To appear in: The Oxford Handbook of Computational Perceptual Organization
S. Gepshtein, L. Maloney, and M. Singh (Eds.), Oxford University Press

Perceptual grouping as Bayesian mixture estimation

Jacob Feldman, Manish Singh, and Vicky Froyen
Dept. of Psychology, Center for Cognitive Science
Rutgers University - New Brunswick

Abstract

Perceptual grouping is the process by which the visual system organizes the image into distinct objects or clusters. Here we briefly describe a Bayesian approach to grouping, formulating it as an inverse probability problem in which the goal is to estimate the organization that best explains the observed set of visual elements. We pose the problem as an instance of *mixture modeling*, in which the image configuration is assumed to have been generated by a set of distinct data-generating components or sources ("objects"), whose locations and structure we seek to estimate. We illustrate the approach with three classes of source models: dot clusters, contours, and axial shapes. We show how this approach to the problem unifies and gives natural accounts of a number of perceptual grouping problems, including contour integration, shape representation, and figure/ground estimation.

Highlight: A novel framework for perceptual grouping uses Bayesian mixture estimation to provide a unifying account of grouping problems, including contour integration, shape representation, and figure-ground estimation.

Keywords: perceptual organization; Bayesian inference; mixture estimation; dot clusters; contours; shape skeleton

A Bayesian approach to perceptual grouping

Perceptual grouping is the process by which the visual system organizes the image into distinct clusters or units. The grouping problem is inherently ambiguous, in that the system must select among an enormous number of potential grouping interpretations (e.g., the number of partitions of N items is exponential in N). The situation is ripe for Bayesian

We are grateful to Seha Kim, Sung-Ho Kim, and John Wilder for stimulating discussions. Please direct correspondence to Jacob Feldman, Department of Psychology, Center for Cognitive Science, Rutgers University - New Brunswick, 152 Frelinghuysen Rd., Piscataway, NJ 08854, or by e-mail at jacob@ruccs.rutgers.edu.

inference, which is a uniquely rational method for interpreting data under conditions of uncertainty.[1] In a Bayesian framework, one assumes that the data (here, the image) are consistent with a variety of hypothetical causes (here, scene models). The fit of each model to the data is expressed by the likelihood, the probability of the image conditioned on each model, and the inherent plausibility of each hypothesis (scene model) is expressed by its prior probability.[2] In the Bayesian framework, a rational observer should believe each model (or, in the continuous case, parameter value) in proportion to its probability conditioned on the data (its posterior probability). Bayes' rule says that the posterior of a particular interpretation is proportional to the product of its prior and likelihood, a simple but profound observation that powers the Bayesian machine.

In the last two decades, Bayesian inference has been applied to a wide array of problems in visual perception (see Knill & Richards, 1996; Kersten, Mamassian, & Yuille, 2004 for overviews). Most applications have involved estimating an "objective" (independently measurable) physical characteristic of the scene, such as depth (Jacobs, 1999; Knill, 2003), color (Brainard et al., 2006), or motion (Weiss, Simoncelli, & Adelson, 2002). Perceptual grouping differs from these problems in that the scene characteristic that we wish to estimate, the assignment of visual elements to distinct groups or units, is not an independently measurable property of the world per se, but instead is an organizational framework imposed upon the elements in the image. That is, while perceptual grouping sometimes involves estimating literal physical bonds, like rigid attachment, it can and often does involves more abstract notions such as "perceptual unit," which entail more abstract types of commonality (such as common origins). Such inferences in turn reflect mental models of how objects tend to be created, and need not be verifiable by any directly measurable physical property. As an example, if a large number of pebbles are laid out in a rectangular

[1] More precisely, Bayesian inference is the provably unique way to assign beliefs in an internally consistent manner (Cox, 1961; Jaynes, 2003), and when coupled with a loss function (forming Bayesian decision theory) is the provably unique rational way to select actions (Savage, 1954; Maloney, 2002).

[2] Note that this includes the case where the set of models forms a continuously parameterized family, in which case each value of the parameters constitutes a distinct model, with its own prior and likelihood. In this case the distribution of posterior probability over models (parameter values) forms is called the posterior distribution or posterior density function.

array (say with an aspect ratio of 2:1), there is no "fact of the matter" regarding whether they "really" lie in rows or columns in the physical world. Rather, the organization into rows or columns is a perceptual construct based on the assumptions that the visual system makes about its environment.[3] Nonetheless, the basic logic of Bayesian inference applies: given generative models of groups, which allow likelihood functions to be specified, and suitable priors to accompany them, we can apply Bayes' rule to attach a posterior belief to each way that the visual elements can be grouped.

This brief paper gives a synopsis of a principled, conceptually simple, and internally coherent Bayesian approach to the perceptual grouping problem. We sketch the mathematics very briefly, aiming primarily to show how the approach unifies the treatment of a number of standard problems in perceptual organization, including dot grouping, contour integration, figure/ground assignment, and shape representation. We explicitly distinguish competence (theory of the computation) from performance (algorithmic) aspects of the problem, briefly sketching an approach to the latter that allows the computation to be embedded in a network formalism.

Mixtures

We begin by making an analogy between the grouping problem and the problem of estimating a mixture. A *mixture density* (or simply *mixture*) is a probability density function that is composed of a weighted sum of K component distributions or sources,

$$p(x) = \sum_{i}^{K} p_i g_i(x), \tag{1}$$

where each g_i is a distinct generative source, and p_i denotes the probability with which the i-th source is chosen. The components g_i generally each have distinct parameters θ_i, such

[3] Examples such as these highlight the fact that the correspondence between the objective world and perceptual representations can be arbitrarily complex—far from the homomorphism that is often assumed. Perceptual representations are tuned by evolution to optimize not veridicality or similarity to the objective world per se, but rather the fitness consequences of the actions that these representations support. These two constraints are different and can lead to very different predictions (see Mark, Marion, & Hoffman, 2010; Koenderink, 2011; Hoffman & Singh, 2012; Feldman, 2013; Singh & Hoffman, 2013).

