A Notion of Comparative Probabilistic Entropy based on the Possibilistic Specificity Ordering

Didier Dubois[1] and Eyke Hüllermeier[2]

[1] Institut de Recherche en Informatique de Toulouse, France
dubois@irit.fr
[2] Faculty of Computer Science, University of Magdeburg, Germany
eyke.huellermeier@iti.cs.uni-magdeburg.de

Abstract. In this paper, we reconsider the problem of deciding whether one probability distribution is more informative (in the sense of representing a less indeterminate situation) than another one. Instead of using well-established information measures such as the Shannon entropy, however, we take up the idea of comparing probability distributions in a qualitative way. More specifically, we focus on a natural partial ordering induced by what is called the "peakedness" of a distribution. Moreover, there is a close connection between this ordering between probability distributions and the standard specificity ordering on possibility distributions that can be constructed from them. The main result of the paper is a proof showing that possibilistic specificity is consistent with probabilistic entropy in the sense that the (total) ordering defined by the latter refines the (partial) ordering defined by the former.

1 Introduction

The principle of maximum entropy plays an important role in probability theory, especially in the case of incomplete probabilistic models (see e.g. Paris [14]). It is instrumental in selecting a probability distribution in agreement with the available constraints, preserving as much indeterminateness as possible and verifying as many independence assumptions as possible. More precisely, entropy faithfully accounts for existing dependencies and only assumes independence where no justification to the contrary can be found. There are axiomatic characterizations of the Shannon entropy function, and Paris [14] has strongly advocated the selection of the maximum entropy probability as being a reasonable default choice under basic principles.

In possibility theory, a similar kind of "least commitment" information principle exists (e.g. Dubois et al. [8]): When a set of constraints delimits a set of possibility distributions, the least committed choice is the minimally specific distribution. The underlying idea is to consider any situation as possible as long it is not explicitly ruled out by the constraints. This principle obviously suggests *maximizing* possibility degrees.

There also exists a natural *partial* information ordering between possibility distributions, called the *specificity relation.* This ordering is based on fuzzy set inclusion: If a possibility distribution $\pi : X \to [0, 1]$ is pointwisely dominated by another distribution $\pi' : X \to [0, 1]$, i.e. $\pi(x) \leq \pi'(x)$ for all $x \in X$, the former is said to be more specific than the latter (and strictly more specific if $\pi(x) < \pi'(x)$ for at least one $x \in X$). The natural measure of non-specificity in agreement with this partial ordering is the sum of the possibility degrees.[3]

There is a close connection between maximal entropy and minimal specificity principles, especially in the light of the Laplace indifference principle: In the possibilistic framework, the case of complete ignorance is adequately represented by the uniform distribution $\pi \equiv 1$ (all x are completely possible). Likewise, if a unique probability distribution must be picked, the aforementioned indifference principle suggests selecting the uniform distribution $p \equiv |X|^{-1}$. For these distributions, the Shannon entropy and the additive possibilistic measure of non-specificity coincide with the Hartley entropy of a set (Higashi and Klir [10]), that is the logarithm of the number of elements in the set. These authors use an additive index of possibilistic non-specificity that looks like Shannon entropy.

The temptation to compare specificity and entropy is great. For instance, Klir [11] has tried to equate numerical entropy and (additive) non-specificity indices for the purpose of transforming possibility distributions into probability distributions and conversely. This is debatable, however, because the entropy scale and the specificity scale are not commensurate. Maung [12] has tried to justify the principle of minimal specificity by adapting Paris' rationality axioms to the possibilistic setting.

Regarding the information-based comparison of distributions, there is an important difference between the probability and possibility settings. In the uncertainty literature, the comparison between probability distributions is always based on a type of entropy index without reference to an underlying *partial* ordering, which is directly defined between probability distributions. There are actually several entropy indices and similar ones (such as the Gini index) but they never stem from a partial ordering that decides if a probability measure is more informative than another one.

