Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments

Craig K. Abbey
Department of Biomedical Engineering, University of California, Davis, CA, USA

Miguel P. Eckstein
Department of Psychology, University of California, Santa Barbara, CA, USA

We consider estimation and statistical hypothesis testing on classification images obtained from the two-alternative forced-choice experimental paradigm. We begin with a probabilistic model of task performance for simple forced-choice detection and discrimination tasks. Particular attention is paid to general linear filter models because these models lead to a direct interpretation of the classification image as an estimate of the filter weights. We then describe an estimation procedure for obtaining classification images from observer data. A number of statistical tests are presented for testing various hypotheses from classification images based on some more compact set of features derived from them. As an example of how the methods we describe can be used, we present a case study investigating detection of a Gaussian bump profile.

Keywords: classification image, linear template, detection, two-alternative forced-choice

Introduction

There has been considerable recent and historical interest in understanding how human observers perform visual tasks in noisy images. Image noise is often used as a tool for teasing out basic information about how the visual system works (e.g., Burgess, Wagner, Jennings, & Barlow, 1981; Pelli, 1981; Legge, Kersten, & Burgess, 1987; Pelli & Farell, 1999). Additionally, in some applied fields, image noise can limit the utility of images acquired for medical or scientific purposes (e.g., Revesz, Kundel, & Graber, 1979; Barrett, 1990). For several years, investigators have been building models of how basic tasks are performed using noisy image stimuli. These models are typically validated through comparisons with performance. A good model is one that matches the performance of human observers, either in terms of absolute performance or in terms of performance trends.

Performance comparisons are very useful for rejecting models that do not qualitatively match human performance. Furthermore, models that are outperformed on an absolute scale by human observers have likely failed to capture some vital image component used by the visual system. However, validating models by performance comparisons alone generally does not uniquely determine how a visual task is performed. Different models that yield indistinguishable performance indices over a range of conditions can often be found. One consequence of this situation is that mechanisms derived from a model that fits human data may not be representative of the actual mechanisms used by human observers. Additionally, the ambiguity of performance comparisons makes it unclear that an observer that predicts performance effects of human observers in one circumstance will generalize to other situations.

Studies involving classification images - an alternative to performance comparisons for visual tasks - have been used recently in increasing numbers to study visual strategies in a variety of tasks. The basis for these methods comes from works of Ahumada and coworkers in audition in the 1970s (Ahumada & Lovell, 1971; Ahumada, Marken, & Sandusky, 1975). The basic idea is that the stimuli used in an experiment, along with an observer's decisions based on those stimuli, contain information about how the task is performed. Averaging stimuli, grouped by whether the observer made a correct or incorrect decision, yields a profile of how the observer weighted the stimuli that Ahumada termed the "classification image." Ahumada (1996) first applied the classification image methodology to a visual task in vernier acuity. Since then, there has been a steadily growing number of studies reporting experimental results of classification image studies for vernier acuity tasks (Beard & Ahumada, 1998; Barth, Beard, & Ahumada, 1999), detection in temporal image sequences (Knoblauch, Thomas, & D'Zmura, 1999), orientation discrimination (Solomon, 2000), and illusory contours (Gold, Murray, Bennett, & Sekuler, 2000). There have also been efforts to extend the estimation of classification images to other experimental paradigms, such as

DOI 10:1167/2.1.5 Received July 1, 2001; published January 28, 2002 ISSN 1534-7362 © 2002 ARVO

multiclass identification (Watson, 1998) and two-alternative forced-choice tasks (Abbey, Eckstein, & Bochud, 1999), as well as to images with correlated noise textures (Abbey & Eckstein, 2000; Edwards, Kupinski, Nishikawa, & Metz, 2000).

One aspect that has been less developed in this growing body of literature is a more rigorous analysis of the estimation problem at the core of methods to obtain classification images. The benefit of such an analysis is that casting the problem as a formal estimation problem allows the methods of statistical point estimation and inference to be brought to bear on the problem. Our goal here is to begin filling this gap. We considered two-alternative forced-choice (2AFC) classification tasks in which a known target is to be distinguished from a known alternative.[1] The target and alternative are presumed to be masked by noise with a multivariate Gaussian distribution. This class of images includes commonly used white-noise images as well as Gaussian-distributed textures that contain spatial or temporal correlations.

Throughout this work we often appeal to the notion of a linear observer that performs a 2AFC task by some weighted linear integration of each stimulus in the spatial and/or temporal domain. In this case, the observer strategy for performing the task is encoded in the weights used in the integration. The linear model is a useful starting point for classification image analysis because, as we shall see below, the classification image is closely related to the weights used by a linear observer. Furthermore, linear observer models have a history of use for modeling human-observer performance in noise-limited[2] simple detection and discrimination tasks (Rose, 1948; Burgess & Ghandeharian, 1984a and 1984b; Ahumada & Watson, 1985; Barrett, Yao, Rolland, & Myers, 1993). However, the visual system is demonstrably nonlinear in many circumstances, and nonlinear models of visual processes are widespread. Linear models may still be applicable in these cases. For example, in the presence of nonlinear transduction (Foley & Legge, 1981; Legge et al., 1987; Lu & Dosher, 1999), there may still be regions in which the transducer function is linear. Tasks that operate predominantly in such a region will behave much like a linear observer. When spatial uncertainty is posited as the nonlinear mechanism (Tanner, 1961; Nachmias & Kocher, 1970; Pelli, 1985; Eckstein, Ahumada, & Watson, 1997), then tasks that have a low degree of uncertainty will limit to a linear observer (Cohn, Thibos, & Kleinstein, 1974). Finally, nonlinearities in the visual system may be well approximated locally by a linear function. In this case, a linearized observer may be a good approximation for a given task (Ahumada, 1987).

We begin by reviewing a general model of 2AFC detection and discrimination. This allows us to describe a general theoretical framework in which to analyze classification images, and to define consistent notational conventions. We then turn to the main results of this work, which are procedures for estimating and performing statistical hypothesis testing on classification images. The estimation procedure presented here is modified somewhat from previous work (Abbey et al., 1999; Abbey & Eckstein, 2000) to be more efficient, and we describe sample methods for estimating the magnitude of errors in classification images. The hypothesis tests use feature vectors as a way to reduce degrees of freedom. Averaging these feature vectors over many 2AFC trials leads to a number of Hotelling T^2 tests. The tests include departure from a hypothesized mean classification image, two tests for differences between classification images, and a test for a nonlinear observer response function. A simple case study is presented as an example of how these methods can be used to make inferences from classification image data.

Modeling Simple Forced-Choice Tasks

Here we review a general framework for how two-stimuli 2AFC visual tasks are performed on noisy images. We restrict our attention to simple tasks involving the discrimination of a known target profile from a known alternative. The approach is based on the formation of scalar internal response variables that reflect the visual strategy used by the observer. The case of a response variable that is a linear function of the image with some associated internal noise is given special attention because linear models lead to a direct interpretation of classification images. We describe how a decision, and hence the outcome of each experimental trial, is made from the response variables and how this decision relates to figures of merit for task performance.

In this work, we consider an image stimulus to be a vector of pixel intensities, denoted generically by $\mathbf{g}$. We use the convention of bold lowercase symbols to indicate vector quantities, bold uppercase to indicate matrix quantities, and nonbold symbols to indicate scalars. We denote the images corresponding to each alternative of a forced-choice trial as $\mathbf{g}^+$ and $\mathbf{g}^-$for signal-present and signal-absent images, respectively. When necessary, we use an index, j, to denote the experimental trial. In this case, $\mathbf{g}_j^+$ denotes the signal-present image vector for the jth trial.

