F_e;\ G) \leq 0$, unless the two functionals are identical. In both cases we conclude that $G = G^*$.

We have shown that G^* is the unique value of G in $\mathcal{F}$ such that $I(F_e;\ G) \leq En(G^*) - L(G^*)$ for all extremal points F_e. This, together with the fact that $I(F_e;\ G^*) \leq En(G^*) - L(G^*)$ for all extremal points F_e, implies that G^* is the unique minimizer in $\mathcal{F}$ of $\max I(F_e;\ G)$. (Here max is taken with respect to all extremal points F_e, which suffices because of the convexity of $I(F;\ G)$ with regard to F.) □

3.3 *Strongly unimodal shapes close to the Gaussian*

We will apply Proposition 3.3 in the case of $\mathrm{co}(SU(\epsilon, \Phi))$, the convex hull of the strongly unimodal shapes. The extremal points of $\mathrm{co}(SU(\epsilon, \Phi))$ must already be contained in $SU(\epsilon, \Phi)$ and it is easy to guess their shape. Let $\lambda \geq 0$ and define

$$f_\lambda(x) = \begin{cases} (1-\epsilon)\varphi(x), & \text{if } |x| > x_u \text{ or } |x| < x_l, \\ (1-\epsilon)\varphi(x_l)\exp(-x_l(|x|-x_l)), & \text{if } x_l < |x| < \lambda, \\ (1-\epsilon)\varphi(x_u)\exp(-x_u(|x|-x_u)), & \text{if } \lambda < |x| < x_u. \end{cases}$$

For any choice of $x_l < x_u$ with $\lambda = (x_l + x_u)/2$, this function is continuous, symmetric and positive. One finds that

$$\begin{aligned}\int_0^\infty f_\lambda(x)dx = 1/2 - \epsilon/2 &+ (1-\epsilon)(\Phi(x_l) - \Phi(x_u)) \\ &+ (1-\epsilon)(\varphi(x_l)/x_l - \varphi(x_u)/x_u) \\ &- (1-\epsilon)\varphi(x_l)\exp(-x_l(x_u - x_l)/2)(1/x_l - 1/x_u).\end{aligned}$$

Setting this expression equal to 1/2 leads to an equation which must be satisfied by (x_l, x_u) in order to make f_λ a density. It turns out that for λ large enough, the density f_λ belongs to $\mathrm{co}(SU(\epsilon, \Phi))$. (And it is certainly an extremal point, since by construction $f_\lambda(\lambda) > f(\lambda)$, $\forall f \in SU(\epsilon, \Phi)$, $f \neq f_\lambda$. This is therefore also true for all $f \in \mathrm{co}(SU(\epsilon, \Phi))$, $f \neq f_\lambda$.) When the value of λ is close to zero, the point x_l is negative and f_λ is no longer strongly unimodal. In that case the proper definition is

$$f_\lambda(x) = \begin{cases} (1-\epsilon)\varphi(x), & \text{if } |x| > x_u, \\ (1-\epsilon)\varphi(x_u)\exp(-x_u(|x|-x_u)), & \text{if } \lambda < |x| < x_u, \\ (1-\epsilon)\varphi(x_u)\exp(-x_u(\lambda - x_u)), & \text{if } |x| < \lambda, \end{cases}$$

with

$$\begin{aligned}1/2 = (1-\epsilon) - (1-\epsilon)\Phi(x_u) &- (1-\epsilon)\varphi(x_u)/x_u \\ &+ (1-\epsilon)\varphi(x_u)\exp(-x_u(\lambda - x_u))(\lambda + 1/x_u).\end{aligned}$$

We conjecture that the family of densities $\{f_\lambda : \lambda \geq 0\}$ contains all the extremal points of $\mathrm{co}(SU(\epsilon, \Phi))$. To determine the solution G^* of

$$\min_{G \in \mathrm{co}(SU(\epsilon,\Phi))} \max_{F \in \mathrm{co}(SU(\epsilon,\Phi))} I(F;\ G)$$

we resort to numerical analysis, since an analytic solution seems impossible. We have chosen a discrete subset of $\{f_\lambda : \lambda \geq 0\}$ and considered the problem

$$\min_{G \in \langle F_{\lambda 1}, \ldots, F_{\lambda n}\rangle} \max_{F \in \{F_{\lambda 1}, \ldots, F_{\lambda n}\}} I(F;\ G), \tag{3.3}$$

where $\langle F_{\lambda 1}, \ldots, F_{\lambda n}\rangle$ denotes the convex set spanned by $F_{\lambda 1}, \ldots, F_{\lambda n}$. In (3.3) we made use of convexity which implies that for any fixed G, the maximum of $I(F;\ G)$ must be achieved at one of the extremal points. The function $I(F;\ G)$ was evaluated numerically using Romberg's integration. All the densities in $\mathrm{co}(SU(\epsilon, \Phi))$

