Huber's Minimax Approach in Distribution Classes with Bounded Variances and Subranges with Applications to Robust Detection of Signals

Georgy Shevlyakov*

Department of Information and Communications,
Gwangju Institute of Science and Technology (GIST),
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Republic of Korea
E-mail: shev@gist.ac.kr

Kiseon Kim

Department of Information and Communications,
Gwangju Institute of Science and Technology (GIST),
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Republic of Korea
E-mail: kskim@gist.ac.kr

Abstract

A brief survey of former and recent results on Huber's minimax approach in robust statistics is given. The least informative distributions minimizing Fisher information for location over several distribution classes with upper-bounded variances and subranges are written down. These least informative distributions are qualitatively different from classical Huber's solution and have the following common structure: (i) with relatively small variances they are short-tailed, in particular normal; (ii) with relatively large variances they are heavy-tailed, in particular the Laplace; (iii) they are compromise with relatively moderate variances. These results allow to raise the efficiency of minimax robust procedures retaining high stability as compared to classical Huber's procedure for contaminated normal populations. In application to signal detection problems, the proposed minimax detection rule has proved to be robust and close to Huber's for heavy-tailed distributions and more efficient than Huber's for short-tailed ones both in asymptotics and on finite samples.

Keywords: minimax approach, least informative distributions, robust detection.

AMS Classification: 62F35, 62P35

Running Head: Huber's Minimax Approach ...

* Corresponding author

1 Introduction

1.1 Preliminaries

About forty years has passed since publishing of the pioneer works of J. W. Tukey [1], P. J. Huber [2], and F. R. Hampel [3]. These outstanding personalities defined a new area of mathematical statistics called robust statistics which has been intensively developed since the sixties and is rather definitely formed by present.

The basic reasons of research in this field are of a general mathematical character. "Optimality" and "stability" are the mutually complementary characteristics of nearly all mathematical procedures. It is a well-known fact that the behavior of many optimal decisions is rather sensible to "small deviations" from prior assumptions. In mathematical statistics, the remarkable example of such unstable optimal procedure is given by the LS method: its performance may become extremely poor under small deviations from normality.

Roughly speaking, robustness means stability of statistical inference under the variations of the accepted distribution models.

Nearly at the same time with robust statistics, there appeared another direction in statistics called exploratory or probability-free data analysis that also partly originated from J. W. Tukey [4]. By definition, data analysis techniques aim at practical problems of data and signal processing. Although robust statistics involves mathematically highly refined asymptotic tools, robust methods exhibit a satisfactory behavior on small samples being quite useful in applications.

This paper represents new results having definite accents both on theoretical aspects of robustness and practical needs of signal and data analysis technologies. In particular, we restrict ourselves to Huber's minimax approach in robustness [5] that is applied to the simplest problem of detection of a constant signal in not so simple distribution models with the restrictions upon noise variances and subranges. We show that the minimax approach has certain reserves both in theory and applications: first, for improving its "optimality" composition with retaining of the "stability" one, and second, for effective performing of "on line" signal processing procedures.

1.2 Huber's minimax approach

Now we briefly comment on Huber's minimax approach in robust statistics. In general, the minimax principle aims at the least favorable situation for which it suggests the best solution. Thus, in some sense, this approach provides a guaranteed result, possibly too pessimistic.

To estimate a location parameter θ, Huber introduced the class of M-estimators $\widehat{\theta}_N$ such that

$$\widehat{\theta}_N = \arg\min_{\theta} \sum_{i=1}^{N} \rho(x_i - \theta) \quad \text{or} \quad \sum_{i=1}^{N} \psi(x_i - \widehat{\theta}_N) = 0, \tag{1}$$

where the sample $x_1, \ldots, x_N$ is taken from the distribution with density $f(x-\theta)$ belonging to a certain class $\mathcal{F}$, $\rho(z)$ is a loss function with the derivative $\psi(z) = \rho'(z)$ called a score function [2, 5].

Under rather general conditions of regularity put upon the distribution class $\mathcal{F}$ and the class Ψ of admissible score functions ψ (these conditions will be specified later), the M-estimators are consistent and asymptotically normal with variance

$$V(\psi, f) = \int_{-\infty}^{\infty} \psi^2(x) f(x)\, dx \left[\int_{-\infty}^{\infty} \psi'(x) f(x)\, dx \right]^{-2}. \tag{2}$$

Huber proposed to use the supremum of asymptotic variance $V(\psi^*, f^*) = \sup_{f \in \mathcal{F}} V(\psi^*, f) \geq V(\psi^*, f)$ as a measure of robustness of the optimal M-estimator with the score function ψ^*: the less the range of the optimal estimator variance $V(\psi^*, f)$ over the class $\mathcal{F}$, the more robust is this estimator, and vice versa.

The proposed measure of robustness allows to design reasonable robust estimators by minimizing the supremum of asymptotic variance $V(\psi^*, f^*) = \inf_{\psi \in \Psi} \sup_{f \in \mathcal{F}} V(\psi, f)$.

For M-estimators of location, Huber's minimax approach is reduced to the determination of a least informative distribution density f^* minimizing the Fisher information for location $I(f)$ over a given class $\mathcal{F}$ of distribution densities with the subsequent application of the maximum likelihood principle

$$f^* = \arg\min_{f \in \mathcal{F}} \int_{-\infty}^{\infty} \left[f'(x)/f(x) \right]^2 f(x)\, dx, \quad \rho^*(z) = -\log f^*(z), \quad \psi^*(z) = -{f^*}'(z)/f^*(z). \tag{3}$$

Furthermore, the saddle-point pair (ψ^*, f^*) provides the guaranteed level of the accuracy of estimation

$$V(\psi^*, f) \le V(\psi^*, f^*) = 1/I(f^*) \quad \text{for} \quad \text{all} \quad f \in \mathcal{F}.$$

1.3 Least informative distributions: classical results

Within the minimax approach, the choice of a distribution class determines all the subsequent stages and the qualitative character of the corresponding robust procedure. In its turn, the choice of a distribution class depends either on the available prior information about data distributions, or on the possibilities of getting this information from the data sample.