Distribution of Images

We divide an image into as many as three distinct additive components. These components are a background, a noise field, and possibly a signal profile. The background component, denoted $\mathbf{b}$, is presumed to be identical in both alternatives. In many cases the background component is simply a uniform luminance that boosts the image to the middle of the display range. However, our formulation is general enough to allow for a background that varies from trial to trial. The noise component is presumed to be independent, and hence a different vector, in each of the two alternatives. The noise field is therefore denoted $\mathbf{n}^+$ for the signal-present image and $\mathbf{n}^-$ for the signal-absent image to indicate this dependence. Finally, the signal profile is denoted $\mathbf{s}$. This profile is added only to the target image. The analysis in this work is confined to the signal-known-exactly paradigm, and hence the signal vector is fixed throughout all trials. Note that for contrast-discrimination experiments the contrast pedestal is incorporated into $\mathbf{b}$, and hence $\mathbf{s}$ is actually the difference signal. The 2AFC images can be written mathematically as

$$\begin{aligned}\mathbf{g}^+ &= \mathbf{b} + \mathbf{n}^+ + \mathbf{s} \\ \mathbf{g}^- &= \mathbf{b} + \mathbf{n}^- \quad .\end{aligned} \tag{1}$$

As stated in the ”Introduction,” we assume that the noise in each image is a realization of a Gaussian random process, often referred to in statistical texts as multivariate normal. Hence a Gaussian probability density function (pdf) describes both noise fields. Furthermore, we assume that the noise process is zero-mean because any mean effect can be attributed to the background vector $\mathbf{b}$. However, we allow for a general noise covariance matrix, $\mathbf{K_n}$, requiring only that this matrix be known and nonsingular. The covariance matrix governs the noise-correlation structure in each image. If white noise is used, then $\mathbf{K_n} = \sigma^2\mathbf{I}$ where σ^2 is the pixel variance and $\mathbf{I}$ is the identity matrix. The pdf of the noise vectors is given by (Mardia, Kent, & Bibby, 1979)

$$p(\mathbf{n}) = \frac{1}{\sqrt{(2\pi)^N |\mathbf{K_n}|}} \exp\left(-\frac{1}{2}\mathbf{n^t K_n^{-1} n}\right).$$

We write $\mathbf{n} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{K_n})$ to indicate that $\mathbf{n}$ is distributed according to this pdf.

Internal Response Variables

It is common to assume that an observer responds individually to each alternative in a forced-choice experiment. This assumption leads to the well-known interpretation of the proportion of correct responses as the area under a receiver-operating characteristic (ROC) curve (Green & Swets, 1966). We can model an observer performing a 2AFC task as formulating scalar responses to each image stimuli in an experimental trial, and then choosing the alternative with the maximal response. The formation of a response variable from an image can be described mathematically by a scalar-valued function of the image vector, $\lambda = w(\mathbf{g})$. The responses to the signal-present and signal-absent images are defined by $\lambda^+ = w(\mathbf{g}^+)$ and $\lambda^- = w(\mathbf{g}^-)$, respectively. Human observers will often give different decisions from the same set of images in repeated trials, a characteristic of internal noise in the observer (Pelli, 1981; Burgess & Colborne, 1988). Internal noise is incorporated into the internal response by allowing random components in $w(\mathbf{g})$.

A linear observer can be defined by a response function that is a linear function of the image intensity. When the observer is subject to internal noise, the linear relationship becomes probabilistic. A convenient way to introduce internal noise is simply to add a random variable to the output of the linear operation. The resulting signal-present and signal-absent internal responses are defined as

$$\begin{aligned}\lambda^+ &= \mathbf{w}^t\mathbf{g}^+ + \varepsilon^+ \\ \lambda^- &= \mathbf{w}^t\mathbf{g}^- + \varepsilon^- \quad ,\end{aligned} \tag{2}$$

where the vector $\mathbf{w}$ is the set of weights used to create the response variable. As such, $\mathbf{w}$ (often called an observer template or filter) represents the summation strategy used by the observer to perform the task. The ε^+ and ε^- terms on the right-hand side of Equation 2 are scalar internal noise components. These components are presumed to be independent, zero-mean Gaussian random variables. We will specify the variance of ε^+ and ε^- to be σ_ε^2. The value of σ_ε^2 is not presumed to be known nor is it necessary for computing a classification image. Even though the internal noise term is specified as a scalar random variable, it is general enough to include noise from multiple independent sources. If we adopt the approach of equating internal noise in the observer with an equivalent noise source in the stimulus (Ahumada, 1987), then the internal noise component is defined by $\varepsilon = \mathbf{w}^t\mathbf{n}_{\mathrm{Eqv}}$, where $\mathbf{n}_{\mathrm{Eqv}}$ is a vector of equivalent noise in the stimulus domain. In this case, the variance of the internal noise component is given by $\sigma_\varepsilon^2 = \mathbf{w}^t\mathbf{K}_{\mathrm{Eqv}}\mathbf{w}$, where $\mathbf{K}_{\mathrm{Eqv}}$ is the covariance matrix associated with the equivalent noise.

Forced-Choice Decisions

To make a decision in a given experimental trial, an observer indicates the image believed most likely to contain the signal profile. If the response to the signal-present image is larger than that of the signal-absent image, then a correct decision is made, and if not, an incorrect decision is made. Let us define the observer score (or trial outcome), o, for a given trial as one if the observer correctly identifies the signal-present image and zero if the observer makes an incorrect choice. The score is defined in terms of the internal responses by

$$o = \mathrm{step}\left(\lambda^{+} - \lambda^{-}\right) \tag{3}$$

where the step function is defined as one for arguments greater than zero and zero for arguments less than zero. We will assume continuous distributions for the internal responses, and hence the probability of a tie ($\lambda^{+} = \lambda^{-}$) can be neglected. In terms of the linear response model given in Equation 2, and the image generating equations in Equation 1, the trial score is defined as

$$\begin{aligned} o &= \mathrm{step}\left(\mathbf{w}^{t}\mathbf{g}^{+} + \varepsilon^{+} - \mathbf{w}^{t}\mathbf{g}^{-} - \varepsilon^{-}\right) \\ &= \mathrm{step}\left(\mathbf{w}^{t}\mathbf{s} + \mathbf{w}^{t}\left(\mathbf{n}^{+} - \mathbf{n}^{-}\right) + \left(\varepsilon^{+} + \varepsilon^{-}\right)\right) \\ &= \mathrm{step}\left(\mathbf{w}^{t}\mathbf{s} + \mathbf{w}^{t}\Delta\mathbf{n} + \Delta\varepsilon\right) \end{aligned} \tag{4}$$

where $\Delta\mathbf{n} = \mathbf{n}^{+} - \mathbf{n}^{-}$ is the vector difference between the noise fields, and $\Delta\varepsilon = \varepsilon^{+} - \varepsilon^{-}$ is the difference between internal noise components. Given the Gaussian assumptions we have made on $\mathbf{n}^{+}$ and $\mathbf{n}^{-}$, the difference is $\Delta\mathbf{n} \sim \mathrm{MVN}\left(\mathbf{0}, 2\mathbf{K}_{\mathbf{n}}\right)$. For independent Gaussian-distributed internal-noise components, $\Delta\varepsilon \sim \mathrm{N}\left(0, 2\sigma_{\varepsilon}^{2}\right)$.

Note that in the second step of Equation 4, the background component, $\mathbf{b}$, cancels out of the expression. Hence the mean background does not directly influence the trial score in the linear model. However, this does not imply that the background is irrelevant because the observer may accommodate the background indirectly by modifying the template, or the background may influence the magnitude of the internal noise.

