PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 20, Fasc. 2 (2000), pp. 223–235

STATISTICAL INFERENCE FROM SET-VALUED OBSERVATIONS

BY

TOMASZ SCHREIBER* (TORUŃ)

Abstract. Consider a random experiment whose true (unknown) outcome is modelled by a certain random element X and the available imprecise observations are modelled by some random set A such that $X \in A$ almost surely. The purpose of the paper is to propose a statistical procedure for estimation of the real distribution of X. The asymptotic properties of the suggested procedure are then investigated in both nonparametric and parametric settings. So far, only the results for a finite sample space are available.

1. INTRODUCTION AND GENERAL RESULTS

One of the interests underlying the development of the random sets theory (see [5] and [7] for extensive reference) is that it provides a natural framework allowing one to represent in an elegant way the imprecision of the data available to a statistician. It happens very often, due to the imperfection of the data acquiring procedures (inaccuracy of the measuring instruments, influence exerted upon the state of the observed system by the act of measurement etc.), that it may be preferable to represent the outcome of an experiment as a set to which the real value of the measured quantity belongs rather than trying to ascribe to it some unique value (e.g. the centre of the uncertainty interval) or a subjectively chosen probability distribution. From the mathematical point of view, in such situations it is particularly convenient to model the observation as a random set containing almost surely the random element corresponding to the true result of the experiment. A question which arises naturally in these circumstances is how to infer from such 'set-valued' observations some information about the distribution of the true value of the random experiment. In other words, if the observation is modelled by a certain random set A, the question is how to find the distribution of its selector X which would be the optimal within the prespecified statistical model. Note that by a *selector of the random set* A we mean each random element X such that $X \in A$ almost surely.

* Faculty of Mathematics and Computer Science, Nicholas Copernicus University, Toruń.

The main goal of this paper is to provide a tool suitable for the above-mentioned purposes and to investigate the quality of estimation it provides. Although we aim at extending the suggested procedure for the continuous case, at the moment we have the rigorous results in the case of finite sample spaces only.

It seems particularly convenient to use in the discussion of the afore-mentioned topics the standard language of the random sets theory (see ibidem), even though there exist some alternatives such as the theory of evidence and belief functions (see [6] and the references therein). Our choice is motivated by the fact that the theory of random sets is particularly rich and it provides a well-founded formalism which can turn out indispensable for suitable generalisations of the theorems presented in this paper.

Consider a finite sample space $\mathcal{X} = \{x_1, \ldots, x_k\}$, where $x_1, \ldots, x_k$ represent the possible true values of a certain experiment. Let $\mathcal{S} := 2^{\mathcal{X}} \setminus \{\emptyset\}$ be the space of all the nonempty subsets of $\mathcal{X}$, which model the possible set-valued observations, always assumed to contain the real outcome. To represent the random mechanism governing the observations we consider the random elements taking values in $\mathcal{S}$, which will be referred to as random nonempty subsets of $\mathcal{X}$ (for short, random sets, provided it does not lead to confusion). Throughout the paper it is assumed without a further mention that all the considered random elements are defined on a common probability space $(\Omega, \mathfrak{F}, \boldsymbol{P})$.

Each random set A is uniquely determined by its hitting functional T_A defined on the power set $T_A: 2^{\mathcal{X}} \to [0, 1]$ by

$$T_A(\mathcal{E}) := \boldsymbol{P}(A \cap \mathcal{E} \neq \emptyset), \qquad \mathcal{E} \subseteq \mathcal{X} \tag{1}$$

(see Section 1.1 in [7] or Chapter 1 in [5]). It is easily verified that T_A satisfies the following conditions:

(T1) $T_A(\emptyset) = 0$, $T_A(\mathcal{X}) = 1$.

(T2) The following functionals recursively defined by

$$S_1(K_0; K_1) := T_A(K_0 \cup K_1) - T_A(K_0),$$

. . . . . .

$$S_n(K_0; K_1, \ldots, K_n) := S_{n-1}(K_0; K_1, \ldots, K_{n-1}) - S_{n-1}(K_0 \cup K_n; K_1, \ldots, K_{n-1})$$

are nonnegative for all $n \geqslant 0$ and $K_0, K_1, \ldots, K_n \subset \mathcal{X}$.

Indeed, (T1) follows from the nonemptiness of A, and to see that (T2) holds it suffices to note that $S_n(K_0; K_1, \ldots, K_n)$ is the probability that A misses K_0 but hits $K_1, \ldots, K_n$. In particular, T_A is monotone because S_1 is nonnegative. Further, it turns out that (T1) and (T2) characterise the class of the hitting functionals of random nonempty subsets of $\mathcal{X}$. More precisely, if some function $T: 2^{\mathcal{X}} \to [0, 1]$ satisfies (T1) and (T2), then there exists a random nonempty subset A of $\mathcal{X}$ such that $T = T_A$ (see ibidem; note that in view of the finiteness

of $\mathscr{X}$ we could omit some conditions of topological nature taken into account in the general case).

Denote by $\mathscr{C}$ the class of all the functionals T satisfying (T1) and (T2) (such functionals are called *alternating Choquet capacities of infinite order*, see ibidem). Clearly, $\mathscr{C}$ can be regarded as a compact convex subset of $[0, 1]^{2^k}$. Further, let $\mathscr{P}$ denote the space of all the probability measures on $\mathscr{X}$. Obviously, $\mathscr{P}$ can also be identified with a compact convex subset of $[0, 1]^k$. We endow $\mathscr{P}$ and $\mathscr{C}$ with the respective Euclidean metrics, denoted both by 'dist'.

