Frequency tuning of shape perception revealed by classification image analysis

John Wilder
Department of Computer Science, University of Toronto, Toronto, Canada

Ingo Fruend
Centre for Vision Research, York University, Toronto, Canada

James H. Elder
Centre for Vision Research, York University, Toronto, Canada

Classification image analysis is a powerful technique for elucidating linear detection and discrimination mechanisms, but it has primarily been applied to contrast detection. Here we report a novel classification image methodology for identifying linear mechanisms underlying shape discrimination. Although prior attempts to apply classification image methods to shape perception have been confined to simple radial shapes, the method proposed here can be applied to general 2-D (planar) shapes of arbitrary complexity, including natural shapes. Critical to the method is the projection of each target shape onto a Fourier descriptor (FD) basis set, which allows the essential perceptual features of each shape to be represented by a relatively small number of coefficients. We demonstrate that under this projection natural shapes are low pass, following a relatively steep power law. To efficiently identify the observer’s classification template, we employ a yes/no paradigm and match the spectral density of the stimulus noise in FD space to the power law density of the target shape. The proposed method generates linear template models for animal shape detection that are predictive of human judgments. These templates are found to be biased away from the ideal, overly weighting lower frequencies. This low-pass bias suggests that higher frequency shape processing relies on nonlinear mechanisms.

Introduction

A considerable portion of primate visual cortex appears to be involved in the coding of object shape (Connor, Brincat, & Pasupathy, 2007). Although objects in our visual world are generally three dimensional, the boundary of a 3-D object projects to the retina as a closed contour, and the shape of this planar contour provides an important cue for object detection and recognition (Elder & Velisavljević, 2009). In this paper, we focus on the visual coding of this planar shape information.

There are many theories of planar shape representation but little consensus on which is the best account of human shape perception. Early theories (Attneave, 1954; Hoffman & Richards, 1984) emphasized local features of the contour, such as curvature extrema and inflections, and later work generalized these features across scale-space (Mokhtarian & Mackworth, 1986; Dubinskiy & Zhu, 2003). Alternatively, shape can be encoded as a sum over global basis functions of the contour (Granlund, 1972; Pavlidis, 1980).

An alternative to a contour-based representation is a symmetry axis representation in which a planar shape is represented by its skeleton and an associated distance function (Blum, 1973; Feldman & Singh, 2006; Kimia, Tannenbaum, & Zucker, 1995). This approach has the advantage of making perceptually salient shape symmetries explicit and capturing regional properties of shape that are not directly represented by contour methods.

A third class of theory considers shape as a process of transformation or growth (Elder, Oleskiw, Yakubovich, & Peyr´e, 2013; Grenander, Srivastava, & Saini, 2007; Jain, Zhong, & Lakshmanan, 1996; Leyton, 1989; Sharon & Mumford, 2006; Thompson, 1917). These theories have a natural generative expression that can be used to support inference with noisy or incomplete visual data and are more able to capture topological properties of objects (Elder et al., 2013).

In order to test these theories as models for human shape perception, a general method for measuring human sensitivity to shape features is required. Ideally, the method will not be biased toward a specific theory and can be applied to a wide variety of shapes.
including natural shapes. One candidate is the classification image methodology, which measures how human judgments vary with random perturbations in the stimulus in order to identify selectivity for visual features. This methodology has been used effectively to reveal aspects of linear (Beard & Ahumada, 1998; Murray, Bennett, & Sekuler, 2002) and nonlinear (Morgenstern & Elder, 2012; Nandy & Tjan, 2007; Solomon, 2002) spatial contrast encoding. The question we ask in this paper is whether it can be applied to shape perception.

Classification images

In a standard yes/no classification image experiment, the observer is required to discriminate between two deterministic visual signals (signal 1 and signal 0; Figure 1). On each trial, a random sampler selects one of these two signals, and a random sample of white Gaussian spatial pixel noise is added. The resulting noisy stimulus is then displayed to the observer, who must decide which of the two signals is present.

It can be shown (Green & Swets, 1966) that, to generate the largest proportion of correct responses, an ideal observer must have an accurate model, called a template, of the difference (signal 1 - signal 0) of the two (noiseless) signals (Figure 2). The ideal observer then computes the inner product of this template with the noisy stimulus and forms a decision by comparing this decision variable with a fixed threshold.

If a human uses the same strategy but with an imperfect template and internal additive Gaussian noise, that template can be estimated with the classification image methodology. Specifically, an unbiased estimate of the template is obtained by computing the mean of the stimulus noise for each of the four possible (signal, response) pairs, adding the two means for the response = signal 1 trials and subtracting the two means for the response = signal 0 trials (Ahumada, 2002; Beard & Ahumada, 1998; Murray et al., 2002). We henceforth refer to this method for template estimation as the noise-averaging method.

It has also become increasingly popular to view this problem within the framework of the generalized linear model (GLM) and to estimate the template by maximizing the likelihood of the observer’s responses (Abbey & Eckstein, 2001; Knoblauch & Maloney, 2008; Murray, 2011; Solomon, 2002). Once again, we assume that an observer’s judgment is formed by computing the inner product of an imperfect template with the noisy stimulus and then comparing against a fixed threshold. Logistic regression generates a maximum likelihood estimate of the template under the assumption that internal noise follows a logistic distribution, and probit regression assumes that internal noise follows a normal distribution.

Although the GLM approach is not guaranteed to be unbiased, by explicitly maximizing the likelihood of the data, it may be more efficient than the noise-averaging method. Because studies of the relative efficiency of the two methods have yielded mixed results (Abbey & Eckstein, 2001; Knoblauch & Maloney, 2008; Murray, 2011), we evaluate both approaches to the estimation of shape templates (see Linear systems identification method).

Applying the classification image methodology to shape

In the standard form of the classification image methodology, noise is added in the luminance domain in the form of a perturbation in the gray level at each pixel, and the method estimates the weight assigned to the gray level at each pixel. This is useful for estimating human contrast sensitivity but is not appropriate for understanding shape perception. This is because shape is defined not in the luminance domain, but in the spatial domain as a sequence of spatial (x, y) coordinates in the image. Thus, to understand shape perception, noise must be added not to the gray levels of the image but to the spatial coordinates of the shape.
Kurki, Saarinen, and Hyvärinen (2014) have recently adapted the classification image methodology to explore the perception of an interesting class of planar shapes called radial frequency (RF) patterns. An RF pattern is a closed planar contour that can be conveniently represented in polar coordinates with origin at the center of the shape. RF patterns are, by definition, shapes that can be represented as a sum of simple radial basis shapes whose radial coordinates are sinusoidal functions of polar angle.

To explore human perceptual sensitivity to features of these shapes, Kurki et al. (2014) displayed each shape as a pattern of bright spots and then added random noise to the radial coordinates of the elements. Under the assumption that human shape discrimination is based upon an inner product of the stimulus with a stored template of the difference in the radial position of these spots for the two signals, the classification image technique can be used to estimate this template. Specifically, Kurki et al. used classification image methodology to examine the perceptual sensitivity to perturbations of the so-called RF4 pattern, which has the appearance of a rounded square. By sampling these patterns at a relatively small number of locations, Kurki et al. were able to keep the dimensionality of the stimulus low, which allowed reasonably accurate estimates of the classification image. The results of these experiments suggest that human observers integrate information globally over the shape in order to form a judgment.

We are inspired by this successful application of the classification image technique to simple shapes and in this paper seek to extend this success in a number of important ways. First, the stimuli employed by Kurki et al. (2014) consisted of a disconnected pattern of blobs that must be perceptually integrated into a coherent object. In order to dissociate this process of perceptual organization from the perception of shape, it is desirable to employ smooth, continuous contours. Second, although RF patterns are an interesting class of shape, shapes that are not roughly circular cannot generally be represented as RF patterns, and this excludes many natural shapes (e.g., animals, human bodies). Given that our visual system has evolved to process natural shapes, it is important to be able to directly measure how these more complex shapes are processed.

