LETTER Communicated by Richard Zemel

The Costs of Ignoring High-Order Correlations in Populations of Model Neurons

Melchi M. Michel
mmichel@cvs.rochester.edu
Robert A. Jacobs
robbie@bcs.rochester.edu
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, U.S.A.

Investigators debate the extent to which neural populations use pairwise and higher-order statistical dependencies among neural responses to represent information about a visual stimulus. To study this issue, three statistical decoders were used to extract the information in the responses of model neurons about the binocular disparities present in simulated pairs of left-eye and right-eye images: (1) the full joint probability decoder considered all possible statistical relations among neural responses as potentially important; (2) the dependence tree decoder also considered all possible relations as potentially important, but it approximated high-order statistical correlations using a computationally tractable procedure; and (3) the independent response decoder, which assumed that neural responses are statistically independent, meaning that all correlations should be zero and thus can be ignored. Simulation results indicate that high-order correlations among model neuron responses contain significant information about binocular disparities and that the amount of this high-order information increases rapidly as a function of neural population size. Furthermore, the results highlight the potential importance of the dependence tree decoder to neuroscientists as a powerful but still practical way of approximating high-order correlations among neural responses.

1 Introduction

The left and right eyes of human observers are offset from each other, and, thus, the visual images received by these eyes differ. For example, an object in the visual environment may project to one location in the left eye image but project to a different location in the right eye image. Differences in left eye and right eye images that arise in this manner are known as *binocular disparities*. Disparities are important because they are often among the most reliable cues to the relative depth of a surface or object in space. Observers with normal stereo vision are typically able to

make fine depth discriminations because they can resolve differences in horizontal disparities below 1 arc minute (Andrews, Glennerster, & Parker, 2001). How this is accomplished is a matter of current research.

Neurophysiological and modeling studies have identified binocular simple and complex cells in primary visual cortex as a likely source of disparity information, and researchers have developed a computational model known as a binocular energy filter to characterize the responses of these cells to visual scenes viewed binocularly (DeAngelis, Ohzawa, & Freeman, 1991; Freeman & Ohzawa, 1990; Ohzawa, DeAngelis, & Freeman, 1990). Based on analyses of binocular energy filters, Qian (1994), Fleet, Wagner, and Heeger (1996), and others have argued, however, that the response of an individual simple or complex cell is ambiguous. In addition to uncertainty introduced by neural noise, ambiguities arise because a cell's preferred disparity depends on the distribution of stimulus frequencies, a cell's tuning response has multiple false peaks (i.e., the cell gives large responses to disparities that differ from its preferred disparity), and image features in a cell's left eye and right eye receptive fields may influence a cell's response even when the features do not arise from the same event in the visual world. These points suggest that in order to overcome the ambiguity of an individual neuron's responses, the neural process responsible for estimating disparity must pool the responses of a large number of neurons.

Researchers studying neural codes often use statistical techniques to interpret the activities of neural populations (Abbott & Dayan, 1999; Oram, Földiàk, Perrett, & Sengpiel, 1998; Pouget, Dayan, & Zemel, 2003). A matter of current debate among these investigators is the relative importance of considering dependencies, or correlations, among cells in a population when decoding the information that the cells convey about a stimulus. Correlations among neural responses have been investigated as a potentially important component of neural codes for over 30 years (Perkel & Bullock, 1969). Unfortunately, determining the importance of correlations is not straightforward. For methodological reasons, it is typically feasible only to experimentally measure pairwise or second-order correlations among neural responses, meaning that high-order correlations are not measured. Even if correlations are accurately measured, there is no guarantee that these correlations contain useful information: correlations can increase, decrease, or leave unchanged the total information in a neural population (Abbott & Dayan, 1999; Nirenberg & Latham, 2003; Seriès, Latham, & Pouget, 2004). To evaluate the importance of correlations, researchers have often compared the outputs of statistically efficient neural decoders, based on maximum likelihood or Bayesian statistical theory, that make different assumptions regarding correlations. Neural decoders are not models of neural mechanisms, but rather statistical procedures that help determine how much information neural responses contain about a stimulus by expressing this information as a probability distribution (Abbott & Dayan, 1999; Oram et al., 1998; Pouget et al., 2003). Statistically efficient neural decoders are

useful because they provide an upper bound on the amount of information about a stimulus contained in the activity of a neural ensemble. Researchers can evaluate the importance of correlations by comparing the value of this bound when it is computed by a neural decoder that makes use of correlations with the value of this bound when it is computed by a decoder that does not. Alternatively, researchers can compare the performances of neural decoders that use or do not use correlations on a stimulus-relevant task.

Several recent studies have suggested that correlations among neurons play only a minor role in encoding stimulus information (e.g., Averbeck & Lee, 2003; Golledge et al., 2003; Nirenberg, Carcieri, Jacobs, & Latham, 2001; Panzeri, Schultz, Treves, & Rolls, 1999; Rolls, Franco, Aggelopoulos, & Reece, 2003), and that the independent responses of neurons carry more than 90% of the total information available in the population response (Averbeck & Lee, 2004). An important limitation of these studies is that they considered only pairwise or second-order correlations among neural responses and thus ignored high-order correlations either by assuming multivariate gaussian noise distributions (e.g., Averbeck & Lee, 2003) or by using a short-time scale approximation to the joint distribution of responses and stimuli (e.g., Panzeri et al., 1999; Rolls et al., 2003). These studies therefore did not fairly evaluate the information contained in the response of a neural population when correlations are considered versus when they are ignored. In a population of n neurons, there are on the order of n^p pth-order statistical interactions among neural response variables. In other words, computing high-order correlations is typically not computationally feasible with current computers. This does not mean, of course, that the nervous system does not make use of high-order correlations or that researchers who fail to consider high-order correlations are justified in concluding that nearly all the information in a neural code is carried by the independent responses of the neurons comprising the population. What is needed is a computationally tractable method for estimating high-order statistics, even if this is done in only an approximate way.

This letter addresses these issues through the use of computer simulations of model neurons, known as binocular energy filters, whose binocular sensitivities resemble those of simple and complex cells in primary visual cortex. The responses of the model neurons to binocular views of visual scenes of frontoparallel surfaces were computed. These responses were then decoded in order to measure how much information they carry about the binocular disparities in the left eye and right eye images. Three neural decoders were simulated. The first decoder, referred to as the full joint probability decoder (FJPD), did not make any assumptions regarding statistical correlations. Because it considered all possible combinations of neural responses, it is the gold standard to which all other decoders were compared. The second decoder, known as the dependence tree decoder (DTD), is similar to the FJPD in the sense that it regarded all correlations as potentially important. However, it used a computationally tractable method to estimate

high-order statistics, albeit in an approximate way (Chow & Liu, 1968; Meilă & Jordan, 2000). The final decoder, referred to as the independent response decoder (IRD), assumed that neural responses are statistically independent, meaning that all correlations should be zero and thus can be ignored. Via computer simulation, we measured the percentage of information that is lost in a population of disparity tuned cells when high-order correlations are approximated and when all correlations are ignored. We also examined the abilities of the DTD and IRD (and a decoder limited to second-order correlations) to correctly estimate the disparity of a frontoparallel surface.