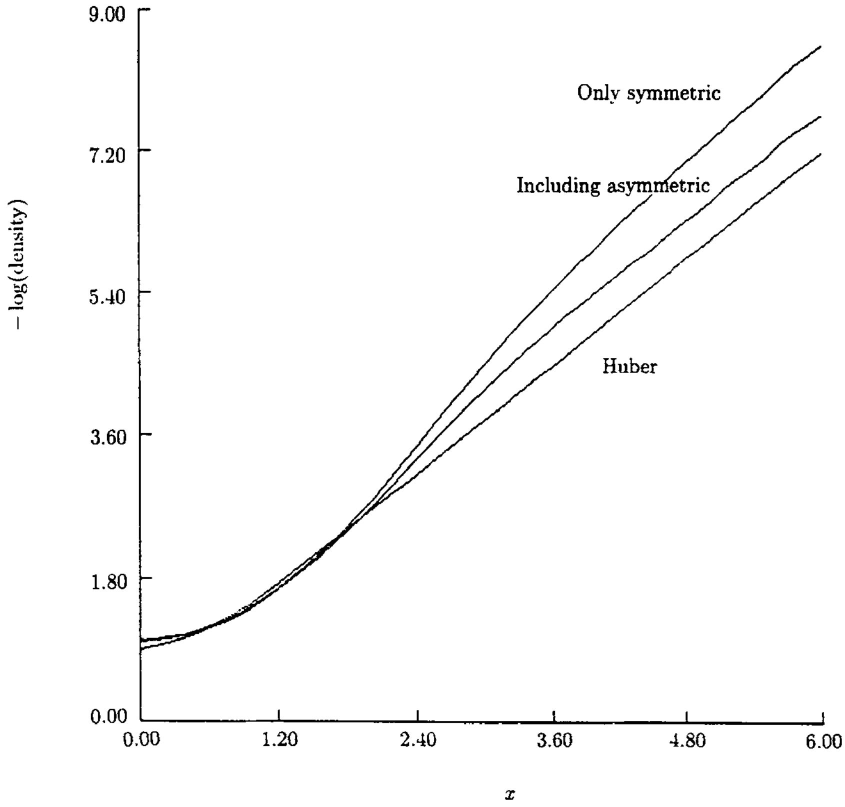

Fig. 3. Approximate solutions in the classes SU and AS.

are well behaved, so that the evaluation of the Kullback-Leibler distance is not tricky.

Figure 3 shows the function $-\log g^*$ for G^* being the solution of (3.3) when $n = 24$, $\lambda \in \{0(0.01)0.05,\ 0.1,\ 0.15,\ 0.25,\ 0.5,\ 0.75,\ 1(0.5)6.5,\ \infty\}$ and $\epsilon = 0.1$ (the symbol $a(h)b$ denotes $a,\ a+h,\ a+2h, \ldots, b$).

Also shown in the plot is Huber's shape (1964), which, as we know, solves the entropy maximization problem. It can be seen from this plot that the Kullback-Leibler shape is more squeezed in the shoulders of the density. The third shape shown in the plot will be explained in Section 4.

3.4 *The gross-error model*

Proposition 3.3 cannot be applied to the set $GE(\epsilon, \Phi)$ because that set is too large. The maximal disparity is always ∞. It is, however, interesting to note that the Kullback-Leibler divergence leads to redescending M-estimates when applied to a smallish subset of $GE(\epsilon, \Phi)$ that contains shapes that are not strongly unimodal.

Example. We choose the triangle spanned by $\{\Phi, H, S\} \in GE(1/3, \Phi)$, where H denotes Huber's least-favorable distribution (1964) and S is the slash shape with density $s(x) = 0.75(1 - \exp(-x^2/(2(0.75)^2)))/(x^2\sqrt{2\pi})$. The following minimax shapes are found with the various disparities discussed in this paper.