Generalizing to complex natural shapes

In the present study, we use animal shapes drawn from the Hemera photo object database as visual stimuli. Animal shapes are of particular ecological relevance, figuring prominently in the earliest known examples of cave art (Aubert et al., 2014) and are known to be processed efficiently by the primate visual system (Elder & Velisavljević, 2009; Fabre-Thorpe, Richard, & Thorpe, 1998; Thorpe, Fize, & Marlot, 1996). We represent these shapes as high-resolution polygons, which appear as smooth continuous contours when rendered on a digital display.

An immediate issue arises: These smooth continuous contour stimuli generally consist of thousands of points—too many dimensions to estimate efficiently and reliably using a classification image methodology. To address this issue, we employ a Fourier descriptor (FD) representation of shape (Granlund, 1972; Pavlidis, 1980) in which each vertex \((x, y)\) of the polygon is represented as a coordinate \(x + yi\) in the complex plane. By taking the Fourier transform of the complex vector representing these vertices, we obtain the FD representation, which represents the complex amplitude of the shape at each frequency over the index space of the vector. We stress that by coding a shape as a function of arc length rather than polar angle, the FD representation generates a complete (i.e., invertible) description of an arbitrary polygon in the plane and, thus, can faithfully represent general shapes, including natural shapes. This is quite distinct from radial basis functions, which can only represent shapes that are functions of polar angle (e.g., convex shapes, star-like shapes). Importantly, one can capture the main features of a natural shape using only a small number of the lowest FD frequency components, limiting the dimensionality of the stimulus to a manageable level.

To be precise, suppose a deterministic polygonal shape (the signal) consists of the complex \(N\)-vector \(s_j\) representing the sequence of \(N\) vertices \(s_j\), \(j \in [0, \ldots, N - 1]\). The low-pass FD representation of the shape is then the complex \(2M\)-vector \(S\), where

\[
S_k = \frac{1}{N} \sum_{j=0}^{N-1} s_j e^{-2\pi ijk/N}, \quad k \in [-M, \ldots, M - 1]. \tag{1}
\]

Here \(k\) indexes frequency: \(k = 0\) is the DC component that determines the location of the stimulus (fixed to zero), the sequence \(k = -1, -2, \ldots, -M\) represents low to high negative frequencies and the sequence \(k = 1, 2, \ldots, M - 1\) represents low to high positive frequencies. (Although for real input signals Fourier coefficients for negative frequencies are simply the conjugates of the coefficients for the corresponding positive frequencies, this is not the case for complex signals.)

To limit the dimensionality of the estimation problem, we require that \(M \ll N\). In the section on dimensionality, we demonstrate through simulation how using a smaller value for \(M\) leads to more accurate estimation of template coefficients. In our psychological experiments, we set \(M = 16\); Figure 3 shows an...
example. Because we must represent both real and imaginary FD coefficients, this shape is described by a total of 64 parameters. These low-pass animal shapes will be used as signal 1 for our discrimination experiments. For signal 0, we construct a stimulus consisting only of the first fundamental \( (S_{-1} + S_{+1}) \) of the FD representation for the corresponding animal shape with all other coefficients (including DC) set to zero. The resulting signal 0 stimulus traces an ellipse roughly approximating the animal shape. Note that, because the DC component and first fundamental are matched between signals 1 and 0, discrimination must be based on the 29 positive and negative higher frequency components \( k = \pm 2, \ldots, \pm (M - 1), -M \).

In our experiments, the visual stimulus \( \tilde{S} \) will consist of one of the two possible signal shapes \( S \) corrupted by a complex Gaussian noise vector \( N \) added in the FD domain: \( \tilde{S} = S + N \). The real and imaginary components of the noise are drawn independently. In the spatial domain, our noisy stimulus \( \tilde{s} \) will be given by \( \tilde{s} = s + n \), where \( n \) is the inverse Fourier transform of \( N \). Note that because the Fourier transform is linear, \( n \) will also be Gaussian.

Matching the noise to the signal

Although the standard classification image method uses white noise, natural images tend not to be white but low pass, containing more energy at low frequencies than high frequencies (Field, 1987). Natural shapes, when represented in the FD domain, also tend to be low pass. In particular, we find that, for animal shapes drawn from the Hemera photo object database, spectral density decreases roughly linearly in log–log space, indicating a power law: \( S_k \propto |k|^{-a} \). Figure 4 shows the spectral amplitudes for the three animal stimuli we employ in this study. We see that the amplitude spectra are roughly approximated by power laws with exponents \( a \) in the range 1.3 to 1.7. (Power law fits explain between 48% and 65% of the variance in these three examples.)

Due to the low-pass nature of the shapes, adding sufficient white stimulus noise to reduce human performance to 75% correct would drive the signal-to-noise ratio (SNR) for higher frequencies to zero, rendering them useless and preventing any estimation of higher frequency coefficients of the human template. Also, there would be a large difference in the amplitude spectrum of signal 0 (white) and signal 1 (low-pass) stimuli, which might dominate the observer’s decision at the expense of phase information. For these reasons, we elected to roughly match the spectral density of the added noise to the spectral amplitude of the signal difference, using low-pass power-law noise with exponent matching the power-law fit to the amplitude spectrum of the animal shape. This roughly equalizes the expected SNR of low and high shape frequencies for the ideal observer although deviations of the stimulus amplitude spectrum from the power-law model mean that the exact utility of each frequency will vary. The FD representation makes the addition of low-pass noise straightforward as the noise added to each FD coefficient of the stimulus remains independent.
The ideal observer

The ideal observer uses the matched filter $W$ that maximizes SNR and, hence, discriminability $d'$. If the noise were white, the ideal observer would form a real-valued scalar decision variable $R$ given by

$$R = \text{Re}(W^H S),$$

where $W = S_1 - S_0$ is the signal difference, $W^H$ is the Hermitian conjugate (i.e., the conjugate transpose) of $W$ and the operator $\text{Re}()$ takes the real part of its argument.

For our nonwhite noise, the matched filter is also shaped by the noise:

$$W = \frac{R}{C_1} D S,$$

where $R$ is the real-valued covariance of the added noise over frequencies $k$ (Kay, 1998). Because in our case the noise at each frequency is independent, $R$ is diagonal, and the ideal template can be written as

$$W_k = S_k / \sigma_k^2 \times S_k |k|^{\alpha}, k = \pm 2, \ldots, \pm (M - 1), -M. \quad (2)$$

Because the spectral amplitude of the signal $S$ falls roughly as $|k|^{-\alpha}$, the ideal observer’s template is actually high pass, increasing in amplitude with frequency, roughly as $|k|^{\alpha}$. This makes sense because, although the SNR is by design equalized across frequencies, the signal amplitudes are falling roughly as $|k|^{-\alpha}$, and so higher frequency components must be boosted to contribute equally to the decision variable.

By Parseval’s theorem, the decision variable $R$ can also be computed in the spatial domain: $R = \text{Re}(w^H S)$, where $w$ is the inverse Fourier transform of $W$. In other words, the observer’s computation can be thought of as an inner product in either the FD or spatial domains, and the corresponding templates are related through the Fourier transform.