The results reveal several interesting findings. First, relative to the amount of information about disparity calculated by the FJPD, the amounts of information calculated by the IRD and DTD were proportionally smaller when more model neurons were used. In other words, the informational cost of ignoring correlations or of roughly approximating high-order correlations increased as a function of neural population size. This implies that there is a large amount of information about disparity conveyed by second-order and high-order correlations among model neuron responses. Second, the informational cost of ignoring all correlations (as in the IRD) rose as the number of neural response levels increased. For example, relative to the amount of information calculated by the FJPD, the amount of information calculated by the IRD was smaller when neuron responses were discretized to four levels (2 bits of information about each neural response) than when they are discretized to eight levels (3 bits of information about a neural response). This trend was less evident for the DTD. Third, when used to estimate the disparity in a pair of left eye and right eye images, the DTD consistently outperformed the IRD, and the magnitude of its performance advantage increased rapidly as the neural population size increased and as the number of response levels increased. Because the DTD also outperformed a neural decoder based on a multivariate gaussian distribution, our data again indicate that high-order correlations among model neuron responses contain significant information about binocular disparities.

These results have important implications for researchers studying neural codes. They suggest that earlier studies indicating that independent neural responses carry the vast majority of information conveyed by a neural population may be flawed because these studies limited their investigations to second-order correlations and thus did not examine high-order correlations. Furthermore, these results highlight the potential importance of the DTD to neuroscientists. This decoder uses a technique developed in the engineering literature (Chow & Liu, 1968; Meilă & Jordan, 2000), but seemingly unknown in the neuroscientific literature, to approximate high-order statistics. Significantly, it does so in a way that is computationally tractable—the calculation of the approximation requires only knowledge about pairs of neurons. This fact, in the context of the results summarized above, suggests that the DTD can replace the IRD as a better, but still practical, approximation to the information contained in a neural population.

2 Simulated Images

The simulated images were created in a manner similar to the method used by Lippert & Wagner (2002), with the difference that the texture elements used by those authors were random black and white dots, whereas the elements that we used were white noise (luminances were real-valued as in Tsai & Victor, 2003). Each image depicted a one-dimensional frontoparallel surface on which were painted dots whose luminance values were chosen from a uniform distribution to take values between 0 (dark) and 1 (light). A virtual observer who maintained fixation at a constant depth and horizontal position in the scene viewed the surface as its depth was varied between 15 possible depth values relative to the fixation point. One of these depth values was the depth of the fixation plane; of the remaining depths, 7 were located farther than the fixation point from the observer, and 7 were located nearer the observer.

Each image of a scene extended over 5 degrees of visual angle and was divided into 186 pixels per degree. Because each pixel's luminance value was chosen randomly from a uniform distribution, an image contained approximately equal power at all frequencies between 0 cycles per degree and 93 cycles per degree (the Nyquist frequency). For each stereo pair, the left image was generated first; then the right image was created by shifting the left image to the right by a particular number of pixels (this was done with periodic borders; e.g., pixel values that shifted past the right border were assigned to pixels near the left border). This shift varied between –7 and 7 pixels so that the shift was negative when the surface was nearer the observer, zero when the surface was located at the fixation plane, and positive when the surface was located beyond fixation.

3 Model Neurons

Model neurons were instances of binocular energy filters, which are computational models developed by Ohzawa et al. (1990). We used binocular energy filters because they provide a good approximation to the binocular sensitivities of simple and complex cells in primary visual cortex. The fidelity of the energy model with respect to the responses of binocular simple and complex cells has been demonstrated in both cat area 17 (Anzai, Ohzawa, & Freeman, 1997; Ohzawa et al., 1990; Ohzawa, DeAngelis, & Freeman, 1997) and in macaque V1 (Cumming & Parker, 1997; Perez, Castro, Justo, Bermudez, & Gonzalez, 2005; Prince, Pointon, Cumming, & Parker, 2002). Although modifications and extensions to the model have been proposed by different researchers (e.g., Fleet et al., 1996; Qian & Zhu, 1997; Read & Cumming, 2003; Tsai & Victor, 2003), the basic form of the energy model remains a widely accepted representation of simple and complex cell responses to binocular stimuli. A simple cell is modeled as comprising left eye and right eye receptive subfields. Each subfield is modeled as a

Gabor function, which is a sinusoid multiplied by a gaussian envelope. We used the phase-shift version of the binocular energy model, meaning that the retinal positions of the gaussian envelopes for the left eye and right eye Gabor functions are identical, though the sinusoidal components differ by a phase shift. Formally, the left (g_l) and right (g_r) simple cell subfields are expressed as the following Gabor functions:

$$g_l = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-x^2/2\sigma^2)} \sin(2\pi\omega x + \phi) \tag{3.1}$$

$$g_r = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-x^2/2\sigma^2)} \sin(2\pi\omega x + \phi + \delta\phi), \tag{3.2}$$

where x is the distance to the center of the gaussian, the variance σ^2 specifies the width of the gaussian envelope, ω represents the frequency of the sinusoid, ϕ represents the base phase of the sinusoid, and $\delta\phi$ represents the phase shift between the sinusoids in the right and left subfields. The response of a simple cell is formed in two stages: first, the convolution of the left eye image with the left subunit Gabor is added to the convolution of the right eye image with the right subunit Gabor; next, this sum is rectified. The response of a complex cell is the sum of the squared outputs of two simple cells whose parameter values are identical except that one has a base phase of 0 and the other has a base phase of $\pi/2$.[1]

In our simulations, the gaussian envelopes for all neurons were centered at the same point in the visual scene. The parameter values that we used in our simulations were randomly sampled from the same distributions used by Lippert and Wagner (2002); these investigators picked distributions based on neurophysiological data regarding spatial frequency selectivities of neurons in macaque visual cortex. Preferred spatial frequencies were drawn from a log-normal distribution whose underlying normal distribution had a mean of 1.6 cycles per degree and a standard deviation of 0.7 cycle per degree. The range of these preferred frequencies was clipped at a ceiling value of 20 cycles per degree and a floor value of 0.4 cycle per degree. The simple cells' receptive field sizes were sampled from a normal distribution with a mean of 0.5 period and a standard deviation of 0.25 period, with a floor value of 0.1 period. A cell's preferred disparity, given by $2\pi\delta\phi/\omega$, was sampled from a normal distribution with a mean of 0 degrees of visual angle and a standard deviation of 0.5 degree.

Figure 1 shows the normalized responses of a typical model complex cell to three different scenes, each using a different white noise pattern to cover the frontoparallel surface. Each of the lines in the figure represents the

[1] Note that binocular energy filters are deterministic. The probability distributions we use have nonzero variances because the white noise visual stimuli are stochastic.

