Visual learning with reliable and unreliable features

A. Emin Orhan
Center for Visual Science, Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA

Melchi M. Michel
Center for Perceptual Systems, Department of Psychology, University of Texas at Austin, Austin, TX, USA

Robert A. Jacobs
Center for Visual Science, Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA

Existing studies of sensory integration demonstrate how the reliabilities of perceptual cues or features influence perceptual decisions. However, these studies tell us little about the influence of feature reliability on visual learning. In this article, we study the implications of feature reliability for perceptual learning in the context of binary classification tasks. We find that finite sets of training data (i.e., the stimuli and corresponding class labels used on training trials) contain different information about a learner's parameters associated with reliable versus unreliable features. In particular, the statistical information provided by a finite number of training trials strongly constrains the set of possible parameter values associated with unreliable features, but only weakly constrains the parameter values associated with reliable features. Analyses of human subjects' performances reveal that subjects were sensitive to this statistical information. Additional analyses examine why subjects were sub-optimal visual learners.

Keywords: computational modeling, learning, categorization

Citation: Orhan, A. E., Michel, M. M., & Jacobs, R. A. (2010). Visual learning with reliable and unreliable features. *Journal of Vision, 10*(2):2, 1–15, http://journalofvision.org/10/2/2/, doi:10.1167/10.2.2.

Introduction

For perceptual scientists interested in how people combine information from multiple sensory signals, the notion of cue or feature "reliability" is important. Reliability is typically defined in a statistical manner, as illustrated in Figure 1. Consider the information that a sensory cue or feature provides about a scene property, as quantified by the probability distribution $p(scene\ property \mid feature\ value)$. If this distribution has a small variance, then the feature provides highly precise or diagnostic information about the scene property and, thus, is regarded as a reliable feature. In contrast, if this distribution has a large variance, then the feature provides imprecise information about the scene property and is regarded as an unreliable feature.

Several studies have shown that people's estimates of a scene property based on multiple perceptual features can often be modeled as a weighted average of estimates based on individual features, where the weight associated with a feature is related to a feature's reliability (e.g., Battaglia, Jacobs, & Aslin, 2003; Ernst & Banks, 2002; Jacobs, 1999; Johnston, Cumming, & Landy, 1994; Knill & Saunders, 2003; Landy, Maloney, Johnston, & Young, 1995; Maloney & Landy, 1989; Young, Landy, & Maloney, 1993). For example, consider a person attempting to estimate the curvature of a surface that is both seen and touched. Information about the surface's curvature is provided by a visual stereo cue and by a haptic cue. Suppose that the visual stereo cue provides precise information about curvature (i.e., $p(curvature \mid stereo\ cue)$ has a small variance) and, thus, is reliable, but the haptic cue provides imprecise information (i.e., $p(curvature \mid haptic\ cue)$ has a large variance) and, thus, is unreliable. In this case, the model will form its estimate of curvature as a weighted average of the estimate based on the visual cue and the estimate based on the haptic cue. Because the stereo cue is more reliable, the curvature estimate based on this cue will be assigned a large weight. In contrast, the haptic cue is less reliable, meaning that the curvature estimate based on it will be assigned a small weight.

Existing studies of sensory integration demonstrate how the reliabilities of perceptual cues or features influence perceptual decisions. However, these studies tell us little about the influence of feature reliability on visual learning. Here, we study the implications of feature reliability for perceptual learning in the context of binary classification tasks. A main point of this article is that finite sets of training data (i.e., the stimuli and corresponding class labels used on training trials) contain different information about a learner's parameters associated with reliable

doi: 10.1167/10.2.2 Received June 23, 2009; published February 4, 2010 ISSN 1534-7362 © ARVO

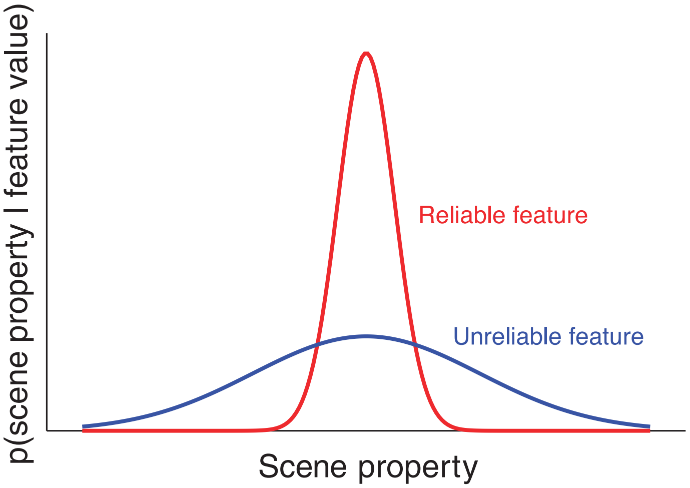

Figure 1. The probability distribution of a scene property given the value of a feature for reliable (red) and unreliable (blue) features.

versus unreliable features. In particular, the statistical information provided by a finite number of training trials strongly constrains the set of possible parameter values associated with unreliable features, but only weakly constrains the parameter values associated with reliable features.

To illustrate this point, consider a learner attempting to learn to perform a binary classification task. The learner performs, for instance, 600 training trials in which, on each trial, he or she views a stimulus and decides whether it belongs to class *A* or class *B*. Auditory feedback indicates the correctness of the learner's decision.

As shown in the left panel of Figure 2, each class of stimuli is represented by a two-dimensional normal distribution in which the mean vector is the class prototype and the covariance matrix characterizes the spread of exemplars around the prototype. In this example, feature X_1 is an unreliable indicator of class membership, whereas X_2 is a reliable indicator.

Suppose that the learner can be characterized as follows. On each trial, the learner calculates a sum of weighted feature values, $S = w_1x_1 + w_2x_2$, where x_1 and x_2 are the current stimulus values of the features X_1 and X_2, respectively, and w_1 and w_2 are the learner's weights or parameters. If the sum S is positive, the learner is likely to decide that the stimulus belongs to class *A*; otherwise, the learner is likely to decide that the stimulus belongs to class *B*.

To perform well on the classification task, the learner needs to discover good values for its parameters w_1 and w_2. For us, an important question is: How much information does the training data (i.e., the 600 stimuli and their corresponding class labels, which were presented on the training trials) provide about good values of the parameters? To address this question, we examine the probability distributions of the parameters given the training data, $p(w_1 \,|\, \{data\})$ and $p(w_2 \,|\, \{data\})$.

The middle and right panels of Figure 2 show hypothetical distributions $p(w_1 \,|\, \{data\})$ and $p(w_2 \,|\, \{data\})$ for the classification task illustrated in the left panel. For parameter w_1, the parameter associated with the unreliable feature X_1, the distribution is centered at zero and has a small variance. In other words, the training data indicate with high certainty that the value of this parameter should be zero. For parameter w_2, the parameter associated with the reliable feature X_2, the distribution is centered at a positive value and has a large variance. That is, the data indicate that feature X_2 should be positively weighted, but there is significant uncertainty as to the exact value to which w_2 should be set. Thus, according to the distributions in Figure 2, the training data provide very different statistical information about the parameters associated with reliable versus unreliable features.

In this article, we study the implications of feature reliability for visual learning in the context of binary classification tasks. We do so by examining the distribution of a learner's weights or parameters given a finite amount of training data. We analyze learning performances on two

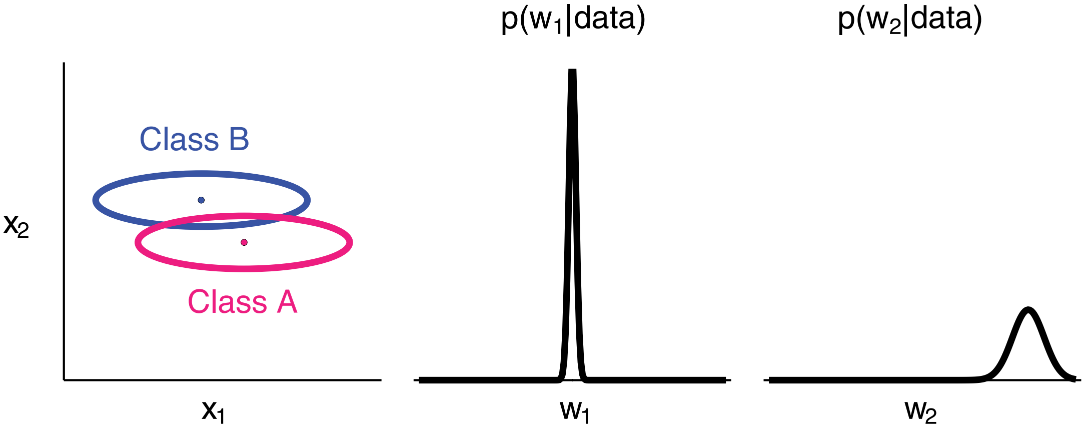

Figure 2. (Left panel) A two-dimensional binary classification task. (Middle and right panels) The probability distributions of weights w_1 and w_2, respectively, given a finite set of training data.

sets of tasks. One task is a simple two-dimensional binary classification task, useful for explanatory purposes (as illustrated by the discussion above and in Figure 2). The second task is a complex twenty-dimensional task that was used in Experiment 2 of Michel and Jacobs (2008). The performances of both human subjects and of computational models known as "ideal observers" are examined.