Figures of Merit for Task Performance

The basic measure of performance in a forced-choice experiment is the proportion of correct responses, denoted P_C (Green & Swets, 1966). The proportion correct is equivalent to the ensemble mean score,

$$P_C = \langle o \rangle \tag{5}$$

where the angled brackets, $\langle \cdots \rangle$, indicate a mathematical expectation of the enclosed quantity. In this case, expectation is taken with respect to random variability in the images as well as random variability due to observer internal noise. Equation 5 forms the basis for analysis of forced-choice data with human observers. With human observers, the internal response variables are not observable. But the score in each trial of the experiment can be observed, allowing the proportion correct to be estimated as the observed proportion of correct responses,

$$\hat{P}_C = \frac{1}{N_T}\sum_{j=1}^{N_T} o_j \quad , \tag{6}$$

where o_j is the score in the jth trial, and N_T is the total number of trials in the experiment. As a sample average, it is well known that $\hat{P}_C$ is an unbiased estimate of the ensemble mean in Equation 5 (Dudewicz & Mishra, 1988).

A second measure of performance, the detectability index d', is defined from the mean and variance of the internal response variables under the assumption of common variance, $\mathrm{Var}\left(\lambda^{+}\right) = \mathrm{Var}\left(\lambda^{+}\right) = \sigma_{\lambda}^{2}$. The detectability index is defined as

$$d' = \frac{\left\langle \lambda^{+} \right\rangle - \left\langle \lambda^{-} \right\rangle}{\sigma_{\lambda}} \quad . \tag{7}$$

Under the assumption of independent Gaussian-distributed responses, d' and P_C are directly related to one another by $P_C = \Phi\left(d' / \sqrt{2}\right)$ (Green & Swets, 1966), where $\Phi(\cdots)$ is the standard cumulative normal distribution function defined as

$$\Phi(x) = \int_{-\infty}^{x} dz \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^{2}\right) \quad .$$

For Gaussian distributed images defined in Equation 1, and the observer responses defined in Equation 2, the detectability index can be written

$$d' = \frac{\mathbf{w}^{t}\mathbf{s}}{\sqrt{\mathbf{w}^{t}\mathbf{K}_{\mathbf{n}}\mathbf{w} + \sigma_{\varepsilon}^{2}}} \quad . \tag{8}$$

Analysis of Classification Images

We now turn to estimating and performing statistical inference on classification images.

Estimation Procedure

As a way of motivating the estimation procedure, let us presume the linear observer model of Equation 2. Now consider a trial in which the noise-field difference ($\Delta\mathbf{n}$) happens to look like the observer template ($\mathbf{w}$). Looking at Equation 4, we would then expect $\mathbf{w}^t\Delta\mathbf{n}$ to take on a large positive value, leading to a high probability of a correct response. We might then imagine that when the observer gets the trial correct, it is because the noise-field difference in the trial (on average) looks something like the observer's template. Conversely, if the noise-field difference looks like the negative of the observer template, then we would expect $\mathbf{w}^t\Delta\mathbf{n}$ to take on a large negative value, leading to a high probability of an incorrect response. We might then surmise that when the observer gets the trial incorrect, it is because the noise-field difference in the trial tends to look something like the negative of the observer's template. In this case, the negative of $\Delta\mathbf{n}$ would tend to look like $\mathbf{w}$. This heuristic suggests weighting $\Delta\mathbf{n}$ by a positive value when the observer gets the trial correct and a negative value when the observer gets the trial incorrect, and then averaging the results.

Let us now take a more quantitative view of this weighting scheme. Consider a weight defined as $o_j - a$, where a is some constant between zero and one. When the observer makes a correct decision ($o_j = 1$), the weight assumes a positive value, and when the observer makes an incorrect decision ($o_j = 0$), the weight assumes a negative value as alluded to above. In previous works (Abbey et al., 1999; Abbey & Eckstein, 2000, 2001a), we have used a weighting scheme in which $a = 1/2$. However, it can be shown that letting $a = P_C$ minimizes the covariance matrix of the estimated classification images; in particular, it minimizes the variance of each element in the classification image. Because we do not have access to the ensemble proportion correct, we propose setting the constant to $a = \hat{P}_C$, the estimated proportion correct defined in Equation 6. Using this weighting scheme, we can define a score weighted difference in noise fields as

$$\Delta\mathbf{q}_j = \frac{N_T}{N_T - 1}\left(o_j - \hat{P}_C\right)\mathbf{K}_\mathbf{n}^{-1}\Delta\mathbf{n}_j . \qquad (9)$$

The $N_T/(N_T - 1)$ factor will be seen below (Equation 11) to be convenient for removing dependencies on the number of experimental trials from the expected value of $\Delta\mathbf{q}_j$. This factor is negligibly different from one in most cases because the number of trials is typically quite large in classification-image experiments.

One disadvantage of the weighting used in Equation 9 is that because $\hat{P}_C$ is defined over all the experimental trials, it introduces the possibility of trial-to-trial correlations among the vectors $\Delta\mathbf{q}_j$. However, the magnitude of these correlations can be shown to be of order $1/N_T^2$. Typically, more that 1,000 trials are used in a classification image experiment, and hence sequential correlations can be neglected for practical purposes.

The $\mathbf{K}_\mathbf{n}^{-1}$ term in Equation 9 accommodates pixel-to-pixel noise correlations. However, in the case of white noise where $\mathbf{K}_\mathbf{n} = \sigma_\mathbf{n}^2\mathbf{I}$ ($\mathbf{I}$ is the identity matrix and $\sigma_\mathbf{n}^2$ is the pixel variance), the formula simplifies to

$$\Delta\mathbf{q}_j = \frac{N_T}{\sigma_\mathbf{n}^2(N_T - 1)}\left(o_j - \hat{P}_C\right)\Delta\mathbf{n}_j . \qquad (10)$$

Expectation of $\Delta\mathbf{q}$

The rationale for defining $\Delta\mathbf{q}$ becomes clearer when we assume the linear observer model of Equation 4 and compute the expectation of Equation 9. We denote this expectation by $\langle\Delta\mathbf{q}\rangle_{\Delta\mathbf{n},\Delta\varepsilon}$, where the subscripts emphasize that the expectation encompasses both the external-noise variability in $\Delta\mathbf{n}$ and internal-noise variability in $\Delta\varepsilon$. We will not derive the expectation here because the derivation is lengthy and has been published previously in a simpler form (Abbey & Eckstein., 2001b). We will simply state the value of the expectation as

$$\langle\Delta\mathbf{q}\rangle_{\Delta\mathbf{n},\Delta\varepsilon} = \frac{\exp\left(-(\mathrm{d}'/2)^2\right)}{\sqrt{\pi\left(\mathbf{w}\mathbf{K}_\mathbf{n}\mathbf{w} + \sigma_\varepsilon^2\right)}}\mathbf{w} \quad , \qquad (11)$$

where d' is the detectability index of Equation 8. The expected value is equivalent to the observer template up to a positive scalar factor. Because the magnitude of the observer template is somewhat arbitrary (scaling the template and the internal noise component yields an equivalent detection strategy), obtaining the observer template with a normalized magnitude is an acceptable stand-in for $\mathbf{w}$. More importantly, we see below that working with a normalized version of the observer template does not hinder our ability to perform statistical inference.

Estimation Formula

A simple approach to estimating the classification image is to replace the mathematical expectation in Equation 11 with a sample average. Let $j = 1, \cdots, N_T$ where N_T is the number of trials. The classification image estimate is then

$$\begin{aligned} \Delta\bar{\mathbf{q}} &= \frac{1}{N_T}\sum_{j=1}^{N_t} \Delta\mathbf{q}_j \\ &= \frac{1}{N_T - 1}\sum_{j=1}^{N_t} \left(o_j - \hat{P}_C\right)\mathbf{K}_{\mathbf{n}}^{-1}\Delta\mathbf{n}_j . \end{aligned} \qquad (12)$$

Sample averages have a number of beneficial properties as estimators of a mean value including unbiasedness, minimum variance, and asymptotic normality (Dudewicz & Mishra, 1988).

