Ising models for neural activity inferred via Selective Cluster Expansion: structural and coding properties

John Barton[1,*,**] **and Simona Cocco**[2]

[1] Department of Physics, Rutgers University, Piscataway, NJ 08854 USA

[2] CNRS-Laboratoire de Physique Statistique de l'ENS, 24 rue Lhomond, 75005 Paris, France

[*] Present address: Department of Chemical Engineering, MIT, Cambridge, MA 02139

[**] Present address: Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Boston, MA 02129

E-mail: jpbarton@mit.edu, cocco@lps.ens.fr

Abstract. We describe the Selective Cluster Expansion (SCE) of the entropy, a method for inferring an Ising model which describes the correlated activity of populations of neurons [1, 2]. We re-analyze data obtained from multielectrode recordings performed *in vitro* on the retina and *in vivo* on the prefrontal cortex. Recorded population sizes N range from $N = 37$ to $N = 117$ neurons. We compare the SCE method with the simplest mean field methods (corresponding to a Gaussian model) and with regularizations which favor sparse networks (L_1 norm) or penalize large couplings (L_2 norm). The network of the strongest interactions inferred via mean field methods generally agree with those obtained from SCE. Reconstruction of the sampled moments of the distributions, corresponding to neuron spiking frequencies and pairwise correlations, and the prediction of higher moments including three-cell correlations and multi-neuron firing frequencies, is more difficult than determining the large-scale structure of the interaction network, and apart from a cortical recording in which the measured correlation indices are small, these goals are achieved with the SCE but not with mean field approaches. We also find differences in the inferred structure of retinal and cortical networks: inferred interactions tend to be more irregular and sparse for cortical data than for retinal data. This result may reflect the structure of the recording. As a consequence the SCE is more effective for retinal data when expanding the entropy with respect to a mean field reference $S - S_{\rm MF}$, while expansions of the entropy S alone perform better for cortical data.

Keywords: statistical inference, neuronal networks

1. Introduction

Recent years have seen the advance of *in vivo* [3] and *in vitro* [4] recording of populations of neurons. The quantitative analysis of these data presents a formidable computational challenge.

Bialek and coworkers [5] and Chichilnisky [6] and coworkers have introduced the idea of using the Ising model to describe patterns of correlated neural activity. The Ising model is the maximum entropy model, *i.e.* the least constrained model [7], capable of reproducing observed spiking frequencies and pairwise correlations in the neural population. This model has proved to be a good model for the activity, in the sense that even though one only constrains the model using the first and second moments, one can predict features of the activity such as third moments and multi-neuron firing patterns.

The inverse Ising approach, consisting of the inference of an effective Ising model from experimental data, goes beyond the cross-correlation analysis of multielectrode recordings by disentangling the network of direct interactions from correlations. The inverse Ising approach can then be used to extract structural information on effective connections, such as how they change with the distance, time scale, and different types of cells. One can also explore how effective connections change during activity due to adaptation, short term memory, and long term learning.

Another key feature of the Inverse Ising approach is the ability to take into account that measured correlations are affected by finite sampling noise. Experiments are limited in time and their ability to thoroughly sample the experimental system, therefore it is important to avoid overfitting of data. Overfitting can have drastic effects on the structure of the inferred interaction network. One way to avoid overfitting is to regularize the inference problem by introducing *a priori* information about the model parameters in the log-likelihood of the Ising model giving the data.

However, inferring an Ising model from a set of experimental data is a challenging computational problem with no straightforward solution. Typical methods of solving the inference problem include the Boltzmann learning method, which involves iterative Monte Carlo simulations followed by small updates to the interaction parameters [8]. This method can be very slow for large systems, though recent advances have notably improved the speed, and this method has proved to be effective in the analysis of neural data [9–11]. Other methods such as iterative scaling algorithms [12, 13], pseudo-likelihood approximation [14, 15], various perturbative expansions [16–19] and mean field (or Gaussian model) approximations [20] have also been developed which attempt to solve the Inverse Ising problem in certain limits. Such approximations are often computationally simple, but suffer from a limited range of validity. One method that has recently proven to effectively infer the parameters of spin glass models is the method of minimum probability flow [21].

Here we review a statistical mechanical approach [1, 2] based on a selective cluster expansion which improves upon mean field methods. Moreover it avoids overfitting by

selecting only clusters with significant contributions to the inverse problem, minimizing the impact of finite sampling noise. We compare this approach with the mean field or Gaussian approximation and we test different forms of regularization, in particular the norm L_1 regularization, which preferentially penalizes small but nonzero couplings, and norm L_2 regularization, which penalizes large couplings in absolute value more heavily.

Our aim is to compare the methods using the structure and sparsity of the inferred interaction network, and the ability of the inferred Ising model to reproduce the multi-neuron recorded activity. These two features correspond respectively to the capacity to give structural information about the system which has generated the data and the ability of the Ising model to encode neural activity.

The outline of this paper is as follows. In the remainder of the introduction we give a description of the neural data, then detail two methods for analyzing the data: cross-correlation histogram analysis, which focuses on correlations between pairs of cells, and the inverse Ising approach, which attempts to infer an effective network of interactions characterizing the whole system. In Section 2 we give an overview of the effects of finite sampling noise in the data on the Ising model inference problem. This includes, in particular, quantifying the expected fluctuations of the empirical correlations and the inferred Ising model couplings and fields, as well as introducing methods of regularizing the inference procedure to minimize problems due to finite sampling. Section 3 gives a discussion of the difficulties of the inverse Ising problem. In Sections 4 and 5, we describe two methods for solving the inverse problem: the mean field or Gaussian approximation, and the Selective Cluster Expansion (SCE). Here we discuss regularization conventions for both the mean field and SCE, as well as practical computational methods and questions of convergence for the SCE.

Beginning with Section 6 we focus on applications to real data. In this Section we assess the performance of the SCE algorithm on retinal and cortical data. We then describe the properties of the inferred Ising models following the analysis performed previously in [17] on retinal data, showing that the inferred models reconstruct the empirical spiking frequencies and pairwise correlations, and evaluating their ability to predict higher order correlations and multi-neuron firing frequencies. Structural properties of the interaction networks in retinal and cortical data, including maps of inferred couplings and the reliability of the inferred positive and negative couplings are also discussed. We illustrate the importance of network effects by comparing the couplings obtained through SCE with the those obtained from the correlation indices, which only depend on properties of pairs of cells rather than on the entire system. We also compare couplings inferred with SCE on a subset of each experimental system with those for the full system. In Section 7 we explore the performance of the mean field approximation using a variety of regularization methods and strengths. Couplings inferred via the mean field approximation and those obtained from SCE are also compared. Finally in Section 8 we discuss the limitations of the algorithm and some possible extensions of the static inverse Ising approach in treating neural data.

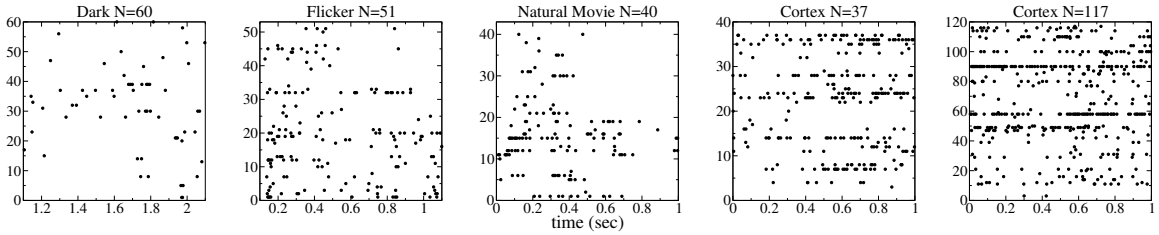

Figure 1. Raster plot of spike train data from recordings of neurons populations: in the retina in dark conditions ($N = 60$ cells) and with a random flickering stimulus ($N = 51$ cells), data by M. Meister [22]; in the retina with a natural movie stimulus ($N = 40$ cells), data by M. Berry [5]; in the prefrontal cortex of a rat ($N = 32$ cells), data by A. Peyrache and F. Battaglia [23]; in the medial prefrontal cortex of a rat ($N = 117$ cells), data by G. Buzsáki and S. Fujisawa [24]. Recordings last for 30 minutes – 1 hour, here we plot only 1 s of the recording. One can translate the continuous time data in the raster plot into binary patterns of activity by arranging the time interval into bins of size Δt and recording whether or not each neuron spikes within each time bin. The probability p_i that a neuron i spikes in a time window of size Δt is the number of windows in which the neuron i is active divided by the total number of time windows. The probability p_{ij} that two neurons i, j are active in the same time window is given by the number of time windows in which both the neurons are active divided by the total number of time windows.

1.1. Description of neural data

We begin by briefly describing the *in vitro* and *in vivo* multielectrode recordings (see Fig. 1) of neural activity used in our analysis.

The first set of recording is done *in vitro* on the retina [5, 22, 25]. Here the retina is extracted from the eye of an animal – typically a salamander, guinea pig, or primate – and placed on a multielectrode array with all five layers of cells (photoreceptors, horizontal, bipolar, amacrine, and ganglion cells) responsible for detecting and preprocessing visual signals. This allows for the simultaneous recording of tens to hundreds of cells in the last layer consisting of ganglion cells while the retina is subjected to various visual stimuli.

The neural network of the retina is relatively simple: it has a feed-forward architecture and preprocesses visual information, which is then transmitted through the optical nerve to the visual cortex [6, 26]. The retina thus presents an ideal environment for studying how a stimulus is encoded in the ganglion cell activity and how information is processed [27].

The first experiments on the retina [28] defined the concept of receptive fields, which are regions in the visual plane to which a ganglion cell is sensitive. Light which is applied to the receptive field of a ganglion cell stimulates a response from the cell. Early studies also identified different ganglion cell types [27], denoted as ON and OFF. ON cells respond when light is switched on in the center of the cell's receptive field, while OFF cells respond when light in the receptive field is switched off.

In the experimental work of Arnett [29] and Mastronarde [30], the correlated

activity of ganglion cells has been studied via the simultaneous recording of pairs of ganglion cells. Neighboring cells tend to spike in synchrony if they are of the same type and tend to be de-synchronized if they are of different types. Cross-correlation histograms, which describe the correlated firing of pairs of neurons [27], have been particularly useful for identifying circuits of connections between ganglion cells in the retina and for determining the distance dependence of functional connections between ganglions [4, 6, 31].

The second set of recordings is done *in vivo* on rats. These recordings are obtained by implanting tetrodes or silicon probes which simultaneously record neural activity on several layers of the prefrontal cortex [3]. Probes can be implanted for several weeks in a rat, and allowing for the study of mechanisms of memory formation. Rats are trained in a working memory task and recordings are typically performed before, during, and after the learning of the task [24, 32].

In the work of Peyrache, Battaglia and collaborators, recordings are performed on the medial prefrontal cortex (prelimbic and infralimbic area) with 6 tetrodes, each of which consists of 4 electrodes. The analysis of cross-correlations between cells through Principal Component Analysis has shown that during sleep neural patterns of activity appearing in the previous waking experiences are replayed. This could be a mechanism for consolidating memory formation [23, 32, 33].

In the work of Fujisawa, Buzsáki and collaborators, recordings are performed with silicon probes which record from either the superficial (layers 2 and 3) or deep (layer 5) layers of the medial prefrontal cortex. Cross-correlation analysis of these recordings has identified neurons which differentiate between different trajectories of the rat in a maze, as well as signatures of short term plasticity in the working memory task.

In all cases, raw data is obtained in the form of voltage measurements from each electrode in the experimental apparatus. The raw data is then analyzed to determine the number of neurons being recorded and to assign each voltage spike to a particular neuron, in a process known as spike sorting [25, 34, 35]. The end result is a set of spike trains – a list of all of the times at which a particular neuron fired – for each neuron observed in the experiment.

1.2. Definition of correlation index

Spatial and temporal correlations in the data can be described in terms of a cross-correlation histogram, which examines the correlations in the firing of pairs of neurons over time. Cross-correlation analysis has been an important analytical tool in the study of neural activity. Let us denote the set of recorded times of each spike event for a neuron labeled by i by $\{t_{i,1}, \ldots, t_{i,N_i}\}$, with N_i the total number of times that the neuron spikes during the total recording time T. This data can be extracted from the spike trains depicted in Fig. 1. One can then determine the cross-correlation histogram of spiking

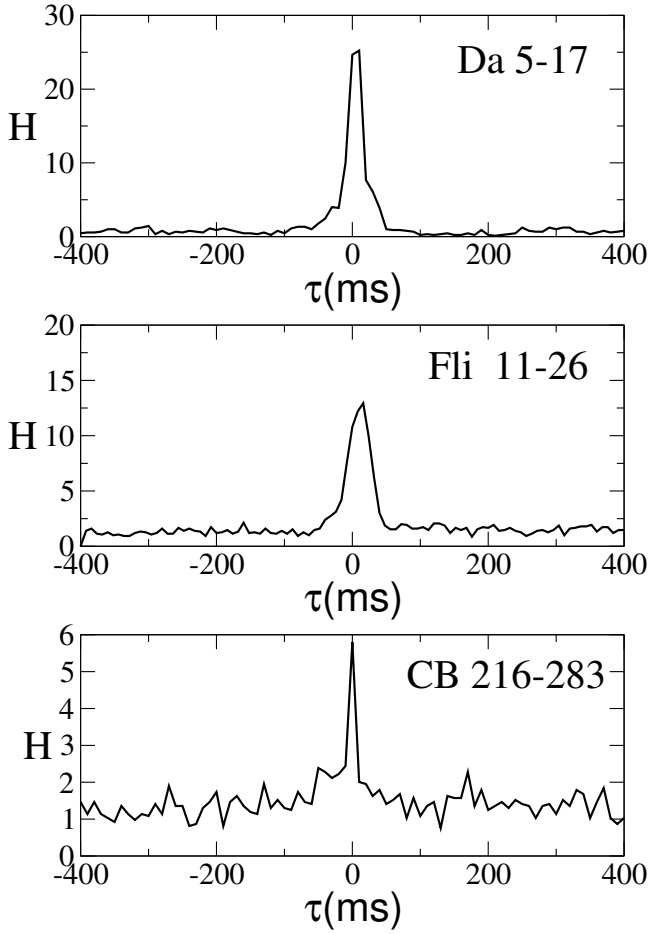

Figure 2. Examples of cross-correlation histograms between two cells with an inferred positive coupling (see also discussion in Section 6), from recordings of 60 cells in retina in dark (Da 5-17), and 51 cells in retina with a random flickering stimulus (Fl 11-26); 117 cells in medial prefrontal cortex of a rat (CB 216-283). A characteristic correlation time of approximately 20 ms corresponding to the central peak is visible.

delays between each pair of cells i, j,

$$H_{ij}(\tau, \Delta t) = \frac{T}{N_i\, N_j\, \Delta t} \sum_{a=1}^{N_i} \sum_{b=1}^{N_j} \theta_{\tau,\Delta t}(t_{i,a}, t_{j,b}) \tag{1}$$

Here Δt is the bin width of the histogram, and $\theta_{\tau,\Delta t}(t_{i,a}, t_{j,b})$ is an indicator function which is equal to one if $|\tau + t_{i,a} - t_{j,b}| < \Delta t/2$ and zero otherwise. The cross-correlation histogram can be interprered as the probability that a cell j spikes in the interval $\tau \pm \Delta t/2$, conditioned on the cell i spiking at time zero, and normalized by the probability the cell j spikes in some time window Δt. The cross-correlation histogram approaches one at very long times τ when the firing of the cell j is independent from the fact that the cell i has fired at time zero.

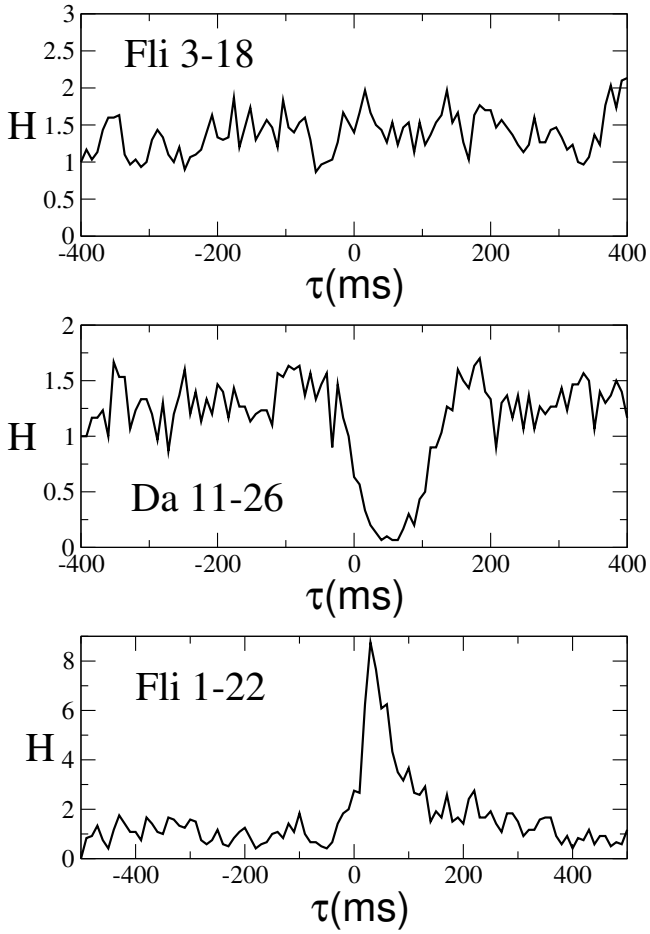

Figure 3. Examples of cross-correlation histograms between two cells with an inferred negative coupling (see also discussion in Section 6), from recordings of 51 cells in retina with a random flickering stimulus, (Fl 3-18) shows no large anticorrelation bump, (Fl 1-22) shows a correlation peak delayed by about 50 ms; 60 cells in retina in dark, (Da 11-26) displays an anticorrelation bump with the characteristic time scale of 100 ms.

The correlation index is a measure of synchrony between two cells, and is defined as the function H_{ij} at its central peak ($\tau = 0$), with bin size Δt

$$CI_{ij}(\Delta t) = H_{ij}(0, \Delta t) \approx \frac{p_{ij}(\Delta t)}{p_i(\Delta t)\, p_j(\Delta t)}. \tag{2}$$

$C_{i,j}(\Delta t)$ is the number of spikes emitted by the cells i and j with a delay smaller than Δt, normalized by the number of times the two cells would spike in the same time window if they were independent. For Δt small with respect to the typical interval between subsequent spikes of a single cell, there are never two spikes of the same cell in the same time bin and therefore the correlation index is exactly the probability that two cells are active in the same bin $p_{ij}(\Delta t)$ divided by the probability $p_i(\Delta t)\, p_j(\Delta t)$ they would be active in the same time window assuming no correlation between them.

Some cross-correlation histograms for the data sets we consider are shown in Figs. 2–4. The time bin Δt is a parameter in the correlation index analysis which defines

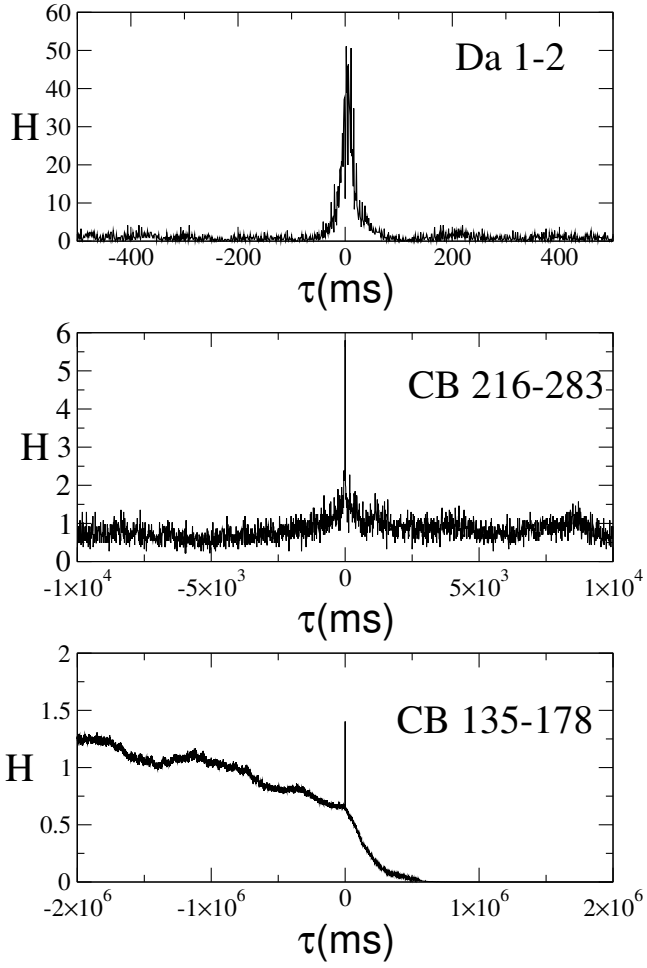

Figure 4. Examples of cross-correlation histograms between two cells at long time scales, from recordings of 60 cells in retina in dark, (Da 1-2) corresponds to a positive coupling and no long time effects; 117 cells in medial prefrontal cortex of a rat, (CB 216-283) corresponds to a positive coupling and no long time effects, (CB 135-178) displays nonstationary effects because cell 178 is active only in the first half of the recording.

the relevant time scale for correlations. A characteristic time scale of some tens of milliseconds is shown in Fig. 2 [36]. However, in retinal recordings long time patterns appear especially in the presence of stimulus, and *in vivo* cortical recordings show long time correlations which can be due, for example, to different firing rhythms or trajectory-dependent activation of neurons (see Fig. 4)), which are observed in the cross-correlations obtained from cortical recordings by Fujisawa and collaborators.

Despite its usefulness as an analytical tool, cross-correlation analysis is unable to account for network effects in a controlled way. That is, simply by treating each pair independently it is not possible to determine whether the correlated activity of a single pair of neurons is due to a direct interaction between them, or whether it results from indirect interactions mediated by other neurons in the network. In the following section

we introduce an artificial Ising network, inferred using the spiking probabilities $p_i(\Delta t)$ and pairwise correlations $p_{ij}(\Delta t)$ on a fixed time scale Δt, which goes in this direction by reproducing neural activity observed in the data beyond that which was used to infer the model. This approach was introduced to analyze retinal data in [5, 6]. In Section 7 we clarify the relationship between the couplings J_{ij} inferred by the Ising approach and the two-cell approximation $J_{ij}^{(2)}$ derived from cross-correlation analysis. The Ising model we apply in the following aims to go beyond cross-correlation analysis in the sense it that disentangles direct couplings from correlations, correcting for network effects.

1.3. Ising model encoding of the activity

Neural activity represented in the form of spike trains can also be encoded in a binary form. We first divide the total time interval of a recording of the activity into small time bins of size Δt. The activity is then represented by set of binary variables s_i^τ, hereafter called spins, where $\tau = 1, \ldots, B$ labels the time bin and $i = 1, \ldots, N$ is a label which refers to a particular neuron. In the data we consider N spans a few decades (30 to 120). If in bin τ neuron i has spiked at least once then we set $s_i^\tau = 1$, otherwise $s_i^\tau = 0$.