Being historically the first [2], various ε-neighborhoods of normal distribution are not the only models of interest. Firstly, we may consider the ε-neighborhoods of other distributions, say, the uniform, Laplace, or Cauchy. Certainly, the reasons to introduce such classes are obviously weaker as compared to that based on normal distribution, but nevertheless, they can be. Secondly, in applications rather often there exist a prior information about the dispersion of a distribution, about its central part and/or its tails, about the moments and/or subranges of a distribution. The empirical distribution function and relative estimators of a distribution shape (quantile functions and their approximations, histograms, kernel estimators) along with their confidence boundaries give other examples. It seems reasonable to use such information in the minimax setting by introducing the corresponding distribution classes $\mathcal{F}$ in order to increase the efficiency (the "optimality" composition) of robust minimax procedures.

In what follows, we deal with distribution densities satisfying the following regularity conditions:

($F1$) f is symmetric and unimodal.

($F2$) f is twice continuously differentiable and $f(x) > 0$ for all x in $\mathbf{R}^+ = (0, \infty)$.

($F3$) the Fisher information for location $I(f) = \int_{-\infty}^{\infty} [f'(x)/f(x)]^2 f(x)\, dx$ satisfies $0 < I(f) < \infty$.

Obviously, also the conditions of non-negativeness and normalization hold $f(x) \ge 0, \quad \int_{-\infty}^{\infty} f(x)\, dx = 1$. For sake of brevity, we will not write these conditions any time we define a distribution class.

Now we enlist the classical results on the least informative distributions for several distribution classes which seem natural and convenient for the description of a prior knowledge about data distributions.

The class $\mathcal{F}_1$ of nondegenerate distributions (with a bounded density value at the center of symmetry):

$$\mathcal{F}_1 = \{f\colon\ f(0) \ge 1/(2a) > 0\}, \qquad f_1^*(x) = L(x; 0, a) = (2a)^{-1} \exp(-|x|/a). \tag{4}$$

In this case, the least informative density is the Laplace, hence we have the sign score function $\psi_1^*(z) = \mathrm{sgn}(x)/a$ and the sample median as the optimal M-estimator. This class is introduced in [6]. It is one of the most wide classes: any unimodal distribution density with a nonzero value at the center of symmetry belongs to it. The parameter a of this class characterizes the dispersion of the central part of a distribution, in other words, a is the upper bound for that dispersion. The condition of belonging to this class is very close to the complete lack of information about an underlying distribution.

The class $\mathcal{F}_2$ of distributions with an upper-bounded variance:

$$\mathcal{F}_2 = \left\{f\colon\ \sigma^2(f) = \int_{-\infty}^{\infty} x^2 f(x)\, dx \le \overline{\sigma}^2\right\}, \qquad f_2^*(x) = N(x; 0, \overline{\sigma}) \tag{5}$$

The least informative density is normal with the corresponding linear score function $\psi_2^*(z) = z/\overline{\sigma}^2$ and the sample mean as the optimal M-estimator. This class is considered in [7]. All distributions with upper-bounded variances are members of this class. Obviously, the heavy-tailed Cauchy-type distributions do not belong to it.

The class $\mathcal{F}_3$ of approximately normal distributions, or the gross error model, or the class of ε-contaminated normal distributions [2]:

$$\mathcal{F}_3 = \{f\colon f(x) = (1-\varepsilon)N(x; 0, \sigma_N) + \varepsilon h(x),\ 0 \le \varepsilon < 1\}, \tag{6}$$

where $N(x; 0, \sigma)$ is normal density, $h(x)$ is an arbitrary density, and ε is a contamination parameter characterizing the fraction of contamination and the level of the uncertainty of information about the shape of the underlying distribution.

In this case, the least informative density consists of two parts: the normal in the center and the exponential tails given by

$$f_3^*(x) = \begin{cases} (1-\varepsilon)N(x;0,\sigma), & \text{for} \quad |x| \le k\,\sigma, \\ (1-\varepsilon)(2\,\pi)^{-1/2}\sigma^{-1}\exp\left(-k\,\sigma|x| + k^2\sigma^2/2\right), & \text{for} \quad |x| > k\,\sigma, \end{cases} \tag{7}$$

where the dependence $k = k(\varepsilon)$ is tabulated (see [5], p. 87). The optimal score function has the following limited linear form $\psi_3^*(z) = \max\left[-k\,\sigma, \min\left(z/\sigma^2, k\,\sigma\right)\right]$ with the trimmed mean as the optimal M-estimator of location [2]. Note that the limit cases of Huber's optimal score function $\psi_3^*(z)$ are the linear function $\psi_3^*(z) = z/\sigma^2$ with $\varepsilon = 0$ and the sign function $\psi_3^*(z) = k\,\sigma\,\mathrm{sgn}(z)$ as $\varepsilon \to 1$ implying the sample mean and sample median as the optimal statistics, respectively.

The class $\mathcal{F}_4$ of finite distributions:

$$\mathcal{F}_4 = \left\{ f\colon \int_{-l}^{l} f(x)\,dx = 1 \right\}, \qquad f_4^*(x) = \begin{cases} l^{-1}\cos^2(\pi z/(2l)), & |z| \le l, \\ 0, & |z| > l. \end{cases}$$

The restriction on this class defines the boundaries of the data (i.e., $|x| \le l$ holds with probability one), and there is no more information about the distribution. The score function is unbounded: $\psi_4^*(z) = \tan(\pi z/(2l))$ for $|z| \le l$.

The class $\mathcal{F}_5$ of approximately finite distributions:

$$\mathcal{F}_5 = \left\{ f\colon \int_{-l}^{l} f(x)\,dx \ge 1-\beta \right\}. \tag{8}$$

The parameters $l > 0$ and β, $0 \le \beta < 1$, are given; the latter characterizes the degree of closeness of $f(x)$ to a finite distribution density. The restriction on this class means that the inequality $|x| \le l$ holds with probability at least equal $1-\beta$. Moreover, the characterization condition (8) of approximate finiteness can be rewritten as the restriction upon the distribution subrange: $F^{-1}(1-\beta/2) - F^{-1}(\beta/2) \le 2\,l$.

The least informative density is given by

$$f_5^*(x) = \begin{cases} A_1\cos^2(B_1 x), & |x| \le l, \\ A_2\exp(-B_2|x|), & |x| > l, \end{cases} \tag{9}$$

where the constants A_1, A_2, B_1, and B_2 are determined from the simultaneous equations characterizing the restrictions of the class $\mathcal{F}_5$, namely the conditions of normalization and approximate finiteness, and the conditions of smoothness at $x = l$:

$$\int_{-\infty}^{\infty} f_5^*(x)\,dx = 1, \quad \int_{-l}^{l} f_5^*(x)\,dx = 1-\beta, \quad f_5^*(l-0) = f_5^*(l+0), \quad {f_5^*}'(l-0) = {f_5^*}'(l+0).$$

The details on this system of equations for the parameters A_1, A_2, B_1, and B_2 can be found in [5, 10]. Note that the optimal score function $\psi_5^*(z)$ is bounded:

$$\psi_5^*(z) = \begin{cases} \tan(B_1 z), & |z| \le l, \\ \tan(B_1 l)\mathrm{sgn}(z), & |z| > l. \end{cases}$$

Apparently, the class $\mathcal{F}_4$ of finite distributions is the particular case of the class $\mathcal{F}_5$. These classes are considered in [5, 8, 9].

