Sparse low-order interaction network underlies a highly correlated and learnable neural population code

Elad Ganmor[a], Ronen Segev[b,1], and Elad Schneidman[a,1]

[a]Department of Neurobiology, The Weizmann Institute of Science, Rehovot 76100, Israel; and [b]Department of Life Sciences and The Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

Edited* by Nikos K. Logothetis, The Max Planck Institute for Biological Cybernetics, Tuebingen, Germany, and approved April 19, 2011 (received for review January 4, 2011)

Information is carried in the brain by the joint activity patterns of large groups of neurons. Understanding the structure and function of population neural codes is challenging because of the exponential number of possible activity patterns and dependencies among neurons. We report here that for groups of ∼100 retinal neurons responding to natural stimuli, pairwise-based models, which were highly accurate for small networks, are no longer sufficient. We show that because of the sparse nature of the neural code, the higher-order interactions can be easily learned using a novel model and that a very sparse low-order interaction network underlies the code of large populations of neurons. Additionally, we show that the interaction network is organized in a hierarchical and modular manner, which hints at scalability. Our results suggest that learnability may be a key feature of the neural code.

high-order | correlations | maximum entropy | neural networks | sparseness

Sensory and motor information is carried in the brain by sequences of action potentials of large populations of neurons (1–3) and, often, by correlated patterns of activity (4–11). The detailed nature of the code of neural populations, namely the way information is represented by the specific patterns of spiking and silence over a group of neurons, is determined by the dependencies among cells. For small groups of neurons, we can directly sample the full distribution of activity patterns of the population; identify all the underlying interactions, or lack thereof; and understand the design of the code (12–15). However, this approach cannot work for large networks: The number of possible activity patterns of just 100 neurons, a population size that already has clear functional implications (16), exceeds 10^{30}. Thus, our understanding of the code of large neural populations depends on finding simple sets of dependencies among cells that would capture the network behavior (17–19).

The success of pairwise-based models in describing the strongly correlated activity of small groups of neurons (19–25) suggests one such simplifying principle of network organization and population neural codes, which also simplifies their analysis. Using only a quadratic number of interactions, out of the exponential number of potential ones, pairwise maximum entropy models reveal that the code relies on strongly correlated network states and exhibits distributed error-correcting structure (19, 21). It is unclear, however, if pairwise models are sufficient for large networks, particularly when presented with natural stimuli that contain high-order correlation structure. Here we show that in this case pairwise models capture much, but not all, of the network behavior. This implies a much more complicated structure of population codes (26, 27). Because learning even pairwise models is computationally hard (28–33), this may seem to suggest that population codes would be extremely hard to learn.

We show here that this is not the case for neural population codes. The sparseness of neuronal spiking and the highly correlated behavior of large groups of neurons facilitate the learning of the functional interactions governing the population activity. Specifically, we found that a very sparse network of low-order interactions gives an extremely accurate prediction of the probability of the hundreds of thousands of patterns that were observed in a long experiment and, in particular, captures the highly correlated structure in the population response to natural movies. We uncover the dominant network interactions using a novel approach to modeling the functional dependencies that underlie the population activity patterns. This "reliable interaction" (RI) model learns the dominant functional interactions in the network, of any order, using only the frequent and reliably sampled activity patterns of the network. Although this approach would not be useful for general networks, for neural populations that have a heavy-tailed distribution of occurrences of activity patterns, the result is an extremely accurate model of the network code. Our results indicate that because of the sparse nature of neural activity (34), the code of large neural populations of ganglion cells is learnable in an easy and accurate manner from examples. Finally, we demonstrate that large network models can be constructed from smaller subnetworks in a modular fashion, which renders the approach scalable and perhaps applicable to much larger networks.

Results

To study the code of large networks of neurons, we recorded simultaneously the spiking activity of groups of ∼100 ganglion cells in a 2-mm^2 patch of Salamander retina responding to long natural and artificial movies (Fig. 1*A* and *Methods*). To give a general representation of network activity, we discretized time using a window of size Δt; for a small enough Δt, the responses of each neuron, x_i, are binary [i.e., the cell either spikes (1) or is silent (0)]. Here, we used 20 ms, which reflects the temporal structure of correlations between cells; different bin sizes did not qualitatively affect our results.

The distribution of occurrences of activity patterns over the population was heavy-tailed for both spatial and spatiotemporal patterns (Fig. 1*B*). Similar to what was found for smaller populations, typical pairwise correlations were weak (e.g., ref. 35) (Fig. 1*B*, *Inset*), yet the network as a whole was strongly correlated. Assuming that cells were independent $P^{(1)}(\{x_i\}) = P(x_1)P(x_2)\dots P(x_n)$ resulted in orders-of-magnitude errors in describing the network synchrony (Fig. 1*C*) and in predicting specific spatial population activity patterns (Fig. 1*D*) as well as spatiotemporal ones (Fig. 1*E* and *Methods*). We conclude that an accurate description of the population code must take into account correlations between cells.

Pairwise Models Are Not Sufficient to Describe the Distribution of Activity Patterns of Large Networks Responding to Natural Stimuli. The most general description of the population activity of n neurons, which uses all possible correlation functions among cells, can be written using the maximum entropy principle (36, 37) as:

Author contributions: E.G., R.S., and E.S. designed research; E.G., R.S., and E.S. performed research; E.G., R.S., and E.S. analyzed data; and E.G., R.S., and E.S. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence may be addressed. E-mail: ronensgv@bgu.ac.il or elad.schneidman@weizmann.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1019641108/-/DCSupplemental.

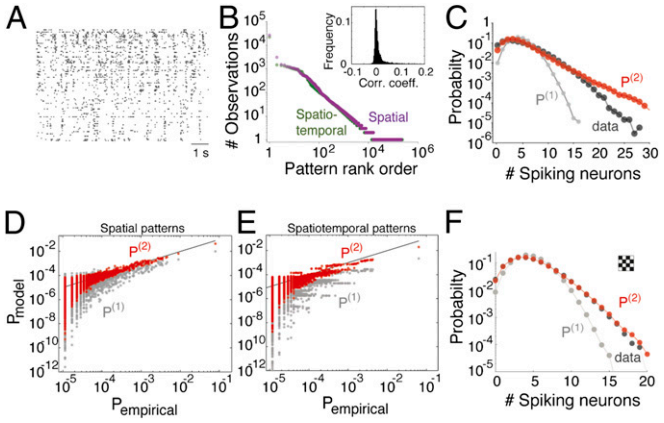

Fig. 1. Pairwise maximum entropy models for spatial and spatiotemporal activity patterns of a large network responding to natural stimuli and signs of higher-order interactions. (*A*) Segment of the simultaneous action potential sequences of 99 ganglion cells in the salamander retina responding to a natural movie clip. Each line corresponds to a single neuron, and each tick represents a single spike. (*B*) All activity patterns of the 99-neuron network described in *A* that were observed in the experiment were sorted and plotted according to the number of times they were observed (*y* axis) for both spatial patterns over all cells (purple) and for spatiotemporal activity patterns of 10 neurons over 10 time steps (green). (*Inset*) Histogram of correlation coefficients between all pairs of neurons in *A*. Corr. coeff., correlation coefficient. (*C*) Probability distribution of synchronous spiking events in the 99-cell population in response to a long natural movie (black). The distribution of synchronous events for the same 99 cells after shuffling each cell's spike train to eliminate all correlations among neurons ($P^{(1)}$, gray) and the synchrony distribution predicted from the second-order maximum entropy model ($P^{(2)}$, red) are also shown. (*D*) Probability of occurrence of each simultaneous (spatial) population activity pattern that appeared in the experiment as predicted if all cells are independent ($P^{(1)}$, gray) or by the second-order maximum entropy model, which takes into account pairwise correlations ($P^{(2)}$, red), are plotted against the measured rate. Although most rare patterns fall within a 95% confidence region (not included), frequently observed patterns are misestimated by the pairwise model. (*E*) Same as in *D* but for spatiotemporal activity patterns. We define temporal population patterns as the 100-bit binary words of 10 retinal ganglion cells from *A* over 10 time steps (200 ms) and fit the independent model ($P^{(1)}$) and second-order maximum entropy model ($P^{(2)}$) to the population spatiotemporal patterns. Plot details are as in *D*. (*F*) Probability distribution of synchronous spiking events in the 99-cell population in response to artificial white noise stimuli (black); second order model and independent model are also shown (as described in *C*). Unlike the responses to natural movies, the pairwise model provides a very good fit to the responses to white noise stimuli, reflecting a negligible role for higher-order interactions in the population activity evoked by such stimuli.

$$\hat{P}(x_1,x_2,\ldots,x_n)=\frac{1}{Z}\exp\left\{\sum_i \alpha_i x_i+\sum_{i<j}\beta_{ij}x_ix_j+\sum_{i<j<k}\gamma_{ijk}x_ix_jx_k+\sum_{i<j<k<l}\delta_{ijkl}x_ix_jx_kx_l+\ldots\right\} \quad [1]$$

where $\{\alpha_i,\beta_{ij},\gamma_{ijk},\delta_{ikjl},\ldots\}$ must be found numerically, such that the corresponding averages, such as $<x_i>$, $<x_ix_j>$, or $<x_ix_jx_k>$ over $\hat{P}$, agree with the experimental ones; the partition function Z is a normalization factor. However, $\hat{P}$ has an exponential number of parameters (interactions). Thus, this representation would be useful only if a small subset of parameters would be enough to capture the network behavior.