as (in the case of Gaussian components) means μ_i and standard deviations σ_i. Individual mixture components are often assumed to have a simple unimodal form (e.g. Gaussian), but because the various components have distinct parameters (including locations), the resulting mixture can be highly multimodal and irregular in structure. The problem of mixtures is how to interpret a complex and hetereogeneous dataset as the result of a combination of simple, internally homogeneous components (see McLachlan & Basford, 1988). More specifically, the problem faced by the observer is to estimate the mixture, meaning to estimate the parameters of the component sources based on a sample of data drawn from the mixture. Estimating a mixture is a difficult problem in part because the observer generally does not know which datum came from which source, but instead must guess, while simultaneously estimating the parameters of the sources. Naturally these two problems interact, as the assignment of data to sources influences the estimate of the parameters of the sources, and vice versa. Mixtures are a natural way of modeling situations that involve a set of distinct data-generating processes that have been intermingled—such as perceptual grouping, where the observed configuration might comprise data drawn from a variety of distinct sources, such as contours, surfaces, and objects. Solving a mixture estimation problem entails partitioning the data into distinct sources—just like perceptual grouping.

We assume that the data consist of a set $X = \{x_1 \dots x_N\}$ of points drawn from a dataspace $\mathcal{X}$. It is convenient to think of each datum x_i as having a missing or hidden label z_i representing its source. A solution to the problem consists of an estimate $\hat{Y} = \{(p_1, \theta_1) \dots (p_{\hat{K}}, \theta_{\hat{K}})\}$ of the weights and parameters of the mixture components (generally with $K \ll N$) along with an estimate $\hat{Z} = \{\hat{z}_1 \dots \hat{z}_N\}$ of the latent source labels for each datum.

In a Bayesian formulation (Stephens, 2000), each of the parameters to be estimated would have a prior density, which taken together determine the prior $p(Y)$ on a given mixture model Y. For example, one might have a prior that favors few mixture components,

or favors narrow standard deviations on mixture components, and so forth.[4] The fit of a particular mixture model Y to the data X is expressed by the likelihood $p(X|Y)$, which quantifies how well the observed data can be explained by the given model. By Bayes' rule, the posterior probability $p(Y|X)$ of a particular mixture model is then proportional to the product of the prior and likelihood for that model $p(Y)p(X|Y)$. A substantial body of theory is devoted to the problem of determining or approximating this posterior (see Stephens, 2000), which often cannot be determined analytically even if the distributional form of the sources is known.

Grouping as a mixture estimation problem

Perceptual grouping can be thought of as an unusually elaborate and geometrically complex mixture estimation problem. The data consists of the ensemble of visual elements present in the image, e.g. points or edges. The mixture components are probabilistic processes that generate visual elements in the image—that is, "objects," or projections of objects—with parameters that govern the spatial patterns by which those elements tend to be distributed. More precisely, we assume that the data have been generated by a mixture of K data sources $g_1 \dots g_K$ (objects), each of which generates visual elements with some probability distribution over 2D space.[5] The goal of the observer is then to estimate the parameters of the mixture components, while (as in any mixture estimation problem) simultaneously estimating which data (visual elements) were generated by which component. (In what follows, we assume for simplicity that each element has exactly one source, but the mixture framework can be extended to encompass "mixed ownership" with probabilistic weighting.) The assignment of visual elements to sources—inferring which elements were generated by which source—is, by this definition, perceptual grouping.

[4] Adopting a particular prior means making an assumption about probable structure of the environment. However, exactly what is entailed by the adoption of a particular prior depends on exactly what is meant by "probability," which is notoriously controversial. In a nutshell, to some theorists (frequentists), priors correspond to factual assumptions about the environment, while to others (subjectivists, including most Bayesians), a prior merely characterizes the observer's state of knowledge, and thus does not amount to an affirmative claim about the actual properties of the environment. See Feldman (2013) for discussion.

[5] More comprehensively, we might assume the g_i generate visual elements in 3D space, which is then projected down to 2D, but we defer the subtleties appropriate for this generalization to a future paper.

Mixture components in the grouping problem

The first step in realizing this program is to define classes of data-generating processes. Many such classes may be imagined, given the infinite variety of visual patterns that may exist in nature. Here we briefly discuss three simple and well-studied types: clusters, contours, and axial shapes. While certainly not exhaustive, these three types reflect a very fundamental classification into respectively location-based, orientation-based, and hierarchically organized generative processes.

Dot clusters

A very basic type of visual pattern, occasionally encountered in nature (e.g. flocks of birds) but especially important as an object of study (e.g. Cohen, Singh, & Maloney, 2008; Juni, Singh, & Maloney, 2010), are simple clusters of isotropic elements (e.g. points or dots) (Fig. 1a). For simplicity, we assume that each cluster is generated by a circular Gaussian process with a spatial mean $\mu_{x,y}$ and standard deviation σ. If the K means are all chosen independently, and the separation between the sources (inter-mean distance divided by component standard deviation) is sufficiently large, the resulting mixture looks like a set of distinct clouds of visual elements. Such a process is a simple model of any situation in which data is generated in spatial proximity to a localized source, so that more closely spaced data are more likely to have a common source.

The probability that a point x should be classified as having been generated by g_1 rather than by g_2 is given by the posterior ratio $p(g_1|x)/p(g_2|x)$. Assuming Gaussian sources g_1 and g_2 with equal priors and equal variances σ^2,

$$p(x|g_i) \sim \mathcal{N}(\mu_i, \sigma^2), \tag{2}$$

the posterior ratio can easily be seen to decay exponentially as x ranges from near g_1 to near g_2,