Constraints on the classification image

Although the estimation procedure in Equation 12 works well for estimating the entire classification image, it is often desirable to restrict attention to regions of the classification image and to employ averaging across elements of the classification image as a way to reduce measurement noise that arises from a finite number of trials. For example, radial symmetry can be used to justify radial averaging of the classification images to reduce the effects of noise (Abbey et al., 1999). These constraints can be particularly valuable for conducting statistical hypothesis testing because they reduce degrees of freedom and hence lead to more powerful tests.

Both averaging and subregion extraction can be implemented as linear functions of the classification image. Let us consider a general linear function of the form

$$\Delta\mathbf{y}_j = \mathbf{R}\Delta\mathbf{q}_j \quad . \qquad (13)$$

The matrix $\mathbf{R}$ can be thought of as reducing the classification image to a set of linear features (specific pixels, spatial averages, etc.) of interest, and hence $\mathbf{R}$ is an $N_Y \times N$ matrix where N is the number of pixels in the stimulus and N_Y is the number of features in $\Delta\mathbf{y}$. Although it will generally be the case that N_Y will be much smaller than N, it is still possible to consider the case where $\mathbf{R}$ is the identity matrix. In this case, $\Delta\mathbf{y}_j = \Delta\mathbf{q}_j$ and $N_Y = N$.

To estimate the constrained classification image, we can use the sample mean of the $\Delta\mathbf{y}$ vectors,

$$\Delta\bar{\mathbf{y}} = \frac{1}{N_T}\sum_{j=1}^{N_T} \Delta\mathbf{y}_j \quad . \qquad (14)$$

We can also compute a sample error covariance matrix for the $\Delta\mathbf{y}$ vectors as

$$\mathbf{S}_{\Delta\mathbf{y}} = \frac{1}{N_T - 1}\sum_{j=1}^{N_T} \left(\Delta\mathbf{y}_j - \Delta\bar{\mathbf{y}}\right)\left(\Delta\mathbf{y}_j - \Delta\bar{\mathbf{y}}\right)^t . \qquad (15)$$

These sample quantities form the basis for most of the hypothesis testing below.

Statistical Inference on Classification Images

We can perform statistical hypothesis testing on classification image data using sample statistics derived from Equations 14 and 15 above. Because of the generally large number of trials needed to get a good estimate of the classification image, asymptotic results can be used to justify a number of Hotelling's T^2 tests.

Hotelling's T^2 distribution is closely tied to the more commonly found F distribution, and this relation is useful for obtaining significance levels from tables. If T^2 has a Hotelling's $T^2_{P,M}$ distribution where P and M are the two degrees of freedom associated with the distribution, then $T^2 \times (M - P + 1)/MP$ has an $F_{P,M-P+1}$ distribution. Hence we can take any of the Hotelling's T^2 tests derived below, multiply the test statistic by $(M - P + 1)/MP$, and then look up critical values or *p* values for the test from published tables (e.g., Mardia et al., 1979). Many programming environments supply procedures to compute these values as well.

Difference from a known profile

If we wish to test the hypothesis that the mean value of $\Delta\mathbf{y}$ is different from some fixed vector, $\Delta\mathbf{y}_0$, we can use Hotelling's one-sample T^2 statistic,

$$T^2 = N_T\left(\Delta\bar{\mathbf{y}} - \Delta\mathbf{y}_0\right)^t \mathbf{S}_{\Delta\mathbf{y}}^{-1}\left(\Delta\bar{\mathbf{y}} - \Delta\mathbf{y}_0\right) \quad . \qquad (16)$$

Under the null hypothesis of $\langle\Delta\mathbf{y}\rangle = \Delta\mathbf{y}_0$, T^2 has a Hotelling's $T^2_{N_Y, N_T - 1}$ distribution.

In order for Hotelling's T^2 distribution to be defined, we must have that $N_T > N_Y$. This is equivalent to requiring that the sample covariance be of full rank. Here we see the advantage of working with a reduced set of classification image features. For the full classification image, N_Y is equal to the number of pixels in each image stimulus. In the case of 64 by 64 pixel images we have 4,096 free parameters that require at least 4,097 trials in order to perform the statistical test.

Difference between two classification images: independent image sets

In some cases we may wish to test for differences between two classification images derived from independent sets of images. For example, we may have classification images for an observer in two different tasks. Here we can use a Hotelling's two-sample test for differences.

Let the first data set have N_{T1} trials and the second have N_{T2} trials. We will denote the sample means and covariance matrices of the two classification images by $\Delta\bar{\mathbf{y}}_1$, $\mathbf{S}_{\Delta\mathbf{y},1}$, $\Delta\bar{\mathbf{y}}_2$, and $\mathbf{S}_{\Delta\mathbf{y},2}$, respectively. For testing the null hypothesis of a common mean, we can use Hotelling's two-sample test statistic,

$$T^2 = \frac{N_{T1}N_{T2}}{N_{T1}+N_{T2}}\left(\Delta\bar{\mathbf{y}}_1 - \Delta\bar{\mathbf{y}}_2\right)^t \mathbf{S}_{1,2}^{-1}\left(\Delta\bar{\mathbf{y}}_1 - \Delta\bar{\mathbf{y}}_2\right), \qquad (17)$$

where

$$\mathbf{S}_{1,2} = \frac{1}{\left(N_{T1}+N_{T2}-2\right)}\left(\left(N_{T1}-1\right)\mathbf{S}_{\Delta\mathbf{y},1} + \left(N_{T2}-1\right)\mathbf{S}_{\Delta\mathbf{y},2}\right).$$

Under the null hypothesis of equal means and equal covariance matrices for the two $\Delta\mathbf{y}$ samples, T^2 has a Hotelling's $T^2_{N_Y,N_{T1}+N_{T2}-2}$ distribution.

Difference between two classification images: common image sets

When two classification images are derived from the same set of images (e.g., for examining the strategy of two different observers on a given image set or a repeated study with the same observer at two different times), the Hotelling's two-sample approach above can be used, but it is overly conservative. In this case, a more efficient test is to look for a significant difference between the individual trial $\Delta\mathbf{y}$ vectors. Let us define

$$\Delta^2\mathbf{y}_j = \Delta\mathbf{y}_{1,j} - \Delta\mathbf{y}_{2,j} \quad , \qquad (18)$$

where $\Delta\mathbf{y}_{1,j}$ and $\Delta\mathbf{y}_{2,j}$ are the individual trial $\Delta\mathbf{y}$ vectors for the first and second observer. The test for differences between the two observers is now defined as a one-sample test for a significant departure from zero in $\Delta^2\mathbf{y}_j$. In this case, the test statistic is defined as

$$T^2 = N_T\left(\Delta^2\bar{\mathbf{y}}\right)^t \mathbf{S}_{\Delta^2\mathbf{y}}^{-1}\left(\Delta^2\bar{\mathbf{y}}\right) \quad , \qquad (19)$$

where $\Delta^2\bar{\mathbf{y}}$ is the sample mean of the $\Delta^2\mathbf{y}_j$ vectors,

$$\begin{aligned}\Delta^2\bar{\mathbf{y}} &= \frac{1}{N_T}\sum_{j=1}^{N_T}\Delta^2\mathbf{y}_j \\ &= \Delta\bar{\mathbf{y}}_1 - \Delta\bar{\mathbf{y}}_2 ,\end{aligned}$$

and $\mathbf{S}_{\Delta^2\mathbf{y}}$ is the sample covariance matrix,

$$\mathbf{S}_{\Delta^2\mathbf{y}} = \frac{1}{N_T-1}\sum_{j=1}^{N_T}\left(\Delta^2\mathbf{y}_j - \Delta^2\bar{\mathbf{y}}\right)\left(\Delta^2\mathbf{y}_j - \Delta^2\bar{\mathbf{y}}\right)^t .$$

Under the null hypothesis of $\left\langle\Delta^2\mathbf{y}_j\right\rangle = \mathbf{0}$, T^2 has a Hotelling's $T^2_{N_Y,N_T-1}$ distribution.

