Dimensionality Reduction on Maximum Entropy Models on Spiking Networks

Rubén Herzog, Maria-Jose Escobar, Adrian Palacios, Bruno Cessac

Dimensionality Reduction on Maximum Entropy Models on Spiking Networks

Rubén Herzog[1,4]*, María-José Escobar[2], Adrián G. Palacios[1,4], Bruno Cessac[3],

1 Centro Interdisciplinario de Neurociencia de Valparaíso, Universidad de Valparaíso, Valparaíso, Chile.
2 Departamento de Electronica, Universidad Tecnica Federico Santa Maria, Valparaíso, Chile.
3 INRIA Biovision team Sophia Antipolis and Universite Cote d'Azur, Sophia-Antipolis, France.
4 Instituto de Sistemas Complejos de Valparaiso, Chile.

¤Current Address: Centro Interdisciplinario de Neurociencia, Universidad de Valparaiso, Pasaje Harrington 287, Playa Ancha, ZIP 2360102, Valparaiso, Chile.
* rubenherzog@ug.uchile.cl

Abstract

Abstract. Will write it after writing the rest of the paper. 300 words

Author Summary

Same than above. 150-200 words

Introduction

The nature and importance of correlations for the neural coding, has been debated since long time, including the role of weak pairwise correlations [14, 15]; for error supression, enhancing channel capacity [**?**]; or to capture the statistical properties of the network activity [1, 2].

The introduction of maximum entropy models (MEM) to the analysis of neural networks started to gain popularity recently. MEM have been proposed to capture and predict the collective spatio-temporal pattern activity from a vertebrate retina network [4–6, 14, 15] or from a cortex cultures cells [7]. However, most of the studies has been limited to capture *spatial* correlations by using an Ising model [4–6], where events occurring at different times are independent. Thus, these approaches have neglected spatio-temporal correlations, memory and causality, which, presumably, play a central role on the nervous system. More recently, an extension of MEM that includes spatio-temporal correlations has been proposed at the theoretical [8, 9]; numerical [10, 11]; and experimental level [12].

In all cases, a MEM is characterized by a function ("energy" or "Hamiltonian") containing functional and/or effective interactions between spiking neurons. In its simplest form this energy contains only self-interactions and the corresponding MEM is similar to a Bernoulli model, where spikes are independent, that well fits the dominant terms in the statistics. However adding interactions, e.g. pairwise or higher, allows in principle to handle higher and higher order statistics. In its simplest form we can

assume that a neural population is coding solely by means of firing rates, while including pairwise spatio-temporal interactions we assumes the neural coding include also the temporal co-activation of the neurons [**?**, **?**]. However, in terms of a computational model, the overwhelming explosion of parameters has corrolary effects: over-fitting and bad reliability in parameters estimations.

This proliferation of parameters (i.e. increase of model dimensionality), especially whith a growing number of neurons, is a major criticism of MEM. Would be it possible to lower the dimensionality of the space where a MEM can be mapped, given the redundancies of the code? [**?**, **?**, **?**, 3] If so, we are facing a compressible MEM that reflects the compressibility of the neural code. So, in analogy with signal compression, where the presence of statistical redundancy can be exploited to map the signal onto a subset of independent channels (i.e. non redundant). In this paper, we propose to exploit the redundancies of the neural code to find an optimal set of independent MEM dimensions that captures the information about the neural code.

Here, we propose a method based on information geometry and MEM that allow us to detect the optimal set of MEM independent dimensions capturing the information contained on neural spiking data. Specifically, we propose a method that is general enough to consider spatio-temporal interactions at several lags between neurons. The core of this approach is a matrix capturing the effect of the change of one parameter of the MEM on the second-order statistics of the network activity. This matrix is known under the name "Fisher metric" in statistics and information geometry [**?**] and Susceptibility Matrix (χ) in statistical physics. Even when this matrix is related to the effect of varying one parameter on the network activity, it can be analytically computed without actually fitting a MEM model being then possible to perform this computation in large set of neurons ($\sim$100 neurons). Based on the spectral properties of this matrix, we can find the optimal number of independent channels that better represent the neural activity.

The method would be first validated in synthetic data, where data would be generated with a known underlying statistics. So, using χ to propose a MEM with more parameters than the underlying data statistics, our method would helps to detect the dimensionality of the underlying statistics, even in the presence of unobserved events. Nevertheless, if the activity is too dense, the underlying statistics hides under the noisy activity, hindering the estimation of the underlying dimensionality.

Additionally, the method was tested in a neural population of retinal ganglion cells (RGC) *in vitro Octodon degus* recorded with 252-MEA (multi electrode array), an experimental set-up that allow to stimulate and record a population of $\sim$ 200-300 of RGC. These cells are part of the retina processing and sent it output trough the optic nerve to the brain. Our method was applied to three different set of neural response obtained after applying three type of visual stimuli: i) spontaneous photopic activity; ii) white-noise checkerboard and iii) a short natural movie. Based on our experimental observations, under i) we shows the activity is sparser: small correlations and silent cells. During ii) and iii) responses are denser, exhibiting highly correlated activity mainly driven by common inputs coming from overlapped receptive fields. As expected, we found that RGC activity is highly compressible ($\sim$ 50% of the imposed model dimensionality) and that the stimuli spatio-temporal modulation increases the number of independent dimensions required to optimally represent the neural activity. This suggest that RGC population activity is able to adapts to a stimuli conditis changing the number of coding channels according to stimuli correlations.

I

Results

Context. We give here the main elements to understand the results described in this section. Technical details are given in the section "Methods". Spike trains are denoted by ω and the state of neuron i in the time bin t is denoted by $\omega_i(t) \in \{0,1\}$ ($=1$ when spiking). MEMs are probability distributions assigning an exponential weight to spike trains. This weight (also called "energy") is a linear combination of specific terms called in the sequel "interactions" or "monomials".

A paradigmatic example is the Ising model where the energy at time t reads $\phi(t) = \sum_i b_i\omega_i(t) + \sum_{i,j} J_{ij}\omega_i(t)\omega_j(t)$. The terms b_i and J_{ij} are parameters tuning the probability, whereas the terms $\omega_i(t)$ and $\omega_i(t)\omega_j(t)$ depend on the spike configuration and are called "interactions" (self-interaction for the term $\omega_i(t)$, and pairwise interactions $\omega_i(t)\omega_j(t)$). In Ising model, there is no coupling of spikes at different times so that the probability of observing a raster of length T factorizes into probabilities of spike states at each time: successive time events are therefore independent. More generally, we can extend energy to a general form:

$$\phi = \sum_{l=1}^{L} h_l m_l. \tag{1}$$

The terms h_l are parameters tuning the probability. There are L such terms where L depends on the number of neurons. They correspond to b_i, J_{ij} in Ising model but they tune more general spikes interactions. The main difference with Ising model is that now interactions involve spikes at different times. They correspond to the terms $m_l(\omega)$, with the general form $\omega_{i_1}(t_1)\ldots\omega_{i_n}(t_n)$, i.e. it involves spike events occurring at different times. As an immediate consequence, successive time events are not independent anymore. As one can show the Gibbs distribution is in this case the invariant probability of a Markov chain [**?**]. Thus, statistics involves memory and has non vanishing time correlations. The interactions correspond thus to the conjunction of events in the raster, varying the space and time. Although interactions can have a very general form (pairwise, triplets of neurons and so on) we shall restrict in this paper to a generalization of Ising model which contains, in addition to Ising terms, pairwise interactions occurring at two successive times (e.g. $\omega_i(t)\omega_j(t+1)$).

All what preceeds is illustrated in Fig 1A. It shows a raster representing neuron population response of four neurons over time, represented as a raster plot. The neuron activity is considered as 1 (or 0) if it generated (or not) spiking activity inside a time window of width equal to the bin size. In this binary representation, we equally call interactions to either self-interaction or spatio-temporal pairwise-interactions between a pair of neurons.

The form of the energy, i.,e. the arbitrary choice of interactions, defines a MEM with which one is attempting to "explain" data, by a suitable tuning of parameters h_l. This is actually achieved by maximizing the statistical entropy under the constraints that, for each l, the average value of m_l predicted by the model (noted $\mu[m_l]$) is equal to the average observed in the experimental raster, ($\pi(m_l)$). Hence the terminology *Maximum Entropy Model* (MEM). Considering each parameter h_l as spanning a direction in a L dimensional space, a MEM is a point in this space which can be very huge. Indeed, the number of parameters increases with the number of neurons leading rapidly to a plethora.