Remark 1 *The least informative density in the class $\mathcal{F}_1$ of nondegenerate distributions is the special limit case of the optimal solution in the class of approximately finite distributions as $l \to 0,\ 1-\beta \to 0,\ \dfrac{1-\beta}{2l} \to \dfrac{1}{2a}$.*

Remark 2 *As it can be seen from the aforementioned results, the minimax approach does not necessarily imply the boundness of an optimal score function and therefore, the robustness of a solution, like in the classes $\mathcal{F}_2$ and $\mathcal{F}_4$. On the other hand, the lack of stability of these solutions in heavy-tailed models is compensated by their higher efficiency in short-tailed ones.*

In the sequel, we consider the classes which are the intersections of the aforementioned: $\mathcal{F}_{12} = \mathcal{F}_1 \cap \mathcal{F}_2$ and $\mathcal{F}_{25} = \mathcal{F}_2 \cap \mathcal{F}_5$. Furthermore, we show that the additional restrictions, in other words, the additional information about data distributions, allows to enhance the efficiency of robust procedures. These models were used for robust estimation of location, regression and autogression parameters [6, 9, 10]. Further we shall apply these models to the problems of signal detection to provide both high robustness of a solution for heavy-tailed noise densities and its high efficiency for short-tailed ones.

1.4 A note on robust detection of signals

Consider the problem of detection of a known signal θ in the additive i.i.d. noise $\{n_i\}_1^N$ with density f from a certain class $\mathcal{F}$. Given $\{x_i\}_1^N$, it is necessary to decide whether the signal θ is observed. The problem of detection is set up as the problem of hypotheses testing: $H_0 : x_i = n_i$ versus $H_1 : x_i = \theta + n_i$, $i = 1, \ldots, N$. Given a density f, the classical theory of hypotheses testing yields various optimal (in the Bayesian, minimax, Neyman-Pearson senses) algorithms for the solution of this problem: all the optimal algorithms are based on the value of the likelihood ratio (LR) statistic $T_N(\mathbf{x}) = \prod_i^N f(x_i - \theta)/f(x_i)$ that must be compared with a certain threshold. The differences between the aforementioned approaches result only in the values of a threshold [11].

In many practical problems of radio-location, acoustics, and communications, noise distributions are only partly known. For instance, it may be known that either the underlying density is approximately normal, or there is some information on its behavior in the central zone and on the tails, or an impulsive noise may distort the observed signal, etc. For these detection problems, Huber's minimax approach to hypotheses testing also can be used with the bounded LR statistics [5], and several robust alternatives to the classical methods have been proposed in [12, 13, 14], in which contaminated normal models were mostly used.

However, we are interested in models containing short-tailed noise densities with small variances as well as the heavy-tailed ones with large or even with infinite variances.

This paper is organized as follows. In Section 2, the least informative densities for different classes with bounded variances and subranges are enlisted. In Section 3, the minimum distance detection rule is proposed, its asymptotic minimaxity in the Huber sense is established, and the probability of detection error is obtained in a closed form. In Section 4, for the distribution class $\mathcal{F}_{12}$ with a bounded variance and density value at the center of symmetry, the performance of the proposed detection rule is compared with Huber's detector in asymptotics and on finite samples for the normal, Cauchy, and short-tailed exponential-power noise densities. In Section 5, the concluding remarks are made.

2 Least informative densities in the classes with bounded variances and subranges

2.1 The least informative density in the class $\mathcal{F}_{12}$

Consider the distribution class containing the restrictions of the both classes $\mathcal{F}_1$ and $\mathcal{F}_2$:

$$\mathcal{F}_{12} = \left\{ f\colon \quad f(0) \geq 1/(2a) > 0, \quad \sigma^2(f) = \int_{-\infty}^{\infty} x^2 f(x)\, dx \leq \overline{\sigma}^2 \right\}. \tag{10}$$

Note that this class comprises qualitatively different distribution densities, for example, the normal, Laplace, Cauchy (with $\overline{\sigma}^2 = \infty$), short-tailed densities close to the uniform, etc.

For the class $\mathcal{F}_{12}$, the least informative density simultaneously depends on the two parameters a and $\overline{\sigma}$ through the ratio $\overline{\sigma}^2/a^2$ naturally having the Laplace and normal densities as the particular cases [10].

Theorem 1 *In the class $\mathcal{F}_{12}$, the least informative density is of the form*

$$f_{12}^*(x) = \begin{cases} N(x;0,\overline{\sigma}), & \text{for } \overline{\sigma}^2/a^2 < 2/\pi, \\ WH(x;0,\nu,\overline{\sigma}), & \text{for } 2/\pi \leq \overline{\sigma}^2/a^2 \leq 2, \\ L(x;0,a), & \text{for } \overline{\sigma}^2/a^2 > 2, \end{cases} \tag{11}$$

where $N(x;0,\overline{\sigma})$ and $L(x;0,a)$ are the normal and Laplace densities, respectively, $WH(x;0,\nu,\overline{\sigma})$ being called the Weber-Hermite density is given by

$$WH(x;0,\nu,\overline{\sigma}) = \frac{\Gamma(-\nu)\sqrt{2\nu+1+1/S(\nu)}}{\sqrt{2\pi}\,\overline{\sigma}\,S(\nu)}\mathcal{D}_\nu^2\left(\frac{|x|}{\overline{\sigma}}\sqrt{2\nu+1+1/S(\nu)}\right) \tag{12}$$

with the real-valued shape parameter ν that takes its values in $(-\infty, 0]$ and depends on the ratio of the parameters $\overline{\sigma}$ and a as follows

$$\frac{\overline{\sigma}}{a} = \frac{\sqrt{2\nu+1+1/S(\nu)}\Gamma^2(-\nu/2)}{\sqrt{2\pi}\,2^{\nu+1}\,S(\nu)\,\Gamma(-\nu)}.$$

Further, $\mathcal{D}_\nu(\cdot)$ are the Weber-Hermite functions or the functions of the parabolic cylinder [15], $S(\nu) = [\psi(1/2-\nu/2) - \psi(-\nu/2)]/2$, and in this context, $\psi(x) = d\ln\Gamma(x)/dx$ is the digamma function.

