PRAGUE STOCHASTICS 2006

Proceedings of the joint session of

7th Prague Symposium on Asymptotic Statistics

and

15th Prague Conference on Information Theory,
Statistical Decision Functions and Random Processes,

held in Prague from August 21 to 25, 2006

Organised by

Charles University
Faculty of Mathematics and Physics
Department of Probability and Mathematical Statistics
and
Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation
Department of Stochastic Informatics

Edited by Marie Hušková and Martin Janžura

ISBN 80-86732-75-4

Preface

Prague Stochastics 2006, held in Prague from August 21 to 25, 2006, is an international scientific meeting that continues the tradition of organising Prague conferences on stochastics, established here five decades ago. The first Prague Conference on Information Theory, Statistical Decision Functions and Random Process was initiated by Antonín Špaček in 1956. Prague Symposia on Asymptotic Statistics were founded by Jaroslav Hájek in 1973. This year, we are commemorating the 80th anniversary of the birth date of this untimely deceased outstanding scientist.

Traditionally, the scope of the proceedings, as well as the conference itself, is quite extensive; the topics range from classical to very up-to date ones. It covers both methodological and applied statistics, theoretical and applied probability and, of course, topics from information theory. We hope that all readers will find valuable contributions and a number of papers of their interest in this rich spectrum of scientific ideas.

The printed part contains the plenary and invited papers, and the list of all contributions published in the volume. The CD disc, attached as an official part of the book with the same ISBN code, contains all accepted papers.

The editors would like to express their sincere thanks to the authors for their valuable contributions, to the reviewers for prompt and careful reading of the papers, and to the organisers of the sections for the help with the entire reviewing process.

Our thanks also go to our colleagues, in particular to Pavel Boček and Tomáš Hobza, for their technical editorial work. Without their devotion and diligence, the proceedings would never be completed.

It is our pleasure to acknowledge that Prague Stochastics 2006 is held under the auspices of the Mayor of the City of Prague, the Bernoulli Society for Mathematical Statistics and Probability, and the Czech Statistical Society.

Prague, June 2006 *Marie Hušková, Martin Janžura*

Table of Contents

Plenary Lectures 1

Invited Papers 55

Contributed Papers (CD ROM) **225**

Probability metrics and robustness: Is the sample median more robust than the sample mean?

Evgueni Gordienko, Andrey Novikov

Abstract: In this paper, we consider a modified concept of qualitative robustness introduced by Hampel (see [5]). The modification takes into account the asymptotic normality of the estimators and allows to obtain quantitative estimates of the robustness in terms of suitable probability metrics. To this end, we measure deviations from the model distribution by metrics rather than by the level of "contamination" with a heavy-tail distribution.

We show that the "robustness" of the sample mean and the sample median statistics depends on the choice of the metrics. In particular, we obtain a quantitative estimate of the "robustness" of the sample mean, which is new in the case of models with non-Gaussian distributions.

MSC 2000: 62F35, 60E15
Key words: probability metric, sample mean, sample median, robust estimation

1 Introduction

In the paper of Hampel [5] the concept of qualitative robustness was introduced (see the definition in Section 3 below). According to his definition (which uses the Prokhorov metric) the sample mean is not robust, but the sample median is ([5], and §2.2 in [6]).

In this paper, we define an alternative concept of robustness based on the Kantorovich metric l (see Section 2). The use of the Kantorovich metric is justified by its closer relation to the mean absolute error of the estimators. To give a quantitative aspect to the robustness, we let it be related to some additional probability metric μ, and call this $(l-\mu)$-robustness.

The goal of the paper is to show that this type of robustness of estimators crucially depends on the choice of the metric μ. We give an example of probability metric μ for which the sample mean is $(l-\mu)$-robust but the sample median is not. The picture can turn over when using another metric (see Section 5).

The main result of the paper is the quantitative estimation of the $(l-\mu)$-robustness of the sample mean when μ is the maximum of the Kantorovich metric and the Zolotarev metric ζ_2 of order 2. More specifically, we show in Section 3 that

$$\sup_{n\geqslant 1}\sqrt{n}\,l(\bar{X}_n,\bar{\tilde{X}}_n)\leqslant c\max\{l(F,\tilde{F}),\zeta_2(F,\tilde{F})\},\tag{1}$$

where $\bar{X}_n$ and $\bar{\tilde{X}}_n$ are the sample means for respective samples from the distributions F and $\tilde{F}$, and where the constant c only depends on the model distribution F. If we admit $c = c(F, \tilde{F})$ (which is not interesting from the statistical point of view), the inequality of type (1) can be relatively easily derived from the properties of the metrics l and ζ_2. Some numerical examples of evaluation of the constant c can be found in Section 4.

