

Common bayesian models for common cognitive issues

Francis Colas, Julien Diard, Pierre Bessiere

▶ **To cite this version:**

Francis Colas, Julien Diard, Pierre Bessiere. Common bayesian models for common cognitive issues. Acta Biotheoretica, 2010, 58 (2-3), pp.191-216. 10.1007/s10441-010-9101-1 . hal-00530356

HAL Id: hal-00530356
https://hal.science/hal-00530356

Submitted on 28 Oct 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Noname manuscript No.
(will be inserted by the editor)

Common Bayesian Models For Common Cognitive Issues

Francis Colas, Julien Diard, and Pierre Bessière

the date of receipt and acceptance should be inserted later

Abstract How can an incomplete and uncertain model of the environment be used to perceive, infer, decide and act efficiently? This is the challenge that both living and artificial cognitive systems have to face. Symbolic logic is, by its nature, unable to deal with this question. The subjectivist approach to probability is an extension to logic that is designed specifically to face this challenge. In this paper, we review a number of frequently encountered cognitive issues and cast them into a common Bayesian formalism. The concepts we review are ambiguities, fusion, multimodality, conflicts, modularity, hierarchies and loops. First, each of these concepts is introduced briefly using some examples from the neuroscience, psychophysics or robotics literature. Then, the concept is formalized using a template Bayesian model. The assumptions and common features of these models, as well as their major differences, are outlined and discussed.

1 Introduction

It is remarkable that a wide variety of common cognitive issues – in the sense that they appear frequently – can be tackled by so small a set of common models – in the sense that these models are shared by these issues. In other words, a few template mathematical constructs, based only on probabilities and Bayes' rule, can be applied to a large assortment of problems that have to be addressed by cognitive systems.

Our purpose, in this paper, is to demonstrate these assertions, by proposing a step-by-step inspection of these cognitive problems, and, for each of them, by describing a candidate Bayesian model. The cognitive issues we cover are ambiguities, fusion, multimodality, conflicts, modularity, hierarchies and loops. Of course, each of these is a large topic, covering several research domains, from neuroscience and experimental psychology to modeling and robotics. Each of these regularly warrant full scale reviews of the relevant literature. We of course do not presume to cover such an extensive body of work here; instead, we only reference a few key papers as illustrations for each model. This larger perspective allows, we believe, to better grasp the big picture: it

Address(es) of author(s) should be given

will appear that these issues and associated models are solved by a graduation of very similar models, each only marginally more complex than the previous.

1.1 Probability as an extension to logic

Both living creatures and artificial cognitive systems have to face the same fundamental difficulty; that is, *how to use an incomplete and uncertain model of their environment to perceive, infer, decide and act efficiently.*

Indeed, any model of a real phenomenon is *incomplete*. Hidden variables, not taken into account in the model, influence the phenomenon. The effect of these hidden variables is that the model and the phenomenon never behave exactly alike. *Uncertainty* is the direct and unavoidable consequence of incompleteness. No model can foresee exactly the future observations of a phenomenon, as these observations are biased by the hidden variables. No model can also provide an exact prediction of the consequences of its decisions.

Logic is the foundation paradigm for rational reasoning. As such, logic is the core of mathematics, scientific methodology, computer science and technology, and is possibly the only common denominator of the various current theories of brain and cognition.

Nevertheless, by its very nature, symbolic logic cannot deal with incomplete and uncertain information. Symbolic logic can only be used to manipulate knowledge that is certain to be either true or false. Logical inference is impossible as soon as part of the required information is missing.

Living organisms must make sensorimotor decisions with quite incomplete knowledge of their environment. However, the creatures that we can observe survive every day as individuals and have survived millions of years as species. This single fact proves that adequate (even if not optimal) sensorimotor decisions can be made with incompleteness and uncertainty.

1.2 The subjectivist approach to probability

The *subjectivist* approach to probability proposes probability theory as an extension to symbolic logic for rational reasoning in the presence of incompleteness and uncertainty (Jaynes, 2003). It is sometimes partially improperly called the "Bayesian approach": indeed, this now ubiquitous expression sometimes refers to different notions in different scientific communities.

The subjectivist approach deals with incompleteness and uncertainty using a two-step process, *learning* and *inference* (see Figure 1).

Learning transforms irreducible incompleteness into quantified uncertainty (*i.e.,* probability distributions). These distributions result from both *preliminary knowledge* of the reasoning subject and *experimental data* obtained by observation of the phenomenon.

Preliminary knowledge, even when it is imperfect and incomplete, is relevant and provides interesting hints about the observed phenomenon. Preliminary knowledge is not a fixed and rigid model purporting completeness. Rather, it is a gauge, with free parameters, waiting to be moulded by experimental data.

Experimental data obtained from physical interaction with the environment reflect all of the complexity of this interaction. This includes the effect of hidden variables that

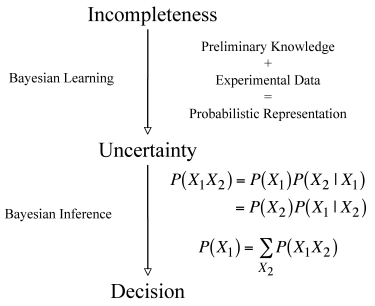


Fig. 1 The subjectivist approach to incompleteness.

are not taken into account by preliminary knowledge. Using the experimental data, the learning process sets the values of the free parameters of the preliminary knowledge. Thus, the influence of the hidden variables is taken into account and quantified.

Inference is performed using the probability distributions obtained in the first step of the procedure.

To do so, we only require the two basic rules of Bayesian inference, the product rule[1] (equation 1) and the marginalization or sum rule (equation 2).

$$P(X_1\ X_2) = P(X_1)P(X_2 \mid X_1) = P(X_2)P(X_1 \mid X_2) \tag{1}$$

$$P(X_1) = \sum_{X_2} P(X_1\ X_2) \tag{2}$$

These two rules are to Bayesian inference what the resolution principle is to logical reasoning (Robinson, 1979). They are sufficient to perform any inference on discrete probability distributions. These inferences may be as complex and subtle as those achieved with logical inference tools. Indeed, it can be shown that logical inference is a particular case of probabilistic inference (Jaynes, 2003).

Learning itself can be described as a probabilistic inference problem in which the values of the learned parameters are obtained by maximizing their probability based on knowing both the preliminary knowledge and the data.

1.3 The Bayesian Programming formalism

The necessary information for completely specifying a probabilistic model consists of: (1) the relevant variables, (2) the conditional dependencies between these variables, (3) the parametric forms of the associated probability distributions, and (4) their associated parameter values.

The usual way of representing such models in the literature uses different graphical forms (see Bayesian networks described by Pearl (1988) and Jensen (1996), or graphical models as presented by Frey (1998) and Jordan (1999)). The relevant variables are nodes, their conditional dependencies represented by vertices between these nodes, and the parametric forms associated with each node.

[1] From which we derive directly Bayes' theorem, which, strictly speaking, reads $P(X_2 \mid X_1) = P(X_1 \mid X_2)P(X_2)/P(X_1)$, provided $P(X_1) \neq 0$. Most of the time though, both names and equations can be used interchangeably.

As probability is an extension to symbolic logic, it is also possible to use an algebraic formalism to define probabilistic models. We proposed such a formalism called *Bayesian Programming* Bessière et al. (2003); Lebeltel et al. (2004); Bessière et al. (2008).

A Bayesian Program (BP) contains two parts:

- a *description* that is the probabilistic model of the studied phenomenon or the programmed behaviour; and
- a *question* that specifies an inference problem to be solved using this model.

A description contains two parts:

- a *specification* part that formalizes the modeler's knowledge ; and
- an *identification* part where free parameters are learned from experimental data.

Finally, the specification is constructed from three parts:

- a selection of *relevant variables* to model the phenomenon;
- a *decomposition*, whereby the joint distribution on the relevant variables is expressed as a product of simpler distributions exploiting conditional independence between variables; and
- the *parametric forms* in which either a given mathematical function or a question for another BP is associated with each of the distributions appearing in the decomposition.

In this paper, we focus mostly on the variables and the decomposition part of this formalism, in which the joint probability distribution is specified as a product of simpler distributions. We will omit the parametric forms and the distributions, which are often problem dependent, in the sense that they are closely related to the problem at hand (whether the system studied is sensory, motor, etc.).

