Classification images: A review

Richard F. Murray

Department of Psychology and Centre for Vision Research, York University, Toronto, Ontario, Canada

Classification images have recently become a widely used tool in visual psychophysics. Here, I review the development of classification image methods over the past fifteen years. I provide some historical background, describing how classification images and related methods grew out of established statistical and mathematical frameworks and became common tools for studying biological systems. I describe key developments in classification image methods: use of optimal weighted sums based on the linear observer model, formulation of classification images in terms of the generalized linear model, development of statistical tests, use of priors to reduce dimensionality, methods for experiments with more than two response alternatives, a variant using multiplicative noise, and related methods for examining nonlinearities in visual processing, including second-order Volterra kernels and principal component analysis. I conclude with a selective review of how classification image methods have led to substantive findings in three representative areas of vision research, namely, spatial vision, perceptual organization, and visual search.

Keywords: categorization, computational modeling, receptive fields, spatial vision

Citation: Murray, R. F. (2011). Classification images: A review. *Journal of Vision, 11*(5):2, 1–25, http://www.journalofvision.org/content/11/5/2, doi:10.1167/11.5.2.

Introduction

Visual psychophysics has the goal of using measures of behavior to develop and test theories of visual processing. These theories may be informed by physiological findings and computational insights, but the hallmark of visual psychophysics is that, experimentally, it is the observer's behavioral responses to visual stimuli that are used to test and constrain theories. Historically, psychophysicists have used a variety of behavioral measures, including appearance matches, response times, and proportion of correct responses. Fifteen years ago, a new experimental tool, the classification image, was introduced into visual psychophysics (Ahumada, 1996). Since then, it has undergone rapid development, and it has been used to examine visual processing in new ways across the full range of vision science, from simple detection tasks to object recognition.

Given the recent progress and increasingly widespread application of classification image methods, a survey of the field may be useful. Here, I review the origins and recent development of classification image methods. I begin with a brief description of the most frequently used method of calculating classification images. I then provide some historical background, describing how classification images and related methods grew out of established statistical and mathematical frameworks and became increasingly common tools for studying biological systems. The next and largest part of this review is an exploration of recent innovations in ways of using and understanding classification images. I then make a selective review of how classification images have led to substantive findings in a few representative areas of vision research, namely, spatial vision, perceptual organization, and visual search. I conclude with some observations on what we have learned about classification image methods and some suggestions on avenues for future research.

The classification image

In a typical classification image experiment, the stimulus on each trial is one of two possible signals, randomly chosen, in a Gaussian noise field that varies from trial to trial (Figure 1). The observer tries to say which signal was shown. Ahumada (1996) introduced the following method of calculating classification images in such experiments:

$$\mathbf{c} = (\bar{\mathbf{n}}^{12} + \bar{\mathbf{n}}^{22}) - (\bar{\mathbf{n}}^{11} + \bar{\mathbf{n}}^{21}). \quad (1)$$

Here, $\bar{\mathbf{n}}^{SR}$ is the sample average of noise fields in a stimulus–response class of trials, e.g., $\bar{\mathbf{n}}^{12}$ is the average of the noise fields over all trials where the stimulus contained signal 1 but the observer identified it as signal 2. Appendix A gives a summary of the notation used throughout the article.

What intuition can we give for the calculation in Equation 1? What do we expect the classification image to reveal about how the observer decides which signal was shown? In a task performed at threshold, the observer sometimes responds correctly and sometimes responds incorrectly. The stimulus noise influences the observer's responses: on some trials, by chance, the noise has features similar to one of the signals, making the observer more likely to identify the stimulus as containing that signal. It seems plausible that $\bar{\mathbf{n}}^{11}$ and $\bar{\mathbf{n}}^{21}$ will show what

doi: 10.1167/11.5.2 Received March 4, 2011; published May 1, 2011 ISSN 1534-7362

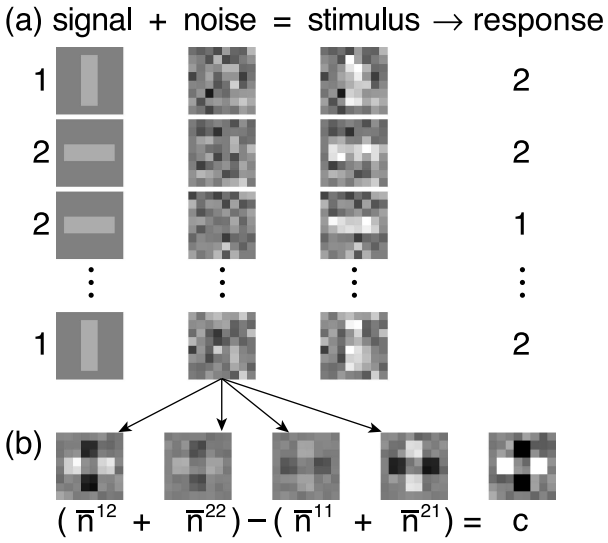

Figure 1. The standard method of calculating a classification image. (a) The experiment: on each trial, a signal and a noise image are summed to produce the stimulus, and the observer generates a response. (b) The analysis: the noise fields from each signal-response category of trials are averaged together, and the averages are combined according to Equation 1 to produce the classification image.

features the observer took to be similar to signal 1 and dissimilar to signal 2, since they are averages of noise fields over trials where the observer identified the stimulus as signal 1. If we expect these two images to be similar, then we can sum them to reduce sampling noise. For the same reasons, we might expect $\bar{\mathbf{n}}^{12}$ and $\bar{\mathbf{n}}^{22}$ to show what features the observer took to be similar to signal 2 and dissimilar to signal 1, and we could sum them as well. If we believe that these two summed images, $\bar{\mathbf{n}}^{11} + \bar{\mathbf{n}}^{21}$ and $\bar{\mathbf{n}}^{12} + \bar{\mathbf{n}}^{22}$, are on average photographic negatives of one another, since they are based on noise fields that led to opposite responses, then we can reduce sampling noise further by adding one to the negative of the other. This sequence of averages, sums, and differences leads to Equation 1.

Another, more immediate way of understanding the classification image is as a correlation map. An image showing the correlation between intensity fluctuations at each stimulus location and the observer's responses would clearly be useful for understanding the observer's decision mechanism. High positive or negative correlations would occur at locations that strongly influenced the observer's responses, and zero correlations would occur at locations that apparently had no influence on the observer's responses. The pixelwise correlation between the noise field $\mathbf{n}$ and the observer's responses r (a random variable where $r = 1$ or $r = 2$ on each trial) is

$$\mathrm{corr}[\mathbf{n}, r] = \frac{E[(\mathbf{n} - E[\mathbf{n}])(r - E[r])]}{\sigma_{\mathbf{n}}\sigma_r}. \quad (2)$$

Here, $\sigma_{\mathbf{n}}$ is the pixelwise standard deviation of the noise field $\mathbf{n}$ and σ_r is the standard deviation of r. With zero-mean noise ($E[\mathbf{n}] = 0$) and an unbiased observer ($E[r] = 1.5$, since an unbiased observer gives responses 1 and 2 equally often), Equation 2 becomes

$$= \frac{E[\mathbf{n}(r - 1.5)]}{\sigma_{\mathbf{n}}\sigma_r}, \quad (3)$$

$$= \frac{E[\mathbf{n}(r-1.5)|r = 1]P(r = 1) + E[\mathbf{n}(r - 1.5)|r = 2]P(r = 2)}{\sigma_{\mathbf{n}}\sigma_r}, \quad (4)$$

$$= \frac{E[\mathbf{n}|r = 2] - E[\mathbf{n}|r = 1]}{4\sigma_{\mathbf{n}}\sigma_r}. \quad (5)$$

Multiplying by $4\sigma_{\mathbf{n}}\sigma_r$ to eliminate the scale factor leads to

$$E[\mathbf{n}|r = 2] - E[\mathbf{n}|r = 1]. \quad (6)$$

The corresponding sample average over a finite number of trials, in a notation like the one used in Equation 1, is

$$\mathbf{c}_{\mathrm{corr}} = \bar{\mathbf{n}}^{*2} - \bar{\mathbf{n}}^{*1}. \quad (7)$$

Here, $\bar{\mathbf{n}}^{*R}$ is the sample average of the noise fields over all trials where the observer gave response R. Thus, $\mathbf{c}_{\mathrm{corr}}$ is the average of the noise fields over all trials where the observer responded $r = 2$, minus the average over all trials where the observer responded $r = 1$, regardless of which signal was shown. Equations 1 and 7 are both weighted sums in which noise fields from trials where the observer responded $r = 2$ are weighted positively, and noise fields from trials where the observer responded $r = 1$ are weighted negatively. Thus, a classification image calculated using the standard method in Equation 1 is, loosely speaking, similar to a map showing the correlations between stimulus fluctuations at each pixel and the observer's responses.[1]

These are intuitive motivations for the classification image method. Later, we will look at these rationales more closely and see what assumptions about visual processing they rely on (see The linear observer model section).

Background

To provide context for later developments, I will review the origins of classification images and related methods.

Volterra and Wiener kernel analysis

Consider a system that has a time-varying input $x(t)$ and a time-varying output $y(t)$, such as a photoreceptor whose input is the luminance at a retinal location and whose output is a membrane potential. The system may be internally complex and may have an intricate relationship between input and output, for example, showing temporal inhibition, gain control, and so on. Volterra (1930) and Wiener (1958) showed that under certain broad conditions (e.g., the system must be time-invariant and have finite memory), such a system can be approximated as a sum of simple subsystems: a zero-order subsystem, plus a first-order subsystem, plus a second-order subsystem, etc. Each subsystem responds to the input in a straightforward way. In Volterra's framework, the output of the zero-order subsystem is a constant H_0, independent of the input. The output $H_1(t)$ of the first-order subsystem is a weighted sum of past inputs, weighted according to a function $h_1(t_1)$ called the first-order *kernel*:

$$H_1(t) = \int_0^\infty h_1(t_1)x(t-t_1)dt_1. \tag{8}$$

The output $H_2(t)$ of the second-order subsystem is a weighted sum of pairwise products of past inputs, weighted according to the second-order kernel $h_2(t_1, t_2)$:

$$H_2(t) = \int_0^\infty \int_0^\infty h_2(t_1,t_2)x(t-t_1)x(t-t_2)dt_1dt_2. \tag{9}$$

The output of the nth-order subsystem is a weighted sum of n-wise products of past inputs, weighted according to the nth-order kernel $h_n(t_1, t_2, \ldots, t_n)$:

$$H_n(t) = \int_0^\infty \int_0^\infty \cdots \int_0^\infty h_n(t_1,t_2,\ldots,t_n)x(t-t_1)x(t-t_2) \cdots x(t-t_n)dt_1dt_2\cdots dt_n. \tag{10}$$

(Note that each subsystem is just an n-dimensional convolution.) The output of the system is approximated as the sum of the outputs of the subsystems:

$$y(t) \simeq H_0 + H_1(t) + H_2(t) + \ldots. \tag{11}$$

Thus, a complex system is described as a sum of simple subsystems. This is similar to a Taylor series expansion, where a function of one variable is expressed as a weighted sum of simple polynomial terms, $(x - x_0)$, $(x - x_0)^2$, $(x - x_0)^3$, etc. In fact, the Volterra series has been called a "Taylor series with memory" (Schetzen, 1980, p. 200), as it allows the estimate of $y(t)$ to depend not only on powers of $x(t)$ at time t, but also on powers and products of past values $x(t - t_1)$.

Wiener's framework is similar to that of Volterra and expresses the system as a sum of subsystems G_i based on kernels g_i. The relationship between G_i and g_i is similar to the relationship between H_i and h_i, but Wiener introduced some refinements that make G_i easier to use for modeling physical systems. For a thorough account of Volterra and Wiener kernel methods, see Schetzen (1980).

To describe a specific system in this framework, we need to find the system's kernels. Lee and Schetzen (1965) showed that we can estimate a system's Wiener kernels g_i simply by giving it a white noise input and measuring correlations between its input and its output. They showed that the zero-, first-, and second-order Wiener kernels can be measured as

$$\hat{g}_0 = E[y(t)], \tag{12}$$

$$\hat{g}_1(t_1) = \frac{1}{K}E[x(t-t_1)y(t)], \tag{13}$$

$$\hat{g}_2(t_1,t_2) = \frac{1}{2K^2}E\left[x(t-t_1)x(t-t_2)\left(y(t)-\hat{g}_0\right)\right]. \tag{14}$$

Here, $x(t)$ is the zero-mean white noise input and $K = E[x(t)^2]$ is its power spectral density. We can find these expected values by averaging over time (i.e., we assume ergodicity): we give the system a white noise input, and over many values of time t, we find the averages of $y(t)$, $x(t - t_1)y(t)$, and $x(t - t_1)x(t - t_2)(y(t) - \hat{g}_0)$. The discovery that Wiener kernels can be estimated this way held out the possibility of using simple physical measurements to completely characterize the input–output patterns of complex systems.

For our purpose, the important points are that the first-order kernel estimate is similar to a classification image measured using the correlation between a white noise input and the system's output (recall Equation 7) and that higher order kernels provide a way of extending the first-order, linear description of the system. There are some superficial differences between classification images and kernel methods as presented here: on each trial, the observer in a classification image experiment gives a single discrete response, whereas in the kernel framework the output is a continuous variable over time; and in a classification image experiment, we examine the influence of discrete

pixels, usually distributed over two-dimensional space, on the observer's responses, whereas in the kernel framework we examine the influence of a single continuous input distributed over time. Later, we will reformulate the kernel framework in a way that is more suitable for psychophysics (see Second-order kernels section).

