



Dimensionality Reduction on Spatio-Temporal Maximum Entropy Models of Spiking Networks

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

► **To cite this version:**

Rubén Herzog, Maria-Jose Escobar, Rodrigo Cofré, Adrian Palacios, Bruno Cessac. Dimensionality Reduction on Spatio-Temporal Maximum Entropy Models of Spiking Networks. 2018. hal-01917485

HAL Id: hal-01917485

<https://hal.inria.fr/hal-01917485>

Preprint submitted on 9 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 2 Dimensionality Reduction on Spatio-Temporal Maximum Entropy Models of 3 Spiking Networks

4
5 Rubén Herzog^{1,5*}, María-José Escobar², Rodrigo Cofre³, Adrián G. Palacios^{1,5}, Bruno Cessac⁴,

6 **1** Centro Interdisciplinario de Neurociencia de Valparaíso, Universidad de Valparaíso, Valparaíso,
7 Chile.

8 **2** Departamento de Electronica, Universidad Tecnica Federico Santa Maria, Valparaíso, Chile.

9 **3** CIMFAV, Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso, Chile.

10 **4** INRIA Biovision team Sophia Antipolis and Universite Cote d’Azur, Sophia-Antipolis, France.

11 **5** Instituto de Sistemas Complejos de Valparaiso, Chile.

12 Current Address: Centro Interdisciplinario de Neurociencia, Universidad de Valparaiso, Pasaje
13 Harrington 287, Playa Ancha, ZIP 2360102, Valparaiso, Chile.

14 * rubenherzog@ug.uchile.cl

15 **Abstract**

16 Maximum entropy models (MEM) have been widely used in the last 10 years to characterize the
17 statistics of networks of spiking neurons. A major drawback of this approach is that the number
18 of parameters used in the statistical model increases very fast with the network size, hindering its
19 interpretation and fast computation. Here, we present a novel framework of dimensionality reduction
20 for generalized MEM handling spatio-temporal correlations. This formalism is based on information
21 geometry where a MEM is a point on a large-dimensional manifold. We exploit the geometrical
22 properties of this manifold in order to find a projection on a lower dimensional space that best
23 captures the high-order statistics. This allows us to define a quantitative criterion that we call the
24 "degree of compressibility" of the neuronal code. A powerful aspect of this method is that *it does not*
25 *require fitting the model*. Indeed, the matrix defining the metric of the manifold is computed directly
26 via the data without parameters fitting. The method is first validated using synthetic data generated
27 by a known statistics. We then analyze a MEM having more parameters than the underlying data
28 statistics and show that our method detects the extra dimensions. We then test it on experimental
29 retinal data. We record retinal ganglion cells (RGC) spiking data using multi-electrode arrays (MEA)
30 under different visual stimuli: spontaneous activity, white noise stimulus, and natural scene. Using
31 our method, we report a dimensionality reduction up to 50% for retinal data. As we show, this is quite
32 a huge reduction compared to a randomly generated spike train, suggesting that the neuronal code, in
33 these experiments, is highly compressible. This additionally shows that the dimensionality reduction
34 depends on the stimuli statistics, supporting the idea that sensory networks adapt to stimuli statistics
35 by modifying the level of redundancy.

36 **Author Summary**

37 Maximum entropy models (MEM) have been widely used to characterize the statistics of networks of
38 spiking neurons. However, as the network size increases, the number of model parameters increases
39 rapidly, hindering its interpretation and fast computation. Here, we propose a method to evaluate the
40 dimensionality reduction of MEM, based on the geometrical properties of the manifold best capturing
41 the network high-order statistics. Our method is validated with synthetic data using independent
42 or correlated neural responses. Importantly, we show that dimensionality reduction depends on the
43 stimuli statistics, supporting the idea that sensory networks adapt to stimuli statistics modifying the
44 level of redundancy.

45 Introduction

46 It is widely admitted that the spikes exchanged between neurons convey information encoded in a, yet
47 unknown, manner. Deciphering these hidden "neural codes" - there is no reason why neurons should
48 use a unique coding strategy - is a contemporary challenge. A natural approach consists of detecting
49 statistical regularities in the spike patterns ("words") produced by a population of neurons. This
50 problem is challenging since the number of possible patterns grows exponentially with the number
51 of neurons and the time window defining the words. A population of N neurons may produce $2^{N \times \tau}$
52 words of time length τ , but experimental recordings fortunately produce a quite smaller subset. In
53 practice, once the number of neurons increases beyond 20, the total number of possible words becomes
54 intractable. Yet, one has to extract, from the empirical statistics of the words displayed, a canonical
55 model predicting the observed statistics, paving a possible way toward decoding. A theoretical frame-
56 work to tackle this issue is based on the Maximum Entropy Principle (MEP) [42, 46]. The objective
57 is to find the least structured model - the one with maximum entropy - given constraints provided by
58 the average values of certain "features" or observables.

59 In its simplest form MEP restricts to spikes occurring at the same time. For example, the Ising
60 model is constrained by the occurrence of spikes emitted at the same time by 2 distinct neurons
61 (Model with "pairwise interactions") [2, 3]. Most extensions with more neurons (triplets-quadruplets)
62 restrict as well to spikes occurring at the same time [8, 9, 10]. The mathematical consequence is that,
63 in these MEM, spikes events occurring at different times are independent. The probability of spike
64 patterns occurrence at time t does not depend on the history. We call these models "MEMs without
65 memory". However, neurons interact together with some delay, inducing spatio-temporal correlations
66 which should be considered as well as constraints for the MEM [8, 9, 10]. The MEP can be extended
67 to handle these spatio-temporal correlations [12, 16] capturing and predicting the collective spatio-
68 temporal pattern activity. This approach has been successfully used to characterize the spike train
69 statistics in cortex cultures [11] and in the vertebrate retina network [33, 14].

70 A strong caveat of MEM is their number of parameters increasing rapidly with the size of the
71 system. But one expects some redundancy in the spike activity of biological neuronal networks (in
72 contrast e.g. with artificial neural networks with random uncorrelated synapses), reflecting a latent
73 structure on the statistics of the activity. From this point of view, the dimensionality of the MEM
74 (the number of its parameters) could be reduced. There might be some analogy with signal processing
75 compression, where the presence of statistical redundancy can be exploited to map the signal onto
76 a subset of independent channels (i.e. non-redundant). As shown in [23, 13], the reduction is more
77 pronounced if the parameters are related by hidden mathematical dependencies.

78 Several tools are available to reduce the dimensionality of a given statistical model, based on the
79 trade-off between dimensionality and likelihood, e.g Akaike Information Criterion (AIC), Bayesian
80 Information Criterion (BIC) or Minimum Description length (MDL) among others. However, these
81 methods necessarily need to fit different models to the same dataset to find the optimal dimension.
82 This can be prohibitive in the case of MEM when considering a large number of neurons. In contrast,
83 a formal framework to find the optimal dimensionality of MEM based on the geometrical properties of
84 the manifold of probability distributions -where a MEM is a point in large dimensional space, whose
85 coordinates are the parameters- was developed in [41]. It generalizes over standard methods (AIC and
86 BIC), taking into account simultaneously the dimensionality, the likelihood and the geometry of the
87 manifold formed by the family of statistical models. Here, we propose to exploit the geometric structure
88 of the manifold formed by the MEM's to find an optimal set of dimensions capturing the information
89 contained on neural spiking data. Our approach generalizes previous approaches considering only i.i.d
90 samples [41], and is general enough to consider spatio-temporal interactions at several lags between
91 neurons. It relies on the spectral properties of a matrix, called "Fisher metric" in statistics and
92 information geometry [24] and "susceptibility Matrix" in statistical physics, capturing the effect of
93 MEM parameters variations on the second-order statistics of the network activity. Remarkably, as we
94 show here this matrix can be numerically computed from data without fitting the MEM parameters,
95 allowing us to apply the method to a large set of neurons (~ 100 neurons). Based on the spectral
96 properties of this matrix, which reflects the local geometry of manifold, we determine the optimal
97 model dimensionality that still preserves the correlation structure in neural activity.

98 The method is first validated using synthetic data generated with a known underlying statistics.
99 It is then applied to the spiking response in a neural population of retinal ganglion cells (RGC) *in*
100 *vitro*, recorded from a diurnal rodent *Octodon degus* [26] using a 252-MEA (multi-electrode array).
101 We consider three sets of neural responses obtained after applying three type of visual stimuli: i)
102 spontaneous photopic activity; ii) white-noise checkerboard and iii) a short natural movie. From the
103 mathematical analysis and from an analogy with signal compression, we define a "compressibility"
104 criterion for the MEM. Based on our experimental observations, we found that RGC activity is highly
105 compressible ($\sim 50\%$ of the imposed model dimensionality) and that the stimuli spatio-temporal
106 modulation increases the number of independent dimensions required to optimally represent the neural
107 activity. This suggests that RGC population activity adapts to stimuli conditions, changing the
108 number of "coding channels" according to stimuli correlations.

109 Methods

110 Spike trains

111 We consider the spiking activity data from a population of N interacting neurons. This data usually
112 comes from experimental recordings using multi-electrode-arrays in the retina or cortical areas. We
113 assume that there is a minimal time interval such that any neuron fires at most one spike within this
114 interval. This provides a time discretization usually referred as "binning". The *spike-state* $\omega_i(n)$,
115 takes the value 1 whenever the i -th neuron emits a spike at the time bin n , otherwise is zero. The
116 *spiking pattern* $\omega(n) := [\omega_k(n)]_{k=1}^N$ is the spike-state of the entire network at time n . The *spike block*
117 $\omega_{t_1}^{t_2} = \{\omega(n)_{t_1 \leq n \leq t_2}\}$ represents the activity of the whole network between time bins t_1 and t_2 . The
118 *length* of a spike block is the number of time steps $t_2 - t_1 + 1$. Experimental data consist of a *spike-block*
119 of finite size T , denoted by ω_1^T . In this paper we also consider infinite spike sequences $\omega_0^{+\infty}$, denoted
120 by ω to alleviate notations. The state space of an infinite sequence of spiking patterns is denoted by
121 Ω .

122 Observables and Monomials

We call *observable* a function $f : \Omega \rightarrow \mathbb{R}$, that associates a real number to a spike-train. We say that
 f has range R , if for every pair of spike trains $\omega, \omega' \in \Omega$ we have that $f(\omega) = f(\omega')$ if and only if
 $\omega_0^{R-1} = \omega_0'^{R-1}$, that is, f only depends on the first R spike patterns of the spike-train. An important
class of observables are the *monomials*, which are binary observables consisting of finite products of
spike states, given by:

$$m_l(\omega) = \prod_{k=1}^q \omega_{i_k}(t_k).$$

123 If one fixes a finite set of pairs $\{(i_k, t_k)\}_{k=1}^q = l$ (neuron index, time index), there are finitely many
124 such possible monomials, which can be indexed by an index l in one-to-one correspondence with the
125 set of pairs (i_k, t_k) . The observable $m_l(\omega) = 1$, if and only if neuron i_k spikes at time $t_k, \forall k \in \{1, \dots, q\}$
126 in the spike-train ω , where q is the number of spike states in the observable, and $m_l(\omega) = 0$ otherwise.
127 In a range $R \geq 1$ monomial, the firing times t_k are constrained within the interval $\{0, \dots, R - 1\}$.

128 Inference of the spike train statistics via Maximum Entropy principle

129 A generalized version of the maximum entropy approach can be framed rigorously using the ther-
130 modynamic formalism of subshifts of finite type [38], which offers a way to build maximum entropy
131 Markov chains (MEMC) from data in a principled way through a variational principle [16, 13]. As
132 we will see, framing the maximum entropy problem in this way is particularly useful, as set up the
133 conceptual framework to exploit ideas from thermodynamics and information geometry.

134 We assume that the spiking data is a sample of a time homogeneous Markov chain of memory R ,
135 i.e., taking values in \mathcal{A}_N^R , in other words, $\mathbb{P}(\omega_1^T)$ can be decomposed according to:

$$\mathbb{P}(\omega_1^T) = \mathbb{P}(\omega_{T-R+1}^T, \dots, \omega_3^{R+2}, \omega_2^{R+1}, \omega_1^R) = P(\omega_{T-R+1}^T | \omega_{T-R}^{T-1}) \dots P(\omega_3^{R+2} | \omega_2^{R+1}) P(\omega_2^{R+1} | \omega_1^R) p(\omega_1^R)$$

136 where $p(\omega_1^R)$ is the probability of the initial state and $P(\omega_i^{i+R-1} | \omega_{i-1}^{i+R-2})$ is the conditional probability
 137 or the transition matrix elements. The goal of the inference problem is to estimate the transition
 138 matrix P .