We write the probability of an N-spin configuration $P_N[\mathbf{s}]$ in the data is the fraction of bins carrying that configuration $\mathbf{s} = \{s_1, s_2, \ldots, s_N\}$, and define the empirical Hamiltonian $H_N[\mathbf{s}]$ as minus the logarithm of P_N. By definition the Gibbs measure corresponding to this Hamiltonian reproduces the data exactly. As the spins are binary variables H_N can be expanded in full generality as

$$H_N[\mathbf{s}] = -\sum_{k=0}^{N} \sum_{i_1<i_2<\ldots<i_p} J^{(k)}_{i_1,i_2,\ldots,i_p}\, s_{i_1}\, s_{i_2}\, \ldots\, s_{i_p}, \tag{3}$$

where $J^{(0)}$ is simply a constant ensuring the normalization of P_N. Notice that, given an experimental multielectrode recording, P_N and H_N depend on the binning interval Δt. Knowledge of the 2^N coefficients J in (3) is equivalent to specifying the probability of each one of the spin configurations. The model defined by (3) provides a complete, but flawed, representation of the experimental data. First, it gives a poor compression of the data, requiring a number of variables which grows exponentially with the system size to exhaustively encode the probability of every configuration of the system. As a predictive model of the future behavior of the experimental system, it will also likely perform poorly compared to other reasonable models due to the overfitting of the data.

It is tempting then to look for simpler approximate expressions of H_N that retain most of the statistical structure of the spin configurations. A sensible approximation to the true interaction model should at least reproduce the values of the empirical one- and two-point correlation functions,

$$p_i = \frac{1}{B}\sum_{\tau=1}^{B} s_i^\tau\,, \quad p_{ij} = \frac{1}{B}\sum_{\tau=1}^{B} s_i^\tau\, s_j^\tau\,. \tag{4}$$

These constraints can be satisfied by the Ising model defined by the Hamiltonian

$$H_2[\mathbf{s}] = -\sum_{i=1}^{N} h_i\, s_i - \sum_{i<j} J_{ij}\, s_i\, s_j\,. \tag{5}$$

The corresponding probability of a configuration in the Ising model is given by the standard equilibrium Gibbs measure,

$$P[\mathbf{s}] = \frac{e^{-H_2[\mathbf{s}]}}{Z[\{h_i\},\{J_{ij}\}]}, \tag{6}$$

where the partition function $Z[\{h_i\},\{J_{ij}\}] = \sum_{\mathbf{s}}\, e^{-H_2[\mathbf{s}]}$ ensures that $\sum_{\mathbf{s}} P[\mathbf{s}] = 1$.

The local fields $\{h_i\}$ and pairwise couplings $\{J_{ij}\}$ are the natural parameters for encoding the one- and two-point correlations because they are the conjugate thermodynamical variables to these correlations. Indeed the $N\,(N+1)/2$ coupled equations (4) can be solved by finding the couplings and fields which minimize the cross entropy between the inferred Ising model and the data

$$S^*[\{h_i\},\{J_{ij}\}] = \log Z[\{h_i\},\{J_{ij}\}] - \sum_i h_i\, p_i - \sum_{i<j} J_{ij}\, p_{ij}\,. \tag{7}$$

This can be verified by computing the derivatives of $S^*[\{h_i\},\{J_{i,j}\}]$ with respect to the couplings and fields,

$$\frac{\partial S^*[\{h_i\},\{J_{ij}\}]}{\partial h_i} = \frac{1}{Z}\sum_{\mathbf{s}} s_i e^{-H_2[\mathbf{s}]} - p_i, \tag{8}$$

$$\frac{\partial S^*[\{h_i\},\{J_{ij}\}]}{\partial J_{ij}} = \frac{1}{Z}\sum_{\mathbf{s}} s_i\, s_j e^{-H_2[\mathbf{s}]} - p_{ij}. \tag{9}$$

At the minimum of (7) the derivatives must vanish, implying that the conditions (4) are satisfied. Moreover the minimal cross entropy $S_2[\{p_i\},\{p_{ij}\}] = \min_{\{h_i\},\{J_{ij}\}} S^*[\{h_i\},\{J_{i,j}\}]$ is the Legendre transform of the free energy $F = -\log Z[\{h_i\},\{J_{ij}\}]$ of the Ising model. It can be shown (see [2]) that the Ising model (6) is the probabilistic model with the maximum entropy

$$S[P] = -\sum_{\mathbf{s}} P[\mathbf{s}] \ln P[\mathbf{s}] \tag{10}$$

which satisfies the constraints that the average value of the one- and two-point correlations coincide with the observed data (4) and that the probability distribution $P[\mathbf{s}]$ is normalized such that all probabilities sum to one.

1.4. Previous results of the inverse Ising approach

It was put forward in [5, 37] that the Ising model defined by (6) not only reproduces the one- and two-point firing correlations extracted from the data, it also captures other features of the data such as higher order correlations and multi-neuron firing frequencies. Thus the energy H_2 provides a good approximation to the whole energy H_N; interactions involving three or more spins in (3) are not quantitatively important, and it suffices to keep in the expansion local fields and pair interactions only. Further simplification,

however, is not generally possible. Hamiltonians H_1 with no multi-spin interactions, corresponding to setting all $J_{ij} = 0$ in (5) with the fields h_i chosen to reproduce the average neural activity, provide a very poor approximation to (3). Such independent neuron approximations fail to reproduce, for example, the probability that k cells spike together in the same bin for $k = 1, \ldots, N$.

The claim in [5, 37] is based on the following points:

- Information-theoretical arguments: call S_1 the entropy of the independent-neuron model H_1, S_2 the entropy of the Ising model (5), and S_N the entropy of the experimental distribution P_N. Then the reduction in entropy of spin configurations coming from pair correlations, $I_2 = S_1 - S_2$, explains much of the reduction in entropy coming from correlations at all orders (or multi-information), $I_N = S_1 - S_N$: $I_2/I_N \simeq 90\%$ [5]. Hence higher order correlations do not contribute much to the multi-information, implying that most of the correlation observed in the data is explained simply through pairwise interactions.
- The multi-neuron firing frequencies, *i.e.* the probability of k neurons spiking together in the same bin, predicted by the Ising model with pairwise interactions are much closer to the experimental values than those of its independent-cell counterpart.
- Higher order correlations (three spin and higher correlations) predicted by the Ising model are in good agreement with the experimental data.

2. Statistical effects of finite sampling

Finite sampling of the experimental systems we consider here introduces fluctuations which complicate the inverse problem. In this Section we discuss the effects of finite sampling noise on the empirical one- and two-point correlations, consequences for the inference procedure, and statistical errors on the inferred couplings and fields.

2.1. Statistical error on empirical correlations

The inverse Ising approach requires measurements of the probability that each cell spikes in a time bin $p_i(\Delta t)$ and that each pair of cells spike in the same time bin $p_{ij}(\Delta t)$. These probabilities can be obtained from the spike train data by counting the number of time windows in which a cell, or two cells for the pair correlations, are active divided by the total number of time windows B. It is important to notice that the number of sampled configurations, while large (typically the total recording time is of the order of one hour, which with a time bin of 20 ms corresponds to roughly 10^6 configurations), is limited. Therefore the empirical correlations $\{p_i\}, \{p_{ij}\}$ we obtain will differ from the ones which would be obtained with an infinite sampling $\{p_i^{\rm true}\}, \{p_{ij}^{\rm true}\}$ by typical fluctuations of the order of $\{\delta p_i\}, \{\delta p_{ij}\}$ which we estimate in the following.

In our model, the spins are stochastic variables with Gibbs average

$$\langle s_i \rangle = p_i^{\rm true} \tag{11}$$

and with variance

$$\sigma_i = \langle s_i^2 \rangle - \langle s_i \rangle^2 = p_i^{\rm true}(1 - p_i^{\rm true}). \tag{12}$$

Their average p_i over B independent samples is also a stochastic variable with the same average as above, but with the standard deviation

$$\delta p_i = \sqrt{\frac{p_i^{\rm true}(1 - p_i^{\rm true})}{B}} \simeq \sqrt{\frac{p_i(1 - p_i)}{B}}, \tag{13}$$

where in the last expression of (13) we have replaced the Gibbs average $p_i^{\rm true}$ with the empirical average, neglecting terms of the order $1/B^2$. Similarly the pair correlations are have the average value

$$\langle s_i \, s_j \rangle = p_{ij}^{\rm true}, \tag{14}$$

and variance

$$\sigma_{ij} = p_{ij}^{\rm true}(1 - p_{ij}^{\rm true}), \tag{15}$$

therefore their sampled average over B samples p_{ij} has a standard deviation

$$\delta p_{ij} \simeq \sqrt{\frac{p_{ij}(1 - p_{ij})}{B}} \, . \tag{16}$$

2.2. The Hessian of the cross entropy: The importance of regularization terms

Because of finite sampling problems the fields and couplings obtained as minimizers of (7) are not necessarily well defined. As a simple example consider a single neuron ($N = 1$) with an average spiking rate r. The probability that the corresponding spin in the binary representation is zero is $\exp(-r \, \Delta t) \simeq 1 - r \, \Delta t$, which is close to one for small bin widths Δt. If the number B of bins is much smaller than $1/(r \, \Delta t)$ the spin is likely to be equal to zero in all configurations. Hence $p = 0$ and the field solution of the inverse problem is $h = -\infty$. In practice the number B of available data is large enough to avoid this difficulty as far as single site probabilities are concerned. However, when two or more spins are considered and higher-order correlations are taken into account incomplete sampling cannot be avoided; some groups of cells are never active simultaneously, which lead to infinite couplings.

This problem is easily cured with a Bayesian approach. We now start with a set of data consisting of B configurations of spins $\{\mathbf{s}^\tau\}$, $\tau = 1, \ldots, B$, and assume that the data come from the Ising model with Hamiltonian (5). The probability or *likelihood* of a spin configuration is given by (6). The likelihood of the whole set of data is therefore

$$\mathcal{P}^{\rm like}[\{\mathbf{s}^\tau\}|\{h_i, J_{kl}\}] = \prod_{\tau=1}^{B} P[\mathbf{s}^\tau|\{h_i, J_{kl}\}] = \exp\left(-B \, S^*[\{h_i\}, \{J_{ij}\}]\right), \tag{17}$$

where the cross entropy S^* is defined in (7).

In order to calculate the *a posteriori* probability of the couplings and fields we need to introduce a prior probability over those parameters. In the absence of any information

(data) let us assume that the couplings have a Gaussian distribution, with zero mean and variance σ^2,

$$\mathcal{P}^{\rm prior}[\{h_i\},\{J_{ij}\}] = (2\pi\sigma^2)^{-N(N+1)/2} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i<j} J_{ij}^2\right)\right] . \tag{18}$$

So far the variance is an unknown parameter. We expect it to be of the order of unity, that is, much smaller than the number of configurations B in the data set. Multiplying (17) and (18) leads to the Bayesian *a posteriori* probability for the fields and couplings given the data,

$$\mathcal{P}^{\rm post}[\{h_i\},\{J_{ij}\}|\{\mathbf{s}^\tau\}] = \frac{\mathcal{P}^{\rm prior}[\{h_i\},\{J_{ij}\}] \times \mathcal{P}^{\rm like}[\{\mathbf{s}^\tau\}|\{h_i\},\{J_{kl}\}]}{\mathcal{P}[\{\mathbf{s}^\tau\}]} , \tag{19}$$

where $\mathcal{P}[\{\mathbf{s}^\tau\}]$ is a normalization constant which is independent of the couplings and fields. The most likely values for the fields and couplings are obtained by maximizing (19), that is, by minimizing the modified cross entropy,

$$S^*[\{h_i\},\{J_{ij}\},\gamma] = S^*[\{h_i\},\{J_{ij}\}] + \frac{\gamma}{2}\left(\sum_{i<j} J_{ij}^2\right) , \tag{20}$$

where

$$\gamma = \frac{1}{B\,\sigma^2} . \tag{21}$$

The value of the variance σ^2 can be determined, once again, using a Bayesian criterion. $\mathcal{P}[\{\mathbf{s}^\tau\}]$ in (19) is the *a priori* probability of the data set over all possible Ising models. The optimal variance is the one maximizing this quantity. The approach is illustrated in detail in the simplest case of a single neuron, then extended to the case of interacting neurons in Appendix A.

The precise form of the regularization term is somewhat arbitrary. For convenience in the calculations we use the regularization

$$\gamma \sum_{i<j} p_i(1-p_i)p_j(1-p_j)\, J_{ij}^2 . \tag{22}$$

This choice corresponds to a Gaussian prior for the interactions in a Bayesian framework. Another regularization term we use is based on L_1 norm rather than L_2, and favors sparse coupling networks, *i.e.* with many couplings equal to zero,

$$\gamma \sum_{i<j} \sqrt{p_i(1-p_i)p_j(1-p_j)}|J_{ij}| , \tag{23}$$

which corresponds to a Laplacian prior distribution. In Section 7 we describe the properties of the L_1 and L_2 norm penalties in more detail, and in Section 4.1 we motivate the p_i, p_ij dependence of the terms in the regularization.

When the number of samples B in the data set becomes large, (21) shows that the regularization strength $\gamma \to 0$; the couplings and fields determined by minimizing (7) always coincide with Bayes predictions in the perfect sampling limit. For finite B the

presence of a quadratic contribution in (20) ensures that S^* grows rapidly far from the origin along any direction in the $\frac{N(N+1)}{2}$-dimensional space of fields and couplings.

Thanks to the regularization there exists a unique and finite minimizer of the cross entropy. The Hessian matrix χ of the cross entropy S^*, also called the Fisher information matrix, is given by

$$\begin{aligned}\chi &= \begin{pmatrix} \frac{\partial^2 S^*}{\partial h_i \partial h_{i'}} & \frac{\partial^2 S^*}{\partial h_i \partial J_{k'l'}} \\ \frac{\partial^2 S^*}{\partial h_{i'} \partial J_{kl}} & \frac{\partial^2 S^*}{\partial J_{kl} \partial J_{k'l'}} \end{pmatrix} \\ &= \left(\begin{array}{c|c} \langle s_i\, s_{i'}\rangle - \langle s_i\rangle\langle s_{i'}\rangle & \langle s_i\, s_{k'}\, s_{l'}\rangle - \langle s_i\rangle\langle s_{k'}\, s_{l'}\rangle \\ \hline \langle s_{i'}\, s_k\, s_l\rangle - \langle s_{i'}\rangle\langle s_k\, s_l\rangle & \langle s_k\, s_l\, s_{k'}\, s_{l'}\rangle - \langle s_k\, s_l\rangle\langle s_{k'}\, s_{l'}\rangle \end{array} \right), \end{aligned} \tag{24}$$

where $\langle\cdot\rangle$ denotes the Gibbs average with measure (6), and $i = 1, \ldots, N$, $1 \le k < l \le N$. The addition of a regularization term with a quadratic penalty on the couplings and fields adds γ times the identity matrix to the Hessian, and as the unregularized χ is a covariance matrix and hence nonnegative‡, this assures that the susceptibility of the regularized model is positive definite. Thus $S^*[\{h_i\}, \{J_{ij}\}, \gamma]$ is strictly convex. This proves the existence and uniqueness of the solution to the inverse problem.

2.3. Statistical error on couplings and fields

In this Section we compute the statistical fluctuations of the inferred couplings and fields due to finite sampling of the experimental system.

As in Section 2.1, let us assume that data are not extracted from experiments but rather generated from the Ising model (5) with the fields $\{h_i^*\}$ and couplings $\{J_{ij}^*\}$ which minimize the cross entropy (7). From Section 2.2 the probability of infering a set of fields $\{h_i\}$ and couplings $\{J_{ij}\}$ is proportional to $\exp(-B\, S^*[\{h_i\}, \{J_{ij}\}])$. When B is very large this probability is tightly concentrated around the minimum of S^*, that is, $\{h_i^*\}, \{J_{ij}^*\}$. The difference between the inferred and the true fields and couplings is encoded in the $\frac{N(N+1)}{2}$-dimensional vector $\vec{\Delta}_{h,J}$ of components $\{h_i - h_i^*\}, \{J_{ij} - J_{ij}^*\}$. The distribution of this vector is asymptotically Gaussian,

$$\mathcal{P}[\vec{\Delta}_{h,J}] \simeq \frac{\sqrt{\det \chi}}{(2\pi B)^{N(N+1)/4}} \exp\left(-\frac{B}{2}\, \vec{\Delta}_{h,J}^{\dagger} \cdot \chi \cdot \vec{\Delta}_{h,J} \right). \tag{25}$$

Statistical fluctuations of the couplings and fields are therefore characterized by the standard deviations

$$\delta h_i = \sqrt{\frac{1}{B}\,(\chi^{-1})_{i,i}}\,, \quad \delta J_{ij} = \sqrt{\frac{1}{B}\,(\chi^{-1})_{ij,ij}}\,. \tag{26}$$

‡ Let $\vec{v} = (\vec{x}, \vec{y})$ where $\vec{x} = (x_1, x_2, \ldots, x_N)$ and $\vec{y} = (y_{12}, y_{13}, \ldots, y_{N-1,N})$ are, respectively N- and $\frac{N(N-1)}{2}$-dimensional vectors. Then

$$\vec{v}^{\dagger} \cdot \mathbf{H} \cdot \vec{v} = \langle \left[\sum_i x_i \big(s_i - \langle s_i\rangle\big) + \sum_{k<l} y_{kl} \big(s_k\, s_l - \langle s_k\, s_l\rangle\big) \right]^2 \rangle \ge 0$$

for any $\vec{v}$.

Note that (13) and (16) can be also directly obtained from the covariance matrix χ. Indeed linear response theory tells us that

$$\vec{\Delta}_{p,c} = \chi \cdot \vec{\Delta}_{h,J} \,. \tag{27}$$

We deduce from (25) that $\vec{\Delta}_{p,c}$ obeys a Gaussian distribution as $\vec{\Delta}_{h,J}$, with the Hessian matrix χ replaced with its inverse matrix in (25). The typical uncertainties of the one- and two-point correlations are given by $\delta p_i = \sqrt{\frac{1}{B}(\chi)_{i,i}}$ and $\delta p_{ij} = \sqrt{\frac{1}{B}(\chi)_{ij,ij}}$, which correspond exactly with (13) and (16).

3. The difficulty of the inverse Ising problem

The inverse problem can be subdivided in three different problems which are of increasing difficulty [2].

Reconstruction of the interaction graph

It is possible to reproduce qualitative features of the underlying interaction graph (*i.e.* the set of interactions which would be inferred if the inverse problem was solved exactly), such as the number of connections to a given cell or the range of interactions, without determining the values of the couplings precisely. Knowing the rank of the couplings in absolute value, for instance, gives an idea of the structure of the interaction graph. We will see in Section 7 that mean field approximations are only useful for finding precisely the network of interactions in the case of weak interactions and not, as in the majority of the analyzed neural data, for a dilute network with large couplings. These approximations generally give, then, wrong values for the couplings but respect the ranking of large couplings by magnitude therefore allowing for the accurate recovery of the structure of the interaction graph of large couplings.

Reconstruction of the interaction graph and values of the couplings

It can be of interest to infer the values of the couplings more precisely, and to identify *reliable* couplings, which are different from zero even when taking into account statistical fluctuations due to finite sampling (26), and *unreliable* couplings which are compatible with zero. A precise inference of the value of the couplings compared with the error bar will be helpful to characterize changes in the values of couplings, which can occur for example in a memory task [24]. Also it would be of interest to characterize reliable negative couplings with the aim of identifying inhibitory connections, which are particularly susceptible to sampling noise and thus difficult to assign reliably.

It has been pointed out in [2] that the difficulty of the inference problem depends on the inverse of the Fisher information matrix χ^{-1} (24). Whereas the susceptibility χ characterizes the response of the Ising model to a small change in the couplings or fields, the inverse susceptibility χ^{-1} gives the change in the inferred couplings and fields due to a perturbation of the observed frequencies or correlations. If data are generated by sparse networks and if the sampling is good, then χ^{-1} is localized, *i.e.* it decays fast with the length of the interaction path between spins in the interaction network. This

property holds even in the presence of long-range correlations and makes the inverse problem not intrinsically hard.

It is important to notice that the sampling fluctuations in frequencies (13) and correlations (16) do not come from a sparse network structure and thus χ^{-1} can lose its localized properties because of sampling noise. In other words, overfitting can generate very complicated and densely connected networks which will have a large couplings-susceptibility. A way to filter the data must then be found to efficiently solve the inverse problem in presence of sampling noise. The selective cluster expansion, which we introduce in Section 5, approaches this problem by including methods of regularization and by building up a solution to the inference problem in a progressive way, including only terms which contribute significantly to the inference so as to avoid the overfitting of noise.

Reconstruction of the empirical correlations

The most difficult inference task is to find the local fields and the network of couplings which truly satisfy the minimum of (7), and which therefore reproduce the frequencies and correlations obtained from data. One can use such an inferred Ising model to then predict, for example, the observed higher moments of the distributions (three body or higher order correlations) or multi-neuron firing frequencies, as discussed in Section 1.4. The Ising model could also be used to predict the response to a local perturbation, which could be due to a stimulation, or the ablation of a connection.

The reconstruction problem is difficult because a small change in the inferred parameters can result in large changes in the correlations and frequencies given by the inferred model. In other words, the susceptibility matrix may be dense, rather than sparse, with many large entries. Let again $\vec{\Delta}_{p,c}$ denote the $\frac{N(N+1)}{2}$-dimensional vector whose components are the differences between the predicted and true values of the frequencies and correlations. Typically inference errors come from the approximations done in the resolution of the inverse problem. Using the linear response theory of (27) we can estimate the order of magnitude of the reconstruction error due to an inference error:

$$\vec{\Delta}_{p,c} \approx |\chi| \vec{\Delta}_{h,J}, \tag{28}$$

where $|\chi|$ is the norm of the matrix defined as the sum over the rows of the maximum element over the columns,

$$|\chi| = \sum_i \max_j \chi_{i,j} \,. \tag{29}$$

Let us imagine that a parameter h_i has been inferred with an inference error of the order $\Delta h \simeq 0.1$, which is a typical order of magnitude of the expected statistical fluctuations in the inferred parameters. This inference error can drastically change the reconstructed p_i. For example for the recording of $N = 40$ neurons in the retina subject to a natural movie stimulus [5] $|\chi| = 0.6$, and the consequent $\delta_p \approx 0.06$ is of the same order of magnitude as the p_i, which is much larger than the statistical errors we expect

on these variables. We discuss in Section 6 how the reconstruction behaves as a function of the inference precision in the cluster expansion. This example explains why even if the network is reconstructed at the precision corresponding to the statistical fluctuations (26) we expect from the finite sampling, and therefore the inverse problem is quite accurately solved, the statistical properties of the data on the inferred model can be much harder to reproduce and may require sophisticated expansion and fine tuning of some parameters.

Note that equation (28) gives only an order of magnitude, the more precise matrix dependence of the response in (27) tells us that the modes which correspond to large eigenvalues of the Hessian matrix χ have to be more precisely determined to solve the reconstruction problem well, while the others, which correspond to small eigenvalues of the Hessian matrix, have less impact on the reconstruction problem.

4. High temperature expansions and the mean field entropy $S_{\rm MF}$

The main problem of the minimization of the cross entropy S^* (7) is the calculation of the partition function Z which, if done exactly, requires the sum over all 2^N possible configurations of the system of N spins. Because this sum becomes prohibitive for systems with more than ≈ 20 spins some approximate solution of the inverse problem must be found.

