From serial to parallel: predicting synchronous firing of large neural populations from sequential recordings

Oleksandr Sorochynskyi,[1] Stéphane Deny,[2] Olivier Marre,[1,3] and Ulisse Ferrari[1,3]

[1]*Sorbonne Université, INSERM, CNRS,*
Institut de la Vision, 17 rue Moreau, 75012 Paris, France.
[2]*Neural Dynamics and Computation Lab, Stanford University, California*
[3]*Equal contribution*

A major goal in neuroscience is to understand how populations of neurons code for stimuli or actions. While the number of neurons that can be recorded simultaneously is increasing at a fast pace, in most cases these recordings cannot access a complete population: some neurons that carry relevant information remain unrecorded. In particular, it is hard to simultaneously record all the neurons of the same type in a given area. Recent progress has made possible to determine the type of each recorded neuron in a given area thanks to genetic and physiological tools. However, it is unclear how to infer the activity of a full population of neurons of the same type from sequential recordings across different experiments. Neural networks exhibit collective behaviour, e.g. noise correlations and synchronous activity, that are not directly captured by a conditionally-independent model that would just pool together the spike trains from sequential recordings. Here we present a method to build population activity from single cell responses taken from sequential recordings, which only requires pairwise recordings to train the model. Our method combines copula distributions and maximum entropy modeling. After training, the model allows us to predict the activity of large populations using only sequential recordings of single cells. We applied this method to a population of ganglion cells, the retinal output, all belonging to the same type. From just the spiking response of each cell to a repeated stimulus, we could predict the full activity of the population. We could then generalize to predict the population responses to different stimuli and even to different experiments. As a result, we were able to use our approach to construct a synthetic model of a very large neuronal population, which uses data combined from multiple experiments. We then predicted the extent of synchronous activity and showed it grew with the number of neurons. This approach is a promising way to infer population activity from sequential recordings in sensory areas.

Introduction

A major goal of neuroscience is to understand how populations of neurons process sensory stimuli. This understanding is limited because, among other reasons, accessing the activity of all neurons of a sensory structure is very challenging. Most techniques only give access to a small fraction of neurons [1, 2] (but see [3, 4]), leaving as hidden variables many neurons that may play a role in information processing but are not recorded.

To overcome this issue, an emerging, 'divide and conquer' approach is to first classify the neurons in a given area into different cell types, where neurons of the same type are supposed to be functionally identical. Then, in a second step, one can characterize the neuronal function of each cell type, to eventually predict how populations composed of all the neurons of the same type will respond to sensory stimuli.

There has been tremendous progress recently in achieving the first step of this approach. Several studies have shown that it is possible to cluster cells in different homogeneous types [5]. This can be done using either the responses of each cell to several standard stimuli [6–8], or using genetic tools [9, 10]. These methods have proven successful in isolating most cell types in the retina [7, 11, 12] and there are several ongoing studies trying to apply these approaches in the cortex [13].

For the second step, many studies have tried to model and predict how neurons of a single type respond to complex stimuli. This strategy has been applied in the retina [14–18] and in many low-level areas [19, 20]. A complementary strategy is to use mouse lines expressing GFP in specific cell types in order to record sequentially (i.e. repeatedly across different experiments) from cells that are functionally identical. This has been performed in the retina [6, 21] and enables one to presents as many stimuli as desired to all cells belonging to the same type. It is thus possible to gather a lot of information about how single neurons of a given well-defined cell type will respond to many different sensory stimuli, using sequential recordings of neurons of the same type taken from different experiments.

Extensive characterization of single cell responses to sensory stimuli is thus possible. The next challenge is to infer how the entire ensemble of neurons of a single type responds together to stimuli. Ideally, one would like to record from all the neurons of a given type, but this is rarely possible.

An open challenge is to use these sequential recordings from cells of the same type to reconstruct how the entire population will respond. Reconstructing the activity of a full population from sequential recordings cannot be done by simply pooling the responses to a given stimulus from many sequential recordings. In many cases, pairs of neurons are correlated due to shared noise (noise correlation), which might significantly reshape the neurons' activity [22], and play an important role in information encoding and transmission [15, 23–28]. Because these noise correlations cannot be predicted from sequential recordings, a model is needed to predict them and therefore to infer the activity of a full population of neurons of the same type.

Here we address this issue and propose a method to infer the activity of an entire population of neurons of the same type from sequential recordings. Our method assumes that we have access to many single cell recordings gathered from different experiments of neurons of the same type, where the same stimulus has been displayed, and additionally to a few recordings of pairs of neurons of the same type. We use these data to reconstruct the activity of a large population of neurons, with a shared noise consistent with the pairwise recordings.

We apply this method in the retina, where the method can be validated with large-scale recordings of neurons of the same type.

Previous works have tried to model and predict the activity of ensembles of neurons in the retina [15, 29–32]. However, they were fitted and tested on ensemble of neurons recorded simultaneously, and it is unclear if they can generalize to predict correlated activity across experiments, a critical feature necessary to reconstruct activity from sequential recordings. Noise correlations are difficult to predict from one experiment to another because overall firing rates will vary between experiments, and this variation will make these models unlikely to predict noise correlations across experiments.

We show that an inference method based on copula distributions [33–36] allows one to predict noise correlations between two neurons of the same type also when the stimulus statistics are changed. By modeling the relation between noise correlations and the physical distance between neurons, we manage to predict noise correlations even across different experimental preparations. From this estimation of pairwise correlations we then use a time-dependent maximum entropy model [30–32] to infer the activity of the full population of neurons of the same type. We show that this method is accurate and reproduces several features of the recorded population activity. We then apply our method to infer the activity of a large population of neurons of the same type, beyond what can be currently recorded experimentally. Thanks to our inference method, we can estimate the extent of synchronous firing in such large population.

As soon as sequential recordings of many cells of the same type will be available also in cortical systems, it will be possible to apply our method to reconstruct the activity of large populations even beyond the retina.

Results

Overview of the inference method

The purpose of our method is to reconstruct the activity of a population of neurons of the same type from their individual responses to a same stimulus. Part of this population activity is directly accessible from sequential recordings, but another part needs to be predicted. For example, if we have recorded sequentially two neurons responding to the same stimulus, a naive solution is to pool together their responses as if they had been recorded at the same time. If the noise present in these responses is independent between the two neurons, this is indeed equivalent to record them together. However, in many cases, the noise between different neurons is correlated. In that case, pooled sequential recordings are not equivalent to simultaneous recordings [22], and the difference is what is usually termed the noise correlation between these two neurons.

Our method aims at inferring these noise correlations from parsimonious pairwise recordings of a few cells, and use them to predict how noise will be correlated across an entire population of hundreds of neurons. From this we can reconstruct how a large population of neurons would respond if they were recorded simultaneously, based on sequential recordings.

Our method is divided in two steps. First, we use a ~~novel~~ model based on copula distributions [33–35] to predict noise correlations between pairs of neurons. Our method allows predicting noise correlations for the same pair of neurons responding to a different type of stimulus, and can generalize to predict noise correlations for another pair of neurons of the same type recorded in a different experiment. This first step thus allows predicting noise correlation between any pair of neurons of this type. In a second step, we use a time-dependent maximum entropy model [30–32] to generalize from pairs of neurons to a full population.

Here we apply this method to population recordings in the rat retina where up to 32 neurons of the same type were recorded simultaneously. These data allow us testing if our reconstruction of the population activity is accurate.

Strong noise correlations between nearby OFF retinal ganglion cells.

We recorded rat retinal ganglion cells (RGCs) in response to different visual stimuli. We used a previously described method [18] to divide them in different types. Briefly, we clustered their responses to a full field flicker and isolated a single type of OFF-alpha ganglion cells. The receptive fields of these cells tiled regularly the visual field (Fig. 1B). All these cells responded reliably to a checkerboard stimulus (Fig. 1A). The cell responses to random checkerboard have been used to estimate their receptive fields, to spatially locate them, and to reconstruct the population mosaic (Fig. 1B).

To estimate noise correlation between pairs of cells, we computed the cross-correlation of their spike count and the cross-correlation of their firing rate (the mean over stimulus repetitions of the spike count, respectively green and black lines in Fig. 1C inset). At short time-scales, spike-count correlation is larger than that of firing rates , but only for nearby pair of cells (see Fig. 1C). We call noise-correlation the difference between the zero-lag cross-correlation of spike counts and firing rate (see Methods). We observed a similar behavior for all visual stimulations and experiments.