Our method proposes an analytic way to find the optimal set of linearly independent dimensions that represents more accurately the inner structure of the neural code. It is based on measuring the interdependence between the MEM parameters and re-mapping those parameters on a different coordinate systems where each dimension is a linear combinations of them and different dimensions are linearly independent. Once we are

on this new coordinate system, we can find an optimal set of dimensions based on the geometric properties of the MEM (see eq. 13). Thus, finding the optimal number of dimensions is equivalent to finding the optimal number of independent coding channels of the neural code under this MEM assumptions: coding is based on firing rates, spatial and temporal correlations.

We have applied the method to two types of data: synthetic and retinal, where the synthetic was defined using different MEM:

1. Independent model: Neurons are independent, i.e. there are only self-interactions (firing rates) of a neuron to itself and the parameter associated to this interaction controls the firing rate of the neuron. A rate model with N neurons has therefore $L = N$ parameters. We call this model Indep., because neurons are independent.

2. Spatio-temporal pairwise interactions model with one time-step delay ($R = 2$ in Eq. 3): two neurons firing at the same time (spatial) and also with one time-step delay between them (temporal). We will call this model PWR2 to alleviate notations. A PWR2 model with N neurons has $L = \frac{N(3N-1)}{2}$ parameters (temporal self-interactions are not considered.

3. Scaled PWR2: 5 modifications of the PWR2 raster were generated multiplying all the model parameters by a given factor $f_{ac} = \{0.4,\ 0.6,\ 0.8,\ 1.2,\ 1.4\}$.

Retinal ganglion cells (RGC) of 4 different *Octodon degus in vitro* retinas were recorded using multi-electrode arrays under 3 different stimuli conditions: photopic spontaneous activity (PSA, also called spatio-temporal uniform full field); white-noise checkerboard (WN) and natural image sequence (NM); yielding a total of 151, 200, 246 and 270 neurons for each recording. We also generated shuffled versions of these rasters, where the neurons firing rates were maintained, but the dependence between them were disrupted by randomizing the individual spike trains.

General Method Description and Applications to Synthetic Data

Our analysis relies on a matrix χ, called *Susceptibility* matrix in the sequel, using here the statistical physics terminology. Its detailed definition and computation is presented in the *Methods*. Here it is sufficient to know that it is numerically computed from the empirical time correlations between monomials (see eq. 12) without needing to fit a MEM on the observable data. Susceptibility matrix has the following properties:

(i) In the MEM each interaction (monomial m_l) has a weight h_l. The set of all weights fixes the statistics predicted by the model. In particular it fixes the predicted average value of m_l, $\mu[m_l]$. A slight variation δh_l of the sole weight h_l induces both a variation $\delta\mu[m_l]$ and $\delta\mu[m_{l'}]$ of $\mu[m_l]$ and $\mu[m_{l'}]$, respectively, i.e. a change on the parameter controlling a given monomial average also affects other monomial average. One can show that $\delta\mu[m_l] = \chi_{ll'}\delta h_{l'}$. Thus, χ is a square matrix whose dimension, L, is the number of parameters fixing the model.

(ii) χ is symmetric and positive, as a consequence, it has real positive eigenvalues[1] $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k \geq \lambda_L > 0$. An example of spectra obtained for synthetic data is presented in Fig 1B.

[1]Note that the positivity holds true for the susceptibility defined exactly by (12), but its estimation from a finite raster can have small negative eigenvalues. When this happened, negative eigenvalues were removed from the spectrum.

(iii) The eigenvectors v_k of χ constitute an orthogonal basis, i.e. linear combinations of model parameters having the following interpretation. Once the set of interactions defining the model has been fixed, each set of parameters $(h_l)_{l=1}^{L}$ defines a probability i.e. assigns a definite value to the predicted probability of events. The set of $(h_l)_{l=1}^{L}$ is a vector of L dimensions, i.e., a point in $\mathbb{R}^L$. Obviously, the closest two points are in the space, the closest are the statistics they predict. As a corollary, trying to fit empirical statistics, two models corresponding to "close" points might be indistinguishable i.e. they describe equally well the empirical statistics. The notion of closeness is however ambiguous here and requires to define a proper metric. This is precisely what χ does: it defines a metric (called Fisher metric).

(iv) From the spectrum of χ one can define an ellipsoid with principal axes the eigenvectors v_k and with extension $\frac{1}{\sqrt{\lambda_k}}$ on the principal axis k (Fig 1C). Clearly, the smaller λ_k the longer the ellipsoid in the direction k. As elaborated in the methods section, this ellipsoid delimitates a volume, called *confidence volume*. Points inside the confidence volume correspond to indistinguishable models reproducing equally well the empirical observations (Fig 1C) within an accuracy ϵ (see eq. 13 for more details). As a consequence, the inverse of the eigenvalues tells us how much a small variation in the estimation of the parameters affects the statistics. For a large eigenvalue λ_k, a tiny variation in the direction v_k has a dramatic impact on statistics. On the opposite, small eigenvalues correspond to sloppy dimension where a big change in the corresponding direction has small impact. The notion of stiff and sloppy dimension in statistics is not new and has been used by several authors, including for the analysis of spike trains [**?**], but the treatment we propose, for MEM with spatio-temporal interactions (in contrast to previous papers dealing with Ising model) is, to our best knowledge, a novelty.

k_c: a parameter representing the structure of the spike train interactions

Independent on the nature of the synthetic data, we computed the χ matrix for a PWR2 statistics, i.e., overfitting the data generated from an independent model. From the χ matrix we obtained their eigenvalues spectrum, which is shown in Fig 1B.

The eigenvalue spectrum presents a particular shape which analysis is the core of our method. Moving along the spectrum from left to right (increasing index k, decreasing eigenvalue λ_k) we observe a first sharp decrease (*cut-off*) at $k = N$, for the Indep. raster (Fig 1B, black) and a corresponding minima on the volume (Fig **??**A, black square). We also observe a second cut-off further in the spectrum. The k first eigenvalues and associated eigenvectors define a k-dimensional ellipsoid with volume $\mathcal{V}(k, \epsilon)$ depending on k and ϵ (eq. 14). The relation between the volume of the ellipsoid and the number of the MEM dimensions given by k, for a given accuracy ϵ, is shown in Fig 1D. The presence of two cut-offs means that there are two values of k where the confidence volume is minimal (Fig 1D), i.e. the model is more accurately determined. The first cut-off occurs at $k = N$ and it is a trivial solution representing only the statistics given by the neuron firing rates and not capturing the spatio-temporal interactions. The second cut-off, from now on called k_c represents a second minimal point of interest, capturing the spatio-temporal interactions, and which value is proposed in this article.

According to our results, the volume of indistinguishable models depends on the number of dimensions considered in the χ matrix (k) and the imposed accuracy ϵ (Eq. 13). A large volume indicates sloppy directions that could be neglected, while a small volume is a straight-forward dimensionality reduction framework. To illustrate this idea Fig 1C shows a system that lives in a 3D space represented as an ellipsoid, whose principal axes (dimensions) are linear combinations of the original axes (x, y, z) and the

amplitude of these dimensions is given by the inverse of the corresponding eigenvalues, representing the degrees of freedom that the system has on each dimension.

Finally, k_c is conditioned to the accuracy value ϵ as shown on Fig 1D. Increasing or decreasing ϵ one order of magnitude yields no convex functions, where the minima would be at $k = 1$ in the former and $k = L$ in the later. This constrains the search of the minima to a subset of ϵ values. Extending the results for different values of ϵ shows that the number of dimensions related to the minima decays monotonically with the accuracy (ϵ) until it reaches an inflection point; this inflection point is the k_c that we use as the number of relevant dimensions (Fig S1).

k_c depending on the the network size and recording length

As we just described, k_c is a global parameter measuring the number of dimensions at a given accuracy that minimizes the confidence volume, representing the minimal set of models that are equally good explaining the observed data. Each of this dimensions, i.e. eigenvectors, are a linear combination of model parameters, as expected from the eigendecomposition process of a symmetric and positive matrix (recall that the dimensionality of the matrix is given by the number of model parameters). So, if the underlying statistics is random (no linear dependences between model parameters) and has the same dimensionality than the imposed model, k_c approximates L for an infinite raster (Fig S2A-B). Otherwise, if there is too much noise or if there are linear relationships on the underlying statistics, we expect to see $k_c < L$. Formally, we can decompose the full MEM dimensionality as $L = k_c + \kappa + N.O.$, where κ is the compressibility of the code and $N.O.$ are the Not Observed events.

[MJ: What about the recording length?, Fig S2A and S2B?]