High temperature expansions [16, 38–40] of the Legendre transform of the free energy are useful when the pairs of variables interact through weak couplings. The Ising model with weak (of the order of $N^{-1/2}$) interactions is the so-called Sherrington-Kirkpatrick model. In this case the entropy $S[\{p_i\},\{p_{ij}\}]$ coincides asymptotically for large N with

$$S_{\rm MF}[\{p_i\},\{p_{ij}\}] = S_{\rm ind}[\{p_i\}] + \frac{1}{2}\log\det M[\{p_i\},\{p_{ij}\}], \tag{30}$$

where

$$S_{\rm ind}[\{p_i\}] = \sum_i \left[-p_i\,\log p_i - (1-p_i)\,\log(1-p_i)\right] \tag{31}$$

is the entropy of independent spin variables with averages $\{p_i\}$, and

$$M_{ij}[\{p_i\},\{p_{ij}\}] = \frac{p_{ij}-p_ip_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}, \tag{32}$$

which can be calculated in $O(N^3)$ time [16, 41], and is consistent with the so-called TAP equations [42]. It is important to underline that what we call the mean field entropy $S_{\rm MF}$ corresponds to the entropy of a Gaussian model of continous spin variables with averages p_i and variances $p_i\,(1-p_i)$.

The derivatives of $S_{\rm MF}$ with respect to the $\{p_{ij}\}$ and $\{p_i\}$ give the value of the couplings and fields,

$$(J_{\rm MF})_{ij} = -\frac{\partial S_{\rm MF}}{\partial p_{ij}} = -\frac{(M^{-1})_{ij}}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}\ ,$$

$$(h_{\rm MF})_i \;=\; -\,\frac{\partial S_{\rm MF}}{\partial p_i} = \sum_{j(\neq i)} (J_{MF})_{ij} \left(c_{ij} \frac{p_i - \frac{1}{2}}{p_i(1-p_i)} - p_j \right) \;, \tag{33}$$

where $c_{ij} = p_{ij} - p_i p_j$ is the connected correlation.

It is possible to get an idea of the goodness of the high temperature expansion from the size of the correlation indices $CI_{ij} = p_{ij}/(p_i p_j)$, which are the expansion parameters of the high temperature series [16]. We note that in neural data, even if the connected correlations are small, the correlation indices CI_{ij} can be large because the probabilities p_i, p_j that individual cells are active in a time bin are also generally small, and CI_{ij} is of order one. Histograms of the correlation indices for the different data sets we analyze are shown in Figs. 13 and 14, and show that the correlation indices are small in absolute value ($|CI_{ij}| \leq 2.2$) only in the data sets of *in vivo* recordings of the medial prefrontal cortex of rats with tetrodes [23, 32]. We therefore expect the high temperature expansion and in particular $S_{\rm MF}$ to be a good approximation to the inverse problem only for this data set.

4.1. L_2-Regularized mean field entropy

A regularized version of the mean field entropy can be computed analytically [2]. The derivation is as follows. First we use the mean field expression for the cross-entropy at fixed couplings J_{ij} and frequencies p_i, see [39], to rewrite

$$S_{\rm Ising}[\{p_i\},\{J_{ij}\}] = S_{\rm ind}[\{p_i\}] - \frac{1}{2}\log\det\left({\rm Id} - J'\right) - \sum_{i<j} J_{ij}\,(p_{ij} - p_i\,p_j), \tag{34}$$

where $J'_{ij} = J_{ij}\sqrt{p_i(1-p_i)p_j(1-p_j)}$ and Id denotes the N-dimensional identity matrix. We consider the L_2-norm regularization (22). The entropy at fixed data $\{p_i\}$, $\{p_{ij}\}$ is

$$\begin{aligned} S^{L_2}_{\rm MF}[\{p_i\},\{p_{ij}\}] &= S_{\rm ind}[\{p_i\}] + \min_{\{J_{ij}\}} \Big[S_{\rm Ising}[\{p_i\},\{J_{ij}\}] \\ &\qquad + \gamma \sum_{i<j} J_{ij}^2\, p_i(1-p_i)p_j(1-p_j) \Big] \\ &= S_{\rm ind}[\{p_i\}] + \min_{\{J'_{ij}\}} \Big[-\frac{1}{2}\log\det\left({\rm Id} - J'\right) \\ &\qquad - \frac{1}{2}{\rm tr}\left(J' \cdot M[\{p_i\},\{p_{ij}\}]\right) + \frac{\gamma}{2}{\rm tr}\left(J'\right)^2 \Big], \end{aligned} \tag{35}$$

where $M[\{p_i\},\{p_{ij}\}]$ is defined in (32). The optimal interaction matrix J' is the root of the equation

$$\left({\rm Id} - J'\right)^{-1} - M[\{p_i\},\{p_{ij}\}] + \gamma\, J' = 0. \tag{36}$$

Hence, J' has the same eigenvectors as $M[\{p_i\},\{p_{ij}\}]$, a consequence of the dependence on p_i we have chosen for the quadratic regularization term in (22). Let j_q denote the q^{th} eigenvalue of J'. Then

$$S^{L_2}_{\rm MF}[\{p_i\},\{p_{ij}\},\gamma] = S_{\rm ind}[\{p_i\}] + \frac{1}{2}\sum_{q=1}^{N} \left(\log\hat{m}_q + 1 - \hat{m}_q\right), \tag{37}$$

where $\hat{m}_q$ is the largest root of $\hat{m}_q^2 - \hat{m}_q(m_q - \gamma) = \gamma$, and m_q is the q^{th} eigenvalue of $M[\{p_i\}, \{p_{ij}\}]$. Note that $\hat{m}_q = m_q$ when $\gamma = 0$, as expected.

4.2. L_1-Regularized mean field entropy

With the choice of an L_1-norm regularization, the entropy at fixed data $\mathbf{p}$ becomes

$$S_{\rm MF}^{L_1}[\{p_i\}, \{p_{ij}\}, \gamma] = S_{\rm ind}[\{p_i\}] + \min_{\{J'_{ij}\}} \Big[-\frac{1}{2} \log\det\left(\mathrm{Id} - J'\right) - \frac{1}{2}\mathrm{tr}\left(J' \cdot M[\{p_i\}, \{p_{ij}\}]\right) + \gamma \sum_{i<j} |J'_{ij}| \Big]. \quad (38)$$

No analytical expression exists for the optimal J'. However, it can be found in a polynomial time using convex optimization techniques. The minimization of (38) is known in the statistics literature as the estimate of the precision matrix with a constraint of sparsity. Several numerical procedures to compute J' efficiently are available [43].

5. Selective Cluster Expansion

When an exact calculation of the partition function is out of reach, and the amplitudes of the correlation indices are large, an accurate estimate of the interactions which solve the inverse Ising problem can be obtained through cluster expansions. Cluster expansions have a rich history in statistical mechanics, *e.g.* the virial expansion in the theory of liquids or cluster variational methods [44].

The selective cluster expansion algorithm (SCE) which we have recently proposed [1, 2] makes use of an assumption about the properties of the inverse susceptibility χ^{-1} to efficiently generate an approximate solution to the inverse Ising inference problem. χ^{-1} is typically much sparser and shorter range than χ, implying that most interactions inferred from a given set of data depend strongly on only a small set of correlations. Thus an estimate of the interactions for the entire system can be constructed by solving the inference problem on small subsets (clusters) of spins and combining the results. The SCE gives such an estimate by recursively solving the inverse Ising problem on small clusters of spins, selecting the clusters which give significant information about the underlying interaction graph according to their contribution to the entropy S of the inferred Ising model, and building from these a new set of clusters to analyze. At the end of this cluster expansion procedure, an estimate of the entropy and of the interactions for the full system is produced based on the values inferred on each cluster.

Let S_Γ denote the entropy of the Ising model defined just on a subset $\Gamma = \{i_1, i_2, \ldots\}$ of the full set of spins, which reproduces the one- and two-point correlations obtained from data for this subset. The entropy S of the inferred Ising model on the full system of N spins can then be expanded formally as a sum of individual contributions from each of the $2^N - 1$ nonempty subsets,

$$S = \sum_\Gamma \Delta S_\Gamma, \qquad \Delta S_\Gamma = S_\Gamma - \sum_{\Gamma' \subset \Gamma} \Delta S_{\Gamma'}\,. \quad (39)$$

The cluster entropy ΔS_Γ, defined recursively in (39), measures the contribution of the cluster Γ to the total entropy. It is calculated by substracting the cluster entropies of all possible smaller subsets of Γ from the subset entropy S_Γ. Each cluster entropy depends only upon the correlations between the spins in that cluster, and for small clusters it is easy to compute numerically. The contribution of each cluster to the interactions, which we denote $\Delta \mathbf{J}_\Gamma$, is defined analogously and computed at the same time as the cluster entropy. The elements of $\mathbf{J}_\Gamma$ are the couplings and fields which minimize the cross entropy S^* restricted to the cluster Γ. Note that in applications to real data, we employ some form of regularization (22), (23), to ensure that the Hessian χ is positive definite and to reduce the effects of sampling noise.

As an example, the entropy of a single-spin cluster is, using (39) with $N = 1$,

$$\Delta S_{(i)}[p_i] \equiv S_{\rm ind}[p_i] = -p_i \log p_i - (1-p_i)\log(1-p_i). \tag{40}$$

The contribution to the field of a single-spin cluster is

$$\Delta h_i^{(1)} \equiv h_i^{(1)} \equiv -\frac{\partial S_{(i)}}{\partial p_i} = \log\left(p_i/(1-p_i)\right). \tag{41}$$

The entropy of a two spin subsystem $\{i, j\}$ is

$$\begin{aligned} S^{(2)}[p_i, p_j, p_{ij}] = & -p_{ij}\log p_{ij} - (p_i - p_{ij})\,\log\left(p_i - p_{ij}\right) \\ & - (p_j - p_{ij})\,\log\left(p_j - p_{ij}\right) \\ & - (1 - p_i - p_j + p_{ij})\,\log\left(1 - p_i - p_j + p_{ij}\right) \end{aligned}$$

Using again (39) with $N = 2$, we obtain $\Delta S_{(i,j)}[p_i, p_j, p_{ij}] = S_{(i,j)}[p_i, p_j, p_{ij}] - \Delta S_{(i)}[p_i] - \Delta S_{(j)}[p_j]$, which measures the loss in entropy when imposing the constraint $\langle s_i\, s_j\rangle = p_{ij}$ to a system of two spins with fixed magnetizations, $\langle s_i\rangle = p_i$, $\langle s_j\rangle = p_j$. The contribution to the field of the two spin cluster is

$$\Delta h_i^{(2)} = \log\left((p_i - p_{ij})/(1 - p_i - p_j + p_{ij})\right) - \Delta h_i^{(1)}, \tag{42}$$

and the contribution to the coupling is

$$\begin{aligned} \Delta J_{ij}^{(2)} &\equiv J_{ij}^{(2)} \\ &= \log p_{ij} - \log\left(p_i - p_{ij}\right) - \log\left(p_j - p_{ij}\right) + \log\left(1 - p_i - p_j + p_{ij}\right) \end{aligned} \tag{43}$$

Note that the two-cell couplings, which do not take into account network effects, reduce to the logarithm of the correlation index for $p_{ij} \ll p_i \ll 1$, which is typically the case when the time bin is much smaller than the typical interval between spikes. In this case $p_i \propto \Delta t$ and $p_{ij} \propto \Delta t^2$, thus

$$J_{ij}^{(2)} \approx \log \frac{p_{ij}}{p_i\, p_j}\,. \tag{44}$$

It is difficult to write the entropy analytically for clusters of more than two spins, but S_Γ can be computed numerically as the minimum of the cross entropy (7). This involves calculating Z, a sum over an exponential number of spin configurations. In practice this limits the size of the clusters which we can consider to those with $\lesssim 20$ spins.

A recursive use of (39), or the Möbius inversion formula, allows us to obtain ΔS_Γ for larger and larger clusters Γ (see also the pseudocode of Algorithm 1). It is important

to note that the form of ΔS is such that the sum of all cluster entropies over all possible subsets, including the whole set, of a cluster Γ is just the total entropy of the cluster S_Γ. The analogous result also holds for $\Delta\mathbf{J}$. This means in particular that by construction the sum over all possible clusters for the full system of N spins yields the exact entropy S and interactions $\mathbf{J}$.

It is also posible to perform an expansion in $S-S_0$, where S_0 is a "reference entropy" approximating S which we expand around. As with S, the reference entropy S_0 should be computable on all possible subsets of the system, and should depend only on the one- and two-point correlations of the spins in the subsets. Again, by the construction rule (39) and independent of the functional form of S_0, summing over all the clusters will give back the exact value of $S-S_0$ for the full system. We will show in applications to neural data (see Section 6) that it can be useful to consider an expansion $S-S_{\rm MF}$ rather than S alone, using the mean field result as a reference entropy S_0. In some cases the expansion of $S-S_{\rm MF}$ converges much faster than the expansion of S alone.

The cluster entropy measures a cluster's contribution to the total entropy, which could not be gained from its sub-clusters taken separately. Intuitively, we expect that clusters with small $|\Delta S_\Gamma|$ contribute little new information about the underlying interaction graph which is not revealed by any of their subsets. Indeed, it has been shown [1, 2] that small cluster entropies have a universal distribution, reflecting fluctuations in the experimentally observed correlations due to finite sampling (13), (16). Cluster entropies that are nonzero due to real interactions between the constituent spins also tend to decrease in magnitude as the cluster size becomes large, decaying exponentially in the size of the shortest closed interaction path between the spins in the cluster. In the SCE therefore all clusters which have $|\Delta S_\Gamma|$ smaller than a fixed threshold T are discarded. Selecting only those clusters which have cluster entropies larger than a chosen threshold helps to avoid the overfitting of noisy data.

Clearly, it is not possible to compute all of the 2^N-1 cluster entropies, corresponding to all the nonempty subsets of the full set of spins, even for rather small systems. To make the algorithm computationally feasible we must have a method for truncating the cluster expansion. This is implemented in the SCE by a recursive construction rule for the clusters included in the expansion (see the pseudocode of Algorithm 2). We begin with the computation of the cluster entropies for all N clusters of size $k=1$. The contribution of each cluster to the interactions is also recorded for later use. Each subsequent step follows the same pattern. First, clusters with $|\Delta S_\Gamma|<T$ are removed. We then include in the next step of the expansion all clusters which are unions of two of the remaining clusters of size k, $\Gamma'=\Gamma_1\cup\Gamma_2$, such that the new cluster Γ' contains $k+1$ spins.

The expansion naturally terminates when no more new clusters can be formed. This approach prevents a combinatorial explosion of the number of clusters considered in the expansion. It is also consistent with the idea of exploring paths of strong interactions in the interaction graph, as new clusters are built up from smaller clusters which have already been found to have significant interactions and which share many spins in

common. Final estimates for the entropy and the interactions are obtained by adding up all of the ΔS_Γ and $\Delta \mathbf{J}_\Gamma$. Contributions of S_0 and J_0 are also added if the reference entropy is used in the expansion.

5.1. Pseudocode of the cluster algorithm

In this section we present pseudo-codes useful for the practical implementation of the inference algorithm, following [1].

The principal routine of the cluster algorithm is the iterative computation of the cluster entropy, given in Algorithm 1. When the routine to compute the entropies

Algorithm 1 Computation of cluster-entropy ΔS_Γ

Require: Γ (of size K), $\{p_i\}$, $\{p_{ij}\}$, routines to calculate S_0 and S
  $\Delta S_\Gamma \leftarrow S_\Gamma - S_{0,\Gamma}$
  for SIZE $= K-1$ **to** 1 **do**
    for every Γ' with SIZE spins in Γ **do**
      $\Delta S_\Gamma \leftarrow \Delta S_\Gamma - \Delta S_{\Gamma'}$
    end for
  end for
Output: ΔS_Γ

of various clusters is called several times a substantial speed-up can be achieved by memorizing the entropies ΔS_Γ of every cluster. In Section 5.2 we will discuss how to calculate the subset entropy S_Γ in more detail.

In Section 6 we discuss the performance of the algorithm and compare the results in of the expansion in S alone, *i.e.* with no $S_{0,\Gamma}$, and with a mean field reference entropy (30) $S_{0,\Gamma} = S^{L_2}_{\mathrm{MF},\Gamma}$ obtained using an L_2 norm regularization (37) and $S_{0,\Gamma} = S^{L_1}_{\mathrm{MF},\Gamma}$ using an L_1 norm regularization (38) on the couplings $\{J_{ij}\}$.

The core of our inference algorithm is the recursive building-up and the selection of new clusters, described in Algorithm 2. The threshold T, which establishes which clusters will be kept in the expansion, is a parameter which is fixed in a run of the selective cluster algorithm. The choice of the optimal threshold T^* is discussed in the Section 5.3.

5.2. Numerical optimization of the cluster algorithm

In this section we review some of the computational challenges of the algorithm and numerical methods for running the algorithm efficiently.

The primary computational bottleneck of the selective cluster expansion algorithm is the repeated solution of the inverse Ising problem and the computation of the entropy on each K–spin system included in the expansion, which is used to calculate the cluster entropy (see Algorithm 1).

Algorithm 2 Adaptive Cluster Expansion

Require: N, T, S_0, routine to calculate ΔS_Γ from $\{p_i\}$, $\{p_{ij}\}$
  LIST $\leftarrow \emptyset$ {All selected clusters}
  SIZE $\leftarrow 1$
  LIST(1) $\leftarrow (1) \cup (2) \cup \ldots \cup (N)$ {Clusters of SIZE=1}
  repeat {Building-up of clusters with one more spin}
    LIST $\leftarrow$ LIST $\cup$ LIST(SIZE) {Store current clusters}
    LIST(SIZE+1) $\leftarrow \emptyset$
    for every pair $\Gamma_1, \Gamma_2 \in$ LIST(SIZE) **do**
      $\Gamma_I \leftarrow \Gamma_1 \cap \Gamma_2$ {Spins belonging to Γ_1 and to Γ_2}
      $\Gamma_U \leftarrow \Gamma_1 \cup \Gamma_2$ {Spins belonging to Γ_1 or to Γ_2}
      if Γ_I contains (SIZE-1) spins **and** $|\Delta S_{\Gamma_U}| > T$ **then**
        LIST(SIZE+1) $\leftarrow$ LIST(SIZE+1) $\cup$ Γ_U {add Γ_U to the list of selected clusters}
      end if
    end for
    SIZE $\leftarrow$ SIZE+1
  until LIST(SIZE) $= \emptyset$
  $S \leftarrow S_0$, $\mathbf{J} \leftarrow -\frac{d}{d\mathbf{p}} S_0$ {Calculation of S, $\mathbf{J}$}
  for $\Gamma \in$ LIST **do**
    $S \leftarrow S + \Delta S_\Gamma$, $\mathbf{J} \leftarrow \mathbf{J} - \frac{d}{d\mathbf{p}} \Delta S_\Gamma$
  end for
Output: S, $\mathbf{J}$ and LIST of clusters

Given a cluster Γ the partition function of the K–spin system restricted to Γ,

$$Z[\{h_i\}, \{J_{ij}\}, \Gamma] = \sum_{\{s_i=0,1;\, i\in\Gamma\}} \exp\left(\sum_{i\in\Gamma} h_i\, s_i + \sum_{i<j;\ i,j\in\Gamma} J_{ij}\, s_i\, s_j \right), \tag{45}$$

can be computed in time $\propto 2^K$. Then one has to find the most likely set of $\{J_{ij}\}$ and $\{h_i\}$ for each cluster given the experimental data, that is, we must solve the convex optimization problem

$$\min_{\{h_i\},\{J_{ij}\}} \left(S^*[\{h_i\}, \{J_{ij}\}, \Gamma] = \log Z[\{h_i\}, \{J_{ij}\}, \Gamma] - \sum_{i\in\Gamma} h_i\, p_i - \sum_{i<j;\, i,j\in\Gamma} J_{ij}\, p_{ij} \right). \tag{46}$$

No analytical solution exists for clusters of more than a few spins. As a single run of the cluster expansion algorithm may include thousands or even millions of clusters, it is critically important that the optimization problem (46) be solved as quickly as possible.

Given a starting value for the $\{J_{ij}\}$ and $\{h_i\}$, we employ a hybrid approach which combines the standard optimization techniques of gradient descent and Newton's method to step progressively closer to the minimum. Gradient descent steps are chosen along the direction of steepest descent, while Newton's method specifies a step direction towards the minimum of a local quadratic approximation of S^*. When far from the

minimum, we use gradient descent for its computational simplicity and numerical stability. Once the $\{J_{ij}\}$ and $\{h_i\}$ are determined to be close to the values which solve (46), we switch to Newton's method, which requires more computational resources but has a much better rate of convergence close to the minimum [45].

A careful choice of the initial conditions is also essential for obtaining a fast solution with minimal computational effort. We begin the optimization problem with an initial guess for $\{J_{ij}\}$ and $\{h_i\}$ based upon the assumption that the couplings and fields minimizing S^* will be similar to those that were found for smaller clusters containing the same sites. In many cases this initial guess works very well, and the optimization routine may find the minimum with just a single step. Just including this choice for the initial interactions can cut the total running time of the algorithm in half.

As described in Section 2.2, we may regularize the couplings $\{J_{ij}\}$ by adding a penalty term to S^* for each coupling which is nonzero. Regularization is useful for controlling spurious large couplings that arise from noise or undersampling of the experimental system, as well as ensuring the convexity of the Hessian χ, so that the inference problem has a unique minimum at a finite value of the couplings. Common choices of the penalty are based on the L_1-norm

$$\gamma \sum_{i<j} \sqrt{p_i(1-p_i)p_j(1-p_j)}|J_{ij}| \tag{47}$$

or the L_2-norm

$$\gamma \sum_{i<j} p_i(1-p_i)p_j(1-p_j)\, J_{ij}^2. \tag{48}$$

Such terms are natural in the context of Bayesian inference. The addition of (47) to S^* is equivalent to assuming a Laplacian prior distribution for the couplings, while the L_2-norm penalty (48) corresponds to a Gaussian prior. Typically we take $\gamma \approx 1/(10\, B\, (p(1-p))^2)$, where p is the average spiking frequency for a given data set.

Use of the L_1-norm penalty makes the optimization problem more difficult, as the function to be minimized is no longer smooth. In particular, in this case the gradient of S^* is undefined when any of the couplings $J_{ij} = 0$. To overcome the lack of differentiability of S^* we use a modified version of the projected scaled sub-gradient method of [46]. This method makes use of the sub-gradient, a generalization of the gradient which is well-defined even when some couplings are zero. It also allows couplings to be set exactly to zero during the step process, unlike a typical gradient descent or Newton's method step. Despite these additional complexities the optimization problem including the L_1-norm penalty can be solved with similar speed and accuracy as in the L_2-norm regularized or unregularized case.

5.3. Convergence and choice of the threshold T^*

In the following we describe how the practical procedure for applying the inference algorithm to a set of data. Because we do not know *a priori* the optimal value of the

threshold T^* we run the algorithm at different values of the threshold following the iterative heuristic below:

- Start with a large value of the threshold T, typically $T = 1$, at which only single-spin clusters are selected.
- Infer the fields and the couplings at that threshold.
- If the number of selected clusters has changed with respect to the previous value of the threshold, run a Monte Carlo simulation to check the reconstruction of the first and second moment. From the Monte Carlo simulation we obtain $\{p_i^{\rm rec}\}, \{p_{ij}^{\rm rec}\}$ and calculate the relative errors on the reconstructed averages and connected correlations $c_{ij}^{\rm rec} = p_{ij}^{\rm rec} - p_i^{\rm rec}\, p_j^{\rm rec}$ with respect to their statistical fluctuations due to finite sampling,

$$\epsilon_p = \left(\frac{1}{N} \sum_i \frac{(p_i^{\rm rec} - p_i)^2}{(\delta p_i)^2} \right)^{\frac{1}{2}}, \quad \epsilon_c = \left(\frac{2}{N\,(N-1)} \sum_{i<j} \frac{(c_{ij}^{\rm rec} - c_{ij})^2}{(\delta c_{ij})^2} \right)^{\frac{1}{2}}. \tag{49}$$

  The denominators in (49) measure the typical fluctuations of the data expected at thermal equilibrium, see (13), (16), and

$$\delta c_{ij} = \delta p_{ij} + p_i \delta p_j + p_j \delta p_i \,. \tag{50}$$

- Iterate this procedure by lowering the threshold T and stopping when the errors $\epsilon_p \simeq 1$, $\epsilon_c \simeq 1$. The corresponding value of the threshold is the optimal threshold T^*.
- Assign error bars to the inferred couplings and fields using (26).