Given a random set A it is crucial for our purposes to ask for a characterisation of the class of the distributions of all its possible selectors (i.e. $\mathscr{X}$-valued random elements belonging almost surely to A), which obviously correspond to all the possible distributions of the true outcome of the experiment. It turns out that this class can be described in a natural way in terms of the hitting functional T_A. We say that a probability distribution $\mu \in \mathscr{P}$ is *dominated by the capacity* $T \in \mathscr{C}$ and we write $\mu \preccurlyeq T$ iff $\mu(\mathscr{E}) \leqslant T(\mathscr{E})$ for all $\mathscr{E} \subset \mathscr{X}$. The following proposition is then a conclusion of Theorem 1 in [8] or Theorem A in [9].

PROPOSITION 1. *Let A be a nonempty random subset of $\mathscr{X}$ and $\mu \in \mathscr{P}$. Then the following conditions are equivalent*:

(1) $\mu \preccurlyeq T_A$.

(2) *There exists a probability space carrying versions of the random set A and of an $\mathscr{X}$-valued random element X with distribution μ such that $X \in A$ almost surely.*

The class of all the probability measures dominated by a given capacity is referred to as its core, that is

$$\operatorname{core}(T) = \{\mu \in \mathscr{P} \mid \mu \preccurlyeq T\}$$

(see, e.g., [4]). Note that $\operatorname{core}(T)$ is a compact and convex subset of $\mathscr{P}$. Obviously, from Proposition 1 it follows immediately that $\operatorname{core}(T)$ is nonempty.

Let us agree to call a *discrepancy measure* on $\mathscr{P} \times \mathscr{P}$ each nonnegative function $\Delta: \mathscr{P} \times \mathscr{P} \to \mathbf{R}_+ \cup \{+\infty\}$ satisfying the following conditions:

(A1) $\Delta(\mu, \nu) = 0$ iff $\mu = \nu$.

(A2) Δ is lower semicontinuous on $\mathscr{P} \times \mathscr{P}$.

(A3) Δ is convex on $\mathscr{P} \times \mathscr{P}$ and strictly convex on its domain of finiteness $\{(\mu, \nu) \in \mathscr{P} \times \mathscr{P} \mid \Delta(\mu, \nu) < +\infty\}$.

Note that we do not require the symmetry. The representative examples of such discrepancy measures are the relative entropy (Kullback–Leibler divergence)

$$\Delta^H(\mu, \nu) := \sum_{i=1}^{k} \log\left(\mu(\{x_i\})/\nu(\{x_i\})\right) \mu(\{x_i\}),$$

the l_∞-distance

$$\Delta^{(\infty)}(\mu, \nu) := \max_{i=1,\ldots,k} |\mu(\{x_i\}) - \nu(\{x_i\})|$$

and the l_p-distance

$$\Delta^{(p)}(\mu, \nu) := \Big(\sum_{i=1}^{k} |\mu(\{x_i\}) - \nu(\{x_i\})|^p\Big)^{1/p} \quad \text{for } p \geqslant 1.$$

Observe that $\Delta^{(2)}$ coincides with the usual Euclidean distance 'dist'.

Let us fix some discrepancy measure Δ and extend it from $\mathscr{P} \times \mathscr{P}$ onto $\mathscr{C} \times \mathscr{P}$ by setting for $T \in \mathscr{C}$ and $\nu \in \mathscr{P}$

$$\Delta(T, \nu) := \inf_{\mu \in \mathrm{core}(T)} \Delta(\mu, \nu).$$

We conclude from (A2) and (A3) that if $\Delta(T, \nu)$ is finite, then there exists a unique probability distribution $\nu[T] \in \mathscr{P}$ such that

$$\Delta(\nu[T], \nu) = \Delta(T, \nu). \tag{2}$$

In other words, for a random set A, $\nu[T_A]$ is the distribution of its selector which is the best approximation of ν with respect to Δ. In particular, if ν itself is a distribution of a certain selector of A, then, by (A1), $\Delta(T_A, \nu) = 0$ and $\nu[T_A] = \nu$.

The problems related to the algorithmic construction of $\nu[T]$ for a given T in the particular case of ν uniform on $\mathscr{X}$ with Δ chosen to be the relative entropy have been considered by several authors (see, e.g., [4] or [6]).

The following lemma, proved below, gives us some properties of the extension of Δ onto $\mathscr{C} \times \mathscr{P}$.

LEMMA 1. *The mapping $\Delta: \mathscr{C} \times \mathscr{P} \to \mathbf{R}_+ \cup \{+\infty\}$ is lower semicontinuous and convex. In addition, it is strictly convex on its domain of finiteness. Further, if $\nu \in \mathscr{P}$ is such that the function $\mathscr{P} \ni \mu \mapsto \Delta(\mu, \nu)$ is continuous, then also $\mathscr{C} \ni T \mapsto \Delta(T, \nu)$ is continuous.*