However, increasing the network size, as stated before, increases the number of possible patterns rapidly, requiring longer and longer rasters in order to observe all the possible patterns. So, for a finite raster, k_c and L will diverge as the network size grows, given that some patterns will never occur in that time. Nevertheless, despite the unavoidable finiteness of the rasters, at $N \leq 80$ and $T = 10^6$ k_c and L are still very close to each other (Fig S2D). [MJ: missing to talk about Fig S2C]

[MJ: I also suggest include Fig S2 as a main figure of the article]

k_c values on the independent and spatio-temporal correlated cases

Using the MEM described above (Indep. and PWR2), we generated 100 different rasters ($N = 20$, $T = 10^6$) for each MEM, looked for the minimal volume and corresponding k_c (Fig **??**A). For Indep. we see two minima, where the global one is the first, showing that we can capture the second-order statistics in a raster without this kind of underlying statistics, but would not correspond to the global minima (i.e. overfitting). In the PWR2 case we see just one minima, beyond which the volume explodes. Thus, we found for Indep. case $k_c = 19 \pm 0$ and for PWR2 $k_c = 447 \pm 7.00$ (mean $\pm$ s.d.) (Fig **??**B). In the Indep. case we found that k_c corresponds almost to the full dimensionality of the underlying model ($L = 20$), while in PWR2 k_c corresponds to approximately 75% of the full underlying model dimensionality (Fig **??**C). In the first case the divergence between k_c and L (just one dimension) is explained by the presence of one low firing rate neuron (data not shown), which according to our framework, is a neglectable dimension. In the second case the divergence between k_c and L is given mainly by the unobserved events, which will be shown in more detail in the next section.

k_c as a measure of code compression

As we previously remarked, the network size does not affect the shape of the spectrum for random PWR2 rasters (Fig S2C), but just its length, i.e. the number of eigenvalues (trivially explained by the increased dimensionality of the MEM). This shape invariant property comes from the fact that the underlying statistics lacks of linear dependencies by construction, so there should be approximately one dimension per parameter. Thus, finding $k_c \ll L$ comes mainly from two facts: (i) unobserved events and (ii) linear dependencies between model parameters. The first case can be easily overcame by adding to k_c the number of unobserved events, making rasters of different recording length comparable. On the second case we are facing *compression*, where given the linear dependencies between model parameters many of them are mapped onto the same dimensions, i.e. the model can be described by less dimensions than the number of parameters. So, for an experimental recording with no unobserved events, the value of k_c shows how much all the variables of the MEM (i.e. our assumptions about the neural code) can be compressed on a lower dimensional space.

Scaling the Underlying Statistics

Five different versions of the PWR2 underlying statistics were generated by scaling the parameters of the MEM by a factor $f_{ac} = \{0.4,\ 0.6,\ 0.8,\ 1.2,\ 1.4\}$. The resulting parameters distributions (split by firing rates and pairwise interactions) and the generated monomials probabilities are shown in Fig **??**A and Fig **??**B, respectively. Recall that very negative parameters defines very unlikely monomials, while very positive the opposite. Parameters very close to zero define events which occurrence probability is near to 0.5.

Values of $f_{ac} < 1$ reduce the magnitude of the MEM parameters, but it increases both the firing rates and the pairwise interactions probabilities with negative associated parameters, compared to the original PWR2 raster, i.e. denser activity (density as the total number of spikes in a raster respect to the recording length) (Fig **??**B). Multiplying the MEM parameter by values of $f_{ac} > 1$ generates a sparser activity with low firing rates probabilities, as expected from very negative rates parameters, and many pairwise interactions vanish (the reduced width of the plot, see Fig **??**B), given the low firing rates and the big amount of negative pairwise interactions parameters (Fig **??**A).

The eigenvalue spectrum of the χ matrix, for all the cases obtained through f_{ac}, were computed (Fig **??**C). Increasing f_{ac} has 3 main effects on the spectrum: (i) the first eigenvalue is decreased, (ii) the offset spectrum is decreased, and (iii) the number of eigenvalues above the minimal observed probability is reduced (1/T, the lower limit of the ordinate on the plot). The first 2 effects come from χ computation based on monomials correlations at different time lags, so the higher the probability of the monomials (density of the raster), the denser χ is and the higher the entries of it, yielding as a consequence bigger eigenvalues. The third fact comes from the unobserved events, that will be related to vanishing eigenvalues.

Differences on the spectra reflect differences on the raster second order statistics: (i) dense rasters are dominated by high-firing rates where almost any kind of interaction is highly probable, (ii) an intermediate situation with moderate firing rates where some interactions are highly probable and some others are not, and (iii) a very sparse activity with low firing rates where many pairwise interactions are unlikely to be observed.

Regarding k_c, we see that none of the modified rasters yields a k_c comparable to the original PWR2 (Fig **??**D). In the case of $f_{ac} < 1$ there are no unobserved events (Fig **??**D), so the reduction of dimensionality should come from a different fact. The high density of the raster makes almost any kind of pattern highly probably, hiding the defined underlying statistics under all the other (not defined) observed patterns,

resulting in a $k_c = N$, like an independent case. Increasing f_{ac}, but keeping it below 1, increases k_c given that the activity become sparser, but still there are many pairwise interactions that are hiding under the increased raster density. In the case of $f_{ac} > 1$, there are many unobserved events, consequence of the very negative underlying rates parameters, yielding smaller k_c values, compared to the original PWR2. However, if we add the number of unobserved events to k_c, PWR2 and the $f_{ac} > 1$ versions become comparable, showing that the dimensionality reduction observed for those raster comes mainly from the unobserved events.

Then, dealing with random underlying statistics the dimensionality reduction will depend on the density of the raster: on one side, very dense rasters are dominated by the firing rates, hiding the underlying second-order statistics on the noisy activity. Under this situation we are not able to recover the original underlying statistics dimensionality. On the other side, very sparse rasters will show many unobserved events, reducing the observed dimensionality. However, this situation can be easily overcame adding the number of unobserved events to k_c, recovering the original underlying statistics dimensionality.

Fig 1. Dimensionality reduction framework overview. A: shows a raster representing the binary activity of $N = 4$ neurons (rows) over time (T). Green square shows a slice of this raster, which is considered as 1 (or 0) if it generated (or not) activity inside a time window of width equal to the bin size. For this raster 2 types of pairwise interactions are defined: spatial interactions (blue) and temporal interactions with $R = 2$, i.e. one time-step between spikes (red). Spatial interaction is exemplified as $w_3(0)w_4(0)$ (both neurons firing at the same time) and $w_3(1)w_4(0)$ (neuron 3 firing one bin after neuron 4). B: Susceptibility matrix eigenvalue spectrum in log-log scale for Independent (black) and Pairwise Range=2 (PWR2, red). Black vertical dashed line denotes the network size ($N = 20$). Indep. raster shows a sharp cut-off close to the network size, while PWR2 shows a monotonic decay of the eigenvalues magnitude, without a clear cut-off. C: Illustrates a 3D system representing a volume of indistinguishable models living in *x, y, z* dimensions which is rotated into a new coordinate system *x', y', z'*, that are linear combinations of the original coordinate system (original coordinate system denoted as dashed lines). Bottom shows the 2D projections, where the amplitude of the ellipsoid on each dimension is the inverse of the corresponding eigenvalue. In this case *x'* and *y'* are the stiff directions, while *z'* is a sloppy direction. D: Shows the value of the volume of indistinguishable models, $\mathcal{V}(k, \epsilon)$ as the degrees of freedom (k) increases. The volume reaches a minimum (orange dot) for certain k named k_c. Increasing k upon k_c makes the volume explode. Increasing/decreasing one order of magnitude the imposed accuracy (ϵ) yields no convex functions, i.e. no minimization.

Dimensionality reduction on retina data

We also tested the dimensionality reduction method (finding k_c) on *in vitro* retina data recorded with multi-electrode arrays under 3 stimuli conditions: photopic spontaneous activity (PSA), white noise (WN) and natural movie (NM), sorted in increasing level of stimuli high-order correlations. The stimuli high-order correlations increases the probabilities both of the firing rates and the pairwise interactions, i.e. increases the raster density. However, the distribution of the pairwise interactions probabilities of the empirical recordings and their shuffled version for all stimuli are not significantly different (bin 10ms, Mann-Whitney test $P > 0.05$, Fig S3), showing that the distributions of the pairwise interactions probabilities are given by the independent neuron firing rates. However, we are not interested in the magnitude of the pairwise

Fig 2. Dimensionality reduction by minimization of the volume of indistinguishable models. A: shows the log of volume of indistinguishable models ($log\mathcal{V}$) as function of the number of dimensions (k) at fixed accuracy (ϵ). Thick solid lines are averages, shaded area is ± 1 s.d. of the 100 temporal subsamples (of a time width equal to half of the recording length) and black vertical dashed line is the network size ($N = 20$). The minimum of this function (squares) is the optimal number of dimensions capturing the raster statistics at the given accuracy, i.e. k_c. Inset shows a zoom-in on the first 100 dimensions, focusing on the Indep. minima. B: Summary of the number of dimensions for both rasters. We found a close relationship between the underlying model dimensionality and the more dimensions required for an optimal model. k_c is 19 ± 0, and 447 ± 7.00 (mean $\pm$ std) for Indep. and PWR2, respectively. C: Is the same than (B), showing k_c as a percentage of the underlying model dimensionality. Indep case show almost 100%, while PWR2 shows $\pm$ 75% which is due to the unobserved events.

