ELSEVIER

Random and fuzzy sets in coarse data analysis

Hung T. Nguyen[a,*], Berlin Wu[b]

[a]*Department of Mathematical Sciences, New Mexico State University, Las Cruces, NM 88003-8001, USA*
[b]*Department of Mathematical Sciences, National ChengChi University, Taipei, Taiwan, ROC*

Abstract

The theoretical aspects of statistical inference with imprecise data, with focus on random sets, are considered. On the setting of coarse data analysis imprecision and randomness in observed data are exhibited, and the relationship between probability and other types of uncertainty, such as belief functions and possibility measures, is analyzed. Coarsening schemes are viewed as models for perception-based information gathering processes in which random fuzzy sets appear naturally. As an implication, fuzzy statistics is statistics with fuzzy data. That is, fuzzy sets are a new type of data and as such, complementary to statistical analysis in the sense that they enlarge the domain of applications of statistical science.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Coarse data analysis; Fuzzy sets; Random fuzzy sets; Random sets

1. Introduction

The theory of fuzzy sets has reached 40 years. While its applications in technology seem widely endorsed, discussions within the probability and statistics communities remain to some extent. This is exemplified by the recent paper (with discussions) by Singpurwalla and Booker (2004) in which probability theory is viewed as a calculus for uncertainty of outcomes whereas fuzzy set theory as a calculus for imprecision. Note that recently there are some emphases on the need of using fuzzy sets in *classical statistics* (e.g. Geyer and Meeden, 2005), namely in providing new exact confidence intervals for parameters in the case of discrete variables.

In gathering information for decision-making, we often face both randomness and imprecision. The typical situation in statistics is *coarse data analysis* (e.g., Little and Rubin, 1987; Heitjan and Rubin, 1991; Gill et al., 1997) in which the random variable of interest X is not directly observable, but instead, we observe a random set S containing X with probability one.

The coexistence of randomness and fuzziness appear when probability values cannot be evaluated precisely such as in situations where we can only assess uncertainty linguistically, e.g. as low, medium or high probability. The concern about meaning and interpretation of such probabilistic expressions is exemplified by works such as Wallsten (1986) and Mosteller and Youtz (1990).

The theory of fuzzy sets is invented precisely to provide mathematical tools for modeling fuzzy concepts in our natural language. While fuzziness is described as a new type of uncertainty, namely the vagueness in the linguistic data, its associated uncertainty known as *possibility* (Zadeh, 1978) seems unconventional within the statistics community.

---

* Corresponding author. Tel.: +1 505 646 2105; fax: +1 505 646 1064.
  E-mail address: hunguyen@nmsu.edu (H.T. Nguyen).

In parallel, another type of non-additive set functions used to describe uncertainty, known as *belief functions* (Dempster, 1967; Shafer, 1976), was not so welcome into the statistics community (see the objection of Lindley, 1982). On the other hand, the foundations of fuzzy set theory were criticized by Elkan (1993). Fortunately, these two main attacks to non-additive set functions (as uncertainty measures) and to the notion of fuzziness were countered on *theoretical* grounds e.g. in Goodman et al. (1990), Nguyen et al. (1996) and Gehrke et al. (1997). Thus, the situation is settled! Now fuzzy sets are a new type of data and as such they should get attention of statisticians, just like *random sets*, since, as G. Watson put it "Modern statistics must be defined as the applications of computers and mathematics to data analysis. It must grow as new types of data are considered and as computing technology advances" (Foreword to G. Matheron's book, 1975).

Now to view fuzzy sets as statistical data (observations), one needs to place *random fuzzy sets* as bona fide *random elements*. The mathematical theory of random fuzzy sets has been the subject of research since the 80's by many researchers, see e.g. the book by Li et al. (2002). As Fréchet (1948) pointed out, nature and technology present various forms of random patterns, and hence probability theory should consider random elements as mathematical objects rigorously defined. We do have such general definition of random elements as measurable mappings on abstract measurable spaces. As random sets are special cases of random fuzzy sets, we start with the rigorous theory of random sets on locally compact, Hausdorff and second countable spaces, such as $\mathbb{R}^d$ (Matheron, 1975), and consider fuzzy sets on $\mathbb{R}^d$ which form a metric space (e.g. Diamond and Kloeden, 1994), so that we are entirely within the standard framework of probability theory on metric spaces.

In the following, we are going to elaborate on some of the above aspects using random sets as a special case of random fuzzy sets in the context of coarse data analysis in order to assess the state-of-the-art of fuzziness in probability and statistics.

2. Random sets and related uncertainty measures

2.1. Probability and belief functions

Using the framework of coarse data in statistics, we are going to show that non-additive uncertainties such as belief functions and possibility measures are derived formally from probability in a pessimistic and optimistic points of view, respectively. The setting of *coarse data analysis* is this. Let X be a random element of interest, defined on a probability space $(\Omega, \mathscr{A}, P)$, and taking values in a set U (equipped with some σ-field $\mathscr{U}$). When performing the experiment or observing X, we cannot observe the values of X, but instead, we can locate each value of X in some set S. The most general model for this type of observations is to view S as a *random set*, defined also on $(\Omega, \mathscr{A}, P)$, with values in the power set 2^U of U such that $P(X \in S) = 1$, i.e. X is an almost sure (a.s.) selector of S, or the other way around, S is a *coarsening* of X.

Let $A \in \mathscr{U}$, since X is not observable, we cannot assess whether A occurs or not. However, we can put bounds on $P(X \in A)$. Clearly, if our observable S is contained in A, then A occurs, so that our lower bound on the "likelihood" of occurrence of A could be $P(S \subseteq A)$. Clearly, since X is an a.s. selector of S, we have that $P(S \subseteq A) \leqslant P(X \in A)$. Let $F(A) = P(S \subseteq A)$. The set function F (here defined on the σ-field $\mathscr{U}$) could be taken as a quantitative measure of belief (in the occurrences of events, from a pessimistic view point). The axiomatization of this set function is known as the Demspter–Shafer theory of *belief functions* (or theory of evidence), Dempster (1967) and Shafer (1976). This non-additive set function thus just measures a weak form of uncertainty in the random occurrences of event when knowledge is only partial. The axiomatic theory is aimed at providing a setting in which no randomness is around, such as subjective evaluations of things. In the case where U is a *finite* set, the belief function F, as defined above in terms of *random sets*, satisfies the following:

(i) $F : 2^U \to [0, 1]$, $F(\emptyset) = 0$, $F(U) = 1$;
(ii) F is monotone of infinite order, i.e. for any $n \geqslant 2$, and $A_1, A_2, \ldots, A_n$, subsets of U.

$$F\left(\bigcup_{i=1}^n A_i\right) \geqslant \sum_{\emptyset \neq I \subset \{1,2,\ldots,n\}} (-1)^{|I|+1} F\left(\bigcap_{i \in I} A_i\right),$$

where $|I|$ denotes the cardinality of the set I.

These properties are characteristic for determining distributions of non-empty random sets on finite spaces. But belief functions are in general subjective, i.e. assigned by "experts" reflecting their own views on uncertainty. As such, the above axioms should be considered for arbitrary spaces, finite or not. Note that the lower probability $F(.)$ induced by a random set is always monotone of infinite order on arbitrary spaces. When U is uncountable, e.g. $U = \mathbb{R}^d$, non-atomic probability measures should be defined on proper subclass of subsets of U since 2^U is too big. This is mainly due to the σ-additivity requirement for probability measures. Now, belief functions are free of this constraint, being even not finitely additive! As such, they could be defined on 2^U, just like the finite case. After all, for any subset of U, measurable or not, subjective assignments of beliefs to it by experts are possible! We recall here a couple of interesting constructions of belief functions on *arbitrary* spaces.

Let U be an *arbitrary* set and $f : U \to [0, 1]$ such that $\sup_{x \in U} f(x) = 1$. Let $T : 2^U \to [0, 1]$ be

$$T(A) = \sup\{f(x) : x \in A\}.$$

Then $F(A) = 1 - T(A^c)$ is a belief function. This is so since T is *maxitive*, i.e., $T(A \cup B) = \max(T(A), T(B))$. As such, T is alternating of infinite order (see Lemma 1 below for ease of reference) and hence F is monotone of infinite order. Note that this result is useful for modeling distributions of random sets via their capacity functionals.

Lemma 2.1. *Let C be a class of subsets of some set Θ, containing $\emptyset$ and stable under finite intersections and unions. Let $T : C \to [0, +\infty)$ be maxitive, i.e., for all $A, B \in \mathscr{C}$*

$$T(A \cup B) = \max\{T(A), T(B)\}.$$

Then T is alternating of infinite order.