We are now ready to give a formal description of the general statistical problem investigated in this paper. Let $\Xi \subset \mathscr{P}$ be a certain nonempty closed (and, therefore, compact) family of distributions on $\mathscr{X}$ and let A be the random nonempty subset of $\mathscr{X}$ with the corresponding hitting functional T_A. As stressed in the informal introduction, we aim at finding a selector X of A so that it would be the closest possible to the class of distributions Ξ specified by the statistical model. Define for each capacity $T \in \mathscr{C}$

$$\hat{\Delta}(T \mid \Xi) := \inf_{\mu \in \Xi} \Delta(T, \mu).$$

It follows from Lemma 1 and the compactness of Ξ that there exists (not necessarily unique) $\nu \in \Xi$ such that

$$\Delta(T, \nu) = \hat{\Delta}(T \mid \Xi).$$

In particular, if $\hat{\Delta}(T \mid \Xi) < \infty$, for such ν the distribution $\nu[T]$ satisfying (2) is well defined and we have $\hat{\Delta}(T \mid \Xi) = \Delta(\nu[T], \nu)$. Here, $\nu[T]$ is to be regarded as the most likely (the closest to the prespecified model Ξ) distribution of the true outcome of the experiment which led to the observations modelled by the hitting functional T. Further, ν is the best approximation, with respect to Δ, of $\nu[T]$ within the model Ξ. In general, it is natural to define the (necessarily nonempty) set $\Xi_A \subset \Xi$:

$$\Xi_A := \{\nu \in \Xi \mid \Delta(T_A, \nu) = \hat{\Delta}(T_A \mid \Xi)\}. \tag{3}$$

Note that from Lemma 1 and from the compactness of Ξ it follows that Ξ_A is compact. The set Ξ_A consists of these distributions from Ξ which 'fit the best' to the observations modelled by A. In particular, if Ξ contains at least one distribution of some selector of A, then $\hat{\Delta}(T_A \mid \Xi) = 0$ and Ξ_A consists of the distributions of all the selectors of A belonging to Ξ.

To proceed we need an additional regularity assumption on Δ. Namely, we require that

(A4) For each $\nu \in \Xi$ the mapping $\mathscr{P} \ni \mu \mapsto \Delta(\mu, \nu)$ is continuous.

Note that (A4) holds automatically if we use $\Delta^{(\infty)}$ or $\Delta^{(p)}$, $p \geqslant 1$, as the discrepancy measure. On the other hand, if we choose Δ to be the relative entropy, an additional assumption on Ξ in needed to guarantee that (A4) is satisfied. For instance, it is enough to require that $\nu(\{x_i\}) > 0$ for all $\nu \in \Xi$ and $i = 1, \ldots, k$.

Our main goal will be to identify the elements of the 'optimal' set Ξ_A. In the particular case, if Ξ_A is a singleton $\{\nu_0\}$, our task reduces to the estimation of ν_0. Let $A, A_1, A_2, \ldots$ be an i.i.d. sequence of i.i.d. copies of the random set A corresponding to the successive observations. The empirical information provided by the observations can be identified with the empirical hitting functional defined, by analogy with (1), as

$$T_A^{(n)}(\mathscr{E}) := \boldsymbol{P}^{(n)}(A \cap \mathscr{E} \neq \emptyset) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{A_j \cap \mathscr{E} \neq \emptyset\}}$$

with $\boldsymbol{P}^{(n)}$ denoting the empirical probability. As in (3) we define

$$\Xi_A^{(n)} := \{\nu \in \Xi \mid \Delta(T_A^{(n)}, \nu) = \hat{\Delta}(T_A^{(n)} \mid \Xi)\}.$$

Observe that, by Lemma 1, $\Xi_A^{(n)}$ is almost surely compact and nonempty. Clearly, $T_A^{(n)}: (\Omega, \mathfrak{F}, \boldsymbol{P}) \to \mathscr{C}$ is measurable. Further, one can show the measurability of $\Xi_A^{(n)}$ regarded as a mapping from $(\Omega, \mathfrak{F}, \boldsymbol{P})$ to the space of all the nonempty compact subsets of Ξ endowed with the usual Hausdorff metric. Since the only interest we have in this fact is to argue that $\Xi_A^{(n)}$ is a well-defined random element, we avoid the details and confine ourselves to a short sketch of the proof, referring the reader to [5] for appropriate techniques. First we note that by Lemma 2 (see Section 3) the mapping $\mathscr{C} \ni T \mapsto \{\nu \in \Xi \mid \Delta(T, \nu) = \hat{\Delta}(T \mid \Xi)\}$ is

upper semicontinuous and, therefore, measurable (see Definition 1.2.2 and Proposition 1.2.4 in [5]). Then, our assertion follows by the measurability of $T_A^{(n)}$ and standard arguments.

It is natural to expect that the elements of $\Xi_A^{(n)}$ are very close to Ξ_A for n large enough. The following theorem confirms this conjecture.

THEOREM 1. *We have almost surely*

$$\lim_{n\to\infty} \hat{\Delta}(T_A^{(n)} \mid \Xi) = \hat{\Delta}(T_A \mid \Xi) \quad \text{and} \quad \lim_{n\to\infty} \sup_{v\in\Xi_A^{(n)}} \operatorname{dist}(v, \Xi_A) = 0,$$

where $\operatorname{dist}(v, \Xi_A)$ *denotes the Euclidean distance between* v *and* Ξ_A.

