Maximum possibility vs. maximum likelihood decisions

Luca Bortolussi[1], Massimo Borelli[1], and Andrea Sgarro[1,2]

[1] Dept. of Mathematics and Computer Science,
University of Trieste,
34100 Trieste, Italia
luca@dmi.units.it, borelli|sgarro@units.it
[2] Consorzio di Biomedicina Molecolare,
Area Science Park, 34100 Trieste.

Consider the following very familiar decision-theoretic situation: a list $\mathcal{L}$ is chosen out of a (finite) input set $\mathcal{X}$, and is communicated to an observer. Further, an input object x, sometimes called a *state of nature*, is chosen inside $\mathcal{L}$. The observer cannot observe directly x, but only a "corrupted version" thereof, y say. He/she makes the following decision: decide for the objects $d(y)$ in the list $\mathcal{L}$ which are "most similar" to what he/she could observe, i.e. to y. Clearly, all this assumes that *similarity measures* $\sigma(x,y)$ are given between input and output objects (between the states of nature and the observables): we shall arrange these measures into a *similarity matrix* Σ with rows headed to $\mathcal{X}$ and columns headed to the (finite) output set $\mathcal{Y}$. The entries of Σ are non-negative real numbers; to avoid trivial situations, at least one entry is strictly positive.

In a coding-theoretic approach, as pursued in [3], the list $\mathcal{L}$ is called the *codebook*, and x and y are the *input codeword* and the *output word*, respectively[3]; then the similarity matrix would describe the *noise* which affects the communication channel. It is in coding theory, and more precisely in possibilistic coding theory and its application to DNA word design [1], that the motivation for this work[4] resides.

Some cases of special matrices follow, which fit into this general frame:

– *Stochastic matrix*: the sum of each row is equal to 1.
– *Joint probability matrix*: the sum of all entries is 1.
– *Possibilistic transition matrix*, or simply *possibility matrix*: the maximum entry in each row is 1.
– *Joint possibility matrix*: the maximum entry in the matrix is 1.

When Σ is a stochastic matrix (then similarities are conditional probabilities), the decision-theoretic principle above is simply *maximum likelihood*, while it is the bayesian

[3] In coding theory, when $|d(y)| \geq 2$ the decoder may either try to guess a single codeword inside $d(y)$, and by so doing increase the probability/possibility of an *undetected error*, or keep $d(y)$ as it is and declare a *detected error*.

[4] What we need in [1] is a communication model which is as unassuming as possible, as we are interested in "negative" results of the type: no noisy channel exists which would justify such and such combinatorial DNA code construction. Since in the sequel we shall concentrate on "singleton events" (elementary events, individual words), rather than compound events (sets of words), we do not even have to specify how we should "aggregate" similarities to obtain similarities between sets of input words and sets of output words. Cf. also the remarks on compound events which conclude this extended abstract.

principle of *maximum posterior probability* in the case of joint probabilities. As for possibility matrices, whose entries are transition possibilities (conditional possibilities), the reader is referred to [4] which deals with a coding-theoretic frame. One may envisage also a "bayesian" possibilistic case, with matrices of joint possibilities whose overall maximum is 1: in this case, each matrix entry is a joint possibility obtained by taking the minimum of the "prior" possibility of the input and the conditional possibility of the output given that input; cf. [2] where the underlying notion of *interactivity* is illustrated.

Assume that Σ is altered to Σ' without changing the orderings between entries. Operationally, nothing would change from the point of view of the decision $d(y)$ made by the observer, whatever the list $\mathcal{L}$, whatever the input object $x \in \mathcal{L}$, and whatever the output object observed y. We shall say in such a case that Σ and Σ' are *equivalent*; an obvious and "limit" case of equivalence is when the two matrices are *proportional*. We shall investigate properties which are *stable* with respect to equivalences. We stress that equivalence concerns only singletons (elementary events) and not an algebra of sets (of compound events).

A problem arises, that of comparing the *representational capacity* or *expressive power* of these approaches, in the sense that on may or may not find equivalent matrices. By just fitting in the suitable proportionality constant, one can prove the following obvious facts: a criterion for a similarity matrix to be equivalent to a possibility matrix is that the maximum similarity in each row is the same; a sufficient but not necessary condition for a similarity matrix to be equivalent to a stochastic matrix is that each row of the similarity matrix sums to the same number. However trivial, we shall stress these facts in the theorem below; in particular, the theorem explains why in the sequel we shall forget about joint probabilities or joint possibilities, and stick instead to similarities: whenever one deals with a similarity matrix, one may well as well think that one is dealing with joint possibilities or joint probabilities, after fitting in the suitable proportionality constant. (Exhibiting possibility matrices which cannot be simulated by means of equivalent stochastic matrices is quite easy; in the lemma below we state a necessary condition.)

Theorem 1. *The representational capacity of similarities, joint possibilities and joint probabilities is the same. The representational capacity of conditional probabilities and transition possibilities are incomparable; both are strictly less than the representational capacity of similarities.*

One may have "odd" similarity matrices, indeed. For example the minimum in row a might be strictly greater than the maximum in row b, which would make the input object b totally "useless". In the sequel, we shall add constraints to the definition of similarity matrices, so as to get rid of "strange" situations, and check how all this shrinks the corresponding representational capacity.

