IMS Collections
Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková
Vol. 7 (2010) 169–181
© Institute of Mathematical Statistics, 2010
DOI: 10.1214/10-IMSCOLL717

Qualitative robustness and weak continuity: the extreme unction?

Ivan Mizera[1,*]

University of Alberta

Abstract: We formulate versions of Hampel's theorem and its converse, establishing the connection between qualitative robustness and weak continuity in full generality and under minimal assumptions.

1. Qualitative robustness

The definition of *qualitative robustness* was given by Hampel [11]. Suppose that t_n is a sequence of *statistics* (estimators or test statistics), that, for each *sample size* n, describe a *procedure*. Let P be a probability measure that identifies the stochastic model we believe that underlies the data, and let $\mathcal{L}_P(t_n)$ be the distribution of t_n under this stochastic model; Hampel [11] implicitly views the data as independent, identically distributed random elements of some *sampling space* $\mathcal{X}$ (assumed to be complete separable metric space), with P then the common distribution of these random elements, a member of $\mathcal{P}(\mathcal{X})$, the space of all probability measures on $\mathcal{X}$ (defined on the Borel σ-field generated by the topology of $\mathcal{X}$). Let π denote the Prokhorov metric on $\mathcal{P}(\mathcal{X})$, as defined in Huber [15]; see also Section 3 below.

Definition 1. Let P be a probability measure from $\mathcal{P}(\mathcal{X})$. A procedure t_n is called *qualitatively robust* at P if for any $\epsilon > 0$ there is $\delta > 0$ such that

$$\pi(P, Q) \le \delta \text{ implies } \pi(\mathcal{L}_P(t_n), \mathcal{L}_Q(t_n)) < \epsilon \tag{1}$$

for all sufficiently large n.

The fact that (qualitative) "robustness is related to some form of continuity", as we can read, for instance, on page 72–73 of Maronna et al. [18], became a part of universal statistical knowledge. It was demonstrated already by Hampel [11] for procedures *representable by functionals* on the space $\mathcal{P}(\mathcal{X})$, the procedures that can be summarized in terms of a functional, T, defined on a subset of $\mathcal{P}(\mathcal{X})$ rich enough to guarantee that for any relevant collection of x_i's,

$$t_n(x_1, \dots, x_n) = T(\mathit{\Delta}_{x_1,\dots,x_n}), \tag{2}$$

where $\mathit{\Delta}_{x_1,\dots,x_n}$ stands for the empirical probability supported by the points x_1, x_2, …, x_n (the probability allocating mass $1/n$ to every of the x_i's). For procedures representable by functionals, qualitative robustness is essentially equivalent to

[*] Research supported by the Natural Sciences and Engineering Research Council of Canada.

[1] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, T6G 2G1, Canada. e-mail: mizera@stat.ualberta.ca

AMS 2000 subject classifications: Primary 62F35, 62B99; secondary 60K35, 60B10.
Keywords and phrases: qualitative robustness, weak continuity, consistency.

weak continuity, the continuity with respect to the weak convergence of probability measures—as defined, for instance, by Billingsley [1].

Possible subtleties arising in this context can be illustrated on a very simple (and already discussed elsewhere) example: median. We define the estimator as the value of t where the graph of the function

$$\psi(t) = \frac{1}{n}\sum_{i=1}^{n} \operatorname{sign}(X_i - t) \tag{3}$$

crosses zero level. This happens when $\psi(t) = 0$; but there may be no such t, as ψ is not continuous. Nonetheless, given that ψ is nondecreasing, we can complete its graph by vertical segments connecting the jumps, and then take t giving the location where such augmented graph intersects the horizontal coordinate axis.

Such a provision takes care of jumps, but still leaves possible ambiguity: sometimes there may be not one, but several t such that $\psi(t) = 0$. (Note that if ψ happens to cross zero level at a jump, then the corresponding location is unique.) To finalize the definition of median, we have to adopt some "ambiguity resolution stance". Roughly, there are three possibilities.

(i) Ignore: that is, consider median defined not for all n-tuples of data, but only for those for which it is defined uniquely. In statistical science, such a strategy is often vindicated by the fact that data configurations yielding ambiguous results happen to be rather a mathematical than practical phenomenon, especially if the underlying stochastic model implies their occurrence with probability zero. While this point of view is pertinent, for instance, for the Huber estimator—which in theory can yield non-unique result, but in practice seldom will—for the median, however, the ambiguity is bound to occur for most of data configurations with even n.

(ii) View the definition as set-valued: instead of uniquely defined median, consider a *median set*—in this case always a closed interval, due to the monotonicity of ψ. This strategy is likely to be successful if it can be pursued without invoking too much of non-standard mathematics—simply as an attitude that instead of ignoring ambiguous data configurations, one can rather admit an occasional possibility of multiple solutions and still maintain some theoretical control over these, as pointed out by Portnoy and Mizera [23] in the discussion of Ellis [8].

(iii) Consider a suitable selection: that is, define the median as a point selected in some specific way from the median set. The often used alternative is the midpoint of the median interval—but minimum or maximum could be considered too. The selection strategy may be naturally suggested by the implementation of the method, when a specific algorithm returns some particular, and thus unique, solution.

A functional representation of the median can be obtained via the straightforward extension of (3): we define the median functional, $T(P)$, to be the location t where the graph of

$$\psi_P(t) = \int \operatorname{sign}(x - t) P(\mathrm{d}x) \tag{4}$$

crosses the zero level (using the same provision as above to define what this precisely means). Any "ambiguity resolution stance" mentioned above can be directly generalized to this situation.