It is worth mentioning that this result is only new in the case of non-Gaussian distribution F. Otherwise, inequality (1) can be obtained as a consequence of the known estimates of the rate of convergence in the central limit theorem.

In Section 5 we discuss the $(l - \mu)$-robustness of the sample median.

2 $(l - \mu)$-robustness of estimators of the mean value

Let $X, X_1, X_2, \dots, X_n$ and $\tilde{X}, \tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$ be random samples from distributions F and $\tilde{F}$, respectively.

For two probability measures F and $\tilde{F}$ let $\pi(F, \tilde{F})$ be the Prokhorov metric:

$$\pi(F, \tilde{F}) := \inf\{\epsilon : F(B) \leqslant \tilde{F}(B^\epsilon) + \epsilon \text{ for all Borel sets } B \subset \mathbb{R}\},$$

where $B^\epsilon = \{x \in \mathbb{R} : d(x, B) < \epsilon\}$. In what follows, we will be applying the notation of probability metric to *random variables* as well (for example $\pi(X, \tilde{X})$, etc.), taking this as the metric applied to their *distributions*, i.e. $\pi(X, \tilde{X}) \equiv \pi(F, \tilde{F})$.

The following Hampel's definition of qualitative robustness deals with a metric neighborhood in the space of distribution functions rather than the standard "contamination" neighborhood (see, for instance, [7], [8]).

Let T_n, $n \geqslant 1$ be some sequence of estimators of θ. According to [6] (§2.2), the sequence $\{T_n, n \geqslant 1\}$ is qualitatively robust at F if for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $\tilde{F}$

$$\pi(F, \tilde{F}) < \delta$$

entails

$$\sup_{n \geqslant 1} \pi(T_n(X_1, X_2, \dots, X_n), T_n(\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n)) < \epsilon.$$

As noted in [5, 6], the sample mean

$$T_n = \bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

is *not* qualitatively robust at any F.

However, the sample median is qualitatively robust at F if $F^{-1}(1/2)$ consists of only one point.

To introduce an alternative concept of qualitative robustness we will use the Kantorovich metric:

$$l(F, G) := \int_{-\infty}^{\infty} |F(x) - G(x)| dx.$$

It is known (see, e.g., [11]) that

$$l(X_n, X) \to 0 \text{ if and only if}$$

$$\pi(X_n, X) \to 0 \ \textit{and} \ E|X_n| \to E|X|, \quad n \to \infty, \tag{2}$$

so we will use a stronger metric for our definition.

Also, we will suppose that the following holds:

Assumption 1. $EX = E\tilde{X} = \theta$, $EX^2 < \infty$, $E\tilde{X}^2 < \infty$.

Let μ be any probability metric.

Definition 2.1. We say that the sequence of estimators $\{T_n, n \geqslant 1\}$ of the mean $\theta = EX$ is $(l - \mu)$-robust at F if for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $\tilde{F}$ satisfying Assumption 1

$$\mu(F, \tilde{F}) < \delta$$

entails

$$\sup_{n \geqslant 1} \sqrt{n} l(T_n(X_1, \ldots, X_n), T_n(\tilde{X}_1, \ldots, \tilde{X}_n)) < \epsilon. \tag{3}$$

Remark 2.2. Since (see, for instance, [10, 11])

$$l(\alpha X + b, \alpha Y + b) = \alpha l(X, Y), \tag{4}$$

where $\alpha \geqslant 0$, $b \in \mathbb{R}$ and X, Y are any random variables , we have

$$\sqrt{n} l(T_n(X_1, \ldots, X_n), T_n(\tilde{X}_1, \ldots, \tilde{X}_n) = l(\sqrt{n} T_n(X_1, \ldots, X_n), \sqrt{n} T_n(\tilde{X}_1, \ldots, \tilde{X}_n))).$$

Hence, condition (3) implicitly involves the asymptotic normality of the statistics T_n.

To advocate the above definition let us consider an example. Let

$$T_n(X_1, \ldots, X_n) = \bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

$$T_n(\tilde{X}_1, \ldots, \tilde{X}_n) = \bar{\tilde{X}}_n = \frac{\tilde{X}_1 + \tilde{X}_2 + \cdots + \tilde{X}_n}{n}$$

be the sample means corresponding to the samples $X_1, X_2, \ldots, X_n$ and $\tilde{X}_1, \tilde{X}_2, \ldots,$ $\tilde{X}_n$.

Suppose we measure the quality of the estimator by the mean absolute error, i.e. we define

$$\delta_n := E|\bar{X}_n - \theta|, \quad \tilde{\delta}_n := E|\bar{\tilde{X}}_n - \theta|,\ n \geqslant 1,$$

and let

$$\delta := \sup_{n \geqslant 1} \sqrt{n} \delta_n, \quad \tilde{\delta} := \sup_{n \geqslant 1} \sqrt{n} \tilde{\delta}_n, \tag{5}$$

By the Hölder inequality and Assumption 1, $\delta < \infty$, $\tilde{\delta} < \infty$.