In the following we use these data to test our method. We first use copulas to predict the noise correlations between pairs of cells. We then use maximum entropy model to reconstruct the activity of a large population of ganglion cells from sequential recordings.

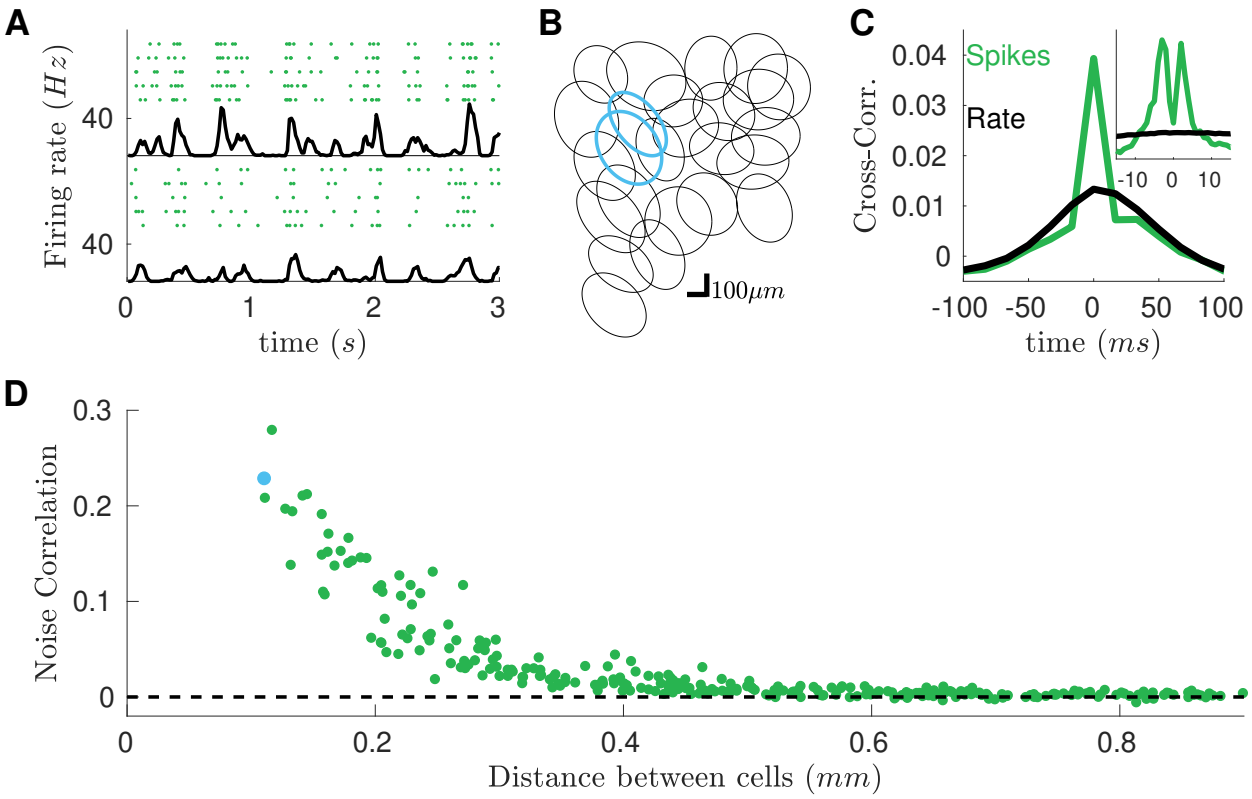

FIG. 1: **High noise correlation between nearby RGCs subject to checkerboard stimulation.** **A**) Raster plots of two example cells in response to checkerboard stimulation. Each line corresponds to a repetition of the same visual stimulation. Black line: averaged firing rate of the cell. **B**) Receptive field mosaic of the recorded OFF cell population. Cyan receptive fields refer to the cells showed in panel A. **C**) Cross-correlation for the two cyan cells (green, $dt = 17ms$), superimposed to the cross-correlation of their firing rates (black). Inset: Cross-correlation at finer time scales ($dt = 1ms$). **D**) *Zero-lag* ($dt = 17ms$) noise-correlation plotted against the distant between cells.

Copula model predicts population response from sequential recordings. A copula is a method to build pairwise probability distributions from pairs of single-variable distributions (see Fig. 2 and Methods). We used this approach to build the joint spike count distribution of pairs of neurons, that, if marginalized, reproduces the empirical single neuron distributions. For each time-bin, for each recorded neuron, we first estimate the distribution of spike count from its response to stimulus repetitions, and from this we obtain its cumulative distribution. Next, for each pair of neurons, we fit one copula distribution to the collection of joint cumulative distributions of activity. We then draw samples from the inferred copula distribution, i.e. pairs of real numbers between 0 and 1 ($(u_1, u_2) \in [0, 1]^2$) with uniform marginal distribution. Finally, we use the inverse cumulative distributions to transform these samples into pair of integers, which follow the predicted joint spike-count distribution.

A copula is characterized by a parameter that tunes the interaction strength of the two variables, and that can be inferred from data. We found that this copula parameter only depended on the distance between the two cells, and can be fitted using a function with just

three parameters (suppl. sect. S2, Fig. S1). We could thus describe the joint activity between all pair of neurons using copulas characterized by only three parameters across the entire population of cells. Once applied on the same response to checkerboard stimulation used for training, our model predicted noise correlations with high accuracy (Pearson's $\rho = 0.99$, $n = 300$ pairs, larger than $\rho = 0.92$ obtained by fitting directly noise correlations, see supplementary sect. S3).

We have built a model with only 3 parameters, that can predict the noise correlation between any pair of neurons from the activity of single cells. We then tested if this model can generalize and predict noise correlations measured in response to different stimuli. We first inferred the copula model from the response to checkerboard stimulation (that of Fig. 1). Then, from the response to repetitions of another type of stimulus, we estimated the spike-count distributions of each neuron in each time-bin. Finally, we used our copula model to first build the joint distribution, and then compute the mean noise-correlations of each neuron pair. We applied this strategy to the RGCs' response to full-field and moving-bar stimuli, see Fig. 2. In both cases, the copula model was able to reproduce the empirical estimates of noise-correlations with high accuracy (Pearson's $\rho = 0.96$ for full-field and $\rho = 0.97$ for moving-bar, $n = 300$ pairs).

To further demonstrate the robustness of our method we tested if our copula model could predict noise correlations in a different RGC population of the same type, recorded in a different experimental preparation. We used the model inferred from the data of the first experiment to predict noise correlations between the same type of RGC, but recorded during a second experiment. Using only single cell responses, our model predictions were accurate (Fig. 3B, Pearson's $\rho = 0.96$, $n = 496$ pairs), and accounted for how noise correlations decrease with distance (Fig. 3C). The functional dependence of the copula parameter with respect to inter-cell distances is thus robust across experiments, and hence corresponds to a general property of OFF-Alpha cells in the rat retina. We obtained similar results for all the 8 testing experiments (averaged Pearson's $\rho = 0.949 \pm 0.017$, for a total of $n = 1632$ pairs).

Note that in order to predict noise correlations (Fig. 3) our model never accessed to the simultaneous recordings, but only to the collection of single neuron responses. We could have thus predicted the noise correlation between pairs of neurons in these new experiments using only the sequential recording of each neuron. Our approach thus allows predicting pairwise noise correlations across experiments without requiring simultaneous recordings.

Time-dependent maximum entropy model reconstructs the activity of large population from the copula's pairwise predictions. Our copula model predicted pairwise synchronous firing. To reconstruct the activity of a large population of neurons from single cell recordings, we then used a *time-dependent* Maximum Entropy population model.

Standard Maximum Entropy models [29] aim at predicting the probability of any spike pattern from the mean firing rate of each neuron and the correlations between each pair of cells. Here we use a recent generalization of this approach [31, 32] that takes into account a time-varying firing rate. This approach built a collection of pairwise Maximum Entropy models (one for each time-bin), which share the same couplings, but with a different mean firing rate for each cell and each time bin [31, 32]. *Time-dependent* Maximum Entropy modelling thus disentagles intrinsic interaction, due to network effects, from extrinsic correlations, due to common inputs [32].