Lee and Schetzen's methods were applied to biological systems almost immediately (de Boer & Kuyper, 1968; Stark, 1969). In an influential early application, Marmarelis and Naka (1972) used these methods to examine a three-neuron chain in the catfish retina. They injected a white noise current into a horizontal cell, which stimulated a bipolar cell that stimulated a ganglion cell whose spike responses were recorded. They repeated a single white noise stimulus several times in order to find the instantaneous spike rate of the system over time in response to the stimulus. They used Equations 13 and 14 to calculate the system's first- and second-order Wiener kernels from these data. They validated their results by calculating the response of the first- and second-order kernels to the white noise stimulus and found that the first-order kernel responded somewhat like the three-neuron chain, but that the first- and second-order kernels together gave a much better characterization. (In this pioneering work, issues of overfitting and generalization beyond the training data were naturally not addressed (Duda, Hart, & Stork, 2000).) Marmarelis and Naka, like later investigators (e.g., Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997), found that usually only enough data to estimate the zero-, first-, and second-order kernels can be collected from biological systems, since the number of correlations that need to be measured increases exponentially with the order of the kernel. This study initiated a vast amount of research using similar methods, which continues to the present day (Marmarelis & Marmarelis, 1977; Pinter & Nabet, 1992; Sakai, 1992; Wu, David, & Gallant, 2006).

Auditory psychophysics

Around the time of Marmarelis and Naka's work, Ahumada and Lovell (1971) independently developed a similar method for auditory psychophysics. The roots of Ahumada and Lovell's work were quite different, and they presented their method as an application of multiple linear regression, not Wiener kernel analysis. They investigated what stimulus features observers used to detect a narrowband auditory signal in noise. Their stimuli contained a sinusoidal signal at a fixed frequency, present on half the trials, and noise at that frequency and at nearby frequencies on all trials. Observers used a four-point rating scale to report their confidence that the target was present. Ahumada and Lovell made a least-squares regression of observers' rating responses against the stimulus energy at each frequency. They interpreted the regression coefficients as weights that observers assigned to various frequencies when judging the presence of the signal. This work introduced several themes that are still active topics of research, including how to validate classification images, how to use differences between signal-present and signal-absent classification images to detect processing nonlinearities, and how to smooth classification images and express them as sums of simple basis functions. Ahumada, Marken, and Sandusky (1975) continued this line of investigation.

Ahumada and his colleagues' work was influential in auditory psychophysics, and auditory researchers solved several problems related to classification images that were later addressed again by visual psychophysicists. Auditory researchers first examined the relationship between classification images, ideal observers, and efficiency (Berg, 1990) and investigated how classification images depend on the template and internal noise power of linear observers (Richards & Zhu, 1994). (Auditory researchers do not use the term "classification image" and typically refer to "weights" or "combination weights.")

Visual psychophysics

Abel and Quick (1978) were the first to use Wiener kernel methods in visual psychophysics. They were apparently unaware of Ahumada's work and described their experiments as an extension of Marmarelis and Naka's (1972) physiological studies. Their experiment was broadly similar to that of Ahumada and Lovell (1971). Their stimuli were sums of ten randomly scaled sinusoidal luminance patterns, and observers judged the stimulus contrast by adjusting the contrast of a nearby sinusoid until it appeared to match the stimulus. Abel and Quick used Lee and Schetzen's method to measure the first- and second-order Wiener kernels of the mapping from the ten spatial frequency amplitudes to observers' responses. Despite its similarity to later visual classification image studies, this work had little impact and has been cited only once in the ensuing 33 years (Logvinenko, 1990). This may be because the results from human observers were described only very briefly and perhaps did not convey the method's potential.

Ahumada and Beard developed the standard classification image method used in visual psychophysics (Equation 1) and used it to test models of human performance in Vernier discrimination tasks (Ahumada, 1996; Beard & Ahumada, 1997, 1998). They calculated classification images in a task where observers judged whether a line segment was aligned or offset relative to another line segment at a fixed position. They found that even though observers often show hyperacuity performance levels in Vernier tasks (Westheimer, 1979), their strategies are nevertheless suboptimal in several ways, e.g., they rely heavily on the line segment that has the same position in both signals and so conveys no information about the correct response. (This probably reflects observers' intrinsic spatial uncertainty

(Zeevi & Mangoubi, 1984).) These studies have been highly influential, and they led to the widespread use of classification images in visual psychophysics over the past fifteen years.

Developments: Methods

The linear observer model

When is a classification image a reasonable way of characterizing how observers identify stimuli? Several early papers addressed this question and concluded that the *linear observer model* is the natural starting point for understanding classification image methods (Abbey, Eckstein, & Bochud, 1999; Ahumada, 2002; Murray, Bennett, & Sekuler, 2002; Solomon, 2002).

The linear observer model is a useful tool for understanding human performance in perceptual tasks (Burgess, Wagner, Jennings, & Barlow, 1981; Green & Swets, 1966/1974; Peterson, Birdsall, & Fox, 1954). Consider a yes–no experiment where there are two signals, $\mathbf{s}^1$ and $\mathbf{s}^2$, shown in a noise field $\mathbf{n}$ that varies from trial to trial. We will let the random variable k be the signal number (1 or 2) on any given trial, so the stimulus is $\mathbf{g} = \mathbf{s}^k + \mathbf{n}$, where the components of vectors $\mathbf{g}$, $\mathbf{s}^k$, and $\mathbf{n}$ encode the stimulus contrast at each pixel. We will represent $\mathbf{g}$, $\mathbf{s}^k$, and $\mathbf{n}$ as $n \times 1$ column vectors, even when the stimuli are shown as two-dimensional images in the experiment. The linear observer model assumes that the observer has two templates, $\mathbf{t}^1$ and $\mathbf{t}^2$, that are internal representations of the signals; we also represent these as $n \times 1$ column vectors. The observer computes decision variables d^1 and d^2 by taking the dot product of the two templates with the stimulus. The observer may also add samples from independent, equal-variance internal noise sources, z^1 and z^2, to the dot products. That is, the decision variables are

$$d^1 = \mathbf{t}^{1T}(\mathbf{s}^k + \mathbf{n}) + z^1, \tag{15}$$

$$d^2 = \mathbf{t}^{2T}(\mathbf{s}^k + \mathbf{n}) + z^2. \tag{16}$$

Here, ^{T} is the matrix transpose operation, so $p^T q = \sum_i p[i]q[i]$ is the dot product of column vectors p and q. (I use brackets to refer to vector components, because later I will use subscripts to refer to samples from a random variable.) The model assumes that the observer identifies the signal as $\mathbf{s}^1$ if d^1 plus some constant a is larger than d^2. That is, the response variable r is

$$r = \begin{cases} 1 & \text{if } d^1 + a > d^2 \\ 2 & \text{otherwise} \end{cases}. \tag{17}$$

The constant a allows the model observer to be biased toward choosing one response more often than the other.

This model is redundant, because only the difference between the two templates influences the observer's responses. The observer responds $r = 2$ if

$$\mathbf{t}^{1T}(\mathbf{s}^k + \mathbf{n}) + z^1 + a < \mathbf{t}^{2T}(\mathbf{s}^k + \mathbf{n}) + z^2, \tag{18}$$

which is equivalent to

$$(\mathbf{t}^2 - \mathbf{t}^1)^T(\mathbf{s}^k + \mathbf{n}) + (z^2 - z^1) > a. \tag{19}$$

That is, the observer's decisions are determined by the difference template $\mathbf{w} = \mathbf{t}^2 - \mathbf{t}^1$ and an internal noise source z with a variance σ_z^2 that is twice the variance of z^1 and z^2:

$$d = \mathbf{w}^T(\mathbf{s}^k + \mathbf{n}) + z, \tag{20}$$

$$r = \begin{cases} 1 & \text{if } d < a \\ 2 & \text{otherwise} \end{cases}. \tag{21}$$

Equations 20 and 21 are the form of the linear observer model that we will use most often. When discussing just the yes–no experiment, there is no need to introduce separate templates $\mathbf{t}^1$ and $\mathbf{t}^2$, since the model depends only on the difference template $\mathbf{w}$. Later, when we discuss experiments with more than two signals, it will be useful to have the multiple-template notation in place (see Multiple response alternatives section).

We can depict the linear observer's strategy in a decision space that represents all possible stimuli $\mathbf{g} = \mathbf{s}^k + \mathbf{n}$ and shows how the observer categorizes them. For simplicity, suppose the stimulus is an image with just two pixels, so that we can represent all possible stimuli on a two-dimensional plane (Figure 2a). If the linear observer has no internal noise, then the decision space is divided into two regions: $\mathbf{w}^T\mathbf{g} < a$, where the observer gives response 1, and $\mathbf{w}^T\mathbf{g} \geq a$, where the observer gives response 2. The border that divides the two, where $\mathbf{w}^T\mathbf{g} = a$, is a line that is perpendicular to the template $\mathbf{w}$ (the black arrow in Figure 2a) and distance $a/|\mathbf{w}|$ from the origin.

This representation makes it clear why the standard weighted sum method in Equation 1 gives an unbiased estimate of the template (Abbey et al., 1999; Ahumada, 2002; Chichilnisky, 2001; Murray et al., 2002; Solomon, 2002). Consider the average of the noise on all trials where the signal was $\mathbf{s}^1$ but the observer identified it as $\mathbf{s}^2$ (Figure 2a, large green circle). Because the noise is circularly symmetric, the expected value of the $\mathbf{s}^1$ stimuli (small green circles) on the $\mathbf{s}^2$ side of the decision line is shifted from the overall mean of stimulus (large white circle) in a

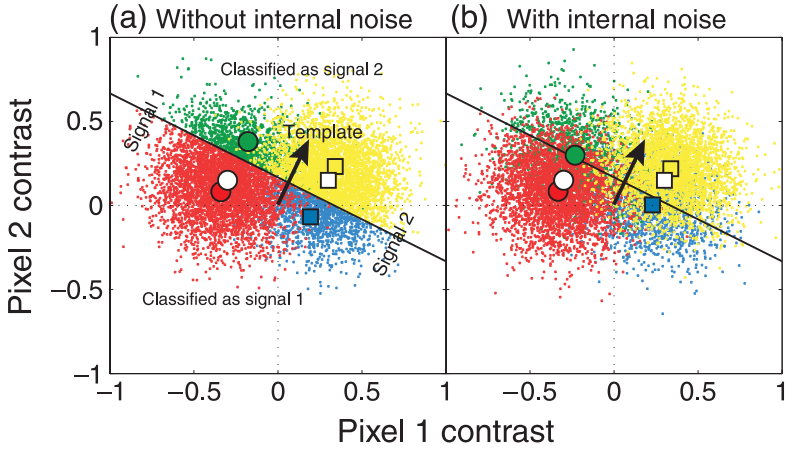

Figure 2. Decision space for a linear observer (a) without internal noise and (b) with internal noise. The white circle is the mean of the stimuli containing signal 1, and the white square is the mean of the stimuli containing signal 2. The small colored data points represent stimuli on individual trials. Red points are stimuli that contained signal 1 and were identified as signal 1, and green points are stimuli that contained signal 1 and were identified as signal 2. Yellow points are stimuli that contained signal 2 and were identified as signal 2, and blue points are stimuli that contained signal 2 and were identified as signal 1. The larger colored symbols are the averages of the corresponding small colored points, e.g., the red circle is the average of the small red points. The black arrow is the hypothetical observer's template, and the oblique black line is the decision line that the observer used to decide whether to identify a stimulus as signal 1 or signal 2.

direction that is perpendicular to the decision line and so in the same direction as the template. Thus, the average of the noise on all such trials is a vector that is proportional to the template (in expected value). The same is true for the averages of the other three stimulus–response classes of noise fields. Taking into account the direction of each shift (some in the direction of the template and some in the opposite direction), we can estimate the template by combining the averages as in Equation 1. That is, the expected value of the classification image is proportional to the linear observer's template.

The same reasoning applies when the stimuli have more than two pixels: then, the template $\mathbf{w}$ and the stimuli $\mathbf{g} = \mathbf{s}^k + \mathbf{n}$ are n-dimensional vectors, the decision surface $\mathbf{w}^T\mathbf{g} = a$ is an n-dimensional hyperplane perpendicular to the template, and the conditional averages of the noise fields are vectors parallel to the template.

This line of reasoning is also valid when the observer has internal noise, as long as the internal noise is independent of the external noise. In this case, some stimuli in the $\mathbf{w}^T\mathbf{g} < a$ region are identified as signal 2, and some in the $\mathbf{w}^T\mathbf{g} \geq a$ region are identified as signal 1 (Figure 2b). However, the conditional expected values of the noise fields are still shifted perpendicular to the decision line, so Equation 1 still gives an unbiased estimate of the template.

Significantly, in all these cases, the experimenter's choice of signals has no influence on the estimate of the linear observer's template (although it may, of course, influence the observer's choice of template). Later, we will see that when the observer uses a nonlinear strategy, things are not so simple: then, the classification image can depend on the signal as well as on the observer's strategy.

The linear observer model is a useful simplification that captures many important aspects of human performance, but it is certainly incomplete. It does not incorporate transduction nonlinearities, contrast normalization, spatial uncertainty, perceptual learning, or many other known properties of human visual processing. Some research on classification images has worked within the linear observer model and aimed at finding better methods for characterizing linear observers. Other work has developed methods that go beyond the linear observer model, taking into account nonlinearities in visual processing. To organize a review of methodological developments, I will treat these two categories separately: first, developments within the linear observer model, and second, developments that go beyond the model.

Developments within the linear observer model

Optimal weighted sums

We have seen that the expected value of a classification image is proportional to a linear observer's template. Several authors have shown that, furthermore, the standard weighted sum method given in Equation 1 is an *efficient* way of calculating classification images under some circumstances. Suppose we wish to calculate a classification image simply by taking a weighted sum of the noise fields in the four stimulus–response categories, and we wish to choose the weights based on the signal-to-noise ratio of the noise fields in each category, in order to maximize the signal-to-noise ratio of the final classification image. If the

observer is unbiased, and if the observer's performance is constant over the course of the experiment, then the standard method is the weighted sum that has the maximum signal-to-noise ratio as an estimate of the observer's template (Abbey & Eckstein, 2002b; Ahumada, 2002; Murray et al., 2002). This means, for instance, that the standard method in Equation 1 is more efficient than the correlation method in Equation 7, which is often used in auditory research and has sometimes been used in vision research as well. Conveniently, the standard method is the optimal weighted sum regardless of the variance of the observer's internal noise.

