Cell assemblies in neuronal recordings: identification and study through the inference of functional network models and statistical physics techniques
Gaia Tavoni

To cite this version:

HAL Id: tel-01358721
https://tel.archives-ouvertes.fr/tel-01358721
Submitted on 1 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
PhD Thesis
Discipline: Statistical Physics

Presented by
Gaia Tavoni

to obtain the degree of
Doctor of the École Normale Supérieure

Title

Cell assemblies in neuronal recordings:
Identification and study through the
inference of functional network models and
statistical physics techniques

Defended on October 30, 2015 in front of the commission composed by:

- Karim Benchenane  Member
- Simona Cocco  Thesis director
- Rémi Monasson  Thesis co-director
- Srdjan Ostojic  Member
- Stefano Panzeri  Referee
- Alessandro Treves  Referee
- Lenka Zdeborová  Member

Thesis prepared at LPS-ENS and LPT-ENS in the framework of the École
Doctorale “Physique en Ile de France”

Identification number of the thesis: 92564
Contents

Abstract 1

Foreword 3

Acknowledgments 5

1 Cell assemblies and memory 7
   1.1 Hippocampal cell assemblies ........................................ 8
      1.1.1 Phase precession and spatial encoding during motion .......... 9
      1.1.2 Replay of past trajectories .................................. 11
      1.1.3 Encoding of a flexible and multimodal cognitive map of the environment 13
   1.2 Cell assemblies in other brain areas ............................... 15
   1.3 Brain rhythms and global cell assemblies .......................... 17
      1.3.1 Synchronization during wakefulness ............................ 18
      1.3.2 Synchronization during sleep ................................... 19
   1.4 Towards a unifying theory of learning and memory .................. 21
      1.4.1 The two-stage model .................................... 22
      1.4.2 The hippocampal memory index theory .......................... 23
      1.4.3 Insights on memory storage and recall: the importance of sparseness in CA3 and DG ................................. 25
      1.4.4 The standard consolidation theory .............................. 27
      1.4.5 The transformation hypothesis ................................ 28

2 Methods for the study of neuronal networks 31
   2.1 Descriptive statistics of correlations ............................. 31
   2.2 Model-based methods .............................................. 33
      2.2.1 Maximum-entropy models ................................... 34
      2.2.2 Inference of ME models ..................................... 39
      2.2.3 Generalized linear models ................................... 52
      2.2.4 State-space models ......................................... 58
   2.3 Methods to identify cell assemblies and replay ..................... 61
      2.3.1 Some introductory examples .................................. 61
      2.3.2 Template matching for hippocampal cell assemblies .......... 63
      2.3.3 PCA-based methods and community detection techniques ...... 63

3 A new method to study cell assemblies 69
   3.1 Effective network model for the neuronal activity .................. 71
   3.2 Inference and validation of the model ............................. 71
   3.3 Choice of the time-bin width $\Delta t$ .............................. 74
3.4 Comparison of the coupling networks across the epochs 76
3.5 Null model for the coupling adjustment 81
3.6 Simulations of the inferred model 82
3.7 Coactivation of the ‘putative cell assemblies’ 85
3.8 Possible scenarios for cell assemblies across sessions 88
3.9 Comparison with PCA–based methods 88
3.10 Quantitative estimates of the replay 93
  3.10.1 A statistical interpretation of the replay 93
  3.10.2 Comparison with previous interpretations 94
  3.10.3 Replay as a function of time in session A 95
  3.10.4 Session-wide replay 97
  3.10.5 Null model for the replay 98
3.11 Insights on sessions A, B, C, D, E, F, G 100
  3.11.1 Session A (181014) 100
  3.11.2 Session B (200208) 107
  3.11.3 Session C (181021) 110
  3.11.4 Session D (150720) 115
  3.11.5 Session E (190228) 118
  3.11.6 Session F (181012) 120
  3.11.7 Session G (200209) 124
4 Simulations of the models in the presence of noise 127
  4.1 Results at T = 0 on slightly modified datasets 127
    4.1.1 Non-uniform inputs in the model of Sleep Post activity 131
  4.2 Simulations at T = 1 132
    4.2.1 Description of the method 132
    4.2.2 Study of the susceptibility maxima and minima 133
    4.2.3 Localized stimulation in the model of Sleep Post activity 138
  4.3 From T = 0 to T = 1 140
    4.3.1 Neuron susceptibilities on varying T 140
    4.3.2 Local susceptibility peaks vs. conditional average variations 141
  4.4 Discussion on the meaning of the method 142
    4.4.1 Is inference of the Ising model necessary to predict rare coactivation events? 142
    4.4.2 What does drive H represent? 144
    4.4.3 Properties of coactivating neurons 149
5 Inference and sampling of a Bernoulli-GLM 153
  5.1 Model inference and goodness-of-fit 153
  5.2 Interactions and spatio-temporal patterns 154
6 Correlation analysis of optogenetic cultures 159
Conclusions 165
Bibliography 171
Abstract

This thesis illustrates a research on cell assemblies, explored through inference methods and statistical physics techniques.

Cell assemblies, groups of closely connected, synchronously activating neurons, are thought to be the units of representation of information in the brain and their activation, in groups or sequences, is thought to underlie perception and high-level cognitive functions.

The first chapter is a review of some of the major discoveries in experimental research on cell assemblies, from Hebb’s intuition of their existence in ’49 to nowadays, passing through the discovery of place cells and hippocampal cell assemblies in ’71, of their role in spatial navigation, abstract spatial encoding, working memory and episodic memory; of their reactivation (replay), during both sleep and wakefulness, associated to memory consolidation; the discoveries of cell assemblies in many other regions of the brain; finally the studies on brain rhythms and on their role in timing and synchronizing the activity of cell assemblies across the brain. The last part of the chapter is dedicated to the theories on learning and memory that this large body of experiments has inspired: in particular, the advances of the systems consolidation theory, a theory about the consolidation and reorganization of memories in the brain during experience, will be presented, from the formulation of the hippocampal memory index theory in ’86 and the two-stage model in ’89, to the developments of the standard consolidation theory in the ’80s and throughout the ’90s, to the more recent hypothesis on the transformations that memory would undergo during consolidation, leading to progressive acquisition of knowledge.

The second chapter is a review of traditional and recent methods for the study of networks of interacting neurons: after a brief introduction on descriptive statistics techniques, the focus is moved onto model-based methods, which consist in mapping the observed spiking data onto abstract graphical models, representing specific structural and functional properties of the real neuronal system. Network models allows one to disentangle direct and indirect correlations (among the recorded neurons) and to assess the effects and the relative importance of different covariates on the neuronal activity; moreover they can be simulated to make predictions about non trivial functional properties of the real system and are in general much more powerful tools compared to descriptive correlation analysis. Different classes of both stationary and non-stationary network models are presented, together with the techniques used to infer their parameters from spike recordings. Particular attention will be given to the inverse Ising problem. The final part of the chapter is a review of state-of-the-art methods to detect and characterize cell assemblies in neuronal recordings and to quantify the phenomenon of replay. Indeed, brute force and exhaustive search for groups of neurons with strongly correlated firings is impossible due to the combinatorial number of possibilities, and the precise characterization of cell assemblies from experimental data requires development of specific techniques.

Chapters from the third to the sixth contain the original contributions of the research work presented in this thesis. The third chapter illustrates a new model-based method to
unveil cell assemblies from neuronal data. The approach is based on the inference of an Ising network of effective interactions between the neurons, which defines a probability distribution over all configurations of neuronal activity. The model is inferred from simultaneously recorded neurons in rat prefrontal cortex, during performance of a decision-making task, and during preceding and following sleep epochs. The probability distribution of activity configurations defined by the model not only reproduces the statistics of the data at the time scale of the inference (10 ms), but also allows exploration of multi-neuron activity patterns which appear at larger time-scales, during salient moments (unknown a priori) of the task and sleep phases, when external or internal inputs drive cell-assembly activations. These multi-neuron activity configurations, corresponding to cell-assemblies, are uncovered simulating the model in the presence of a global uniform drive: as the drive increases, regimes of higher global activity are explored, which accounts for both spanning over different time-scales and simulating a real input transiently feeding the system. Comparison of the inferred interaction networks and of the identified cell assemblies across the three experimental epochs reveals empirical rules for cell assembly modification and allows investigation of the role of learning in re-shaping cell-assemblies and the role of sleep in consolidating memories through replay. The model-based probabilistic framework is also exploited to get quantitative estimates of the replay.

While in the third chapter simulations of the model are performed at zero temperature (i.e., in the absence of noise) to extract the local maxima (or self-sustaining patterns) of the distribution of activity configurations, in the fourth chapter activity fluctuations around those local maxima are taken into account performing Montecarlo simulations of the model at $T = 1$: neurons of cell assemblies extracted with the zero temperature analysis have susceptibility peaks at close values of the drive in the analysis at $T = 1$, meaning that inclusion of noise does not significantly change the model predictions about the identity of the cell assemblies, thus validating the results of the third chapter. The second part of the fourth chapter contains a discussion on the significance of the external drive in the simulations of the neural networks, an interpretation about its possible implementations in prefrontal cortex, and a more general discussion on the meaning and potentiality of this model-based method.

In the fifth chapter, temporal ordering aspects of the neuronal activity are explored through the inference of a Bernoulli-generalized-linear model (GLM) from the same prefrontal cortex recordings. The GLM-couplings, differently from the Ising ones, are not constrained to be symmetric and potentially capture asymmetries in the interactions between the neurons. However, the GLM-couplings inferred from the prefrontal cortex data do not show significant asymmetries, and the distribution of spatio-temporal patterns generated by the inferred model with localized stimulations is also statistically symmetric over all possible orderings of the neurons in the task-related cell assembly, replayed during sleep. These results suggest that information in the prefrontal cortex is encoded in groups of neurons activating synchronously without a specific sequential order.

The sixth chapter moves away from the model-based methods representing the focus of the previous chapters and shows an application of descriptive statistics to the study of in vitro cultures of rat cortical neurons in an optogenetic setting. The effects induced by light stimulation on the genetically transduced cultures are studied through the cross-correlation histograms between the neurons and the comparison of the correlation indices before and after the stimulation periods.
The study that I present in this thesis has been developed under the supervision of and in collaboration with Simona Cocco and Rémi Monasson; the aspects of the research relative to the measure of the effective coupling potentiation and to the formula for the quantification of the replay have been developed with the collaboration of Ulisse Ferrari. The work was funded by the European FP7 FET OPEN project Enlightenment 284801 (“Exploring the neural coding in behaving animals by novel optogenetic, high-density microrecordings and computational approaches: Towards cognitive Brain-Computer Interfaces”), a consortium of 5 research institutions in Europe and Canada aiming at developing theoretical and experimental tools to identify and modify cell assemblies in real-time (http://enlightenment-fp7.eu/).

The neuronal data studied in chapters 3, 4, 5 of this thesis, consisting in multi-electrode recordings of the activity of tens of neurons in the prefrontal cortex of behaving rats, have been collected by Francesco Battaglia’s group, now at the Donders Centre for Neuroscience in Nijmegen, and have been previously analysed in [1–3]. The data studied in chapter 6, consisting of Micro-Electrode Arrays recordings of genetically transduced in vitro cultures of rat cortical neurons have been collected by Michele Giugliano’s research group at the University of Antwerp.

Acknowledgments

I warmly thank Simona Cocco and Rémi Monasson for the generous attention with which they have followed all the stages of this research and for countless suggestions and ideas that have made this work possible and have greatly helped my scientific growth. I am grateful to Francesco Battaglia and Adrien Peyrache for interesting discussions and suggestions about both theoretical and experimental aspects of this study, as well as for having provided the data from which my work started, and to Michele Giugliano and Rocco Pulizzi who have provided the in vitro recordings. I also thank Giuseppe Forte and Tommaso Brotto for helpful informatic tips, and Charles Fisher for interesting discussions.

Finally I want to thank all the friends with whom I have shared this PhD experience for unforgettable moments spent together in these three years, and for many chats about science as well as life: Tommaso Brotto, Tommaso Comparin, Caterina De Bacco, Eleonora De Leonardis, Clélia de Mulatier, Alexis Dubreuil, Giuseppe Forte, Silvia Grigolon, Thomas Gueudré, Alberto Guggiola, Anirudh Kulkarni, Manon Michel, Tomoyuki Obuchi, Volker Pernice, Arthur Prat-Carrabin, Corrado Rainone, Martin Retaux, Sophie Rosay, Tridib Sadhu and Dario Villamaina.
Chapter 1

Cell assemblies and memory

Cell assemblies are closely connected, synchronously activating groups of cells, which are thought to be the units of representation of information in the brain. Their existence was hypothesized for the first time by Hebb in the seminal work [4], where he defined a cell assembly as “a diffuse structure [...] capable of acting briefly as a closed system”. His cell assembly hypothesis proposes that chains of cell assembly activations (or “phase sequences”) are the neuronal substrate of perception and internal cognitive processes, such as thinking, planning, decision making and memory. The idea is based on the assumption, now known as Hebb’s rule and confirmed by the observed mechanisms of spike-timing-dependent plasticity (STDP), that if a cell A repeatedly excites a cell B, the synaptic connection between the two cells become strengthened. The hypothesized consequence of this process is that repeated coactivations of a group of neurons during behaviour induce the formation and stabilization of a cell assembly, whose later “reverberatory activity” can be maintained (at least transiently) by mutual excitation between the cells. Due to inter-assembly connections, activation of a cell assembly can then trigger a phase sequence and evolve according to intrinsic cortical dynamics, becoming substantially decoupled from external stimuli and supporting cognitive processes beyond simple stimulus-response associations. Hebb’s cell assembly hypothesis was revolutionary, in a period in which the behaviourist paradigm was dominant, and it has inspired neuroscience research for over half a century: from the theoretical works on network models, first of all the Hopfield network of auto-associative memory [5], in which the stored cell assemblies are attractors (stable patterns to which network activity evolves), to the numerous experimental studies providing more and more evidence of cell assembly organization in the brain [6].

Recently, G. Buzsáki has proposed a reader-centric definition of cell assemblies [7], pointing out that neuronal synchrony, which is at the core of the cell assembly concept, can be defined objectively only from the point of view of the downstream “observer-reader-classifier-integrator” mechanisms: neurons that fire within the window of the membrane time constant (i.e. the integration time) of a downstream reader neuron define a cell assembly, irrespective of whether assembly members are connected synthaptically or not, despite synaptic connectivity can help the stabilization of the assembly and its future recurrence. This definition is totally centered on the functional meaning of a cell assembly, interpreted as the measurable effect the assembly produces on a reader-actuator neuron. In the following, I will also show experimental studies focusing on the precise definition of the characteristic time-scale for integration of presynaptic spikes by reader neurons in a cell assembly neural code. Reader neurons are not necessarily isolated units, but may themselves be part of other cell assemblies, detected by other downstream neurons to form...
neural words”. The work illustrated in chapters 3, 4 and 5 of this thesis shares the fundamental ideas of both definitions of cell assemblies and provides a method to quantify these concepts and extract precise information about cell assemblies from neuronal recordings. In the following, I will review several experimental works showing the existence of cell assemblies and their functions in information processing and memory; the last part of the chapter will be dedicated to the currently most accepted theory on learning and memory.