Test of a nonlinear observer response function

Barth et al. (1999) have looked at classification images from signal-present images versus signal-absent images as a way to reveal nonlinear effects in the observer response function, such as spatial uncertainty. They used a yes-no task, but the same sort of analysis can be generalized to forced-choice data.

The test for nonlinearity we propose requires breaking $\Delta\mathbf{q}_j$ into two components arising from the signal-present noise field ($\mathbf{n}_j^+$) and the signal-absent noise field ($\mathbf{n}_j^-$). Let us define

$$\begin{aligned}\mathbf{q}_j^+ &= \frac{N_T}{N_T-1}\left(o_j - \hat{P}_C\right)\mathbf{K}_\mathbf{n}^{-1}\mathbf{n}_j^+ \\ \mathbf{q}_j^- &= \frac{N_T}{N_T-1}\left(o_j - \hat{P}_C\right)\mathbf{K}_\mathbf{n}^{-1}\mathbf{n}_j^- \quad ,\end{aligned} \qquad (20)$$

from which it can be seen that $\Delta\mathbf{q}_j = \mathbf{q}_j^+ - \mathbf{q}_j^-$. Under the linear observer response function of Equation 2, the mathematical expectations of $\mathbf{q}_j^+$ and $\mathbf{q}_j^-$ are given by

$$\left\langle\mathbf{q}_j^+\right\rangle = -\left\langle\mathbf{q}_j^-\right\rangle = \frac{\exp\left(-(\mathrm{d}'/2)^2\right)}{2\sqrt{\pi\left(\mathbf{w}\mathbf{K}_\mathbf{n}\mathbf{w} + \sigma_\varepsilon^2\right)}}\mathbf{w} \quad . \qquad (21)$$

Under the linear observer response model, the two components have the same mean except for a change in sign. However, this relationship does not generally hold for nonlinear observer response functions. As a result, we can check for a nonlinear response function by testing the null hypothesis that the means of $\mathbf{q}_j^+$ and $\mathbf{q}_j^-$ sum to zero.

Once again, the high dimensionality of the raw classification images can lead to difficulties with degrees of freedom. Hence it is generally preferable to work with linear functions of the two classification images, as defined in Equation 13. In this case, we define $\mathbf{y}_j^+ = \mathbf{R}\mathbf{q}_j^+$ and $\mathbf{y}_j^- = \mathbf{R}\mathbf{q}_j^-$, and test the null hypothesis that $\left\langle\mathbf{y}_j^+\right\rangle + \left\langle\mathbf{y}_j^-\right\rangle = \mathbf{0}$. Under the null hypothesis,

$$T^2 = N_T\left(\bar{\mathbf{y}}^+ + \bar{\mathbf{y}}^-\right)^t \mathbf{S}_\mathbf{y}^{-1}\left(\bar{\mathbf{y}}^+ + \bar{\mathbf{y}}^-\right), \qquad (22)$$

where $\overline{\mathbf{y}}^{+}$ and $\overline{\mathbf{y}}^{-}$ are the sample means of the $\mathbf{y}_j^{+}$ vectors and the $\mathbf{y}_j^{-}$ vectors, respectively, and the sample covariance matrix is defined as

$$\mathbf{S}_{\mathbf{y}} = \frac{1}{N_T - 1} \sum_{j=1}^{N_T} \left(\mathbf{y}_j^{+} - \overline{\mathbf{y}}^{+} + \mathbf{y}_j^{-} - \overline{\mathbf{y}}^{-}\right)\left(\mathbf{y}_j^{+} - \overline{\mathbf{y}}^{+} + \mathbf{y}_j^{-} - \overline{\mathbf{y}}^{-}\right)^{t}.$$

Under the null hypothesis of a linear observer response function, T^2 has a Hotelling's $T^2_{N_Y, N_T - 1}$ distribution.

It is important to exercise some caution in interpreting the results of this test. As is the case with hypothesis testing in general, we can only reject the null hypothesis that the observer has adopted a linear strategy; we cannot accept it. Furthermore, it is possible that there are nonlinear observer response functions that are not revealed by this test. Nonetheless, we believe that the test is still valuable, despite this limitation. Although it cannot be used to verify that linearity assumptions have been met, if the test rejects, we have a high degree of confidence that the linearity assumptions have not been met.

Case Study: Detection of a Gaussian Bump in White Noise

We now turn to a specific example that shows how the methods described above can be used to analyze classification images from 2AFC data. We considered the detection of a two-dimensional spatial Gaussian profile embedded in white image noise. Figure 1 shows the signal (target) profile as well as example signal-present and signal-absent images with image noise.

Description of Experiment

The procedure for this experiment has been described previously (Abbey & Eckstein, 2000), and hence we will review it briefly. The width of the Gaussian bump target was set by specifying a spatial standard deviation of 3.0 pixels. Each pixel was approximately 0.3 mm on the monitor screen, and observers maintained a viewing distance of approximately 1 m. At these dimensions, the full width at half max of the signal occupied 0.12 degrees (7.2 minutes) of visual angle.

Experiments were conducted on a high-quality monochrome monitor (model M15LMAX; Image Systems Corp., Minnetonka, MN), calibrated to a linear luminance scale in a darkened room. The signal contrast for the experiment was determined by pilot experiments, and set to 6.2% against a mean background luminance of 31.3 cd/m^2. This signal contrast was determined from psychometric function data to give an average human observer performance of approximately 85% correct. The noise contrast (measured as the luminance standard deviation divided by the mean luminance) was fixed at 15%.

The two stimulus alternatives were presented sequentially with a presentation time for each image of 500 ms and a white-noise mask that was displayed for 1,000 ms between them to disrupt any persistence effects. We will consider results from two observers (subjects D.V. and C.H.) who were naïve to the goals of the research but had extensive experience in visual tasks of the sort reported here. The observers participated in a number of training sets before beginning the experiment reported here. A total of 2,000 trials were used in this experiment, and each observer completed all trials.

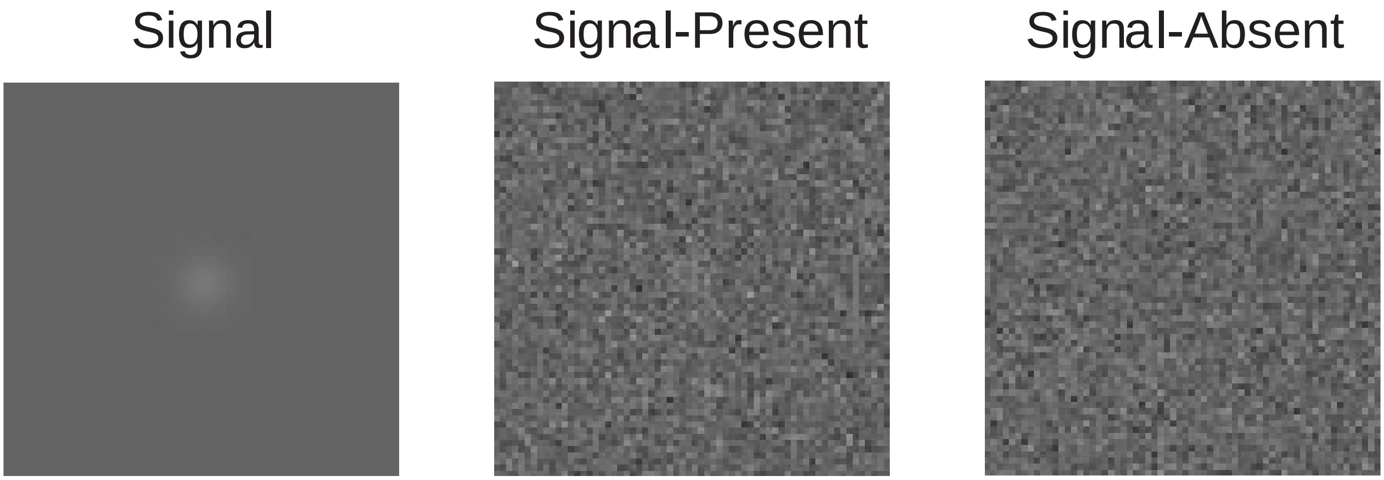

Figure 1. Target profile and sample images. The target contrast is somewhat elevated in these images for clarity of display