It is worth noting that in general it may happen that

$$\limsup_{n\to\infty} \sup_{v\in\Xi_A} \operatorname{dist}(v, \Xi_A^{(n)}) > 0$$

with some positive probability. In other words, although Theorem 1 guarantees that asymptotically each element of the estimator $\Xi_A^{(n)}$ can be approximated by some element of Ξ_A, yet there may exist some elements of Ξ_A which are not approximated by any elements of $\Xi_A^{(n)}$. Clearly, this problem disappears if Ξ_A is a singleton. In this case Theorem 1 yields readily

COROLLARY 1. *If* Ξ_A *contains only the unique element* $v^{(0)}$ *and the sequence of* $\mathscr{P}$*-valued random elements* $(v^{(n)})_{n=0}^{\infty}$ *is such that almost surely* $v^{(n)} \in \Xi_A^{(n)}$, *then with probability* 1

$$\lim_{n\to\infty} \operatorname{dist}(v^{(n)}, v^{(0)}) = 0.$$

The question that immediately arises in this context is what is the accuracy of approximation of the optimal discrepancy $\hat{\Delta}(T_A \mid \Xi)$ by $\hat{\Delta}(T_A^{(n)} \mid \Xi)$ and, more generally, of Ξ_A by $\Xi_A^{(n)}$. Unlike the second one, the first of these questions can be given a partial answer without any additional regularity assumptions. This is done in the following large deviations type theorem. We refer the reader to Section 1.1 in [2] or Section 2.1 in [1] for the terminology used in the large deviations theory.

THEOREM 2. *For each mapping* $\Phi: 2^{\mathscr{X}} \to \boldsymbol{R}$ *define*

$$\mathscr{L}_A(\Phi) := \boldsymbol{E} \exp\Big(\sum_{\mathscr{E}\subset\mathscr{X}, \mathscr{E}\cap A\neq\emptyset} \Phi(\mathscr{E}) \Big) \tag{4}$$

and let

$$\mathscr{I}_A(T) := \sup_{\Phi: 2^{\mathscr{X}}\to\boldsymbol{R}} \Big(\sum_{\mathscr{E}\subset\mathscr{X}} \Phi(\mathscr{E})\, T(\mathscr{E}) - \log \mathscr{L}_A(\Phi) \Big) \tag{5}$$

for every capacity $T \in \mathscr{C}$. *Then the sequence* $\big(\hat{\Delta}(T_A^{(n)} \mid \Xi)\big)_{n=0}^{\infty}$ *satisfies on* $\boldsymbol{R}$ *the*

large deviation principle with the rate function

$$\mathscr{R}_A(h) := \inf\{\mathscr{I}_A(T) \mid T \in \mathscr{C},\ \hat{\Delta}(T \mid \Xi) = h\}, \tag{6}$$

i.e. $\mathscr{R}_A$ *has compact level sets* $\{u \in \boldsymbol{R} \mid \mathscr{R}_A(u) \leqslant M\}$ *for* $M \geqslant 0$, *for each open set* $\mathscr{G} \subset \boldsymbol{R}$

$$\liminf_{n\to\infty} n^{-1} \log \boldsymbol{P}\left(\hat{\Delta}(T_A^{(n)} \mid \Xi) \in \mathscr{G}\right) \geqslant -\inf_{h\in\mathscr{G}} \mathscr{R}_A(h),$$

and for each closed set $\mathscr{F} \subset \boldsymbol{R}$

$$\limsup_{n\to\infty} n^{-1} \log \boldsymbol{P}\left(\hat{\Delta}(T_A^{(n)} \mid \Xi) \in \mathscr{F}\right) \leqslant -\inf_{h\in\mathscr{F}} \mathscr{R}_A(h).$$

In addition, $\mathscr{R}_A(h)$ *is strictly positive for* $h \neq \hat{\Delta}(T_A \mid \Xi)$.

Although in practice it can be very difficult to compute the rate function explicitly, the value of Theorem 2 is that applying the large deviation principle for the closed set $(-\infty, -\varepsilon] \cup [\varepsilon, \infty)$, $\varepsilon > 0$, yields immediately the exponential convergence of $\hat{\Delta}(T_A^{(n)} \mid \Xi)$ to $\hat{\Delta}(T_A \mid \Xi)$ in the following sense:

COROLLARY 2. *For each* $\varepsilon > 0$ *there exists* $L_\varepsilon > 0$ *such that for* n *large enough*

$$\boldsymbol{P}\left(\left|\hat{\Delta}(T_A \mid \Xi) - \hat{\Delta}(T_A^{(n)} \mid \Xi)\right| \geqslant \varepsilon\right) < \exp(-nL_\varepsilon).$$

In the next section we investigate the properties of our statistical procedure under a regular parametric setting, which yields further more detailed asymptotic results.

2. PARAMETRIC CASE

Let $\Theta \subset \boldsymbol{R}^m$ be a certain Borel set of parameters and suppose that

$$\Xi = \{\mu_\theta \mid \theta \in \Theta\}.$$

In classical parametric statistics the maximum likelihood estimators are, under certain regularity conditions, asymptotically normal with variance given by the inverted Fisher's information matrix (see, e.g., [3]). Below we present analogous asymptotic normality results for the procedure investigated in this paper.