A standard argument shows that if $P_n \to P$ weakly, then $\psi_{P_n}(t) \to \psi_P(t)$ for every continuity point t of ψ; further analytical argument based on monotonicity yields that every limit point of a sequence t_n of points giving locations where ψ_{P_n} crosses zero level is a location where ψ_P crosses zero level. In the terminology introduced below, T is weakly semicontinuous at every P, and weakly continuous at every P for which it is uniquely defined. Therefore, by Theorem 6.2 of Huber [15], or by our Theorem 1, median is qualitatively robust at every P for which T is uniquely defined.

The justification of this step—from continuity to qualitative robustness—is the theme of this note, and we will return to it in the next section. Let us illustrate now why uniqueness is necessary, on a simple example (capturing nevertheless the essence of behavior for any P yielding a non-degenerate median interval): let P be concentrated with equal mass $1/2$ in two points, -1, and $+1$. Fix $\epsilon = 1/4$, say. Let Q_α^- and Q_α^+ be concentrated on $\{-1,+1\}$ with corresponding probabilities $1/2+\alpha$, $1/2-\alpha$ and $1/2-\alpha$, $1/2+\alpha$, respectively; given $\delta > 0$, we can always choose $\alpha > 0$ so that both $\pi(P, Q_\alpha^-) < \delta$ and $\pi(P, Q_\alpha^+) < \delta$. A standard probabilistic argument yields that we can find N such that for any $n > N$, the probability of median being $+1$ is bounded from above by $1/4$, if we sample from Q_α^-; and the same bound takes place for the probability of median being -1, if we sample from Q_α^+. Consequently, $\pi(\mathcal{L}_{Q_1}(t_n), \mathcal{L}_{Q_2}(t_n)) > 1/4$ for $n > N$. Given that we arrived to this for fixed ϵ and arbitrary δ, we conclude that median cannot be qualitatively robust at P.

Note that to reach this conclusion, we essentially do not need to know how the estimator is defined in ambiguous situations: albeit data configurations with equal number of -1's and $+1$'s are possible, they occur only with small probability, which further decreases to 0 for $n \to \infty$. The fact that qualitative robustness requires uniqueness of T at P does not depend, and would not change with an adopted "ambiguity resolution stance". The probabilistic behavior of a sample of size n from P is not relevant either: except for the data configuration with equal number of -1's and $+1$'s, which occurs with small probability η_n (tending to 0 with growing n), there are only two possible cases, by symmetry each occurring with probability $1 - \eta_n/2$: either -1's or $+1$ are in majority, and the median is then *unambiguously* equal to -1 or $+1$.

The situation is somewhat different in the functional setting, where the weak continuity of T depends on the adopted "ambiguity resolution stance". If we take in the example above $T(P)$ to be 0, the midpoint of the median set $[-1, 1]$, then we loose continuity: for $\alpha \to 0$ we have $T(Q_\alpha^-) = -1$ for all $\alpha > 0$, which suddenly jumps to 0 for P (which corresponds to $\alpha = 0$). If we adopt the set-valued definition of T, then we have a set-valued *weak semicontinuity* at P: for any sequence Q_n converging weakly to P, the sets $T(Q_n)$ are eventually contained in an ϵ-neighborhood of the set $T(P)$. It might be tempting to consider this as a, possibly extended, definition of robustness, as indicated in Section 1.4 of Huber [15]:

> "We could take this . . . as our definition and call a . . . statistical functional T robust if it is weakly continuous. However, following Hampel [11], we prefer to adopt a slightly more general definition."

Indeed, Definition 1 has an advantage that it is directly based on the procedures rather than on their functional representations, whose existence and form may not be always that clear and intuitive as in our median example, and whose scope is limited to permutation-invariant, exchangeable situations—while Definition 1 exhibits a clear potential for extensions to situations structured beyond such a framework; and also, as we have seen, essentially does not depend on the adopted "ambiguity resolution stance".

Thus, adopting the Hampel [11]'s Definition 1 of the qualitative robustness, we would like to revisit now how it relates to the weak continuity of T at P, in situations when T is uniquely defined at P, but possibly may not be so elsewhere.

2. Weak continuity

Definition 2. A functional T is called *weakly continuous* at P, if for any $\epsilon > 0$ there is $\delta > 0$ such that

$$\pi(P, Q) \leq \delta \text{ implies } d(\theta, \tau) < \epsilon \tag{5}$$

for *any* value θ and τ of T at P and Q, respectively.

The appearance of the word "any" above means that the definition is formulated for set-valued T, without explicitly mentioning this fact; the value of T is considered to be a subset of $\mathcal{X}$. Of course, univalued T (with values that are singletons, sets consisting of precisely one element) are a special case.

For a set-valued functional T, we can also define *weak semicontinuity* of T at P by the requirement that for any $\epsilon > 0$ there is $\delta > 0$ such that $\pi(P, Q) < \delta$ implies that $T(Q) \subseteq T(P)^\epsilon$, the set $T(P)^\epsilon$ containing all points within ϵ distance from the *set* $T(P)$. This seems to be equivalent to Definition 2, but is not: T is weakly continuous at P, if and only if it is weakly semicontinuous and *univalued* at P.

As mentioned above, Hampel [11] pointed out that weak continuity at P implies qualitative robustness at P. However, his Theorem 1 and its Corollary required also an additional assumption of *global pointwise continuity* of all t_n: every t_n had to be continuous as a function of the vector $(x_1, x_2, \ldots, x_n)$, for *all* such vectors. Although Hampel [11] gives also a version (Theorem 1a) which weakens this assumption and allows exceptions from the pointwise continuity if those occur under zero probability P, verifying his condition can be in general burdensome.