Proposition 2.3.
$$|\delta - \tilde{\delta}| \leqslant \sup_{n \geqslant 1} \sqrt{n} l(\bar{X}_n, \bar{\bar{\tilde{X}}}_n).$$

Remark 2.4. If the statistics $\bar{X}_n$ is $(l - \mu)$-robust, and additionally we have a "robustness" inequality (as in the Theorem below):

$$\sup_{n \geqslant 1} \sqrt{n} l(\bar{X}_n, \bar{\bar{\tilde{X}}}_n) \leqslant c\mu(F, \tilde{F}), \tag{6}$$

then we get:
$$|\delta - \tilde{\delta}| \leqslant c\mu(F, \tilde{F}).$$

Below, we will prove (6) with $\mu = \max\{l, \zeta_2\}$, where ζ_2 is Zolotarev's metric of order 2, and we will show (using the same μ) that inequalities as in (6) are impossible if in place of the sample mean $\bar{X}_n$ we take the sample medians $\hat{X}_n$ (for symmetric distributions).

Proof of the Proposition 2.3. Let $n \geqslant 1$ be arbitrary but fixed. Then

$$|\delta_n - \tilde{\delta}_n| \leqslant E||\bar{X}_n - \theta| - |\bar{\bar{\tilde{X}}}_n - \theta|| \leqslant E|\bar{X}_n - \bar{\bar{\tilde{X}}}_n| = l(\bar{X}_n, \bar{\bar{\tilde{X}}}_n). \tag{7}$$

The last equality is true because the left-hand side of (7) depends only on the marginal distributions of X and $\tilde{X}$ and l is the minimal metric for the compound metric $L(X, Y) = E|X - Y|$ (see, e.g., [10, 11]). By the Kantorovich theorem (see [11]), the joint distribution of X and Y can always be chosen in such a way that $L(X, Y) = l(X, Y)$.

Thus,

$$|\sup_{n \geqslant 1} \sqrt{n}\delta_n - \sup_{n \geqslant 1} \sqrt{n}\tilde{\delta}_n| \leqslant \sup_{n \geqslant 1} \sqrt{n}|\delta_n - \tilde{\delta}_n| \leqslant \sup_{n \geqslant 1} \sqrt{n} l(\bar{X}_n, \bar{\bar{\tilde{X}}}_n) \leqslant c\mu(F, \tilde{F}).$$

□

3 Assumptions and the robustness inequality

Assumption 2. $E|X|^3 < \infty$.

Assumption 3. There is an integer $s \geqslant 1$ such that the random variable $X_1 + X_2 + \cdots + X_s$ has a bounded absolutely continuous density p_s. The derivative p_s' is bounded, $p_s' \in L_1(\mathbb{R})$, and for some $\alpha > 0$

$$\int_{|x| > \alpha n} |p_s'(x)| dx = O(n^{-1/2}) \quad \text{as} \quad n \to \infty. \tag{8}$$

Let $\zeta_2(X, Y)$ be the Zolotarev metric of order 2 [10, 12, 13]:

$$\zeta_2(X, Y) := \sup_{\phi \in D_2} |E\phi(X) - E\phi(Y)|,$$

where

$$D_2 := \{\phi : \mathbb{R} \to \mathbb{R} : |\phi'(x) - \phi'(y)| \leqslant |x - y|,\ x, y \in \mathbb{R}\}.$$

Theorem 3.1. *Let Assumptions 1–3 hold. Then there exists a constant $c > 0$, depending only on the d.f. F, such that*

$$\sup_{n \geqslant 1} \sqrt{n} l(\bar{X}_n, \bar{\tilde{X}}_n) \leqslant c \max\{l(F, \tilde{F}), \zeta_2(F, \tilde{F})\}, \tag{9}$$

where l and ζ_2 are, respectively, the Kantorovich and Zolotarev's (of order 2) metrics.

Remark 3.2. The result is new when F is non-Gaussian. If F is a Gaussian distribution, inequality (9) follows from the known estimates of the rate of convergence in the central limit theorem. A large part of the proof is to show that the constant c in (9) is completely determined by the model distribution F.

Remark 3.3. Under the hypotheses taken $l(F, \tilde{F}) < \infty$, $\zeta_2(F, \tilde{F}) < \infty$. Thus, if $\mu = \max\{l, \zeta_2\}$, then the statistics $\bar{X}_n$ is $(l - \mu)$-robust (with an estimate of robustness given in (9)).

For some particular d.f. F the constant c in (9) can be calculated (see the table in Remark 4.4).