Our results reveal at least two important insights, one about the nature of some types of classification tasks and the other about the nature of human learners. In regard to classification tasks, we find results consistent with those described above with respect to the hypothetical example of Figure 2. That is, the posterior marginal distributions of weights associated with relatively unreliable features are centered at near-zero values and have small variances. The distributions of weights associated with reliable features, however, tend to have significantly larger variances. This means that the statistical information provided by the training data (as quantified by the distributions $p(w_i \mid \{data\})$ for all weights w_i, where $\{data\}$ refers to the finite set of visual stimuli and their corresponding class labels used on training trials) indicates with high precision that an unreliable feature is unreliable. In contrast, the information provided by the data indicates with low precision the exact relevance of a reliable feature.

An open research question is whether people are sensitive to this type of statistical information. One possible answer is that we are not. For example, it might be the case that human subjects show the same amount of uncertainty about how to quantify the extent to which a feature should be used regardless of whether the feature is reliable or unreliable. Alternatively, people might be sensitive to this task information. If so, they would learn with high certainty that an unreliable feature should not be used when performing the classification task. However, they would have much greater uncertainty about how to quantify the extent to which a reliable feature should be used. Our analyses based on a logistic regressor that was fit to subjects' responses in Experiment 2 of Michel and Jacobs (2008) suggest that these subjects were indeed sensitive to the task information about the different precisions for parameters associated with reliable and unreliable features.

Additional analyses indicate that subjects showed sub-optimal learning performances because they tended to dramatically underestimate the extent to which they should use reliable features. A possible explanation for this underestimation is that people are highly "regularized" learners, meaning that people are biased toward believing that features tend to be irrelevant (e.g., to a Bayesian statistician, it is as if we use prior distributions on our regression weights that are centered at zero and have small variances). Alternatively, it may be that subjects did not actually underestimate the extent to which they should use reliable features. Instead, it may be that their performances appear to be sub-optimal because they did not always exploit their own beliefs about visual stimuli. For instance, a subject may have believed that the probability that a stimulus belongs to class A is 0.7, but still judged the stimulus as belonging to class B on an experimental trial. If so, this would suggest that the subject engaged in "exploration", a strategy that can be useful in many learning situations (Bellman, 1956; Sutton & Barto, 1998).

This article is organized as follows. In the next section, we describe a two-dimensional binary classification task and our analysis of this task. Our goal is to study a small example where the intuitions underlying our approach can be easily explained and visualized. Following this, we describe Experiment 2 of Michel and Jacobs (2008) that is the source of the experimental data that we reevaluated. We then report the results of applying our techniques to these data. Lastly, we summarize our findings and draw final conclusions.

Two-dimensional binary classification task

In the two-dimensional binary classification task, each class of stimuli was represented by a bivariate normal distribution in which the mean vector was the class prototype and the covariance matrix characterized the spread of exemplars around the prototype (as was the case in Figure 2 described above). Class prototypes were placed at $[-1 \;\; 1]^T$ and $[1 \;\; -1]^T$. As illustrated in the leftmost column of Figure 3, three versions of the task were created differing in their covariance matrices. In all versions, the covariance matrices for classes A and B were identical, diagonal matrices. The covariance structures were isotropic in the first version, meaning that stimulus features X_1 and X_2 had equal variances ($\sigma_{X_1}{}^2 = \sigma_{X_2}{}^2 = 1$). Because of the placement of the mean vectors, and because these variances were equal, the two stimulus features were equally reliable indicators of class membership. The variance of X_1 was relatively large and the variance of X_2 was small in the second version ($\sigma_{X_1}{}^2 = 25$, $\sigma_{X_2}{}^2 = 1$). Consequently, X_1 was an unreliable indicator of class membership, whereas X_2 was reliable. In the final version, the variance of X_1 was small and the variance of X_2 was large ($\sigma_{X_1}{}^2 = 1$, $\sigma_{X_2}{}^2 = 25$), meaning that X_1 was a reliable feature, but X_2 was unreliable.

For a binary classification task in which classes are characterized by normal distributions, a logistic regressor maps a stimulus to the probability that the stimulus belongs to class A (one minus this value is the probability that a stimulus belongs to class B). Let $\vec{x} = [x_1 \; x_2]^T$ denote a stimulus where x_1 and x_2 are the stimulus values for features X_1 and X_2, respectively. Let $y = 1$ denote that the stimulus belongs to class A, and $y = 0$ denote that the stimulus belongs to class B. The logistic regressor works as follows. It first calculates a weighted sum, denoted S, of the stimulus feature values: $S = \sum_i w_i x_i$ where $\{w_i\}$ is the

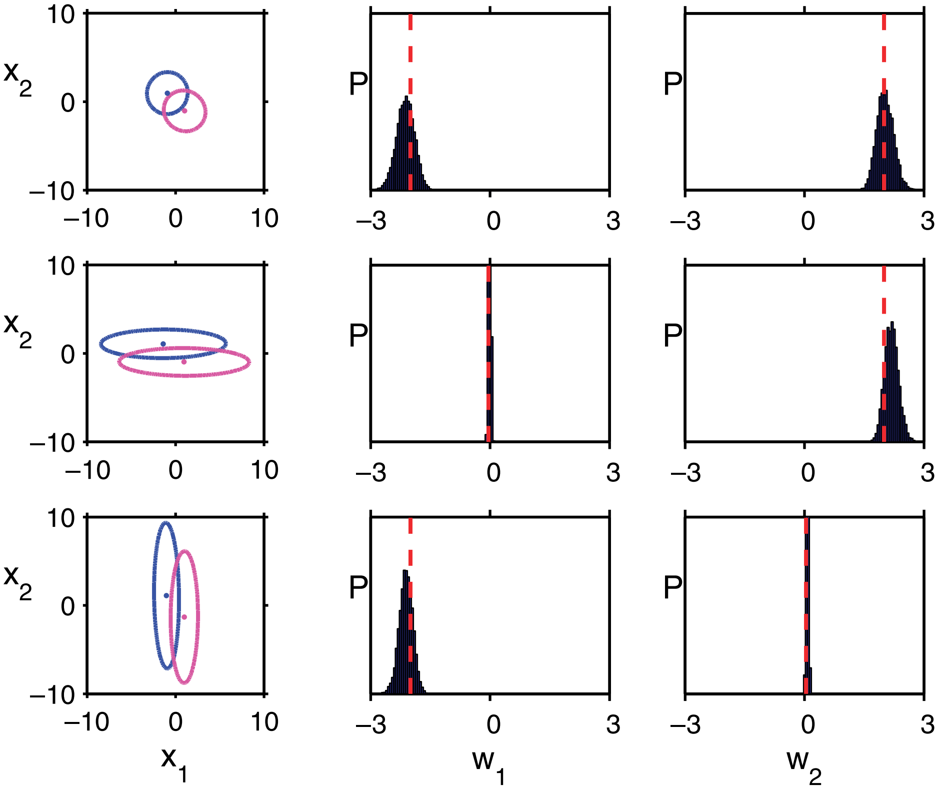

Figure 3. The left column shows three versions of the two-dimensional binary classification task. The middle and right columns show the posterior marginal distributions for parameters w_1 and w_2, respectively, for each task version. The point estimates of the parameter values for the ML model with infinite data are given by the red dashed lines.

set of parameters of the regressor. It then uses this weighted sum and the logistic function to calculate the probability that the stimulus belongs to class *A*: $p(y = 1 | \vec{x}) = 1 / (1 + e^{-S})$.

A maximum likelihood model and a Bayesian model were created for each version of the two-dimensional binary classification task. The models differed in how they inferred values for the weights or parameters $\vec{w} = [w_1\ w_2]^T$ of a logistic regressor. The maximum likelihood model is referred to as the ML model with infinite data. For each task version, its parameters were set to values that maximized the likelihood of a fictional data set containing an infinite number of data items: $w_i = (\mu_i^A - \mu_i^B) / \sigma_i^2$ where μ_i^A and μ_i^B are the values of feature X_i for the prototypes of classes *A* and *B*, respectively, and σ_i^2 is the variance of feature X_i (Bishop, 2006).