1.1 Hippocampal cell assemblies

Most of experimental research on cell assemblies has been made in the hippocampus of rats and mice. Hippocampus is a subcortical structure present in humans and other vertebrates, which plays a major role in spatial navigation and declarative memory, the type of memory that can be consciously recalled, such as memory of experiences and facts. Fig. 1.1 is a representation of the anatomy of the entorhinal-hippocampal circuitry: cells in layers 2 and 3 of the entorhinal cortex (EC) project densely to individual granule cells of the dentate gyrus (DG) and have sparser projections to CA3 and CA1; each granule cell of the DG is connected with approximately 10–15 pyramidal cells in CA3 (and in turn each CA3 pyramidal cell receives inputs from about 50–100 granule cells); CA3 is characterized by auto-associative recurrent connections between its pyramidal neurons and also projects to the pyramidal neurons of CA1 (and to other subcortical structures); finally CA1 neurons have afferents mainly to the subiculum (sub), the terminal region of the hippocampus, which is connected to the deep layer of the EC, thereby closing the entorhinal-hippocampal-entorhinal chain. The EC receives highly processed information from other cortical areas, especially associational, perirhinal, and parahippocampal cortices, as well as prefrontal cortex, and forms reciprocal connections with them; therefore it represents the main interface of the hippocampus with the cortex.

Studies on this hippocampus have proliferated after the discovery reported in 1957 [9] that removal of the hippocampal formation and the surrounding medial temporal lobe components in an epileptic patient, known as patient H.M., caused severe orientation and memory impairments. The growing interest in hippocampal functions led to the discovery in 1971 [10] of a class of neurons in the rat hippocampus, which responded maximally whenever the animal was in a particular region of the environment. These neurons were called place cells and have been found in all regions of the hippocampus. The particular location in the environment where a place cell fires is called its place field. A given place cell has only one, or a few, place fields in a typical small laboratory room, but more in a larger environment [11]. The spatial distribution of the place cells does not reflect the spatial distribution of their place fields: unlike other brain areas such as visual cortex, neighboring place cells are as likely to have nearby fields as distant ones. All place cells with the same place field fire synchronously and can be regarded as members of a fundamental cell assembly. During motion, place cells with nearby place fields reach their maximum firing rate one after the other, reflecting the ongoing trajectory at the behavioural time-scale. Temporal compression mechanisms, both during motion and sleep, allow the sequence of place cell assemblies to activate in the same order but on shorter time-scales. A temporally compressed sequence of place cell assemblies can be regarded as a neural word [7] or simply as a larger cell assembly. Time compression during motion triggers plasticity processes that strengthen synaptic connections between the place cells in the sequence and allow the formation of an initial memory trace of the experienced
Figure 1.1: **Representation of the entorhinal-hippocampal circuit** (taken from Cajal and Cajal [8]). Cells in layers 2 and 3 of the entorhinal cortex (EC) project to granule cells of the dentate gyrus (DG), to cells in CA3 and in CA1; granule cells of the DG are connected with pyramidal cells in CA3; CA3 projects to pyramidal neurons of CA1; finally CA1 neurons have afferents mainly to the subiculum (sub), which is connected to the deep layer of the EC. The flow of information in this hippocampal-entorhinal circuit is largely unidirectional.

trajectory. As I will discuss later in this paragraph, the same sequence can be subsequently replayed during rest and sleep, supporting memory consolidation.

### 1.1.1 Phase precession and spatial encoding during motion

During motion, temporal compression is allowed by the concomitance of two factors: the overlap between nearby place fields and a mechanism called *phase precession*, reported for the first time in [12] and characterized in a detailed way in [13]. To understand this mechanism the first point to make is that since nearby place fields largely overlap with each other, the spike trains of the respective place cells are actually intermingled at a time-scale faster than the time needed to run the distance between the centers of the place fields. The activation order in each compressed sequence is maintained constant thanks to a precise phase relationship between place cell spikes and the hippocampal theta rhythm, a strong neural oscillation with a frequency of $\sim 6$–$10$ Hz, observed especially during active behavior in the hippocampus and many other brain structures (see 1.3). In particular, as the rat moves through a place field, the corresponding place cell assembly discharges earlier and earlier in successive theta cycles (see Fig. 1.2), and the theta phase of a place cell assembly at a particular time represents distance travelled through the place field [14]. Indeed, at faster running speeds the phase shift from one cycle to the next one is larger, suggesting that hippocampal place cell assemblies may be speed-controlled oscillators [15]. This phase precession, the underlying mechanism of which is still debated, enables the activation within each theta cycle of an ordered sequence of place cell assemblies, representing a segment of trajectory centered at the current location. Interestingly, the number of assemblies that can nest in each theta cycle ($\sim 7$, see Fig. 1.2) reflects the number of gamma cycles (faster neural oscillations) per theta period [16, 17]. This is a sign of the central role of brain rhythms in coordinating the activity of populations of neurons and in defining the syntax of cell assembly organization [7]. The number of cell
The dependence of spike times of hippocampal place cells on the animal position in space is well established; however there is evidence for cell assembly organization in the hippocampus reflecting not only external stimuli, but also internal processes which increase the synchronicity between cells beyond the coordination induced by spatial inputs: indeed prediction of the spike times of a place cell improves when the spike times of the other cells are taken into account, in addition to the animal location and the theta phase, and this prediction is optimal when a time dependence of 10–30 ms on peer activity is considered [18]. This time-scale is again coherent with the period of gamma oscillations, and also with the neuron membrane time constant (the time window for input integration) and with the time interval between spikes required to induce long-term potentiation in the synapses. Another signal of a non-trivial cell assembly organization in the hippocampus is provided in [19], where the authors point out that spike times of place cells at each theta phase during motion in a two-dimensional environment do not simply encode the location in space, but are the product of a more complex internal mechanism, reflecting the flow of information about position from the sensory areas to the hippocampus and from the hippocampus back to the sensory and motor areas to produce an output: this conjecture is based on the observation that spike times of place cells are predicted better from the immediate future or past locations of the animal (depending on the phase) than from the current location: according to [19] at the start of the theta cycle neuron spikes
1.1. HIPPOCAMPAL CELL ASSEMBLIES

reflect information about the immediate past locations, possibly due to the time needed for this information to reach the hippocampus; on the contrary, at the end of the theta cycle neuron spikes carry information about the immediate future locations, that is about how the world will appear when this information will become available to the sensory and motor areas.

As discussed in the following sections, the meaning of hippocampal cell assemblies goes beyond the representation of the animal current trajectory. Since their discovery, several experiments have shown that cell assemblies of place cells can also reflect past or upcoming trajectories, or even non-spatial features of the environment [20].

1.1.2 Replay of past trajectories

The reactivation of place cell sequences corresponding to previously visited trajectories is called replay. Replay can occur both in the awake state, during periods of relative immobility, and during sleep [21], usually in concomitance with sharp-wave-ripple (SWR) complexes. SWRs are brief (~100 ms) large amplitude waves in hippocampal local field potential (LFP, see par. 1.3) associated with fast field oscillations (150–250 Hz) and occur during a period of sleep, called slow wave sleep (SWS). It is thought that sharp waves can drive the network into attractor states [22], reflecting previous experiences.

Awake replay is predominant in brief periods of stillness during salient experiences, for example it is prevalent in a novel environment than in a familiar one [23], and soon after reward [24], possibly associating experiences with rewarding outcomes. Moreover it is more frequent immediately after an experience and decays with time [22]. Awake replay can not only reinstate trajectories within the current environment, but also trajectories previously experienced in another environment [21, 25]. Local and remote replay are often (but not always [26]) triggered by the local sensory input: place cells active at the animal’s current location can act as initiator cells, starting the reactivation of a sequence which moves away from the animal in the forward or reverse direction [27]. In other words, local inputs can act as cues for retrieval of a recent experience.

Forward replay (Fig. 1.3, top) is thought to be due to plasticity processes [28] which, during several one-way traversals of a linear track, strengthen the synapses between cells with neighboring place fields in the order in which they appear in the trajectory. As a result, when the first cell of the sequence is activated, firing is likely to propagate along these synapses, reinstating the forward sequence [29]. The proposed explanation for reverse replay is the excitation-spread mechanism (Fig. 1.3, bottom): when the animal arrives at the end of the track, excitability of place cells decreases in reverse order of their activation during the run; an input, such as that represented by sharp waves, can bring excitation of these cells above the firing threshold, starting with the most excitable neuron and progressing to the less excitable ones [30]. However this mechanism does not explain reverse replay of trajectories starting far away from the animal’s current location [31].

A first signature of sleep replay during SWR events after active behaviour was observed in [32] and from it a series of experiments has followed, showing that segments of place cell sequences activated during motion are replayed in the same temporal order during sleep [33]; moreover each segment is replayed within a single sharp-wave event (~ 100 ms) and is temporally compressed by 10–20-fold compared to the behavioural phase [34–36]. The neuronal mechanism inducing sleep replay may be analogous to that inducing awake forward replay, that is potentiation during behaviour of the synaptic connections corresponding to the experienced trajectories.
CHAPTER 1. CELL ASSEMBLIES AND MEMORY

Figure 1.3: Forward replay and reverse replay (adapted from Battaglia et al. [29]).

Forward replay (top): after several one-way traversal of a linear track, synapses between cells with neighboring place fields are unidirectionally potentiated in the order in which their place fields are traversed (from the red one to the green one). Consequently, when the first cell of the sequence is activated, firing can propagate along these synapses reinstating the forward sequence. Reverse replay (bottom): at the end of the track, excitability of place cells progressively decreases; an input (vertical dashed arrow) can bring excitation of these cells above the firing threshold, starting from the cell activated most recently (green) and ending with the cell activated early on (red).
Both awake and sleep replay, occurring on compressed time-scales, induce in turn long-term potentiation of the same synapses and are therefore crucial for consolidation of information about experiences into long-term memory [30, 32, 37]. A confirmation of the role played by replay associated with SWRs in memory consolidation is that disruption of SWRs during sleep after learning impairs performance in spatial memory tasks [38, 39]. Memory consolidation is thought to be concomitant with a transfer of its content from hippocampus to neocortex, as suggested by the synchronous occurrence of cell assembly replay in cortical circuits during SWR events [29, 36, 40, 41] (see par. 1.2). Recent studies propose that replay during sleep may also play a role in the formation of cognitive schemata and abstract concepts by mostly strengthening common (overlapping) elements of related memories [42] (see also par. 1.4), and in associating related memories together [43]. The reactivation of remote trajectories, experienced in a previous environment, may also serve as a mechanism to link aspects of past and present experiences.

Sleep replay has been mainly observed during SWS, but in [44] the authors report replay events during REM sleep. However in this case the time-scale of reactivation is not compressed compared to the behavioural time-scale, and the question of what is the function of REM replay is still largely open.

1.1.3 Encoding of a flexible and multimodal cognitive map of the environment

Cell assemblies in the hippocampus can also represent future trajectories in navigational tasks. In [45], the authors have observed that in a spatial alternation task on a W-track, synchronous neural activity during SWRs is stronger before correct, as compared to incorrect, trials and this coordinated activity represents both correct and incorrect possible future trajectories. This observation suggests another potential function of hippocampal cell assemblies: place cell sequences activated during awake SWRs may support memory-guided decision making, by recapitulating all possible choices at the decision point. Similar forward shifted trajectories on a spatial decision task have been observed in [46]; however in this case future paths represented at the decision point turns out to be concomitant with theta and gamma oscillations rather than SWRs, which seems to indicate that also other brain states may support activation of this kind of cell assemblies. In agreement with the hypothesis of a function in navigational planning, representations of future paths were found at locations where the rat paused and re-oriented, and their amount and content varied according to task demand. Another experiment reported in [26] shows that, during SWRs activity, trajectories representing routes towards a remembered goal location are more represented compared to random trajectories and they predict immediate future behaviour, while trajectories towards locations that are known to be unrewarded are even less represented than random trajectories. Interestingly, these goal-directed trajectories do not start at the animal current location and they can not be interpreted as replay events triggered by local inputs, while they may represent a non-trivial mechanism to help the construction of a task dependent cognitive map of the environment and to guide behaviour. Activation of sequences representing trajectories leading to (and away from) the goal has been observed also during sleep, when goal location is visible but inaccessible and therefore unexplored in the previous behavioural phase [47]. Such pre-activation of new trajectories, which will be explored in the future, is called preplay. Some experiments also report completely ‘de novo’ preplay of trajectories that have not even been seen during previous behaviour [48]: the explanation for the ‘de novo’ preplay is still debated, but it
has been proposed that a new experience can engage cell assemblies that are, at least partially, pre-configured in the place cell synaptic matrix, in such a way that the new sensory cues in the environment will be bound to those cell assemblies and encoded in them. However in [47] preplay is restricted to trajectories that the animal has already seen and associated with reward, and preplay seems to reflect thinking and planning of the future rewarded route. In [31] it is suggested that preplay of trajectories not yet experienced may be important for learning and maintaining a cognitive map of the entire environment.

The hypothesis of a role of hippocampus in building abstract and flexible cognitive maps of experiences dates back to the first studies by O’Keefe and Nadel [49] and has been strongly supported by the discovery of remapping [50]: place cells can start firing, stop firing, or change their place fields in response to changes in the environment, and these modifications are expressed extensively across the place cell population, such that a new map is established whenever a new situation is encountered. A few years after the discovery of remapping, Fenton and collaborators [51] pointed out that the variation in firing rate of a place cell across several traversals of the same place field substantially exceeded that of a random model with Poisson variance. Recently it has been shown [52] that this variability, or overdispersion, is correlated across the population of place cells, that is it reflects an ensemble-level modulation or a dynamical cell assembly organization. The origin of this overdispersion is precisely remapping. Place cell assemblies can switch between different maps according to task parameters, even when the environment does not change, and it seems that switches between maps are due to changes in transient goals: the simplest example is represented by navigation on a linear track, where motion towards the two different ends (which constitute transient targets) is represented by different maps in the place cell population. In [52] the authors observe that switches between maps also occur during foraging or goal directed tasks in two-dimensional environments. Interestingly, the map switching rate increases following reward in the goal directed task, where reward delivery shifts the navigation target, but not in the foraging task, where reward does not change the task goal. When a task requires to process information according to two competing reference frames (e.g. one stationary and the other rotating) the ensemble of place cells switch coherently between two self-consistent maps, with a preference for the map which represents the behaviourally most relevant reference frame [53]: this dynamical cell assembly organization is thought to be important for cognitive control of competing information streams.