As remarked before, $\mathscr{P}$ and $\mathscr{C}$ can be identified with convex compact subsets of $[0, 1]^k$ and $[0, 1]^{2^k}$, respectively. Thus, the notion of differentiation with respect to $T \in \mathscr{C}$ and $\mu \in \mathscr{P}$ is well defined on the relative interiors $\mathrm{ri}(\mathscr{C})$ and $\mathrm{ri}(\mathscr{P})$ of $\mathscr{C}$ and $\mathscr{P}$, respectively, where $\mathrm{ri}(\mathscr{C})$ is the interior of $\mathscr{C}$ which results when $\mathscr{C}$ is regarded as a subset of its affine hull and $\mathrm{ri}(\mathscr{P})$ is defined analogously (see, e.g., [10]). Let us agree to denote the operation of taking these derivatives

by ∇_T and ∇_μ, respectively. It is clear that the affine hull aff $(\mathscr{P})$ has dimension $k-1$. Moreover, it is easily proved that

$$\text{aff}(\mathscr{C}) = \{\Phi\colon 2^{\mathscr{X}} \to \boldsymbol{R} \mid \Phi(\emptyset) = 0,\ \Phi(\mathscr{X}) = 1\}, \tag{7}$$

so, in particular, $\dim(\text{aff}(\mathscr{C})) = 2^k - 2$. To verify (7) consider the random set $B := \mathscr{X}\backslash\{x_\eta\}$ with η distributed uniformly on $\{1, 2, \ldots, k\}$ and check that in (T2) the quantities $S_n(K_0; K_1, \ldots, K_{n-1})$ are strictly positive provided none of $K_1, \ldots, K_{n-1}$ is contained in K_0. This means that $\mathscr{C}$ contains some ball in $\{\Phi\colon 2^{\mathscr{X}} \to \boldsymbol{R} \mid \Phi(\emptyset) = 0,\ \Phi(\mathscr{X}) = 1\}$ centred in T_B. Relation (7) allows us to identify the differentiation with respect to $T \in \mathscr{C}$ with the differentiation with respect to $T(\mathscr{E}_1), \ldots, T(\mathscr{E}_{2^k-2})$, where $(\mathscr{E}_1, \ldots, \mathscr{E}_{2^k-2})$ is a fixed sequence of all the subsets of $\mathscr{X}$ except for $\emptyset$ and $\mathscr{X}$.

Throughout this section we will assume that the following regularity conditions are satisfied:

(U) $\hat{\Delta}(T_A \mid \Xi) < \infty$ and there exists unique $\theta_0 \in \Theta$ such that $\Delta(T_A, \mu_{\theta_0}) = \hat{\Delta}(T_A \mid \Xi)$. In addition, θ_0 lies in the topological interior of Θ.

(C) If a sequence $(\theta_n)_{n=0}^{\infty} \subset \Theta$ is such that $\mu_{\theta_n} \to \mu_{\theta_0}$, then also $\theta_n \to \theta_0$.

(D) The function $\Delta_\Theta\colon \mathscr{C} \times \Theta \to \boldsymbol{R} \cup \{+\infty\}$ defined by $\Delta_\Theta(T, \theta) := \Delta(T, \mu_\theta)$ is twice continuously differentiable in a neighbourhood of (T_A, θ_0).

(I) The $(m \times m)$-matrix of the second derivative of Δ_Θ in (T_A, θ_0) with respect to θ, denoted by $V_0 := [\nabla^2_{\theta,\theta} \Delta_\Theta](T_A, \theta_0)$, is invertible.

Note that condition (D) is satisfied for instance if Δ is twice continuously differentiable with respect to μ and T in some neighbourhood of (T_A, μ_{θ_0}) and the parametrisation mapping $\theta \mapsto \mu_\theta(\{x_i\})$ is twice continuously differentiable in a neighbourhood of θ_0 for $i = 1, \ldots, k$.

To proceed, consider a sequence of Θ-valued random vectors $\hat{\theta}_n$ such that almost surely

$$\mu_{\hat{\theta}_n} \in \Xi_A^{(n)}$$

for all $n \in \boldsymbol{N}$ (thus, if $\Xi_A^{(n)}$ contains more than one element, $\hat{\theta}_n$ is chosen arbitrarily, the only requirement being the measurability of $\hat{\theta}_n$). It follows immediately from condition (C) and Corollary 1 that $\hat{\theta}_n$ is a consistent estimator of θ_0, i.e.

COROLLARY 3. *With probability* 1

$$\lim_{n \to \infty} \hat{\theta}_n = \theta_0.$$

The following theorem, stating the asymptotic normality of $\hat{\theta}_n$, is the main result of this section.

THEOREM 3. *Assume that conditions* (U), (C), (D) *and* (I) *are satisfied and denote by* W_0 *the* $m \times (2^k - 2)$*-matrix*

$$W_0 := [\nabla^2_{\theta,T} \Delta_\Theta](T_A, \theta_0),$$

where $\nabla^2_{\theta,T}$ is the second order partial derivative with respect to θ and T. Further, let Π be the $(2^k-2)\times(2^k-2)$-matrix with entries

$$\Pi_{i,j} = \Pi_{j,i} := \boldsymbol{P}(A\cap\mathscr{E}_i \neq \emptyset \wedge A\cap\mathscr{E}_j \neq \emptyset) - T_A(\mathscr{E}_i)\, T_A(\mathscr{E}_j). \tag{8}$$

Then the sequence $\sqrt{n}(\hat{\theta}_n-\theta_0)$ converges in distribution to the mean zero normal law with covariance matrix $V_0^{-1} W_0 \Pi W_0^T V_0^{-1}$.

As in Theorem 2, also here the character of the result is mainly qualitative, because the explicit computation of the covariance matrix $V_0^{-1} W_0 \Pi W_0^T W_0^{-1}$ is, even for moderate k, a lengthy and difficult task due to both the size of the matrices involved and the complicated character of Δ on $\mathscr{C}\times\mathscr{P}$. Nevertheless, in some particular cases one may try to estimate the covariance matrix by means of Monte-Carlo methods.

3. PROOFS

We will proceed as follows. First we give the proofs of Lemmas 1 and 2 in which we establish the continuity and other relevant properties of Δ and $\hat{\Delta}$, respectively. The next step is to deduce Theorems 1 and 2 from Lemma 2. Finally, we establish Theorem 3.