Note that T^* is chosen as the first value of the threshold such that $\epsilon_p \simeq 1, \epsilon_c \simeq 1$ in order to reconstruct the data with the simplest possible network of interactions, *i.e.* that which has the smallest inverse susceptibility. Through the cluster expansion one "fills in" entries of the inverse susceptibility matrix $\chi^{-1}_{ijkl} = \frac{\partial J_{ij}}{\partial c_{kl}}$ of the inferred Ising model in a progressive way, starting with the largest contributions and stopping when the one- and two-point correlations are reconstructed within the uncertainty provided by the finite sampling fluctuations. By decreasing the threshold, the number of clusters which contribute to a given coupling increases, or in other words each inferred parameter $\{h_i\}, \{J_{ij}\}$ depends on increasingly more experimental correlations, including those which are poorly sampled. As a consequence, even if the underlying interactions which have generated the data have a simple structure and a corresponding small inverse susceptibility, poorly sampled correlations may lead to reconstructed networks with complicated structure and large inverse susceptibility (see Fig. 5D-E). This is especially true for small correlations, where the sampling fluctuations (13), (16) can be of a similar order of magnitude as the correlations themselves. Decreasing the threshold more than is necessary to fit the one- and two-point correlations is undesirable not only because the complicated structure of interactions due to overfitting will not necessarily correspond well with the underlying interaction network, but also because progressively lower values of the threshold increase the computational difficulty of the inference problem.

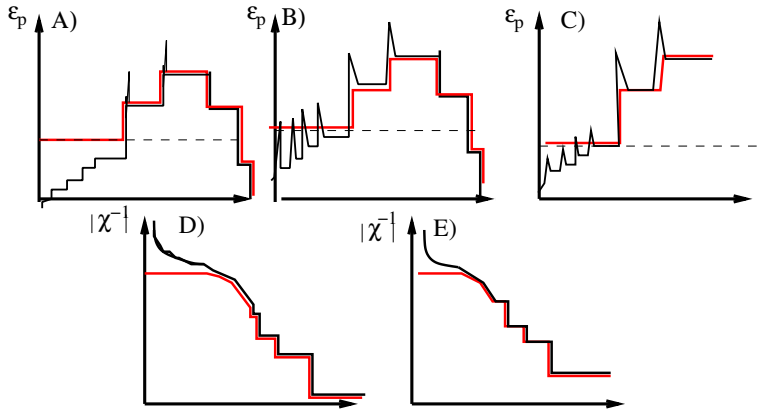

Figure 5. A) Typical pattern of the relative reconstruction error ϵ_p for small correlation lengths of the system, and without a reference entropy. At small values of the threshold the reconstruction error decreases monotonically because the entropy expansion is absolutly convergent. Note that without the reference entropy $\epsilon_p \sim 0$ at large values of the threshold because when only clusters of size one are included in the expansion, the fit is equivalent to that of an independent cell model and the experimental frequencies are matched perfectly. When $\epsilon_p \simeq 1$ (dashed line) and $\epsilon_c \simeq 1$ (not shown) a network which is able to reproduce the data within the statistical uncertanties is found. At smaller values of the threshold, errors $\epsilon_p, \epsilon_c < 1$ corresponding to overfitting of the data. For systems which are very well sampled (red line), the problem of overfitting is less severe. B) Typical pattern for ϵ_p for systems with longer correlation lengths, larger fluctuations of ϵ_p, ϵ_c are present. C) Thanks to the the reference entropy the fluctuations of ϵ_p, ϵ_c at small threshold are smaller but ϵ_p at large threshold can be large. D) By lowering the threshold the susceptibility of the inverse problem increases: we reconstruct networks less and less sparse while making the inverse problem more computationally challenging. At small values of the threshold the relative errors can become smaller than one, suggesting overfitting. When the data is overfit the susceptibility of the inverse problem increases because the noise has no simple network structure. E) With the reference entropy at large values of the threshold the inverse susceptibility is nonzero even if the underlying interactions are short range, thanks to the properties of the Gaussian model [2].

We note however that there is no risk of overfitting data which has been perfectly sampled. In this case, as there is no sampling noise, one is justified in fitting the model to the data as tightly as is practically possible.

We now consider the typical pattern for the convergence of the reconstruction error and its relationship with the convergence of the entropy expansion in more detail. First let us imagine that the largest inference error is $\delta h_i^{\rm in}$ on a field h_i; the inference error on the entropy $\delta S^{\rm in}$ can be related to $\delta h_i^{\rm in}$ because h_i and p_i are conjugate variables (see (6)) and therefore $\frac{\partial S}{\partial p_i} = -h_i$, or

$$p_i \, \delta h_i^{\rm in} \sim \delta S^{\rm in} \, . \tag{51}$$

The relationship between the convergence error on the fields and the reconstruction error on the one- and two-point correlations is, as discussed in Section 3, $\delta p_i \approx |\chi| \delta h_i$, where $|\chi|$ is the norm of the susceptibility matrix as defined in (29). Using (51) the

order of magnitude of the typical reconstruction error is related to the convergence error of the cluster expansion,

$$\delta p^{\rm rec} \approx \frac{|\chi|}{p} \delta S^{\rm in} \,, \tag{52}$$

where p is the average spiking probability of the cell populations for a given choice of the time bin. As mentioned in Section 3 for $|\chi|/p$ small the reconstruction problem is easy, while for $|\chi|/p$ large, the reconstruction problem is difficult and small errors on the inferred parameters give large errors in the reconstructed observables. Reconstruction of the correlations within the sampling precision (13) requires

$$\frac{\delta p^{\rm rec}}{\delta p} = \frac{|\chi|\, \delta S^{\rm in} \sqrt{B}}{p\sqrt{p(1-p)}} \approx 1. \tag{53}$$

The same reasoning holds for the reconstruction error ϵ_c, but again it gives only a rough estimate, because the matrix dependence of (27) has to be properly taken into account. A sketch of the typical pattern for the reconstruction errors by lowering the threshold T is shown in Fig. 5.

To understand how the inference errors $\delta h_i^{\rm in}, \delta J_{ij}^{\rm in}$ behave as a function of the threshold, it is necessary to start from the understanding of the convergence of the cluster entropy expansion as a function of the threshold. At large values of the threshold important clusters, corresponding to pairs with large correlation indices, are progressively added up, resulting in a large change in the relative errors. As the threshold is lowered large fluctuations of the relative errors may still appear. As already discussed in [2], dilute interaction networks display a cancellation property of the cluster entropies. Clusters which share the same interaction path, and which are not first neighbors in the coupling network, have similar cluster entropies in absolute value but with different signs. When all of the cluster entropies corresponding to a given interaction path are summed, their overal contribution is typically much smaller than the contribution of any individual cluster. Because such collections of clusters have cluster entropies which are similar in magnitude, they are typically added to the cluster expansion in packets when the threshold is lowered. Large fluctuations of the entropy error $\delta S^{\rm in}$ arise when the threshold is chosen between the largest and smallest values of the cluster entropies for clusters belonging to a given interaction path, thus "cutting" the packet by including only part of the collection in the expansion, such that the cluster entropies cannot cancel each other. Larger fluctuations of the reconstruction error which are amplified by the direct susceptibility of the system (52) arise in correspondence to these values of the threshold.

As shown in [2] the entropy expansion can be absolutely convergent as the threshold is lowered (see Fig. 5A). When the correlation length of the system is small $\delta S^{\rm in}$ decreases smoothly as the threshold is lowered. On the contrary when the correlation length of the system is large, fluctuations of $\delta S^{\rm in}$ appear (see Fig. 5B). When taking $S_{\rm MF}$ as the reference entropy, as shown in Fig. 5C, the relative error of the one-point correlations ϵ_p is not small even when the threshold is large; the network corresponding to $S_{\rm MF}$ is not a

network of independent spins, and the inverse susceptibility of the reconstructed system is different from zero even for large values of the threshold (Fig. 5D). Fluctuations of the error at small values of the threshold are smaller, because the entropies of nonzero cluster entropies arising purely from sampling fluctuations (13), (16) are smaller [2].

When a threshold T^* is reached such that the relevant clusters corresponding to the structure of the inverse susceptibility matrix in the real interaction graph have been summed up, the entropy error $\delta S^{\rm in}$ goes to zero in the limit of perfect sampling. For finite sampling, it goes to a residual value corresponding to the difference between the entropy of the perfectly sampled system and the system which has been finitely sampled. At such values of the threshold $\epsilon_p = \epsilon_c \simeq 1$. We have verified in the analyzed data sets that the inverse susceptibility χ^{-1} for the inferred Ising model at the threshold T^* is sparse. For example for the retinal recording of 60 cells in dark conditions only 20% of the elements of χ^{-1} are different from zero (within a precision of 10^{-9}). However, we note that the convergence of the expansion is not guaranteed. When the sampling noise, which has no simple network structure, is very large it can mix up different packets and obscure the structure of the inferred interaction graph.

6. Application to real data

As an example of potential applications of the Selective Cluster Expansion algorithm (SCE) we have re-analyzed, following [17], several sets of real data from multielectrode recordings of collections of neurons. We examine *in vitro* recordings of salamander retinal ganglion cells:

- A 4450 second recording of 51 ganglion cells in a retina illuminated with randomly flickering bright squares [22]. The recording is made with a multielectrode array on a surface of about 1 mm^2, and approximately 20% of the ganglion cells on that surface are recorded. The position of the receptive field of these cells is known.
- A 2000 second recording of the spontaneous activity of 60 cells of the same retina as above observed in total darkness, of which 32 cells are common to the recording with randomly flickering stimulus [22].
- A recording of 40 cells in a salamander retina presented with a 120 second natural movie repeated 20 times. This recording is much denser; approximately 90% of the ganglion cells on the analyzed surface are recorded [5].

We also study several *in vivo* recordings of neurons in the medial prefrontal cortex of a rat during a working-memory task:

- A 1500 second recording of 37 cells with tetrodes, each consisting of of four electrodes, which record both superficial and deep layers of the medial prefrontal cortex [32].
- A 2800 second recording of 117 cells with silicon probes which record both superficial (layers 2 and 3) and deep (layer 5) layers of the medial prefrontal cortex [24].

Fig. 1 shows the raster plots displaying spike trains for the first second of recordings of the different data sets. Unless otherwise indicated, the time bin we use to analyze the data is $\Delta t = 20$ ms. Spiking frequencies and pairwise correlations for a fixed time window are shown in Fig. 9 and Fig. 10. We observe from these figures that in cortical data the frequency of the activity is higher, and in particular some cells spike very rapidly. The retinal recordings are stationary in the sense that the spiking frequencies and pairwise correlations of the data do not change over time. In the recordings with a natural movie stimulus this holds at least on a time scale larger than 120 s. The cortical data sets, however, are nonstationary. In particular some cells are active only in part of the recording of the 117 cells in medial prefrontal cortex, see Fig. 4. Note that in the analysis we present here data are considered to be stationary and the time dependence is not explicitly taken into account. We discuss the consequences of this assumption for nonstationary data and how to go beyond this limitation in the Conclusion.

The correlation index histograms for some cells of the different data sets are shown in Figs. 2–4. As discussed in Section 1.2 the correlation index is the expansion parameter in the high temperature expansion, and when it is small, as for the $N = 37$ Cortical data set (CA), it means that the interactions are weak and that $S_{\rm MF}$ gives a good approximation of the couplings. This implies also that the cluster expansion, with or without a reference entropy, will converge easily. When the correlation indices are large the magnitude of the couplings will be also be larger. In this case the convergence of the SCE and the optimal expansion variable (*i.e.* $S - S_{\rm MF}$ or S with no reference entropy) will depend on the structure of the interactions. We explore this point in more detail below.

6.1. Performance of the algorithm for retinal data

We show the behavior of the reconstruction errors ϵ_p and ϵ_c as a function of the threshold T on the retinal data in Fig. 6 for the expansion of $S-S_{\rm MF}$. Results on the convergence of the algorithm and on the value of the inferred entropy are summarized in Table 1 for the different procedures tested, including expansions without the reference entropy. These results include the value of the optimal threshold T^*, as well as size of the largest cluster in the expansion $K_{\max}$, the total number of clusters processed $N_{\rm totcl}$, the total number of clusters selected at $N_{\rm totsel}$, and the inferred value of the entropy S, all evaluated at T^*.

As shown in the figures and in the table the reference entropy S_0 is helpful for the inference problem when the SCE is applied to retinal data. The threshold T^* is lower for the expansion of S alone with no S_0, implying an increase in the number of processed clusters and a larger value for $K_{\max}$. For the flickering stimulus (Fl) the procedure without reference entropy did not converge even at a low value of the threshold. Here small couplings, which would have been obtained through the mean field couplings in the expansion of $S - S_{\rm MF}$, are important for the proper reconstruction of the correlations, as mentioned in Section 3. Decreasing the threshold enough to fit many small couplings

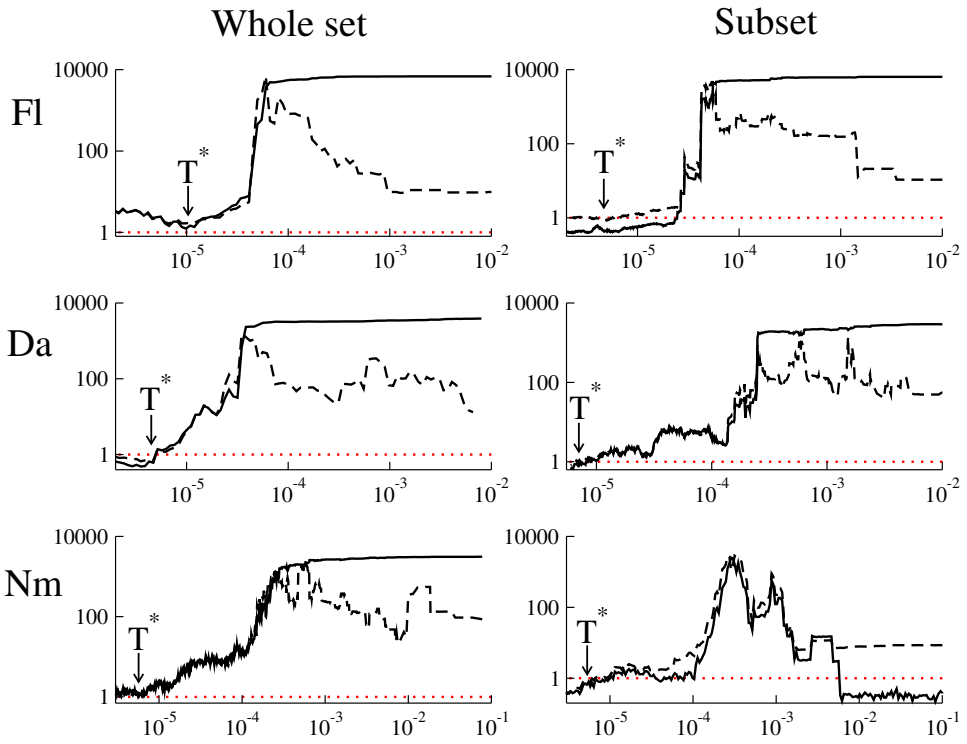

Figure 6. Performance of the SCE of $S - S_{\rm MF}$ with L_2 norm regularization on retinal data as a function of the threshold T, analyzed with a time bin of $\Delta t = 20$ ms. The reconstruction errors ϵ_p (solid) and ϵ_c (dashed) (49) are computed from Monte Carlo simulations. The value $\epsilon = 1$ is indicated by a red dotted line, and the selected optimal threshold T^* is marked in each plot with an arrow. Fl : flickering stimulus, left whole set of 51 cells, right subset of 32 cells; Da : dark, left whole set of 60 cells, right subset of 32 cells; Nm : natural movie stimulus, left whole set of 40 cells, right subset of 20 cells.

drives the SCE algorithm to consider very large clusters ($K_{\rm max} = 17$ at the smallest value of T tested), which slows the algorithm considerably.

Both L_1 and L_2 norm regularizations on S and $S_{\rm MF}$ worked similarly well. In each case we found the value of $K_{\rm max}$ to be around $6-8$ and the threshold T^* varied between 10^{-5} and 10^{-6}.

We have also studied the inference problem on subsets of the full data sets described above. As shown in the right column of Fig. 6, the value of the theshold T^* and the maximum cluster size $K_{\rm max}$ for large subsets of spins are of the same order of magnitude as for the full set. Moreover the number of clusters selected increases approximately linearly with the system size. This property was related in [1, 2] to the *locality* of the coupling-susceptibility χ^{-1}, and was shown to hold on artificial data of unidimensional and bidimensional Ising models. This suggests that for the chosen subset the structure of the interaction graph has not changed much locally.

Retinal data							
		Fl 51	Fl 32	Da 60	Da 32	Nm 40	Nm 20
	T^*	$9.8\,10^{-6}$	$6.5\,10^{-6}$	$3.5\,10^{-6}$	$6.0\,10^{-6}$	$5.6\,10^{-6}$	$3.7\,10^{-5}$
$S - S_{\rm MF}$	$K_{\rm max}$	6	7	11	8	9	5
L_2 norm	$N_{\rm totcl}$	$1.1\,10^4$	4700	$7.6\,10^4$	9200	$6.8\,10^4$	930
	$N_{\rm selcl}$	1450	840	$1.0\,10^4$	1700	$1.0\,10^4$	190
	T^*	$6.2\,10^{-6}$	$6.7\,10^{-6}$	$5\,10^{-6}$	$8.2\,10^{-6}$	$6.6\,10^{-6}$	$3.5\,10^{-5}$
$S - S_{\rm MF}$	$K_{\rm max}$	7	7	9	8	9	5
L_1 norm	$N_{\rm totcl}$	$1.5\,10^4$	4500	$5.1\,10^4$	7300	$5.8\,10^4$	940
	$N_{\rm selcl}$	2000	830	7500	1400	8500	190
	T^*	$< 6.0\,10^{-7}$	$1.2\,10^{-6}$	$2.7\,10^{-6}$	$6.1\,10^{-6}$	$1.1\,10^{-6}$	$6.4\,10^{-6}$
S	$K_{\rm max}$	> 17	9	11	9	15	7
L_1 norm	$N_{\rm totcl}$	$> 3.0\,10^5$	$3.8\,10^4$	$3.1\,10^5$	$2.6\,10^4$	$1.7\,10^6$	5900
	$N_{\rm selcl}$	$> 3.2\,10^4$	6400	$3.3\,10^4$	4100	$2.4\,10^5$	1300
Same for all	S	3.5	2.2	4.6	2.5	3.7	1.9

Table 1. Convengence of the Selective Cluster Expansion on retinal data, analyzed with a time bin of $\Delta t = 20$ ms. In cases where the algorithm did not easily converge, bounds on the minimum or maximum quantities necessary to obtain a good fit to the data are given, determined by the lowest value of the threshold considered.

6.2. Performance of the algorithm for cortical data

In Fig. 7 we show the performance of the algorithm on cortical data, using no reference entropy and with L_1 regularization on the cluster entropies, and in Table 2 we give the results also for the cases with L_2 regularization and for the expansion of $S - S_{\rm MF}$. As expected from the small values of the correlation indices the SCE applied to the cortical recording of $N = 37$ neurons (CA) works very well for expansions both with and without the mean field reference entropy. Indeed, the expansion converges already at just $K_{\rm max} = 2$ with the reference entropy, and $K_{\rm max} = 4$ without it.

For the cortical recording of $N = 117$ cells we have tested the performance of the algorithm with time bins of size $\Delta t = 5$ ms (C5) and $\Delta t = 20$ ms (CB). In both cases the convergence of the algorithm was faster in without the reference entropy (see Table 2). The convergence of the algorithm for both $\Delta t = 20$ ms and $\Delta t = 5$ ms when expanding with respect to the mean field reference entropy was very slow, and the relative errors ϵ_p, ϵ_c did not approach one at a threshold $T \approx 10^{-7}$.

The poor performance of the expansion of $S - S_{\rm MF}$ for cortical data, and the success of the expansion in S alone, is in stark contrast with the analysis of retinal recordings. This phenomenon is related to the fact that the interaction network which is capable of fitting the cortical data is relatively dilute and consists of larger couplings, for which the mean field result is a poor approximation. The presence of a reference entropy makes reconstructing the empirical correlations more difficult in this case, as

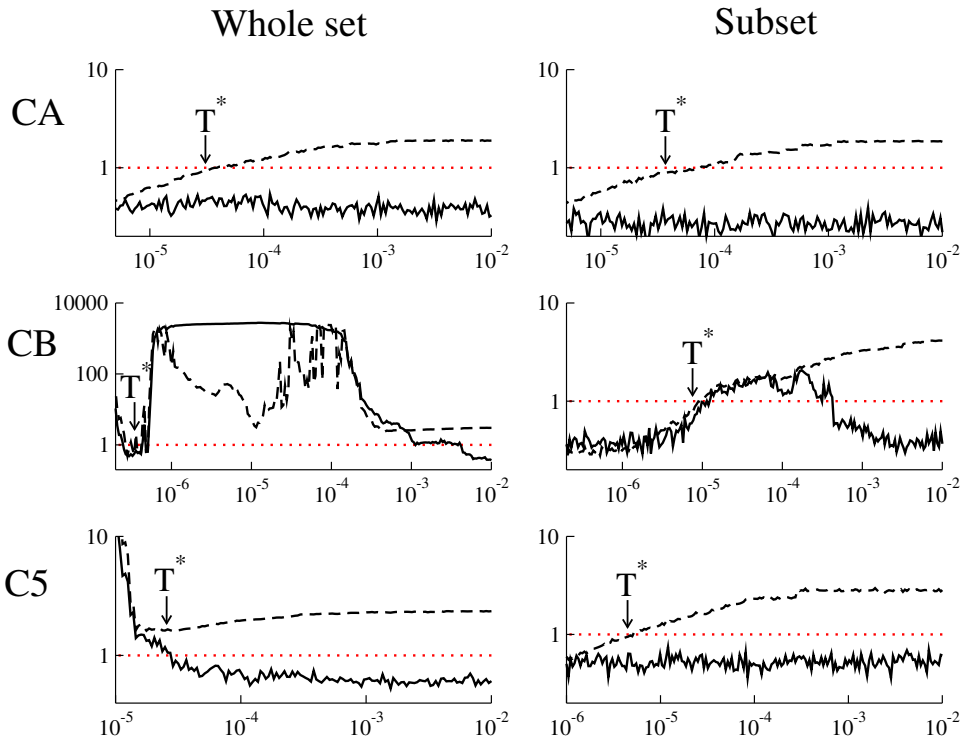

Figure 7. Performance of the SCE of S with L_1 norm regularization on cortical data as a function of the threshold T. The reconstruction errors ϵ_p (solid) and ϵ_c (dashed) (49) are computed from Monte Carlo simulations. The value $\epsilon = 1$ is indicated by a red dotted line, and the selected optimal threshold T^* is marked in each plot with an arrow. CA : cortex recording of 37 cells analyzed with a time bin of 20 ms; CB : cortex recording of 117 cells analyzed with a time bin of 20 ms; C5 : cortex recording of 117 cells analyzed with a time bin of 5 ms.

the inferred interaction network is then fully connected (even if many of these couplings are small).