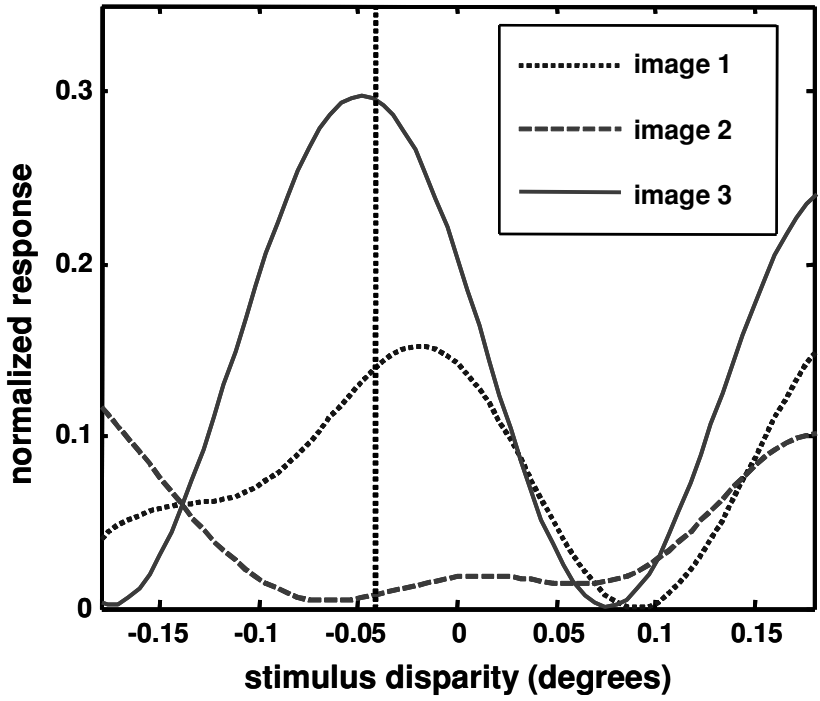

Figure 1: Characteristic responses of an individual model neuron as a function of the disparity (in degrees of visual angle) of the presented surface. The three curves show the normalized responses of a single model binocular energy neuron to each of three sample surfaces presented along a range of disparities (from -0.2 to 0.2 degree). The vertical dotted line indicates the cell's preferred disparity (-0.0417 degree). This figure illustrates the fact that an individual model neuron's response depends on many factors and thus is an ambiguous indicator of stimulus disparity.

responses of the model neuron as the disparity of a surface was varied. The neuron responded differently to different surfaces, illustrating that a single neuron's response is an ambiguous indicator of stimulus disparity. This finding motivates the importance of decoding the activity of a population of neurons rather than that of a single neuron (Fleet et al., 1996; Qian, 1994).

4 Neural Decoders

Neural decoders are statistical devices that estimate the distribution of a stimulus parameter based on neural responses. Three different decoders evaluated $p(d \mid \vec{r})$, the distribution of disparity, denoted d, given the

responses of the model complex cells, denoted $\vec{r}$. The decoders differ in their assumptions about the importance of correlations among neural responses.

4.1 Full Joint Probability Decoder. The FJPD is the simplest of the decoders used, but also has the highest storage cost since it requires representing the full joint distribution of disparity and complex cell responses $p(d, \vec{r})$. This distribution has sb^n states, where s is the number of possible binocular disparities, b is the number of bins or response levels (i.e., each complex cell response was discretized to one of b values), and n is the number of complex cells in the population. The conditional distribution of disparity was calculated as

$$p_{full}(d \mid \vec{r}) = \frac{p(d, \vec{r})}{p(\vec{r})}, \tag{4.1}$$

where the joint distribution $p(d, \vec{r})$ and marginal distribution $p(\vec{r})$ were computed directly from the complex cell responses to the visual scenes (histograms giving the frequencies of each of the possible values of $\vec{r}$ and $(d, \vec{r})$ were generated and then normalized; see below). The result of equation 4.1 represents the output of the FJPD.

4.2 Dependence Tree Decoder. The DTD makes use of a data structure and learning algorithm originally proposed in the engineering literature (Chow & Liu, 1968; see also Meilă & Jordan, 2000). It can be viewed as an instance of a graphical model or Bayesian network, a type of model that is currently popular in the artificial intelligence community (Neapolitan, 2004). The basic idea underlying Bayesian networks is that a joint distribution over a set of random variables can be represented by a graph in which nodes correspond to variables and directed edges between nodes correspond to statistical dependencies (e.g., an edge from node x_1 to node x_2 means that the distribution of variable x_2 depends on the value of variable x_1; as a matter of terminology, node x_1 is referred to as the parent of x_2). Dependence trees are Bayesian networks that are restricted in the following ways: (1) the graphical model must be a tree (i.e., ignoring the directions of edges, there are no loops in the graph, meaning that there is exactly one path between every pair of nodes); (2) there is one node that is the root of the tree—this node has no parents; and (3) all other nodes have exactly one parent. A dependence tree is a graphical representation of the following factorization of a joint distribution:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid pa(i)), \tag{4.2}$$

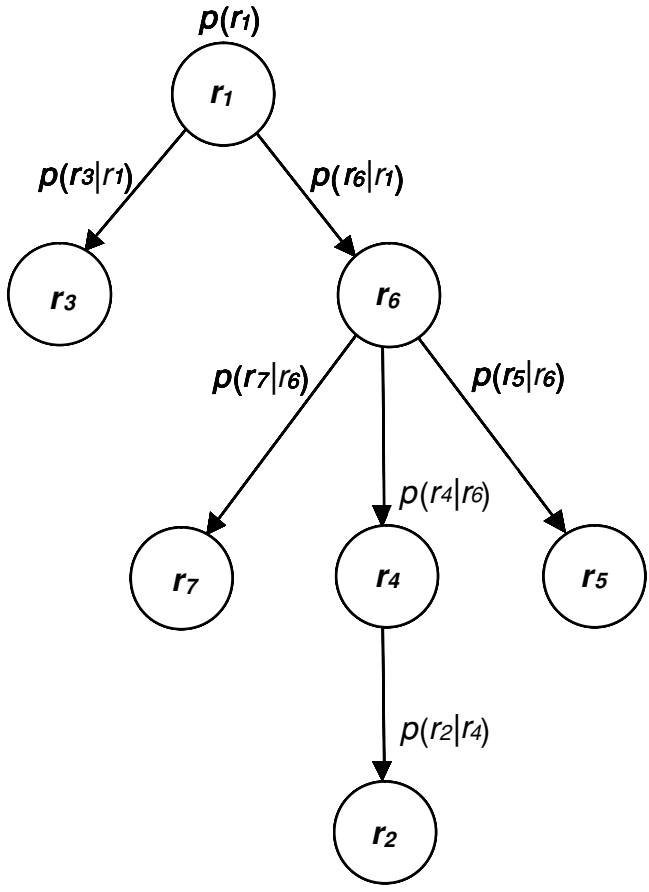

Figure 2: An example of a dependence tree. Each of the nodes $r_1, \ldots, r_7$ represents a random variable, such as the response of a model neuron. The edges (depicted as arrows) represent the conditional dependencies between variables and are labeled with the conditional distribution of a child variable given its parent $p(child|parent)$. According to this tree, the joint distribution of these variables is factorized as follows: $p(r_1, \ldots, r_7) = p(r_1)p(r_3 \mid r_1)p(r_6 \mid r_1)p(r_7 \mid r_6)p(r_4 \mid r_6)p(r_5 \mid r_6)p(r_2 \mid r_4)$.

where $p(x_1, \ldots, x_n)$ is the joint distribution of variables $x_1, \ldots, x_n$ and $p(x_i \mid pa(i))$ is the conditional distribution of variable x_i given the value of its parent (if x_i is the root of the tree, then $p(x_i \mid pa(i)) = p(x_i)$). Figure 2 depicts an example of a dependence tree. Of course, not all joint distributions can be factorized in this way. In this case, the right-hand side of equation 4.2 gives an approximation to the joint distribution. How can good approximations be discovered?

