Population model learned on different stimulus ensembles predicts network responses in the retina

Ulisse Ferrari,[1,*] Stéphane Deny,[2] Matthew Chalk,[1] Gašper Tkačik,[3] Olivier Marre,[1,4] and Thierry Mora[5,4]

[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]*Institute of Science and Technology, Klosterneuburg, Austria*
[4]*Equal contribution*
[5]*Laboratoire de physique statistique, CNRS, Sorbonne Université, Université Paris-Diderot and École normale supérieure (PSL), 24, rue Lhomond, 75005 Paris, France*

A major challenge in sensory neuroscience is to understand how complex stimuli are collectively encoded by neural circuits. In particular, the role of correlations between output neurons is still unclear. Here we introduce a general strategy to equip an arbitrary model of stimulus encoding by single neurons with a network of couplings between output neurons. We develop a method for inferring both the parameters of the encoding model and the couplings between neurons from simultaneously recorded retinal ganglion cells. The inference method fits the couplings to accurately account for noise correlations, without affecting the performance in predicting the mean response. We demonstrate that the inferred couplings are independent of the stimulus used for learning, and can be used to predict the correlated activity in response to more complex stimuli. The model offers a powerful and precise tool for assessing the impact of noise correlations on the encoding of sensory information.

I. INTRODUCTION

An important goal in sensory neuroscience is to build models that predict how sensory neurons respond to complex stimuli—a necessary step towards understanding the computations carried out by neural circuits. The collective activity of sensory neurons has been modeled using disordered Ising models to account for correlations between cells [1–3]. However this approach cannot predict the activity of the network in response to a particular stimulus, as it ignores stimulus information. In these models, the correlations captured by the Ising couplings arise from shared cell dependencies to the stimulus as well as from shared variability across cells (called noise correlations). Ising models have been generalized to include stimulus information through "fields" that modulate the spiking activity and fluctuate as the stimulus changes over time [4, 5], as well as effective couplings between cells to reflect noise correlations. However, there exists no systematic way to predict how those fields depend on the stimulus.

In the retina, models have been proposed to predict the response of single retinal ganglion cells to a given visual stimulation. A common but limited approach is the the Linear-Non-linear-Poisson (LNP) model [6, 7]. This model is only accurate in restricted cases, e.g. when the stimulus is a full field flicker [8], or has a low spatial resolution, and is not able to predict the responses of ganglion cells to stimuli more complex or with fine spatial structure [9, 10]. Recent studies have used deep convolutional neural networks to predict the responses of ganglion cells to complex visual scenes [11–13]. These models have predicted well the responses of single cells to stimuli with complex spatio-temporal structure [13], and improved significantly the performance in predicting the responses to natural scenes [12]. However, these models ignore noise correlations, and it is unclear if they can predict the joint activity of a population of ganglion cells. Ganglion cells are not deterministic, and show intrinsic variability in the response, even when presented with identical repetitions of the same stimulus. Some of this variability is correlated between different cells, as was shown for neighboring cells of the same type [14–17]. A possible explanation for this correlated noise is that ganglion cells of the same type are connected to each other through gap junctions, at least in the rodent retina [18]. These direct connections between ganglion cells are not taken into account by convolutional models, which treat each cell independently.

Here we build a model that combines a feed-forward multi-layer network to process the stimulus [13], and an interaction network of couplings between output neurons to account for noise correlations. We develop a strategy to learn this model on physiological data and to fit the couplings to reproduce solely noise correlations. We validate the approach by showing that couplings learned on one stimulus ensemble can successfully predict the correlated activity of the population in response to a very different stimulus. Thus, inferred couplings are independent of the stimulus used to learn them, and can be used to predict the correlated activity in response to more complex stimuli. This strategy is not restricted to the specific stimulus-processing model considered here in the context of retinal ganglion cells, and can be applied to generalize any model predicting the mean spiking activity of single cells.

*Correspondence should be sent to `ulisse.ferrari@gmail.com`.

II. NOISE CORRELATIONS IN CELLS OF THE SAME TYPE

We re-analyzed [13] a dataset of *ex-vivo* recording of the activity of retina ganglion cells (RGCs) in Long-Evans adult rats using a multi-electrode array (see Fig. 1A top for an example response). The stimulus was a movie of two parallel bars, whose trajectories followed the statistics of two independent overdamped stochastic oscillators. Additionally a random binary checkerboard and a sequence of specifically devoted stimuli were projected for one hour to allow for receptive field estimation and RGC type identification. Raw voltage traces were stored and spike-sorted off-line through a custom spike sorting algorithm [19].

We applied a standard clustering analysis based on the cell response to various stimuli to isolate a population of OFF ganglion cells of the same type. Their receptive fields tiled the visual field to form a mosaic (Fig. 1B).

The ganglion cell spike times were binned at $\sim 17ms$ (locked to the stimulus frame rate) to estimate the empirical spike counts, $n_i(t) \in [0, 1, 2, \dots]$ for cell i during time-bin t (see Fig. 1A bottom). The stimulus alternates between non-repeated sequences of random bar trajectories, and a repeated sequence of randomly moving bars, displayed 54 times. Non-repeated sequences were divided into training (two thirds of the total) and testing (one third) sets. Repeated sequences were equally divided into training and testing sets by splitting the repeated trajectory in two different halves. We estimate the Peri-Stimulus Time Histogram (PSTH), i.e. average firing rate over time in response to the repeated stimulus sequence, as $\lambda_i(t) \equiv \langle n_i(t) \rangle_{\text{repeat}}$ for cell i, (Fig. 1A bottom) and its average, $\lambda_i \equiv (1/T) \sum_{t=1}^{T} \lambda_i(t)$.