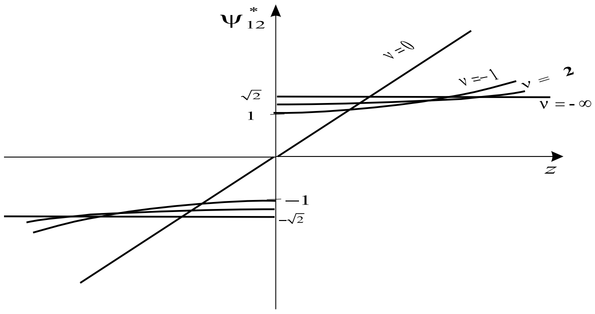

Fig. 1. The minimax score function for the class $\mathcal{F}_{12}$.

Now we comment on the structure of the optimal density. Note that the Weber-Hermite densities defined by Eq. (12) appear as the solution to the Euler-Lagrange equation for the variational problem of minimizing Fisher information (for details, see [10]). Further, the normal and Laplace densities arise as the particular cases of Eq. (12) when $\nu = 0$ and $\nu \to -\infty$, respectively. Different branches of $f_{12}^*(x)$ appear due to the degree in which the constraints are taken into account:

- For $\overline{\sigma}^2/a^2 < 2/\pi$ or with relatively small variances, only the restriction upon a variance does matter (it becomes the equality $\sigma^2(f_{12}^*) = \overline{\sigma}^2$, and the restriction on the density value at the center of symmetry takes the form of the strict inequality $f_{12}^*(0) > 1/(2a)$. Thus, the latter restriction can be omitted, and the class $\mathcal{F}_{12}$ is reduced to the class $\mathcal{F}_2$ with the corresponding least informative normal density.
- For $\overline{\sigma}^2/a^2 > 2$ or with relatively large variances, only the restriction upon the density value is essential: $f_{12}^*(0) = 1/(2a)$, and the restriction on a variance $\sigma^2(f_{12}^*) < \overline{\sigma}^2$ is omitted. Therefore, the class $\mathcal{F}_{12}$ is reduced to the class $\mathcal{F}_1$ with the corresponding least informative Laplace density.
- Finally, for $2/\pi \le \overline{\sigma}^2/a^2 < 2$ or with relatively moderate variances, the both restrictions become the equalities: $f_{12}^*(0) = 1/(2a)$ and $\sigma^2(f_{12}^*) = \overline{\sigma}^2$. Thus, they both must be taken into account, and the least informative density takes the most general Weber-Hermite form.

The optimal minimax score function is defined by the maximum likelihood principle (3)

$$\psi_{12}^*(z) = \begin{cases} z/\overline{\sigma}^2, & \overline{\sigma}^2/a^2 < 2/\pi, \\ -WH'(z;0,\nu,\overline{\sigma})/WH(z;0,\nu,\overline{\sigma}), & 2/\pi \le \overline{\sigma}^2/a^2 \le 2, \\ a^{-1}\mathrm{sgn} z, & \overline{\sigma}^2/a^2 > 2. \end{cases} \tag{13}$$

From Fig. 1 we can see the qualitative character of these compromise score functions: (i) they have a discontinuity at the origin; (ii) this discontinuity jump $2\psi_{12}^*(+0)$ is less than the corresponding jump of the limit score function $\psi_1^*(z) = a^{-1}\mathrm{sgn}(z)$ and it vanishes as $\nu \to \infty$; (iii) it can be shown that the asymptotes of the curves $\psi^* = \psi_{12}^*(z)$ go through the origin having lesser slopes than the slope of the limit score function $\psi_2^*(z) = z/\overline{\sigma}^2$; (iv) the asymptotic slope ${\psi^*}'_{12}(+\infty)$ vanishes as $\nu \to -\infty$.

Some numerical results are: (i) with $\nu = -1$ we have ${\psi^*}'_{12}(0) = 0.1$ and ${\psi^*}'_{12}(\infty) = 0.44$; (ii) with $\nu = -2$ we have ${\psi^*}'_{12}(0) = 0.04$ and ${\psi^*}'_{12}(\infty) = 0.25$.

This score function qualitatively differs from Huber's score function $\psi_3^*(z)$ that is optimal in the class $\mathcal{F}_3$, though both have the linear and sign forms as the limit cases.

2.2 The least informative density in the class $\mathcal{F}_{25}$

Consider the intersection of the classes $\mathcal{F}_2$ and $\mathcal{F}_5$ with the constraints on the variance and on the mass of the central part of a distribution

$$\mathcal{F}_{25} = \left\{ f\colon \sigma^2(f) \le \overline{\sigma}^2,\ \int_{-l}^{l} f(x)\,dx \ge 1 - \beta \right\}. \tag{14}$$

A lower bound upon the mass of the central zone of a distribution is equivalent to an upper bound upon its dispersion, or more precisely, upon the subrange of a symmetric distribution. In this case, the following result is true [10].

Theorem 2 *In the class of distributions $\mathcal{F}_{25}$, the least informative density is of the form*

$$f_{25}^*(x) = \begin{cases} N(x;0,\overline{\sigma}), & \overline{\sigma}^2 \le k_1 l^2, \\ \overline{f}_{25}^*(x), & k_1 l^2 < \overline{\sigma}^2 \le k_2 l^2, \\ f_5^*(x), & \overline{\sigma}^2 > k_2 l^2, \end{cases} \tag{15}$$

where

- $N(x;0,\overline{\sigma}) = f_2^*(x), f_5^*(x)$, *and* $\overline{f}_{25}^*(x)$ *are the least informative distribution densities in the classes* $\mathcal{F}_2$, $\mathcal{F}_5$, *and* $\overline{\mathcal{F}}_{25}$ *(the class* $\overline{\mathcal{F}}_{25}$ *is defined as in Eq. (14) but with the restrictions in the form of equalities);*
- *the switching parameters* k_1 *and* k_2 *depend on the parameters of the class* $\mathcal{F}_{25}$

$$\sigma^2(f_5^*) = \int_{-\infty}^{\infty} x^2 f_5^*(x)\,dx = k_2 l^2, \qquad \frac{1}{\sqrt{2\pi}\sqrt{k_1}\,l}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2k_1 l^2}\right) dx = 1-\beta;$$

