Bayesian comparators: a probabilistic modeling tool for similarity evaluation between predicted and perceived patterns

Alexandra Steinhilber, S. Valdois, Julien Diard

Bayesian comparators: a probabilistic modeling tool for similarity evaluation between predicted and perceived patterns

Alexandra Steinhilber (alexandra.steinhilber@univ-grenoble-alpes.fr)
Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France

Sylviane Valdois (sylviane.valdois@univ-grenoble-alpes.fr)
Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France

Julien Diard (julien.diard@univ-grenoble-alpes.fr)
Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France

Abstract

A central component of the predictive coding theoretical framework concerns the comparison between predictions and sensory decoding. In the probabilistic setting, this takes the form of assessing the similarity or distance between probability distributions. However, such similarity or distance measures are not associated with explicit probabilistic models, making their assumptions implicit. In this paper, we explore an original variation on probabilistic coherence variables; we define a probabilistic component, that we call a "Bayesian comparator", that mathematically yields a particular similarity measure. A geometrical analogy suggests two variants of this measure. We apply these similarity measures to simulate the comparison of known, predicted patterns to patterns from sensory decoding, first in a simple, illustrative model, and second, in a previous model of visual word recognition. Experimental results suggest that the variant that is scaled by the norms of both predicted and perceived probability distributions yields better robustness and more desirable dynamics.

Keywords: probabilistic modeling; probabilistic similarity; coherence variable; pattern matching; lexical decision

Introduction

In cognitive science, predictive coding has become a major framework, offering an overall theoretical conception of cognitive processes. In this framework, prediction and prediction errors are cornerstones of cognitive architectures: high-level representations would essentially generate predictions, to be sent to lower-level representations in a top-down fashion. Lower-level representations, in turn, would compare the predictions they received with their input from sensory decoding (Rao & Ballard, 1999; Friston, Kilner, & Harrison, 2006). The result of this comparison (i.e., the error signal), if any, would be propagated back to higher-level representations, in a bottom-up manner. Predictive coding provides useful interpretations of information exchange in neuronal architectures (Friston & Kiebel, 2009; Huang & Rao, 2011).

A central component of architectures based on predictive coding, therefore, is comparing between prediction and sensory decoding. In the deterministic setting, both prediction and sensory decoding would be "point-like", that is to say, values in some representational space (e.g., a time-interval is predicted to be 2 s long, whereas it is perceived as being 2.5 s). In this case, a straightforward candidate would be to compute the difference, or the length of the difference vector, in the multidimensional case, between prediction and sensory decoding. For instance, such a measure is used throughout artificial neural networks-based approaches, to compute error signals during learning.

In the probabilistic framework, one would compare probability distributions instead. In the free-energy principle framework, for instance, under some Gaussian assumptions about noise, the prediction error would take the form of a precision-weighted difference between prediction mean and signal mean (Friston, 2010). In more general settings, a widespread measure is based on the Kullback-Leibler (KL) divergence (Bishop, 2006), or its symmetrized variant. In some contexts, the KL divergence is "theoretically justified". For instance, in variational inference, it appears in mathematical derivations for computing log-marginal likelihoods (Neal & Hinton, 1998; Girin et al., 2021). However, this does not imply that the precision-weighted differences between means, or the distance measure based on the KL divergence, would be "theoretically evident" in all contexts. Indeed, many distance measures between probability distributions have been proposed (Cha, 2007), each with specific properties (as is the case for distance measures in general).

To the best of our knowledge, the more common practice seems to mainly focus on building probabilistic models to compute probability distributions, both for encoding predictions and sensory decoding, and then, select a distance measure, from the wide array of existing distance measures. In that sense, selecting the distance measure "comes after the fact", and it is not part of the probabilistic model *per se*. Therefore, possible assumptions that accompany the choice of one measure over another are neither made explicit, nor represented in the model itself.

In this paper, we propose to establish a link between a probabilistic model and a specific similarity measure between probability distributions. (Note that similarity and distance measures are conceptually equivalent, and easily linked mathematically, e.g., with a reciprocal, $1/x$ relationship). More precisely, we explore a previously defined probabilistic model, based on coherence variables (Gilet, Diard, & Bessière, 2011; Bessière, Mazer, Ahuactzin, & Mekhnacha, 2013), that yields, thanks to Bayesian inference, a particular similarity measure between probability distributions. This measure is thus theoretically derived from the probabilistic model, from the rules of Bayesian inference. We call "Bayesian comparator" the probabilistic model, when it is used in such a fashion.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Coherence variables can be used and interpreted as "Bayesian switches", that is to say, they allow explicitly connecting or disconnecting portions of models during inference (Gilet et al., 2011), or reasoning with soft evidence (Bessière et al., 2013). In this paper, we explore the mathematical properties of coherence variables in a novel context, involving probabilistic computations that were not considered before.