Computing the shape-classification image

We model human perceptual shape discrimination using the linear template model (Figure 2) in the FD domain. In the standard noise-averaging classification image method the added noise is white, and an unbiased estimate $\hat{W}$ of the observer’s inner template $W$ can be computed as

$$\hat{W} = (N_{11} + N_{01}) - (N_{10} + N_{00})$$

(White stimulus noise), \quad (3)

where $N_{ij}$ is the mean of the added noise over all trials in which the stimulus contained signal $i$ and the observer indicated signal $j$ (Ahumada, 2002; Beard & Ahumada, 1998; Murray et al., 2002).

In our case, the noise is low pass, not white. Abbey and Eckstein (2002) considered the discrimination of two distinct signals embedded in additive nonwhite Gaussian noise within a two-alternative, forced choice (2AFC) experimental paradigm. In particular, they showed that, for a linear observer with additive Gaussian internal noise, generalizing from white to nonwhite stimulus noise involves normalization of the estimated template by the covariance of the added noise. Murray (2016) extended this result to a yes/no paradigm, but because that proof is embedded in the context of an analysis of more general decision rules, we provide in Appendix B a proof specific to the yes/no task with a threshold decision rule. In particular, we prove that an unbiased estimate $\hat{W}'$ of the observer’s inner template can be computed for general multivariate Gaussian stimulus noise, added independently to real and imaginary coefficients by

![Graph showing amplitude spectrum of three animal shapes used in this study.](image)
dividing the biased estimate by the real-valued noise covariance \( \Sigma \):

\[
\tilde{\mathbf{W}} = \Sigma^{-1}\mathbf{W} \quad \text{(General case)}.
\]  (4)

In our case, the covariance matrix is diagonal, and the coefficients of the observer template are given by

\[
\hat{W}_k = \hat{W}_k / \hat{\sigma}_k^2 = |k|^{2\beta} \hat{W}_k. \quad \text{(5)}
\]

Because the GLM framework does not depend upon the distribution of the stimulus, no adjustments to it are required to estimate the observer template \( \mathbf{W} \).

Assessing the shape-classification image

In a white-noise framework, the ideal template is simply the signal difference, and deviations of the human observer from ideal can be assessed by comparing the estimated human template with the signal difference. However, in our low-pass stimulus noise framework, the ideal observer must also normalize the stimulus by the noise covariance:

\[
W_k^I \times S_k / \sigma_k^2 = S_k |k|^{2\beta}, \text{ resulting in a high-pass template.}
\]

In order to visualize human template tuning to the low-pass FD frequency structure of natural shapes, we therefore scale estimated templates by the noise covariance \( \sigma_k^2 = |k|^{-2\beta} \). For the noise-averaging method, this cancels the earlier normalization to the calculation of the classification image (Equation 5), and so the similarity of human and ideal observer tuning can be assessed by simply comparing the un-normalized classification image \( \mathbf{W} \) (Equation 3) with the signal difference \( \mathbf{AS} \). However, for the GLM method, the estimated template \( \hat{\mathbf{W}} \) must be multiplied by the noise covariance \( \Sigma \) prior to comparison with the signal difference \( \mathbf{AS} \).

The human observer template \( \mathbf{W} \) is only identifiable up to a scale factor. To facilitate comparison with the ideal observer, we scale the classification image \( \mathbf{W} \) by the positive scale factor \( \beta \) that minimizes a measure of deviation from the signal difference \( \mathbf{AS} \). We expect uncertainty in the estimated coefficients of the observer template \( \hat{\mathbf{W}} \) to scale with the standard deviation of the stimulus noise \( \sigma_k \). We therefore determine the scale factor \( \beta \) that minimizes the weighted squared deviation between the observer template and the signal difference:

\[
\beta = \arg \min_{\beta} \sum_{k=\pm 2, \ldots, \pm M, M+1} \left( \frac{1}{\sigma_k^2} \right) |\beta \hat{W}_k - S_k|^2
\]

\[
= \arg \min_{\beta} \sum_{k=\pm 2, \ldots, \pm M, M+1} |k|^{2\beta} |\beta \hat{W}_k - S_k|^2. \quad \text{(6)}
\]

Scaling estimated templates by the noise covariance also allows us to visualize the templates in the spatial domain. The estimated spatial template is simply the inverse Fourier transform \( \hat{\mathbf{w}} \) of the estimated (and noise-scaled) FD template \( \hat{\mathbf{W}} \), and tuning can be assessed by comparing \( \hat{\mathbf{w}} \) to the signal difference \( \mathbf{AS} \). Note that the fundamental of the FD representation forms an ellipse in the spatial domain that acts as a kind of “scaffold” that higher FD frequencies modulate. Because the fundamental for signal 0 and signal 1 are matched, the fundamental for the signal difference is zero. As a result, rendering the signal difference or estimated human template without the fundamental generates an uninterpretable squiggle. This is rectified by adding the fundamental to the ideal and estimated human templates prior to displaying in the spatial domain.

Simulations

Before reporting psychophysical results, we present the results of three simulation studies that have informed the design and validation of our method.

Linear systems identification method

To compare noise-averaging and GLM approaches to template estimation, we conducted a simulation experiment using the rabbit shape as signal. The stimulus noise gain was set to the mean value required for our three human observers to perform at 75% correct (see Experiment 2 below). We simulated a noisy ideal observer model, adding internal Gaussian noise to bring performance down to 75% correct. We generated estimates \( \mathbf{W} \) of the simulated observer template using both noise-averaging and probit GLM methods for 100–5,000 trials, repeating the experiment 30 times. These estimates were evaluated by computing the squared error (Equation 6) of the (noise-scaled) estimated template \( \hat{\mathbf{W}} \) relative to the signal difference. We found that the two methods led to very similar accuracies with the GLM having a slight edge for experiments with between 500 and 2,500 trials (Figure 5a). Because the experiments described below employ 1,500 trials, we employ the GLM method in most of the analyses to follow.

Dimensionality

To evaluate how the dimensionality \( M \) of the low-pass FD shape representation affects the accuracy of template estimation, we conducted a second simulation, again using the rabbit shape as signal but varying the
dimensionality $M$ of both the signal and the noise components of the stimulus (i.e., the number of low-frequency harmonics) and adjusting the gain of the stimulus noise to maintain 75\% correct performance. For each value of $M$, we ran 30 simulated experiments of 1,500 trials each and used the GLM method to estimate the observer template. Figure 5b shows that, as the stimulus dimensionality $M$ increases, the mean error of estimated observer template coefficients also increases for all estimated FD frequencies. This result demonstrates the importance of using a low-dimensional FD subspace to obtain accurate shape template estimation.

### Amplitude and phase information

The Gaussian noise added to the FD coefficients induces noise in both the FD amplitude and phase domains. To interpret the psychophysical results that follow, it is helpful to know how this noise determines the information available in these two domains for our shape-discrimination task. The amplitudes of the noisy FD coefficients follow a smooth, positively skewed distribution defined on $[0, \infty)$ known as a Rice distribution (Rice, 1945, equations 3.7–10). In the FD phase domain, the stimulus noise induces a smooth, symmetric, unimodal distribution centered on the signal phase. When conditioned on amplitude, the phase distribution is von Mises; however, the marginal phase distribution is more complicated; we derive an analytical expression for this distribution in Appendix A.

To understand how this low-pass noise affects the information available in the FD amplitude and phase domains, we employed our analytical models for the marginal amplitude and phase distributions of the noisy FD coefficients (Appendix A) to implement.

![Figure 5](image.png)

**Figure 5.** Results of template estimation simulations using the rabbit shape as signal. The simulated observer used an ideal template with added internal Gaussian noise. Plots show mean and standard error over 30 repetitions. (a) Total weighted squared error (Equation 6) for the noise averaging and probit GLM template estimation methods as a function of the number of trials. (b) Weighted squared error at each frequency (Equation 6) for a 1,500-trial experiment as a function of the dimensionality $M$ of the stimulus.