Proof. Clearly T is monotone increasing on $\mathscr{C}$. Note that for any $A_1, \ldots, A_n$ in $\mathscr{C}, n \geqslant 2$,

$$T\left(\bigcup_{i=1}^{n} A_i\right) = \max\{T(A_i), \ i = 1, 2, \ldots, n\}.$$

We need to show that

$$T\left(\bigcap_{i=1}^{n} A_i\right) \leqslant \sum_{\emptyset \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} T\left(\bigcup_{i \in I} A_i\right). \tag{1}$$

Without loss of generalities, we may assume that

$$0 \leqslant \alpha_n = T(A_n) \leqslant \alpha_{n-1} = T(A_{n-1}) \leqslant \cdots \leqslant \alpha_1 = T(A_1).$$

For $k \in \{1, 2, \ldots, n\}$, let $J(k) = \{I \subseteq \{1, 2, \ldots, n\} : |I| = k\}$. For $I \in J(k)$, let $m(I) = \min\{i : i \in I\}$, and for $i = 1, 2, \ldots, n - k + 1$, let

$$J_i(k) = \{I \in J(k) : m(I) = i\}.$$

Then we have

$$T\left(\bigcup_{j \in I} A_j\right) = \alpha_i$$

for every $I \in J_i(k), \ i = 1, 2, \ldots, n - k + 1$. Also,

$$|J_i(k)| = \binom{n-i}{k-1} = \frac{(n-i)!}{(k-1)!(n-i-k+1)!}$$

for every $i = 1, 2, \ldots, n - k + 1$. Thus

$$\sum_{I \in J(k)} T\left(\bigcup_{i \in I} A_i\right) = \sum_{i=1}^{n-k+1} \sum_{I \in J_i(k)} T\left(\bigcup_{j \in I} A_j\right) = \sum_{i=1}^{n-k+1} \sum_{I \in J_i(k)} \alpha_i = \sum_{i=1}^{n-k+1} \binom{n-i}{k-1} \alpha_i$$

and therefore

$$
\begin{aligned}
\sum_{\emptyset \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} T\left(\bigcup_{i \in I} A_i\right) &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{I \in J(k)} T\left(\bigcup_{i \in I} A_i\right) \\
&= \sum_{k=1}^{n} (-1)^{k+1} \sum_{i=1}^{n-k+1} \binom{n-i}{k-1} \alpha_i \\
&= \sum_{i=1}^{n} \left(\sum_{k=0}^{n-i} \binom{n-i}{k} (-1)^k\right) \alpha_i \\
&= \alpha_n = T(A_n) \geqslant T\left(\bigcap_{i=1}^{n} A_i\right)
\end{aligned}
$$

by observing that for any $i = 1, 2, \ldots, n-1$,

$$
\sum_{k=0}^{n-i} \binom{n-i}{k} (-1)^k = (1-1)^{n-i} = 0. \qquad \square
$$

Belief functions can be induced by probability measures as follows: (i) Let $(U, \mathscr{U})$ be an arbitrary measurable space, μ be a probability measure on $\mathscr{U}$ and $F : 2^U \to [0, 1]$ be the *inner measure* of μ, i.e. for $B \subseteq U$,

$$
F(B) = \sup\{\mu(A) : A \in \mathscr{U}, A \subseteq B\}.
$$

Then, F is a belief function on U. Clearly, $F(\emptyset) = 0$ and $F(U) = 1$. For completeness, we reproduce the results of Fagin and Halpern (1991), see also Halpern (2003) and Walley (1991).

Lemma 2.2. *For each $B \subseteq U$, $\sup\{\mu(A) : A \in U, A \subseteq B\}$ is attained.*

Proof. For each $n \geqslant 1$, let $A_n \in \mathscr{U}$, $A_n \subseteq B$ such that

$$
\mu(A_n) \geqslant F(B) - 1/n,
$$

where $F(B) = \sup\{\mu(A) : A \in \mathscr{U}, A \subseteq B\}$. Let $A = \bigcup_{n \geqslant 1} A_n$. Then $A \in \mathscr{U}$ and $A \subseteq B$. Thus,

$$
F(B) \geqslant \mu(A) \geqslant \mu(A_n) \geqslant F(B) - 1/n.
$$

Since n is arbitrary, it follows that $F(B) = \mu(A)$. $\square$

Lemma 2.3. *For any $n \geqslant 1$, let $B_1, B_2, \ldots, B_n$ in 2^U and $A_1, A_2, \ldots, A_n$ in U with $A_i \subseteq B_i$, $F(B_i) = \mu(A_i)$, $i = 1, 2, \ldots, n$. Then $F\left(\bigcap_{i=1}^{n} B_i\right) = \mu\left(\bigcap_{i=1}^{n} A_i\right)$.*

Proof. By Lemma 2, there is a $C \in \mathscr{U}$, $C \subseteq \bigcap_{i=1}^{n} B_i$ such that $F\left(\bigcap_{i=1}^{n} B_i\right) = \mu(C)$. Thus, it suffices to show that $\mu(C) = \mu\left(\bigcap_{i=1}^{n} A_i\right)$. Without loss of generality, we can assume that $\bigcap_{i=1}^{n} A_i \subseteq C$ (otherwise, replace C by $C \cup \left(\bigcap_{i=1}^{n} A_i\right)$ by noting that $C \cup \left(\bigcap_{i=1}^{n} A_i\right) \subseteq \bigcap_{i=1}^{n} B_i$ and

$$
F\left(\bigcap_{i=1}^{n} B_i\right) = \mu(C) \leqslant \mu\left(C \cup \left(\bigcap_{i=1}^{n} A_i\right)\right) \leqslant F\left(\bigcap_{i=1}^{n} B_i\right)
$$

so that $\mu\left(C \cup \left(\bigcap_{i=1}^{n} A_i\right)\right) = F\left(\bigcap_{i=1}^{n} B_i\right)$.

Let $D = C \backslash \bigcap_{i=1}^{n} A_i = \bigcup_{i=1}^{n} \left(C \cap A_i^c\right)$. Now, for each $i \in \{1, 2, \ldots, n\}$, $D \subseteq C \subseteq B_i$ with $D \cup A_i \subseteq B_i$ implying that $\mu(D \cup A_i) = F(B_i) = \mu(A_i)$, so that $\mu(D \cap A_i^c) = 0$.

Since $D \subseteq \bigcup_{i=1}^{n} A_i^c$, we have

$$
\mu(D) = \mu\left(D \cap \left(\bigcup_{i=1}^{n} A_i^c\right)\right) = \mu\left(\bigcup_{i=1}^{n} \left(D \cap A_i^c\right)\right) \leqslant \sum_{i=1}^{n} \mu\left(D \cap A_i^c\right) = 0.
$$

Thus, $\mu\left(\bigcap_{i=1}^{n} A_i\right) = \mu(C)$. $\square$

Corollary 2.1. *F is infinitely monotone.*

Proof. For $n \geqslant 2$, let $B_1, B_2, \ldots, B_n$ in 2^U. By Lemma 2, let $A_1, A_2, \ldots, A_n$ in $\mathscr{U}$ such that $A_i \subseteq B_i$ and $\mu(A_i) = F(B_i)$, $i = 1, 2, \ldots, n$. We have, since $\bigcup_{i=1}^n A_i \in \mathscr{U}$ and $\bigcup_{i=1}^n A_i \subseteq \bigcup_{i=1}^n B_i$,

$$F\left(\bigcup_{i=1}^n B_i\right) \geqslant \mu\left(\bigcup_{i=1}^n A_i\right) = \sum_{\emptyset \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} \mu\left(\bigcap_{i\in I} A_i\right) \sum_{\emptyset \neq I \subseteq \{1,2,\ldots,n\}} (-1)^{|I|+1} F\left(\bigcap_{i\in I} B_i\right)$$

by Lemma 3. $\quad \square$

(ii) Let U be a *finite* set and P be a probability measure defined on 2^U. It can be checked that the Mobius inverse of the set function $F(A) = P^2(A)$ is a probability density function on 2^U so that F is a belief function. More generally, for any integer $k \geqslant 1$, the set function $P^k : 2^U \to [0, 1]$, defined as $P^k(A) = [P(A)]^k$, is a belief function. Indeed, it suffices to observe that P^k is of the form

$$Q^k(A) = \sum_{B \subseteq A} f(B),$$

where $f : 2^U \to [0, 1]$ is a probability density (on 2^U). Using the multinomial theorem

$$(x_1 + x_2 + \cdots + x_n)^k = \sum \frac{k!}{k_1! k_2! \ldots k_n!} \cdot x_1^{k_1} x_2^{k_2} \ldots x_n^{k_n},$$

where the summation extends over all non-negative integral solutions $k_1, k_2, \ldots, k_n$ of $k_1 + k_2 + \cdots + k_n = k$, it is easy to verify that the following f will do:

$$f(A) = \begin{cases} 0 & \text{if } A = \emptyset \text{ or } |A| > k, \\ \displaystyle\sum_{n_a \neq 0 \mid \sum n_a = k} \prod_{a \in A} \frac{k!}{\prod n_a!} [P(a)]^{n_a} \end{cases}$$

by noting that $P^k(A) = [\sum_{a \in A} P(a)]^k$.