Optimal weighted sums have also been worked out for a broader range of conditions. Optimal weighted sums are known for the cases where the observer is biased, where signals are shown at multiple contrast levels, where the observer gives confidence rating responses, and where some noise fields are repeated in order to measure the observer's internal-to-external noise ratio (Abbey & Eckstein, 2002b; Ahumada, 2002; Murray et al., 2002). In all these cases, there are more efficient ways of using the data than the method given by Equation 1. Abbey and Eckstein (2002b) developed similar methods for the 2AFC design and also showed how these methods can be modified to allow non-white Gaussian stimulus noise.

These findings justify using the standard method of calculating classification images, under appropriate conditions. However, they only establish that of all the methods that estimate the observer's template by taking a weighted sum of the noise fields based on the signal-to-noise ratios of the four stimulus–response categories, the standard method is most efficient. They leave open the possibility that there are better methods whose estimate of the template is not simply a weighted sum of the noise fields. In the next section, we discuss one method where the classification image is not a weighted sum.

The generalized linear model

An important development in understanding classification images has been the realization that they fit naturally into the generalized linear model (GLM) statistical framework (Dobson & Barnett, 2008; McCullagh & Nelder, 1989). Several investigators have estimated visual classification images using instances and variants of the GLM (Abbey & Eckstein, 2001; Ludwig, Gilchrist, McSorley, & Baddeley, 2005; Solomon, 2002). Knoblauch and Maloney (2008) were the first to highlight and explore in detail the relationship of visual classification images to the GLM, and Mineault, Barthelmé, and Pack (2009) continued with valuable work along these lines. Auditory psychophysicists have used logistic regression, an instance of the GLM, to calculate classification images for several years (e.g., Alexander & Lutfi, 2004), and variants of logistic regression have been used to estimate activation patterns in fMRI data (Yamashita, Sato, Yoshioka, Tong, & Kamitani 2008), a task that is similar in some ways to estimating classification images (Victor, 2005).

In the *general* linear model, the dependent variable is a normal random variable y whose variance is fixed and whose mean $\mu_y = E[y]$ depends linearly on the covariates:

$$\mu_y = \mathbf{x}^T\beta. \tag{22}$$

Here, $\mathbf{x} = (x[1], \ldots, x[p])^T$ is a column vector of covariates and $\beta = (\beta[1], \ldots, \beta[p])^T$ is a column vector of regression coefficients. (As noted earlier, I use brackets to refer to vector components.) This model underpins many common statistical methods for handling continuous data, including multiple linear regression and ANOVA. Binary response probabilities have sometimes been modeled in this framework; in such a model, the covariates $\mathbf{x}$ represent the stimulus, the regression coefficients β represent the observer's template, the dependent variable y represents the observer's responses, and the expected value of the dependent variable $E[y]$ represents the observer's response probabilities. However, difficulties arise from the fact that the variance of a Bernoulli random variable depends on its mean, and that probabilities are limited to the range [0, 1] whereas the dependent variable y in this model can take on any value (Dobson & Barnett, 2008). Ahumada and Lovell (1971) found a partial solution to this problem by having observers make four-point confidence rating responses instead of just yes–no detection responses and calculating classification images by regressing the rating responses against noise fields. However, rating responses are not a simple linear transformation of the decision variable (Egan, Schulman, & Greenberg, 1959), so this approach does not completely resolve the mismatch between the general linear model and the linear observer model. Nevertheless, Levi and Klein (2002) reported that this method gave higher signal-to-noise ratios than a weighted sum method, although they did not give details of the comparison.

The *generalized* linear model (GLM) is an extension of the general linear model that allows the dependent variable to be continuous or categorical and allows the mean of the dependent variable to be a nonlinear function of the covariates. In the GLM, the dependent variable y is a random variable with mean $\mu_y = E[y]$, and μ_y is a possibly nonlinear function of a linear transformation of the covariates:

$$g(\mu_y) = \mathbf{x}^T\beta. \tag{23}$$

The GLM assumes that y is a random variable in the exponential family, which includes the Bernoulli, binomial, multinomial, Poisson, exponential, and normal distributions, among others. The function g is a smooth, monotonic function, called the link function, that relates the mean of the dependent variable to the covariates. Methods for finding maximum likelihood estimates of the GLM regression coefficients β are available in most statistical software packages. Dobson and Barnett (2008) give a

clear introduction to the GLM, and McCullagh and Nelder (1989) give a more thorough treatment.

The GLM is highly relevant to perceptual modeling: much of signal detection theory can be seen as a special case of the GLM, and the GLM offers promising ways of extending classical detection theory models (DeCarlo, 1998). In particular, the linear observer model implies that observers' responses can be modeled with the GLM. For a specific stimulus on trial number i, $\mathbf{g}_i = \mathbf{s}^{k_i} + \mathbf{n}_i$, the observer responds $r = 2$ with some probability p and $r = 1$ with probability $1 - p$. (Here, we subscript $\mathbf{g}$, $\mathbf{k}$, and $\mathbf{n}$ to indicate that, as components of a stimulus shown on a specific trial number i, they are now samples from random variables, not random variables.) That is, the dependent variable is a Bernoulli random variable and, thus, belongs to the exponential family. We can define $y = r - 1$, so that y encodes the observer's responses as 0 and 1, and then, the mean of the dependent variable is $\mu_y = E[y] = p$, which according to the linear observer model is

$$\mu_y = P(r = 2|\mathbf{g}_i = \mathbf{s}^{k_i} + \mathbf{n}_i), \tag{24}$$

$$= P(\mathbf{w}^T(\mathbf{s}^{k_i} + \mathbf{n}_i) + z > a), \tag{25}$$

$$= P(-z < \mathbf{w}^T(\mathbf{s}^{k_i} + \mathbf{n}_i) - a). \tag{26}$$

Introducing the normal cumulative distribution function $\Phi(x, \mu, \sigma)$, this becomes

$$= \Phi(\mathbf{w}^T(\mathbf{s}^{k_i} + \mathbf{n}_i) - a, 0, \sigma_z), \tag{27}$$

$$= \Phi(\mathbf{w}^T(\mathbf{s}^{k_i} + \mathbf{n}_i)/\sigma_z - a/\sigma_z, 0, 1). \tag{28}$$

We can rewrite these terms to show that this model is an instance of the GLM:

$$g(\mu_y) = \mathbf{x}_i^T\beta, \text{ where } g(u) = \Phi^{-1}(u, 0, 1),$$
$$\mathbf{x}_i = \begin{bmatrix} \mathbf{s}^{k_i} + \mathbf{n}_i \\ \\ 1 \end{bmatrix}, \text{ and } \beta = \begin{bmatrix} \mathbf{w}/\sigma_z \\ \\ -a/\sigma_z \end{bmatrix}. \tag{29}$$

Thus, the linear observer model leads directly to a generalized linear model of observer responses, with a Bernoulli dependent variable and a link function that is the inverse of the standard normal cumulative distribution function.[2]

Equipped with a GLM model of the observer, we can use maximum likelihood methods associated with the GLM to estimate the regression parameters β, which are the observer's template and criterion, expressed as multiples of the internal noise standard deviation. This is a very different approach to calculating classification images than the weighted sum method in Equation 1. Knoblauch and Maloney (2008) compared the two methods in simulations of linear model observers. They compared the methods by examining the mean squared residual of the least-squares fit of classification images to the simulated observer's template. They found that when the model observer had no internal noise, the GLM consistently had lower residuals, but when the model observer had realistic amounts of internal noise (e.g., internal-to-external noise ratio ≥0.5 (Neri, 2010a)), the two methods performed equally well.[3] Abbey and Eckstein (2001) reported that a similar maximum likelihood method performed better than a weighted sum method on data from human observers, but they used a suboptimal weighted sum method similar to the correlation method in Equation 7, and their maximum likelihood method incorporated a smoothing prior, so there is not necessarily a contradiction between their results and those of Knoblauch and Maloney. In auditory research, Tang and Richards (2005) found little difference between correlation methods, least-squares regression, logistic regression, and probit regression applied to psychophysical data.

The GLM is a promising approach for estimating and validating classification images in a well-established statistical framework. Some of its appealing features are that it makes no assumptions about the stimulus distribution (i.e., the noise need not be Gaussian), so it can be used to estimate classification images from natural images (Abbey & Eckstein, 2001); it has been studied in detail and has many associated statistical tools; and it is usually used in a maximum likelihood framework that can easily be extended to incorporate priors on classification images (see Dimensionality reduction section). However, there is a need for studies evaluating how well this approach works in practice, with realistic amounts of data from human observers. Maximum likelihood estimates based on the GLM can be biased even when the observer matches the model perfectly, whereas human observers are known to depart from the GLM in important ways (e.g., perceptual learning, spatial uncertainty, response nonlinearities).

Statistical tests

Sometimes we can draw conclusions about visual processing from obvious and robust features of classification images (e.g., Caspi, Beutter, & Eckstein, 2004; Gold, Murray, Bennett, & Sekuler, 2000), but often more careful analysis and statistical testing are necessary.

Sometimes we are interested in the precise profile of a classification image, e.g., the strength of an inhibitory surround in a detection template. Abbey and Eckstein (2002a) showed that the Hotelling T^2 test is useful in such cases. The T^2 test is a generalization of Student's t-test and relies on the fact that classification images based on weighted sums are, to a very good approximation, multivariate normal. Abbey and Eckstein showed how to use the

T^2 test to determine whether a classification image's mean is equal to a hypothesized image (such as an ideal observer's template (Geisler, 1989)) and whether two classification images are significantly different. The latter test is useful for detecting nonlinearities in observers' decision strategies, because as we will see later, nonlinearities often lead to differences between classification images calculated from subsets of trials where different signals were shown, e.g., the averages $\bar{\mathbf{n}}^{12} - \bar{\mathbf{n}}^{11}$ and $\bar{\mathbf{n}}^{22} - \bar{\mathbf{n}}^{21}$, which according to the linear observer model should have identical expected values when the observer is unbiased. Abbey and Eckstein also showed how to apply T^2 tests to any linear transformation of a classification image, such as a classification image that has been downsampled or averaged along some dimension in order to increase its signal-to-noise ratio.

Ahumada (2002) developed a method of testing the hypothesis that the observer's template is a specific, known image, e.g., the signal in a detection task. This test compares two estimates of the observer's internal noise power, measured relative to the power of the external stimulus noise. First, the internal-to-external noise ratio is estimated by measuring how consistently the observer responds to repeated presentations of identical stimuli (Ahumada, 2002; Burgess & Colborne, 1988): high consistency indicates low internal noise, and low consistency indicates high internal noise. Second, the internal-to-external noise ratio is estimated by assuming that the observer uses the hypothesized template and calculating how much internal noise would then be necessary in order to account for the observer's actual performance level. If these two noise estimates are inconsistent, then we can reject the hypothesis that the human observer is a linear observer who uses the template in question.

Sometimes we are not interested in the exact profile of a classification image, but instead we would just like to know what stimulus regions observers use for a task, e.g., whether observers use the eye region to identify faces. A *t*-test can show which pixels in a classification image are significantly different from zero, but with hundreds or thousands of pixels, it is important to correct for multiple comparisons, ideally with a method not as conservative as Bonferroni correction. This problem is compounded by the common practice of blurring classification images to increase their signal-to-noise ratio, which introduces correlations among neighboring pixels. Chauvin, Worsley, Schyns, Arguin, and Gosselin (2005) showed that statistical tests based on random field theory, common in neuroimaging studies, can be used for this purpose. They described methods for testing whether a single pixel is significantly different from the image mean and for testing whether a cluster of pixels above some intensity level is larger than expected by chance. These methods can be applied to smoothed or unsmoothed classification images.

Another important goal of statistical testing is to validate the classification image itself, i.e., test whether it gives a reasonably good description of the observer's decision strategy. Neri and Levi (2006) developed such a test, based on using the human observer's classification image as the template of a simulated linear observer and examining how well the simulated observer predicts the human observer's trial-by-trial responses. This test requires measuring the human observer's internal-to-external noise ratio (Ahumada, 2002; Burgess & Colborne, 1988). Neri and Levi derived upper and lower bounds for how well the simulated observer should predict the human observer's responses. Murray, Bennett, and Sekuler (2005) developed another validation test, based on the idea that by comparing the human observer's classification image to the ideal observer's template (Geisler, 1989), one should be able to predict the human observer's performance. A convenient feature of Murray et al.'s test is that it does not require measuring the human observer's internal-to-external noise ratio.

Knoblauch and Maloney (2008) noted that one advantage of calculating classification images with the GLM is that many statistical tests are available for this model. They illustrated one such test, using nested models to test whether a classification image in a detection task was significantly different on signal-present and signal-absent trials (as mentioned above, a sign of nonlinearity). Point estimation, interval estimation, inference, and goodness-of-fit methods have been developed for the GLM and should be useful for testing hypotheses about classification images.

Dimensionality reduction

We typically need several thousand trials to create a classification image with an adequate signal-to-noise ratio. Sometimes this is merely inconvenient, but it can also make it difficult to study transient phenomena like perceptual learning and sensitive populations like children and clinical groups.

One solution is simply to use a small number of pixels, either by using stimuli with few pixels or by combining neighboring pixels during data analysis. Abbey and Eckstein (2002a) studied detection of two-dimensional Gaussian bumps and analyzed radial averages of the full two-dimensional classification images. With 2,000 trials, they were able to estimate the radial profile of observers' templates quite precisely. A related approach, sometimes referred to as using "dimensional noise," is to add noise along a small number of task-relevant stimulus dimensions, such as the position or orientation of stimulus elements, instead of using pixelwise Gaussian noise (e.g., Li, Klein, & Levi, 2006; Neri & Parker, 1999).