For instance, if we have a model with three variables, M (the motor commands for speech), G (the geometry of the articulatory apparatus) and F (the produced sound characterized by its formants), we can assume a conditional independence between F and M when G is known. Indeed, knowledge of the shape of the vocal tract provides sufficient information to compute the formants. The joint distribution, $P(M\ G\ F)$, is then decomposed as:

$$P(M\ G\ F) = P(M)P(G \mid M)P(F \mid G).$$

When the decomposition of a joint distribution is available, the answer to any question you might want to ask of the model may be computed.

For instance, the answer to the question, $P(M \mid F)$, which searches for the motor commands to produce a given sound, is computed as:

$$P(M \mid F) = \frac{P(M\ F)}{P(F)} \tag{3}$$

$$= \frac{\sum_G P(M\ G\ F)}{\sum_{G,F} P(M\ G\ F)} \tag{4}$$

$$= \frac{\sum_G P(M)P(G \mid M)P(F \mid G)}{\sum_{G,F} P(M)P(G \mid M)P(F \mid G)} \tag{5}$$

$$P(M \mid F) = \frac{1}{Z}\sum_G P(M)P(G \mid M)P(F \mid G) \tag{6}$$

where equation 3 is obtained by the application of Bayes' rule, equation 4 is obtained by the marginalization rule applied to both the numerator and the denominator, equation 5 is obtained by replacing the joint distribution by its decomposition and finally, equation 6, as the denominator, is a normalization constant independent of M.

After this short introduction to the formalism, in order to advance our knowledge about the relevance of the subjectivist probabilistic approach to modeling and building either living or artificial cognitive systems, we review, in the sequel, a number of concepts common to cognitive systems, such as *ambiguities*, *fusion*, *multimodality*, *conflicts*, *modularity*, *hierarchies* and *loops*.

First, each of these concepts is introduced briefly and illustrated by examples from the literature in neuroscience, psychophysics or robotics. Finally, we present a template Bayesian model to account for these concepts, emphasizing their most prominent assumptions.

2 Ambiguities

Natural cognitive systems are immersed in rich and widely variable environments. It would be difficult to assume that such systems apprehend their environments in all their details, all the time, if only because of limited sensory or memory capacities. As a consequence, relations between the characteristics of external phenomena and internal states cannot always be bijections. In other words, internal states will sometimes be ambiguous with respect to external situations.

2.1 Inverse problem

Description A problem is said to be *inverse* when we know a direct (or forward) relation and we seek the reverse relation. The inversion of a deterministic function, which often does not have a closed-form solution, can be very difficult.

Example Sensation is commonly defined as the effect of some phenomenon on the senses. Perception involves recovering information about the phenomenon, given the sensation. Perception is an inverse problem (Poggio, 1984; Yuille and Bülthoff, 1996; Pizlo, 2001). Indeed, it is often easy to predict the sensations corresponding to a particular phenomenon (see section 3). In this case, the direct function yields the sensation given the phenomenon, whereas perception is the inverse problem of extracting the phenomenon given the sensation. For example, when we know the shape of an object, basic geometry allows us to derive its projection on the retina. Conversely, it is difficult to reconstruct the shape of an object given only its projection on the retina.

Model In the Bayesian framework, an inverse problem is addressed using the symmetry of Bayes' rule. In a generic example of perception, let Φ be a variable representing some characteristics of the phenomenon and let S be a variable representing the sensation. The joint probability distribution is typically factored as:

$$P(\Phi\ S) = P(\Phi)P(S \mid \Phi). \tag{7}$$

In this expression, $P(\Phi)$ is a *prior* on the phenomenon; that is, the expectation about the phenomenon before any observation has occurred. $P(S \mid \Phi)$ is the probability distribution over sensations, given the phenomenon, which is also known as the *likelihood* of the phenomenon (when considered not as a probability distribution but as a function of Φ); it is the direct model.

The probabilistic question of perception is $P(\Phi \mid S)$, the probability distribution on the phenomenon, based on a given sensation. This question, which is the *posterior* distribution on the phenomenon after some observation, is solved by Bayesian inference:

$$P(\Phi \mid S) = \frac{P(\Phi\ S)}{P(S)} \tag{8}$$

$$= \frac{P(\Phi\ S)}{\sum_{\Phi} P(\Phi\ S)} \tag{9}$$

$$P(\Phi \mid S) = \frac{P(\Phi)P(S \mid \Phi)}{\sum_{\Phi} P(\Phi)P(S \mid \Phi)}. \tag{10}$$

Equation 8 is the application of Bayes' rule, and equation 9 is the application of the marginalization rule. In equation 10, the joint distribution is replaced by its decomposition (equation 7). This expression can be computed because it only involves the prior and the likelihood functions, which are specified in the decomposition.

2.2 Ill-posed problem

Description A problem is said to be *well posed* when it admits a unique solution. Conversely, an *ill-posed* problem admits either many solutions or no solution at all. In most of the non-trivial inverse problems, the direct functions are not injective. Therefore, the inverse relation is not properly a function. When the direct function is not injective, the inverse problem is ill-posed.

Example Perception is often an ill-posed problem. One illustrative example is the well-known Necker cube, in which a wire-frame 2-D drawing of a cube is often perceived as a cube in one of two possible positions (even if it can actually correspond to the projection of an infinite number of 3-D structures).

There are many other examples of ill-posed perception, including recovering the shape of an object from lighting or motion. Such examples have led to the development of Bayesian models of perception, which are reviewed in the book by Knill and Richards (1996) and in the article by Kersten et al. (2004). It has also been argued that illusions arise when ill-posed problems are solved by choosing a solution that does not correspond to the reality of the percept (an example can be found in Geisler and Kersten (2002)).

In robotics, a common instance of an ill-posed problem is *perceptual aliasing*, which occurs in robot localization when two different locations in an environment produce identical sensor readings (Kuipers, 2000; Thrun, 2000).

In speech recognition or generation, recovering the seven-dimensional shape of the vocal tract (Maeda, 1990) from the first four dimensions (formants) of the acoustic signal alone would be another example of an ill-posed problem. More generally, the control of any redundant motor system (*i.e.,* one with more degrees of freedom than the dimension of the space to be accessed) is an ill-posed problem.

Model In the Bayesian framework, the focus is not on finding a solution matching all the constraints of the problem. Instead, a probability distribution is computed. The analogy of a well-posed problem is a distribution with exactly one mode. An ill-posed problem will typically yield a multimodal distribution, or a distribution with a plateau.

The theoretical issue with inverse problems is that a direct function does not always admit an inverse function. In the Bayesian framework, the probability distribution can always be reversed, even if it is difficult to compute. Notice that the property of an ill-posed problem is defined by the result of the inference. On the other hand, a problem is inverse by virtue of its structure. Therefore, an ill-posed problem is not necessarily an inverse problem.

2.3 Discussion

An ambiguity is a difficulty in finding a unique interpretation for a given stimulus. This often leads to multistable percepts, in which case, different interpretations can be actually perceived in the same conditions. Perceptual reversals can be involved when a transition occurs from one percept to another. Ambiguity often arises as a consequence of an ill-posed, inverse problem.

However, an ambiguous likelihood for a given stimulus does not always induce multiple percepts. Indeed, as equation 10 shows, the likelihood is multiplied by the prior, which can filter out part of the solutions. An illusion can occur precisely when the likelihood produces multiple interpretations and the prior rules out the correct one.

3 Fusion, multimodality, conflicts

Natural cognitive systems are equipped with a variety of rich sensors: rich, as they continuously provide several measurements about a given phenomenon (*e.g.,* multiple cells in the retina), and various, as they can measure different manifestations of the same phenomenon (*e.g.,* both hearing and seeing a falling object). The difficulty arises when several of these measurements are to be used together to recover characteristics of the phenomenon. We argue that the concepts of fusion, multimodality and conflict can generally be cast into the same model structure. This model is structured around the conditional independency assumption. We also present extensions of this model that alleviate this assumption, thus trading model simplicity for expressiveness.

3.1 Fusion

Description Often, there are multiple sources of information that can be exploited for any given phenomenon. *Fusion* is the process of forming a single percept starting from multiple sources of information.