In the following, we first provide the mathematical definition of Bayesian comparators, and demonstrate that they yield an inner product expression. Interpreting this geometrically suggests variants, which we define. Then, we illustrate Bayesian comparators on a small, abstract model, to evaluate how a perceived pattern is similar to memorized patterns. Finally, we describe how Bayesian comparators have been applied in BRAID, a Bayesian word recognition model (Phénix, 2018; Ginestet, Phénix, Diard, & Valdois, 2019), to assess stimulus familiarity, yielding novel models of lexical decision and of novelty detection in the context of orthographic learning. We experimentally evaluate the properties of the Bayesian comparator and its variants, in the context of gradual accumulation of perceptual evidence.

Model

Here, we define the base case model of Bayesian comparators. Let A and B be any two probabilistic variables, that share the same discrete, finite domain $\mathcal{D}$. Variable λ is said to be a *coherence variable* (Bessière et al., 2013) if it is binary (domain $\{0,1\}$), and associated to a conditional probability distribution defined by:

$$P([\lambda = 1] \mid [A = a]\ [B = b]) = \begin{cases} 1 \text{ if } a = b \\ 0 \text{ otherwise .} \end{cases} \quad (1)$$

There are no constraints on the rest of the probabilistic model, although it is simpler to examine the case where the probability distributions over A and B are independent. Therefore, as our base case, we consider the model:

$$P(A\ B\ \lambda) = P(A)P(B)P(\lambda \mid A\ B)\ . \quad (2)$$

In previous works, coherence variables have been interpreted as "Bayesian switches" (Gilet et al., 2011), that are either "closed" when assumed to be equal to 1 (i.e., computing $P(A \mid [\lambda = 1])$ involves $P(B)$), or "open" when their value is left unspecified (i.e., computing $P(A)$ does not involve $P(B)$).

Here instead, we consider computing the probability distribution over the coherence variable λ itself. In the base case model of Eq. (2), computing $P([\lambda = 1])$ yields:

$$\begin{aligned} P([\lambda = 1]) &= \sum_{A,B} P(A\ B\ \lambda) \\ &= \sum_{A,B} P(A)P(B)P([\lambda = 1] \mid A\ B) \\ P([\lambda = 1]) &= \sum_{a \in \mathcal{D}} P([A = a])P([B = a])\ . \end{aligned}$$

The first line results from the marginalization rule, the second rewrites the joint probability distribution according to Eq. (2) and the third recognizes that, in the joint summation over the domains of A and B, that is, over the square domain $\mathcal{D}^2$, only the diagonal remains because $P([\lambda = 1] \mid A\ B)$ is 0 otherwise.

The last expression can be rewritten, using the notation of the inner product between $P(A)$ and $P(B)$. We note this with P_{inner} and obtain:

$$P_{inner}([\lambda = 1]) = \langle P(A), P(B) \rangle\ . \quad (3)$$

Since λ is a binary variable, we of course also have:

$$P_{inner}([\lambda = 0]) = 1 - \langle P(A), P(B) \rangle\ .$$

This suggests a geometrical interpretation, in which we consider probability distributions as vectors. Indeed, the set of probability distributions defined on domain $\mathcal{D}$, of cardinal $n \in \mathbb{N}$, is defined by $\mathcal{P} = \{\mathbf{p} \in \mathbb{R}^n, \sum_{i=1}^{n} p_i = 1\}$. In other words, in the discrete case, probability distributions can be seen as vectors of positive values, that are normalized in the 1-norm sense: $\|p\|_1 = \sum_i p_i = 1$ (making the 1-norm explicit here; everywhere else, $\|\cdot\|$ refers to the 2-norm). However, the 2-norm length of a probability distribution, seen as a vector, is not constant. Consider for instance an arbitrary probability distribution over variable X with binary domain $\{0,1\}$; it is entirely defined by a single parameter: $P([X = 1]) = p_X$, $P([X = 0]) = 1 - p_X$. Then, its 2-norm length is $\|P(X)\| = p_X^2 + (1 - p_X)^2 = 1 - 2p_X + 2p_X^2$, that is, a quadratic function of p_X. This expression behaves in a manner similar to the entropy measure: it varies continuously between its different extremum values, that are for the Dirac Delta and Uniform distributions. (We apply loose terminology and refer to the $\{0,1\}$ and $\{1,0\}$ distributions over binary domains as Dirac Delta in the following.) In other words, the 2-norm and entropy are both measures that are sensitive to the uncertainty of probability distributions.