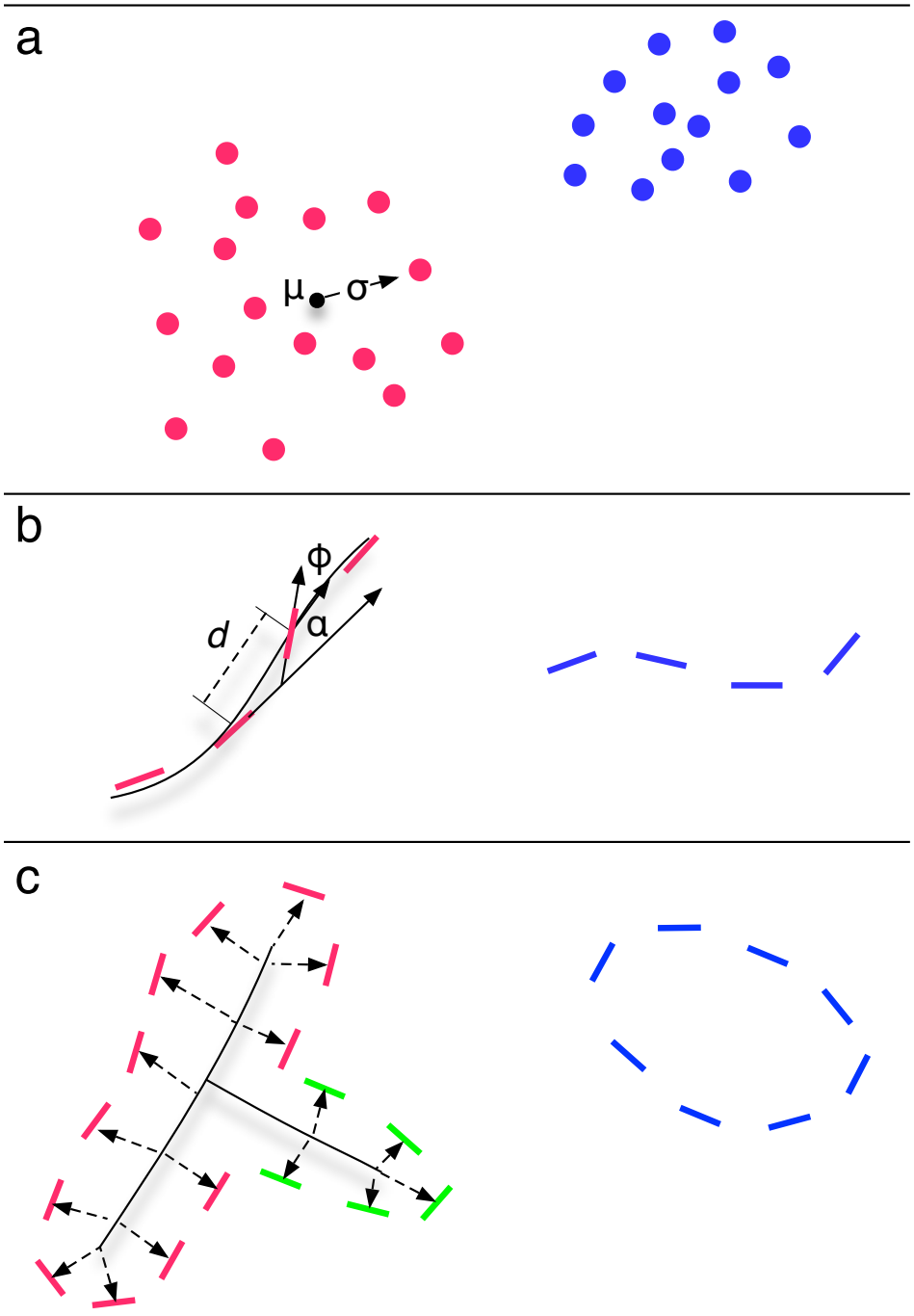

Figure 1. Schematic illustrations of the three generating source models discussed in the text: (a) dot clusters (b) contours and (c) axial shapes. Distinct colors indicate distinct mixture sources.

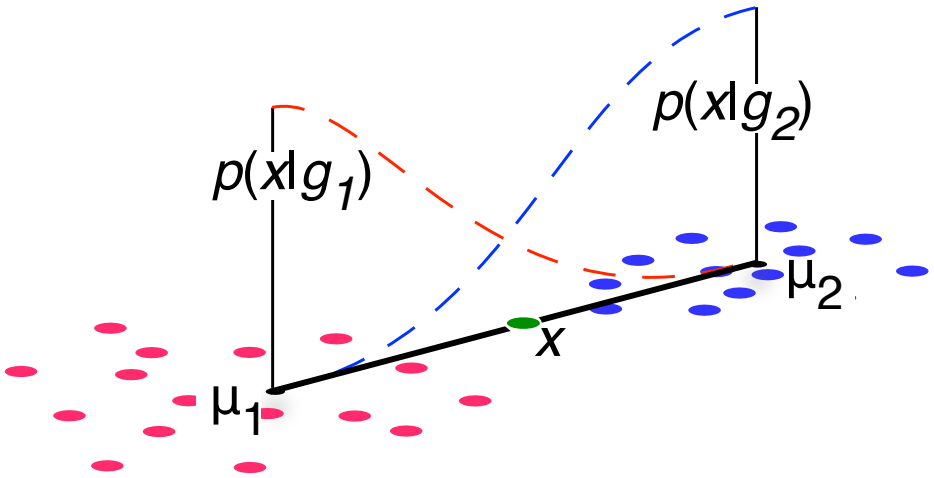

Figure 2. The Bayesian rationale for the "pure distance law" of Kubovy & Wagemans (1995) For any dot x between the two clusters g_1 and g_2 (with means respectively μ_1 and μ_2), the evidence that x belongs to g_1 rather than g_2 depends on the likelihood ratio $p(x|g_1)/p(x|g_2)$ (the ratio of the red curve to the blue curve). If $p(x|g_1)$ and $p(x|g_2)$ are both Gaussian (normal), then the ratio decays exponentially as x moves from μ_1 to μ_2.

(Eq. 3).

$$\frac{p(x|g_1)}{p(x|g_2)} = \exp\Big[-\frac{1}{\sigma^2}d + \frac{1}{2\sigma^2}\Big], \tag{3}$$

where $d = \|x - \mu_1\|/\|\mu_1 - \mu_2\|$ is the distance between x and the first component relative to the distance between the two components (Fig. 2). Indeed, a simple exponential decay in grouping strength was discovered empirically by Kubovy and Wagemans (1995) and Kubovy, Holcombe, and Wagemans (1998), who called it the "pure distance law." Eq. 3 gives a Bayesian rationale for this law, demonstrating its optimality under the assumed conditions. Thus the Bayes optimal strategy for estimating Gaussian mixtures actually entails—and, we would argue, *explains*—the Gestalt principle of proximity. The broader connection between Gestalt laws and Bayesian inference will be discussed below.