We inferred the *time-dependent* Maximum Entropy population model from the activity of single neurons and the pairwise correlations predicted by the copula model. The inferred

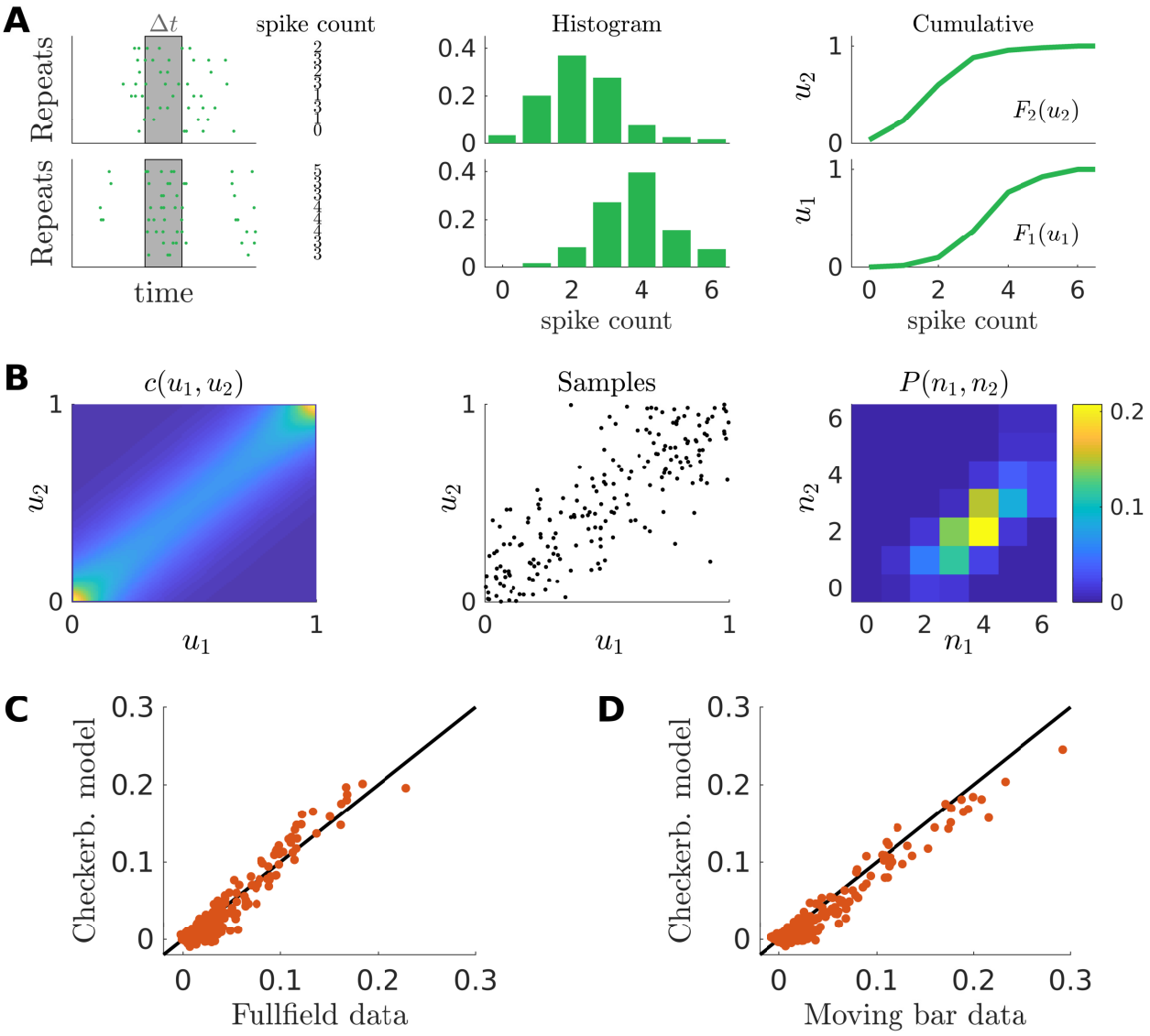

FIG. 2: **Copula model predicts noise correlations across stimulus ensembles.** **A**) Illustration of our copula model: first step. Spike counts are estimated across stimulus repetition for a given pair of cells, in a given time-bin. Empirical histogram of the spike counts are then estimated and later used to compute the empirical cumulative probability distribution. **B**) Second step. A copula distribution accounts for the mutual dependency of two random variables. From it we draw many samples of pairs of real numbers $(u_1, u_2) \in [0, 1]^2$ with uniform marginals. Inverse of the cumulative distributions transform these samples into pairs of positive integer numbers, whose distribution matches the empirical spike count marginals by construction and accounts for their mutual dependency. **C**) Scatterplot of the empirical and model predicted noise correlations fro the response to full-field stimulation. **D**) As **C** but for the moving-bar stimulation. In both case the model has been inferred from the response to checkerboard stimulation.

couplings are large only between nearby cells, see Fig. 4A, and the model reconstructs a "nearest-neighbor" interaction network, see Fig. 4B. As expected the model reproduced well

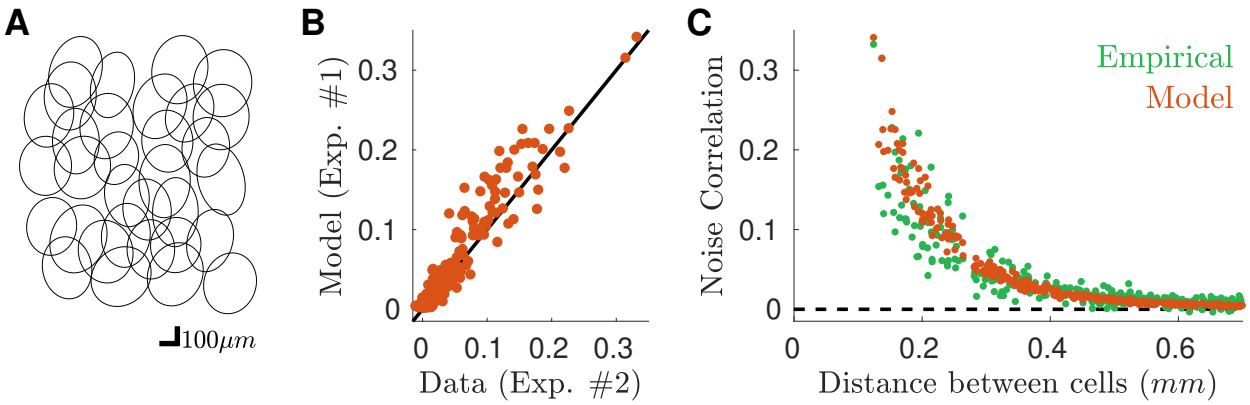

FIG. 3: **Copula model predicts noise correlations across experimental preparations.** **A**) Receptive field mosaic of the recorded OFF cell population for a new dataset, different from the one used for training the model. **B**) Scatterplot of the empirical and model predicted noise correlations from the response to checkerboard stimulation. **C**) Behavior of the empirical and model predicted noise correlations plotted against the distance between cells. Data from a second dataset (#2), different from the one used for training the model (#1).

the pairwise noise correlation (Fig. 4C, Pearson's $\rho = 0.96$, $n = 496$ pairs), which were already finely predicted by the copula model (Fig. 3B). Remarkably, Fig. 4D shows that the model also accounts for the triplet noise correlation (see Methods) (Pearson's $\rho = 0.6$, $n = 4960$ triplets). In order to show that our model captures the synchronous behavior of the neuronal population, we compute the probability distribution of the population rate, i.e. the total number of spikes emitted by the entire population in a given time-bin. We compared this distribution with the "shuffled" distribution, which destroys noise correlations, and is equivalent to the prediction made by a conditionally-independent model (i.e. a model that would assume there are no noise correlations). The distribution computed after shuffling the data overestimated the probability of number of spikes close to the average population rate, and underestimated the occurrence of transients of very large or very low activity (see Fig. 4E and F). Remarkably, our model captured the empirical behavior of the population rate averaged over the whole recording (see Fig. 4E). It also performed well when focusing only on highly active time-bins (Fig. 4F).

High synchrony in a large population of ganglion cells reconstructed from multiple experiments Thanks to our model, we could reconstruct the activity of a large population of neurons using only single cell activity. Since the model can generalize across experiments, it means that the activity of the different single cells can be taken from different experiments. Our method only needs a few pairs of neurons recorded simultaneously to fit the three parameters of the model. In the following we illustrate how this model can be used to reconstruct the activity from a very large population of neurons, bigger than what could be recorded experimentally. We illustrate how this inference of the activity of a large population of cells can be useful by measuring synchronous activity over increasing number of cells.