Chow and Liu (1968) developed an algorithm for finding approximations and proved that this approximation maximizes the likelihood of the data over all tree distributions. In short, the algorithm has three steps: (1) compute all pairwise marginal distributions $p(x_i, x_j)$ where x_i and x_j are a pair of random variables; (2) compute all pairwise mutual informations I_{ij}; and (3) compute the maximum weight spanning tree using I_{ij} as the weight for the edge between nodes x_i and x_j. This spanning tree is the dependence tree.[2] Importantly for our purposes, the algorithm has quadratic time complexity in the number of random variables, linear space complexity in the number of random variables, and quadratic space complexity in the number of response levels. That is, discovering the dependence tree that approximates the joint distribution among a set of variables will often be computationally tractable.

The dependence tree decoder computes a dependence tree to approximate the joint distribution of complex cell responses given a binocular disparity value.[3] This approximation is denoted $p_{tree}(\vec{r} \mid d)$. Using Bayes' rule, the distribution of disparity given cell responses is given by

$$p_{tree}(d \mid \vec{r}) = \frac{p_{tree}(\vec{r} \mid d)\, p(d)}{p(\vec{r})}, \tag{4.3}$$

where $p(d)$, the distribution of disparities, is a uniform distribution (i.e., all disparities are equally likely), and $p(\vec{r})$, the distribution of cell responses, is computed by marginalizing $p_{tree}(\vec{r} \mid d)$ over d. Equation 4.3 is the output of the DTD.

4.3 Independent Response Decoder. Using Bayes' rule, we can rewrite the probability of a disparity d given a response $\vec{r}$ as

$$p(d \mid \vec{r}) = \frac{p(\vec{r} \mid d)\, p(d)}{p(\vec{r})}, \tag{4.4}$$

where $p(d)$ is the prior distribution of binocular disparities and $p(\vec{r})$ is a distribution over complex cell responses. Because all disparities were equally likely, we set $p(d)$ to be a uniform distribution. Consequently,

$$p(d \mid \vec{r}) = k p(\vec{r} \mid d), \tag{4.5}$$

[2] The spanning tree is an undirected graph. Our simulations used an equivalent directed graph obtained by choosing an arbitrary node to serve as the root of the tree. The directionality of all edges follows from this choice (all edges point away from the root).

[3] Our data structure can be regarded as a mixture of trees in which there is one mixture component (i.e., one dependence tree) for each possible disparity value (Meilă & Jordan, 2000).

where k is a normalization factor equal to $p(d)/p(\vec{r})$. The distinguishing feature of the independent response decoder (IRD) is that it assumes that the complex cell responses are statistically independent given the binocular disparity. In other words, the conditional joint distribution of cell responses is equal to the product of the distributions for the individual cells, that is, $p(\vec{r} \mid d) = \prod_i p(r^i \mid d)$, where r^i is the response of the ith complex cell. Equation 4.5 can therefore be rewritten as

$$p_{ind}(d \mid \vec{r}) = k \prod_i p(r^i \mid d). \tag{4.6}$$

The distribution of disparity as computed by equation 4.6 is the output of the IRD. The conditional distributions for individual cells $p(r^i \mid d)$ were approximated in our simulations by normalized histograms based on cell responses to visual scenes.

4.4 Response Histograms. Normalized relative frequency histograms were used in our simulations to approximate the distributions of cell responses. In these histograms, each cell's response was discretized to one of b bins or response levels. This discretization was based on a cell's maximum observed response value. Our procedure was similar to that used by Lippert and Wagner (2002), with one important difference. Because the probability of a response was a rapidly decreasing function of response magnitude, Lippert and Wagner created bins representing responses from zero to half of the maximum observed response value and grouped all responses greater than half-maximum into the final bin. This was necessary to avoid bins corresponding to response values that never (or rarely) occurred. To deal with this same problem, we created histograms whose bin values were a logarithmic function of cell response.[4]

5 Simulation Results

Two sets of simulations were conducted. The goal of the first set was to compute the informational costs of using the approximate distributions calculated by the IRD, $p_{ind}(d \mid \vec{r})$, or the DTD, $p_{tree}(d \mid \vec{r})$, instead of the exact distribution calculated by the FJPD, $p_{full}(d \mid \vec{r})$. To quantify these costs, we used an information-theoretic measure, referred to as $\Delta I/I$, introduced

[4] Specifically, histograms were created as follows. A cell's responses were first linearly normalized by dividing each response by that cell's maximum response across all stimuli. Next, each normalized response was discretized into one of b bins where boundaries between bins were logarithmically spaced. To get probabilities of responses given a disparity, bin counts were appropriately normalized and then smoothed using a gaussian kernel whose standard deviation equaled one-quarter of a bin width. This was done to avoid probabilities equal to zero.

by Nirenberg et al. (2001). We chose this measure because, unlike other measures of information difference such as $\Delta I_{shuffled}$ (Nirenberg & Latham, 2003; Panzeri et al., 2002) and $\Delta I_{synergy}$ (Brenner, Strong, Koberle, Bialek, & de Ruyter van Steveninck, 2000), this measure is sensitive only to dependencies that are relevant for decoding (Nirenberg & Latham, 2003).[5] In brief, $\Delta I/I$ can be characterized as follows. The numerator of this measure is the Kullback-Leibler distance between the exact distribution and an approximating distribution. This distance is normalized by the mutual information between a stimulus property (e.g., the disparity d) and the neural responses $\vec{r}$ based on the exact distribution. A small value of $\Delta I/I$ means that the decoding produced by an approximate distribution contains similar amounts of information about the stimulus property as the decoding produced by an exact distribution, whereas a large value means that the approximate decoding contains much less information than the exact decoding.

Simulations were conducted with a variety of neural population sizes (denoted n) and bins or response levels (denoted b). Neural populations sizes were kept small because of the computational costs of computing the exact distribution $p_{full}(d \mid \vec{r})$. Note that the possible values of $\vec{r}$ equals b^n—for example, if $n = 8$ and $b = 8$, then $\vec{r}$ can take 16,777,216 possible values. Fortunately, in practice, $\vec{r}$ took a smaller number of values by a factor of about 100, allowing us to compute $p_{full}(d \mid \vec{r})$ using fewer presentations of visual scenes than would otherwise be the case. We used the responses of model neurons to a collection of 3×10^6 visual scenes in which the frontoparallel surface was located at all possible depths (15 possible depths $\times$ 200,000 scenes per depth) to compute each of the probability distributions $p_{full}(d \mid \vec{r})$, $p_{tree}(d \mid \vec{r})$, and $p_{ind}(d \mid \vec{r})$. This process was repeated six times for each combination of neural population size n and number of bins b. The repetitions differed in the parameter values (e.g., spatial frequencies, receptive field sizes) used by the model neurons.

The results are illustrated in Figure 3. The horizontal axis represents the simulation condition (combination of n and b), and the vertical axis represents the measure $\Delta I/I$. Dark bars give the value of this measure for the IRD, and light bars give the value for the DTD. The error bars indicate the standard errors of the means based on six repetitions of each condition.