Remapping can also involve firing rates only: in particular, it has been observed [54] that when the place of the recording chamber is changed, place cells undergo a global remapping, in which both their firing rates and place fields change, while when the recording chamber is varied (e.g. in shape or wall color) but its location is kept constant, place cells maintain their place fields but, especially in CA3, they undergo rate remapping, that is their firing rates change substantially. Moreover, in two distinct rooms with common spatial elements, firing rates and place field positions of CA1 neurons are correlated and the overlap between the two representations increase with increasing similarity between the enclosures; on the contrary, there is no overlap in the activated populations of CA3 cells [55]. Both global and rate remapping can be observed upon transient changes in task targets [52]. Rate remapping may be the neuronal mechanism by which hippocampal cell assemblies support both a spatial and a non-spatial code for episodic memory [56]: there is increasing evidence that strictly spatial information is encoded by the neuron place fields, and on top of this representation non-spatial variables are often encoded by the firing rate
of the same place cells. The extent to which each kind of coding is expressed seems to vary according to the type of cues (spatial or non-spatial) that are emphasized in the task and behaviorally more salient [57, 58]. The non-spatial variables represented by hippocampal cell assemblies can be not only visual (like shape and color [59]), but also tactile [57] and olfactory [58, 60]; and in [61] the authors also report the presence of place cells responding to odor valence rather than odor identity, in a task in which the rat had to discriminate positive and negative odors. Olfactory signals can also help the formation of a stable map of place fields when visual information is lacking, e.g. in dark environments [62, 63].

All these experimental observations strongly support the theory that hippocampal cell assemblies play a central role in abstract spatial encoding and in linking spatial with non-spatial information about events to form coherent representations of experiences.

### 1.2 Cell assemblies in other brain areas

Cell assemblies are not only found in the hippocampus but seem to be ubiquitous in the brain. Since Hebb’s cell assembly theory, the hypothesis that cortical cognitive functions may depend on the propagation and transformation of synchronous activity sequences (called synfire chains [64] and reminiscent of Hebb’s phase sequences) has been further investigated both from a theoretical point of view in models of neural networks [65, 66] and from an experimental point of view.

In [67], the authors report the presence of repeated spatio-temporal patterns in frontal areas of behaving monkeys. Repeated synchronous sequences have also been observed in the spontaneous activity of populations of neurons in slices from mouse visual cortex [68] and in cat primary visual cortex in vivo [69]. These sequences have a specific topographic structure: differently from sequences of place cells representing trajectories, which generally involve non-neighboring place cells, cortical synfire chains observed in [69] are formed by neurons which occupy the same cortical layer or vertical column, or which are in other ways spatially clustered. Moreover series of sequences are also repeated several times in the same temporal order, forming modular assemblies, the so called cortical songs. Repetition of both sequences and series of sequences are more and more compressed in time, probably reflecting synaptic potentiation processes which progressively increase neuron synchrony. These observations suggest that the neocortex can indeed generate spontaneously a temporally precise dynamics of cell assembly activation, similar to that hypothesised by Hebb.

An indirect confirmation seems to come from a study of the local field potentials in the inferior convexity of the macaque prefrontal cortex (icPFC) [70], which identifies a phase gradient in coherent oscillatory activity across different spatial locations within this region. This gradient unveils the presence of travelling waves of electrical activity, which may reflect highly coordinated cortical processing.

Another well known example of internally induced neural words are the cell assembly sequences recorded in the high vocal centre of birds, which generate the stereotypical bird songs [71].

Neuronal synchrony is also observed in response to specific stimuli. For example, populations of neurons within orientation columns in cat visual cortex respond with a synchronous oscillatory activity in the gamma frequency range (25–65 Hz) to optimally oriented moving bars [72]. Synchronous activity with a fast oscillatory temporal structure in response to specific stimuli has been observed in many other studies of the visual cortex,
like [73], where the authors show that in early stages of visual processing synchronization of oscillatory responses, rather than neuron firing rates, correlates with signal perception in binocular rivalry. In the middle temporal area of monkey visual cortex, long-range synchronization with a gamma oscillatory structure seem to encode the concept of stimulus coherence according to Gestalt criteria [74, 75]: neuron groups which respond independently to two non-aligned contours engage in a synchronous oscillation when stimulated with a single contour, while neuron firing rates are again not affected by stimulus coherence. These studies support the hypothesis that synchronization in visual cortex represents the coding paradigm of relations between elementary features of the visual scene and is important to form coherent perceptions [76].

Examples of externally triggered cell assembly sequences are also found in the olfactory system: transient gamma oscillations are induced in the antennal lobe of insects in response to odor stimuli, with different groups of neurons firing in each gamma cycle; multiple presentations of the same odor reliably trigger the same sequence of cell assembly activations, while different odors elicit different sequences [77, 78].

In the superficial layers of mice auditory cortex [79], cell assemblies are spatially localized: neurons responding to a particular sound are spatially clusterized and each cell assembly is segregated (in its center of mass) from other cell assemblies responding to different sounds, though a neuron can take part in more than one cell assembly. Moreover these assemblies undergo a discrete dynamics when the stimulus is a weighted superimposition of two sounds, known to elicit the activation of different neuron ensembles: at each moment, only one neuron group responds, and for a particular value of the weight the activity abruptly shifts to the other group. Spatial distance between different cell assemblies represents perceived dissimilarity between sounds and discrete cell assembly dynamics reflects classification of sounds into discrete categories. As noticed in [80], this kind of cell assembly organization is likely to be specific of the superficial cortex, while in deeper layers neural activity seems to be more spatially distributed.

Studies on population coding in auditory cortex have come to coherent and complementary results: in [81], the authors prove that, while the mutual information conveyed about the stimulus by a randomly sampled population of neurons increases monotonically with population size, the mutual information conveyed by an optimized, maximally informative population reaches its maximum at a relatively small size, denoting that a code based on small groups of neurons (cell assemblies) is potentially an optimal code.

Overlapping cell assemblies are also present in motor cortex: in [82] the authors report synchronized activity in the primary motor cortex of monkeys at specific moments of a sensorimotor task, when a behaviourally relevant signal occurs, inducing a motor response, or when such a signal is expected to occur. When synchronization is concomitant with signal occurrence, it is accompanied with an increase in neuron firing rates, while when it is concomitant with signal expectancy, firing rates remain constant. In analogy with the place coding and the rate coding in the place cell population, in the motor cortex synchronization and firing rate modulations seem to operate as complementary codes, which permit to process different kinds of information at the same time, such as the behavioural relevance of the signal and its internal vs. external origin. Complementary codes like these may be a mechanism to increase the representational power of neuronal ensembles. In other recordings in primary motor cortex of monkeys [83] during a task consisting in pointing to target directions, synchrony has been observed to be unrelated to the neuron tuning properties: neurons responding maximally during movement preparation and neurons responding maximally during movement execution can be synchronized significantly over
1.3. BRAIN RHYTHMS AND GLOBAL CELL ASSEMBLIES

Brain rhythms are oscillations observed in the local field potential (LFP), that is the electric potential recorded (typically using micro-electrodes) in the extracellular brain.

chance level at the end of the first period, i.e. the beginning of the second; conversely, no spiking synchrony was observed in this study between neurons with similar tuning properties.

In medial prefrontal cortex (mPFC) of rats during a working memory task involving odor-place matching, cell assemblies have been found to have features similar to hippocampal cell assemblies of place cells [84]: some mPFC neurons (both pyramidal cells and interneurons) have been shown to fire preferentially at specific locations of the maze, in particular in one of the two arm (either the left or the right), and while the rat is smelling, at the beginning of the task, the odor of the food that he knows he will find at the end of the same arm. Moreover, similarly to place cells, these neurons tend to fire sequentially, covering, one after the other, the entire trajectory from the starting box to the end of either the left or the right arm. Such cell assemblies may therefore encode a representation of the goal and of the trajectory to reach it. In [84] it is suggested that this kind of cell assembly chain dynamics, in which each elementary cell assembly, after a relatively short activation time (corresponding to a specific location in the maze), transfers its information content to another transiently active cell assembly, may be explained by short-term synaptic plasticity processes. Functional synaptic efficacy (i.e. short-latency correlations between pre- and post-synaptic neurons) is indeed observed to vary as a function of the rat’s position in the maze, probably reflecting facilitation/depression mechanisms dependent on the spiking history of the pre-synaptic neuron and the supralinear effect of coincident firing of presynaptic neurons on the activity of a postsynaptic neuron. These mechanisms have the potentiality to generate the sequential activation of cell assemblies observed during the task.

In [1, 2], Peyrache and collaborators have also found behavioural correlates of cell assemblies in mPFC of rats during learning of a rule in a Y-maze. They identify cell assemblies as correlation modes in the neuronal activity with principal component analysis and they show that the first principal component (PC1) of the correlation matrix during task execution defines a cell assembly which is mostly active right after trial onset, while the second principal component (PC2) activates just before the central platform and the third principal component (PC3) activates later on. On a longer time-scale, PC1 and PC2 increase their activity as the rat changes his strategy to solve the task, while PC3 decreases its activity. Moreover, the authors show that the cell assembly defined by PC1 is strongly replayed during sleep after the task, and this reactivation is triggered by hippocampal SWRs. However, by doing a reverse analysis, they show that the principal correlation pattern of SWRs events of sleep post task is represented by neurons mostly active when the rat is at the decision point of the maze after rule learning. As discussed better in the next paragraph, in [3] the authors show that these periods coincide with periods of increased coherence between mPFC and hippocampal theta rhythms, i.e. increased communication between these two structures.

We have re-analyzed these data and extended these results, using very different techniques, illustrated in 3, 4 and 5. More details about this experiment and the techniques used in [1, 2], as well as in other studies described in this chapter, to identify cell assemblies and replay will be provided in 2.

1.3 Brain rhythms and global cell assemblies

Brain rhythms are oscillations observed in the local field potential (LFP), that is the electric potential recorded (typically using micro-electrodes) in the extracellular brain.
tissues, reflecting the sum of the local synaptic currents. Brain rhythms, both during behaviour and during sleep, are thought to play and important role in orchestrating the activity of cell assemblies in spatially widespread brain structures.

1.3.1 Synchronization during wakefulness

As already mentioned in par. 1.1, during active behaviour (and also during REM sleep) hippocampal activity is entrained by theta waves, oscillations with a frequency range of 6–10 Hz, which are probably driven by extrinsic generators (burst firing patterns of cells) located in the medial septum and in the EC [85, 86], together with intrinsic generators in the CA1 region of the hippocampus [87]. EC layers 2 and 3 feed input to the CA3 and CA1 regions of the hippocampus respectively (Fig. 1.1), where the firing preferences of place cells move towards phases which are earlier and earlier in the theta cycle compared to the firing phase of the EC neurons, as the animal crosses each place field [29]; therefore hippocampal firing dynamics seems to be initiated by entorhinal inputs and to lately evolve independently from them through mechanisms, like phase precession, which are closely related to the theta rhythm. Together with cell assembly sequences built up through phase precession, groups of strongly synchronized neurons, possibly carrying non-sequential information, have also been observed in the hippocampus [18], as previously noticed: these Hebbian-like cell assemblies repeatedly activate at the troughs of theta cycles, on times-scales (∼30 ms, corresponding to a gamma period) shorter than those typical of place cell sequences (which typically span ∼7 gamma periods).

Timing of activity in medial prefrontal cortex can also be biased by theta rhythm. A significative example is illustrated in [3]: in a task in which a rule has to be learned in a Y-maze, theta-coherence between mPFC and hippocampus peaks at the decision point and after learning; interestingly, during these high theta-coherence periods, pyramidal neurons in mPFC shift their firing phase towards the troughs of the theta cycle, probably due to increased efficacy of interneurons. As a result, highly synchronized cell assemblies emerge in mPFC cortex, which match the theta phase of Hebbian cell assemblies observed (in other experiments) in the hippocampus. This finding suggests that hippocampal theta rhythm may play a central role in coordinating activity in cortical structures, and in generating global, inter-structure cell assemblies, which reflect multiple features of episodic memories (see also par. 1.4).

This hypothesis is supported by the observation that not only PFC, which has monosynaptic excitatory connections with the hippocampus, but also other neocortical areas, even many synapses away, are modulated by hippocampal theta oscillations: for example, both pyramidal cells and interneurons in parietal cortex of rats and mice show theta phase-locking during running on a track and REM sleep [88]. However, in this region neurons preferentially fire at the peak/descending phase of the hippocampal theta cycle, that is with a phase shift compared to neurons in PFC. LFP gamma oscillations, recorded locally in different areas of parietal cortex, are also modulated and linked together by the hippocampal theta rhythm. Several studies show that coherent gamma oscillations in many different brain regions correlate with learning of associations between different sensory stimuli: gamma coherence in the frequency range of 20–40 Hz has been observed, for example, in entorhinal and hippocampal activity in rats during encoding and retrieval of olfactory-spatial associative memory [60], and in human visual and somatosensory cortex during a visuo-tactile classical conditioning task [89].

Brain rhythms may be important not only to coordinate activity in different brain
regions, but also to provide an internal reference frame for decoding responses relative
to sensory stimuli: it has been shown in [90] that a decoding scheme which exploits
the phase angle of spike times with respect to theta oscillations is much more effective in
discriminating different stimuli than a spike count decoding scheme (based on the total
number of spikes in response to each stimulus) and achieves performance similar to a
decoding scheme based on the time elapsed from stimulus onset. Phase intervals within
theta oscillations could represent integration epochs for downstream neurons, thereby
providing an internal clock for decoding the temporal dynamics of sensory inputs. The
authors illustrate the high performance of the phase-based decoding scheme in auditory
and visual cortices and attribute the success of this mechanism to the alignment of theta
oscillations to sensory stimuli, observed in these areas: thanks to this alignment the phase
angle represents indeed an intrinsic copy (generated by the network itself) of the stimulus
time reference.

1.3.2 Synchronization during sleep

Synchronization between hippocampal and cortical activity has been observed also
during sleep, though brain rhythms allowing activity coordination are different in this case.
As shown in Fig. 1.4, hippocampal-cortical communication takes place during SWS, when
neocortex engages in slow oscillations (<1 Hz) between periods of generalized elevated
activity (“up-states”) and silence (“down-states”); up-states, the onset of which is probably
facilitated by firing of neurons in the locus coeruleus [91], can encompass slightly faster
oscillations (2–4 Hz), called delta-waves, and bursts of oscillatory activity (7–14 Hz),
called sleep spindles. In the hippocampus, SWS is characterized by the occurrence of
brief sharp-wave-ripple events, associated with the replay of past activity, as seen in
section 1.1.2. Replay has been widely observed in cortex as well, immediately after
hippocampal sharp-waves.