Fig 3. Dimensionality reduction on scaled PWR2 statistics. A: shows the underlying parameters distribution, split by firing rates (blue) and pairwise interactions (red). The bigger the scaling factor *fac*, the more negative the rates parameters and the more wide the interactions parameters distribution. B: shows the corresponding monomials probabilities for the scaled rasters, split as firing rates and interactions, as A. We see that increasing *fac* has the effect of decreasing both rates and pairwise interactions probabilities, reaching the point where many pairwise interactions vanish ($fac > 1$). C: are the corresponding χ eigenvalue spectra (average out of 10 rasters) for the scaled rasters. Increasing fac decreases the first eigenvalue, the spectrum offset and the number of eigenvalues above the minimal observed probability (1/T). D: k_c values for the scaled rasters. None of the scaled rasters shows k_c value as the one obtained for the original PWR2 raster. Dost are averages and errorbars 1 s.d. of 10 different rasters with the same underlying parameters. E: k_c values plus the number of unobserved events. For $fac > 1$ adding the unobserved events yields values close to the original PWR2 raster, showing that the dimensionality reduction obtained for those rasters is given mainly by the unobserved effects. For $fac < 1$ we see no unobserved events, so the dimensionality reduction obtained for those cases is given mainly by the increased density of the raster hiding the underlying statistics.

interactions, but instead, we are interested on the inner structure of the neural code, its dimensionality and compressibility, which is captured by χ and k_c.

To this end, we computed the χ eigenvalue spectrum of 30 random sub-networks of $N = 50$ from the total number of neurons recorded in each of the four experiments, under the 3 stimuli conditions. According to our synthetic rasters experiments, for $N = 50$ and $T \sim 10^6$ we can get good k_c estimates (Fig S2B and S2B), which fits with our experimental recordings. We applied the same procedure to the shuffled version of the rasters, to compare with rasters having exactly the same firing rates distribution, but with no dependency between neurons.

Experimental versus shuffled spectrum

Similar to what we obtained for scaled synthetic rasters (Fig **??**C), we see that the first eigenvalue, the offset and the vanishing eigenvalues (below 1/T) of χ spectrum increases with the raster density, which in this case is driven by the stimuli (Fig **??**A). In addition, we only see a clear cut-off close to N for WN, while the other two conditions the cut-off is not clear. Counter wise, all the shuffled rasters show a cut-off close to N, suggesting that in the experimental recordings there are significant linear dependences

between monomials that are not present in the shuffled version. Specially, NM shows a smooth decay of the spectrum, similar to the observed for PWR2, which is highly modified when the raster is shuffled, having a sharp cut-off close to N.

k_c for empirical data

Following our work flow, we minimized the volume and computed k_c for all the rasters (experimental and shuffled) at 2 different time scales (bin size): fast (1 and 5 ms) and medium (10 and 20 ms. Larger values of bin sizes were discarded because they significantly reduced the total bin number biasing the k_c estimation). As a global picture, the stimuli spatio-temporal modulation increases k_c for all bin sizes (i.e. comparing PSA with the other two stimuli) (Fig **??**C).

The k_c analysis on retinal data revealed that RGC activity is not random, showing almost the half of dimensionality compared to a random PWR2 with the same network size. Furthermore, k_c corrected by the number of unobserved events (Fig **??**D) decreases with the bin size as a consequence of a larger window in which spikes can be correlated, increasing the interdependence of RGC population activity (both for empirical and shuffled data).

Specifically, we see that k_c increases with the stimuli high-order correlation for fast time scales (Mann-Whitney test, $P < 10^{-5}$ for all comparisons). This suggests that for fast time scales the retina increases the number of coding channels as the stimuli high-order correlation increases. However, for medium time scales the pictures changes, showing no significant differences between WN and NM for 10ms (Mann-Whitney test, $P > 0.1$) and showing higher k_c values for WN than NM for 20ms (Mann-Whitney test, $P < 10^{-5}$). This suggest that at larger time scales the RGC activity under NM becomes more interdependent than for WN.

On the other hand, the shuffled rasters show higher k_c values for all the bin sizes and conditions (Mann-Whitney test, $P < 10^{-4}$), except for 1ms where is significantly small (Mann-Whitney test, $P < 10^{-4}$) and for PSA at 5ms, where they are not significantly different (Mann-Whitney test, $P > 0.1$). However, when corrected by the number of unobserved events (Fig **??**D), the picture is the same for all bin sizes: shuffled rasters have always higher k_c values than the empirical ones (Mann-Whitney test, $P < 10^{-6}$ and $P < 0.01$ for WN at 1ms). This confirms that the RGC neural code has interdependences that can be mapped onto a lower dimensional space (i.e. compressed), compared to the shuffled version, that lacks of interdependences by construction. Yet, the k_c obtained for the shuffled raster is still very small (almost a half) compared to the random PWR2 raster, which suggests that just the firing rates distribution introduces some non-random interdependences in the neural code, allowing compression.

In sum, RGC neural code is highly compressible compared to a random raster of same network size. Even compared to rasters with exactly the same firing rates distribution, that also reproduces the pairwise interactions probability distribution (Fig S3), the RGC neural code is more compressible. Nevertheless, a significant compression can be achieved considering just the firing rates distribution. Besides, stimuli high-order correlations increases the raster density and the number of independent coding channels increases with the stimuli spatio-temporal modulation. For fast time scales, k_c increases with the stimuli high-order correlations. Finally, at medium time scales the code for NM becomes more compressible than WN, suggesting higher redundancy on the code under NM.

Fig 4. Dimensionality reduction on RGC data. A: shows the average eigenvalue spectrum (solid line) of RGC data ($N = 50$) under 3 different stimuli conditions, with 10ms bin size. Stimuli high-.order statistics increases the first eigenvalue, the spectra offset and the number of eigenvalues above 1/T. Except from WN, there is no clear cut-off close to N. Shaded area is 1 s.d. out of 30 sub networks. Black line is a random PWR2 rasters of the same network size than the RGC raster. B: shows the same than A, but for the shuffled version of the empirical rasters. We see that all of them show a clear cut-off close to N and that they preserve effects induced by stimuli high-order correlations, i.e. change on the first eigenvalue, offset and number of eigenvalues above 1/T. The differences on the cut-off suggests that experimental data has an inner structures that is not merely explained by firing rates. C: show box plots for k_c values for empirical and shuffled rasters. Except from 1ms, k_c is higher for shuffled data in all bin sizes, under all conditions. Stimuli spatio-temporal modulation significantly increases k_c. Also, neither of both type of rasters reaches the k_c values obtained for a PWR2 rasters, which shows almost no dimensionality reduction. Note that k_c decreases as the bin size increases, showing that for bigger bin sizes the inner structure of the code becomes more interdependent. For fast time scales (1 and 5ms), k_c increases with the stimuli-high order correlations, but at 10ms WN and Nm are not significantly different and for 20 ms WN is larger than NM, suggesting that bigger bin sizes capture more redundancies on the code. D: shows box plots k_c values corrected by the number of unobserved events. Now for all bin sizes the shuffled data has bigger values than the empirical one, showing that the RGC data has an inner structure with interdependences between variables that is not present on a shuffled raster. However, shuffled data has values almost a half than a PWR2 raster of the same size, suggesting that the firing rates distribution only is enough to perform compression on the neural code. The effect of stimuli spatio-temporal modulation and stimuli high-order statistics remain.

Discussion

In this paper we have proposed a method to reduce the dimensionality of MEM on artificial and biological spiking networks. It is grounded on information geometry of the matrix χ, which characterizes how a small variation of parameters impacts the statistical estimations. The χ matrix captures the interdependences between the neural code variables. After an eigendecomposition process, the eigenvalue spectrum of χ exhibit two cut-offs. The first one shows that, both in synthetic as well as in retina data, large part of statistics is "explained" by the neurons firing rates. On the opposite, the second cut-off (here called k_c) reflects a non trivial effect associated with higher order statistics. As the eigendirections on the right part of the cut-off correspond to noise, the spectrum lying between the two cut-off contains a relevant information associated to statistics of second order.

