Cognition 123 (2012) 61–83

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Symbolic representation of probabilistic worlds

Jacob Feldman

Department of Psychology, Center for Cognitive Science, Rutgers University – New Brunswick, 152 Frelinghuysen Rd., Piscataway, NJ 08854, United States

ARTICLE INFO

Article history:
Received 3 August 2010
Revised 10 December 2011
Accepted 12 December 2011
Available online 24 January 2012

Keywords:
Mental representation
Symbols
Probabilistic models
Mixture models

ABSTRACT

Symbolic representation of environmental variables is a ubiquitous and often debated component of cognitive science. Yet notwithstanding centuries of philosophical discussion, the efficacy, scope, and validity of such representation has rarely been given direct consideration from a mathematical point of view. This paper introduces a quantitative measure of the effectiveness of symbolic representation, and develops formal constraints under which such representation is in fact warranted. The effectiveness of symbolic representation hinges on the *probabilistic structure of the environment* that is to be represented. For arbitrary probability distributions (i.e., environments), symbolic representation is generally *not* warranted. But in *modal* environments, defined here as those that consist of mixtures of component distributions that are narrow ("spiky") relative to their spreads, symbolic representation can be shown to represent the environment with a relatively negligible loss of information. Modal environments support propositional forms, logical relations, and other familiar features of symbolic representation. Hence the assumption that our environment is, in fact, modal is a key tacit assumption underlying the use of symbols in cognitive science.

1. Perspectives on symbolic representation

The structure, function, and even existence of symbolic representations has been a central issue in cognitive science ever since its inception, and often a contentious one. Philosophical perspectives on this issue have centered on the sufficiency of internal symbolic mechanisms to afford a genuinely representational ("intensional") status with respect to the world. For example Putnam (1988) has rejected a completely computational account of mental representation, meaning one that depends only on the form of symbolic expressions inside the head, on the grounds that the truth conditions of such expressions inevitably relate to conditions *outside* the head. Some connectionists have taken as a founding premise that symbolic representations are inadequate to model the dynamic, variegated and intrinsically continuous world (Harnad, 1993; Rumelhart, McClelland, & Hinton, 1986). Others (e.g. Holyoak & Hummel, 2000) have argued that symbols remain an essential and ineliminable component of mental representation. But, like many foundational controversies, this debate has featured a wide variety of conceptualizations of key terms, impeding a clear understanding of exactly *how* symbols might contribute (or, alternatively, fail to contribute) to mental representation. In this paper, I consider this problem from a mathematical point of view, attempting to quantify the *information* that symbolic representations capture about the environment, and the fidelity with which they capture it (cf. Dretske, 1981; Usher, 2001). To preview the argument, the fidelity of symbolic representations turns out to depend heavily on what we assume about the environment. In some environments, symbolic representations make demonstrably faithful models, while in others, they do not. This paper attempts to understand the factors that modulate the degree of fidelity, and thus to shed light on the foundations of the symbolic representations that are so ubiquitous in cognitive science.

Roughly speaking, symbols are discrete mental representations that reliably correspond to stable, distinguishable entities in the world. But very little in this vague phrase admits to a precise definition. A particularly basic case of symbolic representation, which nevertheless retains most of the definitional difficulties of the general

E-mail address: jacob@ruccs.rutgers.edu

0010-0277/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.12.008

case, is that of Boolean or other discrete-valued features.[1] These are variables that take on several distinct levels or values, like *big/little*, *on/off*, or *animal/vegetable/mineral*. Discussions of symbolic representations in cognitive science often devolve into debate about the aptness or "naturalness" of such discrete features, as compared to corresponding continuous ones (*tall/short* for *height*, *heavy/light*) for *weight*, etc.). Discrete-valued features are sometimes derided as unnatural on the grounds that classical physics employs continuous variables exclusively as its underlying parameters (e.g. position, mass, time). Yet this accusation lacks a firm empirical foundation; how, after all, do we know exactly what is objectively "natural" independently of the choice of variables we use to measure it? Conversely, affirmations that *some* naturally-occurring variables seem essentially Boolean (*male/female, inside/outside*) seem equally feeble, for precisely the same reason. What is the principle at work here? When are variables intrinsic to the environment, and when are they "merely" approximations?

In what follows I pursue a mathematical rather than a philosophical perspective on this question, though I will occasionally draw attention to salient connections to traditional philosophical questions. The main emphasis will be on whether, and to what degree, symbolic representations preserve *functionally useful* information about the outside world. Nevertheless the thrust of the argument shares with many philosophical treatments an emphasis on the *environment* as the source of validation for mental representations. But in contrast to many accounts, here the question "Are symbolic representations legitimate?" will turn out to have a range of answers, which depend on the probabilistic structure of the environment.

2. Mixtures and modality

We begin with the intuition that some Boolean variables seem more "natural" than others, in the sense that they represent more effective summaries of their continuous counterparts. Many Boolean variables are derived from related continuous variables, e.g. by dividing them at some threshold; *tall*/*short* might really mean (*height* $\geqslant$ *six feet*)/(*height* < *six feet*), and so forth, although such thresholds are notoriously context-sensitive (Shapiro, 2006). The key idea in all of what follows is that how useful such a classification is cannot be determined in a vacuum, but rather depends on the way the continuous variable is distributed in the environment—that is, on the structure of the probability distribution that governs it. For example, if this distribution of *height* is conspicuously bimodal (Fig. 1a), that is, has two distinct peaks, then it seems well-justified to treat *height* as approximately dichotomous. But conversely if height is unimodal or uniform (Fig. 1b), such a Booleanization seems arbitrary.

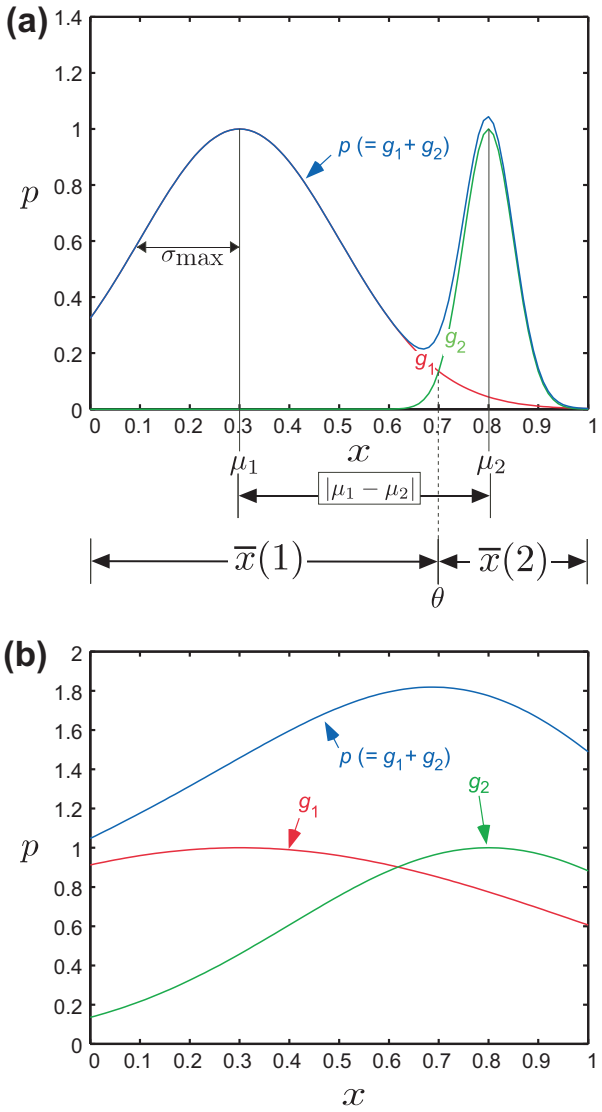

Fig. 1. Given a conspicuously bimodal density (panel a), one feels justified in treating the underlying variable x as an approximately Boolean variable $\bar{x}$ having two distinct values or bins $\bar{x}(1)$ and $\bar{x}(2)$ (indicated along the x axis). But discretization does not seem similarly reasonable with a unimodal density (panel b), even though as shown it could also be the sum of distinct (but poorly separated) sources. The figure illustrates how either observed distribution $p(x)$ could be the mixture (sum) of two component distributions, labeled g_1 and g_2. In panel a, the components are well-separated, in that the distance $|\mu_1 - \mu_2|$ between their respective means μ_1 and μ_2 is large compared to the larger of their standard deviations $\sigma_{\max}$. This results in a visibly bimodal mixture in which each of the symbols ($\bar{x}(1)$ and $\bar{x}(2)$) refers to one of the mixture components (g_1 and g_2). In panel b, the distributions are poorly separated, resulting in a unimodal mixture, no discretization, and no such reference. These mathematical aspects will be explained more thoroughly later in the paper.

Extending this idea, in what follows we develop the *degree of modality* ("spikiness") as the parameter that modulates the effectiveness of symbolic representation. The basic reasoning is illustrated by this simple one-dimensional example, but nevertheless the most interesting aspect of this approach turns out to be what happens

[1] The intended scope of the term "symbol" in this paper will become more clear as the argument unfolds, but it should be understood that not all senses of the word necessarily fall within it. What is meant here is *discrete* symbols, meaning mental tokens that are intended to correspond to individual phenomena or attributes in the world. The symbols used in algebraic operations, such as the x and y in the expression $x + y$, correspond to continuous variables, and thus to infinite collections of states of the world; these fall outside the intended scope. Of course, these various senses of "symbol" are related. The section below entitled *Observability* develops some of these connections, suggesting how the representation of discrete states informs the choice of continuous parameters.

in higher dimensions, where the situation expands richly. In higher dimensions, instead of a simple probability distribution we have a potentially more complicated *joint* probability distribution among multiple variables, and the symbols derived from these variables must *combine* to represent it. Broadly speaking, the paper is simply an extended elaboration of the idea of modality—that is, "spikiness" in the probability distribution at work in the environment—and its role in rendering mental symbol systems coherent and useful.

In keeping with the viewpoint conventional in the natural sciences, we will begin by regarding the environment as a system of bounded (non-infinite) continuous variables $X = \{x_1, x_2, \ldots, x_D\}$ governed by a probability density function (PDF) $p(X)$, which assigns probabilities to states of the world. In order to proceed I will assume that the set of parameters X constitutes a closed, comprehensive definition of the world under consideration. Of course it should be borne in mind that this is merely a working assumption, and that other parameters outside the range of our analysis may well exist; we can hope that other parameters will be systematically related to those in X, but we cannot guarantee it. (Partly for this reason, most of the properties we will be concerned with below are invariant to smooth transformations of the parameters, which protects us from too much dependence on our choice of parameterization.) But with this caveat in mind, we assume that X fully encodes the environment under consideration, and $p(X)$ describes *what tends to happen* in this world, in the sense that it says how often each state tends to happen. Most of the following discussion concerns what is reasonable or useful to assume about the structure of this PDF, and how symbolic representations of this structure succeed or fail as a result of these assumptions.

The term *modality* has often been used to refer to environmental regularity and its role in cognition, perhaps most overtly in Richards' "Principle of natural modes" (Jepson, Richards, & Knill, 1996; Richards, 1988; Richards & Bobick, 1988), closely akin to Shepard's notion of environmental regularities (Shepard, 1994) and to Barlow's (1961, 1974, 1990, 1994) ideas about their role in informing the neural code. The grandparent of all these ideas is perhaps Hume's Principle of the Uniformity of Nature, with which they share an overarching emphasis on structure and regularity in the environment as the ultimate source of its comprehensibility by the mind. The exact meaning of "regularity" in this context is potentially vague, but has received a number of different technical definitions, including the tendency for natural parameters to "clump" at special values (i.e., modes); and the tendency for natural parameters to correlate with one another. However these tendencies are difficult to quantify precisely, and the precise nature of their relationship to mental structures has never been fully explicated.

In this paper I attempt to realize the idea of modality via the technical instrument of *mixture distributions*.[2] A mixture is a probability density function that is composed of some number K of distinct *components* or *sources* from which observations may be drawn (see McLachlan & Basford, 1988; McLachlan & Peel, 2000; Titterington, Smith, & Makov, 1985). Typically, each source has a distinct mean $\mu_i \in R^D$, variance σ_i^2, and prior probability w_i of being chosen as the source of any given observation $\mathbf{x}$. Thus for example a set of fish drawn from the river might be a mixture of two species, and thus the observed distribution of lengths and weights might be a population-weighted mixture of the two individual species' distributions. Similarly, the set of objects in front of you on your desk might be a mixture of books, papers, and pens, with a corresponding multicomponent mixture of shapes, sizes, colors, or whatever other parameter you choose to measure. We will generally assume that each source is unimodal (like a Gaussian or normal density), meaning that each has a single most probable value, with other values diminishing in probability around it. Pearson (1894) was perhaps the first to clearly recognize the importance of decomposing observations into their distinct generating sources, when he decomposed a set of crab forehead measurements into a mixture of two distinct Gaussians, inferring (as it happens, correctly) the emergence of two distinct species. Similarly, in what follows I will argue that mental symbols "effectively" represent the environment when they (or in higher dimensions, combinations of them) correspond to individual components of the mixture present in the environment.

Technically, mixtures provide a variety of interesting challenges. In general the observer does not know the true source for any given observation, nor even the number of sources, but rather must estimate them from observations, which makes it difficult to formulate an accurate predictive model of future observations. Mixtures make a convincing model of many natural cognitive situations, because they capture the general idea of a heterogeneous combination of sources that the world confronts us with: a disjunction of categories, events, objects, and other stable aspects of the world that are all combined into in a single complex stream as they are propelled at our senses. In order for the observer to understand the environment and reason about it, it is first necessary to disentangle this stream into the coherent ensemble of environmental regularities that actually generated it. The main idea of this paper is that mental symbols make sense when they correspond reliably to distinct components of the mixture that governs the environment.

I will refer to PDFs generated by mixtures as "modal", meaning literally that they are constructed from a set of distinct modes or peaks. However it is essential in what follows to note that separate peaks are often *not* actually distinguishable in mixture densities or in the samples drawn from them, when the modes are close to each other relative to their spreads and thus collapse into a single peak. Fig. 2 shows several examples of mixtures of various levels of modality or separability. (The figure also illustrates the measure of modality introduced later in the text.)

A note on ontology. It is natural to ask whether the distributions we will speak of as defining the environment—and in particular the mixture components—are in fact "in

[2] Also called *mixture densities* (when the parameters are continuous) or simply *mixtures*. The most common term is *mixture models*, but this is more properly reserved for methods for estimating a mixture rather than the mixture itself.