2.2. Probability and possibility measures

Now in our framework of coarse data analysis, for $A \in \mathscr{A}$, if $S \cap A \neq \emptyset$, then it is *possible* that A occurs. A quantitative measure for possibility could be taken to be $T(A) = P(S \cap A \neq \emptyset)$. In particular, for singleton sets, $T(u) = P(u \in S)$ is the *covering function* of S, just like in *Survey Sampling* (see e.g. Nguyen, 2004) which can be interpreted as the *possibility* for u to happen. This is what Zadeh (1978) called a *possibility distribution,* see also Dubois and Prade (1988). The axioms for *possibility measures*, as proposed by Zadeh, are these. A possibility measure on a set U is a map $T : 2^U \to [0, 1]$ such that $T(\emptyset) = 0$, $T(U) = 1$ and for any family of subsets $(A_i, i \in I)$ of U, $T\left(\bigcup_{i \in I}\right) = \sup_{i \in I} T(A_i)$. As such, for any A, $T(A) = \sup_{u \in A} T(u)$. In coarse data analysis or survey sampling, given a covering function π on a *finite* population U, i.e. $\pi : U \to [0, 1]$, among all probability sampling plans (finite random sets on U) having π as their common inclusion probability function, there exists precisely one random set S (a coarsening scheme) having π as its covering function and such that

$$P(S \cap A \neq \emptyset) = \sup_{u \in A} P(u \in S).$$

Indeed, let $\alpha : (\Omega, \mathscr{A}, P) \to [0, 1]$ be a random variable, uniformly distributed. Consider the randomized level-set map:

$$S : \Omega \to 2^U, \quad S(\omega) = \{u \in U : \pi(u) \geqslant \alpha(\omega)\}.$$

Then, S is a random set on U with $P(u \in S) = P(\omega : \alpha(\omega) \leqslant \pi(u)) = \pi(u)$, and

$$P(S \cap A \neq \emptyset) = P\left(\omega : \alpha(\omega) \leqslant \max_{u \in A} \pi(u)\right) = \max_{u \in U} \pi(u).$$

Thus, possibility measures are closely related to probability. In fact, as opposed to belief functions, possibility is an optimistic view on uncertainty. In the context of coarse data analysis, since X is an a.s. selector of S, we have $P(X \in A) \leqslant P(S \cap A \neq \emptyset)$, i.e. possibility is always larger than probability. See also Goodman (1982), Dubois and Prade (1987) and Miranda et al. (2004).

3. Fuzzy sets in statistics

3.1. Membership functions

The so-called intelligent technologies aim at duplicating remarkable human strategies in decision-making or control. If we look at look-up tables in the field of *fuzzy control*, or at reports on, say, risk assessments of scenarios, we see rules containing linguistic labels, e.g. rules of the form "if the obstacle is 'near' and the wind is 'strong' then 'slow' down", assessments of probabilities of occurrences of events as 'high', 'medium', 'low', and predicting rates of interest as 'low', 'moderate', 'high'.

These are *perception-based information* (Zadeh, 2002). These linguistic terms clearly form a fuzzy partition of some measurement space of interest. For example, 'high', 'medium' and 'low' probabilities form a fuzzy partition of [0, 1]. They are in fact precisely coarsening schemes of a more general form, namely forming *fuzzy partitions* of the measurement spaces. Being unable with naked eyes to measure with accuracy, humans use coarsenings instead, in order to extract useful information. As Zadeh put it, humans use *linguistic variables*.

Our thesis is that humans use coarsenings as their perception-based information gathering processes. Thus, fuzzy sets appear in so many important applications in science and technology. Putting models on observed data is a practice in statistics. The statistical models on fuzzy data are *random fuzzy sets* extending random sets (for recent texts in Random sets, see Molchanov, 2005; Nguyen, 2006). While fuzzy set theory (and its associated logics) is an appropriate analytic tool for modeling fuzzy concepts in human perception-based data, the debate in the statistical community seems to center on the practical question: How to obtain *membership functions* for fuzzy sets? Of course, a more fundamental question is: What is the difference between degrees of membership and probabilities?

At this present stage, this fundamental question has been, to some extent, answered satisfactorily. Degrees of membership have a clear semantic as so stated! They provide a quantitative concept of fuzziness for applications. As such, it is clear why degrees of memberships should not be confused with probabilities! The practical question is somewhat similar to statistics. Membership functions are obtained from available information. Thus, they can be given either by experts, as in Bayesian statistics, or by empirical investigations.

The connection between random sets and fuzzy sets (e.g., Goodman and Nguyen, 2000) has been used as a way to obtain membership functions of fuzzy concepts: if π is a membership function on a space U, then π is written as $\pi(u) = P(u \in S)$ for some random set on U. A more general *formula* is proposed by Orlowski (1994) by replacing the probability measure P by a *fuzzy measure* μ (e.g. Nguyen and Walker, 2000) and S by a multi-valued mapping. Orlowski's formula is intended to provide a way to mimic experts' assignments of degrees of membership. To obtain μ we need data, just like in machine learning, either under supervised or unsupervised learning, i.e. carrying out the problem of *identification of fuzzy measures*. This is also essential in multi-criteria decision-making (e.g. Grabisch et al., 1994), especially for ranking alternatives based upon interacting criteria where *Choquet integrals* with respect to fuzzy measures seem to be appropriate aggregation operators generalizing linear operators such as expected utilities.

As mentioned in the introduction, recently, statisticians started asking whether fuzzy set theory is useful for classical statistics such as confidence interval (or set) estimation of population parameters. In this context, membership functions arise naturally from testing procedures. The crucial point is this. In view of the duality between tests and confidence intervals, the randomized test for a *discrete* variable should correspond to a *fuzzy confidence interval* whose membership function is precisely the test function. For details, see Geyer and Meeden (2005).

3.2. Random sets in probability density estimation

As pointed out earlier, the theory of statistics with fuzzy data should be based upon a rigorous theory of random fuzzy sets where the space of fuzzy sets forms a metric space. Random fuzzy sets are random elements with values in a such metric space. In this direction, it is obvious that the use of α-level sets of membership functions is essential.

This is exemplied by the analogy with a basic estimation problem in classical statistics, namely the non-parametric estimation of a probability density function.

Let X be a random vector with values in $\mathbb{R}^d$, with unknown density f. We desire to estimate f from a random sample $X_1, X_2, \ldots, X_n$ drawn from X. The α-level sets of f is, for $\alpha > 0$,

$$A_\alpha(f) = A_\alpha = \{x \in \mathbb{R}^d : f(x) \geqslant \alpha\}.$$

We outline here the problem of density estimation via the estimation of level sets. This is an alternative approach to kernel method and orthogonal functions in non-parametric density estimation when qualitative information about the density (such as its shape, geometric properties of its contour clusters) is available rather than analytic information.

Let $f : \mathbb{R}^d \to \mathbb{R}^+$ be the unknown probability density of a random vector X of interest. Then f can be written in terms of its α-level sets as

$$f(x) = \int_0^{+\infty} A_\alpha(x)\, \mathrm{d}\alpha$$

for all $x \in \mathbb{R}^d$, where $A_\alpha(x) = I_{A_\alpha}(x)$ denotes the *indicator function* of the set A_α. Thus, if each A_α is estimated by a random set $A_{\alpha,n}$ (measurable with respect to the sample $X_1, \ldots, X_n$), then it is natural to consider the plug-in estimator

$$f_n(x) = \int_0^{+\infty} A_{\alpha,n}(x)\, \mathrm{d}\alpha$$

for $f(x)$. Note that here $A_{\alpha,n}$ is a random set of more general nature than finite random sets.

The point is this. In order to estimate the density f, we are led to set estimation and random sets, as set estimators, arise naturally.

It is interesting to note that the *excess mass approach* (Hartigan, 1987, see also Polonik, 1995) to level-set estimation bears some resemblance with maximum likelihood principle in statistics. Indeed, for a fixed level α, the target parameter is the set A_α. The qualitative information about f leads to the statistical model: $A_\alpha \in \mathscr{C}$, where $\mathscr{C}$ is a specified class of subsets of $\mathbb{R}^d$, e.g. closed convex subsets, ellipsoids.