![Figure 6](image.png)

**Figure 6.** Sensitivity ($d'$) of FD coefficients, their amplitudes, and phases for the shape discrimination task (rabbit shape).
“limited ideal” observer models that are allowed to use either only one complex FD frequency component or just the amplitude or the phase of that one FD component. We then ran these models on the experiment, using the rabbit shape as signal, to determine their sensitivity ($d'$) (Figure 6). We found that although both FD amplitude and phase carry information, phase is, on average, somewhat more informative than amplitude (mean $d'$ of 1.2 for phase vs. 0.75 for amplitude). A similar pattern is seen for the other two shapes used in this study.

**Experiment 1**

**Method**

**Participants**

There were three observers: author JW (O1), author IF (O2), and a third naïve observer (O3). Observers gave informed consent prior to participation.

**Stimuli**

The observer was required to discriminate between two signal shapes: signal 1 and signal 0. For signal 1, we employed three different animal shapes in a blocked design: a deer, a rabbit, and a turtle (Figure 4). The shapes were derived from the Hemera object data set, which consists of blue-screened images of isolated objects, roughly $600 \times 500$ pixels in size. The bounding contours of the objects were extracted at pixel resolution and represented in the FD domain using only the $16$ lowest negative and $15$ lowest positive frequencies. For signal 0, we employed the first fundamental ($S_{-1} + S_{+1}$) of the FD representation for the corresponding animal shape (signal 1), which traces an ellipse roughly approximating the animal shape.

On each trial, the visual stimulus consisted of one of these two signals corrupted by independent low-pass Gaussian noise added in the FD domain. In particular, we set $\sigma_k \propto |k|^{-\alpha}$, where $\sigma_k$ is the standard deviation of the added Gaussian noise at frequency $k$ and $\alpha$ is the power-law exponent of the spectral density of the natural shape being discriminated (signal 1). Note that independent noise was added to real and imaginary components of each FD coefficient.

Each shape was shown as a white contour centered at the center of the screen, roughly $16^\circ \times 16^\circ$ in visual angle on a dark gray background. At the viewing distance of 57 cm, each pixel subtended roughly $1.5$ arcmin.

**Design and procedure**

The experiment adhered to the tenets of the Declaration of Helsinki.

The experiment was blocked by animal. There were 1,500 trials per animal, run in three sessions of 500 trials each. Before each session, the observer was (re)familiarized with the two signal shapes. First the (noise-free) animal shape (signal 1) was shown until the observer pressed a button and then the (noise-free) ellipse (signal 0) shape was displayed. A second button press ended the familiarization procedure.

On each trial, one of the two possible signals was randomly selected with uniform probability, a random sample of low-pass Gaussian noise was added in the FD domain, and the inverse Fourier transform was applied to this noisy FD signal to generate the displayed spatial stimulus (Figure 7). The observer was given unlimited time to indicate with left/right arrow keys which of the two signals was shown. Feedback was
An adaptive psychometric procedure (Quest, Watson, & Pelli, 1983) was used to adjust the gain of the noise after each trial in order to maintain performance near 75% correct.

Results

We found that the Quest procedure reliably maintained performance near 75% correct (Table 1). Figure 8 shows an example of the estimated human shape template in both FD and spatial domains for one observer and one shape. Note that the noise-averaging and GLM methods generate very similar templates. Qualitatively, the estimated human template seems to roughly track the ideal at low frequencies but is attenuated at high frequencies. Figures 9a and d show that this low-pass bias is consistent across observers and shapes.

To quantitatively assess this apparent low-pass bias, we fit power law models to the estimated human template coefficients $\hat{W}_k$ (Figure 10). Observe that the fall-off in spectral density as a function of frequency is much steeper for the estimated human templates than for the corresponding ideal templates.

One prediction of the linear template model is that templates estimated using signal 0 trials only should not be systematically different from templates estimated using signal 1 trials only. Further, it has been shown that nonlinearities in human detection and discrimination mechanisms tend to bias templates

### Table 1. Experiment 1. Performance (percentage correct) for each observer and shape.

<table>
<thead>
<tr>
<th>Observer</th>
<th>Rabbit</th>
<th>Deer</th>
<th>Turtle</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>75.7</td>
<td>76.7</td>
<td>76.7</td>
</tr>
<tr>
<td>O2</td>
<td>72.7</td>
<td>74.5</td>
<td>78.7</td>
</tr>
<tr>
<td>O3</td>
<td>79.8</td>
<td>77.5</td>
<td>73.1</td>
</tr>
</tbody>
</table>

Figure 8. Experiment 1. Example shape classification image estimated with noise-averaging and GLM methods.

---

Wilder, Freund, & Elder (2018) 18(8):9, 1–21

Downloaded from jov.arvojournals.org on 07/12/2022
toward the signal (Beard & Ahumada, 1998; Goris, Zanten, & Wagemans, 2008; Morgenstern & Elder, 2012; Nandy & Tjan, 2007; Solomon, 2002). Thus, if human shape-discrimination mechanisms involve substantial nonlinearities, templates estimated from signal 1 trials only are predicted to look more like the animal shapes than templates estimated from signal 0 trials only. Conversely, any resemblance between the animal shapes and templates estimated from signal 0 trials only cannot be due to such bias because the animal shapes were not present in the stimulus for any of these trials.

Spatial observer templates estimated from signal 0 trials only and signal 1 trials only are shown for the GLM method in Figure 9b and c, respectively, and for the noise-averaging method in Figure 9e and f, respectively. We observe that templates derived from signal 0 trials only still track the low frequencies of the animal stimuli, indicating a significant linear component to this shape-discrimination task. At the same time, systematic differences between templates derived from signal 0 trials and those derived from signal 1 trials indicate the presence of nonlinearities. Qualitatively, the signal 1 templates seem to exhibit sharper features than the signal 0 templates.

To assess whether these nonlinearities can be explained in terms of differences in shape-frequency tuning, we fit power-law models to the templates estimated from signal 0 trials only and from signal 1 trials only; Figure 11 summarizes the results. Despite the apparent difference in the spatial templates, there is no consistent difference in the spectral slopes (exponents of the best-fitting power laws) for templates estimated from all trials, signal 0 trials only or signal 1 trials only, \( F(1, 12) = 0.29, p = 0.6 \). This suggests that these nonlinearities have more to do with the phase tuning of human shape-discrimination mechanisms.

Do these nonlinearities bias the signal 1 templates toward the animal shape? To test this, we computed the weighted squared deviation of estimated templates...
from ideal (i.e., the objective minimized by Equation 6). Figure 12 shows that templates estimated from signal 1 trials are indeed, on average, slightly closer to the ideal templates, and a three-way ANOVA (signal 0 vs. 1 vs. shape vs. observer) reveals that this main effect is statistically significant, $F(1, 12) = 23.5$, $p = 0.0004$.

Evaluating the model

We employ two methods to evaluate the estimated linear shape templates as models of human shape discrimination. Specifically, we compare the agreement between a linear model $M_H$ based upon the estimated human shape template and a linear model $M_I$ based upon the ideal template. To avoid overfitting, $M_H$ templates and responses were computed using leave-one-out cross-validation over the 1,500 trials of the experiment.

First, we use the $t$ score method employed by Morgenstern and Elder (2012), which is related to the measure of choice probability introduced by Britten, Newsome, Shadlen, Celebrini, and Movshon (1996). Although the original choice probability method was nonparametric, in our experiments the Gaussian nature of the stimulus noise means that model responses will also be Gaussian distributed, making a parametric approach appropriate.