Contours

A contour is a set of oriented visual elements that form a "smooth" trajectory through visual space. The Gestaltists referred to the tendency towards collinearity as "good contin-

uation," a vague term that has since been formalized in a variety of ways (e.g. Uttal, 1973; Zucker, 1985; Smits, Vos, & van Oeffelen, 1985; Field, Hayes, & Hess, 1993; Pizlo, Salach-Goyska, & Rosenfeld, 1997). To formalize contours as a probabilistic generating process, we imagine a smooth contour that stochastically generates discrete visual elements at random intervals along its arclength. We assume that inter-sample distances d are drawn from a normal distribution, $p(d) \propto N(d_0, \sigma_d^2)$, and samples each have an orientation drawn from the local contour tangent plus a random orientation error (Fig. 1b). Because the generating contour has some local curvature κ, successive tangent samples will differ by an angle with expectation approximately κd. Assuming that turns in clockwise and counterclockwise directions are equally probable (corresponding to the assumption of an open contour; see Feldman & Singh, 2005), and that perturbations have mean 0°, the resulting angle between successive samples, called the *turning angle* and usually denoted α, will have expectation 0° and angular variance σ^2. We assume that this net turning angle has a von Mises distribution centered on 0°,

$$p(\alpha) \propto e^{\beta \cos \alpha}. \tag{4}$$

(see Feldman, 1995, 1997, 2001; Feldman & Singh, 2005; Singh & Feldman, 2012). The von Mises is the analog of a normal distribution suitable for angular measurements, with parameter β acting approximately like $1/\sigma^2$ (see Mardia, 1972).

This simple model yields a sequence of approximately equal-spaced and approximately collinear visual elements, with von Mises distributed inter-element turning angles. The elements technically form a Markov chain, because successive turning angles (changes in orientation) will be independent, meaning that non-adjacent orientations will be independent conditioned on the intervening orientations. (That is, each sample's orientation is independent of non-adjacent samples except via dependencies conveyed by intervening samples). A simple extension of this model is to assume that non-successive turning angles are positively correlated rather than independent, which introduces a bias towards

cocircularity in addition the bias towards collinearity, for which there is ample evidence (Singh & Fulvio, 2005, 2007; see also Singh & Feldman, 2012).

Given this contour-generating process, the mixture estimation problem consists of estimating the sources of the N edges: as a single smooth chain of N edges, two distinct chains each containing a subset, or any other partition conveyed by a set $\hat{Z}$ of estimated source labels (Fig. 3b). The prior inherently favors fewer components, because each additional contour g_i entails additional parameters θ_i, whose probabilities $p(\theta)$, when multiplied by the priors on other parameter, inevitably decrease the overall prior. The likelihood inevitably favors more components, and is maximized when each contour (perfectly) explains just one edge. Each grouping interpretation Y has a posterior proportional to the product of its prior and likelihood, which may be maximized at some intermediate number of components. A similar model was tested in Feldman (2001), in which subjects were asked to group sets of dots into some number of distinct smooth contours, e.g. grouping all of them into one smooth contour, or breaking them into two or more. In those studies, to a high degree of precision, each interpretation was chosen with a probability approximately proportional to its posterior.

Axial shapes

Finally, a more structurally complex class of visual patterns consists of complete closed shapes bounded by smooth contours. There are an infinite number of ways in which the shape of such objects might be parameterized, each potentially leading to a distinct probabilistic generative model. Many authors have argued that many natural shapes can be understood as combinations of distinct parts centered on elongated axes (Blum, 1973; Marr & Nishihara, 1978; Biederman, 1987), and the resulting axial representation has known neural correlates (Lee, Mumford, Romero, & Lamme, 1998; Hung, Carlson, & Connor, 2012; Lescroart & Biederman, 2012). Putting this in a Bayesian context, Feldman and Singh (2006) proposed that shape boundaries can be understood as data generated stochastically from a skeletal or axial model (see Feldman et al., 2013). In this framework,

axes are generated via a smooth curve process similar to that described for contours above, with a von Mises distribution of turning angle along each axis. Additional axes branch out from other axes in random directions, each being born with some constant probability p_C, resulting in a potentially complex, hierarchically organized skeletal structure (Fig. 1c). From this skeleton, random deviates (called *ribs*) sprout laterally from both sides, extending a distance that is normally distributed about a continuously varying mean, in a direction that is perpendicular on either side plus a von Mises distributed directional error. The endpoints of the ribs form a two-dimensional shape surrounding the original skeletal structure, which constitutes the data available to the observer. The resulting contours tend to form articulated shapes consisting of a set of interconnected axial parts. In this framework, each individual shape is a sample from an axially structured mixture component, and an image containing several shapes is a mixture of samples drawn from several distinct components. That is, just as an image is a mixture of objects, each object is a mixture of parts. (For example, Fig. 1c shows a configuration decomposed into two objects, one of which has multiple parts.) Grouping an image thus entails both decomposing it into objects and decomposing the objects into parts.

Feldman and Singh (2006) present many more details about shape representation and skeleton estimation in this framework, including an account of how the recovered axial structure leads to a breakdown of the shape into intuitive parts (Singh, Froyen, & Feldman, 2014). Natural classes of shapes, such as animals and leaves, have distinguishable skeletal parameters (Wilder, Feldman, & Singh, 2011), leading to specialized formulations of the prior model. In the current paper we restrict our attention to the generic model sketched above, focusing on the role such a generative model might play in a broader account of perceptual grouping.

The three generative classes given above—clusters, contours, and axial shapes—constitute a set of assumptions about the generative processes at work in the environment. Beyond these three one can imagine a nearly infinitely diverse set of more complex models suited to particular environments. Generally, alternative assumptions would lead to

alternative computational mechanisms. To Bayesians, inferential procedures inevitably reflect contingent knowledge and assumptions (they are "tuned" to the world; see Feldman, 2013), rather than supposedly *a priori* or domain-independent laws (as Gestalt principles were sometimes argued to be). It should be borne in mind that the explicitness of the link between the assumptions we have made and the computational procedures they entail is—at least to Bayesians—a feature, not a bug.