Interestingly, it turns out that an old paper by Birnbaum [1] suggested such a qualitative comparison of probabilities on the real line in terms of what is called their *peakedness*, independently of the notion of entropy. It basically consists of checking the nestedness of confidence intervals of various confidence levels extracted from the probability distribution. Interestingly, the nestedness property of confidence intervals strongly suggests a similarity between the relative peakedness of probability distributions and the relative specificity of possibility distributions. On the other hand, the more peaked a probability distribution, the less indeterminate it is, and the lower its entropy should be.

The aim of this paper is to prove that these intuitions are mathematically valid, thereby bridging the gap between Birnbaum peakedness adapted to the finite setting, and Shannon entropy. We show that the former is a partial ordering on probabilities compatible with the latter index (and other related information

[3] Of course, here we assume the domain X to be finite or at least countable. Otherwise, the sum must be replaced by an integral.

measures). The connection uses possibility theory because checking the peakedness relation between two probability distributions comes down to comparing, in terms of specificity, possibility distributions whose cuts are the confidence intervals of the original probability distributions. These possibility distributions are in fact the most specific transforms from probability to possibility, already proposed by Dubois and Prade [3], and Delgado and Moral [2] in the eighties.

The paper thus establishes a new link between possibilistic and probabilistic traditions, and proposes a qualitative comparison test between probability distributions that may arguably be considered as the natural information ordering between probability functions, something that, to the best of our knowledge, is apparently missing in the uncertainty literature.

The next section introduces the basic notions of possibilistic specificity and probabilistic peakedness and discusses a particular type of probability-possibility transformation. Our main result, establishing the consistency between the possibilistic specificity ordering and the probabilistic entropy measure, is stated and proved in section 3. A discussion of related works follows in section 4. The paper concludes with a summary and an outlook in section 5.

2 Specificity, Peakedness and Probability-Possibility Transforms

Consider two probability distributions (probability vectors) $\alpha = (\alpha_1 \ldots \alpha_n)$ and $\beta = (\beta_1 \ldots \beta_n)$ defined over a finite domain X of cardinality n; α_i resp. β_i denote the probability $\Pr(x_i)$ of the i-th element x_i. We denote by $a = O(\alpha)$ the *ordered* probability vector obtained from the vector α by rearranging the probability degrees α_i in a non-increasing order. That is,

$$a = (a_1 \ldots a_n) = (\alpha_{\sigma(1)} \ldots \alpha_{\sigma(n)}),$$

where σ is a permutation of $\{1 \ldots n\}$ such that $\alpha_{\sigma(i)} \geq \alpha_{\sigma(j)}$ for $i < j$. Likewise, we denote by $b = (b_1 \ldots b_n) = O(\beta)$ the ordered probability vector associated with β. Since both a and b are still probability vectors, they do of course satisfy the characteristic properties $a, b \geq 0$, $\sum_{j=1}^{n} a_j = \sum_{j=1}^{n} b_j = 1$.

A possibility distribution π is a mapping from X to the unit interval such that $\pi(x) = 1$ for some $x \in X$. A possibility degree $\pi(x)$ expresses the absence of surprise about x being the actual state of the world, and can be viewed as an upper bound of a probability degree [6]. Let $\pi = T(a)$ be the possibility distribution derived from the (ordered) probability vector a according to the following probability-possibility transformation suggested by Dubois and Prade [3]:

$$\pi_i = \sum_{j=i}^{n} a_j, \qquad i = 1 \ldots n. \tag{1}$$

Obviously, $1 = \pi_1 \geq \ldots \geq \pi_n$. We note that this possibility function is also a (de-)cumulative distribution function with respect to the ordering induced by the probability values. Moreover, the possibility measure $\mathit{\Pi}$ associated with π dominates the corresponding probability measure Pr, that is,

$$\forall A \subseteq X \, : \, \mathit{\Pi}(A) \geq \Pr(A)$$

with $\mathit{\Pi}(A) = \max_{x_i \in A} \pi_i$ and $\Pr(A) = \sum_{x_i \in A} a_i$.