For instance, the condition of pointwise global continuity holds true, and is not difficult to verify for every data vector, if we define the median as the midpoint of the median interval. However, when exploring, in Mizera and Volauf [20], the same topic for a multivariate generalization of the median called the Tukey median, we realized that this route would lead to serious complications. First, specifying the appropriate selection from a convex set in $\mathbb{R}^k$ is not that straightforward for $k > 1$; second, we realized that the Tukey median may not be always continuous—so, third, we would have to show that such configurations occur with probability zero under P yielding the weak continuity of the Tukey median.

Such complications are not necessary: Huber [15], while giving the result the name of Hampel, also noted that weak continuity of T at P is all what is needed. A somewhat related global version was given already by Hampel [11]: weak continuity at an empirical probability $\varDelta_{x_1,\ldots,x_n}$ implies the pointwise continuity of t_n at $(x_1, \ldots, x_n)$; therefore, if weak continuity is postulated at *all* P, the pointwise continuity then follows. The Hampel [11]'s proof suggests that some version of local pointwise continuity, or even local boundedness would suffice; but it is not obvious how such a condition would have to be formalized.

So, we could use Theorem 6.2, Section 2.6 of Huber [15] to conclude that the Tukey median is qualitatively robust whenever it is weakly continuous—if not for the following. Huber [15]'s formulation and proof uses for the first π in (1) the Lévy metric, instead of the Prokhorov one. This means that Theorem 6.2 is formally valid only for $\mathcal{X} = \mathbb{R}$; that is good enough for our median example, but does not apply to the Tukey median, when $\mathcal{X} = \mathbb{R}^k$. Actually, Huber [15] allows t_n to assume values in $\mathbb{R}^k$; but P is clearly restricted to $\mathcal{P}(\mathbb{R})$.

We did not consider this minor detail to be of major importance; it is clear that Huber [15] envisioned the broad validity of his Theorem 6.2—only for educational or practical reasons he preferred the simple argument based on the uniformity in the Glivenko–Cantelli theorem (with a direct consequence for the Lévy metric) to possibly more technical treatment required for the general case (which can be nowadays carried in the language of the modern theory of empirical processes, which Huber [15] pioneered in his works). Thus, writing Mizera and Volauf [20], we believed that we could limit our focus to continuity questions, their statistical consequences for robustness being well known.

However, the reviewers of Mizera and Volauf [20] did not initially share this view—until we introduced in the revised version a theorem, which up to some technical details is identical with Theorem 1 below. Its proof, however, was considerably out of scope of Mizera and Volauf [20]; in lieu of it we rather promised that "the proof of the theorem will appear elsewhere in the literature"—hoping that somebody (a referee or anybody else) would argue that this is not really necessary, because the theorem appears to be an obvious consequence of Huber [15], Hampel [11], or some other reference.

However, it seems that our hope has not materialized, and it is time to fulfill our promise now. Before formulating the theorem and showing how the original proof of Huber [15] can be altered to cover rigorously also the multidimensional case, we need to discuss one formal subtlety. Thinking of our functionals and procedures as of set-valued mappings, we are not completely sure whether we may still speak in a mathematically consistent manner about their distribution. There are ways to formalize the notion of law for set-valued random functions—however, we would prefer to stay away from this level of abstraction. In practice, a lot of procedures consist of functions yielding unique values with probability one—we will call such set-valued functions *lawful*, as we can speak about their distributions without ambiguities. For instance, the ℓ_1 regression estimator is lawful as long as the distribution of covariates is continuous. However, the case of median—as well as that of the Tukey median—is different; the median is not lawful for even n, unless we consider its *lawful version*: a univalued selection from the estimator, that is a univalued function picking always one value from the set of all possible ones. This resembles the selection strategy for the "ambiguity resolution stance", with one important distinction: now the selection does not have to be deterministic, but may be also randomized: a lawful version of the sample median may be a point selected at random according to the uniform distribution on the median interval. We stress that lawful versions are introduced exclusively for "law enforcement", to ensure that the symbol $\mathcal{L}(t_n)$ in the definition of qualitative robustness is well-defined; as far as other aspects are concerned, we will consider functionals in their original deterministic expression.

Theorem 1. *Suppose that a procedure t_n is represented by a functional T. If T is weakly continuous at P, then any lawful version of t_n is qualitatively robust at P.*

The proof—which is that of Huber [15], only the argument using the Lévy metric is replaced by a more general one—is given in Section 3. The rest of this section presents the converse to Theorem 1, to make this note self-contained; we essentially follow Hampel [11], the proofs are given in Section 3.

The appropriate formulation of the converse requires some insights into the nature how the procedure is represented by a functional. We remark that the general question of representability by functionals may involve some delicate aspects; Hampel [11] and Huber [15] addressed the question to some extent; see also Mizera [19]. For example, such representation exist only when the t_n's exhibit some mutual consistency—if an empirical probability for a given n arises as an empirical probability for some other n, the corresponding t_n should yield the same result. Again, we do not want to go into more depth than needed here.

Developing all the theory in the set-valued context, we have to include an appropriate definition of convergence in probability: for the purposes of Definition 3, we say that a sequence of random sets E_n converge to E in probability, if for any selected subsequence $x_n \in E_n$, the distance of x_n to E converges to 0 in probability. In the set-valued terminology, this may be called rather "upper convergence", but for the present purpose, the name and definition are good enough; the interesting cases will be those when $E = \{x\}$ is a singleton, and then the term "convergence" is justified, and means that x_n converges to x in probability for any sequence t_n selected from the E_n's.