The Bayesian model used a finite data set for each task version. A data set consisting of 600 data items was created as follows. For each data item, a class was randomly selected, and then the normal distribution representing the selected class was sampled. The sample was assigned a class label in a stochastic manner using the probabilities from the ML model with infinite data (i.e., the true posterior probabilities $p(y = 1 | \vec{x})$ and $p(y = 0 | \vec{x})$). A data item consisted of the sample and the assigned class label. The Bayesian model inferred the joint distribution of the logistic parameters using a Markov chain Monte Carlo sampling method due to Holmes and Held (2006). (A summary of this method is provided in Appendix A.) Each parameter was assigned a vague prior distribution, namely $p(w_i) \sim N(0, 100^2)$. A single chain was run, and 100,000 samples were collected. The first 10,000 samples were discarded as burn-in. After examining the autocorrelation function of the samples, the chain was then thinned to every 10th sample to reduce correlations among nearby samples. Thus, the results for the Bayesian model were based on 9,000 samples.[1]

The middle and rightmost columns of Figure 3 show the results for parameters w_1 and w_2, respectively. The point estimates of the parameter values for the ML model with infinite data are given by the red dashed lines. The distributions are the posterior marginal distributions calculated by the Bayesian model.

These results reveal a number of important findings. First, the models show similar behaviors; the expected values of the parameters computed by the Bayesian model are very close to the point estimates of the ML model with infinite data. Second, the results of the Bayesian model show that the distributions for parameters associated with unreliable features are centered at near-zero values and have small variances. This means that the training data constrain or specify the values of these parameters with high precision. In contrast, the distributions for parameters associated with reliable features have larger variances,

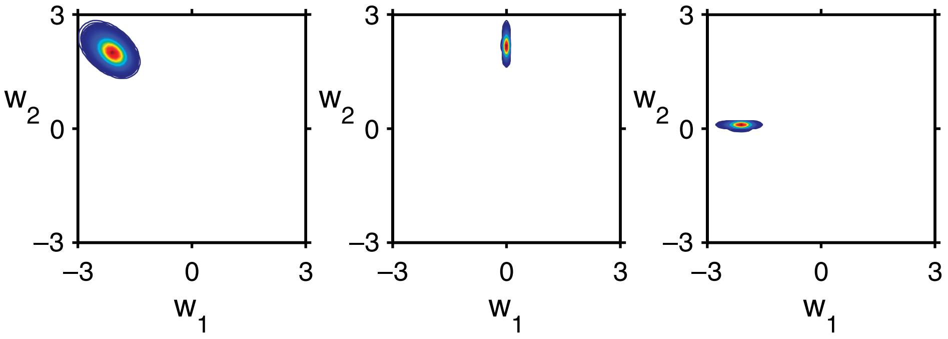

Figure 4. The likelihood of the data, $p(\{data\}\,|\,w_1,\ w_2)$, for each version of the two-dimensional binary classification task.

meaning that the data constrain the values of these parameters with significantly less precision.

To illustrate more clearly that the training data constrain the parameters associated with unreliable features with high precision, and constrain the parameters associated with reliable features with low precision, Figure 4 plots the likelihood function, $p(\{data\}\,|\,w_1,\ w_2)$, for each version of the task. For the first task version (left graph), in which stimulus features X_1 and X_2 are equally reliable, contours of equal likelihood are diagonally oriented ellipses. For the second task version (middle graph), in which X_1 was an unreliable feature and X_2 was reliable, the likelihood function in the local region near its peak is relatively steep along dimension w_1 and flat along dimension w_2. In other words, the likelihood changes quickly as the value of w_1 is perturbed. However, it changes slowly as the value of w_2 is perturbed. For the final task version (right graph), in which X_1 was a reliable feature and X_2 was unreliable, the likelihood changes slowly along w_1 and quickly along w_2.

Based on the results reported in this section, it seems that classification tasks of the type studied here place different constraints on parameters associated with reliable and unreliable features. Will people be sensitive to these task constraints? Our prediction is that the answer is yes. That is, we expect that a model that is fit to human subjects' responses while learning to perform a similar binary classification task will behave like the Bayesian model. It will learn with high certainty that unreliable features are indeed unreliable and, thus, should not be used for classification. However, it will have much greater uncertainty about how to quantify the extent to which reliable features should be used. If so, then the model suggests that people are sensitive to the task information about the different precisions for parameters associated with reliable and unreliable features. These predictions are evaluated below.

Experimental data set

We summarize Experiment 2 of Michel and Jacobs (2008) in this section. These investigators examined how people learn to combine information from arbitrary visual features when performing a set of perceptual discrimination tasks.

Visual stimuli were linear combinations of an underlying set of visual "basis" features or primitives (see Li, Levi & Klein, 2004; Olman & Kersten, 2004, for related approaches). These basis features are illustrated in Figure 5. At first glance, these features should seem to be arbitrary texture blobs. In fact, they are not completely arbitrary. They were created using an optimization procedure that yielded features that are orthogonal to each other (if features are written as vectors of pixel values, then the vectors are orthogonal to each other), relatively smooth (the optimization procedure minimized the sum of the Laplacian across each image), and equally salient (feature luminance-contrast values were normalized based on a feature's spatial frequency content).

Subjects performed a binary classification task. The prototype for each class was a linear combination of the basis features. The linear coefficients for class A were randomly set to either 1.0 or -1.0. The coefficients for class B were the negative of the coefficients for class A. In addition, a matrix K was added to each prototype where K consisted of the background luminance plus an arbitrary image constructed in the null space of the basis feature set (the addition of this arbitrary matrix prevented the prototypes from appearing as contrast-reversed versions of the same image). In summary, a prototype was computed using the following equation:

$$\text{prototype} = K + \sum_i c_i F_i, \tag{1}$$

where F_i is basis feature i and c_i is its corresponding linear coefficient.

Exemplars from a class were created by randomly perturbing the linear coefficients $\{c_i\}$ defining the prototype for that class. This was done using the following equation:

$$\text{exemplar} = K + \sum_i (c_i + \varepsilon_i) F_i, \tag{2}$$

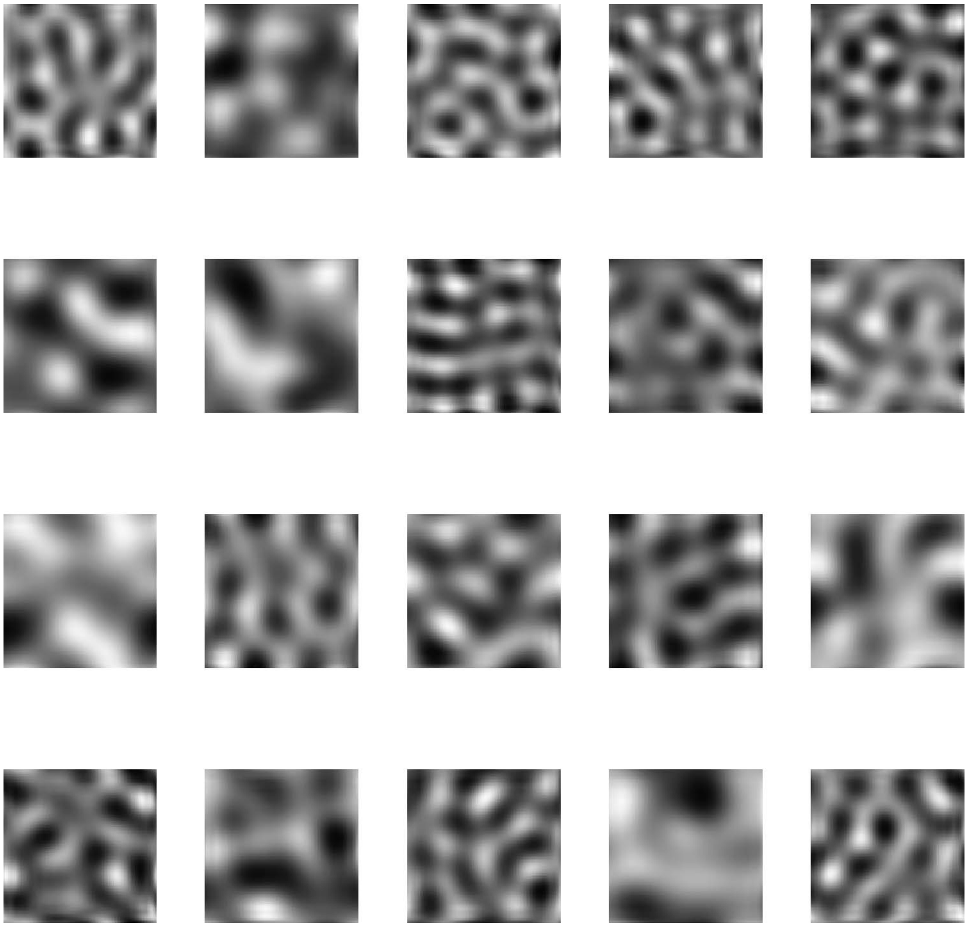

Figure 5. Set of 20 visual "basis" features or primitives.

where ε_i is a random sample from a normal distribution with mean zero and variance ${\sigma_i}^2$. This variance is referred to as a feature's noise variance. Importantly, each feature had its own noise variance, and the magnitude of this variance determined the reliability of a feature. Features with small noise variances tended to have coefficient values near one of the class prototypes. Therefore, these features were highly diagnostic of whether an exemplar belonged to class A or B. In contrast, features with large noise variances tended to have coefficient values far from the class prototypes. These features were less diagnostic of an exemplar's class membership. To avoid outliers, if a feature's coefficient value was more than two standard deviations from the corresponding value for the prototype, then this value was discarded and a new value was sampled. Consequently, the exemplars from the two classes were linearly separable.