139 Entropy Rate

140 If the spike train ω is characterized by a stationary ergodic Markov measure denoted by $\mu(p, P)$ taking
 141 values in \mathcal{A}_N^R with homogeneous transition matrix P and unique stationary distribution p , the entropy
 142 rate of ω is referred as *Kolmogorov-Sinai entropy (KSE)* of $\mu(p, P)$ and takes the simple form [39]:

$$\mathcal{S}_{KS}(p, P) = - \sum_{i,j \in \mathcal{A}_N^R} p_i \sum_j P_{ij} \log P_{ij}. \quad (1)$$

143 where $P_{ij} = P(j|i)$. It is easy to see that when the stochastic process is i.i.d ($P_{ij} = p_j$) we recover the
 144 classical definition of Shannon entropy.

145 Variational Principle

146 We now introduce the maximum entropy principle (MEP) in the context of Markov chains (it extends
 147 to chain with infinite memory). In comparison to the standard statistical physics formulation of
 148 MEP, the present formalism extends to time-dependent interactions including with an infinite range
 149 (requiring then appropriate summability conditions [37]).

150 Given a set of observables (monomials) $m_l, l = 1 \dots L$, we denote $\nu[m_l]$ the expectation of m_l under
 151 the probability ν . We also denote c_l the empirical average of m_l (i.e. measured from an experimental
 152 raster). The MEP find the unique invariant probability μ satisfying $\mu[m_l] = c_l, l = 1 \dots L$ that
 153 maximize the KSE. This is equivalent to solve the following problem considering observables of range
 154 $R \geq 2$:

$$\sup_{\nu \in \mathcal{M}_{inv}} \left\{ \mathcal{S}_{KS}[\nu] : \nu[m_l] = c_l, \quad \forall l \in \{1, \dots, L\} \right\}. \quad (2)$$

155 where \mathcal{M}_{inv} is the set of translational invariant measures (stationary). Since the function $\nu \rightarrow \mathcal{S}_{KS}[\nu]$
 156 is strictly concave, there is a unique maximizing Markov measure $\mu(p, P)$ given the constraints c_l . This
 157 probability is uniquely defined by the *potential*:

$$\mathcal{H} = \sum_{l=1}^L h_l m_l, \quad (3)$$

158 This is a linear combination of the monomials m_l associated with constraints. The coefficients h_l
 159 are called the parameters of the model as there variation change the probability of events. They also
 160 formally correspond to interactions between neurons (e.g.; in the Ising model h_l are either the pairwise
 161 interaction J_{ij} between two neurons - monomials of type $\omega_i(0)\omega_j(0)$ - or the local external magnetic
 162 field - monomials of type $\omega_i(0)$).

163 Equation (2) is equivalent to the following unconstrained problem, which is a particular case of
 164 the so-called *variational principle* of the thermodynamic formalism [38]:

$$\mathcal{P}[\mathcal{H}] = \sup_{\nu \in \mathcal{M}_{inv}} \left\{ \mathcal{S}_{KS}[\nu] + \nu[\mathcal{H}] \right\} = \mathcal{S}_{KS}[\mu] + \mu[\mathcal{H}_\beta], \quad (4)$$

165
 166 where $\mathcal{P}[\mathcal{H}]$ is called the *free energy of \mathcal{H}* , and $\nu[\mathcal{H}] = \sum_{l=1}^L h_l \nu[m_l]$ is the average value of \mathcal{H} with
 167 respect to ν .

168

169 The average value of the observables, their correlations, as well as their higher cumulants can be
 170 obtained by taking the successive derivatives of the free energy with respect to the parameters h . This
 171 outlines the important role played by the free energy in this framework. In particular, taking the first
 172 derivative:

$$\frac{\partial \mathcal{P}[\mathcal{H}]}{\partial h_l} = \mu[m_l], \quad \forall l \in \{1, \dots, L\} \quad (5)$$

173 where $\mu[m_l]$ is the average of m_l with respect to μ . The second derivative of the free energy w.r.t a
 174 single parameter h_l gives the second cumulant or the variance of the observable m_l :

$$\frac{\partial^2 \mathcal{P}[\mathcal{H}]}{\partial h_l^2} = \mu[m_l^2] - \mu^2[m_l] \quad \forall l \in \{1, \dots, L\}. \quad (6)$$

176 Susceptibility matrix

177 Let ϕ be the shift map $\phi : \omega \rightarrow \omega$, defined by $\phi^i(\omega) = \omega_{i+1}^\infty$. Let m_l be an arbitrary observable. We
 178 may consider the sequence $\{m_l \circ \phi^i(\omega)\}$ as a random variable whose statistical properties depend on
 179 those of the process producing the samples of ω . For a pair $m_l, m_{l'}$ of monomials, the *time covariance*
 180 *of order r* of the stationary processes $\{m_l \circ \phi^n; n \geq 0\}$ and $\{m_{l'} \circ \phi^n; n \geq 0\}$ is defined as:

$$C_{l,l'}(r) := \int m_l \cdot m_{l'} \circ \phi^r d\mu - \int m_l d\mu \int m_{l'} d\mu \quad (7)$$

181 In particular, the auto-covariance of order r is:

$$C_l(r) := \int m_l \cdot m_l \circ \phi^r d\mu - \left(\int m_l d\mu \right)^2 \quad (8)$$

182 For a pair of finite range observables $m_l, m_{l'}$, the susceptibility can be obtained from the free energy
 183 as follows:

$$\chi_{ll'} = \frac{\partial^2 \mathcal{P}[\mathcal{H}]}{\partial h_l \partial h_{l'}} = \frac{\partial \mu[m_{l'}]}{\partial h_l} = \frac{\partial \mu[m_l]}{\partial h_{l'}} = \chi_{l'l} \quad (9)$$

184 It is also a standard result in ergodic theory (Green-Kubo formula) that the elements of the matrix χ
 185 can be obtained via time correlations:

$$\chi_{ll'} = C_{l,l'}(0) + \sum_{s=1}^{\infty} C_{l,l'}(s) + \sum_{s=1}^{\infty} C_{l',l}(s). \quad (10)$$

186 Especially, for the diagonal term:

$$\chi_{ll} = C_l(0) + 2 \sum_{s=1}^{\infty} C_l(s) \quad (11)$$

187 **Remark 1:** The matrix χ is at the core of our analysis and can be computed directly from data,
 188 assuming that μ is the empirical measure, that is without fitting the maximum entropy parameters.

189 **Remark 2:** In the case of memory independent MEM (e.g. Ising), and *only in this case*, (9) reduces
 190 to $\chi_{ll'} = C_{ll'}(0)$. In general, χ involves correlations at all times

192

193 Properties of the susceptibility matrix

194 Our method relies on the analysis of this matrix, which has the following properties:

- 195 (i) The set of all parameters fixes the statistical model. A slight variation δh_l of the sole weight h_l
 196 affects all the other monomials average. One can show that $\delta \mu[m_{l'}] = \sum_l \chi_{l'l} \delta h_l + O(\delta h^2)$. Thus,
 197 χ is a linear response matrix.

- 198 (ii) It is symmetric and positive, thus with real positive eigenvalues, which can be arranged increas-
 199 ingly $\lambda_1 \geq \lambda_2 \geq \dots \lambda_k \geq \lambda_L > 0$.
- 200 (iii) The set of $(h_l)_{l=1}^L$ is a vector of L dimensions, which can be seen as a point in $\mathcal{E} = \mathbb{R}^L$. The set
 201 of eigenvectors \mathbf{v}_k of χ constitute an orthogonal basis of this space, where the k -th eigenvector
 202 \vec{v}_k is the k -th direction in the space \mathcal{E} .
- 203 (iv) The closest two points are in \mathcal{E} , the closest are the statistics they predict. As a corollary, trying to
 204 fit empirical statistics, two models corresponding to “close” points might be indistinguishable as
 205 they describe equally well the empirical statistics. The notion of closeness is however ambiguous
 206 here and requires to define a proper metric. This is precisely what χ does: it defines a metric
 207 (called Fisher metric).
- 208 (v) From a statistical perspective, $\frac{1}{\sqrt{\lambda_k}}$ gives the amplitude of second-order fluctuations in the es-
 209 timation of coefficients h_l projected on direction \vec{v}_k . The smaller the eigenvalue, the larger the
 210 amplitude. Finally, from the linear response perspective, $\frac{1}{\sqrt{\lambda_k}}$ tells us how much a small variation
 211 in the estimation of average of monomials affects the estimation of h_l .

212 The Susceptibility matrix (or Fisher matrix) is used here in order to study *both* the geometric
 213 structure and the statistical fluctuations of the family of statistical models [24, 36, 23].

214 Distinguishable MEMC

215 Consider a spike train data set ω_1^T consisting of T spike patterns generated by a biological neuronal
 216 network. Given a set of L arbitrary observables $\{m_l\}_{l=1}^L$ (possibly non-synchronous) we compute their
 217 empirical averages from ω_1^T in order to set the constraints and infer the maximum entropy parameters
 218 $\mathbf{h} = (h_1, \dots, h_L)$ characterizing the MEMC. We may use these parameters to generate a sample $\tilde{\omega}_1^T$ of
 219 the inferred MEMC of same size T as the original data set. Considering the same set of observables
 220 we can apply again the MEP to infer a new set of parameters \mathbf{h}' from $\tilde{\omega}_1^T$, which is expected to be
 221 different from \mathbf{h} due to finite size sampling. Thus, the MEMC specified by $\mathcal{H}_{\mathbf{h}}$ and $\mathcal{H}_{\mathbf{h}'}$ cannot be
 222 distinguished on the basis of a dataset of finite length. However, increasing the sample size, one ex-
 223 pects the MEMC specified by the potential $\mathcal{H}_{\mathbf{h}'}$ to get “closer” to the one characterized by $\mathcal{H}_{\mathbf{h}}$. This
 224 idea can be rigorously formulated using large deviations techniques (see appendix).

225

226 **Definition** Consider two MEMC specified by $\mathcal{H}_{\mathbf{h}}$ and $\mathcal{H}_{\mathbf{h}'}$ within the same family of observables.
 227 Then, given a dataset of length T and empirical averages sampled from the model specified by $\mathcal{H}_{\mathbf{h}}$
 228 and tolerance $\epsilon > 0$, we say that the MEMC with parameters \mathbf{h} and \mathbf{h}' are ϵ -indistinguishable if:

$$-\ln \mathbb{P}(\mathbf{h} \approx \mathbf{h}') \leq \epsilon \quad (12)$$

229 The notation \approx stands for “inside a ball of radius δ , for some small $\delta > 0$, as explained in
 230 the supplementary material. This last property identifies an approximatively elliptical region of ϵ -
 231 indistinguishable models around each MEMC specified by \mathbf{h} , whose volume can be easily calculated
 232 in the large T limit [41, 36].

233 Volume of indistinguishable models

234 Following Balasubramanian [41] two distributions $\mu^{(1)}$, $\mu^{(2)}$ are indistinguishable with tolerance ϵ if
 235 $-\ln \mathbb{P}[h^{(1)} \approx h^{(2)}] \leq \epsilon$. If T is large enough, the set of indistinguishable distributions defines an
 236 ellipsoid with a volume denoted by \mathcal{V} called *confidence volume*. Points inside the confidence volume
 237 correspond to indistinguishable models within a tolerance ϵ . When the sample size T , the model
 238 dimension L (number of observables-parameters) and the tolerance $\epsilon = -\log \kappa$ are fixed the volume
 239 is:

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

240 Therefore $\log \mathcal{V} \propto -\frac{1}{2} \sum_{i=1}^L \log \lambda_i$. We observe that the model estimation is better if the eigenvalues
241 of χ are larger, which implies that there exists a set of eigenvalues that can be neglected given their
242 small magnitude, i.e. huge fluctuations impairing the model estimation. Then, instead of considering
243 the volume in the space of all parameters, let us consider the volume $\mathcal{V}(k)$ of the projection in the
244 subspace spanned by the k first eigenvectors of χ . We have:

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

245 with:

$$S_{\mathcal{H}}(k) = - \sum_{i=1}^k \log \lambda_i, \quad (15)$$

246 where the eigenvalues λ_i are ordered decreasingly. In eq. (14), the second term, $\log \Gamma(\frac{k}{2} + 1)$, depends
247 only of k . The third term, $\log(\frac{2\pi\kappa}{T})$ characterizes the effect of accuracy and finite sampling. Therefore,
248 the only term which depends on the statistical model (here characterized by the potential \mathcal{H}) is $S_{\mathcal{H}}(k)$.
249 In particular, it depends on the number of neurons N and the range R of the potential.