Definition 3. A representation of a procedure t_n by a functional T is called *consistent* at P, if t_n converges in probability to $T(P)$ whenever the data are independent and identically distributed according to the law P.

Proposition 1. *If a procedure t_n is represented by a functional T weakly continuous at P, then this representation is consistent at P.*

Definition 4. A representation of a procedure t_n by a functional T is called *regular*, if (i) it is consistent for every P in the domain of T; and (ii) for every P and every $\tau \in T(P)$, there is a sequence P_ν of empirical probabilities weakly converging to P, the functional T is univalued at every P_ν, and $T(P_\nu)$ converges to τ.

The following result serves as a "prototype" of the converse part of Hampel's theorem. It can be used for disproving qualitative robustness in nonregular cases—in particular, when T is not univalued at P.

Proposition 2. *Suppose that a procedure t_n is represented by a functional T. If there are Q_ν^-, Q_ν^+ such that (i) both Q_ν^- and Q_ν^+ weakly converge (in n) to P; (ii) $T(Q_\nu^-)$ converges to θ and $T(Q_\nu^+)$ to τ, where $\theta \neq \tau$; (iii) T is univalued at every Q_ν^- and Q_ν^+; (iv) the representation of t_n by T is at every Q_ν^- and Q_ν^+ consistent—then no lawful version of t_n is qualitatively robust at P.*

The converse to Theorem 1 is formulated for regular representations.

Theorem 2. *Suppose that the representation of a procedure t_n by a functional T is regular. If some lawful version of t_n is qualitatively robust at P, then T is weakly continuous (in particular, uniquely defined) at P.*

3. Proofs

We assume that S is a Polish space, a complete and separable metric space with a metric d. For $E \subset S$, E^ϵ denotes the ϵ-fattening of E, the set of all $x \in S$ within ϵ distance from E. The Prokhorov metric, $\pi(P,Q)$, is defined as the infimum of all $\epsilon > 0$ such that $P(E) \leq Q(E^\epsilon) + \epsilon$ for all measurable E. It is uniformly equivalent to the bounded Lipschitz metric β,

$$\frac{2}{3}\pi^2(P,Q) \leq \beta(P,Q) \leq 2\pi(P,Q). \tag{6}$$

The bounded Lipschitz metric is defined as

$$\beta(P,Q) = \sup_{f \in BL(S)} \left| \int f \,\mathrm{d}P - \int f \,\mathrm{d}Q \right|,$$

where $BL(S)$ stands for the set of all real functions on S satisfying

$$\sup_{u \in S} |f(u)| + \sup_{u,v \in S} \frac{|f(u) - f(v)|}{d(u,v)} \le 1;$$

in particular, $|f| \le 1$ for all f from $BL(S)$. A set F is called *totally bounded*, if for any $\epsilon > 0$ there is a finite collection of ϵ-balls, balls with radius ϵ in metric ℓ, covering F; the symbol $N(\epsilon, F, \ell)$ then denotes the minimal cardinality of such a collection, the ϵ-covering number of F in metric ℓ. Symbols L^p_E denote the usual metrics on spaces of functions defined on E.

Let X_1, X_2, …, X_n be independent random variables, each with the distribution Q; it can be arranged that all X_i are defined on the same probability space $(\Omega, \mathcal{S}, \mathbb{P}_Q)$ (depending on Q). Let $\mathbb{Q}_n$ be the (random) empirical probability measure supported by the random variables Z_i; note that the distribution of $\mathbb{Q}_n$ depends on Q.

Lemma 1. *Let K be a totally bounded subset of S. For every $\epsilon > 0$,*

$$\mathbb{P}_Q \left[\sup_{f \in BL(K^\epsilon)} \left| \int f \,\mathrm{d}\mathbb{Q}_n - \int f \,\mathrm{d}Q \right| > 48\epsilon \right] \tag{7}$$

tends to 0 *uniformly in all $Q \in \mathcal{P}(K^\epsilon)$.*

Proof. Proceeding as in the proof of Theorem 6 of Dudley et al. [7], we obtain an upper bound for (7),

$$2N(6\epsilon, BL(K^\epsilon), L^1_{\mathbb{Q}_n})\ \mathrm{e}^{-18n\epsilon^2} \le 2N(6\epsilon, BL(K^\epsilon), L^\infty_{K^\epsilon})\ \mathrm{e}^{-18n\epsilon^2}. \tag{8}$$

The inequality, obtained by approximating the functions in $BL(K^\epsilon)$ by stepwise functions and using their analytical properties,

$$N(6\epsilon, BL(K^\epsilon), L^\infty_{K^\epsilon}) \le \left(\frac{1}{2\epsilon}\right)^{N(2\epsilon, K^\epsilon, d)} \tag{9}$$

and the fact that $N(2\epsilon, K^\epsilon, d) \le N(\epsilon, K, d)$ together imply, given the total boundedness of K, that the covering numbers in (8) are bounded uniformly in n. Hence the expressions in (8) and consequently in (7) tend to 0, uniformly in Q. □

Lemma 2. *For fixed $E \subseteq S$ and any $\epsilon > 0$, the sequence $\mathbb{P}_Q\left[\mathbb{Q}_n(E) > 2\epsilon\right]$ converges to* 0 *as $n \to \infty$, uniformly in all $Q \in \mathcal{P}(S)$ such that $Q(E) \le \epsilon$.*