A first signature of hippocampal/cortical coordination during SWS is represented by the
co-occurrence of hippocampal ripples and cortical spindles: in [92], the authors show that
the cross-correlations between ripples in the hippocampus and spindles in rat prefrontal
and visual cortex have a peak very close to zero delay, but with a slight asymmetry in
the tails, indicating that ripples often precede spindle–ripple events; this is also reflected
in the firing of single neurons: a correlation between hippocampal and cortical neurons
is observed close to ripple-spindle episodes, with a tendency for hippocampal spikes to
immediately precede cortical neuron spikes.

Though the origin of hippocampal/cortical communication is debated and some studies
seem to fully identify this origin in the hippocampus [93], recent works have revealed, by
precise temporal analysis, that neocortex can also affect hippocampal activity; in particular,
several neurons in the DG, in CA1 and in CA3 are modulated by cortical up-down states:
the membrane potentials of DG granule cells and CA1 inhibitory interneurons are phase-
locked to neocortical up-down states with a small delay; coherently CA1 pyramidal cells
show an up-down state modulation of opposite sign [94]; CA3 pyramidal neurons show
significant, but mixed up-down state modulation, with some cells depolarized and other
cells hyperpolarized during cortical up-states, probably reflecting a different balance
between excitatory entorhinal and CA3 recurrent inputs on the one hand and inhibitory
inputs from CA3 interneurons on the other hand [94, 95]. Moreover, SWRs mostly occur
after the onset of cortical up-states [96, 97]. One hypothesis, suggested by Battaglia et al.
[29] (see Fig. 1.4), is that a strong excitatory drive during up-states is conveyed from
neocortex to CA3, via the entorhinal cortex; in the CA3 region, which is characterized by an auto-associative network of recurrent synapses [98], this cortical drive may induce activity bursts, constituting sharp-waves; sharp waves then propagate back to CA1 and outside the hippocampus, triggering in turn both hippocampal and cortical (probably spindle-correlated) replay. A more complete discussion of possible cortico-hippocampal interactions during both memory encoding and consolidation will be the focus of next paragraph; the idea that excitatory inputs, like sharp waves or the cortical drive in the up-states, may induce the very rapid activation of neuron ensembles, as observed during replay, is also the fundamental inspiration of the method we developed to unveil cell assemblies in neuronal recordings (see chapters 3,4). Some experiments have found a connection between spindles and cortical replay: similarly to what observed in human EEG, spindle density has been shown to increase during the first hour of sleep in rats after learning of an odor-reward association task and after retrieval of remote memories [99], indicating a possible role of spindles in memory consolidation of learned information; this is supported by the proof that synthetic spindle oscillations induce synaptic potentiation in neocortical pyramidal cells in vitro [100].

Synchronous replay of correlation patterns representing a previous experience has been observed, during sleep, in several brain structures. Similarly to [1], in [101] the authors study the evolution of neuron correlations in CA1 and in the posterior parietal neocortex, from a sleep phase before performance of a task, to the task and a sleep phase after the task: a greater similarity is reported between neuron correlations during sleep after the task and correlations during the task than between sleep before the task and the task, both
for within-structure pairs and for between-structure pairs, indicating that the two brain regions reactivate a representation of the same preceding behavioural experience, and that this reactivation can be synchronized in the two structures. However, the temporal order of correlation is preserved from the task to the sleep post task only for within-structure pairs; between hippocampus and parietal cortex temporal order can change, possibly reflecting alternation of states in which the hippocampus acts as either information source or receiver with respect to the cortex. In [102], reactivation during sleep of the behaviourally induced correlation structure is reported within and between motor, somatosensory, and parietal cortex, with some degree of similarity between the temporal ordering of neuron activity in the task and in the sleep post task; in [103], multi-neuron firing sequences in the visual cortex and in the hippocampus are observed to be replayed during SWS, concurrently to abrupt increases in the population activity (likely corresponding to SWRs).

Simultaneous replay can also occur between the hippocampus and some subcortical centres, probably allowing consolidation of nondeclarative, procedural memory, and of the reward-expectancy component of procedural as well as episodic memory, functions in which structures not residing in hippocampus and neocortex are strongly implicated. For instance, during sleep after a reward-searching task, the ventral striatum, a subcortical structure receiving direct inputs from the hippocampus and involved in the evaluation of the motivational value of actions, has been shown to re-activate task-related correlation patterns, mainly in temporal association with hippocampal ripples, but with a longer persistence time [104]. In agreement with this study, in [105] the authors report coordinated replay of hippocampal and striatal correlation patterns during sleep SWRs, after a place-reward association task: interestingly, they observe that replay is stronger for neuron pairs encoding information about reward and that the emotional (reward-related) information encoded in the ventral striatum is replayed shortly after the information about place encoded in the hippocampus: this finding supports the hypothesis that synchronous replay in different brain regions is important for learning and consolidating associations between salient pieces of information, and it agrees with a key principle of the systems consolidation theory, stating that the hippocampus initiates and coordinates replay in several brain areas (par. 1.4).

However, recent studies indicate that activity in most subcortical centres is suppressed during SWRs. A global and fascinating picture of the neuronal activity during sleep brain rhythms has been provided by Logothetis et al. [106]: exploiting both electrophysiological techniques and functional magnetic resonance imaging, the authors show that hippocampal SWRs immediately follow suppression of thalamic activity and are concomitant with up-states in association and primary cortical areas, while most subcortical structures are silenced. The thalamus may therefore play an important role in establishing a favorable condition for cortico-hippocampal interaction and replay of episodic memory, minimizing interference with other brain centres.

1.4 Towards a unifying theory of learning and memory

Inspired by the original work of Marr [107], who first proposed in '71 a layered model for memory storage and retrieval, a series of studies in the 80s laid the foundations of the systems consolidation theory, a theory about the process of reorganization of memory in brain-wide neuronal networks, taking place with time and experience and leading to
long-term storage and gradual acquisition of knowledge. The cornerstones of this theory are represented by the hippocampal memory index theory, formulated by Teyler and DiScenna [108] in '86, by the two-stage model for memory consolidation, proposed by Buzsáki [30] in '89, both focusing mainly on the central role of the hippocampus in memory formation and recall, and by the works of Squire and colleagues [109–112], started in the 80s and pursued throughout the 90s, which clearly articulated the so called standard consolidation theory (SCT). The SCT is a theory about the interplay between hippocampus and neocortex in memory processes. Several subsequent studies have focused on the changes that memories would undergo over time, in a dynamic process that would lead not only to consolidation but also to memory transformation [42, 113–115].

1.4.1 The two-stage model

The two-stage model [30] is a promising interpretation about the functional role the specific entorhinal-hippocampal formation may play in memory processes, in light of its topological anatomy and of the closed circuit it forms with the cortex (Fig. 1.1).

According to the two-stage model, during active behaviour, when hippocampus is entrained by theta oscillations, granule cells in the DG, reaching their highest firing rates, would induce weak and transient synaptic potentiation in the group of CA3 neurons to which they project. In this first stage, information about the behavioural experience would be therefore transmitted from sensory areas in the neocortex to a specific group of CA3 cells, where this information would be transiently encoded. Different sensory inputs, causing the activation of different granule cells, would produce the potentiation of different CA3 groups; moreover, inputs related to the most frequent and most recent experiences would determine the identity of the CA3 group whose potentiation will persist at the end of the behavioural phase, while the potentiating effects of the cues explored less or earlier would vanish. In the second stage, the potentiated CA3 neurons would initiate a reverberation of excitation in the CA3 region, exploiting the CA3 auto-associative synaptic network: this reverberation would give rise to sharp wave bursts. It is indeed an experimental evidence that the identity of neocortical inputs determines the identity of the CA3 cells that trigger sharp-wave events (initiator cells). The mechanism of reverberation in CA3 hypothesized by Buzsáki is the excitation-spread mechanism, illustrated in section 1.1.2 as a plausible explanation of reverse replay: when some (even weak) external drive perturbs the system, the most excitable CA3 cells would fire first, followed by the less excitable ones (in reverse order compared to that in which they have been activated in the exploratory phase). This external drive is hypothesized to be some subcortical input in [30]; however, more recent works reviewed in the previous paragraph identify in the cortical excitation at the onset of up-states a more plausible detonator of the excitation spread mechanism. Moreover, as pointed out in [30], after its initiation, the sharp wave itself may prolong the excitatory reverberation; the recurrent excitation will be stronger on the most excitable cells, i.e. the initiator cells, and will induce long-term potentiation (LTP) in their synapses.

In summary, while in the first stage weak potentiation of an experience dependent group of CA3 neurons would permit initial memory encoding during behaviour, in the second stage, the weakly potentiated CA3 neurons would initiate sharp waves and undergo LTP. Finally, the synaptic connections of the CA3 initiator cells with their CA1 targets would also undergo LTP. Importantly, these CA1 neurons are predicted to be the same cells that fired maximally during the exploratory phase, since the CA3 cells are the same, and information is reliably transmitted from CA3 to CA1. Indeed, it is shown in [30] that
an electrical stimulation of the perforant path (between the EC and the DG) can evoke two response cycles (each cycle consisting in the flow from the EC to the DG, and through CA3, CA1 and the subiculum, back to the EC, see Fig. 1.1) with the same or closely overlapping spatial distributions: this means that, despite the large divergence of synaptic projections in the CA3 region, information is not mixed, but is reliably conveyed to the successive stages of the hippocampal circuitry until the cells of origin in the entorhinal cortex are re-activated at the beginning of the second cycle. The mechanism hypothesized in the two-stage model for the reactivation of the same cells in CA3 and CA1 agrees with the observation of reverse replay in both regions; forward replay can be produced, for example, if the excitation of the CA3 group in the first behavioural stage is already strong enough to induce LTP; in this way, when the most excitable cell is activated at the arrival of an external input, the CA3 potentiated synapses will be re-activated in the forward direction.

1.4.2 The hippocampal memory index theory

The reliability of information transmission from the EC to the hippocampus and back to the EC, highlighted in the two-stage model, is a fundamental property of this neuronal circuit, necessary to support the indexing function of the hippocampus, hypothesized by Teyler and DiScenna [108]. According to the hippocampal memory index theory, the potentiated group in CA3, determined by the pattern of neocortical activity evoked during an experience, would serve as an index to the same pattern of neocortical activity, in the sense that if an input activates the index at a later time (e.g. in cued recall or replay), the index will re-activate in turn those unique neocortical areas [116]. In other words, from an anatomical point of view, the central tenet of indexing theory is that an episode generates an index in the hippocampus which encodes locations in cortical space [117].

The indexing property would allow the hippocampus to accomplish some important functions, namely pattern completion, pattern separation and binding.

Pattern completion is illustrated schematically in Fig. 1.5, top: when a subset of cues representing a previous experience is received by the neocortex, only a few neocortical neurons representing that experience will be activated; however they can trigger the activation of the entire index in the hippocampus, thanks to a reverberation of the activity, originally called ‘collateral effect’ by Marr [107], through the previously potentiated synapses within the CA3 group; the index will activate in turn (retrieve) the entire neocortical pattern representing the original experience.

Pattern separation is shown in Fig. 1.5, bottom left: the hippocampus forms sparse, non-overlapping representations (indices) for similar, overlapped neocortical patterns, allowing retrieval of a large number of similar memories with minimal interference (a more detailed discussion on the importance of sparseness in the hippocampus and on some related theoretical studies will be given in the next section).

Different indices should be maximally activated by distinct input patterns for pattern separation to work optimally: the diffuse random connectivity observed between EC and neocortical areas is compatible with this constraint. The theory predicts a tradeoff between pattern separation and pattern completion as a function of the overlap in the neocortical input patterns: for high levels of overlap, pattern completion dominates over pattern separation [118]. Indeed, when input similarity is extremely marked, the CA3 representations may lose their orthogonality [55]. However, for moderate levels of overlap in the neocortical patterns, distinct hippocampal indices enable reactivation of each
neocortical pattern separately, which would be impossible in a one-level system, with information uniquely stored in the neocortex (Fig. 1.5, bottom right). Moreover, in [119] the authors observe that the transition from pattern completion to pattern separation, or in other words the remapping in the CA3 network, is smoother and accompanied by hysteresis when the input is varied slowly and progressively between two extremes than when it changes abruptly from one extreme to the other.

The random connectivity between EC and neocortical areas also suggests that the hippocampus works as a simple binding device of widespread neocortical information, instead of forming more complex relational encodings. Both subcortical inputs (e.g. from the amygdala, the thalamus, the medial septum) and neocortical inputs (from association cortices) converge into the hippocampus, suggesting that the hippocampal index may be an important mechanism to rapidly bind multi-modal information into a unified representation, which does not contain that information but can retrieve it.
1.4.3 Insights on memory storage and recall: the importance of sparseness in CA3 and DG

Since D. Marr’s work [107], sparseness has been highlighted as a fundamental property of the neuronal representations formed in the hippocampus, particularly in CA3. As seen in the previous section, sparseness is thought to be important for the indexing function to be carried out successfully, in particular it seems necessary to decorrelate overlapping cortical representations and to keep distinct during retrieval memories which are similar (pattern separation). In other words, the fraction of neurons in CA3 activated by each memory should be small and neuronal representations corresponding to different memories should be orthogonal to minimize interference among the widest possible set of memories. This section is dedicated to both experimental and theoretical studies which strongly support and articulate this idea.

As observed in connection with remapping, ensemble representations of distinct but similar environments are indeed orthogonal in CA3 [55] and they remain independent when the animal is exposed to a large number of environments (e.g. eleven) [121]: the majority of neurons active in one environment are silent in the other environments, and the few neurons which are active in more environments change completely their place fields, suggesting that the CA3 network is indeed capable of storing, in distinct memories, many variations of an experience.