- *the density* $\overline{f}_{25}^*(x)$ *is expressed through the Weber–Hermite functions*

$$\overline{f}_{25}^*(x) = \begin{cases} A_1[\mathcal{D}_{\nu_1}(B_1 x) + \mathcal{D}_{\nu_1}(-B_1 x)]^2, & |x| \le l, \\ A_2 \mathcal{D}_{\nu_2}^2(B_2|x|), & |x| > l; \end{cases} \tag{16}$$

The values of the parameters A_1, A_2, B_1, B_2, ν_1, *and* ν_2 *in (16) are determined from the equations:*

- *the normalization condition*

$$\int_{-\infty}^{\infty} \overline{f}_{25}^*(x)\,dx = 1, \tag{17}$$

- *the characterization restrictions of the class* $\overline{\mathcal{F}}_{25}$

$$\int_{-l}^{l} \overline{f}_{25}^*(x)\,dx = 1-\beta, \qquad \int_{-\infty}^{\infty} x^2 \overline{f}_{25}^*(x)\,dx = \overline{\sigma}^2; \tag{18}$$

- *the conditions of smoothness of the optimal solution at* $x = l$

$$\overline{f}_{25}^*(l-0) = \overline{f}_{25}^*(l+0), \quad \overline{f}^{*\prime}_{25}(l-0) = \overline{f}^{*\prime}_{25}(l+0); \tag{19}$$

- *the additional condition of optimality connecting the solutions in the zones* $|x| \le l$ *and* $|x| > l$

$$\int_{-l}^{l} x^2 \overline{f}_{25}^*(x)\,dx = {d_1^*}^2 = \arg \min_{{d_1^*}^2 \le \overline{\sigma}^2} I(f). \tag{20}$$

As with the solution of a similar problem in Subsection 2.1, the three branches of solution (15) appear according to the degree in which the restrictions of the class $\mathcal{F}_{25}$ are taken into account:

- for the first branch $N(x;0,\overline{\sigma}) = f_2^*(x)$, only the restriction on the variance in (14) does matter taking the form of the equality: $\sigma^2(f_2^*) = \overline{\sigma}^2$;
- for the third branch, the restriction on the central part of a distribution in (14) is essential, and the restriction on the variance has the form of the strict inequality: $\int_{-l}^{l} f_5^*(x)\,dx = 1-\beta, \quad \sigma^2(f_5^*) < \overline{\sigma}^2$;
- for the second, both restrictions have the form of equalities.

From Eqs. (17)–(20) we have the following particular cases of solution (16):

- for $\overline{\sigma}^2 = k_1 l^2$,

$$\overline{f}_{25}^*(x) = f_2^*(x) = N(x;0,\overline{\sigma}) = \frac{1}{\sqrt{2\pi}\overline{\sigma}} \exp\left(-\frac{x^2}{2\overline{\sigma}^2}\right), \qquad \nu_1 = \nu_2 = 0, \quad B_1 = B_2 = 1/\sigma;$$

- for $\overline{\sigma}^2 = k_2 l^2$,

$$\overline{f}^*_{25}(x) = f^*_5(x) = \begin{cases} A_1 \cos^2(B_1 x), & |x| \le l, \\ A_2 \exp(-B_2|x|), & |x| > l. \end{cases}$$

We now turn directly to the restrictions of the class $\mathcal{F}_{25}$:

- as $\overline{\sigma}^2 \to \infty$, the first restriction in (14) is inessential, and, in this case, we have the optimal solution in the class $\mathcal{F}_5$: $f^*_{25} = f^*_5$;
- as $l \to 0, 1-\beta \to 0$, and $(1-\beta)/(2l) \to 1/(2a)$, we obtain the restriction of the class $\mathcal{F}_{12}$: $f(0) \ge 1/(2a) > 0$, and the optimal solution $f^*_{25} = f^*_{12}$, respectively.

Consider the important particular case of the class $\mathcal{F}_{25}$ when the restriction on the central part of the distribution has the form of an upper bound upon the value of the distribution interquartile range

$$F^{-1}(3/4) - F^{-1}(1/4) \le \bar{b}.$$

Then from Theorem 2 we obtain the following result.

Corollary 1 *In the class* $\widetilde{\mathcal{F}}_{25} = \left\{f\colon F^{-1}(3/4) - F^{-1}(1/4) \le \bar{b},\ \sigma^2(f) \le \overline{\sigma}^2\right\}$, *the least informative density is of the form*

$$\tilde{f}^*_{25}(x) = \begin{cases} N(x; 0, \overline{\sigma}), & \overline{\sigma}^2 \le 0.548\overline{b}^2, \\ \overline{f}^*_{25}(x), & 0.548\overline{b}^2 < \overline{\sigma}^2 \le 0.903\overline{b}^2, \\ f^*_5(x), & \overline{\sigma}^2 > 0.903\overline{b}^2, \end{cases} \qquad (21)$$

where the parameters of the density $\overline{f}^*_{25}(x)$ *(16) are determined from Eqs. (17)–(20) with* $\beta = 1/2$.

The score functions $\widetilde{\psi}^*_{25}(z)$ for the three branches of solution (21) are exhibited in Fig. 2.

3 The minimax robust detection rule

To detect the constant signal θ (see the set-up of detection problem in Subsection 1.4), we use the following minimum distance detection rule

$$\sum_{i=1}^{N} \rho(x_i) \quad \lessgtr \quad \sum_{i=1}^{N} \rho(x_i - \theta), \qquad (22)$$

where $\rho(z)$ is a loss function characterizing the assumed form of a distance. This choice of a detection rule is mainly determined by the fact that it allows the direct use of Huber's minimax theory on $M-$ estimators of location. Further, it can be seen that the choice $\rho(z) = -\log f(z)$ makes the optimal LR test statistic minimizing the Bayesian risk with equal costs and prior probabilities of hypotheses.

To formulate the result, we must specify the regularity conditions put on densities f and score functions ψ. These conditions can be formulated in many ways, say, strengthening the conditions imposed on score functions and weakening those put on densities, and vice versa. We use the following set of assumptions sufficient for our aims, first, the conditions $(F1)$ - $(F3)$ on densities from Subsection 1.3 and, second, the conditions on score functions (for details, see [12], pp. 125-127):

(Ψ1) ψ is well-defined and continuous on $\mathbf{R}^+ \setminus C(\psi)$, where $C(\psi)$ is finite. At each point of $C(\psi)$ there exist finite left and right limits of ψ.