Each trial of the experiment began with the presentation of a fixation square, followed by an exemplar, referred to as a test stimulus, followed by the prototypes of classes A and B. Subjects were instructed to decide which of the two prototypes had appeared in the test stimulus and responded by pressing the key corresponding to the selected prototype. Subjects received immediate auditory feedback after every trial indicating the correctness of their response. In addition, after every 15 trials, a printed message appeared on the screen indicating their (percent correct) performance on the previous 15 trials.

Each subject performed two classification tasks, Task 1 on days 1–3 (blocks 1–6) and Task 2 on days 4–6 (blocks 7–12). Importantly, the exemplars (but not the prototypes) were manipulated across the two tasks. This was accomplished by modifying the feature noise variances. In Task 1, half the features were randomly chosen to serve as reliable features for determining class membership. These features had a small noise variance (σ^2 = 1). The remaining features served as unreliable features and were assigned a large noise variance (σ^2 = 25). In Task 2, the roles of the two sets of features were swapped such that the reliable features were made unreliable, and the unreliable features were made reliable.

The authors predicted that people would learn to integrate information from the basis features based on the relative reliabilities of these features. Consequently, they expected subjects to successfully track the reliable versus unreliable features during the course of the experiment. When performing Task 1, they expected subjects to make their visual judgments on the basis of half the features—the reliable features—and ignore the remaining features. When performing Task 2, they expected subjects to flip their use of each feature. That is, they expected subjects' judgments to be based on the newly reliable features (the features that were previously ignored) and to ignore the newly unreliable features (the features that were previously the basis of subjects' judgments).

Subjects' data were analyzed using logistic regression in which the regression weights were estimated using maximum likelihood estimation. It was found that subjects successfully tracked the reliabilities of the visual basis features by tracking the noise variances of these features, and preferentially used the reliable features when performing each task. The results suggest that an expanded perspective on the standard model of cue combination described in the Introduction section is warranted. The model is applicable to tasks involving arbitrary perceptual signals that need to be learned, not just conventional perceptual cues that are highly familiar, to tasks involving many information sources, not just two sources, and to multi-task settings in which different cue combinations are optimal for different tasks, not just single-task settings (Michel, Brouwer, Jacobs, & Knill, 2010).

Simulation and analysis of ideal observers

We implemented computational models of the experimental data collected from each subject. The modeling results were qualitatively identical across subjects, and thus, for the sake of brevity, this article focuses on models of one subject's data (subject MSB). This section considers models that can be regarded as "ideal observers" in the sense that the models are based on the true posterior probabilities that a stimulus belonged to class A or B, as opposed to the subject's responses or estimates of the correct class labels (the latter is considered in the next section).

Computational models were logistic regressions. A maximum likelihood model and a Bayesian model were implemented. In the ML model with infinite data, denoted ML_{IO}^{∞}, the parameters were set to values that maximized the likelihood function based on a fictional data set containing an infinite number of data items. As described in the Two-dimensional binary classification task section, parameter w_i was set using the equation $w_i = (\mu_i^A - \mu_i^B) / \sigma_i^2$, where μ_i^A and μ_i^B are the values of feature X_i for the prototypes of classes A and B, respectively, and σ_i^2 is the variance of feature X_i (Bishop, 2006).

The Bayesian model, denoted BM_{IO}, used finite data sets based on the subject's experimental trials. Recall that the experiment contained two tasks in which the sets of reliable and unreliable features were swapped between tasks. The trials devoted to each task were divided into 6 blocks of 600 trials each. A data item used when estimating BM_{IO}'s parameter values consisted of representations of a test stimulus displayed on an experimental trial along with a class label for that stimulus. A stimulus was encoded by its representation in the space of visual basis features (i.e., the 20 linear coefficients used to construct the stimulus). The class label was set in a stochastic manner using the probabilities from the ML model with infinite data (i.e., the true posterior probabilities $p(y = 1 | \vec{x})$ and $p(y = 0 | \vec{x})$).

BM_{IO} used the set of data items associated with a single block of trials. Thus, it was simulated 12 times, once for each experimental block. On each simulation, the model inferred the joint distribution of its parameters using a Markov chain Monte Carlo sampling method (see Appendix A). Because the two classes of data items in a data set were linearly separable in the space defined by the visual basis features, there are many different logistic regressors that could be fit to a data set. That is, the data did not provide a strong constraint on the model's distributions of parameters. As a result, the sampling procedure of a model with a vague prior distribution [e.g., $p(w_i) \sim N(0, 100^2)$] often did not converge within a reasonable number of iterations. We used, therefore, a prior distribution on each parameter with a small variance [$p(w_i) \sim N(0, 2)$].[2] Three Markov chains were run, and 100,000 samples were collected from each chain (see Appendix A for details on how the chains were initialized). The Gelman–Rubin scale reduction factor was used to diagnose convergence (Gelman, 1996).[3] Based on this factor, the initial 10,000 samples from the first chain were discarded as burn-in. To reduce correlations among nearby samples, this chain was then thinned to every 10th sample. Thus, the posterior joint distributions of BM_{IO} were based on 9,000 samples.

The results are shown in Figure 6. The graphs on the left and right of Figure 6 are based on the trials in blocks 6 and 12, the final blocks for Tasks 1 and 2, respectively. The 20 sub-graphs within each graph correspond to the 20 parameters of a model. The point estimates of the parameter values for ML_{IO}^{∞} are given by the red dashed lines. The distributions are the posterior marginal distributions calculated by BM_{IO}.

Perhaps the most important outcome is that the distributions of parameters associated with unreliable features have relatively small variances, whereas the distributions of parameters associated with reliable features have large variances. Significantly, this outcome is identical to the outcome reported above when studying the two-dimensional binary classification task. It seems that in both the experimental task of Michel and Jacobs (2008) and the two-dimensional binary classification task, the information in a finite set of training data strongly constrains the set of possible parameter values associated with unreliable features, but only weakly constrains the possible parameter values associated with reliable features.

Recall that the experiment consisted of 12 blocks of trials. Figure 7 shows the absolute values of the means (left graph) and standard deviations (right graph) of BM_{IO}'s parameters across all experimental blocks. The black lines correspond to parameters associated with reliable features in Task 1 (unreliable in Task 2), and the red lines correspond to parameters associated with

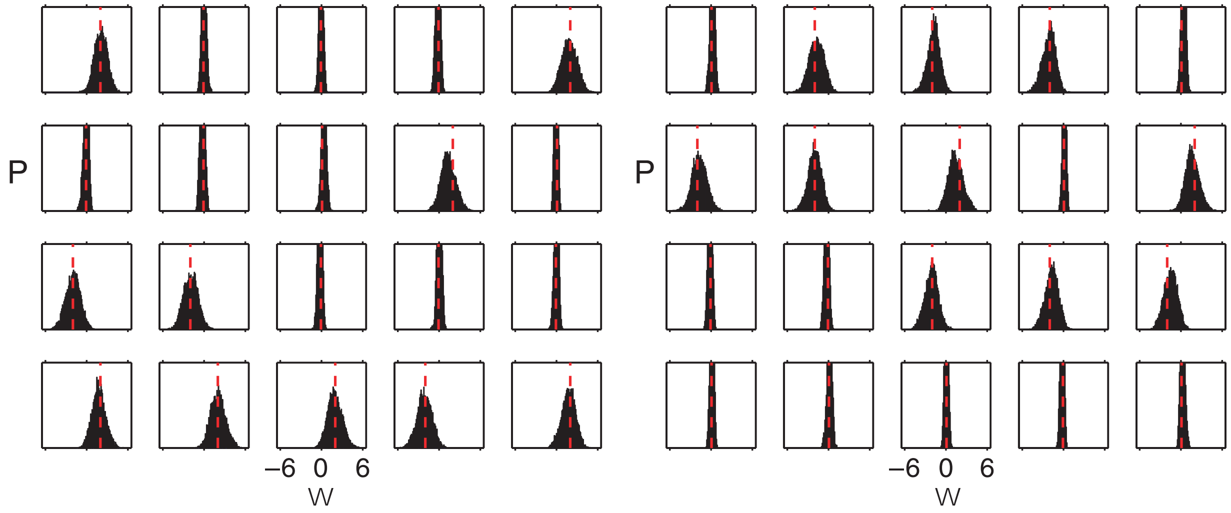

Figure 6. The posterior marginal distributions of the parameters of BM_{IO}. The graphs on the left and right are based on the trials in blocks 6 and 12, respectively. The 20 sub-graphs within each graph correspond to the 20 parameters of a model. The point estimates of the parameter values for ML_{IO}^{∞} are given by the red dashed lines.

unreliable features in Task 1 (reliable in Task 2). It seems that there are enough trials within a single block for BM_{IO} to learn the reliabilities of the features.