Example A common example of fusion in the study of perception is *cue combination* (or intramodal fusion). For instance, Jacobs (1999) studied the perception of the depth of a simulated rotated cylinder. The depth of the cylinder can be perceived by the motion of the texture, the deformation of the texture, or both. Weiss et al. (2002), who studied the perception of the motion of a rhombus, proposed a model of fusion of

the velocity estimates obtained from the edges that, in particular, can account for an illusion in the perception of the motion of a flat rhombus. Hillis et al. (2004) proposed a model for combining texture and disparity cues for perceiving the slant of a plane. One last example is given by Drewing and Ernst (2006), who modeled the perception of curvature from haptic feedback, using both position and force information.

In engineering and robotics, fusion is a common tool that is involved in many applications (*e.g.,* sensor fusion, medical diagnosis and antispam filters).

Model All these examples follow the classical model of *naive Bayesian fusion* (sometimes called *weak fusion*). A variant of this model is the naive Bayes' classifier (Mitchell, 1997). This model relies on the assumption that each piece of information is independent from the others, conditioned on knowledge of the phenomenon. Each piece of information is related to the phenomenon using an inverse model similar to that discussed in section 2.1. For instance, once the cause is known, the consequences follow independently. This very strong hypothesis is far from always being satisfied. However, it often gives satisfactory results.

With $S_1, \ldots, S_N$ representing the variables corresponding to the N pieces of information that are to be part of the fusion, and Φ being the variable corresponding to characteristics of interest in the phenomenon, the naive fusion model assumes the following decomposition of the joint probability distribution:

$$P(\Phi\ S_1 \cdots S_N) = P(\Phi) \prod_{i=1}^{N} P(S_i \mid \Phi). \tag{11}$$

This model can be used to compute the probability distribution on the phenomenon, Φ, given all the sensations, S_1, …, S_N:

$$P(\Phi \mid S_1\ \cdots\ S_N) = \frac{P(\Phi) \prod_{i=1}^{N} P(S_i \mid \Phi)}{P(S_1\ \cdots\ S_N)} \propto P(\Phi) \prod_{i=1}^{N} P(S_i \mid \Phi).$$

This model offers the advantage of being much simpler than a complete model without any independence, as the joint is written as a product of low dimensional factors. The size of a complete joint probability distribution is exponential in the number of variables, whereas it is only linear when assuming naive fusion.

Furthermore, the use of many information sources generally allows for an increase in the signal. For example, in the case of Gaussian uncertainty for the sensor models, $P(S_i \mid \Phi)$, and the prior, $P(\Phi)$, the uncertainty of the posterior distribution, $P(\Phi \mid S_1 \cdots S_N)$, can be proven to be smaller with such a model than if the pieces of information are not fused at all.

Note that some models rely on *maximum likelihood estimation* (MLE). That is, they do not compute the posterior probability distribution, $P(\Phi \mid S_1 \cdots S_N)$, but instead they are concerned with the likelihood of the phenomenon, $P(S_1 \cdots S_N \mid \Phi)$. For a given set of sensations, this is a function of Φ that reflects how well the phenomenon explains the sensations. Thus, the MLE selects the phenomenon that best matches the data.

Moreover, Bayes' rule is written as: $P(\Phi \mid S_1 \cdots S_N) \propto P(\Phi) P(S_1 \cdots S_N \mid \Phi)$. With a uniform prior, this states that the posterior is proportional to the likelihood. Therefore, the MLE is tantamount to assuming a uniform prior or dismissing it altogether.

3.2 Multimodality

Description Fusion is often considered within a sensory modality, such as vision or touch. However, it is often beneficial to consider multiple sources of information from various sensory modalities when forming a percept in what is known as *multimodality*.

Example Bayesian modeling has often been used to study multimodal fusion. For example, Anastasio et al. (2000) have proposed a model of multisensory enhancement in the superior colliculus. Enhancement occurs when the neural response to a stimulus in one modality is augmented by a stimulus in another modality. In their work, they proposed a model of the probability of a target in the colliculus with respect to the presence or absence of visual and auditory input. Zupan et al. (2002) have proposed a model for visuo-vestibular interaction in which they use a sensory weighing model, interpreted as an MLE. Gepshtein and Banks (2003), who studied visuo-haptic integration, used a maximum likelihood model in which the reliability of the visual cues is related to the projection. Körding and Wolpert (2004a) performed experiments on sensorimotor learning. Using a Bayesian model of integration between visual feedback and proprioception, they inferred and manipulated the prior distributions assumed for the participants in their experiments. Jürgens and Becker (2006) used Bayesian fusion for the combination of vestibular, optokinetic, podokinesthetic, and cognitive information for the perception of rotation. Finally, Haith et al. (2008) looks into sensorymotor adaptation in a visually guided reaching experiment and proposes a model that can also account for perceptual after-effects. This model predicts a visual shift after adaptation of reaching in a force field, that has been confirmed by an experiment.

Model These models match the assumptions of the naive Bayesian fusion model presented above. The main assumption is that the probability distribution over each sensation is independent of the others given the phenomenon. Some of these models use MLE, while others compute the complete posterior distribution.

Although the models are the same as those for intramodal fusion, they differ in robustness. The conditional independence hypothesis is never totally valid. Part of the uncertainty that exists between sensations may not be explained by Φ. These additional correlations are more likely smaller between different modalities than within the same modality, as different physical processes are involved. For example, while fog or lack of illumination can jointly degrade different types of visual information, such as colour or motion, hearing performance is not affected.

3.3 Conflicts

Description When various sources of information are involved, each can sometimes lead to significantly different individual percepts. In experiments, a *conflict* arises when sensations from different modalities induce different behaviour compared with when each modality is observed individually.

Introducing conflicts offers good experimental leverage for providing a lot of information on the fusion process. For instance, the relative importance of different cues can be assessed by observing the frequencies of behaviour corresponding to each of the cues. Sometimes, the response to conflicting situations is mixed behaviour that is also interesting to study when compared with that occurring in the original situations.

Examples To study the integration of visual and haptic information, Ernst and Banks (2002) designed experiments in which the height of a bar had to be evaluated either by simulated visual input, simulated tactile input or both. In this latter condition, when different heights for both visual and haptic stimuli were simulated, a maximum-likelihood model provided a good match to their data. Battaglia et al. (2003) designed a localization task with conflicting visual and auditory input to compare the classical visual capture model with a maximum likelihood model. They argued for taking into account the observers' perceptual biases. One special case of this conflict is the ventriloquist effect (Alais and Burr, 2004; Banks, 2004).

Model The models accounting for conflicts are still naive fusion and maximum likelihood (with a naive fusion decomposition). Conflicts arise when the probability distributions corresponding to each sensation are significantly different.

The concept of a conflict is of a similar nature to that of an ill-posed problem. Both are defined with respect to characteristics of the result. Conversely, inversion and fusion are both structural properties of a problem and its associated model.

3.4 Less naive fusion

Description Many models assume independence between the sensations conditionally to knowledge of the phenomenon. This is sometimes too strong an assumption. A model could assume no conditional independence between the cues. This is sometimes called *strong fusion*, and leads to arbitrarily complex models that are not tested easily. Less naive fusion models lie between strong fusion and naive (or weak) fusion.

Example For instance, Landy et al. (1995) introduced a *modified weak fusion* framework. They tackled the cue combination problem in depth estimation and argued for the addition of so-called *ancillary cues*. These cues do not provide direct information on depth but help to assess the reliability of the various cues that appear in the fusion process. Ancillary cues are the basis for a *dynamic reweighting* of the depth cues.

Yuille and Bülthoff (1996) propose the term *strong coupling* for non-naive Bayesian fusion. They consider the examples of shape recovery from shading and texture, and the coupling of binocular depth cues with monocular cues obtained from motion parallax.

Models Recall that the naive fusion model is based on the following factorization of the joint probability distribution (equation 11):

$$P(\Phi\ S_1 \cdots S_N) = P(\Phi) \prod_{i=1}^{N} P(S_i \mid \Phi).$$

The inclusion of ancillary cues can be achieved by the addition of a variable, A, to represent them.[2] Typically, the joint distribution is then factored as:

$$P(\Phi\ A\ S_1 \cdots S_N) = P(\Phi) P(A \mid \Phi) \prod_{i=1}^{N} P(S_i \mid \Phi\ A).$$

[2] We note that Landy et al. (1995) did not find this type of Bayesian model to be useful.