For small groups of 10–20 neurons, it has been shown that the minimal model of the network that relies only on pairwise interactions, namely, the maximum entropy pairwise $P^{(2)}$ model, with only $\{\alpha_i\}$s and $\{\beta_{ij}\}$s in Eq. **1**, captures more than 90% of the correlation structure of the network patterns (19, 20). Learning $P^{(2)}(x_1,x_2,\ldots,x_n)$, which has only $O(n^2)$ parameters, is a known hard computational problem (29). For 100 cells, we cannot even enumerate all network states and must rely on a Monte Carlo algorithm to find the parameters of $P^{(2)}$ (*Methods*). The resulting model gave a seemingly accurate description of the network spatial patterns (Fig. 1*D*), correcting the orders-of-magnitude mistakes that $P^{(1)}$ makes. We obtained similar results for the spatiotemporal activity patterns of 10 neurons over 10 consecutive time bins, which we learned using the temporal correlations between cells $\langle x_i(t)x_j(t+\Delta t)\rangle_t$ (Fig. 1*E* and *Methods*).

However, closer inspection reveals that the most common population activity patterns were actually misestimated (Fig. 1 *D* and *E*), whereas the rare patterns were predicted well by $P^{(2)}$. This reflects a real failure of the pairwise model in this case and not a numerical artifact of our algorithm (Fig. S1). Moreover, $P^{(2)}$ is an accurate model of the population's response to artificial movies without spatial structure, such as spatiotemporal white noise (24) (Fig. 1*F* and Fig. S1). Thus, the inaccuracy of the pairwise model for the retina responding to natural stimuli is the signature of the contribution of higher-order interactions to the neural code of large populations.

High-order interactions, which reflect the tendency of groups of triplets and quadruplets, for example, to fire synchronously (or to be silent together) beyond what can be explained from the pairwise relations between cells, have been studied before in small networks (4, 26, 27, 36, 37). These synchronous events are believed to be crucial for neural coding and vital for the transmission of information to downstream brain regions (4, 6, 7, 10). However, unlike the case of small groups of neurons, where one can systematically explore all interaction orders, how can we find the high-order interactions that define the neural population code in large networks among the exponential number of potential ones?

Learning the Functional Interactions Underlying the Neural Code from Reliably Sampled Activity Patterns. We uncovered the structure of the population code of the retina using a network model that gives an extremely accurate approximation to the empirical distribution of population activity patterns. We note that if we were given the true distribution, $P(\{x_i\})$, over all the network activity patterns, we could then infer the interactions among neurons simply by solving a set of linear equations for the parameters. For a specific activity pattern $\{x_i\}$, substituting the appropriate 0's and 1's in Eq. **1**, we get that $\log P(\{x_i\})=-\log Z+\sum_i\alpha_i+\sum_{i<j}\beta_{ij}+\ldots$, where the sums are only over terms where all x_is are 1. The interaction parameters that govern the network could then be calculated recursively, starting from the low-order ones and going up (36, 38):

$$\begin{aligned}\hat{Z}&=1/P(000\ldots0),\\ \hat{\alpha}_1&=\log P(100000..)+\log\hat{Z},\\ \hat{\beta}_{12}&=\log P(110000\ldots)+\log\hat{Z}-\hat{\alpha}_1-\hat{\alpha}_2,\text{ etc.}\end{aligned} \quad [2]$$

In general, this approach would not be useful for learning large networks, because the number of possible activity patterns grows exponentially with network size. We would therefore rarely see any pattern more than once, regardless of how long we observe the system, and as a result we would not be able to estimate the parameters in Eq. **2**. However, the distribution of pattern occurrences in large networks of neurons had many patterns, both spatial and spatiotemporal ones, that appeared many times during the experiment (Fig. 1*B*). This feature of the pattern count distribution is a direct result of the sparseness of the neural code and the correlation structure of neural activity, and does not hold for other kinds of distributions (Fig. S2).

We can then use the frequently occurring patterns to estimate the interactions governing the joint activity of neural populations accurately and to build the RI model that approximates the distribution of neural population responses:

$$f_{RI}(\{x_i\}) = \frac{1}{\hat{Z}} \exp\left(\sum_{\mu} \hat{g}_{\mu} O_{\mu}(\{x_i\}) \right), \qquad [3]$$

where $O_{\mu}(\{x_i\})$ are the dependencies of any order among neurons, whose interaction parameters, $\hat{g}_{\mu}$, are inferred, as in Eq. **2**, from patterns μ that appeared more than n_{RI} times (our learning threshold) in the training data. The order of interactions in the model is therefore determined by the nature of the patterns in the data and not by an arbitrarily imposed finite order; higher-order interactions beyond pairs, if needed, would stand for synchronous events that could not be explained by pairwise statistics.

Sparse Low-Order Interaction Network Model Captures the Code of Large Populations of Neurons. We learned an RI model for the data from Fig. 1, and the resulting model had far fewer parameters than the pairwise maximum entropy model but included a small set of higher-order interactions. This RI model predicted the probability of the spatial population activity patterns in a 1-h test data extremely well (Fig. 2*A*). Its accuracy was actually within the experimental sampling error of the data itself, which we quantified using the log-likelihood ratio of the model and the empirical data, $\log_2 \frac{P_{empirical}(x_1, x_2, \ldots, x_n)}{P_{model}(x_1, x_2, \ldots, x_n)}$. The RI model was equally accurate in predicting the spatiotemporal patterns in the population (Fig. 2*B*). For comparison, we also present the poor performance of the independent $P^{(1)}$ model and the systematic failure of the pairwise $P^{(2)}$ model (Fig. 2 *A* and *B*). We found similar behavior of the models when we used only the subpopulation of "Off" cells in the salamander retina recorded in our experiment (Fig. S3). Moreover, the interactions among the Off cells network were similar to the ones we found using the full network of all ganglion cells recorded (Fig. S3). The RI model remained a highly accurate model for different time bin values used to analyze neural activity (Fig. S4), whereas $P^{(2)}$ became even less accurate for larger time bins of 40 ms.

Because of the sparseness of the neural code, the number of interaction parameters that the RI model relies on is very small, much smaller than the number of all possible pairs, and these parameters are low-order ones. The functional architecture of the retina responding to a natural movie, inferred from the RI model, is shown in Fig. 3*A*, overlaid on the physical layout of the receptive field centers of the cells. We found that beyond a selective group of single-cell biases (α_i) and pairwise interactions (β_{ij}), there were significant higher-order interactions (e.g., γ_{ijk}, δ_{ijkl}). However, importantly, the model was very sparse, using only hundreds of pairwise and triplet interactions, tens of quadruplets, and a single quintuplet (with a learning threshold of $n_{RI} = 10$). The pairwise interactions were mostly positive; that is, cells tended to spike together or to be silent together, and their strength decayed with distance (Fig. 3*D*). These higher-order interactions had a similar decay with distance, but the interactions among nearby cells were mostly negative, indicating that neurons were more silent together than would be expected from pairwise relations alone (Fig. 3*D*). We note that the decay of interactions with distance is stronger than the decay of spatial correlations in the natural movies we showed the retina (Fig. 3*D*, *Inset*). We found a similar decay of pairwise and higher-order temporal interactions among cells with the time difference between them as well as a similar polarity of the interaction sign (Fig. 3*E*).