(Ψ2) The set $D(\psi)$ of points at which ψ is continuous but in which ψ' is not defined or not continuous is finite.

(Ψ3) $\int_{-\infty}^{\infty} \psi(x) f(x)\, dx = 0$.

(Ψ4) $\int_{-\infty}^{\infty} \psi^2(x) f(x)\, dx < \infty$.

(Ψ5) $0 < \int_{-\infty}^{\infty} \psi'(x) f(x)\, dx = -\int_{-\infty}^{\infty} \psi(x) f'(x)\, dx < \infty$.

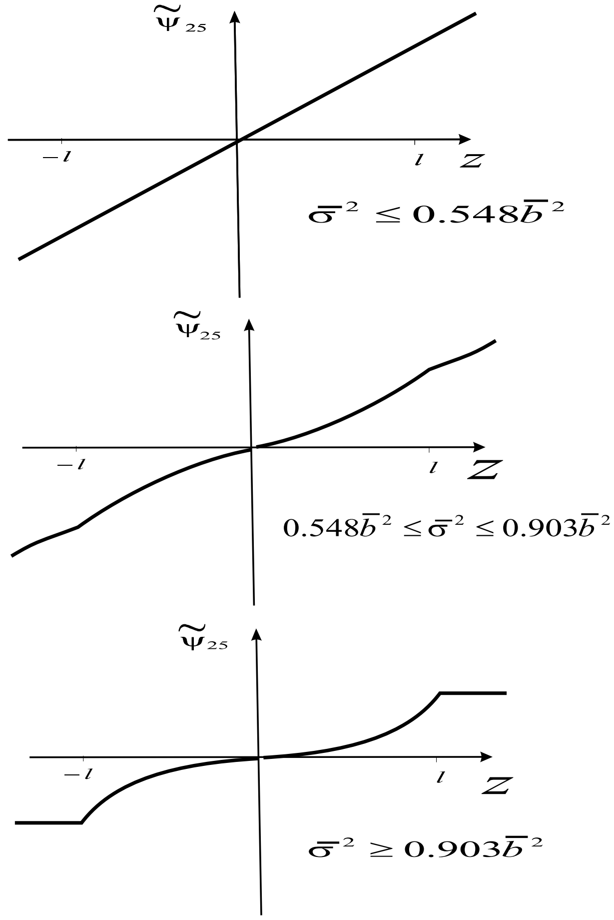

Fig. 2. The minimax score function for the class $\mathcal{F}_{25}$: $l = \bar{b}/2$.

THEOREM 3 *Let a density $f \in \mathcal{F}$ and a score function $\psi \in \Psi$ satisfy the assumptions (F1) – (F3), and (Ψ1) – (Ψ5), respectively.*

Then the probability of detection error for the minimum distance detection rule (22) takes the following form as $N \to \infty$:

$$P_E = \mathcal{Q}\left(\frac{A}{2}\frac{\int_{-\infty}^{\infty}\psi'(x)f(x)\,dx}{\sqrt{\int_{-\infty}^{\infty}\psi^2(x)f(x)\,dx}}\right), \tag{23}$$

where $\mathcal{Q}(z)$ is the complementary error function and the parameter A determines the amplitude of a weak signal as a decreasing sequence $\theta = \theta_N = A/\sqrt{N}$.

PROOF Now we derive Eq. (23), using the key idea of derivation of the asymptotic variance for M-estimators [2].

Write down the expression for the probability of detection error

$$P_E = P\left(\sum_i \rho(x_i) < \sum_i \rho(x_i - \theta)|H_1\right) = P\left(\sum_i \rho(\theta + n_i) - \sum_i \rho(n_i) < 0\right) \tag{24}$$

and denote the expression in parentheses as the function

$$q_N(\theta) = \sum_i \rho(\theta + n_i) - \sum_i \rho(n_i).$$

Its Taylor expansion has the form

$$q_N(\theta) = q(0) + q'(0)\theta + \frac{q''(\xi\theta)}{2}\theta^2, \qquad 0 \le \xi \le 1.$$

Furthermore,

$$q_N(\theta) = \theta \sum_i \psi(n_i) + \frac{\theta^2}{2} \sum_i \psi'(n_i + \xi\theta).$$

Now consider the case of a weak signal (the infinitesimal alternative)

$$\theta = \theta_N \to 0 \quad \text{as} \quad N \to \infty: \quad \theta_N = A/\sqrt{N}, \quad A > 0,$$

rewrite Eq. (24) as

$$P_E = P(q_N(\theta) < 0) = P\left(\frac{1}{\sqrt{N}} \sum_i \psi(n_i) < -\frac{A}{2}\frac{1}{N} \sum_i \psi'(n_i + \xi\, A/\sqrt{N})\right),$$

and consider the asymptotic behavior of the statistics:

$$T_N = \frac{1}{\sqrt{N}} \sum_i \psi(n_i) \quad \text{and} \quad T'_N = \frac{1}{N} \sum_{i=1}^N \psi'\left(n_i + \xi \frac{A}{\sqrt{N}}\right).$$

From assumptions (Ψ3) and (Ψ4) it follows that each summand $\psi(n_i)$ has mean 0 and variance $\int \psi^2(x) f(x)\, dx$. Hence by the CLT, the statistic T_N is asymptotically normal with mean 0 and variance $\int \psi^2(x) f(x)\, dx$. Further, by the law of large numbers, the statistic T'_N tends in probability to the positive constant $\int \psi'(x) f(x)\, dx > 0$ under assumption (Ψ5).

Therefore, we obtain the probability of detection error for the minimum distance detection rule in the form of Eq. (23)

$$P_E = P\left(T_N < -\frac{A\, T'_N}{2}\right) \to \mathcal{Q}\left(\frac{A}{2} \frac{\int_{-\infty}^{\infty} \psi'(x) f(x)\, dx}{\sqrt{\int_{-\infty}^{\infty} \psi^2(x) f(x)\, dx}}\right) \quad \text{as} \quad N \to \infty. \quad \text{Q.E.D.}$$

From Eq. (23) it follows that the minimax problem with respect to the probability of detection error

$$\min_{\psi \in \Psi} \max_{f \in \mathcal{F}} P_E(\psi, f) \tag{25}$$

is equivalent to Huber's minimax problem $\min_{\psi \in \Psi} \max_{f \in \mathcal{F}} V(\psi, f)$,.