It is easy to show that the sample mean $\bar{X}_n$ is not $(l - \mu)$-robust when μ is the total variation metric $\mu(X, \tilde{X}) = \sigma(X, \tilde{X})$ (and consequently, not for the uniform metric $\mu(F, \tilde{F}) = \sup_x |F(x) - \tilde{F}(x)|$, nor for the Prokhorov metric $\mu(F, \tilde{F}) = \pi(F, \tilde{F})$).

The following example shows that the same can be said in the case when $\mu(X, \tilde{X}) = l(X, \tilde{X})$.

Example 3.4. Let $X \sim N(0, 1)$ and $\epsilon > 0$ be arbitrary but fixed. Define

$$\tilde{X} = \begin{cases} X & \text{with probability} \quad 1 - \epsilon^{4/3} \\ \xi & \text{with probability} \quad \epsilon^{4/3}, \end{cases}$$

where

$$\xi = \begin{cases} 1/\epsilon & \text{with probability} \quad 1/2 \\ -1/\epsilon & \text{with probablity} \quad 1/2, \end{cases}$$

and ξ is independent of X (one could also use a ξ with a density).

It is easy to see that $\tilde{X} \Rightarrow X$ as $\epsilon \to 0$. Moreover,

$$E|\tilde{X}| = E|X|(1 - \epsilon^{4/3}) + \epsilon^{1/3} \to E|X|, \quad \epsilon \to 0.$$

Therefore $l(X, \tilde{X}) \to 0$ as $\epsilon \to 0$.

Obviously (in this example $\theta = 0$),

$$\delta = \sqrt{n} E|\bar{X}_n| = \sqrt{\frac{2}{\pi}}.$$

On the other hand,

$$\mathrm{Var}(\tilde{X}) = \sigma^2(\epsilon) = (1-\epsilon^{4/3}) + \epsilon^{-2/3} < \infty,$$

and for any fixed $\epsilon > 0$

$$\sqrt{n}\bar{\tilde{X}}_n \Rightarrow \eta_\epsilon$$

where $\eta_\epsilon \sim N(0, \sigma^2(\epsilon))$. Since the variances of the summands are finite, we have

$$E|\sqrt{n}\bar{\tilde{X}}_n| \to E|\eta_\epsilon| = \sqrt{\frac{2}{\pi}}\sigma(\epsilon), \quad \text{as} \quad n \to \infty.$$

so that

$$\tilde{\delta} = \sup_{n \geqslant 1} E|\sqrt{n}\bar{\tilde{X}}_n| \geqslant \sqrt{\frac{2}{\pi}}\sigma(\epsilon) \to \infty \quad \text{as} \quad \epsilon \to 0.$$

So we get $l(X, \tilde{X}) \to 0$, but $|\delta - \tilde{\delta}| \to \infty$ as $\epsilon \to 0$.

Now take into account the above Proposition.

4 The proof of the Theorem

Lemma 4.1. *Let X, Y, ξ be random variables such that*

a) ξ is independent of X and Y;

b) $EX = EY$; $EX^2 < \infty$, $EY^2 < \infty$;

c) the random variable ξ has a bounded absolutely continuous density f_ξ such that $f'_\xi \in L_1(\mathbb{R})$.

Then

$$l(X+\xi, Y+\xi) \leqslant ||f'_\xi||_{L_1}\zeta_2(X,Y), \tag{10}$$

where ζ_2 is Zolotarev's metric of order 2.

Proof.

$$\begin{aligned} l(X+\xi, Y+\xi) &= \int_{-\infty}^{\infty} dx \Big| \int_{-\infty}^{\infty} f_\xi(x-t)[F_X(t) - F_Y(t)]dt \Big| \\ &= \int_{-\infty}^{\infty} dx \Big| \int_{-\infty}^{\infty} f_\xi(x-t) d\Big[\int_{-\infty}^{t} [F_X(\tau) - F_Y(\tau)]d\tau\Big]\Big|. \end{aligned} \tag{11}$$

Since f_ξ is bounded and $EX = EY$ the integration by parts on the right-hand side of (11) yields:

$$l(X+\xi, Y+\xi) = \int_{-\infty}^{\infty} dx \Big| \int_{-\infty}^{\infty} f'_\xi(x-t)\Big[\int_{-\infty}^{t} [F_X(\tau) - F_Y(\tau)]d\tau\Big]dt\Big|$$

$$\leqslant \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} |f'_\xi(x-t)|| \int_{-\infty}^{t} [F_X(\tau) - F_Y(\tau)]d\tau|dt$$

by Fubini's theorem

$$= \int_{-\infty}^{\infty} dt| \int_{-\infty}^{t} [F_X(\tau) - F_Y(\tau)d\tau| \int_{-\infty}^{\infty} |f'_\xi(x-t)|dx$$

$$= ||f'_\xi||_{L_1} \int_{-\infty}^{\infty} dt| \int_{-\infty}^{t} [F_X(\tau) - F_Y(\tau)]d\tau|. \tag{12}$$

The integral term on the right-hand side of (12) is the known (see [10], §14.1) representation of the metric ζ_2. □

The next lemma is the main part of the proof of Theorem 1. Hopefully it is useful in a wider context than that of the present paper.