In the usual Euclidean geometrical context, the inner product between vectors is the product of the cosine of their angle and of their lengths (in the 2-norm sense). This suggests two variants of similarity measure P_{inner}. In the first variant, called P_{cos}, the inner product P_{inner} is scaled by the product of 2-norm lengths of the considered distributions:

$$P_{cos}([\lambda = 1]) = \frac{\langle P(A), P(B) \rangle}{\|P(A)\| * \|P(B)\|}\ , \quad (4)$$

(We refrain from noting 2-norm lengths as inner products of distributions with themselves to avoid confusion, and reserve below the inner product notation for application between different probability distributions.) This equation can be interpreted as computing the "cosine of the angle between probability distributions $P(A)$ and $P(B)$".

In the second variant, called P_{proj}, the inner product is only scaled by the 2-norm of one of the distributions. Here, the inspiration is the computation of the length of the projection of one vector onto the other. This yields a measure related to the "compatibility" of one vector with a reference vector: projecting vector A onto B amounts to removing, from A, its

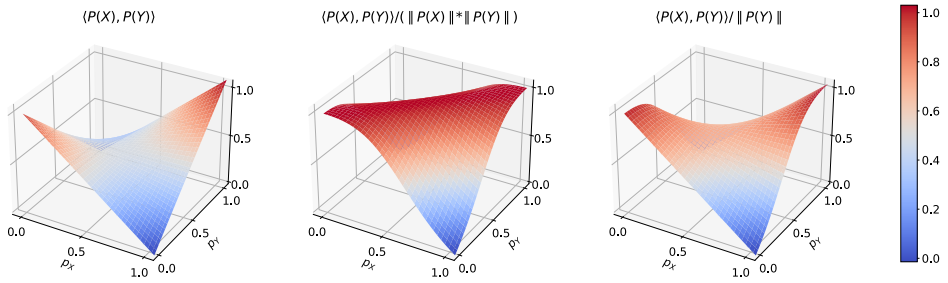

Figure 1: Three variants of a similarity measure between probability distributions over binary domains, parametrized by p_X (x-axis) and p_Y (y-axis). Color coding is redundant with the value, on the z-axis, of the similarity measure. Left: similarity measure P_{inner} of Eq. (3); Middle: similarity measure P_{cos} of Eq. (4); Right: similarity measure P_{proj} of Eq. (5).

portion that is orthogonal to B. When this "orthogonal component" is small, the length of the projection of A onto B is large. We define:

$$P_{proj}([\lambda = 1]) = \frac{\langle P(A), P(B)\rangle}{\|P(B)\|} \ , \tag{5}$$

which can be interpreted as computing the "length of the projection of distribution $P(A)$ onto distribution $P(B)$".

The P_{inner} measure is referred to as the cosine similarity measure (Cha, 2007). It is commonly applied in natural language processing models, such as, for instance, in the word2vec model of word semantics (Mikolov, Chen, Corrado, & Dean, 2013); however here it cannot be negative, since, in the probabilistic setting, probability values are positive. Additionally, in the context of decision theory, the P_{proj} measure is called the spherical scoring rule (Jose, 2009).

Figure 1 illustrates the three similarity measures, in the probabilistic setting. We consider any two distributions $P(X)$ and $P(Y)$ over binary variables, parametrized by p_X and p_Y, respectively, and compute and show P_{inner}, P_{cos} and P_{proj} in the left, middle and right plots of Figure 1, respectively.

Eq. (3) and P_{inner} yield a saddle shape: P_{inner} is minimal and equal to 0 when comparing "opposite" Dirac Delta distributions ($p_X = 0, p_Y = 1$ and vice versa); it is maximal and equal to 1 when comparing identical Dirac Delta distributions ($p_X = p_Y = 1$ and $p_X = p_Y = 0$), and it is locally maximum when comparing identical distributions ($p_X = p_Y$).