In the following definition, we recall a basic notion from possibility theory (e.g. Dubois et al. [8]) already mentioned in the introduction.

Definition 1. *We say that a possibility distribution π is* more specific *than a possibility distribution ρ iff $\pi_i \leq \rho_i$ for all $1 \leq i \leq n$. It is strictly more specific if $\pi_i < \rho_i$ for at least one index $i \in \{1 \ldots n\}$.*

Clearly, the more specific π, the more informative it is. If $\pi(x_i) = 1$ for some i and $\pi(x_j) = 0$ for all $j \neq i$, then π is maximally specific (full knowledge); if $\pi(x_i) = 1$ for all i, then π is minimally specific (no information).

It turns out that $T(a)$ is a maximally specific element of the family of possibility measures that dominate the probability function Pr induced by the distribution a (see Dubois and Prade [3], and Delgado and Moral [2]). Moreover, if the ordering induced by a on X is linear (i.e. $a_i \neq a_j$ for all $i \neq j$) then $T(a)$ is the *unique* maximally specific dominating possibility distribution and respecting the ordering induced by the probability assignment. When there are elements of equal probability, the uniqueness of the maximally specific dominating possibility distribution can be recovered if the ordering induced by π on X is requested to be the same as the ordering induced by a (but then the equation defining $T(a)$ must be adjusted accordingly).

Probability-possibility transformations have been extended to the real line by Dubois et al. [7] (see also Dubois et al. [9]). Let p be a unimodal continuous probability density. It is first proved that the most narrow prediction interval I such that $\Pr(I) \geq \lambda$, where λ is a fixed confidence level, is of the form $I_\lambda = \{ \, x \,|\, p(x) \geq \theta \, \}$ for some threshold θ. Then the most specific possibility transform (inducing the same ordering as p on the real line) is $\pi = T(p)$ such that

$$\forall \, x \in R \, : \, \pi(x) = \pi(y) = 1 - \Pr([x, y]),$$

where $[x, y] = I_{p(x)}$.

In 1948, Birnbaum dealt with what he called the *quality* of a probability distribution, referring to its peakedness. Considering that the fourth moment of a distribution is not an appropriate measure of peakedness he proposed a definition of the relative peakedness of distributions as follows:

Definition 2. *Let Y and Z be real random variables and y_1 and z_1 real constants. Y is said to be more peaked about y_1 than Z about z_1 if and only if*

$$\Pr(\mid Y - y_1 \mid \geq t) \; \leq \; \Pr(\mid Z - z_1 \mid \geq t)$$

holds for all $t \geq 0$.

It is clear that the function

$$\pi_y(y_1 - t) = \pi_y(y_1 + t) = \Pr(\mid x - y_1 \mid \geq t) = 1 - \Pr([y_1 - t, y_1 + t])$$

is a possibility distribution, and easy to show that for any choice of y_1, its possibility measure dominates Pr (see Dubois et al. [9]). In this paper, we shall adapt this definition in two ways: First, the results on the probability-possibility transforms clearly indicate that for unimodal densities, choosing y_1 as the mode of the distribution is reasonable. Moreover, Birnbaum [1] considers intervals whose common midpoint is y_1, yielding a symmetric possibility distribution even if the density is not symmetric by itself. Instead of intervals of the form $[y_1 - t, y_1 + t]$, we shall use intervals of the form $\{x \,|\, p(x) \geq \theta\}$, since they lead to a possibility distribution of the same shape as the probability density (and peakedness refers to the shape of this density anyway). This change enables peakedness to be defined for any referential set, not just the reals. Indeed, the set $\{x \,|\, p(x) \geq \theta\}$ makes sense in general, if measurability is ensured, while $[y_1 - t, y_1 + t]$ assumes the real line as an underlying domain. Here, we nevertheless restrict ourselves to the case of a finite referential set, because entropy indices are usually applied to such domains.