The $t$ score method measures the agreement between the scalar values of the model decision variable and the binary responses of the human observer. The premise is that if the model decision variable is causal on the human responses, its value should be predictive of those responses. To assess this, the trials are first partitioned into a subset in which the stimulus contained signal 0 and a subset in which the stimulus contained signal 1. Then, for each of these subsets, the $t$ score for the difference in the mean model response when the observer responded signal 1 versus when they responded signal 0 is computed:

$$t = \frac{E[R_M|R_H = 1] - E[R_M|R_H = 0]}{\sqrt{\text{Var}[R_M|R_H = 1]/n_1 + \text{Var}[R_M|R_H = 0]/n_0}}.$$  (7)
Here $R_M$ is the model response, $R_H$ is the human response, and $n_1$ and $n_0$ are the number of trials in which the human observer responded signal 1 and signal 0, respectively. To be consistent with human judgments, the model should generate high values when the observer responds signal 1 and low values when the observer responds signal 0, thus producing a large positive $t$ score.

Figure 13 compares the $t$ scores for a linear template model $M_H$ based on the estimated human template with those for a linear template model $M_I$ based on the ideal template; $t$ scores are generally higher for $M_H$, $F(29, 1) = 10.3, p = 0.0033$, indicating that the model based upon the estimated human template is more consistent with human behavior than the ideal template model.

Our second method of evaluation takes into account the presence of internal noise in our human observers. Our human observers perform in the 73%–79% correct range on our shape-discrimination task. Simulations reveal that the ideal observer performs at 100% correct for the same stimuli, and a linear model based upon the estimated human templates performs in the 76%–81% correct range. This suggests that at least part of the inefficiency in human observer responses derives from internal noise, and to be complete, a model of human performance must take this noise into account.

To this end, we revise our models $M_H$ and $M_I$ to include added internal zero-mean Gaussian noise. $M_H$ employs the estimated observer template, and $M_I$ employs the ideal template. The gain of the internal noise was adjusted so that the proportion correct of the
model matched that of the human observer. Specifically, we measured proportion correct for the model over a range of noise gains, fit a sigmoid function, and then from this function estimated the noise gain that would generate the proportion correct attained by the human observer.

To assess these models, we measured trial-by-trial agreement with human psychophysical responses. Figure 14a shows the trial-by-trial agreement of the two models with the human data (i.e., the proportion of model responses matching human responses). These are compared against the agreement expected by chance for an observer that matches human performance but is otherwise statistically independent, given by $p_c^2 + (1 - p_c)^2$, where $p_c$ is proportion correct. The estimated observer template model $M_H$ is consistently more predictive of human judgments than the ideal template model $M_I$, and a three-way ANOVA (model $\times$ shape $\times$ observer) reveals that this difference (main effect of model) is significant, $F(1, 12) = 26.42, p = 0.0002$. This result clearly indicates the utility of the shape-classification image method: It produces a model that is significantly more predictive of human performance than could be obtained by simply degrading an ideal observer model with noise.

The trial-by-trial agreement of model $M_H$ with human judgments is in the 70%–76% range. Should this be considered good? If $M_H$ were a perfect model of the human observer, we would expect its agreement with the human observers to be comparable to its internal consistency, i.e., the agreement between its responses to identical stimuli (same stimulus noise sample, but different internal noise samples). The results of this analysis are shown in Figure 14b. We observe that the internal consistency of the noisy ideal observer $M_I$ is at chance levels, indicating that internal noise is the dominant factor limiting its performance. In contrast, we observe much higher internal consistencies for our noisy observer template model $M_H$, in the range of 72%–90%, indicating that stimulus noise and internal noise jointly determine its performance. Importantly, we note that the internal consistency of $M_H$ is considerably higher than its agreement with our human observers. This discrepancy shows that $M_H$ is not a perfect model of human shape detection; deviations could include both inaccuracies in the estimated template as well as unmodeled nonlinearities in the human visual detection mechanism.

To gain further insight into how far $M_H$ is from being a perfect model, it would be helpful to also know the internal consistency of the human observer when presented with repeated trials of exactly the same stimulus (same signal, same noise sample). This internal consistency yields an upper bound on the agreement any model with the same internal consistency could hope to achieve with the human data. If model $M_H$ achieves agreement near to this upper bound, it should be judged a good model. This motivates our second experiment.

### Experiment 2

Our second experiment was identical to our first with the exception that each stimulus (signal + noise) was repeated twice, in separate trials, separated by a random interval. This double-pass technique allows us to measure human internal consistency and compare this with model–observer agreement.

#### Method

Methods are identical to Experiment 1 except where noted below.

#### Participants

Three human observers participated in this experiment. O1 and O2 were the authors (JW and IF) who were also observers in Experiment 1. O3 was a new observer, naïve to the purpose of the experiment.

#### Stimuli

Stimuli were the same as for Experiment 1.

#### Design and procedure

The procedure was identical to Experiment 1 with the following exceptions. Although for Experiment 1 we used Quest to adapt the gain of the noise from trial to trial, in Experiment 2 the gain of the noise was set independently for the three shapes to match the average noise threshold over observers from Experiment 1. The noise was then fixed at this gain for all trials and for all three observers.

Observers performed 3,000 trials per shape: two passes of 1,500 trials each. Each trial of the first pass involved an independent sample of the stimulus noise. These same stimuli (with the same noise samples) were presented again in the second pass, but in a different random order. Exactly the same noise samples were
used for all observers, allowing consistency both within an observer and agreement between observers to be computed. Although the order of the trials was randomized so that the delay between repeated presentations varied widely, all observers saw the stimuli in the same order.

Results

We found that the fixed noise gains selected for each shape yielded performance close to 75% correct for all observers (Table 2). Figure 15 shows the estimated templates. The results are qualitatively similar to the results of Experiment 1.

As for Experiment 1, we employed two methods to evaluate the estimated linear shape templates as models of human shape discrimination. (For details of the methods, please see the earlier section on Evaluating the model.) Figure 16 compares the t scores for a (noiseless) linear template model $M_I$ based on the estimated human template with those for a (noiseless) linear template model $M_H$ based on the ideal template; $t$ scores are generally higher for $M_H$, $F(29, 1) = 25.1, p = 2.5 \times 10^{-5}$, indicating that the model based upon the estimated human template is more consistent with human behavior than the ideal template model.

As for Experiment 1, we also added internal noise to our observer models $M_H$ and $M_I$ to match the performance of our human observers. Figure 17a shows the trial-by-trial agreement of the two resulting models with the human data (i.e., the proportion of model responses matching human responses) together with the average agreement between different human observers as a reference. A three-way ANOVA (model $\times$ shape $\times$ observer) reveals a significant main effect of model, $F(2, 20) = 31.78, p = 6.2 \times 10^{-7}$, and a (protected) least significant difference (LSD) post hoc test indicates that the model based on the estimated observer templates $M_H$ is significantly more predictive of human responses than the noisy ideal observer model $M_I$, $t(2) = 5.36, p = 3.12 \times 10^{-5}$. At the same time, we see a significant effect of shape, $F(2, 40) = 43.2, p < 10^{-5}$, indicating that the templates for different shapes also differ significantly.
time, the agreement between different human observers exceeds that between $M_H$ and human observers, $t(2) = 2.43, p = 0.025$. We conclude from this analysis that, although the noisy observer template model $M_H$ provides a better account of human judgments than the noisy ideal template model $M_I$, there is an important aspect of human shape discrimination, common to our three observers, that is captured by neither model.