One reason classification image methods need so much data is that we usually allow them far more flexibility than necessary. Using a 32 × 32 pixel array of white noise to cover a 1-degree square stimulus, for instance, allows the possibility that the observer will perform the task using an alternating black-and-white, pixel-by-pixel checkerboard template. This is unlikely. We can use data more efficiently modifying the estimation process to incorporate

this kind of prior knowledge about what templates are likely or unlikely. Of course, one appeal of classification images is that they offer a highly flexible, open-ended approach to probing observers' decision strategies. To incorporate prior knowledge successfully, it is necessary to strike a balance between eliminating unlikely templates and avoiding strong biases. Such tradeoffs between variability and bias are pervasive in statistical modeling (Bishop, 2006; Duda et al., 2000).

Abbey and Eckstein (2001) measured classification images using a method that incorporated priors. They used a prior that penalized large classification image pixel values, so that only pixels that played a strong role in explaining observers' responses were assigned large values (a form of *shrinkage* (Duda et al., 2000)). They also tested a prior that penalized high spatial frequencies in the classification image, thereby incorporating a form of smoothing into the estimation process. They made maximum a posteriori estimates of classification images by maximizing likelihood functions that incorporated one of these two priors. In a Gaussian bump detection task with human observers, they found that both priors increased the signal-to-noise ratio of the classification image over a maximum likelihood estimate with no prior and did not introduce obvious artifacts.

Knoblauch and Maloney (2008) estimated classification images using the generalized additive model (GAM) framework. This approach allowed them to impose a smoothness prior on the classification image by representing it as a sum of splines and penalizing splines that contributed little to explaining observers' responses (again, a form of shrinkage). As with the GLM, there are advantages to using an established statistical framework like the GAM. Interestingly, though, Knoblauch and Maloney found in simulations of a linear model observer that, with realistic levels of internal noise, the weighted sum, GLM, and GAM approaches all produced classification images with about the same signal-to-noise ratio.

Mineault et al. (2009) developed a similar approach to incorporating priors into classification image estimation. They noted that classification images typically consist of a small number of simple, blobby features, which suggests that the appropriate prior is that classification images should be *sparse* and *smooth*. They developed a framework that can accommodate a wide range of sparse and smooth priors. To illustrate their method, they implemented a prior that represents classification images in an overcomplete Gaussian pyramid,[4] a multiscale representation that codes an image as a sum of two-dimensional Gaussians at various scales (Burt & Adelson, 1983). The prior penalized large coefficients in the pyramid representation (again, shrinkage), so that only Gaussians that played a strong role in explaining observers' responses appeared in the classification image. Mineault et al. made maximum a posteriori estimates of classification images by maximizing a likelihood function that incorporated this prior. Using simulated and human observer data, they showed that the prior substantially improved the signal-to-noise ratio of classification images.

Mineault et al. gave a useful discussion of potential problems arising from the biases that priors can introduce. They suggested that the choice of a prior should depend on the purpose of the experiment: the appropriate prior is one whose biases are irrelevant to the hypothesis being tested. If we are interested in whether an observer's template has an inhibitory surround, for instance, then a prior based on Gaussians is more appropriate than one based on Gabors, since Gaussians do not have inhibitory lobes, whereas Gabors do. It may not always be obvious what biases a prior will introduce, but Mineault et al.'s suggestion is a sensible starting point for exploring priors on classification images.

Multiple response alternatives, part 1

All the methods reviewed so far were designed for experiments with just two response alternatives: the observer judges the presence or absence of a signal, classifies a face as male or female, and so on. Recent work has opened the way to measuring classification images in more flexible designs, where the observer identifies a stimulus as one of multiple alternatives.

Watson (1998) calculated classification images in a three-alternative letter identification experiment. This work was the first treatment of more than two response alternatives, but it was reported only in an abstract, so it is difficult to evaluate. Briefly, though, Watson measured the dot product of each letter signal with the average of the noise fields in each of the nine stimulus–response classes of trials and used these dot products to estimate the observer's templates as weighted sums of the letter signals.

Knoblauch and Maloney (2008) investigated classification images based on the GLM, as discussed earlier. They mostly examined the GLM with a Bernoulli dependent variable, which allows only two response categories. However, they pointed out that the GLM family also includes models with multinomial dependent variables, and they suggested that this should make it possible to estimate classification images in *m*-alternative experiments.

It is worth expanding on this suggestion. We can estimate *m*-alternative classification images using the GLM as follows. Suppose we have signals $\mathbf{s}^1$, …, $\mathbf{s}^m$, and on each trial, the observer views a noisy stimulus $\mathbf{s}^k + \mathbf{n}$ and tries to identify the signal. We can extend the linear observer model by giving the observer templates $\mathbf{t}^1$, …, $\mathbf{t}^m$. The observer calculates decision variables d^1, …, d^m by finding the dot product of each template with the stimulus and adding internal noise, $d^i = \mathbf{t}^{iT}(\mathbf{s}^k + \mathbf{n}) + z^i$. The observer identifies the stimulus by choosing the template that elicits the largest decision variable d^i, adjusted by a criterion a^i, so the observer's response is $r = \text{a rgmax}_i(d^i - a^i)$ (Van Trees, 1968/2001, p. 154). The response variable r is

now a multinomial random variable: for each stimulus, there are m possible responses with probabilities p^1, …, p^m (which sum to one). The template with the largest dot product is most likely to be chosen, but because the decision variables include internal noise there is always some probability that another template will be chosen instead.

As with the two-template model observer for yes–no tasks, this formulation of the m-alternative model is redundant, because only the differences between templates affect the observer's responses. We can remove this redundancy by taking one template, say $\mathbf{t}^1$, as a reference[5] and subtracting it from all templates: $\mathbf{w}^i = \mathbf{t}^i - \mathbf{t}^1$. Thus, the difference templates $\mathbf{w}^2$, …, $\mathbf{w}^m$ are free to vary, but $\mathbf{w}^1$ is always zero. The model observer with decision variables $d^i = \mathbf{w}^{iT}(\mathbf{s}^k + \mathbf{n}) + z^i$ gives the same responses as the original model observer. Thus, the most we can expect from a classification image experiment is to recover the linear observer's difference templates $\mathbf{w}^i$, not the individual templates $\mathbf{t}^i$. For similar reasons, we can assume without loss of generality that criterion $a^1 = 0$.

The most common form of multiple category regression based on the GLM is multinomial logistic regression (Dobson & Barnett, 2008), which uses a decision rule that is similar but not identical to the one we have just described. We can define $\tilde{d}_i$ to be the observer's decision variables without internal noise, adjusted by the criteria: $\tilde{d}^i = \mathbf{w}^{iT}(\mathbf{s}^k + \mathbf{n}) - a^i$. Multinomial logistic regression assumes that the response probabilities p^i are softmax functions of $\tilde{d}^i$:

$$p^i = \frac{e^{\tilde{d}^i}}{\sum_{j=1}^{m} e^{\tilde{d}^j}}. \tag{30}$$

The softmax rule is qualitatively similar to the noisy max rule: the template with the highest $\tilde{d}_i$ has the greatest probability of being chosen, but templates with lower values have some probability of being chosen instead.

Figure 3 shows results from a simulation where four-alternative classification images were calculated using multinomial logistic regression. A linear model observer identified a noisy stimulus as one of four 5 × 9 pixel signals (Figure 3, column 1), using the max rule and four templates (Figure 3, column 2). Classification images were estimated from 10,000 trials by making a multinomial logistic regression of the observer's identification responses against the noise fields (Figure 3, column 3). (Appendix B provides full details of the simulation.) Template 1 was taken as the reference template, so the ith classification image contains a positive image of template i and a negative image of template 1. Clear estimates of the target and reference templates appear in each classification image. There are also faint images of other templates, e.g., a ghost of template 2 in classification

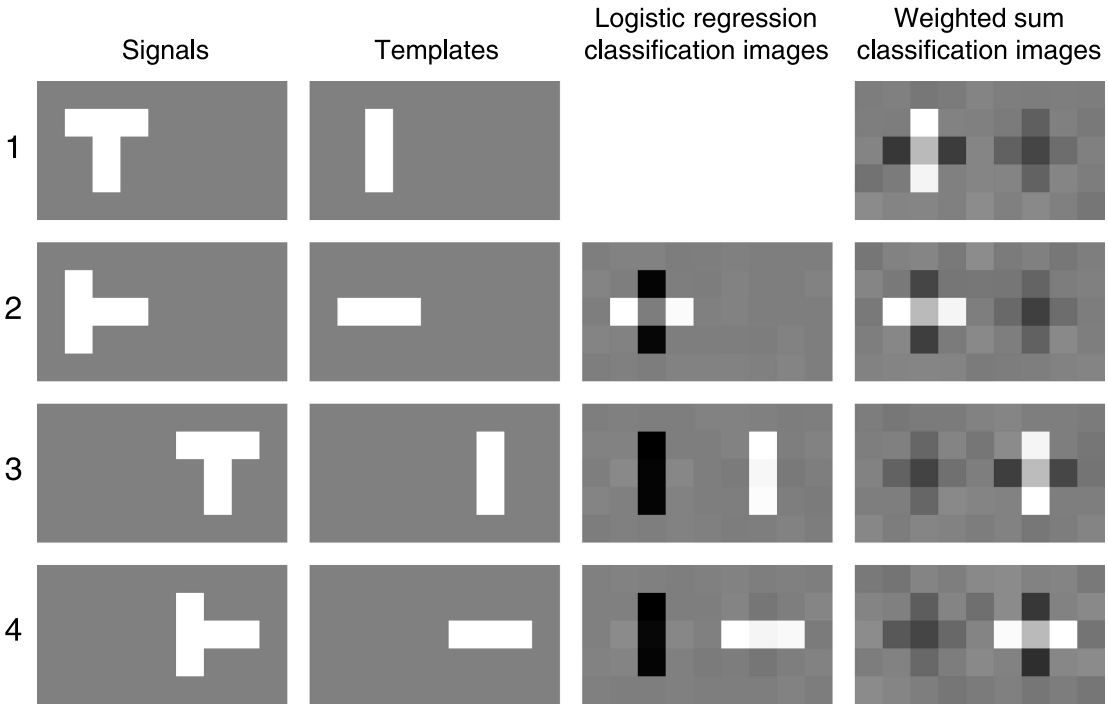

Figure 3. Four-alternative classification images from simulations of a linear model observer. Column 1 shows the signals. Column 2 shows the simulated observer's templates. Column 3 shows classification images estimated using multinomial logistic regression. Column 4 shows classification images estimated using Dai and Micheyl's (2010) method. See Appendix B for details of the simulation.

image 3, suggesting that the estimates are weakly biased. This bias might be due to the small discrepancy between the simulated observer (whose responses were based on the max rule with Gaussian noise) and the assumptions of multinomial logistic regression (which models responses using the softmax rule).

Thus, routines for estimating maximum likelihood classification images in *m*-alternative experiments are already available in statistical software packages that incorporate the GLM, such as R and the MATLAB Statistics Toolbox (R Development Core Team, 2010; The MathWorks, Natick, MA). This illustrates one benefit of formulating observer models in terms of established statistical frameworks. As with the GLM in yes–no experiments, it remains to be seen how well this method works with data from human observers.

Multiple response alternatives, part 2

Dai and Micheyl (2010) proposed a different method for estimating classification images in *m*-alternative experiments. They divided trials into m groups according to which signal was shown and analyzed the groups separately. Within each group, they measured the pixel-wise correlation between the noise fields and the correctness of the observer's responses, coded as correct = 1 and incorrect = 0. They took the correlation map $\mathbf{c}^i$ for each group to be an estimate of the observer's template for the signal $\mathbf{s}^i$ shown in that group. (That is, they used linear discriminant analysis in a one-against-the-rest fashion (Duda et al., 2000).) To validate this method, they reported simulations of a linear model observer that used five templates and made responses by finding the template that had the largest dot product with the stimulus. Their simulations showed that the correlation maps were similar to the simulated observer's templates.

However, I have found that classification images calculated with this correlation method do not actually converge to one template. Instead, each classification image is a mixture of all m templates. To see why, consider a group of trials that show a particular signal. Within this group, the observer will tend to give correct responses when the noise is similar to the template for that signal and incorrect responses when the noise is similar to one of the other templates. Accordingly, the classification image will have a component similar to the target template but also components similar to the negative images of the other templates. The amplitudes of the negative template images will depend on the probabilities of the various types of errors (i.e., stimulus $k = i$, response $r = j$), and so the mixture of negative templates will differ from one classification image to another. It also seems unlikely that the template images will combine additively. This means that contrary to Dai and Micheyl's suggestion, the classification image $\mathbf{c}^i$ does not converge to template $\mathbf{t}^i$. Further work on this method may find a way of separating the components, but as discussed earlier, the most we can expect is to recover a linear observer's template differences $\mathbf{t}^i - \mathbf{t}^j$, not the templates $\mathbf{t}^i$ themselves.

Dai and Micheyl's simulations seemed to support their method, but this was because of their choice of model observer templates: each template was proportional to the negative of the average of the $m - 1$ other templates, so the unwanted negative images of the other templates simply changed the amplitude of the classification image and were not visible as distinct artifacts. The fourth column in Figure 3 shows classification images calculated using Dai and Micheyl's method in the simulation discussed earlier (see Appendix B for details). Here, it is apparent that each classification image contains both the target template and negative images of the other templates.