Now, for $\pi = T(a)$ it is clear that $\pi_i = 1 - \Pr(\{x \mid \Pr(\{x\}) \geq \theta\})$ if $a_{i-1} \geq \theta > a_i$, so that the above considerations motivate the following variant of the original peakedness relation due to Birnbaum.

Definition 3. *Let $\pi = T(a)$ be the transformation (1) of an ordered probability vector a, i.e. $\pi_i = \sum_{j=i}^{n} a_j$. We say that a probability distribution α on a finite set X is* more peaked *than a distribution β on X iff $\pi_i \leq \rho_i$ for all $1 \leq i \leq n$, where $\pi = T(O(\alpha))$ and $\rho = T(O(\beta))$. We say that α is* strictly *more peaked than β if it is more peaked and $\pi_i < \rho_i$ for at least one index $i \in \{1 \ldots n\}$.*

Subsequently, the peakedness relation is understood in the sense of this definition. It is clear that it compares probability distributions by means of the specificity relation applied to their optimal possibility transforms. The *less peaked* relation is obviously invariant under permutations of the involved probability vectors. Therefore, we restrict our attention to ordered probability or possibility vectors in the next section.

Example 1. For the two probability distributions specified by the probability vectors

$$\alpha = (\, .05 \ .20 \ .25 \ .25 \ .20 \ .05 \,),$$
$$\beta = (\, .30 \ .15 \ .05 \ .05 \ .15 \ .30 \,)$$

(see Fig. 1 for a graphical illustration) we obtain

$$\pi = (\, 1.0 \ .75 \ .50 \ .30 \ .10 \ .05 \,),$$
$$\rho = (\, 1.0 \ .70 \ .40 \ .25 \ .10 \ .05 \,).$$

Since $\pi \geq \rho$ (and $\pi_2 > \rho_2$), α is (strictly) less peaked than β.

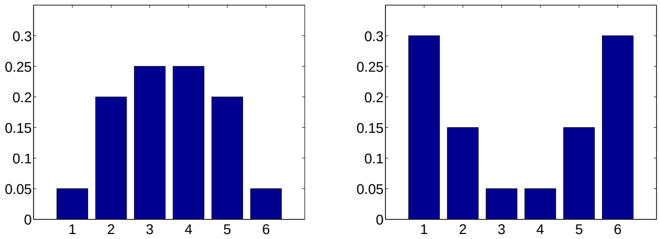

Fig. 1. The probability distribution on the left is (strictly) less peaked than the one on the right.

3 From Peakedness to Entropy

The aim of this section is to prove that the peakedness relation, which is expressed in terms of possibilistic specificity, is consistent with the ordering of probability distributions induced by Shannon entropy.

Definition 4. *The* entropy *of a probability distribution a is defined by*

$$E(a) = -\sum_{j=1}^{n} a_j \cdot \log a_j . \tag{2}$$

The main result of this paper claims that the entropy ordering refines the peakedness relation.

Theorem 1. *If a probability vector a is less peaked than a vector b, then $E(a) \geq E(b)$; if a is strictly less peaked than b, then $E(a) > E(b)$.*

Below, we shall prove this theorem in the following way: We construct a sequence of probability vectors $a^0, a^1, \ldots, a^m$ such that $a^0 = a$, $a^m = b$ and a^{k+1} is more peaked than a^k. Moreover, this sequence will satisfy $E(a^k) \geq E(a^{k+1})$ (resp. $E(a^k) > E(a^{k+1})$) for all $1 \leq k \leq m-1$.

Remark 1. Simple counterexamples can be constructed showing that an implication in the other direction, namely that $E(a) \geq E(b)$ implies a to be less peaked than b, does not hold. In fact, such an implication cannot be expected since the entropy measure induces a total ordering on the class of probability measures, whereas the peakedness relation defines only a partial ordering. In other words, the former ordering is a proper refinement of the latter one.