In summary, this section has reported the results of a Bayesian model that is an ideal observer in the sense that it is based on the true posterior probabilities that a stimulus belonged to class *A* or *B*. The most interesting result is that the posterior marginal distributions of the model's parameters had small variances for parameters associated with unreliable features, and large variances for parameters associated with reliable features. In other words, the information in the training data constrains the values of parameters associated with unreliable features with high precision but constrains the values of parameters associated with reliable features with low precision. We next report the data of a Bayesian model based on the subject's experimental data. That is, this model estimates the subject's response, or estimate of the class label, on each experimental trial.

Analysis of the experimental subject

We implemented a Bayesian model, denoted BM_{subj}, that used finite data sets based on the subject's trials in an experimental block. A data item consisted of representations of a test stimulus displayed on a trial along with the subject's response or estimate of the correct class label for

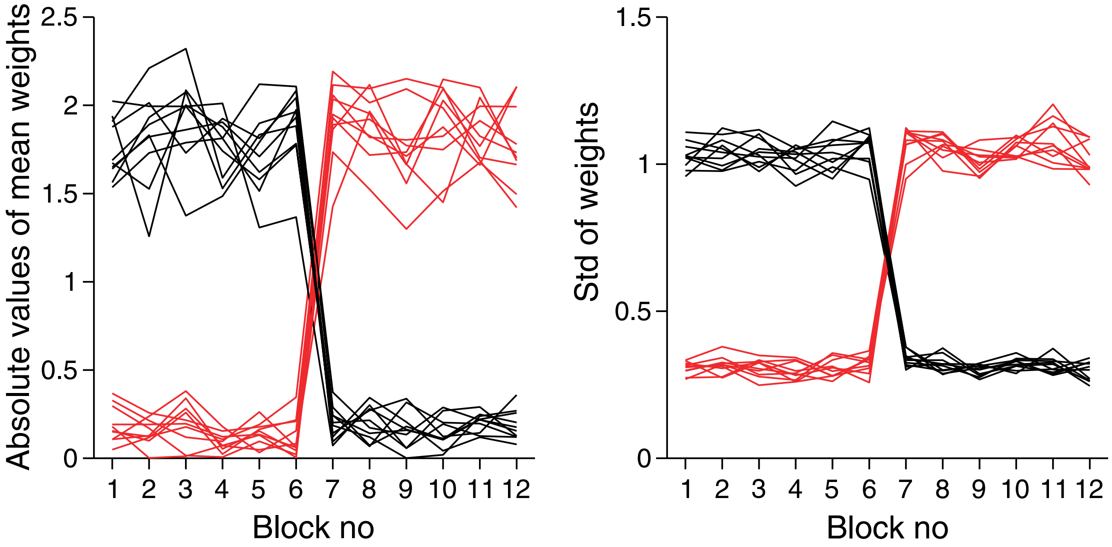

Figure 7. The absolute values of the means (left graph) and standard deviations of BM_{IO}'s parameters across all experimental blocks. Black lines correspond to parameters associated with reliable features in Stage 1 of the experiments (unreliable in Stage 2), and the red lines correspond to parameters associated with unreliable features in Stage 1 (reliable in Stage 2).

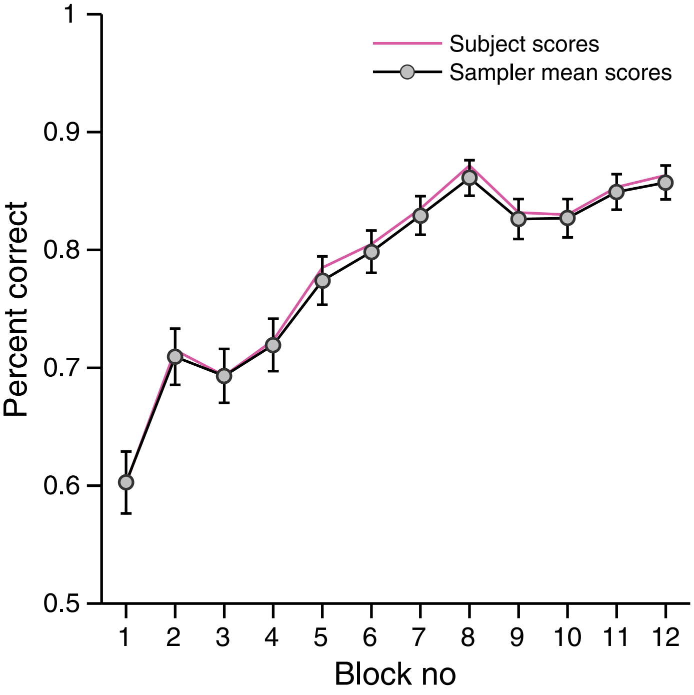

Figure 8. Subject's (percent correct) performance (magenta line) and the distribution of BM_{IO}'s performances (black dots and line; a dot indicates the mean and error bars denote one standard deviation around the mean) on each experimental block.

that stimulus. The model used a vague prior distribution [$p(w_i) \sim N(0, 100^2)$]. Three Markov chains were run, and 100,000 samples were collected from each chain (see Appendix A for further details). The Gelman–Rubin scale reduction factor was used to diagnose convergence (Gelman, 1996). Based on this factor, the first 10,000 samples from the first chain were discarded as burn-in. After examining the autocorrelation functions for the samples, the first chain was then thinned to every 10th sample to reduce correlations among nearby samples. The remaining samples were used to estimate the posterior joint distribution of BM_{subj}'s parameters.

Figure 8 shows the subject's (percent correct) performance (magenta line) and the distribution of BM_{subj}'s performances (black dots and lines; a dot indicates the mean and error bars denote one standard deviation around the mean) on each experimental block. The distribution of BM_{subj}'s performances on a block was obtained by sampling from its joint distribution of parameters. Clearly, BM_{subj} provides a good fit to the subject's performances.

Figure 9 shows the relationship between the parameter distributions for BM_{subj} and the point estimates of the ideal observer ML_{IO}^{∞}. Define the "normalized dot product" to be the quantity:

$$\frac{\vec{w}_{subj}^{T}\vec{w}_{IO}}{\|\vec{w}_{subj}\|\|\vec{w}_{IO}\|}, \tag{3}$$

where $\vec{w}_{subj}$ is a sample of parameter values drawn from the joint distribution of parameters for BM_{subj} and $\vec{w}_{IO}$ is the parameter point estimates of ML_{IO}^{∞}. This quantity is analogous to a correlation coefficient (Michel & Jacobs, 2008, referred to the square of this quantity as "template efficiency"). It is near one when $\vec{w}_{subj}$ and $\vec{w}_{IO}$ are similar, and near zero when $\vec{w}_{subj}$ and $\vec{w}_{IO}$ are unrelated. Figure 9 shows the median normalized dot product (error bars show the 25th and 75th percentiles of the distribution of normalized dot products) at each experimental block. The black points and line show the data based on the ideal observer ML_{IO}^{∞} for Task 1 of the experiment, whereas the red points and line are based on the ideal observer for Task 2. Clearly, the parameter values of BM_{subj} are closer to the optimal point estimates based on Task 1's stimulus noise structure during the first half of the experiment. They are closer to the optimal estimates based on Task 2's noise structure during the second half of the experiment.

Figure 10 shows the absolute values of the means (left graph) and standard deviations (right graph) of the parameter distributions for model BM_{subj} across all experimental blocks. Black lines correspond to parameters associated with reliable features in Task 1 (unreliable in Task 2), and red lines correspond to parameters associated with unreliable features in Task 1 (reliable in Task 2).