In contrast, Eq. (4) and measure P_{cos} also yield a saddle shape, but with a constant maximal value of 1 whenever $p_X = p_Y$. Indeed, thanks to the scaling by the norms of the probability distributions, it is only sensitive to the angle between distributions, and thus it is maximal whenever they are identical (their angle is 0 so that their cosine is 1). On the contrary, the measure is 0 when they are maximally different (they are orthogonal, their cosine is 0). Finally, we observe that Eq. (5) and measure P_{proj} also yield a saddle shape, with a more complicated trajectory for its maximal manifold.

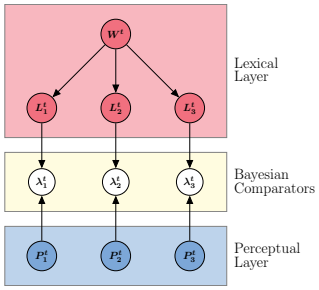

Figure 2: Graphical representation of the dependency structure of the pattern matching model. Nodes represent variables, and arrows dependencies between variables, following the classical convention for graphical models.

Application for similarity evaluation between memorized and perceived patterns

We first apply Bayesian comparators, in a very simple model, illustrated Figure 2. We consider an arbitrary, discrete and finite domain in which to represent perceived and predicted patterns. We note $\mathcal{D} = \{\text{“}a\text{”}, \text{“}b\text{”}, \text{“}c\text{”}, \text{“}d\text{”}\}$ this domain. In the perceptual layer of the model, we define three probabilistic variables, P_1^t to P_3^t, to represent probability distributions over perceived inputs of length 3. These distributions vary over time: at time instant 0, they are uniform, and evolve so that probability accumulates in favor of one value of domain $\mathcal{D}$. The probability of this value follows a sigmoid function of time t, to mimic what the mathematics of a temporal model of perceptual accumulation of sensory evidence would yield (Phénix, 2018). Therefore, in the perceptual layer, we can simulate the model perceiving, from sensory input, any pattern in $\mathcal{D}^3$ (e.g., "abc", "abb", etc.).

In the "lexical" layer of the model, we define predicted patterns with a naïve Bayes model (Russell & Norvig, 1995; Norris, 2006). Variables L_1^t to L_3^t have domain $\mathcal{D}$, and variable W^t has a discrete, finite domain $\mathcal{D}_W$ to index known patterns. We assume that prior distribution $P(W^t)$ is uniform. Probability distributions $P(L_i^t \mid [W^t = w])$ define the pattern

for word w ($i \in \{1,2,3\}$), by assigning a very high probability (0.91) to a point of $\mathcal{D}$ (i.e., the correct letter for word w at position i), and uniformly distributing probability on alternatives (0.03). In other words, $P(L_i^t \mid W^t)$ are "quasi-Dirac Delta" distributions. For instance, if a known word w_1 is "abc", $P([L_1^t = \text{“}a\text{”}] \mid [W^t = w_1])$ is 0.91, $P([L_1^t = \text{“}b\text{”}] \mid [W^t = w_1])$ is 0.03, and so on.

The perceptual and lexical layers of the model are connected by three Bayesian comparators. Therefore, the whole model is defined by:

$$P(W^t\ L_{1:3}^t\ \lambda_{1:3}^t\ P_{1:3}^t) = P(W^t)\prod_{i=1}^{3} P(L_i^t \mid W^t) P(\lambda_i^t \mid L_i^t P_i^t) P(P_i^t)\ ,$$

with $X_{1:3}^t$ referring to the set of three variables, X_1^t, X_2^t and X_3^t; in other words the model is 3*3+1=10 dimensional.