3.1 Auxiliary Result

Let a and b denote two (ordered) probability vectors such that a is strictly less peaked than b. Starting with $a^0 = a$, a distribution a^{k+1} will be obtained from a distribution a^k by shifting a part of the probability mass a^k_j to a^k_i for appropriately defined indices $j > i$. More generally, a shifting operation $S(a, i, j, c)$ will transform an ordered vector $a = (a_1 \dots a_i \dots a_j \dots a_n)$ into the ordered vector

$$a^c = (a_1 \dots a_i + c \dots a_j - c \dots a_n).$$

Note that if $\pi = T(a)$ and $\pi^c = T(a^c)$ denote, respectively, the possibilistic transforms of a and a^c, then

$$\pi^c_k = \begin{cases} \pi_k & \text{if} \quad k \leq i \\ \pi_k & \text{if} \quad j < k \\ \pi_k - c & \text{if} \quad i < k \leq j \end{cases} \tag{3}$$

Thus, $\pi^c \leq \pi$ does obviously hold true, and a^c is strictly more peaked than a in the case where $c > 0$.

To guarantee a shifting operation $S(a, i, j, c)$ to be valid in the scope of turning a into b, the choice of c must satisfy the following conditions:

(i.) Proper ordering : $a_{i-1} \geq a_i + c$ and $a_j - c \geq a_{j+1}$
(ii.) Limited increase of specificity: $\pi^c \geq \rho$

Recalling (3), the latter item means that

$$\pi^c_k = \sum_{i=k}^{n} a_i - c \geq \sum_{i=k}^{n} b_i = \rho_k$$

for all $i < k \leq j$. Define $d_k = a_k - b_k$. Since $\pi = T(a) \geq T(b) = \rho$ by assumption, we have $\sum_{m=k}^{n} d_m \geq 0$ for all $1 \leq k \leq n$. The condition $\pi^c \geq \rho$ can thus be written as

$$\forall i < k \leq j \, : \, c \leq \sum_{m=k}^{n} d_m.$$

To satisfy both (i.) and (ii.), we hence need

$$c \leq \min\left(\min_{i<k\leq j} \sum_{m=k}^{n} d_m, \, a_{i-1} - a_i, \, a_j - a_{j+1} \right). \tag{4}$$

Since $a \neq b$, there exists $j = \max\{k \,|\, a_k \neq b_k\}$. Of course, $a_j > b_j$ since $\pi \geq \rho$. By definition, we also have $d_j = a_j - b_j = \pi_j - \rho_j$. Since a and b are probability distributions, there must be some $i < j$ such that $b_i > a_i$. So, let

$$i = \max\{k \,|\, 1 < k \leq n, \, b_k > a_k \text{ and } a_{k-1} > a_k\} \tag{5}$$

if the set on the right-hand side is not empty (as will be assumed for the time being).

In order to simplify the upper bound on the number c, we first derive a lower bound on the quantity $\min_{i<k\leq j} \sum_{m=k}^{n} d_m$ that appears as the first argument on the right-hand side of (4).

Lemma 1. $\min_{i<k\leq j} \sum_{m=k}^{n} d_m \geq \min\{a_j - b_j, b_i - a_i\}$

Proof: Define $D(k) = \sum_{m=k}^{n} d_m$ and $D = \min_{i<k\leq j} D(k)$. We consider two cases:
(a) $D = D(j)$. In this case, the lemma obviously holds, since $d_m = a_m - b_m = 0$ for $m > j$ and hence $D(j) = a_j - b_j$.
(b) $D < D(j)$. In this case, there must be an index k_0 with $i < k_0 < j$ and such that $D(k_0) < D(k_0 + 1)$. We claim that

$$D(i + 1) < D(i + 2) < \ldots < D(k_0). \tag{6}$$

In fact, since $D(k_0) < D(k_0 + 1)$ we have $a_{k_0} < b_{k_0}$. Thus, either $i = k_0 - 1$ (in which case (6) does trivially hold), or $a_{k_0-1} = a_{k_0}$ (since if $a_{k_0-1} > a_{k_0}$ and $a_{k_0} < b_{k_0}$, the index k_0 is a potential candidate for the choice of i). In the latter case, $a_{k_0-1} = a_{k_0} < b_{k_0} \leq b_{k_0-1}$ and therefore