We collected the (marginal) response of many cells recorded during multiple experiments, and constructed a large population of $n = 256$ cells. We first built a synthetic lattice repre-

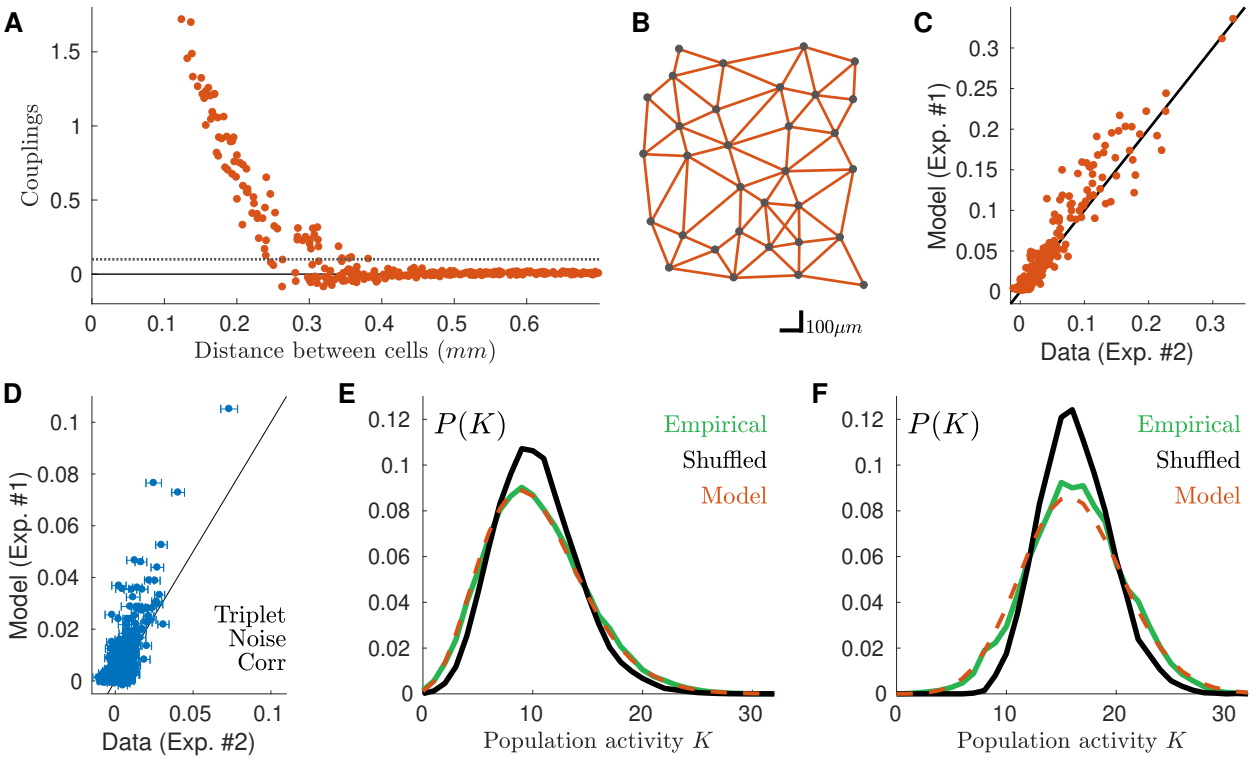

FIG. 4: **Time-dependent Maximum Entropy model predicts population synchronous firing across experimental preparations. A**) Behavior of the inferred couplings with the distance between cells. Grey dotted line threshold for identifying stong couplings (see panel **B**). **B**) Position of the cells on the retinal surface, and stong couplings linking them. In the model, only nearest neighbor cells are directly interacting **C**) Noise correlation prediction of the model against empirical value. As expected by construction, the model reproduces the copula estimations, and hence the empirical values (see Fig. 3B). **D**) Model prediction of triplet noise correlations against empirical value (Pearson correlation $\simeq 0.6$, $n = 4960$). **E**) Empirical, shuffled (cond. independent) and model distributions of the population activity averaged over all time-bins. **F**) Same as **E** but for the 5% of time-bins with the highest mean population firing rate. Data from a second dataset (#2), different from the one used for training the model (#1).

senting the positions of the cells reproducing the empirical statistics of inter-cells distances (Fig. 5A and Methods). Then, we associated to each lattice position the response to the checkerboard stimulus of a randomly chosen cell among the ones recorded in all experiments (excluding those from the experiment used to learn the model parameters). Finally we applied our two-step approach to predict how this large population of cell would have responded to a checkerboard stimulus.

At first, we looked at the population rate as previously defined. In comparison with shuffled data, our model predicted a more frequent occurrence of transients with either very high or very low population activity (Fig. 5B). This is a signature that synchronous activity extends up to large populations. To study how correlated firing grows with the population size, we sub-sampled the synthetic model (Fig. 5A) and computed the variance of the population rate. The variance predicted by the model grows with the number of neurons

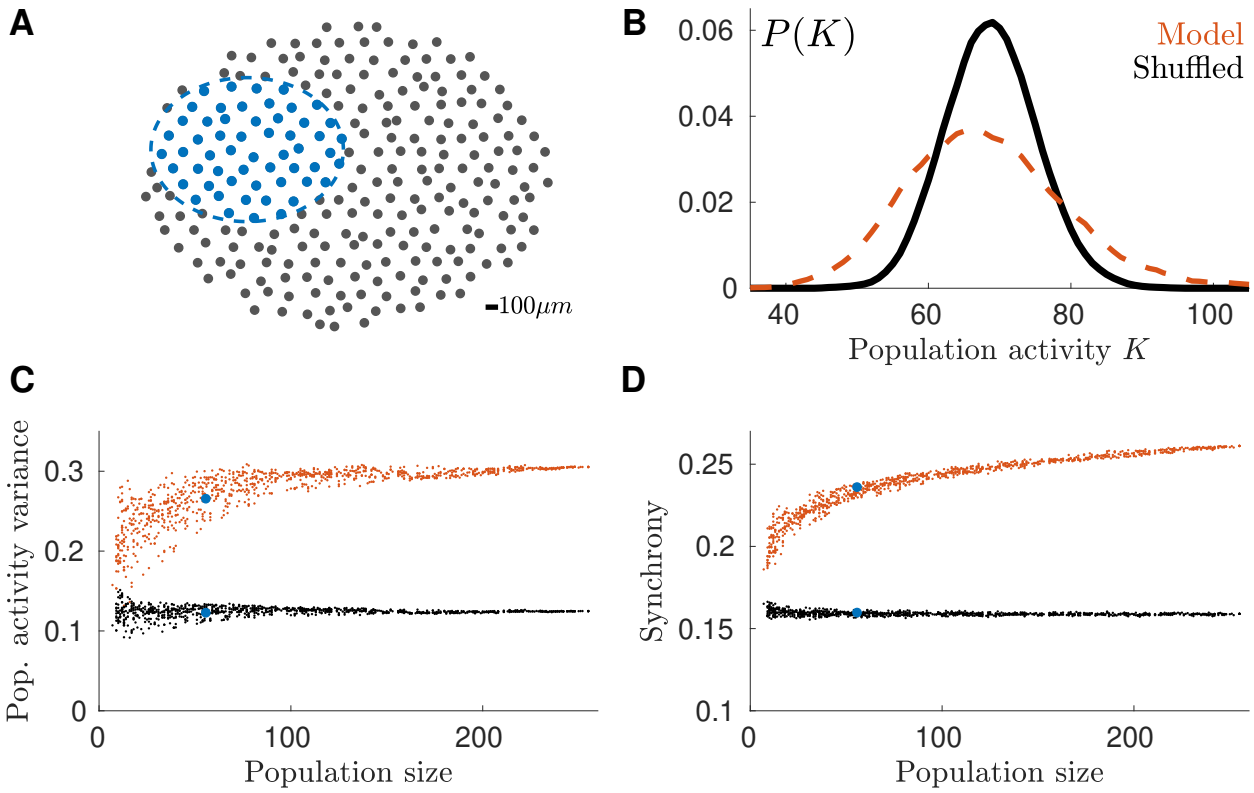

FIG. 5: **Large population model.** **A**) Synthetic mosaic of $n = 256$ cells. Blue: example of sub-population. **B**) Probability distribution of the population activity for the time-bin with the highest firing rate. **C**) Average over time of the population activity variance for many sub-populations. **D**) Behavior of Synchrony with the number of cell in the sub-population.