In this model, pattern recognition, that is, computing the probability distribution over known three-letter words given a sensory input, can be performed by assuming that coherence variables $\lambda_{1:3}^t$ are equal to 1 and computing $P(W^t \mid P_{1:3}^t\ [\lambda_{1:3}^t = 1])$. Bayesian inference yields:

$$\begin{aligned} P(W^t \mid P_{1:3}^t\ [\lambda_{1:3}^t = 1]) &= \frac{P(W^t\ P_{1:3}^t\ [\lambda_{1:3}^t = 1])}{P(P_{1:3}^t\ [\lambda_{1:3}^t = 1])} \\ &\propto \sum_{L_{1:3}^t} P(W^t\ L_{1:3}^t\ [\lambda_{1:3}^t = 1]\ P_{1:3}^t) \\ &\propto P(W^t)\prod_{i=1}^{3} \left\langle P(L_i^t \mid W^t), P(P_i^t) \right\rangle\ . \end{aligned}$$

In this derivation, the $\propto$ symbol indicates equality up to a proportional constant (indeed, the denominator can be considered a constant, and the result can be re-normalized afterwards). In this computation, setting coherence variables to 1 can be interpreted as assuming that the perceived pattern corresponds to a known pattern. This allows collapsing the summation over all possible letters $L_{1:3}^t$: since coherence variables are 1, the only non-zero value inside the sum is when the considered value for $L_{1:3}^t$ is the same as for $P_{1:3}^t$, which is a given value. The resulting computation assigns highest probability to the known pattern that most resembles the perceived input.

Bayesian comparators yield another inference, to assess whether the assumption that the input pattern corresponds to a known one is true. Indeed, following Eq. (3), we compute the probability that the three coherence variables are 1:

$$P_{inner}([\lambda_{1:3}^t = 1]) = \sum_{W^t} \left(P(W^t) \prod_{i=1}^{3} \left\langle P(L_i^t \mid W^t), P(P_i^t) \right\rangle \right)\ . \quad (6)$$

Then, $1 - P_{inner}([\lambda_{1:3}^t = 1])$ is the probability that all coherence variables are not simultaneously equal to 1, meaning that the input pattern does not correspond to any known pattern: there is an error, in at least one position, between the perceived pattern and any known pattern. Eq. (6) thus provides the basis for models of familiarity assessment (or novelty detection), in the probabilistic framework, based on similarity computations given by Bayesian comparators.

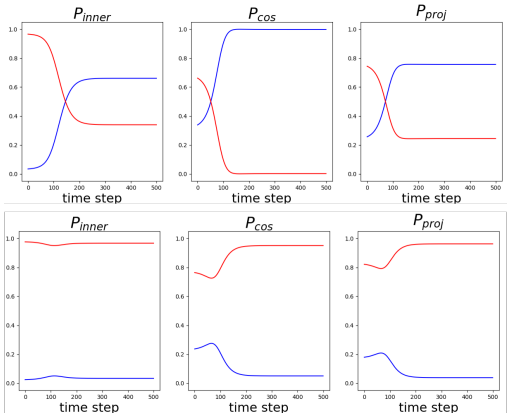

Figure 3: Familiarity assessment simulation: probability (y-axis) that the stimulus is a known pattern (blue curves), or not a known pattern (red curves), as a function of simulated time (x-axis). Top row: the perceived pattern is "abc", which is a known word. Bottom row: the perceived pattern is "abd", which is not a known word. Left (resp. middle, right) column features the P_{inner} (resp., P_{cos}, P_{proj}) measure.

Following Eqs. (4) and (5), we further consider two variants, depending on whether inner products are scaled by the norms of one, or both of the probabilities involved:

$$P_{cos}([\lambda_{1:3}^t = 1]) = \frac{P_{inner}([\lambda_{1:3}^t = 1])}{\|P(L_i^t \mid W^t)\| * \|P(P_i^t)\|} \quad (7)$$

$$P_{proj}([\lambda_{1:3}^t = 1]) = \frac{P_{inner}([\lambda_{1:3}^t = 1])}{\|P(P_i^t)\|}\ . \quad (8)$$

Therefore, in P_{proj}, we consider the projection of the predicted pattern onto the perceived pattern.

We simulate familiarity assessment in the model of Figure 2, with the three variants provided by Eqs. (6–8), first, for the perceived pattern "abc" that corresponds to a known word, then for the pattern "abd" that does not correspond to a known word. Results are shown in Figure 3.

We observe that, for all similarity measures, familiarity assessment based on Bayesian comparators performs as expected: the probability that the λ variables are 1 is high when the perceived pattern is a known one (Figure 3, top row), and low when the perceived pattern is a novel one (Figure 3, bottom row). In the case where the perceived pattern is a known one, similarity measures "are wrong" during the first few time steps, in the sense that they are in favor of the perceived pattern being a novel one. This results from the comparison between a quasi-Dirac Delta predicted probability distribution and an almost uniform perceived distribution. The three proposed measures are affected differently, since they are mathematically scaled differently by the 2-norm of distributions. The P_{cos} measure appears as the most robust in this regard (Figure 3, top row, middle plot), at the cost of a slower convergence for the opposite situation, in which the perceived pattern is novel (Figure 3, bottom row, middle plot).