$$D(k_0 - 1) = D(k_0) + (a_{k_0-1} - b_{k_0-1}) < D(k_0).$$

This argument can be repeated, showing that (6) does indeed hold. This in turn shows that $D = D(i + 1)$. Moreover, we then have

$$D = D(i + 1) = D(i) - (a_i - b_i) = D(i) + (b_i - a_i) \geq b_i - a_i$$

since $D(i) \geq 0$.
Overall, we get $D \geq a_j - b_j$ in case (a) and $D \geq b_i - a_i$ in case (b). Thus, the lemma does indeed hold. Q.E.D.

Now, if we let

$$c = \min\left(a_j - b_j,\, b_i - a_i,\, a_{i-1} - a_i \right) \tag{7}$$

then the above results and the fact that

$$a_j - a_{j+1} = (a_j - b_j) + (b_j - a_{j+1}) \geq (a_j - b_j) + (b_{j+1} - a_{j+1}) = (a_j - b_j)$$

guarantee that (4) is satisfied. Moreover, the constant c is strictly positive, since $a_{i-1} - a_i > 0$, $a_j - b_j > 0$, $b_i - a_i > 0$ by construction.

Let us now turn to the case where the right-hand side of (5) is empty.

Lemma 2. *Suppose that a is less peaked than b, and that the right-hand side on (5) is empty. Then $b_1 > a_1$.*

Proof: Suppose that a is less peaked than b. There is some $k < j$ such that $b_k > a_k$. Since the right-hand side on (5) is empty, it holds that $b_u > a_u$ implies $a_u = a_{u-1}$ for all $u < j$. Moreover, since $b_k > a_k$, this implies in turn $b_{k-1} \geq b_k > a_{k-1}$. The fact that $b_1 > a_1$ follows immediately by repeating this argument. Q.E.D.

Regarding the choice of c in the case of an empty right-hand side in (5), the only difference concerns the condition $a^c_{i-1} \geq a^c_i$ which simply becomes unnecessary. Hence, one can define

$$c = \min\left(a_j - b_j,\, b_1 - a_1 \right) \tag{8}$$

and apply the shifting operation $S(a, 1, j, c)$ in the same way as before.

3.2 Proof of the Main Result

Obviously, if the quantity c as defined in (7) (resp. (8)) is shifted from position j to position i (resp. position 1) , then either $a_j^c = b_j$ or $a_i^c = b_i$ or $a_i^c = a_{i-1}$. In any case, at least one of the indices i or j will have a smaller value in the next iteration. Hence, the process of repeating the shifting operation, with i, j, and c as specified above, is well-defined, admissible and turns a into b in a finite number of steps.

Given the above results, Theorem 1 follows immediately from the next lemma (recall that in each step of our iterative procedure, the constant c shifted from index j to index i is strictly positive):

Lemma 3. *Let* $E(a) = -\sum_{j=1}^{n} a_j \cdot \log(a_j)$*. Then* $E(a) > E(a^c)$ *for* $c > 0$*.*

Proof: It is easy to see that $E(a) > E(a^c)$ is equivalent to

$$(a_i + c)\log(a_i + c) - a_i \log(a_i) \; > \; a_j \log(a_j) - (a_j - c)\log(a_j - c).$$

Noting that $a_i > a_j$, this inequality can be secured by showing that the function $x \mapsto x\log(x)$ is strictly convex on $(0,1)$. This is indeed the case, since the second derivative of this function is given by $x \mapsto 1/x$. Q.E.D.

Let us finally note that Theorem 1 can be generalized to informativeness measures other than the standard entropy. In fact, it is easily verified that the logarithm $\log(\cdot)$ in (2) can be replaced by any monotone increasing function $F(\cdot)$ the second derivative $F''(\cdot)$ of which exists on $(0,1)$ and satisfies $F''(x)/F'(x) > -2/x$ for all $0 < x < 1$ (where $F'(\cdot)$ denotes the first derivative).