250 As stated above, the inverse of the eigenvalues tells us how much a small variation in the estimation
251 of the parameters affects the statistics. For a large eigenvalue λ_k , a tiny variation in the direction
252 v_k has a dramatic impact on the general statistics. On the opposite, small eigenvalues correspond to
253 sloppy dimension where even a big change in the corresponding direction has small impact. The notion
254 of stiff and sloppy dimension in statistics is not new and has been used by several authors, including
255 for the analysis of spike trains [25], but the treatment we propose, for MEM with spatio-temporal
256 interactions (in contrast to previous papers dealing with Ising model) is, to our best knowledge, a
257 novelty. Thus, using χ as a metric to distinguish between models, we can find the optimal number of
258 dimensions of a MEM given data, which is the main goal in this paper.

259 Dimensionality reduction

260 In information geometry, one extends a family of parametric probability distributions to a manifold
261 \mathcal{M} such that the points in \mathcal{M} are in a one to one relation with the probability distributions. The
262 parameters of the distributions can thus also be used as coordinates on \mathcal{M} .

263 For a fixed potential form (3), the MEMC parametrized by the coefficients \mathbf{h} , corresponds to a
264 point in the space \mathcal{E} . Equivalently, this point corresponds to a unique Markov measure. If the h_{ls} are
265 tuned independently from each others, the set of MEMC fully spans \mathcal{E} . However, when fitting data
266 from neuronal networks, either artificial or biological, one expects hidden relations between the h_{ls} ,
267 constrained by dynamics [13]. In this case, the MEMC spans a manifold \mathcal{M} in \mathcal{E} , of (presumably)
268 quite lower dimension.

269 To estimate the dimension of \mathcal{M} , i.e, the dimension of the manifold of \mathcal{E} sufficient to explain the
270 data with a minimal redundancy we use the following argument. Considering the equation (15) as
271 the eigenvalues λ_k increases, there will be one or more ks at which $S_{\mathcal{H}}(k)$ is expected to become
272 bigger than the sum of the other two terms of (14), which are both negative. This means that, for
273 increasing k , $\log \mathcal{V}(k)$ will first decrease then it will increase, yielding at least one minima on the
274 function. There is therefore a critical value of k , denoted by $k_c \equiv k_c(\epsilon, T)$, which characterizes the
275 optimal dimension for which the *volume is minimal*. The value of ϵ belongs to some interval; outside
276 this interval $\log \mathcal{V}(k)$ is convex or concave having only trivial minima. Inside this interval, we obtain a
277 set of $k_c \equiv k_c(T)$ minimizing the volume, which characterize the number of dimensions ensuring that
278 the model indeterminacy is minimized (see figure 1). Additionally, k_c provides a rough estimate of
279 the dimension of \mathcal{M} .

280 General Method Description and Applications to Synthetic Data

281 In order to test the method in a context where the ground truth is known, we artificially generated
282 spike trains from MEMs using different sets of observables:

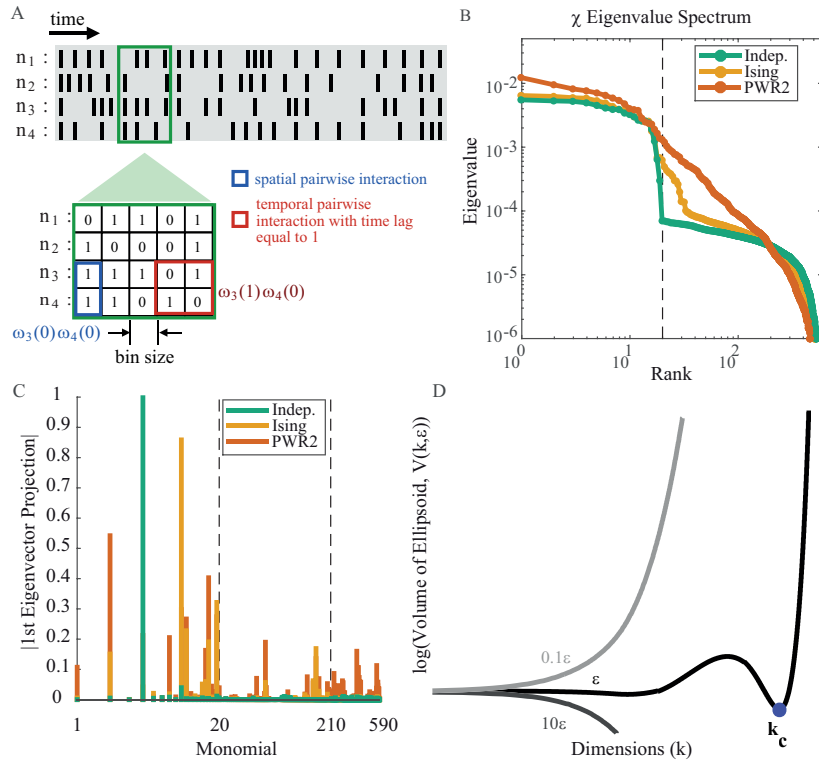


Figure 1: Dimensionality reduction framework overview. A: Raster representing the binary activity of $N = 4$ neurons (rows) over time (T). Green square shows a slice of this raster. For this raster 2 types of pairwise interactions are defined: spatial interactions (blue) and temporal interactions with $R = 2$, i.e. one time-step between spikes (red). Spatial interaction is exemplified as $\omega_3(0)\omega_4(0)$ (both neurons firing at the same time) and $\omega_3(1)\omega_4(0)$ (neuron 3 firing one bin after neuron 4). B: Susceptibility matrix eigenvalue spectrum in log-log scale for Independent (green), Ising (orange) and Pairwise Range=2 (PWR2, red). Black vertical dashed line denotes the network size ($N = 20$). Indep. raster has a sharp cut-off close to the network size; Ising has a cut-off few eigenvalues beyond the network size ($k_c = 29$); PWR2 shows a monotonic decay of the eigenvalues magnitude, without a clear cut-off. C: First eigenvector absolute projections for Independent (green), Ising (orange) and Pairwise Range=2 (PWR2, red) rasters. Left-most vertical dashed line is the limit between spike rates and spatial interactions monomials while the right-most is the limit between spatial and temporal interactions monomials. Indep. raster projects only on rates monomials, while Ising projects on rates and spatial interactions, but not on temporal ones. PWR2 projects on the tree types of monomials. D: Log of the volume of indistinguishable models, $\log \mathcal{V}(k, \epsilon)$ as the degrees of freedom (k) increases. The volume reaches a minimum (blue dot) for $k = k_c$. Increasing k beyond k_c makes the volume explode. Increasing/decreasing the error (ϵ) one order of magnitude yields to strictly increasing or decreasing functions, where volume minimization is trivial and uninformative.

- 283 1. Independent model: This model considers that neurons are independent, i.e. it only contains
 284 self-interactions (firing rates). The parameters of the potential control the firing rate of each
 285 neuron. A rate model with N neurons has therefore $L = N$ parameters. We call this model
 286 Indep. to alleviate notations. The potential for this model reads:

$$\mathcal{H}(\omega(0)) = \sum_{i=1}^N h_i \omega_i(0)$$

- 287 2. Ising model: This model considers firing rates and pairs of neurons firing at the same time
 288 (spatial). There is no interaction between spikes at different times. Considering N neurons, the
 289 model has $L = N + \frac{N(N-1)}{2}$ parameters. The potential for this model reads:

$$\mathcal{H}(\omega(0)) = \sum_{i=1}^N h_i \omega_i(0) + \sum_{i,j=1}^N J_{ij} \omega_i(0) \omega_j(0)$$

- 290 3. Spatio-temporal pairwise interactions model with memory depth 1: This model considers firing
 291 rates, pairs of neurons firing at the same time (spatial) and also pairs of neurons firing with one

292 time-step delay between them (temporal). We call this model PWR2. Considering N neurons,
293 this model has $L = N^2 + \frac{N(N-1)}{2}$ parameters (temporal self-interactions are not considered).
294 The potential for this model reads:

$$\mathcal{H}(\omega_0^1) = \sum_{i=1}^N h_i \omega_i(0) + \sum_{i,j=1}^N [J_{ij} \omega_i(0) \omega_j(0) + J'_{ij} \omega_i(0) \omega_j(1)]$$

295 Note that, in contrast to J_{ij} s, J'_{ij} are not symmetric.

296 4. Scaled PWR2: We multiply the potential \mathcal{H} by a factor (corresponding to an "inverse tempera-
297 ture"). In our experiments we use $\beta = \{0.4, 0.6, 0.8, 1.2, 1.4\}$.

298 Results

299 We apply our method to study the critical dimension (k_c) of synthetic spike trains of size ($N = 20$,
300 $T = 10^6$) generated from random potentials corresponding to the Independent, Ising and PWR2
301 models. We generate 100 different rasters of each using the same set of parameters for each MEM. For
302 each spike train, we compute the χ matrix *using the observables of the PWR2 model*, that is, we over
303 fit when data are generated by an Independent or an Ising and, then, obtain the respective eigenvalue
304 spectra.

305 We show in Fig 1B, the spectrum of χ . The entries of this matrix are estimated from data
306 considering three different cases. First data was generated by a PWR2 model (red curve); in the
307 second case, data comes from an Ising model (orange); in the last case data comes from an Independent
308 model (green). In all cases, the dimension of χ is the same, but, in the independent and Ising case,
309 we are overfitting the estimation, as for the independent model firing rates are enough to fit the data,
310 and for the Ising model rates and pairwise interactions are enough.

311 The difference in the three cases is clearly seen in the spectrum of χ . Moving along the spectrum
312 from left to right (increasing index k , decreasing the magnitude of the eigenvalue λ_k) we observe a
313 first sharp decrease (*cut-off*) at $k = N$, for the Indep. (Fig 1B, black) and Ising rasters.

314 In addition to the differences on the eigenvalue spectrum, the functional relationships between the
315 monomials change depending on the underlying statistics, as exemplified by the first eigenvector of
316 the corresponding χ matrix for each raster (Fig 1C). The Independent model shows large projections
317 only on the monomials related to spike rates, while showing negligible projections on the pairwise
318 monomials. The Ising model shows large projections as well on the spike rates monomials but in
319 addition shows some projections on the spatial interactions monomials. Finally, the PWR2 raster has
320 projections on all the monomials, reflecting in part its underlying statistics. Thus, the differences on
321 the energy function are captured both by the eigenvalues spectrum and also by the structure of the
322 corresponding eigenvectors.

323 In order to illustrate the volume of indistinguishable MEM, $\mathcal{V}(k, \epsilon)$, we computed it at different ϵ
324 and k finding some functions with non trivial minima, suggesting a reducible MEM (i.e. $k_c < L$). The
325 presence of a cut-off on the eigenvalue spectrum shows that there is a value of k where the confidence
326 volume is minimal (Fig 1D), i.e., where the model is more accurately determined. This point is highly
327 non trivial and captures somewhat the role of spatio-temporal interactions on second order statistics¹.

328 Finally, k_c is conditioned to the tolerance value ϵ as shown on Fig 1D. Increasing or decreasing ϵ one
329 order of magnitude yields functions with trivial minima. This constrains the search of the minima to a
330 subset of ϵ values. Extending the results for different values of ϵ shows that the number of dimensions
331 related to the minimal volume decays monotonically with the accuracy ($\kappa = -\log \epsilon$) until it reaches
332 an inflection point; this inflection point is the k_c that we use as the number of relevant dimensions
333 (Fig. 2).

¹Christoffel coefficients can be computed from the metric χ , giving thus access to the local information geometry and higher order statistics [24]

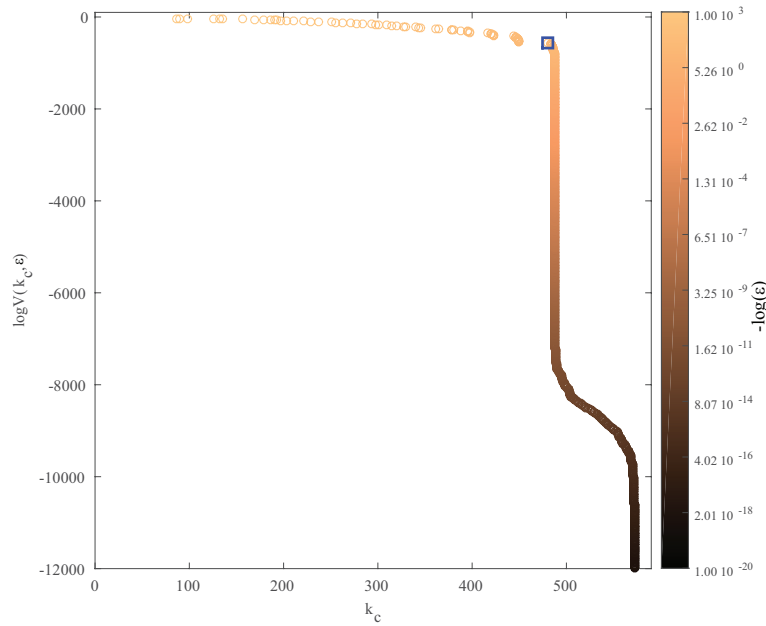


Figure 2: **Finding k_c on the set of convex volume functions.** The volume at the minimum and its corresponding k_c and κ ($-\log \epsilon$) values. There are many convex functions and corresponding k_c that we could use, but given that we are looking for a cut-off, the actual k_c used is the last k_c before the inflection point (blue square) on the $\log \mathcal{V}(k_c, \epsilon)$ vs k_c curve, representing a trade-off between maximal number of dimensions, the highest accuracy and the minimal volume as possible. This is the method used to choose the k_c for all the data analyzed.