The relation between sparseness of neural coding and storage capacity has been studied for the most analytically on recurrent artificial neural networks, models in which a set of patterns (memories) are stored in recurrent Hebbian-type connections between the units and specific dynamics rules, dependent on the values of those connections, make the network evolve to the stored pattern that is more correlated with the input cue (memory recall). The first and most popular of such models of autoassociative memory is the Hopfield model [5], in which the network units are binary variables \( \{\sigma_i\} \) and the Hebb rule is implemented by equaling the connection \( J_{ij} \) between each pair of units \( i \) and \( j \) to the average value of \( \sigma_i\sigma_j \) over the patterns to be stored. Within its capacity limit, the model can work as an autoassociative memory if a dynamics is imposed on the network such that each unit \( i \) becomes active (\( \sigma_i = 1 \)) if \( \sum_j J_{ij}\sigma_j > \theta_i \) (\( \theta_i \) activation threshold for unit \( i \)), and becomes silent (\( \sigma_i = -1 \)) in the opposite case. Subsequent studies have shown that the capacity of similar autoassociative networks (i.e. the maximum number \( p \) of retrievable patterns per synapse) increases with the sparseness of those patterns, represented by parameter \( a = \langle \eta \rangle^2 / \langle \eta^2 \rangle \), where \( \langle \cdot \rangle \) denotes the average over the distribution of the neuron activities in the stored patterns \( \{\eta_i^\mu\} \) (for 0,1 binary units, \( a \) reduces to \( \langle \eta \rangle \)). One of the first derivations of the storage capacity is found in [122] for an Hopfield-type model, with both positive and negative Hebbian synapses and binary neurons; a generalization of this formula for more realistic graded response (threshold-linear) neurons and different types of connectivity, has been proposed by Treves and Rolls [123] and it shows that \( p \) is approximately proportional to \( (a \log(1/a))^{-1} \): the smaller \( a \) (i.e. the sparser the coding), the greater the number of retrievable patterns, due to the decrease of interference among them. Golomb, Rubin, and Sompolinsky [124] have studied the dependence of the capacity on the sparseness in a different autoassociative network, a modified version of the Willshaw model. In the Willshaw model memories are stored using a very simplified version of the Hebb’s rule, which assigns to a synapse \( J_{ij} \) value 1 if neurons \( i \) and \( j \) are both active in at least one pattern, value 0 otherwise. In [124] a uniform negative term is added to all synapses to represent global inhibition. The attractive feature of this model is that it not
only has sparsely coded memories, that is memories with low global activity, but it also shows partially ordered low temperature phases (highly correlated with the memories) with low local activity, that is low firing rates of the active neurons, a property observed experimentally in cortical patterns during short-term memory task. For this kind of model, however, the storage capacity is smaller than that derived in [123]. Indeed, in [125], it is shown that in models with subtractive inhibition (like that of [124]), the steady states of the dynamics (stored memories) can be retrieved only keeping the excitatory recurrent connections low, thus reducing the network capacity. The authors prove that this conflict is potentially solved with a faster (not subtractive but divisive) inhibition, which is more realistic for recurrent neuronal networks.

There is increasing evidence that the sparse, orthogonal property of the representations formed in CA3 derives from sparseness of the neuronal activity in the DG. Indeed, most granule cells (\(\sim 95\%\)) in the DG are silent during exploration of any environment, though (differently from what happens in CA3) representations are not orthogonal, that is the active sets observed in different environments explored at close times are highly correlated. It has been proposed in [126] that most granule cells become non functional after some time from their generation; therefore the active set at any given time would correspond to the most recently generated (most excitable) cells. A similar but not equivalent hypothesis [127] is that the large, non functional set would be produced by LTD of a large portion of synapses between granule cells in the DG and grid cells in the EC: most of these synapses are indeed too weak to convey to the granule cells an input strong enough to make them fire, thereby undergoing LTD. This effect would be particularly strong since activity in the EC is almost perpetual. The few functional granule cells would be those with strongest afferent synapses from the EC and higher excitability. The sparseness of the activity in the DG, together with the particular strength of the mossy fiber synapses connecting the DG with CA3, have been hypothesized in [128] to be fundamental for effective storage of new information in CA3; in particular, it would allow storage of new information in a pattern of activity as independent as possible from any patterns previously stored in the CA3 recurrent network. As suggested in [128], this would also justify the presence of the mossy fibers as an additional afferent system to CA3 compared to the perforant path, which conveys information to CA3 directly from the EC and would have a complementary function, more related to memory retrieval.

The particular connectivity between DG and CA3 also seems to enable encoding of sequential events: it is argued in [129] that the delay with which information on the \(n\) element of a sequence to be learned is transmitted from DG to CA3 and back to DG is the same as the delay with which the \(n+1\) element of the sequence reaches DG from cortex (this consideration comes from the idea that the EC activity is temporally compartmentalized into information packets, items of the sequence, corresponding to the gamma cycles within each theta cycle). This temporal coincidence would induce LTP in the feedback synapses from the CA3 cells representing the \(n\) item to the DG cells representing the \(n+1\) item. This kind of heteroassociative mechanism suggests a suggestive interpretation of phase precession as a cued recall of a memory sequence [129], in which the cue is the current animal position and subsequent positions are recalled, or predicted, on the basis of the memorized trajectory; update of the cue at each new theta cycle would produce progressive anticipation of the firing phases of the place cells.

In conclusion, sparseness, which begins to be considered a general property of the activity in the entire medial temporal lobe [130], together with the specific types of connectivity within CA3 and between CA3, DG and EC, strongly supports the hippocampal memory
1.4. TOWARDS A UNIFYING THEORY OF LEARNING AND MEMORY

1.4.4 The standard consolidation theory

The hippocampal indexing property is at the basis of the standard consolidation theory (SCT), which extends the two-stage model, elucidating the role of the neocortex in memory consolidation and re-organization. The SCT proposes that the hippocampus is important for the initial storage of declarative memory, that is episodic memory about events and semantic memory about abstract concepts or generic facts. In the early stages of memorization, the hippocampus would rapidly provide a conjunctive representation of distributed sites in the neocortex that altogether represent a memory: whenever the index is activated, e.g. during rest or sleep, in conjunction with sharp waves, or when (even partial) cues of the same experience are encountered, the neocortical array of activity patterns representing that experience would be reinstated and the memory retrieved. On each reinstatement, cortico-cortical synapses would be gradually strengthened, and after several reinstatements a long-term memory of the experience would be formed in the cortex, becoming independent from the hippocampus. Moreover, non-declarative, procedural memory (necessary to perform automatic actions without the need for conscious control) would be acquired independently from the hippocampus.

The temporary role of the hippocampus in declarative memory is supported by several experimental observations. In particular, medical studies shows that patients with hippocampal lesions are subject to severe anterograde amnesia (the inability to create new memories after the event that caused the amnesia) and a temporally graded retrograde amnesia, in which recently acquired memories are forgotten while the older ones are more resistant to disruption; in these patients technical skills and general intelligence also seem to be spared [9], confirming that the hippocampus would not be fundamentally involved in non-declarative memory or other cognitive functions. Moreover, some functional magnetic imaging studies show that hippocampal activity decreases over time during consolidation, whereas activity in neocortical areas like PFC increases; these changes are observable over a few months [131].

The SCT requires two different forms of LTP in the hippocampus and in the neocortex: synaptic plasticity in the hippocampus should support rapid and reversible strengthening of connections to form transient memory traces (the indices); in the neocortex, synapses should require more input to be strengthened, but their potentiation should be more stable to permanently store remote memories. Indeed, two different forms of LTP are observed in experiments: a rapidly induced and easily reversible form of LTP, in which a moderate input makes calcium (the key element of synaptic plasticity) enter the post-synaptic cell through the NMDA receptor (NMDA-dependent LTP); and a slowly induced, long-lasting form of LTP in which a strong input allows calcium to enter the post-synaptic cell via voltage dependent calcium channels (VDCC-dependent LTP). Both NMDA-dependent and VDCC-dependent LTP can be found in both hippocampus and neocortex [132]. However, in vivo studies confirm the SCT prediction that inducing LTP is in general more difficult in the cortex than in the hippocampus [133, 134]. Though the picture is still incomplete, it seems plausible from plasticity studies that the hippocampus can generate both transient and long-lasting representations, while cortical synapses, requiring stronger afferent inputs to be changed, can form stable memories more slowly [117].
1.4.5 The transformation hypothesis

The implementation of memory by means of two complementary, interacting systems, a fast-learning module represented by the hippocampus and a slower one represented by the neocortex, can allow the brain to avoid “catastrophic interference”. Catastrophic interference means the retroactive interference of a set of newly learned associations (e.g. AB) on recall of a set of associations previously acquired (e.g. AC), or in other words the impossibility to learn new information without forgetting the specifics of information learned in a previous experience. This problem affects all standard artificial neural networks or connectionist models [135], i.e. networks that can learn to store knowledge in their connections, but do not have a parallel system for the rapid acquisition and storage of specific pieces of information. These networks are probably good models of the kind of processing carried out by the neocortex, in that they are able to discover generalities, or the shared structure of multiple experiences, by interleaved learning, in which a new piece of information about a class of objects is slowly acquired through presentations interleaved with exposure to other samples of the same class (a paradigm which seems to represent well the way we accumulate experience throughout life) [136]. These models, however, cannot retain the specifics of single experiences, which causes catastrophic interference.

The bipartite hippocampal-cortical system has the potential to solve this problem [113, 118]: the hippocampus can rapidly learn the specifics of new experiences and retain them at least temporarily, while the neocortex can gradually incorporate the items that are more frequently encountered (and more relevant) into a structure that will contain the salient aspects of multiple experiences. It is argued that during this process general features of experiences would be gradually extracted from several related episodes to form a knowledge of the world. Therefore, during consolidation, it is likely that memories are not just stored in their original form, but they are transformed and reorganized into more semanticized, decontextualized representations. Moreover, the transformed memory, residing in the neocortex, does not necessarily replace the initial, more detailed memory, residing in the hippocampus, but can coexist and interact with it until the specific memory of the episode is reinstated. This transformation hypothesis [114, 115] extends the SCT. In particular, it clearly distinguishes the mechanisms underlying the two types of declarative memory, episodic and semantic, and the fate of the two (whereas this distinction remains implicit in the original works of the SCT): until episodic memories are retained, they would continue to depend on the hippocampus, while semantic memories, once formed, would be stored in the cortex and would not require the hippocampus to be retrieved. This hypothesis is in agreement with several reports on amnesic patients, with known or presumed damage to the hippocampus, showing severe and temporally ungraded retrograde amnesia for episodic memories, but spared semantic memories.

This idea of a transformation of memory during consolidation is also the focus of a recent theory [42], called “information overlap to abstract” (iOtA), illustrated in Fig. 1.6, which identifies in the replay during sleep the key mechanism enabling abstraction and formation of cognitive schemata. The authors propose that when two related memories (represented by two overlapping groups of neurons in the neocortex) are replayed during SWS, the overlapping part (representing the common aspects of the two memories), which is replayed each time the first or the second memory is evoked, is re-activated more than the neurons unique to each memory. Therefore synapses within the shared group would be strengthened more than those not involved in the overlap. As a result of synaptic downscaling (which takes place throughout SWS to avoid synaptic saturation) only these strongly potentiated synapses between the shared neurons will survive, representing the
Figure 1.6: The iOtA mechanism (taken from Lewis and Durrant [42]). A larger neuron size indicates greater neural activation and a thicker line indicates a stronger synaptic connection. During sleep, the hippocampus triggers replay of two overlapping memories in the neocortex; the common part of the two memories is re-activated more (a); as a result, connections between the shared neurons are strengthened more than the other connections (b); following synaptic downscaling, only these strongly potentiated synapses are preserved, and a scheme of the common aspects extracted from the two memories is formed in the neocortex (c).

generalities extracted from the two memories. Through the replay of new overlapping memories, new related information would be progressively integrated into this initial scheme.
Chapter 2

Quantitative methods for the study of networks of interacting neurons

A central objective of neuroscience is to understand how neurons or different areas of the brain interact, what are the specific functions of these interactions [137], and how they ultimately give rise to behaviour. This is a challenging task because anatomical synapses are difficult to determine and simple correlations in the activity of different neurons are not always indicative of direct interactions between them: indeed, neurons can be correlated because they are synaptically connected, or because they receive inputs from a common pool of other neurons, or because they receive the same external (sensory) inputs. In this chapter, I will review some traditional and more recent methods for the study of neuron correlations, of how disentangling their different sources, and more generally of how extracting non-trivial interaction structures from spike recordings. The last part of the chapter will be focused on state-of-the-art methods to detect and characterize cell assemblies in neuronal data and to quantify the phenomenon of replay.

2.1 Descriptive statistics of correlations

For many decades, neuroscientists have used simple descriptive statistics to study populations of interacting neurons. Traditional methods for the study of correlations are cross-correlation histograms and joint peri-stimulus time histograms.

The cross-correlation histogram between neurons $i$ and $j$ is defined as:

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

(2.1)

where $N_i$ and $N_j$ are the numbers of spikes of neurons $i$ and $j$ during the recording time $T$, $\Delta t$ is the bin-width of the histogram, and $\mathcal{I}_{\tau, \Delta t}(t_{i,a}, t_{j,b}) = 1$ if $|\tau - (t_{j,b} - t_{i,a})| < \Delta t/2$, and $\theta_{\tau, \Delta t}(t_{i,a}, t_{j,b}) = 0$ otherwise. Eq. 2.1 can also be written as:

$$H_{ij} (\tau, \Delta t) = \frac{p_{ij}(\tau, \Delta t)}{p_i(\Delta t)p_j(\Delta t)}$$

(2.2)

where $p_i(\Delta t)$ is the probability that neuron $i$ spikes in a time-bin of width $\Delta t$ and $p_{ij}(\tau, \Delta t)$ is the probability that the delay between a spike of neuron $i$ and a spike of neuron $j$ is in the range $[\tau - \Delta t/2, \tau + \Delta t/2]$.  

The value the cross-correlation histogram takes in $\tau = 0$ is called correlation index and is a measure of synchrony between two neurons:

$$CI_{ij}(\Delta t) = \frac{p_{ij}(\Delta t)}{p_i(\Delta t)p_j(\Delta t)}$$  \hspace{1cm} (2.3)

The cross-correlation histogram is flat and (within the sampling errors) equal to 1 when neurons are independent (for instance, it always approaches 1 when $\tau$ is large enough). Cross-correlation histograms do not allow us to distinguish correlations due to a direct synaptic connection between two neurons from correlations mediated by a third neuron, or induced by a common input. All these types of correlations produce a peak in the cross-correlation histogram and it is difficult (or in most cases impossible) to judge which kind of correlation is at the origin of the peak by the value of the delay $\tau$ corresponding to the peak. Moreover, cross-correlation histograms may show spurious correlations when the spikes of a neuron in different time-bins are not independent [138]: for instance, refractoriness reduces the probability that a neuron will spike soon after a previous spike of the same neuron, and this introduces deviations from the flatness of the cross-correlation of the neuron with another independent neuron. However, the simple analysis of cross-correlation histograms and correlation indices is in some cases sufficient to show interesting properties, like the change of correlations in response to particular stimulations: an example will be illustrated in chapter 6 of this thesis.