As an example, consider the case of the well-known Gini measure

$$G(a) \; = \; \sum_{j=1}^{n} (a_j)^2.$$

Since $G(\cdot)$ thus defined is an informativeness index rather than a measure of indeterminateness (such as entropy), we actually have to consider its negation $-G(a) = -\sum_{j=1}^{n}(a_j)^2 = -\sum_{j=1}^{n} a_j F(a_j)$ with $F : x \mapsto x$. Here, we have

$$a \text{ (strictly) less peaked than } b \quad \Rightarrow \quad -G(a)\,(>) \; \geq \; -G(b)$$

since $F'' \equiv 0$.

4 Related Work

Even though the proposed notion of relative informativeness, based on possibilistic specificity and Birnbaum peakedness, seems to be unknown in the uncertainty literature, there is a subfield of the social sciences where similar notions have apparently been developed for some twenty years or so:[4] The study of social welfare orderings.

[4] The authors are grateful to Jérôme Lang for pointing out this connection.

We refer to the book by Moulin [13]. In this framework, X is a set of agents, whose welfare under some life conditions is measured by a utility function over X. The problem is to compare the quality of utility vectors $(u_1 \dots u_n)$ from the standpoint of social welfare. Under an egalitarian program of redistribution from the rich to the poor, the so-called Pigou-Dalton principle of transfer states that transferring some utility from one agent to an other one so as to reduce inequalities of utility values improves the social welfare of the population.[5]

Formally, the transformation of a vector a into a vector a^c as in section 3.1 is known as a Pigou-Dalton transfer. The sequence of transformations we propose here is also used in this literature. Moreover, the role of entropy is played by so-called inequality indices. The counterpart to the possibility transform of a probability vector is called the Lorentz curve of the utility vector, and the counterpart of the peakedness ordering is called the Lorentz dominance relation.

It seems that counterparts to our main results already exist in this literature, and this point would be worth studying in more detail. One difference is that utility vectors do not sum to 1. But Lorentz dominance is precisely making sense for the comparison of utility vectors with equal sum. Note that it would not be the first time that possibility-probability transformations find counterparts in the social sciences. For instance, a transformation from a belief function to a probability measure (obtained by generalizing the Laplace indifference principle) introduced in [3] and called *pignistic transformation* by Smets [16] is known in social science as the Shapley value of cooperative games (see again Moulin [13]).

5 Conclusions and Perspectives

The contribution of this paper is mainly to lay bare a notion of relative information content that can decide if a probability distribution represents more or less uncertainty than another one (or whether the two distributions are not directly comparable). The test we offer appears to be natural in the sense that it exactly captures the notion of relative peakedness of distributions, thus meeting our intuition. The fact that Shannon entropy as well as the Gini index (and many other ones, potentially) refine the peakedness relation corroborates this intuition. It sheds light on the meaning of these indices, that were sometimes dogmatically proposed as natural ones, even if axioms or properties that justify the entropy index were proposed in order to its use for uncertain reasoning more transparent. The peakedness ordering offers a minimal robust foundation for probabilistic information indices. The surprise is that it comes down to comparing two possibility distributions in the sense of their relative specificity (using fuzzy set inclusion!). Finding an extension of these results to continuous probability distributions, using differential entropy for instance, is an obvious next task.

Our discussion also shows that there is a range of arbitrariness in the choice of these indices, namely in the case of two distributions that cannot be compared

[5] This principle does not seem to be popular nowadays.

by the peakedness relation but are ranked in opposite orders by, say, the entropy and the Gini index. This point needs further study. We note, however, that the situation is the same with the specificity relation in possibility theory where several non-specificity indices have been proposed (Higashi and Klir [10], Dubois and Prade [4], Yager [17], Ramer [15]) that disagree with each other. The same difficulty can be observed in the case of belief functions (Dubois and Prade [5]). Besides, the close relationship between peakedness and Lorentz dominance also comforts the legitimacy of the proposed relative probabilistic informativenes notion.