Multiplicative noise

Gosselin and Schyns (2001) independently developed a method similar to classification images, which they called *bubbles*. This method uses multiplicative noise instead of additive noise and identifies the stimulus regions that help an observer to identify a stimulus correctly. In a typical bubbles experiment, the noise field is a sum of many small, randomly placed two-dimensional Gaussian bumps. The stimulus is one of m signals multiplied pointwise by the noise field, with the result that most of the signal is eliminated, and only randomly placed fragments of the signal remain visible (i.e., the fragments at the locations of the Gaussian bumps). The observer attempts to identify the signal from the visible fragments. The experiment is analyzed by calculating a *bubbles image* that shows the extent to which each visible stimulus fragment increases the probability that the observer makes a correct response. Currently, the properties of the bubbles method are best understood in the context of the linear observer model. Bubbles images have been shown to recover less information about a linear observer's template than a classification image does, so the bubbles method is most promising for investigating nonlinear decision strategies (Gosselin & Schyns, 2004; Murray & Gold, 2004a, 2004b).

The bubbles method should be able to benefit from some of the advances made with classification images. For instance, the key idea of the bubbles method is that each stimulus region contributes to a greater or lesser degree to an observer's correct responses: each stimulus has an associated map of *potent information* (Gosselin & Schyns, 2002). It may be possible to formalize this notion in the GLM framework, using the noise fields as covariates, the correctness of the observer's response as the dependent variable, and potent information maps as the regression coefficients to be estimated. This approach would permit maximum likelihood estimates of bubbles images and could be developed into a GAM (following Knoblauch and Maloney) or modified to incorporate a smooth and sparse prior on bubbles images (following Mineault et al.).

Beyond the linear observer model

Nonlinearly transformed stimuli

The most straightforward way of extending classification image methods beyond the linear observer model is to apply the same methods to properties that are nonlinear functions of the stimulus, such as contrast energy or power spectral density. This transformation does not change the class of statistical model (e.g., generalized linear model), because it transforms the covariates, not the regression coefficients. In the psychophysical sense, however, it transforms a linear model into a nonlinear model, because it allows the observer's decision variable to be a nonlinear function of the stimulus as measured in standard physical units. (Thus, what is psychophysically "linear" is ambiguous until we specify units, as there are nonlinearly related but equally valid ways of measuring stimuli, such as contrast and log contrast.)

Neri and Heeger (2002) used this approach to measure spatiotemporal classification images in a task where observers detected a thin vertical bar in dynamic noise. They measured first-order classification images, combining the means of the noise fields in each stimulus–response category according to Equation 1. They also measured second-order classification images, combining the *variance* of the noise fields in each stimulus–response category in a manner analogous to Equation 1:

$$\mathbf{c}_{\mathrm{VAR}} = \left(\mathbf{n}^{12}_{\mathrm{VAR}} + \mathbf{n}^{22}_{\mathrm{VAR}}\right) - \left(\mathbf{n}^{11}_{\mathrm{VAR}} + \mathbf{n}^{21}_{\mathrm{VAR}}\right). \tag{31}$$

Here, $\mathbf{n}^{SR}_{\mathrm{VAR}}$ is the variance of the noise fields in a stimulus–response category of trials, e.g., $\mathbf{n}^{12}_{\mathrm{VAR}}$ is the variance of the noise fields on trials where the stimulus contained signal $\mathbf{s}^1$ and the observer responded $r = 2$. Neri and Heeger's first- and second-order classification images revealed a surprising detection strategy. Three types of events made the observer more likely to say that the target was present: (a) there was a burst of high-energy noise, positive or negative in contrast, at the target location just before the time when the target might appear; (b) there was a burst of positive-contrast noise at the time and location of the target; or (c) there was a burst of negative-contrast noise at the time of and spatially adjacent to the location of the target. Neri and Heeger proposed a neural circuit composed of simple and complex cells as a detection mechanism that is consistent with these classification images. A further experiment suggested that the first-order mechanism is responsible for identification, and the second-order mechanism is responsible for detection. Other researchers have also used second-order classification images (Knoblauch & Maloney, 2008; Murray, 2002), and classification images based on the Fourier power spectrum have been informative as well (Gold, Cohen, & Shiffrin, 2006; Solomon, 2002; Taylor, Bennett, & Sekuler, 2009).

Note that a *late*, monotonic nonlinearity does not interfere with the usual method of calculating classification images. Suppose we modify the linear observer model in Equations 20 and 21 by introducing a monotonic nonlinearity f on the decision variable:

$$d^* = f(\mathbf{w}^T(\mathbf{s}^k + \mathbf{n}) + z), \tag{32}$$

$$r = \begin{cases} 1 & \text{if } d^* < a^* \\ 2 & \text{otherwise} \end{cases}. \tag{33}$$

This model is equivalent to a linear model with decision variable $d = f^{-1}(d^*)$ and criterion $a = f^{-1}(a^*)$, so methods appropriate for linear observers can be used to estimate the template $\mathbf{w}$. (The observation that a late, monotonic nonlinearity does not affect the information carried by a decision variable is sometimes called Birdsall's theorem (Lasley & Cohn, 1981).)

Second-order kernels

Neri and Heeger showed that measuring classification images based on both contrast and squared contrast can reveal important properties of observers' detection strategies. A natural generalization is to also consider products of contrasts at distinct locations as predictors of observers' responses. This is the second-order Volterra and Wiener kernel approach.

We can convert the linear observer model into a second-order model by giving the observer a linear template $\mathbf{w}$ as before and also a second-order Volterra kernel, represented as a symmetric matrix $W = W^T$. The decision rule is then

$$d = \mathbf{w}^T(\mathbf{s}^k + \mathbf{n}) + (\mathbf{s}^k + \mathbf{n})^T W(\mathbf{s}^k + \mathbf{n}) + z, \tag{34}$$

$$r = \begin{cases} 1 & \text{if } d < a \\ 2 & \text{otherwise} \end{cases}. \tag{35}$$

That is, the decision variable is a weighted sum of the stimulus elements (with weights in $\mathbf{w}$), plus a weighted sum of all products of pairs of stimulus elements (with weights in W), plus internal noise.

As discussed earlier, Abel and Quick (1978) estimated first- and second-order kernels in a task where observers judged stimulus contrast. Their dependent variable was continuous (a contrast level), so they were able to use Lee and Schetzen's methods directly. Neri (2004) developed methods for estimating the second-order kernel in tasks

where observers make yes–no responses. For instance, he showed that we can estimate the second-order kernel as follows:

$$\hat{W} = (\hat{C}^{12} + \hat{C}^{22}) - (\hat{C}^{11} + \hat{C}^{21}). \quad (36)$$

Here, $\hat{C}^{SR}$ is the sample covariance matrix of the stimuli in a stimulus–response class of trials, e.g., $\hat{C}^{12}$ is the sample covariance matrix over all trials where the stimulus contained signal $\mathbf{s}^1$ but the observer gave response $r = 2$. Neri showed that this method is strictly correct only when the observer is unbiased, but using simulations of model observers he showed that it is reasonably accurate over a wide range of biases. In subsequent work, Neri demonstrated that characteristic patterns in second-order kernels can be used to understand mechanisms of brightness perception and texture perception (Neri, 2009) and to identify simple mechanisms that often appear in models of visual processing, such as divisive normalization and max-rule uncertainty mechanisms (Neri, 2010b, 2010c).

Nandy and Tjan (2007) developed a related method. In a letter discrimination task, they examined the extent to which observers' responses depended not just on the contrast at a given pixel but also on correlations between contrast at pairs of pixels, e.g., whether the noise had positive contrast at both pixels. They used this approach to deduce how observers' templates were subdivided into component features. They did not describe their method as an instance of second-order kernel analysis, but the two approaches are broadly similar, and given the substantial amount of theory developed for kernel analysis, this may be a useful viewpoint for developing Nandy and Tjan's method further.

Uncertainty, part 1

Ahumada and Beard (1999) measured classification images in a task where observers detected a high spatial frequency (16 cycle/degree), sine-phase Gabor signal in noise. Classification images calculated from signal-present trials showed a template similar to the signal, as might be expected, but classification images from signal-absent trials were empty, consisting only of sampling noise. Ahumada and Beard noted that this is what one would expect from an observer who was uncertain about the phase of the Gabor to be detected and gave a "signal-present" response when a Gabor-like pattern of any phase at all appeared in the stimulus. On signal-present trials, such an observer's decisions would be driven largely by whether a noise pattern similar to the signal nudged the signal intensity above or below the detection criterion, and Gabor-like noise patterns at other phases would have little influence on the observer's responses; hence, on signal-present trials, the classification image would show a pattern much like the signal. On signal-absent trials, a phase-uncertain observer would respond "signal present" when a Gabor-like pattern of *any* phase appeared in the noise; over trials, Gabor-like noise patterns at different phases would average to zero, and the signal-absent classification image would be empty. Solomon (2002) reported further experiments along these lines (including a classification image analysis on the power spectrum of the stimuli) and reached similar conclusions.

It is instructive to examine the decision space for this task. Suppose signal $\mathbf{s}^1$ is blank and signal $\mathbf{s}^2$ is a sine-phase Gabor. A phase-uncertain observer could use an energy detection strategy, in which the decision variable is the squared dot product of the signal with a sine-phase template $\mathbf{t}^S$, plus the squared dot product of the signal with an orthogonal cosine-phase template $\mathbf{t}^C$:

$$d = (\mathbf{t}^{ST}(\mathbf{s}^k + \mathbf{n}))^2 + (\mathbf{t}^{CT}(\mathbf{s}^k + \mathbf{n}))^2, \quad (37)$$

$$r = \begin{cases} 1 & \text{if } d < a \\ 2 & \text{otherwise} \end{cases}. \quad (38)$$

This is the ideal strategy in a task where the observer is completely uncertain about the phase of the Gabor to be detected (Peterson et al., 1954; Van Trees, 1968/2001, p. 335). We can depict this strategy in a decision space where the first axis x_1 is in the direction of the sine-phase template $\mathbf{t}^S$, the second axis x_2 is in the direction of the cosine-phase template $\mathbf{t}^C$, and the remaining axes are in orthogonal directions (Figure 4). In this representation, the observer's decision rule is to respond "present" if $x_1^2 + x_2^2 \geq a$,

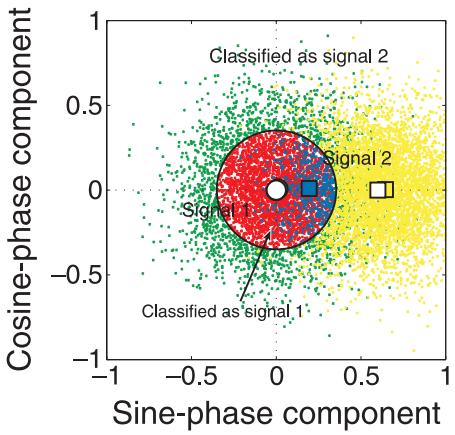

Figure 4. Decision space for a phase-uncertain observer without internal noise. The symbols and colors have the same meaning as in Figure 2. The large red and green circles (the means of signal 1, response 1 trials and signal 1, response 2 trials, respectively) are mostly hidden behind the white circle (the mean of signal 1 trials).

and "absent" otherwise. That is, the response depends on whether the stimulus is inside or outside a circle in the x_1x_2 plane and does not depend on the stimulus position along the remaining axes: the decision surface is an n-dimensional cylinder. (To simplify the exposition, we have assumed that the observer has no internal noise, but this does not affect the conclusions we will reach.)

The explanation for Ahumada and Beard's findings that we outlined above can be made clearer in terms of this decision space (Figure 4). On signal-absent trials, the expected value of the noise fields on trials where the observer responds "present" (green points) is clearly zero, and the expected value on trials where the observer responds "absent" (red points) is zero as well, so the expected value of the signal-absent classification image $\overline{\mathbf{n}}^{12} - \overline{\mathbf{n}}^{11}$ is zero. On signal-present trials, the expected value of the noise fields on trials where the observer responds "present" (yellow points) is proportional to the signal, and the expected value on trials where the observer responds "absent" (blue points) is proportional to the negative of the signal, so the expected value of the signal-present classification image $\overline{\mathbf{n}}^{22} - \overline{\mathbf{n}}^{21}$ is proportional to the signal.

Interestingly, if we change the phase of the signal in this task, the signal-present classification image changes to match the new signal, even though the observer's strategy has not changed. This is true even if we use a Gabor signal that has a phase somewhere between sine phase and cosine phase, in which case the signal-present classification image indicates a template that does not appear anywhere in the implementation of the observer's strategy, which is based on sine-phase and cosine-phase templates.

For simplicity, we have assumed that the signal falls in the subspace spanned by the observer's templates. If it does not, we can decompose the signal as $\mathbf{s}^2 = \mathbf{s}^{\parallel} + \mathbf{s}^{\perp}$, where $\mathbf{s}^{\parallel}$ is the component in the x_1x_2 plane and $\mathbf{s}^{\perp}$ is the component perpendicular to the x_1x_2 plane. (Here, $\parallel$ means "parallel" and $\perp$ means "perpendicular.") In this case, only the component $\mathbf{s}^{\parallel}$ that has a nonzero dot product with the observer's templates affects the observer's responses, and the expected value of the classification image on signal-present trials is proportional to $\mathbf{s}^{\parallel}$, the projection of the signal into the subspace spanned by the observer's templates.

An important lesson to be learned from the analysis of this phase-uncertain observer is that when we move beyond the linear observer model, the classification image can depend on interactions between the signal, the noise, and the observer's decision strategy.

Uncertainty, part 2

The case of the phase-uncertain observer suggests that classification images can be useful for understanding nonlinear decision strategies, as long as we have an adequate model of the relevant nonlinearities. Tjan and Nandy (2006) drew on a long-standing model of uncertainty (Pelli, 1985; Tanner, 1961) to develop a method of using classification images to measure observers' uncertainty in detection and discrimination tasks.

The uncertainty model that Tjan and Nandy used proposes that instead of using a single template for each stimulus, an uncertain observer uses multiple templates. For instance, if the observer is uncertain about the phase of a Gabor to be detected, then they will have many Gabor templates at a range of different phases. The observer will respond "signal present" if the maximum template response exceeds a threshold. Similarly, if the observer is uncertain about the spatial location of the stimulus in a discrimination task with letters O and X, then they will use multiple O templates at various locations, as well as multiple X templates at various locations, and their response will be based on the largest template response.