We measured the correlated variability between pairs of cells (example in Fig. 1C) computed from the repeated dataset between two cells. The total pairwise covariance is the sum of stimulus ($c^{(S)}$) and noise ($c^{(N)}$) covariances. The first can be computed as the cross-covariance of the PSTH (black line), which estimates the correlation between cells induced by the stimulus:

$$c_{ij}^{(S)}(t-t') \equiv \frac{1}{T} \sum_{t=1}^{T} (\lambda_i(t) - \lambda_i)(\lambda_j(t') - \lambda_j) \qquad (1)$$

The difference between the total cross-covariance and the stimulus covariance is the noise covariance:

$$c_{ij}^{(N)}(t-t') \equiv \frac{1}{T} \sum_{t=1}^{T} \langle (n_i(t) - \lambda_i(t))(n_j(t') - \lambda_j(t')) \rangle_{\text{repeat}} \qquad (2)$$

Noise covariances are significantly different from zero only at zero lag ($t = t'$), meaning that they happen at a short time scale, which suggests they may be due to gap-junctions [14, 20]. Only cells that were physically close had large noise correlations (Fig. 1D), and their values strongly decreased with distance.

III. INDEPENDENT CELL MODEL

We aim at developing a population model of the RGC response to visual stimuli. First we develop a feed-forward model to predict how each cell responds individually to the stimulus. Second, we will add recurrent interactions to the model to account for noise correlations. We build LN^{2}E , a model composed of a cascade of two layers of Linear-Non-linear units. This model aims at predicting the cell firing rate $\lambda_i(t)$ as a function $\hat{\lambda}_i[S_t]$ of the stimulus S. Spikes are then generated as draws from the Effective (E) model described in [21]:

$$P_{\text{LN}^2\text{E}}\big(n_i(t) \,\big|\, S_t \big) = \exp\Big\{ \hat{h}_i[S_t] n_i(t) - \gamma\, n_i(t)^2 - \delta\, n_i(t)^3 - \log n_i(t)! - \log Z_t \Big\} \qquad (3)$$

$$\hat{h}_i[S_t] \qquad \text{such that} \quad \langle n_i(t) \rangle_{P_{\text{LN}^2\text{E}}} = \hat{\lambda}_i[S_t] \ . \qquad (4)$$

where $Z_t = Z[S_t, \gamma, \delta]$ is the partition function. γ and δ are parameters of a simple model that allows efficient modeling of the sub-Poisson variability, and accurately reproduce the spike count variance [21]. $\hat{h}_i[S_t]$ is a uniquely defined function of γ and δ that constraints the average of $n(t)$ to equal $\hat{\lambda}_i[S_t]$, much like in classical physics a chemical potential can be tuned to impose the mean number of particle in the system.

A. LN^{2}E model

Within the LN^{2}E model [13, 21], the stimulus $S(x,t)$, representing the time behavior of each pixel, is first convolved with a Gaussian and biphasic factorized kernel $K^{\text{BP}}(x,t)$ and then passed through Rectified Quadratic Units with the two possible polarities:

$$\Theta[S_t] = \Theta(x,t) \equiv \Big[\ \hat{\Theta}(x,t)^{\text{ON}} \ , \ \hat{\Theta}(x,t)^{\text{OFF}} \ \Big] \qquad (5)$$

$$= \Big[\text{NL}^{\text{ON}}\big(K^{\text{BP}} \star S \big) \, , \, \text{NL}^{\text{OFF}}\big(K^{\text{BP}} \star S \big) \Big] \qquad (6)$$

where $\star$ means spatial and temporal convolution and NL^{ON} and NL^{OFF} correspond to rectified square

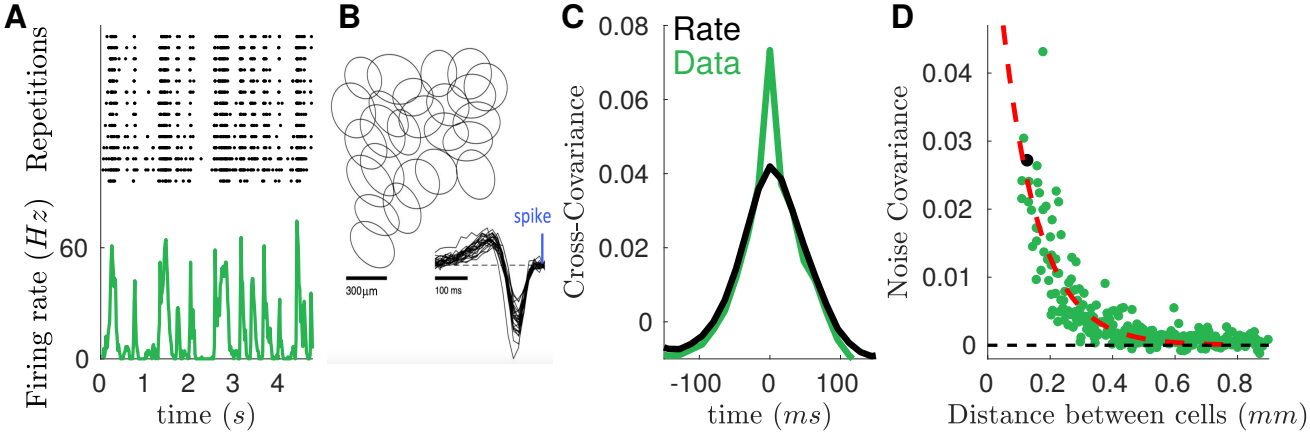

FIG. 1: ***Ex-vivo* recording of retinal ganglion cell shows short time-scale noise correlations.** A) Top: activity of one ganglion cell over repeated presentations of the same stimulus. Bottom: firing rate of the cell across time. B) Receptive field mosaic of the isolated OFF ganglion cells C) Empirical cross-covariance between two example cells (green trace) superimposed on the covariance between their average firing rate (black trace). The two differ significantly at short time-scales. D) Zero-lag noise correlation as a function of cell pair distance. Black point refers to the pair of neurons plotted in panel C. Red dashed line shows an exponential fit with spatial scale of $0.11mm$.