$V(F;\ G) \to H,$

$I(F;\ G) \to 0.58 \cdot \Phi + 0.29 \cdot H + 0.13 \cdot S,$

$i(F;\ G) \to S.$

This shows again the extreme sensitivity of the entropy maximizer to tail thickness. And it shows that the slash shape which has Pareto tails contributes to the Kullback-Leibler solution.

4. Asymmetric shapes and selection of pairs

The restriction to symmetric shapes is artificial and is usually justified by the need for a unique definition of parameters, applicable to classes of shapes (Huber (1964)). In the present context, where parameter estimation is more in the background, one can easily treat families of shapes, which include asymmetric ones.

Example. A mild model for asymmetry would be the class

$$\begin{aligned} AS(\epsilon,\ \Phi) = \{F = (1-\epsilon)\Phi + \epsilon H\colon F \text{ absolutely continuous,}\\ -\log(f) \text{ convex and support of } F \text{ equal to } \mathbb{R}\}. \end{aligned}$$

We may again represent $AS(\epsilon,\ \Phi)$ by a subset of $L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$. The set $AS(\epsilon,\ \Phi)$ is then compact and its closed convex hull is equal to the convex hull of the extremal points of $AS(\epsilon,\ \Phi)$. These now have densities of the form

$$f_\lambda(x) = \begin{cases} (1-\epsilon)\varphi(x), & \text{if } x < x_l \text{ or } x_u < x, \\ (1-\epsilon)\varphi(x_l)\exp(-x_l(x - x_l)), & \text{if } x_l < x < \lambda, \\ (1-\epsilon)\varphi(x_u)\exp(-x_u(x - x_u)), & \text{if } \lambda < x < x_u. \end{cases}$$

Again, $\lambda = (x_l + x_u)/2$ and in order for f_λ to be a density, we require

$$\begin{aligned} (1-\epsilon) + (1-\epsilon)\varphi(x_u) + (1-\epsilon)(\varphi(x_l)/x_l - \varphi(x_u)/x_u) \\ - (1-\epsilon)\varphi(x_l)\exp(-x_l(\lambda - x_l)/2)(1/x_l - 1/x_u) = 1. \end{aligned}$$

It is obvious that if $F \in \mathcal{F}$ implies $F^m \in \mathcal{F}$, then the shape selected by (3.2) will be symmetric (define F^m by $f^m(x) = f(-x)$). This is the case for $AS(\epsilon,\ \Phi)$ and Fig. 3 shows $(-\log g^*)$, the approximate solution to (3.2) when $\mathcal{F} = \text{co}(AS(\epsilon,\ \Phi))$ with $\epsilon = 10\%$. An approximation was again computed using a discrete set of extremal points. We took 37 distinct values for λ, namely 0.0(0.1)0.5, 1.0(0.5)6.5, as well as their negative counterparts. As we just remarked, the solution is symmetric. When asymmetry is not ruled out, the selected shape has heavier shoulders than was the case with the solution from last section (see Fig. 3).

The asymmetric model can also serve as an example for the more general approach to the selection of models. Suppose we wish to select two shapes G_1, G_2 simultaneously. We have already indicated in Section 3 that in this case a minimax problem for pairs of shapes is appropriate. A possible choice for the corresponding disparity function is $d(F;\ G_1) \wedge d(F;\ G_2)$. In the case of the Kullback-Leibler divergence this leads to

$$\min_{(G_1,G_2)\in\mathcal{F}\times\mathcal{F}} \max_{F\in\mathcal{F}} I(F;\ G_1) \wedge I(F;\ G_2). \tag{4.1}$$

This problem can be tackled with the methods of Section 3. The solution of (4.1) is equal to the solution of

$$\min_{G_1\in\mathcal{F}_1} \max_{F\in\mathcal{F}_1} I(F;\ G_1), \qquad \min_{G_2\in\mathcal{F}_2} \max_{F\in\mathcal{F}_2} I(F;\ G_2) \tag{4.2}$$

with $\mathcal{F}_1 = \{F \in \mathcal{F}: I(F;\ G_1) < I(F;\ G_2)\}$, and $\mathcal{F}_2$ correspondingly. We know that each of the two subproblems in (4.2) has a unique solution for any choice of two compact convex subsets $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$. However, the solution of (4.2) only solves (4.1) if

$$I(F;\ G_1) = I(F;\ G_2), \qquad \forall F \in \mathcal{F}_1 \cap \mathcal{F}_2.$$