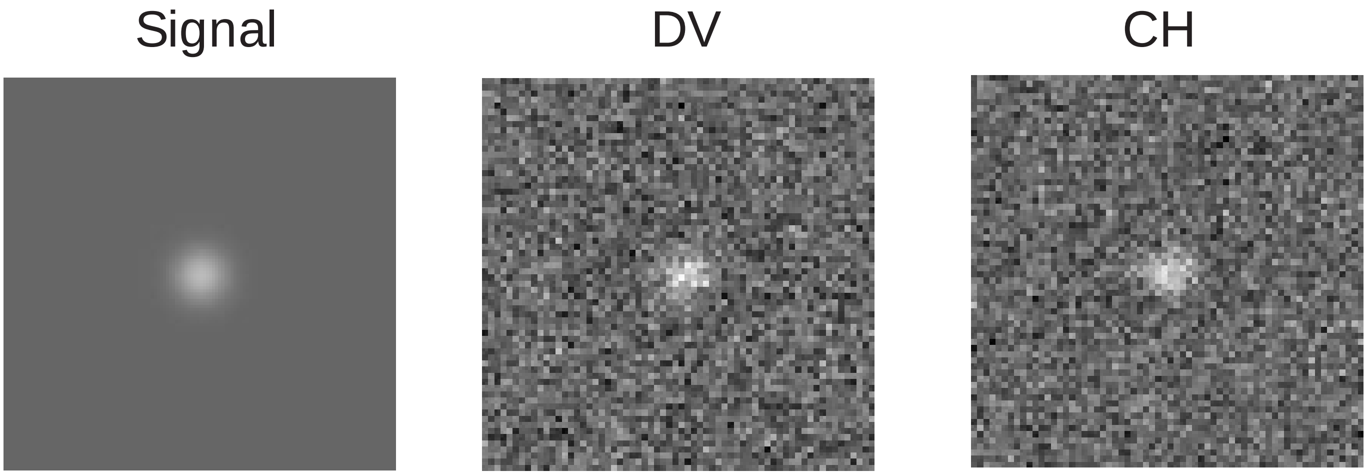

Figure 2. Signal profile and classification images for both subjects.

Classification Images

Classification images estimated using Equation 12 for the two observers are shown in Figure 2. The estimates are clearly somewhat noisy, but nevertheless an area of activation can be seen at the target location. There appears to be a mild inhibitory region surrounding this central area. These effects are more clearly seen in radial average plots of the classification images.

Radial averages

Figure 3 shows plots of the two classification images averaged over pixels of equal radius from the center. The apparent radial symmetry of the classification images in Figure 2 indicated that this could be a good way to reduce the degrees of freedom in the data without losing important features. Radial averaging is a linear operation and hence fits the general form of Equation 13. The radial profile of the signal is plotted as well for reference. Because the image noise is uncorrelated and Gaussian, the signal profile is also the profile of the ideal observer. The magnitude of the signal profile is set using the relation in Equation 11 and choosing the internal noise variance so that performance matches each subject.

The radial average data are plotted with error bars consisting of +/−1 SE in each radial bin. The error bars are largest near the origin because the radial bins accumulate fewer pixels there. The data appear to agree reasonably well with the signal profile. However, from approximately 0.1 to 0.2 degrees from the signal center, the classification image profiles dip down slightly below the signal profile, indicative of the inhibitory surround alluded to above.

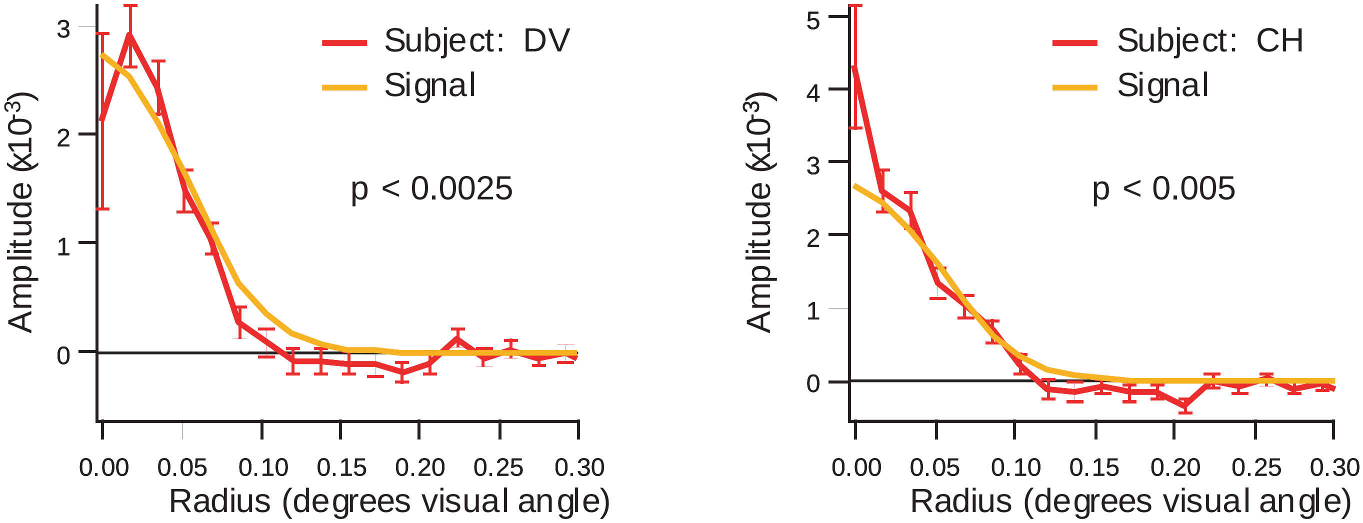

Figure 3. Radial averages of the classification images for both subjects. Human observer data is plotted with error bars of ± 1 standard error. The signal profile is plotted as well. The *p* values for significant departures from the signal profile are given on each plot. Differences between the two observer profiles were not significant (*p* > .36).