There are at least two interesting trends in the data. First, for both the IRD and DTD approximations, the information cost grows with the size of the neural population. In other words, the approximate distributions provided by these decoders become poorer relative to the exact distribution as the neural population grows in size. A two-way (decoder by population size) ANOVA across the $b = 3$ conditions confirmed that this effect is significant ($F(2,35) = 22.15$; $p < 0.001$), with no significant decoder

[5] The best way to measure the distance between two distributions for the purposes of neural decoding is a topic of ongoing scientific discussion (e.g., Nirenberg & Latham, 2003; Schneidman, Bialek, & Berry, 2003).

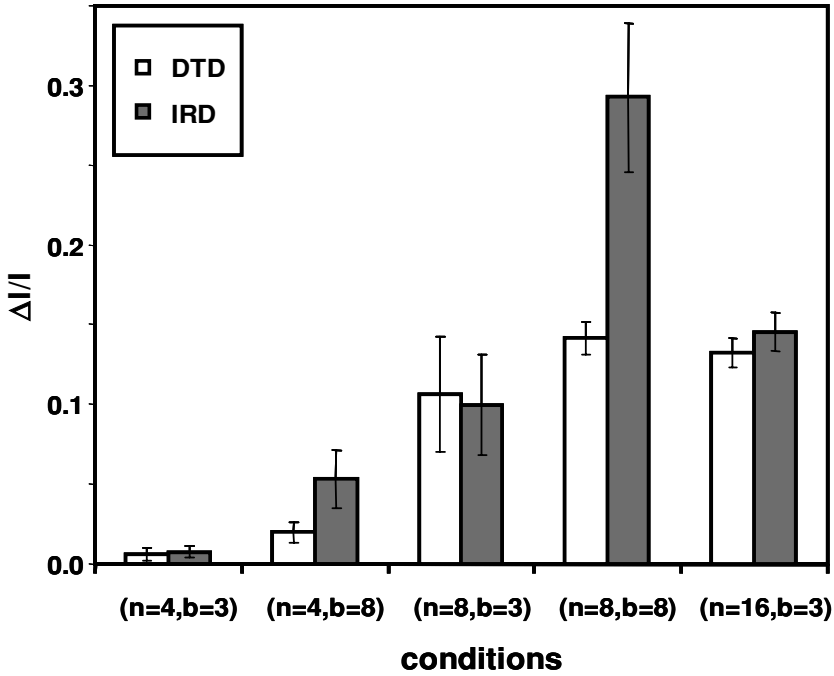

Figure 3: The informational cost $\Delta I/I$ of using the dependence tree decoder (DTD; light bars) or the independent response decoder (IRD; dark bars) as a function of population size (n) and the number of discretized response levels (b). Error bars represent the standard errors of the means.

by population size interaction. This trend is not surprising given that the number of possible high-order correlations grows rapidly with the number of neurons in a population. This result has important implications. Many studies that have attempted to measure the information lost by assuming independence among neural responses have approximated the exact joint distribution with a distribution that takes into account only second-order dependencies (e.g., Abbott & Dayan, 1999; Averbeck & Lee, 2003; Golledge et al., 2003; Nirenberg et al., 2001; Panzeri et al., 1999; Rolls et al., 2003; Seriès et al., 2004). Our results suggest that the difference in relevant information between an approximation based on the assumption that responses are independent and an approximation based on second-order correlations may greatly underestimate the information difference that investigators actually care about: the difference between an approximation based on statistical independence and the exact distribution. If so, this may account for why previous investigators concluded that most of the useful information is in the independent responses of individual neurons. A

second trend in our data is an increase in information cost as the number of discrete response levels increases. This trend is unsurprising as we would expect the differences between exact and inexact distributions to increase as the resolution of neuron responses increases. A three-way ANOVA (decoder by population size by response levels) confirmed that this trend is significant ($F(1,47) = 9.49$; $p < 0.01$), along with a main effect for decoder type ($F(1,47) = 4.35$; $p < 0.05$) and a decoder by response levels interaction ($F(1,47) = 5.05$; $p < 0.05$), which indicate that the effect is significantly greater and more pronounced, respectively, for the IRD than the DTD. In summary, the results of the first set of simulations suggest that the cost of ignoring or approximating statistical dependencies becomes greater with larger populations and also may tend to increase with more neural response levels.

A limitation of the first set of simulations is that the excessive computational cost of calculating the exact distribution $p_{full}(d \mid \vec{r})$ prevented us from examining large population sizes. Therefore, a second set of simulations was conducted in which we evaluated the IRD and DTD with large populations using a performance measure that compared the disparity predicted by a decoder with the true disparity present in a pair of left eye and right eye images. The disparity predicted by a decoder was the disparity with the highest conditional probability (i.e., the disparity that maximized $p(d \mid \vec{r})$, known as the maximum a posteriori estimate).

The distributions $p_{ind}(d \mid \vec{r})$ and $p_{tree}(d \mid \vec{r})$ generated by the IRD and DTD, respectively, were computed on the basis of 150,000 visual scenes in which the frontoparallel surface was located at all possible depths (15 possible depths $\times$ 10,000 scenes per depth). However, the performances of the decoders were measured using a different set of scenes. This set consisted of 1400 scenes in which the surface was located at the central seven depths (possible disparities ranged from -3 to 3 pixels $\times$ 200 scenes per disparity).

The simulation results are illustrated in Figure 4. The horizontal axis indicates the simulation condition (neural population size n and number of response levels b), and the vertical axis indicates the root mean squared (RMS) error of the disparity estimate. Dark bars give the RMS error value for the IRD, and light bars give the value for the DTD. The error bars indicate the standard errors of the means based on six repetitions of each condition. A three-way ANOVA showed significant main effects for population size ($F(2,107) = 9.83$; $p < 0.0001$), for decoder ($F(1,107) = 343.55$; $p < 0.0001$), and for the number of discretized response levels ($F(2,107) = 12.71$; $p < 0.0001$), along with significant effects ($p < 0.0001$) for all two-way interactions. Three primary trends can be gleaned from these combined effects. First, performance for the DTD improved as the population size increased. This was also found for the IRD in the $b = 5$ condition. This trend is unsurprising, as we would expect the amount of information to increase with the size of a neural population. Second, the performance of the DTD became