334 k_c values on the independent and spatio-temporal correlated cases

335 Using the synthetic data, we evaluated how the optimal dimension given by k_c depends on the un-
 336 derlying statistics of neural data. Considering the k biggest eigenvalues, we computed the log of the
 337 volume $\mathcal{V}(k, \epsilon)$ (14) and compute its minimal value obtaining k_c (Fig 3A). The values of the parame-
 338 ters of the underlying MEM related to firing rates and to pairwise interactions were randomly chosen
 339 from a normal distribution with mean -5 and -1, respectively, and 1 standard deviation. We took one
 340 set of parameter for a PWR2 MEM and to obtain the Independent and Ising rasters, we kept only
 341 the parameters related to firing rates in the former and the ones relates to firing rates and spatial
 342 interaction in the latter. For this example we observe the following:

343 In the case of independent statistics, the value of k_c is closely related to the network size ($k_c =$
 344 19 ± 0). In this example, one neuron has a very low firing rate ($\sim 10^{-5}$) compared to the others
 345 ($> 10^{-4}$). This explain the sharp cut-off at this neuron value. On the opposite, in the case of PWR2
 346 the number of optimal dimension is $k_c = 447 \pm 7.00$ (see Fig 3B). The intermediate case, Ising, shows
 347 few number of dimensions more than N , but much less than $L(590)$.

348 Interestingly, even when the number of dimension for the PWR2 and Ising is larger than the
 349 Independent case, the percentages of the optimal number of dimension relative to the total number
 350 of dimensions is smaller for them, compared to Indep. as shown in Fig 3C. This shows that our
 351 method can detect how many dimensions are needed for a MEM to fit data respect to underlying
 352 raster statistics.

353 For the Indep. case we found that k_c corresponds almost to the full dimensionality of the underlying
 354 model ($L = 20$), while Ising can be highly reduced respect to its full dimensionality. For PWR2
 355 k_c corresponds to approximately 75% of the full underlying model dimensionality (Fig 3CB, inset),
 356 possibly related to finite size sample effects (events that are unlikely to be observed in finite time). So,
 357 for these examples, the effective dimensionality (k_c) found by this dimensionality framework increases
 358 with the number of terms of the underlying statistics defined by the energy function.

359 k_c depending on the the network size and recording length

360 We evaluated the impact of the number of neurons N and the raster duration T on the estimation of
 361 k_c . We focus only on the Independent and PWR2 case, leaving Ising outside this analysis.

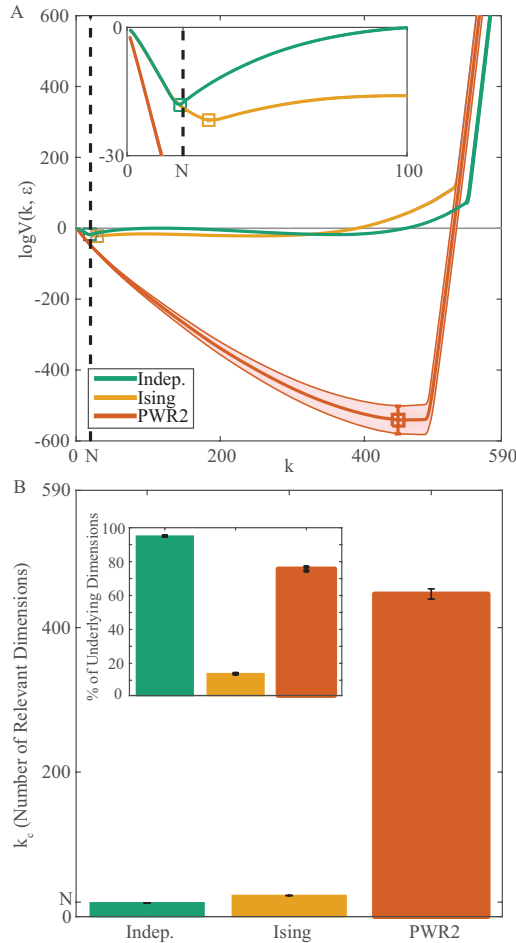


Figure 3: Dimensionality reduction by minimization of the volume of indistinguishable models. A: log of volume of indistinguishable models ($\log \mathcal{V}$) as function of the number of dimensions (k) at the maximal accuracy that yields no trivial minima ($\kappa = 1336.6 \pm 16.00$ for Indep., 1250.6 ± 101.86 for Ising and $108.1 \pm$ for PWR2). Thick solid lines are averages, shaded area is ± 1 s.d. of the 100 samples 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. and Ising. minima. B: Summary of the number of dimensions for both statistical models. We found a close relationship between the underlying model dimensionality and the number of dimensions required for an optimal model. In this example, k_c is 19 ± 0 for Indep., 29.16 ± 0.37 , and 447 ± 7.00 (mean \pm std) for Indep. and PWR2, respectively. Inset: k_c as a percentage of the underlying model dimensionality. Indep case show almost 100%, while Ising has $\pm 14\%$ of reduction and PWR2 $\pm 75\%$.

362 Fixing the neural population size ($N = 20$), we first varied the recording length from $T = 10^2$
 363 up to $T = 2 \cdot 10^6$ bins and observed the effect obtained (see Fig 4A) on the χ spectrum for the
 364 Independent (left) and PWR2 (right) cases. In the independent case, the first cut-off of the spectrum
 365 is increased while the value of T increases. On the opposite, in the case of PWR2 the cut-off is
 366 not observed, indeed, the separation in the spectrum of dimensions related to neuron firing rates or
 367 combined activation is less evident than the Indep. case. Additionally, we obtained the k_c value for
 368 each value of T as shown in Fig 4B. In both cases, we can consider that the estimation of k_c converges
 369 for values of T over 10^6 bins.

370 Fixing now the value of $T = 10^6$, we evaluated the effect of the network size N on the estimation
 371 of k_c for the Independent and PWR2 case. As shown in Fig 4C, increasing the number of neurons
 372 increases the model dimensionality as well as k_c . Remarkably, the shape of the spectrum remains the
 373 same, suggesting finite size scaling [43]. Finally, Fig 4D shows how the maximal dimensionality of the
 374 model L and the value of k_c depends on the network size N .

375 For small neural populations sizes, $k_c \sim L$, but as N increases, the number of possible interactions
 376 (L) grows, requiring longer recording lengths. Black circle denotes the network size used in retinal
 377 spike train data shown in this article ($N = 50$).

378 The fact that a larger number of neurons N requires longer recording lengths brings as consequence
 379 a value of k_c departing from L . As either the simulated or real data has a finite length T , increasing

380 N generate an increasing number of unobserved monomials, or, monomials with very low occurrence
 381 probability, reducing the effective χ matrix rank. The effect of unobserved or very low empirical
 382 probability events on χ and, consequently, on k_c estimation is detailed in the next subsection.

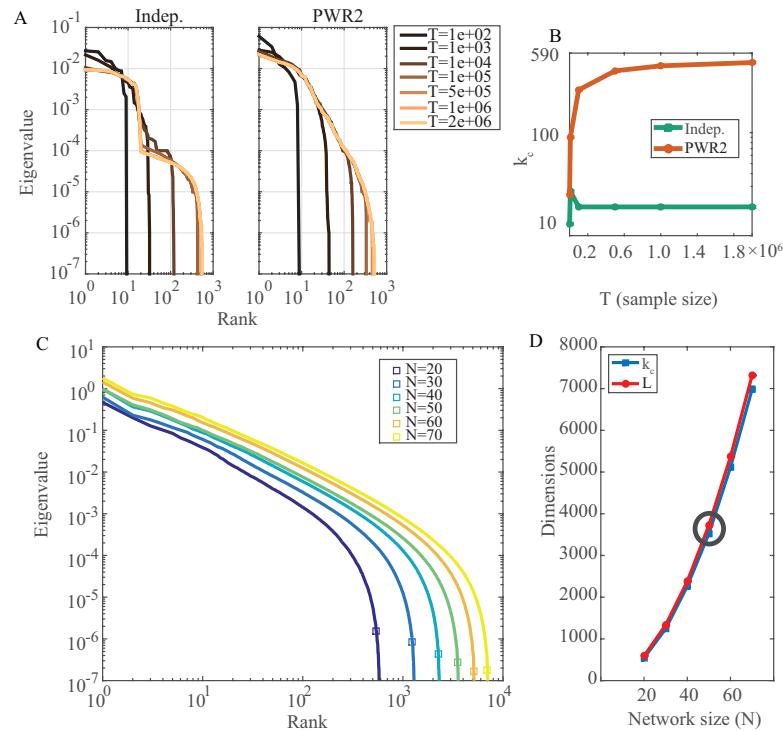


Figure 4: **Method dependence on recording length and network size.** A: Spectra of Indep. and PWR2 rasters using different recording lengths (colors) for $N = 20$. B: k_c values for both rasters at different recording length. Using $T \geq 10^6$ is enough to have good k_c estimations. C: Spectra for random PWR2 rasters with increasing network size (N , colors) with the corresponding k_c on it (squares). The length of the spectrum (number of eigenvalues) changes because of the increase of model parameters, but the shape of the spectrum remains the same given the constant underlying model. D: k_c value for each network size (blue) and corresponding maximal dimensionality (L , red). Black circle denotes the raster size that we use in biological recordings ($N = 50$). $k_c \sim L$ at almost all network sizes, but as N increases, the number of possible interactions grows, requiring longer recording lengths, making k_c diverge from L .

383 A measure of "code compressibility"

384 We now relate k_c to a notion of "neural code compressibility". We start from an important remark:
 385 as a sum of covariance matrices, χ is non negative. Nevertheless, it can have many zero eigenvalues
 386 for two distinct reasons:

- 387 1. There are hidden linear dependencies between the coefficients h_l . This is typically due to an
 388 hidden structure in the dynamics which has generated the data. An explicit example is provided
 389 in [13] where the h_l of a neuronal network model are computed as a function of the W_{ij} s (synaptic
 390 weights) and stimulus. These eigenvalues are intrinsic to the dynamics and constitute somewhat
 391 the redundant part of the information that we want to remove to "explain" data. We note d_N
 392 the number of these eigenvalues.
- 393 2. χ is computed from finite rasters. Here, some monomials have zero empirical value when the
 394 corresponding event do not appear in the empirical raster. If m_l is one of these unobserved
 395 monomials we have $\pi(m_l) = 0$ and, $\forall l', \pi(m_l m_{l'}) = 0$, where π is empirical probability. We call
 396 this type of events Unobserved Events (U_E). As a consequence, $\chi_{ll'} = 0, \forall l'$ and the row l of χ
 397 has zero entries. Consequently, this row generates, in the spectrum of χ , a zero eigenvalue.

398 We have therefore $\mathcal{R} + d_N + U_E = L$ where \mathcal{R} is the dimension of the image of χ . In general $k_c < \mathcal{R}$
 399 because, after the cut-off there are eigenvalues, small, but nevertheless positive. If the cut-off is sharp,
 400 as we observe, \mathcal{R} is very close to k_c though. On this basis we define the "compressibility" of the code

401 as:

$$\mathcal{C} = L - k_c - U_E \quad (16)$$

402 For sharp cut-off it is very close to d_N which precisely reflects the irrelevant dimensions in the space
403 of parameters, corresponding to hidden dependencies.

404 **How do k_c and \mathcal{C} depend on the raster density.**

405 Using the aforementioned approach to dissect the relevant from the irrelevant dimensions, we aim to
406 study the dependency between the raster density and code compressibility given a spatio-temporal
407 MEM imposed to data. We consider neural density (or, conversely, sparsity) as the number of spikes
408 of a raster divided by the number of bins of the raster. This density depends on many factors
409 as the excitatory-inhibitory balance, network connectivity and stimuli, among others. Very dense
410 neural responses could produce artificial neural correlations beyond the underlying statistics. On the
411 opposite, very sparse responses could not provide enough information to fairly recover the correct
412 empirical statistics of complex events (pairwise with time delays). Having this in mind we generated
413 synthetic data with different levels of density and estimated k_c under these different regimes.