The reduction of the MEM dimensionality is directly linked with data compression, obtained from the linear dependencies between variables of the MEM, i.e., higher order interactions between neurons. For example our analysis in the case of synthetic rasters, where both the firing rates and the pairwise interactions are defined randomly, show a k_c value very close to L (maximal dimensionality). This demonstrates that if there are no linear relationships between the parameters by construction, the code is not compressible and we have almost one dimension per parameter. Counter wise, retina data show a significant compression of $\sim 50\%$, as expected from a neural tissue where cells are driven by common inputs and cells are electrically coupled [**?**], increasing the level of dependency between them. This compression reduces the dimensionality of the MEM to a lower dimensional space, where each dimension is a linear combination of model parameters, characterizing the population activity by a set of independent

dimensions representing the inner structure of the network activity.

Limits of the method

[MJ: BRUNO: from where this method comes from? or this manner to compute the susceptibility matrix is a result of this work?] The first limitation of our method comes from the numerical approximation used to compute χ matrix, which is obtained by summing monomials correlations at different time lags (Eq. 12). In general, this function decays exponentially with time, but adding more time lags on this sum adds more noise to the matrix, reaching a point where the matrix is highly dominated by noise. To truncate this approximation we need to consider a trade-off between temporal resolution and reduction of noise. To this end we use 4 time lags (i.e. 4 terms of the sum), which is equivalent to the double memory depth used in the model ($R = 2$). This numerical estimation of χ also imposes limits on the method, given that considering the infinite memory case ($R \to \infty$) will generate a χ matrix that is governed by noise.

On the other hand, it is possible to compute χ from the model parameters, requiring a previous model fitting step, as done by [**?**]. However, for $N > 20$ and $R > 1$ this computation becomes more and more prohibitive as the network size and the memory depth of the model increases. So, despite the numerical approximation and its intrinsic errors, computing χ from the empirical monomials time correlations is the best approach we found to work with medium size networks and spatio-temporal constrains.

Furthermore, computing χ this way, instead of fitting the MEM in advance, give us geometrical information about the model and, ultimately, about the linear dependences of the neural code. So, it is possible to have insights about how a neural population modulates its activity given the stimuli, increasing or decreasing the number of independent coding channels, without a fiting procedure. Not less important, fitting the model will give information about the sign (positive or negative) and the magnitude of the interactions (weak or strong), giving detailed information about the network topology. Nevertheless, the scope of this work was not to fit different models on data and test its performance (e.g. Bayesian or Akaike information criterion that takes into account the model parameters and likelihood [**?**]) neither study changes in network topology under different stimuli. Instead, we focus on exploring the geometrical properties of the MEM and its meaning in terms of the neural code redundancy and compressibility.

k_c is not one value, but a set of values

The main challenge this methodology introduces is the selection of k_c. The selection of k_c is related to the minimization of the so-called confidence volume, which not only depends on the number of dimensions given by k_c, but also on the imposed accuracy ϵ. As shown on Fig S1, the confidence volume ($log\mathcal{V}(k_c, \epsilon)$) decays monotonically as k_c increases and ϵ decreases, reaching an inflection point where even if we decrease the accuracy by orders of magnitude k_c remains almost unchanged, so adding more dimensions only reduces the accuracy. It could be also possible to choose the first cut-off as k_c (or any other k), based on the difference between consecutive eigenvalues (another criteria to find the cut-off on eigenvalue spectra), but from our observations on synthetic data we know that the first cut-off is related mainly to the firing rates, so it misses the raster high-order statistics. We also know that even when the underlying statistics defines pairwise interactions but the firing rates are too high, the density of the raster hides the higher order statistics, yielding $k_c = N$, considering the dimensions related to higher order moments as noise.

Even though we found two minima on the volume function, the global minima for experimental and shuffled data was always at the second minima. The only case where

the first minima was the global minima was for the Indep. synthetic raster (Fig **??**A, black trace), which is in agreement with the underlying statistics.

Comparison with similar analysis

To our knowledge, there is no previous work related to the analysis of the χ matrix considering spatio-temporal pairwise interactions applied to neuronal networks. The authors in [**?**] proposed a similar analysis considering only spatial interactions, i.e. the Fisher Information Matrix (FIM) for the Ising model on *in vitro*, *in vivo* and *in silico* networks. The work of [**?**] studies small neural networks (10 neurons), they looked for stiff neurons, which are related to stiff dimensions (the largest FIM eigenvalues), proposing that those neurons are the ones giving stability to the network, while the neurons involved on the sloppy dimensions (dimensions where the parameters can have significant changes without affecting the model) are the ones involved on plasticity, allowing the network to remodel its connections. On our approach, which involves larger networks ($N = 50$), we deal with neurons and their spatio-temporal interactions, exploiting the linear dependences between them to find a minimal set of dimensions that better represents the neural code. To this end, we found analytically two set of stiff dimensions: the ones before the first cut-off, related mainly to neuron firing rates and the second set, after the first cut-off, related to spatio-temporal interactions. According to our framework, the sloppy dimension would be the ones beyond k_c, but we could also interpret the first N dimensions as the stiff dimensions, the ones between N and k_c as the sloppy dimensions and the ones beyond k_c just noisy dimensions. This re-interpretation arises from the large magnitude difference between the first and second set of dimensions (at least one order of magnitude).

In our case, the analysis was extended to spatio-temporal interactions in larger neural networks, and more importantly, we used χ as a tool for finding linear dependences between the neural code variables, compressing the observed data and characterizing the neural activity on the basis of independent dimensions.

Comparing the inner structure of the relevant dimensions between different stimuli conditions could give us insights about which channels are invariant between stimuli and which ones are adaptive/plastic. However, this is not the scope of this article and the analysis of invariant or adaptive channels will be kept for a future work.

Recently, Battaglia et al [**?**], studying large scale networks between brain areas, proposed a concept called *Meta Connectivity*, which instead of analysing the correlation between nodes of a network (the usual functional/effective connectivity analysis), they focus on the correlations between the network interaction along time. This means focusing not on the coupling $w_i(0)w_j(0)$ between neurons w_i and w_j, but focusing on the interactions between the couplings $w_i(0)w_j(0)$ and $w_k(0)w_l(0)$. This provides information about high-order correlation for at least 3 nodes of the network (e.g. case of $i = k$) and captures the relationships between modules of network activity. Thus, χ matrix is both a functional/effective connectivity matrix (the matrix entries related to the correlations between firing rates) and also a meta connectivity matrix (the matrix entries related to correlations between pairwise interactions), providing information about the spatio-temporal interaction between network nodes and network modules. The extension of our analysis towards the understanding of the meta connectivity has never been applied to networks of neurons and we proposed it as a future research direction, where the focus is on the variability and dependence of the interactions.

Stimuli-induced changes on RGC population activity

Retina data has significant high-order correlations, including pairwise spatial [14, 15], temporal [12] interactions, triplets [5] and groups of neurons [6]. This correlations have

been widely studied under MEM, which can accurately reproduce the raster spatio-temporal patterns. Nevertheless, there has been no work devoted to reduce the MEM dimensionality considering the inner dependences between the population activity variables.

Here, as a proof of concept, we used retina data of a diurnal rodent under 3 different stimuli with different statistics, from photopic spontaneous activity (spatio-temporal uniform full field, no second order statistics), a spatio-temporal white noise (gaussian statistics) to a repeated natural movie (high-order correlations both on time and space). From our analysis, we know that this stimuli high-order correlations increases the magnitude of both the firing rates and the raster high-order correlations, even making silent cells to fire. This could be related to recruitment of specific cell types by the stimuli features (e.g. local contrast [**?**], optic flow, color [**?**], among others).

In general, most of the correlated activity observed in the retina could be attributed to the receptive field overlap between recorded RGCs and to shared common noise [**?**]. Additionally, electrical coupling are highly present in the O. degus retina [**?**] inducing fast correlations between close cells. These three effects modifies the pairwise spatio-temporal interactions observed in real data. We explored the presence of these correlated activity in the neural code. To do this we compared the results of recorded data with a suffled version of it, which preserves the firing rates distribution. Interestingly, both data sets share the pairwise interactions probability distribution suggesting that statistics up to the second order are preserved. Nevertheless, our method exhibits significant differences in the compression capability of each data set. RGC presents a maximal compression compared to the shuffled data, suggesting that statistics of higher order are needed to fully characterize the response of RGCs.