The double-pass design also allows us to compute the internal consistency of each of our human observers and compare against the internal consistency of our models (Figure 17b). As in Experiment 1, the internal consistency of the noisy ideal model $M_I$ hovers around chance levels, indicating that internal noise is the dominant factor limiting its performance, and the internal consistency of the noisy observer template model $M_H$ is consistently higher, indicating that both stimulus and internal noise jointly limit its performance. A three-way ANOVA (model $\times$ shape $\times$ observer) reveals a significant effect of model on internal consistency, $F(2, 20) = 33.96, p = 3.7 \times 10^{-7}$. Post hoc tests, again using Fisher’s protected LSD, reveal that model $M_H$ has significantly higher internal consistency (lower internal noise) than $M_I$, $t(8) = 5.54, p = 2.0 \times 10^{-5}$. At the same time, our human observers have higher internal consistency than $M_H$, $t(8) = 2.50, p = 0.02$.

Figure 18 provides a more detailed comparison of the human data with our noisy models. Here we varied the internal noise gain for the two models over a broad range to sweep out curves that relate performance (proportion correct) to internal consistency; note that the gain of the internal noise decreases from left to right as these curves are traversed. We also plot our nine experimental measurements (3 observers $\times$ 3 shapes) as points in this space. The substantial downward displacement of these points relative to the ideal observer highlights
the substantial degree of systematic inefficiency present in the human visual system: for the same level of internal noise, proportion correct for our human observers may be up to 10% lower than for the noisy ideal. The fact that the human data points fall near to the estimated classification image observer ($M_H$) curves suggests that $M_H$ does a reasonable job in capturing this systematic inefficiency.

**Discussion**

**Contributions**

In this paper, we have introduced a new yes/no classification image methodology to explore the discrimination of 2-D (planar) shape. In contrast to prior work (Kurki et al., 2014), our method can be applied to fully general shapes, including ecologically important stimuli, such as animals. Critical to the success of the technique is the projection of shape stimuli onto an FD basis, which allows the essential features of the shapes to be captured by a small number of coefficients, limiting the dimensionality of the observer template that must be estimated.

Because natural shapes are steeply low pass (Figure 4), efficient identification of these weights also depends upon the use of low-pass (correlated) stimulus noise. We have developed a methodology for analyzing responses from a yes/no shape discrimination task using low-pass stimulus noise and proven that an unbiased estimate of the observer template can be estimated by the standard noise-averaging method as long as mean noise fields are normalized by the noise covariance (Appendix B; see also Murray, 2016). Alternatively, templates can be estimated by maximizing the likelihood of a generalized linear model.

There are several aspects of our results attesting to the utility of this new methodology. First, the method yields observer templates bearing a resemblance to the target shapes (Figures 9 and 15), demonstrating that the method is picking up information relevant to the task. Second, templates estimated from signal 0 trials only are similar, although not identical, to templates estimated from signal 1 trials, showing that the resemblance of estimated templates to target animal shapes is not just an artifact of a nonlinearity in the human shape-discrimination system (Ahumada & Beard, 1999; Morgenstern & Elder, 2012; Nandy & Tjan, 2007; Solomon, 2002). Third, using two different evaluation methods, we find that the trial-by-trial agreement between human responses and models based upon estimated observer templates is significantly higher than for models based upon ideal templates. Fourth, the internal consistency of the estimated noisy observer template model $M_H$ approaches the internal consistency of human observers when matched to human performance (Figure 17). Finally, a plot of performance (proportion correct) versus internal consistency (Figure 18) reveals that human performance is limited by a substantial degree of systematic inefficiency, roughly matched by the systematic inefficiency of our estimated observer templates (model $M_H$).

In summary, it seems that the proposed method is sound and can help characterize human shape-discrimination performance. In particular, what it appears to tell us here is that human shape discrimination can be modeled, to some approximation, within the linear template framework and that discrimination is based primarily on lower shape frequencies. Consistent with this selectivity for lower shape frequencies, Figure 18 shows that the human data and the $M_H$ model are closest to ideal for the turtle shape, which is the most low pass of the three shapes tested (Figure 4).

**Open questions**

Why would human shape-discrimination mechanisms revealed by our method be biased toward lower shape frequencies? One possibility is that spatial uncertainty generates greater phase uncertainty for high shape frequencies than low frequencies. For example, suppose that spatial uncertainty can be modeled as a Gaussian process $\sim N(0, \sigma^2)$ in the arc length coordinate of the shape. This would cause phase uncertainty to scale roughly as $k\sigma$, where $k$ is the FD frequency. To take this into account, the observer should attenuate these higher shape frequencies in the linear template. This hypothesis could be tested in the future by measuring performance as a function of the variance of added phase noise to identify the equivalent internal phase noise for both high and low frequencies (Pelli & Farell, 1999).

The small but apparent differences between shape templates estimated from signal 0 and signal 1 trials (Figures 9 and 15) show that there are significant nonlinearities in human shape-discrimination mechanisms. What is the nature of these nonlinearities? One possibility is that the human visual system deals with the increase in phase uncertainty with FD frequency by shifting from linear to phase-invariant mechanisms at higher FD frequencies. To test this idea, we modified the input to our linear template model to include not only the complex FD stimulus coefficients $S_k$, but also their phase-invariant moduli $|S_k|$. We reasoned that if the human visual system shifts from linear to phase-invariant encoding at higher frequencies, this expanded model should yield higher agreement with the human data, and we should see a shallower fall off in the $|S_k|$ template coefficients relative to the fall off in the $S_k$ template coefficients. We found that, in fact, the
agreement with the human data for the two models was very similar, and the low-pass fall off was nearly identical for the linear and phase-invariant coefficients. We conclude from this analysis that nonlinearities in human shape discrimination cannot be accounted for by a simple shift from linear to phase-invariant encoding of the stimulus at higher frequencies.

If not a simple shift to phase-invariant mechanisms at higher frequencies, what could explain the evidence for the nonlinear encoding we see in our data? Prior work in the domain of spatial vision may provide insight.

Classification image analysis in the power-spectrum domain (Morgenstern & Elder, 2012) suggests that large-field contrast grating detection is based on an incoherent (phase-invariant) energy pooling but over highly localized linear filters. Similarly, it is possible that in our experiments higher shape frequencies are coded at least partially incoherently by localized shape mechanisms and combined through nonlinear (e.g., energy) pooling. Candidate localized shape encoding mechanisms include shapelets (Dubinskiy & Zhu, 2003) and formlets (Elder et al., 2013). It is also quite possible that higher FD frequency components are not processed independently from other components. For example, coding of higher FD frequencies may be conditioned upon phase alignment with lower FD frequencies.

Given these possibilities, we should be careful in our interpretation of the low-pass bias in estimated linear templates; this bias does not necessarily mean that higher shape frequencies are not important, only that they are not used in a linear way. Repeating the experiments reported here but with higher shape frequencies removed could serve to better quantify the nonlinear role of higher frequencies in shape discrimination. A modest decline in performance would suggest that these higher frequencies are not critical to human performance (whether processed linearly or nonlinearly), and a substantial drop in performance would suggest that the high frequencies are being used but in a nonlinear way.

We must also ask to what degree the task we set for our observers is a reasonable approximation of how we use shape “in the wild.” In our experiments, observers were asked to repeatedly discriminate the same noisy animal shape from a noisy ellipse, always presented at exactly the same location, orientation, and size. This is quite unlike the typical way we process shape information in day-to-day tasks in our normal visual environment, in which we must be prepared to discriminate between thousands of different shapes that may appear at arbitrary locations, poses, and scales.

We note that our method could easily be adapted to vary the location, orientation, and scale of the shape, and this would serve to make the task more natural. Although there have been some attempts to generalize the classification image methodology to more than two-way discrimination (Dai & Micheyl, 2010; Knoblauch & Maloney, 2008; Watson, 1998), as with the two-class method, these multiclass methods estimate difference templates, not the templates themselves. This problem is magnified in the multiclass case because the template for a category contains the negative images from each other class (Murray, 2011). In addition, the number of trials needed to get reliable estimates of the templates increases with the number of alternative categories.