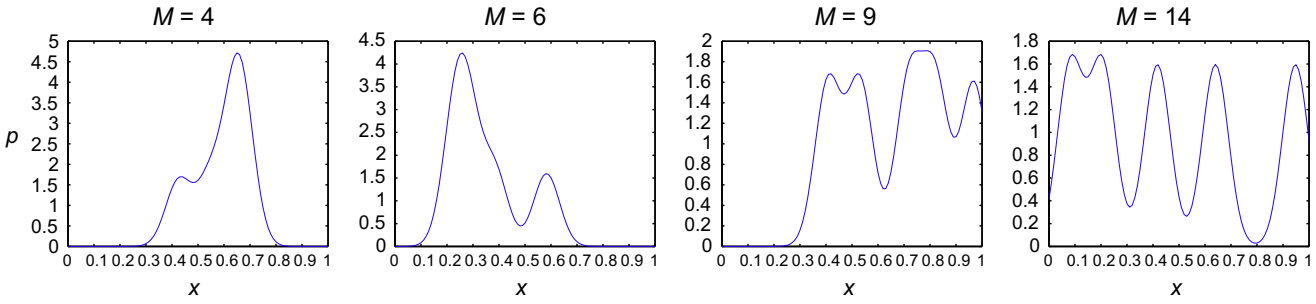

Fig. 2. Examples of mixture PDFs, each with five components, showing various levels of modality M (see text). Each PDF tabulates probability along the parameter x. Note that at low levels of modality (left), the five components tend to blend together (fewer than five modes are visible), while at higher levels of modality (right), all 5 are increasingly distinguishable, leading to a symbolic representation that corresponds closely to the generating sources.

the world" or are better thought of simply as descriptions of reality and thus "in the head". For ease of exposition, I will speak of them as if they were real, but it is probably best understood that this is only a convenient way of getting started. In my own view, pronouncements about the "true" properties of Nature are generally meaningless, as every description of reality embodies a system of assumptions about what it actually consists of (see Hoffman, 2009). In this sense probability models in general, and mixtures in particular, are best thought of simply as familiar and technically well-developed tools by which science may describe reality. Thus in what follows when I write "if the world is composed of a mixture... " the reader should understand something like "if the world can be described with reasonable observational fidelity by a mixture... ". I have chosen to use mixture distributions to play the role of "objective reality" in this paper because they are intuitive and conventional, and will already be familiar to statistically-minded readers, not because they are "correct". The aim is to show that this simple premise leads some interesting conclusions.

2.1. Synopsis of the paper

The goal of the paper is to investigate the consequences of modality in the environment—meaning its generation by a mixture—for an observer attempting to represent and comprehend it. The main conclusion is that such environments are capable of being represented by symbols and combinations of symbols, while *arbitrary* environments—that is, worlds that are statistically typical of the set of all possible worlds—generally are not. In this Section 1 give a brief conceptual overview of the paper, emphasizing basic principles and intuitions. Subsequent sections flesh out the argument in more mathematical detail, although in a way that is still intended to be readily comprehensible by a wide audience. Mathematical details, including derivations of the results presented in the text, are in the appendices (labeled Appendices A.1–A.6).

To preview the flow of the argument, first consider the one-dimensional case. As in the example above, if one takes samples of the world along a *single* dimension—selecting a set of objects and measuring them along some fixed yardstick x, and tabulating the results—one may find multiple modes or peaks in the resulting distribution. It is natural to refer to the distinct modes by distinct symbols, such as the discrete values of a discrete variable, and assume that they correspond to distinct phenomena in the world. (Mixture distributions are simply a formalism for expressing this idea.) In this sense a symbolic description represents a compact summary of the information in the original measurements. The first part of the paper below describes technical conditions for this reduction to be reasonably faithful. The main result is that the more modal or "spiky" the environment, in a technical sense defined below, the more faithful is the corresponding symbolic representation.

The situation becomes much more interesting when one considers *multiple* dimensions, where the topography of the modes grows substantialy more complex (Ray & Lindsay, 2005). Fig. 3 shows several examples of multimodal worlds in two dimensions. As in 1D, while some worlds are cleanly separable into their component peaks (Fig. 3a), others are less so (Fig. 3b), because their components are either too broad or two closely spaced to be easily discriminated. Now, each of the individual dimensions has a PDF of its own, the *marginal PDF*, which tabulates probability along that dimension while integrating over the other one (see figure). The marginal PDF $p(x_1)$ can be thought of as the *projection* of the world onto x_1 (Diaconis & Freedman, 1984; Friedman & Tukey, 1974), or, equivalently, the viewpoint of an observer who is sensitive only to x_1 and cannot "see" x_2.

Consider what each of the worlds pictured in Fig. 3 looks like from the viewpoint of one axis, say x_1. Because the joint PDF is modal, the marginal density $p(x_1)$ will also tend to be modal (and likewise $p(x_2)$). But some components that are plainly visible in the "God's-eye view" of the joint PDF will no longer be distinguishable in the marginal density $p(x_1)$, because they align or nearly align along the "line of sight", and thus collapse in the marginal PDF. By the argument sketched above, the modality along x_1 means that it can be approximately reduced to a discrete symbolic variable. Likewise, the other dimension x_2 will also reduce to a (different) symbolic variable. The full joint PDF then corresponds to *some logical combination* of these two symbolic variables. But *what* combination best represents the joint PDF, and how well does it do so? That depends on how the geometry along one dimension relates to the geometry along the other dimension, that is, it depends on the relative placement of the modes in the joint PDF. The effectiveness of the entire system of symbols in

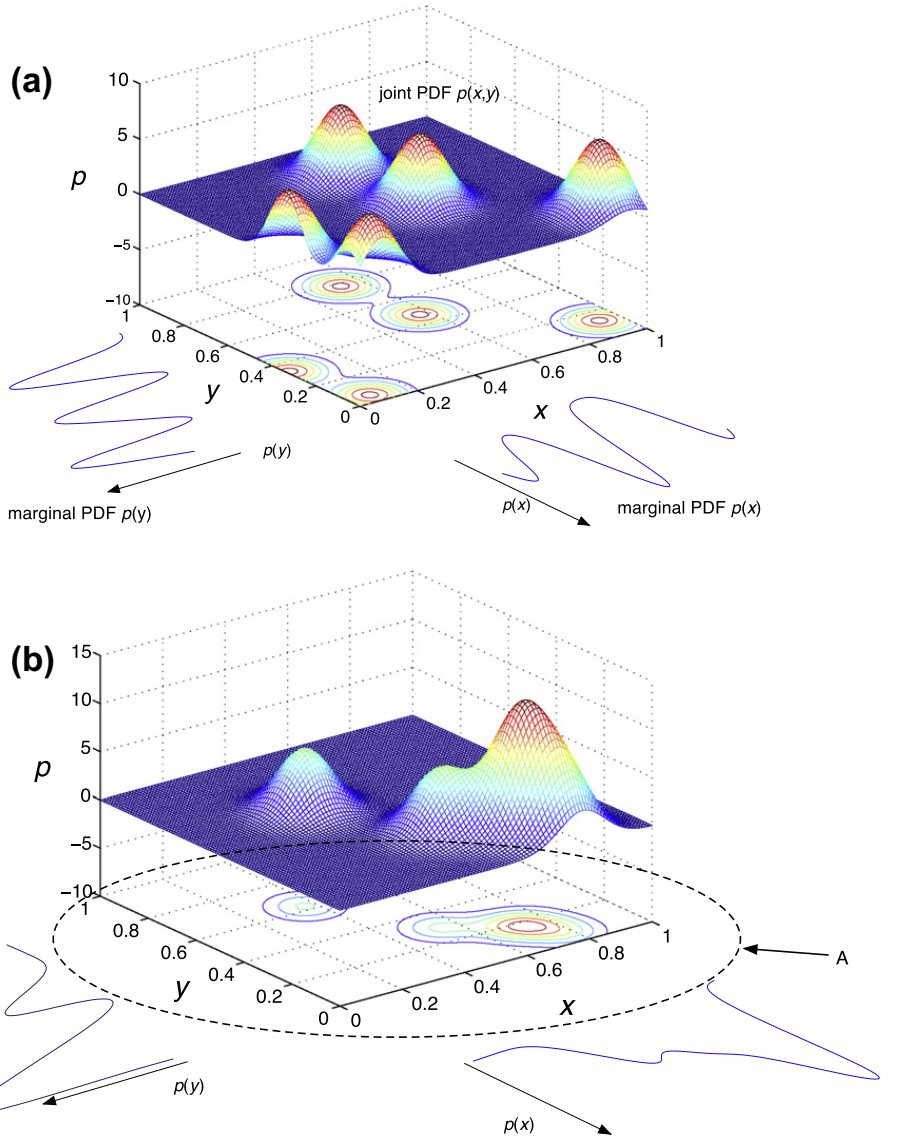

Fig. 3. Two worlds in two dimensions, each containing five modes ($K = 5$) with different levels of modality M: (a) a more modal one ($M = 135$), and (b) a less modal one ($M = 75$). Note that all five components cannot be distinguished in panel (b). The figure also illustrates the marginal PDFs $p(x)$ and $p(y)$, which collapse several of the modes. Panel (b) also illustrates the difference in perspective when the mixtures is "viewed" (marginally projected) from an alternative direction (A) instead of the axes. The dotted circle indicates the range of possible viewpoints, i.e. the "observer hypersphere" explained later in the text.

representing the entire probabilistically defined world depends on the nature of geometrical relations among the modes in the mixture. Much of this paper is devoted to exploring these geometrical relations, and the impact they have on symbolic representations of the joint PDF.

Before proceeding further, it is worth remarking on how profoundly ambiguous this situation is. To an observer attempting to discover the structure of the world by sampling only individual features—like the fabled blind men and the elephant—the geometry of the modes inside the space is unknown. A simple metaphor helps make the nature of the ambiguity clear. Think of the world as a "cloud" of unknown internal structure, which we are attempting to probe by taking a set of measurements (distinct yardsticks; see Fig. 4). Prior to taking measurements, not only is the world's structure unknown, but—without some strong assumptions—so is the *relationship* among the various measurements. Do they represent substantially similar information, or substantially independent information? One cannot know *a priori*. This is Davidson (1973)'s notion of *radical interpretation*, the puzzle of translation across languages, applied to the interpretation of PDFs. Like speakers of foreign languages without intermediaries, the multiple observers of the cloud (multiple yardsticks) cannot be sure if their referents correspond.

How can the readings taken from the various probes be combined or compared? If the first measurement x_1 reveals (say) three modes (as illustrated in Fig. 4), suggesting three distinct structures within the cloud, and the second x_2 *also* reveals three modes, can we feel confident that the three modes in x_1 correspond to the three in x_2? Without any assumptions, the answer is no. The world might contain as few as three modes (if the modes on x_1 corresponded exactly to the modes on x_2); or six (if the three on x_1 were

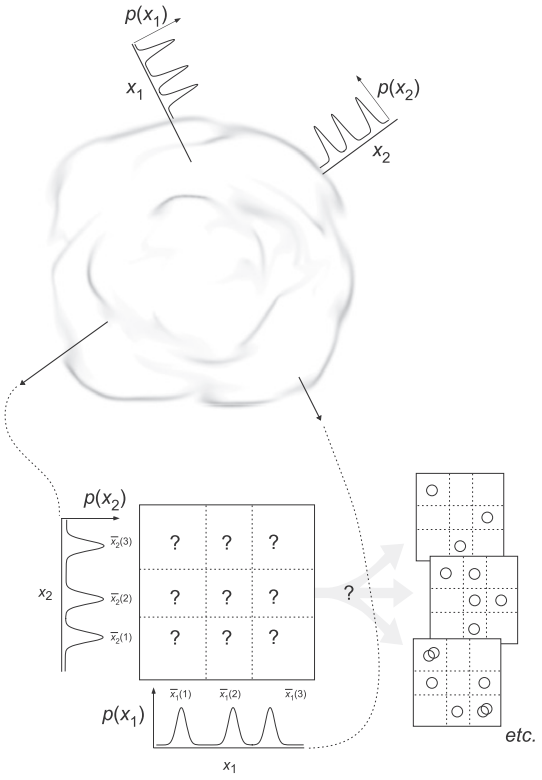

Fig. 4. The "cloud" metaphor. Prior to observation, the world consists of joint PDF of totally unknown structure. We probe it by collecting some number of measurements (here, x_1 and x_2), sampling the frequency of values along them. We may find modes along each of them (three each in the example), which can be represented by discrete symbols ($\overline{x_1}$ and $\overline{x_2}$). But *a priori* we do not know the relation among the various symbols, because we do not know the relations among the various modes inside the cloud (the PDF). The goal of the paper is to understand the conditions under which combinations of symbols effectively represent what is going on within the cloud.

distinct from those on x_2); or, in fact, many more (see Fig. 4). (Recall that what appears to be a single mode on one dimension might resolve to two more on the other.) Multiplying this situation with larger numbers of measurements, with potentially more complex relations among them, simply adds to the perplexity. Without some assumptions about the structure inside the cloud, we simply cannot make sense of the measurements we make. To solve this problem—and thus to establish how representations derived from individual yardsticks can be combined to create an accurate representation of the world—requires an understanding of how the *logical relations* among the various symbols relate to the *geometric relations* among the modes hidden within the cloud. On the main goals of the paper is to unravel the possible geometric relations and establish their correspondence to familiar logical forms.

The remainder of the paper lays out the above argument in more technical terms, developing the mathematical substance underlying it. The key idea is to establish mathematical criteria for a world, i.e., a probability distribution of unknown structure, to be representable by symbols. Again, it turns out that we can relate the degree of modality of the world to the degree of fidelity of the corresponding symbolic reduction. In one dimension, the basic idea was that a feature is symbolically representable if the best symbolic representation of it represents it sufficiently faithfully—that is, loses little enough information. This is generally possible if the world is sufficiently modal, generally impossible if it is not. In multiple dimensions, the situation is more complicated. In order for the world to be symbolically represented by a system of symbols, not only does it have to have modes, but the modes have to relate to each other in a particular way.

The next section of the paper treats the 1D case, developing an information-theoretic criterion (called ϵ-representation) under which continuous features can be effectively discretized; this establishes conditions for the validity of individual symbols. Subsequent sections treat the multidimensional case, establishing analogous criteria for the representation of multidimensional joint probability densities. As mentioned, this turns out to be a far richer and more interesting situation, involving how the symbols that arise from the individual dimensions *combine* with each other to form complex representations of the joint PDF. Exactly how well they can represent it depends on the qualitative geometric relations among the mixture components within the joint PDF, which can get complicated in multiple dimensions. The basic contribution of this paper is to show how symbolic representations are related to the geometry of mixtures—how the combinatoric possibilities of symbols relate to the geometric relations among the corresponding modes. Then we consider how standard elements of symbolic representations, such as propositional formulae and logical relations, relate to the multidimensional geometry of the mixture components in the probabilistic environment that they help represent. Finally, we consider how modality in the multidimensional distribution relates to the "meaningfulness" of the features themselves.