Tjan and Nandy showed that the signals have a predictable influence on classification images in such tasks. In the letter discrimination task, for instance, on trials where the letter X signal is present, the maximum letter X template response will almost always come from the letter X template at the location of the signal. Whether the observer responds "O" or "X" will then depend on whether the stimulus noise causes the response of that one letter X template to exceed the maximum of all the letter O template responses. Thus, a single letter X template will appear in the classification image from these trials. On the other hand, the letter X signal will evoke a small-to-moderate response from many letter O templates, so many letter O templates will appear in the classification image from these trials. Spatial uncertainty, in effect, blurs the letter O template, making it appear smeared over the region of uncertainty. On trials where the letter O signal is shown, the roles are reversed, so the letter O appears clearly in the classification image, and the letter X appears blurred over the region of uncertainty. Tjan and Nandy described this effect as "signal clamping" of the templates.

In letter discrimination experiments, Tjan and Nandy showed that this model qualitatively predicts the effect of uncertainty on classification images. Remarkably, they also showed that it is possible to use the differences between the classification images from the four signal-response categories to quantify the observer's uncertainty, e.g., the size of the region that the observer monitors for letters in a letter discrimination task.

Principal component analysis

Rajashekar, Cormack, and Bovik (2002) tested a novel method for examining nonlinear decision strategies. They tracked observers' eye movements during visual search for simple geometric patterns in a large noise field and used classification images to determine what features attracted observers' saccades. Previous research had led them to expect that the mechanisms guiding observers'

search trajectories would be highly nonlinear, for instance making saccades equally often to regions of high positive and high negative contrast. As discussed earlier, such strategies can result in classification images with an expected value of zero. To examine the strategies that guided observers' saccades, Rajashekar et al. did a principal component analysis (PCA) of the noise field samples in a 3.7-degree square around each fixation location (Jolliffe, 2010). Several strong principal components emerged, resembling sinusoids in various phases and orientations. In this experiment, it seems likely that these components emerged mostly because the noise field had a $1/f$ spectrum (Rajashekar, personal communication), but the method is intriguing nevertheless.

What do principal components mean in this analysis? A simple case is illustrative. Consider the phase-uncertain observer we discussed earlier (Figure 4). Consider the two response regions, "signal-absent" and "signal-present," on trials where there is no signal, only noise. The "signal-absent" response region is a long cylinder, with minimum variance across the width of the cylinder and maximum variance along the length of the cylinder. Thus, we expect to find the two smallest principal components in the x_1x_2 plane and the remaining principal components in orthogonal directions. The "signal-present" response stimuli form a Gaussian cloud with a cylinder removed, which will have maximum variance in the x_1x_2 plane and smaller variance along the remaining axes. Here, we expect to find the two largest principal components in the x_1x_2 plane and smaller principal components in orthogonal directions. Rajashekar et al. reported simulations showing that PCA does in fact recover the orthogonal templates of such a phase-uncertain observer.

For any realistic decision strategy, linear or nonlinear, the distribution of noise fields in each stimulus–response category of trials is not multivariate Gaussian, so some caution is necessary when using PCA. Nevertheless, this method does seem to have the potential to characterize an observer's decision strategy by revealing one or more directions in decision space that affect the observer's responses, and neurophysiologists have successfully used similar methods to characterize neural responses (Horwitz, Chichilnisky, & Albright, 2007; Prenger, Wu, David, & Gallant, 2004).

Methodological developments in other fields

I have mostly discussed classification image methods that have been developed and used by visual psychophysicists. However, as we have seen, similar methods have been used in other fields for a much longer time, and there is undoubtedly a great deal to be learned from approaches developed for auditory psychophysics, neurophysiology, neuroimaging, and statistical learning (Victor, 2005).

Research on classification-image-like methods has been more extensive in neurophysiology than in psychophysics. This may be because the potential payoff is higher: complete system identification is a plausible goal for research on early sensory neurons (Wu et al., 2006), whereas in psychophysics, no one would maintain that a classification image completely characterizes, say, letter identification mechanisms, except possibly over a tiny range of stimuli. Whatever the reason, the result is that over several decades physiologists have examined these methods in detail and have developed approaches that may be suitable for psychophysics as well. Higher order kernel methods (Neri, 2004, 2009, 2010b, 2010c) and random field theory (Chauvin et al., 2005), both previously used by physiologists, have already been shown to be useful in visual psychophysics. There has been extensive work in neurophysiology on using non-Gaussian noise to characterize receptive fields, for example, on using m sequences to estimate receptive fields more efficiently (Sutter, 1987) and using natural images to characterize neural responses under realistic stimulus conditions (David, Vinje, & Gallant, 2004), whereas there has been little work along these lines in psychophysics (but see Dobres & Seitz, 2010 on m sequences and Abbey & Eckstein, 2001 on natural images). Victor (2005) discusses similarities between classification image methods, neural receptive field mapping methods, and neuroimaging analysis methods and also gives a clear overview of how nonlinearities, internal noise, and high dimensionality pose difficult problems that these methods must overcome. Wu et al. (2006) give an in-depth review of recent work on methods for neural receptive field mapping.

Another promising area for cross-fertilization is statistical learning (Bishop, 2006; Duda et al., 2000; Hastie, Tibshirani, & Friedman, 2001). The goal of statistical learning algorithms is to discover structure in complex and often noisy data sets. Applied to data from classification image experiments, this means discovering what factors determine whether the observer identifies a stimulus as belonging to one category or another. In a simple but illustrative example of this approach, Cohen, Shiffrin, Gold, Ross, and Ross (2007) applied a Gaussian mixture model to classification image data, to test whether pattern identification is mediated by features detected in an all-or-none fashion, as many models of visual processing have suggested (e.g., Pelli, Burns, Farell, & Moore-Page, 2006). The statistical learning literature is vast and growing, and it is likely to contain useful ideas for extending classification image methods beyond the linear model that has dominated psychophysical applications so far.

Developments: Psychophysics

It would be impractical to give a complete review of the substantive research that has been done with classification

images, both because there is simply too much of it and because evaluating any given finding requires knowing the state of the relevant research area. Nevertheless, this review has so far been almost entirely about methods, and it is fair to ask what questions classification images have allowed researchers to answer. With this goal in mind, I will review a sample of work done on three representative topics: spatial vision, perceptual organization, and visual search. I will not attempt to review the substantial amount of relevant background material in each area, although I will sometimes mention directly related studies.

Spatial vision

Visual classification images were first developed to test models of Vernier acuity (Ahumada, 1996), so given their provenance and the role of simple, well-defined observer models in their development, it is not surprising that they have been used extensively to study low-level detection and discrimination tasks. We have already discussed several spatial vision experiments, including Ahumada and Beard's (1999) Gabor detection task, Neri and Heeger's (2002) line detection task, and Tjan and Nandy's (2006) letter identification experiment.

Classification images are ideally suited to studying the effect of noise on visual processing: in this case, the stimulus noise *is* the experimental manipulation. Such studies are useful for understanding perception of cluttered scenes, including natural scenes and medical images. They can also help refine our use of other methods that use visual noise, such as band-pass masking (Solomon & Pelli, 1994) and noise masking functions (Pelli & Farell, 1999). Abbey and Eckstein (2007) and Conrey and Gold (2009) measured classification images in white, low-pass, high-pass, and band-pass noise and found that observers' templates were different in different types of noise. The most notable pattern was that in low-pass noise, observers shifted to using higher spatial frequencies, but in high-pass noise they were unable to shift to lower spatial frequencies. Furthermore, Abbey and Eckstein (2009) found that observers' templates changed as a function of noise amplitude: in stronger noise, observers shifted to lower spatial frequencies. Classification image studies and other types of noise-based studies typically assume that noise does not substantially change observers' strategies, but these findings show that it can sometimes have a moderate effect. These findings are consistent with earlier performance-based studies (e.g., Burgess, 1999), but classification images have revealed in greater detail how different types of noise affect observers' templates.

Mareschal, Dakin, and Bex (2006) measured spatiotemporal classification images to examine the time course of observers' decision strategies in a simple orientation discrimination task. Observers discriminated between two Gabor patterns at slightly different orientations, in a 500-ms movie of spatiotemporal white noise. Physiological studies have suggested that the receptive fields of orientation-selective V1 neurons evolve rapidly after stimulus onset (e.g., Ringach, Hawken, & Shapley, 2003), but Mareschal et al. found no evidence of tuning changes in spatial frequency, orientation, or bandwidth over the course of a trial. They did find that the influence of the stimulus on observers' responses was significant only during the interval 20–300 ms after stimulus onset and peaked at around 150 ms after onset. They also confirmed Solomon's (2002) finding that classification images in an orientation discrimination task are predictably different from the ideal observer's classification image: human observers attend to orientations that are more widely separated than the orientations of the two Gabors being discriminated, which is consistent with some theories of multiple spatial frequency channels (Itti, Koch, & Braun, 2000).

Perceptual organization

The fields of spatial vision and perceptual organization both have the goal of understanding shape perception, but historically they have taken very different paths, and until recently they have proceeded in parallel with little communication. In the past two decades, there has been more dialogue, for instance, as spatial vision models have begun to incorporate contour grouping mechanisms (e.g., Elder & Goldberg, 2002; Field, Hayes, & Hess, 1993; Geisler, Perry, Super, & Gallogly, 2001).

Gold et al. (2000) used classification images to study perception of illusory and occluded contours. They used a task where observers judged the direction of curvature of illusory contours and partly occluded contours. They found that even though the stimulus regions where illusory and occluded contours were perceived contained no physical signal that could help observers to give correct responses, classification images nevertheless showed strong correlations between noise contrast in these regions and observers' responses. That is, observers' templates had large weights along illusory and occluded contours, just as they did along luminance-defined contours. This finding suggested that illusory and occluded contours are not simply epiphenomena, but that they play a key role in perception of shape. Ringach and Shapley's (1996) performance-based study had already given behavioral evidence for this idea. Gold et al. provided strong confirmation of Ringach and Shapley's findings and also revealed several novel, idiosyncratic features of observers' shape discrimination strategies, such as using mainly vertical contours, using mainly contours in the left visual field, and exhibiting substantial individual differences.

Several later classification image studies expanded on these findings. Murray (2002) showed that noise along

contours defined purely by grouping, where no illusory or occluded contours are perceived, also has a strong influence on observers' responses, and suggested that Gold et al.'s results primarily reflect grouping, not perceptual completion. Gold and Shubel (2006) measured spatiotemporal classification images in the same task used by Gold et al. and found evidence for a rapid time course for visual completion, lasting around 130 ms, consistent with earlier performance-based studies. Keane, Lu, and Kellman (2007) used spatiotemporal classification images to show that the influence of noise along illusory contours and the time course of completion are largely unchanged even when observers must interpolate illusory contours over both space and time. Nagai, Bennett, and Sekuler (2008) investigated the strong bias to use vertical contours that Gold et al. revealed, and they confirmed this bias but found that observers nevertheless performed about equally well when forced to use horizontal or vertical contours.

Visual search

Visual search requires fast, automatic saccadic targeting guided by stimuli in the periphery, where visual processing is limited by low spatial resolution (Banks, Sekuler, & Anderson, 1991), high internal noise (Pelli & Farell, 1999), intrinsic spatial uncertainty (Michel & Geisler, 2011), imperfect phase encoding (Bennett & Banks, 1987), and crowding (Pelli & Tillman, 2008), in ways still being elucidated. The linear observer model is clearly incomplete under these conditions. Nevertheless, classification image studies have contributed to our understanding of visual search and, in particular, to discovering how the saccadic targeting system gathers information over space and time in order to direct saccades.

Rajashekar, Bovik, and Cormack (2006) and Tavassoli, van der Linde, Bovik, and Cormack (2007), in a pair of closely related studies, used classification images to estimate the template that guides saccades during visual search for a simple visual target, such as a triangle or a circle. Observers' eye movements were tracked while they searched freely for a known target in a large noise field. Rajashekar et al. calculated classification images by averaging the noise in a small square region around each saccade location during search. Tavassoli et al. used a similar procedure but introduced some improvements that made the classification images dramatically less noisy. In both studies, the classification images had roughly the same shapes as the targets, demonstrating that the saccadic targeting system can use shape information with some precision to guide saccades. Eckstein, Beutter, Pham, Shimozaki, and Stone (2007) used similar methods and reached the even stronger conclusion that the template that guides saccades is the same as the template that mediates explicit perceptual decisions (e.g., keypress responses) when judging stimuli in the periphery.

Ludwig, Eckstein, and Beutter (2007) showed that there are limits to how precisely the saccadic template can match the target, however. In their experiments, the target was a Gaussian luminance bump and the distractors were slightly lower contrast but otherwise identical Gaussian luminance bumps. In separate conditions, the Gaussian bumps were small (σ = 0.175°) or large (σ = 0.8°). Classification images showed that the saccadic template was a Gaussian bump with a weak inhibitory surround that was approximately the optimal size in the small-bump condition but much too small in the large-bump condition. Thus, although the template that guides saccades is flexible, it seems to be sufficiently constrained that it cannot even correctly match circular patterns of various sizes.

Caspi et al. (2004) measured temporal classification images to investigate how the saccadic targeting system integrates information over time. Their observers searched for a high-contrast Gaussian bump among lower contrast Gaussian bumps. The classification image noise was temporal noise: all the Gaussian bumps flickered rapidly throughout the trial. Caspi et al. examined the effect of the flicker noise up to the time of the first saccade. They calculated a first-saccade temporal classification image by averaging the noise leading up to the time of the first saccade, at the Gaussian bump that was the location of the observer's first saccade. They found that bright flicker tended to attract the observer's first saccade up to around 100 ms before the saccade. The last 100 ms before the first saccade was "dead time," during which flicker did not influence the location of the first saccade. Caspi et al. also measured a second-saccade temporal classification image, by averaging the noise leading up to the time of the first saccade, at the Gaussian bump that was the location of the observer's second saccade. Interestingly, they found that bright flicker during the 100 ms of dead time before the *first* saccade had a strong influence on the location of the *second* saccade, indicating that the last 100 ms was dead time only for the first saccade and that information continued to be gathered during this time to target the second saccade.