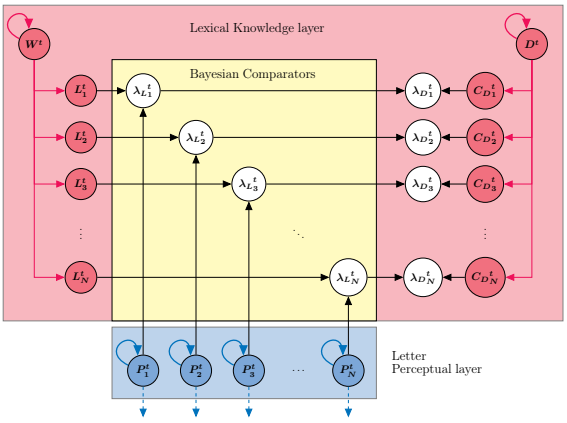

Figure 4: Graphical representation of a portion of the BRAID model. Graphical convention is the same as in Figure 2. Dashed arrows indicate that P_i^t variables connect to other layers (visuo-attentional and sensory) of the complete model.

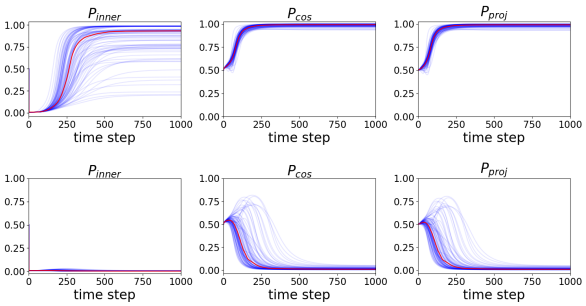

Figure 5: Familiarity assessment simulation in the BRAID model: evolution of probability (y-axis) that the stimulus is a known pattern, as a function of simulated time (x-axis). Each blue curve corresponds to one stimulus, and the red curve is the median curve across all stimuli. (The probability curves that the input would not be a word are not shown; they would be "1 minus the blue curves".) Top row: stimuli are known words. Bottom row: stimuli are non-words. Left (resp. middle, right) column: familiarity assessment according to the P_{inner} (resp., P_{cos}, P_{proj}) measure.

Application in the BRAID model

The BRAID model (*Bayesian word Recognition with Attention, Interference and Dynamics*) is a probabilistic and hierarchical model to simulate word recognition and lexical decision (Phénix, 2018; Phénix, Valdois, & Diard, 2018; Ginestet et al., 2019). The model's architecture contains a letter perceptual layer, where accumulation of sensory evidence about letters in the visual stimulus occurs and a lexical knowledge layer, where sequences of letters are associated with known words. These two layers are connected by a layer of Bayesian comparators. The dependency structure of this portion of the BRAID model is shown Figure 4. The complete model also features a letter sensory layer, where the stimulus letter-sequence is processed, and a visuo-attentional layer, that controls how much sensory information is transferred and accumulated in the perceptual layer (not shown in Figure 4). These layers also represent gaze position, the spatial distribution of visual attention, and their effects on sensory processing; this is beyond the scope of the current paper.

The BRAID model we consider in Figure 4 is similar to the "simple" model of Figure 2, with a few differences. First, in BRAID, possible letters are the 26 letters of the Latin alphabet, and letter sequences and words are of length N (so that variables are indexed from position 1 to N in subscript). Second, an explicit model of "known similarity patterns" is expressed with binary variables C_{D1}^t to C_{DN}^t and Boolean variable D^t: when D^t is *True*, all variables $C_{D1:N}^t$ are expected to be 1 (Bayesian comparators match in all positions), whereas when D^t is *False*, one of the variables $C_{D1:N}^t$ is expected to be 0 (one of the Bayesian comparators does not match). These "similarity patterns" are then connected to the Bayesian comparators, to perform familiarity assessment and novel detection proper. Finally, variables W^t and D^t of BRAID are integrated in a Markov chain-like model, to perform temporal integration of perceptual evidence: this yields dynamically evolving probability of words in $P(W^t)$, to simulate word recognition, and dynamically evolving probability that the stimulus is a known word in $P(D^t)$, to simulate both the lexical decision task (i.e., deciding whether the input is a known word or not) and orthographic learning (i.e., create and update orthographic representations in the word space $\mathcal{D}_W$; (Ginestet, Valdois, & Diard, 2022)).