Example. In the case of $\text{co}(AS(\epsilon,\ \Phi))$ it is easy to guess the proper partition $\mathcal{F}_1 \cup \mathcal{F}_2 = \text{co}(AS(\epsilon,\ \Phi))$. Recall that the extremal points of $\text{co}(AS(\epsilon,\ \Phi))$ are $\{f_\lambda: \lambda \in \mathbb{R}\}$. These will be separated into $\{f_\lambda: \lambda \leq 0\} \cup \{f_\lambda: \lambda \geq 0\}$, thus defining the partition $\mathcal{F}_1 \cup \mathcal{F}_2$. The convex set $\mathcal{F}_1$ has, of course, additional extremal points besides $\{f_\lambda: \lambda \leq 0\}$. All of these additional extremals have the form $wf_\lambda + (1-w)f_\mu$ with $\lambda < 0$ and $\mu > 0$ and it is easy to find an equation for w. First of all, note that because of symmetry we know that the solution of (4.1) consists of a pair $(G,\ G^m)$. (Recall that $g^m(x) = g(-x)$.) The newly created extremal points must satisfy

$$I(F;\ G) = I(F;\ G^m).$$

In other words

$$\int_{-\infty}^{\infty} \log gf\,dx = \int_{-\infty}^{\infty} \log g^m f\,dx = \int_{-\infty}^{\infty} \log gf^m dx.$$

When $f = wf_\lambda + (1-w)f_\mu$, this yields

$$w = \frac{\displaystyle\int_{-\infty}^{\infty} \log g(f_{-\mu} - f_\mu)dx}{\displaystyle\int_{-\infty}^{\infty} \log g((f_\lambda - f_{-\lambda}) + (f_{-\mu} - f_\mu))dx},$$

using the fact $f_\lambda^m = f_{-\lambda}$. An approximate solution for $n = 10\%$ was obtained using the discrete extremal set $\{f_\lambda: \lambda \in \Lambda \text{ and } -\lambda \in \Lambda\}$ with $\Lambda = \{0.05,\ 0.3,\ 0.6,\ 1.0,\ 2.0,\ 3.0,\ 5.0,\ 6.0,\ \infty\}$. Because we now have to deal with a partitioned set, there are 90 extremal points to be considered, even in this discrete version of the problem. To gain an impression of the degree of asymmetry present in the solution, one has to consult Fig. 4, where the score functions for the various estimators discussed in this paper are shown. The asymmetry present in the solution is quite marked, the score function is closer to the linear score function of the arithmetic mean on one side.

5. Score functions

If we intend to use the models proposed in this paper, it becomes necessary to not only know $-\log(\text{density})$, but also the derivative of that function. Since we were unable to solve the minimax problems analytically, and had to use numerical analysis on a discretized set of extremals, we only have approximate solutions. These approximations will typically not have smooth densities due to the discretization. Some nonparametric smoothing is appropriate in this case. Figure 4 shows plots of the score function associated with the minimax problems we discussed in the previous sections. The method of smoothing we used was a (cubic) spline smoother with smoothing parameter (sum of squared error)/(sum of squared error from the least squares line) = 0.01. Care was taken, not to evaluate the approximate solutions near points, where these solutions have discontinuous first derivative.