We can still ask the same question of the phenomenon given the sensations:

$$P(\Phi \mid S_1 \cdots S_N) \propto P(\Phi) \sum_A \left[P(A \mid \Phi) \prod_{i=1}^{N} P(S_i \mid \Phi\ A) \right] \tag{12}$$

or we can know the value of the ancillary cue:

$$P(\Phi \mid A\ S_1 \cdots S_N) \propto P(\Phi) P(A \mid \Phi) \prod_{i=1}^{N} P(S_i \mid \Phi\ A). \tag{13}$$

This model is the same as a naive fusion model but with a phenomenon augmented by the ancillary cues, $\Phi' = \Phi \wedge A$. The only difference is in the questions asked, this new model only being concerned with a part, Φ, of the global phenomenon, Φ', either with knowledge of ancillary cues, A, (this question is illustrated by equation 13) or not (this question is illustrated by equation 12).

3.5 Discussion

Fusion is often seen as a product of models. It can be used when the underlying models are defined independently so that they can be combined to form a shared variable. Each model links this variable to distinct properties, and the fusion operates on the conjunction (logical product) of these properties.

This idea of fusion as a product of independent models is also mirrored by the inference process. Indeed, most of the time, the result of the fusion process is proportional to the product of each individual result obtained by the underlying models, $P(\Phi \mid S_1 \cdots S_n) \propto \prod_{i=1}^{N} P(\Phi \mid S_i)$. However, this may not be the case, depending on the exact specification of the underlying models. There are some more complex fusion models that ensure that this inference product holds (Pradalier et al., 2003).

It is also interesting to note that, when all the probability distributions in the product are Gaussian, the resulting probability distribution is also Gaussian. Its mean is a weighted sum of the means of each Gaussian weighted according to a function of their variance. Therefore, many weighted models proposed in the literature can be interpreted as Bayesian fusion models with Gaussian uncertainty. The weights also acquire the meaning of representing uncertainty, which can sometimes be manipulated.

Finally, a conflict can only occur when it is assumed that a unique object is to be perceived. When the discrepancy is too large, segmentation occurs leading to the perception of separate objects (*e.g.,* perception of transparency). Likewise, fusion is a process of combining information from different sources for one object or feature. Therefore, to account for segmentation, there is a need for more complex models. Such models can deal with either one or multiple objects, as well as provide a mechanism for deciding whether there is one or more objects. This theoretical issue is called *binding*, *unity assumption*, or *pairing* (Sato et al., 2007), and has received recent attention using hierarchical Bayesian models.

4 Modularity, hierarchies

It would be difficult to assume that natural cognitive systems process their complex sensory inputs in a single layer of computation. Therefore, computations can probably

be apprehended as processes that communicate intermediate results. This is similar to the notion of modularity, which is central to structured computer programming. A flow of computation can be broken down into structured sequences. First, we present Bayesian models that amount to probabilistic versions of subroutine calls and conditional switches. More complex constructs occur when parts of the computations are based on observing other computation processes. By so doing, we also describe Bayesian model recognition and Bayesian model abstraction.

4.1 Subroutine

Description Hierarchies rely on the notion of modularity, the simplest instance of which is the subroutine; that is, the use of part of a model (submodel) within another model. A model can be seen as a resource that can be exploited by other models.

Example For example, Laskey and Mahoney (1997) propose the *network fragments* framework, inspired by *object-oriented* software analysis and design to define submodules of a global probabilistic model. They apply this tool to military situation assessment, where a military analyst has to reason on various levels (basic units, regiment, overall situation, etc.)

Koller and Pfeffer (1997) also propose *object-oriented Bayesian networks* as a modeling language for Bayesian models based on an object-oriented framework.

Model Here, we define the operation of calling a subroutine in terms of the Bayesian programming framework. Let A and B be variables of interest for the global model. We add an additional variable, Π, to represent the possible models. Let π_1 be a value of Π that represents the particular model to be written. When defining π_1, the model Π is known; that is, the probability distributions are written given $\Pi = \pi_1$. For example:

$$P(A\ B \mid [\Pi = \pi_1]) = P(A \mid [\Pi = \pi_1])P(B \mid A\ [\Pi = \pi_1]).$$

A call to a subroutine can be performed simply by specifying that a distribution for this model is the same as that for another model:

$$P(B \mid A\ [\Pi = \pi_1]) = P(B \mid A\ [\Pi = \pi_2]) \tag{14}$$

where π_2 is another model concerned with variables A and B. Equation 14 is an assumption that the distribution over B given A in model π_1 is the same as in π_2.

Naturally, this requires model π_2 to be specified. In this submodel, $P(B \mid A\ [\Pi = \pi_2])$ can be a factor in the decomposition of the joint probability distribution over all the variables of model π_2. In this case, the use of a submodel is not required. However, most of the time, $P(B \mid A\ [\Pi = \pi_2])$ is the result of an inference in model π_2. As a Bayesian Program, π_2 can be asked any probabilistic question related to its variables and can be arbitrarily complex. Moreover, many different models can question π_2. In this sense, π_2 can be considered as a resource.

4.2 Probabilistic conditional switches

Description Mixture models are probability distributions with multiple components. Such a model is generally a weighted sum of unimodal probability distributions in order to yield a desired multimodal probability distribution.

Example Mixture models are basic tools with which to build a model. They are used as the parametric forms of probability distributions, along with uniform or Gaussian distributions.

One common use of mixture models is in clustering and classification. A data set is represented by a mixture of some basis distributions. The aim of clustering is to learn the parameters of the mixture, whereas classification is usually interested in recovering the element of the mixture from which a particular observation originates. Classical examples of mixture models are the *Gaussian mixture models*.

Mixture of experts is another type of mixture model. Initially, such models were defined for neural networks (Jacobs et al., 1991). Waterhouse et al. (1996) proposed a Bayesian learning algorithm for mixture models, the aim of which was to infer the parameters of the mixture. Bishop and Svensén (2003) proposed variational methods of inference for mixture of experts models and applied their algorithm to the prediction of the position of the end effector of a robotic arm.

Model Mixture models are usually presented directly as weighted sums of distributions:

$$P(A) = \sum_{i=1}^{N} w_i \times P_i(A) \tag{15}$$

where N is the number of components, w_i is the weight of each component, $P_i(A)$, and $\sum_{i=1}^{N} w_i = 1$.

Another way to present it involves introducing a variable, Π, which can take integer values between 1 and N. We can build the following decomposition:

$$P(A\ \Pi) = P(\Pi)P(A \mid \Pi)$$

where $P(A \mid [\Pi = \pi_i]) = P_i(A)$, with $P_i(A)$ as defined above and $P([\Pi = \pi_i]) = w_i$.

The probabilistic question corresponding to the mixture model is:

$$P(A) = \sum_{i=1}^{N} P([\Pi = \pi_i])P(A \mid [\Pi = \pi_i]) = \sum_{i=1}^{N} w_i \times P_i(A).$$

This definition of a mixture places the emphasis on variable Π, which can be considered as the index on the component $P_i(A)$ to be chosen. It is hierarchical in the sense that a collection of models, $P_i(A)$, can be chosen using a variable external to each of these models.

This mixture model can be generalized by a probabilistic conditional statement. Let C be the variable representing the test condition. We can build the following joint probability distribution:

$$P(A\ \Pi\ C) = P(C)P(\Pi \mid C)P(A \mid \Pi)$$

where $P(C)$ is a prior on the condition (not used in the following question), $P(\Pi \mid C)$ is the relative weighting among each of the models that depend on the condition, and $P(A \mid \Pi)$ is a collection, $P_i(A)$, of various models of A, possibly given condition C.

The question can be written as:

$$P(A \mid C) = \sum_{\Pi} P(\Pi \mid C)P(A \mid \Pi). \tag{16}$$

With Dirac distributions for $P(\Pi \mid C)$, only one model, $P_i(A)$, can be selected in the expression 16, this question operating like a switch between the models. Otherwise, this is a probabilistic generalization of a conditional statement; that is, the different models are weighted according to the condition.

4.3 Model recognition

Description The mixture model allows for a combination of different models into one. Another issue that can be addressed using Bayesian modeling is model recognition. In particular, when the class of models is a parameter space and recognition is based on experimental data, this recognition amounts to a *machine learning* problem.