For $n = 1, 2, \dots$ let $S_n = X_1 + X_2 + \cdots + X_n$; $\tilde{S}_n = \tilde{X}_1 + \tilde{X}_2 + \cdots + \tilde{X}_n$; $\sigma^2 = \mathrm{Var}(X) > 0$ and $EX = E\tilde{X} = \theta$.

For $n \geqslant s$ (see Assumption 3) denote by g_n the density of the r.v. $S_n/(\sigma\sqrt{n})$.

Lemma 4.2. *Under Assumptions 1-3*

$$l(S_n, \tilde{S}_n) \leqslant cn^{1/2} \max\{l(F, \tilde{F}), \zeta_2(F, \tilde{F})\}, \tag{13}$$

where

$$c = \max\{(10s-1)^{1/2}, 5.4d/\sigma\} \tag{14}$$

and

$$d = \sup_{n \geqslant s} \int_{-\infty}^{\infty} |g'_n(x)|dx < \infty. \tag{15}$$

Remark 4.3. By (4)

$$l(S_n, \tilde{S}_n) = \sqrt{n} l(\frac{S_n - n\theta}{\sqrt{n}}, \frac{\tilde{S}_n - n\theta}{\sqrt{n}}).$$

For $EX^2 \neq E\tilde{X}^2$ by the central limit theorem we get that

$$\liminf_{n\to\infty} l(\frac{S_n - n\theta}{\sqrt{n}}, \frac{\tilde{S}_n - n\theta}{\sqrt{n}}) > 0.$$

Thus, the rate $n^{1/2}$ on the right-hand side of (13) can not be reduced.

Remark 4.4. The constant c in (13), (14) is completely determined by the distribution function F of X. The constant d in (15) can be calculated in some particular cases numerically, using the fact that under our hypotheses

$$\int_{-\infty}^{\infty} |g'_n(x)|dx \to \int_{-\infty}^{\infty} |(\frac{1}{\sqrt{2\pi}} e^{-x^2/2})'|dx \quad \text{as} \quad n \to \infty.$$

For, instance we have the following estimates of the constants for normal, gamma, and uniform distributions:

Distribution	$d<$	$c<$
$N(0,\sigma)$ (s=1)	0.798	$\max\{3, 4.31/\sigma\}$
$\Gamma(2,\lambda)$ (s=1)	1.040	$\max\{3, 3.98/\lambda\}$
$U[0,a]$ (s=2)	0.817	$\max\{4.36, 15.29/a\}$

The proof of Lemma 2 uses an improved version of the technique given in [3] and the standard methods of probabilistic metrics in the framework of the central limit theorem (see, for instance, [10], [12]).

Making use of the resuts in [4] it is possible to prove a version of inequality (13) for i.i.d. random vectors.

Proof. Let the integer s and $\alpha > 0$ be from Assumption 3.

For $n \geqslant s$ we denote by f_n and g_n the densities of the random variables S_n and $S_n/(\sigma\sqrt{n})$, respectively. First, we show the existence of a finite constant d (depending on the distribution function F of the r.v. X) such that

$$\sup_{n\geqslant s} ||g_n'|| \leqslant d, \tag{16}$$

where (here and in what follows)

$$||\phi|| = ||\phi||_{L_1} = \int_{-\infty}^{\infty} |\phi(x)|dx.$$

We have $(n \geqslant s)$

$$||g_n'|| = \int_{|x|\leqslant 2\sqrt{n}\alpha/\sigma} |g_n'(x)|dx + \int_{|x|>2\sqrt{n}\alpha/\sigma} |g_n'(x)|dx. \tag{17}$$

Assumption 3 implies the conditions of Theorem 7 in [9], Ch. VI. Therefore

$$g_n'(x) = \frac{1}{\sqrt{2\pi}}\frac{d^2}{dx^2}\left[\int_{-\infty}^{x} e^{-t^2/2}dt - e^{-x^2/2}\frac{EX^3(x^2-1)}{6\sigma^3\sqrt{n}}\right] + O(n^{-1/2}) \quad \text{as} \quad n\to\infty, \tag{18}$$

where the term $O(n^{-1/2})$ can be chosen independent of $x \in \mathbb{R}$.

From (18) it follows that the first summands in (17) are uniformly bounded in n.