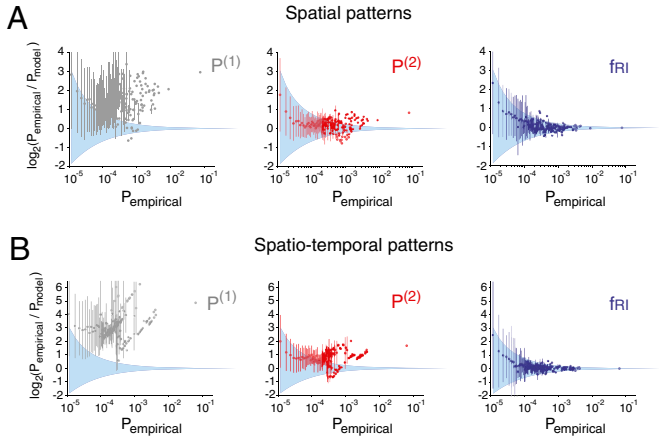

Fig. 2. RI model is a highly accurate and easy-to-learn model of large neural population activity. (*A*) For all spatial activity patterns of the cells from Fig. 1*A* that occur with probability *P* (abscissa), we plot the mean and SD (error bar) of the log-likelihood ratio of the model and data (ordinate): $\log_2 \frac{P_{empirical}(x_1, x_2, \ldots, x_n)}{P_{model}(x_1, x_2, \ldots, x_n)}$. The blue funnel marks the estimated 95% confidence interval of the empirical measurement. Independent model $P^{(1)}$(*Left*), pairwise maximum entropy model $P^{(2)}$(*Center*), and RI model (*Right*). (*B*) Model predictions for population spatiotemporal activity patterns of 10 neurons over 10 time steps. Other details are as in *A*.

Higher-Order Interactions in the Neural Code of the Retina Responding to Natural Scenes. We found that the RI model was much better than the pairwise $P^{(2)}$ model in describing the retinal population's response to different natural movies (Fig. 4 *A* and *B*). The benefit of the high-order interactions that the RI models identify was even more pronounced when the whole retina patch was presented with the same stimulus, as we found for a "natural pixel" (NatPix) movie in which the stimulus shown was the luminance of one pixel from a natural movie, retaining natural temporal statistics, and for Gaussian full-field movies (*Methods*). The RI model's accuracy in describing the population activity patterns was similar to that of the $P^{(2)}$ model (but with far fewer parameters) for spatiotemporal white noise stimuli that had no temporal or spatial correlations (Fig. 4 *A* and *B*). The RI model was also similar to the $P^{(2)}$ model in capturing the retinal response to a manipulated natural movie, which retained the pairwise correlations in the movie but with randomized phases (*Methods*). Fig. 4*B* summarizes the average gain in accuracy of the RI model compared with $P^{(2)}$ for many groups of neurons in salamander and archer fish retinae responding to artificial, naturalistic, and natural stimuli. Our results indicate that as the stimulus becomes more correlated, the network response becomes more correlated with a growing contribution of high-order interactions; thus, the improvement introduced by the RI model becomes more pronounced. We conclude that the higher-order interactions in the neural response to natural movies are driven by the higher-order statistics in natural scenes. Furthermore, most of the high-order interactions that we found for widely correlated stimuli were gone when we constructed the conditionally independent response of the cells (Fig. 4*C*). This is done by shuffling the responses of each neuron to a repetitive movie, thereby retaining only stimulus-induced correlations among neurons. Thus, beyond the stimulus statistics, the retinal circuitry does play an important role in generating the functional higher-order interactions in the response.

We emphasize that f_{RI} was much more accurate than $P^{(2)}$ using a considerably smaller number of parameters. In particular, although the full-pairwise model requires almost 5,000 parameters for 99 neurons, the RI model was already more accurate with just 450 parameters (Fig. 4*D*). The model continued to improve as we added more parameters or lowered the learning threshold n_{RI}, or when we added training samples (Fig. S5). The cross-validation performance continued to improve even when we used patterns that appeared only twice in the data to train the model, indicating that the model was not overfit (other measures of model performance and validation are shown in Fig. S6). We found similar behavior of the RI and $P^{(2)}$ models for spatiotemporal patterns (Fig. S5).

We note that the RI model is not guaranteed to be a normalized probability distribution over all the population activity patterns. The reason is that by using only the reliable patterns to

NEUROSCIENCE

learn the model, we are inevitably missing nonzero interactions, corresponding to less common patterns that are not used in the training. Unlike other such pseudolikelihood models, which are common in machine learning and statistics (29, 38, 39), the RI model does give an extremely accurate estimate of the actual probability for almost all patterns observed in the test data. The patterns for which it would fail are typically so rare that we would not see them even in a very long experiment.

Hierarchical Modular Organization of the Interaction Network Underlying the Code of Large Neural Populations. For networks much larger than 100 neurons, we expect that no activity pattern will appear more than once even in a very long experiment because of neural noise; thus, the RI approach could not be simply applied. However, because the interactions of all orders decayed with distance between cells, we hypothesized that the RI model might be applied in a hierarchical manner. Indeed, we found that the interactions inferred from smaller overlapping subnetworks of spatially adjacent neurons (based on their receptive fields) were very similar to those estimated using the full network (Fig. 4*E*). This surprising result reflects an overlapping modular structure of neuronal interactions (40, 41); namely, neurons directly interact only with a relatively small set of adjacent neurons, known as their "Markov blanket" (28), and the interaction neighborhood of each neuron overlaps with that of its neighbors. This is not an ad hoc approximation but rather a mathematical equality based on the local connectivity structure of the network (42). This suggests that the RI approach may be scalable and useful to model and analyze much larger networks, also beyond the retina.

Contribution of Higher-Order Interactions to Stimulus Coding. The average stimulus preceding the joint spiking events of pairs and of triplets of ganglion cells (pattern-triggered average) was significantly different from what would be expected from conditionally independent cells (Fig. 5 *A*–*C*), reflecting that the population codebook carried information differently than what could be read from the single cells or pairs (4, 7). To demon-

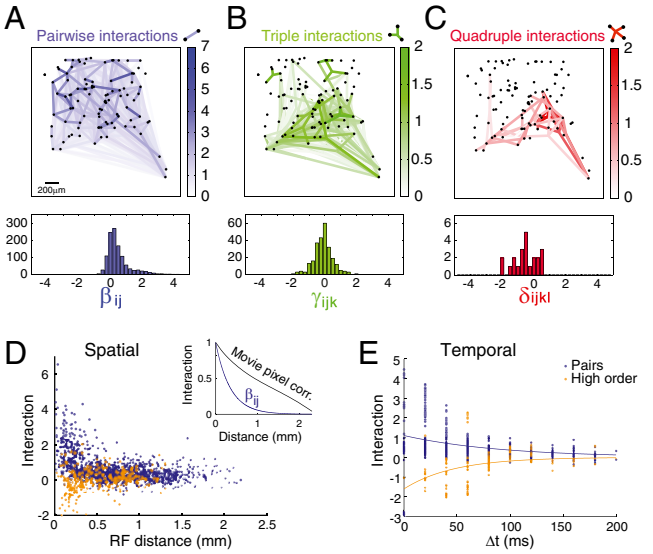

Fig. 3. Functional interaction maps of retinal ganglion cells responding to a natural movie, based on the RI model. (*A*) Pairwise interactions, β_{ij}, learned by the RI model for 99 retinal ganglion cells responding to a natural movie. (*Upper*) The center of the receptive field of each of the neurons is marked by a black dot (mapped using flickering checkerboard stimuli; *Methods*). The saturation level of an edge between vertices is set according to the absolute value of the interaction. (*Lower*) Histogram of the pairwise interaction terms. (*B*) Same as in *A*, except for the triple-wise interaction terms, γ_{ijk}, of the RI model. Note that each of the interactions is shown as a three-line star that connects each of the participating cells and the center of mass of their locations. (*C*) Same as in *A*, except for the quadruple-wise interaction terms, δ_{ijkl}, of the RI model; interactions are shown as a four-line star that connects the cells and the center of mass of their locations. (*D*) RI model pairwise interaction values are plotted as a function of distance between the receptive field centers of the interacting neurons (blue). Higher-order interaction values are plotted as a function of their average pairwise distance (yellow). (*Inset*) Decay of pairwise interaction terms (blue; line is a scaled exponential fit to the blue dots in *D*) and of the correlations between pixels in the movie (black) as a function of the distance on the retina. corr., correlation. (*E*) RI model pairwise interaction values are plotted as a function of the temporal delay between the interacting neurons (blue dots; solid line corresponds to exponential fit). Higher-order interaction values are plotted as a function of the maximal lag between pairs (yellow dots; solid line corresponds to exponential fit).