Example For example, Gopnik and Schulz (2004) studied the learning of causal dependencies by young children. The experiments included trying to decide which objects are "blickets" (imaginary objects that are supposed to illuminate a given machine). Some objects are put on a machine that lights up depending on which objects are placed on it. The patterns of response were predicted well by a causal Bayesian network, even after adding some prior knowledge ("blickets are rare"). The learning phase involved selecting the causal dependency structure that matches the observations among all the possibilities.

Another example is the application of *embedded Hidden Markov Models*. These are models in which a top-level Bayesian model reasons on nodes that are themselves Bayesian models. Nefian and Hayes (1999) proposed such a formalism in the context of face recognition. Each submodel is responsible for the recognition of a particular feature of the face (forehead, eyes, nose, mouth and chin). The global model ensures perception of the facial structure by recognizing each feature in the correct order. Neal et al. (2003) applied embedded HMMs to the tracking of 3-D human motion from 2-D tracker images. The high dimensionality of the problem proved to be less an issue for their algorithm than it was for the previous approach.

Model These models can be fit using the general framework of Bayesian model recognition. Let $\Delta = \{\Delta_i\}$ be the variables corresponding to the data used for learning (Δ_i, a variable for each datum), and let Π be the variable corresponding to the model. Generally, model recognition is performed using the following decomposition:

$$P(\Pi\ \Delta) = P(\Pi)P(\Delta \mid \Pi)$$

where $P(\Pi)$ is a prior on the various models and $P(\Delta \mid \Pi)$ is the probability of observations given the model (*i.e.*, the likelihood of models).

Typically, the data are assumed to be independently and identically distributed (the i.i.d. assumption); that is, $P(\Delta \mid \Pi) = \prod_{i=1}^{N} P(\Delta_i \mid \Pi)$, where $P(\Delta_i \mid \Pi)$ does not depend on index i of each datum. For each possible model, π_j, the distribution, $P(\Delta_i \mid [\Pi = \pi_j])$, is a call to the submodel, π_j, as defined in section 4.1.

The probabilistic question for model recognition is:

$$P(\Pi \mid \Delta) \propto P(\Pi) \prod_{i=1}^{N} P(\Delta_i \mid \Pi). \tag{17}$$

The expression 17 enables the computation of a probability distribution on the various models based on some of the data. This mechanism is hierarchical in the sense that we build a model for reasoning about an underlying set of models.

Parameters A common instance of model recognition is parameter learning. In this case, the models, Π, share a common parametric form, Π', and the recognition occurs on the parameters, Θ: $\Pi = \Theta \wedge \Pi'$. We can modify the decomposition presented above by including knowledge of the parametric form:

$$P(\Theta\ \Delta \mid \Pi') = P(\Theta \mid \Pi')P(\Delta \mid \Theta\ \Pi') = P(\Theta \mid \Pi')\prod_{i=1}^{N} P(\Delta_i \mid \Theta\ \Pi')$$

where $P(\Theta \mid \Pi')$ is a prior distribution on the parameters and $P(\Delta \mid \Theta\ \Pi')$ (or $P(\Delta_i \mid \Theta\ \Pi')$ with the i.i.d. assumption) is the likelihood of parameters.

So, learning can be accomplished using the following question:

$$P(\Theta \mid \Delta\ \Pi') \propto P(\Theta \mid \Pi')P(\Delta \mid \Theta\ \Pi') \propto P(\Theta \mid \Pi')\prod_{i=1}^{N} P(\Delta_i \mid \Theta\ \Pi').$$

The likelihood functions are usually completely specified by the parametric form, Π', and the parameters, Θ. However, the prior on the parameters, $P(\Theta \mid \Pi')$, may need some additional parameters called *hyperparameters*. The Bayesian formalism allows for these hyperparameters in the same way. Let Λ be the variable representing these hyperparameters. We write the joint probability distribution as:

$$P(\Lambda\ \Theta\ \Delta \mid \Pi') = P(\Lambda \mid \Pi')P(\Theta \mid \Lambda\ \Pi')P(\Delta \mid \Theta\ \Pi')$$

where $P(\Lambda \mid \Pi')$ is a prior on hyperparameters, $P(\Theta \mid \Lambda\ \Pi')$ is the distribution on parameters, Θ, according to hyperparameters and $P(\Delta \mid \Theta\ \Pi')$ is the likelihood function, as above. As a result, inference on the parameters is modified slightly:

$$P(\Theta \mid \Delta\ \Pi') \propto \sum_{\Lambda} \left[P(\Lambda \mid \Pi')P(\Theta \mid \Lambda\ \Pi')\right]P(\Delta \mid \Theta\ \Pi').$$

The prior on the hyperparameters could also be parametric, and it is possible to add another set of parameters. It all amounts to what knowledge the modeler wants to include in the model. Moreover, it can be shown that deep layers of priors have much less influence on parameter estimation.

Entropy principles All the entropy principles (maximum entropy principle, minimum relative entropy principle, Kullback–Leibler divergence) and their numerous applications are closely and simply related to the above models.

Indeed, to take the simplest case of equation 17, if we consider that there are K different possible observations (*i.e.,* the variable Δ_i can take K different values), by gathering the same observations, we can restate this equation as:

$$P(\Pi \mid \Delta) \propto P(\Pi)\prod_{k=1}^{K} P([\Delta_i = k] \mid \Pi)^{n_k}$$

where n_k is the number of times that the observation, $[\Delta_i = k]$, has been made. To a first approximation, this number is proportional to the probability, $P([\Delta_i = k] \mid \Pi)$, itself and we obtain, with N is still the total number of observations:

$$P(\Pi \mid \Delta) \propto P(\Pi) \prod_{k=1}^{K} P([\Delta_i = k] \mid \Pi)^{NP([\Delta_i=k] \mid \Pi)} \tag{18}$$

Finally, if we assume a uniform prior on the different models, and if we take the logarithm of equation 18, we obtain the Maximum Entropy Principle:

$$\log\left(P(\Pi \mid \Delta)\right) = N \sum_{k=1}^{K} \left[P([\Delta_i = k] \mid \Pi) \log\left(P([\Delta_i = k] \mid \Pi)\right)\right] + C.$$

4.4 Abstraction

Description Usually, a modeler uses learning in order to select a unique model or set of parameter values, which is then applied to the problem at hand. That is, the learning process computes a probability distribution over models (or their parameters) to be applied, and a decision, based on this probability distribution, is used to select only one model or parameter set.

Another way to use model recognition is to include it as part of a higher-level program in order to maintain the uncertainty on the models at the time of their application. This is called model *abstraction*.

Example Diard and Bessière (2008) used abstraction for robot localization. They defined several *Bayesian maps* corresponding to various locations in the environment. Each map is a model of sensorimotor interactions with a part of the environment. Then, they built an *abstracted map* based on these models. In this new map, the location of the robot is defined in terms of the submap that best fits the observations obtained from the robot's sensors. The aim of their abstracted map was to navigate in the environment. Therefore, they were more interested in the action to be taken than in the actual location. However, the choice of an action was made with respect to the uncertainty of the location.

A similar model was also recently applied to the domain of multimodal perception, under the name of causal inference (Körding et al., 2007; Sato et al., 2007). When sensory cues are close, they could originate from a single source, and the small spatial discrepancies could be explained away by noise; on the other hand, when cues are largely separated, they more probably originate from distinct sources, instead, and their spatial positions are not correlated. The optimal strategy, when estimating the positions of these cues is then to have both alternative models coexist and to integrate over the number of sources during the final estimation.

Model Let Π be the variable representing the submodels, let Δ be the variable representing the data and let X be the sought-after variable that depends on the model. The joint probability distribution can be decomposed in the following way:

$$P(X\ \Delta\ \Pi) = P(\Pi)P(\Delta \mid \Pi)P(X \mid \Pi)$$

where $P(\Pi)$ and $P(\Delta \mid \Pi)$ are the priors and likelihood functions defined as before and $P(X \mid \Pi)$ describes the influence of the model, Π, on the variable of interest, X.