(Fig. 5C) and saturates at ~ 200 neurons (population activity variance is 0.24 ± 0.02 at $N \simeq 30$ cells and 0.299 ± 0.002 at $N \simeq 200$ cells , mean $\pm$ s.d.), much larger than the typical number of cells recorded in an experimental session. On the contrary, in the reshuffled control, the variance is smaller and roughly constant when the number of neurons increases. Next we estimated the synchrony in the population as the probability of observing a transient with large population activity (see Methods). This synchrony grew fast for small number of cells and stopped increasing at ~ 200 cells (synchrony is 0.221 ± 0.004 at $N \simeq 30$ cells and 0.257 ± 0.001 at $N \simeq 200$ cells , mean $\pm$ s.d., Fig. 5D) .

Discussion

We have developed a model to predict noise correlations across experiments. For any new pair of cells we just needed the single cell response of each cell, which can be accessed in the sequential recordings. Thanks to this prediction we could then reconstruct the activity of large populations of neurons of a single type at a scale beyond what can be recorded experimentally. We have applied our method in the retina, where recording populations of neurons of the same type is possible. Our method could thus be validated by comparing the prediction of our model to real data. Using this method, we have shown that synchrony in the population of OFF-alpha ganglion cells grows with the number of cells. Our method allows thus giving a new insight about the coding of information by large population of neurons, at a scale currently not accessible with experimental recordings.

Previous works have shown how different methods can be used to model and predict noise correlations. GLM model [15] uses spike history filter to couple the spiking activity between different neurons. Stimulus dependent maximum entropy models [31, 32] have coupling terms to model synchronous activity between pairs of neurons. These models have been successfully used to model how a population of neurons responds to a stimulus ensemble. However, they usually fail in trying to predict the population activity in responses to different stimuli. This is because the parameters generating the correlations usually change when learned on different stimulus statistics [37]. Moreover, they were never used to predict the activity of a population across different experiments. A main obstacle for this is the fact that the firing rates can vary from experiments to experiments, which should induce parameter changes in most of these models, and impeding generalization. Here we have found that the noise correlation could be predicted using our model knowing just the distance between the two cells and their single cell activity. We could then predict noise correlations across experiments. Having a model that generalizes across experiments is crucial to pool together recordings from neurons of the same type and get a picture of the full population activity.

Our approach relies on the ability to predict noise correlations across different conditions. A naive approach could be to assume that noise correlations remain similar across experiments, as a fixed stereotyped behavior. We constructed this naive model by fitting an exponential function over our training dataset. This approach, however, gave significantly lower results than our method based on copulas (see supplementary sect. S3), demonstrating that our approach capture non-trivial properties of the retinal correlated firing. Another striking feature of our model is that the copula parameters do not depend on time: when we relaxed this assumption, copular parameters for a given pair of neurons varied little with time (see supplementary sect. S2). This effect shows that the same copula distribution accounts for the mutual dependency of the cells spike-counts, independently of the instantaneous firing rate.

Copulas have rarely been used in neuroscience studies [34–36], but none of them applied this method to predict noise correlations. In [34], Pillow and co-workers proposed for the first time discrete copula distributions to model the total spike-count correlation in pre-motor cortex neurons. However, they did not distinguish stimulus from noise correlations as we have done here.

Our approach is general and could be applied in any sensory area. We applied it in the retina, where it is possible to have recordings of many neurons of the same type. We could thus validate our method and show that excess synchrony (i.e. beyond what could

be predicted by a conditionally-independent model) increased with the number of neurons, and will become less negligible when looked at larger population sizes. Recent technological advances will make this method relevant to understand cortical populations, where it should be soon possible to define the cell type of each recorded cell using genetic (e.g. single cell transcriptomics [9, 10]) and physiological (e.g. clustering of responses [7]) tools.

Our method can be applied on neurons in sensory cortices, provided that sequential recordings of the same type of cells are available, as well as some pairwise recordings to fit the model parameters. One issue could arise if noise correlations strongly depend on the stimulus, as it has, for example, been reported for V1 [38]. In our data, noise correlations depend very little on the stimulus, which allowed us to reduce our model and let copula parameters depend only on the distance between cells. However, if noise correlations depend largely on the stimulus, the model can be extended. The simplest solution would be to make the copula parameters depend on the stimulus. If this stimulus dependence can be explicitly modeled, our method would still manage to predict the activity of large ensembles of neurons. Our method has thus the potential to be useful in the cortex to infer the activity of large populations from sequential recordings.

Finally, here we predicted the responses of a population of neurons to a stimulus for which we have access to single cell responses across stimulus repetitions. If a model is available to predict the responses of single cell to other stimuli, it would be possible to combine this model with our approach to predict the activity of large populations. This makes our method complementary to recent efforts trying to model and predict accurately the responses of single neurons in sensory cortices [17, 39, 40].

Methods

a Multi-electrode array recordings. We analyze the response of rat RGCs to visual stimulation recorded in 9 multi-electrode array *ex-vivo* experiments [3], and spike sorted with *SpyKING CIRCUS* [41]. This dataset and the experimental methods have been already previously described [18]. In one experiment we probed the retinal response to three different visual stimuli: (i) a random black and white checkerboard, with spatio-temporal uncorrelated checkers; (ii) a full-field stimulus with fluctuating luminance and (iii) two gray horizontal bars performing an independent Brownian motion along the vertical direction [18]. In the other 8 experimental sessions, only the response to random black and white checkerboard and full-field was retained and analyzed here. Each of these stimulations lasted about $10sec$ and have been repeated at least $R = 79$ times. Spiking times have been binned with a window of about $17ms$, corresponding to a bin rate of $60Hz$. With a custom algorithm [18] -similar to that of [7]- we used the cell's response to the full-field stimulation to identify the type of the recorded RGCs. Across the 9 experiments, we identified populations of 20 ± 6 (mean $\pm$ s.d.) OFF-Alpha cells.

b Stimulus and noise correlations. After binning the spiking response of each cell, we estimate $n_i^{(t,r)}$, the number of emitted spike by cell i, in the time-bin t, during repetition r, and its mean across repetitions $\mu_i^{(t)}$. Then we calculate the total covariance between two neurons (i, j) as follows:

$$Cov_{total}(n_i, n_j) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{R}\sum_{r=1}^{R}(n_i^{(t,r)} - \mu_i)(n_j^{(t,r)} - \mu_j) \tag{1}$$

Where $\mu_i = \sum_{t=1}^{T}\mu_i^{(t)}/T$ is the mean number of spikes across repetitions, and then averaged in time. It is possible to decompose the total covariance into a sum of the so called "stimulus" and "noise" covariances. We calculated these quantities as follows

$$Cov_{noise}(n_i, n_j) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{R}\sum_{r=1}^{R}(n_i^{(t,r)} - \mu_i^{(t)})(n_j^{(t,r)} - \mu_j^{(t)}) \tag{2}$$

$$Cov_{stimulus}(n_i, n_j) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{R}\sum_{r=1}^{R}(\mu_i^{(t)} - \mu_i)(\mu_j^{(t)} - \mu_j) \tag{3}$$

Noise correlations are then estimated as:

$$Corr_{noise}(n_i, n_j) = \frac{Cov_{noise}(n_i, n_j)}{\sqrt{V_i V_j}} \tag{4}$$

where $V_i = Cov_{total}(n_i, n_i)$.

Triplet noise correlations are instead defined as:

$$Corr_{noise}(n_i, n_j, n_k) = \frac{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{R}\sum_{r=1}^{R}(n_i^{(t)} - \mu_i^{(t)})(n_j^{(t)} - \mu_j^{(t)})(n_k^{(t)} - \mu_k^{(t)})}{\sqrt{V_i V_j V_k}} \tag{5}$$

c Copulas. Copula-based modeling allows for disentangling the marginal distributions of two random variables from their mutual dependency, that can therefore be modeled alone, without the additional difficulties due to potentially complicated marginal distributions. Consider two random variables X and Y, with joint distribution $f_{X,Y}$, marginal distributions f_X and f_Y and marginal cumulative density function (c.d.f.) F_X and F_Y respectively. By construction, the random variables $U_X \equiv F_X(X), X \sim f_X$ and $U_Y \equiv F_Y(Y), Y \sim f_y$ have uniform distributions over $[0,1]$. Consequently the joint distribution of (U_X, U_Y) has uniform marginal, yet it contains all the information about the mutual dependence between X and Y. This property can be used to model the dependency of U_X and U_Y *instead* of that of X and Y. Specifically, a copula is the c.d.f. of the joint variable (U_X, U_V), i.e. a function $C(\cdot,\cdot) : [0,1]^2 \to [0,1]$. and it can be used to reconstruct the joint distribution of (X, Y), via its c.d.f.:

$$F_{X,Y}(x,y) = C\left(F_X(x), F_Y(y)\right). \tag{6}$$

Sklar's theorem (see supplementary sect. S1) ensures the existence and uniqueness of C, and this allows for modeling the mutual dependency between X and Y, independently from their marginal distributions.