3.1. Proof of Lemma 1. To see that Δ is convex on $\mathscr{C}\times\mathscr{P}$ take some $\alpha\in(0,1)$, $T_1, T_2\in\mathscr{C}$ and $\nu_1, \nu_2\in\mathscr{P}$. We are to show that

$$\Delta(\alpha T_1+(1-\alpha)T_2, \alpha\nu_1+(1-\alpha)\nu_2) \leqslant \alpha\Delta(T_1,\nu_1)+(1-\alpha)\Delta(T_2,\nu_2). \tag{9}$$

Clearly, we can assume that both $\Delta(T_1,\nu_1)$ and $\Delta(T_2,\nu_2)$ are finite, for otherwise (9) holds trivially. Thus, $\nu_1[T_1]$ and $\nu_2[T_2]$ are well defined (see (2)) and, obviously, $\alpha\nu_1[T_1]+(1-\alpha)\nu_2[T_2]\in\operatorname{core}(\alpha T_1+(1-\alpha)T_2)$. Thus, by the convexity of $\Delta\colon \mathscr{P}\times\mathscr{P}\to\boldsymbol{R}_+\cup\{\infty\}$ we get

$$\begin{aligned}(10)\quad &\Delta(\alpha T_1+(1-\alpha)T_2, \alpha\nu_1+(1-\alpha)\nu_2)\\ &\leqslant \Delta(\alpha\nu_1[T_1]+(1-\alpha)\nu_2[T_2], \alpha\nu_1+(1-\alpha)\nu_2)\\ &\leqslant \alpha\Delta(\nu_1[T_1],\nu_1)+(1-\alpha)\Delta(\nu_2[T_2],\nu_2) = \alpha\Delta(T_1,\nu_1)+(1-\alpha)\Delta(T_2,\nu_2),\end{aligned}$$

which yields (9) and, therefore, the convexity of Δ on $\mathscr{C}\times\mathscr{P}$. If in addition $T_1\neq T_2$ or $\nu_1\neq\nu_2$, inequality in (10) becomes strict by the strict convexity of $\Delta\colon\mathscr{P}\times\mathscr{P}\to\boldsymbol{R}\cup\{+\infty\}$ on its domain of finiteness, which gives us in turn the strict convexity of $\Delta\colon\mathscr{C}\times\mathscr{P}\to\boldsymbol{R}\cup\{+\infty\}$ on its domain of finiteness.

To prove the lower semicontinuity of Δ on $\mathscr{C}\times\mathscr{P}$ choose an arbitrary sequence $(T_n,\nu_n)_{n=0}^{\infty}$ convergent to some $(T_\infty,\nu_\infty)\in\mathscr{C}\times\mathscr{P}$. Since the only nontrivial case we have to consider is $\Delta(T_n,\nu_n)<\infty$ infinitely often, we can assume without loss of generality that $\Delta(T_n,\nu_n)<\infty$ for all n, optionally passing to an appropriate sequence. Thus, the distributions $\nu_n[T_n]$ given by (2) are well defined. Taking into account the compactness of $\mathscr{P}$ we can assume that $\nu_n[T_n]$

converges to some μ_∞. Thus, in view of the lower semicontinuity of Δ on $\mathscr{P} \times \mathscr{P}$ (see (A2)) and taking into account that $\mu_\infty \preccurlyeq T_\infty$, we conclude that

$$\liminf_{n\to\infty} \Delta(T_n, \nu_n) = \liminf_{n\to\infty} \Delta(\nu_n[T_n], \nu_n) \geqslant \Delta(\mu_\infty, \nu_\infty) \geqslant \Delta(T_\infty, \nu_\infty),$$

as required.

It remains to establish the continuity of $\mathscr{C} \ni T \mapsto \Delta(T, \nu)$ for each ν for which the function $\mathscr{P} \ni \mu \mapsto \Delta(\mu, \nu)$ is continuous. Fix such ν and note that $T \mapsto \Delta(T, \nu)$ is lower semicontinuous by the preceding part of the proof. Thus, it suffices to prove the upper semicontinuity of this mapping, i.e.

$$\limsup_{n\to\infty} \Delta(T_n, \nu) \leqslant \Delta(T_\infty, \nu) \tag{11}$$

for $T_n \to T_\infty$. Assume that $\Delta(T_\infty, \nu) < \infty$, for otherwise (11) becomes obvious. Then the distribution $\nu[T_\infty]$ is well defined and we can choose a sequence $(\mu_n)_{n=0}^\infty \subset \mathscr{P}$ convergent to $\nu[T_\infty]$ and such that $\mu_n \preccurlyeq T_n$. Therefore, by the choice of ν,

$$\limsup_{n\to\infty} \Delta(T_n, \nu) \leqslant \limsup_{n\to\infty} \Delta(\mu_n, \nu) = \Delta(\nu[T_\infty], \nu) = \Delta(T_\infty, \nu),$$

which yields (11). The proof is complete. ∎

3.2. Lemma 2. The following lemma is crucial for Theorems 1 and 2.

LEMMA 2. *Let $(T_n)_{n=0}^\infty \subset \mathscr{C}$ be a sequence of capacities converging to a certain capacity $T_\infty \in \mathscr{C}$. Let further*

$$\Sigma^{(n)} := \{\nu \in \Xi \mid \Delta(T_n, \nu) = \hat{\Delta}(T_n | \Xi)\}$$

for $n = 1, 2, \ldots, \infty$. Then $\Sigma^{(n)} \subseteq \Xi$ are compact and nonempty,

$$\lim_{n\to\infty} \hat{\Delta}(T_n | \Xi) = \hat{\Delta}(T_\infty | \Xi) \tag{12}$$

and

$$\lim_{n\to\infty} \sup_{\nu\in\Sigma^{(n)}} \operatorname{dist}(\nu, \Sigma^{(\infty)}) = 0. \tag{13}$$

Proof. The compactness and nonemptiness of $\Sigma^{(n)}$ follow immediately from the compactness of Ξ and from Lemma 1.