Thus, all the results on the minimax variance estimation of location are also applicable in this case: the optimal loss function ρ^* in the minimum distance detector is defined by the maximum likelihood choice for the least informative density f^* minimizing the Fisher information for location $I(f)$ over the given class $\mathcal{F}$. Furthermore, the saddle-point pair (ψ^*, f^*) provides the guaranteed maximal level of the probability of detection error P_E

$$P_E(\psi^*, f) \le P_E(\psi^*, f^*) \quad \text{for all} \quad f \in \mathcal{F}.$$

Remark 3 *Since the signal energy E is equal to $\theta^2 N$, we have $A = \sqrt{E}$, and moreover, in the particular case of the unit noise variance, it can be written down as $A = \sqrt{SNR}$.*

4 Minimax detector performance

In this paper, we apply the obtained results for the class $\mathcal{F}_{12}$ to the aforementioned detection problem. As the optimal minimax detector (22) involves the loss function $\rho^*_{12}(u)$ which is based on the analytically complicated Weber–Hermite functions, we use the low-complexity L_p-norm approximations to it in the form

$$\sum_{i=1}^N |x_i|^{p^*} \quad < \quad \sum_{i=1}^N |x_i - \theta|^{p^*}.$$

The optimal values of the exponent p^* are expressed through the ratio $\overline{\sigma}^2/a^2$ and given by

$$p^* = \begin{cases} 5.333 - 7.610\, x + 3.731\, x^2, & 2/\pi \le x \le 1.35, \\ 2.656 - 1.646\, x + 0.409\, x^2, & 1.35 < x \le 2, \end{cases} \tag{26}$$

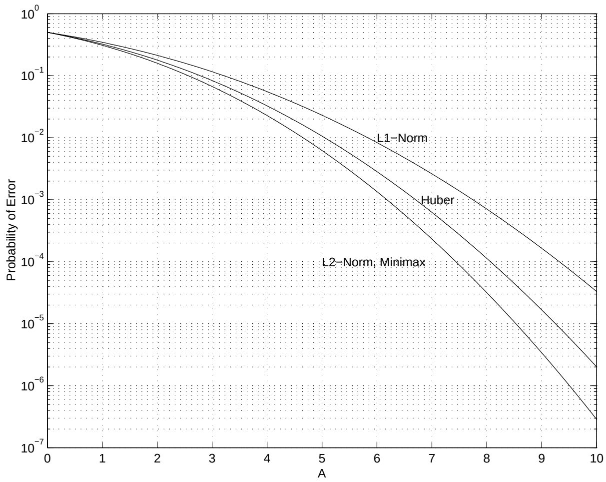

Fig. 3. Detection in the normal noise: asymptotics, $A = \sqrt{SNR}$.

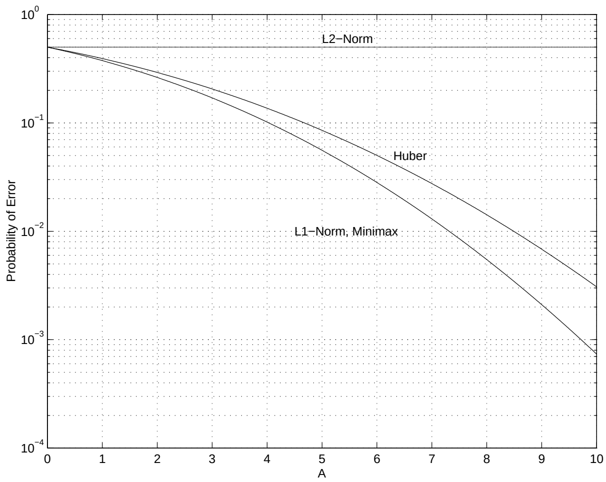

Fig. 4. Detection in the Cauchy noise: asymptotics, $A = \sqrt{E}$.

where $x = \overline{\sigma}^2/a^2$.

It can be shown that the decrease in power of the low-complexity detector as compared to the detector based on the Weber-Hermite functions does not exceed 2.5%.

The asymptotic performance of detection is measured by the probability of detection error P_E given by Eq. (23) that can be rewritten as

$$P_E = \mathcal{Q}\left(A\, I_1/(2\sqrt{I_2}\,)\right), \tag{27}$$

where $I_1 = \int_{-\infty}^{\infty} \psi'(x) f(x)\, dx$ and $I_2 = \int_{-\infty}^{\infty} \psi^2(x) f(x)\, dx$.

We compute the probability of error for the normal noise with the density $f(x) = N(x; 0, 1)$, the heavy-tailed Cauchy noise with the density

$$f(x) = C(x; 0, 1) = 1/[\pi(1 + x^2)],$$

and for the exponential-power noises with the densities

$$f(x; q, \beta) = \frac{q}{2\beta\Gamma(1/q)} \exp\left(-\frac{|x|^q}{\beta^q}\right),$$

where β is a scale parameter and q is a shape parameter. This formula describes a wide collection of symmetric unimodal densities: the Laplace density with $q = 1$, the Gaussian with $q = 2$, and the uniform with $q \to \infty$.

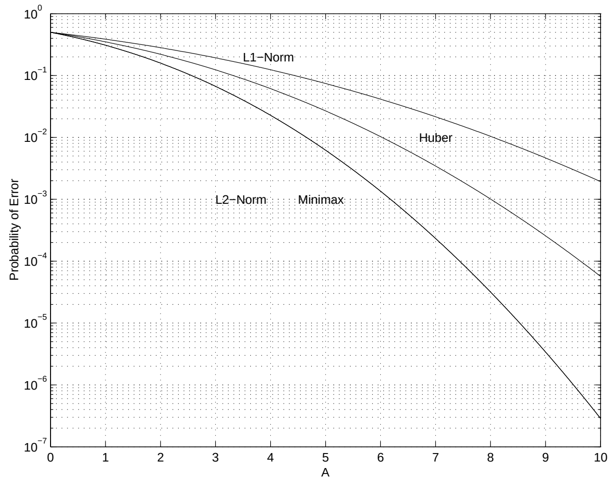

Fig. 5. Detection in the exponential-power noise: asymptotics, $q = 100$, $A = \sqrt{SNR}$.

Further, we compare the performance of the minimax detector with the L_1-, L_2-norm, and Huber's detectors. In the latter case, we consider the minimum distance detection rule (22) with Huber's optimal loss function $\rho_3^*(x) = -\log f_3^*(x)$ in the model of ε-contaminated normal distributions (6) with $\varepsilon = 0.1$.