The question is concerned with the distribution over X given the data, Δ:

$$P(X \mid \Delta) = \sum_{\Pi} P(\Pi \mid \Delta) P(X \mid \Pi) \propto \sum_{\Pi} \left[P(\Pi) P(\Delta \mid \Pi) P(X \mid \Pi) \right].$$

This inference is similar to model recognition except for the factor $P(X \mid \Pi)$. With respect to the question, $P(X \mid \Delta)$, the details of the models can be abstracted.

When applied to classes of models and their parameters (*i.e.* when X is Θ), this abstraction model yields the *Bayesian Model Selection* method (BMS). It can also be used to jointly compute the distribution over joint models and parameters, using $P(\Pi\ \Theta \mid \Delta)$ (Kemp and Tenenbaum, 2008).

4.5 Discussion

Hierarchy is a means of controlling a given set of models by a higher-level model. This model uses each lower-level model as a subroutine providing probabilistic relations as a resource. One common example is behaviour choice. Several models propose some type of behaviour depending on the partial knowledge that they need, a higher-level model playing the role of an arbiter between them. There are two common strategies considered for the arbiter, weighing and switching.

Weighing consists of mixing the various behaviours, usually by summation according to some measure. In a Bayesian framework, this is exactly what is achieved by mixture models (see section 4.2 and in particular equation 15), in which each individual model contributes to the result of the inference in proportion to its probability. If the process evolves with time, the probability distribution over the various behaviours can also evolve and, for example, allow for a smooth transition between behaviours.

On the other hand, switching consists of choosing a behaviour, discarding the others, and then applying that behaviour (Stocker and Simoncelli, 2008). This can be seen as the approximation of the summation in equation 15 by only one term. In this case, the decision, which is also called an intermediate decision in the model, helps reduce the complexity of the model by computing only one term instead of all the subroutine calls in the submodels. Conversely, this is no longer an approximation if the weights are zero for all except one submodel (a Dirac probability distribution). In this case, a mixture model is a switching process.

Therefore, the Bayesian formulation unifies these two strategies into a common hierarchical inference mechanism. Furthermore, the difference between weighing and switching depends on the probability distribution over the models, switching if it is a Dirac function (no uncertainty) and weighing otherwise. Therefore, for low uncertainty, weighing and switching strategies cannot be distinguished.

5 Loops

It would be hard to assume that natural cognitive systems process their complex sensory systems in a single direction, uniquely from sensory input toward motor outputs. Indeed, neurophysiology highlights a variety of ways that neural system activity can be fed back in loops. Models that include loops are mainly temporal in nature and deal

with memory systems. Examples are mostly taken from models of artificial systems, and especially from the robotics community.

5.1 Temporal series of observations

Description In order to model the temporal behaviour of a phenomenon, it is usually assumed that a sequential series of observations is available. These are usually used for *state estimation*: recovering the evolution of some internal state that is not observed.

Example Kalman filters are the most common examples of models in this category, probably because of strong assumptions that lead to a closed-form solution for state estimation (Kalman, 1960; Ghahramani et al., 1997). They are widely used in robotics (Thrun et al., 2005) and in the life sciences (van der Kooij et al., 1999; Kiemel et al., 2002). When the state space can be assumed to be discrete, Hidden Markov Models can be applied (Rabiner, 1989; Rabiner and Juang, 1993). A common technique for approximating the inference required for state estimation is to model the state probability distribution using particle filters; this can also be seen as the application of a mixture model to a loop model (Arulampalam et al., 2002). More generally, these models are instances of Bayesian filters (Leonard et al., 1992; Bessière et al., 2003). When the independence assumptions do not vary over time, this class of models is called Dynamic Bayesian Networks (Dean and Kanazawa, 1989; Murphy, 2002). These models have also been extensively covered in the statistics literature, sometimes using different vocabularies (Harvey, 1992; Brockwell and Davis, 2000).

Model Let $O^{0:T}$ denote a time series of observation variables from time 0 to T, *i.e.*, $O^{0:T} = O^0, O^1, \ldots, O^T$. Let $S^{0:T}$ represent the state variables over the same time period. The global probabilistic model is defined by:

$$P(S^{0:T}\ O^{0:T}) = P(S^0\ O^0) \prod_{t=1}^{T} P(S^t\ S^{t-1}\ O^t) \tag{19}$$

$$P(S^{0:T}\ O^{0:T}) = P(S^0\ O^0) \prod_{t=1}^{T} P(S^t \mid S^{t-1}) P(O^t \mid S^t). \tag{20}$$

Here, equation 19 results from the application of the *Markov assumption*: a given state only depends on a number, n, of previous states, and not on those that occur previously in the state history. In this case, $n = 1$ implies a *first-order* Markov model. In other words, S^{t-1} is assumed to be sufficient memory for the system. Equation 20 states further that, for all t, the dependency structure between variables is the same. This is the *stationarity hypothesis*.

In equation 20, $P(S^t \mid S^{t-1})$ is usually called the transition or *dynamic model*, and $P(O^t \mid S^t)$ is the *observation model*. When these models have the same form for all time steps, t, the local model, $P(S^t\ S^{t-1}\ O^t) = P(S^t \mid S^{t-1}) P(O^t \mid S^t) P(S^{t-1})$, is said to be time invariant or *homogeneous*. In this case, the model can be seen to be an application of the subroutine construct; that is, the global model, which deals with the complete time sequence, is a result of the iteration of the local model.

State estimation, which corresponds to the question, $P(S^T \mid O^{0:T})$, is solved by:

$$P(S^T \mid O^{0:T}) \propto \sum_{S^{T-1}} P(S^T \mid S^{T-1}) P(S^{T-1} \mid O^{0:T-1}).$$

By so doing, state estimation at time T, $P(S^T \mid O^{0:T})$, is computed recursively, based on $P(S^{T-1} \mid O^{0:T-1})$. In most practical implementations of this computation, the state evaluation is recomputed each time a new observation is acquired using the past state estimate as well as the prediction and observation models. In other words, state estimation occurs in a loop over time.

The same model can be used to predict future states (*prediction*: $P(S^{t+k} \mid O^{0:t}), k > 0$), or to refine some past estimate given observations that occur later in time (*filtering*: $P(S^{t-k} \mid O^{0:t}), k > 0$). The larger the value of k, the more computationally expensive these questions become, as the inferences require summations over $k - 1$ variables.

5.2 Efferent copy

Description Usually, in robotics and control contexts, the state of the observed system can be not only observed but also acted upon. The previous models are enriched with control variables as additional inputs for state estimation. "Reading" the values of these control variables after they have been decided, in order to ameliorate state estimation or prediction, is one possible reason for there being efferent copies of motor variables in animal central nervous systems.

Example Input/output HMMs (Bengio and Frasconi, 1995) have been introduced in the machine learning literature and applied as benchmarks in grammatical inference. In the robotic localization context, the Markov Localization model is the most common example of this model category (Thrun et al., 2005).

In the life sciences, Laurens and Droulez (2007, 2008) applied Bayesian state estimation using efferent copies to the modeling of 3-D head orientation in space in humans and showed that such models could account for a variety of known perceptual illusions.

Model Let $S^{0:T}$ and $O^{0:T}$, respectively, denote state and observation time series, as previously, and let $A^{0:T}$ denote the control time series. The control variables are usually used to refine the dynamic model into $P(S^t \mid A^{t-1}\ S^{t-1})$:

$$P(S^{0:T}\ O^{0:T}\ A^{0:T}) = P(S^0\ O^0\ A^0) \prod_{t=1}^{T} P(S^t \mid A^{t-1}\ S^{t-1}) P(O^t \mid S^t) P(A^t).$$

5.3 Decision theory

Description The temporal loop models presented so far are usually used for state estimation, whether this state is a past, current or future one. They can also be used to compute probability distributions over future actions. For instance, this can allow the choice of an action or a sequence of actions that are most likely to help progress toward a given goal. The goal is usually not defined in probabilistic terms but with the use of a deterministic function that describes those states and actions that are either beneficial or detrimental. This is called a *reward, cost or loss function*. The process of computing future actions based on this function is known as *decision theoretic planning*.

Example A number of current models apply decision theory to robotic and control planning. For instance, when the above Markov Localization model is enriched with a reward function, it becomes a Partially Observable Markov Decision Process (Kaelbling et al., 1998; Boutilier et al., 1999), which has been widely applied in mobile robotics. If the internal state is not "hidden", in the sense that it can be measured directly, the sensor model, $P(O^t \mid S^t)$, is removed from the dependency structure; the resulting formalism is a Markov Decision Process (Fully Observable Markov Decision Process).