We are in the standard framework of statistical estimation theory: the parameter space is $\mathscr{C}$, so that estimators of A_α should be random sets with values in $\mathscr{C}$. Due to the nature of the target parameters A_α, it is possible to find a general principle to suggest estimators for it.

Let $\mathrm{d}F$ denote the probability law of X on $\mathbb{R}^d$ (i.e. the Stieltjes measure associated with f), and again, μ denotes the Lebesgue measure on $\mathbb{R}^d$.

Clearly, $(\mathrm{d}F - \alpha\mu)(A_\alpha)$ is the "excess mass" of the set A_α at level α. Thus, we can consider the signed measure $\mathrm{d}F - \alpha\mu = \mathscr{E}_\alpha$ on $\mathscr{B}\left(\mathbb{R}^d\right)$, with $\mathscr{E}_\alpha(A)$ as the excess mass of A at level α. Writing $A = AA_\alpha \cup AA_\alpha^c$, we see that

$$\mathscr{E}_\alpha(A) \leqslant \mathscr{E}_\alpha(A_\alpha)$$

for all $A \in \mathscr{B}\left(\mathbb{R}^d\right)$, i.e., the level set A_α has largest excess mass, at level α, among all Borel sets. This suggests a way to estimate A_α using the empirical counter-part of the signed measure $\mathrm{d}F - \alpha\mu$.

Let $\mathrm{d}F_n$ denote the empirical measure associated with the sample $X_1, X_2, \ldots, X_n$, i.e.,

$$\mathrm{d}F_n = \frac{1}{n} \sum_{j=1}^n \delta_{X_j}$$

with δ_x being the Dirac measure at $x \in \mathbb{R}^d$. Then the empirical excess mass, at level α, of $A \in \mathscr{B}\left(\mathbb{R}^d\right)$ is

$$\mathscr{E}_{\alpha,n}(A) = (\mathrm{d}F_n - \alpha\mu)(A).$$

Thus, it is natural to hope that "good" estimators $A_{n,\alpha}$ of A_α can be obtained by maximizing $\mathscr{E}_{\alpha,n}(A)$ over $A \in \mathscr{C}$. Note that this optimization problem needs special attention: the variable is neither a vector, nor a function, but a set. This type of optimization of set-functions occurs in many areas of applied mathematics such as shape optimization. For a variational calculus of set-functions which seems appropriate for optimization of set-functions, see e.g. Nguyen and Kreinovich (1999).

As a routine in statistical practice, desirable properties of the above random set estimator $A_{\alpha,n}$ need to be assessed. In particular, for large sample statistics, consistency and limiting distribution problems need to be examined. This can be carried out by considering statistical convergence in distribution of sequences of random sets (see e.g. Molchanov, 2005; Nguyen, 2006).

The literature on fuzzy statistics contains of course all aspects of statistical inference in which fuzziness is involved under one form or another. For example, in hypothesis testing, the hypotheses could be imprecise, due to the nature of the problem, e.g. expressed in linguistic forms. Here is a sample of works on fuzzy statistics: Bandemer and Näther (1992), Buckley (2004), Gil and Lopez-Diaz (1998), Kruse and Meyer (1987), Kruse et al. (1999), Nguyen and Wu (2000, 2006) and Montenegro et al. (2001, 2004).

4. Some statistics with imprecise data

4.1. Coarse data analysis

We choose to discuss here some statistical aspects of statistics with set-valued observations (random sets) because, in our view, they should form the basis for statistics with fuzzy set-valued observations. The analysis is useful for coarse data analysis in standard statistics as well. A framework for set-valued observations in the *finite case* was laid out by Schreiber (2000) whose extension to the *compact*, *metric spaces* was recently established in Feng and Feng (2004).

Coarse data are a typical situation where the observations are sets rather than points in a sample space. By coarse data we mean rough data or data with low quality. This happens, for example, when the available data are imprecise, say, due to imperfection of the data acquiring procedure (e.g., inaccuracy of the measuring instruments). In such cases, rather than trying to ascribe unique values to the observations, it might be more preferable to represent the outcomes of the random experiment as subsets containing the "true" observation values. Familiar examples of coarse data are missing data, censored or grouped data in biostatistics. A general framework for set-valued observations was proposed using random sets. Specifically, let X be the random variable of interest. The observation process is modeled by a random set S on the range of X. Each unobservable outcome X_j, $j = 1, 2, \ldots, n$, is in the observed random set outcomes S_j, $j = 1, 2, \ldots, n$, which is an i.i.d. sample from S. Thus, X is an *almost sure selector* of S. The statistical inference problem about X, say, estimating the probability density functions of X, will be based upon the *random set data* S_j, $j = 1, 2, \ldots, n$. The point is this. In order to study the above estimation problem, we need a rigorous theory of random sets, especially their distributions. See e.g. Wagner (1977) for a survey on measurable selections.

The practical situation in coarse data analysis is this. Being unable to observe with accuracy the values of a random sample $X_1, X_2, \ldots, X_n$ from X, the statistician tries to locate these observations in random sets S_j, $j = 1, 2, \ldots, n$. There are of course various ways for doing so. Each way represents a coarsening of the data. The observed random sets S_j are viewed as a random sample from a coarsening S which is a random set. A random set S is called a coarsening of X if S contains X almost surely (i.e. with probability one), i.e. X is an almost sure selector of S. Note that here X is given first, and S is a *random set model* for X. Thus, selector or coarsening depends on which is given first!

A useful model for coarsening is the CAR model (Heitjan and Rubin, 1991; Gill et al., 1997), where CAR stands for *coarsening at random*. Here is an example. Let $U \subseteq \mathbb{R}$ be the range of X, and $\{A_1, A_2, \ldots, A_k\}$ be a (measurable) partition of U. Consider the coarsening scheme

$$S : (\Omega, \mathscr{A}, P) \rightarrow \{A_1, A_2, \ldots, A_k\}.$$

Suppose the unknown probability density function of X is of a parametric form, i.e., $f(x|\theta), \theta \in \Theta$. Then the likelihood function based on the random set sample S_j, $j = 1, 2, \ldots, n$, is

$$L(\theta | S_1, S_2, \ldots, S_n) = \prod_{j=1}^{n} \int_{S_j} f(x|\theta) \, \mathrm{d}x.$$

Thus, the maximum likelihood estimator of θ can be computed using only the observed $S_1, S_2, \ldots, S_n$. However, the investigation of large sample properties of the estimator requires also distributional aspects of the random set model S.

4.2. A statistical framework

Coarse data analysis can be viewed as a special procedure in perception-based information gathering process by humans. This type of data is used in the field of *artificial intelligence* to imitate remarkable human intelligent behavior, say, in decision making and control. The search for the source of how humans gather the information in their environments is of course useful in many applications, such as robotics. For example, being unable with naked eyes to measure accurately the distance to some referential location, humans try to extract useful information by considering some simple schemes over the range of the measurements, such as a partition of it. In other words, when we cannot measure exactly the values of some variable of interest X, we coarsen it, e.g. use some random set S such that $P(X \in S) = 1$ to extract information about X.

Consider the standard problem of statistics, namely the estimation of the unknown probability law π_0 of a random variable X with values in U. But unlike the conventional situation, the data is an i.i.d. sample of random sets $S_1, S_2, \ldots, S_n$, drawn from a random S, rather than an i.i.d. sample from X. The relation between X and S is that $X \in S$ almost surely, i.e. X is an a.s. measurable selection (or *selector*) of S. This clearly extends the structure of standard data analysis in statistics. By the strong law of large numbers, the distribution function F of S is estimated consistently by the empirical distribution function

$$F_n(A) = \frac{1}{n} \, |\{S_i : S_i \subseteq A\}|$$

we are led to describe the statistical model in terms of F. Let $\mathbb{P}$ denote the set of all probability measures on U (finite here). The parameter space is some subset $\mathscr{P} \subseteq \mathbb{P}$ containing π_0, where before having additional information to narrow down $\mathscr{P}$, $\mathscr{P}$ is simply the set of probability laws of all possible selectors of copies of S. We would like to specify further the set $\mathscr{P}$. Now observe that since X is a selector of S (the observed sets $S_1, S_2, \ldots, S_n$ contain unobserved values $X_1, X_2, \ldots, X_n$), we have that $F \leqslant \pi_0$. Indeed, let $(\Omega, \mathscr{A}, P)$ be a probability space on which are defined both S and X. Let $D \in \mathscr{A}$ such that $P(D') = 0$ and $X(\omega) \in S(\omega)$ for all $\omega \in D$. Then,

$$F(A) = P(\omega : S(\omega) \subseteq A) = P((\omega : S(\omega) \subseteq A) \cap D) \leqslant P(\omega : X(\omega) \in A) = \pi_0(A).$$

Thus $\pi_0 \in \mathscr{C}(F) = \{\pi \in \mathbb{P} : F \leqslant \pi\}$. Borrowing a name from game theory, $\mathscr{C}(F)$ is called the *core* of F or of S. Note that the core of F is the same as the core of the *capacity functional* $T(A) = 1 - F(A^c)$ of the random set S, where A^c denotes the set-complement of A, i.e. $\mathscr{C}(F) = \mathscr{C}(T) = \{\pi \in \mathbb{P} : \pi \leqslant T\}$. The structure of the core $\mathscr{C}(F)$ is very nice, and the question is whether $\mathscr{P} = \mathscr{C}(F)$? This amounts to check the converse of the above fact, namely, given F on 2^U (or equivalently its associated probability measure dF on the power set of 2^U), and $\pi \in \mathbb{P}$ with $F \leqslant \pi$, can we find a probability space $(\Omega, \mathscr{A}, P)$ and $S : \Omega \to 2^U$, $X : \Omega \to U$ such that $P(X \in S) = 1$ and $PS^{-1} = dF$, $PX^{-1} = \pi$? i.e. is π *selectionable* with respect to F?