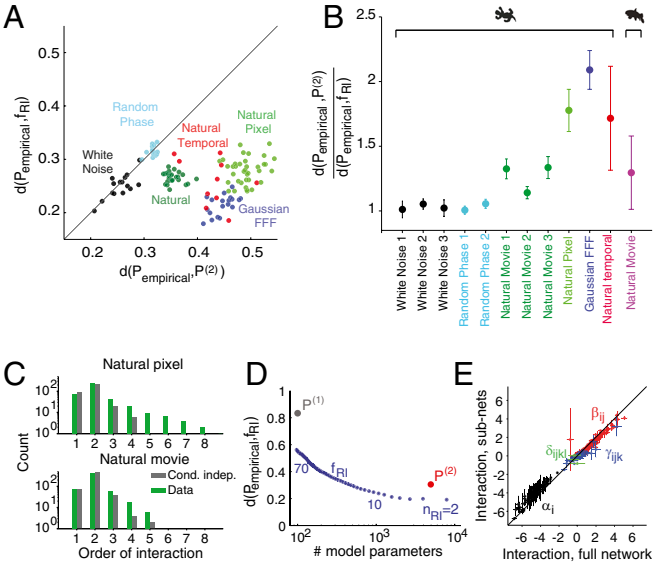

Fig. 4. Higher-order interactions in neural response are driven by long-range stimulus correlations and give more accurate, easier to learn, and hierarchical models of the neural population activity. (*A*) Comparison of the accuracy of the second-order maximum entropy model and the RI model, measured by the absolute log-likelihood ratio $d(P_{empirical}, f_{RI})$, where $d(p,q) = \sum_{\{x_i\}} p(\{x_i\})|\log(p(\{x_i\})/q(\{x_i\}))|$ is summed only over patterns x, which appear more than once in the experiment. $d(P_{empirical}, f_{RI})$ is plotted against $d(P_{empirical}, P^{(2)})$, such that dots below the equality line (black) demonstrate better performance of the RI model. Each dot of the same color represents one network of 50 neurons taken from one experiment. Different natural and artificial visual stimuli were used. (*B*) Overall summary of average and STD of $d(P_{empirical}, P^{(2)})/d(P_{empirical}, f_{RI})$ for many groups of 50 neurons from different retinae responding to many different kinds of movies. Datasets used for the salamander retinas had about 100 cells recorded simultaneously (*Methods*). In the archer fish retina (purple), 38 neurons were recorded; averages are over randomly selected groups of 30 neurons. (*C*) Histograms of the number of interactions of different orders observed in the real data (green bars) compared with the number observed in conditionally independent surrogate data (gray bars) for a NatPix movie (*Upper*) and a natural movie (*Lower*). Cond. indep., conditionally independent. (*D*) Accuracy of the RI model as a function of the number of parameters in the model. $d(P_{empirical}, f_{RI})$ is plotted against the number of parameters in the model for the spatial activity patterns of the 99-neuron network. Results were averaged over 10 randomly selected train and test data (STD error bars are smaller than marker size). The maximum entropy pairwise model (red dot) and the independent model (gray dot) are also shown for comparison. The number of parameters was varied by changing the learning threshold, n_{RI}. (*E*) RI model parameters averaged over all 30-neuron subnetworks containing the 29 nearest neighbors of each neuron are plotted against their value estimated from the full 99-neuron network. Error bars represent SDs over different randomly selected subsets of data (x axis) and SDs across different subnetworks (y axis). sub-nets, subnetworks.

strate the role of higher-order interactions in stimulus coding under natural conditions, we presented the retina repetitively with a 50-s natural movie and trained an RI model on the responses of groups of 10 neurons to each 1-s clip of the movie using half of the repeats. We found that these models could discriminate between the clips very accurately (Fig. 5*D*). For larger networks, more samples are required; thus, we presented the same retina with two long and different full-field stimuli: one with natural temporal statistics (NatPix) and one in which each frame was sampled from a Gaussian distribution [full field flicker (FFF); *Methods*]. For each of the stimuli, an RI model was learned from the population responses f_{NatPix} and f_{FFF}, respectively. New stimuli (not used for the training) were sampled from both stimulus classes and presented to the retina, and they were then classified according to the log-likelihood ratio, $\log(f_{NatPix}(x)/f_{FFF}(x))$. Using the RI models, we found that we could accurately classify new stimulus segments within a few hundred milliseconds (Fig. 5*E*) far better than $P^{(2)}$ (almost three-fold faster) or the independent model (almost 10-fold faster). Finally, we also show that a limited memory-based estimation of f_{RI}, which is learned using only frequent patterns that appeared at least once every few minutes in the training data (Fig. 5*F*), proves to be highly accurate. This limited memory approach can be readily converted into an online learning algorithm: As new samples are presented, only the parameters corresponding to patterns that appear frequently are retained, whereas parameters learned from patterns that have not appeared in the recent past are completely forgotten.

Discussion

Using a novel nonnormalized likelihood model, we have found that a sparse low-order interaction network captures the activity of a large neural population in the retina responding to natural stimuli. This RI model gave an extremely accurate prediction of the observed and unobserved patterns using a small number of interaction parameters. The network interaction model further revealed the modular nature of the code, which could also be learned using limited memory/history.

Studies of the "design principles" of the neural code have suggested efficiency, optimality, and robustness as central features (3, 43–45). Our results suggest a possible new feature following (43), namely that the neural code is learnable from examples. What makes the code learnable in our case is the combination of several characteristics. First, the sparseness of neuronal spiking, which has long been suggested as a design principle of the neural code (34), means that even for large networks many population activity patterns appear often enough so that interactions governing the joint activity of the network can be estimated reliably. Second, the network activity as a whole is strongly correlated. Third, although higher-order interactions are present and meaningful (e.g., ref. 27), they are few and do not go to very high orders. Finally, the limited range of direct interactions (small Markov blankets) results in the ability to learn the model from small subnetworks, thus rendering the approach scalable and perhaps applicable to much larger networks.

Based on these observations, we expect that the RI approach may be applicable to brain regions beyond the retina. Any neural population that exhibits highly correlated and sparse activity, as well as a limited range of interaction, may be a good substrate for the RI approach.

The accuracy and simplicity of the RI model in the case of the retinal neural code are related to the fact that although f_{RI} predicts the actual probability of almost all patterns we observe in the experiment, it is not a normalized probability distribution over all pattern space. Sacrificing this mathematical completeness means that we can then overcome the computational difficulty of learning probability distributions exactly. Moreover, we note that in many decision or classification problems, it is like-

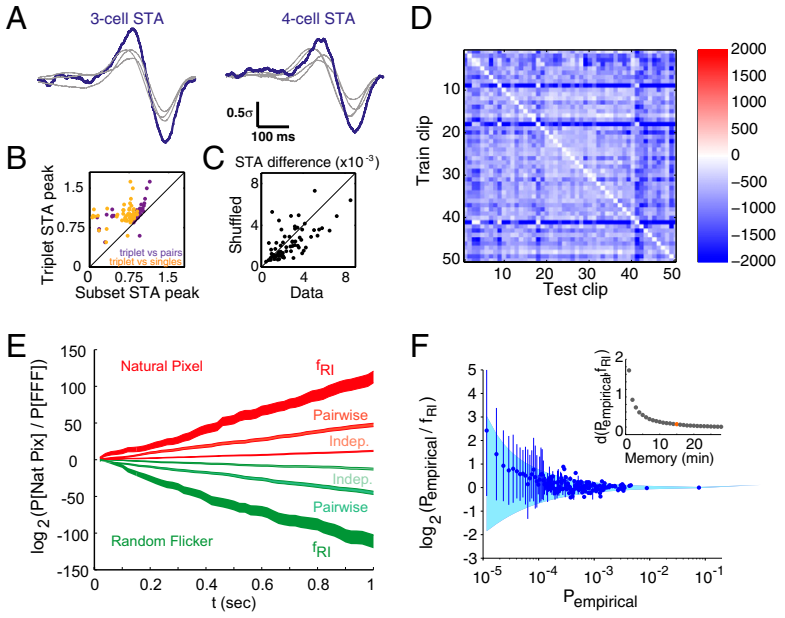

Fig. 5. Role of high-order interactions in encoding and decoding of natural stimuli. (*A*) Examples of the average stimulus preceding a specific triplet (*Left*, blue) or quadruplet (*Right*, blue) of cells spiking within 20 ms for FFF stimulus; stimuli are shown in contrast-normalized units. The STAs of the single neurons comprising the triplet/quadruplet are shown in gray. The vertical scale bar is expressed in units of stimulus STD. STA, spike-triggered average. (*B*) Absolute value of the triplet STA peak is plotted against the absolute value of the peaks of the average STAs of the subsets comprising the triplet (single neurons, yellow; pairs, purple; normalized contrast units). As in *A*, triplet STAs correspond to significantly greater changes in luminance ($P < 1e\text{-}9$, $n = 70$, two-sided paired Wilcoxon signed rank test). (*C*) Norm of the difference between pair STAs (in normalized contrast units) and the average STAs of the single neurons is plotted for real neural data (*x* axis) vs. conditionally independent surrogate data (*y* axis). Conditionally independent pair STAs were significantly more similar to the single-neuron STAs ($P < 1e\text{-}3$, $n = 57$, two-sided paired Wilcoxon signed rank test). (*D*) Accurate classification of short natural movie clips by the RI model. A 50-s natural movie was presented repetitively to the retina 101 times. We divided the movie into 50 consecutive 1-s clips, and an RI model was constructed for each clip, based on training data, for 10 randomly chosen groups of 10 neurons. Later, the likelihood of each trained model was estimated on test data taken from the different clips. The (*i*, *j*)th entry in the matrix is the log ratio between the likelihood of the data taken from clip *j* according to the model trained on clip *i* and the likelihood of the same data according to the model trained on clip *j*. Negative values indicate that the model trained on the tested clip gave a higher likelihood. (*E*) For large groups of neurons, the RI model rapidly outperforms the independent and pairwise models. The same group of neurons was presented with naturalistic (NatPix) and artificial (FFF) stimuli (*Methods*). RI models were trained on each stimulus. The average log-likelihood ratio, $\left\langle \log_2 \frac{f_{NatPix}\{x_i\}}{f_{FFF}\{x_i\}} \right\rangle_{\{x_i\}}$, is plotted for test data taken from the NatPix (red) or FFF (green) stimuli as a function of time (shaded area represents SEM over 100 repeats). Classification by $P^{(1)}$ and $P^{(2)}$ models is shown for comparison. Note that a difference of 10 in the ordinate means a likelihood ratio of over 1,000. Indep., independent. (*F*) Performance of a limited memory ("online") version of the RI model, where each interaction is estimated from the corresponding activity pattern only if it has appeared at least once every $\Delta\tau = 15$ min in the experiment (*Methods*). (*Inset*) $d(P_{empirical}, f_{RI})$ of the limited-memory RI model and the data as a function of the time window $\Delta\tau$ used for training the model. Data from the main panel are marked by the orange dot.