Differently from cross-correlations, joint peri-stimulus time histograms (JPSTH) display the dynamics of correlations (not simply their time average) and they also allow one to distinguish the case of correlations induced by a stimulus (provided the stimulus is known and repeated several times) from correlations of different origin. The JPSTH [139, 140] is the two-dimensional extension of the single neuron peri-stimulus time histogram (PSTH), which is the spike count per unit time of a single neuron at each time $t$. The JPSTH is a matrix of square bins of area $\Delta t^2$: for each stimulus repetition, if a neuron emits a spike in time-bin $x$ after stimulus onset and the other neuron emits a spike in time-bin $y$ after stimulus onset, one count is added to the matrix element $(x,y)$. The most significant parts of the JPSTH are its main diagonal, which displays the time course after stimulus onset of the number of spike time coincidences of the two neurons, and the other parallel lines, which represent the time course of the number of spike events producing time-lag correlations between the two neurons (where the time-lag increases as the line is more and more distant from the main diagonal). In order to remove from this diagram correlations due to a common stimulus, the ‘shift predictor’ is computed. This is a matrix built like the JPSTH, but with a shift of one or more stimulus trials in the spike trains of a neuron with respect to the other neuron: the matrix thus compares the response of one neuron in a trial with the response of the other neuron in another trial. The shift predictor retains correlations induced by the stimulus (since it still displays responses aligned with stimulus onset), but destroys correlations due to a synaptic connection between the two neurons or mediated by other neurons, since these correlations act on relatively small time scales and do not last for different stimulus trials. Subtracting bin by bin the shift predictor from the JPSTH, one obtains a corrected JPSTH, in which the stimulus-locked covariation is eliminated (a less noisy shift predictor is derived as the average over all possible shifts). The normalized JPSTH can be finally obtained dividing the corrected JPSTH, again bin by bin, by the product of the individual neurons’ PSTH standard deviations (computed for the set of bin counts across stimulus trials): the element $(x,y)$ of the normalized JPSTH is the Pearson correlation (computed across trials) of the activity of the two neurons at
times x and y, respectively, after stimulus onset. The normalized JPSTH can be used to compare correlations of different data sets.

The significance of the results obtained from a JPSTH are usually assessed assuming that the bin counts have Poisson distributions with means given by the shift predictor matrix elements (and standard deviations by their squared root), and evaluating the distance of each matrix element from its mean. A drawback of this test is that it assumes that the first and second order statistics of the bin counts are constant across trials (this is usually a good approximation but it may not be always true). Moreover, information obtained from a JPSTH may require large data sets with many trial repetitions to be evaluated as statistically reliable (a problem that is less onerous for the cross-correlation histograms, where correlations are averaged throughout the stimulus duration).

The JPSTH technique has been used successfully in several works: an example is illustrated in [141], where it is applied to the study of correlations between tactile interneurons in the crayfish. The corrected JPSTH shows that correlations between these interneurons in the later part of the response to a tactile stimulus (after \( \sim 40 \text{ ms} \)) are due to the electrical synapses between them; these synapses increase synchronization between the spikes of these cells, which is necessary to elicit a response in the afferent lateral giant cell, a decision fibre responsible for escape behaviours with a very large discharge threshold. However, this is a case in which it is known that excitatory inputs to tactile interneurons can only come from tactile afferents, or other tactile interneurons. In more complex neuronal networks, where neurons may also share inputs from different sources, independent from the stimulus, the JPSTH technique is not able to tell apart correlations due to a direct synaptic connection from correlations mediated by another source unrelated to the stimulus.

A more accurate and complete picture of the interaction network of the recorded neurons can be obtained from the inference of functional network models, which is the topic of next paragraph.

### 2.2 Model-based methods

Model-based methods consist in mapping the observed spiking data onto abstract network models, representing specific structural and functional properties of the real system: in a very general case, the recorded neuronal system is represented as a network or graph of nodes (neurons) and links (neuron interactions) with external inputs entering into each node (representing the neuron tendencies to fire, independently from the network interactions). Other features can be added to this representation in more complex network models. A network supports a set of functional states, representing configurations of neuronal activity, and each state is generated by the network with a certain probability. The specific structure of the network, set by the model parameters, defines a probability distribution over all these functional states. Parameter values are inferred from the data in such a way the probability distribution of the activity configurations (states) generated by the network model is as close as possible to the probability distribution of the activity configurations experimentally observed. Inference is therefore an inverse problem, because it consists in reconstructing some microscopic properties of the system (represented by the model parameters) from macroscopic measurable quantities (functions of the observed neuronal activity), as it will become clearer later in this chapter. Different models define different probability distributions over the states, characterizing different properties of the neuronal activity, and the choice of the model to be used should be guided by the specific
CHAPTER 2. METHODS FOR THE STUDY OF NEURONAL NETWORKS

In general, network models can disentangle direct and indirect correlations (among the recorded neurons), by explicitly including parameters which represent direct effective interactions and enabling to fit these - and the other - parameters all at once [142]. The inferred interactions (or couplings) are effective or functional, in the sense that if two neurons have a strong coupling, either they are synaptically connected or they are functionally correlated through inputs that are not included in the model (e.g. common inputs from unrecorded neurons or external, not modeled, inputs). The difference with respect to correlations is that if two neurons A and B are correlated through a third neuron C which is also recorded and included in the model, pairs CA and CB will have a positive coupling, but pair AB will have a zero coupling (despite a positive correlation is measured between them). In [143], for instance, the authors show that inference of an Ising model (see 2.2.1 and 2.2.2) from neuronal activity in the salamander retina may uncover negative couplings between cells with small but positive correlation indices.

Model-based methods have several advantages compared to descriptive statistics of correlations: not only they improve estimation of connectivity, but they can also be used to assess the effects and relative importance of different covariates on a neuron activity (like the neuron past spiking history, the activity of the other neurons and external variables, as will be illustrated in some detail in section 2.2.3); this could not be achieved by studying the single covariates separately. Moreover, inferred models can be used as generative models of the neuronal activity, that is they can be simulated to make predictions about hidden functional properties of the real system, which would be impossible to uncover from the raw recording data (this topic will be the core of the central chapters 3, 4 and 5 of this thesis).

In the following part of this paragraph I will present different classes of functional network models, the methods developed by the community to infer their parameters from spike recordings, and some examples of techniques used to assess a model goodness-of-fit, that is its ability to represent the data or to predict certain properties of them. I will first present the class of maximum-entropy models, which are in general stationary, that is the probability distribution of the activity configurations generated by these models is, in most cases, constant over time. I will present some of their recent applications to neuronal data, focusing in particular on the Ising model, and I will discuss in some detail the inverse Ising problem. Then I will illustrate several classes of non-stationary models (defining time-varying probability distributions), like generalized-linear models, linear-nonlinear-Poisson models and state-space models. These classes are not disjoint: a schematic representation of these model categories and of their overlaps is given in Fig. 2.1.

2.2.1 Maximum-entropy models

Maximum-entropy (ME) models are the least-constrained models reproducing some set of average properties of the system. The entropy of a distribution is indeed a measure of its ‘uncertainty’ or ‘uniformity’ [144], and ME models represent the most general (least structured) models consistent with a set of experimental observations. The potential of these models to reproduce known and relevant properties of a system, without making any implicit assumption about the system beyond those properties, and the relative simplicity with which their parameters can be inferred from experimental data, have made them popular in many different fields of research, including neuroscience.

In the context of neuronal modeling, a spike train is usually described as a set of binary
variables \{\sigma_{it}\}, which represent the state of each neuron \(i\) in each time-window \(t\) of the activity: if the duration \(dt\) of these time-bins is chosen to be small enough, a neuron either does not spike \((\sigma_{it} = 0)\) or it spikes just once \((\sigma_{it} = 1)\) in a time-bin. ME models give a time-independent probability distribution for the configurations \{\sigma_i\}, \(i = 1, \ldots, N\), of the \(N\) recorded neurons in a time-bin.

The entropy of a distribution \(p(\{\sigma_i\})\) is defined as:

\[
S[p(\{\sigma_i\})] = - \sum_{\{\sigma_i\}} p(\{\sigma_i\}) \log p(\{\sigma_i\})
\]  

(2.4)

The ME distribution satisfying no constraint at all is the uniform distribution over all configurations \{\sigma_i\}. In this model, each neuron has equal probability of being active or silent in a time-bin, which is clearly very far from reality. Constraints are added to this maximally unstructured model to reproduce significative properties of the real system. The ME distribution which is constrained to reproduce the experimental average values of some functions of the state of the system \(f^\mu(\{\sigma_i\}), \mu = 1, \ldots, L\) and which satisfies the normalization constraint is the function \(p(\{\sigma_i\})\) which maximizes the Lagrangian:

\[
\Lambda(\{\sigma_i\}, \{\lambda^\mu\}) = S[p(\{\sigma_i\})] + \sum_{\mu=1}^{L} \lambda^\mu \left( \langle f^\mu(\{\sigma_i\}) \rangle_p - \langle f^\mu(\{\sigma_i\}) \rangle_{obs} \right) + \lambda^{L+1} \left( \sum_{\{\sigma_i\}} p(\{\sigma_i\}) - 1 \right)
\]  

(2.5)

with \(\langle f^\mu(\{\sigma_i\}) \rangle_p\) and \(\langle f^\mu(\{\sigma_i\}) \rangle_{obs}\) average values of \(f^\mu(\{\sigma_i\})\) computed with respect to the model distribution \(p(\{\sigma_i\})\) and over the observed activity configurations, respectively.

The solution of this maximization problem is the Boltzmann distribution:

\[
p(\{\sigma_i\}) = \frac{e^{-\mathcal{H}}}{Z}
\]  

(2.6)
with Hamiltonian
\[ H = - \sum_{\mu=1}^{L} \lambda^\mu f^\mu(\{\sigma_i\}) \] (2.7)
\{\lambda^\mu\} are the model parameters and \(Z\) is the normalization constant: 
\[ Z = e^{1-\lambda^{L+1}} = \sum_{\{\sigma_i\}} e^{-H}. \]

When modeling a neuronal system, typical properties that can be easily computed from the data and used to build a ME model of the activity are the mean spike probabilities \(\{\langle \sigma_i \rangle_{obs} = p_i\}\) for every neuron in the time-bin, the mean pairwise correlations between neurons \(\{\langle \sigma_i \sigma_j \rangle_{obs} = p_{ij}\}\), or higher order correlations (e.g. among triplets of neurons \(\{\langle \sigma_i \sigma_j \sigma_k \rangle_{obs} = p_{ijk}\}\)), and the distribution of synchrony \(P_N(k) = \langle \delta(\sum_i \sigma_i, k) \rangle_{obs}\) (that is the probability that \(k\) out of \(N\) neurons spike in the same time-bin). Different ME models (with more and more levels of structure) are obtained by adding these constraints one at a time [145], as illustrated below.

**Independent and Ising models**

Taking the single-neuron spike probabilities as constraints one obtains the independent model:
\[ p_{\text{ind}}(\{\sigma_i\}) = \frac{e^{\sum_i h_i \sigma_i}}{Z} \] (2.8)
with parameters \(\{h_i\}\).

The pairwise or Ising model is obtained constraining both single-neuron probabilities and pairwise correlations:
\[ p_{\text{Ising}}(\{\sigma_i\}) = \frac{e^{\sum_i h_i \sigma_i + \sum_{i<j} J_{ij} \sigma_i \sigma_j}}{Z} \] (2.9)
Parameters are \(\{h_i\}\) (called ‘fields’ in statistical physics), representing local inputs to each neuron, or their average tendency to fire (independently from the other neurons), and \(\{J_{ij}\}\), representing couplings or effective interactions between neurons.

The Ising model has been largely applied to investigate the nature of correlations, especially in the retina.

Schneidman et al. [146] have proved that, compared to the independent model, the pairwise model is much more efficient in predicting the probability of occurrence of multi-neuron firing patterns in the vertebrate retina: the predicted and observed rates of commonly occurring 10-neuron patterns are indeed very close under the pairwise model, and larger deviations are found only for rare patterns, whose empirical probabilities are also uncertain due to sampling errors. The authors also argue that pairwise interactions are dominant compared to higher order interactions. This consideration is based on the computation of the contribution of different orders of interactions to the multi-information \(I_N = S_1 - S_N\), where \(S_1\) is the entropy of the independent model and \(S_N\) the true entropy of the system (potentially including all orders of interactions up to the number of recorded neurons \(N\)). The contribution of interactions of order \(k\) to the multi-information is given by \(I_k = S_{k-1} - S_k\), which represents the reduction in entropy produced by interactions of order \(k\) (or, in other words, the amount of structure added by those interactions to the distribution of activity configurations). \(I_N\) results from the sum of all contributions \(I_k\), with \(k = 2, \ldots, N\). In the retina, under many different conditions, \(I_2 \sim 90\% I_N\) [146], meaning that pairwise interactions are much more present in the data than higher order interactions.
In [147] this consideration is further developed: the authors show that in parasol retinal ganglion cells of primate retina, not only pairwise interactions are dominant over higher order interactions, but pairwise interactions restricted to adjacent cells can explain accurately multi-neuron firing patterns within a time-bin; since these synchronous, multi-neuron patterns extend to non-adjacent cells, it is likely that they are produced by the propagation of signals through chains of neighboring cells. In [147], the accuracy of different models in describing the correlation content of the data is quantified in a different way, exploiting the concept of cross-entropy between a model and the observed distribution of neuronal configurations:

\[ S^*_{\text{model}} = - \sum_{\{\sigma_i\}} p_{\text{obs}}(\{\sigma_i\}) \log p_{\text{model}}(\{\sigma_i\}) = -\langle \log p_{\text{model}} \rangle_{\text{obs}} \] (2.10)

with \[ p_{\text{obs}}(\sigma) = \frac{1}{B} \sum_{t=1}^{B} \delta_{\sigma, \sigma_t} \] (B number of time-bins in the recordings). The greater the cross-entropy, the greater the divergence between the model and the empirical distribution. In particular, the cross-entropy between the data and the independent model \( S^*_{\text{ind}} \) represents the departure of the data from statistical independence, and gives another possible estimate of the overall interaction content of the observed neuronal configurations. Similarly, \( S^*_{\text{Ising}} \) can be interpreted as a measure of the amount of interactions of order higher than 2 present in the data. Therefore, \( (S^*_{\text{ind}} - S^*_{\text{Ising}})/S^*_{\text{ind}} \) is an index of the amount of pairwise interactions relative to the overall interaction content, and \( (S^*_{\text{ind}} - S^*_{\text{adj-Ising}})/S^*_{\text{ind}} \) of the relative amount of adjacent pairwise interactions. This index computed in [147] for the Ising model is \( \sim 99\% \) (approximately equal to the benchmark value \( (S^*_{\text{ind}} - S^*_{\text{obs}})/S^*_{\text{ind}} \)) and for the adjacent-pairwise model it is \( \sim 98\% \).

However, the contributions of different levels of interactions to bridge the cross-entropy between the data and the independent model, or to account for the multi-information \( I_N \), do not tell anything about the importance of those interactions for the neural code, and they should only be interpreted as estimates of the extent to which these interactions are present in the data. Rarity of a neuronal pattern is not a sign of its lack of significance: rare patterns may indeed be crucial in the neural code, as the work presented in chapters 3, 4 of this thesis suggests.