The above converse problem is the problem of existence of a measurable selection with given image measure. Let us say a few words about the selection problem in this specific setting.

Let S be a set-valued function, defined on some set Ω, with values as subsets of a set U (arbitrary). Recall that a selection of S is a function $X : S \to U$ such that $X(\omega) \in S(\omega)$, for all $\omega \in \Omega$. The existence of a selection is the axiom of choice. In our case, there is more mathematical structure involved, namely, a probability space $(\Omega, \mathscr{A}, P)$ and U together with some σ-field $\mathscr{B}$ on it. We seek selections which are $\mathscr{A} - \mathscr{B}$-measurable as well as "almost sure selections" in the sense that the selection X of S is measurable and $X \in S$ except on a P-null set of Ω.

Now, from the given structure (U, π), $(2^U, F)$ with $F \leqslant \pi$, we consider the probability space

$$\left(2^U \times [0, 1], dF \otimes dx \right)$$

and the random set $S : 2^U \times [0, 1] \to 2^U$ defined by $S(A, t) = A$ for all $t \in [0, 1]$. The random set S has F as its distribution function. The solution to the converse problem consists of showing that the condition $F \leqslant \pi$ is sufficient for the existence of a random variable

$$X : 2^U \times [0, 1] \to U$$

such that $(dF \otimes dx)(X \in S) = 1$ and π is the probability law of X on U, i.e. X is a selector of S.

Without going further into technical details at this stage, let us mention that, the result is expected since the framework *has the flavor* of an old selection problem, called the *marriage problem* started out with Hall (1935) as a problem of choosing distinct representatives elements for a class of subsets of some set. The marriage problem is this. Consider two finite sets B (boys) and G (girls), of same cardinality, say. Each boy $b \in B$ is acquainted with a set of girls $S(b)$, so that $S : B \to 2^G$. Suppose we are interested in the question "under what conditions is it possible for each boy to marry one of his acquaintances?". This is a selection problem of a particular type, namely an injective (one-to-one) selection. Specifically, we seek conditions under which there exists a function $X : B \to G$ with $X(b)$ being the girl in $S(b)$ who is chosen by b for marriage. Clearly, if $b_1 \neq b_2$ then $X(b_1) \neq X(b_2)$!

Suppose $S(b_1) = \{g_1\}$, $S(b_2) = \{g_2\}$, $S(b_3) = \{g_1, g_2\}$ then it is not possible for these boys $\{b_1, b_2, b_3\}$ to marry their acquaintances. Thus, the *necessary condition* for the marriage problem is that any k boys should know, collectively, at least k girls, i.e.,

$$|A| \leqslant \left| \bigcup_{b \in A} S(b) \right| \tag{2}$$

for all $A \subseteq B$.

The remarkable result of Hall is that (2) is also sufficient.

The "analogy" of the marriage problem with the selection problem in coarse data analysis can be seen as follows. Let U be a finite set (take $B = G = U$). Let μ be the counting probability measure on 2^U, i.e. $A \subseteq U$, $\mu(A) = |A|/|U|$. Let S be a (non-empty) random set U, defined on the probability space $(U, 2^U, \mu)$. Let $X : U \to U$ be the map $X(u) = u$, for all $u \in U$, so that $P_X = \mu$. Clearly X is an a.s. selector of S if for all $u \in U$, $u \in S(u)$. But then X is also an a.s. of the random set $S' : U \to 2^U \setminus \{\emptyset\}$ where $S'(u) = \{v \in U : u \in S(v)\}$. This implies that for all $u, v \in U$, $u \in S(v) \Leftrightarrow v \in S'(u)$. The necessary and sufficient condition of *Hall's theorem* takes the form, for all $A \subseteq U$,

$$|A| \leqslant \left| \bigcup_{u \in A} S'(u) \right| = |\{u \in U : S(u) \cap A \neq \emptyset\}|$$

is equivalent to $\mu(A) \leqslant \mu(S \cap A \neq \emptyset)$ i.e. $P_X(A) \leqslant T_S(A)$.

The fact that the core of T (or F) consists precisely of probability laws of all a.s. selectors of a non-empty random set S (i.e. the core of T is the parameter space for statistical inference) can be proved probabilistically in a fairly general setting, using the concept of ordered coupling of random closed sets on topological spaces (Norberg, 1992). In the following, we investigate the finite case in which we describe elements of the core of T as *allocations* of the probability density function of S.

But first, since here F is monotone of infinite order (equivalently, T is alternating of infinite order), $\mathscr{C}(F) \neq \emptyset$. In fact, this follows from a more general result of Shapley (1971) for *convex games* on U. In our context of random sets, i.e. when F is a distribution function (and hence monotone of infinite order), we can show that $\mathscr{C}(F) \neq \emptyset$ by simply constructing a $\pi \in \mathbb{P}$ such that $\pi \geqslant F$, as follows.

Let f be the distribution function of F, i.e.,

$$f(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} F(B).$$

For a fixed $\emptyset \neq A \subseteq U$, define

$$g_A : U \to \mathbb{R}$$

by

$$g_A(u) = \begin{cases} \displaystyle\sum_{u \in B \subseteq A} \frac{f(B)}{|B|} & \text{for } u \in A, \\ \displaystyle\sum_{u \in B} \frac{f(B)}{|B \setminus (A \cap B)|} & \text{for } u \notin A. \end{cases}$$

Then $g_A(\cdot) \geqslant 0$ and

$$\sum_{u \in U} g_A(u) = \sum_{u \in A} g_A(u) + \sum_{u \notin A} g_A(u).$$

But

$$\sum_{u \in A} g_A(u) = \sum_{u \in A} \sum_{u \in B \subseteq A} \frac{f(B)}{|B|} = \sum_{B \subseteq A} \frac{f(B)}{|B|} \sum_{u \in B} 1 = \sum_{B \subseteq A} f(B)$$

and

$$\sum_{u \notin A} g_A(u) = \sum_{B \nsubseteq A} \frac{f(B)}{|B \setminus A \cap B|} \sum_{u \in B \setminus A \cap B} 1 = \sum_{B \nsubseteq A} f(B).$$

Thus,

$$\sum_{u \in U} g_A(u) = \sum_{B \subseteq} f(B) = 1.$$

Hence $g_A(\cdot)$ is a probability density on U. Let π_A denote the probability measure on U associated with g_A, then for $B \subseteq U$, we have

$$\pi_A(B) = \sum_{u \in B} g_A(u) = \sum_{u \in A \cap B} g_A(u) + \sum_{u \in A' \cap B} g_A(u) = \sum_{u \in A \cap B} \sum_{u \in D \subseteq A} \frac{f(D)}{|D|} + \sum_{u \in D \cap A'} \sum_{u \in D \cap A' \neq \emptyset} \frac{f(D)}{|D \cap A'|}$$

$$\geqslant \sum_{D \subseteq A \cap B} f(D) + \sum_{D \cap A' \neq \emptyset} f(D).$$

Thus, π_A is a probability measure in $\mathbb{P}$ such that $\pi_A \geqslant F$, i.e. $\pi_A \in \mathscr{C}(F)$. Moreover, the above construction of π_A exhibits the fact that, in particular, $\pi_A(A) = F(A)$, so that for each $A \subseteq U$,

$$F(A) = \inf\{\pi(A) : \pi \in \mathscr{C}(F)\}$$

i.e., F is the lower envelop of its core $\mathscr{C}(F)$.