The Bayesian grouping interpretation

With a set of data-generating source classes at hand, we can now give a Bayesian statement of the grouping problem. Given a set $X = \{x_1, \ldots x_N\}$ of image elements, the degree of belief in a particular set of hypothesized generating sources $Y = \{(p_1, g_1) \ldots (p_K, g_K)\}$ is given by the posterior probability

$$\begin{aligned} p(Y|X) &= \frac{p(X|Y)p(Y)}{p(X)} \\ &\propto p(X|Y)p(Y) \end{aligned} \tag{5}$$

For fixed image data X, the posterior distribution $p(Y|X)$ assigns a probability to each possible grouping interpretation Y, proportional to the product of its prior probability $p(Y)$ and its likelihood (fit to the data) $p(X|Y)$. Generally interpretations with more sources (larger K) will have higher likelihood, because they allow the parameters of each source to be fit more closely to a (smaller) subset of the image data. In the limit a solution with $K = n$, in which each datum is interpreted as the product of its own individual source, will maximize the likelihood. But this tendency will be counterbalanced by any reasonable prior, because the prior $p(Y)$ is the product of the priors of the parameters of all the component sources, and hence generally diminishes with larger numbers of components. That is, the prior automatically favors fewer sources, while the likelihood favors more; the posterior, which

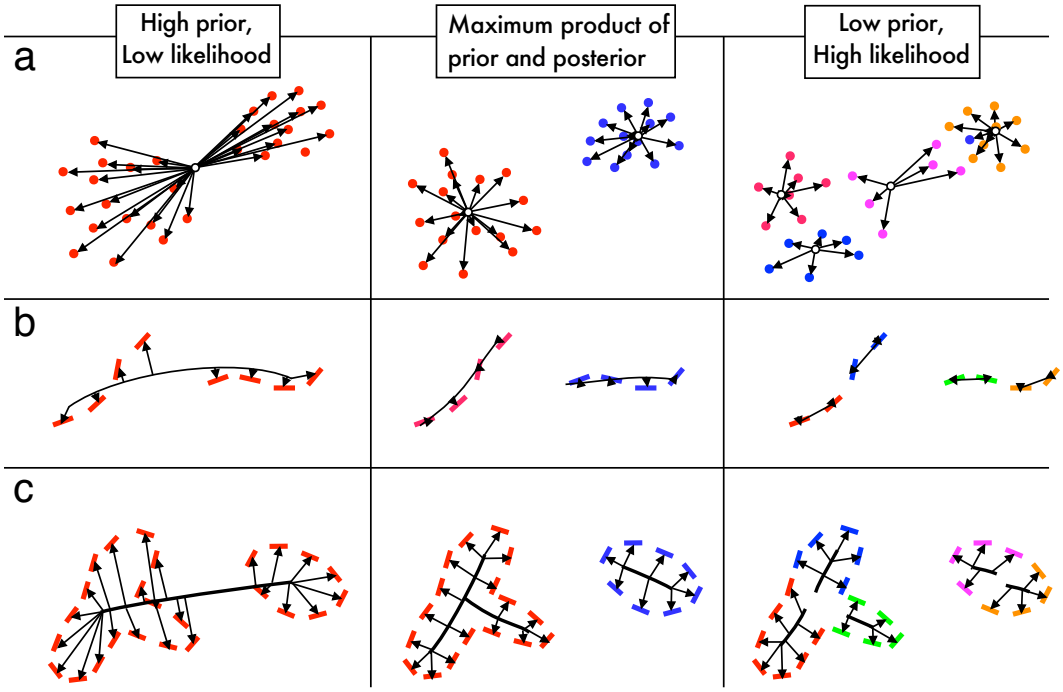

Figure 3. Schematic illustration of the range of possible mixture estimates for each of the three source classes. On the left, the estimated mixture has only a single component, which has high prior but concomitantly fits the image data poorly (low likelihood). On the right, the mixture has many components, which has lower prior but concomitantly fits the image data well (high likelihood). In the middle, the product of the prior and likelihood is maximal (maximizing the posterior), yielding a more intuitive mixture with a psychologically reasonable number of components.

is their product, favors interpretations that balance complexity with fit to the data (Jeffreys, 1939/1961; MacKay, 2003). Fig. 3 gives schematic examples of this tradeoff for the three source classes.

In perception, it is often assumed that the observer draws a single, unique interpretation (the "percept"). If we wish to restrict attention to one interpretation, a natural choice is the maximum a posteriori (MAP) interpretation,

$$Y_{MAP} = \arg\max_{Y} p(Y|X) \tag{6}$$

which maximizes the product of the prior and likelihood. Still, it should be kept in mind that the MAP is at best an imperfect substitute for the full posterior distribution, which includes posterior probabilities for *all* interpretations, and its use discards potentially

useful information. In many contexts in perception, such as any involving multistability or competing interpretations, one might want to sample from the posterior distribution rather than maximizing it (Moreno-Bote, Knill, & Pouget, 2011), which requires that the full posterior distribution (or an approximation thereof) be retained.

Description-length formulation

Shannon (1948) showed that any set of messages can be encoded with maximum efficiency (that is, minimum expected code length) if each message is assigned a code whose length is proportional to the negative logarithm of that message's probability. As a result the quantity $-\log p$ is often referred to as the *description length* (DL—although it should be kept in mind that the DL is only in fact the length of the description if the code is optimal). As perhaps first noticed by Rissanen (1978), it follows that the MAP interpretation of a dataset is also the interpretation with minimum DL (because the maximum posterior is also the minimum negative log posterior). In our setting, the MAP grouping interpretation Y_{MAP} is also the minimum DL interpretation, because its DL

$$\mathrm{DL}(Y_{MAP}|X) = \mathrm{DL}(X|Y_{MAP}) + \mathrm{DL}(Y_{MAP}) + \text{constant} \tag{7}$$

is smaller than the DLs of all other interpretations. (The constant, $-\log p(X)$, is independent of Y.) In this sense, the most likely grouping interpretation of the image I, given the assumed ensemble of potential generative sources, is also the *simplest* (minimum DL) interpretation.