Compressibility of the RGC code

Then, in order to study the compressibility of the RGC code, we studied χ spectrum and k_c for RGC under three stimuli conditions, finding that RGC population code adapts to stimuli conditions by changing the number of independent channels.

The first difference we found between stimuli was on χ spectrum, which in correspondence with the stimulus-induced increase of the monomials probability (Fig S3), shows an increase on the offset, i.e. the eigenvalues increases their magnitude as the stimuli high-order correlation increases. But given the way χ is computed (see Eq. 11), the shape of the spectrum comes not only from the increased monomials probabilities, but also from the dependence between firing rates and spatio-temporal interactions and the dependence between spatial and temporal interactions, all of them captured by the matrix. On the other hand, shuffled data, where there are no dependences between the neural code variables, exhibits these differences on the eigenvalue magnitudes (Fig **??**B), but changing the first cut-off, showing that eigenvalues magnitude are closely related to the raster density (also shown for synthetic data on fig **??**C), while the cut-off is related with the linear dependences between the monomials.

The second difference we found between stimuli is on k_c, i.e. our approximation to the dimensionality of the neural code. As expected from the stimuli statistics, k_c (and also its corrected version by the unobserved events) is always lower for PSA than for the dynamic stimuli, suggesting that the network optimizes the number of dimensions required for coding the stimuli depending on the stimuli statistics. In terms of metabolic cost, a very redundant stimulus as PSA (which has the same spatial and temporal information all over the stimuli space), should be coded with less dimensions than a stimuli with more independent components (less redundant), thus, optimizing the metabolic resources.

However, for bin sizes of 1 and 5ms we observe that k_c is higher for NM than for WN. This relation varies for larger values of bin sizes, such as 10 or 20ms. For bin sizes

of 10ms WN has the same number of relevant dimensions than NM, while for 20ms WN has more dimensions than NM. So, for fast time scales, we face a non-optimal situation, because NM has more redundancies than the WN. It could be possible that at this time scales we are not capturing the inter-dependences of the neural code that are relevant for the brain, given that at 20 ms bin size we see the expected effect: the system exploits the stimuli redundancies and exhibit more compression for NM than for WN. Coincidently, many MEM on retina have been done using 20 ms as bin size [14, 15], which in our case is the bin that allow the highest compression. In addition, synthetic data shows that if the raster is too dense the underlying statistics hides under the noisy activity, which could also be possible at large values of bin sizes. To control this situation we used shuffled rasters, which preserves the same raster density. In the shuffled rasters we see that k_c increases with the raster density, discarding the effect of density on the changes of dimensionality at higher bin sizes. Thus, at large bin sizes the interdependences of the neural code are responsible for the compression effect and not the raster density hiding some events. Nevertheless, we do not know in advance what time scale(s) is(are) actually relevant for the brain and neither if our assumptions about the neural code (firing rate and spatio-temporal interactions) are right, so the choice of the bin size and code variables is still an open question and somewhat arbitrary.

Finally, our work is related to the idea that a stimuli-dependent network of *in silico* noisy spiking neurons adapts its code according to noise and stimuli correlations [**?**], instead of using just one way of coding. On our case, the stimili-dependent network is a biological one, so we don't have access to modify the noise of each neurons nor the network noise. Instead, we can just modify the stimuli correlations, which changes the dimensionality of the code. This change in dimensionality could reflect the smooth interpolation between encoding strategies: highly redundant stimuli evokes less code dimensions than stimuli which presents high-order correlations, suggesting that the MEM dimensionality is also a measure of the code redundancies. However, the analytic relationship between dimensionality reduction and the coding strategies requires an extensive mathematical and computational research that is not developed here, but we give an indirect way of studying the interdependences of the neural code as the stimuli conditions changes.

Materials and Methods

Model Definition

A spike train, ω, is a discrete-time sequence of events that represent the activity of a neural network of size N over a time period T time-steps. The spiking activity of neuron i at discrete time t is mathematically defined by a variable:

$$\omega_i(t) = \begin{cases} 1, & \text{if neuron } i \text{ fires at time } t, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

A *spike pattern* $\{\omega(t)\}_{i=1\ldots N}$, represents the activity of the whole network at a given time t. Finally, a *spike block*, $\omega_{t_1}^{t_2} = \{\omega(t)_{t_1 \leq t \leq t_2}\}$, represents the activity of the whole network between two instants t_1 and t_2. The *range* of a spike block is the number of time steps $t_2 - t_1 + 1$; the *degree* is its number of spike-events.

A monomial is a function which associates to a raster a real value:

$$m_l(\omega) = \prod_{k=1}^{m} \omega_{i_k}(t_k),\ 0 \leq i_k < N, 0 \leq t_k < R. \tag{3}$$

One fixes a set of pairs (i_k, t_k) (neuron, time), and $m_l(\omega) = 1$ if and only if neuron i_k spikes at time t_k in the *raster block* ω, and $m_l(\omega) = 0$ otherwise. The simplest example

is $\omega_i(t)$ which is 1 if and only if neuron i fires at time t in spike train ω. In (3), the monomial value depends in fact on the $R-1$ first time steps in the raster, i.e. on the block ω_0^{R-1}, instead of the whole raster. R is called the *range* of the monomial. From now on we note $D = R-1$.

We note $\pi_\omega^{(T)}$ the empirical probability constructed from the raster ω, where the average is performed by time average over a time interval of length T. This way of averaging is relevant only if one assumes that the underlying process that generated the raster is *stationary*. Thus, the empirical average of a monomial m_l is noted $\pi_\omega^{(T)}[m_l]$. We want to construct a stationary Markov chain (i.e. where transition probabilities are time-translation invariant) whose invariant distribution μ "approaches at best" the empirical data. Approaching at best means here minimizing Kullback-Leibler divergence $d_{KL}\left[\pi_\omega^{(T)} \mid \mu\right]$ between the empirical probability of the spike train ω of size, $\pi_\omega^{(T)}$, and the invariant probability distribution of the estimated Markov chain, μ. For this, we use the Maximum Entropy Principle (MEM).

As opposite to most existing MEM, like Ising model [14–18] or extensions with triplets and quadruplets interactions [5], we consider here models having a memory so as to treat time correlations. The main developments of the method and application to retina data have been published in [10, 11, 19, 20]. The Maximum Entropy Principle corresponds to finding a stationary probability distribution μ which maximizes the statistical entropy $S(\mu)$ given the constraints:

$$\mu[m_l] = \pi_\omega^{(T)}[m_l], l \in \mathcal{S}. \tag{4}$$

This means the average $\mu[m_l]$ of monomial m_l predicted by the model, equals the raster empirical average $\pi_\omega^{(T)}[m_l]$, where m_l belongs to a prescribed list of monomials $\mathcal{S}$, where $\mathcal{S}$ defines the model-type. If $\mathcal{S}$ contains only monomials of the form $\omega_i(0)$ the model is Bernoulli. If $\mathcal{S}$ contains monomials of the form $\omega_i(0)$ and $\omega_i(0)\omega_j(0)$, $i \neq j$ the model is Ising. If $\mathcal{S}$ contains monomials of the form $\omega_i(0)$; $\omega_i(0)\omega_j(r)$, $i \neq j$, $r = 0 \ldots R$ the model is pairwise with range R. Here, a model with range $R = 2$ is used.

It can be shown that there is a unique probability distribution $\mu_\mathcal{S}$, depending on the set of constraints $\mathcal{S}$, maximizing the statistical entropy under the constraints (4). $\mu_\mathcal{S}$ has the following properties:

$$\phi = \sum_{l \in \mathcal{S}} h_l m_l. \tag{5}$$

is called "potential" such that $\mu_\mathcal{S}$ is a Gibbs distribution for ϕ. Being a Gibbs distribution means that the probability of a spike block ω_0^n obeys:

$$\mu_\mathcal{S}\left[\omega_0^{n-1}\right] \sim e^{-(n-D)\mathcal{P}}\, e^{\phi\omega_0^{n-1}} \tag{6}$$

where:

$$\mathcal{P} = \lim_{n\to\infty} \frac{1}{n} \log Z_n, \tag{7}$$

is called *free energy density* and

$$Z_n = \sum_{\omega_0^{n-1}} e^{\phi\omega_0^{n-1}}. \tag{8}$$

is called n *time steps partition function*

Susceptibility Matrix, χ

$\mathcal{P}$ is a function of h_ls. One can show that the average of m_l predicted by the Gibbs distribution $\mu_{\mathcal{S}}$ is given by the derivative of the free energy with respect to h_l [21] :

$$\mu_{\mathcal{S}}\left[m_l\right] = \frac{\partial \mathcal{P}}{\partial h_l}. \tag{9}$$

Therefore, the second derivative:

$$\chi_{ll'} = \frac{\partial^2 \mathcal{P}}{\partial h_l \, \partial h_{l'}}, \tag{10}$$

tells us how much the average of the monomial m_l is modified when we slightly modify the coefficient of monomial $m_{l'}$. We call the matrix χ with entries $\chi_{ll'}$ the *susceptibility matrix*.