The above construction leads to a detailed description of $\mathscr{C}(F)$. For $\emptyset \neq A \subseteq U$, and $u \in A$, let the $\alpha(u, A)$ be non-negative numbers such that $\sum_{u \in A} \alpha(u, A) = f(A)$. For example, take $\alpha(u, A) = f(A)/|A|$ but there are many other choices. Such an α is called an *allocation* of F density. It can be checked that the probability measure $\pi_\alpha(A) = \sum_{u \in A} g_\alpha(u) \geqslant F(A)$, and if $\pi \geqslant F$ then $\pi = \pi_\alpha$ for some allocation α.

Let $\mathbb{P}$ denote the class of all probability measures on U. The core of T, denoted by $\mathscr{C}(T)$, is $\mathscr{C}(T) = \{P \in \mathbb{P} : P \leqslant T\}$, where $P \leqslant T$ means for all $A \subseteq U$, $P(A) \leqslant T(A)$ with set-valued observations is this. Consider the situation in *coarse data analysis*. The random variable X with values in U is unobservable. Instead, a coarsening S of X is observable. X is an almost sure selector of S which is a (non-empty) random set on U with unknown capacity functional T. The true (unknown) probability law of X is an element of $\mathscr{C}(T)$ which plays the role of a "parameter space". However, $\mathscr{C}(T)$ is unknown. It can be consistently estimated from a random sample $S_1, S_2, \ldots, S_n$ drawn from S. Indeed, let T_n denote the empirical capacity functional based on $S_1, S_2, \ldots, S_n$, i.e.,

$$T_n(A) = \frac{1}{n} \# \left\{ 1 \leqslant j \leqslant n : S_j \cap A \neq \emptyset \right\}.$$

Then by the strong law of large numbers, $T_n(A) \to T(A)$, with probability 1, as $n \to +\infty$. Moreover, the empirical probability dF_n based on the unobservable $X_1, X_2, \ldots, X_n$ from X belongs to $\mathscr{C}(T_n)$ almost surely. As such, inference about the true probability law of X can be based upon the approximation of $\mathscr{C}(T)$ by $\mathscr{C}(T_n)$, for n sufficiently large.

4.3. Some techniques inspired from game theory

We proceed to investigate the structure of the core of a capacity functional T, denoted by $\mathscr{C}(T)$, or equivalently, the core of its associated distribution function F (of a non-empty random set on the finite set U), denoted by $\mathscr{C}(F)$, where

$\mathscr{C}(F) = \{P \in \mathbb{P} : F \leqslant P\}$. In the following f always denoted the Möbius transform of F, which is a probability density on 2^U with $f(\emptyset) = 0$.

Given f, there is a natural way to construct densities on U. These are *allocations*.

Definition 4.1. Let f be a density on 2^U with $f(\emptyset) = 0$. An *allocation* of f is a function $\alpha : U \times (2^U \setminus \{\emptyset\}) \to [0, 1]$ such that for all $A \subseteq U$, $\sum_{u \in A} \alpha(u, A) = f(A)$.

Each allocation α gives rise to a density g_α on U, namely $g_\alpha(u) = \sum_{u \in A} \alpha(u, A)$ where the sum is over all sets A containing u. Indeed, $g_\alpha(\cdot) \geqslant 0$, and

$$\sum_{u \in U} g_\alpha(u) = \sum_{u \in U} \sum_{u \in A} \alpha(u, A) = \sum_{A \subseteq U} \sum_{u \in A} \alpha(u, A) = \sum_{A \subseteq U} f(A) = 1.$$

Example 1. Let $p(\cdot) : U \to [0, 1]$ be a density such that $p(u) \neq 0$, for all $u \in U$. We write $P(A) = \sum_{u \in A} p(u)$ for $A \subseteq U$. Then clearly

$$\alpha(u, A) = \frac{f(A)}{P(A)} p(u)$$

is an allocation. In particular, if $p(\cdot)$ is the uniform probability density on U, i.e., $p(u) = 1/|U|$, for all $u \in U$, then

$$\alpha(u, A) = \frac{f(A)}{|A|}.$$

Example 2. Let $\{U_1, U_2, \ldots, U_n\}$ be a partition of U with $U_i \neq \emptyset$, $i = 1, 2, \ldots, k$. Let P be a probability measure on U and $p(\cdot)$ its associated density on U, i.e. $p(\cdot) : U \to [0, 1]$, $p(u) = P(\{u\})$.

Let $f : 2^U \to [0, 1]$ be

$$f(A) = \begin{cases} P(U_i) & \text{if } A = U_i, \\ 0 & \text{if } A \text{ is not one of the } U_i\text{'s.} \end{cases}$$

Then f is a density on 2^U with $f(\emptyset) = 0$. Let

$$\alpha(u, A) = \begin{cases} p(u) & \text{if } u \in A = U_i, \\ 0 & \text{otherwise.} \end{cases}$$

Then clearly α is an allocation of f.

Example 3. Let U with $|U| = n$ and f be a density on 2^U with $f(\emptyset) = 0$. Let $U = \{u_1, u_2, \ldots, u_n\}$ be an *ordering* of the set U. Let $\alpha(u, A) = f(A)$ if $u = u_j \in A$ with $j = \max\{i : u_i \in A\}$, and zero otherwise. The associated density g_α is

$$g_\alpha(u_i) = \sum_{u_i \in A} \alpha(u, A) \sum_{u_i \in A \subseteq \{u_1, u_2, \ldots, u_i\}} f(A) = F(\{u_1, u_2, \ldots, u_i\}) - F(\{u_1, u_2, \ldots, u_{i-1}\})$$

for $i = 1, 2, \ldots, n$ (if $i = 1$, then $F(\{u_1, \ldots, u_{i-1}\}) = F(\emptyset) = 0$) where F is the corresponding distribution function with density f, i.e. $F(A) = \sum_{B \subseteq A} f(B)$.

Let U with $|U| = n$, and $\mathbb{P}$ be the set of all probability measures on U. Since each $P \in \mathbb{P}$ is uniquely determined by its density p, we identify P with a vector $p = (p_1, p_2, \ldots, p_n)$, where $p_i = P(\{u_i\})$. Thus, we identify $\mathbb{P}$ with the unit simplex $\mathbb{S}_n$ of $\mathbb{R}^n$, where

$$\mathbb{S}_n = \left\{ p = (p_1, p_2, \ldots, p_n) \in [0, 1]^n, \sum_{i=1}^n p_i = 1 \right\}.$$

In the following, F denotes a distribution function (of a non-empty random set) on 2^U with density f on 2^U. The *capacity functional* T is dual to F as $T(A) = 1 - F(A^c)$. The cores of F and of T are the same (in fact, for arbitrary spaces U):

$$\{P \in \mathbb{P} : P \leqslant T\} = \text{core of } T = \{P \in \mathbb{P} : F \leqslant P\} = \text{core of } F,$$

denoted as $\mathscr{C}(F)$.

It can be checked that $\mathscr{C}(F)$, by identification, is a compact convex subset of the simplex $\mathbb{S}_n$ ($|U| = n$).

There are $n!$ different orderings of the elements of a set U with $|U| = n$. Specifically, let Σ denote the set of all permutations of $\{1, 2, \ldots, n\}$. For $\sigma \in \Sigma$, the elements of U are indexed as $\{u_{\sigma(1)}, u_{\sigma(2)}, \ldots, u_{\sigma(n)}\}$. For each σ, we associate a density g_σ on U as follows:

$$g_\sigma\left(u_{\sigma(1)}\right) = F\left(\{u_{\sigma(1)}\}\right) \text{ and for } i \geqslant 2,$$

$$g_\sigma\left(u_{\sigma(i)}\right) = F\left(\{u_{\sigma(1)}, u_{\sigma(2)}, \ldots, u_{\sigma(i)}\}\right) - F\left(\{u_{\sigma(1)}, \ldots, u_{\sigma(i-1)}\}\right).$$

For example, to simplify notations, take $\sigma(i) = i$, for all $i = 1, \ldots, n$, we write $U = \{u_1, u_2, \ldots, u_n\}$, and dropping σ from our writing (but not from our mind!). The associated probability measure P is defined as $P(A) = \sum_{u \in A} g(U)$, $A \subseteq U$, then $F(A) \leqslant P(A)$ when $A = U$. Suppose $A \neq U$. Since $U \setminus A = A^c \neq \emptyset$, let $j = \min\{i : u_i \in A^c\}$. For $B = \{u_1, u_2, \ldots, u_j\}$, we have $A \cap B = \{u_1, u_2, \ldots, u_{j-1}\}$, $A \cup B = A \cup \{u_j\}$. Since F is monotone of order 2, we have $F(A \cup B) \geqslant F(A) + F(B) - F(A \cap B)$, i.e.