3. Discrete symbols in one dimension

This section considers the situation depicted in Fig. 1 in more mathematical detail, establishing criteria for a continuous variable to be treated "effectively" as a discrete symbolic variable. The basic idea, ϵ-representation, is that symbolic representation serves as a reasonably faithful approximation if the underlying probability distribution is modal in the sense discussed above.

Consider a single continuous feature x (e.g. *height* or *weight* in the examples mentioned above). We assume for concreteness that x runs over the unit interval $[0,1]$, which for ease of calculation we quantize by sampling at a large number N of small equal intervals. (This is simply a mathematical convenience and should not be confused with the coarser discretization which is the main focus below.[3]) A mixture consists of a set of K ($\ll N$) sources with means $\mu_i \in [0,1]$, standard deviations σ_i, and weights (mixing proportions) w_i which sum to one $\left(\sum_i w_i = 1\right)$. We make no assumptions about the functional form of these sources (e.g. Gaussian or otherwise) except that they are unimodal

[3] We could avoid the quantization by using uncertainty as defined over continuously defined probability density functions, called differential entropy. But differential entropy has a number of peculiarities, like dependence on the parameterization, and the possibility of negative values, which would unnecessarily muddy the exposition here.

and have finite means and variances. The mixture p is then simply the weighted sum of the components,

$$p(x) = \sum_{i=1}^{K} w_i g_i(x). \quad (1)$$

How well can x be approximated by a discrete variable? First, it is important to see that if the components are closely spaced relative to their spreads, the mixture may be very difficult to separate. McLachlan and Basford (1988) suggest as a rule of thumb that when two components' means differ by less than twice their common standard deviation, their mixture will actually be unimodal (have only one maximum). Exactly where between two components a boundary will be found, and indeed whether two sources can be separated at all, will depend on the nature of the discretization method employed. Many methods have been developed (e.g. Bay, 2000; Dougherty, Kohavi, & Sahami, 1995; Fayyad & Irani, 1993; see Dy, 2008 for a useful review). Here though we are not concerned with the details of the method, but with the effectiveness of the resulting discretization, in a sense to be defined. All that we need to assume for the ensuing argument is that in general sources are more easily distinguished when they are further apart, and that as they are spaced more closely they eventually become practically impossible to separate given finite data, which is true for all known methods.

3.1. A measure of modality

Intuitively, mixtures are natural candidates for discretization, with the values of the discrete variable corresponding to the distinct components of the mixture. The aim of this section is to unpack and quantify this intuition.

A variable x acts something like a discrete variable when the mixture $p(x)$ is very "spiky", that is, when the σs are narrow relative to the intervals between the means μ_i (Fig. 1a). With $K = 2$, a conventional measure of the degree of separation between the modes is Cohen's d

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}, \quad (2)$$

(Cohen, 1988), often used as a measure of the size of a statistical effect, here the size of a mean difference relative to the noise in the measurements (signal to noise ratio).[4] The demoninator σ represents the common standard deviation of the two modes, or usually their root mean square if they are unequal (usage in the literature varies depending on the situation). Cohen's d is high when the two distributions are well-separated relative to their spreads, and low when they substantially overlap, going down to 0 if they coincide. Intuitively, a mixture of two well-separated modes (high Cohen's d) is effectively Boolean, and in fact is treated so by human subjects (Aitkin & Feldman, submitted for publication). Our immediate aim is to quantify this idea and generalize it to larger numbers of modes.

For more than two modes ($K > 2$), a natural generalization of Cohen's d is to replace the distance between the μs in the numerator with the overall *spread* among the component means, quantified by the standard deviation of the ensemble (the μ_i), defined by

$$S = E[(\mu_i - \bar{\mu})^2]^{1/2}, \quad (3)$$

where $E(\cdot)$ indicates the expectation or average value. We then define the *modality* M as

$$M = \frac{2S}{\sigma_{\max}}, \quad (4)$$

that is, twice the ratio of the spread S to the largest component standard deviation $\sigma_{\max}$. Loosely speaking, M measures how spread out the modes are relative to their internal spreads, which determines how cleanly separated they are. Note that M reduces to Cohen's d in the case of two modes (see Appendix A.1). Fig. 2 shows several examples of worlds with various levels of M. At high values the mixture is very modal or "spiky" and all the modes are clearly visible, while at low values the components tend to blend together and are no longer plainly separable.

3.2. Mixtures can be effectively discretized

We next ask how well the a continuous variable x can be discretized—binned and treated as a discrete variable—as a function of the modality M of its governing density $p(x)$. Intuitively, when M is low, the resulting mixture becomes very homogeneous and difficult to separate (Fig. 1b), because the components overlap (so any given x has a substantial probability of having been generated by more than one source). At the other extreme, when M is high, the distribution become extremely spiky: each x can be readily classified as originating from a particular distinct source, and the mixture more and more closely approximates a discrete variable (Fig. 1a). In intermediate cases, sources may overlap to an intermediate degree, making the resulting distribution *somewhat* but not perfectly discrete. This means that the modality parameter M modulates the degree to which it is "reasonable" to treat x as a discrete variable.

A *discretization* of x is a partitioning of x into "bins", not necessarily of equal width, with each bin treated as a distinct value of a new discrete variable denoted $\bar{x}$, with the bins denoted $\bar{x}(1), \bar{x}(2) \ldots \bar{x}(K)$. Our aim is to quantify the degree to which the continuous variable x can be effectively captured by its discrete counterpart $\bar{x}$. When we discretize, we lose some information, because we are throwing away the precise original value of x. But depending on the nature of the PDF $p(x)$, we may not be throwing away *very much* information. To quantify this more precisely, we measure the *Shannon uncertainty* of the value of x once the discretized value $\bar{x}$ is known. Shannon uncertainty, the basic measure of information in the modern theory of information (see Cover & Thomas, 1991) quantifies the degree to which a signal improves the receiver's state of knowledge. (See Dretske, 1981; Usher, 2001 for other applications of

[4] Cohen's d is closely related to the response measure d' (d prime) familiar from signal detection theory (Green & Swets, 1966). Like Cohen's d, d' is a ratio of signal (separation between peaks) to noise (variability), but it usually is computed from a sample of responses rather than from the parameters of the generating distribution, as here. Hence to minimize confusion I will henceforth avoid this terminology.

information theory to problems of mental representation, and Feldman & Singh (2005) and Resnikoff (1985) for applications to perceptual representations.) For any PDF p, the uncertainty $H(p)$ is given by $H(p) = -\sum p \log p$. In our context, the uncertainty contained in our original parameter x before discretization is

$$H[p(x)] = H(x) = -\sum_{j}^{N} p(x_j) \log p(x_j). \quad (5)$$

(From here on, when the choice of PDF is unambiguous we will abbreviate $H[p(x)]$ to $H(x)$.) But once x has been discretized into bins $\bar{x}(i)$, and the value i of the discretized variable is known, the uncertainty becomes

$$H[x|\bar{x}(i)] = -\sum_{j \in \bar{x}(i)} p(x_j|\bar{x}(i)) \log p(x_j|\bar{x}(i)), \quad (6)$$

which sums up the uncertainty within the ith bin—i.e., the uncertainty that remains once we know what bin x falls in. This represents a reduction in uncertainty compared to before the binning, because the remaining possible values of x have been narrowed. In what follows we will focus on the *average* value of this uncertainty across bins, referred as the *symbol uncertainty*, and defined as

$$E[H(x|\bar{x}(i))] = -\sum_{i=1}^{K} p[\bar{x}(i)] H[p(x|\bar{x}(i))], \quad (7)$$

that is, the sum of the uncertainties in each bin $H[p(x|\bar{x}(i))]$ weighted by the probability of each bin $p[\bar{x}(i)]$. For brevity the symbol uncertainty will be notated $H(x|\bar{x})$, with the omission of the subscript i indicating that we have taken the expectation across all bins.

The symbol uncertainty is the uncertainty about the state of the world *after* we know what its symbolic representation is: for example, your uncertainty about a person's height after you have been told that he or she is "tall", (or "short"—averaging across both cases); or your uncertainty about the location of your car keys once you find out that they are somewhere in the kitchen (or in some other room, etc.). The symbol uncertainty quantifies how much, on average, is still unknown about the true state of the world x once we know what bin x falls in—that is, how much uncertainty remains about the world after we know its symbolic representation. If this expected uncertainty is sufficiently small, it is reasonable to say that the discrete variable $\bar{x}$ "effectively captures" the true state of the world x, in that it represents it with only a negligible residual uncertainty.

3.3. ϵ-representation

This motivates the following definition. We say that the discrete variable $\bar{x}$ ϵ-represents the continuous feature x if the symbol uncertainty is less than ϵ,

$$H(x|\bar{x}) < \epsilon. \quad (8)$$

(for some arbitrary threshold ϵ), and likewise we say that a particular world $p(x)$ is ϵ-representable if there exists a non-trivial discretization $\bar{x}$ that ϵ-represents it.[5] A world $p(x)$ that is ϵ-representable is capable of being symbolically represented with negligible loss of information (that is, with loss bounded by ϵ). This means that an observer who represents it that way is approximately "right".

In the optimal discretization of a modal world, each of the K values of the discrete variable would correspond to one of the K generating sources. This tends to happen as modality M grows large, and the components each become increasingly spiky and well-separated. In the limit, as M goes to infinity, each mode becomes an infinitely narrow spike, and the uncertainty within each bin goes to zero, because all the probability mass within it is located at one position (μ). In this case, once one knows the bin, no uncertainty remains about the actual value of x. In technical terms, as M goes to infinity, $H(x|\bar{x})$ goes to zero.

At the other extreme, when M is 0, the components, though they exist, overlap completely with each other, so knowing the value of $\bar{x}$ provides *no* information about the true value of x. In this case no symbolic representation of p is more useful than any other, and this world is not effectively representable by symbols.

Less obvious, but more revealing, is the general case, where M is somewhere in between 0 and infinite; here the components overlap somewhat but not completely. Here, the goal is to express the symbol uncertainty of p as a function of its modality M—that is, to quantify just how effectively the world can be represented symbolically, as a function of how modal it is. It can be shown (see Appendix A.2) that for a mixture with K sources and modality M, the symbol uncertainty is bound by an expression of order

$$H(x|\bar{x}) \leqslant O(K, -\log M). \quad (9)$$

meaning that the bound on symbol uncertainty rises as a linear function of the number of mixture components K, and decreases with the logarithm of the modality M. (The notation $O(\cdot)$ means "on the order of"; see Appendix A.2 for details of the bound.) This limit depends on the fact that though the components may overlap, the *magnitude* of the overlap is guaranteed to decrease as modality M increases and the modes get more separated. Eventually as M gets large enough this results in a highly modal distribution with small residual symbol uncertainty. Intuitively, the more components there are (larger K), the more they tend to overlap; but the spikier they are (larger M), the less they tend to overlap. Hence the more modal the world is, and the fewer components it has, the more effectively it can be represented by symbols.

3.4. Uncertainty of a mixture

The previous section established that once we know the state of a modal world symbolically, relatively little uncertainty about it remains. This argument can be extended to show that the *total* uncertainty $H(p)$ of a modal world tends to be low. This is because the total uncertainty of p is simply the sum of the symbol uncertainty plus the uncertainty in the symbolic representation itself (that is, how much you know when you know what symbol applies). The total amount of information in p is the information in its symbolic representation plus the infor-

[5] A trivial discretization is one with K approaching N, in which case the discretization is not really "discrete" at all; see below.

mation the remains once its symbolic representation is known.

In mathematical terms, because $p(x) = p(x|\bar{x})p(\bar{x})$, the total uncertainty of p is simply the sum of the average uncertainty within each bin (that is, the symbol uncertainty) plus the uncertainty about which bin x is falls in, $H[p(\bar{x})]$, which has expectation $\log K$ (averaging over all distributions). This latter quantity is the uncertainty inherent in the symbolic variable itself ($\bar{x}$), which the observer automatically takes on by representing the environment symbolically. This value is small as long as the number of levels K is small, while the symbol uncertainty (the remaining uncertainty after the symbol value is known) will be small if the environment is modal, i.e. M is high. Specifically, for a mixture p of K sources g_i and modality M, the average (expected) total uncertainty is bound by

$$E[H(p)] \leqslant H(x|\bar{x}) + \log K, \tag{10}$$

which again rises about linearly with K and decreases with $\log M$.

As a concrete illustration, Fig. 5 shows the actual computed uncertainty of 500,000 simulated mixtures of Gaussians, plotted alongside the theoretical bound derived above, both plotted as a function of K and M. The bound plainly tracks the computed uncertainty values, confirming that the derived dependencies on K and M are correct when applied to real PDFs.[6]

In summary, modal worlds can be effectively discretized; and the more modal they are, the more *effectively* they can be discretized. For high M and low K, mixtures can be effectively represented by symbols; they are ϵ-representable with ϵ dependent on M and K. With environments in the real world, modes may overlap, and consequently symbolic representation may be imperfect, because the symbols will not refer quite perfectly to the corresponding sources. But if the world is sufficiently modal, the imprecision of symbolic representation is modest (bounded by ϵ), and symbolic representation is effective. Modal worlds have low uncertainty because they have good models—namely, symbolic ones.

3.5. But most distributions cannot be effectively discretized

But cannot *any* distribution $p(x)$ be discretized? Yes, but not effectively. Mixtures are, in this sense, very atypical. To see this, consider an *arbitrary* density $p(x)$. By "arbitrary" we mean one without any particular special structure—such as being a Gaussian, being a mixture, etc.—but that instead is statistically "typical" of the entire set of possible distributions. This is an enormous set exhibiting a vast variety of structures, but information theory allows us to characterize in general terms the *average* properties of its members.

Specifically, it can be shown (see Appendix A.4) that most possible PDFs are approximately uniform, and thus have uncertainty about $\log N$. (Recall that N is the number of steps into which we have quantized the parameter x.)

[6] The bound is substantially higher than the computed values because it makes no distributional assumptions, whereas the simulation contains mixtures of Gaussians.