To clarify the procedures of the choice of the minimax detector structure and of computing the probability of detection error, we consider the following examples.

Example 1. *The probability of detection error for the minimax and L_2-norm detectors in the normal noise.* The choice of the optimal structure is defined by the ratio σ^2/a^2. Subsequently, we have $f(0) = 1/\sqrt{2\,\pi}$, $\sigma^2(f) = 1$, $a = \sqrt{\pi/2}$, $\sigma^2/a^2 = 2/\pi$. Hence, the minimax detector is the minimum L_2-norm distance, the score function is linear $\psi^*(x) = x$, the integrals are $I_1 = 1$ and $I_2 = 1$, and thus, its probability of error is given by $P_E = \mathcal{Q}(A/2)$.

Example 2. *The probability of detection error for Huber's and L_1-norm detectors in the normal noise.* For $\varepsilon = 0.1$, the score function $\psi(x) = \psi_3^*(x) = \max\,[-1.14, \min\,(x, 1.14)]$, the integrals are evaluated numerically, and $P_E = \mathcal{Q}(0.461\,A)$. For the L_1-norm detector, the score function is the sign function $\psi(x) = \mathrm{sgn}(x)$, the integrals are $I_1 = 2f(0) = \sqrt{2/\pi}$ and $I_2 = 1$, and thus, we get $P_E = \mathcal{Q}(A/\sqrt{2\pi}) = \mathcal{Q}(0.399\,A)$.

The results of computing for the normal, Cauchy, and the close to the uniform the exponential-power ($q = 100$) density are exhibited in Fig. 1 – Fig. 3. These results are analyzed in Section 5.

The structure of the minimax detector is determined by Eq. (26) through the ratio $\overline{\sigma}^2/a^2$, and contrary to Huber's detector, the parameters $\overline{\sigma}^2$ and a^2 of the class $\mathcal{F}_{12}$ can be directly estimated from the sample. Taking into account the form of the middle branch of the minimax detection rule with the parameter $1 < p^* < 2$, when the both restrictions of the class $\mathcal{F}_{12}$ hold as equalities, we choose the estimates of variance and density as the characteristics of this class.

For the variance estimate, it is the sample variance $\widehat{\overline{\sigma}}^2 = N^{-1}\sum_{i=1}^N (x_i - \overline{x})^2$. For the parameter a, its estimate is based on the central order statistics $x_{(k)}$ and $x_{(k+1)}$ ($N = 2k$ or $N = 2k + 1$) and is given by

$$\widehat{a} = 1/[2\widehat{f}(0)] = [(N + 1)(x_{(k+1)} - x_{(k)})]/2.$$

On finite samples when $N = 20$ and $N = 100$, the performance of the minimax, Huber's, L_1- and L_2-norm detectors under the normal, Cauchy, and uniform noises was studied by Monte Carlo technique. The detection model was chosen consistently with the initial assumption of detection of a weak signal:

$$H_0 : x_i = n_i \quad \text{versus} \quad H_1 : x_i = \theta + n_i, \quad i = 1, \dots, N,$$

where the useful signal $\theta = \theta_N = A/\sqrt{N}$.

On small samples with $N = 20$, the results of modelling are discussed in Section 5. On large samples with $N = 100$, these results are close to the asymptotic results given by Eq. (27).

5 Concluding Remarks

The normal noise: large samples and asymptotics. As it can be seen from Fig. 1, the minimax detector coincides with the optimal L_2-norm detector both being better than Huber's.
The normal noise: small samples. On the contrary, the minimax detector is close in performance to the robust L_1-norm detector being slightly inferior to Huber's on small samples. Here we observe the so-called small size sample effect that arises due to the bias of the threshold statistic $\widehat{\widehat{\sigma}}^2/\widehat{a}^2$ on small samples [10].
The Cauchy noise. From Fig. 2 it can be seen that the L_2-norm detector naturally has the extremely poor performance in asymptotics (the qualitatively similar performance is also observed on small samples). The minimax, L_1-norm and Huber's detectors exhibit their good robust properties, the latter being inferior to the former.
The short-tailed noises (the exponential-power with $q = 100$ and uniform). In asymptotics, the L_2-norm and minimax detectors prove their superiority over Huber's and the L_1-norm detectors, but on small samples, the aforementioned small size sample effect reveals itself: the minimax detector is close in performance to the L_1-norm detector, and thus it is slightly inferior to Huber's.
Final remark. Our main aim is to show some new possibilities of Huber's minimax approach. In the problem of robust detection, the minimax detector designed for the distribution class with an upper-bounded variance demonstrates both high robustness in heavy-tailed noises and good efficiency in short-tailed noises on small and large samples.

References

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

[2] Huber, P.J.: Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101 (1964)

[3] Hampel, F.R., Contributions to the Theory of Robust Estimation, Ph.D. Thesis, University of California, Berkeley, 1968

[4] Tukey, J.W.: The future of data analysis. *Ann. Math. Statist.* **33**, 1-67 (1962)

[5] Huber, P.J., Robust Statistics, Wiley, New York, 1981

[6] Polyak, B.T., Tsypkin, Ya.Z.: Robust identification. In: *Identif. Syst. Parameter Estim., Part 1, Proc. 4th IFAC Symp.* Tbilisi, 203-224 (1976)

[7] Kagan, A.M., Linnik, Yu.V., and Rao, C.R., Characterization Problems in Mathematical Statistics, Wiley, New York, 1973

[8] Sacks, J., Ylvisaker, D.: A note on Huber's robust estimation of a location parameter. *Ann. Math. Statist.* **43**, 1068-1075, (1972)

[9] Tsypkin, Ya.Z., Foundations of Information Theory of Identification, Nauka, Moscow, 1984 (in Russian)

[10] Shevlyakov, G.L. and Vilchevski, N.O., Robustness in Data Analysis: criteria and methods, VSP, Utrecht, 2002

[11] Van Trees, H.L., Detection, Estimation and Modulation Theory, Part I, Wiley, New York, 1981

[12] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A., Robust Statistics, The Approach Based on Influence Functions, Wiley, New York, 1986

[13] Kassam, S.A., Poor, H.V.: Robust techniques for signal processing: a survey, *Proc. IEEE*, **73**, 433-481 (1985)

[14] Kassam, S.A., Signal Detection in Non-Gaussian Noise, Springer, Berlin, 1988

[15] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, New York: Dover, 1972.