414 We generated different synthetic sets of neural data where the density of the response varies. To
415 do this, starting from PWR2 model, we generated five new responses scaling the parameters of the
416 MEM by a factor $\beta = \{0.4, 0.6, 0.8, 1.2, 1.4\}^2$

417 The effect of increasing the parameter β has therefore a tendency to spread the distribution of
418 parameters h_l as shown in Fig 5A enlarging its variance, where the parameters h_l associated to firing
419 rates are represented in blue, and those associated to spatio-temporal interactions in red. As expected,
420 scaling the model parameters by low values of β condense the parameters distribution.

421 Similarly, monomial probabilities are also affected by the β parameter (see Fig 5B). As expected,
422 decreasing the value of $\beta < 1$ tends to equalize the probabilities of the monomials generating a very
423 dense neural response. On the contrary, higher values of $\beta > 1$ tend to maximize the energy generating
424 configurations of the response with less variability (sparse distribution). High values of β diminishes
425 therefore the probability of occurrence of those monomials whose h_l value is negative while it increases
426 the probability of monomials (spike events) with a positive h_l .

427 Increasing β value causes two main effects on the eigenvalue spectrum of the χ matrix (see Fig 5C):
428 (i) the spectrum is flatter as β decreases, (ii) there are more zero eigenvalues, i.e. U_E increases. Both
429 effects come from the fact that the raster sparsity decreases with β , decreasing the events probabilities
430 as well. Thus, this low probability events are reflected on the increase of U_E due to finite sampling
431 and also on the flatness of the spectrum, given that the few observed events have similar probabilities.

432 Decreasing the value of β below 1 decreases the value of k_c . In a similar manner, values of β larger
433 than 1 monotonically decrease the value of k_c . Then, moving away from $\beta = 1$ reduces k_c , but the
434 reasons for the first case are different from the second. To illustrate this, Fig 5D shows the mean
435 k_c value obtained for 10 different rasters (sharing the same underlying parameters and β) as a circle
436 and 1 s.d as error bars. As we previously mentioned, small values of β produces a very dense neural
437 response, that apparently can be condensed in a few dimensions ($k_c = N$ for $\beta = 0.4$). On the second
438 case, for $\beta > 1$ many events have very low probability, increasing U_E , which will decrease k_c . This
439 difference is illustrated on figure 5E), where larger values of β generates responses with large U_E .

440 For values of $\beta < 1$ we see almost no difference between Fig 5D and Fig 5E, meaning that k_c is an
441 indicator of data compression not affected by the unobserved events (that are negligible). Nevertheless,
442 the data compression detected by a low value of k_c is artificial because most of the correlations are
443 induced by the effect of a high activation in the neural population. On the opposite, in the cases
444 where $\beta > 1$ the unobserved events become significant and the low value of k_c obtained in Fig 5D was
445 mainly due to the number of non-zero dimensions of the χ matrix without compression.

²Increasing β , known in physics as inverse temperature, amounts to favor configurations maximizing the energy, i.e., monomials with high positive terms. On the opposite, decreasing β (increasing temperature) tends to equalize probabilities of patterns. Note that, in contrast to statistical physics, energy is maximized, not minimized, because we don't have the minus sign in front of the potential, in the Gibbs distribution.

446 Thus, from this approach, the full dimensionality of the model imposed to data can be dissected
 447 on the effective dimensionality, k_c , and the compressibility, \mathcal{C} , reporting the level of redundancy that
 448 the model can capture from data.

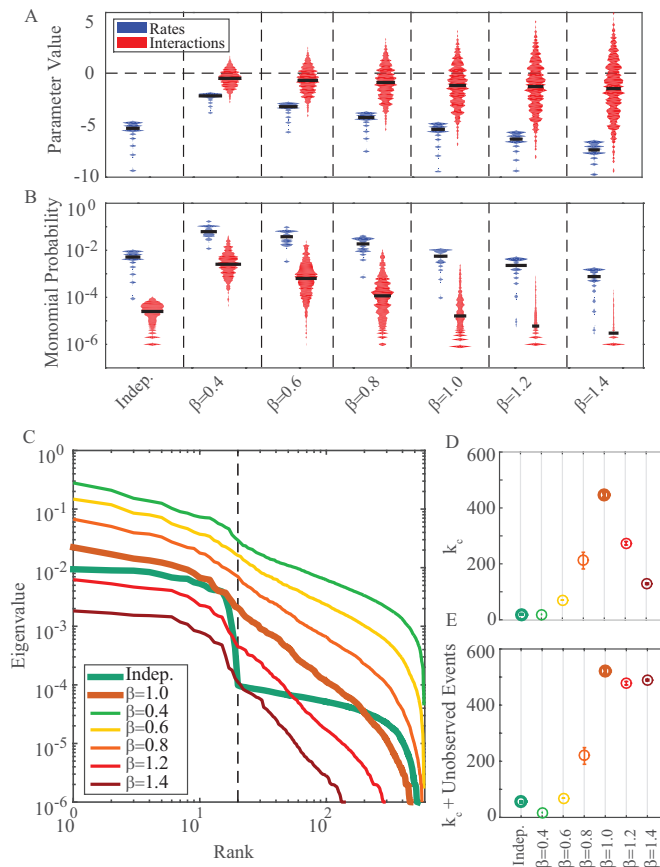


Figure 5: **Dimensionality reduction on scaled PWR2 statistics.** A: Underlying parameters distribution, split by firing rates (blue) and pairwise interactions (red). The bigger the scaling factor β , the more negative the rates parameters and the wider the interactions parameters distribution. $\beta = 1$ is the original PWR2 raster. The first column show the Independent raster as a reference. B: Corresponding monomials probabilities for the scaled rasters, split between firing rates and interactions, as in A. We see that increasing β has the effect of decreasing both rates and pairwise interactions probabilities, reaching the point where many pairwise interactions vanish ($\beta > 1$). C: Corresponding χ eigenvalue spectra (average out of 10 rasters for the scaled rasters). Increasing β flattens the spectrum, 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. Dots represent the average of k_c over 10 different rasters with the same underlying parameters. Error bars represent 1 s.d. E: Average k_c values plus the number of unobserved events (U_E). For $\beta > 1$ 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 $\beta < 1$ we have less unobserved events, so the dimensionality reduction obtained for those cases is given mainly by the increased density of the raster.

449 Dimensionality reduction on retina data

450 We are now interested in verifying if the properties found in synthetic data are scalable to real neural
 451 data. We did it on retina data obtained for three different conditions: photopic spontaneous activity
 452 (PSA), white noise (WN) and natural movie (NM).

453 Retina data was recorded in a 252-MEA system obtaining the response of 867 retinal ganglion cells
 454 in four different pieces of *O. degus* retina. Additionally, we generated shuffled version of the real data
 455 to compare the value of k_c and \mathcal{C} in the two cases, where the shuffled version maintained the firing
 456 rates of each neuron destroying the spatio-temporal correlations. In Fig 6 we show the distribution
 457 of the empirical monomials present in real versus shuffled data. Monomials related to firing rates
 458 are shown in blue and those related to spatio-temporal correlations are shown in red. Interestingly,
 459 distribution of the pairwise interactions probabilities of the empirical recordings and their shuffled
 460 version for all stimuli are not significantly different (bin 10ms, Mann-Whitney test $P > 0.05$), so we

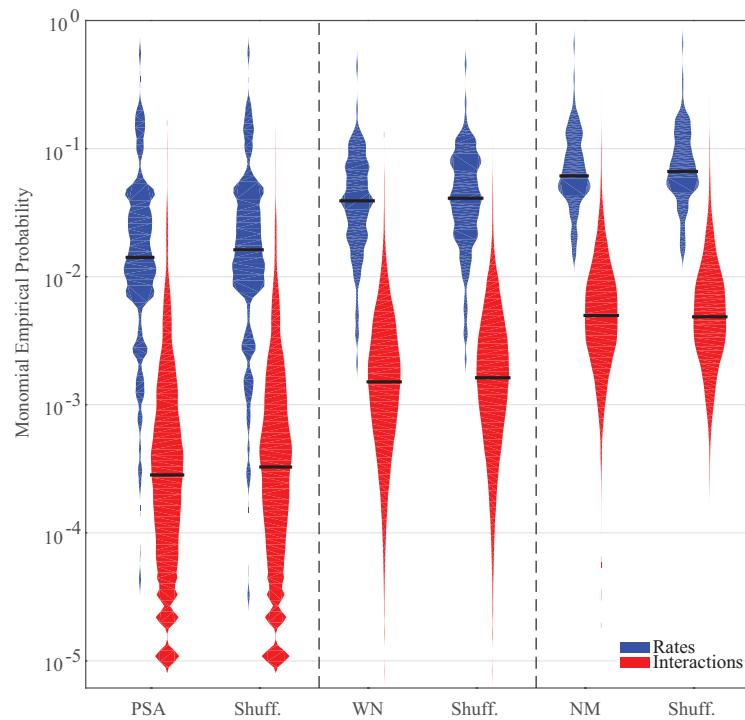


Figure 6: **Distribution of firing rates and pairwise interactions.** Violin plot of the firing rates (blue) and pairwise interactions (red) are shown for the 3 stimuli used and their corresponding shuffled version using a bin of 10ms (black horizontal line is the median). Both type of monomials increases with the stimuli high-order correlation. Note that the shuffled version reproduces the distribution of the pairwise interactions observed probabilities, even when the temporal structure of the raster is destroyed. This doesn't imply that the linear dependencies between MEM parameters are kept.

461 don't expect big differences on monomials empirical probabilities, but we do expect differences on the
 462 linear dependencies between them. Also from Fig 6 we observe that the highest activity, both for
 463 firing rates and spatio-temporal interactions, is obtained for NM, followed by WN and then by PSA.
 464 Thus, we modify both the raster density and the hidden dependencies of the neural activity by means
 465 of stimuli, where, in this case, the raster density increases with the stimuli high-order correlations.

466 **Experimental versus shuffled spectrum**

467 We computed the χ eigenvalue spectrum of 30 random sub-networks (i.e. sub-samples of the entire
 468 population) of $N = 50$ from the total number of neurons recorded in each of the four experiments,
 469 under the 3 stimuli conditions. According to our synthetic rasters experiments, for $N = 50$ and
 470 $T \sim 10^6$ we can get reliable k_c estimates (Fig 4D), which fits with our experimental recordings. Same
 471 procedure was applied to shuffled data.

472 Similar to what we obtained for scaled synthetic rasters (Fig 5C), we see that the spectrum is
 473 flatter, and the vanishing eigenvalues (below $1/T$) of χ spectrum increases with the raster density,
 474 which is driven by the stimuli (Fig 7A). The size of the network ($N = 50$) is represented by a vertical
 475 dashed line. In real data, only WN condition shows a cut-off in the spectrum close to N . Interestingly,
 476 in shuffled data both NM and WN present this cut-off near N , suggesting that in the experimental
 477 recordings there are significant linear dependencies between monomials that are not present in the
 478 shuffled version (see Fig 7B). Specially, NM shows a smooth decay of the spectrum, similar to the
 479 observed for PWR2, which is highly modified when the raster is shuffled, having a sharp cut-off close
 480 to N . So, even when the distribution of the probabilities of firing rates and pairwise interaction remain
 481 the same after the shuffling procedure (Fig 6), the linear dependencies between them are modified,
 482 changing the shape of the spectrum, showing sharper cut-offs for all conditions.