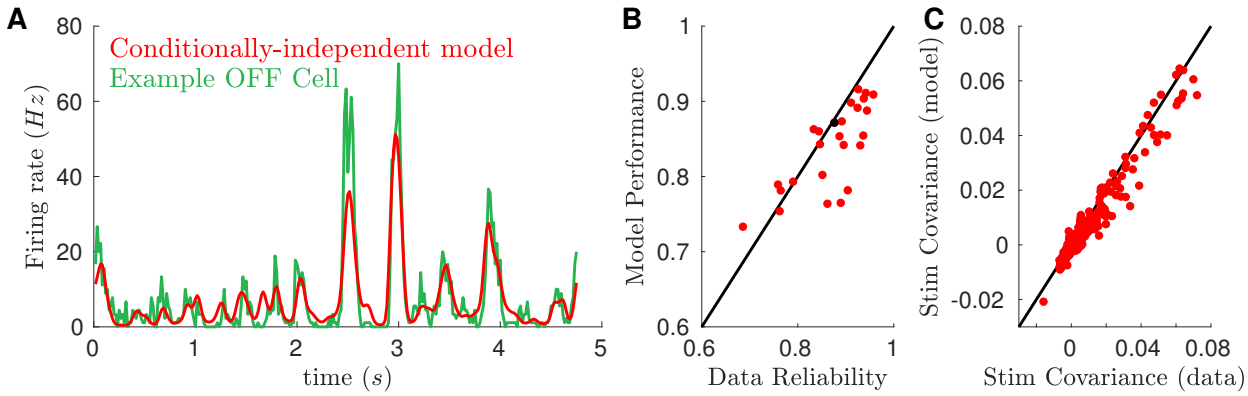

FIG. 2: **Prediction for the LN^{2}E model.** A) For an example cell, the model-predicted firing rate (red trace) is super-imposed to the empirical one (green trace). By computing the correlation between the two traces ($\rho = 0.87$) we can estimate the model performance. B) Scatter-plot of the LN^{2}E performance against the cell reliability, computed as the correlation between two halves of the repeated dataset. Black point refers to the cell of panel A. C) Model-predicted stimulus covariances account for more than 84% of stimulus covariance (slope of a linear fit), but systematically under-estimate their empirical value. Black line is the identity.

$(\max(0, \cdot^2))$ non linearities [13]. $\Theta(x,t)$ is then fed into a second non-linear stage: for each cell i, it is first convolved with a receptive field $K_i^{\mathrm{RGC}}(x,t)$ and then passed through a non-linear function $\mathrm{NL}_i^{\mathrm{RGC}}$:

$$\hat{\lambda}_i[S_t] = \mathrm{NL}_i^{\mathrm{RGC}}\Big[\, K_i^{\mathrm{RGC}} \star \Theta[S_t] \,\Big] \quad (7)$$

$$\mathrm{NL}_i^{\mathrm{RGC}}\Big[\, x \,\Big] = a_i \log\Big(1 + \exp\big\{b_i(x + c_i)\big\}\Big) \,. \quad (8)$$

B. Inference and results

To infer the parameters of the model we follow a simplified version of [22], where only the second non-linear stage is learned. We keep fixed the first stage on a setting that has been shown to work efficiently [13]. For the second stage we apply an iterative procedure where we maximize the log-likelihood of the data under the model given by Eq. (3), penalized by a weighted sum of the L1 and L2 norms of the parameter vector, with respect to all parameters. We used the non-repeated training set to compute the log-likelihood gradient. To avoid overfitting we early-stopped the iterative procedure when the

log-likelihood computed on the non-repeated testing set stopped to increase. L1 and L2 penalties were optimized by maximizing the performance on the repeated dataset that later we will use for training the population model.

In Fig. 2A we compare the time course of the empirical firing rate, $\lambda_i(t)$, with the prediction of the inferred LN^{2}E model for an example cell. By computing the Pearson correlation among these two temporal traces ($\rho = 0.87$), we can estimate the performance of the model for each cell. In Fig. 2B we compare this performance with the reliability of the retinal spike activity, estimated as the correlation between two disjoint subsets of responses to the repeated stimulus and found that they were comparable. In Fig. 2C we show how the model predicts the empirical stimulus covariances. Even if a small under-estimation is present, the model accounts for more than the 84% of the empirical value (slope of a linear fit).

IV. POPULATION MODEL

We then build a population model that accounts for noise correlation, starting from the LN^{2}E . From the model in the form Eq. (3) we build a two layers Linear-Non-linear Population Effective model, LN^{2}PE , by adding a pairwise interaction term:

$$P_{\mathrm{LN^2PE}}\left(\mathbf{n}(t)\,\middle|\,S,\mathbf{J}\right) = \exp\left\{\sum_i \left(\hat{h}_i[S_t]n_i(t) - \gamma\, n_i(t)^2 - \delta\, n_i(t)^3 - \log n_i(t)!\right) + \sum_{i\leq j} J_{ij} n_i(t) n_j(t) - \log Z_t\right\} \tag{9}$$

$$\hat{h}_i[\,S_t] \qquad \text{such that} \quad \langle n_i(t)\rangle_{P_{\mathrm{LN^2PE}}} = \hat{\lambda}_i[S_t]\ . \tag{10}$$

where $Z_t \equiv Z\big[\boldsymbol{S}_t, \gamma, \delta, \mathbf{J}\big]$ is the normalization constant. The terms J_{ii} are assumed to be independent of i, and are a correction to γ due to the impact of the population. The model is thus designed such that the conditional probability distribution of any spike count $n_i(t)$ reduces to an Effective model (Eq. (3)), when the activity of all the others neurons is fixed.

A. Inference of the interaction network

One possibility to infer the LN^{2}PE model could have been to estimate the couplings by maximizing likelihood on the non-repeated dataset, similar to [4]. However, the issue with this approach is that it does not distinguish between signal and noise correlation. As a result, if the feed-forward model does not reproduce perfectly the signal correlation, the interaction parameters will try to compensate this error. We observed this effect by inferring a Generalized Linear Model (GLM) [7], see App. A. Straightforward likelihood maximization cannot guarantee that coupling terms only reflect noise correlations; rather, they reflect an uncontrolled combination or mixture of signal and noise correlations, which have very different biological interpretations. An extreme version of this phenomenon can be found in [4], Fig. 15, where the response of one neuron to a natural scene is predicted from the responses of other neurons, without any use of the stimulus. In our case we want to have couplings reflecting solely noise correlations. We therefore developed an alternative inference method.