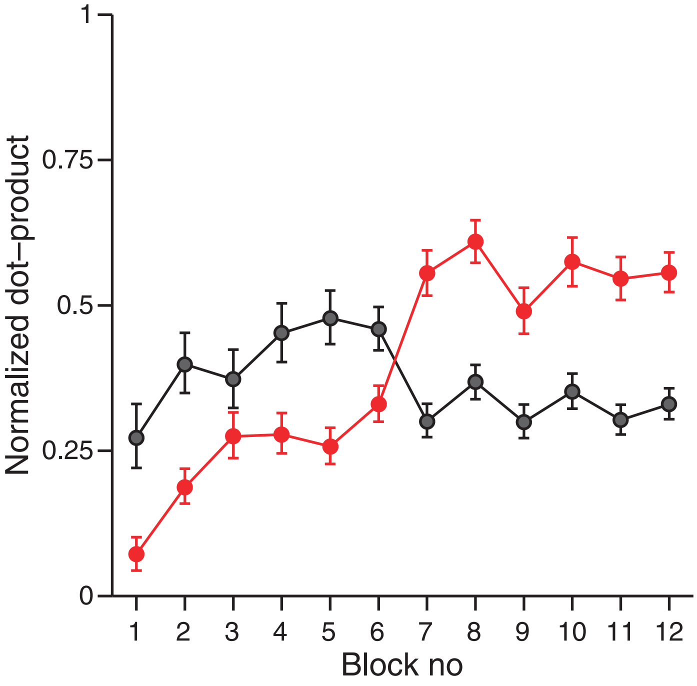

Figure 9. The median normalized dot product (error bars show the 25th and 75th percentiles of the distribution of normalized dot products) between BM_{subj}'s parameter values (or classification image) and the point estimates of the ideal observer ML_{IO}^{∞} at each experimental block. The black points and line show the data based on the ideal observer for Stage 1 of the experiment, and the red points and line show the data based on the ideal observer for Stage 2.

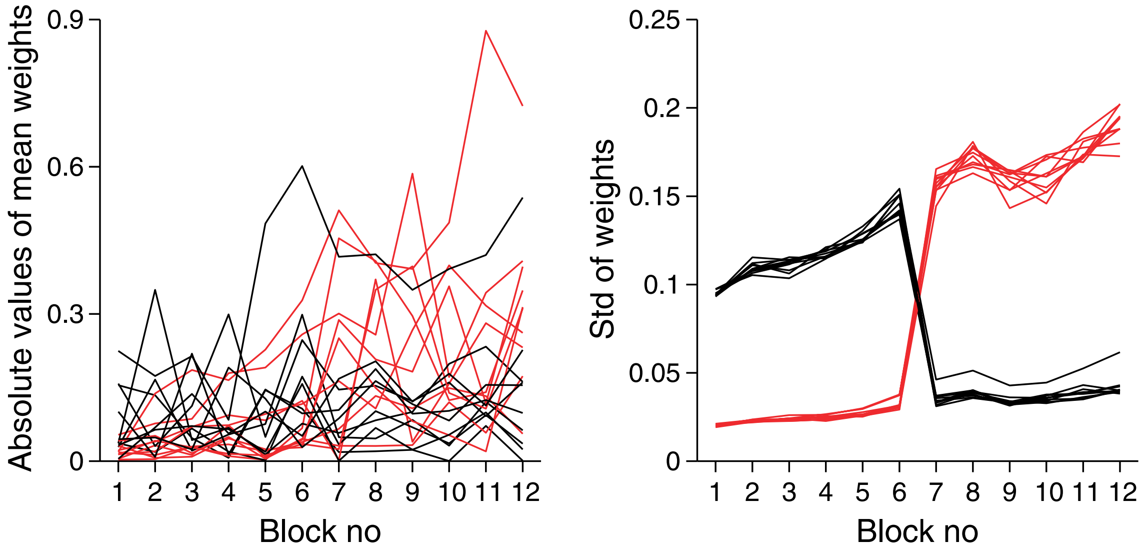

Figure 10. The absolute values of the means (left graph) and standard deviations of BM_{subj}'s parameters across all experimental blocks. Black lines correspond to parameters associated with reliable features in Stage 1 of the experiments (unreliable in Stage 2), and the red lines correspond to parameters associated with unreliable features in Stage 1 (reliable in Stage 2).

Although there is considerable noise in the mean data, the overall trend is expected; the black lines in the left graph tend to be at larger values in the first half of the experiment, and the red lines are at larger values in the second half. Importantly, the standard deviations are larger for parameters associated with reliable features, and smaller for parameters associated with unreliable features.

Figure 11 shows the posterior marginal distributions calculated by BM_{subj}. The graphs on the left and right are based on the trials in blocks 6 and 12, the final blocks for Tasks 1 and 2, respectively. The red lines show the parameter point estimates from ML_{IO}^{∞}, the ideal observer with infinite data described above (the red lines in Figures 6 and 11 are identical although the scales of the graphs are different).

It is informative to compare the distributions of BM_{subj} and BM_{IO}, the Bayesian models trained with the subject's responses and with the true posterior probabilities over class labels, respectively. Recall that BM_{IO}'s parameter distributions associated with unreliable features have small variances, and its distributions associated with reliable features have large variances. Above, we reasoned that this outcome follows from the nature of the constraints imposed by the training data. If people are sensitive to these constraints, then models that are fit to

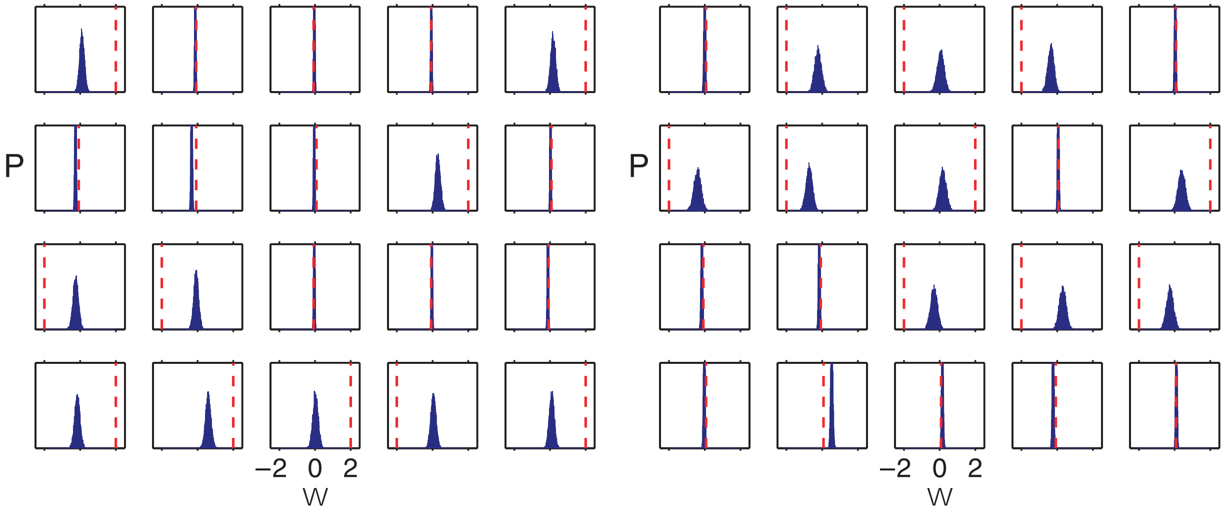

Figure 11. The posterior marginal distributions of the parameters of BM_{subj}. The graphs on the left and right are based on the trials in blocks 6 and 12, respectively. The 20 sub-graphs within each graph correspond to the 20 parameters of a model. The point estimates of the parameter values for ML_{IO}^{∞} are given by the red dashed lines.

human subjects' responses will show similar behaviors. The results of BM_{subj} displayed in Figure 11 verify that this is indeed the case. The distributions of BM_{subj}, like those of BM_{IO}, have significantly larger variances for parameters associated with reliable features.

Overall, the variances of BM_{subj}'s distributions are smaller than those of BM_{IO}. This can be explained by the fact that the set of stimuli that the subject labeled as class A and the set that he or she labeled as class B overlapped (in the space defined by the visual basis features), whereas the true classes did not. As a consequence, the training data for BM_{subj} placed strong constraints on BM_{subj}'s possible parameter values. The constraints placed by the training data for BM_{IO} were comparatively weaker.