In the BRAID model, we simulated familiarity assessment on a set of 5-letter words and non-words. All model parameters were set to their default values; notably, gaze and the focus of visual attention were positioned on the central letter, and the lexical knowledge layer was configured with the 79,673 English words from the British Lexicon Project (BLP) (Keuleers, Lacey, Rastle, & Brysbaert, 2012). We randomly selected 100 5-letter words from the BLP to serve as word stimuli. From another set of 100 5-letter words randomly drawn from the BLP, word-like non-words were generated using Wuggy (Keuleers & Brysbaert, 2010). Example words are "sheet", "clock", "brush"; example non-words are "ropat", "loors", "squay". Simulations were carried out for 1,000 time steps.

Experimental results are shown Figure 5. We first observe that all variants, almost always, successfully perform familiarity assessment: when the stimulus is a known word, the model assigns to the fact that it is a word a probability larger than .5. This is the case except for less than 10 words for the P_{inner} variant. The P_{inner} measure also yields, for the first time steps, dynamics that are worth noting: at time step 0, $P(D^0)$ is uniform, but, right from time step 1, the probability that the input is a known word becomes very low. This yields very fast convergence in the case where the stimulus is

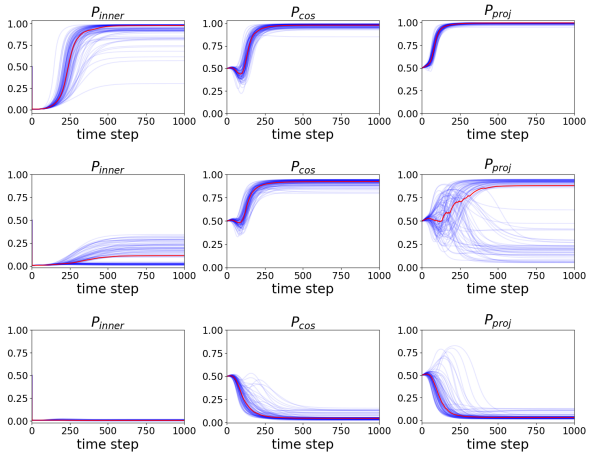

Figure 6: Familiarity assessment simulation in the BRAID model, for "perfectly known" words (top row), "not very-well known words" (middle row) and non-words (bottom row). Information is graphically represented in the same manner as in Figure 5, please refer to its caption.

novel (Figure 5, bottom left plot); however, this slows down, and possibly terminally impairs, recognition that the stimulus is known when indeed it is a word (Figure 5, top left plot). This initial behavior is not observed for the other two variants, P_{cos} and P_{proj}, which rapidly and successfully recognize, for all stimuli, whether they are known or not. Overall, these variants appear to have more stable dynamics, especially for the initial portion where the distributions on perceived letter are still close to uniform. This is mathematically clear, since they both are scaled by the 2-norms of probability distributions on perceived letters.

However, P_{cos} and P_{proj} also differ by whether or not they involve the 2-norms of the probability distributions of the known words. To explore this, we have conducted an additional experiment, in which we manipulated the quality lexical representations. In the previous experiments, words were either known, with quasi-Dirac Delta distributions, or unknown (not in $\mathcal{D}_W$). Here, we add a set of "not very-well known" words, that is to say, with more uncertain lexical representations ($P(L_i^t \mid W^t)$ is 0.48 for the correct letter, and 0.02 for all 26 alternatives). Two new sets of 5-letter words were randomly selected from the BLP, providing 100 "perfectly known" words and 100 "not very-well known" words; non-words were the same as in the previous experiment.

Simulation results are shown Figure 6. Overall, we observe that the P_{inner} variant has difficulty recognizing words associated with uncertain distributions as being words (Figure 6, middle row, left plot). Indeed, at the end of simulation, when an input pattern is very well perceived, and thus of very low uncertainty, it is considered as not matching the predicted pattern for the corresponding word, since this is of higher uncertainty (even though they match with respect to the letters indexed by the peaks of probability distributions). Measures that instead correct for the uncertainty of compared distributions do not feature this issue, with P_{cos} being the more robust in this regard (Figure 6, middle plot).