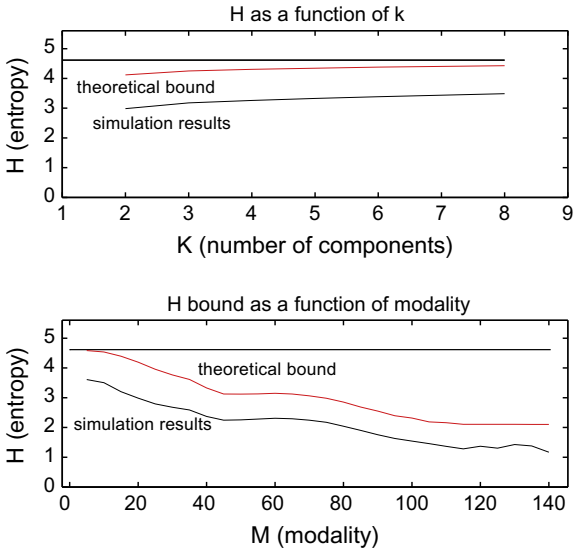

Fig. 5. Uncertainty of mixtures of Gaussians as a function of the number of components (K, top) and the modality (M, bottom). The black curves illustrate the numerically calculated Shannon uncertainty drawn from 500,000 randomly chosen mixtures, while the red curves illustrate the theoretical bound derived in the text (Eq. (26), Appendix A.2). The bound is generally higher than the simulated values because it does not presume Gaussian sources. Note that as K increases or M decreases, both the mixtures and the theoretical bound eventually hit the absolute theoretical bound of $\log N$ (black line), at which point they are effectively non-modal and symbolic representation ceases to be valuable. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Obviously, many distributions are extremely non-uniform and will thus have much lower uncertainty—including the modal ones that are the main focus of the paper—but statistically such cases are greatly outnumbered by the nearly uniform ones. Putting this another way, if one were to imagine choosing a world "at random", with no constraint (see Appendix A.4 for an explanation of exactly what this means), the world would most likely have nearly maximal uncertainty. This means that statistically typical environments are not generally ϵ-representable—they cannot be effectively represented by symbols.

But as shown above, the uncertainty of a mixture is relatively small (Eq. (10)). The contrast becomes more and more extreme as N increases (we quantize the world more finely), K decreases (the world has fewer modes) or M increases (the world gets more modal). The more modal the world is, and thus the more effectively it can be represented by symbols, the more statistically atypical it is. In this sense, representing the environment symbolically—as in many theories of cognition—constitutes a substantial commitment to the assumption that it is modal.

This is really a special case of the idea of Kolmogorov complexity (Chaitin, 1966; Kolmogorov, 1965; see Chater & Vitányi, 2003; Li & Vitányi, 1997), which entails that only a small fraction of possible structures are compressible (capable of being represented by short strings of symbols) while the vast majority are incompressible and thus essentially random. Analogously, here we have shown that only a small fraction of *worlds* are ϵ-representable (capable of being faithfully represented by symbols) while the vast

majority are not. This is a counting argument, not a probabilistic assumption; we are *not* making any assumptions about how often either type of world actually occurs. In fact, ϵ-representable worlds, though outnumbered in the set of all worlds, seem to occur all the time in practice—which is why symbolic representations are so often useful. Mixture distributions are ubiquitous in the world, and symbolic representations are pervasive in many accounts of mental representation. This paper argues that these two facts are fundamentally connected. Symbol systems are an effective "compression" of many real phenomena because natural probabilistic systems tend to be modal.

Summarizing, the argument is:

(i) mixtures are ϵ-representable, meaning they can be symbolically represented with small loss of information if M is sufficiently high and K sufficiently low, but
(ii) most worlds (PDFs) are *not* ϵ-representable, meaning that if they are represented by symbols, the approximation is poor.

4. Discrete symbols in multiple dimensions

The previous section established the basic logic of symbolization in 1D, showing how the modality of the environment licenses the conversion of a continuous feature into a discrete symbolic feature with qualitatively distinct values. This section extends this logic into multiple dimensions, meaning multiple interacting symbols. With multiple dimensions, the probability distribution governing the world becomes a multidimensional joint density describing the probabilistic interaction among a number of continuous variables; and the symbolic description of the world becomes a system of logically interacting symbolic features. This section describes how the logical relations among these symbols relate to the probabilistic relations among their continuous counterparts, and in particular (as in the 1D case) on the modality of the joint density.

The leap to multiple dimensions introduces several new issues. Given a PDF $p(x_1, x_2)$ defined over two dimensions x_1 and x_2 having discretizations $\overline{x_1}$ and $\overline{x_2}$ respectively, we must now consider (a) how $\overline{x_1}$ and $\overline{x_2}$ *relate* (compare and contrast) with one another, and (b) how they may be *combined* to create an effective representation of the joint density $p(x_1, x_2)$. As in one dimension, the answers to these questions hinge on the modal structure of the PDF.

In general, we consider a D-dimensional space $X = \{x_1, x_2 \ldots, x_D\}$ over which is defined a PDF $p(X)$. The parameters $x_1, x_2 \ldots, x_D$ are said to be *conjointly modal* if $p(X)$ is a mixture

$$p(X) = \sum_{i=1}^{K} w_i g_i(X), \tag{11}$$

where components g_i have means μ_i, and covariance matrices Σ_i. For example, in two dimensions, a PDF would be conjointly modal if it was produced by a mixture of (say) three Gaussian sources, though as before we will generally make no assumptions about the functional form of the components, except (as above) that they are unimodal and have finite means and covariance matrices. This definition exactly parallels the 1D definition: a PDF is conjointly modal if it is produced by a mixture.

The main new idea in multiple dimensions is that the *logical relations* among the various separate symbols, $\overline{x_1}, \overline{x_2}, \ldots$ depend on the *geometrical relations* among the modes in the mixture. The key issue is how the modes "line up" with respect to the axes. Recall that modes that are sufficiently aligned with respect to one feature will collapse in the marginal PDF, projecting to a single broad peak (Fig. 6). This in turn determines how that feature will be discretized, since the projected mode, a blend of two modes in the other dimension, will correspond to a single level of the resulting discrete variable. Modes that are distinct in multiple dimensions collapse into single symbol values whenever such an alignment occurs. This in turn alters the structure of the individual component symbols and the logical relations among them.

The geometry of modes in multiple dimensions can thus be broken down into qualitative cases depending on how the modes align. Fig. 6 shows the possible cases in the simplest possible multidimensional situation, two modes ($K = 2$) in two dimensions ($D = 2$). The main focus from here on is on the modal structure in the marginal densities, i.e. the projections of $p(X)$ on each of the x_i, relates to the full joint density $p(X)$. Each marginal density is subject to discretization, inducing an alphabet of symbols on that dimension, e.g. dividing x_i into bins $\overline{x_i}(1), \overline{x_i}(2), \ldots$. The Cartesian product of these bins forms a grid $\overline{X}^D$, each cell of which is one possible combination of symbols (Fig. 7). The main issue from here on is how (or whether) the

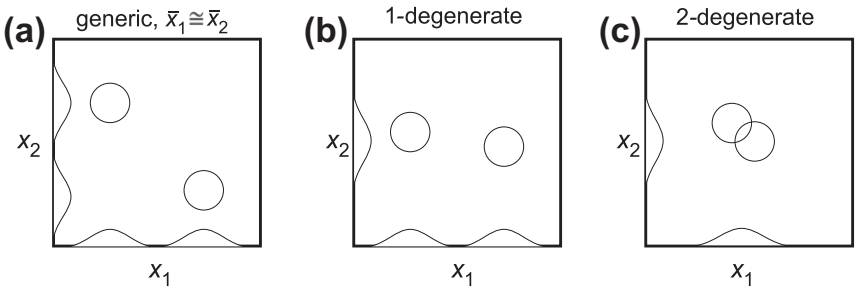

Fig. 6. Qualitative configurations of two modes with two parameters ($K = 2$, $D = 2$), illustrating marginal projections (with implied discretizations, not shown). Modes are illustrated schematically as circles (i.e. contour plots of circular Gaussians), though in general they need not be circular and need not be the same size.

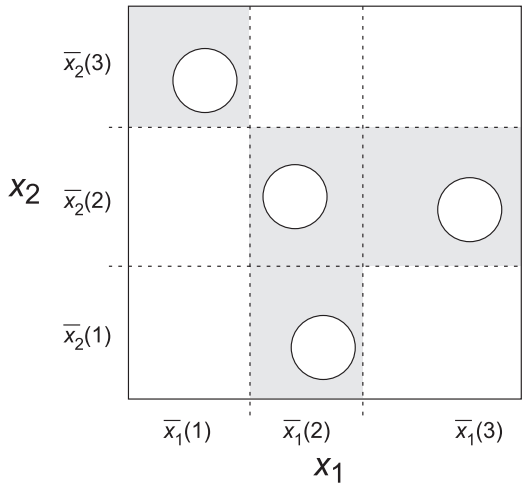

Fig. 7. The grid $\overline{X}^D$ of combinations of the alphabet $\overline{X}$, here with $D = 2$, $K = 4$, and 3 discretized bins on each of the dimensions. The shaded areas are the theory $\phi(\overline{X})$, which in this case ϵ-represents the PDF. Here $\phi = [\overline{x_1}(1) \wedge \overline{x_2}(3)] \vee [\overline{x_1}(2) \wedge (\overline{x_2}(1) \vee \overline{x_2}(2))] \vee [\overline{x_1}(3) \wedge \overline{x_2}(2)]$.

symbols drawn from the various dimensions can be combined to form an effective symbolic representation of the joint density $p(X)$.

Fig. 6a shows the "typical" case, which is said to be *generic* (in general position) in that none of the modes' means align along any single dimension.[7] In the generic case, K modes in two dimensions project to K modes in *each* of the dimensions. Alternatively, two modes may line up in a given dimension, in which case they are said to be *degenerate* in that dimension. (Again, degeneracy does not require perfect alignment, but simply that the two modes project to a marginal density that is so unimodal that they cannot be recovered by the discretization method in use.) In the generic case (Fig. 6a), the two discrete variables $\overline{x_1}$ and $\overline{x_2}$ "agree" with each other: each of them discretizes the world in the same way, in that one level of $\overline{x_1}$ corresponds precisely to one level of $\overline{x_2}$, and to one of the two modes. If p is conjointly modal, then this is the typical way that the symbol $\overline{x_1}$ "unfolds" when $\overline{x_2}$ is considered: that the levels of $\overline{x_1}$ and the levels of $\overline{x_2}$ are isomorphic, although we may not know the correspondence.

A useful way to imagine this, which will be developed below, is to think of the two variables as independent *observers* of the world, e.g. two agents using different measurements or "yardsticks" (as mentioned above in connection with the cloud metaphor; cf. Bennett, Hoffman, & Prakash, 1989). In the generic case, these two observers, after rendering their worlds symbolically, would find that the their representations agree: both carve up the world in the same manner. Like Davidson (1973)'s interlocutors, these two observers' symbols might in principle refer to different phenomena, but if the world is assumed to be a generic conjoint mixture in the above sense, the referred phenomena will be *approximately* the same. In the generic case, the two representations $\overline{x_1}$ and $\overline{x_2}$ are redundant with one another, and each of them fully expresses the modal structure of p.

This is *not* true in degenerate cases, as when the modes align in one dimension (Fig. 6b) or in both dimensions (Fig. 6c). In the 1-degenerate case (Fig. 6b), $\overline{x_1}$ is capable of adequately representing the world, in that its distinct levels are isomorphic to the modes; but $\overline{x_2}$ is not, because it conflates the modes. In the 2-degenerate case (Fig. 6c), the modes are conflated in both dimensions, and neither variable captures the structure. Such a world is not ϵ-representable, because no symbol nor combination of symbols allows a representation that is isomorphic to the multidimensional modes.

With three modes (Fig. 8) the situation becomes slightly more complex. Again there is a single generic case (Fig. 8a), in which the two features are isomorphic to each other and to the modes. Fig. 8b shows the 1-degenerate case, and Fig. 8c shows the 2-degenerate case. Fig. 8d shows a different kind of partly degenerate case that will be taken up in the next section.

Notice that it is possible for two marginal densities, say $p(x_1)$ and $p(x_2)$, to each be individually modal without $p(x,y)$ being conjointly modal; in this case we say x_1 and x_2 are *disjointly modal*. Fig. 9 shows an example. In the figure, x_1 and x_2 can each be seen to be modal by itself, but the joint PDF is *not* a mixture of 2-dimensional sources; instead, each parameter separately is the product of independent modal sources. Parameters that are either conjointly modal or disjointly modal are called *jointly modal*. The main question in the rest of the paper is how jointly modal multidimensional PDFs can be represented by combinations of symbols drawn from their one-dimensional projections—in other words, whether modal PDFs can be represented symbolically. As in the 1D case, the main conclusion is that jointly modal PDFs generally can be effectively represented by symbols, whereas statistically typical multidimensional PDFs generally cannot.

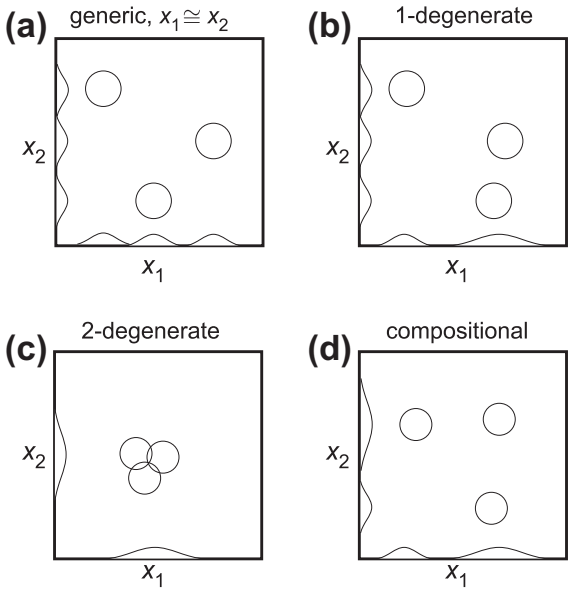

Fig. 8. Qualitative configurations of three modes and two parameters ($K = 3$, $D = 2$).

[7] Later I will use a different notion of general position, requiring that no three modes be collinear, no four coplanar, etc. Here we only require only general position with respect to the axes.

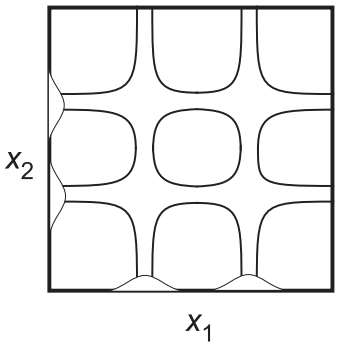

Fig. 9. A disjointly modal density in two dimensions.

4.1. The importance of degeneracy

One trend that becomes increasingly salient as the dimension increases is that (perhaps ironically) generic cases are actually statistically unusual. Recall that alignment between modes does not need to be perfect in order to break genericity, but only sufficiently close to induce two modes to project to a unimodal marginal in any one dimension. By McLachlan and Basford (1988)'s rule of thumb, this will happen whenever any two modes venture within approximately 2σ of each other (see Ray & Lindsay, 2005 for more detailed analysis of the conditions supporting multimodality). "Merging" of modes has increasingly high probability as M decreases or K increases (see Appendix A.5). The result is that fully generic cases are very unusual. For example with $M = D = K = 3$ the probability of complete genericity is less than 1%. Statistically speaking, degenerate cases are the norm, not the exception. For this reason the next section considers varieties of degeneracy in more detail.