χ is numerically computed without the need of fit the model. This is achieved by computing the correlations between monomial m_l at time 0 and monomial l' at time n is:

$$\mathcal{C}_{l,l'}\left(n\right) = \mu[m_l(0)\, m_{l'}(n)] - \mu[m_l]\mu[m_{l'}]. \tag{11}$$

In the case where $m_l = \omega_i(0), m_{l'} = \omega_j(0)$, this gives the time pairwise correlations between neurons i,j at time n. $\mathcal{C}_{l,l'}\left(n\right)$ depends on the model $\mathcal{S}$.

One can show that:

$$\chi_{ll'} = \mathcal{C}_{l,l'}\left(0\right) + 2\sum_{n=1}^{+\infty} \mathcal{C}_{l,l'}\left(n\right). \tag{12}$$

Thus, $\chi_{ll'}$ integrates *all time correlations* between monomial m_l and $m_{l'}$. χ is symmetric positive thus it has real positive eigenvalues, but given its numerical computation, as more terms are considered on the sum of (12) more noise is captured by χ, yielding several negative eigenvalues (data not shown). For that reason, n was constrained to 4, being the double of the range of the model. When the potential ϕ only contains range 1 monomials successive time steps are uncorrelated and (12) reduces to

$$\chi_{ll'} = \mathcal{C}_{l,l'}\left(0\right)$$

the so-called fluctuation-dissipation theorem in physics.

Volume of indistinguishable models

Functions of the form (5), parametrized by the coefficients h_l, live in a L dimensional space $\mathcal{E}$. Since each set of h_l defines a maximum entropy (Gibbs) probability distribution, each point of $\mathcal{E}$ represents such a probability. This space is very huge, and a *generic* probability in this space have all h_ls different from 0. Depending on the dynamics and specific properties of the studied system, the h_ls are typically distributed onto a sub-manifold $\mathcal{W}$ in $\mathcal{E}$, corresponding to functional relations between the h_ls parameters, [9]. The right object to study *both* the geometric structure (metric) of the manifold $\mathcal{W}$ and the second order statistical fluctuations associated with the central limit theorem is the Fisher matrix or Susceptibility matrix [22–25]. Given χ properties, it has eigenvalues $\lambda_k, k = 1 \ldots L$ ordered by decreasing value (i.e. the spectrum) without loss of generality and corresponding eigenvector $\vec{v}_k$ associated with λ_k. The $\vec{v}_k$s define an orthogonal eigenbasis composed of a linear combinations of model parameters. From the statistical perspective, $\frac{1}{\sqrt{\lambda_k}}$ gives the amplitude of second-order fluctuations in the estimation of coefficients h_l projected on direction $\vec{v}_k$. The smaller the eigenvalue, the larger the amplitude. When projecting back to the original basis, these errors

propagate to h_ls estimation. From the geometrical perspective, as χ is a metric, i.e. the inverse of a square distance, $\frac{1}{\sqrt{\lambda_k}}$ is a distance on the manifold $\mathcal{W}$. This distance can be interpreted as a scale where, in direction k, $\mathcal{W}$ is essentially flat. This is also a characteristic distance on which models (projected on direction k) are indistinguishable. From the linear response perspective, $\frac{1}{\lambda_k}$ tells us how much a small variation in the estimation of average of monomials (a linear combination of these variations parallel to $\vec{v}_k$) affects the estimation of h_l. Thus, following Mastromateo [23] two distributions $\mu^{(1)}$, $\mu^{(2)}$ are indistinguishable with accuracy ϵ if $-\log \mu^* \left[h^{(1)} = h^{(2)}\right] \leq \epsilon$. If T is large enough, the set of indistinguishable distributions defines an ellipsoid s in $\mathcal{E}$ with a volume $\mathcal{V}$:

$$\mathcal{V} = \frac{1}{\sqrt{\det \chi}} \left[\frac{1}{\Gamma(\frac{L}{2}+1)} \left(\frac{2\pi\epsilon}{T} \right)^{\frac{L}{2}} \right]. \tag{13}$$

Thus, when the sample size T, the model dimension L and the accuracy ϵ are fixed the log of the volume is proportional to $\log \mathcal{V} \propto -\frac{1}{2}\sum_{k=1}^{L} \log \lambda_k$. Thus, model estimation is better if λ_hs are larger, which implies that there exists a set of eigenvalues that can be neglected given their small magnitude, i.e. huge fluctuations impairing the model estimation. Then, instead of considering the volume of s in the space of all parameters, let us consider the volume $\mathcal{V}(k)$ of the projection of s in the subspace spanned by the k first eigenvectors of χ. We have:

$$\log \mathcal{V}(k) = \frac{1}{2} S_\phi(k) - \log \Gamma(\frac{k}{2}+1) + \frac{k}{2} \log \left(\frac{2\pi\epsilon}{T} \right), \tag{14}$$

with:

$$S_\phi(k) = -\sum_{i=1}^{k} \log \lambda_i, \tag{15}$$

where λ_is are ordered by decreasing values. In eq. (14), The second term, $\log \Gamma(\frac{k}{2}+1)$ is purely combinatory, whereas the third term, $\log\left(\frac{2\pi\epsilon}{T}\right)$ characterizes the effect of precision accuracy and finite sampling. Therefore, the only term which depends on the statistical model (here the potential $\mathcal{H}$) is $S_{\mathcal{H}}(k)$. In particular, it depends on the number of neurons N and memory depth R.

Finding the cut-off, k_c

As eigenvalues k increases, there will be one or more ks at which $S_\phi(k)$ is expected to become bigger than the sum of the other two terms of (14), which are both negative. This means that, for increasing k, $\log \mathcal{V}(k)$ will first decrease then it will increase, yielding at least one minima on the function. There is therefore a k, $k_c \equiv k_c(\epsilon, T)$, which characterizes the number of eigenvectors (i.e. dimension) for which the volume is minimal. Precisely, as we observed, ϵ belongs to some interval; outside this interval $\log \mathcal{V}(k)$ is not convex any more. On this range over ϵ we obtain a set of $k_c \equiv k_c(T)$ minimizing the volume, which characterize the number of eigenvectors ensuring an optimal fit with the model: the model indeterminacy is minimized. Among the set of possible k_c, we chose the smallest one, representing the minimum set of dimensions necessary to reliable fit the model given data. The corresponding eigenvectors span a subspace $\mathcal{E}^r$ of $\mathcal{E}$, embedded in the tangent space of $\mathcal{W}$ at the point corresponding to the empirical measure. This subspace corresponds to linear combination of h_ls giving the most reliable model. Getting outside this space, by adding/subtracting degree of freedom/eigenvectors leads to a worse approximation of the empirical distribution.

Synthetic Raster Generation

Synthetic rasters ($T = 2 * 10^6$ time-points, $N = 20$ neurons) were generated using different underlying statistics: **Independent**, where only firing rates are defined, $L = N$; **Pairwise R=2** (PWR2), with firing rates and spatio-temporal correlations, $L = N(3N - 1)/2$. Using the underlying potential of the PWR2, 5 more rasters were generated scaling the magnitude of the model parameters by a factor $fac = [0.4\ 0.6\ 0.8\ 1.2\ 1.4]$. This scaling mimics the effect of the stimuli high-order statistics, which increases both the firing rate and the spatio-temporal correlations. In addition, 6 more random PWR2 rasters were generated with $N = [30\ 40\ 50\ 60\ 70\ 80]$ to study the dependence between k_c estimation and the network size. Each raster was generated using the ENAS software (https://enas.inria.fr/). For the first 3 rasters, 100 random subsamples with half duration of the whole recording were taken for each raster and the χ matrix associated with a Pairwise R=2 model was computed. Then, k_c was found by volume minimization, yielding 100 k_c values per raster. For the scales rasters, the same procedure was applied, but using 10 temporal subsamples.

Shuffling

In order to destroy the dependencies between the empirical raster monomials, we generated random raster where the number of neurons and firing rates was exactly the same than observed on the recordings (i.e. on each retina under each condition), but the spikes times were otherwise random, avoiding violations of the refractory period (2ms).