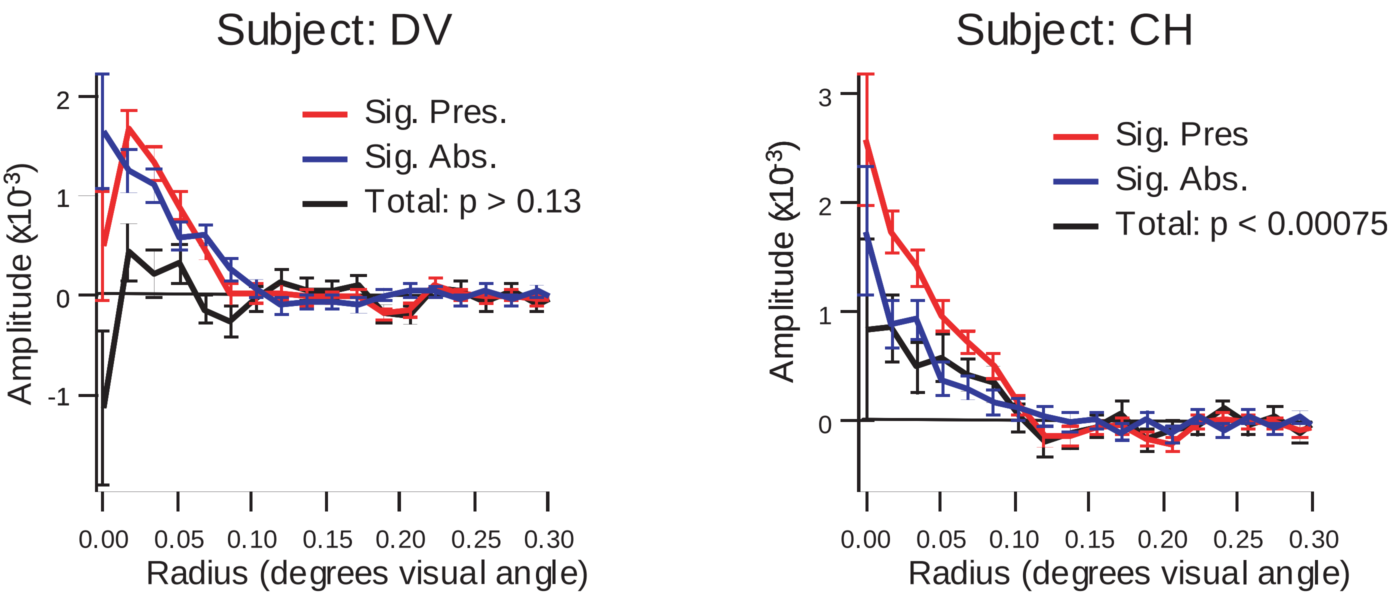

Figure 4. Radial average plots of the signal-present and signal-absent classification images with error bars consisting of ±1 standard error. The negative of the signal-absent plot is used here in order to highlight its differences with the signal-present plot. The sum of the two radial averages (the difference between the signal-present and signal-absent functions in this figure) is plotted in black. The test of a nonlinear observer response function consists of testing this plot for significant departures from zero.

Statistical Inference

Hypothesis testing on the radial averages reveals significant departures from the signal profile in the human-observer data. We tested for significance on the radial bins from 0 to 0.3 degrees of visual angle from the signal center using Hotelling's one-sample test defined in Equation 16. There was a total of 18 data points in this angular range ($N_y = 18$), and all 2,000 trials were used to compute the test statistics ($N_T = 2000$). The test is significant at the 1% level for both observers (D.V., $p <$ .0025; C.H., $p <$ .005) even when Bonferroni-corrected for multiple comparisons (Altman, 1999) across the two observers; therefore, we can conclude that the classification images of both observers depart significantly from the signal profile.

We tested for significant differences between the two observers via the test defined in Equation 19. In this case, the test was not significant ($p >$ .36). It should be noted that because both observer templates are subject to estimation error, the resulting hypothesis test is generally less powerful than a test of one observer against a known classification-image profile. It seems reasonable to suppose that at some point, if we collected enough trials, we would find observer differences. Nonetheless, the fact that the templates are not significantly different after 2,000 trials does imply some degree of consistency between the two subjects.

We also tested for nonlinear observer response functions in both observers using Equation 22. Figure 4 contains plots of the radial averages of the signal-present classification image ($\bar{\mathbf{q}}^+$), the negative of the signal-absent classification image ($-\bar{\mathbf{q}}^-$), and the sum of the two ($\bar{\mathbf{q}}^+ + \bar{\mathbf{q}}^-$) with standard errors. The negative of the signal-absent classification image is plotted to better visualize the differences between it and the signal-present classification image. In this test, subject D.V. showed no significant difference between the two ($\bar{\mathbf{q}}^+ + \bar{\mathbf{q}}^-$ not significantly different from $\mathbf{0}$, $p > .13$), whereas subject C.H. did show a significant effect ($p <$ .00075). It is possible that a significant effect for subject D.V. would have been found had a more restrictive range of visual angle been used.

Discussion and Summary

Nonlinear Observer Response Functions

Even when observers use a nonlinear strategy, classification images may still be illuminating and worth obtaining. An excellent example of this is the work of Gold et al. (2000), who used classification images to examine illusory contours in Kaniza squares. The classification images they observed extended out along the illusory contours, even though these regions had no useful information for performing the task. With the broad spatial extent of the signal used in these experiments, it is doubtful that human observers adopt a linear strategy to perform the task. Nonetheless, the classification images observed by Gold et al. (2000) show that human observers make heavy use of the illusory contours to perform the task.

The apparent nonlinearity found for subject C.H. in Figure 4 is of interest for understanding how this observer is performing the detection task. We can imagine looking at specific nonlinear effects, such as intrinsic spatial uncertainty or nonlinear signal transduction, to see if they account for the divergence from linearity in this observer.

Other Approaches to Analyzing Classification Images

Maximum likelihood approaches

For estimating classification images, one alternative to Equation 12 is a more standard categorical regression approach (McCullagh & Nelder, 1989; Abbey & Eckstein., 2001a). If we assume the linear observer model for Equation 4, then the observer score in a given trial can be modeled as a binomial random variable,

$$o_j \sim B\left(\Phi\left(\frac{\mathbf{w}^t\left(\mathbf{s}+\Delta\mathbf{n}_j\right)}{\sqrt{2}\sigma_\varepsilon}\right),1\right) \quad . \qquad (23)$$

where $B(p,N)$ indicates the binomial probability function

$$\Pr\left(o_j\right)=\frac{N!}{o_j!\left(N-o_j\right)}p^{o_j}\left(1-p\right)^{N-o_j} \quad .$$

Note that this model can easily accommodate data that consists of multiple passes through the same set of images by letting N be greater than one.

The functional form of p in Equation 23, often referred to as the link function, is based on the assumption of independent Gaussian distributions for each internal noise component. From the binomial distribution, it is possible to derive the likelihood of the observer scores given a specific choice of the observer template $\mathbf{w}$. The maximum-likelihood (ML) estimate of the classification image is then found by optimizing the likelihood function.

ML estimates have a number of attractive properties, including asymptotic efficiency. However, there are a number of issues generally having to do with model assumptions that need to be resolved before the approach can be applied reliably to observer data. One problem occurs if there are more free parameters in $\mathbf{w}$ than there are observed trials. In this case, there will not be a unique maximum of the likelihood function and hence no unique ML estimate. This problem can be reduced by using some sort of regularizing function (Abbey & Eckstein, 2001a), but it is not clear at this stage how the choice of a regularizer will influence the resulting estimates.

A second issue is the dependence on the assumption of Gaussian distributed internal noise. It is not clear what the effect on the estimate is if internal noise does not follow this distribution.

Analytic approximations to the covariance matrix

Recently, Abbey and Eckstein (2001b) proposed an analytic approximation to the covariance matrix of the estimated classification image. Such an approximation could in principle be used in place of the sample covariance matrices for hypothesis testing. Tests based on analytic (known) covariance matrices use the chi-square distribution instead of Hotelling's T^2 and generally have more statistical power. The analytic approximation was derived for a somewhat different (and less efficient) estimate of the classification image. It remains to be seen if the approximation will still be good for the estimate defined in Equation 12.

Summary

The main purpose of this work has been to provide a rigorous framework analyzing classification images derived from the 2AFC experimental paradigm. The methodology we describe includes procedures for estimating classification images and testing hypotheses on the resulting estimates. These procedures can be used to make inferences about how observers perform basic visual tasks. The estimation procedure we propose here differs somewhat from what has been described earlier. The principle difference is that, in this work, incorrect trials are given more weight than correct trials. This can be shown to result in a more precise estimate of the classification image. This revised estimation procedure and the statistical inference we present provide a more efficient and complete methodology for analyzing classification images in 2AFC experiments in the presence of correlated noise.

The hypothesis tests derived in this work consist of testing for significant differences with a known mean (e.g., a classification image that is significantly different from zero), significant differences in intra- and inter-observer classification images, and a test of significance between signal-present and signal-absent estimates that serves as a test for nonlinearity in the observer response.