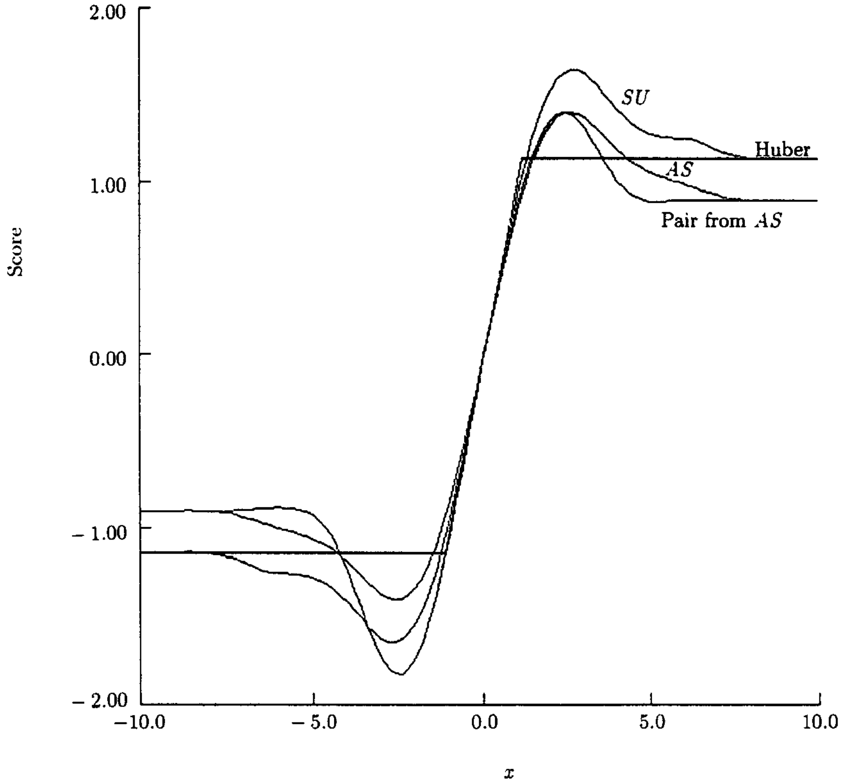

Fig. 4. Various smoothed score functions.

The comments that were made when we discussed plots of $-\log(\text{density})$ are confirmed in these graphs.

6. Concluding remarks

Robust statistics as used in the literature deals with the stability of inferential procedures. The theoretical paradigms for robustness have been almost exclusively concerned with tail heaviness of error distributions in measurement models. Furthermore, they have concentrated on defining desirable properties of inferential procedures. In this framework of thinking one chooses, from all the estimators having these desirable properties, the one which is optimal. In this paper we discuss a different approach to the robustness problem which is similar to the idea in Huber (1964). In our framework, the theoretical statistician must "model" the possible deviations from the idealistic, i.e. he must know what he wants to be robust against. The resulting model is typically very large, it usually contains an infinite-dimensional nuisance parameter. Given such circumstances, robust behavior gives an alternative to the typical reaction of statisticians, namely to estimate every unknown parameter. Robust behavior consists in choosing a single, representative value for that nuisance parameter. It is clear from this discussion that robustness is useful only when the estimation of the nuisance parameter is difficult. In addition, the nuisance parameter must be such that the outcome of the inference process is influenced by its value. We think that the robust selection can conveniently be described with the help of a disparity function, i.e. a real-valued function defined for pairs of nuisance parameters. Both of these elements of robust theory, namely a nonparametric supermodel, as well as a disparity function, are necessary to gain full generality.

It may often be relatively straightforward to define appropriate supermodel and disparity. But solving the corresponding minimax problems is far from easy. We saw some examples involving the Kullback-Leibler divergence as a disparity function defined for pairs of distributional shapes. Note however that once we solved such a model selection problem, the solution may be useful for many similar situations.

If one accepts the idea of a representing model, then the concept of a representing pair of models is a straightforward generalization. That point is also raised in the paper. We show that in the case of a supermodel containing asymmetric shapes, the single representer will often be symmetric, whereas the representing pair will be a pair of asymmetric shapes.

Acknowledgement

The authors would like to thank a referee for the detailed comments which led to a substantial improvement in the presentation, as well as a correction of the proof of Proposition 3.2.

REFERENCES

Beran, R. (1978). An efficient and robust adaptive estimator of location, *Ann. Statist.*, **6**, 292–313.

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

Hampel, F. R. (1971). A general qualitative definition of robustness, *Ann. Math. Statist.*, **42**, 1887–1896.

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

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

Huber, P. J. (1964). Robust estimation of a location parameter, *Ann. Math. Statist.*, **35**, 73–101.

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

Kullback, S. (1968). *Information Theory and Statistics*, Dover, New York.

Morgenthaler, S. (1986*a*). Asymptotics for configural location estimators, *Ann. Statist.*, **14**, 174–187.

Morgenthaler, S. (1987*b*). Confidence intervals for location, *J. Amer. Statist. Assoc.*, **81**, 518–525.

Rockafellar, R. T. (1971). Integrals which are convex functionals, II, *Pacific J. Math.*, **39**, 439–468.

Stone, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.*, **3**, 267–284.