A third issue concerns the nature of the noise. We employed low-pass noise here in order to match the spectral density of the signal (the shape) and to improve the efficiency of template estimation. However, prior work (Abbey & Eckstein, 2007) suggests that human observers can, under some conditions, adapt their templates to be more efficient depending upon the spectral density of the noise, e.g., by up-weighting higher frequencies in low-pass noise. This issue points clearly to new experiments: estimate the observer shape template with different types of noise and assess how estimated templates adapt.

Keywords: classification image analysis, ideal observer analysis, shape perception, shape discrimination

Acknowledgment

The authors would like to acknowledge funding by the Natural Sciences and Engineering Research Council of Canada (NSERC) CREATE Program in Vision Science & Applications, and the NSERC Discovery Grant Program. IF also acknowledges receiving York University Bridging Funds.

Commercial relationships: none.
Corresponding author: John Wilder.
Email: jdwilder@cs.toronto.edu.
Address: Department of Computer Science, University of Toronto, Toronto, Canada.

Footnote

1 Although we defined correctness in terms of the signal (0 or 1) used to generate the stimulus and not in terms of the ideal observer response, for the range of noise gains used in our psychophysical experiments, the ideal observer always generated the correct response, and so the two were equivalent.

References

Abbey, C. K., & Eckstein, M. P. (2001). Maximum-


Appendix A: Phase distribution of a complex normal variable

Let \( z = x + iy \) be a complex normal random variable, where the real and imaginary components may have different means but identical variance:

\[
x \sim N(x_0, \sigma^2), \quad y \sim N(y_0, \sigma^2).
\]

\( z \) can also be represented in polar coordinates (\( r, \theta \)), where \( x = r \cos \theta, \ y = r \sin \theta \). We wish to identify the probability density of \( z \) in this polar coordinate frame.

By the multivariate change of variables theorem, we have that

\[
p_{R, \Theta}(r, \theta) = \left|\frac{\partial(x, y)}{\partial(r, \theta)}\right| p_{X, Y}(r \cos \theta, r \sin \theta),
\]

where \( \left|\frac{\partial(x, y)}{\partial(r, \theta)}\right| \) is the Jacobian of the inverse transformation and evaluates to \( \left|\frac{\partial(x, y)}{\partial(r, \theta)}\right| = r \).

Thus, we have

\[
p_{R, \Theta}(r, \theta) = r p_{X, Y}(r \cos \theta, r \sin \theta)
\]

\[
= \frac{r}{2\pi \sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left((r \cos \theta - r_0 \cos \theta_0)^2 + (r \sin \theta - r_0 \sin \theta_0)^2\right)\right\}
\]

\[
= \frac{r}{2\pi \sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left(r^2 - 2rr_0 \cos(\theta - \theta_0) + r_0^2\right)\right\}
\]

\[
= \frac{r}{2\pi \sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left((r - r_0 \cos(\theta - \theta_0))^2 + r_0^2 \sin^2(\theta - \theta_0)\right)\right\}
\]

\[
= \frac{r}{2\pi \sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left(r - r_0 \cos(\theta - \theta_0))^2\right)\right\} f(\theta),
\]

where \( f(\theta) \triangleq \exp\left\{-\frac{\theta_0^2}{2\sigma^2} \sin^2(\theta - \theta_0)\right\} \).

Defining \( r' = r - r_0 \cos(\theta - \theta_0), \ r' \in [-r_0 \cos(\theta - \theta_0), \infty) \), this can be written as

\[
p_{R', \Theta}(r', \theta) = \frac{1}{2\pi \sigma^2} \left(r' + r_0 \cos(\theta - \theta_0)\right) \times \exp\left(-\frac{\theta^2}{2\sigma^2}\right) f(\theta).
\]
Marginalizing over $r'$ on the domain $r' \in [-r_0 \cos(\theta - \theta_0), \infty]$ then yields

$$p_{\theta}(\theta) = \frac{1}{2\pi} \exp \left( -\frac{r_0^2 \cos^2(\theta - \theta_0)}{2\sigma^2} \right) \times \exp(-v^2/2). \quad (12)$$

Thus, the marginal distribution of the phase $\theta$ is symmetric about $\theta_0$. We have verified this equation by sampling $z$ and comparing against the resulting empirical density of the phase.

### Appendix B: Yes/no classification images with correlated stimulus noise

In the standard classification image method with additive white Gaussian noise, an unbiased estimate $\hat{w}$ of the observer’s inner template $w$ can be computed as

$$\hat{w} = (\bar{w}_{01} + \bar{w}_{11}) - (\bar{w}_{00} + \bar{w}_{10}), \quad (13)$$

where $\bar{w}_{ij}$ is the mean of the added noise over all trials in which the stimulus contained signal $i$ and the observer indicated signal $j$ (Ahumada, 2002; Murray et al., 2002).

What if the noise is Gaussian but not white? Abbey and Eckstein (2002) considered the discrimination of two distinct real signals embedded in additive nonwhite (correlated) Gaussian noise, within a 2AFC experimental paradigm. In particular, they showed that, for a linear observer with additive Gaussian internal noise, generalizing from white to nonwhite noise involves normalization of the estimated template by the covariance of the noise. Here we show that this result generalizes to a yes/no task. We first prove the claim for real signals and then show that it easily generalizes to complex-valued stimuli.

Let $s = s + n$ be an $m$-dimensional real-valued random vector representing a visual stimulus, where $s$ is a binary signal variable taking one of two values $s_0$ or $s_1$ and $n \sim \mathcal{N}(0, \Sigma)$ is added zero-mean multivariate normal stimulus noise with covariance matrix $\Sigma$. Following Ahumada (2002), we model the internal source of variability by assuming that the criterion $\beta$ is also a random variable. We let $R \in \{0, 1\}$ represent the two possible responses of the observer.

Let $\bar{w}_{ij} = \mathbb{E}[n s = s_i, R = j]$. Then

$$\bar{w}_{00} = \mathbb{E}_{n \beta}[n | w^\top (s_i + n) < \beta], \quad (14)$$

and

$$\bar{w}_{11} = \mathbb{E}_{n \beta}[n | w^\top (s_i + n) > \beta], \quad (15)$$

where $w$ is an $m$-vector representing the observer template. Without loss of generality, we assume that $||w|| = 1$.

Let $\mathbf{U}$ be an orthonormal rotation matrix with first column $w$, so that $\mathbf{U}^\top w = \mathbf{e}_1 \triangleq [1, 0, \ldots, 0]^\top$, and let $\mathbf{n}' \triangleq \mathbf{U}^\top n$ and $\mathbf{s}' \triangleq \mathbf{U}^\top s_i$. Note that $\mathbf{n}' \sim \mathcal{N}(0, \Sigma')$ is also zero-mean multivariate normal with covariance matrix $\Sigma' = \mathbf{U}^\top \Sigma \mathbf{U}$. Note also that $w^\top (s_i + n) = (\mathbf{U}^\top w)^\top \mathbf{U}^\top (s_i + n) = \mathbf{e}_1^\top (s_i' + n') = s_i' + n_1'$, where $s_i'$ and $n_1'$ are the first elements of $s_i'$ and $n'$, respectively.