lihood ratios rather than the likelihood values that matter. However, unlike other pseudolikelihood models commonly used in statistics and machine learning, the RI model gives an excellent estimate of the partition function Z, because the all-zero pattern is so common.

We reiterate that the functional interactions inferred by the model do not necessarily correspond to physical connections. They may also result from joint input circuitry between cells, unrecorded neurons, or stimulus correlations. In particular, our results indicate that the higher-order interactions are strongly driven by the stimulus statistics. In addition, the abundance of higher-order interactions in the data compared with conditionally independent surrogate data indicates a contribution of retinal processing to higher-order interactions. Regardless of how higher-order interactions arise from the physical retinal circuitry, they are crucial for any observer of the retina to capture the statistics of the neural circuit response accurately. Our recordings did not capture all neurons in the retinal patch. It will be interesting to explore how dense recordings affect both higher-order interactions and the size of the Markov blankets we have found here. Based on the full-field stimuli experiments, we expect that dense populations will exhibit more highly correlated activity with a larger contribution of higher-order interactions, thus increasing the advantage of the RI model over pairwise approaches, similar to our results for NatPix stimuli.

Finally, we point out that the modular structure of the RI model, its computational efficiency, and the fact that it can be updated continuously using limited memory suggest that the RI model may also be a biologically plausible approach to learning and representing the joint activity of large neural populations (34).

Methods

Full details on experimental procedures, analysis, and modeling are presented in *SI Methods*.

Electrophysiology. Experiments were performed on adult tiger salamanders and archer fish in accordance with Ben-Gurion University of the Negev and government regulations. Retinal responses were recorded using a multi-electrode array with 252 electrodes (salamander) or 196 electrodes (archer fish). Extracellularly recorded signals were amplified and digitized at 10 kSamples/s. Datasets used in this study had simultaneous recordings of 99, 115, and 101 ganglion cells responding to natural movies; 99, 99, and 68 cells responding to spatiotemporal white noise; 90 cells responding to full-field stimuli; 99 and 83 cells responding to random phase movie from salamander retina; and 38 cells in archer fish.

Visual Stimulation. Natural movie clips were acquired using a video camera and converted to gray scale. Stimuli were projected onto the retina from a cathode ray tube video monitor. NatPix stimuli were generated by selecting a random pixel from a natural movie and displaying the intensity of that pixel uniformly on the entire screen.

ACKNOWLEDGMENTS. We thank M. Tsodyks, I. Lampl, R. Paz, O. Barak, M. Shamir, Y. Pilpel, G. Tkacik, and W. Bialek for discussions and comments on the manuscript. This work was supported by the Israel Science Foundation and The Center for Complexity Science. E.S. is supported by the Minerva Foundation, ERASysBio+ program, The Clore Center for Biological Physics, and The Peter and Patricia Gruber Foundation.

1. Perkel DH, Bullock TH (1968) Neural coding. *Neurosci Res Prog Bull* 6:221–348.
2. Georgopoulos AP, Schwartz AB, Kettner RE (1986) Neuronal population coding of movement direction. *Science* 233:1416–1419.
3. Rieke F, Warland D, de Ruyter van Steveninck RR, Bialek W (1997) *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA).
4. Schnitzer MJ, Meister M (2003) Multineuronal firing patterns in the signal from eye to brain. *Neuron* 37:499–511.
5. Stopfer M, Bhagavan S, Smith BH, Laurent G (1997) Impaired odour discrimination on desynchronization of odour-encoding neural assemblies. *Nature* 390:70–74.
6. Riehle A, Grün S, Diesmann M, Aertsen A (1997) Spike synchronization and rate modulation differentially involved in motor cortical function. *Science* 278:1950–1953.
7. Dan Y, Alonso JM, Usrey WM, Reid RC (1998) Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. *Nat Neurosci* 1:501–507.
8. Hatsopoulos NG, Ojakangas CL, Paninski L, Donoghue JP (1998) Information about movement direction obtained from synchronous activity of motor cortical neurons. *Proc Natl Acad Sci USA* 95:15706–15711.
9. Brown EN, Frank LM, Tang D, Quirk MC, Wilson MA (1998) A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *J Neurosci* 18:7411–7425.
10. Harris KD, Csicsvari J, Hirase H, Dragoi G, Buzsáki G (2003) Organization of cell assemblies in the hippocampus. *Nature* 424:552–556.
11. Averbeck BB, Lee D (2004) Coding and transmission of information by neural ensembles. *Trends Neurosci* 27:225–230.
12. Schneidman E, Bialek W, Berry MJ, 2nd (2003) Synergy, redundancy, and independence in population codes. *J Neurosci* 23:11539–11553.
13. Narayanan NS, Kimchi EY, Laubach M (2005) Redundancy and synergy of neuronal ensembles in motor cortex. *J Neurosci* 25:4207–4216.
14. Puchalla JL, Schneidman E, Harris RA, Berry MJ, 2nd (2005) Redundancy in the population code of the retina. *Neuron* 46:493–504.
15. Chechik G, et al. (2006) Reduction of information redundancy in the ascending auditory pathway. *Neuron* 51:359–368.
16. Huber D, et al. (2008) Sparse optical microstimulation in barrel cortex drives learned behaviour in freely moving mice. *Nature* 451:61–64.
17. Hopfield JJ, Tank DW (1986) Computing with neural circuits: A model. *Science* 233: 625–633.
18. LeMasson G, Marder E, Abbott LF (1993) Activity-dependent regulation of conductances in model neurons. *Science* 259:1915–1917.
19. Schneidman E, Berry MJ, 2nd, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440:1007–1012.
20. Shlens J, et al. (2006) The structure of multi-neuron firing patterns in primate retina. *J Neurosci* 26:8254–8266.
21. Tkacik G, Schneidman E, Berry MJ, 2nd, Bialek W (2006) Ising models for networks of real neurons. arXiv:0611072 [q-bio].
22. Tang A, et al. (2008) A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J Neurosci* 28:505–518.
23. Marre O, El Boustani S, Frégnac Y, Destexhe A (2009) Prediction of spatiotemporal patterns of neural activity from pairwise correlations. *Phys Rev Lett* 102:138101.
24. Shlens J, et al. (2009) The structure of large-scale synchronized firing in primate retina. *J Neurosci* 29:5022–5031.
25. Pillow JW, et al. (2008) Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* 454:995–999.
26. Martignon L, et al. (2000) Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Comput* 12:2621–2653.
27. Ohiorhenuan IE, et al. (2010) Sparse coding and high-order correlations in fine-scale cortical networks. *Nature* 466:617–621.
28. Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Francisco).
29. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. *Neural Comput* 18:1527–1554.
30. Swendsen RH, Wang JS (1987) Nonuniversal critical dynamics in Monte Carlo simulations. *Phys Rev Lett* 58:86–88.
31. Broderick T, Dudík M, Tkačik G, Schapire RE, Bialek W (2007) Faster solutions of the inverse pairwise Ising problem. arXiv:0712.2437 [q-bio.QM].
32. Cocco S, Leibler S, Monasson R (2009) Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proc Natl Acad Sci USA* 106:14058–14062.
33. Bethge M, Berens P (2008) *Near-Maximum Entropy Models for Binary Neural Representations of Natural Images* (MIT Press, Cambridge, MA), pp 97–104.
34. Olshausen BA, Field DJ (2004) Sparse coding of sensory inputs. *Curr Opin Neurobiol* 14:481–487.
35. Bair W, Zohary E, Newsome WT (2001) Correlated firing in macaque visual area MT: Time scales and relationship to behavior. *J Neurosci* 21:1676–1697.
36. Amari S (2001) Information geometry on hierarchy of probability distributions. *IEEE Trans Inf Theory* 47:1701–1711.
37. Schneidman E, Still S, Berry MJ, 2nd, Bialek W (2003) Network information and connected correlations. *Phys Rev Lett* 91:238701.
38. Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. *J R Stat Soc Series B Stat Methodol* 36:192–236.
39. Besag J (1977) Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika* 64:616–618.
40. Santos GS, Gireesh ED, Plenz D, Nakahara H (2010) Hierarchical interaction structure of neural activities in cortical slice cultures. *J Neurosci* 30:8720–8733.
41. Ganmor E, Segev R, Schneidman E (2011) The architecture of functional interaction networks in the retina. *J Neurosci* 31:3044–3054.
42. Abbeel P, Koller D, Ng AY (2006) Learning factor graphs in polynomial time and sample complexity. *J Mach Learn Res* 7:1743–1788.
43. Barlow H (2001) Redundancy reduction revisited. *Network* 12:241–253.
44. Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. *Annu Rev Neurosci* 24:1193–1216.
45. Turrigiano G, Abbott LF, Marder E (1994) Activity-dependent changes in the intrinsic properties of cultured neurons. *Science* 264:974–977.