In most practical situations, the copula C is chosen from a parametric copula family. In this work we chose to work with Frank copulas:

$$C_{\text{Frank}}(u,v|\theta) = -\theta^{-1}\log\big(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\big), \tag{7}$$

where $\theta \in \mathcal{R}$ is the copula parameter, that can be estimated by log-likelihood maximization. We chose the Frank copula because this family has already been showed to perform well in modeling spike counts [34]. Once θ has been inferred, the marginal distributions can in turn be approximated either with some model, or, as we will do in this work, empirically. We refer to the Mathematics section, the literature and textbooks (see, for example, [33]) for more explanations and details on copula models and/or other copula families.

d Copula-based model. We constructed a copula model able to predict $P(n_i^{(t)}, n_j^{(t)})$, the joint distribution of pair of spike counts in time:

$$P\left(\{n_i^{(t)}, n_j^{(t)}\}_{t=1}^T\right) = \prod_{t=1}^T c_{\text{Frank}}\left(F_i^{(t)}(n_i^{(t)}), F_j^{(t)}(n_j^{(t)}) \;\middle|\; \hat{\theta}(d_{ij})\right), \tag{8}$$

where c_{Frank} is the copula density function, corresponding to the c.d.f. of Eq. (7), and $F_i^{(t)}(n_i^{(t)})$ is the *empirical* c.d.f. $n_i^{(t)}$, that we estimate across repetitions, d_{ij} is the distance between neurons i and j, and $\hat{\theta}(d_{ij})$ is a parametric function: $\hat{\theta}(d) = \exp(a + bd + cd^2)$. To infer this function from data, we first inferred a copula parameter for each pair of neurons, θ_{ij}, by log-likelihood maximization, and then we obtained a, b, c by fitting the behavior of θ_{ij} with respect to the distance d_{ij}. See supplementary sect. S2 for further information.

e Time-dependent Maximum Entropy model. The time-dependent maximum entropy model we used is:

$$P\left(\{\{n_i^{(t)}\})_{i=1}^N\}_{t=1}^T\right) = \prod_{t=1}^T \left(\exp\left\{\sum_i h_i^{(t)} n_i^{(t)} + \sum_{i\le j} n_i^{(t)} J_{ij} n_j^{(t)} - \sum_i \ln(n_i^{(t)}!)\right\} \Big/ Z^{(t)}\right) \tag{9}$$

where $n_i^{(t)} \in [0, 1, \ldots, n^{\text{Max}}]$ is an integer spike-count, with n^{Max} matched from data. The index "$^{(t)}$" expresses the time dependence. $Z^{(t)}$ is a normalization constant (the partition function). $h_i^{(t)}$ is the local field for neuron i at time t imposing the firing probability and J_{ij} is the couplings network that allows for reproducing the system's correlations. Note that J_{ij} does not depend on time and it includes also the diagonal terms J_{ii} which set each neurons variance equal to its empirical value [32, 42]. The log-factorial term allows for matching the single neurons statistics [42], as by taking $J = 0$ the model reduces to a collection of independent Poisson distributions.

The inference of the model (9) is done by log-likelihood maximization using an iterative algorithm with adaptive learning rate similar to that of [43]. Thanks to its exponential form, the model inference requires only the average value of $n_i^{(t)}$ across repetitions and the value of the noise covariances (see Methods). In our case we estimate the first from the marginal response of each neurons and we used the copula model to predict the second.

f Synthetic lattice. In order to construct a synthetic lattice that respects the inter-cells distance of empirical recordings, we started by a triangular regular lattice with side $194\mu m$, and then we added a Gaussian noise with a standard deviation of $22\mu m$ to the x and y coordinate of each cells. These parameters are optimized in order to match the distribution of cells distances measured in real experimental recordings.

g Population activity variance and synchrony. In this paper we compute the population activity as the sum of all spike-counts, $\sum_i n_i(t)$. To compute the "population activity variance" of Fig. 5C we first estimate the variance of the population activity for each time-bin. Then we averaged over time and finally we normalized by the population size. "Synchrony" of Fig. 5D is instead the probability of observing an event with a population activity larger than the mean plus one standard deviation of the population activity of the shuffled model. We compute this for each time-bin and then we averaged over time.

S1. Supplementary mathematics

a Copula's definition. A copula function is the c.d.f. of a bi-variate distribution with $Uniform(0,1)$ marginals. This mathematical construct allows to model the dependency structure of bi-variate random variables separately from their marginal distributions, as the following theorem shows.

b Sklar's theorem. Let X and Y be any two, mutually dependent, real random variables. Let F_X, F_Y, and $F_{(X,Y)}$ be the c.d.f.s of X, Y, and (X,Y) respectively. Note that for any X we have $F_X(X) \equiv Uniform(0,1)$, *idem* for Y. The Sklar theorem asserts that given such X, and Y:

$$\exists! \text{ a copula } C, \text{such that } F_{X,Y}(x,y) = C(F_X(x), F_Y(y)) \tag{10}$$

c Discrete copulas. The proof of existence of Sklar's theorem holds for both continuous, and discrete random variables (as is our case). However, in practice, to apply copula models in the discrete case we need to make adjustments. In particular we turn a continuous copula, into a discrete distribution that may take values in a countable set of points in $[0,1]$. By doing so we define a so called "pseudo density" function, that describes a discrete counterpart to copulas. Assuming that $(X,Y) \in \mathcal{N}^2$, without loss of generality, the "pseudo density" of a copula defined by a copula density c is given by:

$$f_{pseudo}(F_X(x), F_Y(y)) = \int_{F_X(x-1)}^{F_X(x)} \int_{F_Y(y-1)}^{F_Y(y)} c(u,v)\, du\, dv \tag{11}$$

S2. Supplementary information: model construction

Here we fully justify how we simplified the copula model, from its bare version with one parameter for each neuron pair in each time-bin, to the final version with just three parameters in total.

As explained in the Methods section, our starting point is a copula model where for each neuron pair (i,j) and each time-bin t a copula parameter $\theta_{ij}^{(t)}$ accounts for the correlation between the two spike-counts. Despite being very accurate, this model has little capacities to generalize across stimulus conditions or experiments. Moreover, because of the large number of parameters is potentially pruned to overfitting. In Fig. S1A, for an example neuron pair, we show the values of the inferred parameters in all time-bins plotted against the mean firing rate of the two cells -in the corresponding time-bin and computed across repetitions. As can be observed, at high firing rate, that is when the statistics is large and the inference error is small, the inferred parameters tend to accumulate around a single value. This result suggested us that a model where the copula parameters does not depend on time could have a similar performance, yet requiring much less parameters.