In this experiment, the behaviors of the P_{cos} and P_{proj} variants are different. The P_{cos} variant appears as the more robust, recognizing almost equally well words, independently of the quality of their probabilistic representations in the lexicon.

Discussion

In this paper, we have proposed an original use of coherence variables, that we call "Bayesian comparators", to define a new class of similarity operators in the probabilistic framework. We have shown how this yields a model of familiarity assessment based on the similarity between perceived and predicted probability distributions. This model has also been applied in a model of word recognition, where Bayesian comparators assess familiarity to detect whether a stimulus corresponds to a known word or not. Experimental results suggest that all proposed variants perform successfully; however, the P_{cos} measure appears to yield the more desirable dynamics and performance for familiarity assessment, overall.

Throughout this paper, we have defined and experimentally illustrated three variants of the similarity measure provided by Bayesian comparators. Our main goal was to anchor similarity measures in probabilistic models, in order to make explicit assumptions that could underlie the measures. We have shown that the inner product based similarity measure P_{inner}, and thus familiarity assessment, could be interpreted as evaluating the probability that coherence variable λ was equal to 1. An open issue remains, to anchor the two variants P_{cos} and P_{proj} theoretically, in the same manner. Another issue concerns the relation between the similarity measure and its use in the predictive coding framework. Indeed, predictive coding assumes a precise temporal organization between predictions, error computation and error propagation, whereas in this paper, we have assumed that all these components would happen at all time steps. This makes our model compatible with predictive coding at an algorithmic level, although it is not an implementation level model of predictive coding.

The probabilistic similarity measures that we have defined suggest intriguing relations with other domains. We observe that they only differ for non-Dirac Delta distributions. Indeed, Figure 1 shows that they all consider that identical Dirac Delta distributions are maximally similar, and different Dirac Delta distributions are maximally different. In other words, our similarity measures can be seen as probabilistic extensions of the logical XNOR operator (the exclusive NOR, True if and only if its two inputs are both True or both False). This suggests that coherence variables implicitly involve a probabilistic extension of one of the core logical operators. Indeed, their definition, in Eq. (1), involves an equality constraint between the values of the variables they connect. Whether other probabilistic constructs, or variations on coherence variables, extend other logical operators is an open issue.

Acknowledgments

This work was supported by a French Ministry of Research (MESR) Ph.D. grant to AS.

References

Bessière, P., Mazer, E., Ahuactzin, J. M., & Mekhnacha, K. (2013). *Bayesian programming*. Boca Raton, Florida: CRC Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY, USA: Springer.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *4*(1), 300–307.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Phil. Trans. R. Soc. B*, *364*, 1211–1221.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology – Paris*, *100*, 70–87.

Gilet, E., Diard, J., & Bessière, P. (2011). Bayesian action-perception computational model: Interaction of production and recognition of cursive letters. *PLoS ONE*, *6*(6), e20387.

Ginestet, E., Phénix, T., Diard, J., & Valdois, S. (2019). Modeling the length effect for words in lexical decision: The role of visual attention. *Vision Research*, *159*, 10–20.

Ginestet, E., Valdois, S., & Diard, J. (2022). Probabilistic modeling of orthographic learning based on visuo-attentional dynamics. *Psychonomic Bulletin & Review*.

Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., & Alameda-Pineda, X. (2021). Dynamical variational autoencoders: A comprehensive review. *Foundations and Trends in Machine Learning*, *15*(1–2), 1–175.

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *WIREs Cognitive Science*, *2*, 580–593.

Jose, V. R. (2009). A characterization for the spherical scoring rule. *Theory and Decision*, *66*, 263–281.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. doi: 10.3758/BRM.42.3.627

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. doi: 10.3758/s13428-011-0118-4

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models* (pp. 355–368). Springer.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Phénix, T. (2018). *Modélisation bayésienne algorithmique de la reconnaissance visuelle de mots et de l'attention visuelle*. Unpublished doctoral dissertation, Univ. Grenoble Alpes.

Phénix, T., Valdois, S., & Diard, J. (2018). Reconciling opposite neighborhood frequency effects in lexical decision: Evidence from a novel probabilistic model of visual word recognition. In T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 2238–2243). Austin, TX: Cognitive Science Society.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs, New Jersey: Prentice Hall Series in Artificial Intelligence.