Supporting Information

Ganmor et al. 10.1073/pnas.1019641108

SI Methods

Electrophysiology. Experiments were performed on the adult tiger salamander (*Ambystoma tigrinum*) and the archer fish (*Toxotes chatareus*). Before the experiment, the tiger salamanders were adapted to bright light for 30 min. Retinas were isolated from the eye and peeled from the sclera, together with the pigment epithelium. Retinas were placed with the ganglion cell layer facing a multielectrode array with 252 electrodes (Ayanda Biosystems) and superfused with oxygenated (95% O_2/5% CO_2) Ringer's medium, which contains 110 mM NaCl, 22 mM $NaHCO_3$, 2.5 mM KCl, 1 mM $CaCl_2$, 1.6 mM $MgCl_2$, and 18 mM glucose, at room temperature (1). The electrode diameter was 10 μm, and electrode spacing varied between 40 and 80 μm. Recordings lasting 24–30 h were achieved consistently. Archer fish retinas were isolated from the eye in the dark. Each retina was peeled from the sclera, together with the pigment epithelium, and placed in a Petri dish with a glass bottom, with the ganglion cell layer facing down. Retinas were superfused with oxygenated (97% O_2/3% CO_2) Ringer's medium at room temperature (2). An array of 196 electrodes in a fakir bed-like formation was produced by placing two Cyberkinetics 3D multielectrode arrays side by side. The array was lowered onto the retina from above using a standard mechanical manipulator. Extracellularly recorded signals were amplified (Multi Channel Systems), digitized at 10 kSamples/s on four personal computers, and stored for off-line spike sorting and analysis. Spike sorting was done by extracting amplitude and width from each potential waveform, followed by manual clustering using an in-house written MATLAB program (MathWorks). Datasets used in this study from salamander retinas had simultaneous recordings of 99, 115, and 101 ganglion cells responding to natural movies; 99, 99, and 68 cells responding to spatiotemporal white noise; 90 cells responding to full-field stimuli; and 99 and 83 cells responding to random phase movies. The archer fish retina experiment had 38 recorded cells.

Visual Stimulation. Natural movie clips were acquired using a Sony video camera (Handycam DCR-HC23) at 30 frames/s. The stimulus was projected onto the retina from a CRT video monitor (ViewSonic G90fB) at a frame rate of 60 Hz, such that each acquired frame was presented twice using standard optics (3). The original color movies were converted to gray scale using a γ-correction for the computer monitor. The checkerboard stimulus was generated by setting each checker (~100 μm on the retina) randomly every 33 ms to be either black or white. In all cases, the visual stimulus covered the retinal patch that was used in the experiment entirely. The receptive field of each cell was mapped by calculating the average stimulus pattern preceding a spike under the random checkerboard stimulation. The location of the center of each receptive field was found by fitting a 2D Gaussian field to the spatial profile of the individual cell's response. FFF stimuli were generated by sampling the projected gray level at each time point from a normal distribution in an independent and identical manner. NatPix stimuli were generated by selecting one random pixel from a natural movie and displaying the intensity of that pixel uniformly on the entire screen. The phase-scrambled movie was obtained from the original movie in the following way. First, we segmented the movie into non-overlapping segments of 2,048 frames. The phase of the Fourier transform of each segment was then randomized, such that the inverse transform was real (i.e., maintaining symmetry about the origin).

Data Analysis. Analysis was carried out in MATLAB. Spikes were binned at 20 ms (bins of 10 and 40 ms did not affect our results; Fig. S4). When model accuracy was estimated, data were separated into disjoint train and test sets, with each consisting of 50% of the relevant experimental data.

Exact Solution for the Maximum Entropy Pairwise Distribution. The maximum entropy pairwise distribution is known to take the form $P^{(2)}(x) = \frac{1}{z}\exp(\sum_{i=1}^{N}\alpha_i x_i + \sum_{i<j}\beta_{ij}x_i x_j)$ (4, 5); the parameters (α_i, β_{ij}) can be found by maximizing:

$$\begin{aligned}\Lambda(\alpha,\beta) = H\left(P^{(2)}\right) &+ \textstyle\sum_i \alpha_i\left(<x_i>_{p_{empirical}} - <x_i>_{p^{(2)}}\right)\\ &+ \textstyle\sum_{i<j}\beta_{ij}\left(<x_i x_j>_{p_{empirical}} - <x_i x_j>_{p^{(2)}}\right)\\ &+ \gamma\left(\textstyle\sum_x P^{(2)}(x) - 1\right),\end{aligned}$$

where H denotes the entropy function. This function is concave with the following derivatives: $\frac{\partial\Lambda}{\partial\alpha_i} = <x_i>_{p_{empirical}} - <x_i>_{p^{(2)}}$ and $\frac{\partial\Lambda}{\partial\beta_{ij}} = <x_i x_j>_{p_{empirical}} - <x_i x_j>_{p^{(2)}}$. The parameter values were found using gradient ascent.

The spatiotemporal pairwise maximum entropy models were built for groups of cells, such that the firing rates of each cell, $<x_i^{t_k}>$, and temporal correlation functions between them, $<x_i^{t_k}, x_j^{t_l}>$, were constrained to be identical to the ones observed experimentally (t_k indexes discrete time bins; in our case, ten 20-ms time bins from 0 to 180 ms time lag). The result is a unique maximum entropy model of spatiotemporal activity that satisfies all single and pairwise marginal distributions of the form $P(x_i^{t_k}), P(x_i^{t_k}, x_j^{t_l})$ (6).

Estimating Maximum Entropy Distributions for Large Networks. To find the maximum entropy distribution using exact methods, one must calculate the expected values of the constrained observables under the model, which are $<x_i>_{p(2)}$ and $<x_i x_j>_{p(2)}$ in our case Because the number of patterns is exponential in the number of neurons in the population, it quickly becomes unfeasible to calculate the expected values exactly. To estimate the expected values for such networks, we used Monte Carlo methods (7). We used a gradient ascent algorithm, applying a combination of Gibbs sampling and importance sampling (7) to estimate the gradient efficiently. Sampling was carried out in parallel on a 21-node cluster with 160 Intel cores in total. Our algorithm is similar to the one described by Broderick et al. (8) and was much more efficient for this specific task than previously suggested algorithms (9, 10). The optimization was terminated when the average error in firing rates and coincident firing rates reached below 1% and 5%, respectively, which is within the experimental error. To estimate the partition function, we generated samples from the desired distribution using Gibbs sampling and defined our estimate as $\widehat{Z} = \frac{1}{p(00\ldots0)}$, because it holds that $P(00\ldots0) = \frac{1}{Z}$. This estimate proved better than several other estimators we tested (11, 12).

Simulated Data. Simulated data were generated by sampling from a pairwise maximum entropy distribution that was initially defined to have similar expected values to those observed in experiments. Samples were generated using Gibbs sampling. Sample size was similar to that of our experimental dataset.

Error Measures. Error estimates for pairwise model parameter values were generated using a bootstrap approach. The dataset was resampled with repetitions, and the optimization procedure was carried out for each bootstrap sample. Error estimates for the RI model parameters were obtained by using different random train and test data.