Recordings and Stimuli

Animals and Recordings

4 Adult male and female Octodon degus (3-6 months) were maintained in the animal facility of the Universidad de Valparaiso, at 20–25°C on a 12-h light-dark cycle, with access to food and water ad-libitum. The methods of MEA recording has previouslly been described [**?**]. In brief, experiments were approved by the bioethics committee of the Universidad de Valparaiso, in accordance with the bioethics regulation of the Chilean Research Council (CONICYT) and international protocols. Animals were euthanized under deep isofluorano or halothane anesthesia and both eyes were extracted. Then, one of the extracted retinas was diced into quarters while the other was stored in oxygenated in oxygenated (O_2 95% CO_2 5%) AMES medium at 32°C in the dark for further experiments. The same AMES media was used for continuous perfusion during extracellular recordings. For MEA recordings (MEA USB-256, 20kHz sample, Multichannel Systems GmbH, Germany), one piece of retina was mounted onto a dialysis membrane placed into a ring device mounted in a traveling (up/down) cylinder, which was moved to contact the electrode surface of the MEA recording array. Data were processed off-line using Plexon Offline Sorter (Plexon Instruments). Further, spikes were detected using a threshold of -4.5 to -5.5 S.D. from the mean voltage value and then were manually classified using the 2D space of the first two principal components on each electrode. Only somatic spikes were kept. Refractory period violations were detected and discarded if two or more spikes of the same neuron occur in a 2ms period. We recorded on 3 stimuli conditions (see next section): i) Photopic Spontaneous Activity (PSA), ii) Spatio-temporal white Noise (WN) and iii) Natural Movie (NM), obtaining 151, 200, 246 and 270 RGC for each of the 4 experiments, respectively. For each experiment 30 random subsamples of 50 neurons were taken and χ matrix corresponding to a Pairwise R=2 model were computed considering 4 bin sizes (1, 5, 10 and 20 ms) for each subsample, yielding 30 k_c values per experiment, per condition and per bin size. The shuffled version of these rasters were submitted to the same analysis.

Visual Stimuli

Visual stimuli were generated by a custom software created with PsychoToolbox (Matlab) on a MiniMac Apple computer and projected onto the retina with a LED projector (PLED-W500, Viewsonic, USA) equipped with an electronic shutter (Vincent Associates, Rochester, USA) and connected to an inverted microscope (Lens 4x, Eclipse TE2000, NIKON, Japan). The image was conformed by 380 x 380 pixels, each covering $5\mu m^2$. Since rodents are dichromatic (green and blue/UV cones), in our experiments only the B (blue) and G (green) beams of the projector were used, while the R (red) channel was used for signal synchronization. Dark spontaneous activity where recorded in order to monitor the stabilization of the activity. The, stimuli where applied. For PSA a space-time invariant stimuli with G and B intensities equal to the mean intensity of the NM stimulus were presented for 15 mins. WN stimulus with a block size of $50\mu m$ was used at rate of 60 fps and presented for 20 mins, with each block taking independently 0 or 255 (max value) in the pixel value scale. NM consisted on a 1800 frames movie recorded on the natural habitat of the rodent using a robotic solution to capture the natural visual environment of degus, including grass, trees, optic flow, head-like movements. This short movie were presented 40 times at a refresh rate of 60fps, yielding a total duration of 20 mins. Optical density filters in the optical path were used to control final light intensity. A CCD camera (Pixelfly, PCO, USA) attached to the microscope was used for online visualization and calibration of the light stimuli projected onto the recording array.

Supporting Information

S1 Fig. Finding k_c on the set of convex volume functions.

S2 Fig. Most of the pairwise interactions are both spatial and temporal.

S3 Fig. Logistic fit to porcentage of remaining pairwise parameters.

Acknowledgments

Financial support: CONICYT-FONDECYT 1140403 and 1150638, CONICYT-Basal Project FB0008, Millennium Institute ICM-P09-022-F, ECOS-Conicyt C13E06, ONR Research Grant #N62909-14-1-N121.

References

1. Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature. 2006;440(7087):1007–12. doi:10.1038/nature04701.

2. Shlens J, Field GD, Gauthier JL, Grivich MI, Petrusca D, Sher A, et al. The structure of multi-neuron firing patterns in primate retina. J Neurosci. 2006;26(32):8254–66. doi:10.1523/JNEUROSCI.1282-06.2006.

3. Tkacik G, Prentice JS, Balasubramanian V, Schneidman E. Optimal population coding by noisy spiking neurons. Proc Natl Acad Sci U S A. 2010;107:14419–14424. doi:10.1073/pnas.1004906107.

4. Ganmor E, Segev R, Schneidman E. The architecture of functional interaction networks in the retina. The journal of neuroscience. 2011;31(8):3044–3054.

5. Ganmor E, Segev R, Schneidman E. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. PNAS. 2011;108(23):9679–9684.

6. Tkacik G, Marre O, Mora T, Amodei D, 2nd MJB, Bialek W. The simplest maximum entropy model for collective behavior in a neural network. J Stat Mech. 2013; p. P03011.

7. Marre O, El Boustani S, Frégnac Y, Destexhe A. Prediction of spatiotemporal patterns of neural activity from pairwise correlations. Physical review letters. 2009;102(13).

8. Cofré R, Cessac B. Dynamics and spike trains statistics in conductance-based Integrate-and-Fire neural networks with chemical and electric synapses. Chaos, Solitons and Fractals. 2013;50(8):13–31.

9. Cofre R, Cessac B. Exact computation of the maximum-entropy potential of spiking neural-network models. Phys Rev E. 2014;89(052117).

10. Nasser H, Marre O, Cessac B. Spatio-temporal spike train analysis for large scale networks using the maximum entropy principle and Montecarlo method. Journal of Statistical Mechanics: Theory and Experiment. 2013;2013(03):P03006.

11. Nasser H, Cessac B. Parameters estimation for spatio-temporal maximum entropy distributions: application to neural spike trains. Entropy. 2014;16(4):2244–2277. doi:10.3390/e16042244.

12. Vasquez JC, Marre O, Palacios A, Berry MJ, Cessac B. Gibbs distribution analysis of temporal correlation structure on multicell spike trains from retina ganglion cells. J Physiol Paris. 2012;106(3-4):120–127.

13. Ganmor E, Segev R, Schneidman E. The architecture of functional interaction networks in the retina. J Neurosci. 2011;31(8):3044–54. doi:10.1523/JNEUROSCI.3682-10.2011.

14. Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature. 2006;440(7087):1007–1012.

15. Shlens J, Field GD, Gauthier JL, Grivich MI, Petrusca D, Sher A, et al. The Structure of Multi-Neuron Firing Patterns in Primate Retina. Journal of Neuroscience. 2006;26(32):8254.

16. Tkačik G, Schneidman E, Berry II MJ, Bialek W. Ising models for networks of real neurons. arXiv q-bio/0611072. 2006;.

17. Tkačik G, Schneidman E, Berry II MJ, Bialek W. Spin glass models for a network of real neurons. arXiv preprint arXiv:09125409. 2009;.

18. Tkacik G, Prentice JS, Balasubramanian V, Schneidman E. Optimal population coding by noisy spiking neurons. PNAS. 2010;107(32):14419–14424.

19. Vasquez JC, Palacios A, Marre O, II MJB, Cessac B. Gibbs distribution analysis of temporal correlation structure on multicell spike trains from retina ganglion cells. J Physiol Paris. 2012;106(3-4):120–127.

20. Cessac B, Palacios A. Spike train statistics from empirical facts to theory: the case of the retina. In: Cazals F, Kornprobst P, editors. Modeling in Computational Biology and Biomedicine: A Multidisciplinary Endeavor. Lectures Notes in Mathematical and Computational Biology (LNMCB). Springer-Verlag; 2012.

21. Keller G. Equilibrium States in Ergodic Theory. Cambridge University Press; 1998.

22. Nakahara H, Amari S. Information-Geometric Decomposition in Spike Analysis. In: NIPS; 2001. p. 253–260.

23. Mastromatteo I. On the typical properties of inverse problems in statistical mechanics; 2013. Available from: http://arxiv.org/abs/1311.0190.

24. Sessak V. Inverse problems in spin models [Theses]. Université Pierre et Marie Curie - Paris VI; 2010. Available from: https://tel.archives-ouvertes.fr/tel-00525040.

25. Sessak V, Monasson R. Small-correlation expansions for the inverse Ising problem. Journal of Physics A: Mathematical and Theoretical. 2009;42(5):055001.