Certain input objects (codewords, states of nature) in a similarity matrix may be "redundant" in the sense of row domination: row a is *dominated* by row b when $a_i \leq b_i$. General similarity matrices or even possibility matrices may freely have domination between their rows, while stochastic matrices have it only in a limit case, since they verify the obvious property: if row a is dominated by row b, then $a = b$. Actually, stochastic matrices verify a stronger ordinal property, which involves domination for rows after *re-ordering* the row entries: in two rows of a similarity matrix there is an *inversion* when, after re-ordering the rows with respect to the non-decreasing order, say, there are two positions i and j with $a_i < b_i$, while $a_j > b_j$. Now, two rows exhibit no inversion iff a permutation of one of the two is dominated by the other.

Lemma 1. *For a similarity matrix to be equivalent to a stochastic matrix, there must be at least one inversion in each couple of rows, apart from couples of rows which are equal up to a permutation of their entries. This condition is also sufficient for two-row matrices.*

(Proof omitted in this extended abstract.) When a matrix satisfies the condition as in the lemma, for convenience' sake we shall say that the matrix is *regular*; we stress that regularity is a topological property which is *stable* with respect to matrix equivalence. The following three-line counter-example shows that this condition is *not* sufficient to have stochasticity up to an equivalence. Take the three-row similarity matrix

$$\begin{matrix} a & a & d & d \\ b & c & c & c \\ a & c & c & d \end{matrix}$$

with $a < b < c < d$; the three rows are already properly ordered. In rows 1 and 2 there is an inversion in positions (columns) 1 and 3, in rows 1 and 3 there is an inversion in positions 2 and 3, while in rows 2 and 3 there is an inversion in positions 1 and 4. However, the linear programming problem which one has to solve (details omitted in this extended abstract) is

$$a < b < c < d \ , \ \ 2a + 2d = 1 \ , \ \ b + 3c = 1 \ , \ \ a + 2c + d = 1 \ ,$$

whose solution set is empty: actually, the last two equations (after replacing $a+d$ by 1/2, cf. the first equation) give $b = c = 1/4$, while one should have $b < c$. Assuming $d < 1$ and adding an all-1 column shows that one can as well start from a possibility matrix.

Theorem 2. *The representational capacity of regular similarities (and so of regular joint possibilities) strictly exceeds that of stochastic matrices. The representational capacity of regular transition possibilities and that of stochastic matrices are not comparable; however, they are the same for two-row matrices.*

All this leaves open the following *open problem*, at least when the number of states of nature is at least 3: *find a simple criterion to ensure that a similarity matrix is equivalent to a stochastic matrix*. Unfortunately, at this point we are only able to provide a sufficient condition which ensures the equivalence, based on a suitable "geometry" of inversions, as will be given in the final version.

Compound events. If one moves from singletons (individual words) to compound events (sets of words), one would have to specify a suitable *aggregator*, which is the sum in the case of probabilities and the maximum for possibilities, and would presumably be an "abstract" aggregator in the general case of similarities. By the way, restricting ourselves to singletons, as we do below, makes it difficult to re-cycle classical results on qualitative probabilities [3], which e.g. require that the intersection of conditioning events is not void, unlike what happens when intersecting distinct singletons. Considering only singletons (elementary events, be they states of nature or codewords), is of no consequence as far as *decoding* (decision making) is concerned, since this depends only on how similarities are ordered in the similarity matrix; however, it does matter when it comes to evaluate the *error* that the decoder might make, which is an additive error of the form $\mathrm{Prob}(E|x)$ in the case of probabilities and a maxitive error of the form $\mathrm{Poss}(E|x)$ in the case of possibilities, with E made up of several[5] output objects (more general aggregators might be used to evaluate the error in the case of similarities). In other words, our concern here is only how decisions are *made*, and not also how decisions should be *evaluated*. If one wants a notion of equivalence such as to be significant also for error evaluation, one should require that the ordering is preserved also for compound events. This is definitely more assuming than above; e.g., it is quite easy to give two-rows examples where an inversion is not enough to have equivalence in this strong sense between a possibilistic and a stochastic matrix. Take the joint possibilities

$$\begin{matrix} a & b & b & c \\ a & a & d & d \end{matrix}$$

with $0 < a < b < c < d = 1$; there is an inversion e.g. in columns 2 and 3. ¿From the second row one has $2a + 2d = 1$, and so $b < d < 1/2$; instead, from the first row one has $a + 2b + c = 1$ and so, after subtracting b, $a + b + c = \mathrm{Prob}(a,b,c) > 1/2 > d$, while $\max(a,b,c) = \mathrm{Poss}(a,b,c) < d$. Add an all-1 column if you want to start from a possibilistic matrix.

References

1. L. Bortolussi and A. Sgarro. Possibilistic channels for dna word design. In *Proceedings of SMPS 2006*, 2006.
2. D. Dubois and H. Prade, editors. *Fundamentals of Fuzzy Sets*. Kluwer, 2000.
3. T. L. Fine. *Theories of Probability*. Academic Press, 1973.
4. A. Sgarro. Possibilistic information theory: a coding-theoretic approach. *Fuzzy Sets and Systems*, 132:11–32, 2002.

[5] By the way, it is a moot point how to define terms of the form $\mathrm{Poss}(y|E)$ in the case of possibilities, let alone in the "abstract" case of similarities.