Proof. Use the Chebyshev inequality for the Bernoulli sequence of independent events with $p = Q(E) \le \epsilon$,

$$\begin{aligned} \mathbb{P}_Q\left[\mathbb{Q}_n(E) \ge 2\epsilon\right] &= \mathbb{P}_Q\left[\mathbb{Q}_n(E) - p \ge 2\epsilon - p\right] \\ &\le \mathbb{P}_Q\left[|\mathbb{Q}_n(E) - p| \ge \epsilon\right] \le \frac{p(1-p)}{n\epsilon^2} \le \frac{1}{n\epsilon}. \end{aligned}$$

The lemma follows. □

Lemma 3. *Let $P \in \mathcal{P}(S)$. For any $\epsilon > 0$, there exists a totally bounded subset K of S such that*

$$\mathbb{P}_Q\left[\sup_{f\in BL(S)}\left|\int_{K^\epsilon} f\,\mathrm{d}\mathbb{Q}_n - \int_{K^\epsilon} f\,\mathrm{d}Q\right| > 96\epsilon\right] \to 0, \tag{10}$$

uniformly in all $Q \in \mathcal{P}(S)$ such that $\pi(P,Q) \le \epsilon/2$.

Proof. Given $\epsilon > 0$, choose a compact subset K of S such that $P(K) \ge 1 - \epsilon/2$; here we use the fact that a probability measure on a Polish space is tight, in the terminology of Theorem 1.4 of Billingsley [1]. Fix $\eta > 0$ and choose n_0 such that (7) in Lemma 1 is bounded by $\eta/3$ for all $n \ge n_0$. Choose n_1 such that

$$\frac{n_1(1-\epsilon)}{2} \ge n_0, \qquad \frac{4\epsilon}{n_1(1-\epsilon)} \le \frac{\eta}{3}, \quad \text{and } \frac{1}{2304\,n_1\epsilon^2} \le \frac{\eta}{3}\ ; \tag{11}$$

note that the first inequality also implies $n_1 \ge n_0$. Let Q be an element from $\mathcal{P}(S)$ such that $\pi(P,Q) \le \epsilon/2$; then

$$1 - \frac{\epsilon}{2} \le P(K) \le Q(K^{\epsilon/2}) + \frac{\epsilon}{2} \le Q(K^\epsilon) + \frac{\epsilon}{2}$$

and consequently $1 - \epsilon \le Q(K^\epsilon)$. Let N_Q be the (random) number of $X_i \in K^\epsilon$; let Q_{K^ϵ} denote the conditional probability on K^ϵ defined by $Q_{K^\epsilon}(E) = Q(E \cap K^\epsilon)/Q(K^\epsilon)$. Using again the Chebyshev argument as in Lemma 2, for the Bernoulli series of events with $p = Q(K^\epsilon) \ge 1 - \epsilon$, we obtain, using the first two inequalities in (11), that for any $n \ge n_1$,

$$\begin{aligned}\mathbb{P}_Q\left[N_Q \le n_0\right] &\le \mathbb{P}_Q\left[N_Q \le \tfrac{1}{2}n_1(1-\epsilon)\right] \le \mathbb{P}_Q\left[N_Q \le \tfrac{1}{2}n(1-\epsilon)\right] \\ &\le \mathbb{P}_Q\left[\left|\frac{N_Q}{n} - p\right| \ge \frac{p}{2}\right] \le \frac{4(1-p)}{np} \le \frac{4\epsilon}{n_1(1-\epsilon)} \le \frac{\eta}{3}\,,\end{aligned} \tag{12}$$

uniformly in Q. The Chebyshev inequality yields once again, now together with the third inequality in (11), that for $n \ge n_1$,

$$\mathbb{P}_Q\left[\left|\frac{N_Q}{n} - p\right| > 48\epsilon\right] \le \frac{p(1-p)}{48^2\,n\epsilon^2} \le \frac{p\,\epsilon}{2034\,n_1\epsilon^2} \le \frac{1}{2034\,n_1\epsilon} \le \frac{\eta}{3}\,, \tag{13}$$

again uniformly in Q. Dividing the expression within (10) by $p = Q(K^\epsilon)$, we obtain that for $n \ge n_1$,

$$\begin{aligned}&\mathbb{P}_Q\left[\sup_{f\in BL(S)}\left|\frac{1}{np}\sum_{X_i\in K^\epsilon} f(X_i) - \frac{1}{p}\int_{K^\epsilon} f\,\mathrm{d}Q\right| \ge \frac{96\epsilon}{p}\right] \\ &\le \mathbb{P}_Q\left[\sup_{f\in BL(S)}\left|\frac{1}{np}\sum_{X_i\in K^\epsilon} f(X_i) - \frac{1}{N_Q}\sum_{X_i\in K^\epsilon} f(X_i)\right| \ge \frac{96\epsilon}{2p}\right] \\ &\qquad + \mathbb{P}_Q\left[\sup_{f\in BL(S)}\left|\frac{1}{N_Q}\sum_{X_i\in K^\epsilon} f(X_i) - \frac{1}{Q(K^\epsilon)}\int f\,\mathrm{d}Q\right| \ge \frac{96\epsilon}{2p}\right] \\ &= \mathbb{P}_Q\left[\left|\frac{N_Q}{n} - p\right| \sup_{f\in BL(S)}\left|\frac{1}{N_Q}\sum_{X_i\in K^\epsilon} f(X_i)\right| \ge 48\epsilon\right] \\ &\qquad + \mathbb{P}_Q\left[\sup_{f\in BL(K^\epsilon)}\left|\frac{1}{N_Q}\sum_{X_i\in K^\epsilon} f(X_i) - \int f\,\mathrm{d}Q_{K^\epsilon}\right| \ge \frac{48\epsilon}{p}\right] \\ &\le \mathbb{P}_Q\left[\left|\frac{N_Q}{n} - p\right| > 48\epsilon\right] + \mathbb{P}_Q\left[\sup_{f\in BL(K^\epsilon)}\left|\frac{1}{N_Q}\sum_{X_i\in K^\epsilon} f(X_i) - \int f\,\mathrm{d}Q_{K^\epsilon}\right| > 48\epsilon\right].\end{aligned}$$