483 **Computing k_c for retinal ganglion cells spike trains**

484 We computed k_c for spike trains of retinal ganglion cells obtained for different stimulus. For each
 485 stimulus we generate different spike trains (experimental and shuffled) and we use different time bins

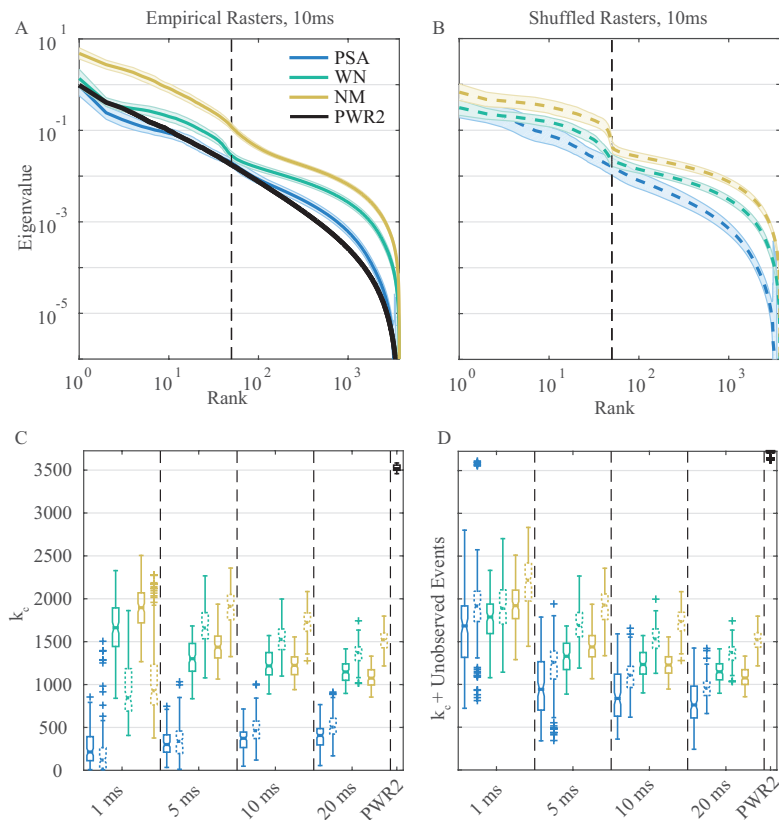


Figure 7: Dimensionality reduction on RGC data. A: Average spectrum (30 random sub networks plus 1 s.d., solid line and shaded area, respectively) of RGC data ($N = 50$) under 3 different stimuli conditions, with 10ms bin size. Stimuli high-order statistics increases the spectra offset (the magnitude of the eigenvalues). Except for WN, there is no clear cut-off close to N . Black line is a random PWR2 raster of the same network size than the RGC raster plot. B: Same than A, but for the shuffled version of the empirical rasters in dashed lines. All of them show a clear cut-off close to N and, preserving the effects induced by stimuli high-order correlations, i.e. change on the first eigenvalue and offset. C: Box plots for k_c values for empirical (solid) and shuffled rasters (dashed). Given the absence of prior knowledge about the relevant timescales the brain uses to integrate retinal signals, we studied several bin sizes. For example, for 1ms, k_c is higher for shuffled data in all bin sizes, under all conditions. Stimuli high-order correlations 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. 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. D: Same as C, but corrected by the number of unobserved events. Now for all bin sizes the shuffled data has bigger values than the empirical one. Shuffled data has values almost a half than a PWR2 raster of the same size, regardless of the absence of linear dependences by construction. The effect of stimuli high-order correlations on the effective dimensionality remains.

486 to binarize the data: 1, 5, 10 and 20ms. As a global picture, the value of k_c in the real and shuffled
 487 data increases as the activity of the network does (Fig 7C). This behavior is replicated for all the bin
 488 sizes.

489 The k_c analysis on retinal data reveals that the RGC activity is not random, showing almost the
 490 half of dimensionality compared to a random PWR2 with the same network size, which is shown in
 491 black in the upper right corner of Fig 7C. Furthermore, even if the firing rates and spatio-temporal
 492 events have the same distribution between real and shuffled data, the values of k_c obtained in the two
 493 cases differ, the real value being always smaller than the shuffled version. For all the time scales (from
 494 1 up to 20 ms) we observe that the value of k_c in the shuffled data is always higher than the one
 495 obtained for real data. Additionally, for the shuffled data the value of k_c monotonically increases with
 496 the activity. Similarly, the real data also presents a tendency to increase the value of k_c as the raster
 497 density does; nevertheless the relation is inverted between WN and NM for bin sizes 10 and 20ms: in
 498 these two cases the k_c value obtained for NM is smaller than the values obtained for WN.

499 In order to verify if the k_c values are actually associated to linear dependencies between neurons,
 500 and not to unobserved event, we did the correction $k_c + U_E$ as it was shown in Fig 7D. PSA has
 501 the larger U_E values, showing a big difference between k_c and $k_c + U_E$. In general, k_c corrected by
 502 U_E decreases with the bin size, because the larger the bin, the more likely is to observe two spikes

503 on the same bin and less likely is to have unobserved events. As a consequence, this increases the
504 apparent interdependence of RGC population activity (both for real and shuffled data). Specifically,
505 we see that k_c increases with the stimuli high-order correlation for fast time scales (Mann-Whitney
506 test, $P < 10^{-5}$ for all comparisons). This suggests that for fast time scales the retina increases the
507 number of coding channels as the stimuli high-order correlation increases. However, for medium time
508 scales the picture changes, showing no significant differences between WN and NM for 10ms (Mann-
509 Whitney test, $P > 0.1$) and showing higher k_c values for WN than NM for 20ms (Mann-Whitney
510 test, $P < 10^{-5}$). This suggests that at larger time scales the RGC activity under NM becomes more
511 interdependent than for WN, which could be related to the increased level of redundancy of NM
512 compared to WN and the time scale that this redundancy is captured by the retina generating a
513 redundant activity.

514 On the other hand, the shuffled data show higher k_c values for almost all the bin sizes and conditions
515 (Mann-Whitney test, $P < 10^{-4}$), except for 1ms where it is significantly smaller (Mann-Whitney test,
516 $P < 10^{-4}$) and for PSA at 5ms, where they are not significantly different (Mann-Whitney test,
517 $P > 0.1$). However, when corrected by the number of unobserved events (Fig 7D), the picture is
518 the same for all bin sizes: shuffled data have always higher k_c values than the real ones (Mann-
519 Whitney test, $P < 10^{-6}$ and $P < 0.01$ for WN at 1ms). This confirms that the RGC neural code has
520 interdependences that can be mapped onto a lower dimensional space (i.e. compressed), compared
521 to the shuffled version, that lacks of interdependences by construction. Yet, the k_c obtained for
522 the shuffled raster is still very small (almost a half) compared to the random PWR2 raster, which
523 suggests that just the firing rates distribution, i.e. the diversity of firing rates, introduces some non-
524 random interdependences in the neural code, allowing compression. Finally, for medium time scales
525 (10-20ms), shuffled data shows an increase of k_c (after the correction by U_E) with the stimuli high-
526 order correlation, on the opposite to empirical data, where at this time scales NM induces a lower
527 dimensionality, compared to WN. This supports the idea that there are time scales where the stimuli
528 redundancies are reflected on the retinal activity and that this timescale is not present anymore on
529 shuffled data.

530 In sum, the density of the RGC rasters increases with the stimuli high-order correlations. Also,
531 RGC neural code is highly compressible compared to a random raster of same network size. Even
532 compared to shuffled data, the RGC neural code is more compressible. Furthermore, a significant
533 compression can be achieved considering just the firing rates distribution of empirical data, suggesting
534 an effect of the diversity of firing rates on the code compressibility. Although stimuli high-order
535 correlations increase the raster density, this activity can be compressed on a lower dimension for
536 NM than WN using a bin of 20ms. Thus, an increased raster density induced by stimuli does not
537 imply necessarily more coding channels. Instead, an increased raster density could be mapped onto
538 a low-dimensional hidden structure, as in the case of concomitance of dense activity and oscillations.
539 Regarding the time scales, $k_c + U_E$ is inversely proportional to the bin size for all stimuli. On the one
540 hand, this could be due to the specific time scales at which more redundancy is present on the neural
541 activity. On the other hand, this could be due to artifactual correlations induced by binning. Both
542 scenarios are possible and not mutually excluding, however, the problem of binning neural data is far
543 from being solved.

544 Discussion

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

554 The reduction of the MEM dimensionality is directly linked with data compression, obtained from
555 the linear dependencies between variables of the MEM that reflects hidden (non linear) interactions
556 between neurons. For example, our analysis in the case of synthetic rasters, where in contrast to retinal
557 data both the firing rates and the pairwise interactions are defined randomly, shows a k_c value very close
558 to L (maximal dimensionality). This demonstrates that if there are no linear relationships between
559 the parameters by construction, the code is not compressible and we have almost one dimension per
560 parameter. On the opposite, retina data shows a significant compression of $\sim 50\%$, as expected
561 from a neural tissue where cells are driven by common inputs and cells are electrically coupled [26],
562 increasing the level of dependency between them. This compression reduces the dimensionality of
563 the MEM to a lower dimensional space, where each dimension is a linear combination of model
564 parameters, characterizing the population activity by a set of independent dimensions representing
565 the inner structure of the network activity.

566 Limits of the method

567 The first limitation of our method comes from the numerical approximation used to compute χ matrix,
568 which is obtained by summing monomials correlations at different time lags (Eq. 11). An exact
569 exponential decay of correlation function with time will ensure the convergence of the series. But,
570 because we are estimating correlations from finite rasters, it is hard to estimate correlations with large
571 time lags and errors accumulate.

572 To truncate this approximation we need to consider a trade-off between temporal resolution and
573 reduction of noise. To this end we use 4 time lags (i.e. 4 terms of the sum), which is equivalent to
574 the double of memory depth used in the model ($R = 2$). This numerical estimation of χ also imposes
575 limits on the method, given that considering too large R s will generate a χ matrix that is governed
576 by noise.

577 On the other hand, it is possible to compute χ from the model parameters, requiring a previous
578 model fitting step, as done by [25]. However, for $N > 20$ and $R > 1$ the MEM computation becomes
579 prohibitive as the network size and the memory depth of the model increases. So, despite the numerical
580 approximation and its intrinsic errors, computing χ from the empirical monomials time correlations
581 is the best approach we found to work with medium size networks and spatio-temporal constrains.

582 Nevertheless, without fitting the MEM model we miss information about the sign (positive or
583 negative) and the magnitude of the interactions (weak or strong) defining the the network topology
584 and statistics. However, the scope of this work was not to fit different models on data and test
585 its performance (e.g. Bayesian or Akaike information criterion that takes into account the model
586 parameters and likelihood [27]) neither study changes in network topology under different stimuli.
587 Instead, we focus on exploring the geometrical properties of the MEM and its meaning in terms of
588 the neural code redundancy and compressibility. As we presented here, these geometrical properties
589 can be directly extracted from the χ matrix without fitting a MEM.

590 k_c is not one value, but a set of values

591 The main challenge this methodology introduces is the selection of k_c . The selection of k_c is related
592 to the minimization of the so-called confidence volume, which not only depends on the number of
593 dimensions given by k_c , but also on the imposed tolerance ϵ . So, strictly speaking, the selection of
594 k_c depends on the tolerance ϵ that the analyst decide to use based on equation 22 . However, here
595 we used a criterion based on a trade-off between dimensionality and the minimal confidence volume
596 ($\log \mathcal{V}(k_c, \epsilon)$): as shown on Fig 2, $\log \mathcal{V}(k_c, \epsilon)$ decays monotonically as k_c increases and ϵ decreases,
597 reaching an inflection point where even if we decrease the accuracy by orders of magnitude k_c remains
598 almost unchanged. This means that we look for the minimal dimensions with the highest possible
599 accuracy. Thus, our choice of k_c was not based on fixing a tolerance, but rather on relationship
600 between the minimal confidence volume, the dimensionality and imposed accuracy.

601 Comparison with similar analysis on spike trains

602 To our knowledge, there is no previous work related to the analysis of the χ matrix considering
603 *spatio-temporal* pairwise interactions applied to neuronal networks.

604 The authors in [25] proposed a similar analysis considering only spatial interactions, i.e. the Fisher
605 Information Matrix (FIM) for the Ising model on *in vitro*, *in vivo* and *in silico* networks. The work of
606 [25] studies small neural networks (10 neurons), they looked for stiff neurons, which are related to stiff
607 dimensions (the largest FIM eigenvalues), proposing that those neurons are the ones giving stability to
608 the network, while the neurons involved on the sloppy dimensions (dimensions where the parameters
609 can have significant changes without affecting the model) are the ones involved on plasticity, allowing
610 the network to remodel its connections.

611 Similarly, using larger network sizes ($N = 50$) and spatio-temporal interactions, our approach
612 exploits the linear dependencies of neuron interactions (given by the χ matrix) to find a minimal set
613 of dimensions that better represents the neural code, which could be considered an equivalent to stiff
614 dimensions present in [25].

615 To this end, we found two set of stiff dimensions: the ones before the first cut-off, related mainly to
616 neuron firing rates and the second set, after the first cut-off, related to spatio-temporal interactions.
617 According to our framework, the sloppy dimension would be the ones beyond k_c , but we could also
618 interpret the first N dimensions as the stiff dimensions, the ones between N and k_c as the sloppy
619 dimensions and the ones beyond k_c just noisy dimensions. This re-interpretation arises from the
620 large magnitude difference between the first and second set of dimensions and is compatible with
621 our previous description of compressibility: to represent data with the minimal set of dimensions we
622 require both stiff and sloppy dimensions. So, we consider the sloppy dimensions as relevant, because
623 we need them to "explain" data, while we consider the irrelevant dimensions as noise.