Further, using (A4) and Lemma 1 we conclude that $\hat{\Delta}(\cdot | \Xi)$ is upper semicontinuous as the minimum of the family $\{\Delta(\cdot | \nu) \mid \nu \in \Xi\}$ of continuous functions.

Hence, to establish the continuity of the mapping $\hat{\Delta}(\cdot | \Xi)$ asserted in (12) it is enough to prove its lower semicontinuity, i.e.

$$\liminf_{n\to\infty} \hat{\Delta}(T_n | \Xi) \geqslant \hat{\Delta}(T_\infty | \Xi). \tag{14}$$

Take an arbitrary sequence $v_n \in \Sigma^{(n)}$. By the compactness of Ξ we can assume without loss of generality that this sequence converges to some $v_\infty \in \Xi$. Thus, by Lemma 1,

$$\hat{\Delta}(T_\infty \mid \Xi) \leqslant \Delta(T_\infty, v_\infty) \leqslant \liminf_{n\to\infty} \Delta(T_n, v_n) = \liminf_{n\to\infty} \hat{\Delta}(T_n \mid \Xi),$$

which yields (14), and hence (12) as well.

To prove (13) take a sequence $v_n \in \Sigma^{(n)}$ for $n = 1, 2, \ldots$ and let $v_\infty \in \Xi$ be its arbitrary cluster point such that $v_\infty = \lim_{n'\to\infty} v_{n'}$ for some subsequence n'. Taking into account (12), Lemma 1 and the definition of $\Sigma^{(n)}$ we deduce that

$$\Delta(T_\infty, v_\infty) \leqslant \liminf_{n'\to\infty} \Delta(T_{n'}, v_{n'}) = \liminf_{n'\to\infty} \hat{\Delta}(T_{n'} \mid \Xi) = \hat{\Delta}(T_\infty \mid \Xi).$$

This means that $v_\infty \in \Sigma^{(\infty)}$. Relation (13) follows now by standard arguments. ∎

3.3. Proof of Theorem 1. Theorem 1 is an easy conclusion of Lemma 2. By the law of large numbers we have almost surely

$$\lim_{n\to\infty} T_A^{(n)} = T_A.$$

Hence the assertions of Theorem 2 follow immediately from Lemma 2. ∎

3.4. Proof of Theorem 2. For each $i \in \boldsymbol{N}$ define the random function $S_A^{(i)}: 2^{\mathcal{X}} \to \boldsymbol{R}$ by

$$S_A^{(i)}(\mathcal{C}) := \begin{cases} 1 & \text{if } A_i \cap \mathcal{C} \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

It is easily seen that $S_A^{(i)}$ are i.i.d. and

$$T_A^{(n)} = \frac{1}{n} \sum_{i=1}^{n} S_A^{(i)}.$$

Note also that the functional $\mathscr{L}_A$ defined in (4) is the Laplace transform of $S_A^{(1)}$. Therefore, using the standard Cramer's result (see [1], Corollary 3.1.7 and Exercise 3.1.11 below or Theorem 3.3.11, and [2], Theorem 3.5.1) we conclude that the sequence $(T_A^{(n)})_{n=0}^{\infty}$ satisfies on $\mathscr{C}$ the large deviation principle with the rate function $\mathscr{I}_A$ given in (5), i.e. $\mathscr{I}_A$ has compact level sets, for each open set $\mathscr{G} \subset \mathscr{C}$

$$\liminf_{n\to\infty} n^{-1} \log \boldsymbol{P}(T_A^{(n)} \in \mathscr{G}) \geqslant -\inf_{T\in\mathscr{G}} \mathscr{I}_A(T)$$

and for every closed set $\mathscr{F} \subset \mathscr{C}$

$$\limsup_{n\to\infty} n^{-1} \log \boldsymbol{P}(T_A^{(n)} \in \mathscr{F}) \leqslant -\inf_{T\in\mathscr{F}} \mathscr{I}_A(T).$$

Applying Lemma 2 and the standard contraction principle (see [2], Theorem 1.3.2) we obtain the large deviation principle for $(\hat{\Delta}(T_A^{(n)} | \Xi))_{n=0}^{\infty}$ with rate function $\mathscr{R}_A$ defined in (6), which is the first assertion of Theorem 2.