The tests yield a set of rigorously defined tools for evaluating the visual strategies employed by human observers in simple detection and discrimination tasks masked by Gaussian-distributed image noise.

Footnotes

[1] We take the definition of a 2AFC experiment (Green & Swets, 1966) as an experiment in which two stimuli are shown in a given trial, and the observer is asked to identify the stimulus that contained the target of interest. The term is sometimes used to describe experiments in which a single stimulus is shown, and the observer is asked to identify one of two target profiles as being present in the image (sometimes referred to as two-alternative forced-response experiments). However, these latter experiments are more closely related to "yes-no" tasks, and methods for estimating classification images for them fit directly into the methodology developed by Ahumada and coworkers (Ahumada & Lovell, 1971; Ahumada et al., 1975; Ahumada, 1996).

[2] We use the term noise limited to designate visual tasks in which independent trial-to-trial stimulus variability between the two alternatives limits observer performance. A noise-limited task yields a much higher level of performance if the external noise was removed from the stimuli. Alternatively, contrast-limited tasks result in imperfect performance in the absence of any external image noise. Additionally, background-limited tasks are limited by masking induced from variability in a background component that is common to the two alternatives (sometimes referred to as twin noise studies [Burgess & Colborne, 1988; Ahumada & Beard, 1997; Eckstein et al., 1997]).

Acknowledgments

This work was supported by a National Aeronautics and Space Administration grant (NASA NAG-1157) and a National Institutes of Health grant (NIH-HL 53455). Commercial Relationships: None.

References

Abbey, C. K., Eckstein, M. P., & Bochud, F. O. (1999). Estimation of human-observer templates for 2 alternative forced choice tasks. *Proceedings of SPIE, 3663*, 284-295.

Abbey, C. K., & Eckstein, M. P. (2000). Estimates of human-observer templates for simple detection tasks in correlated noise. *Proceedings of SPIE, 3981*, 70-77.

Abbey, C. K., & Eckstein, M. P. (2001a). Maximum-likelihood and maximum a-posteriori estimates of human-observer templates. *Proceedings of SPIE, 4324*, 114-122.

Abbey, C. K., & Eckstein, M. P. (2001b). Theory for estimating human-observer templates in two-alternative forced-choice experiments. In M. F. Insana and R. Leahy (Eds.), *Proceedings of the 17th International Conference on Information Processing in Medical Imaging*, (pp. 24-35). Berlin: Springer-Verlag.

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America*, *49*, 1751-1756.

Ahumada, A. J., Marken, R., & Sandusky, A. (1975). Time and frequency analyses of auditory signal detection. *Journal of the Acoustical Society of America*, *2*, 1133-1139.

Ahumada, A. J., & Watson, A. B. (1985). Equivalent-noise model for contrast detection and discrimination. *Journal of the Optical Society of America A*, 57, 385-390. [PubMed]

Ahumada, A. J. (1987). Putting the visual system noise back in the picture. *Journal of the Optical Society of America A*, *4*, 2372-2378. [PubMed]

Ahumada, A. J. (1996). Perceptual classification images from Vernier acuity masked by noise [Abstract]. *Perception*, *26*(Suppl. 18), 18.

Ahumada, A. J., & Beard, B. L. (1997). Image discrimination models: Detection in fixed and random noise. *Proceedings of SPIE, 3016*, 34-43.

Altman, D. G. (1999). *Practical statistics for medical research*. New York: Chapman and Hall/CRC.

Barrett, H. H. (1990). Objective assessment of image quality: Effects of quantum noise and object variability. *Journal of the Optical Society of America A*, *7*, 1266-1278. [PubMed]

Barrett, H. H., Yao, J., Rolland, J. P., & Myers, K. J. (1993). Model observers for assessment of image quality. *Proceedings of the National Academy of Sciences of the United States of America*, *90*, 9758-9765. [PubMed]

Barth, E., Beard, B. L., & Ahumada, A. J. (1999). Nonlinear features in Vernier acuity. *Proceedings of SPIE, 3644*, 88-96.

Beard, B. L., & Ahumada, A. J. (1998). Technique to extract relevant image features for visual tasks. *Proceedings of SPIE, 3299*, 79-85.

Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of human visual signal discrimination. *Science*, *214*, 93-94. [PubMed]

Burgess, A. E., & Ghandeharian, H. (1984a). Visual signal detection. I. Ability to use phase information. *Journal of the Optical Society of America A*, *1*, 900-905. [PubMed]

Burgess, A. E., & Ghandeharian, H. (1984b). Visual signal detection. II. Signal-location identification. *Journal of the Optical Society of America A*, *1*, 900-905. [PubMed]

Burgess, A. E., & Colborne, B. (1988). Visual signal detection. IV. Observer inconsistency. *Journal of the Optical Society of America A*, *5*, 617-627. [PubMed]

Cohn, T. E., Thibos, L. N., & Kleinstein, R. N. (1974). Detectability of a luminance increment. *Journal of the Optical Society of America A*, *64*, 1321-1327. [PubMed]

Dudewicz, E. J., & Mishra, S. N. (1988). *Modern mathematical statistics*. New York: Wiley.

Eckstein, M. P., Ahumada, A. J., & Watson, A. B. (1997). Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise. *Journal of the Optical Society of America A, 14*, 2406-2419. [PubMed]

Edwards, D. C., Kupinski, M. A., Nishikawa, R. M., & Metz, C. E. (2000). Estimation of linear observer templates in the presence of multi-peaked Gaussian noise through 2AFC experiments. *Proceedings of SPIE, 3981*, 86-96.

Foley, J. M., & Legge, G. E. (1981). Contrast detection and near-threshold discrimination in human vision. *Vision Research, 21*, 1041-1053. [PubMed]

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioral receptive fields for visually completed contours. *Current Biology, 10*, 663-666. [PubMed]

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

Knoblauch, K., Thomas, J. P., & D'Zmura, M. (1999). Feedback temporal frequency and stimulus classification [Abstract]. *Investigative Ophthalmology and Visual Science, 40*, 4171.

Legge, G. E., Kersten, D., & Burgess, A. E. (1987). Contrast discrimination in noise. *Journal of the Optical Society of America A*, *4*, 391-404. [PubMed]

Lu, Z. -L., & Dosher, B. A. (1999). Characterizing human perceptual inefficiencies with equivalent internal noise. *Journal of the Optical Society of America A*, *16*, 764-778. [PubMed]

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. San Diego: Academic.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York: Chapman and Hall/CRC.

Nachmias, J., & Kocher, E. C. (1970). Visual detection and discrimination of luminous increments. *Journal of the Optical Society of America A*, *60*, 382-389. [PubMed]

Pelli, D. G. (1981). *Effects of visual noise* (Doctoral dissertation, Cambridge University, Cambridge).

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A*, *2*, 1508-1530. [PubMed]

Pelli, D. G., & Farell, B. (1999). Why use noise? *Journal of the Optical Society of America A, 16*, 647-653. [PubMed]

Revesz, G., Kundel, H. L., & Graber, M. A. (1974). The influence of structured noise on the detection of radiologic abnormalities. *Investigative Radiology*, *9*, 479-486. [PubMed]

Rose, A. (1948). The sensitivity performance of the human eye on an absolute scale. *Journal of the Optical Society of America A, 38*, 196-208.

Solomon, J. A. (2000). A picture of orientation discrimination [Abstract]. *Investigative Ophthalmology and Visual Science, 41*(ARVO Suppl. 1), 4241.

Tanner, W. P. (1961). Psychological implications of psychophysical data. *Annals of the New York Academy of Science*, *89*, 752-765.

Watson, A. B. (1998). Multi-category classification: Template models and classification images [Abstract]. *Investigative Ophthalmology and Visual Science, 39*(ARVO Suppl. 4), S912