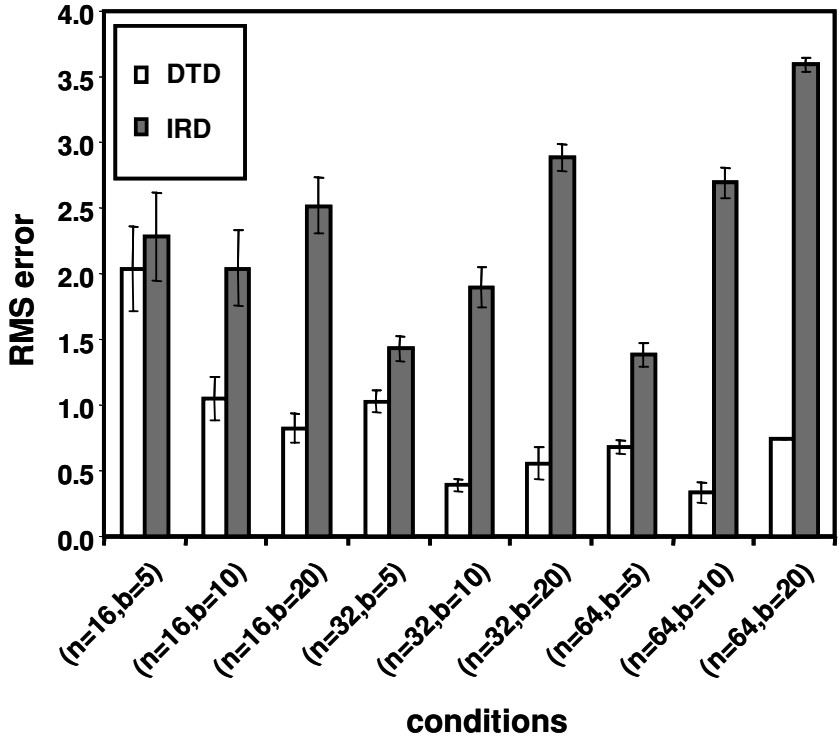

Figure 4: Root mean squared (RMS) error (in pixels) for the DTD (light bars) and IRD (dark bars) as a function of population size (n) and number of discretized response levels (b). RMS errors were calculated by comparing the maximum a posteriori estimates of disparity given by the decoders with the true disparities over 1400 test trials (or novel visual scenes). Error bars indicate the standard errors of the means.

significantly better than that of the IRD with increases in population size, suggesting that the proportion of information about disparity contained in high-order correlations increases with population size compared with the proportion stored in the independent responses of model neurons. Third, the performance of the IRD decreased as the number of discretized response levels increased. In contrast, the performance of the DTD showed the opposite trend—for example, its performance improved slightly from the $b = 5$ to $b = 10$ conditions. This trend may seem surprising given that the number of parameters estimated by the DTD grows quadratically with b while the number of parameters estimated by the IRD grows only linearly. However, the DTD is capable of representing much richer distributions than the IRD. Increasing the number of discretized response levels, like increasing the

number of neurons in a population, increases the possible complexity of correlations. To the extent that information about a stimulus is contained in the possibly high-order response correlations of a neural population, we should expect that any decoder that takes into account these correlations will perform better than the IRD, which, by definition, discards all information in these correlations.

Similar to the results of the first set of simulations, the results of the second set of simulations suggest that much of the information about disparity is carried by statistical dependencies among model neuron responses. These results do not, however, indicate whether the information carried by response dependencies is limited to second-order dependencies or whether higher-order dependencies also need to be considered. To examine this issue, we evaluated the performance of a decoder that was limited to second-order statistics; it approximated the distribution of neural responses given a disparity, $p(\vec{r} \mid d)$, with a multivariate gaussian distribution whose mean vector and covariance matrix were estimated using a maximum likelihood procedure. The performance of this decoder is not plotted because the decoder consistently generated a prediction of disparity equal to 0 pixels (the frontoparallel surface is at the depth of the fixation point) regardless of the true disparity in the left eye and right eye images. A decoder that was forced to use a diagonal covariance matrix produced the same behavior. The poor performances of these decoders are not surprising given the fact that the marginal distributions of an individual neuron's response given a disparity, $p(r^i \mid d)$, are highly nongaussian. The horizontal axis of the graph in Figure 5 represents a normalized response of a typical model neuron, and the vertical axis represents the probability that the neuron will give each response. The light bars indicate the probability when the disparity in a pair of images equals the preferred disparity of the neuron, and the dark bars indicate the probability when the image disparity is different from the neuron's preferred disparity. In both cases, the probability distributions are highly nongaussian; the distributions peak near a response of zero (the neuron most frequently gives a small response) and have relatively long tails (especially the distribution for when the image and preferred disparities are equal). This finding is consistent with earlier results, such as those reported by Lippert and Wagner (2002; see Figure 3).

A possible objection to the simulations discussed so far is that the simulations used a very large number of training stimuli. In contrast, neuroscientists use much smaller data sets, and there is no guarantee that the results that we have found will also be found when using fewer data items. To address this issue, we conducted new simulations with a relatively small data set (100 training samples for each disparity). Figure 6 shows the results for the IRD and the DTD when population sizes were set to 64 neurons, and the number of response levels was set to either 5, 10, or 20. Again, the DTD consistently outperformed the IRD, and the trends described above for the large training set appear to hold for the small training set too.

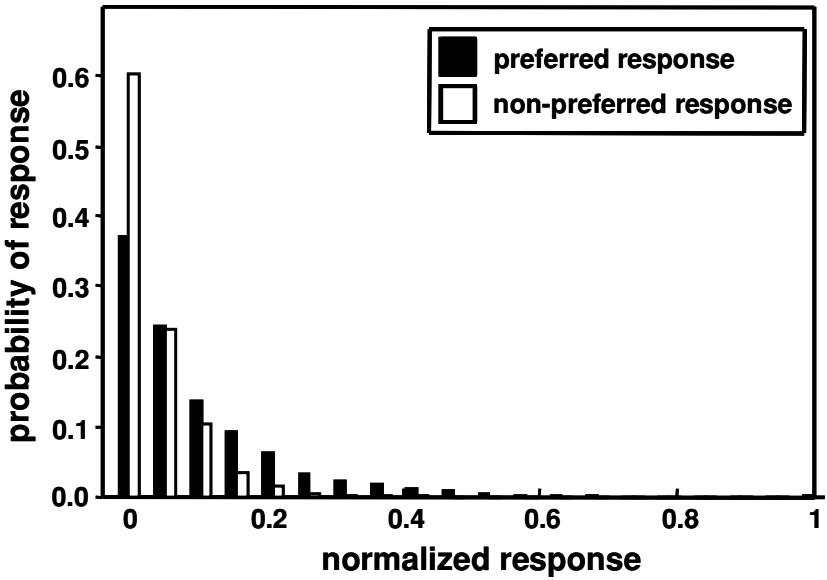

Figure 5: Sample response histograms for a typical model neuron. The black bars indicate the probability of a normalized response to an image pair with the cell's preferred disparity, and the white bars indicate the probability of a response to an image pair with an arbitrarily selected nonpreferred disparity. Note that cell responses are highly nongaussian; the probability distributions are skewed with a peak at very low responses and tails at higher response values. In general, as the selected disparity deviates from the preferred disparity, the mass of the response distribution becomes increasingly concentrated at zero.