In order to infer such time-independent copula parameters, for each neuron pair, we first select the "active" time-bins where the two neurons spiked synchronously in at least one repetition: $\sum_{r=1}^{R} n_{i,r}^t n_{j,r}^t > 0$, where $n_{i,r}^t$ is the number of spikes emitted by neuron $i \in [1,\ldots,N]$, in time bin $t \in [1,\ldots,T]$ during repetition $r \in [1,\ldots,R]$. Once the inactive time bins are filtered out, we estimate the model parameter by maximizing the likelihood:

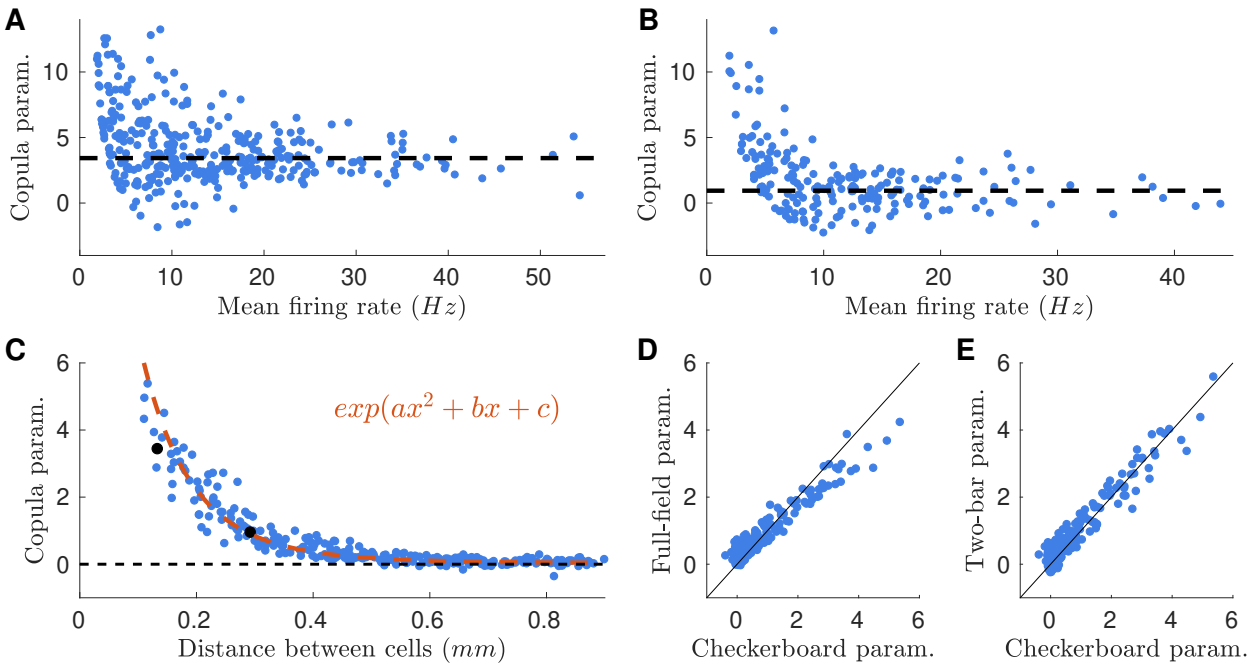

FIG. S1: **Stimulus-conditioned copula model is robust across different stimulus ensembles.** **A**) Values of the inferred copula parameter in all time-bins for an example neuron pair, plotted against the mean firing rate of the two cells. Horizontal line correspond to the inferred parameter in the *time-independent* copula model (see text) **B**) As **A**, but for another example pair. **C**) Inferred Frank copula's parameters plotted against the distance between cells for checkerboard stimulation. Parameters of each pairs are inferred independently. Orange: exponential fit used for estimating the copula parameter from the cell distance in the final version of the model. **D**) Scatterplot of the *time-independent* (see text) parameters inferred from checkerboard and full-field stimulations **E**) Scatterplot of the parameters inferred from checkerboard and two-bar stimulations

$$l(\theta|X) = \sum_t \sum_{r=1}^{R} \log(f^{(t)}_{pseudo}(n^{(t)}_{i,r}, n^{(t)}_{j,r}|\theta)) = \sum_t \sum_{r=1}^{R} \log \int_{F_i^{(t)}(n^{(t)}_{i,r}-1)}^{F_i^{(t)}(n^{(t)}_{i,r})} \int_{F_j^{(t)}(n^{(t)}_{j,r}-1)}^{F_j^{(t)}(n^{(t)}_{j,r})} c_\theta(u,v)\, du\, dv \tag{12}$$

where the summation over t runs over the active time-bins for the neuron pair. Note how F, and thus f_{pseudo}, depends on t, as the empirical marginals are estimated separately for every time bin. We infer the time-independent copula parameters θ_{ij} by log-likelihood maximization for each neurons' pair and for checkerboard, full-field and two-bar stimuli. Figs. S1B and C compare the parameters inferred from different stimuli and show how the inference is robust across changes of visual stimulation. The copula parameter thus reflect some properties of the retinal network, independent of the current stimulus ensemble.

Fig. S1C show the behavior of the inferred checkerboard parameters with respect to the physical distance between cells. Furthermore these parameters are independent of the visual stimulus (see Figs. S1D and E). These results suggested us that a simple fit of copula parameters may account for most of the variability of the parameter values across neuron

pairs.

To further simplify our copula model, and reduce the number of its parameters, we hence fitted the inferred copula parameters with a parametric function of the inter-cell distance d: $\theta(d) = exp(a + bd + cd^2)$. The copula model takes now as input only the distance between the cells, uses it to estimate the copula parameter, and then construct the joint spike-count distribution using Eq. (8).

S3. Supplementary information: simplest model for noise correlations

In this section we compare the performance of our copula-based approach in predicting noise correlations with a straightforward model that assumes distant-dependent noise correlations.

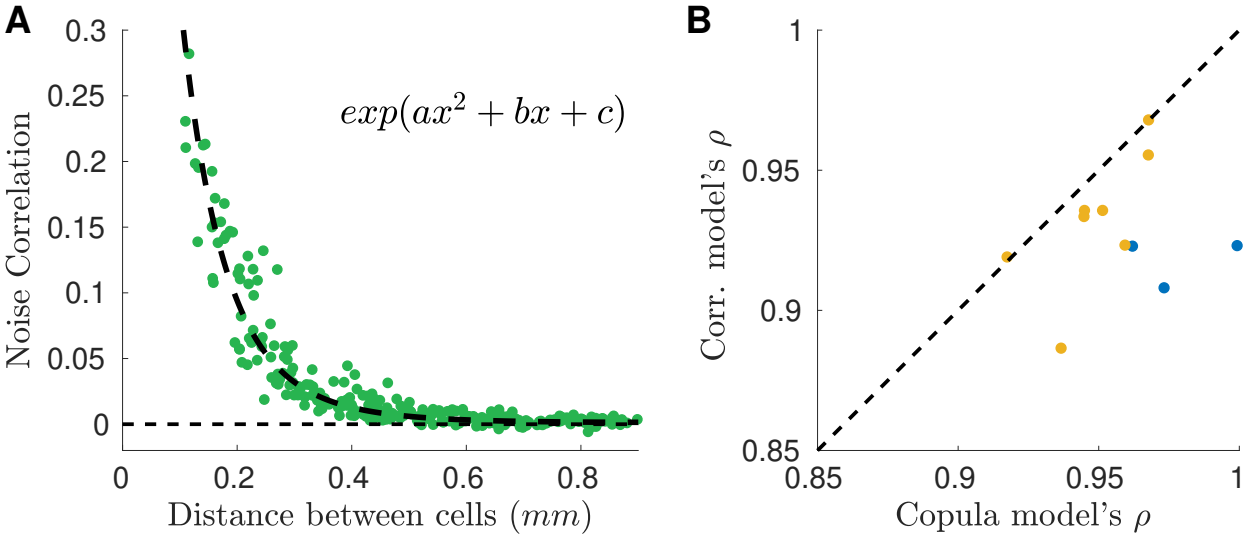

FIG. S2: **Copula model outperforms simpler model with distant dependent noise-correlations. A**) Construction of the model: noise correlations observed in one experiment (Exp. #1) are fitted with an exponential function of the distance between neurons. Such fit is then used to predict noise correlations in other experiments. **B**) Performance in predicting noise correlation for our copula model against the model presented here. Blue points: performance for the first dataset with three different stimulus ensembles. Yellow points: performance for the other experimental sessions.

Noise correlations decrease with the distance between the corresponding neurons (see Fig. 1 and Fig. S2A). We fit this relation with an exponential function, and we asked to what extent this behavior is conserved across experiments. To estimate this, we used this simple method to predict noise correlations in all the experiments described before. Although the predictions were accurate, our copula model outperforms this simpler approach (see Fig. S2B).

Acknowledgments

We like to thank M. Chalk for useful discussions. This work was supported by ANR Trajectory, the French State program Investissements dAvenir managed by the Agence Nationale de la Recherche (LIFESENSES; ANR-10-LABX-65), EC Grant No. H2020-785907 from the Human Brain Project, NIH Grant No. U01NS090501, and an AVIESAN-UNADEV grant to O.M.

[1] G. Buzáki. *Large-scale recording of neuronal ensembles* . Nat Neurosci. , **7(5)**:446–51, 2004.

[2] Adam M Packer, Lloyd E Russell, Henry WP Dalgleish, and Michael Häusser. Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo. *Nature methods*, 12(2):140, 2014.