4.2. Compositionality: mixtures and logical forms

Fig. 8d presents a new wrinkle. There, both $\overline{x_1}$ and $\overline{x_2}$ are degenerate (they are both bimodal while the joint PDF is trimodal), but they are not isomorphic to one another (they conflate different modes). This means that neither $\overline{x_1}$ nor $\overline{x_2}$ effectively represents p; nor does the Cartesian product $\overline{x_1} \times \overline{x_2}$, because it has four cells while p has only three modes. What is needed is a *selection function* that identifies which combinations of $\overline{x_1}$ and $\overline{x_2}$ correspond to the modes g_i, and thus effectively capture p. Such a function would map the grid $\overline{X}^D$ to (0, 1), with 1 indicating a legal combination and 0 an illegal one. Such a function is conveniently viewed as a *Boolean function*, easily represented by a propositional form $\phi(\overline{X})$ defined over the alphabet $\overline{x_1}, \overline{x_2}, \ldots$ (Fig. 7). Following logical terminology, I will henceforth refer to ϕ as a *theory* and the legal combinations of $\overline{X}^D$ that satisfy it (i.e for which $\phi(\overline{X}) = 1$) as its *models* $\mathcal{M} = m_1, m_2 \ldots$.

This suggests an appropriate multidimensional generalization of definition of ϵ-representation. By analogy to the 1-D case, the symbol uncertainty of $\phi(\overline{X})$ is the expected symbol uncertainty

$$H[X|\phi(\overline{X})] = E_{\phi(\overline{X})=1}[H(X|\overline{X})], \tag{12}$$

where the expectation is now taken over legal cells (i.e. such that $\phi(\overline{X}) = 1$). As in the 1D case, this is the uncertainty that remains about $p(X)$ once we know its symbolic representation $\phi(\overline{X})$, including both its multivariate discretization $\overline{X}$ and the theory ϕ defined over it. Note that this definition encompasses the 1D definition if we tacitly regard the theory there as a trivial one in which $\phi(\bar{x}) = 1$ for all values of $\bar{x}$.

With this definition, the representation $\phi(\overline{X})$ (defined over the discretized alphabet $\overline{X}$) ϵ-represents the PDF $p(X)$ (defined over the continuous space X) if

$$H[X|\phi(\overline{X})] < \epsilon. \tag{13}$$

Conceptually, this definition is just like the 1D version. A world p is ϵ-representable if the uncertainty that remains once you know its symbolic representation is small (less than ϵ). Moreover, all the main properties of ϵ-representation carry over: the magnitude of the symbol uncertainty increases linearly with K and decreases linearly with $\log M$ (see Appendix A.6).

More specifically, $\phi(\overline{X})$ will ϵ-represent $p(X)$ if each component g_i of p falls in a distinct legal cell of ϕ, and all legal cells contain exactly one mode. If so, there will be an isomorphism between the modes g_i and the legal cells of ϕ. (It also follows that any two theories that both ϵ-represent p must themselves be isomorphic, a point that will be developed below.) In this situation, the theory ϕ exactly expresses the structure of the PDF, and all the properties of ϵ-representation follow. As M increases, each of the modes will increasingly predominate within its cell, contributing an increasingly large proportion of the probability mass, with a decreasing proportion coming from other modes. As M increases each legal cell of ϕ becomes an increasingly "pure" product of a single generating source.

It is clear that a necessary condition for such a theory ϕ to exist is that every *pair* of modes in p be resolved by at least one feature in $\overline{X}$, i.e. that any distinct modes project to distinct values of some symbol. If so, then the grid $\overline{X}^D$ will contain a distinct cell for each mode g_i. Some subset of these cells actually contain a mode, and this subset defines a ϕ that ϵ-represents p. Conversely, if two modes are fully degenerate—conflated in all dimensions—then this condition is not met, and p is not ϵ-representable. Note though that any two modes can be resolved if M is large enough (unless they coincide exactly); any mixture of K distinct modes can be effectively represented if the modes are sufficiently narrow.

As in the 1D case, an observer who represents a modal world $p(X)$ via a symbolic representation $\phi(\overline{X})$ that ϵ-represents it is approximately "right". Note that ϵ-representation does *not* mean that the world obeys the propositional description perfectly. There may be, and generally is, some non-zero probability of exceptional cases, meaning values of $\overline{X}$ that do not obey ϕ. But the conditions of ϵ-representation mean that the probability of such cases is small, or more specifically, that an observer whose representation discounts such cases will not be surprised too often, with expected total surprise (i.e. uncertainty) bound by ϵ.

The remainder of this section shows that representations of modal worlds obey all the familiar properties of

logical forms: they correspond to propositional formulae, they compose, and they support logical relations. That ϵ-representation composes means that combinations of ϵ-representable worlds are themselves ϵ-representable; and moreover the theory representing the joint world is a composition (specifically, a conjunction) of the component theories. What is more, we can generate *all and only* ϵ-representable worlds by combinations of simpler ϵ-representable worlds. This means that any symbolically representable world can be thought of as systematic combinations of simpler worlds. This means that ϵ-representable world can generally be reasoned about via logic—while, again, arbitrary worlds generally cannot.

4.3. Modal worlds correspond to propositional formulae

The intimate relationship between mixtures (modal PDFs) and propositional formulae can be more fully appreciated by observing that *all* mixtures (other than fully degenerate ones) are ϵ-represented by some formula on a suitable alphabet; and that *all* propositional formulae ϵ-represent some mixture. We establish this correspondence separately in each direction.

The first direction, that any non-fully-degenerate mixture is ϵ-represented by some ϕ, was already established above. To see the other direction, that any formula represents some world, simply observe that any propositional formula ϕ is equivalent to a disjunction of conjunctions each of which includes a positive or negative mention of each variable (called a complete disjunctive normal form formula or complete DNF). This complete DNF defines the legal cells of ϕ over $\overline{X}$. We can place one component (with sufficiently small σ) in the center of each bin corresponding to one term of the DNF, and their mixture will be a PDF that is ϵ-represented by ϕ.

Clearly an infinite number of other mixtures (e.g. slight perturbations of this one) will also be ϵ-represented by ϕ. Similarly, an infinite number of distinct formulae can ϵ-represent the same mixture. This relation between formulae, which will be important below, is referred to as *metagruence*. Two formulae ϕ_1 and ϕ_2 are metagruent, denoted $\phi_1 \cong \phi_2$, if they both ϵ-represent the same mixture p.

4.4. Modal worlds compose

It follows immediately from the above that if two PDFs are ϵ-representable, then their combination (joint density) is also ϵ-symbolically representable, specifically by a theory that is a conjunction of the two component theories. Specifically, if X_1 and X_2 are jointly modal, and $p(X_1)$ is ϵ_1-represented by $\phi_1(\overline{X_1})$, and $p(X_2)$ is ϵ_2-represented by $\phi_2(\overline{X_2})$, then the joint density $p(X_1,X_2)$ will be ϵ-represented by $\phi_1 \wedge \phi_2$ (defined over the alphabet $\overline{X_1} \cup \overline{X_2}$), with $\epsilon = \max(\epsilon_1, \epsilon_2)$. That is, modal worlds compose to form modal worlds. Note however that the fidelity of the representation, the magnitude of ϵ, only gets worse, never better, as worlds combine (and the dimension increases). As the world gets more complex, the effectiveness of any symbolic representations of it tend to degrade.

In the specific case of conjointly modal worlds, we can be more specific about the nature of the conjoined representation. If $\phi_1(\overline{X_1})\epsilon$-represents p, then its models m_{1-1}, $m_{1-2}\ldots$ each correspond to one of the modes g_i, meaning that each legal cell contains one and only one μ_i. If $\phi_2(\overline{X_2})$ also ϵ-represents p, then its models m_{2-1}, $m_{2-2}\ldots$ must also correspond to the same modes. From this it follows that the two representations ϕ_1 and ϕ_2 must be isomorphic to each other; they pick out exactly the same K modes. This means that the conjoined theory has form

$$\{m_{1-i}\} \leftrightarrow \{m_{2-i}\}, \tag{14}$$

after suitable renumbering of the models. This is a *biconditional* relation between ϕ_1 and ϕ_2. For example, the simple 2-mode configuration in Fig. 6a has two features, $\overline{x_1}$ and $\overline{x_2}$, each of which ϵ-represent p. Each of them has two models, corresponding to the two modes, and the conjoined theory ϕ is just

$$\phi = [\overline{x_1}(1) \wedge \overline{x_2}(2)] \vee [\overline{x_1}(2) \wedge \overline{x_2}(1)], \tag{15}$$

which is equivalent to the biconditional

$$\phi = \{\overline{x_1}(i)\} \leftrightarrow \{\overline{x_2}(i)\} \tag{16}$$

(after renumbering of the values). More complex examples with more dimensions and more modes would work similarly. If both theories ϵ-represent the same world, then when one representation takes a particular set of values, it implies that the other takes a corresponding set of values, and vice versa. In this sense the two theories are mutually redundant; they contain the same information and represent the same world in perfectly isomorphic ways.

This isomorphism is an example of metagruence as defined above. Putting this in the language of observers, this means that if two observers independently observe a common world p through distinct measurement languages X_1 and X_2 respectively, if they assume only that X_1 and X_2 are conjointly modal, they can reasonably infer that their observations are essentially equivalent—that phenomena in p referred to by symbols in the X_1 language are the *same* phenomena referred to in the X_2 language. Again, such an isomorphism does not hold in principle, and is not valid in non-modal worlds (indeed it is not even true in worlds that are disjointly modal but not conjointly modal). This suggests that conjoint modality is a key assumption underlying the mutual intercomprehensibility of distinct representational systems—solving, at least for the case of PDFs, the problems of radical translation or radical interpretation posed respectively by Quine (1960) and Davidson (1973). Martians and Earthlings, observing the same universe via completely incommensurate measurements and conceptual structures, can nevertheless assume common referents—*if* they assume conjoint modality, but generally not otherwise.

If we assume that every set of mixture components has a uniform dimension[8] (is generated across a fixed number of features), then it follows that every jointly modal PDF p can be divided into conjointly modal "bubbles", within which all dimensions are conjointly modal, but between which all dimensions are only disjointly modal. *Within* each bubble, features are conjointly modal, so non-trivial logical

[8] The situation with mixed-dimension mixture components is more complicated, and will be deferred to a future paper.

relations between features may exist, and distinct ϵ-representations are metagruent, meaning that (like the Martians and Earthlings in the example above) they share common referents (the modes within the common bubble). But *between* bubbles, since there is no conjoint modality, there are no common referents, no logical connections, and representations that are fundamentally incommensurate because there are no common structures to refer to. This creates a "Rashomon"-like situation in which distinct observers viewing what is nominally the same world will nevertheless draw totally unrelated conclusions about it (cf. Breiman, 2001). For such observers, mutual translation—even while talking about the same world—is indeed impossible (again see Davidson, 1973; Quine, 1960).

4.5. Modal worlds support logical relations

All the above arguments taken together suggest that modal mixtures support logic, in the sense that they are capable of *approximately* satisfying logical inferences (see Ali, Chater, & Oaksford, 2011). More specifically, ϵ-representation respects logical implication. If a formula $\phi\epsilon$-represents a world p, and ϕ logically entails another formula ϕ',

$$\phi \Rightarrow \phi', \tag{17}$$

then ϕ' ϵ-represents p as well. Specifically, such an implication will hold whenever the models of ϕ' are a subset of the models of ϕ, and thus correspond to a subset of the mixture components of p. Logical statements of the form *if A then B* can be regarded as approximately true when referring to modal worlds—or, putting this more strictly, an observer who believes them to be true will rarely be surprised (expected surprise, i.e. uncertainty, less than ϵ). Modal worlds are capable of being the approximate extensions (models) of logical implications, but statistically typical worlds generally are not.

A narrow but important example of this is an implication entailed by a single discrete symbol, say $\bar{x}(v)$ (see Feldman, 2006). If the world under observation is not modal, then such an observation cannot in general be assumed to imply *anything* about any other variable, and in this sense is quite literally meaningless. If you know that the object you are holding is a blicket (not a dax), but the blicket/dax distinction is not conjointly modal with *any* other variable, then this knowledge has literally no value. But if x is conjointly modal with some other set of variables X, then the world p will generally be ϵ-representable by some formula over the alphabet $x \cup X$. It is still possible that $\bar{x}$ may happen to be logically independent of the other variables $\overline{X}$. But more generally it may not be, in which case there will be some formula(e) ψ such that

$$\bar{x}(v) \rightarrow \psi, \tag{18}$$

"if $\bar{x}(v)$ then ψ is true", e.g. "if blicket then edible". In this case observing $\bar{x}(v)$ means *something* potentially important.

4.6. Multidimensional modality: summary

In one dimension, mixtures can be effectively represented by discrete features, though most PDFs cannot. In multiple dimensions, mixtures can generally be effectively represented by potentially complex *combinations* of features drawn from their 1D projections. This composition generally corresponds to a propositional formula, the exact form of which depends on the pattern of degeneracy (conflationary alignments) along individual dimensions in the mixture.

Syllogisms and other law-like logical relations lie at the heart of symbolic reasoning, but in a complex stochastic context, one might imagine they would rarely hold perfectly. Indeed this doubt is central to skepticism about the cognitive validity of symbolic representations. But a logical law need not be perfectly valid to be useful; it only needs to be accurate enough so that adopting it rarely leads one astray. This is the criterion captured by ϵ-representation. Modal worlds are potential extensions for logical laws, albeit imperfect ones; while arbitrary worlds generally do not satisfy logical relations to any reliable degree.

5. The choice of features

Above we have regarded the choice of dimensions as given, i.e. we have assumed a fixed set of subspaces x_1, $x_2 \ldots$ of X through which p is observed. But more generally, if X can rotate freely (in psychological terminology, if its dimensions are *integral*), we might imagine other choices of dimension through the space, e.g. linear combinations of the x_i which correspond to diagonal slices through X. Indeed, assuming the x_i to be given begs the question of exactly why these dimensions make more sense than others in the first place. So relaxing this assumption allows us to ask more basic questions about *feature selection*: which dimensions of X are most helpful in contributing to an effective representation of p?

Geometrically, what this means is that we will now spin the space X freely, sampling *arbitrary* one-dimensional features and combinations thereof, instead of being limited to the arbitrary coordinate frame we began with (Fig. 10). This way the question of how p can be represented by combinations of features can be expanded beyond the original symbol vocabulary. As before, the main focus is on the marginal densities projected onto these subspaces, and how discretizations of them combine to form effective symbolic representations of p.