A second possible objection to the simulations discussed above is that they used white-noise stimuli; frontoparallel surfaces were covered with dots whose luminance values were independently sampled from a uniform distribution ranging from 0 (dark) to 1 (light). We chose these stimuli for several reasons. White noise stimuli have simple properties that make them amenable to mathematical analyses. Consequently, they have played an important role in engineering, neuroscientific, and behavioral studies. In addition, for our current purposes, we are interested in how binocular disparities can be evaluated in the absence of form information. Furthermore, white noise stimuli do not contain correlations across spatial frequency bands, and thus their use should not introduce biases into our evaluations of the role of high-order correlations when decoding populations of model neurons. Despite the motivations for the use of white noise stimuli, natural visual stimuli contain very different properties. Images of natural scenes usually contain a great deal of form information and contain energy in

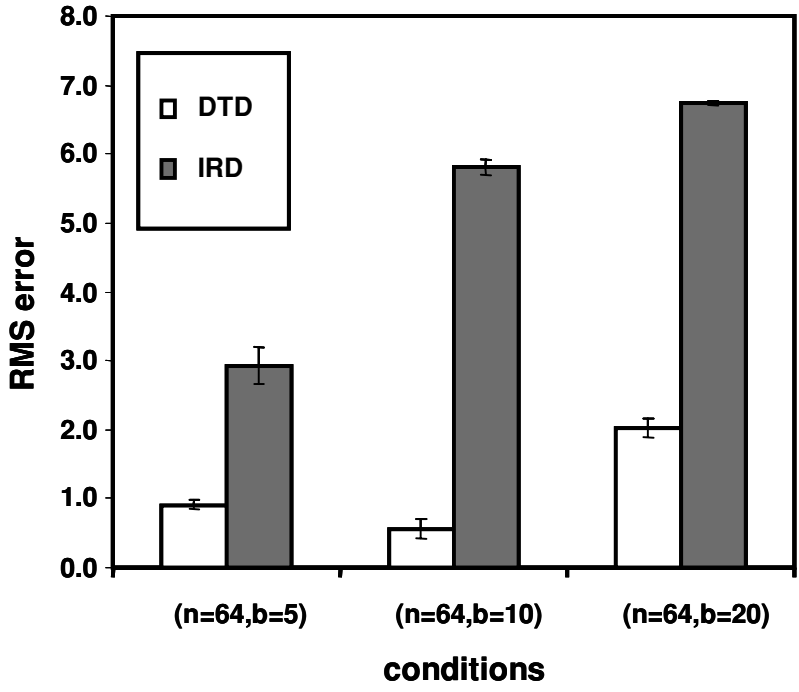

Figure 6: RMS error of the maximum a posteriori disparity estimate provided by the DTD (light bars) and IRD (dark bars) as a function of the number of discretized response levels (b) in the small training sample case. These data were generated using a fixed population size ($n = 64$), and using only 100 training samples per disparity rather than the 10,000 training samples per disparity used to generate the data in Figure 4.

a large range of spatial frequency bands. Because of dependencies in the energies across frequency bands, we expect that high-order correlations in model neuron responses to natural stimuli should be important during neural decoding, as was found when using white noise stimuli. To partially evaluate this prediction, we repeated some of the preceeding simulations using more "naturalistic" stimuli.[6]

[6] Ideally, we would have liked to conduct simulations using left eye and right eye images of natural scenes. Unfortunately, this was not possible for a variety of reasons. Perhaps most important, there are no available databases, to our knowledge, of large numbers of left eye and right eye images of natural scenes taken by well-calibrated camera systems that include ground truth information (e.g., true disparity or depth at each point in the scene).

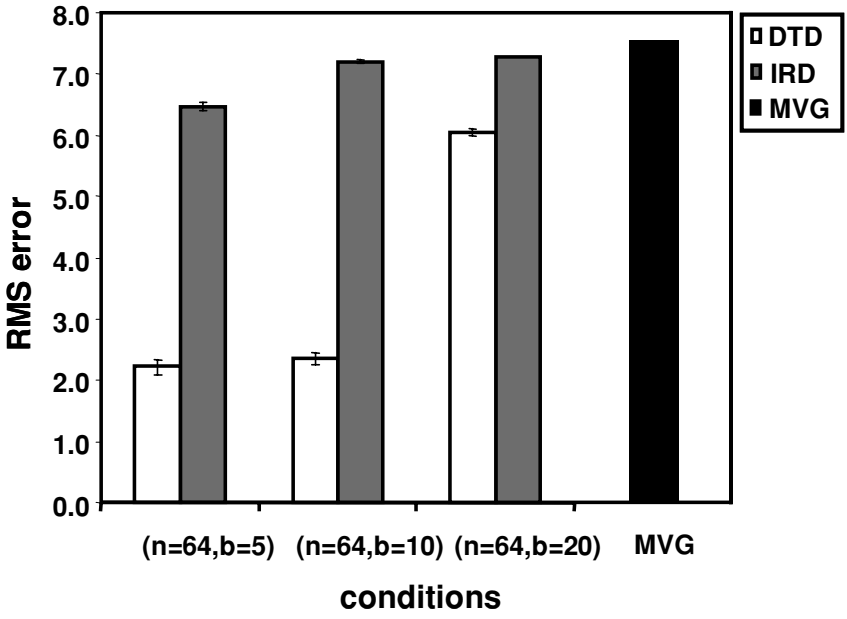

Figure 7: RMS error of the maximum a posteriori disparity estimate provided by the DTD (white bars) and IRD (gray bars) as a function of the number of discretized response levels (b), along with the performance of a multivariate gaussian fitted to the training data (black bar) when the training and test surfaces were painted with $1/f$ noise rather than white noise. These data were generated using a fixed population size ($n = 64$) and using 10,000 training samples per disparity.

In these new simulations, we exploited the fact that the amplitude spectra of natural images fall as approximately $1/f$ (Burton & Moorhead, 1987; Field, 1987, 1994; Tolhurst, Tadmor, & Tang, 1992), We generated left eye and right eye images in the manner described above for the white-noise stimuli, with the exception that each image was a "noise texture" with $1/f$ amplitude spectra; the luminance values of the dots on a surface were independently sampled from a uniform distribution and then passed through a $1/f$ filter (i.e., the luminance values were Fourier transformed, the amplitude at each frequency f was multiplied by $1/f$, and the result was inverse Fourier transformed; in addition, the images resulting from this process were normalized so that their luminance values fell in the range from 0 to 1). The graph in Figure 7 shows the results for the IRD and the DTD based on a population of 64 neurons. As was the case with white noise stimuli, the DTD consistently outperformed the IRD, though the performance of both decoders was markedly worse with the $1/f$-noise stimuli. These results are

consistent with our earlier conclusions that high-order correlations among model neuron responses contain significant information about binocular disparities.

6 Summary

Investigators debate the extent to which neural populations use pairwise and higher-order statistical dependencies among neural responses to represent information about a visual stimulus. To study this issue, we used three statistical decoders to extract the information in the responses of model neurons about the binocular disparities present in simulated pairs of left eye and right eye images. The full joint probability decoder (FJPD) considered all possible statistical relations among neural responses as potentially important. The dependence tree decoder (DTD) also considered all possible relations as potentially important, but it approximated high-order statistical correlations using a computationally tractable procedure. Finally, the independent response decoder (IRD) assumed that neural responses are statistically independent, meaning that all correlations should be zero and thus can be ignored. Two sets of simulations were performed. The first set examined the informational cost of ignoring all correlations or of approximating high-order correlations by comparing the IRD and DTD with the FJPD. The second set compared the performances of the IRD and DTD on a binocular disparity estimation task when neural population size and number of response levels were varied.