The issue of decision, that is, passing from a probability distribution to a value, is actually not specific to loop models. For example, Körding and Wolpert (2004b) apply Bayesian modeling to infer what loss function might be used by the central nervous system in an open-loop motor control problem.

Model The reward function, R, in its most general notation, associates states and actions with a number that quantifies their interest or cost, $R : S^t \times A^t \rightarrow \mathbb{R}$.

The reward function helps drive the planning process. Indeed, the aim of this process is to find an optimal plan in the sense that it maximizes a certain measure based on the reward function. This measure is most frequently the expected discounted cumulative reward, $\langle \sum_{t=0}^{\infty} \gamma^t R^t \rangle$, where γ is a discount factor (less than 1), R^t is the reward obtained at time t and $\langle \cdot \rangle$ is the mathematical expectation[3]. Given this measure, the goal of the planning process is to find an optimal mapping from probability distributions over states to actions (a *policy*).

This planning process, which leads to intractable computations, is sometimes approximated by iterative algorithms called *policy iteration* or *value iteration*. These algorithms start with random policies and improve them at each step until some numerical convergence criterion is met. Another approach is to try to automatically find a hierarchical decomposition of the state space, so as to alleviate the combinatorial explosion (Hauskrecht et al., 1998; Pineau and Thrun, 2002).

5.4 Action selection

Description Instead of trying to find a structural decomposition of the state space automatically, alternative approaches can be pursued in order to reduce the computational complexity of the planning and control processes. It is assumed that the model already incorporates knowledge about the task or domain.

For instance, modeling cognitive systems requires including in the model knowledge about the environment structure or task progression structure, so as to separate elementary filters, and to reduce the complexity and dimensionality of the internal state space. Attention systems must also be defined so that computations are performed only in those subsets of the state space that are thought to be relevant. Finally, elementary motor tasks can be identified and modeled as behaviours, so that they are not planned completely through each time period.

Example Koike (2005); Koike et al. (2008) applied this approach in the domain of autonomous mobile sensorimotor systems. Koike showed how time and space limits on computations, and limited on-board processing power, could be accommodated,

[3] Please note that, although the notation is the same, γ^t refers to γ elevated to the power of t, while R^t refers to R at time index t.

$$
\begin{aligned}
&P(A^{0:t}\ S^{0:t}\ O^{0:t}\ B^{0:t}\ C^{0:t}\ \lambda^{0:t}\ \beta^{0:t}\ \alpha^{0:t} \mid \pi) \\
&= \left[\begin{array}{l}
\prod_{j=1}^{t} \left[\begin{array}{l}
\prod_{i=1}^{N_i} P(S_i^j \mid S_i^{j-1}\ A_i^{j-1}\ \pi_i) \prod_{i=1}^{N_i} P(O^j \mid S_i^{j-1}\ C^j\ \pi_i) \\
P(B^j \mid \pi) \prod_{i=1}^{N_i} P(\beta^j \mid B^j\ S_i^j\ B^{j-1}\ \pi_i) \\
P(C^j \mid \pi) \prod_{i=1}^{N_i} P(\alpha^j \mid C^j\ S_i^j\ B^j\ \pi_i) \\
P(A^j \mid \pi) \prod_{i=1}^{N_i} P(\lambda^j \mid A^j\ B^j\ S_i^j\ A^{j-1}\ \pi_i)
\end{array} \right] \\
P(A^0\ S^0\ O^0\ B^0\ C^0\ \lambda^0\ \beta^0\ \alpha^0 \mid \pi)
\end{array} \right]
\end{aligned}
$$

Fig. 2 Joint probability factorization for the full model as designed by Koike.

thanks to simplifying assumptions such as the stationarity of the temporal models, partial independence between sensory processes, domain-of-interest selection at the processing stages (attention focusing) and behaviour selection.

Model Incorporating such knowledge into models makes them less generic and more sophisticated. Figure 2 shows the joint distribution of the full model from Koike.

In Figure 2, π_i refers to the i elementary filters, which are assumed to be independent (hence all the $\prod_{i=1}^{N_i}$ products). $P(S_i^j \mid S_i^{j-1}\ A_i^{j-1}\ \pi_i)$ are their dynamic models, $P(O^j \mid S_i^{j-1}\ C^j\ \pi_i)$ are their sensor models, $P(\beta^j \mid B^j\ S_i^j\ B^{j-1}\ \pi_i)$ are their behaviour models, $P(\alpha^j \mid C^j\ S_i^j\ B^j\ \pi_i)$ are their attention models, and $P(\lambda^j \mid A^j\ B^j\ S_i^j\ A^{j-1}\ \pi_i)$ are their motor command models. Finally, $P(A^0\ S^0\ O^0\ B^0\ C^0\ \lambda^0\ \beta^0\ \alpha^0 \mid \pi)$ encodes the initial state of the system. This model is then used to compute the probability distribution over the next action to be performed, given the past history of observation and control variables: $P(A^t \mid O^{0:T}\ A^{0:T-1})$.

5.5 Discussion

There is no clear definition of a loop. In this section, we have only presented the definition and examples of temporal loops. These loops can be compared to loops in the field of computer science, occurring when the execution flow gets several times through the same set of instructions. Such instructions are specified once for all the executions of the loop, and the global program is its replication through time.

This replication often occurs with fixed time spans. However, in biological systems, multiple loops may take place simultaneously with different and sometimes varying time constants. In robotics, many processes are run concurrently with different levels of priority. There is a need in Bayesian modeling for a proper way of integrating and synchronizing loops with different time scales. Finally, loops can also be considered without reference to time. Bayesian filters are a single model that is replicated at each time step, with an optional temporal dependency on preceding time steps. Models have also been proposed for spatial replication of models, with dependencies occurring over a neighbourhood. One interesting difference is that temporal relations between instances are oriented according to the passage of time, whereas models of spatial loops, such as the Markov Random Field, rely on a symmetrical relation between neighbours.

6 Conclusion

In this document, we reviewed a number of cognitive issues that both living and artificial systems have to face. We grouped these issues according to their global nature,

in terms of ambiguities, fusion, hierarchies and loops. For each issue, we proposed a template Bayesian model, according to its general specification, which was based on examples from the literature.

These groups of issues correspond to a presentation of the models in terms of their increasing complexity. It can be noted that, in the beginning of the paper, the examples are taken mostly from the life sciences, whereas toward the end, they are taken from computer science and robotics. This last point is not a choice but mirrors the relative shortage of more complex models in biology. However, the recent literature shows that Bayesian methodology is gaining momentum in the life sciences, even if the proposed models are still simpler than those developed in computer science (Wolpert, 2007).

One important reason for this difference is that, contrary to the situation in robotics, the models presented in the life sciences should be falsifiable. That is, there must be a way to decide whether the model is false or not. By so doing, the assumptions underlying the model can be eliminated. In an uncertain environment, this can be difficult. In the Bayesian formalism, the analogous paradigm is better replaced by a comparison between models. Various models corresponding to different assumptions are constructed. Then, their respective probabilities can be assessed relative to observations using the recognition mechanism. This process yields a probability distribution over the chosen models, and the final decision for eliminating models is left to the modeler.

References

Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14:257–262.

Anastasio, T. J., Patton, P. E., and Belkacem-Boussaid, K. (2000). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation*, 12(5):1165–87.

Arulampalam, S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filter for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2).

Banks, M. S. (2004). Neuroscience: what you see and hear is what you get. *Current Biology*, 14(6):236–238.

Battaglia, P. W., Jacobs, R. A., and Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, 20(7):1391–1397.

Bengio, Y. and Frasconi, P. (1995). An input/output HMM architecture. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems 7*, pages 427–434. MIT Press, Cambridge, MA.

Bessière, P., Ahuactzin, J.-M., Aycard, O., Bellot, D., Colas, F., Coué, C., Diard, J., Garcia, R., Koike, C., Lebeltel, O., LeHy, R., Malrait, O., Mazer, E., Mekhnacha, K., Pradalier, C., and Spalanzani, A. (2003). Survey: Probabilistic methodology and techniques for artefact conception and development. Technical Report RR-4730, INRIA Rhône-Alpes, Montbonnot, France.