It remains to prove that $\mathscr{R}_A(h) > 0$ for $h \neq \hat{\Delta}(T_A | \Xi)$. Assume that for some $h_0 \neq \hat{\Delta}(T_A | \Xi)$ we have $\mathscr{R}_A(h_0) = 0$. Then there exists a sequence T_n with $\mathscr{I}_A(T_n) \to 0$ and $\hat{\Delta}(T_n | \Xi) = h_0$. In view of the compactness of $\mathscr{C}$ we can assume without loss of generality that T_n converges to some T_∞. By Lemma 2 we have $\hat{\Delta}(T_\infty | \Xi) = h_0$. On the other hand, $\mathscr{I}_A$ is lower semicontinuous (because it has compact level sets) so that $0 \leqslant \mathscr{I}_A(T_\infty) \leqslant \liminf_{n\to\infty} \mathscr{I}_A(T_n) = 0$. Since the only element of $\mathscr{C}$ at which $\mathscr{I}_A$ equals 0 is T_A, we conclude that $T_\infty = T_A$. On the other hand, however, $\hat{\Delta}(T_\infty | \Xi) = h_0 \neq \hat{\Delta}(T_A | \Xi)$, so we come to a contradiction. This completes the proof of Theorem 2. ∎

3.5. Proof of Theorem 3. Applying the standard central limit theorem we conclude that the sequence $\sqrt{n}(T_A^{(n)} - T_A)$ converges in distribution to the mean zero normal law with covariance matrix Π given by (8). Hence, setting

$$Z_n := \sqrt{n}\, W_0 (T_A^{(n)} - T_A),$$

we see that

$$Z_n \to_{\mathscr{D}} \mathscr{N}(0,\ W_0 \Pi W_0^T). \tag{15}$$

Further, from condition (D) we obtain the following Taylor expansion:

$$\begin{aligned} \Delta_\Theta(T, \theta) - \Delta_\Theta(T_A, \theta_0) &= (T - T_A)^T \big([\nabla_T \Delta_\Theta](T_A, \theta_0)\big) \\ &+ (\theta - \theta_0)^T W_0 (T - T_A) + \tfrac{1}{2}(\theta - \theta_0)^T V_0 (\theta - \theta_0) \\ &+ \tfrac{1}{2}(T - T_A)^T \big([\nabla^2_{T,T} \Delta_\Theta](T_A, \theta_0)\big)(T - T_A) + o(|\theta - \theta_0|^2 + |T - T_A|^2), \end{aligned} \tag{16}$$

where the absence of term linear in $(\theta - \theta_0)$ follows from (U), which yields $[\nabla_\theta \Delta_\Theta](T_A, \theta_0) = 0$. Therefore, writing

$$M(\theta) := \Delta_\Theta(T_A, \theta) \quad \text{and} \quad M_n(\theta) := \Delta_\Theta(T_A^{(n)}, \theta),$$

we conclude from (16) that

$$(M_n - M)(\theta) - (M_n - M)(\theta_0) = (\theta - \theta_0)^T W_0 (T_A^{(n)} - T_A) + o(|\theta - \theta_0|^2 + |T_A^{(n)} - T_A|^2).$$

Taking into account (15) we get

$$\begin{aligned} &\sqrt{n}(M_n - M)(\theta) - \sqrt{n}(M_n - M)(\theta_0) \\ &\qquad = (\theta - \theta_0)^T Z_n + o_P\left(\sqrt{n}\,|\theta - \theta_0|^2 + \frac{1}{\sqrt{n}}\right). \end{aligned} \tag{17}$$

In view of (15) and (17) we can apply Theorem 3.2.16 in [11] with $r_n = \sqrt{n}$ to obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_0^{-1} Z_n + o_P(1).$$

This yields immediately the assertion of Theorem 3. The proof is complete. ∎

4. CONCLUDING REMARKS

The further objectives aimed at by the author are twofold. The first one is to elaborate appropriate algorithms and to implement the estimation procedures suggested in the paper in relevant particular cases. It is worth noting that although related algorithms have been already investigated in the literature (see [4] and [6]), they performed the relative entropy minimisation over the core of a certain capacity with respect to a fixed (usually uniform) measure, while in our case an additional complication arises due to the fact that the reference measure is also subject to the optimisation procedure.

The second objective is to extend the established results to the case of continuous sample space.

REFERENCES

[1] J.-D. Deuschel and D. W. Stroock, *Large Deviations*, Academic Press, Boston 1989.

[2] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, Chichester 1997.

[3] I. A. Ibrahimov and R. Z. Khasminskii, *Asymptotic Estimation Theory* (in Russian), Nauka, Moscow 1979.

[4] J.-Y. Jaffray, *On the maximum of conditional entropy for upper/lower probabilities generated by random sets*, in: *Random Sets: Theory and Applications*, J. Goutsias, R. P. S. Mahler and H. T. Nguyen (Eds.), *The IMA Volumes in Mathematics and its Applications*, Springer, New York 1997, pp. 105–127.

[5] G. Matheron, *Random Sets and Integral Geometry*, Wiley, New York 1975.

[6] A. Meyerowitz, F. Richman and E. Walker, *Calculating maximum-entropy probability densities for belief functions*, Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 2, No. 4 (1994), pp. 377–389.

[7] I. S. Molchanov, *Limit Theorems for Unions of Random Closed Sets*, Lecture Notes in Math. 1561, Springer, Berlin 1993.

[8] T. Norberg, *An ordered random set coupling*, Probab. Theory Related Fields 37 (1992), pp. 161–163.

[9] T. Norberg, *On the existence of ordered couplings of random sets — with applications*, Israel J. Math. 77 (1992), pp. 241–264.

[10] R. Schneider, *Convex bodies: the Brunn–Minkowski theory*, Encyclopaedia Math. Appl. 44, Cambridge University Press, 1993.

[11] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer, New York 1996.

Faculty of Mathematics and Computer Science
Nicholas Copernicus University
Toruń, Poland
E-mail: tomeks at mat.uni.torun.pl

Received on 26.4.2000