In order to obtain the parameters values J that account solely for noise correlations, we need a model that reproduces perfectly the time course of the firing rate. To construct such a model we use repetitions of the same stimulation, to estimate the empirical rate and impose it with Lagrange multipliers. To this aim we recast the LN^{2}PE model in a Time Dependent Population Effective (TDPE) model [4]:

$$P_{\mathrm{TDPE}}\big(\mathbf{n}(t)\,\big|\,\mathbf{J},\boldsymbol{h}\big) \sim \exp\left\{\sum_i \left(h_i(t)n_i(t) - \gamma\, n_i(t)^2 - \delta\, n_i(t)^3 - \log n_i(t)!\right) + \sum_{i\leq j} J_{ij} n_i(t) n_j(t)\right\} \tag{11}$$

where we have substituted the stimulus dependent input-rate $\hat{h}_i[S_t]$ with Lagrange multipliers $h_i(t)$ enforcing the empirical mean spike count, $\langle n_i(t)\rangle = \lambda_i(t)$, for each cell in each time-bin of the repeated stimulation. TDPE is not a real stimulus model, as it can not predict the spiking activity in response to stimuli different from the one

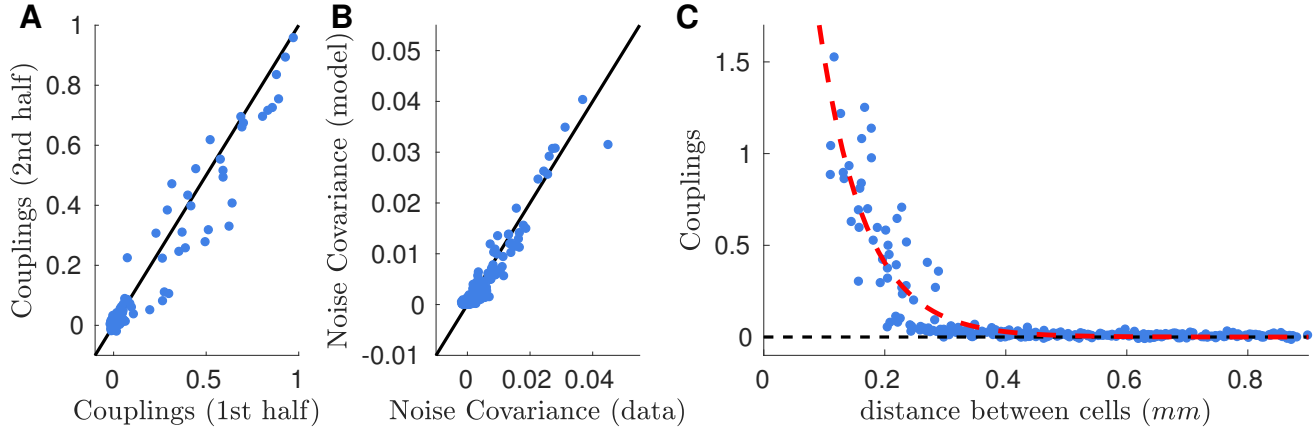

FIG. 3: **Time-dependent inference provides good estimates of the interaction network.** A) Comparison between the inferred interaction from first (later used as training) and second (later used as testing) halves of the repeated dataset. Black line is the identity. B) Comparison of the predicted noise correlations when the interaction matrix $\boldsymbol{J}$ learned using the training set is applied to the testing set. Black line is the identity. C) The behavior with distance of the inferred interactions scales similarly to that of noise correlations (see Fig. 1D), although it goes to zero at shorter distances. Red dashed line is an exponential fit with spatial constant $0.08mm$.

used for training. However the learned J will account solely for noise correlations as stimulus correlations are reproduced by construction.

We will proceed as follows: At first we will learn J and $\boldsymbol{h}$ of the model (11) on repeated sequences of the stimulus. As the model (11) belongs to the exponential family, the inference can be easily performed by maximizing the log-likelihood (see for example [23]). We add a small L2 penalty (with coefficient $\eta_{\text{L2}} \sim 2 \cdot 10^{-6}$) over the biases $\mathbf{h}$ simply to avoid divergences to $-\infty$ when $\langle n_i(t)\rangle_{\text{repeat}} = 0$. In order to avoid spurious non-zero values, we also add a L1 penalty (with coefficent $\eta_{\text{L1}} = 0.04$). Fig. 3 shows the results of the TDPE inference. In order to evaluate the robustness of the inference with respect to a change of dataset, in Fig. 3A we plot the interactions inferred from the training set against those inferred from another training set of the same size. The comparison shows that inferred networks are robust against a change of stimulus.

We will use the values found for the parameters J_{ij} when inferring the model (11) in the first step, and keep these parameters fixed during the second step of the inference, where we will learn the rest of the parameters of the population model (Eq. (9)).

To check the validity of this approach, in Fig. 3B we first compare empirical noise correlations obtained with the test dataset with those predicted by the model. To obtain this prediction, we freeze the $\mathbf{J}$ value obtained from the inference on the training set and we re-infer the $h_i(t)$ to match the firing rates of the testing set. The inferred parameters J were able to predict well the noise correlation on a part of the recording that had not been used to learn them. In panel C we show the behavior of the interaction parameters as a function of the distance between the corresponding cell couple. J decreases with distance slightly faster than noise correlation, see Fig. 1.

The good quality of the inference allows us now to apply the inferred interaction network on top of the single cell model.