A second interesting result about BM_{subj} is illustrated in Figure 11. BM_{subj}'s parameters typically have expected values with correct signs. On both blocks 6 and 12, the expected values of 8 of the 10 parameters associated with reliable features have the same signs as the optimal point estimates of the ideal observer ML_{IO}^{∞}. However, these values are much smaller (in magnitude) than the optimal point estimates. This result is surprising because the (percent correct) performance of BM_{subj} would be significantly improved if its parameter distributions were located at larger values.[4] There are at least two possible explanations for this outcome (see Eckstein, Abbey, Pham, & Shimozaki, 2004; Jacobs, 2009, for other discussions of sub-optimal visual learning).

One possible explanation is that the subject was a highly "regularized" learner. Within the fields of statistics and machine learning, it is typically the case that biases or constraints are added to a learning agent so that its parameter values remain within a desirable region. In probabilistic models, this is often achieved through the use of prior distributions on the parameters that are centered at zero and have small variances. Biased or regularized agents tend to be less sensitive to the idiosyncratic properties of the particular set of data items that they receive. That is, they tend to learn about the "signal" in their training data rather than the "noise" (Bishop, 2006). If the subject was a regularized learner constrained by a prior belief that visual features tend to be unreliable (as if the subject had a prior distribution on its parameters that was centered at zero and had a small variance), then this would provide an explanation as to why BM_{subj}'s posterior marginal parameter distributions are located at small values.

A second possible explanation is that a modified version of a logistic regressor is a better characterization of the subject's responses than the version we have studied so far. In this modified version, the weighted sum of a regressor's inputs, $S = \sum_i w_i x_i$, is mapped to the probability that the subject judged a stimulus as belonging to class A ($y = 1$) using a modified logistic function: $p(y = 1 | \vec{x}) = 1 / (1 + e^{-S/\beta})$ (the original logistic function is recovered by setting $\beta = 1$). In this new model, the parameter β is analogous to a variance parameter. If β is a small value (e.g., $\beta = 0.1$), then the model will tend to always believe that a stimulus belongs to class A with a probability of either 1 or 0 (intermediate probabilities will be rare). In this case, the model is essentially deterministic, and the model is said to "exploit" its current knowledge. If β is a large value (e.g., $\beta = 10$), the model will tend to always believe that a stimulus belongs to class A with an intermediate probability (extreme probabilities near 1 or 0 will be rare). It will appear to be at least partially random. For example, if the model believes that the probability that a stimulus belongs to class A is 0.6, then it will judge the stimulus as belonging to class A with a probability of 0.6 and will judge the stimulus as belonging to class B with a probability of 0.4. In this case, the model is said to "explore". In the field of machine learning, there is a lot of discussion about the advantages and disadvantages of exploration and exploitation. Exploration is often thought to be useful when a learner has incomplete knowledge of its environment or when an environment is non-stationary (Bellman, 1956; Sutton & Barto, 1998; note that the exploitation/exploration trade-off is closely related to a sub-optimal decision-making strategy known as "probability matching" [e.g., Newell, Lagnado, & Shanks, 2007]).

Importantly, there is a trade-off between the values of the parameters $\{w_i\}$ and the parameter β in the modified logistic function. Consider this new model where the expected values of the parameter values are relatively large in magnitude. In fact, suppose they are roughly equal to the optimal point estimates of the ideal observer ML_{IO}^{∞}. However, the parameter β in the new model is set to a moderately large value, meaning that the model is moderately random. This new model would show the same (percent correct) performance as the original model BM_{subj} (and as was shown by the subject). However, it leads to different implications about the subject's behavior. According to the original model, the subject was sub-optimal because he or she under-estimated the information carried by each reliable feature about a stimulus category. Based on the new model, the subject properly estimated the information carried by each feature, but the subject's performance was sub-optimal because he or she did not exploit this knowledge but rather engaged in exploratory behavior. Future research will need to design experiments to distinguish the predictions of these two models.

Summary and conclusions

In summary, we have studied the implications of feature reliability for perceptual learning in the context of binary classification tasks. We developed Bayesian ideal observer models, first for a two-dimensional binary classification task and then for a pattern discrimination task that was used in Experiment 2 of Michel and Jacobs (2008). Our results indicate that the marginal posterior distributions of parameters associated with unreliable features have relatively small variances, whereas the distributions of parameters associated with reliable features have large variances. That is, in both classification tasks, statistical information provided by the training data (as quantified by the distributions $p(w_i \,|\, \{data\})$ for all parameters w_i, where $\{data\}$ refers to the finite set of visual stimuli and their corresponding class labels used on training trials) strongly constrains the set of possible parameter values associated with unreliable features but only weakly constrains the possible parameter values associated with reliable features.

We then sought to determine if human observers performing the pattern discrimination task were sensitive to this statistical information. To this end, we applied the Bayesian model to a human subject's experimental data (i.e., the visual stimuli that the subject was exposed to and the subject's responses to these stimuli). We found that the subject was indeed sensitive to this type of task constraint. In addition, we found that for reliable features, parameter values inferred from the subject's data were significantly smaller (in magnitude) than the optimal point estimates of the same parameters. Two possible explanations for this result were provided. One possible explanation is that people performing this task might be "regularized" learners incorporating a strong bias toward small parameter values. Another possible explanation is that people might be engaging in exploratory behavior, rather than exploiting their potentially near-optimal knowledge regarding the parameter values associated with visual features.

An important aspect of the research reported here is that it makes use of Bayesian methods that, we believe, have important advantages over other approaches such as maximum likelihood estimation methods (Gelman, Carlin, Stern, & Rubin, 1995). For the study of human visual perception, where it is important to characterize the ambiguities of visual stimuli and the perceptual uncertainties underlying observers' actions, Bayesian methods are becoming essential research tools (e.g., Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996; Kuss, Jäkel, & Wichmann, 2005; Najemnik & Geisler, 2005; Yuille & Kersten, 2006). Bayesian inference computes full posterior distributions over parameters of a model, as opposed to single point estimates that are obtained through maximum likelihood estimation of the same parameters. Full posterior distributions provide more information about these parameters than point estimates, such as the expected values and variances of parameters, correlations between parameters, and the shape of the distributions over parameters. Bayesian methods also allow the use of prior information or expectations regarding parameters. For instance, if it is known *a priori* that a parameter is unlikely to have a large value, this information can be incorporated by placing an appropriately chosen prior distribution (one that has a small mass over large values) over that parameter. The use of prior information makes inference more robust and less variable by constraining the set of possible values that parameters can take.

In this article, we performed Bayesian inference using a Markov chain Monte Carlo sampling procedure. However, there is no reason to believe that the results reported here depend on the use of this specific procedure. Similar results would occur with other Bayesian inference procedures, such as the use of Laplace approximations, variational approximations, or the expectation propagation algorithm (Gelman et al., 1995; Jordan, Ghahramani, Jaakkola & Saul, 1999; Minka, 2001). Non-parametric sampling procedures, such as bootstrapping (Efron & Tibshirani, 1993), would also yield similar results (albeit with the necessity of alternative mathematical assumptions and possibly greater computational expense; see Hastie, Tibshirani, & Friedman, 2009).

Classical methods in both statistics (e.g., linear discriminant analysis) and machine learning (e.g., perceptrons) typically perform binary classification tasks by forming discriminant functions that are positive for stimuli in one class and negative for stimuli in the other class. When used to model human performance, these methods implicitly assume that a subject uses a single, deterministic discriminant function for making classification decisions. Because we have used Bayesian methods, we have taken a different approach by thinking of a subject's weights or parameters, and thus a subject's discriminant function, as random variables. That is, we have assumed that subjects maintained full distributions over discriminant functions. In future research, it might prove useful to think of other perceptual and cognitive variables as random variables too. Although we used a Bayesian logistic regression model here, the Bayesian approach can be applied to many other models attempting to explain other perceptual or cognitive phenomena.

Appendix A

This appendix provides details about the simulations that were not included in the main body of the text. We simulated logistic regressors in which the distributions of the regressors' weights or parameters were inferred using a Markov chain Monte Carlo (MCMC) sampler due to Holmes and Held (2006), henceforth referred to as H&H.