For $n > s$ (see Assumption 3):

$$f_n(x) = \int_{-\infty}^{\infty} p_s(x-t)f_{n-s}(t)dt \tag{19}$$

and since p'_s is bounded and it belongs to $L_1(\mathbb{R})$ we can differentiate under the sigh of integral in (19) (almost everywhere on $\mathbb{R}$, see [2], Appendix A)

$$f'_n(x) = \int_{-\infty}^{\infty} p'_s(x-t) f_{n-s}(t)dt.$$

Also $g'_n(x) = \sigma^2 n f'_n(\sigma\sqrt{n}x)$. Thus (applying Fubini's theorem),

$$\int_{|x|>2\sqrt{n}\alpha/\sigma} |g'_n(x)|dx = \sigma\sqrt{n}\int_{|z|>2\alpha n} dz|\int_{\infty}^{\infty} p'_s(z-t)f_{n-s}(t)dt|$$

$$\leqslant \sigma\sqrt{n}\int_{|t|\leqslant\alpha n} f_{n-s}(t)dt\int_{|z|>2\alpha n}|p'_s(z-t)|dz$$

$$+\sigma\sqrt{n}\int_{|t|>\alpha n} f_{n-s}(t)dt\int_{|z|>2\alpha n}|p'_s(z-t)|dz = I_{1,n} + I_{2,n}.$$

By (8)

$$I_{1,n} \leqslant \sigma\sqrt{n}\int_{|t|\leqslant\alpha n} f_{n-s}(t)dt\int_{|y|>\alpha n}|p'_s(y)|dy = O(1). \tag{20}$$

Also,

$$I_{2,n} \leqslant \sigma\sqrt{n}||p'_s||P(|X_1+X_2+\cdots+X_{n-s}|>\alpha n)$$

$$\leqslant \sigma\sqrt{n}||p'_s||\frac{(n-s)\sigma^2}{\alpha n^2} = O(n^{-1/2}). \tag{21}$$

Finally, from (17), (20), (21) it follows (16).

Let us turn now to the proof of inequality (13). Because of the regularity of the metric l (see, e.g. [11]):

$$l(S_n, \tilde{S}_n) \leqslant nl(X_1, \tilde{X}_1).$$

Thus, inequalities (13) hold for $n \leqslant 10s-1$, provided that

$$c \geqslant (10s-1)^{1/2}. \tag{22}$$

For $n \geqslant 10s$ let $m = [9n/10]$.

For independent r.v.'s X, Z, U, V we have ([11], §8.1):

$$l(X+U, Z+U) \leqslant l(X,Z)\sigma(U,V) + l(Z+V, Z+V). \tag{23}$$

Let $S_{k,n} = X_{k+1}+\cdots+X_n$, $\tilde{S}_{k,n} = \tilde{X}_{k+1}+\cdots+\tilde{X}_n$, $0\leqslant k\leqslant n$.

Applying the triangular inequality and (23) to $X = S_{m,n}$, $Z = \tilde{S}_{m,n}$, $U = \tilde{S}_{0,m}$, $V = S_{0,m}$ we get

$$l(S_n, \tilde{S}_n) = l(S_{0,m}+S_{m,n}, \tilde{S}_{0,m}+\tilde{S}_{m,n})$$

$$\leqslant l(S_{0,m}+S_{m,n}, \tilde{S}_{0,m}+\tilde{S}_{m,n}) + l(\tilde{S}_{0,m}+S_{m,n}, \tilde{S}_{0,m}+\tilde{S}_{m,n})$$

$$\leqslant l(S_{0,m}+S_{m,n}, \tilde{S}_{0,m}+S_{m,n}) + l(S_{0,m}+S_{m,n}, S_{0,m}+\tilde{S}_{m,n})$$

$$+l(S_{m,n},\tilde{S}_{m,n})\sigma(S_{0,m},\tilde{S}_{0,m}). \tag{24}$$

Applying (10) and (16) we get

$$T_1 = l(S_{0,m}+S_{m,n},\tilde{S}_{0,m}+S_{m,n})$$

$$= \sigma\sqrt{n-m}\,l(\frac{S_{0,m}}{\sigma\sqrt{n-m}}+\frac{S_{m,n}}{\sigma\sqrt{n-m}},\frac{\tilde{S}_{0,m}}{\sigma\sqrt{n-m}}+\frac{S_{m,n}}{\sigma\sqrt{n-m}})$$

$$\leqslant \sigma\sqrt{n-m}\,d\zeta_2(\frac{S_{0,m}}{\sigma\sqrt{n-m}},\frac{\tilde{S}_{0,m}}{\sigma\sqrt{n-m}})$$

$$\leqslant d\frac{1}{\sigma\sqrt{n-m}}m\zeta_2(X,\tilde{X}), \tag{25}$$

since (see, e.g. [10],[12],[13])

$$\zeta_2(a\sum_1^k X_i, a\sum_1^k \tilde{X}_i) \leqslant a^2\sum_1^k \zeta_2(X_i,\tilde{X}_i)$$

($a>0,\quad X_1,X_2,\dots,X_k;\quad \tilde{X}_1,\tilde{X}_2,\dots\tilde{X}_k$ are independent).