B. Complete model inference

Now that we have inferred the network $\mathbf{J}$ we can build the LN^{2}PE model, presented in Eq. (9). The last step to perform is to adjust the parameters that specify the fields as a function the stimulus $\hat{\mathbf{h}}[\mathbf{S_t}]$

Because of the interaction term, each cell receives a random input $\sum_j J_{ij} n_j(t)$ from the rest of the network. This random contribution is responsible for the noise correlation. However it generates also a mean contribution $\Delta\hat{h}_i[S_t]$ that may affect the cell firing rate and if not removed, will impact the firing rate prediction. Because of network effects, this mean contribution differs from its mean value: $\Delta\hat{h}_i[S_t] \neq \sum_{i\neq j} J_{ij}\hat{\lambda}_j(t)$. In order to estimate it, we compute the Thouless, Anderson and Palmer (TAP) free energy formalism of the model (9) [24] and use it to derive an approximation for $\Delta\hat{h}_i[S_t]$. We followed strictly [25] and [26] and apply a second order Plefka expansion [27] (see App.B for more details). The result is:

$$\begin{aligned}\Delta\hat{h}_i[S_t] \approx & \sum_{j\neq i} J_{ij}\hat{\lambda}_j[S_t] + J_{ii}\Big(V'\big(\hat{\lambda}_i[S_t]\big) + 2\hat{\lambda}_i[S_t]\Big) \\ & + \frac{1}{2}V'\big(\hat{\lambda}_i[S_t]\big)\sum_{j\neq i} J_{ij}^2 V\big(\hat{\lambda}_j[S_t]\big) \\ & + \frac{1}{2}J_{ii}^2 W'\big(\hat{\lambda}_i[S_t]\big) \qquad (12)\end{aligned}$$

where the first two terms are the mean-field contributions

of network and self couplings, whereas the lasts two are their Onsager [28] reaction terms. Together with their derivatives, in (12) we have introduced

$$V(\lambda) \equiv P_\lambda^{(2)} - P_\lambda^{(1)} P_\lambda^{(1)} \quad (13)$$

$$W(\lambda) \equiv P_\lambda^{(4)} - P_\lambda^{(2)} P_\lambda^{(2)} - \frac{\left(P_\lambda^{(3)} - P_\lambda^{(2)} P_\lambda^{(1)}\right)^2}{V(\lambda)} \quad (14)$$

where $P_\lambda^{(s)} \equiv \langle n^s \rangle_E$ are the moments of n computed within an Effective model with the same structure as Eq. (3), but with parameter $\hat{h}_i$ tuned to enforce the mean spike count, $\langle n \rangle_E = \lambda$. Note that the correction (12) is a stimulus dependent quantity because it depends only on inputs $\hat{\lambda}[S_t]$ and not on the spike counts $\mathbf{n}(t)$. Consequently it can be reabsorbed within the form (9) by substituting:

$$\hat{h}_i^{\mathrm{PE}}[S_t] = \hat{h}_i^{\mathrm{E}}[S_t] - \Delta \hat{h}_i[S_t] \quad (15)$$

$$\hat{h}_i^{\mathrm{E}}[S_t] \text{ such that } \langle n_i(t) \rangle_{P_{\mathrm{LN^2E}}} = \hat{\lambda}_i[S_t] \ , \quad (16)$$

where $\langle n_i(t) \rangle_{P_{\mathrm{LN^2E}}}$ is computed with the conditionally independent model LN^{2}E , not the population model as in (9). Computing this correction requires only some matrix multiplications and can thus be evaluated easily for a new stimulus.

In Fig. 4A and B, we check that the inferring the population model still preserves the quality of the prediction of single cell activity obtained with the independent model. We compare the performance of the two models in reproducing the firing rate of the recorded cells (same quantity as Fig. 2B). Firing rate is a stimulus dependent quantity and accordingly the two models have similar performance. The fact that LN^{2}PE has not smaller performance than LN^{2}E ensures the quality of the estimation of the corrections $\mathbf{\Delta h}[S_t]$ with the TAP approach. Whereas the fact that it is neither larger ensures that the couplings do not help the model in predicting stimulus-dependent, average quantities but can account only for noise fluctuations. Panel B shows that thanks to the self interaction term, LN^{2}E and LN^{2}PE have similar performance also for spike count variance.

Fig. 4C shows how the LN^{2}PE model outperforms the LN^{2}E in predicting the population activity. The coupled LN^{2}PE model accounts well for noise covariances on a testing set (blue points). Because it is an independent cell model, the LN^{2}E predicts vanishing noise covariance (red points).

C. Generalizing to other types of stimuli

A major challenge with fitting complex encoding models to neural responses is that they rarely generalize well. Here we ask if the interaction network can be inferred from a qualitatively different stimulus and then applied to our two-bars stimulation. To test for this, we infer the couplings as in the section IV A, but on the response to repeated checkerboard stimuli. We then use this new set of couplings to equip the conditionally-independent LN^{2}E we have previously inferred on the moving bar stimulus and obtain a new LN^{2}PE model. In panel D we show that this LN^{2}PE model predicts noise covariances when applied to the two-bars testing set. This shows that our inference method allows to generalize from one stimulus type to the other, because the couplings between cells seem invariant to the stimulus here.

V. DISCUSSION

We have shown a method to infer a model from population recordings, which takes into account both the impact of the stimulus, through a feed-forward network, and the correlated noise between cells, thanks to a coupling network. Note that the interactions between cells are here described as "instantaneous". The activity of one cell at time t will affect other cells at time bin t, not $t+1$. This is due to the fact that the noise correlations between cells are very fast. We chose a time bin of the magnitude of the time constant of these correlations.

Previous approaches [4, 7] have aimed at reproducing noise correlations, but the stimulus could only influence the response through a linear filter. When reduced to a single cell, these models were similar to the classical LN model. Conversely, non-linear models derived from convolutional networks [12, 13] did not try to model noise correlations. Here we take the best of both approaches, and developed a method to infer both the parameters describing the influence of the stimulus, and the coupling values. Our method is general and could be applied to other structures than the retina. Furthermore, it can be straightforwardly extended to models where the influence of the stimulus on single cell activity is described by different non-linear models, e.g. a deep network with more layers [12].