This mathematically straightforward observation provides a rational basis for the idea that the best perceptual interpretation is also the simplest (although again it should be kept in mind that the connection depends on the assumption of an optimal code, a condition that can never be confirmed in practice). The idea that perception favors the simplest interpretation has a long history, ranging from the original Gestalt idea of *Prägnanz* (Kanizsa, 1979; Koffka, 1935) to a range of more concrete complexity measurement procedures (Attneave & Frost, 1969; Hochberg & McAlister, 1953; Leeuwenberg, 1971). Historically, simplicity principles have often been contrasted with those based on probability

maximization (Hatfield & Epstein, 1985; van der Helm, 2000). But as argued by Chater (1996), the current formulation suggests that the two principles are intimately connected or even identical (see Feldman, 2009).

The estimated number of objects

A simple but important application of this model concerns the estimation of the number of components $\hat{K}$, that is, the number of "groups" or objects apparently present in the visual field. The estimate $\hat{K}$ may reflect the marginal posterior distribution of K, $p(K|X)$, which gives the probability distribution over K given observed image X (that is, the posterior probability of each value K after observing X). The mean or expectation of this distribution is simply

$$E(\hat{K}) = \int_Y K_Y p(Y|X) dY, \tag{8}$$

where $K_Y = |Y|$ denotes the number of components in the interpretation Y. That is, the estimated number of components (or, more strictly, the average estimate over all interpretations) would be the number of components in each possible interpretation Y weighted by the posterior probability of that interpretation $p(Y|X)$. This is a probabilistically-weighted average of many integer-valued estimates, and thus need not itself be an integer. Alternatively, as discussed above, if we reduce the full posterior to the MAP, then the estimated number of components is simply the (integer) number of components in this interpretation, $K_{Y_{MAP}} = |Y_{MAP}|$. Note that the full posterior distribution of K retains information not present in this single estimate, such as the degree of belief in the most likely estimate, which may be far less than certainty in ambiguous cases. For example, in an array containing a number of imperfectly separable, overlapping groups, in which the true number of distinct groups is unclear to observers, the full posterior $p(K|X)$ allows predictions about the relative probability of various numerical estimates.

The close connection between numerical estimation and perceptual grouping has been well established (Compton & Logan, 1993; van Oeffelen & Vos, 1982). Indeed, any

assessment of the number of "units" present in the scene rests on some decomposition of the image into perceived objects (Feldman, 2003), and in this sense perceptual grouping inherently underlies the determination of visual numerosity (Juni et al., 2010) as well as perception of the properties of multi-part objects (Cohen et al., 2008). The mathematical connection given above simply expresses this connection formally, showing how intuitions about visual number can be related to a rational estimate of the number of groups.

Towards a performance theory

In this paper we have focused on the *theory* of the computation, Marr's (1982) term for an account that is abstracted away from details of the algorithm and implementation (also called the *competence* theory after a similar idea due to Chomsky). To build a comprehensive account it is essential to consider *performance* as well, that is, to describe a computational mechanism that approximates the competence within the constraints of available neural hardware. Froyen (2013) describes a tractable algorithmic framework that computes a hierarchically organized mixture estimate along the lines described above. Here we give a few remarks about how the grouping mixture posterior might be estimated in a parallel computational architecture consisting of a collection of nodes that (i) each receive evidence from a restricted neighborhood of the image, and (ii) communicate with only a limited set of neighboring nodes.

First, notice that computation of the likelihood in the Bayesian account is substantially *local* in nature. That is, imagine we divide the image into a set of (possibly overlapping) neighborhoods $\{X_1, X_2, \ldots\}$, that is, subsets of the image data ($X_j \subseteq X$) whose union is the full image ($X = \cup_j X_i$). Each of these neighborhoods can be explained by a local generative model Y_j with local likelihood function $p(X_j|Y_j)$. Because of the nature of the global generative model described above, these likelihoods will be approximately independent, meaning that the global likelihood function $p(X|Y)$ can be approximated by the product of the local likelihoods $\prod_j p(X_j|Y_j)$; or, equivalently, the global log likelihood is approximately the sum of the local log likelihoods, $\log p(X|Y) \approx \sum_j \log p(X_j|Y_j)$. This suggests a neural-

like arrangement in which nodes integrate local evidence within their "receptive fields" in favor of a particular grouping interpretation (cf. Doya, Ishii, Pouget, & Rao, 2007; Sotani & Wang, 2010; Yang & Shadlen, 2007), which can then propagate to neighboring nodes in a fashion organized so as to approximate the Bayesian posterior.

We illustrate this idea by showing how it can be used to give an estimate of figure and ground, a solution to which is implied by the mixture conceptualization. In the mixture model, the winning mixture estimate entails an interpretation of the *ownership* of each image element, that is, the generative source interpreted as most likely to have generated it. In our context, this in turn determines perceived figure/ground (border ownership), because the generative model assumes that shapes are generated "from the inside." That is, the figural status of each image element is interpreted so that the side with the winning skeleton is the interior of the shape. Hence a parallel estimate of the maximum posterior mixture ownership labels $\hat{Z}$ associated with fixed mixture model $\hat{Y}$ constitutes an interpretation of border ownership (figure/ground status) for all the contours present in the image. Note that when skeletons lose these competitions, their "explanation pool" (the set of elements they explain) is reduced, requiring re-estimation of the skeleton to better explain the remaining elements. In the extreme, some skeletons—those that lie in what are perceived as ground areas—lose *all* the competitions in which they are involved, in which case they effectively "drop out" and play no role in explaining the image.