In this new coordinate frame, $\bar{w}_{ij}$ can be expressed as

$$\bar{w}_{ij} = \mathbb{E}_{n \beta} [n | w^\top (s_i + n) > \beta] = \mathbb{E}_{n \beta} [n | s_i' + n_1' > \beta]. \quad (16)$$

Let us now consider the conditional expectation of each element $n_k'$ of the noise $n'$ in this new coordinate frame. Consider first the conditional expectation of the first element $n_1'$:

$$\mathbb{E}_n [n_1' | s_1' + n_1' > \beta - s_1'] = \int_{\beta - s_1'}^{\infty} n_1' p(n_1') dn_1'$$

$$= \frac{1}{\sqrt{2\pi} \sigma_1} \int_{\beta - s_1'}^{\infty} n_1' \exp\left( -\frac{n_1'^2}{2\sigma_1^2} \right) dn_1'$$

$$= \frac{\sigma_1}{\sqrt{2\pi}} \exp\left( -\frac{(\beta - s_1')^2}{2\sigma_1^2} \right), \quad (17)$$

where $\sigma_1$ is the standard deviation of $n_1'$.

Now consider the conditional expectations of the remaining elements $n_k', k \neq 1$:

$$\mathbb{E}_n [n_k' | s_1' + n_1' > \beta - s_1'] = \mathbb{E}_n [n_k' | s_1' > \beta - s_1']$$

$$= \int_{\beta - s_1'}^{\infty} n_k' p(n_k' | s_1') dn_k' \cdot \int_{\beta - s_1'}^{\infty} \cdot \int_{\beta - s_1'}^{\infty} n_1' p(n_1' | s_1') dn_1' \cdot \int_{\beta - s_1'}^{\infty} \cdot \int_{\beta - s_1'}^{\infty} n_k' p(n_k' | s_1') dn_k'. \quad (18)$$

Because $n_1'$ and $n_k'$ are jointly normal, the conditional random variable $n_k' | s_1' \sim \mathcal{N}(\mu_k' | s_1', \sigma_k^2 | s_1')$ is univariate normal with mean and variance given by

$$\mu_k' | s_1' = (\sigma_k' / \sigma_1') n_1', \quad (19)$$

$$\sigma_k^2 | s_1' = \sigma_k^2 - \sigma_{1k}^2 / \sigma_1^2, \quad (20)$$

where $\sigma_k^2$ is the variance of $n_k'$ and $\sigma_{1k}^2$ is the covariance.
Thus, we have
\[
\mathbb{E}_w \left[ n'_k | s'_k + n'_1 > \beta \right] = \int_{\beta - s'_1}^{\infty} \mu_{k|1} p(n'_1) \, dn'_1 \\
= (\sigma_{1k} / \sigma'_1)^2 \int_{\beta - s'_1}^{\infty} n'_1 p(n'_1) \, dn'_1 \\
= \frac{\sigma_{1k}^2}{2\pi \sigma'_1} \exp \left( - \frac{(\beta - s'_1)^2}{2\sigma'_1} \right). \tag{21}
\]

As a result, we can write
\[
\mathfrak{n}_1 = \mathbb{E}_w \left[ \mathfrak{n}' | s'_1 + n'_1 > \beta \right] = \frac{1}{\sqrt{2\pi \sigma'_1}} \mathbb{E}_\beta \left[ \exp \left( - \frac{(\beta - s'_1)^2}{2\sigma'_1} \right) \right] \\
\quad \times \mathcal{U}[\sigma'_1^2, \sigma_{12}^2, \ldots, \sigma_{1m}^2]^T \\
= c_{i1} \mathbb{E}_w \left[ \mathfrak{n}' n'_1 \right], \tag{22}
\]
where
\[
c_{i1} = \frac{1}{\sqrt{2\pi \sigma'_1}} \mathbb{E}_\beta \left[ \exp \left( - \frac{(\beta - s'_1)^2}{2\sigma'_1} \right) \right] \text{ is a positive proportionality constant.}
\]

This result can be transformed back to the original pixel coordinates by applying the inverse rotation \( \mathbf{U} \) and making the substitution \( n'_1 = \mathbf{w}^T \mathfrak{n} = \mathfrak{n}^T \mathbf{w} \) before taking the expectation:
\[
\mathfrak{n}_1 = c_{i1} \mathbb{E}_w \left[ \mathbf{Un}'_1 \right] = c_{i1} \mathbb{E}_w \left[ \mathbf{nn}^T \right] \mathbf{w} = c_{i1} \Sigma \mathbf{w}. \tag{23}
\]

Thus, we have that an unbiased estimate of the observer template \( \mathbf{w} \) can be obtained by premultiplying \( \mathfrak{n}_1 \) by the inverse covariance of the stimulus noise:
\[
\mathbf{w} = (c_{i1} \Sigma)^{-1} \mathfrak{n}_1, \quad \text{with}
\]
\[
c_{i1} = \frac{1}{\sqrt{2\pi \sigma'_1}} \mathbb{E}_\beta \left[ \exp \left( - \frac{(\beta - s'_1)^2}{2\sigma'_1} \right) \right]. \tag{24}
\]

It is straightforward to show that for \( \mathfrak{n}_0 \) an analogous equation holds but with a negative proportionality constant. Specifically,
\[
\mathbf{w} = (c_{i0} \Sigma)^{-1} \mathfrak{n}_0, \quad \text{with}
\]
\[
c_{i0} = -\frac{1}{\sqrt{2\pi \sigma'_1}} \mathbb{E}_\beta \left[ \exp \left( - \frac{(\beta - s'_0)^2}{2\sigma'_1} \right) \right]. \tag{25}
\]

We now generalize this result to complex-valued signals. In our complex-valued linear template model, discrimination is based on a real-valued scalar decision variable \( r \) given by \( r = \text{Re}(\mathbf{w}^H \tilde{s}) \), where \( \mathbf{w} \) is the complex-valued observer template and \( \tilde{s} \) is the complex-valued noisy stimulus. This can be re-expressed as a sum of two real-valued inner products:
\[
r = \mathbf{w}_x^T \tilde{s}_x + \mathbf{w}_y^T \tilde{s}_y, \tag{26}
\]
where \( \mathbf{w}_x \) and \( \mathbf{w}_y \) are the real and imaginary components of \( \mathbf{w} \) and \( \tilde{s}_x \) and \( \tilde{s}_y \) are the real and imaginary components of \( \tilde{s} \), respectively. This can be reduced to a single real-valued inner product if we stack the real and imaginary components of the template and stimulus:
\[
\begin{align*}
\mathbf{w}_{xy} &= \begin{bmatrix} \mathbf{w}_x^T \mathbf{w}_y^T \end{bmatrix}^T, \\
\tilde{s}_{xy} &= \begin{bmatrix} \tilde{s}_x^T \tilde{s}_y^T \end{bmatrix}^T \\
\rightarrow r &= \mathbf{w}_{xy}^T \tilde{s}_{xy}. \tag{27}
\end{align*}
\]

From the proof above, we know that an unbiased estimate \( \mathbf{w}_{xy} \) of the real-valued template \( \mathbf{w}_{xy} \) can be obtained by normalizing the biased estimate (Equation 13) by the covariance of \( \tilde{s}_{xy} \):
\[
\mathbf{w}_{xy} = \Sigma_{xy}^{-1} \mathbf{w}_{xy}. \tag{28}
\]

Because there is a 1:1 identification of the real-valued coefficients of the template \( \mathbf{w}_{xy} \) with the real and imaginary coefficients of the complex-valued template \( \mathbf{w} \), Equation 28 also yields an unbiased estimate \( \mathbf{w}' \) of the latter. In our particular case, because the same real-valued independent and identically distributed noise process \( \sim \mathcal{N}(0, \Sigma) \) is used to generate both real and imaginary components of the stimulus, \( \Sigma_{xy} \) is block-diagonal and can be written as
\[
\Sigma_{xy} = \begin{bmatrix} \Sigma & 0_m \\ 0_m & \Sigma \end{bmatrix}, \tag{29}
\]
and so
\[
\mathbf{w}' = \Sigma^{-1} \mathbf{w}. \tag{30}
\]