[3] O. Marre, D. Amodei, N. Deshmukh, K. Sadeghi, F. Soo, T. Holy, and M.J. Berry. *Recording of a large and complete population in the retina.* Journal of Neuroscience , **32(43)**:1485973, 2012.

[4] Diwakar Turaga and Timothy E Holy. Organization of vomeronasal sensory coding revealed by fast volumetric calcium imaging. *Journal of Neuroscience*, 32(5):1612–1621, 2012.

[5] Anna L Vlasits, Thomas Euler, and Katrin Franke. Function first: classifying cell types and circuits of the retina. *Current opinion in neurobiology*, 56:8–15, 2019.

[6] Karl Farrow and Richard H Masland. Physiological clustering of visual channels in the mouse retina. *American Journal of Physiology-Heart and Circulatory Physiology*, 2011.

[7] T. Baden, P. Berens, K. Franke, M.R. Rosn, M. Bethge, and T. Euler. *The functional diversity of retinal ganglion cells in the mouse.* Nature , **529**:345–50, 2016.

[8] G.L.B. Spampinato, E. Ronzitti, V. Zampini, U. Ferrari, F. Trapani, H. Khabou, D. Dalkara, S. Picaud, E. Papagiakoumou, O. Marre, and V. Emiliani. All-optical interrogation of a direction selective retinal circuit by holographic wave front shaping. *bioRxiv*, page 513192, 2019.

[9] Michael N Economo, Sarada Viswanathan, Bosiljka Tasic, Erhan Bas, Johan Winnubst, Vilas Menon, Lucas T Graybuck, Thuc Nghi Nguyen, Kimberly A Smith, Zizhen Yao, et al. Distinct descending motor cortex pathways and their roles in movement. *Nature*, 563(7729):79, 2018.

[10] Mean-Hwan Kim, Petr Znamenskiy, Maria Florencia Iacaruso, and Thomas D Mrsic-Flogel. Segregated subnetworks of intracortical projection neurons in primary visual cortex. *Neuron*, 100(6):1313–1321, 2018.

[11] Rava Azeredo da Silveira and Botond Roska. Cell types, circuits, computation. *Current opinion in neurobiology*, 21(5):664–671, 2011.

[12] Karthik Shekhar, Sylvain W Lapan, Irene E Whitney, Nicholas M Tran, Evan Z Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z Levin, James Nemesh, Melissa Goldman, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5):1308–1323, 2016.

[13] Xiaolong Jiang, Shan Shen, Cathryn R Cadwell, Philipp Berens, Fabian Sinz, Alexander S Ecker, Saumil Patel, and Andreas S Tolias. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science*, 350(6264):aac9462, 2015.

[14] E.J. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12(2):199–213, 2001.

[15] J.W. Pillow, J. Shlens, L. Paninski, A. Sher, A.M. Litke, E. J. Chichilnisky, and E.P. Simoncelli. *Spatio-temporal correlations and visual signalling in a complete neuronal population.* Nature , **454**:995–999, 2008.

[16] J.M. McFarland, Y. Cui, and D.A. Butts. Inferring nonlinear neuronal computation based on physiologically plausible inputs. *PLoS Comput Biol*, 9(7):e1003143, 2013.

[17] L. McIntosh, N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. Baccus. Deep learning models of the retinal response to natural scenes. In *Advances in Neural Information Processing Systems*, pages 1361–1369, 2016.

[18] S. Deny, U. Ferrari, E. Mace, P. Yger, R. Caplette, S. Picaud, G. Tkačik, and O. Marre. Multiplexed computations in retinal ganglion cells of a single type. *Nature communications*, 8(1):1964, 2017.

[19] AMHJ Aertsen, Peter IM Johannesma, and DJ Hermes. Spectro-temporal receptive fields of auditory neurons in the grassfrog. *Biological Cybernetics*, 38(4):235–248, 1980.

[20] Maneesh Sahani and Jennifer F Linden. How linear are auditory cortical responses? In *Advances in neural information processing systems*, pages 125–132, 2003.

[21] Thomas A Münch, Rava Azeredo Da Silveira, Sandra Siegert, Tim James Viney, Gautam B Awatramani, and Botond Roska. Approach sensitivity in the retina processed by a multifunctional neural circuit. *Nature neuroscience*, 12(10):1308, 2009.

[22] E. Schneidman, W. Bialek, and M.J. Berry. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–11553, 2003.

[23] P. K. Trong and F. Rieke. Origin of correlated activity between parasol retinal ganglion cells. *Nature neuroscience*, 11(11):1343–1351, 2008.

[24] J. Shlens, F. Rieke, and E.J. Chichilnisky. Synchronized firing in the retina. *Current opinion in neurobiology*, 18(4):396–402, 2008.

[25] M. Greschner, J. Shlens, C. Bakolitsa, G.D. Field, J. L. Gauthier, L. H Jepson, A. Sher, A. M. Litke, and E.J. Chichilnisky. Correlated firing among major ganglion cell types in primate retina. *The Journal of Physiology*, 589(1):75–86, 2011.

[26] Dmitry R Lyamzin, Samuel J Barnes, Roberta Donato, Jose A Garcia-Lazaro, Tara Keck, and Nicholas A Lesica. Nonlinear transfer of signal and noise correlations in cortical networks. *Journal of Neuroscience*, 35(21):8065–8080, 2015.

[27] X. Pitkow and M. Meister. Decorrelation and efficient coding by retinal ganglion cells. *Nat. Neurosci.*, 15(4):628–635, 2012.

[28] F. Franke, M. Fiscella, M. Sevelev, B. Roska, A. Hierlemann, and R.A. da Silveira. Structures of neural correlation and how they favor coding. *Neuron*, 89(2):409–422, 2016.

[29] E. Schneidman, M. Berry, R. Segev, and W. Bialek. *Weak pairwise correlations imply strongly correlated network states in a population* . Nature , **440**:1007, 2006.

[30] Hideaki Shimazaki, Shun-ichi Amari, Emery N Brown, and Sonja Grün. State-space analysis of time-varying higher-order spike correlation for multiple neural spike train data. *PLoS computational biology*, 8(3):e1002385, 2012.

[31] E. Granot-Atedgi, G. Tkacik, R. Segev, and E. Schneidman. Stimulus-dependent maximum entropy models of neural population codes. *PLOS Computational Biology*, 9(3):1–14, 2013.

[32] U. Ferrari, S. Deny, M. Chalk, G. Tkačik, O. Marre, and T. Mora. Separating intrinsic interactions from extrinsic correlations in a network of sensory neurons. *Physical Review E*,

98(4):042410, 2018.

[33] Pravin K Trivedi, David M Zimmer, et al. Copula modeling: an introduction for practitioners. *Foundations and Trends® in Econometrics*, 1(1):1–111, 2007.

[34] Pietro Berkes, Frank Wood, and Jonathan W Pillow. Characterizing neural dependencies with copula models. In *Advances in neural information processing systems*, pages 129–136, 2009.

[35] Arno Onken, Steffen Grünewälder, Matthias HJ Munk, and Klaus Obermayer. Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flashlight transformation. *PLoS computational biology*, 5(11):e1000577, 2009.

[36] Houman Safaai, Arno Onken, Christopher D Harvey, and Stefano Panzeri. Information estimation using nonparametric copulas. *Physical Review E*, 98(5):053302, 2018.

[37] S. Cocco, S. Leibler, and R. Monasson. *Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods* . Proc. Natl. Acad. Sci. USA , **106**:14058, 2009.

[38] Mihály Bányai, Andreea Lazar, Liane Klein, Johanna Klon-Lipok, Marcell Stippinger, Wolf Singer, and Gergő Orbán. Stimulus complexity shapes response correlations in primary visual cortex. *Proceedings of the National Academy of Sciences*, page 201816766, 2019.

[39] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *BioRxiv*, page 201764, 2018.

[40] D.L.K. Yamins and J.J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[41] Pierre Yger, Giulia LB Spampinato, Elric Esposito, Baptiste Lefebvre, Stéphane Deny, Christophe Gardella, Marcel Stimberg, Florian Jetter, Guenther Zeck, Serge Picaud, et al. A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo. *ELife*, 7:e34518, 2018.

[42] U Ferrari, S. Deny, O Marre, and T Mora. A simple model for low variability in neural spike trains. *Neural Computation*, 30(11), 2018.

[43] U. Ferrari. Learning maximum entropy models from finite-size data sets: A fast data-driven algorithm allows sampling from the posterior distribution. *Phys. Rev. E*, 94:023301, 2016.