Froyen, Feldman, and Singh (2010) demonstrated an implementation of this approach, in which the border ownership estimate was computed using units that communicate locally via Bayesian belief propagation (Pearl, 1988; more specifically see Weiss, 1999). This approach necessarily implies a network that includes both contour nodes, which represent local border ownership, as well as axial nodes, which represent ownership of contours by skeletons, both classes that have intriguing neural analogs (Craft, Schutze, Niebur, & von der Heydt, 2007). Fig. 4 shows some results of our implementation, showing figure/ground estimates drawn from a mixture-of-skeletons description for several critical shape configurations. In each case, the border ownership estimate derived

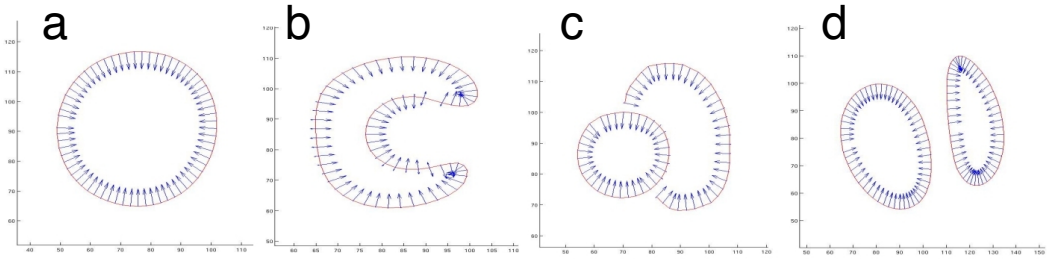

Figure 4. Figure/ground estimates (arrows point towards perceived interior of shape) from network model (Froyen et al., 2010). In the model, which implements part of the performance model described in the text, border ownership esimates spread by Bayesian belief propagation over networks of skeleton nodes. (a) simple closed shape (b) shape with deep indentation showing f/g reversal (c) overlapping shapes (d) non-overlapping shapes.

from the local skeletal posteriors corresponds to the perceived figural status point on each contour (indicated by inward-pointing arrows in the figure). These results are suggestive of the wide scope of problems in perceptual organization that can be simultaneously solved in a Bayesian framework once the decision problem (here, mixture estimates with entailed border ownership) has been suitably posed.

Discussion and conclusion

The Gestalt psychologists proposed a wide range of principles to describe human perceptual organization: depending on the author, as many as 114 (see Pomerantz, 1986) or as few as one (the unifying but vague Gestalt principle *Prägnanz*). The Bayesian approach to grouping replaces the heterogeneous "bag of tricks" exemplified by 114 separate rules—not to mention the hundreds of putatively distinct rules that have been proposed in the literature since—with a single unifying principle, Bayes' rule. Broadly speaking, we would argue that Bayesian mixture estimation can be seen as a realization of the comprehensive Gestalt principle *Prägnanz*. Viewed more carefully, the Bayesian approach is as diverse as the generative models it assumes. Indeed, as mentioned above, several of these generative models show obvious parallels with specific Gestalt rules. For example, the principle of proximity—nearby items should be grouped together—is in effect a strategy for decomposing isotropic (e.g. Gaussian) mixture components; the principle of good continuation

is a strategy for decomposing smooth contour components; and so forth. The Bayesian approach replaces a diversity of principles with a diversity of generative models—albeit all united under a single unifying principle, Bayes' rule. In this sense, each of these narrow Gestalt rules can be seen as a heuristic that helps the system achieve the Bayes optimal mixture estimate. Nevertheless our argument is that it only takes a handful of generative models—each of which has a natural, intuitive interpretation such as contours, shapes, etc.—to handle a wide range of pattern classes. Fundamentally, understanding perceptual grouping in terms of mixture estimation helps clarify the formal justification for these rules, and points the way towards more complete understanding of the computations that underlie them.

The Bayesian approach has several substantial advantages over the traditional approaches, including the Gestalt tradition. First, as just mentioned, it unifies many distinct rules of grouping—which have sometimes seemed like a dizzying collection of unrelated heuristic tendencies—under a common mathematical framework. The framework is internally coherent and motivated by well-defined goal: the attribution of belief in proportion to the Bayesian posterior (or, more broadly, the selection of action by minimization of a suitable loss function; Maloney & Zhang, 2010). Second, despite occasional criticism that the Bayesian approaches can encompass virtually any inference mechanism, Bayesian models can generate a wide range of quantitatively precise predictions from a relatively small set of assumptions. Third, the rationality of Bayesian inference means that the inferences drawn represent optimal use of the information and assumptions available to the observer (Jeffreys, 1939/1961; Jaynes, 2003). This rationality gives the Bayesian approach unique explanatory power, because in principle it can show how the particular perceptual mechanisms it posits actually serve to further the goals of the organism.

References

Attneave, F., & Frost, R. (1969). The determination of perceived tridimensional orientation by minimum criteria. *Perception & Psychophysics*, *6*, 391–396.

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychological Review*, *94*, 115–147.

Blum, H. (1973). Biological shape and visual science (Part I). *Journal of Theoretical Biology*, *38*, 205–287.

Brainard, D. H., Longere, P., Delahunt, P. B., Freeman, W. T., Kraft, J. M., & Xiao, B. (2006). Bayesian model of human color constancy. *Journal of Vision*, *6*(11), 1267–1281.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*(3), 566–581.

Cohen, E. H., Singh, M., & Maloney, L. T. (2008). Perceptual segmentation and the perceived orientation of dot clusters: the role of robust statistics. *Journal of Vision*, *8*(7), 1–13.

Compton, B. J., & Logan, G. D. (1993). Evaluating a computational model of pereptual grouping by proximity. *Perception & Psychophysics*, *53*(4), 403–421.

Cox, R. T. (1961). *The algebra of probable inference*. London: Oxford Universitiy Press.

Craft, E., Schutze, H., Niebur, E., & von der Heydt, R. (2007). A neural model of figure-ground organization. *Journal of Neurophysiology*, *97*(6), 4310–4326.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (2007). *Bayesian brain: probabilistic approaches to neural coding*. Cambridge, MA: M.I.T. Press.

Feldman, J. (1995). Perceptual models of small dot clusters. In I. J. Cox, P. Hansen, & B. Julesz (Eds.), *Partitioning data sets* (pp. 331–357). (DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 19)

Feldman, J. (1997). Curvilinearity, covariance, and regularity in perceptual groups. *Vision Research*, *37*(20), 2835–2848.

Feldman, J. (2001). Bayesian contour integration. *Perception & Psychophysics*, *63*(7), 1171–1182.

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences*, *7*(6), 252–256.

Feldman, J. (2009). Bayes and the simplicity principle in perception. *Psychological Review*, *116*(4), 875–887.