Let the ith data item consist of a vector of covariate variables, denoted $\vec{x}_i$, and a scalar response variable, denoted y_i. In addition, let $\vec{w}$ denote a logistic regressor's parameters. H&H introduced a latent variable, denoted z_i, such that

$$z_i = \vec{x}_i^T \vec{w} + \varepsilon_i, \tag{A1}$$

where ε_i is a sample from a standard logistic distribution. The response variable y_i is related to the latent variable z_i by the following equation:

$$y_i = \begin{Bmatrix} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{Bmatrix}. \tag{A2}$$

Let the prior distribution for the parameter vector $\vec{w}$ be a Gaussian distribution, denoted $N(\vec{m}, \sigma^2 I)$, with mean vector $\vec{m}$ and covariance matrix $\sigma^2 I$. In this case, it is difficult to construct an efficient Gibbs sampler because the full conditional distribution of $\vec{w}$ is not Gaussian. (This problem does not arise in probit regression where the noise variable ε_i is distributed according to a Gaussian distribution.) H&H solved this problem by introducing an additional latent variable, denoted λ_i, and by making the noise variable dependent on this new latent variable as follows:

$$\begin{aligned} &\varepsilon_i|\lambda_i \sim N(0,\lambda_i) \\ &\lambda_i = (2\psi_i)^2 \quad , \\ &\psi_i \sim KS \end{aligned} \tag{A3}$$

where KS is the Kolmogorov–Smirnov distribution. Importantly, the conditional distribution of ε_i given λ_i is Gaussian, whereas the marginal distribution of ε_i is logistic (Andrews & Mallows, 1974).

H&H used the following equations in their Gibbs sampler:

$$\begin{aligned} &z_i|\vec{w},\vec{x}_i,y_i \sim Logistic(\vec{x}_i^T\vec{w},1,y_i) \\ &\vec{w}|\vec{z},\vec{\lambda} \sim N(\vec{\mu},V) \\ &\vec{\mu} = V(\sigma^{-2}\vec{m} + X^T W \vec{z}) \\ &V = (\sigma^{-2}I + X^T W X)^{-1} \\ &W = diag(\lambda_1^{-1},\ldots,\lambda_n^{-1}). \end{aligned} \tag{A4}$$

Here $Logistic(\vec{x}_i^T\vec{w}, 1, y_i)$ is a truncated logistic distribution with mean $\vec{x}_i^T\vec{w}$, scale 1, and the side of the distribution that is truncated is determined by y_i: if y_i = 1, the distribution is truncated below 0; otherwise, it is truncated above 0. In these equations, X is a matrix whose ith row is the covariate variable $\vec{x}_i$, and $\vec{z}$ and $\vec{\lambda}$ are vectors containing the variables $\{z_i\}$ and $\{\lambda_i\}$, respectively. H&H used a rejection sampling method to sample from the conditional distribution of λ_i because this distribution does not have a standard form.

For the two-dimensional binary classification task, the Bayesian model (BM_{IO}) produced a single chain of 100,000 samples. The variables $\{\lambda_i\}$ were initialized to 1, and the variables $\{z_i\}$ were initialized to values sampled from a truncated logistic distribution with mean parameter 0 and scale parameter 1 (the side of truncation depended on y_i). The first 10,000 samples of the chain were discarded as burn-in, and the remaining samples were then thinned to every 10th sample.

For the experimental data set, BM_{IO} and BM_{subj} each produced three chains of 100,000 samples for each experimental block. In Chain 1, the variables $\{\lambda_i\}$ were initialized to 1, and the variables $\{z_i\}$ were initialized to values sampled from a truncated logistic distribution with mean parameter 0 and scale parameter 1. In Chain 2, the variables $\{\lambda_i\}$ were initialized to values sampled from a uniform distribution on the interval [0.5, 1.5], and the variables $\{z_i\}$ were initialized to values sampled from a truncated logistic distribution whose mean was sampled from a uniform distribution on the interval [0, 1] and whose scale was set to 5. Chain 3 was initialized in the same manner as Chain 2. Relative to Chain 2, however, it reversed the update order of the variables $\{z_i\}$ and $\{\lambda_i\}$. The first 10,000 samples of Chain 1 were discarded as burn-in, and the remaining samples were thinned to every 10th sample.

Acknowledgments

We thank P. Dayan for helpful discussions. This work was supported by NSF Research Grant DRL-0817250.

Commercial relationships: none.

Corresponding author: Robert A. Jacobs.
Email: robbie@bcs.rochester.edu.
Address: Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.

Footnotes

[1] In our research, we also considered models containing lapse parameters (Wichmann & Hill, 2001). These models are useful when subjects' responses seem to be random (stimulus-independent) guesses on significant numbers of trials. However, we found that the subjects in Michel and Jacobs (2008) had small lapse rates, and thus, we omit models with lapse parameters from this article.

[2] For a binary classification task with linearly separable classes, a maximum likelihood estimator of a logistic regressor's weights is not well defined because the likelihood function can always be increased by increasing the magnitudes of the weights. To circumvent this problem,

practitioners typically seek weights that maximize the likelihood function and are not too large in magnitude (so-called maximum penalized likelihood estimation). In a Bayesian setting, this corresponds to placing a relatively restrictive prior distribution on the logistic weights.

[3] Roughly, the Gelman–Rubin scale reduction factor is a mathematical tool designed to detect when multiple chains, each initialized in its own way, are showing similar statistical properties, meaning that the chains have converged to the same distribution. The time period prior to convergence is referred to as "burn-in", and the chains' samples during burn-in are discarded.

[4] The subject's performance (and, thus, BM_{subj}'s performance) was sub-optimal. To better understand why, we did the following. We fit a logistic regressor to the subject's responses using maximum likelihood estimation. It could be that the vector of parameter estimates is too small in magnitude, points in the wrong direction, or both. We scaled the magnitude of this vector, maintaining its direction, and measured the performance of a logistic regressor whose parameter values were set to this scaled vector. By increasing the magnitude of the vector, a logistic regressor could increase its performance from about 77% correct to 83% correct on block 6, and from 83% correct to 90% correct on block 12. The remaining error is due to the fact that this vector points in the wrong direction.

References

Andrews, D., & Mallows, C. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B, 36,* 99–102.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 20,* 1391–1397. [PubMed]

Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhya, 16,* 221–229.

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

Eckstein, M. P., Abbey, C. K., Pham, B. T., & Shimozaki, S. S. (2004). Perceptual learning through optimization of attentional weighting: Human versus optimal Bayesian learner. *Journal of Vision, 4*(12):3, 1006–1019, http://journalofvision.org/4/12/3/, doi:10.1167/4.12.3. [PubMed] [Article]

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415,* 429–433. [PubMed]

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 131–143). London: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman and Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.

Holmes, C. C., & Held, L. (2006). Bayesian auxiliary variable methods for binary and multinomial regression. *Bayesian Analysis, 1,* 145–168.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research, 39,* 3621–3629. [PubMed]

Jacobs, R. A. (2009). Adaptive precision pooling of model neuron activities predicts the efficiency of human visual learning. *Journal of Vision, 9*(4):22, 1–15, http://journalofvision.org/9/4/22/, doi:10.1167/9.4.22. [PubMed] [Article]

Johnston, E. B., Cumming, B. G., & Landy, M. S. (1994). Integration of stereopsis and motion shape cues. *Vision Research, 34,* 2259–2275. [PubMed]

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 105–162). Cambridge, MA: MIT Press.

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

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

Knill, D. C., & Saunders, J. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research, 43,* 2539–2558. [PubMed]

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision, 5*(5):8, 478–492, http://journalofvision.org/5/5/8/, doi:10.1167/5.5.8. [PubMed] [Article]

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

Li, R. W., Levi, D. M., & Klein, S. A. (2004). Perceptual learning improves efficiency by re-tuning the decision "template" for position discrimination. *Nature Neuroscience, 7,* 178–183. [PubMed]

Maloney, L. T., & Landy, M. S. (1989). A statistical framework for robust fusion of depth information. *Visual Communications and Image Processing IV: Proceedings of the SPIE, 1199,* 1154–1163.

Michel, M. M., Brouwer, A.-M., Jacobs, R. A., & Knill, D. C. (2010). Optimality principles apply to a broad range of information integration problems in perception and action. In J. Trommershäuser, M. S. Landy, & K. Körding (Eds.), *Sensory cue combination*. New York: Oxford University Press.

Michel, M. M., & Jacobs, R. A. (2008). Learning optimal integration of arbitrary features in a perceptual discrimination task. *Journal of Vision, 8*(2):3, 1–16, http://journalofvision.org/8/2/3/, doi:10.1167/8.2.3. [PubMed] [Article]

Minka, T. (2001). Expectation propagation for approximate Bayesian inference. In J. Breese & D. Koller (Eds.), *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 362–369). New York: Morgan Kaufmann.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434,* 387–391. [PubMed]

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2007). *Straight choices: The psychology of decision making*. New York: Psychology Press.

Olman, C., & Kersten, D. (2004). Classification objects, ideal observers, and generative models. *Cognitive Science, 28,* 227–240.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press. [PubMed]

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness-of-fit. *Perception and Psychophysics, 63,* 1293–1313. [PubMed] [Article]

Young, M. J., Landy, M. S., & Maloney, L. T. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Research, 33,* 2685–2696. [PubMed]

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Science, 10,* 301–308. [PubMed]