In his book [14], Jeff Paris advocates the use of conditional probability statements as a natural means for expressing knowledge and the maximal entropy principle as a natural tool for selecting a reasonable default probabilistic model of this knowledge. The above results suggest that the maximal entropy principle can be replaced by a minimal peakedness principle in problems with incompletely specified probability distributions. Of course, the minimally peaked distribution in agreement with the constraints may fail to be unique, and the issue of choosing between them is an intriguing one. Anyway, the peakedness relation can be used in all reasoning problems where the information content of a distribution is relevant, for example in machine learning techniques à la decision tree induction where measures of that kind are used for selecting (hopefully) optimal attributes according to which the data is partitioned in a recursive manner. The notion of peakedness is easy to understand, but, compared to entropy and other numerical indices, quite weak and its efficiency in probabilistic reasoning and decision making is still unclear. These issues constitute interesting topics of future research.

Acknowledgements: The authors are grateful to Jürgen Beringer and Jérôme Lang for helpful comments.

References

1. Birnbaum Z. W. On random variables with comparable peakedness, *Annals of Mathematical Statistics*, 19, 1948, 76-81.
2. Delgado M. and Moral S. On the concept of possibility-probability consistency, *Fuzzy Sets and Systems* , 21, 1987 311-318.
3. Dubois D. and Prade H. On several representations of an uncertain body of evidence, in *Fuzzy Information and Decision Processes*, M.M. Gupta, and E. Sanchez, Eds., North-Holland, Amsterdam, 1982, pp. 167-181.
4. Dubois D. and Prade H. A note on measures of specificity for fuzzy sets, *Int. J. of General Systems*, 10, 1985, 279-283.
5. Dubois D. and Prade H.: The principle of minimum specificity as a basis for evidential reasoning, In: *Uncertainty in Knowledge-Based Systems* (B. Bouchon, R.R. Yager, eds.), Springer Verlag, 1987, 75-84.
6. Dubois D. and Prade H. When upper probabilities are possibility measures, *Fuzzy Sets and Systems* , 49,1992 65-74.

7. Dubois D., Prade H. and Sandri S. On possibility/probability transformations. In: *Fuzzy Logic. State of the Art*, (R. Lowen, M. Roubens, eds.), Kluwer Acad. Publ., Dordrecht, 1993, 103-112.
8. Dubois D., Nguyen H. T., Prade H. Possibility theory, probability and fuzzy sets: misunderstandings, bridges and gaps. In: *Fundamentals of Fuzzy Sets*, (Dubois, D. Prade,H., Eds.), Kluwer, Boston, Mass., The Handbooks of Fuzzy Sets Series, 2000 343-438.
9. Dubois D., Foulloy L., Mauris G., Prade H. Possibility/probability transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing* 10, 2004, 273-297.
10. Higashi and Klir G. Measures of uncertainty and information based on possibility distributions, *Int. J. General Systems*, 8, 1982, 43-58.
11. Klir G. A principle of uncertainty and information invariance, *Int. J. of General Systems*, 17, 1990, 249-275.
12. Maung I.Two characterizations of a minimum-information principle in possibilistic reasoning *Int. J. of Approximate Reasoning*, 12, 1995, 133-156.
13. H. Moulin. *Axioms of Cooperative Decision Making.* Cambridge University Press, Cambridge, MA, 1988.
14. Paris J. *The Uncertain Reasoner's Companion.* Cambridge University Press, Cambridge, UK, 1994.
15. Ramer A. Possibilistic information metrics and distances: Characterizations of structure, *Int. J. of General Systems*, 18, 1990, 1-10.
16. Smets P. Constructing the pignistic probability function in a context of uncertainty, *Uncertainty in Artificial Intelligence* 5 (Henrion M. et al., Eds.), North-Holland, Amsterdam, 1990, 29-39.
17. Yager R.R. On the specificity of a possibility distribution, *Fuzzy Sets and Systems*, 50, 1992, 279-292.