As we expect from the fact that by increasing the time bin one unveils more of the network structure and indirect couplings, the value of $K_{\max} = 4$ we obtain is smaller at $\Delta t = 5$ ms than at $\Delta t = 20$ ms. $K_{\max} = 10$ is quite large in the latter case, indicating that some cells are connected with many others. This hub effect has been noted in [24, 47], and it can be related to the structure of the probes which record cells from different layers. For this data set, moreover, the value of $K_{\max}$ and of the threshold T^* changes from the subset of 30 neurons to the whole set, indicating a change in the local connectivity due to the removal of many of the neurons in the full data set. A more detailed analysis of the inferred structure will be performed below.

6.3. Computational time at T^*

As shown in Fig. 8, for all data sets for which the algorithm has converged, the computational time at the threshold T^* was not more than a few tens of minutes of calculation on a single core of a 2.8 GHz Intel i7 processor. The computational time

Cortical data							
		CA 37	CA 20	CB 117	CB 30	C5 117	C5 30
	T^*	$7.8\,10^{-6}$	$7.8\,10^{-6}$	$< 8.0\,10^{-8}$	$1.0\,10^{-5}$	$< 5.0\,10^{-8}$	$5.3\,10^{-6}$
$S - S_{\rm MF}$	$K_{\rm Max}$	2	3	> 11	4	> 8	3
L_2 norm	$N_{\rm totcl}$	910	230	$> 5.6\,10^5$	940	$> 2.1\,10^5$	570
	$N_{\rm selcl}$	91	32	$> 2.3\,10^4$	110	$> 5.4\,10^4$	64
	T^*	$7.8\,10^{-6}$	$7.8\,10^{-6}$	$< 8.0\,10^{-8}$	$9.9\,10^{-6}$	$< 5.0\,10^{-8}$	$7.8\,10^{-6}$
$S - S_{\rm MF}$	$K_{\rm Max}$	3	3	> 10	4	> 8	3
L_1 norm	$N_{\rm totcl}$	920	230	$> 5.4\,10^5$	920	$> 2.0\,10^5$	530
	$N_{\rm selcl}$	93	32	$> 2.2\,10^4$	110	$> 5.2\,10^4$	55
	T^*	$3.4\,10^{-5}$	$5.7\,10^{-5}$	$2.5\,10^{-7}$	$2.9\,10^{-5}$	$2.0\,10^{-5}$	$4.8\,10^{-6}$
S	$K_{\rm Max}$	4	3	10	4	4	3
L_1 norm	$N_{\rm totcl}$	1300	270	$4.4\,10^6$	5000	$1.6\,10^4$	1200
	$N_{\rm selcl}$	140	40	$1.7\,10^5$	580	630	140
Same for all	S	6.6	3.8	14.7	3.3	5.9	1.2

Table 2. Convengence of the Selective Cluster Expansion on cortical data. In cases where the algorithm did not easily converge, bounds on the minimum or maximum quantities necessary to obtain a good fit to the data are given, determined by the lowest value of the threshold considered.

incereases with the number of processed clusters $N_{\rm totcl}$ (see Fig. 8) and the maximum cluster size $K_{\rm max}$ as

$$\text{time} \simeq \sum_{K=1}^{K_{\rm max}} N_{\rm totcl}(K)\; 2^K \,, \tag{54}$$

which is roughly proportional to the number of clusters given the same value of $K_{\rm max}$.

However, when the number of clusters becomes large the construction rule can become slow, and in fact this can become the limiting step for the computational time of the algorithm. In the cluster construction step we test whether the union of each pair of clusters of size K in the list of selected clusters forms a new cluster of size $K+1$. The time necessary to complete this step thus scales quadratically with the number of selected clusters, which may be quite large particularly when expanding S without the reference entropy S_0. While the vast majority of attempts to form a new cluster may fail, the simple act of checking each possibility becomes computationally demanding.

6.4. Reconstruction of the first and second moments of the activity

The spiking probabilities and connected correlations obtained from Monte Carlo simulations with the inferred parameters at T^* are close to those obtained from experiments as shown in Figs. 9–10. Indeed the conditions $\epsilon_c = 1$, $\epsilon_p = 1$ specify that the difference between the inferred and empirical correlations is approximately the same as the statistical uncertainty of the empirical correlations due to sampling a

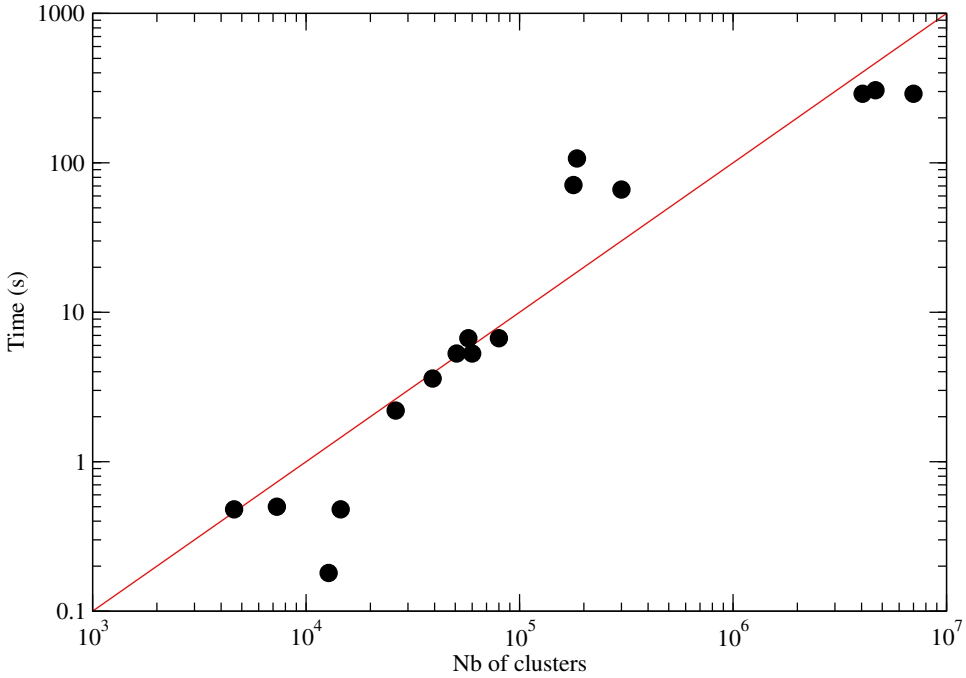

Figure 8. Computational time of the SCE at threshold T^* for the different data sets as a function of the number of processed clusters $N_{\rm totcl}$. All times are for computed on one core of a 2.8 GHz Intel i7 processor. The computational time grows with the number of processed clusters, which depends more strongly on the structure of the interaction graph and on the number of sampled configurations than on the total system size N. For a fixed maximal size of clusters, $K_{\rm max}$, the running time is roughly proportional to the number of clusters.

finite number of configurations. We find that the reconstruction of the $\{p_i\}$ and $\{c_{ij}\}$ is equally good for the all different reference entropy and regularizations choices, provided of course that they reach the treshold T^* at which ϵ_c, $\epsilon_p = 1$.

6.5. Reconstruction of third moments and probability k cells spike in the same time window

We have verified that the three cell connected correlations c_{ijk} and the probability $P(k)$ that k cells spike in the same bin are also quite well reproduced by the inferred Ising model, as shown in Figs. 11 and 12. Note that for the cortical recording of 37 cells (CA) and of 117 cells analyzed with $\Delta t = 5$ ms (C5) the connected 3 cell correlations are for the majority of the cells so small that it is difficult to separate them from the sampling noise. We have therefore also plotted the correlations p_{ijk}, which is the probability that the three cells spikes in the same time bin, in Fig. 12. We emphasize that, unlike the reconstruction of the first and second moments, there are no *a priori* reasons that the Ising model should also reconstruct the higher moments of the experimental distribution, because it is not the true model which has generated the data, and these measurements are not used as constraints in the inference. Nevertheless as discussed in Section 1.4

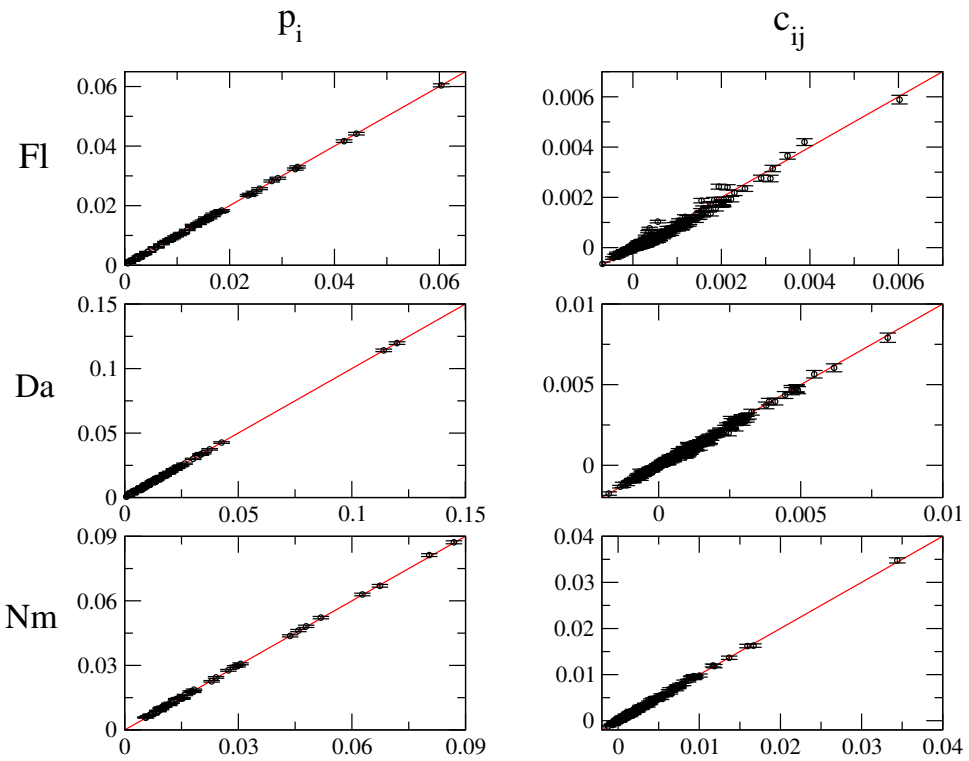

Figure 9. Spiking frequencies p_i and connected correlations c_{ij} for retinal data sets calculated from a Monte Carlo simulation of the inferred Ising model (vertical) versus experimental values $p_i^{\rm exp}$, $c_{ij}^{\rm exp}$ (horizontal). The error bars are given by the statistical fluctuations δp_i, δc_{ij}. Fl : flicker stimulus, 51 cells; Da : dark, 60 cells; Nm : natural movie stimulus, 40 cells.

this model seems to reproduce fairly well the statistics of the data, at least for the c_{ijk} and the $P(k)$.

Intuitively the reason for such a success can be simply due to the fact that the number of effective configurations that the system explore is 2^S, and this number should be compared with the number of parameters $N\,(N+1)/2$ that we use to fit the distribution (see discussion in [17]). In principle any model with a number of parameters similar to the effective number of configurations could be equally good. For the data sets we have analyzed with $N \leq 60$, (see Tables 1 and 2) we have that $N\,(N+1)/2 > 2^S$. For larger systems with more than one hundred cells, the exponential growth of 2^S with N, due to the fact that $S \propto N$, will be difficult to compensate with the increase of the number of parameters $\propto N^2$. For the Fujisawa data sets with time bins of size 5 ms we have $S = 5.9$ so $2^S < N\,(N+1)/2$, but for 20 ms for we obtain $N\,(N+1)/2 = 6900$ while $2^S = 2.6\,10^4$. Moreover the reconstruction of the higher moments is still quite good. This could indicate that the Ising model, which is the maximal entropy model, is really a good model to reproduce the data.

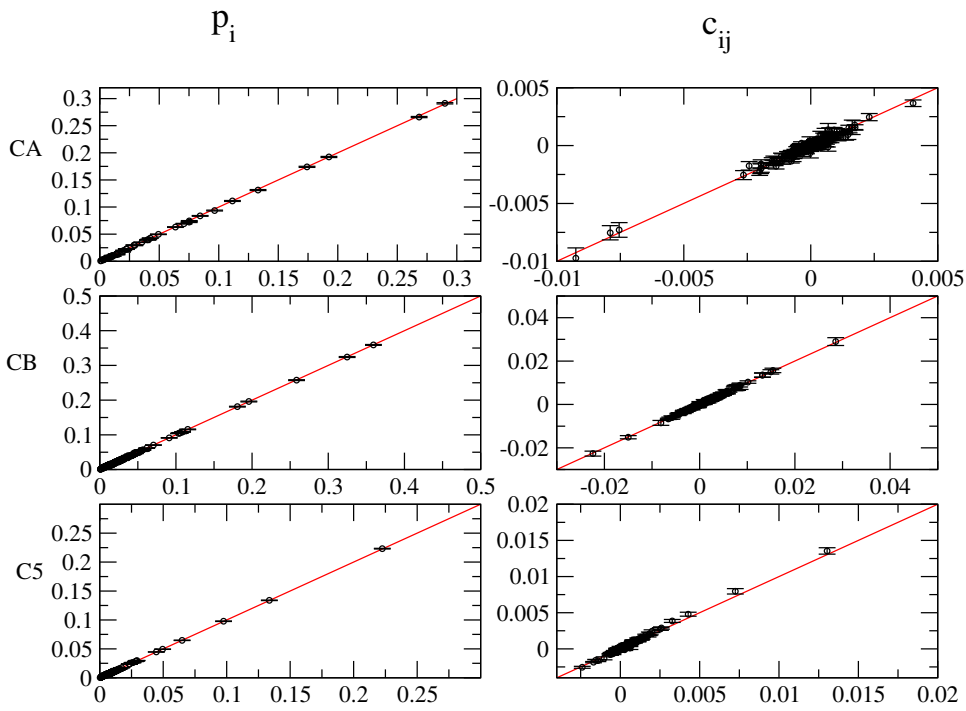

Figure 10. Spiking frequencies p_i and connected correlations c_{ij} for cortical data sets calculated from a Monte Carlo simulation of the inferred Ising model (vertical) versus experimental values $p_i^{\rm exp}$, $c_{ij}^{\rm exp}$ (horizontal). The error bars are given by the statistical fluctuations δp_i, δc_{ij}.

6.6. Histogram of couplings, negative couplings and error bars on couplings

Once the couplings are inferred, the errors bars δJ_{ij} can be calculated from (26). Fig. 13 and Fig. 14 show the the histogram of couplings in which we distinguish reliable couplings, for which $|J_{ij}|/\delta J_{ij} > 3$, from unreliable couplings, which are compatible with zero within the error bars. As expected, because of the sampling fluctuations there are many unreliable couplings (particularly small couplings), but these may be still necessary to accurately reproduce the activity, as discussed in Section 3. Large reliable couplings correspond generally to large correlation indices (see the cross-correlation histograms in Fig. 2), but the converse is not necessarily true.

An interesting issue is to separate inhibitory and excitatory connections. It is important to notice that negative couplings can be due to undersampling. Indeed, as already noted, they can correspond to cells which never spike together in the recorded activity, but often these couplings also have large error bars and therefore will be classified as unreliable (see the histograms in Figs. 13 and 14). Therefore to properly define reliable negative couplings it is important to know the error bars on the inferred couplings.

As shown in the histograms Fig. 13 and Fig. 14 both in retinal and cortical data there are some large negative couplings which are reliable. There are fewer reliable large negative couplings in dark and flicker retinal data, which can be related to the fact the

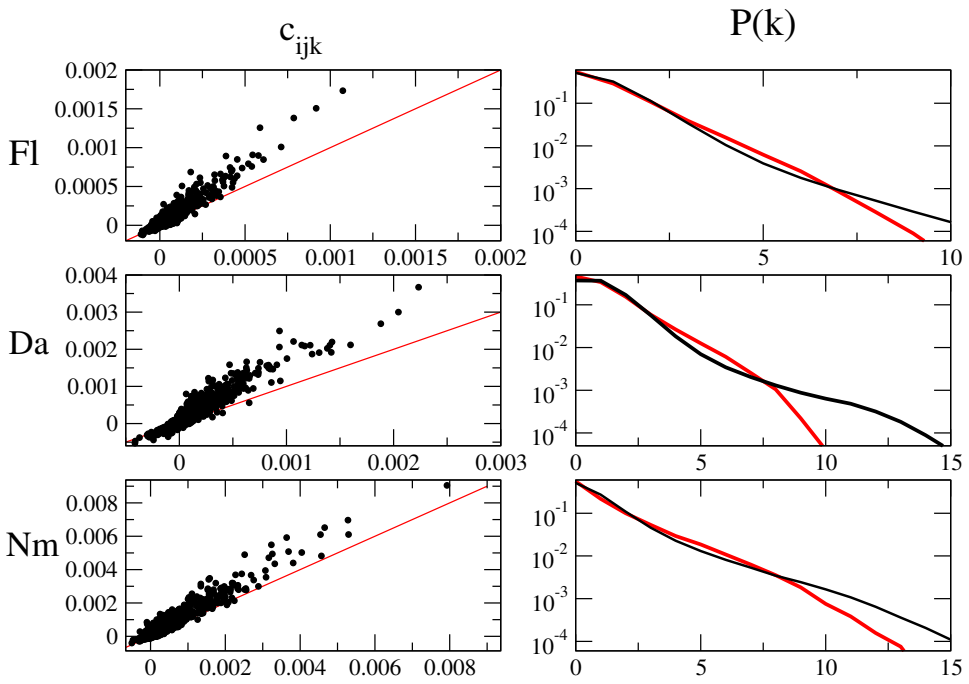

Figure 11. Reconstructed three-cell connected correlations c_{ijk} (vertical), and probability $P(k)$ that k cells spike in the same time bin (black), versus the experimental ones $c^{\rm exp}_{ijk}$ (horizontal) and $P^{\rm exp}(k)$ (red) for retinal data. The reconstructed values are obtained from a Monte Carlo simulation of the inferred Ising model. Fl : flicker stimulus, 51 cells; Da : dark, 60 cells; Nm : natural movie stimulus, 40 cells.

the majority of the recorded cells in this data sets are of OFF cells.

In Fig. 3 we show the cross correlation histograms of pairs of cells corresponding to negative couplings. Cells with negative couplings can correspond to those with a large central anticorrelation peak and a large correlation index, as is the case for Da 11-26, or they can display a small negative correlation index as in the example of Fli 3-18, which is perhaps attributable to network effects. Negative couplings can also be due to the fact that the Ising approach reproduces pairwise correlations in a fixed time bin and does not include the delay in response between neurons. Some couplings can change sign when increasing the bin size is changed, as shown in the example of Fli 1-22, which displays a correlation peak shifted by about 33 ms with respect to zero. As noted in [17] the corresponding pair recorded in dark conditions has an unshifted peak and the inferred coupling is positive. Another case which is present in $N = 117$ cortical recording is the pair CB 135-178, shown in Fig. 4, which has a negative couplings and a negative cross correlation index for $\Delta t = 5$ ms and $\Delta t = 20$ ms because the neuron labeled 135 is active only in the first part of the recording, therefore important nonstationary effects are present.

As shown in Fig. 13 correlation indices are large in the retinal data and many couplings have large values (between -4 and 4). The histogram of couplings is similar for the different data sets.

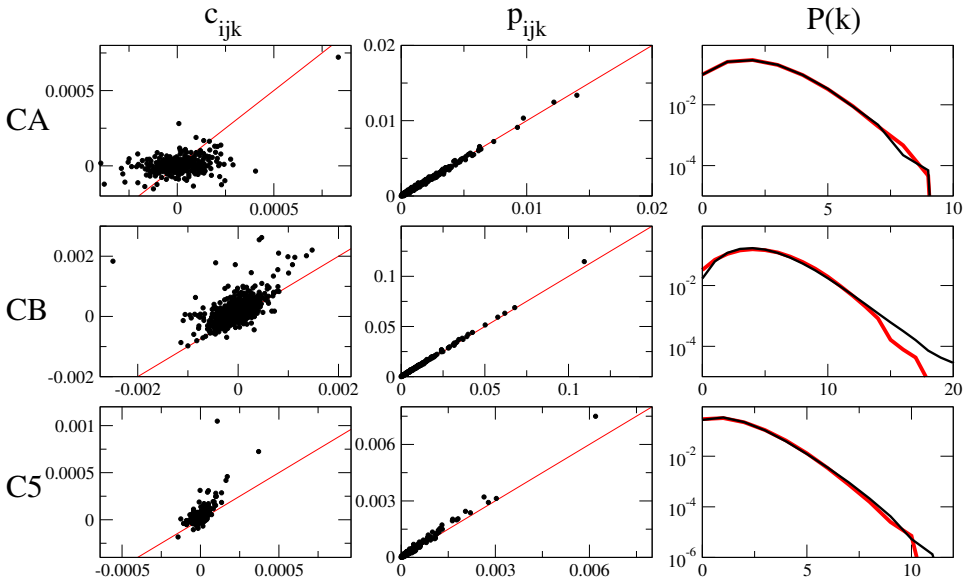

Figure 12. Reconstructed three-cell correlations p_{ijk} and connected correlations c_{ijk} (vertical), and probability $P(k)$ that k cells spike in the same time bin (black) versus the experimental ones $p^{\rm exp}_{ijk}$, $c^{\rm exp}_{ijk}$ (horizontal) and $P^{\rm exp}(k)$ (red) for cortical data. The reconstructed values are obtained from a Monte Carlo simulation of the inferred Ising model. CA : cortex recording of 37 cells; CB : cortex recording of 117 cells analyzed with a time bin of 20 ms; C5 : cortex recording of 117 cells analyzed with a time bin of 5 ms.

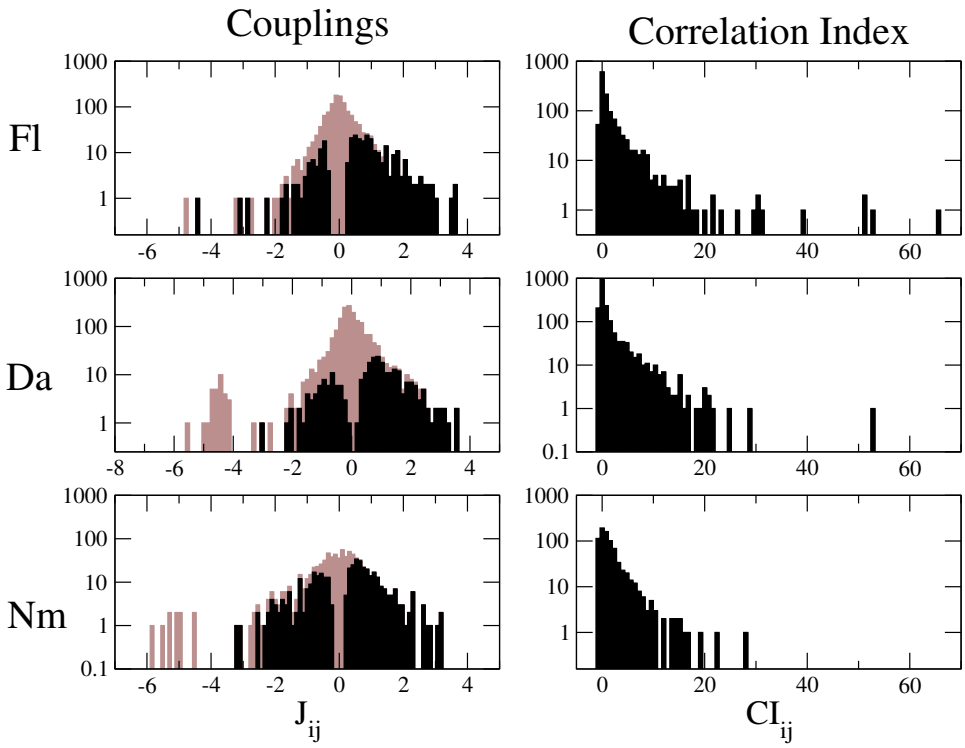

Figure 13. Histogram of couplings for retinal data, compared to the histogram of correlation indices. Reliable couplings are marked in black, and unreliable couplings are marked in brown.