By (25) we get

$$T_1 \leqslant \frac{d}{\sigma}\frac{[\frac{9}{10}n]}{(n-[\frac{9}{10}n])^{1/2}}\mu,$$

where $\mu=\max\{l,\zeta_2\}$. Or, by simple calculations ($n\geqslant 10s$)

$$T_1 \leqslant 2.85\frac{d}{\sigma}\sqrt{n}\mu. \tag{26}$$

Similarly,

$$T_2 = l(S_{m,n}+S_{0,m},\tilde{S}_{m,n}+S_{0,m})$$

$$\leqslant \frac{d(n-[\frac{9}{10}n])}{\sigma\sqrt{[\frac{9}{10}n]}}\zeta_2(X,\tilde{X}) \leqslant 0.21\frac{d}{\sigma}\sqrt{n}\mu, \quad n\geqslant 10s. \tag{27}$$

Combining inequalities (24)–(27) and taking into account that $\sigma(S_{0,m},\tilde{S}_{0,m})\leqslant 2$ we get for $n>10s$.

$$l(D_n,\tilde{S}_n) \leqslant 3.06\frac{d}{\sigma}\sqrt{n}\mu + 2l(S_{m,n},\tilde{S}_{m,n}).$$

Making the induction assumption

$$l(S_k,\tilde{S}_k) \leqslant c\sqrt{k}\mu$$

we get

$$l(S_n,\tilde{S}_n) \leqslant \mu\sqrt{n}\left[3.06\frac{d}{\sigma}+\left(1-\frac{1}{n}[\frac{9}{10}n]\right)^{1/2}c\right] \leqslant \mu\sqrt{n}\left[3.06\frac{d}{\sigma}+0.43c\right]$$

Thus, the induction would be fulfilled if

$$c \geqslant (10s-1)^{1/2} \quad \text{and} \quad c \geqslant 3.06\frac{d}{\sigma}+0.43c.$$

Finally, we take $c=\max\{(10s-1)^{1/2}, 5.4\frac{d}{\sigma}\}$.

The end of the proof of the Theorem. □

5 Some robust properties of sample median

The common opinion is that "the sample median is more robust than the sample mean" (see, for example, [5, 6, 7]). This is true in many different senses but right now we will give an example showing something opposite.

Let$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the order statistics corresponding to the sample $X_1, X_2, \dots X_n$.

The sample median is defined as

$$\hat{X}_n = \begin{cases} X_{(k+1)} & \text{if} \quad \text{n=2k+1}, \\ \frac{1}{2}(X_k + X_{(k+1)}) & \text{if} \quad \text{n=2k}. \end{cases}$$

We will show that in the situation of the above Theorem the sample median $\hat{X}_n$ may not be $(l-\mu)$-robust with respect to the metric $\mu=\max\{l,\zeta_2\}$ (while the sample mean is $(l-\mu)$-robust).

Example 5.1. Let $X \sim N(0,1)$ and $1>\epsilon>0$. We choose the density $f_{\tilde{X}}$ of F as follows:

$$f_{\tilde{X}}(x)=\frac{1}{c(\epsilon)}(\epsilon+\frac{x^2}{\epsilon^3+x^2})\exp\{-x^2/2\},$$

where $c(\epsilon)=\int_{-\infty}^{\infty}(\epsilon+\frac{x^2}{\epsilon^3+x^2})\exp\{-x^2/2\}dx$.

It is easy to see that:

1. $f_{\tilde{X}} \in C^1(\mathbb{R})$;
2. $\sup_{\epsilon>0} f_{\tilde{X}}(x) \leqslant b < \infty$;
3. $E\tilde{X}=0$;

and that $\mu(X,\tilde{X}) \to 0$ as $\epsilon \to 0$ (using the definition of the metric l and the inequality

$$\zeta_2(X,Y) \leqslant \frac{1}{2}\int_{-\infty}^{\infty} x^2|f_X(x)-f_Y(x)|dx \tag{28}$$

which is valid provided that $EX = EY$ and r.v.'s X and Y have densities f_X and f_Y respectively, see [10]).

On the other hand, it is known that (see, for example, [1], §1.8)

$$\sqrt{n}\hat{\tilde{X}}_n \Rightarrow \eta_\epsilon \sim N(0, 1/(4f^2_{\tilde{X}}(0))) = N(0, \frac{c^2(\epsilon)}{4\epsilon^2}) \tag{29}$$

as $n \to \infty$, where $\hat{\tilde{X}}_n$ is the sample median corresponding to $\tilde{X}_1, \ldots, \tilde{X}_n$.