By (13), the left-hand expression is dominated by $\eta/3$; the right-hand one can be written as

$$\sum_{m=1}^{\infty} \mathbb{P}_Q\biggl[\sup_{f\in BL(K^\epsilon)}\biggl|\frac{1}{N_Q}\sum_{X_i\in K^\epsilon} f(X_i) - \int f\,\mathrm{d}Q_{K^\epsilon}\biggr| > 48\epsilon \biggm| N_Q = m\biggr]\mathbb{P}_Q[N_Q = m]$$

which can be split to two sums: the first is dominated by

$$\sum_{m\le n}\mathbb{P}_Q[N_Q = m] = \mathbb{P}_Q\bigl[N_Q \le n_0\bigr] \le \tfrac{1}{3}\eta,$$

by (12); the second is

$$\begin{aligned}
&\sum_{m>n} \mathbb{P}_Q\biggl[\sup_{f\in BL(K^\epsilon)}\biggl|\frac{1}{N_Q}\sum_{X_i\in K^\epsilon} f(X_i) - \int f\,\mathrm{d}Q_{K^\epsilon}\biggr| > 48\epsilon \biggm| N_Q = m\biggr]\mathbb{P}_Q[N_Q = m]\\
&= \sum_{m>n}^{\infty} \mathbb{P}_{Q_{K^\epsilon}}\biggl[\sup_{f\in BL(K^\epsilon)}\biggl|\frac{1}{m}\sum_{i=1}^{m} f(Z_i) - \int f\,\mathrm{d}Q_{K^\epsilon}\biggr| > 48\epsilon\biggr]\mathbb{P}_Q[N_Q = m]\\
&\le \tfrac{1}{3}\eta\sum_{m>n}^{\infty}\mathbb{P}_Q[N_Q = m] \le \tfrac{1}{3}\eta,
\end{aligned}$$

where $Z_1, Z_2, \ldots, Z_m$ are independent random variables (different for each m), each with distribution Q_{K^ϵ}, so that Lemma 1 applies. As η was arbitrary, the lemma follows. □

Lemma 4. *For any $\alpha, \eta > 0$, there exists $\delta > 0$ and ν such that*

$$\mathbb{P}_Q\bigl[\pi(\mathbb{Q}_n, P) > \alpha\bigr] < \eta \tag{14}$$

whenever $n \ge \nu$ and $\pi(P, Q) < \delta$.

Proof. Given P and α, choose $\epsilon < \alpha^2/12$ such that $96\epsilon \le \alpha^2/3$. As in the proof of Lemma 3, we take a compact K such that $Q(K^\epsilon) \ge 1 - \epsilon$ whenever $\pi(Q, P) < \delta = \epsilon/2$. By (6), we obtain

$$\begin{aligned}
&\mathbb{P}_Q\bigl[\pi(\mathbb{Q}_n, P) > \alpha\bigr] \le \mathbb{P}_Q\bigl[\beta(\mathbb{Q}_n, P) > \tfrac{2}{3}\alpha^2\bigr]\\
&\quad\le \mathbb{P}_Q\biggl[\sup_{f\in BL(S)}\int_{S\setminus K^\epsilon}|f|\,\mathrm{d}\mathbb{Q}_n + \sup_{f\in BL(S)}\int_{S\setminus K^\epsilon}|f|\,\mathrm{d}Q > \tfrac{1}{3}\alpha^2\biggr]\\
&\qquad\quad + \mathbb{P}_Q\biggl[\sup_{f\in BL(S)}\biggl|\int_{K^\epsilon} f\,\mathrm{d}\mathbb{Q}_n - \int_{K^\epsilon} f\,\mathrm{d}Q\biggr| > \tfrac{1}{3}\alpha^2\biggr]\\
&\quad\le \mathbb{P}_Q\bigl[\mathbb{Q}_n(S\setminus K^\epsilon) > \tfrac{1}{4}\alpha^2\bigr] + \mathbb{P}_Q\biggl[\sup_{f\in BL(S)}\biggl|\int_{K^\epsilon} f\,\mathrm{d}\mathbb{Q}_n - \int_{K^\epsilon} f\,\mathrm{d}Q\biggr| > 96\epsilon(1-\epsilon)\biggr]
\end{aligned}$$

By Lemma 3, there exists n_1 such that the second term is bounded by $\eta/2$ for $n \ge n_1$. Since $Q(S\setminus K^\epsilon) \le \epsilon < \alpha^2/12$, Lemma 2 yields n_2 such that for $n \ge n_2$,

$$\mathbb{P}_Q\bigl[\mathbb{Q}_n(S\setminus K^\epsilon) > \tfrac{1}{4}\alpha^2\bigr] \le \mathbb{P}_Q\bigl[\mathbb{Q}_n(S\setminus K^\epsilon) > \tfrac{1}{6}\alpha^2\bigr] \le \tfrac{1}{2}\eta.$$