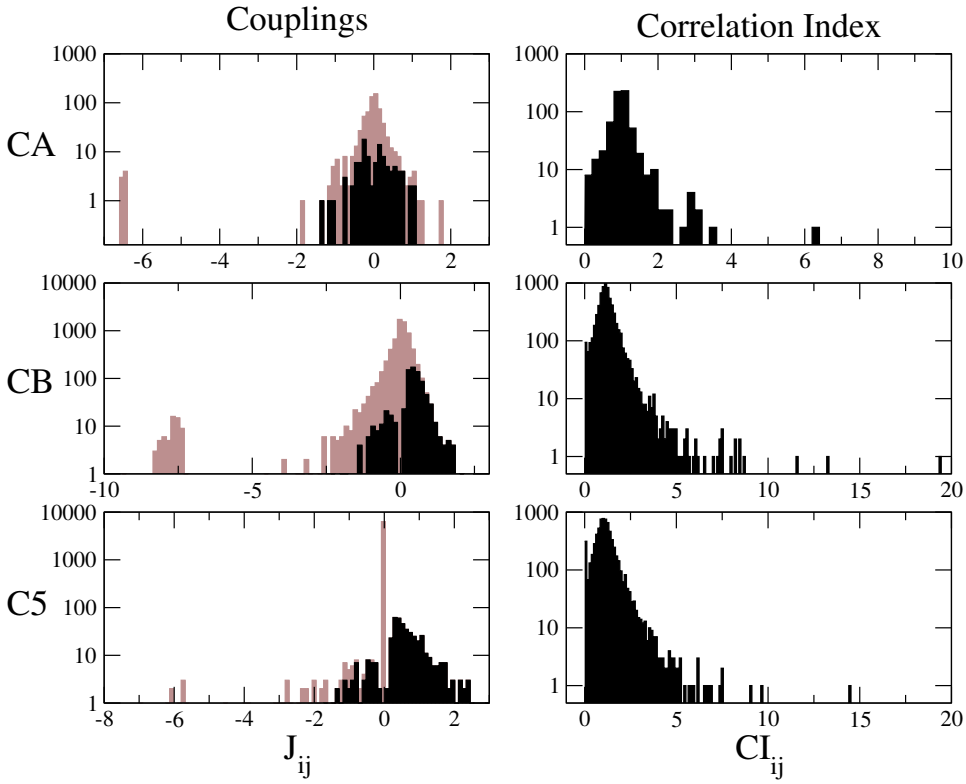

Figure 14. Histogram of couplings for cortical data, compared to the histogram of correlation indices. Reliable couplings are marked in black, and unreliable couplings are marked in brown.

Fig. 14 shows that large correlation indices and couplings are also presents in the cortical data, with the exception of the cortical system of 37 cells, for which the correlation indices and couplings are the smallest. In the histograms the largest positive couplings are of order one. In cortical data there are large reliable negative couplings which have the same magnitude as the positive couplings. The network for the 117 neuron cortical recording and $\Delta t = 5$ ms is sparser than the one for $\Delta t = 20$ ms, as shown by the large peak at zero. Indeed, as already noticed, for $\Delta t = 20$ ms the structure of the inferred network becomes more complicated with larger cluster sizes.

6.7. Network effects: comparison with correlation index and with inferred couplings on a subset of data

We analyze the importance of the network effects in the inference by comparing the couplings inferred via SCE with the two-cell couplings $J_{ij}^{(2)}$, and by comparing the couplings obtained for a full set of data to those obtained when only considering a subset of the whole system. As shown in Fig. 15 for retinal data the inferred couplings are generally different from two-cell couplings, indicating that network effects are important. Thus the correlations do not accurately reflect the structure of the interactions, and it is particularly important to disentangle the couplings from the correlations. As shown

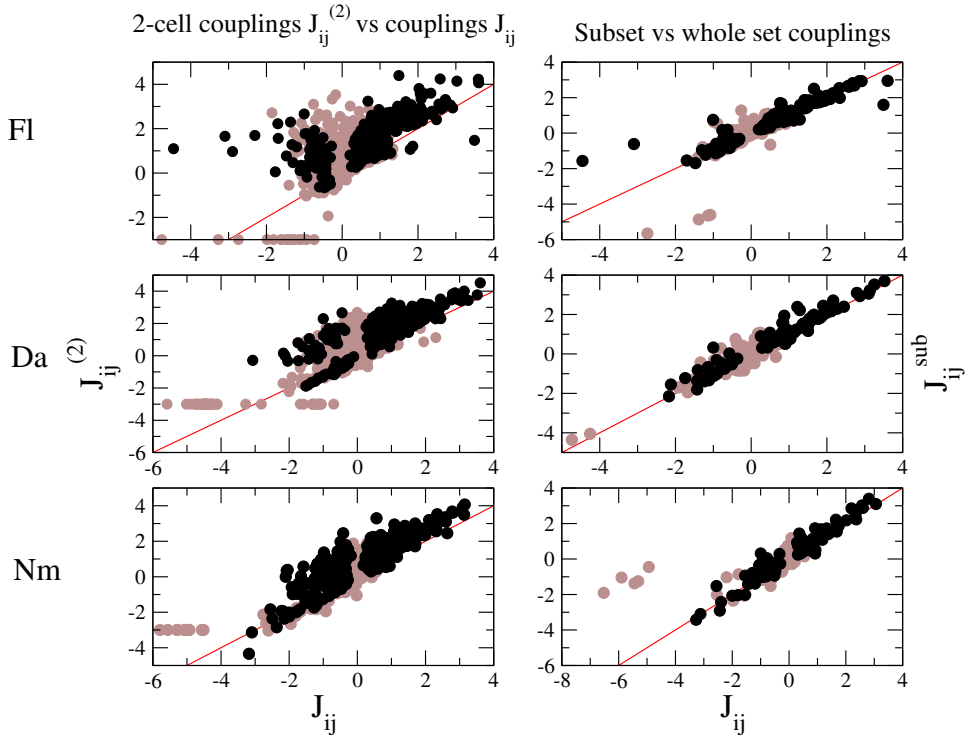

Figure 15. Left: two-cell couplings $J^{(2)}$ versus couplings, right: couplings inferred from a subset of cells versus couplings inferred from the whole set. Reliable couplings are marked in black, and unreliable couplings are marked in brown.

for the pair Fli 3-18 in Fig. 3, there can be negative couplings between a pair of cells even if positive correlations are observed. In this case the positive correlation is due to couplings with other neurons.

As shown in Fig. 15 (right) most of the couplings calculated on the subsets of cells are similar to the ones calculated on the whole set, though some couplings may change their amplitude. The fact that many couplings are similar implies that the inverse susceptibility is local, and therefore each coupling can be inferred from the knowledge of the activity of its neighboring cells. Only if one neighboring cell is removed will the coupling change.

As shown in Fig. 16 the correlation indices and the couplings for the 37 cell cortical recording (CA) are very are similar because not only is the network sparse, but also the correlation length is small. The couplings calcuated on a subset are also similar to the ones for the whole set, except for some negative couplings. For the cortical data of 117 cells set analyzed at $\Delta t = 5$ ms, only some two spin clusters are enough to infer the couplings in the cluster expansion. Neverthless it is important to select the right ones rather than simply including all the two spin clusters in the inference procedure. The inclusion of all the two spin clusters would yield a fully connected interaction graph, much denser than the observed network of couplings, which would not reproduce the observed correlations. Typically the large couplings $J_{ij}^{(2)}$ correspond to real interactions,

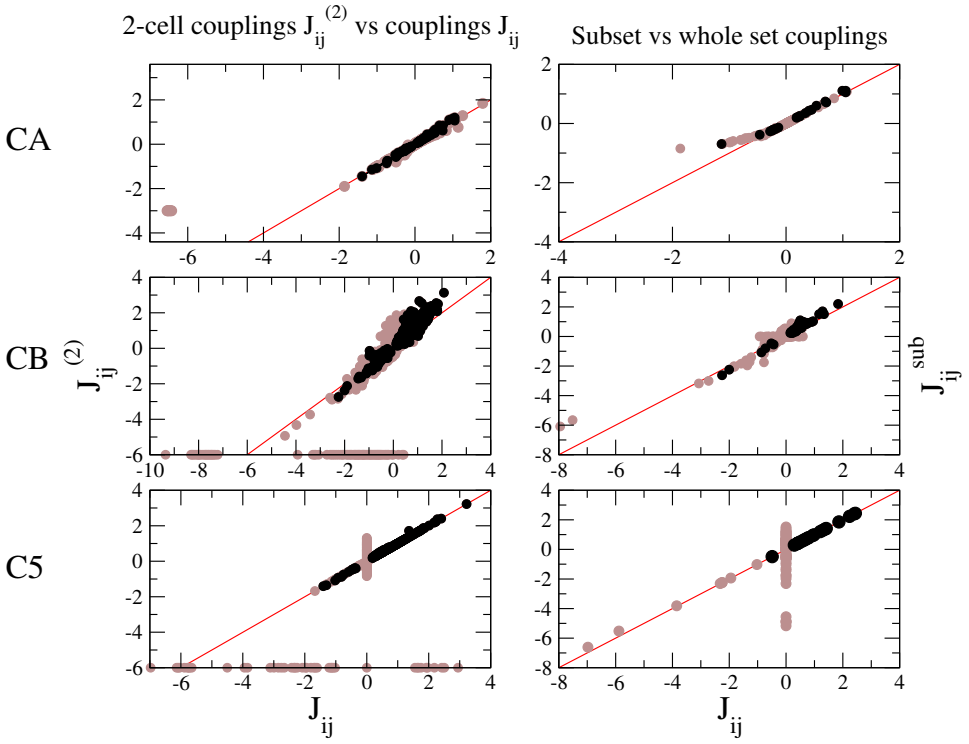

Figure 16. Left: two-cell couplings $J^{(2)}$ versus couplings, right: couplings inferred from a subset of cells versus couplings inferred from the whole set. Reliable couplings are marked in black, and unreliable couplings are marked in brown. Network effects are present especially for the CB set.

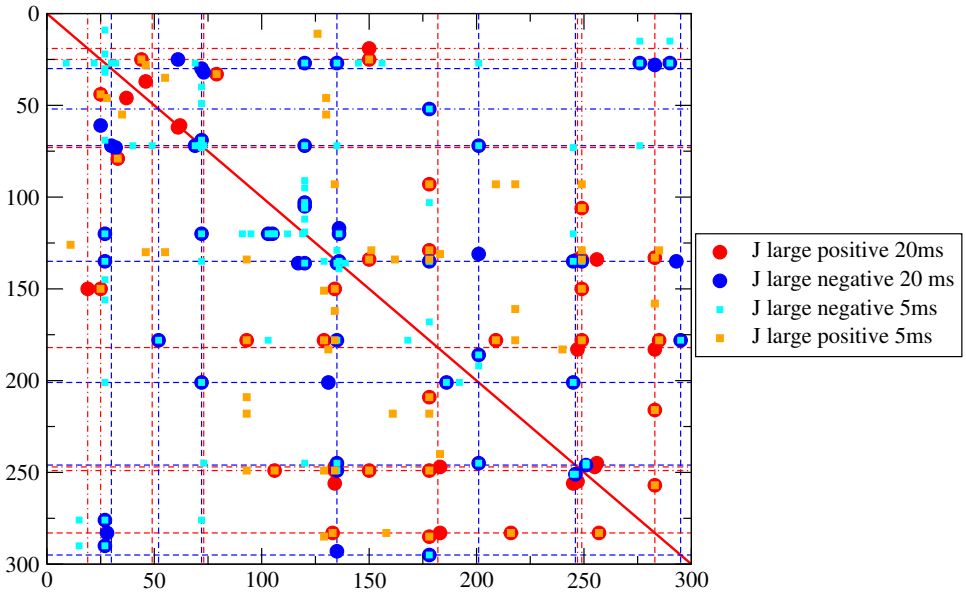

Figure 17. Coupling map for Fujisawa data with 20 ms and 5 ms time bins. Large positive couplings are shown in red, and large negative couplings in blue. Lines indicate excitatory or inhibitory neurons which have been identified by Fujisawa *et al.* [24]. For more details see Section 6.8.

while the correlations leading to smaller $J_{ij}^{(2)}$ reflect network effects rather than direct interactions.

6.8. Map of couplings for the $N = 117$ cortical data set

We show the contact map for the recording of $N = 117$ cortical neurons in Fig. 17, a representation of the inferred coupling matrix. Note that the 117 cells are labeled from 1 to 300 as in the original recordings [24], which reflects more closely the position of the recorded cells. Here large reliable couplings are indicated by dots. Positive couplings are indicated by red (orange) dots while negative couplings are indicated by blue (light blue) dots for the inference performed with 20 ms (5 ms) time bins. Several couplings are similar when inferred with both 5 ms and 20 ms time bins. Fujisawa *et al.* have analyzed monosynaptic couplings in [24] corresponding to approximatively 5 ms delayed peaks with a characteristic width of about 1 ms. The coupling matrix we obtain has some similarities with the connections identified by Fujisawa *et al.*; there are 'hub' effects, *i.e.* some neurons are connected to many others. The inhibitory or excitatory hubs found in [24] are indicated in Fig. 17 by dashed lines. Neurons which are not recognized as hubs but which have been identified as inhibitory or excitatory neurons are indicated by a dot-dashed line. It is worth noticing that hubs and nonlocal effects could also depend on global states or interactions between different layers. Indeed recorded neurons are in different layers. Few connections between cells which are far apart in the numbering order seem to be present.

6.9. Map of couplings for the $N = 32$ retina recording in dark and flickering stimulus

The large positive ($J_{ij} > 0.3$) and reliable couplings are shown in the plane of the receptive fields of the $N = 32$ cells in the retina shared by the recordings in dark (Fig. 18) and with a flickering stimulus (Fig. 19). The centers of the receptive fields of the recorded cells are identified in the receptive field plane. As pointed out in [17] the dark map is short range in the receptive field plane. The map of couplings obtained with a flickering stimulus is similar, but with some additional long-range large positive couplings. For example, the coupling Da 11-26 is negative in dark conditions (see Fig. 3), but it is positive and has a large correlation index in the data taken with the retina subject to a flickering stimulus.

7. Comparison with mean field model results

In this section we compare the properties of the effective Ising model inferred via SCE with the one inferred using the mean field model (Gaussian approximation) alone. We discuss the performance of the mean field model both with respect to the inference of the structure of the couplings and the reconstruction of the statistics of the spiking activity in the population (see Section 3). Here we consider the mean field model with L_1 and L_2 regularizations, varying the value of the penalty parameter γ over a wide range which

extends far above the optimal value in the Bayesian formalism, see (21) and Appendix A. Moreover we discuss how the inference of the structure and the reconstruction of the statistics of the spiking populations changes as a function of γ.

To study the difference in the inferred couplings with respect to the cluster expansion as a function of γ, we define the average inference error ϵ_J with respect to the cluster expansion as

$$\epsilon_J = \left(\frac{2}{N(N-1)} \sum_{i<j} (J_{ij}^{\rm MF} - J_{ij})^2 \right)^{\frac{1}{2}} . \tag{55}$$

The couplings $\{J_{ij}\}$ in (55) are the couplings obtained via SCE which accurately reproduce the empirical correlations.

7.1. Mean field model with L_2 regularization

The penalty parameter which is optimal in the Bayesian framework and which we have used in the reference entropy for the cluster expansion is $\gamma = \frac{1}{10\,B\,(p(1-p))^2}$ (see (22) and (23)), where p is the average spiking probability over the whole popuation. With this choice the network of the largest couplings is very similar to the one inferred with the cluster expansion, as shown in the map of the large positive couplings in the receptive field plane for 32 retinal cells recorded in darkness (Fig. 18) and with a flickering stimulus (Fig. 19). However with this choice of γ the mean field method overestimates the magnitude of the large couplings (they are of order 10), and this gives a large error for the inferred value of the couplings, see Fig. 20 obtained for the flicker data set (Fl) of $N = 32$ cells and for $\gamma = 0.0014$. Furthermore the frequencies and correlations are not reconstructed at all; we find the reconstruction errors $\epsilon_p = 7000$ and $\epsilon_c = 11$ from the expansion of $S - S_{\rm MF}$ with L_2 norm regularization at threshold $T = 1$, such that only single spin clusters are taken into account (see Fig. 6).

The error on the couplings (55) achieves its minimum at $\gamma_{\rm opt} = 7.5$, well above the optimal value γ as determined in the Bayesian framework, see Fig. 20. At this larger value of the regularization strength the magnitude of the large couplings is better inferred and the reconstruction for the single site probabilities is good, even if the pairwise correlations are loosely reconstructed as shown in Figs. 21 and 22. This large regularization term has no meaning in Bayesian approach; indeed the regularization strength should be of the order $1/B$ (21). However it seems to give an effective correction to the mean field inference. The equivalent of a large regularization strength has, moreover, been succesfully applied to predict from large couplings the contact between sites of proteins from coevolving sequences of the same family [48]. The map of couplings in flicker and dark conditions for the large regularization strength are shown in Fig. 19 and Fig. 18, respectively, and are in good agreement with the ones obtained with a Bayesian regularization strength and the one obtained with the SCE.

For the other data sets the inference error obtained from $S_{\rm MF}$ with L_2 regularization is not shown but it behaves similarly with a minimum for $\gamma_{\rm opt} \gg \gamma$ (with $\gamma_{\rm opt} = 5$ for (Da

32), $\gamma_{\rm opt} = 1$ for (CB), and $\gamma_{\rm opt} = 4$ for (Nm 32)), with the sole exception of the cortical recording of 37 cells (CA) for wich $\gamma_{\rm opt} = \gamma = 0.00023$. As already meantioned the Ca data is characterized by small correlation indices and the mean field approximation works well in this case.

The reconstruction of the empirical spiking probabilities is shown in Fig. 21 where we have chosen $\gamma_{\rm rec} = \gamma_{\rm opt} = \gamma$ for (CA), $\gamma_{\rm rec} = \gamma_{\rm opt}$ for (Fl) and $\gamma_{\rm rec} > \gamma_{\rm opt}$ for the other sets for which the spiking probabilities were not reconstructed at the value $\gamma_{\rm opt}$ (see the caption for the values of $\gamma_{\rm opt}$). As shown in Figs. 22 and 23 two cell correlations and probabilities $P(k)$ that k cells spike in the same bin are not well reconstructed. Interestingly, even if the magnitudes of the correlations are not reconstucted, their relative ordering is respected.

7.2. Mean field model with L_1 regularization

In this Section we discuss the inference with the mean field model using L_1 norm regularization. The problem is the same as solving the inverse problem for the Gaussian model with L_1 regularization which has become very popular in the field of graphical learning in recent years, and efficient methods to solve this problem, such as the Lasso algorithm, have been found [20].

In the regularization strength of $S_{\rm MF}$ used in the expansion of $S - S_{\rm MF}$ we have used the penalty (23) with a Bayesian penalty strength $\gamma \propto 1/B$. Fig. 24 shows that, as in the case of the L_2 penalty the value of large couplings with the Bayesian penalty ($\gamma = 0.033$ for the Fl 32 data set) is overestimated; when the penalty parameter is small the L_1 and L_2 regularizations yield very similar results. Again to obtain couplings of the right order of magnitude and with a smaller error ϵ_J, the penalty parameter should be larger, *e.g.* the value $\gamma = 0.11$. However, around this value of the regularization strength the inference is very sensitive to changes in γ. As shown in Fig. 24 a small change to $\gamma = 0.15$ (bottom panel) results in a large change in the number of inferred couplings which are different from zero.

We have therefore studied a simpler penalty with the form

$$\gamma' \sum_{i<j} |J_{ij}| \,. \tag{56}$$

Increasing the vaue of γ' makes the inferred network sparser. It has been proven [49] that the choice of the regularization strength

$$\gamma'_{\rm opt} \simeq \frac{\sqrt{\log N}}{B} \tag{57}$$

selects only couplings which are statistically significant. This choice corresponds to zeroing couplings which could be due to statistical fluctuations of the sampled correlations (16).

In Fig. 24 (bottom left) we show ϵ_J as a function of γ', and indeed the value of γ' which corresponds to the minimal ϵ_J (55) is of the order of $\gamma'_{\rm opt}$. For this choice of

the regularization the inference is less sensitive to changes in γ', and couplings are more progressively unveiled as γ' is decreased.

The map of couplings obtained with a large L_1 penalty strength for the 32 retinal cells is compared to the mean field map with (L_2 and L_1) Bayesian penalty strength in Fig. 18 and Fig. 19. The penalty strength is such that nearly all the nonzero couplings are represented in the map (apart from three couplings with $|J_{ij}| < 0.3$ which are not shown). The map of couplings obtained here is roughly similar to the map of couplings obtained with the cluster expansion. It is worth noticing that, unlike the case with small penalty, the $S_{\rm MF}$ with large L_1 penalty reproduces essentially the large correlations through a sparse network of couplings, and therefore it seems that couplings are less disentangled from the correlations. For example, in the flicker map there are several largely connected sites which are not present in the small penalty strength map or in the SCE.

The reconstructed p_i, c_{ij}, and $P(k)$ are shown in Fig. 25, Fig. 26, and Fig. 28, for the values of $\gamma'_{\rm opt}$. As shown in Fig. 26 the very sparse network which is inferred for these values of $\gamma'_{\rm opt}$ allows for the reconstruction of the largest correlations. The reconstruction of the c_{ij} is quite good for CB, which is in agreement with the very sparse network of large couplings inferred via SCE. The c_{ijk}, however, are not well reconstructed, as shown in Fig. 27.

As a conclusion, the inference with the mean field entropy, which is very simple to compute, is able to successfully reconstruct the structure of the interaction graph of large couplings, especially when using a small Bayesian regularization strength. The magnitude of the strongest interactions, however, is often overestimated; a larger value of the regularization strength is useful for obtaining couplings that more closely match those inferred via SCE, at least for the L_2 norm regularization. The ability of these methods to reproduce the empirical correlations is limited. Notably, the L_1 regularization with large penalty parameter reconstructs quite well the spiking probabilities and pairwise correlations for the cortical data set CB, but not the higher order statistics.