We have developed an inference strategy that allows the coupling term to reflect solely shared noise across neurons, without the interference of stimulus effects due to artifacts of the learning procedure. Such effects can arise when the inference procedure tries to use the activity of other cells as a surrogate of the stimulus itself to improve the predictability of a given cell, compensating for non-linearities in the stimulus dependence that are unaccounted by the model. Consequently, couplings show a weak dependence on the actual stimulus ensemble used to learn them. This allowed us to to show that their value can be inferred even from a qualitatively different stimulus. Note that checkerboard and two-bars stimulations drive very different responses of the retina. It is thus a remarkable result that noise correlations in the response to a complex video of moving objects can be predicted by couplings inferred from responses to stimuli. This result can thus be seen as a first step toward the construction of a full model that accounts for large and heterogeneous stimulus ensembles.

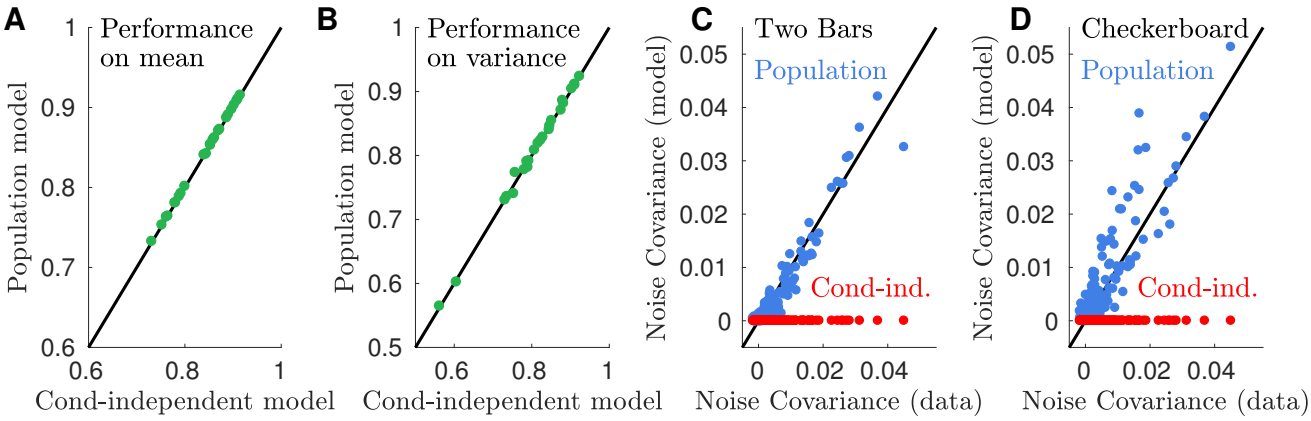

FIG. 4: **Population model (LN^{2}PE) predicts empirical noise covariances and performs as conditional-independent model (LN^{2}E) for stimulus dependent quantities.** A) Performance in estimating the cell firing rate (same as Fig. 2B) of the LN^{2}PE equals that of the LN^{2}E . B) LN^{2}PE and LN^{2}E have the same performance also in predicting the spike count variance. C) The LN^{2}PE predicts empirical noise covariances when applied on a testing set (blue points). Conditionally-independent LN^{2}E predicts zero noise covariances (red). D) same as C, but when the couplings $\boldsymbol{J}$ are inferred from response to checkerboard stimulation.

Having models that account for noise correlations is a crucial first step also to study the impact of noise correlations on coding complex stimuli. Because we develop our inference strategy such that the coupling terms only reflect the noise correlation and do not impact the response prediction of single cells, we can easily compare the population model with the case where we simply set the coupling values to 0. By comparing the computations performed by these two models, future works will be able to understand how these fast noise correlations affect coding in the retina. The same strategy can be used in other sensory structures or any other noisy input/output system.

VI. ACKOWNLEDGEMENTS

We like to thank C. Gardella for useful discussion. This work was supported by ANR TRAJECTORY, the French State program Investissements d'Avenir managed by the Agence Nationale de la Recherche [LIFESENSES: ANR-10-LABX-65], a EC grant from the Human Brain Project (FP7-604102)), NIH grant U01NS090501 and AVIESAN-UNADEV grant to OM.

Appendix A: Generalized linear model analysis

We fitted a generalized linear model (GLM) to the responses of $N_r = 25$ off-cells in response to the moving bar stimuli. We discretized the response and stimulus using temporal bins of length $1.667ms$ (i.e. 600Hz; 10 times smaller than the temporal bins used for the model described in the main paper). The response of the ith neuron at time t is described by an integer $n_i(t)$ denoting the number of spikes fired by the ith neuron in the tth time bin. The spatial location of the presented bar stimuli were discretized into $N_x = 100$ equi-sized bins. The stimulus at time t was denoted by a binary vector $\boldsymbol{x}_t$, of which the ith component $(x_i(t))$ was set to 1 if one of the bars was centred on the corresponding spatial location at time t, and 0 otherwise.

In a GLM, the spiking response of each neuron is assumed to be Poisson distributed, with the mean number of spikes of the ith neuron at time t given by $\exp(r_i(t))$, where:

$$r_i(t) = \sum_{j=1}^{N_x} \sum_{k=0}^{N_w-1} w_{ijk} x_j(t-k) + \sum_{j=1}^{N_r} \sum_{k=1}^{N_v} v_{ijk} n_j(t-k) + b_i. \tag{A1}$$

In this equation, b_i denotes a constant bias term, w_{ijk} is an element of the temporal stimulus filter (of size $[N_r, N_x, N_w]$)), and v_{ijk} is an element of the recurrent filter (of size $[N_r, N_r, N_v]$)). We used a stimulus filter of length $N_w = 200$ (i.e. ∼300ms), and recurrent filters of length $N_v = 15$ (i.e. ∼25ms).

Model parameters were fitted by maximizing an objective function consisting of the log-likelihood (which, for a Poisson distribution, is given up to constant by $L = \sum_{i,t} n_i(t) r_i(t) - e^{r_i(t)}$), minus an L2 norm regularization term that promoted smooth filters and thus prevented overfitting. The regularization parameters were chosen to maximize the log-likelihood on test data, held out during model fitting. Further, to reduce the number of parameters, and thus further reduce over–fitting we assumed that the stimulus filter could be described as the sum of $N_{rank} = 3$ spatio-temporally separable filters, each given by a spatial filter multiplied by a temporal filter (i.e. $w_{ijk} = \sum_{l=1}^{N_{rank}} u_{ij}^l a_{ik}^l$, where u and a here denote the spatial and temporal filters, respectively). Relaxing

this assumption (by increasing N_{rank}) did not improve the quality of the model fit.