$$F\left(A \cup \{u_j\}\right) \geqslant F(A) + F\left(\{u_1, u_2, \ldots, u_j\}\right) - F\left(\{u_1, u_2, \ldots, u_{j-1}\}\right),$$

or $g\left(u_j\right) \leqslant F\left(A \cup \{u_j\}\right) - F(A)$. But $g\left(u_j\right) = P\left(\{u_j\}\right) = P\left(A \cup \{u_j\}\right) - P(A)$, so that

$$F(A) - P(A) \leqslant F\left(A \cup \{u_j\}\right) - P\left(A \cup \{u_j\}\right).$$

On the right-hand side of the above inequality, viewing $A \cup \{u_j\}$ as another set A', and using the same argument, we have

$$F\left(A \cup \{u_j\}\right) - P\left(A \cup \{u_j\}\right) = F(A') - P(A') \leqslant F\left(A \cup \{u_k\}\right) - P\left(A \cup \{u_k\}\right),$$

where $u_k \neq u_j$. Continuing this process, we arrive at

$$F(A) - P(A) \leqslant F\left(A \cup \{u_j\}\right) - P\left(A \cup \{u_j\}\right) \leqslant F\left(A \cup \{u_j, u_k\}\right) - P\left(A \cup \{u_j, u_k\}\right)$$
$$\leqslant \cdots \leqslant F(U) - P(U) = 1 - 1 = 0,$$

i.e., for all $A \subseteq U$, $F(A) \leqslant P(A)$, so that $P \in \mathscr{C}(F)$.

For each permutation σ of $\{1, 2, \ldots, n\}$, we obtain an element of $\mathscr{C}(F)$ as above, denoted as P_σ. There are $n!$ (not necessarily distinct) P_σ (see Miranda et al., 2003, for the case of 2-alternating capacities). These elements of $\mathscr{C}(F)$ are very special. For example, as indicated in Example 3, they all come from allocations, namely, for every σ, allocate $f(A)$ to the element of A with highest rank. Secondly, we have, for all $A \subseteq U$,

$$F(A) = \inf\{P(A) : P \in \mathscr{C}(F)\},$$

and the infimum is attained for some $P_\sigma(A)$. Indeed, by definition of $\mathscr{C}(F)$,

$$F(A) \leqslant \inf\{P(A) : P \in \mathscr{C}(F)\}.$$

Now for given A, choose σ so that $A = \{u_1, u_2, \ldots, u_k\}$, say. Then

$$P_\sigma(A) = \sum_{i=1}^{k} \left[F\left(\{u_1, \ldots, u_i\}\right) - F\left(\{u_1, \ldots, u_{i-1}\}\right)\right] = F\left(\{u_1, \ldots, u_k\}\right) = F(A).$$

Thus, $F(A) \geqslant \inf\{P(A) : P \in \mathscr{C}(F)\}$.

But the most important fact about these P_σ is that they are the only *extreme points* of $\mathscr{C}(F)$ (when $\mathscr{C}(F)$ is identified as a convex subset of the simplex $\mathbb{S}_n$ in $\mathbb{R}^n$), a remarkable result of Shapley (1971). We state Shapley's result in the following special form:

Theorem 4.1 (*Shapley, 1971*). *Let F be the distribution function of some non-empty random set U with $|U| = n$. Then $C(F)$ is a compact convex polyhedron with at most n! extreme points which are precisely the P_σ's.*

With Shapley's theorem, we are now in position to describe the structure of $\mathscr{C}(F)$.

Theorem 4.2. $C(F)$ *consists of probability measures coming from allocations.*

Proof. Let A denote the subset of P consisting of all probability measures P_α on U coming from allocations α (of the Möbius inverse f of F). Let α be an allocation of f. Then, for all $A \subseteq U$,

$$F(A) = \sum_{B \subseteq A} f(B) = \sum_{B \subseteq A} \sum_{u \in B} \alpha(u, B) \leqslant \sum_{u \in A} \sum u \in B\alpha(u, B) = P_\alpha(A).$$

Thus, $\mathbb{A} \subseteq \mathscr{C}(F)$.

Conversely, by Shapley's theorem, $\mathscr{C}(F)$ has the P_σ's as extreme points, and thus $\mathscr{C}(F)$ is the set of convex combinations of these extreme points. But the P_σ's are elements of $\mathbb{A}$ so that, since $\mathbb{A}$ is clearly convex, $\mathbb{A}$ contains all convex combinations of its elements, in particular, convex combinations of these P_σ's, which is $\mathscr{C}(F)$. $\quad\square$

Example 4 (*Shapley value*). As a set function, a distribution function F can be viewed as a *coalitional game*. The *Shapley value* of the game F is the center of gravity of the extreme points P_σ of $\mathscr{C}(F)$, i.e. it is a probability measure μ on U ($|U| = n$) whose density on U is given by

$$h(u) = \frac{1}{n!} \sum_{\sigma \in \Sigma} g_\sigma(u),$$

where g_σ are densities defined, in terms of F as

$$g_\sigma\left(u_{\sigma(i)}\right) = F\left(\left\{u_{\sigma(1)}, u_{\sigma(2)}, \ldots, u_{\sigma(i)}\right\}\right) - F\left(\left\{u_{\sigma(1)}, \ldots, u_{\sigma(i-1)}\right\}\right),$$

$u \in U, i = 1, 2, \ldots, n$ (for $\sigma(1)$, $F(\emptyset) = 0$, so that $g_\sigma\left(u_{\sigma(1)}\right) = F\left(\left\{u_{\sigma(1)}\right\}\right)$.

As an element of $\mathscr{C}(F)$, μ comes from some allocation α. To see this, we need to write $h(\cdot)$ in terms of the Möbius transform f of F.

Using the relation $F(A) = \sum_{B \subseteq A} f(B)$, a direct computation leads to: for all $u \in U, h(u) = \sum_{u \in A} f(A)/|A|$ where, as understood, the sum is over A. Thus, the Shapley value μ comes from the allocation

$$\alpha(u, A) = \frac{f(A)}{|A|}.$$

It is interesting to observe that the Shapley value μ, viewing as a point $\mu = \left(\mu_1, \mu_2, \ldots, \mu_n\right)$ in the simplex $\mathbb{S}_n$, maximizes the function $H : \mathbb{S}_n \to \mathbb{R}$,

$$H(p) = \sum_{\emptyset \neq A \subseteq U} f(A) \log \left[\prod_{u \in A} p(u)\right]^{1/|A|}$$

where $p = (p_1, p_2, \ldots, p_n) \in \mathbb{S}_n$ and $p_i = p(u_i), i = 1, 2, \ldots, n$.

Example 5 (*Allocations for the CAR model*). Recall the CAR model in coarse data analysis. Let X be a random variable, defined on a probability space $(\Omega, \mathscr{A}, Pr)$ with values in the finite set U with $|U| = n$. Recall again that a coarsening of X is a non-empty random set S with values in $2^U \setminus \{\emptyset\}$ such that X is an almost sure selector of S, i.e. $Pr(X \in S) = 1$. Let f be the density on 2^U of S. The random set S is called a coarsening at random (CAR) of X if its distribution is related to the probability law P of X in some special way. The existence of CAR is proved in Gill et al. (1997). Specifically, there exist CAR probabilities $\pi(A), A \subseteq U$, and a density $g(\cdot)$ on U such that

$$\sum_{u \in A} \pi(A) = 1 \quad \text{for all } u \in U, \tag{3}$$

where the sum is over A, and

$$f(A) = P_g(A)\pi(A) \quad \text{for all } u \in U,$$ (4)

where, as usual, $P_g(A) = \sum_{u \in A} g(u)$. Define

$$\alpha(u, A) = \begin{cases} g(u)\dfrac{f(A)}{P_g(A)} & \text{if } P_g(A) \neq 0, \\ 0 & \text{if } P_g(A) = 0. \end{cases}$$

In view of (3), we see that $\alpha(\cdot, \cdot)$ is an allocation. Now, if $g(u) = 0$, then $\sum_{i \in A} \alpha(u, A) = 0$, and if $g(u) \neq 0$, then $P_g(A) \neq 0$ for any $A \ni u$, thus,

$$\sum_{u \in A} \alpha(u, a) = \sum_{u \in A} g(u)\frac{f(A)}{P_g(A)} = g(u)\sum_{u \in A} \pi(A) = g(u),$$

in view of (4). Thus the CAR model of X, i.e. P_g, comes from the above $\alpha(\cdot, \cdot)$ allocation (and hence belongs to $\mathscr{C}(F)$! where F is the distribution function with density f on 2^U).