Bessière, P., Laugier, C., and Siegwart, R., editors (2008). *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 of *STAR*. Springer Verlag.

Bishop, C. M. and Svensén, M. (2003). Bayesian hierarchical mixtures of experts. In *Proceedings of the ninteenth conference on uncertainty in artificial intelligence*.

Boutilier, C., Dean, T., and Hanks, S. (1999). Decision theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research (JAIR)*, 11:1–94.

Brockwell, P. J. and Davis, R. A. (2000). *Introduction to Time Series and Forecasting (Second Edition)*. Springer-Verlag.

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 5(3):142–150.

Diard, J. and Bessière, P. (2008). Bayesian maps: probabilistic and hierarchical models for mobile robot navigation. In Bessière, P., Laugier, C., and Siegwart, R., editors, *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 of *Springer Tracts in Advanced Robotics*, pages 153–176. Springer-Verlag.

Drewing, K. and Ernst, M. (2006). Integration of force and position cues for shape perception through active touch. *Brain Research*, 1078:92–100.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–33.

Frey, B. J. (1998). *Graphical Models for Machine Learning and Digital Communication*. MIT Press.

Geisler, W. S. and Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience*, 5(6):598–604.

Gepshtein, S. and Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current Biology*, 13(6):483–488.

Ghahramani, Z., Wolpert, D. M., and Jordan, M. I. (1997). Computational models of sensorimotor integration. In Morasso, P. G. and Sanguineti, V., editors, *Self-organization, computational maps and motor control*, pages 117–47. Elsevier.

Gopnik, A. and Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, 8(8):371–377.

Haith, A., Jackson, C., Miall, C., and Vijayakumar, S. (2008). Unifying the sensory and motor components of sensorimotor adaptation. In *Advances in Neural Information Processing Systems (NIPS 2008)*.

Harvey, A. C. (1992). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge.

Hauskrecht, M., Meuleau, N., Boutilier, L., Kaelbling, L., and Dean, T. (1998). Hierarchical solution of markov decision processes using macro-actions. In *Proceedings of the 14-th Conference on Uncertainty in Artificial Intelligence*, pages 220–229.

Hillis, J. M., Watt, S. J., Landy, M. S., and Banks, M. S. (2004). Slant from texture and disparity cues: optimal cue combination. *Journal of Vision*, 4:967–992.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research*, 39:3621–9.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.

Jaynes, E. T. (2003). *Probability Theory: the Logic of Science*. Cambridge University Press.

Jensen, F. (1996). *An Introduction to Bayesian Networks*. UCL Press.

Jordan, M. I. (1999). *Learning in Graphical Models*. MIT Press. Edited Volume.

Jürgens, R. and Becker, W. (2006). Perception of angular displacement without landmarks: evidence for bayesian fusion of vestibular, optokinetic, podokinesthetic, and cognitive information. *Experimental Brain Research*, 174:528–543.

Kaelbling, L. P., Littman, M., and Cassandra, A. (1998). Planning and acting in partially observable stochastic domain. *Artificial Intelligence*, 101(1-2):99–134.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Kemp, C. and Tenenbaum, J. (2008). The discovery of structural form. *Proc. Natl. Acad. Sci. USA*, 105(31):10687–10692.

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, 55:271–304.

Kiemel, T., Oie, K., and Jeka, J. (2002). Multisensory fusion and the stochastic structure of postural sway. *Biological Cybernetics*, 87:262–277.

Knill, D. C. and Richards, W. (1996). *Perception as bayesian inference*. MIT Press, Cambridge, MA.

Koike, C. (2005). *Bayesian Approach to Action Selection and Attention Focusing. Application in Autonomous Robot Programming*. Thèse de doctorat, Inst. Nat. Polytechnique de Grenoble, Grenoble (FR).

Koike, C., Bessière, P., and Mazer, E. (2008). Bayesian approach to action selection and attention focusing. In Bessière, P., Laugier, C., and Siegwart, R., editors, *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 of *STAR*. Springer Verlag.

Koller, D. and Pfeffer, A. (1997). Object-oriented Bayesian networks. In *Proceedings of the thirteenth conference on uncertainty in artifical intelligence*, pages 302–313. Morgan Kaufmann publishers.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS one*, 2(9):e943.

Körding, K. P. and Wolpert, D. M. (2004a). Bayesian integration in sensorimotor learning. *Nature*, 427:244–7.

Körding, K. P. and Wolpert, D. M. (2004b). The loss function of sensorimotor learning. *Proc. Natl. Acad. Sci. USA*, 101(26):9839–9842.

Kuipers, B. J. (2000). The spatial semantic hierarchy. *Artificial Intelligence*, 119(1–2):191–233.

Landy, M. S., Maloney, L. T., Johnston, E. B., and Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, 35:389–412.

Laskey, K. B. and Mahoney, S. M. (1997). Network fragments: representing knowledge for constructing probabilistic models. In *Proceedings of the thirteenth conference on uncertainty in artifical intelligence*, pages 334–341. Morgan Kaufmann publishers.

Laurens, J. and Droulez, J. (2007). Bayesian processing of vestibular information. *Biological Cybernetics*, 96:389–404.

Laurens, J. and Droulez, J. (2008). Bayesian modeling of visuo-vestibular interactions. In Bessière, P., Laugier, C., and Siegwart, R., editors, *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, volume 46 of *STAR*. Springer Verlag.

Lebeltel, O., Bessière, P., Diard, J., and Mazer, E. (2004). Bayesian robot programming. *Advanced Robotics*, 16(1):49–79.

Leonard, J., Durrant-Whyte, H., and Cox, I. (1992). Dynamic map-building for an autonomous mobile robot. *The Intl. J. of Robotics Research*, 11(4):286–298.

Maeda, S. (1990). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. J. and Marchal, A., editors, *Speech production and speech modelling*, pages 131–149. Dordrecht: Kluwer.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Murphy, K. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, University of California, Berkeley, Berkeley, CA.

Neal, R. M., Beal, M. J., and Roweis, S. T. (2003). Inferring state sequences for non-linear systems with embedded hidden Markov models. In Thrun, S. and al, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Nefian, A. and Hayes, M. (1999). Face recognition using an embedded hmm. In *Proceedings of the IEEE Conference on Audio and Video-based Biometric Person Authentication*, pages 19–24.

Pearl, J. (1988). *Probabilistic reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pineau, J. and Thrun, S. (2002). High-level robot behaviour control with POMDPs. In *AAAI Workshop on Cognitive Robotics*.

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research*, 41(24):3141–61.

Poggio, T. (1984). Vision by man and machine. *Scientific American*, 250:106–116.

Pradalier, C., Colas, F., and Bessiere, P. (2003). Expressing bayesian fusion as a product of distributions: Applications in robotics. In *Proc. IEEE Int. Conf. on Intelligent Robots and Systems*.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286.

Rabiner, L. R. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*, chapter Theory and implementation of Hidden Markov Models, pages 321–389. Prentice Hall, Englewood Cliffs, New Jersey.

Robinson, J. A. (1979). *Logic : Form and Function*. North-Holland, New York, USA.

Sato, Y., Toyoizumi, T., and Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: Identification of common sources of audiovisual stimuli. *Neural Computation*, 19(12):3335–3355.

Stocker, A. and Simoncelli, E. (2008). A Bayesian model of conditioned perception. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 1409–1416. MIT Press, Cambridge, MA.

Thrun, S. (2000). Probabilistic algorithms in robotics. *AI Magazine*, 21(4):93–109.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. MIT Press.

van der Kooij, H., Jacobs, R., Koopman, B., and Grootenboer, H. (1999). A multisensory integration model of human stance control. *Biological Cybernetics*, 80:299–308.

Waterhouse, S., MacKay, D., and Robinson, T. (1996). Bayesian methods for mixtures of experts. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 351–357. The MIT Press.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604.

Wolpert, D. (2007). Probabilistic models in human sensorimotor control. *Human Movement Science*, 26:511–24.

Yuille, A. L. and Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In Knill, D. C. and Richards, W., editors, *Perception as bayesian inference*, pages 123–161. MIT Press, Cambridge, MA.

Zupan, L. H., Merfeld, D. M., and Darlot, C. (2002). Using sensory weighting to model the influence of canal, otolith and visual cues on spatial orientation and eye movements. *Biological Cybernetics*, 86(3):209–230.