Feldman, J. (2013). Tuning your priors to the world. *Topics in Cognitive Science*, *5*(1), 13–34.

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

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Science*, *103*(47), 18014–18019.

Feldman, J., Singh, M., Briscoe, E., Froyen, V., Kim, S., & Wilder, J. D. (2013). An integrated Bayesian

approach to shape representation and perceptual organization. In S. Dickinson & Z. Pizlo (Eds.), *Shape perception in human and computer vision: an interdisciplinary perspective.* Springer.

Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field". *Vision Research*, *33*(2), 173–193.

Froyen, V. (2013). *Bayesian mixture estimation for perceptual grouping*. Unpublished doctoral dissertation, Rutgers University.

Froyen, V., Feldman, J., & Singh, M. (2010). A Bayesian framework for figure-ground interpretation. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems 23* (pp. 631–639).

Hatfield, G., & Epstein, W. (1985). The status of the minimum principle in the theoretical analysis of visual perception. *Psychological Bulletin*, *97*(2), 155–186.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural "goodness". *Journal of Experimental Psychology*, *46*, 361–364.

Hoffman, D., & Singh, M. (2012). Computational evolutionary perception. *Perception*, *41*, 1073–1091.

Hung, C. C., Carlson, E. T., & Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, *74*(6), 1099–1113.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Researh*, *39*(21), 3621–3629.

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

Jeffreys, H. (1939/1961). *Theory of probability (third edition)*. Oxford: Clarendon Press.

Juni, M. Z., Singh, M., & Maloney, L. T. (2010). Robust visual estimation as source separation. *J Vis*, *10*(14), 2.

Kanizsa, G. (1979). *Organization in vision: essays on Gestalt perception*. New York: Praeger Publishers.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271—304.

Knill, D. C. (2003). Mixture models and the probabilistic structure of depth cues. *Vision Res.*, *43*(7), 831–854.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Koenderink, J. J. (2011). Vision as a user interface. In B. E. Rogowitz & T. N. Pappas (Eds.), *Human vision and electronic imaging XVI; proceedings of the spie* (Vol. 7865, pp. 1–13). Bellingham, WA: SPIE.

Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt.

Kubovy, M., Holcombe, A. O., & Wagemans, J. (1998). On the lawfulness of grouping by proximity. *Cognitive Psychology*, *35*, 71–98.

Kubovy, M., & Wagemans, J. (1995). Grouping by proximity and multistability in dot lattices: a quantitative gestalt theory. *Psychological Science*, *6*(4), 225–234.

Lee, T. S., Mumford, D., Romero, R., & Lamme, V. A. F. (1998). The role of the primary visual cortex in higher level vision. *Vision Research*, *38*, 2429–2454.

Leeuwenberg, E. L. J. (1971). A perceptual coding language for visual and auditory patterns. *American Journal of Psychology*, *84*(3), 307–349.

Lescroart, M. D., & Biederman, I. (2012). Cortical representation of medial axis structure. *Cerebral Cortex*.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Maloney, L. T. (2002). Statistical decision theory and biological vision. In D. Heyer & R. Mausfeld (Eds.), *Perception and the physical world: Psychological and philosophical issues in perception* (pp. 145–189). New York: Wiley.

Maloney, L. T., & Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Research*, *50*, 2362–2374.

Mardia, K. V. (1972). *Statistics of directional data*. London: Academic Press.

Mark, J. T., Marion, B. B., & Hoffman, D. D. (2010). Natural selection and veridical perceptions. *J. Theoretical Biology*, *266*, 504–515.

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

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B.*, *200*, 269–294.

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

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Science*, *108*(30), 12491–12496.

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

Pizlo, Z., Salach-Goyska, M., & Rosenfeld, A. (1997). Curve detection in a noisy image. *Vision*

Research, *37*(9), 1217–1241.

Pomerantz, J. R. (1986). Visual form perception: an overview. In *Pattern recognition by humans and machines, vol. 2: Visual perception.* Orlando, FL: Academic Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465-471.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Singh, M., & Feldman, J. (2012). Principles of contour information: a response to Lim and Leek (2012). *Psychological Review*, *119*(3), 678–683.

Singh, M., Froyen, V., & Feldman, J. (2014). *Unifying parts and skeletons: a Bayesian approach to part decomposition.* (This volume)

Singh, M., & Fulvio, J. M. (2005). Visual extrapolation of contour geometry. *Proceedings of the National Academy of Sciences, USA*, *102*(3), 939–944.

Singh, M., & Fulvio, J. M. (2007). Bayesian contour extrapolation: Geometric determinants of good continuation. *Vision Research*, *47*, 783–798.

Singh, M., & Hoffman, D. D. (2013). Natural selection and shape perception. In S. Dickinson & Z. Pizlo (Eds.), *Shape perception in human and computer vision: An interdisciplinary perspective* (pp. 171–185). New York: Springer Verlag.

Smits, J. T., Vos, P. G., & van Oeffelen, M. P. (1985). The perception of a dotted line in noise: a model of good continuation and some experimental results. *Spatial Vision*, *1*(2), 163–177.

Sotani, A., & Wang, X.-J. (2010). Synaptic computation underlying probabilistic inference. *Nature Neuroscience*, *13*(1), 112–119.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Annals of statistics*, *28*(1), 40–74.

Uttal, W. R. (1973). The effect of deviations from linearity on the detection of dotted line patterns. *Vision Research*, *13*, 2155–2163.

van der Helm, P. (2000). Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, *126*(5), 770–800.

van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: counting by groups. *Memory & Cognition*, *10*(4), 396–404.

Weiss, Y. (1999). *Bayesian belief propagation for image understanding.* (In Workshop on Statistical and Computational Theories of Vision 1999)

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nat. Neurosci.*, *5*(6), 598–604.

Wilder, J., Feldman, J., & Singh, M. (2011). Superordinate shape classification using natural shape statistics. *Cognition*, *119*, 325–340.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, *447*, 1075–1082.

Zucker, S. W. (1985). Early orientation selection: Tangent fields and the dimensionality of their support. *Computer Vision, Graphics, and Image Processing*, *32*, 74–103.