A remarkable result of the works reviewed above is that the Ising model is indeed a generative model of the neuronal activity (at least for the retina): it not only reproduces neuron firing rates and pairwise correlations, but it also reconstructs higher order correlations in the population activity, and can in principle be used to make predictions about multi-neuron firing patterns (this idea will be developed in chapters 3 and 4). A discussion on the applicability of the Ising model to the study of neuronal interactions is found in [143], where the authors address the important issue of the reliability of couplings inferred from recordings of the retina, considering not only the temporal sampling problem (limited durations of the recordings), but also the more complex problem of spatial sampling: failure in recording some cells in the retina may in principle result in inferring couplings which do not represent the true neuron connections, but indirect correlations, mediated by non-recorded cells. The authors show that removal of a cell only affects local couplings (within a radius of 600 µm) and therefore couplings can be considered reliable as long as the electrode arrays are dense enough to record most cells in a small region. This issue, inherent to all inverse problems, will be further discussed in the next paragraph. Reliability of the couplings inferred in [143] allows the authors to prove that neuron interactions in the retina change according to light conditions: in particular positive long-range couplings (between neurons distant more than 500 µm) are present in flicker conditions but not in
the dark.

Some other examples of applications of the Ising model to the study of neuronal recordings will be mentioned later, in relation to the work presented in the following chapters.

### K-pairwise and higher order ME models

Another example of a ME model is the \( k \)-pairwise model, which reproduces single-neuron probabilities, pairwise correlations and the distribution of synchrony:

\[
p_{\text{pair}}(\{\sigma\}) = e^{\sum_i b_i \sigma_i + \sum_{i<j} J_{ij} \sigma_i \sigma_j + \sum_{k=1}^N \lambda_k \left( \sum_{i=1}^N \sigma_i \right)^k} / Z
\] (2.11)

Despite this model is not fitted to reproduce correlations of specific triplets of neurons, it is able, in some cases, to predict triplet correlations better than the pairwise model, as shown in [145] for populations of neurons in the salamander retina.

A third order ME model reproducing first, second and third-order moments can be easily obtained adding the terms \( \sum_{i<j<k} J_{ijk} \sigma_i \sigma_j \sigma_k \) to the Ising Hamiltonian. Third order and higher order ME models are also used to evaluate the effects of different orders of interaction: for example in [148], the authors show that high order interactions (at least of order 3) are necessary to explain the empirical probability distribution of synchronous discharge \( p(k) \) (with \( k \), number of neurons active in the same time-bin, up to \( \sim 20 \)) in rat somatosensory cortex in response to a stimulus. Exploiting techniques of information theory, based on the concept of entropy of a probability distribution, it is also possible to quantify the impact of interactions of different orders in the transmission of information about a stimulus. The fundamental quantity which estimates the amount of information transmitted about a certain stimulus \( s \) by a system with a given distribution \( p(\{\sigma\}) \) of activity configurations is called mutual information [149] and is defined by the following equation:

\[
I(s, \{\sigma\}) = S[p(\{\sigma\})] - \langle S[p(\{\sigma\} | s)] \rangle_{p(s)}
\] (2.12)

where \( S[p(\{\sigma\})] \) (called response entropy) and \( \langle S[p(\{\sigma\} | s)] \rangle_{p(s)} \) (called noise entropy) are defined as:

\[
S[p(\{\sigma\})] = - \sum_{\{\sigma\}} p(\{\sigma\}) \log_2 p(\{\sigma\})
\]

\[
\langle S[p(\{\sigma\} | s)] \rangle_{p(s)} = - \sum_{\{\sigma\}, s} p(s)p(\{\sigma\} | s) \log_2 p(\{\sigma\} | s)
\]

\( I(s, \{\sigma\}) \) is a measure of the average reduction in entropy produced by knowledge of the stimulus on the distribution of the neuronal responses. If neurons respond deterministically to the stimulus (i.e. \( p(\{\sigma\} | s) \) is either 0 or 1, meaning that there is no loss of information during neuronal processing), the noise entropy is zero, and the mutual information is maximized, as expected. An estimate of the effects of different orders of interaction on information transmission can be obtained by comparing the mutual information between the stimulus and the empirical distribution of activity configurations \( p_{\text{obs}}(\{\sigma\}) \) with the mutual information between the stimulus and the distributions defined by ME models of higher and higher orders: in [148], for instance, the authors show that third order correlations not only are necessary to explain the probability of synchronous discharge, but they also play an important role in transmitting information about the stimulus, since the
third order (but not the pairwise) ME model well approximates the mutual information computed with the empirical distribution.

A Markovian ME model

The ME principle has also been used with a Markovian assumption to study spatio-temporal patterns of limited duration (2 time-bins) in [150]: a Markovian ME model for 2-time-bin patterns is developed with the constraints on the single neuron spike probabilities \( \langle \sigma_i,t \rangle_t \), on the zero-time-lag pairwise correlations \( \langle \sigma_i,\sigma_j,t \rangle_t \) and on the one-time-lag pairwise correlations \( \langle \sigma_i,\sigma_j,t+1 \rangle_t \). The stationary ME distributions for the 1-time-bin and 2-time-bin patterns satisfying these constraints are

\[
p_{\text{Markov}}(\{\sigma_i\}) = \frac{e^{\sum_i h_i^{(1)} \sigma_i + \sum_{i,j} j_{ij}^{(1)} \sigma_i \sigma_j}}{Z^{(1)}}
\]

\[
p_{\text{Markov}}(\{\sigma_i\},\{\sigma'_i\}) = \frac{e^{\sum_i h_i^{(1)} \sigma_i + \sum_{i,j} j_{ij}^{(2)} \sigma_i \sigma'_j + \sum_{i,j} h_i^{(3)} \sigma'_i + \sum_{i,j} j_{ij}^{(3)} \sigma'_i \sigma'_j + \sum_{i,j} j_{ij}^{(4)} \sigma_i \sigma'_j}}{Z^{(2)}}
\]

and the transition probability is given by \( p_{\text{Markov}}(\{\sigma_i\}|\{\sigma'_i\}) = p_{\text{Markov}}(\{\sigma_i\},\{\sigma'_i\})/p_{\text{Markov}}(\{\sigma'_i\}) \). The marginalization condition \( p_{\text{Markov}}(\{\sigma_i\}) = \sum_{\{\sigma'_i\}} p_{\text{Markov}}(\{\sigma_i\},\{\sigma'_i\}) \) binds parameters \( \{h_i^{(1)}, j_{ij}^{(1)}\} \) to parameters \( \{h_i^{(2)}, j_{ij}^{(2)}, h_i^{(3)}, j_{ij}^{(3)}, j_{ij}^{(4)}\} \), and the normalization condition on the transition probability binds \( \{h_i^{(3)}, j_{ij}^{(3)}\} \) to \( \{h_i^{(2)}, j_{ij}^{(2)}, j_{ij}^{(4)}\} \). Therefore, the only free parameters are \( \{h_i^{(2)}, j_{ij}^{(2)}, j_{ij}^{(4)}\} \), which are determined with the Boltzmann Machine learning algorithm (presented in the next paragraph). As shown in [150], the probability of occurrence of spatio-temporal patterns up to 120 ms long in the cat parietal cortex during SWS is predicted much better by this Markov ME model than by the Ising model, despite temporal correlations used to fit the model are only one-time-lag pairwise correlations. No significative improvement with respect to the Ising model prediction is found for spatio-temporal patterns recorded during REM sleep, coherently with the smaller correlation time constants measured during these periods compared to SWS.

2.2.2 Inference of ME models

As anticipated in 2.2, every inverse problem is an optimization problem: it consists in searching in the space of model parameters those values for which the model reproduces the data as closely as possible. From a mathematical point of view, the problem can be formulated in several equivalent ways, using, for instance, the concepts of log-likelihood, cross-entropy or Kullback-Leibler divergence.

Several possible formulations

The log-likelihood of a dataset composed by \( B \) configurations \( \sigma_{\text{obs}} = \{\{\sigma_1^1\}, \ldots, \{\sigma_t^1\}, \ldots, \{\sigma_B^B\}\} \) (\( B \) number of time-bins) is simply the logarithm of the probability \( p_{\text{like}}(\sigma_{\text{obs}}|\lambda) \) of sampling the observed neuronal configurations \( \sigma_{\text{obs}} \) with the model distribution, having parameters \( \lambda = \{\lambda^1, \ldots, \lambda^M, \ldots, \lambda^L\} \). Inferring the optimal model parameters means finding the values of \( \{\lambda^t\} \) which maximize the log-likelihood. In general \( p_{\text{like}}(\sigma_{\text{obs}}|\lambda) = \prod_{t=1}^B p_{\text{model}}(\{\sigma_t^t\}|\lambda), \) with \( t' < t \). For stationary models, the likelihood simplifies to \( p_{\text{like}}(\sigma_{\text{obs}}|\lambda) = \prod_{t=1}^B p_{\text{model}}(\{\sigma_t^t\}|\lambda) \), from which we obtain the following simple relation
between the log-likelihood and the cross-entropy \( S^\ast \) (defined by eq. 2.10) between the model and the observed distribution of neuronal configuration:

\[
\log p^{\text{like}}(\sigma|\lambda) = B(p_{\text{model}}(\{\sigma_i\}|\lambda))_{\text{obs}} = -BS^\ast
\]  

(2.15)

The cross-entropy is also strictly related to the Kullback-Leibler divergence of the model distribution from the empirical one, defined as:

\[
D_{KL}(p_{\text{obs}}(\{\sigma_i\})|p_{\text{model}}(\{\sigma_i\})) = \sum_{\{\sigma_i\}} p_{\text{obs}}(\{\sigma_i\}) \log \frac{p_{\text{obs}}(\{\sigma_i\})}{p_{\text{model}}(\{\sigma_i\})} = S^\ast - S_{\text{obs}}
\]  

(2.16)

with \( S_{\text{obs}} = -\sum_{\sigma} p(\sigma)_{\text{obs}} \log p(\sigma)_{\text{obs}} \) is the entropy of the empirical distribution and does not depend on the model. \( D_{KL} \) represents the information lost when \( p_{\text{model}} \) is used to approximate \( p_{\text{obs}} \) and is often used as a measure of the difference between two distributions. \( D_{KL} \) is always non negative for Jensen’s inequality:

\[
- D_{KL}(p_{\text{obs}}(\{\sigma_i\})|p_{\text{model}}(\{\sigma_i\})) = \sum_{\{\sigma_i\}} p_{\text{obs}}(\{\sigma_i\}) \log \frac{p_{\text{model}}(\{\sigma_i\})}{p_{\text{obs}}(\{\sigma_i\})} \leq
\]

\[
\leq \log \sum_{\{\sigma_i\}} p_{\text{obs}}(\{\sigma_i\}) \frac{p_{\text{model}}(\{\sigma_i\})}{p_{\text{obs}}(\{\sigma_i\})} = \log \sum_{\{\sigma_i\}} p_{\text{model}}(\{\sigma_i\}) = 0
\]

In conclusion, the problem of finding the optimal values for the model parameters can be formulated as the problem of maximizing the log-likelihood, or of minimizing \( S^\ast \), or equivalently of minimizing \( D_{KL} \). In ME models, this also means finding the parameter values for which the model predictions for the quantities to be reproduced \( \{f^\mu(\{\sigma_i\})\}_{\text{model}} \) match the empirical averages of those quantities \( \{f^\mu(\{\sigma_i\})\}_{\text{obs}} \). Indeed, substituting 2.6 and 2.7 into 2.10 one obtains:

\[
S^\ast = \log Z - \sum_{\mu=1}^{L} \lambda^\mu \langle f^\mu(\{\sigma_i\}) \rangle_{\text{obs}}
\]  

(2.17)

So, for a ME model, the gradient of \( S^\ast \) with respect to \( \lambda \) is:

\[
\nabla S^\ast = \langle f(\{\sigma_i\}) \rangle_{\text{model}} - \langle f(\{\sigma_i\}) \rangle_{\text{obs}}
\]  

(2.18)

with \( f = \{f^1, \ldots, f^L\} \). Therefore, if the solution of the minimization problem exists and is unique, this solution is the parameter set \( \lambda \) for which \( \langle f(\{\sigma_i\}) \rangle_{\text{model}} = \langle f(\{\sigma_i\}) \rangle_{\text{obs}} \).

The Hessian of \( S^\ast \) is the \( L \times L \) matrix of elements:

\[
\lambda_{\mu\nu} = \frac{\partial^2 S^\ast}{\partial \lambda^\mu \partial \lambda^\nu} = \langle f^\mu(\{\sigma_i\}) f^\nu(\{\sigma_i\}) \rangle_{\text{model}} - \langle f^\mu(\{\sigma_i\}) \rangle_{\text{model}} \langle f^\nu(\{\sigma_i\}) \rangle_{\text{model}}
\]  

(2.19)

Since the Hessian is a covariance matrix, it is positive semi-definite; therefore, \( S^\ast \) is a convex (but not necessarily strictly convex) function of the model parameters and the minimum of \( S^\ast \) may not be unique or finite. In practice, this happens when sampling is poor: if a quantity to be reproduced, e.g. \( \langle f^\mu(\{\sigma_i\}) \rangle_{\text{obs}} \), is zero because the number of time bins in the recording is not large enough compared to \( \langle f^\mu(\{\sigma_i\}) \rangle_{\text{obs}}^{-1} \) then \( \nabla S^\ast = 0 \) for \( \lambda^\mu = -\infty \). An inverse problem is well-posed only if its solution is unique and finite. To ensure the uniqueness and finiteness of the minimum of \( S^\ast \), a standard approach is to add a regularization to \( S^\ast \), which makes the Hessian positive definite, that is the cross-entropy strictly convex.
Adding a regularization to $S^*$ not only makes the inverse problem well-posed, but also slightly changes the purpose of inference: if sampling is poor, the aim should not be anymore to find the parameter set $\lambda$ for which the model reproduces the data as close as possible (i.e., to maximize the log-likelihood), since this would lead to fit not only the signal but also the noise present in the data; the aim becomes that of reaching a compromise between finding the values which maximize the log-likelihood and at the same time match our prior knowledge (or some reasonable assumptions) about the parameter distribution. This kind of inference is called Bayesian inference. More precisely, Bayes rule expresses the posterior probability distribution of the model parameters given the data in terms of the likelihood of the data given the model and the prior probability distribution over model parameters:

$$p_{\text{Post}}(\lambda | \sigma_{\text{obs}}) = \frac{p^\text{like}(\sigma_{\text{obs}} | \lambda) p^\text{prior}(\lambda)}{p(\sigma_{\text{obs}})} \propto p^\text{like}(\sigma_{\text{obs}} | \lambda) p^\text{prior}(\lambda)$$ (2.20)

where $p(\sigma_{\text{obs}})$ is a normalization constant, independent of $\lambda$. Bayesian inference consists in searching for the vector $\lambda$ which maximizes the posterior probability. A typical prior assumption about model parameters is that they are distributed as a normal law with zero mean and variance $a$; this Gaussian prior leads to:

$$\log p_{\text{Post}}(\lambda | \sigma_{\text{obs}}) = \log p^\text{like}(\sigma_{\text{obs}} | \lambda) + \log \left( e^{-\frac{1}{2a} \sum_{\mu} (\lambda_{\mu})^2} \right) + C(\sigma_{\text{obs}}, a) =$$

$$= -B \left( S^* + \frac{1}{2aB} \sum_{\mu} (\lambda_{\mu})^2 \right) + C(\sigma_{\text{obs}}, a)$$

with function $C(\sigma_{\text{obs}}, a)$ independent of the parameter vector $\lambda$. Maximizing this quantity with respect to $\lambda$ is equivalent to minimizing the regularized cross-entropy $S^\text{reg}$:

$$S^\text{reg} = S^* + \frac{1}{2aB} \sum_{\mu} (\lambda_{\mu})^2$$ (2.21)