Conclusion

I will conclude with some general observations on what we have learned about classification images and suggestions for avenues for future research.

What have we learned? The linear observer model that underpins the most straightforward interpretation of classification images, which some researchers expressed misgivings about when classification images began to be used in visual psychophysics, has turned out not to put a strong limit on the usefulness of these methods. The

linearity assumption can be tested, and it often turns out to be valid, at least over the very small range of stimuli used in a typical classification image experiment (Abbey & Eckstein, 2002a; Murray et al., 2005). Furthermore, departures from linearity are sometimes unimportant, as when we draw conclusions simply from the fact that a classification image shows strong correlations between a stimulus region of interest and the observer's responses. Finally, there are many ways of modifying the method to incorporate nonlinearities in visual processing, including the general-purpose Volterra and Wiener kernel frameworks and more specific modifications based on models of nonlinearities in visual processing.

In early presentations of the classification image method, researchers (including the present writer) sometimes described classification images as "revealing the observer's strategy," offering a more "direct" view of decision mechanisms than traditional measures like proportion correct, and so on. While it is true that classification images provide a distinct and useful kind of information about visual processing, it is important not to overstate these claims. Our review of methods for studying nonlinear strategies made it clear that classification images do not simply lay bare the observer's decision rule. Different decision rules can produce the same classification image, and a single decision rule can produce different classification images in experiments with different stimuli. Furthermore, our review of substantive research suggests that classification image studies have usually not led to discoveries that were inaccessible, in principle, to other behavioral methods. More often, they have served as a new test of hypotheses that had already been suggested by previous behavioral studies. They have provided an additional source of converging evidence and sometimes have also allowed researchers to estimate the properties of visual processing more precisely than was feasible with other methods.

Are classification images "system identification" tools? Volterra and Wiener kernel methods are system identification tools that provide a way of describing the input–output behavior of a system in terms of a flexible but limited framework (Ljung, 1999; Schetzen, 1980). Classification image methods are similar to kernel methods, and they are sometimes described as system identification tools for psychophysics. This may not be a useful description. As discussed earlier, complete system identification is one aim of current research on early sensory neurons, but in psychophysics such comprehensive goals for classification images are not credible. A classification image does not provide enough information to "identify" the visual system in any useful sense. Instead, it is more appropriate to treat a classification image as a behavioral measure, like a threshold or a median reaction time albeit more complex, that can be used to test the predictions of competing theories of visual processing. Victor (1992) suggests this interpretation for kernel analyses of physiological systems as well. On this view, the accuracy with which a classification image predicts an observer's trial-by-trial responses is generally less important than qualitative features, such as inhibitory surrounds, that may have little impact on predicting performance and yet may be significant for our understanding of visual processing.

Where to from here? One promising line for future work, already begun by some investigators, is to explore the relationship between classification images and mainstream statistical frameworks, such as the general linear model, the generalized linear model, and the generalized additive model. This work may lead to better ways of estimating classification images and new ways of broadening the linear observer model. It should also allow psychophysicists to draw on the findings of professional statisticians and reduce the need to develop ad hoc methods.

There is a need for more conclusive studies on which methods of calculating classification images are least noisy while remaining reasonably unbiased. I have mentioned several studies on this topic, but their approaches were so diverse that they are difficult to compare with confidence: some used rating responses, and some used yes–no responses; some had many more trials than parameters, and some did not; some used various types of priors, and some used none; some used human data, and some used simulated data. It may be that no one method will be best in all circumstances, but nevertheless further studies should be able to give us a better idea than we have at present of which methods work well under which conditions.

We have seen that classification images based on the linear observer model can be estimated either using weighted sums of noise fields, which is an approach closely related to kernel methods, or using maximum likelihood estimates based on the generalized linear model and its extensions. Each approach has its advantages. Classification images based on weighted sums are linear functions of the stimulus noise, so they are conceptually straightforward and their statistical properties are generally easy to work out. They are also computationally undemanding, which was no doubt one appeal of correlation-based kernel methods when they were developed in the mid-1960s. Classification images based on maximum likelihood estimates are statistically less transparent and computationally more intensive, but because they are based on an explicit likelihood function, they are highly flexible. They can easily be modified to incorporate priors. More significantly, they can also be modified to incorporate nonlinearities in visual processing. In principle, it is straightforward to write down a probabilistic model of visual processing and use observers' responses to noisy stimuli to make a maximum likelihood estimate of its parameters. In practice, difficulties abound, including the problems of finding the global maximum of the likelihood function instead of local maxima, showing that the available data ensure a single global maximum, estimating bias and variability, and testing the model for robustness

against departures from its assumptions. Despite these obstacles, recent progress suggests that using observers' responses to noisy stimuli in order to constrain probabilistic, nonlinear models of visual processing is a promising direction for future work on classification images.

Appendix A

Notation

Conventions: Vectors are in bold. Matrices are in upper case. Samples of random variables are subscripted. Superscripts correspond to alternative signals. Images are represented as column vectors.

The experiment	
m	number of signals and response alternatives
k	signal number (1, …, m)
$\mathbf{g}$	stimulus
$\mathbf{s}^i$	signal
$\mathbf{n}$	external noise
r	response number (1, …, m)
The linear observer model	
$\mathbf{t}^i$	template
$\mathbf{w}$, $\mathbf{w}^i$	template difference, $\mathbf{t}^a - \mathbf{t}^b$
z, z^i	internal noise
d, d^i	decision variable
a, a^i	response criterion
The classification image	
$\mathbf{c}$, $\mathbf{c}^i$	classification image
$\overline{\mathbf{n}}^{ij}$	sample average of external noise images over trials where the signal was $\mathbf{s}^i$ and the observer gave response $r = j$
$\overline{\mathbf{n}}^{*j}$	sample average of external noise images over trials where the observer gave response $r = j$
General	
$E[x]$	expected value of random variable x
corr$[x, y]$	Pearson correlation between random variables x and y
$\Phi(x, \mu, \sigma)$	normal cumulative distribution function
$\mathbf{v}^T$, A^T	transpose of vector $\mathbf{v}$ or matrix A

Appendix B

Multiple response alternatives

Simulation. The simulation used four 5 × 9 pixel signals (Figure 3, first column). The background pixels had value 0.0, and the foreground pixels had value 1.0. On each trial, the stimulus was one of the four signals, chosen randomly, in Gaussian white noise with mean 0.0 and standard deviation 1.0. The model observer identified the stimulus by taking its dot product with four templates (Figure 3, second column) and choosing the template with the largest dot product. The model observer had no internal noise. (Further simulations with an internal-to-external noise ratio of 1.0 gave similar results, though with more sampling noise evident in the classification images.) The simulation ran for 10,000 trials. The model observer gave 72% correct responses and was unbiased.

Logistic regression. One set of classification images was calculated using multinomial logistic regression (Figure 3, third column). The dependent variable was the observer's response number (1 through 4). The covariates were the 45 noise pixel values and three dummy variables that encoded the signal number. Signal 1 was encoded as (0, 0, 0), signal 2 as (1, 0, 0), signal 3 as (0, 1, 0), and signal 4 as (0, 0, 1). Thus, the covariate vectors had 48 components. The regression coefficient vector for template 1 was taken as the reference, so each classification image shows an estimate of $\mathbf{t}^i - \mathbf{t}^1$. The regression coefficients were estimated using the mnrfit multinomial logistic regression function in the MATLAB Statistics Toolbox (The MathWorks, Natick, MA).

Correlation method. A second set of classification images was calculated using Dai and Micheyl's (2010) correlation method (Figure 3, fourth column). Trials were divided into four groups according to which signal was shown. The classification image for each group was calculated as the pixelwise correlation between the 45 noise pixel values and the correctness of the model observer's responses, coded as correct = 1 and incorrect = 0.

Each classification image in Figure 3 has been affinely transformed so that its average value is mid-gray (0.5) and it occupies as much of the black–white range as possible, i.e., either its minimum value is darkest black (0.0) or its maximum value is brightest white (1.0).

Acknowledgments

I thank Peter Neri and Jonathan Victor for discussions about kernel methods, Carl Gaspar for bringing Abel and Quick (1978) to my attention, Virginia Richards for information on related methods in auditory research, Jack Gallant for references to methods in fMRI research, Huanping Dai, Miguel Eckstein, Patrick Mineault, and Umesh Rajashekar for clarifications of their work, and Yaniv Morgenstern, Minjung Kim, Christopher Taylor, Al Ahumada, Craig Abbey, and an anonymous reviewer for comments on drafts of this article. This work was funded by the Canada Foundation for Innovation, the Natural Sciences and Engineering Research Council of Canada, and a Faculty of Arts Research Grant from York University.

Commercial relationships: none.
Corresponding author: Richard F. Murray.
Email: rfm@yorku.ca.
Address: Centre for Vision Research, York University, 4700 Keele Street, CSE 0009, Toronto, Ontario M3J 1P3, Canada.

Footnotes

[1]Equation 7 can also be seen as an instance of Fisher's linear discriminant for classifying noise fields according to the observer's responses, $r = 1$ or $r = 2$ (Duda et al., 2000, p. 120, Equation 106; Fisher, 1936).

[2]In practice, it may be better to use just the noise pixels **n** as covariates and to use dummy variables to encode the signals as levels of a factor, instead of using the pixels of the full stimulus $\mathbf{g} = \mathbf{s}^k + \mathbf{n}$ as covariates (Knoblauch & Maloney, 2008; also see Appendix B). One reason for doing so is that if the signals do not appear in the covariates, then any signal-like patterns in the classification image must reflect the observer's decision strategy. If the signals appear in the covariates, then signal-like patterns in the classification image could be artifacts of the estimation procedure (e.g., the result of bias).

[3]Knoblauch and Maloney's simulations leave some room for doubt, because (a) they gave their model observer a very liberal response criterion (equal to the mean of the decision variable on signal-absent trials, so $c' = -0.5$ (Macmillan & Creelman, 2005)), incorporated the exact value of the criterion into the model used to make the GLM estimate, and used the weighted sum estimate that is appropriate for unbiased responses, thus giving a potential advantage to the GLM estimate; and (b) they held signal contrast constant as they varied internal noise strength, so increases in internal noise were confounded with decreases in performance. I have repeated their simulations with an unbiased model observer at a constant performance level, and I found very similar results.

[4]Mineault et al. called their representation a Laplacian pyramid (Burt & Adelson, 1983), but since they applied the shrinkage prior to the coefficients of Gaussian basis functions, not difference-of-Gaussian basis functions, it may be more appropriate to call it a Gaussian pyramid (Mineault, personal communication).

[5]Other ways of removing the redundancy are also possible, e.g., we could estimate m classification images and require that they sum to zero.

References

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

Abbey, C. K., & Eckstein, M. P. (2002a). Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments. *Journal of Vision, 2*(1):5, 66–78, http://www.journalofvision.org/content/2/1/5, doi:10.1167/2.1.5. [PubMed] [Article]

Abbey, C. K., & Eckstein, M. P. (2002b). Optimal shifted estimates of human-observer templates in two-alternative forced-choice experiments. *IEEE Transactions on Medical Imaging, 21,* 429–440.

Abbey, C. K., & Eckstein, M. P. (2007). Classification images for simple detection and discrimination tasks in correlated noise. *Journal of the Optical Society of America A, 24,* B110–B124.

Abbey, C. K., & Eckstein, M. P. (2009). Frequency tuning of perceptual templates changes with noise magnitude. *Journal of the Optical Society of America A, 26,* B72–B83.

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

Abel, L. A., & Quick, R. F., Jr. (1978). Wiener analysis of grating contrast judgments. *Vision Research, 18,* 1031–1039.

Ahumada, A. J., Jr. (1996). Perceptual classification images from Vernier acuity masked by noise [Abstract]. *Perception, 25,* ECVP Abstract Supplement.

Ahumada, A. J., Jr. (2002). Classification image weights and internal noise level estimation. *Journal of Vision, 2*(1):8, 121–131, http://www.journalofvision.org/content/2/1/8, doi:10.1167/2.1.8. [PubMed] [Article]

Ahumada, A. J., Jr., & Beard, B. L. (1999). Classification images for detection [Abstract]. *Investigative Ophthalmology and Visual Science, 40,* S572.

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

Ahumada, A. J., Jr., Marken, R., & Sandusky, A. (1975). Time and frequency analyses of auditory signal detection. *Journal of the Acoustical Society of America, 57,* 385–390.

Alexander, J. M., & Lutfi, R. A. (2004). Informational masking in hearing-impaired and normal-hearing listeners: Sensation level and decision weights. *Journal of the Acoustical Society of America, 116,* 2234–2247.

Banks, M. S., Sekuler, A. B., & Anderson, S. J. (1991). Peripheral spatial vision: Limits imposed by optics, photoreceptors, and receptor pooling. *Journal of the Optical Society of America A, 8,* 1775–1787.

Beard, B. L., & Ahumada, A. J., Jr. (1997). Relevant image features for Vernier acuity [Abstract]. *Perception, 26,* ECVP Abstract Supplement.

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

Bennett, P. J., & Banks, M. S. (1987). Sensitivity loss in odd-symmetric mechanisms and phase anomalies in peripheral vision. *Nature, 326,* 873–876.

Berg, B. G. (1990). Observer efficiency and weights in a multiple observation task. *Journal of the Acoustical Society of America, 88,* 149–158.

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

Burgess, A. E. (1999). Visual detection with two-component noise: Low-pass spectrum effects. *Journal of the Optical Society of America A,* 694–704.

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

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

Burt, P. J., & Adelson, E. H. (1983). The Laplacian Pyramid as a compact image code. *IEEE Transactions on Communications, 31,* 532–540.