624 Recently, Battaglia et al [28], studying large scale networks between brain areas, proposed a concept
625 called *Meta Connectivity*, which instead of analyzing the correlation between nodes of a network
626 (the usual functional/effective connectivity analysis), focuses on the correlations between the network
627 interaction along time. This means focusing not on the coupling $\omega_i(0)\omega_j(0)$ between neurons ω_i and
628 ω_j , but focusing on the interactions between the couplings $\omega_i(0)\omega_j(0)$ and $\omega_k(0)\omega_l(0)$. This provides
629 information about instantaneous high-order correlation for at least 3 nodes of the network (e.g. case
630 of $i = k$) and captures the relationships between modules of network activity. Thus, χ matrix is both
631 a functional/effective connectivity matrix (the matrix entries related to the correlations between firing
632 rates) and also a meta connectivity matrix (the matrix entries related to correlations between pair-
633 wise interactions), which also includes information about the temporal interactions between network
634 nodes and network modules. The extension of our analysis towards the understanding of the meta
635 connectivity has never been applied to networks of neurons. It is a future research direction, where
636 the focus is on the variability and dependence of the interactions of a network.

637 Stimuli-induced changes on RGC population activity

638 Retina data has significant high-order correlations, including pairwise spatial [2, 3], temporal [16]
639 interactions, triplets [9] and groups of neurons [10]. These correlations have been widely studied
640 under MEM, which can accurately reproduce the raster spatio-temporal patterns. Nevertheless, there
641 has been no work devoted to reduce the MEM dimensionality considering the inner dependences
642 between the population activity variables.

643 Here, as a proof of concept, we used retina data of a diurnal rodent under 3 different stimuli with
644 different statistics, from photopic spontaneous activity (spatio-temporal uniform full field, no second
645 order statistics), a spatio-temporal white noise (gaussian statistics) to a repeated natural movie (high-
646 order correlations both on time and space). From our analysis, we know that these stimuli high-order
647 correlations increases the magnitude of both the firing rates and the raster high-order correlations,
648 even making silent cells to fire. This could be related to recruitment of specific cell types by the stimuli
649 features (e.g. local contrast [29], optic flow, color [30], among others).

650 In general, most of the correlated activity observed in the retina could be attributed to the receptive
651 field overlap between recorded RGCs and to shared common noise [31]. Additionally, electrical coupling

652 are highly present in the *O. degus* retina [26] inducing fast correlations between close cells. Also, there
653 could be correlations driven by amacrine cells, but the mechanisms and roles of those cells on the
654 *O. degus* retina is still unknown. The aforementioned causes of correlations modifies the pairwise
655 spatio-temporal interactions observed in real data. Furthermore, we explored the presence of these
656 correlated activity in the neural code. To do this we compared the results of recorded data with a
657 shuffled version of it, which preserves the firing rates distribution. Interestingly, both data sets share
658 the pairwise interactions probability distribution suggesting, in this case, that second order statistics
659 can be preserved just fixing the firing rates distribution.

660 Finally, as a main conclusion, our method suggest that RGC activity has significant high-order
661 statistics that are modified by stimuli, compared to shuffled data. Thus, this significant interactions
662 on RGC data are the base of the increased compressibility compare to shuffled data.

663 Compressibility of the RGC code

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

667 The first difference we found between stimuli was on χ spectrum, which shows an increase on the
668 offset, i.e. the eigenvalues increase their magnitude as the stimuli high-order correlation increases.
669 We recall that the stimuli-high order correlation increases the raster density. But given the way χ
670 is computed (see Eq. 11), the shape of the spectrum comes not only from the increased monomials
671 probabilities, but also from the dependence between the set of MEM monomials, all of them cap-
672 tured by the matrix. On the one hand, shuffled data also exhibit these differences on the eigenvalue
673 magnitudes (Fig 7B), showing that eigenvalues magnitude are closely related to the raster density
674 (also shown for synthetic data on Fig 5C). On the other hand, the differences between the cut-off for
675 empirical and shuffled data would be related to the linear dependences between the monomials, that
676 in the latter case are destroyed. Then, using shuffled data as a control, we suggest that the magnitude
677 of the eigenvalue spectrum highly depends on the monomials probability (raster density) while the
678 cut-off depends on the hidden linear dependencies between them.

679 The second difference we found between stimuli is on k_c and \mathcal{C} i.e. our approximation to the
680 compressibility of the neural code. As expected from the stimuli statistics and retinal stimuli inte-
681 gration, k_c and \mathcal{C} , are always lower for PSA than for the dynamic stimuli. This suggests that the
682 network optimizes the number of dimensions required for coding the stimuli depending on the stimuli
683 statistics. In terms of metabolic cost, a very redundant stimulus as PSA (which has the same spatial
684 and temporal information all over the stimuli space), may be coded with less dimensions than stimuli
685 with more independent components (less redundant), thus, optimizing the metabolic resources.

686 However, for bin sizes of 1 and 5ms we observe that k_c is higher for NM than for WN. This relation
687 varies for larger values of bin sizes, such as 10 or 20ms. For bin sizes of 10ms WN has the same number
688 of relevant dimensions than NM, while for 20ms WN has more dimensions than NM. So, for fast time
689 scales, we face a non-optimal situation, because NM has more redundancies than the WN. It could be
690 possible that at this time scales we are not capturing the inter-dependences of the neural code that are
691 relevant for the brain. For example, at 20 ms bin size, we see the expected optimal effect: the system
692 exploits the stimuli redundancies and exhibit more compression for NM than for WN. Coincidentally,
693 many MEM on retina have been done using 20 ms as bin size [2, 3], which in our case is the bin that
694 allows the highest compression. In addition, synthetic data shows that if the raster is too dense, the
695 underlying statistics hides under the noisy activity, which could also be possible at large values of bin
696 sizes. To control this situation we used shuffled rasters, which preserves the same raster density. In
697 the shuffled rasters we see that k_c increases with the raster density, discarding the effect of density on
698 the changes of dimensionality at higher bin sizes.

699 Thus, at large bin sizes, the interdependences of the neural code are responsible for the compression
700 effect and not the raster density hiding some events. Nevertheless, we do not know in advance what
701 time scale(s) is(are) actually relevant for the brain and neither if our assumptions about code optimality
702 and the neural code variables (firing rate and spatio-temporal interactions) are right, so the choice of
703 the bin size and code variables is still an open question and somewhat arbitrary.

704 Finally, our work is related to the idea that a stimuli-dependent network noisy spiking neurons
 705 adapt its code according to noise and stimuli correlations [21], instead of using just one way of coding.
 706 In our case, the stimuli-dependent network is a biological one, so we don't have access to modify the
 707 noise of each neuron nor the network noise. Instead, we can just modify the stimuli correlations,
 708 which changes the dimensionality of the code. This change in dimensionality could reflect the smooth
 709 interpolation between encoding strategies: highly redundant stimuli evokes fewer dimensions than
 710 stimuli which presents high-order correlations.

711 This suggests that the MEM dimensionality is also a measure of the code redundancies. The
 712 analytic relationship between dimensionality reduction and the coding strategies require an extensive
 713 mathematical and computational research that is not developed here, however, we provide an indirect
 714 way of studying the interdependences of the neural code as the stimuli conditions changes.

715 Supplementary Material

716 **Gibbs measures in the sense of Bowen.** Consider a potential \mathcal{H}_h of range $R \geq 2$. A shift invariant
 717 probability measure μ is called a Gibbs measure (in the sense of Bowen) if there are constants $M > 1$
 718 and $\mathcal{P}[\mathcal{H}_h] \in \mathbb{R}$ s.t.

$$M^{-1} \leq \frac{\mu[\omega_1^n]}{\exp(\sum_{k=1}^{n-R+1} \mathcal{H}_h(\omega_k^{k+R-1}) - (n+R-1)\mathcal{P}[\mathcal{H}_h])} \leq M \quad (17)$$

719 It is easy to see that the classical Boltzmann-Gibbs distribution is a particular case of (17), when
 720 $M = 1$ and \mathcal{H} is a potential of range 1.

721
 722 **Ruelle-Follmer theorem:** Suppose μ' is a Gibbs measure for some potential $\mathcal{H}_{h'}$, and μ is another
 723 Gibbs measure. Then the relative entropy density:
 724

$$d(\mu | \mu') = \mathcal{P}[\mathcal{H}_{h'}] - S(\mu) - \mu(\mathcal{H}_{h'}) \quad (18)$$

725 if $d(\mu | \mu') = 0$, we obtain the variational characterization of Gibbs measures (4).

726 Following [36], consider the potential $\mathcal{H}_h = \sum_{l=1}^L h_l m_l$ associated with an ergodic Markov Chain
 727 $\mu(P, \pi)$. Consider a sample of $\mu(P, \pi)$ of length n and the observables $\{m_l\}_{l=1}^L$. We may obtain
 728 from the sample new maximum entropy parameters \mathbf{h}' . The probability that the maximum entropy
 729 parameters \mathbf{h}' associated with an ergodic Markov Chain $\mu'(P', \pi')$ get close to \mathbf{h} follow the asymptotic
 730 relationship:

$$\lim_{\delta \rightarrow 0} \lim_{n \rightarrow \infty} -\frac{1}{n} \ln \mathbb{P} \left(|\mathbf{h} - \mathbf{h}'| \in \Delta\delta \right) = d(\mu | \mu') \quad (19)$$

731 where $\Delta\delta = [-\delta, \delta]^K$. Choosing $\Delta\delta$ close to 0 we may formally rewrite the above relationship in the
 732 form:

$$-\frac{1}{n} \ln \mathbb{P} \left(|\mathbf{h} - \mathbf{h}'| \in \Delta\delta \right) \xrightarrow{n \rightarrow \infty} d(\mu | \mu') \quad (20)$$

733 Thus, for large n ,

$$\mathbb{P} \left(|\mathbf{h} - \mathbf{h}'| \in \Delta\delta \right) \approx e^{-nd(\mu | \mu')}$$

734 Meaning that close-by parameters index very similar distributions [41].

735 Consider two maximum entropy Markov chains $\mu(P, \pi)$ and $\mu'(P', \pi')$ specified by \mathcal{H}_h and $\mathcal{H}_{h'}$ respec-
 736 tively. As both satisfy the variational principle, we have that the relative entropy (18) reads:

$$d(\mu | \mu') = \mathcal{P}[\mathcal{H}_{h'}] - \mathcal{P}[\mathcal{H}_h] + \mu(\mathcal{H}_h) - \mu(\mathcal{H}_{h'}) \quad (21)$$

737 Taking the expansion of $d(\mu | \mu')$ around $\mathbf{h}' = \mathbf{h}$ we obtain:

$$d(\mu | \mu') \approx d(\mu | \mu) + \sum_k \frac{\partial d(\mu | \mu')}{\partial h'_k} \Big|_{h'=h} (h_k - h'_k) + \frac{1}{2} \sum_{k,j} \frac{\partial^2 d(\mu | \mu')}{\partial h'_k \partial h'_j} \Big|_{h'=h} (h_k - h'_k)(h_j - h'_j)$$

738 Since $d(\mu | \mu')$ is minimized at $h' = h$:

$$d(\mu | \mu') \approx \frac{1}{2} \sum_{k,j} \frac{\partial^2 d(\mu | \mu')}{\partial h'_k \partial h'_j} \Big|_{h'=h} (h_k - h'_k)(h_j - h'_j)$$

739 Taking the second derivative of $d(\mu | \mu')$ from (21), we obtain:

$$\chi^{kj} = \frac{\partial^2 d(\mu | \mu')}{\partial h'_k \partial h'_j} = \frac{\partial^2 \mathcal{P}[\mathcal{H}_{h'}]}{\partial h'_k \partial h'_j}$$

740 Given two maximum entropy Markov chains specified by \mathcal{H}_h and $\mathcal{H}_{h'}$ in the limit of large T they are
741 ϵ -indistinguishable if:

$$\frac{1}{2} [(\mathbf{h} - \mathbf{h}')^T \chi (\mathbf{h} - \mathbf{h}')] \leq \frac{\epsilon}{T} \quad (22)$$

742 where χ is the Fisher information matrix, which represents the curvature of the relative entropy.

743 **Materials and Methods**

744 **Ethics Statement**

745 Animal manipulation and breeding and corresponding experiments were approved by the bioethics
746 committee of the Universidad de Valparaiso, in accordance with the bioethics regulation of the Chilean
747 Research Council (CONICYT) and international protocols.