A useful way of thinking about this situation is to imagine each measurement as a different *observer* of the same world. Each observer measures the same world from a distinct, unique vantage point, assessing a distinct characteristic—technically, sampling along a distinct subspace. The set of possible observers corresponds to the set of distinct subspaces, which form a hypersphere, which I will refer to as the *observer hypersphere* (Fig. 11). (Technically it is a hemihypersphere, because each viewpoint v is interchangeable with the inverse viewpoint $-v$). The main question now becomes: how can a set of observers combine their measurements to adequately represent the structure of the world "inside the cloud" (Fig. 4)? More specifically, how can the discrete symbols drawn from their measurements be combined to form an effective

symbolic representation of the structure within the cloud? A broader view of this problem is introduced when one considers arbitrary (non-orthogonal) viewpoints.

5.1. Representational equivalence

Looking at Fig. 10, one can see that there will be many dimensions that provide essentially the same information about the modes. For example, any sufficiently slight perturbation (rotation) of one feature yields another that resolves the same set of modes (Fig. 12), and thus provides approximately the same representational benefit. This sort of equivalence is not limited to slight perturbations, but is shared by any alternate variable whose discretization plays the some role in compositional formulae that represent p, which can include broad swaths of the observer hypersphere. We adopt the following definition. Assume p is ϵ-represented by some formula ϕ over an alphabet $\overline{X}$ which includes a symbol $\bar{x}$. Construct ϕ' by replacing $\bar{x}$ with another symbol $\bar{x}'$ wherever it appears in ϕ. Then if ϕ' also ϵ-represents p, then x and x' are *representationally equivalent*, denoted $x \sim x'$. Loosely speaking, $x \sim x'$ means that features drawn from x and features drawn from x' are interchangeable in descriptions of the world; they serve the same role in symbolic descriptions because they pick out the same modes. Representationally equivalent symbols in modal worlds, while *not* precisely equivalent, are mutually interpretable in the sense of Davidson (1973).

Representational equivalence is obviously an equivalence relation (it is reflective, symmetric, and transitive), so it divides the observer hypersphere into equivalence classes of symbols, referred collectively to as the *observer chart* (Fig. 13). The observer chart is an exhaustive map of the possible qualitatively alternative symbols for representing p. Intuitively, each symbol class in the chart contains an infinity of symbols which, while not exactly the same, resolve exactly the same modes in p, and are thus interchangeable with respect to symbolic descriptions of it. The boundaries between the classes represent those points in observer space where the projection shifts from resolving one set of modes to resolving a different set of modes (larger, smaller, or simply different). That is, within each class the pattern of degeneracy is uniform, but between classes it changes. Exactly where the boundaries lie depends on the discretization method in use. But the qualitative structure of the chart of classes depends only on the geometry of the modes.

The observer chart summarizes what meaningfully distinct symbols are available to any observer of a given world. Symbols within each class are qualitatively equivalent, in the sense the choice makes no difference in how the world is symbolically represented. Symbols in different classes are qualitatively distinct.

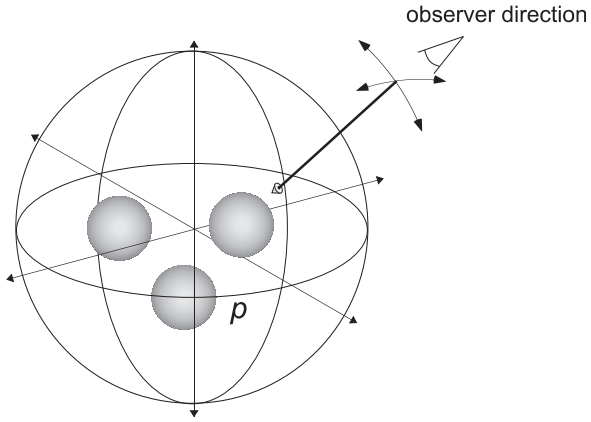

Fig. 11. The observer hypersphere in three dimensions.

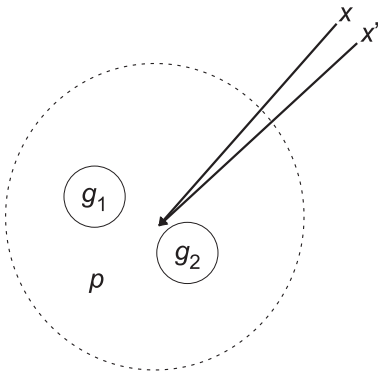

Fig. 12. Representational equivalence ($x \sim x'$). The two features x and x', while distinct, provide qualitatively the same contribution to a representation of p because they resolve the same two modes. The dotted circle is the observer hypersphere in two dimensions.

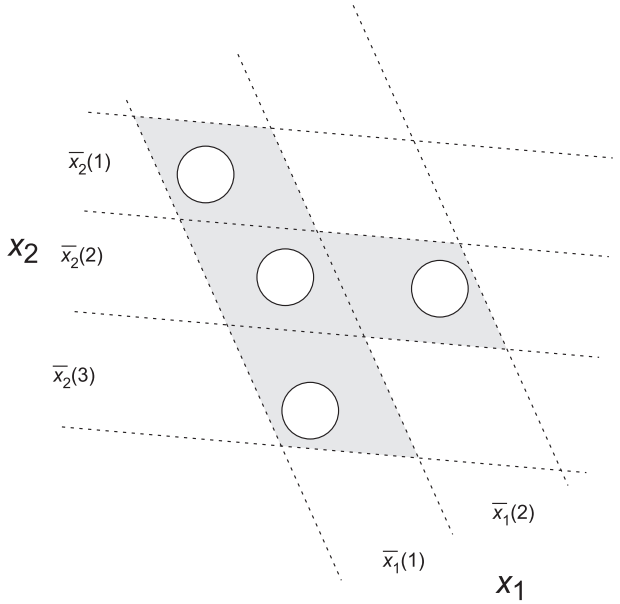

Fig. 10. Non-orthogonal dimensions of observation induce a non-perpendicular grid $\overline{X}^D$, upon which may be built an ϵ-representation ϕ (shaded) just as in the orthogonal case.

5.2. Bases

This naturally raises the question of which *combinations* of symbols (or really symbol classes, since symbols within a class are interchangeable) are sufficient to represent p.

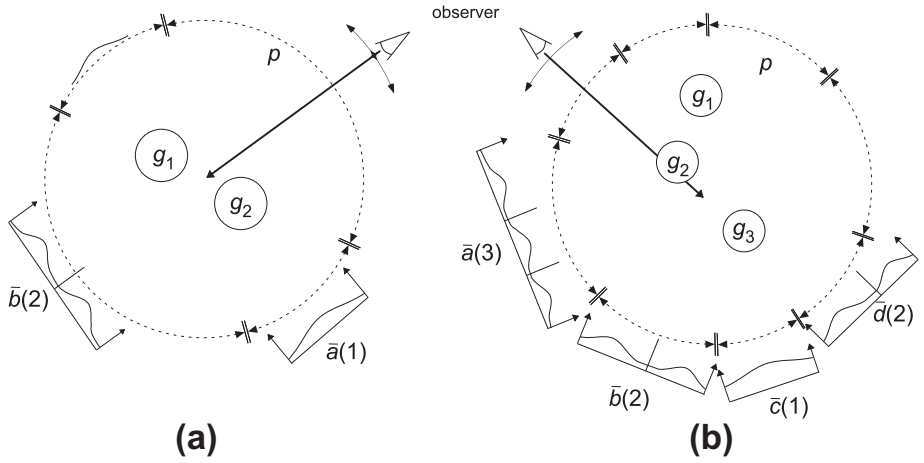

Fig. 13. Observer chart with (a) $K = 2$ and (b) $K = 3$. In each case, the space of observing directions (the observer hypersphere, dotted circle) is divided into a set of regions (equivalence classes). In (a), there two classes, $\bar{a}$ (within which all measurements have one mode) and $\bar{b}$ (within which all have two). Here $\bar{b}$ ϵ-represents p, but $\bar{a}$ does not. In (b), there are four classes, $\bar{a}$, $\bar{b}$, $\bar{c}$ and $\bar{d}$, with respectively 3, 2, 1, and 2 modes. Some combinations jointly ϵ-represent the joint distribution, but others do not.

For example, looking at the observer chart in Fig. 13a, symbol $\bar{b}$ is sufficient by itself to represent p ($\bar{b}$ ϵ-represents p), while a is unhelpful because it conflates the two modes. (Generally, any symbol that conflates all K modes contributes nothing to representation.) With three modes (Fig. 13b), the situation is more complicated. One can see that $\bar{b}$ and $\bar{d}$ are jointly sufficient to represent p, because they define a 2×2 grid in which each mode has its own cell. $\bar{a}$ is sufficient all by itself, because it resolves the three modes. Symbol $\bar{c}$ (like $\bar{a}$ in Fig. 13a) is unhelpful because it conflates all three modes. With more complex worlds with more modes, one can easily imagine that there would be a variety of combinations that would represent p in distinct ways, some with more symbols, others with fewer, but always jointly sufficient to resolve all K modes. As explained above, any combination that induces a grid fine enough for each mixture component to inhabit a distinct cell is sufficient to represent p.

A set of symbol classes that are jointly sufficient to ϵ-represent a PDF p will be called a *basis* for p. A basis can be thought of as an alphabet sufficiently expressive to serve as a representation.[9] Obviously, every ϵ-representable PDF has a basis, while other PDFs—i.e. most PDFs—do not. In fact, most modal worlds will have many distinct bases. At one extreme, every mixture of K modes in general position (no three modes collinear, no four modes coplanar, etc.), and with sufficiently high M, will have one (unique) basis consisting of a single variable with K levels (referred to as a 1-basis). This is the direction of maximum "projection index" defined by Friedman and Tukey (1974), i.e. the direction which maximally reveals clusters in the data. In addition, there will be various 2-bases, 3-bases, etc. In general, for every composite number $K' > K$, there will be at least one basis for every factorization of K', up to and including the next power of two above K. For example, five sufficiently separated modes in general position will have one 1-basis, one 2-basis (a subset of the 2×3 grid), and one 3-basis (a subset of a $2 \times 2 \times 2$ grid). The largest basis is the one in which all the variables are Boolean. If the modes are not sufficiently separated (M too low), not all of these bases will necessarily exist, though at least one of the must since we have assumed that p is ϵ-representable.

Representations of p over distinct bases are metagruent (see definition above). Again this is a very abstract notion of equivalence, more so than logical equivalence (which requires provably equivalent formulae), or congruence, meaning equivalence after permutation of variables (see Feldman, 2003). Metagruent representations must have the same number of models, as they are isomorphic to each other and to the components of the mixture they represent. In the example given above, a mixture of $K = 5$ modes has several bases including one 2-basis (2×3) and one 3 basis ($2 \times 2 \times 2$). Representations expressed over these bases will be metagruent because they ϵ-represent the same world, even though they are obviously neither equivalent nor congruent. Indeed, unlike equivalence or congruence, metagruence is not a syntactic concept at all, but a semantic one. It cannot be defined solely in terms of the relations among the symbols in ϕ_1 and ϕ_2, but rather depends on the relationship of ϕ_1 and ϕ_2 to the world p to which they both refer.

It should be added that from this perspective there is no meaningful distinction between *atomic* features and others, nor between simple and complex concepts. The same set of modes can be represented by a single feature (a 1-basis) or by a (metagruent) composition of multiple features (a B-basis for $B > 1$). No feature is intrinsically primitive, though some features play the role of primitive features in particular representations. No one basis is intrinsically superior to all others regardless of modal structure, though some features may resolve more modes in the environment of a particular organism, perhaps making them desirable choices as basic perceptual features.

[9] The use of the term *basis* here should not be confused with its use elsewhere in Boolean algebra, where it refers to a set of operators jointly sufficient to represent all propositional forms. Here it is a set of symbols, not a set of operators, and more similar in meaning to the basis of a vector space.

5.3. What makes a feature meaningful?

Jepson and Richards (1992) have argued that a perceptual feature is "meaningful"—conveys functional significance above and beyond an arbitrary measurement—when its taps the natural modes of the environment. This paper expands on this point, with the notion of modes explicitly identified with components in a mixture. Above, features were described as meaning "essentially the same thing" (being representationally equivalent) when they were mutually interchangeable in representations of the world. Expanding on this idea, we can postulate that the meaning of a feature lies in its ability to aid in representations of the world, and thus to participate in reliable inference about the world—which as developed above means its ability to resolve modes. This suggests that any feature's "true" meaning consists in its ability to carve up the world in a way that relates to the carving due to *other* features. Some have argued that features take on meaning in virtue of their role within a larger system of other features (*conceptual role semantics*; see Harman, 1987). Here we have grounded this idea by understanding these interacting roles in terms of formal relations among features in a multidimensional PDF.

To be concrete, imagine we observe an apple to be *red*, not *green*. Why is this meaningful to us? If the color itself is a random variable, generated as a random deviate from a single "apple" color mode, then learning its value conveys nothing of value; the only stable property of the object is that it is an apple, while its exact color value is inconsequential. But if the color feature is conjointly modal with another feature—say, species (*Macintosh* or *Granny Smith*), or ripeness (*ripe* or *unripe*)—then its color thereby takes on meaning. As in conceptual role semantics, color is not *inherently* meaningful in the absence of assumptions about what color pertains to; it acquires meaning in virtue of its relationship (specifically, conjoint modality) with other features. More abstractly, consider an observer viewing Fig. 6a (a generic configuration of two modes in two dimensions) viewed through feature $\overline{x_1}$. The symbol $\overline{x_1}$ has two values, corresponding to the two modes. Why do we *care* which value it takes? Without an assumption of conjoint modality, there is actually no reason to. But if we assume that x_1 is conjointly modal with other, unknown dimensions (such as it in fact is, with x_2), then knowing the value of $\overline{x_1}$ conveys something potentially useful about their values. In other words, if having observed x_1 we assume that the world p unfolds into higher dimensions as a multidimensional conjointly modal mixture (as it actually does, from the "God's eye" vantage point in Fig. 6a), x_1 thereby takes on tangible meaning.

The vast majority of cognitive science presumes simple physical features (size, shape, luminance) as the naive basis for psychological representations of the physical world. Nevertheless, more ontologically skeptical authors (e.g. Hoffman, 2009) have argued forcefully that the physical features subjectively available to cognition constitute a infinitesimal, and in some ways arbitrary, fraction of the physically possible features. Indeed, contemporary physics assumes only a handful of truly atomic features (length, duration, charge, spin, and a few others; see Richards, 1988)—which correspond poorly to human perception. Some authors (Feldman & Richards, 1998; Goldstone & Steyvers, 2001; Koenderink, 1993; Richards & Koenderink, 1995; Schyns, Goldstone, & Thibaut, 1998) have attempted to articulate criteria for the creation of perceptual features, but the principles separating "meaningful" features from others are still foggy.