We fitted two versions of the GLM model to the recorded neural responses: first, a 'coupled model' with firing rate described by equation A1; second, an uncoupled model, in which the recurrent filters, v, were set to zero. Figure 5 plots the firing rate of a single recorded OFF-cell (also plotted in Fig. 2A of the main text), alongside the predictions by the coupled and uncoupled GLM models. The coupled model gave a different prediction for the cell's firing rate, compared to the uncoupled model. Further, plotting the Pearson correlation between empirical firing rate and the predictions by both model variants for each cell (Figure 5b), shows that the coupled model gives an improved prediction of the recorded PSTH for nearly all recorded OFF-cells, compared to the uncoupled model. Thus, in addition to fitting the interactions between different neurons (i.e. the noise correlations), the coupled GLM model used the recurrent filters, v, to improve the prediction of how each neuron responded to the stimulus (i.e. the stimulus correlation).

Appendix B: Construction of the mean-field theory and Thouless-Anderson-Palmer correction

We are interested in computing the TAP correction to the fields $h_i(t)$ due to the addition of the coupling term $\mathbf{J}$ to the conditionally-independent LN^{2}E model (3). Because we are not interested in the TAP expression for couplings nor in that for covariances, we can construct the mean-field theory for a single time-bin. Otherwise, because the couplings are constant in time we should had considered the whole model. To apply the Plefka expansion we introduce:

$$F[\alpha, \boldsymbol{h}, J] \equiv -\log \sum_{\mathbf{n}} \exp\left\{ \sum_{i=1}^{N} \Big(h_i n_i - \gamma\, n^2 - \delta\, n^3 - \log n! \Big) + \alpha \sum_{i<j} J_{ij} n_i n_j \right\} \tag{B1}$$

and its Legendre transform:

$$G[\alpha, \boldsymbol{\lambda}, J] = \left(\sum_i h_i \lambda_i + F[\alpha, \boldsymbol{h}, J] \right\Big|_{\boldsymbol{h}=\tilde{\boldsymbol{h}}} \tag{B2}$$

where $\tilde{\boldsymbol{h}} = \tilde{\boldsymbol{h}}[\alpha, \boldsymbol{\lambda}]$ is defined implicitly from:

$$\left. \frac{\partial \left(\sum_i h_i \lambda_i + F[\alpha, \boldsymbol{h}, J] \right)}{\partial h_i} \right|_{\boldsymbol{h}=\tilde{\boldsymbol{h}}} = \lambda_i - \langle n_i \rangle^{(\alpha)} \Big|_{\boldsymbol{h}=\tilde{\boldsymbol{h}}} = 0 \tag{B3}$$

where $\langle \dots \rangle^{(\alpha)}$ is the average with respect to the distribution related to the free energy (B1). Our goal is to expand $G(\alpha, \boldsymbol{\lambda}, j)$ around $\alpha = 0$ up to the second order. At first we evaluate the derivatives:

$$G'[\alpha, \boldsymbol{\lambda}, J] = -\sum_{i<j} J_{ij} \langle n_i n_j \rangle^{(\alpha)} \tag{B4}$$

$$G''[\alpha, \boldsymbol{\lambda}, J] = -\sum_{i<j} \sum_{k \le l} J_{ij} J_{kl} \langle n_i n_j n_k n_l \rangle_c^{(\alpha)} + \sum_{i<j} J_{ij} \sum_g \frac{\partial \tilde{h}_g}{\partial \alpha} \langle n_i n_j (n_g - \lambda_g) \rangle^{(\alpha)} \tag{B5}$$

$$\frac{\partial \tilde{h}_g}{\partial \alpha} = \frac{\left\langle \sum_{k \le l} J_{kl} n_k n_l (n_g - \lambda_g) \right\rangle^{(\alpha)}}{\langle n_g^2 \rangle_c^{(\alpha)}} \tag{B6}$$

where to obtain the last equality we applied the implicit function theorem to Eq. (B3). To compute the term of the Plefka expansion we need to evaluate $G[\alpha, \boldsymbol{\lambda}, J]$ and its derivatives at $\alpha = 0$. To this aim we note that $\tilde{h}_i[\alpha = 0, \boldsymbol{\lambda}] = \tilde{h}_i[\lambda_i]$, as for $\alpha = 0$ the system units become independent and consequently $\tilde{h}_i$ depends only on λ_i. For $\alpha = 0$, in fact, the distribution over $\{n_i\}_{i=1}^N$ factorizes over a set of single variable distributions. this allows us to compute model expectations at $\alpha = 0$. For future convenience, we define the moments of such distributions:

$$P^{(s)} \equiv \langle n^s \rangle^{(\alpha=0)} \tag{B7}$$

so that the terms in the expansion become:

$$G[0, \boldsymbol{\lambda}, J] = \sum_i \tilde{h}_i \lambda_i + F[\alpha = 0, \tilde{h}] \tag{B8}$$

$$G'[0, \boldsymbol{\lambda}, J] = -\sum_{i<j} J_{ij} \lambda_i \lambda_j - \sum_i J_{ii} P_i^{(2)} \tag{B9}$$

$$G''[0, \boldsymbol{\lambda}, J] = -\sum_{i<j} J_{ij}^2 \big(P_i^{(2)} - \lambda_i^2\big)\big(P_j^{(2)} - \lambda_j^2\big) - \sum_i J_{ii}^2 W_i \tag{B10}$$

where W has been defined in Eq. (14). We can now obtain an approximation for $G[\boldsymbol{\lambda}, J] = G[\alpha = 1, \boldsymbol{\lambda}, J$ defined in Eq. (B2) as:

$$G[\boldsymbol{\lambda}, J] \approx G[0, \boldsymbol{\lambda}, J] + \alpha\, G'[0, \boldsymbol{\lambda}, J] \tag{B11}$$

$$+ \frac{1}{2} \alpha^2\, G''[0, \boldsymbol{\lambda}, J] + O(\alpha^3) \tag{B12}$$