1. Meister M, Pine J, Baylor DA (1994) Multi-neuronal signals from the retina: Acquisition and analysis. *J Neurosci Methods* 51:95–106.
2. Segev R, Schneidman E, Goodhouse J, Berry MJ, 2nd (2007) Role of eye movements in the retinal code for a size discrimination task. *J Neurophysiol* 98:1380–1391.
3. Puchalla JL, Schneidman E, Harris RA, Berry MJ, 2nd (2005) Redundancy in the population code of the retina. *Neuron* 46:493–504.
4. Jaynes ET (1957) Information theory and statistical mechanics. *Phys Rev* 106:620–630.
5. Schneidman E, Berry MJ, 2nd, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440: 1007–1012.
6. Marre O, El Boustani S, Frégnac Y, Destexhe A (2009) Prediction of spatiotemporal patterns of neural activity from pairwise correlations. *Phys Rev Lett* 102: 138101.
7. Neal RM (1993) *Probabilistic Inference Using Markov Chain Monte Carlo Methods.* (University of Toronto , Toronto, ON, Canada), report no. CRG-TR-93-1.
8. Broderick T, Dudik M, Tkacik G, Schapire RE, Bialek W (2007) Faster solutions of the inverse pairwise Ising problem. arXiv:0712.2437. Available at http://arxiv.org/abs/0712.2437.
9. Swendsen RH, Wang JS (1987) Nonuniversal critical dynamics in Monte Carlo simulations. *Phys Rev Lett* 58:86–88.
10. Wolff U (1989) Collective Monte Carlo updating for spin systems. *Phys Rev Lett* 62: 361–364.
11. Potamianos GG, Goutsias JK (1993) Partition function estimation of Gibbs random field images using Monte Carlo simulations. *IEEE Trans Inf Theory* 39:1322–1332.
12. Potamianos G, Goutsias J (1997) Stochastic approximation algorithms for partition function estimation of Gibbs random fields. *IEEE Trans Inf Theory* 43:1948–1965.

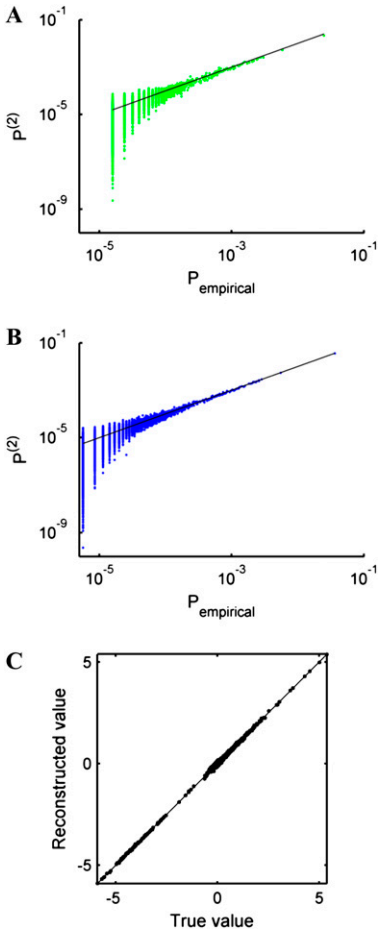

Fig. S1. Underlying model can be accurately reconstructed for spatiotemporal white noise and for surrogate data. (*A*) Fitted model predictions for simultaneous activity patterns of 99 neurons (same as in Fig. 1) responding to artificial white noise stimuli. The observation frequency of each pattern occurring in the recorded data (abscissa) is plotted against the probability assigned to the same pattern by the pairwise model fitted to the firing rates and pairwise correlations (ordinate). Compare with Fig. 1*D*. (*B*) Surrogate dataset of 100 artificial neurons was constructed by sampling from a pairwise distribution. The statistics of the pairwise distribution were similar to those measured experimentally (in response to natural movies), and the dataset was the same size as our experimental data. The firing rates and correlations were then fit with a pairwise model. Other details are as in *A*. (*C*) Comparison of the parameters used to generate the surrogate data (abscissa) with the parameters recovered using our fitting algorithm (ordinate) (*Methods*).

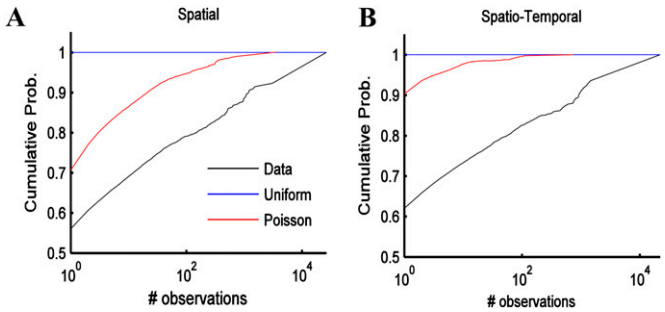

Fig. S2. Pattern counts for different joint activity distributions. (*A*) Cumulative distribution of the number of times each spatial pattern of activity was observed in the experiment is plotted in black. Also plotted are the cumulative distributions expected if neuron spiking activity followed a Poisson distribution with the same firing rate for each neuron as in the data (red) and for a distribution in which all activity patterns had the same probability (blue). (*B*) Same as in *A*, except for spatiotemporal activity patterns. Clearly, the data display many more frequently occurring patterns than any of the model distributions. Prob., probability.

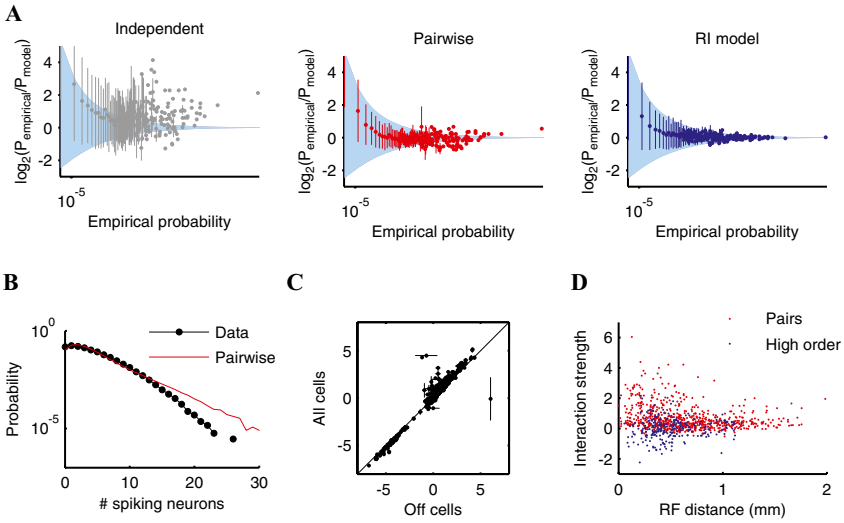

Fig. S3. RI model for the "Off" cell subpopulation. (*A*) Model predictions for simultaneous activity patterns of 64 Off cells (subset of the population in Fig. 2*A*). For all patterns occurring in the data with log probability *P* (abscissa), we plot the mean and STD of the log prediction error (ordinate) of the independent model (*Left*, gray), pairwise model (*Center*, red) and RI model (*Right*, blue). The shaded area marks the estimated 95% confidence interval of the empirical error. Compare with Fig. 2*A*. (*B*) Distribution of synchronous spiking events as measured from the data (black) and as predicted by the pairwise model (red) for the subpopulation of Off cells. Results are very similar to those obtained for the full population in Fig. 1*C*. (*C*) RI model interaction parameters learned from the entire population (as in Fig. 3 *A*–*C*) are plotted against the values of the same interactions, when present, between Off cells inferred from the activity of the Off subpopulation alone. (*D*) RI model interactions are plotted as a function of receptive field distance between neuron pairs (red) and average pairwise distance for higher order interactions (blue). Compare with Fig. 3*D*.