8. Conclusion

We have described a Selective Cluster Expansion (SCE) to infer the fields and couplings which are the parameters of an Ising model from the measured spiking frequency and pairwise correlations of a population of neurons. We have described the performance of the SCE on retinal and cortical multielectrode recordings and checked that the Ising model is capable of reproducing some features of the recorded spiking statistics, *i.e.* the spiking probabilities in a time window of single, pairs, and triplets of cells, and the probability that k cells spike in the same time window. We have compared two different SCE procedures: the expansion of the entropy of the inferred Ising model given the measured spiking probabilities $S[\{p_i\},\{p_{ij}\}]$, and the expansion of the difference $S[\{p_i\},\{p_{ij}\}] - S_{\rm MF}[\{p_i\},\{p_{ij}\}]$ between the entropy S and a reference entropy, which

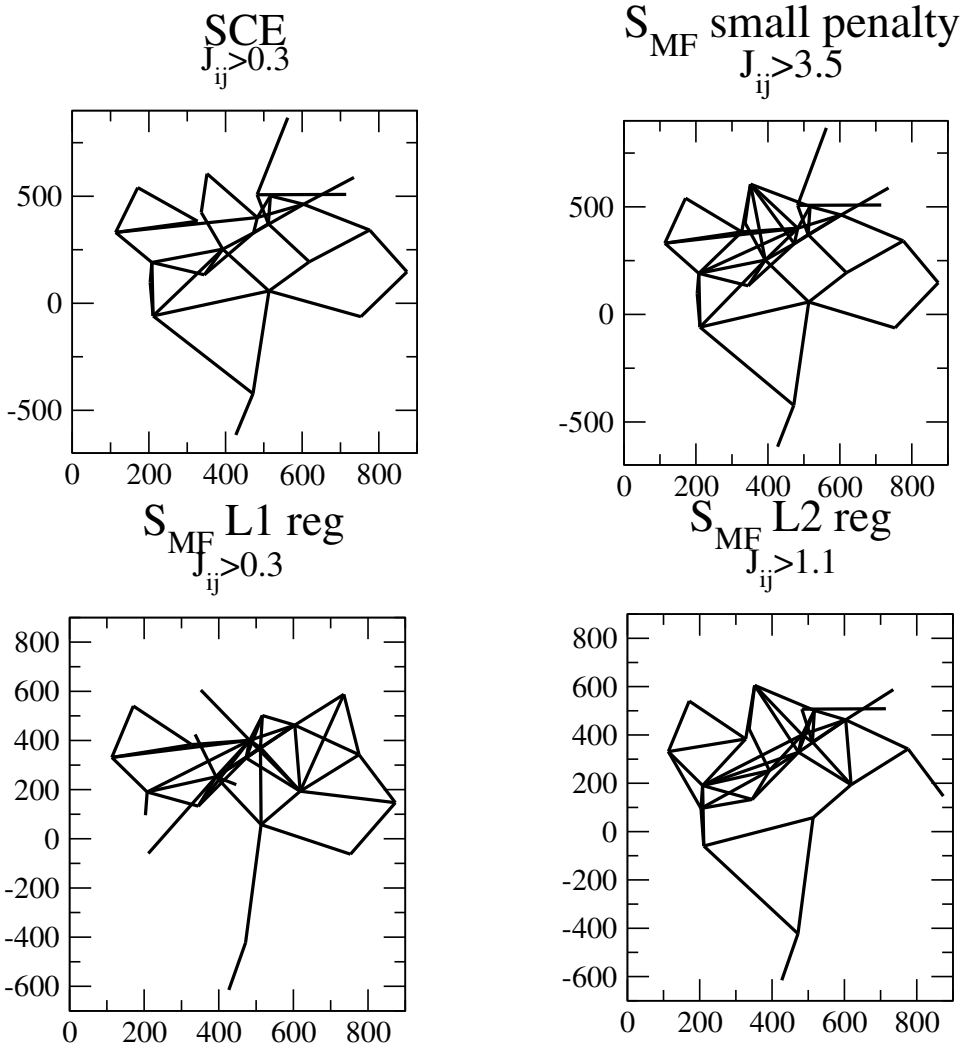

Figure 18. Map of inferred couplings in the receptive field plane of 32 retinal cells recorded in dark. Selective Cluster Expansion (top-left) reliable couplings with $|J_{ij}| > 0.3$; L_1- or L_2-regularized mean field model with small Bayesian penalty strength $\gamma = 0.003$ (top-right, same network for both choices of the penalty), the largest 48 couplings are represented $|J_{ij}| > 3.5$, L_1-regularized mean field model with large penalty strength (bottom-left) $\gamma' = 0.004$, the value of γ' is chosen to have 47 couplings $|J_{ij}| > 0.3$ which are plotted. L_2-regularized mean field model with large penalty strength (bottom-right) $\gamma = 5$, the largest 48 couplings are represented $|J_{ij}| > 1.1$.

we have chosen to be the mean field entropy $S_{\rm MF}$. We have also implemented the use of L_1 or L_2 norm in the regularizations of the entropies.

We found that the use of the reference entropy $S_{\rm MF}$ is helpful in retinal data, while large population ($N = 117$) cortical data are more easily processed without $S_{\rm MF}$. The difference comes from the different structure of the inferred interaction network. In cortical data the inferred interactions tend to be sparse with some neurons largely connected to the others, while the structure of the inferred retinal network is more

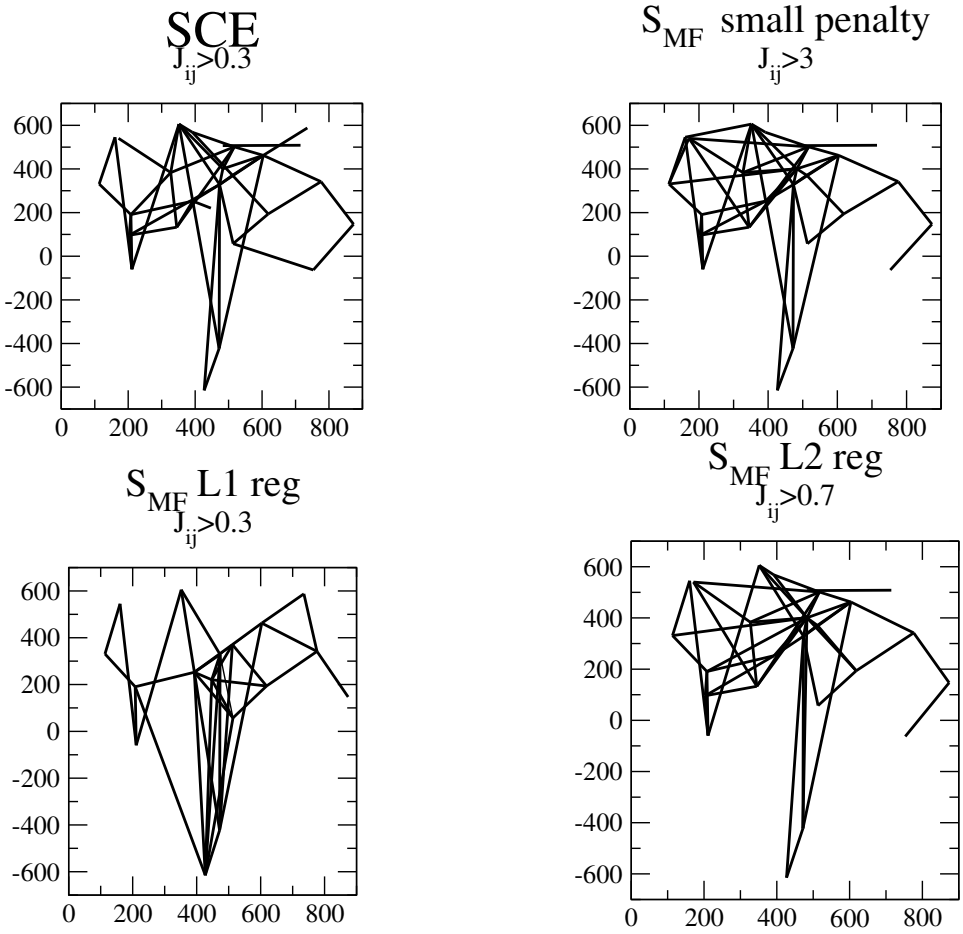

Figure 19. Map of inferred couplings in the receptive field plane of 32 cells recorded in flicker conditions. Selective Cluster Expansion (top-left), reliable couplings with $|J_{ij}| > 0.3$ are represented; L_2-regularized mean field model with small Bayesian penalty strength $\gamma = 0.0015$, with the largest 48 couplings represented $|J_{ij}| > 3$ (top-right), L_1-regularized mean field model with small penalty strength $\gamma' = 8\,10^{-4}$ the value of γ' is chosen to have 49 couplings with $|J_{ij}| > 0.3$ (bottom-left); L_2-regularized mean field model with large penalty strength $\gamma = 7.5$, with the largest 48 couplings represented $|J_{ij}| > 0.7$ (bottom-right).

homogeneous, with on average more connections per cell. It is important to underline that the structure of the inferred network also reflects the structure of the recording, as we discuss in more detail in the following.

The SCE has been sucesfully tested here on large populations, from ranging $N = 37$ to $N = 117$ cells. There are still some limitations which slow down the algorithm in the case of large data sets and which could be improved.

1 When too many clusters are selected the construction rule for new clusters becomes a time limiting step in the algorithm. This can happen, for example, when considering very low values of the threshold, particularly in the expansion of S without the reference entropy $S_{\rm MF}$.

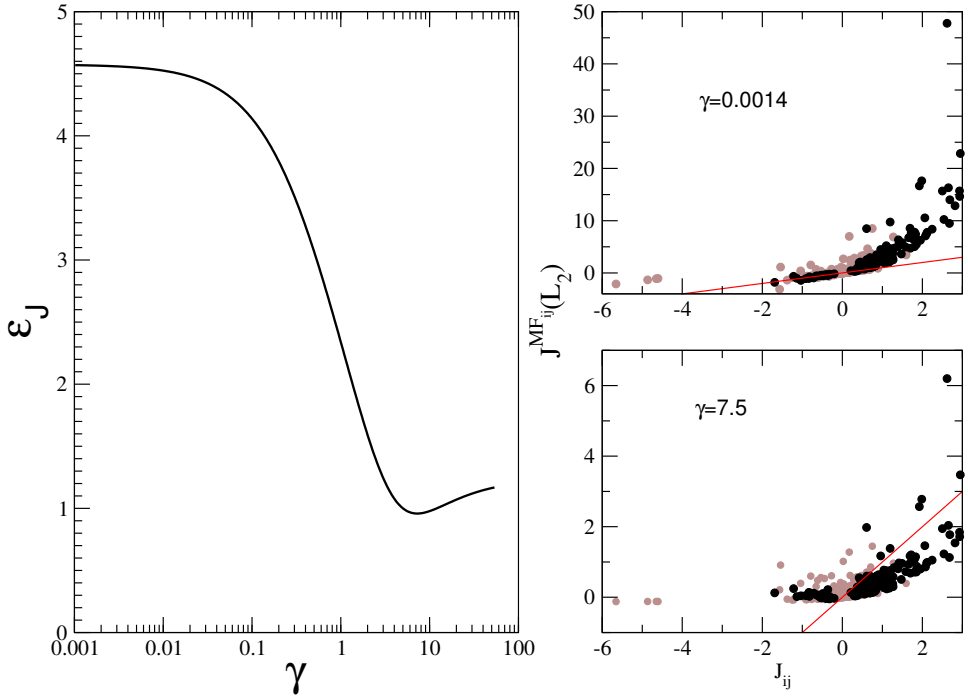

Figure 20. Performance of the L_2-regularized mean field model for the recording of 51 cells in the retina with a flickering stimulus (Fl). Left: mean differences between couplings obtained from the mean field model with the couplings found at T^* via the cluster algorithm, as a function of penalty strength. Right: couplings J_{ij} obtained with the mean field model for the optimal value in a Bayesian framework $\gamma \propto 1/B$ and with a larger penalty $\gamma = 7.5$, compared to the ones obtained with the cluster expansion.

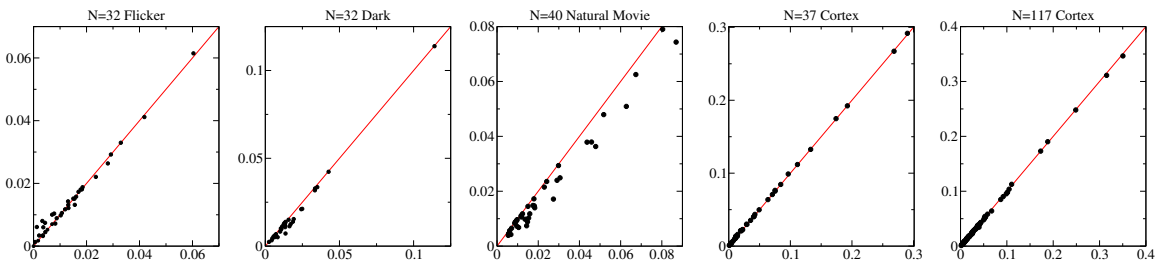

Figure 21. Reconstruction of the spiking probabilities $\{p_i\}$ with the Gaussian model and L_2 norm; the regularization strengths used are $\gamma = 7.5$ (Fl N=32), $\gamma = 20$ (Da N=32), $\gamma = 20$ (Nm), $\gamma = 0.00023$ (CA), $\gamma = 10$ (CB).

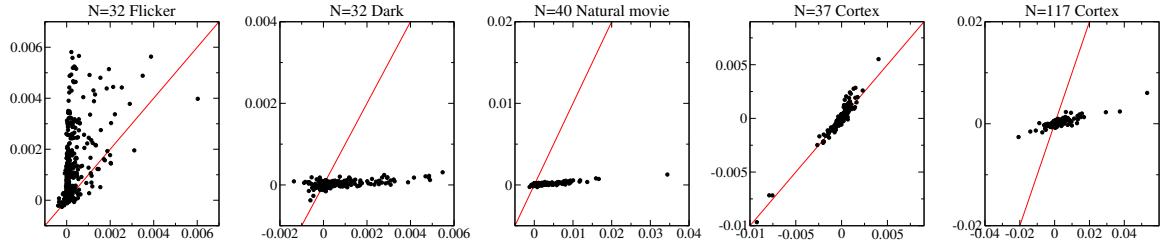

Figure 22. Reconstruction of pairwise connected correlations $\{c_{ij}\}$ with the Gaussian model and L_2 norm; the regularization strengths used are $\gamma = 7.5$ (Fl N=32), $\gamma = 20$ (Da N=32), $\gamma = 20$ (Nm), $\gamma = 0.00023$ (CA), $\gamma = 10$ (CB).

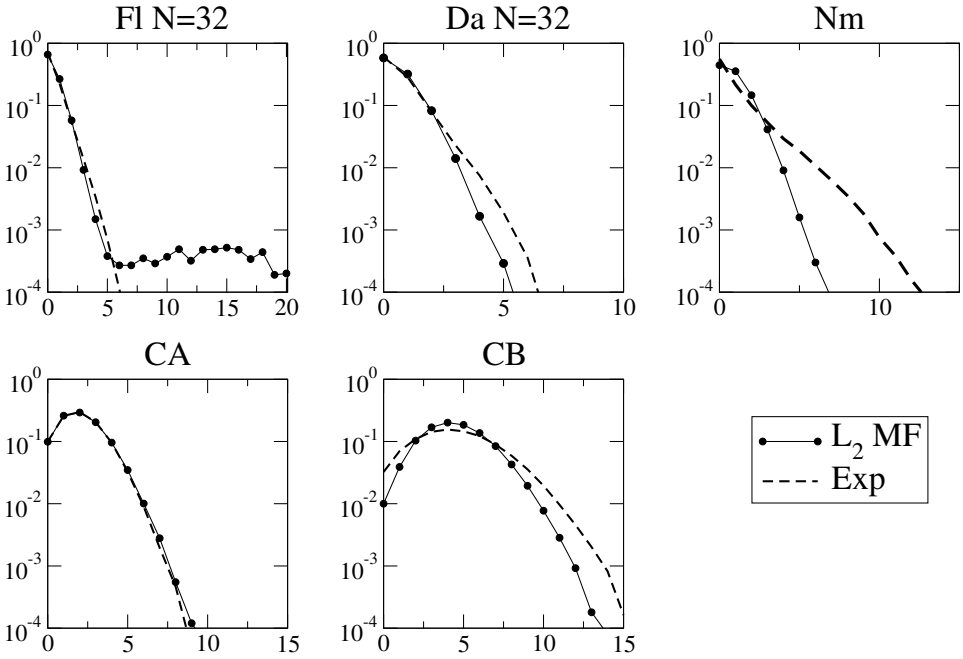

Figure 23. Reconstruction of the probability $P(k)$ that k cells spike in the same bin with the Gaussian model and L_2 norm; the regularization strengths used are $\gamma = 7.5$ (Fl N=32), $\gamma = 20$ (Da N=32), $\gamma = 20$ (Nm), $\gamma = 0.00023$ (CA), $\gamma = 10$ (CB).

2 The necessity of using a Monte Carlo simulation to test the reconstruction properties of the expansion at a fixed value of the threshold is one drawback of the SCE method. Some theoretical arguments to fix the optimal threshold or an approximation in the direct problem could be used to speed up the algorithm, expecially at large N where the thermalization of the Monte Carlo simulation can be slow.

3 Decreasing large fluctuations of the entropy expansion as a function of the threshold avoiding the cutting of packets of clusters sharing the same interaction path could improve the convergence of the entropy expansion.

Finally, we have compared the SCE results with the simpler inference from $S_{\rm MF}$ with L_1 or L_2norm regularization over a wide range of values of the regularization strength.

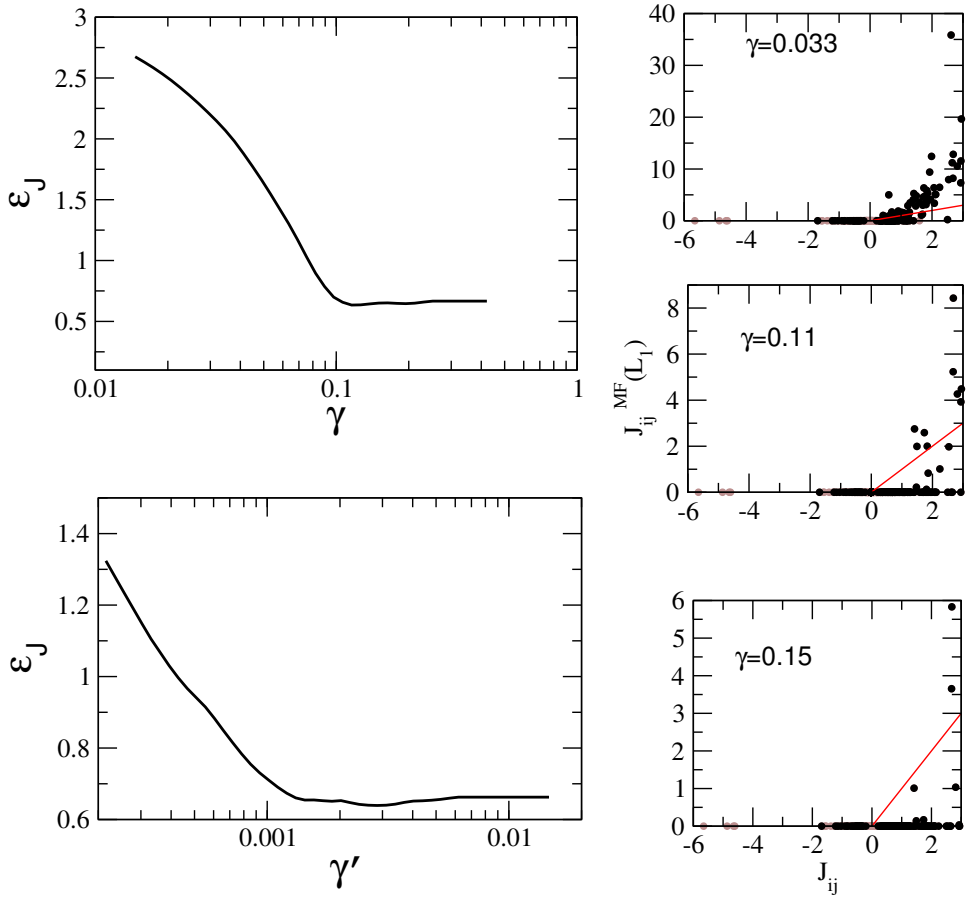

Figure 24. Performance of the L_1-regularized mean field model for a 32 cell subset of the recording of 51 cells in the retina with a flickering stimulus (Fl). Left: mean differences between couplings obtained from the mean field model using penalties γ (top) and γ' (bottom) with the couplings found at T^* via the cluster algorithm, as a function of penalty strength (see main text). Right: couplings J_{ij} obtained with the mean field model with various values of the penalty strength compared to the ones obtained with the cluster expansion.

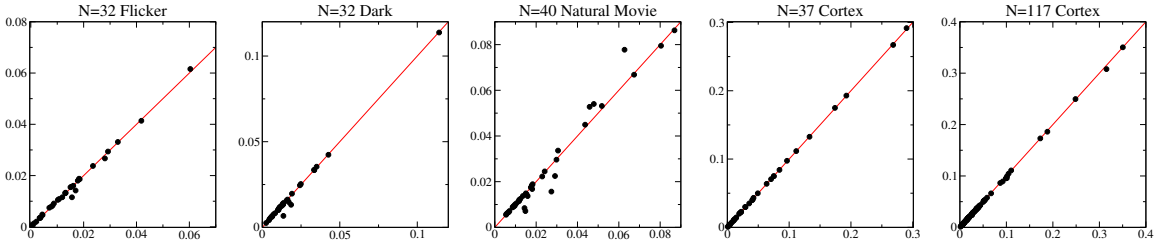

Figure 25. Reconstruction of the spiking probabilities $\{p_i\}$ with the Gaussian model and L_1 norm; the regularization strengths used are $\gamma' = 0.003$ (Fl N=32), $\gamma' = 0.004$ (Da N=32), $\gamma' = 0.0048$ (Nm), $\gamma' = 0.00012$ (CA), $\gamma' = 0.0029$ (CB).

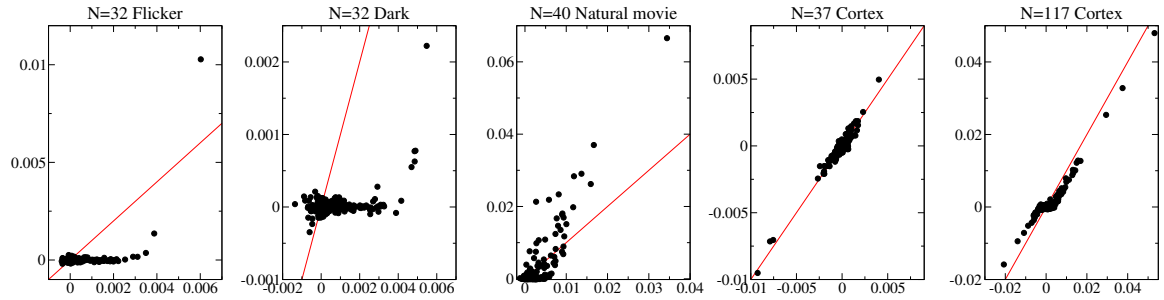

Figure 26. Reconstruction of pairwise connected correlations $\{c_{ij}\}$ with the Gaussian model and L_1 norm; the regularization strengths used are $\gamma' = 0.003$ (Fl N=32), $\gamma' = 0.004$ (Da N=32), $\gamma' = 0.0048$ (Nm), $\gamma' = 0.00012$ (CA), $\gamma' = 0.0029$ (CB).

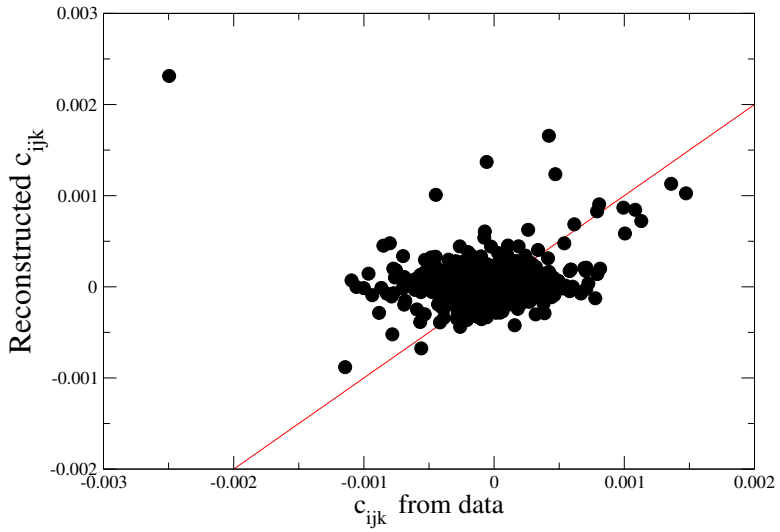

Figure 27. $N = 117$ Cortical recording, reconstruction of three cell connected correlations with the Gaussian model and L_1 norm; the regularization strength used is $\gamma' = 0.0029$ (CB).