The current argument suggests that features' meaning derives, in effect, from the modal structure of the environment being represented (Richards & Bobick, 1988). *Red* vs. *green* is meaningful if it is biconditional with some other feature (e.g. *ripe* vs. *unripe*, *stop* vs. *go*)—or participates in some other non-trivial representation of the world (again cf. conceptual role semantics). But if it does not, it is literally meaningless.

6. Summary and extensions

The main argument of this paper is simply that environments constructed from mixtures can be effectively represented by discrete symbols, while others generally cannot. Modal worlds contain statistically stable structures to which symbols may refer, allowing compact, logic-like representations to be reasonably faithful. In contrast, the vast majority of statistically possible worlds contain no probabilistically stable structures, and thus nothing for symbols to refer to. The degree to which symbolic representation is effective is modulated by its degree of modality, quantified by M (the separation among the modes) and K (the number of modes). As M falls and K rises, the modes becomes broader and more numerous, the mixture increasingly resembles a uniform density, and the effectiveness of symbolic representation progressively degrades. This quantitative relationship is in some ways more revealing than broad philosophical arguments about symbols. Many familiar aspects of symbolic representation, such as compositionality, logical implication, and propositional forms, can be seen as arising from the rich pattern of geometrical relationships among modes that arise in higher dimensions.

As connectionists have argued, the world may indeed be too complex and stochastic for clean symbolic representations to work perfectly—indeed *most* possible worlds are like that, though perhaps not *our* world. Symbolists have argued that symbolic description allow us to comprehend the world at a desirably broad level of abstraction—but obviously this is possible only if only the descriptions are reasonably faithful. The contribution of this paper is to establish conditions under which such symbolic descriptions apply enough to be useful.

The well-known "No free lunch" theorem of Wolpert (1996) (which shows that no inferential procedure is uniformly superior to all others across all possible environments), and the "Ugly duckling theorem" of Watanabe (1969) (which shows that one cannot infer meaningful similarity relations without making some non-trivial assumptions about the environment) suggest, broadly speaking, that one cannot make effective inferences about the environment without first making some assumptions about its form. Broadly speaking, abstract arguments about the validity of symbolic representations—without any stipulations about the nature of the environment—are hopeless. Like

all other methods of inference and representation, symbolic representations work well or poorly depending on the environment in which they are applied. Thus the assumption that the environment is a mixture—i.e. the principle of natural modes—is a critical assumption licensing the use of symbolic representations. Since modal worlds can be effectively represented by symbols, but arbitrary worlds generally cannot, the the use of symbols entails a tacit assumption of modality in the world.

Several extensions and generalizations should be briefly mentioned. Above I have assumed that every mixture component was centered at a point $\mu_i \in X$. But many conceptions of natural regularities assume them to take the form of *correlations* among parameters (Murphy & Wisniewski, 1989; Richards & Bobick, 1988). Correlations, meaning covariance among multiple continuous parameters, can be thought of as components (modes) with dimension higher than 0—space curves, planes, and more generally manifolds with multiple intrinsic dimensions. The mathematics becomes much more complicated in this case, though the basic questions of symbolic representation, discretization, and observability are all fundamentally similar. Another aspect not yet developed involves components of mixed rank, such as when one component is conjointly modal among several dimensions, but another only among a subset. Finally, another important extension is to establish representational conditions for 1st-order logic (rather than simple propositional logic as developed above), a more complex and challenging setting for logical representations. These and other technical extensions await future work.

7. Conclusions

This paper began by asking: what is the difference between a continuous variable and a discrete one? This question is the leading edge of a far broader issue: the relationship between continuously-parameterized stochastic worlds and the discrete compositional symbol systems often used to represent them (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Smolensky & Legendre, 2006). The long tradition of symbolic models in cognitive science, emphasizing composition and crisp logical relations, is often set in opposition to the world of statistical models, including the newly resurgent class of probabilistic Bayesian models (Chater & Oaksford, 2008) which emphasize statistical regularities and messy data. But the argument in this paper is that these worlds are connected by the concept of modality, as formally captured in mixture models. In modal probabilistic worlds, faithful, meaningful symbolic representation is demonstrably possible, though only approximately so; they may well have residual aspects not accurately reflected in their symbol representations, but only up to an identifiable bound (ϵ-representation). But arbitrary (non-modal) worlds generally do *not* support symbolic representation, in the sense that their structure is generally too random to be effectively summarized by symbols. In this regard the conception has much in common with that in various other probabilistic renderings of logic, including default logics (Reiter, 1980), probabilistic non-monotonic reasoning (Pearl, 1988), and probabilistic models of deduction (Oaksford & Chater, 2009), and moreover establishes conditions on the world for these approximations to obtain. The degree to which various logical forms constitute reasonably faithful glosses of the world can be quantified in terms of the degree of modality, in the sense defined above. The contribution of this paper is to quantify this approximation, thus bridging what is often viewed as a chasm between probabilistic and symbolic models.

Because symbolic representation is mathematically justifiable for modal worlds, but generally not otherwise, the assumption of modality must be viewed as a key element of the justification for the use of symbolic representations in cognitive models. That is, symbolic representations cannot be used to represent stable and recurring phenomena in statistically typical (non-modal) worlds, because such worlds generally *do not contain* such phenomena. But symbols generally do work in modal environments, in that the symbol uncertainty (and thus total uncertainty) in such worlds is demonstrably low. So cognitive systems that employ symbolic representations rest on a tacit assumption that the world being represented actually has a modal structure. As mentioned above, this conclusion can be seen as a special case of a standard result from the theory of complexity, namely that most structures (here, PDFs) are incompressible, while a small fraction are compressible in the sense that they can be concisely summarized (here, represented by symbols).

That is, most probabilistically-defined worlds are not, in principle, compressible to the extent that they can be represented by a mental symbol system with a reasonable degree of fidelity. Conversely, we have identified one well-defined class of environments, modal ones, that *are* symbolically compressible in this sense. Our cognitive apparatus is capable of representing the environment because the environment is modal, or, more specifically, *to the extent that* it is modal.

As Marr (1982) famously observed, models of mental mechanisms cannot get far without a substantive consideration of the conditions under which they will be effective. Modality can be regarded as a (very broad) Marrian constraint: an environmental condition that helps makes effective symbolic representation and logical inference possible. Statistical structure abounds in the natural environment, and a complete understanding of cognitive mechanisms is impossible until we fully understand how this fact informs the mind. To paraphase Warren McCulloch (1960), the key question is not *What is a symbol, that it may represent the world?* but rather *What is the world, that a symbol may represent it?*

Acknowledgments

I am grateful to Whitman Richards for the many discussions that inspired this paper, as well as for stimulating insights from Cordelia Aitkin, Nick Chater, David Fass, Noah Goodman, Aaron Kheifets, Michael Kubovy, Mike Oaksford, Manish Singh, and an anonymous reviewer. I am grateful for support from NSF SBR-9875175, EITM-0339062, DGE 0549115, and NIH EY021494.

Appendix A. Derivations and details

A.1. The modality parameter M

As explained in the text, the modality parameter M is defined as twice the ratio of the standard deviation of the component means to the maximum component standard deviation,

$$M = \frac{2S}{\sigma_{\max}}. \tag{19}$$

This makes it a close relative of the traditional F ratio: specifically $M^2 \propto F$ after normalization of degrees of freedom. However I avoid this notation because in the context of this paper there is generally no reason to expect M^2 to be distributed as Fisher's F distribution. If we draw K samples from a common Gaussian distribution (not a mixture), and separately estimate their within-sample standard variances σ_i^2 and between-sample variance S^2, then we would indeed expect the ratio of these variance estimates to be distributed as F. But the premise of the current paper is specifically that the environment is *not* unimodal in this sense, but rather that the distinct sources g_i have independent means μ_i, in which case the measured M^2 would be expected to distributed very high relative to F. In this regard the "Hypothesis of Natural Modes" (that distinct modes have distinct causal sources) can be regarded as the negation of the classical Fisherian null hypothesis in the context of analysis of variance (namely, that apparently separate samples are all drawn from a common distribution.)

A.2. ϵ-representation, one-dimensional case

We seek to establish a bound on the expectation of the uncertainty $H[p(x|\bar{x})]$ after the value of value of the discretized variable $\bar{x}$ is known, expressing this bound as a function of the parameters M and K of the mixture that generated $p(x)$. As in the text, we assume a mixture $p(x) = \sum_i w_i g_i(x)$ having K components $g_i(x)$ with respective means μ_i and standard deviations σ_i, of which the largest is $\sigma_{\max}$. The ensemble of μ_i has standard deviation S, and as in the text we define the modality $M = 2S/\sigma_{\max}$. Fig. 14 illustrates the situation with $K = 3$.

We establish an approximate upper bound on the uncertainty $H[p(x|\bar{x})]$ by observing that "most" of the probability mass within the ith interval is due to the ith mode, while some portion λ of it is due to other sources $j \neq i$. The probability mass within the interval is a mixture of these sources, so we establish a bound by (i) finding a bound on the uncertainty of the ith component, (ii) finding a bound on the uncertainty when a PDF p is mixed with λ of another *PDF*, and (iii) establishing a bound on the magnitude of λ. Combining (i)–(iii) establishes a bound on the uncertainty of the mixed PDF within the ith interval, i.e. on the expectation of $H(p(x|\bar{x}))$.

(**i**) Each component $g_i(x)$ has been assumed to be unimodal, but we have made no other assumptions about its functional form (although note that Diaconis & Freedman (1984) have shown that 1-D projections of higher-dimensional densities tend to be Gaussian under broad assumptions). But even without making any such assumptions, we can still place a bound on its uncertainty by observing that among all densities with a given mean and standard deviation σ, the Gaussian has maximum uncertainty $\log(\sigma\sqrt{2\pi e})$. Because $\sigma \leqslant \sigma_{\max}$, and $M = 2S/\sigma_{\max}$ it follows that $\sigma \leqslant 2S/M$. Hence the expectation of the uncertainty of g_i is bounded by

$$E[H(g_i)] \leqslant \log\left(\frac{2S\sqrt{2\pi e}}{M}\right), \tag{20}$$

This quantity decreases with $\log M$, which means that as modality increases and the mixture gets more spiky, the (bound on the expectation of the) uncertainty due to each mode decreases.

(**ii**) Inside the ith interval the density p is a mixture of g_i with some small quantity of additional probability mass due to other sources. Uncertainty is concave, meaning that when an g_i is mixed with other PDFs the uncertainty generally rises. But it is a continuous function, suggesting that if the total mass of the additional source is small, the rise in uncertainty should also be small.

In general, consider the change in uncertainty of a PDF q when it is mixed with a total quantity λ drawn from another PDF. We establish an upper bound on the uncertainty of the mixture q' using the following

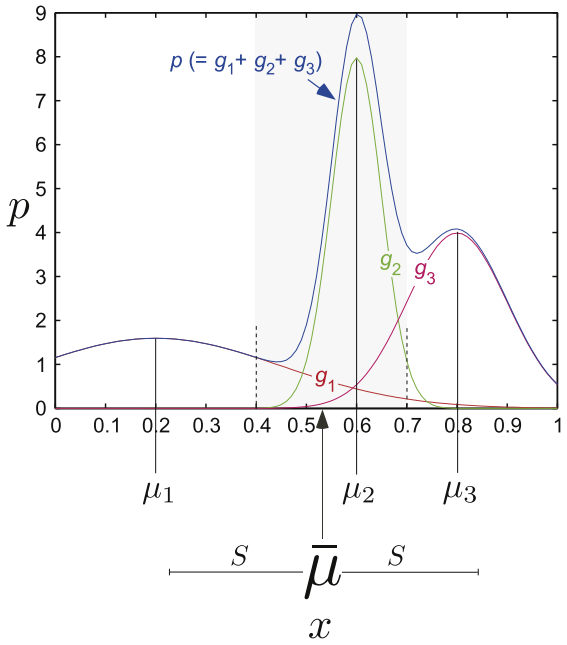

Fig. 14. A mixture of $K = 3$ Gaussian modes, illustrating the computation of uncertainty within each bin of the corresponding discretized variable. The three components have respective means μ_1, μ_2, and μ_3, which have overall mean $\bar{\mu}$ and standard deviation S. Within a given bin (here the middle one, $\bar{x}(2)$, shaded), the PDF is a mixture of the ith component (here g_2) with smaller contributions from the other components. The bound given in the text shows how the expected uncertainty within this bin is bound by the is a sum of the uncertainty of a Gaussian g_2 plus a small contribution (bounded by λ) due to the addition of probability mass from the other $K - 1$ components.

inequality, which guarantees that for arbitrary densities r and s, if

$$\sum_{i=1}^{N} |r_i - s_i| \leqslant \lambda \leqslant \frac{1}{2}, \tag{21}$$

then

$$|H(r) - H(s)| \leqslant -\lambda \log \frac{\lambda}{N}. \tag{22}$$

(See Cover & Thomas, 1991, Thm. 16.3.2.) In our situation the roles of r and s are played by q and q', whose total differential $\sum_i^N |q - q'|$ is bound by λ. This means that

$$H(q') \leqslant H(q) + \lambda \log N\lambda. \tag{23}$$

(**iii**) To place a bound on λ, consider how much of the mass generated by g_i is likely to fall outside the ith interval i. Somewhere between each pair of adjacent means there exists an optimal classification boundary θ that minimizes the expected uncertainty within the bins. We instead choose to divide bins at the midpoints between them (e.g. $\frac{|\mu_i - \mu_{i+1}|}{2}$), knowing that this necessarily yields a higher expected uncertainty. The expectation of the distance between any two modes is $2S$ (because by assumption μ_i have been drawn from a distribution with standard deviation S). The expected radius from the center of each bin to the boundary of another bin is therefore S/σ, which is just $M/2$. Chebyshev's inequality establishes that given any $p(x)$ with standard deviation σ, the total probability mass falling further that $z\sigma$ outside the mean must be less than

$$p(|x - \mu| > z\sigma) \leqslant \frac{1}{z^2}. \tag{24}$$

In the current situation this means that the total mass due to g_i falling more than M away from μ_i (i.e. outside the interval $\bar{x}(i)$) is less than

$$\sum_{x \notin \bar{x}(i)} p(x|g_i) \leqslant \frac{4}{M^2}. \tag{25}$$

This bound holds regardless of the form of the distributions g_i.