The results indicate that high-order correlations among model neuron responses contain significant information about disparity and that the amount of this high-order information increases rapidly as a function of neural population size. In addition, the DTD consistently outperformed the IRD (and also a decoder based on a multivariate gaussian distribution) on the disparity estimation task, and its performance advantage increased with neural population size and the number of neural response levels. These results raise the possibility that previous researchers who have ignored pairwise or high-order statistical dependencies among neuron responses, or who have examined the importance of statistical dependencies in a way that limited their evaluation to pairwise dependencies may not be justified in doing so. Moreover, the results highlight the potential importance of the dependence tree decoder to neuroscientists as a powerful but still practical way of approximating high-order correlations among neural responses.

Finally, the strengths and limitations of this work highlight important areas for future research. For example, future investigations will need to make use of databases of natural images, such as databases with many pairs of right eye and left eye images of natural scenes taken by well-calibrated camera systems, along with ground-truth information about each scene (e.g., depth or disparity information at every point in a scene). Such a database for the study of binocular vision in natural scenes does not

currently exist. In addition, future computational work will need to use more detailed neural models, such as models of populations of neurons that communicate via action potentials and models of individual neurons that include ion kinetics. We expect that the results reported here will generalize to these more realistic situations, but further work is needed to test this prediction.

Acknowledgments

We thank A. Pouget for encouraging us to study the contributions to neural computation of high-order statistical dependencies among neuron responses and thank F. Klam and A. Pouget for many helpful discussions on this topic. This work was supported by NIH research grant RO1-EY13149.

References

Abbott, L. F., & Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural Computation*, *11*, 91–101.

Andrews, T. J., Glennerster, A., & Parker, A. J. (2001). Stereoacuity thresholds in the presence of a reference surface. *Vision Research*, *41*, 3051–3061.

Anzai, A., Ohzawa, I., & Freeman, R. D. (1997). Neural mechanisms underlying binocular fusion and stereopsis: Position vs. phase. *Proceedings of the National Academy of Sciences*, *94*, 5438–5443.

Averbeck, B. B., & Lee, D. (2003). Neural noise and movement-related codes in the macaque supplementary motor area. *Journal of Neuroscience*, *23*, 7630–7641.

Averbeck, B. B., & Lee, D. (2004). Coding and transmission of information by neural ensembles. *Trends in Neurosciences*, *27*, 225–230.

Brenner, N., Strong, S. P., Koberle, R., Bialek, W., & de Ruyter van Steveninck, R. R. (2000). Synergy in a neural code. *Neural Computation*, *12*, 1531–1552.

Burton, G. J., & Moorhead, I. R. (1987). Color and spatial structure in natural scenes. *Applied Optics*, *26*, 157–170.

Chow, C. K., & Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, *14*, 462–467.

Cumming, B. G., & Parker, A. J. (1997). Responses of primary visual cortical neurons to binocular disparity without depth perception. *Nature*, *389*, 280–283.

DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1991). Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature*, *352*, 156–159.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*, 2379–2394.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, *6*, 559–601.

Fleet, D. J., Wagner, H., & Heeger, D. J. (1996). Neural encoding of binocular disparity: Energy models, position shifts, and phase shifts. *Vision Research*, *36*, 1839–1857.

Freeman, R. D., & Ohzawa, I. (1990). On the neurophysiological organization of binocular vision. *Vision Research*, *30*, 1661–1676.

Golledge H. D. R., Panzeri, S., Zheng, F., Pola, G., Scannell, J. W., Giannikopoulos, D. V., Mason, R. J., Tovée, M. J., & Young, M. P. (2003). Correlations, feature-binding and population coding in primary visual cortex. *NeuroReport*, *14*, 1045–1050.

Lippert, J., & Wagner, H. (2002). Visual depth encoding in populations of neurons with localized receptive fields. *Biological Cybernetics*, *87*, 249–261.

Meilă, M., & Jordan, M. I. (2000). Learning with mixtures of trees. *Journal of Machine Learning Research*, *1*, 1–48.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Upper Saddle River, NJ: Prentice Hall.

Nirenberg, S., Carcieri, S. M., Jacobs, A. L., & Latham, P. E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature*, *411*, 698–701.

Nirenberg, S., & Latham, P. E. (2003). Decoding neuronal spike trains: How important are correlations? *Proceedings of the National Academy of Sciences USA*, *100*, 7348–7353.

Ohzawa I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science*, *249*, 1037–1041.

Ohzawa I., DeAngelis, G. C., & Freeman, R. D. (1997). Encoding of binocular disparity by complex cells in the cat's visual cortex. *Journal of Neurophysiology*, *77*, 2879–2909.

Oram M. W., Földiàk, P., Perrett, D. I., & Sengpiel, F. (1998). The "ideal homunculus": Decoding neural population signals. *Trends in Neuroscience*, *29*, 259–265.

Panzeri, S., Golledge, H. D. R., Zheng, F., Pola, G., Blanche, T. J., Tovee, M. J., & Young, M. P. (2002). The role of correlated firing and synchrony in coding information about single and separate objects in cat V1. *Neurocomputing*, *44–46*, 579–584.

Panzeri, S., Schultz, S. R., Treves, A., & Rolls, E. T. (1999). Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society of London Series B*, *266*, 1001–1012.

Perez, R., Castro, A. F., Justo, M. S., Bermudez, M. A., & Gonzalez, F. (2005). Retinal correspondence of monocular receptive fields in disparity-sensitive complex cells from area V1 in the awake monkey. *Investigative Ophthalmology and Visual Science*, *46*, 1533–1539.

Perkel, D. H., & Bullock, T. H. (1969). Neural coding. In F. O. Schmitt, T. Melnechuk, G. C. Quarton, & G. Adelman (Eds.), *Neurosciences research symposium summaries* (pp. 405–527). Cambridge, MA: MIT Press.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Computation and inference with population codes. *Annual Review of Neuroscience*, *26*, 381–410.

Prince, S. J. D., Pointon, A. D., Cumming, B. G., & Parker, A. J. (2002). Quantitative analysis of the responses of V1 neurons to horizontal disparity in dynamic random-dot stereograms. *Journal of Neurophysiology*, *87*, 191–208.

Qian, N. (1994). Computing stereo disparity and motion with known binocular cell properties. *Neural Computation*, *6*, 390–404.

Qian, N., & Zhu, Y. (1997). Physiological computation of binocular disparity. *Vision Research*, *37*, 1811–1827.

Read, J. C. A., & Cumming, B. G. (2003). Testing quantitative models of binocular disparity selectivity in primary visual cortex. *Journal of Neurophysiology*, *90*, 2795–2817.

Rolls, E. T., Franco, L., Aggelopoulos, N. C., & Reece, S. (2003). An information theoretic approach to the contributions of the firing rates and the correlations between the firing of neurons. *Journal of Neurophysiology*, *89*, 2810–2822.

Seriès, P., Latham, P. E., & Pouget, A. (2004). Tuning curve sharpening for orientation selectivity: Coding efficiency and the impact of correlations. *Nature Neuroscience*, *10*, 1129–1135.

Schneidman, E., Bialek, W., & Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, *23*, 11539–11553.

Tolhurst D. J., Tadmor, Y., & Tang, C. (1992). The amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, *12*, 229–232.

Tsai, J. J., & Victor, J. D. (2003). Reading a population code: A multi-scale neural model for representing binocular disparity. *Vision Research*, *43*, 445–466.

Received February 4, 2005; accepted July 20, 2005.