748 **Animals and Recordings**

749 4 Adult male and female Octodon degus (3-6 months) were maintained in the animal facility of the
750 Universidad de Valparaiso, at 20–25°C on a 12-h light-dark cycle, with access to food and water ad-
751 libitum. The methods of MEA recording has previously been described [26]. In brief, animals were
752 euthanized under deep isofluorano or halothane anesthesia and both eyes were extracted. Then, one of
753 the extracted retinas was diced into quarters while the other was stored in oxygenated in oxygenated
754 (O_2 95% CO_2 5%) AMES medium at 32°C in the dark for further experiments. The same AMES
755 media was used for continuous perfusion during extracellular recordings. For MEA recordings (MEA
756 USB-256, 20kHz sample, Multichannel Systems GmbH, Germany), one piece of retina was mounted
757 onto a dialysis membrane placed into a ring device mounted in a traveling (up/down) cylinder, which
758 was moved to contact the electrode surface of the MEA recording array. Data were processed off-line
759 using Plexon Offline Sorter (Plexon Instruments). Further, spikes were detected using a threshold of
760 -4.5 to -5.5 S.D. from the mean voltage value and then were manually classified using the 2D space of
761 the first two principal components on each electrode. Only somatic spikes were kept. Refractory period
762 violations were detected and discarded if two or more spikes of the same neuron occur in a 2ms period.
763 We recorded on 3 stimuli conditions (see next section): i) Photopic Spontaneous Activity (PSA), ii)
764 Spatio-temporal white Noise (WN) and iii) Natural Movie (NM), obtaining 151, 200, 246 and 270
765 RGC for each of the 4 experiments, respectively. For each experiment 30 random subsamples of 50
766 neurons were taken and χ matrix corresponding to a Pairwise R=2 model were computed considering
767 4 bin sizes (1, 5, 10 and 20 ms) for each subsample, yielding 30 k_c values per experiment, per condition
768 and per bin size. The shuffled version of these rasters were submitted to the same analysis.

769 Visual Stimuli

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

787 Generation of Synthetic Data

788 Synthetic rasters ($T = 2.10^6$ time-points, $N = 20$ neurons) were generated using different underlying
789 statistics: **Independent**, where only firing rates are defined, $L = N$; **Pairwise R=2** (PWR2), with
790 firing rates and spatio-temporal correlations, $L = N(3N - 1)/2$. Underlying coefficients related to
791 firing rates and to pairwise interactions were randomly chosen from a normal distribution with mean
792 -5 and -1, respectively, and 1 standard deviation. were generated scaling the magnitude of the model
793 parameters by a factor $\beta = [0.4 \ 0.6 \ 0.8 \ 1.2 \ 1.4]$. In addition, 6 more random PWR2 rasters were
794 generated with $N = [30 \ 40 \ 50 \ 60 \ 70 \ 80]$ to study the dependence between k_c estimation and the
795 network size. Each raster was generated using the PRANAS software (<https://pranas.inria.fr/>) [45].
796 For the first 3 rasters, 100 random subsamples with half duration of the whole recording were taken
797 for each raster and the χ matrix associated with a Pairwise R=2 model was computed. Then, k_c (see
798 text) was found by volume minimization, yielding 100 k_c values per raster. For the scales rasters, the
799 same procedure was applied, but using 10 temporal subsamples.

800 Shuffling

801 In order to destroy the dependencies between the empirical raster monomials, we have generated
802 random rasters where the number of neurons and firing rates was exactly the same than observed
803 on the recordings (i.e. on each retina under each stimuli condition), but the spikes times were taken
804 uniformly at random, avoiding violations of the refractory period (2ms) [44]

805 Acknowledgments

806 Financial support: CONICYT-FONDECYT 1140403 and 1150638, CONICYT-Basal Project FB0008
807 (Chile), CONICYT-PAI Inserción 79160120; Grant ICM-P09-022-F supported by the Millenium Sci-
808 entific Initiative of the Ministerio de Economía, Desarrollo y Turismo (Chile); ECOS-Conicyt C13E06;
809 ONR Research Grant #N62909-14-1-N121, ANR TRAJECTORY CE37 (France). We thank Felipe
810 Olivares and Michael Pizarro to help in the experiments.

811 References

812 [1] Rieke F, Warland D, van Steveninck R, Bialek W, Spikes, Exploring the Neural Code, The M.I.T.
813 Press. 1996.

- 814 [2] Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly corre-
815 lated network states in a neural population. *Nature*. 2006.
- 816 [3] Shlens J, Field GD, Gauthier JL, Grivich MI, Petrusca D, Sher A, Litke AM, Chichilnisky EJ.
817 The Structure of Multi-Neuron Firing Patterns in Primate Retina. *Journal of Neuroscience*. 2006.
- 818 [4] Da Silveira RA, Berry MJ. High-Fidelity Coding with Correlated Neurons. *PLoS Computational*
819 *Biology*. 2014.
- 820 [5] Brunel N, Hakim V. Fast global oscillations in networks of integrate-and-fire neurons with low
821 firing rates. *Neural Computation*. 1999.
- 822 [6] Lindner B, Doiron B, Longtin A. Theory of oscillatory firing induced by spatially correlated noise
823 and delayed inhibitory feedback *Physical Review E*. 2005.
- 824 [7] Trousdale J, Hu Y, Shea-Brown E, Josic K. Impact of network structure and cellular response
825 on spike time correlations. *PLoS Computational Biology*. 2012.
- 826 [8] Ganmor E, Segev R, Schneidman E. The architecture of functional interaction networks in the
827 retina. *Journal of Neuroscience*. 2011.
- 828 [9] Ganmor E, Segev R, Schneidman E. Sparse low-order interaction network underlies a highly
829 correlated and learnable neural population code. *Proceedings of the National Academy of sciences*.
830 2011.
- 831 [10] Tkacik G, Marre O, Mora T, Amodei D, 2nd MJB, Bialek W. The simplest maximum entropy
832 model for collective behavior in a neural network. *Journal Statistical Mechanics*. 2013.
- 833 [11] Marre O, El Boustani S, Frégnac Y, Destexhe A. Prediction of Spatiotemporal Patterns of Neural
834 Activity from Pairwise Correlations. *Physical Review Letters*. 2009.
- 835 [12] Cofré R, Cessac B. Dynamics and spike trains statistics in conductance-based Integrate-and-Fire
836 neural networks with chemical and electric synapses. *Chaos, Solitons and Fractals*. 2013.
- 837 [13] Cofré R. and Cessac B. Exact computation of the maximum-entropy potential of spiking neural-
838 network models. *Physical Review E*. 2014.
- 839 [14] Nasser H, Marre O, Cessac B. Spatio-temporal spike train analysis for large scale networks
840 using the maximum entropy principle and Montecarlo method. *Journal of Statistical Mechanics:*
841 *Theory and Experiment*. 2013.
- 842 [15] Nasser H, Cessac B. Parameters estimation for spatio-temporal maximum entropy distributions:
843 application to neural spike trains. *Entropy*. 2014.
- 844 [16] Vasquez JC, Marre O, Palacios A, Berry MJ, Cessac B. Gibbs distribution analysis of temporal
845 correlation structure on multicell spike trains from retina ganglion cells. *Journal Physiology Paris*.
846 2012.
- 847 [17] Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, Simoncelli EP. Spatio-
848 temporal correlations and visual signalling in a complete neuronal population. *Nature*. 2008.
- 849 [18] Trenholm S, Mclaughlin AJ, Schwab DJ, Turner MH, Smith RG, Rieke F, Awatramani GB. Non-
850 linear dendritic integration of electrical and chemical synaptic inputs drives fine-scale correlations.
851 *Nature Neuroscience*. 2014.
- 852 [19] Puchalla JL, Schneidman E, Harris Ra, Berry MJ. Redundancy in the population code of the
853 retina. *Neuron*. 2005.
- 854 [20] Narayanan NS, Kimchi EY, Laubach M. Redundancy and synergy of neuronal ensembles in motor
855 cortex. *Journal of Neuroscience*. *Neuroscience*. 2005.

- 856 [21] Tkacik G, Prentice JS, Balasubramanian V, Schneidman E. Optimal population coding by noisy
857 spiking neurons. *Proceedings of the National Academy of Sciences*. 2010.
- 858 [22] Simmons KD, Prentice JS, Tkacik G, Homann J, Yee HK, Palmer SE, Nelson PC, Balasubra-
859 manian V. Transformation of stimulus correlations by the retina. *PLoS Computational Biology*.
860 2013;
- 861 [23] Sessak V, Monasson R. Small-correlation expansions for the inverse Ising problem. *Journal of*
862 *Physics A: Mathematical and Theoretical*. 2009.
- 863 [24] Amari Si, Nagaoka H, Harada D. *Methods of information geometry*. Translations of mathematical
864 monographs of the AMS. (Oxford University Press). 2000.
- 865 [25] Panas D, Amin H, Maccione A, Muthmann O, van Rossum M, Berdondini L, Hennig MG. Slop-
866 piness in Spontaneously Active Neuronal Networks. *Journal of Neuroscience*. 2015.
- 867 [26] Palacios-Muñoz A, Escobar MJ, Vielma A, Araya J, Astudillo A, Valdivia G, García IE, Hurtado
868 J, Schmachtenberg O, Martínez AD, Palacios AG. Role of connexin channels in the retinal light
869 response of a diurnal rodent. *Frontiers in Cellular Neuroscience*. 2014.
- 870 [27] Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and
871 its analytical extensions. *Psychometrika*. 1987.
- 872 [28] Battaglia D, Thomas B, Hansen EC, Chettouf S, Daffertshofer A, McIntosh AR, Zimmermann
873 J, Ritter P, Jirsa V. Functional Connectivity Dynamics of the Resting State across the Human
874 Adult Lifespan. *bioRxiv* 107243. 2017.
- 875 [29] Zaghloul KA, Boahen K, Demb JB. Contrast adaptation in subthreshold and spiking responses
876 of mammalian Y-type retinal ganglion cells. *Journal of Neuroscience*. 2005.
- 877 [30] Gollisch T, Meister M. Eye smarter than scientists believed: neural computations in circuits of
878 the retina. *Neuron*. 2010.
- 879 [31] Vidne M, Ahmadian Y, Shlens J, Pillow JW, Kulkarni J, Litke AM, Chichilnisky EJ, Simoncelli
880 E, Paninski L. Modeling the impact of common noise inputs on the network activity of retinal
881 ganglion cells. *Journal of Computational Neuroscience*. 2012.
- 882 [32] Tkacik G, Prentice JS, Balasubramanian V, Schneidman E. Optimal population coding by noisy
883 spiking neurons. *Proceedings of the National Academy of Sciences*. 2010.
- 884 [33] Vasquez JC, Palacios A, Marre O, II MJB, Cessac B. Gibbs distribution analysis of temporal
885 correlation structure on multicell spike trains from retina ganglion cells. *Journal of Physiology*
886 *Paris*. 2012.
- 887 [34] Cessac B, Palacios A. Spike Train Statistics from Empirical Facts to Theory: The Case of the
888 Retina.. In: Cazals F., Kornprobst P. (eds) *Modeling in Computational Biology and Biomedicine*.
889 Springer. 2013.
- 890 [35] Keller G. *Equilibrium States in Ergodic Theory*. Cambridge University Press. 1998.
- 891 [36] Mastromatteo I. On the criticality of inferred models. *J. Stat. Mech. Theory Exp*, 2011.
- 892 [37] Ruelle D. *Statistical Mechanics: Rigorous results*. New York. 1969.
- 893 [38] Ruelle D. *Thermodynamic formalism*. Addison-Wesley, Reading, Massachusetts. 1978.
- 894 [39] Sinai Y. Gibbs measures in ergodic theory. *Russian Mathematical Surveys*. 1972.
- 895 [40] Georgii HO. Gibbs measures and phase transitions. *De Gruyter Studies in Mathematics*. 1988.

- 896 [41] Balasubramanian V. Statistical inference, Occam's Razor, and statistical mechanics on the space
897 of probability distributions. *Neural Computation*. 1997.
- 898 [42] Jaynes ET. Information theory and statistical mechanics. *Physical Review*. 1957.
- 899 [43] Privman V, Fisher MJ. Universal critical amplitudes in finite-size scaling. *Physical Review B*.
900 1984.
- 901 [44] Segev R, Goodhouse J, Puchalla J, Berry MJ. Recording spikes from a large fraction of the
902 ganglion cells in a retinal patch, *Nature Neuroscience*. 2004.
- 903 [45] Cessac B, Kornprobst P, Kraria S, Nasser H, Pamplona D, Portelli G, Viéville T. PRANAS: a
904 new platform for retinal analysis and simulation, *Frontiers in Neuroinformatics*. 2017.
- 905 [46] Savin C, Tkačik G. Maximum entropy models as a tool for building precise neural controls,
906 *Current Opinion in Neurobiology*. 2017.
- 907 [47] Tkačik G, Marre O, Mora T, and Amodei D, and Berry II MJ, and Bialek W. The simplest maxi-
908 mum entropy model for collective behavior in a neural network, *Journal of Statistical Mechanics*.
909 2013.