$S_{\rm MF}$ corresponds to the entropy of a Gaussian model in which spins are treated as continuous variables. L_2 regularization penalizes large couplings preferentially, while L_1 regularization imposes a sparsity constraint on the inferred network for large values of the regularization strength. Fast methods of solution for either choice of the regularization are available, such as the Lasso method for the L_1 norm regularization [43].

We have found that generally the couplings and fields inferred using $S_{\rm MF}$ are not able to reproduce the spiking statistics, including the one and two cell spiking probabilities. The exception to this rule is the $N = 37$ cortical recording data (CA) in which the correlation indices are small, and therefore $S_{\rm MF}$ is a good approximation of S. However, as we have checked for the retinal recordings of 32 cells in dark (Da 32) and with a flickering stimulus (Fl 32), where the positions of the receptive fields are known, $S_{\rm MF}$ correctly reproduces the structure of the network of large couplings found with the SCE

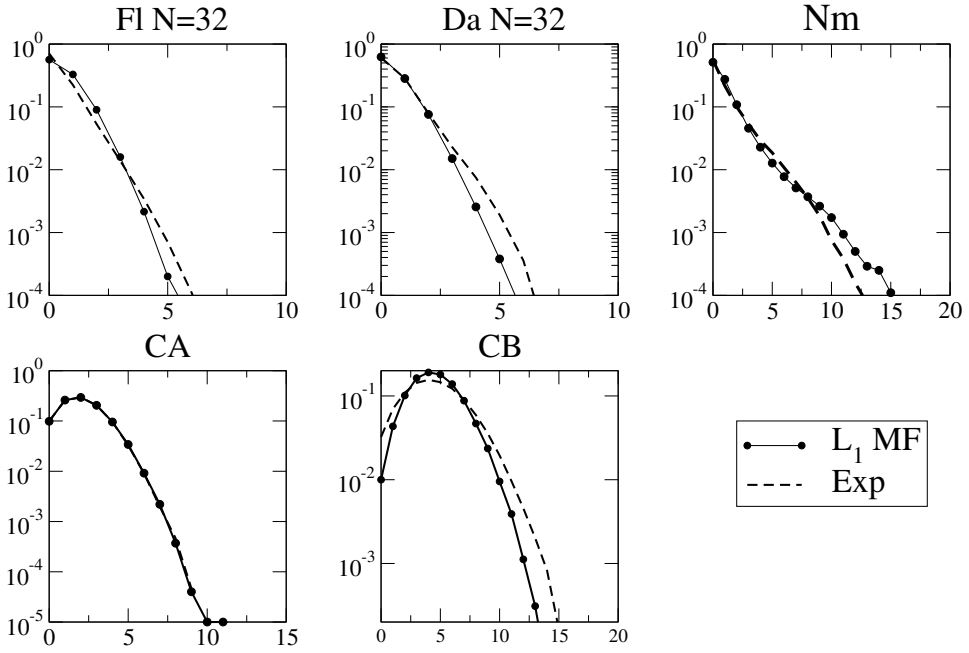

Figure 28. Reconstruction of probability that k cells spike in the same bin with the Gaussian model and L_1 norm; the regularization strengths used are $\gamma' = 0.003$ (Fl N=32), $\gamma' = 0.004$ (Da N=32), $\gamma' = 0.0048$ (Nm), $\gamma' = 0.00012$ (CA), $\gamma' = 0.0029$ (CB).

for small, Bayesian regularization strengths, and for large regularization strengths in the L_2 norm case. The L_1-regularized $S_{\rm MF}$ with a large value of the regularization strength is less effective in reproducing the structure of the network; on the other hand it infers a set of parameters which reproduce quite fairly the spiking frequencies and correlations for the $N = 117$ cortical recording (CB), but not the connected three cell correlations and the probability that k cells spike in the same bin $P(k)$.

8.1. Discussion

The potential long term applications of the inverse Ising approach to describing neural activity from data [5, 6, 50] are very interesting and promising, ranging from unveiling the structure of connections in a neuronal network [11] from data, to understanding, for example in the retina, how a stimulus is encoded [51], to interpreting learning experiments in cortical recordings [24, 32], to having predictive tools for the system from the recorded activity. Even if right now it is difficult to assess the utility and applicability of the stationary inverse Ising approach that we have discussed in the large scale analysis of neural data, different extensions of this model can be possible, some of which have already begun to be explored. We discuss in the following some limitations of the approach we have descibed and possible extensions.

8.2. Sampling problems

Data are affected by different limitations which could affect the outcome and the applicability of the inference.

1 The recorded area is limited and the recorded neurons are often a subpopulation of the whole population of cells. Indeed, this is often true because it is impossible to record all cells in a neuronal system. Even if the recorded area is limited, the inverse problem is well-posed when the inverse susceptibility is short-ranged (decaying within the recorded distance) as shown in [1, 2]. However, even if by solving the inverse problem we remove the network effects between the recorded cells and extract the direct couplings, these couplings are still effective couplings because of the influence of unrecorded cells on the system which is recorded. It would be important to study to what extent the inferred 'effective' models are still predictive of the behavior of the sampled population when, for example, a perturbation is applied to the system.

2 Apart from sampling errors due to finite recording time, which we have previously discussed, data are preprocessed by spike sorting [35, 52] and also in this process there are errors. It would be important to study how these errors affect the outcome of the inference.

3 In the Ising approach, as descibed above we do not take into account dynamic effects or temporal correlations between time windows. This corresponds to the fact that we fix a time bin Δt on the time scale of relevant correlations. But the relevant time scale can differ for different pairs of neurons, and delay effects can be present. The incorporation of dynamical effects can be improved in the Ising model, as in [25, 53].

4 The above Ising approach can be improved by taking into account a nonstationary external drive or stimulus. Explicitly including a time dependent external stimulus could help in removing couplings which arise from this external stimulus combined from the partial sampling of the population [35, 51, 54, 55]. It is crucial to study the importance of these external drive effects in the inferred couplings. In the process of including a time dependent external stimulus it is important to avoid the overfitting of the system by not increasing too much the number of parameters used to describe the system. To this extent it seems promising to combine the inference of the couplings with Generalized Linear models [51, 56] which include the description of the temporal response to a stimulus with a filter function which depends on a limited number of parameters. A more detailed description of the response to the stimulus could be also useful for removing spatial correlations between neurons induced by the structure of the stimulus, expecially in retina recordings.

Acknowledgments: We thank F. Battaglia and A. Peyrache, G. Buzsáki and S. Fujisawa, M. Meister, M. Berry for kindly makings their multielectrode recordings available to us. We thank also F. Battaglia, J. Lebowitz, R. Monasson, O. Marre, and

E. Speer for discussions and encouragements. We are also grateful to V. Oudovenko and G. Kotliar for the use of computing resources. This work was partially supported by the FP7 FET OPEN project Enlightenment 284801 and by NSF grant DMR-1104501 and AFOSR grant FA9550-10-1-0131.

Appendix A. Optimal choice for the regularization parameter γ

In this Appendix we discuss how the optimal value for the parameter γ in the L_2 regularization (22) can be determined. As explained in Section 2.2, the regularization term can be interpreted as a Gaussian prior P_0 over the couplings. Let us call σ^2 the variance of this prior. Parameters γ and $1/\sigma^2$ are related through

$$\gamma\, p^2\,(1-p)^2 = \frac{1}{2\,\sigma^2\,B}\ , \tag{A.1}$$

where we have assumed that the single site frequencies p_i are uniformly equal to p. To calculate the optimal value for γ, or, equivalently, for σ^2, we start with the case of a single spin for the sake of simplicity, and then turn to the general case of more than one spin.

Appendix A.1. Case of $N = 1$ spin

For a unique spin subjected to a field h the likelihood of the set of sampled spin values, $\{s^\tau\}$, is $P_h[s] = \exp(B\,p\,h)/(1+e^h)^B$. Here p denotes the average value of the spin over the sampled configurations (4). We obtain the *a posteriori* probability (19) for the field h given the frequency p,

$$P_{\rm post}[h|p] = \frac{\exp(-h^2/(2\sigma^2) + B\ p\,h - B\log(1+e^h))/\sqrt{2\pi\sigma^2}}{\mathcal{P}(p,B,\sigma^2)} \tag{A.2}$$

where the denominator $\mathcal{P}(p,B,\sigma^2)$ (marginal likelihood) is simply the integral of the numerator over all real-valued fields h. Given p and B we plot $I = -\log\mathcal{P}(p,B,\sigma^2)/B$ as a function of σ^2. The general shape of I is shown in Fig. A1. The value of σ^2 minimizing I is the most likely to have generated the data, and should be chosen on Bayesian grounds.

For more than one spin calculating the marginal likelihood would be difficult. We thus need an alternative way of obtaining the best value for σ^2. The idea is to calculate I through a saddle-point method, and include the Gaussian corrections which turn out to be crucial. This approach is correct when the size of the data set is large. A straightforward calculation leads to

$$I \simeq \log\,(1+\exp(h^*)) - p\ h^* + \frac{\Gamma}{2}\ (h^*)^2 + \frac{1}{2B}\log\left[1+\frac{1}{\Gamma}\,\frac{\exp(-h^*)}{(1+\exp(-h^*))^2}\right] \tag{A.3}$$

where $\Gamma = 1/(B\sigma^2)$ and h^* denotes the root of $(1+\exp(-h^*))^{-1} - p + \Gamma\ h^* = 0$. I decreases from $I(\sigma^2=0) = \log 2$ with a strong negative slope, $dI/d\sigma^2(0) \simeq -B\ p^2$, and increases as $\log\sigma^2/(2B)$ for large values of the variance. Expression (A.3) cannot be distinguished from the logarithm of the true marginal likelihood I shown in Fig. A1.

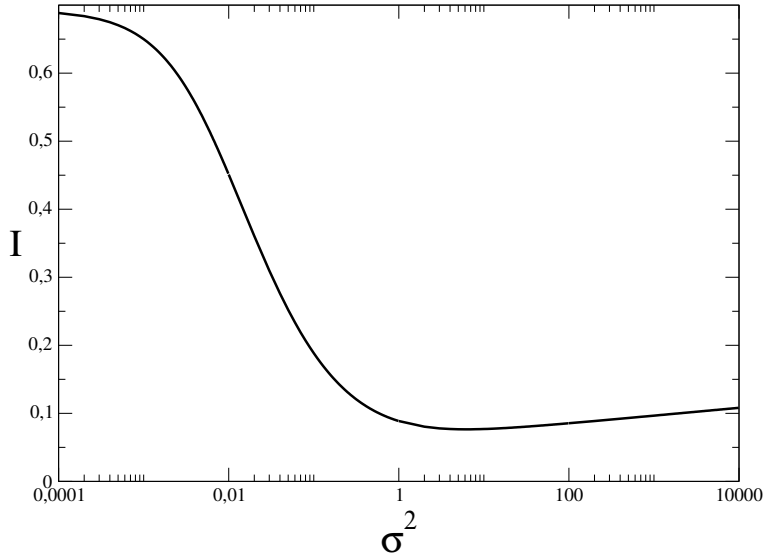

Figure A1. Logarithm of the marginal likelihood (with a minus sign, and divided by the size B of the data set) versus variance σ^2 of the prior distribution of the field. Parameters are $B = 100$, $p = .02$.

Appendix A.2. Case of $N \geq 2$ spins

The above saddle-point approach can be generalized to any number N of spins, with the result

$$I \simeq S^*[\{h_i\}, \{J_{ij}\}|\{p_i\}, \{p_{ij}\}] + \frac{1}{2B} \log \det \left(1 + \frac{\mathbf{H}}{\Gamma}\right) \tag{A.4}$$

where S^* was defined in (7) and χ is the $N+\frac{1}{2}N(N-1) = \frac{1}{2}N(N+1)$-dimensional Hessian matrix composed of the second derivatives of S^* with respect to the couplings and fields (24). In principle χ could be diagonalized and the expression (A.4) calculated. However this task would be time-consuming. As we have seen in the previous subsection we expect I not to increase too quickly with σ^2 (for not too small variances) and approximate calculations of I can be done under some data-dependent hypothesis. We now give an example of such an approximation, valid in the case of multi-electrode recordings of neural cell populations.

A simplification arises when the number B of configurations and the frequency p are such that: (a) each spin i is active ($= 1$) in a number of configurations much larger than 1 and much smaller than B *i.e.* $1 \ll B \times p \ll B$; (b) the number n_2 of pairs of spins that are never active together is much larger than one and much smaller than $\frac{N(N-1)}{2}$. These assumptions are generically true for the applications to neurobiological data. For instance, the recording of the activity of $N = 40$ salamander retinal ganglion cells in [5] fulfills conditions (a) and (b) for a binning time $\Delta t = 5$ ms: a cell i firing at least once in a time bin corresponds to $s_i = 1$, while a silent cell is indicated by $s_i = 0$. More precisely: (a) the least and most active neurons respectively fire 891 and 17,163 times (among $B = 636,000$ configurations); (b) $n_2 = 34$ pairs of cells (among 780 pairs)

are never active together.

Condition (a) allows us to omit the presence of Γ in the calculation of the fields, $h_i \simeq \log p_i$, to the first order of a large (negative) field expansion. Condition (b) forces us to introduce a nonzero Γ to calculate the couplings, with the result that interactions between pairs i, j of cells not active together are equal to $J_{ij} \simeq \log \Gamma + O(\log\log(1/\Gamma))$. Finally we obtain the asymptotic scaling of the entropy when $\Gamma \to 0$,

$$S^* \simeq n_2 \, \frac{\Gamma}{2} \, (\log \Gamma)^2 + O\Big(\Gamma \log \Gamma \log\log \frac{1}{\Gamma}\Big) \; . \tag{A.5}$$

We are now left with the calculation of the determinant in (A.4). From assumption (b) the number of pairs of neurons not spiking together is small with respect to N^2, meaning that most of the eigenvalues λ^a of the Hessian matrix of S^* are nonzero. Hence,

$$\log\det\Big(1 + \frac{\chi}{\Gamma}\Big) = \sum_{a=1}^{N(N-1)/2} \log\Big(1 + \frac{\lambda^a}{\Gamma}\Big) \simeq -\frac{N^2}{2} \log\Gamma. \tag{A.6}$$

Putting both contributions to I together we get

$$I(\Gamma) \simeq n_2 \, \frac{\Gamma}{2} \, (\log \Gamma)^2 - \frac{N^2}{4B} \, \log \Gamma \; . \tag{A.7}$$

The optimal value for the variance σ^2 is the root of

$$\frac{dI}{d\Gamma}(\Gamma) = 0 \simeq \frac{n_2}{2} \, (\log \Gamma)^2 - \frac{N^2}{4 \, B \, \Gamma} \simeq \frac{n_2}{2} \, (\log B)^2 - \frac{N^2}{4} \, \sigma^2 \; . \tag{A.8}$$

We finally deduce the optimal variance

$$\sigma^2 \simeq 2 \, n_2 \, \left(\frac{\log B}{N}\right)^2 \; . \tag{A.9}$$

For the data described above we find $\sigma^2 \simeq 8$.

References

[1] S Cocco and R Monasson. Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data. *Physical Review Letters*, 106(9), March 2011.

[2] S Cocco and R Monasson. Adaptive Cluster Expansion for the Inverse Ising Problem: Convergence, Algorithm and Tests. *Journal of Statistical Physics*, 147(2):252–314, March 2012.

[3] B L McNaughton, J O'Keefe, and C A Barnes. The stereotrode: a new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records. *Journal of Neuroscience Methods*, 8(4):391–397, 1983.

[4] M Meister, L Lagnado, and D A Baylor. Concerted signaling by retinal ganglion cells. *Science*, 270(5239):1207–1210, 1995.

[5] E Schneidman, M J Berry II, R Segev, and W Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[6] G D Field and E J Chichilnisky. Information processing in the primate retina: circuitry and coding. *Annual Review of Neuroscience*, 30:1–30, 2007.

[7] E T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

[8] D H Ackley, G E Hinton, and T J Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

[9] G Tkačik, E Schneidman, M J Berry II, and W Bialek. Ising models for networks of real neurons. *arXiv:q-bio/0611072*, q-bio.NC, November 2006.

[10] T Broderick, M Dudik, G Tkačik, R E Schapire, and W Bialek. Faster solutions of the inverse pairwise Ising problem. *arXiv:0712.2437*, q-bio.QM, December 2007.

[11] E Ganmor, R Segev, and E Schneidman. The architecture of functional interaction networks in the retina. *The Journal of Neuroscience*, 31(8):3044–3054, 2011.

[12] A Tang, D Jackson, J Hobbs, W Chen, J L Smith, H Patel, A Prieto, D Petrusca, M I Grivich, and A Sher. A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *The Journal of Neuroscience*, 28(2):505–518, 2008.

[13] S Pajevic and D Plenz. Efficient network reconstruction from dynamical cascades identifies small-world topology of neuronal avalanches. *PLoS Computational Biology*, 5(1):e1000271, 2009.

[14] P Ravikumar, M J Wainwright, and J D Lafferty. High-dimensional Ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[15] E Aurell and M Ekeberg. Inverse Ising inference using all the data. *Physical Review Letters*, 108(9):090201, 2012.

[16] V Sessak and R Monasson. Small-correlation expansions for the inverse Ising problem. *Journal of Physics A: Mathematical and Theoretical*, 42:055001, 2009.

[17] S Cocco, S Leibler, and R Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences*, 106(33):14058–14062, 2009.

[18] F Ricci-Tersenghi. The bethe approximation for solving the inverse ising problem: a comparison with other inference methods. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08015, 2012.

[19] M T Schaub and S R Schultz. The Ising decoder: reading out the activity of large neural ensembles. *Journal of Computational Neuroscience*, 32(1):101–118, June 2011.

[20] T Hastie, R Tibshirani, and J H Friedman. *The Elements of Statistical Learning*. Data Mining, Inference, and Prediction. Springer Verlag, 2009.

[21] J Sohl-Dickstein, P Battaglino, and M DeWeese. New Method for Parameter Estimation in Probabilistic Models: Minimum Probability Flow. *Physical Review Letters*, 107(22), November 2011.

[22] M J Schnitzer and M Meister. Multineuronal firing patterns in the signal from eye to brain. *Neuron*, 37(3):499–511, 2003.

[23] A Peyrache, K Benchenane, M Khamassi, S I Wiener, and F P Battaglia. Principal component analysis of ensemble recordings reveals cell assemblies at high temporal resolution. *Journal of Computational Neuroscience*, 29(1-2):309–325, June 2009.

[24] S Fujisawa, A Amarasingham, M T Harrison, and G Buzsáki. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nature Neuroscience*, 11(7):823–833, 2008.

[25] O Marre, S El Boustani, Y Frégnac, and A Destexhe. Prediction of Spatiotemporal Patterns of Neural Activity from Pairwise Correlations. *Physical Review Letters*, 102(13), April 2009.

[26] T Hosoya, S A Baccus, and M Meister. Dynamic predictive coding by the retina. *Nature*, 436(7047):71–77, 2005.

[27] R Segev, J Puchalla, and M J Berry II. Functional organization of ganglion cells in the salamander retina. *Journal of Neurophysiology*, 95(4):2277–2292, 2006.

[28] S W Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1):37–68, 1953.

[29] D W Arnett. Statistical dependence between neighboring retinal ganglion cells in goldfish. *Experimental Brain Research*, 32:49–53, 1978.

[30] D N Mastronarde. Correlated firing of retinal ganglion cells. *Trends in Neurosciences*, 12(2):75–80, 1989.

[31] I H Brivanlou, D K Warland, and M Meister. Mechanisms of concerted firing among retinal ganglion cells. *Neuron*, 20(3):527–539, 1998.

[32] A Peyrache, M Khamassi, K Benchenane, S I Wiener, and F P Battaglia. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience*, 12(7):919–926, May 2009.

[33] M A Wilson and B L McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679, 1994.

[34] E N Brown, R E Kass, and P P Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7(5):456–461, 2004.

[35] J S Prentice, J Homann, K D Simmons, G Tkačik, V Balasubramanian, and P C Nelson. Fast, scalable, Bayesian spike identification for multi-electrode arrays. *PLoS One*, 6(7):e19884, 2011.

[36] S Ostojic, N Brunel, and V Hakim. How connectivity, background activity, and synaptic properties shape the cross-correlation between spike trains. *The Journal of Neuroscience*, 29(33):10234–10253, 2009.

[37] J Shlens, G D Field, J L Gauthier, M I Grivich, D Petrusca, A Sher, A M Litke, and E J Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *The Journal of Neuroscience*, 26(32):8254–8266, 2006.

[38] T Plefka. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *Journal of Physics A: Mathematical and General*, 15(6):1971, 1982.

[39] A Georges and J S Yedidia. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, 24:2173, 1991.

[40] A Georges. Lectures on the Physics of Highly Correlated Electron Systems VIII: 8th Training Course in the Physics of Correlated Electron Systems and High-Tc Superconductors. *AIP Conference Proceedings*, 715:3–74, 2004.

[41] M Opper and D Saad. *Advanced Mean Field Methods*. Theory and Practice. MIT Press, February 2001.

[42] D J Thouless, P W Anderson, and R G Palmer. Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

[43] J Friedman, T Hastie, and R Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[44] A Pelizzola. Cluster variation method in statistical physics and probabilistic graphical models. *Journal of Physics A: Mathematical and General*, 38(33):R309–R339, August 2005.

[45] J Nocedal and S J Wright. *Numerical Optimization*. Springer Verlag, 1999.

[46] M Schmidt, G Fung, and R Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. 4701:286–297, 2007.

[47] I E Ohiorhenuan, F Mechler, K P Purpura, A M Schmid, Q Hu, and J D Victor. Sparse coding and high-order correlations in fine-scale cortical networks. *Nature*, 466(7306):617–621, 2010.

[48] M Weigt, R A White, H Szurmant, J A Hoch, and T Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[49] N Meinshausen and P Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[50] E Ganmor, R Segev, and E Schneidman. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of Sciences*, 108(23):9679, 2011.

[51] J W Pillow, J Shlens, L Paninski, A Sher, A M Litke, E J Chichilnisky, and E P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

[52] O Marre. Private communication. 2012.

[53] J Tyrcha, Y Roudi, M Marsili, and J Hertz. Effect of Nonstationarity on Models Inferred from Neural Data. *arXiv:1203.5673*, q-bio.QM, March 2012.

[54] K D Harris, J Csicsvari, H Hirase, G Dragoi, and G Buzsáki. Organization of cell assemblies in the hippocampus. *Nature*, 424(6948):552–556, July 2003.

[55] E Granot-Atedgi, R Segev, and E Schneidman. Stimulus-dependent maximum entropy models of neural population codes. *arXiv:1205.6438*, q-bio.NC, May 2012.

[56] Z Chen, S Vijayan, S N Ching, G Hale, F J Flores, M A Wilson, and E N Brown. Assessing neuronal interactions of cell assemblies during general anesthesia. *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 4175–4178, 2011.