Setting $\nu = \max\{n_1, n_2\}$ concludes the proof. □

Proof of Theorem 1. Let ℓ be the metric on the range of T. Given $\epsilon > 0$, weak continuity of T at P yields α such that $\ell(\tau, T(P)) < \epsilon/3$ whenever $\tau \in T(Q)$ and $\pi(P, Q) < \alpha$. Setting η to $\epsilon/3$ and taking ν and δ yielded by Lemma 4, we obtain that if $n \geq \nu$ and $\pi(P, Q) \leq \delta$, then

$$\mathbb{P}_Q\left[\ell(\tau, T(P)) > \tfrac{1}{3}\epsilon\right] < \tfrac{1}{3}\epsilon$$

whenever $\tau \in T(\mathbb{Q}_n)$. The Strassen theorem — see Huber [15], Chapter 2, Theorem 3.7, or also the original paper Strassen [26]—then gives

$$\pi(\mathcal{L}_Q(t_n), \delta_{T(P)}) \leq \tfrac{1}{3}\epsilon; \tag{15}$$

here $\delta_{T(P)}$ stands, in the spirit of the notation introduced above, for the point (Dirac) measure concentrated in $T(P)$. Using (15) once again for $Q = P$ and then combining both inequalities, we obtain the desired result: if $\pi(P, Q) \leq \delta$, then $\pi(\mathcal{L}_Q(t_n), \mathcal{L}_P(t_n)) < \epsilon$ for $n \geq \nu$, uniformly in Q. □

Proof of Proposition 1. The proposition follows from the Varadarajan theorem, stating that when the data are independently sampled from P, the corresponding empirical probability measures converge weakly to P with probability one. The consistency then follows from the weak continuity of T at P. □

Proof of Proposition 2. Let ℓ be the metric on the range of T, and suppose that $\ell(\theta, \tau) = \epsilon > 0$. Suppose that some lawful version of t_n is qualitatively robust at P. Given $\epsilon/4$, we may pick Q^-, Q^+, out of Q_ν^- and Q_ν^+ satisfying assumptions (ii), (iii), and (iv), such that T is univalued, and the representation of t_n by T is consistent at both Q^- and Q^+; by qualitative robustness, we can pick them so that for some n_1

$$\pi(\mathcal{L}_{Q^-}(t_n), \mathcal{L}_P(t_n)) \leq \tfrac{1}{4}\epsilon, \tag{16}$$

$$\pi(\mathcal{L}_{Q^+}(t_n), \mathcal{L}_P(t_n)) \leq \tfrac{1}{4}\epsilon, \tag{17}$$

for all $n \geq n_1$. The consistency at Q^- and Q^+ yields n_2 such that for all $n \geq n_2$,

$$\mathbb{P}_{Q^-}\left[\ell(T(\mathbb{Q}_n^-), T(Q^-)) \geq \tfrac{1}{4}\epsilon\right] < \tfrac{1}{4}\epsilon, \tag{18}$$

$$\mathbb{P}_{Q^-}\left[\ell(T(\mathbb{Q}_n^-), T(Q^-)) \geq \tfrac{1}{4}\epsilon\right] < \tfrac{1}{4}\epsilon. \tag{19}$$

Take $n \geq \max\{n_1, n_2\}$. Applying the Strassen theorem to (18) and (19) (given that T is univalued at Q^- and Q^+), we obtain that

$$\pi(\mathcal{L}_{Q^-}(t_n), \delta_{T(Q^-)}) < \tfrac{1}{4}\epsilon, \tag{20}$$

$$\pi(\mathcal{L}_{Q^+}(t_n), \delta_{T(Q^+)}) < \tfrac{1}{4}\epsilon. \tag{21}$$

Combining (20), (21) with (16) and (17) yields that $d(\theta, \tau) = \pi(\delta_{T(Q^-)}, \delta_{T(Q^+)}) < \epsilon$, a contradiction. □

Proof of Theorem 2. Suppose that $\theta, \tau \in T(P)$, $\theta \neq \tau$. By the regularity of T, there are Q_ν^- and Q_ν^+ that satisfy the assumptions of Proposition 2. Hence $\theta = \tau$. The same argument yields that θ must be equal to the limit (possibly in a one-point compactification of the range of T) of any other sequence $T(P_\nu)$ such that $P_\nu \to P$. Hence, T has a unique limit at P, equal to $T(P)$. □

4. Final remarks

After the introduction by Hampel [11], which reappeared in the more settled form in Hampel, Ronchetti, Rousseeuw and Stahel (1986), and the influential treatment by Huber [15], all in the context of estimation and independent sampling, qualitative robustness was extended to hypothesis testing framework by Lambert [16] and Rieder [24]; dependent data models of time series flavor were considered by Papantoni-Kazakos [21], Boente et al. [2]; some further theoretical aspects were addressed by Cuevas [3]. It seems that despite these developments, its use for evaluating robustness was not too intense: a few relevant references are Rieder [25], Good and Smith [9], Cuevas and Sanz [4], Machado [17], and He and Wang [13]. The fade-out citation pattern is indicated by the only 21st century exception retrieved from `scholar.google.com`, Daouia and Ruiz-Gazen [5].