Conversely, within the ith interval the contribution from each *other* source $j \neq i$ obeys the same bound. There are $K - 1$ other sources, so the total probability mass within $\bar{x}(i)$ due to sources other than i is less than or equal to

$$\lambda \leqslant \frac{4(K-1)}{M^2}. \tag{26}$$

Finally, combining (i)–(iii) allows us to conclude that the expected symbol uncertainty is bound by

$$E[H(x|\bar{x})] \leqslant \log\left(\frac{2S\sqrt{2\pi e}}{M}\right) - \lambda \log \lambda/N, \tag{27}$$

with $\lambda = 4(K-1)/M^2$. This quantity bounds the expected residual uncertainty about the parameter x after the value of the discrete value $\bar{x}$ is known. By definition this establishes that $p(x)$ is ϵ-representable with ϵ equal to the above expression (Eq. (27)). This quantity is of order $O(K)$ and $O(-\log M)$, justifying the summary given in the text (Eq. (9)).

A.3. Uncertainty of a mixture

We can extend the above reasoning to establish a bound on the uncertainty H of a mixture $p(x)$, by noting as in the text that $p(x) = p(x|g_i)p(g_i)$, which means that the total uncertainty of $p(x)$ is simply the sum of the expected uncertainty in each bin (which is bounded by Eq. (26)) plus the uncertainty in the discrete variable $-\sum_{i=1}^{K} w_i \log w_i$, which has expectation $\log K$. Hence the uncertainty in a mixture is bounded by

$$E[H(p)] \leqslant H(x|\bar{x}) + \log K, \tag{28}$$

with the first term bounded as in Eq. (26). As M decreases or K increases, the mixture generally gets "flatter" and uncertainty rises, and symbolic representation becomes progressively less useful; but note that it can never exceed the theoretical bound of $\log N$. See Fig. 5 for numerical corroboration of these analytical results. Again, the derived bound is very loose because of the loose Chebyshev bound, which makes no distributional assumptions, but the simulation confirms the functional dependence on M and K.

A.4. Uncertainty of an arbitrary distribution

We aim to show that "most" distributions have high uncertainty. A probability distribution is simply an assignment of probability mass $p(x)$ to the N bins along x such that $\sum_i p = 1$. We aim to show that "most" distributions of this form have nearly maximal entropy, about $\log N$.

This is actually just a special case of the Wallis formulation of the maximum-entropy principle discussed in Jaynes (2003), which shows that among all distributions satisfying some fixed set of constraints, almost all have entropy near the maximum. Here we simply set the constraints to be empty, allowing all distributions to qualify. In the Wallis formulation, we imagine creating a quantized probability density function by assigning some large number T of quanta of probability mass to bins along x, in a way that satisfies some set of constraints (here, none). Among all ways of assigning the quanta, almost all have uncertainty near maximal, where the nearness to maximality depends on how small we choose to make the probability quanta. In an arbitrary distribution, no region of the underlying parameter x would have expected density higher than any other.[10] (If one did, it would be a "constraint" of the type we are assuming do not apply.) To be more specific, assume that the T probability quanta are each place in one of the N slots in $[0,1]$, with each slot having equal probability $1/N$ of being chosen. Thus the number actually falling in the ith slot will have a multinomial density, with the expected number in the ith bin having expected mean T/N and variance $T(N-1)/N^2$. We divide by T to normalize and

[10] To be clear, *individual* distributions may certainly have higher density in one region than another. It is the *expected density*, i.e. the distribution from which the probability quanta are drawn, that is uniform. Thus though individual distributions may be non-uniform, the entire ensemble of distributions does not systematically favor one region of x over another.

obtain expected mean $p_i = 1/N$, and variance $T(N-1)/N^2$. Thus the expected absolute deviation (positive square root of the variance, normalized by T) of the p_i's from $1/N$ will be

$$\left[\sqrt{T(N-1)}/N\right]/T \tag{29}$$

or about $1/\sqrt{TN}$. The uncertainty is the expectation of $-p\log p$, and is thus about

$$H(p) = -N\left[\frac{1}{N}+\frac{1}{\sqrt{TN}}\right]\log\left[\frac{1}{N}+\frac{1}{\sqrt{TN}}\right] \tag{30}$$

As $T \to \infty$ the $1/\sqrt{TN}$ terms vanish and the uncertainty approaches that of a uniform density,

$$H \to -N\frac{1}{N}\log\frac{1}{N} = \log N. \tag{31}$$

The total number of quanta T modulates the precision with which we are approximating the distribution. As T grows, a larger and larger portion of the resulting densities are nearly uniform. This can also be seen as a variant of the asymptotic equipartition property (Cover & Thomas, 1991), which guarantees that as a sequence of Bernoulli trials grows in length, a larger and larger fraction of outcomes has nearly the same, typical, uncertainty.

In summary, "most" distributions are approximately uniform and have uncertainty about log$N$. As noted in the text, this is a counting argument, not a probabilistic one. The argument shows simply that the total set of probability distributions is made up primarily of ones with high uncertainty; it says nothing about how often we will encounter any subset of them. In fact, the main argument in the text is that while modal worlds are but a small minority of all possible worlds, they are "over-represented" in reality (Richards & Bobick, 1988), which is why symbolic representations are so often useful.

When the value of $\bar{x}$ is known, the uncertainty about x is reduced by the ratio of total measure (1) to the measure of the ith bin, which has expectation $1/K$. (Again we are taking the expectation over the underlying quantum-generating distribution, not over any one particular distribution. Any one distribution might yield uneven weights w_i larger or smaller than $1/K$; but over the entire ensemble of distributions no one bin will generally be larger than any other.) So for an arbitrary density the expected uncertainty conditioned on knowing the value of the discretized variable (the symbol uncertainty) is about

$$E[H(x|\bar{x})] \approx \frac{\log N}{K}. \tag{32}$$

For large M and small K, this quantity is much larger than Eq. (27), the comparable value for mixtures. Arbitrary PDFs have high uncertainty, regardless of discretization, while modal densities have low uncertainty, because they can be effectively discretized.

A.5. Probability of degeneracy

We approximate the probability of degeneracy by quantifying the probability of degeneracy for any two modes along any single dimension, and then assuming all such degeneracies are independent. Exactly how distant two modes must be before they become separable depends on the discretization method, but as a simple approximation we adopt McLachlan and Basford (1988)'s rule of thumb for bimodality, whereby the mixture of two modes becomes unimodal when their means μ_1 and μ_2 fall within σ of each other. As means themselves have standard deviation S, so does the separation $|\mu_1 - \mu_2|$ between them (as can be seen if one imagines the location of one fixed at 0, in which case the location of the other has standard deviation S.) The modes will be conflated if this distance is less than 2σ, or, putting this the other way, the configuration will be generic if one mode falls in the tail of the other mode out beyond this criterion. This criterion has z-score $2\sigma/S$. As $M = 2S/\sigma$ (Eq. (4)), the critical z is just $4/M$. By Chebyshev's inequality (cf. the very similar argument in Appendix A.2), the probability of a non-degenerate configuration along this dimension is less than

$$p(\text{generic}) < (M/4)^2. \tag{33}$$

Assuming K independently positioned modes in D dimensions, such conflations will all be independent. There are $\binom{K}{2}$ pairs of modes projected to each dimension (for example 6 pairs of 4 modes), so there are $D\binom{K}{2}$ pairs total in all dimensions. So the overall probability of degeneracy will be bound by

$$p(\text{generic}) \leqslant \left(\frac{M^2}{16}\right)^{D\binom{k}{2}}. \tag{34}$$

For $M > 4$, this bound is greater than unity, because the Chebyshev bound is very weak. But for $M < 4$, meaning broad overlapping modes, the bound decreases rapidly with D and K, as a completely generic case becomes increasingly "special" and unlikely to occur by accident. Degenerate cases then become the norm, and representations of them correspondingly essential, justifying the special attention given to them in the text.

A.6. ϵ-representation, multidimensional case

The multidimensional definition of ϵ-representation is a simple generalization of the one-dimensional case. The maximum standard deviation $\sigma_{\max}$ is now over dimensions as well as over modes (i.e. it is now the largest univariate σ in any single dimension in any single mode), though a richer approach might be developed by taking covariances into account. The other main difference is the presence of the theory ϕ which picks out legal cells of $\bar{X}^D$. The expectation (e.g. Eq. (20)) is now taken over these cells only. Inside these cells, uncertainty is low, because they primarily contain probability mass due to the component centered there (and only λ due to other modes). As in the 1D case, uncertainty within each cell shrinks as M increases and the PDF becomes progressively spikier, with one spike in each legal cell. Though there is less probability mass in other cells, the uncertainty conditioned on them can be high, as what mass there is may be uniformly spread throughout the (modeless) cell.

Distinct theories that each ϵ-represent the same PDF will generally define different grids, meaning that expectations will be taken over slightly different regions of X leading to different values of symbol uncertainty and ϵ. But these distinct representations will pick out the same K modes, so the dependence on K and M will be qualitatively similar. Different representations will represent p with different degrees of absolute fidelity, but any one that resolves the K modes will do so "effectively" in the sense defined in the text, with qualitatively similar variation in effectiveness as M and K vary.

References

Aitkin, D., & Feldman, J. (submitted for publication). Discretization of continuous features by human learners.

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition, 119*, 403–418.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.) (pp. 217–234). Cambridge: M.I.T. Press.

Barlow, H. B. (1974). Inductive inference, coding, perception, and language. *Perception, 3*, 123–134.

Barlow, H. B. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research, 30*(11), 1561–1571.

Barlow, H. B. (1994). What is the computational goal of the neocortex? In C. Koch & J. L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 1–22). Cambridge: M.I.T. Press.

Bay, S. D. (2000). Multivariate discretization of continuous variables for set mining. In *Kdd '00: Proceedings of the sixth acm sigkdd international conference on knowledge discovery and data mining* (pp. 315–319). New York, NY, USA: ACM Press.

Bennett, B. M., Hoffman, D. D., & Prakash, C. (1989). *Observer mechanics: A formal theory of perception*. London: Academic Press.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–215.

Chaitin, G. J. (1966). On the length of programs for computing finite binary sequences. *Journal of the Association for Computing Machinery, 13*, 547–569.

Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford: Oxford University Press.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science. *Trends in Cognitive Sciences, 7*(1), 19–22.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (second ed.). Hillsdale, New Jersey: Lawrence Erlbaum.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley.

Davidson, D. (1973). Radical interpretation. *Dialectica, 27*, 313–328.

Diaconis, P., & Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics, 12*(3), 793–815.

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *International conference on machine learning* (pp. 194–202). Los Altos, CA: Morgan Kaufman.

Dretske, F. I. (1981). *Knowledge and the flow of information*. Cambridge: M.I.T. Press.

Dy, J. G. (2008). Unsupervised feature selection. In H. Liu & H. Motoda (Eds.), *Computational methods of feature selection*. Boca Raton, FL: Chapman & Hall.

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th international conference on artificial intelligence* (pp. 1022–1027). Choambery, France: Morgan Kauffman.

Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology, 47*(1), 98–112.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology, 50*, 339–368.

Feldman, J., & Richards, W. A. (1998). Mapping the mental space of rectangles. *Perception, 27*, 1191–1202.

Feldman, J., & Singh, M. (2005). Information along contours and object boundaries. *Psychological Review, 112*(1), 243–252.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers, 9*, 881–890.

Goldstone, R., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology, 130*(1), 116–139.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*(1), 108–154.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Harman, G. (1987). (Non-solipsistic) conceptual role semantics. In E. Lepore (Ed.), *New directions in semantics*. London: Academic Press.

Harnad, S. (1993). Grounding symbols in the analog world with neural nets. *Think, 2*, 57–62.

Hoffman, D. D. (2009). The user-interface theory of perception. In S. Dickinson, M. Tarr, A. Leonardis, & B. Schiele (Eds.), *Object categorization: Computer and human vision perspectives*. Cambridge: Cambridge University Press.

Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Deitrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines*. Cambridge: MIT Press.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Jepson, A., & Richards, W. A. (1992). What makes a good feature? In L. Harris & M. Jenkin (Eds.), *Spatial vision in humans and robots* (pp. 89–125). Cambridge: Cambridge University Press.

Jepson, A., Richards, W. A., & Knill, D. C. (1996). Modal structure and reliable inference. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 63–92). Cambridge: Cambridge University Press.

Koenderink, J. J. (1993). What is a "feature"? *Journal of Intelligent Systems, 3*(1), 49–82.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission, 1*(1), 1–7.

Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. New York: Springer.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

McCulloch, W. S. (1960). What is a number, that a man may know it, and a man, that he may know a number? *General Semantics Bulletin (26/27)* (Reprinted in McCulloch, W. S. (1965). *Embodiments of mind*, Cambridge, MA: M.I.T. Press).

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley.

Murphy, G., & Wisniewski, E. (1989). Feature correlations in conceptual representations. In *Advances in cognitive science*. In G. Tiberghien (Ed.). *Theory and applications* (Vol. 2, pp. 23–45). Chichester: Ellis Horwood.

Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences, 32*, 69–84.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kauffman.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A, 185*, 71–110.

Putnam, H. (1988). *Representation and reality*. Cambridge, MA: M.I.T. Press.

Quine, W. V. O. (1960). *Word and object*. New York: Springer.

Ray, S., & Lindsay, B. G. (2005). The topography of multivariate normal mixtures. *The Annals of Statistics, 33*(5), 2042–2065.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence, 12*, 81–132.

Resnikoff, H. L. (1985). *The illusion of reality: Topics in information science*. New York: Springer-Verlag.

Richards, W. A. (1988). The approach. In W. A. Richards (Ed.), *Natural computation*. Cambridge, MA: M.I.T. Press.

Richards, W. A., & Bobick, A. (1988). Playing twenty questions with nature. In Z. Pylyshyn (Ed.), *Computational processes in human vision: An interdisciplinary perspective* (pp. 3–26). Norwood, NJ: Ablex Publishing Corporation.

Richards, W. A., & Koenderink, J. J. (1995). Trajectory mapping: A new non-metric scaling technique. *Perception, 24*(11), 1315–1331.

Rumelhart, D. E., McClelland, J. L., & Hinton, G. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, Massachusetts: MIT Press.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21*, 1–54.

Shapiro, S. (2006). *Vagueness in context*. Oxford: Oxford University Press.
Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review, 1*(1), 2–28.
Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar* (Vols. 1–2). Cambridge, MA: M.I.T. Press.
Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analyses of finite mixture distributions*. Chichester: Wiley & Sons.
Usher, M. (2001). A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind & Language, 16*(3), 311–334.
Watanabe, S. (1969). *Knowing and guessing: A quantitative study of inference and information*. New York: John Wiley.
Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8*(7), 1341–1390.