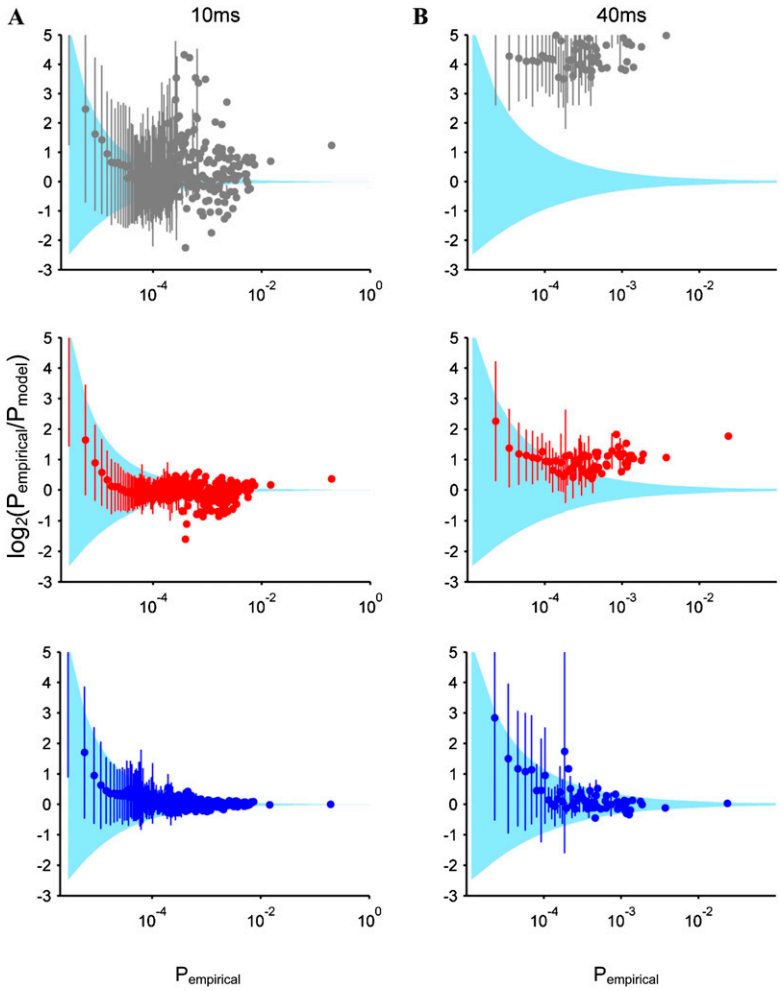

Fig. S4. Performance of the RI model is not sensitive to the size of time bin used. The effect of bin size was tested on the independent, pairwise, and RI models. (*A*) Model predictions for simultaneous activity patterns of 99 neurons binned at 10 ms. For all patterns that occur in the data with log probability *P* (abscissa), we plot the mean and STD of the log prediction error (ordinate) of the independent model (*Top*, gray), pairwise model (*Middle*, red) and the RI model (*Bottom*, blue). The shaded area marks the estimated 95% confidence interval of the empirical error. The learning threshold for the RI model was set to 20 observations (1,049 parameters) to account for the increase in the number of samples compared with the main text (20-ms bin). (*B*) Same as in *A*, except for bin size of 40 ms. The learning threshold for the RI model was set to 5 observations (312 parameters) to account for the reduction in the number of samples compared with the main text (20-ms bin is shown in Fig. 2 *A* and *B*). As bins grow larger, some bins may contain more than one spike; in our case, for 10-ms, 20-ms, and 40-ms bins, 0.12%, 0.59%, and 1.8% of the bins contained more than one spike, respectively.

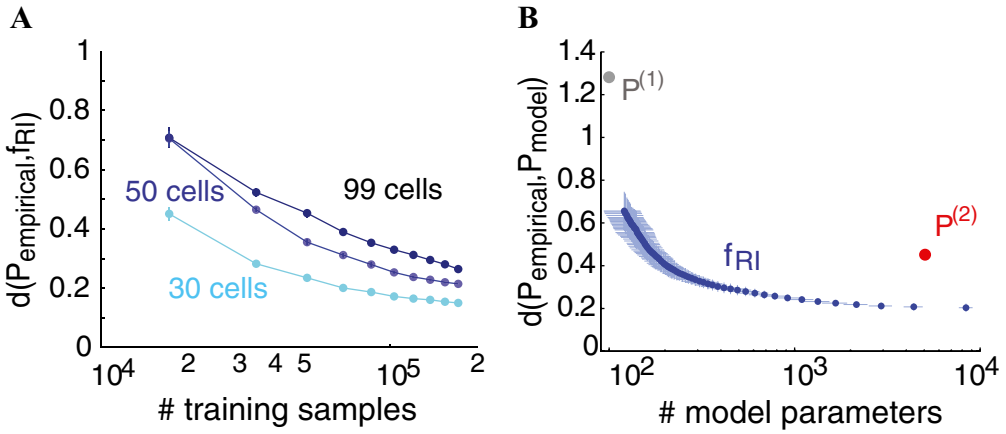

Fig. S5. (*A*) Accuracy of the RI model as a function of the number of training samples. Average $d(P_{empirical}, f_{RI})$ is plotted against the number of samples used to train the model. Error bars represent STD over 10 randomly selected train and test sets. The RI model's accuracy does not saturate, given the number of samples in our dataset. (*B*) Accuracy of the RI model for spatiotemporal patterns as a function of the number of parameters in the model. $d(P_{empirical}, f_{RI})$ is plotted against the number of parameters in the model for the spatiotemporal activity patterns of 10 neurons over ten 20-ms time bins. Results were averaged over 10 randomly selected train and test data (error bars represent STD). The maximum entropy pairwise model (red dot) and the independent model (gray dot) are also shown for comparison. The number of parameters was varied by changing the learning threshold, n_{RI}. There are no signs of overfitting, even when we learn from patterns that appeared only twice in the experiment. Spatial patterns are shown in Fig. 4*D*.

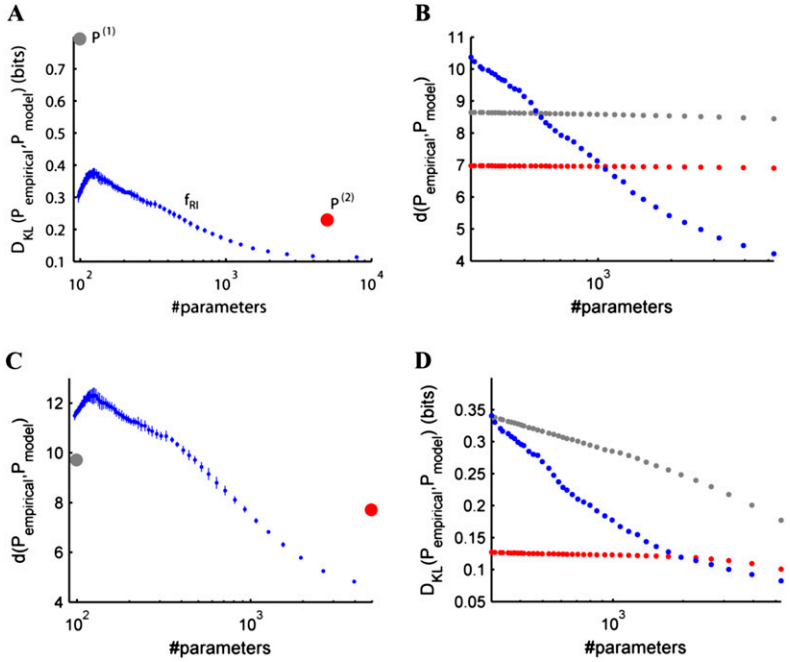

Fig. S6. Superior performance of RI models is similar under different distance measures used for assessment. In the main text, the distance measure used to evaluate the models' performance is $d(P,Q) = \sum_{X:P(X)>\frac{1}{N}} P(X)\left|\log_2\left(\frac{P(X)}{Q(X)}\right)\right|$ (N denotes the number of data samples). This measure differs from the more common Kullback–Leibler divergence $\left(D_{KL}(P\|Q) = \sum_X P(X)\log_2\left(\frac{P(X)}{Q(X)}\right)\right)$ in two ways: The sum is only over patterns that are observed more than once, and the absolute value of the logarithm is taken. The absolute value of the logarithm was taken because the RI model is not normalized; thus, it is important to penalize errors in which the model overestimates the empirical probability (such events will actually reduce the Kullback–Leibler divergence) as well as events in which the model underestimates the probability. We only used patterns that appeared more than once because the error estimates for patterns that appear only once are very large and their probability cannot be accurately estimated or modeled even if the distribution is known a priori. (For the surrogate data generated from a strictly pairwise distribution and fit by a pairwise distribution, presented in Fig. S1, the mean prediction error for patterns that appeared only once was outside our estimated confidence interval. This is not attributable to a bad fit but, rather, to large errors in empirical sampling and very conservative confidence estimates). (*A*) Accuracy of the RI model as a function of the number of parameters as quantified by the estimated Kullback–Leibler divergence (over patterns that appear more than once, train and test data kept separate) as in Fig. 4*D*, only we used the modified distance measure there. The independent model (gray dot) and pairwise model (red dot) are also shown for comparison. Error bars (approximately the same size as the dots) represent STD over different train and test sets. D_{KL}, Kullback–Leibler divergence. (*B*) Accuracy of the RI model was verified using a second method of cross-validation. Here, we measured the accuracy of the RI model using the *d* measure but tested the model only on patterns that did not reach the learning threshold [i.e., were not used to train the model (including patterns that appear only once)]. The accuracy of the independent model (gray) and pairwise model (red) is also shown for the same set of patterns used to test the RI model. (*C*) Accuracy of the RI model as a function of the number of parameters as quantified by the *d* measure defined above, including patterns that appear only once. Compare with Fig. 4*D*, in which patterns that appeared only once were excluded. Error bars represent STD over different train and test sets. (*D*) Same as in *B* but excluding patterns that appear only once.