As the name indicates, and the definition clearly shows, qualitative robustness does not provide any "quantitative" appraisal: the procedure is judged either not robust or robust—and in the latter case we do not know "how much". The rush for "more" and "most" robust methods might have been the reason that other robustness criteria gained more following. Nevertheless, given the multitude of "desirable features" considered in the screening of aspiring data-analytic techniques, qualitative robustness may be just enough to draw a dividing line in the territory of robustness—especially in complex situations where classical criteria modeled in standard circumstances may loose steam.

In the universe of mathematical sciences, qualitative robustness is similar to the notion of stability used in the theory of differential equations: a small change in initial conditions still renders the new solution staying in a tube enclosing the original one. Interestingly, the translation of "qualitatively robust at P" to "solution exists, is unique, and depends continuously on the data", discussed in this note, corresponds exactly to what in applied mathematics is called well-posed problem in the sense of Hadamard [10]. Indeed, continuous dependence on the data is essential for any procedure, in particular for its numerical implementation—which is always based on approximation; it ensures the stability of an algorithm. A referee pointed out that among the references given above, we might have missed some that, like Hildebrand and Müller [14], refer more generally to "robustness" or "continuity" without mentioning explicitly qualitative robustness. In the similar spirit, we may see witness a resurrection of the term (likely under a different name) in learning theory — see Poggio, Rifkin, Mukherjee, and Niyogi (2004).

Of course, numerical stability requires only pointwise continuity; qualitative robustness goes a step further, requiring continuity with respect to the distribution underlying the data. Some may argue that it goes too far, indicating that continuity violated by statistical procedures otherwise in common use may be too stringent a requirement. In the context of well-posedness in the sense of Hadamard [10], the usual mode of requiring continuity is "in some reasonable topology". All this indicates that the most important aspect of qualitative robustness, and robustness theory in general, lies at its very start—as pointed out by Davies [6], and put down already by Huber [15],

> "It is by no means clear whether different metrics give rise to equivalent robustness notions; to be specific we work with Lévy metric for F and the Prokhorov metric for $\mathcal{L}(T_n)$."

We remark that such a choice might came out as natural under the influence of Billingsley [1] in the times of Hampel [11]; the question is whether it still remains such. Of course, the relationship between qualitative robustness and continuity discussed in this note indicates that it is only the induced topology, not a particular metric, that matters for *qualitative* robustness.

References

[1] Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[2] Boente, G., Fraiman, R. and Yohai, V. J. (1987). Qualitative robustness for stochastic processes. *Ann. Statist.* **15** 1293–1312.

[3] Cuevas, A. (1988). Quantitative robustness in abstract inference. *J. Statist. Planning Inference* **18** 277–289.

[4] Cuevas, A. and Sanz, P. (1989). A class of qualitatively robust estimates. *Statistics* **20** 509–520.

[5] Daouia, A. and Ruiz–Gazen, A. (2006). Robust nonparametric frontier estimators: Qualitative robustness and influence function. *Statistica Sinica* **16** 1233–1253.

[6] Davies, P. L. (1993). Aspects of robust linear regression. *Ann. Statist.* **21** 1843–1899.

[7] Dudley, R. M. Giné, E. and Zinn, J. (1991). Uniform and universal Glivenko-Cantelli classes. *J. Theoret. Probab.* **4** 485–510.

[8] Ellis, S. P. (1998). Instability of least squares, least absolute deviation and least median of squares linear regression. *Statist. Sci. 13* 337–350.

[9] Good, I. J. and Smith, E. P. (1986). An additive algorithm analogous to the singular decomposition or a comparison of polarization and multiplicative models: An example of qualitative robustness. *Commun. Statist. B* **15** 545–569.

[10] Hadamard, J. (1902). Sur les problèmes aux dérivées et leur signification physique. *Princeton University Bulletin* 49–52.

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

[12] 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.

[13] He, X. and Wang, G. (1997). Qualitative robustness of S^*-estimators of multivariate location and dispersion. *Statistica Neerlandica* **51** 257–268.

[14] Hildebrand, M. and Müller, C. H. (2007). Outlier robust corner-preserving methods for reconstructing noisy images. *Ann. Statist.* **35** 132–165.

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

[16] Lambert, D. (1982). Qualitative robustness of tests. *J. Amer. Statist. Assoc.* **77** 352–357.

[17] Machado, J. A. F. (1993). Robust model selection and M-estimation. *Econometric Theory* **9** 478–493.

[18] Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods.* Wiley, New York.

[19] Mizera, I. (1995). A remark on existence of statistical functionals. *Kybernetika* **31** 315–319.

[20] Mizera, I. and Volauf, M. (2002). Continuity of halfspace depth contours and maximum depth estimators: Diagnostics of depth-related methods. *Journal of Multivariate Analysis* **83** 365–368.

[21] Papantoni-Kazakos, P. (1984). Some aspects of qualitative robustness in time series. In *Robust and Nonlinear Time Series Analysis,*(J. Franke, W.

Härdle and D. Martin, eds.) *Lecture Notes in Statistics* **26** 218–230. Springer-Verlag, New York.

[22] Poggio, T., Rifkin, R., Mukherjee, S., and Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature* **428** 419–422.

[23] Portnoy, S. and Mizera, I. (1998). Comment of "Instability of least squares, least absolute deviation and least median of squares linear regression". *Statistical Science* **13** 344–347.

[24] Rieder, H. (1982). Qualitative robustness of rank tests. *Ann. Statist.* **10** 205–211.

[25] Rider, H. (1983). Continuity properties of rank procedures. *Statist. Decisions* **1** 341–369.

[26] Strassen, V. (1976). The existence of probability measures with given marginals. *Ann. Math. Statist.* **36**423–439.