Caspi, A., Beutter, B. R., & Eckstein, M. P. (2004). The time course of visual information accrual guiding eye movement decisions. *Proceedings of the National Academy of Sciences of the United States of America, 101,* 13086–13090.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of Vision, 5*(9):1, 659–667, http://www.journalofvision.org/content/5/9/1, doi:10.1167/5.9.1. [PubMed] [Article]

Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Network, 12,* 199–213.

Cohen, A. L., Shiffrin, R. M., Gold, J. M., Ross, D. A., & Ross, M. G. (2007). Inducing features from visual noise. *Journal of Vision, 7*(8):15, 1–14, http://www.journalofvision.org/content/7/8/15, doi:10.1167/7.8.15. [PubMed] [Article]

Conrey, B., & Gold, J. M. (2009). Pattern recognition in correlated and uncorrelated noise. *Journal of the Optical Society of America A, 26,* B94–B109.

Dai, H., & Micheyl, C. (2010). Psychophysical reverse correlation with multiple response alternatives. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 976–993.

David, S. V., Vinje, W. E., & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of V1 neurons. *Journal of Neuroscience, 24,* 6991–7006.

de Boer, E., & Kuyper, P. (1968). Triggered correlation. *IEEE Transactions on Biomedical Engineering, 15,* 169–179.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3,* 186–205.

Dobres, J., & Seitz, A. R. (2010). Perceptual learning of oriented gratings as revealed by classification images. *Journal of Vision, 10*(13):8, 1–11, http://www.journalofvision.org/content/10/13/8, doi:10.1167/10.13.8. [PubMed] [Article]

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). New York: John Wiley & Sons.

Eckstein, M. P., Beutter, B. R., Pham, B. T., Shimozaki, S. S., & Stone, L. S. (2007). Similar neural representations of the target for saccades and perception during search. *The Journal of Neuroscience, 27,* 1266–1270.

Egan, J. P., Schulman, A. I., & Greenberg, G. Z. (1959). Operating characteristics determined by binary decisions and by ratings. *Journal of the Acoustical Society of America, 31,* 768–773.

Elder, J. H., & Goldberg, R. M. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision, 2*(4):5, 324–353, http://www.journalofvision.org/content/2/4/5, doi:10.1167/2.4.5. [PubMed] [Article]

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

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7,* 179–188.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review, 96,* 267–314.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41,* 711–724.

Gold, J. M., Cohen, A. L., & Shiffrin, R. (2006). Visual noise reveals category representations. *Psychonomic Bulletin & Review, 13,* 649–655.

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

Gold, J. M., & Shubel, E. (2006). The spatiotemporal properties of visual completion measured by response classification. *Journal of Vision, 6*(4):5, 356–365, http://www.journalofvision.org/content/6/4/5, doi:10.1167/6.4.5. [PubMed] [Article]

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research, 41,* 2261–2271.

Gosselin, F., & Schyns, P. G. (2002). RAP: A new framework for visual categorization. *Trends in Cognitive Sciences, 6,* 70–77.

Gosselin, F., & Schyns, P. G. (2004). No troubles with bubbles: A reply to Murray and Gold. *Vision Research, 44,* 471–477.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics*. Huntington, NY: R. E. Krieger Publishing. (Original work published 1966)

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Horwitz, G. D., Chichilnisky, E. J., & Albright, T. A. (2007). Cone inputs to simple and complex cells in V1 of awake macaque. *Journal of Neurophysiology, 97,* 3070–3081.

Itti, L., Koch, C., & Braun, J. (2000). Revisiting spatial vision: Toward a unifying model. *Journal of the Optical Society of America A, 17,* 1899–1917.

Jolliffe, I. T. (2010). *Principal component analysis* (2nd ed.). New York: Springer.

Keane, B. P., Lu, H., & Kellman, P. J. (2007). Classification images reveal spatiotemporal contour interpolation. *Vision Research, 47,* 3460–3475.

Knoblauch, K., & Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision, 8*(16):10, 1–19, http://www.journalofvision.org/content/8/16/10, doi:10.1167/8.16.10. [PubMed] [Article]

Lasley, D. J., & Cohn, T. E. (1981). Why luminance discrimination may be better than detection. *Vision Research, 21,* 273–278.

Lee, Y. W., & Schetzen, M. (1965). Measurement of the Wiener kernels of a non-linear system by cross-correlation. *International Journal of Control, 2,* 237–254.

Levi, D. M., & Klein, S. A. (2002). Classification images for detection and position discrimination in the fovea and parafovea. *Journal of Vision, 2*(1):4, 46–65, http://www.journalofvision.org/content/2/1/4, doi:10.1167/2.1.4. [PubMed] [Article]

Li, R. W., Klein, S. A., & Levi, D. M. (2006). The receptive field and internal noise for position acuity change with feature separation. *Journal of Vision, 6*(4):2, 311–321, http://www.journalofvision.org/content/6/4/2, doi:10.1167/6.4.2. [PubMed] [Article]

Ljung, L. (1999). *System identification: Theory for the user* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Logvinenko, A. D. (1990). Nonlinear analysis of spatial vision using first- and second-order Volterra transfer functions measurement. *Vision Research, 30,* 2031–2057.

Ludwig, C. J. H., Eckstein, M. P., & Beutter, B. R. (2007). Limited flexibility in the filter underlying saccadic targeting. *Vision Research, 47,* 280–288.

Ludwig, C. J. H., Gilchrist, I. D., McSorley, E., & Baddeley, R. J. (2005). The temporal impulse response underlying saccadic decisions. *The Journal of Neuroscience, 25,* 9907–9912.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Mareschal, I., Dakin, S. C., & Bex, P. J. (2006). Dynamic properties of orientation discrimination assessed by using classification images. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 5131–5136.

Marmarelis, P. Z., & Marmarelis, V. Z. (1977). *Analysis of physiological systems: The white-noise approach*. New York: Plenum Press.

Marmarelis, P. Z., & Naka, K.-I. (1972). White-noise analysis of a neuron chain: An application of the Wiener theory. *Science, 175,* 1276–1278.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Michel, M., & Geisler, W. S. (2011). Intrinsic position uncertainty explains detection and localization performance in the periphery. *Journal of Vision, 11*(1):18, 1–18, http://www.journalofvision.org/content/11/1/18, doi:10.1167/11.1.18. [PubMed] [Article]

Mineault, P. J., Barthelmé, S., & Pack, C. C. (2009). Improved classification images with sparse priors in a smooth basis. *Journal of Vision, 9*(10):17, 1–24, http://www.journalofvision.org/content/9/10/17, doi:10.1167/9.10.17. [PubMed] [Article]

Murray, R. F. (2002). *Perceptual organization and the efficiency of shape discrimination*. Ph.D. thesis, University of Toronto.

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: Weighted sums. *Journal of Vision, 2*(1):6, 79–104, http://www.journalofvision.org/content/2/1/6, doi:10.1167/2.1.6. [PubMed] [Article]

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2005). Classification images predict absolute efficiency.

Journal of Vision, 5(2):5, 139–149, http://www.journalofvision.org/content/5/2/5, doi:10.1167/5.2.5. [PubMed] [Article]

Murray, R. F., & Gold, J. M. (2004a). Reply to Gosselin and Schyns. *Vision Research, 44,* 479–482.

Murray, R. F., & Gold, J. M. (2004b). Troubles with bubbles. *Vision Research, 44,* 461–470.

Nagai, M., Bennett, P. J., & Sekuler, A. B. (2008). Exploration of vertical bias in perceptual completion of illusory contours: Threshold measures and response classification. *Journal of Vision, 8*(7):25, 1–17, http://www.journalofvision.org/content/8/7/25, doi:10.1167/8.7.25. [PubMed] [Article]

Nandy, A. S., & Tjan, B. S. (2007). The nature of letter crowding as revealed by first- and second-order classification images. *Journal of Vision, 7*(2):5, 1–26, http://www.journalofvision.org/content/7/2/5, doi:10.1167/7.2.5. [PubMed] [Article]

Neri, P. (2004). Estimation of nonlinear psychophysical kernels. *Journal of Vision, 4*(2):2, 82–91, http://www.journalofvision.org/content/4/2/2, doi:10.1167/4.2.2. [PubMed] [Article]

Neri, P. (2009). Nonlinear characterization of a simple process in human vision. *Journal of Vision, 9*(12):1, 1–29, http://www.journalofvision.org/content/9/12/1, doi:10.1167/9.12.1. [PubMed] [Article]

Neri, P. (2010a). How inherently noisy is human sensory processing? *Psychonomic Bulletin & Review, 17,* 802–808.

Neri, P. (2010b). Stochastic characterization of small-scale algorithms for human sensory processing. *Chaos, 20,* 1–19.

Neri, P. (2010c). Visual detection under uncertainty operates via an early static, not late dynamic, non-linearity. *Frontiers in Computational Neuroscience, 4,* 1–17.

Neri, P., & Heeger, D. J. (2002). Spatiotemporal mechanisms for detecting and identifying image features in human vision. *Nature Neuroscience, 5,* 812–816.

Neri, P., & Levi, D. M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research, 46,* 2465–2474.

Neri, P., & Parker, A. J. (1999). Probing the human stereoscopic system with reverse correlation. *Nature, 401,* 695–698.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A, 2,* 1508–1532.

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research, 46,* 4646–4674.

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

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience, 11,* 1129–1135.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory, 4,* 171–212.

Pinter, R. B., & Nabet, B. (1992). *Nonlinear vision: Determination of neural receptive fields, function, and networks.* Boca Raton, FL: CRC Press.

Prenger, R., Wu, M. C.-K., David, S. V., & Gallant, J. L. (2004). Nonlinear V1 responses to natural scenes revealed by neural network analysis. *Neural Networks, 17,* 663–679.

R Development Core Team (2010). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rajashekar, U., Bovik, A. C., & Cormack, L. K. (2006). Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis. *Journal of Vision, 6*(4):7, 379–386, http://www.journalofvision.org/content/6/4/7, doi:10.1167/6.4.7. [PubMed] [Article]

Rajashekar, U., Cormack, L. K., & Bovik, A. C. (2002). Visual search: Structure from noise. In A. T. Duchowski (Ed.), *Proceedings of the 2002 Symposium on Eye Tracking Research and Applications* (pp. 119–123). New York: Association for Computing Machinery.

Richards, V. M., & Zhu, S. (1994). Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients. *Journal of the Acoustical Society of America, 95,* 423–434.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: The MIT Press.

Ringach, D. L., Hawken, M. J., & Shapley, R. (2003). Dynamics of orientation tuning in macaque V1: The role of global and tuned suppression. *Journal of Neurophysiology, 90,* 342–352.

Ringach, D. L., & Shapley, R. (1996). Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision Research, 36,* 3037–3050.

Sakai, H. M. (1992). White-noise analysis in neurophysiology. *Physiological Reviews, 72,* 491–505.

Schetzen, M. (1980). *The Volterra and Wiener theories of nonlinear systems*. New York: John Wiley & Sons.

Solomon, J. A. (2002). Noise reveals visual mechanisms of detection and discrimination. *Journal of Vision,*

2(1):7, 105–120, http://www.journalofvision.org/content/2/1/7, doi:10.1167/2.1.7. [PubMed] [Article]

Solomon, J. A., & Pelli, D. G. (1994). The visual filter mediating letter identification. *Nature, 369,* 395–397.

Stark, L. (1969). The pupillary control system: Its nonlinear adaptive and stochastic engineering design characteristics. *Automatica, 5,* 655–676.

Sutter, E. E. (1987). A practical non-stochastic approach to nonlinear time-domain analysis. In V. Z. Marmarelis (Ed.), *Advanced methods of physiological systems modeling* (vol. 1, pp. 303–315). Los Angeles: Biomedical Simulations Resource.

Tang, Z., & Richards, V. M. (2005). Comparing linear regression models applied to psychophysical data [Abstract]. *Journal of the Acoustical Society of America, 117,* 2597.

Tanner, W. P., Jr. (1961). Physiological implications of psychophysical data. *Annals of the New York Academy of Sciences, 89,* 752–765.

Tavassoli, A., van der Linde, I., Bovik, A. C., & Cormack, L. K. (2007). An efficient technique for revealing visual search strategies with classification images. *Perception & Psychophysics, 69,* 103–112.

Taylor, C. P., Bennett, P. J., & Sekuler, A. B. (2009). Spatial frequency summation in noise. *Journal of the Optical Society of America A, 26,* B84–B93.

Tjan, B. S., & Nandy, A. S. (2006). Classification images with uncertainty. *Journal of Vision, 6*(4):8, 387–413, http://www.journalofvision.org/content/6/4/8, doi:10.1167/6.4.8. [PubMed] [Article]

Van Trees, H. L. (2001). *Detection, estimation, and modulation theory: Part I*. New York: John Wiley & Sons. (Original work published 1968)

Victor, J. D. (1992). Nonlinear systems analysis in vision: Overview of kernel methods. In R. B. Pinter & B. Nabet (Eds.), *Nonlinear vision: Determination of neural receptive fields, function, and networks* (pp. 1–37). Boca Raton, FL: CRC Press.

Victor, J. D. (2005). Analyzing receptive fields, classification images and functional images: Challenges with opportunities for synergy. *Nature Neuroscience, 8,* 1651–1656.

Volterra, V. (1930). *Theory of functionals and of integral and integrodifferential equations*. London: Blakie.

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

Westheimer, G. (1979). The spatial sense of the eye. *Investigative Ophthalmology and Visual Science, 18,* 893–912.

Wiener, N. (1958). *Nonlinear problems in random theory*. New York: John Wiley & Sons.

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience, 29,* 477–505.

Yamashita, O., Sato, M., Yoshioka, T., Tong, F., & Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage, 42,*1414–1429.

Zeevi, Y. Y., & Mangoubi, S. S. (1984). Vernier acuity with noisy lines: Estimation of relative position uncertainty. *Biological Cybernetics, 50,* 371–376.