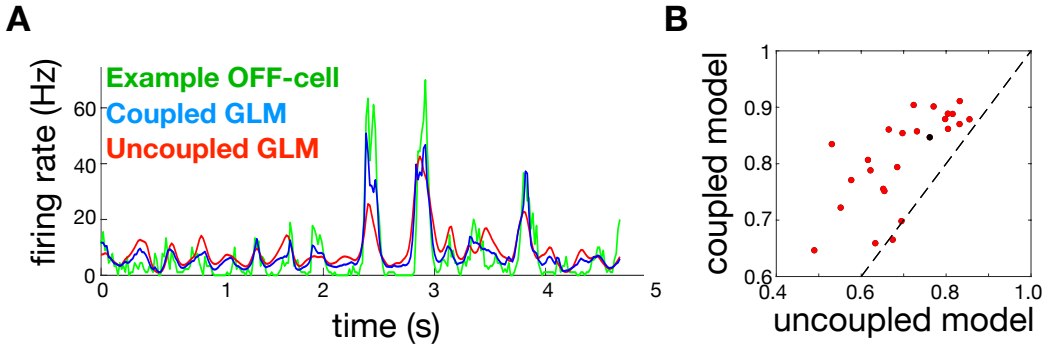

FIG. 5: Effect of adding recurrent filters in a generalised-linear-model (GLM) fit to the recorded neural responses. (A) Example firing rate of a single cell (green), alongside the predictions of the coupled (blue) and uncoupled (red) GLM fits. (B) Correlation coefficient between the firing rate predicted by the coupled (y-axis) and uncoupled (x-axis) GLM models and the data, for each cell. The cell shown in panel A is shown in black.

The mean-field equation for the fields $\boldsymbol{h}$ can be easily obtained by a reverse Legendre transform:

$$h_i = \frac{\partial G[\boldsymbol{\lambda}, J]}{\partial \lambda_i} \tag{B13}$$

$$= \tilde{h}_i[\lambda_i] + \frac{\partial G'[0, \boldsymbol{\lambda}, J]}{\partial \lambda_i} + \frac{1}{2}\frac{\partial G''[0, \boldsymbol{\lambda}, J]}{\partial \lambda_i} \tag{B14}$$

which provides the expression (12) for the TAP correction.

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

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

[3] Tkacik G, et al. (2014) *Searching for collective behaviour in a network of real neurons* . PloS Comput. Biol. **10(1)**:e1003408.

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

[5] Tkacik G, Granot-Atedgi E, Segev R, Schneidman E (2013) Retinal metric: A stimulus distance measure derived from population neural responses. *Phys. Rev. Lett.* 110:058104.

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

[7] Pillow J, et al. (2008) *Spatio-temporal correlations and visual signalling in a complete neuronal population.* Nature **454**:995–999.

[8] Berry II MJ, Meister M (1998) *Refractoriness and neural precision* pp 110–116.

[9] Gollisch T, Meister M (2010) Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron* 65:150–64.

[10] Heitman A, et al. (2016) Testing pseudo-linear models of responses to natural scenes in primate retina. *bioRxiv.*

[11] Vintch B, Movshon JA, Simoncelli EP (2015) A convolutional subunit model for neuronal responses in macaque v1. *Journal of Neuroscience* 35:14829–14841.

[12] McIntosh L, Maheswaranathan N, Nayebi A, Ganguli S, Baccus S (2016) *Deep learning models of the retinal response to natural scenes* pp 1361–1369.

[13] Deny S, et al. (2017) Multiplexed computations in retinal ganglion cells of a single type. *Nature communications* 8:1964.

[14] Brivanlou I, Warland D, Meister M (1998) *Mechanisms of Concerted Firing among Retinal Ganglion Cells.* Neuron **20**:527–539.

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

[16] Greschner M, et al. (2011) Correlated firing among major ganglion cell types in primate retina. *The Journal of Physiology* 589:75–86.

[17] Ala-Laurila P, Greschner M, Chichilnisky E, Rieke F (2011) Cone photoreceptor contributions to noise and correlations in the retinal output. *Nature neuroscience* 14:1309–1316.

[18] Völgyi B, Chheda S, Bloomfield S (2009) Tracer coupling patterns of the ganglion cell subtypes in the mouse retina. *J. Comp. Neurol.* 512:664–687.

[19] Yger P, et al. (2016) Fast and accurate spike sorting in vitro and in vivo for up to thousands of electrodes. *bioRxiv.*

[20] Trenholm S, et al. (2014) Nonlinear dendritic integration of electrical and chemical synaptic inputs drives fine-scale correlations. *Nature neuroscience* 17:1759–1766.

[21] Ferrari U, Dény S, Marre O, Mora T (2018) A sim-

ple model for low variability in neural spike trains. *bioRxiv/2018/243543*.

[22] McFarland JM, Cui Y, Butts DA (2013) Inferring nonlinear neuronal computation based on physiologically plausible inputs. *PLoS Comput Biol* 9:e1003143.

[23] Ferrari U (2016) 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.

[24] Thouless D, Anderson P, Palmer R (1977) *Solvable model of a spin glass*. Phil. Mag. **35**:593.

[25] Kappen H, Rodriguez F (1997) *Efficient learning in boltzmann machines using linear response theory.*. Neural Comput. **10**:1137–1156.

[26] Tanaka T (1998) *Mean-field theory of Boltzmann machine learning*. Phys. Rev. E **58**:2302.

[27] Plefka T (1982) Convergence condition of the tap equation for the infinite-ranged ising spin glass model. *Journal of Physics A: Mathematical and general* 15:1971.

[28] Onsager L (1936) Electric moments of molecules in liquids. *Journal of the American Chemical Society* 58:1486–1493.