For any $\epsilon > 0$ $E\tilde{X}^2 < \infty$, thus from (29) it follows that

$$\sqrt{n}E|\hat{\tilde{X}}_n| \to E|\eta_\epsilon| = \sqrt{\frac{2}{\pi}}\frac{c(\epsilon)}{2\epsilon}.$$

On the other hand,

$$\sqrt{n}E|\hat{X}_n| \to E|\eta| = \sqrt{\frac{2}{\pi}}$$

for a standard normal r.v. η. As in (5) we define (recall that $\theta = 0$)

$$\Delta = \sup_{n \geqslant 1} \sqrt{n}E|\hat{X}_n| < \infty,$$

$$\tilde{\Delta} = \sup_{n \geqslant 1} \sqrt{n}E|\hat{\tilde{X}}_n| \geqslant \sqrt{\frac{2}{\pi}}\frac{c(\epsilon)}{2\epsilon}.$$

So we get $\mu(X, \tilde{X}) = \max\{l(X, \tilde{X}), \zeta_2(X, \tilde{X})\} \to 0$ but $|\Delta - \tilde{\Delta}| \to \infty$ as $\epsilon \to 0$.

Finally, we note that an analogue of the Proposition of Section 2 holds with δ replaced by Δ.

Remark 5.2. The type of the deviation from the model considered in the above example is rather "exotic". Also, this example suggests that the statistic $\hat{X}_n$ can be robust with respect to the total variation distance or the distance

$$d(X, Y) = \text{esssup}_{x \in \mathbb{R}}|f_X(x) - f_Y(x)|$$

(supposing that the density functions f_X and f_Y of X and Y, respectively, exist). Possibly (and likely) this is true if one considers asymptotic robustness. In our setting it is not the case. Indeed,

$$\Delta \geqslant E|\hat{X}_1 - \theta| = E|X_1 - \theta|$$

and a suitable "smooth" modification of Example 1 shows that it can be that $|E|X_1 - \theta| - E|\tilde{X}_1 - \theta|| \to \infty$ while $d(X_1, \tilde{X}_1) \to 0$.

It seems that some positive result on the $(l - \mu)$-robustness of the sample median could be found combining the metrics d and l.

Consider the class Q of r.v.'s such that for $X \in Q$

a) $\theta = EX$ exists;

b) X has a density f_X which is symmetric with respect to θ;

c) for some $q > 0$

$$\inf_{X \in Q} f_X(\theta) \geqslant q;$$

d) f_X is continuous at θ.

For $m = 1, 2, \ldots$ let $X, X^{(m)} \in Q$ and let $\hat{X}_n, \hat{X}_n^{(m)}$ be the sample medians corresponding to $(X_1, X_2, \ldots, X_n)$ and $(X_1^{(m)}, X_2^{(m)}, \ldots, X_n^{(m)})$, respectively. Additionally, we assume that $\theta = EX = E\hat{X}^{(m)}$.

Conjecture. If $\max\{d(X, X^{(m)}), l(X, X^{(m)})\} \to 0$ as $m \to \infty$ then

$$\sup_{n \geqslant 1} \sqrt{n} l(\hat{X}_n, \hat{X}_n^{(m)}) \to 0.$$

References

[1] Borovkov A.A. *Mathematical Statistics.* Nauka, Moscow, 1984 (in Russian).

[2] Dudley R.M. *Uniform Central Limit Theorem.* Cambridge University Press, 1999.

[3] Gordienko E., Ruiz de Chavez J. New estimates of continuity in $M|GI|1|\infty$ queues. *Queueing Systems*, 29:175–188, 1988.

[4] Gordienko E. Comparing the distributions of sums of independent random vectors. *Kybernetika*, 41:519–529, 2005.

[5] Hampel F.R. A general qualitative definition of robustness. *Ann. Math. Statist.* 42:1887–1896,1971.

[6] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. *Robust Statistics: The Approach based an Influence Functions.* Wiley, New York, 1986.

[7] Huber P.J. Robust Estimation of a Location Parameter. *Ann. Math. Statist.*, 35:73–101, 1964.

[8] Jurečková J., Sen P.K. *Robust Statistical Procedures. Asymptotics and Interrelations.* Wiley, 1996.

[9] Petrov V.V. *Sums of independent Random Variables.* Springer-Verlag, Berlin, 1975.

[10] Rachev S.T. *Probability metrics and the Stability of Stochastic Models.* Wiley, Chichester, 1991.

[11] Rachev S.T., Rüschendorf L. *Mass Transportation Problem. V. II: Applications.* Springer, 1998.