Note that, like the Shapley value, the CAR solution P_g, viewed as a point in the simplex $\mathbb{S}_n$, maximizes the following entropy function:

$$G : \mathbb{S}_n \to \mathbb{R},$$

$$G(p) = \sum_{\emptyset \neq A \subseteq U} f(A) \log\left[\frac{P(A)}{|A|}\right],$$

where $P(A) = \sum_{u \in A} P(u)$, $p = (p_1, p_2, \ldots, p_n) \in \mathbb{S}_n$, $p_i = P(u_i)$, $i = 1, 2, \ldots, n$, where the sum is over A containing u, and f is the Möbius inverse of F.

5. Concluding remarks

In the context of coarse data analysis, random sets appear to be a natural mathematical model for set-valued observations. As bonafide random elements taking values in separable metric spaces, the theory of random sets generalizes multivariate statistical analysis in a rigorous fashion. Extending Matheron's hit-or-miss topology on the hyperspace of closed sets (of Hausdorff, locally compact, second countable spaces) to an appropriate topology on the function space of upper semi continuous membership functions (of fuzzy sets), the theory of random fuzzy sets is firmly established within standard probability theory in which statistical inference can be derived along the lines of standard statistical theory. On the application side, taking into account random fuzzy sets, as imprecise data, both in terms of occurrences as well as meaning interpretations, is a realistic approach to contemporary emerging problems. The research in this direction will enlarge the domain of applicability of statistical science.

References

Bandemer, H., Näther, W., 1992. Fuzzy Data Analysis. Kluwer Academic, Boston.

Buckley, J.J., 2004. Fuzzy Statistics. Springer, Berlin.

Dempster, A.P., 1967. Upper and lower probabilities induced from a multivalued mapping. Ann. Math. Statist. 38, 325–339.

Diamond, P., Kloeden, P., 1994. Metric Spaces of Fuzzy Sets. World Scientific, Singapore.

Dubois, D., Prade, H., 1987. The mean value of a fuzzy number. Fuzzy Sets and Systems 24, 279–300.

Dubois, D., Prade, H., 1988. Possibility Theory. Plenum, New York.

Elkan, C., 1993. The paradoxical success of fuzzy logics. Proceedings AAAI-93. Amer. Assoc. AI, 698-703.

Fagin, R., Halpern, J.Y., 1991. Uncertainty, belief, and probability. Comput. Intelligence 7, 160–173.

Feng, D.-J., Feng, D., 2004. On a statistical framework for estimation from random set observations. J. Theoret. Probab. 17, 85–110.

Fréchet, M., 1948. Les éléments aléatoires de nature quelque dans un espace distancié. Ann. Inst. Poincaré IV, 215–310.

Gehrke, M., Walker, C., Walker, E., 1997. A mathematical setting for fuzzy logics. Internat. J. Uncertainty Fuzziness and Knowledge-Based Systems 5, 223–238.

Geyer, C.J., Meeden, G.D., 2005. Fuzzy and randomized confidence intervals and P-values (with discussions). Statist. Sci. 20 (4), 358–366.

Gil, M.A., Lopez-Diaz, M., 1998. Fundamentals and Bayesian analyses of decision problems with fuzzy-valued utilities. Internat. J. Approx. Reasoning 15, 203–224.

Gill, R.D., Van der Laan, M.J., Robins, J.M., 1997. Coarsening at random: characterizations, conjectures, and counterexamples. In: Lecture Notes in Statistics, vol. 123. Springer, Berlin, pp. 255–294.

Goodman, I.R., 1982. Fuzzy sets as equivalence classes of random sets. In: Yager, R. et al. (Eds.), Fuzzy Sets and Possibility Theory. Pergamon Press, New York, pp. 327–432.

Goodman, I.R., Nguyen, H.T., 2000. Fuzziness and randomness. In: Bertoluzza, C. et al. (Eds.), Statistical Modeling Analysis and Management of Fuzzy Data. Physica-Verlag, New York, pp. 3–21.

Goodman, I.R., Nguyen, H.T., Rogers, G.S., 1990. On the scoring rule approach to the admissibility of uncertainty measures in expert systems. Math. Anal. Appl. 159 (2), 550–594.

Grabisch, M., Nguyen, H.T., Walker, E.A., 1994. Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference. Kluwer Academic, Dordrecht.

Hall, P., 1935. On representatives of subsets. J. London Math. Soc. 10, 26–30.

Halpern, J.Y., 2003. Reasoning About Uncertainty. MIT Press, Cambridge, MA.

Hartigan, J.A., 1987. Estimation of a convex density contour in two dimensions. J. Amer. Statist. Assoc. 82, 267–270.

Heitjan, D.F., Rubin, D.B., 1991. Ignorability and coarse data. Ann. Statist. 19, 2244–2253.

Kruse, R., Meyer, K.D., 1987. Statistics with Vague Data. D. Reidel, Dordrecht.

Kruse, R., Gebhardt, J., Gil, M.A., 1999. Fuzzy statistics. In: Webster, J.G., et al. (Eds.), Wiley Encyclopedia of Electrical and Electronics Engineering, vol. 8. Wiley, New York, pp. 181–196.

Li, S., Ogura, Y., Kreinovich, V., 2002. Limit Theorems and Applications of Set-Valued and Fuzzy Set-Valued Random Variables. Kluwer Academic, Boston.

Lindley, D.V., 1982. Scoring rules and the inevitability of probability. Internat. Statist. Rev. 50, 1–26.

Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.

Matheron, G., 1975. Random Sets and Integral Geometry. Wiley, New York.

Miranda, E., Couso, I., Gil, P., 2003. Extreme points of credal sets generated by 2-alternating capacities. Internat. J. Approx. Reasoning 33, 95–115.

Miranda, E., Couso, I., Gil, P., 2004. A random set characterization of possibility measures. Inform. Sci. 168 (1–4), 51–75.

Molchanov, I.S., 2005. Theory of Random Sets. Wiley, New York.

Montenegro, M., Casals, M.R., Lubiano, M.A., Gil, M.A., 2001. Two-sample hypothesis tests of means of a fuzzy random variable. Inform. Sci. 133, 89–100.

Montenegro, M., Colubi, A., Casals, M.R., Gil, M.A., 2004. Asymptotic and bootstrap techniques for testing the expected value of a fuzzy random variable. Metrika 59, 31–49.

Mosteller, F., Youtz, C., 1990. Quantifying probabilistic expressions. Statist. Sci. 5 (1), 2–34.

Nguyen, H.T., 2004. Survey sampling revisited and coarse data analysis. Thailand Statist. 2, 1–19.

Nguyen, H.T., 2006. An Introduction to Random Sets. Chapman & Hall/CRC, Boca Raton, FL.

Nguyen, H.T., Kreinovich, V., 1999. How to divide a territory? A new simple differential formalism for optimization of set-functions. Internat. J. Intell. Systems 14, 223–251.

Nguyen, H.T., Walker, E.A., 2000. A First Course in Fuzzy Logic. Chapman & Hall/CRC, Boca Raton, FL.

Nguyen, H.T., Wu, B., 2000. Fuzzy Mathematics and Statistical Application. Taipei, Taiwan. (in Chinese).

Nguyen, H.T., Wu, B., 2006. Fundamentals of Fuzzy Statistics. Springer, Berlin.

Nguyen, H.T., Kosheva, O., Kreinovich, V., 1996. Is the success of fuzzy logics really paradoxical? J. Intell. Syst. 11 (5), 295–326.

Norberg, T., 1992. On the existence of ordered couplings of random sets with applications. Israel J. Math. 77, 241–264.

Orlowski, S.A., 1994. Calculus of Decomposable Properties, Fuzzy Sets and Decisions. Allerton Press, New York.

Polonik, W., 1995. Density estimation under qualitative assumptions in higher dimensions. J. Multi. Anal. 55, 61–81.

Schreiber, T., 2000. Statistical inference from set-valued observations. Probab. Math. Statist. 20, 223–235.

Shafer, G., 1976. A Mathematical Theory of Evidence. Princeton University Press, Princeton.

Shapley, L.S., 1971. Cores of convex games. Internat. J. Game Theory 1, 11–26.

Singpurwalla, N.D., Booker, J.M., 2004. Membership functions and probability measures of fuzzy sets. J. Amer. Statist. Assoc. 99 (467), 867–889.

Wagner, D.H., 1977. Survey of measurable selection theorems. SIAM J. Control Optim. 15, 859–903.

Walley, P., 1991. Statistical Reasoning with Imprecise Probabilities. Chapman & Hall, New York.

Wallsten, T.S., 1986. Measuring the vague meaning of probability terms. J. Experiment. Psychol. 115 (4), 348–365.

Zadeh, L.A., 1978. Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1, 2–28.

Zadeh, L.A., 2002. Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. J. Statist. Plann. Inference 102, 233–264.