Another popular choice for the prior is the Laplacian prior distribution with zero mean and diversity $b$, from which the following expressions for the posterior distribution and the regularized cross-entropy are obtained:

$$\log p_{\text{Post}}(\lambda | \sigma_{\text{obs}}) = \log p^\text{like}(\sigma_{\text{obs}} | \lambda) + \log \left( e^{-\frac{1}{b} \sum_{\mu} |\lambda_{\mu}|} \right) + C(\sigma_{\text{obs}}, b) =$$

$$= -B \left( S^* + \frac{1}{bB} \sum_{\mu} |\lambda_{\mu}| \right) + C(\sigma_{\text{obs}}, b)$$

$$S^\text{reg} = S^* + \frac{1}{bB} \sum_{\mu} |\lambda_{\mu}|$$ (2.22)

The value of the variance $a$ (or the diversity $b$) in the prior distribution can be determined maximizing the marginal-likelihood of the data (or an approximation of it) $p(\sigma_{\text{obs}}) = \int p(\sigma_{\text{obs}} | \lambda) p(\lambda) d\lambda$, which can be seen as the probability of the data over all possible values of model parameters. The Laplacian (or L1) regularization, compared to the Gaussian (or L2) regularization, favors sparser solutions, that is with many parameters set to zero, and is to be preferred when it is known that the structure generating the data (represented by the parameter matrix) is sparse, or to obtain a compressed representation of the data.
In many cases, the prior is chosen to be not uniform only for some subsets of the model parameters, those which are more affected by sampling noise (as in the example shown in the next section). The regularization added to the cross-entropy $S^*$ is inversely proportional to the duration of the recording $B$: the smaller $B$ (short recordings), the larger the weight of the prior distribution with respect to the likelihood in 2.20, in agreement with the unreliability of the data, which are more and more corrupted by sampling noise. Conversely, for $B \to \infty$ (perfect sampling), the regularization term vanishes and Bayesian inference is equivalent to maximizing the log-likelihood.

It is possible to compute analytically the sampling errors on the inferred parameters from the Hessian of the regularized cross-entropy: indeed, the posterior distribution is proportional to $e^{-BS^*_{\text{reg}}}$, which for large $B$ is concentrated around the minimum of $S^*_{\text{reg}}$, that is the value obtained with the inferred parameter vector $\bar{\lambda}$; expanding $S^*_{\text{reg}}$ up to the second order in $\lambda$ around this minimum, we obtain that $p_{\text{post}}(\lambda | \sigma_{\text{obs}}) \propto \exp\left(-\frac{J^2}{2\chi^{-1}B}\right)$, with $\chi = \frac{\partial^2 S^*_{\text{reg}}}{\partial J^2} \bigg|_{\bar{\lambda}}$. Thus, the posterior distribution is asymptotically Gaussian with covariance matrix $\frac{1}{B}\chi^{-1}$ and the errors on the inferred parameters are:

$$\delta \lambda^\mu = \sqrt{\frac{1}{B} (\chi^{-1})_{\mu,\mu}} \quad (2.23)$$

Matrix $\chi$ is the covariance matrix 2.19, evaluated in $\bar{\lambda}$, with the addition of the term deriving from the regularization of $S^*$; therefore calculating $\chi$ requires calculating the multi-neuron correlations with respect to the inferred model distribution, which can be done with Monte Carlo simulations. In practical applications, a good approximation of $\chi$ is usually computed much faster from the empirical multi-neuron correlations.

In the following part of this paragraph, I will focus on the inverse problem for the Ising model, the most popular among the maximum entropy models.

**The inverse Ising problem**

The cross-entropy 2.17 for the Ising model is:

$$S^* = \log Z - \sum_i h_i p_i - \sum_{ij} J_{ij} p_{ij}$$

with $p_i = \langle \sigma_i \rangle_{\text{obs}}$ and $p_{ij} = \langle \sigma_i \sigma_j \rangle_{\text{obs}}$. In neuronal data, firing rates are generally well sampled and no regularization is needed for the fields $\{h_i\}$, while couplings $\{J_{ij}\}$ have to be regularized to ensure the uniqueness and finiteness of the solution of the inverse problem. The L2 and L1 regularized cross-entropies are therefore:

$$S^*_{L2} = \log Z - \sum_i h_i p_i - \sum_{ij} J_{ij} p_{ij} + \frac{1}{2aB} \sum_{i<j} J_{ij}^2 \quad (2.25a)$$

$$S^*_{L1} = \log Z - \sum_i h_i p_i - \sum_{ij} J_{ij} p_{ij} + \frac{1}{bB} \sum_{i<j} |J_{ij}| \quad (2.25b)$$

The Laplacian (or L1) regularization favors solutions with many zero couplings, that is sparse interaction networks, while the Gaussian (or L2) regularization does not enforce sparsity. Both regularizations penalize large coupling values and are the simplest, most neutral choices to approximate realistic distributions of synaptic strengths, characterized
by the presence of few strong synapses among many weaker ones [151, 152] and by a vast majority of unconnected neurons: in the human cortex, for instance, the estimated number of synapses is $10^{15}$ for $10^{11}$ neurons, that is about $10^{-7}$ the number of synapses there would be in a fully connected network. Indeed, the probability that two neurons are synaptically connected in the brain decreases rapidly with distance: the majority of synapses, developing according to local spatial growth rules, link neurons that are only a few hundred micrometers apart, and form local, closely connected clusters with specific functional properties [153]. Although less in number, long-range connections are also present in real cortical networks, ensuring short average path length (i.e. small number of intermediate synapses) between physically distant cortical sites and allowing rapid integration of different pieces of information. However, precisely modeling these ‘small world’ attributes hypothesized for real cortical networks [154] through a complicated prior distribution of the couplings is impracticable and would strongly bias the inference, resulting in poor exploitation of the information content of the data. In [155] the authors point out that the distribution of synaptic strengths between connected cortical neurons is roughly lognormal. However, if one takes into account the vast majority of unconnected neurons, the distribution of synaptic strengths can be considered Gaussian in good approximation. An L1 or an L2 regularization is in general to be preferred to more sophisticated choices, for its simplicity and given the large amount of information which is still missing on the precise connectivity matrix of the brain [156]. Moreover, as already said, model couplings are not meant to represent anatomical synapses, but functional interactions.

Minimizing $S_{\text{reg}}^*$ with respect to $\{h_i, J_{ij}\}$ means solving the set of equations:

\begin{align}
\langle \sigma_i \rangle_{\text{Ising}} &= p_i \\
\langle \sigma_i \sigma_j \rangle_{\text{Ising}} &= p_{ij} - \frac{\partial R}{\partial J_{ij}} (2.26a, 2.26b)
\end{align}

with $i, j = 1, \ldots, N$ ($R =$ regularization added to $S^*$).

This set of equations can not be solved analytically for systems of more than a few neurons. The problem is the computation of the partition function $Z$ (needed to obtain the model averages $\langle \sigma_i \rangle_{\text{Ising}}$ and $\langle \sigma_i \sigma_j \rangle_{\text{Ising}}$), which requires a time growing exponentially with $N$, since $Z$ is a sum over $2^N$ configurations.

Some inference algorithms compute approximations of the gradient of $S_{\text{reg}}^*$ avoiding the computation of $Z$, and exploit these estimations to find an approximate solution to the minimization of $S_{\text{reg}}^*$ with gradient descent or Newton’s method. Other inference approaches rely on techniques, developed in the statistical physics of spin glasses, which enable the calculation of $\log Z$ under certain assumptions on the distribution of model parameters, and of the corresponding $\langle \sigma_i \rangle_{\text{Ising}}$ and $\langle \sigma_i \sigma_j \rangle_{\text{Ising}}$; the set of equations 2.26 is then solved analytically (when possible) or numerically. However, the assumptions on the parameter space under which $\log Z$ can be computed in polynomial time are often not justified for real systems of interacting neurons, and the resulting approximation of the solution may be poor on real data.

I will present below some representative examples of these two classes of inference methods, and a new efficient technique, based on a cluster expansion of the entropy, recently proposed in [157].

**The Boltzmann Machine learning algorithm**

In the Boltzmann Machine (BM) learning algorithm [158], minimization of $S_{\text{reg}}^*$ is performed with gradient descent: at each step, couplings and fields are updated with the
where $\alpha_h$, $\alpha_J$ are the learning rates, and the regularization $R$ on the cross-entropy is often neglected in common applications. The algorithm halts when $\Delta h_i < \epsilon$ $\forall i$ and $\Delta J_{ij} < \epsilon$ $\forall i,j$. The averages $\langle \sigma_i \rangle_{Ising}$ and $\langle \sigma_i \sigma_j \rangle_{Ising}$ are computed through Monte Carlo simulations (e.g. with the Metropolis algorithm). The interesting feature of the BM learning approach is that it finds a good approximation of the solution of the minimization problem exploiting only locally available information: variations of $h_i$ and $J_{ij}$ at each step depend on units $i$ and $j$ only; nevertheless, they optimize a global measure (function of all $N(N + 1)/2$ variables). Despite this advantage, BM learning can be very slow, especially for large $N$ [159], because at each parameter update 2.27 an entire new set of Monte Carlo samples from the initial condition $h_0, J_{ij}$ is run, and parameter updates, thus reducing the computational time of a factor $\sim 1/T$. The idea of histogram Monte Carlo is based on the following relation, which expresses the average value of a generic function $\Phi$ with respect to a distribution $p_\lambda = e^{-\mathcal{H}_\lambda(\sigma)}/Z(\lambda)$ in terms of the ratio between expectation values with respect to another distribution $p_{\lambda'} = e^{-\mathcal{H}_{\lambda'}(\sigma)}/Z(\lambda')$:

$$
\langle \Phi(\sigma) \rangle_{\lambda'} = \sum_\sigma p_{\lambda'}(\sigma) \Phi(\sigma) = \sum_\sigma \left( \frac{p_{\lambda'}(\sigma)}{p_\lambda(\sigma)} \right) p_\lambda(\sigma) \Phi(\sigma) = \frac{\langle p_{\lambda'}(\sigma) \Phi(\sigma) \rangle_\lambda}{Z(\lambda)} = \frac{Z(\lambda)}{Z(\lambda')} \langle \Phi(\sigma) e^{-(\lambda' - \lambda) \mathcal{H}_\lambda(\sigma)} \rangle_\lambda = \langle \Phi(\sigma) e^{-(\lambda' - \lambda) \mathcal{H}_\lambda(\sigma)} \rangle_{\lambda'} \approx \langle \Phi(\sigma) e^{-(\lambda' - \lambda) \mathcal{H}_\lambda(\sigma)} \rangle_{MC\lambda} \approx \langle \Phi(\sigma) e^{-(\lambda' - \lambda) \mathcal{H}_{\lambda'}(\sigma)} \rangle_{MC\lambda}$$

where $\langle \cdot \rangle_\lambda$ denotes the expectation value with respect to $p_\lambda$. The approximation only comes at the last point, where the expectation values with respect to the distribution $p_\lambda$ are replaced with the expectation values over a set of samples drawn from a Monte Carlo simulation of the same distribution. In this case, $p_\lambda$ is the Ising distribution with parameters inferred at the $kT + 1$ step (an iteration in which a new Monte Carlo simulation is run), and $p_{\lambda'}$ is the Ising distribution at all $kT + q$ steps with $q = 1, \ldots, T$ (iterations in which Monte Carlo samples from the $kT + 1$ step are exploited). A careful choice of the initial condition $h_0, J_{ij}$ (e.g. an estimate of the parameters obtained with faster, though usually less accurate, methods like Mean Field) can also considerably reduce the computational time to reach thermalization.

Despite this improvements, BM algorithms are generally still slow, and they may have convergence problems in the presence of long-range correlations (when the system is critical) because a little change in the $J_{ij}$ affects correlations at large scales and many Monte Carlo steps are required at the successive iteration to reach the thermalization...
condition. Moreover, the issue of overfitting is not controlled perfectly with this method, and can be contained only through the addition of a regularization, as in most available methods. A remarkable improvement in this direction comes from a new algorithm that will be presented in section 2.2.2.

The Pseudo-likelihood-based algorithm

The Pseudo-likelihood method [161] consists in simplifying the original N-body problem into N independent one-body problems, with \( N - 1 \) quenched variables:

\[
p_{\text{Ising}}(\{\sigma_i\}) \approx \prod_i p_{\text{Ising}}(\sigma_i|\{\sigma_j\}_{j \neq i}) = \prod_i \frac{e^{\sigma_i (h_i + \sum_j J_{ij}\sigma_j)}}{1 + e^{h_i + \sum_j J_{ij}\sigma_j}} = p_{\text{PS}}^{\text{Ising}}(\{\sigma_i\})
\]

From this approximation, the Pseudo-entropy 2.10 is derived:

\[
S^{PS} = -\langle \log p_{\text{PS}}^{\text{Ising}}(\{\sigma_i\}) \rangle_{\text{obs}} = \sum_i \left(-h_i p_i - \sum_j J_{ij} p_{ij} + \frac{1}{B} \sum_{t=1}^B \log (1 + e^{h_i + \sum_j J_{ij}\sigma_j(t)}) \right)
\]

\[
= \sum_i S_i^{PS}
\]

(2.29)

Each \( S_i^{PS} \) is a convex function, which becomes strictly convex upon addition of a regularization term, and can be minimized with gradient descent techniques. The gradient of \( S_i^{PS} \) can in fact be calculated directly from the empirical configurations as:

\[
\frac{\partial S_i^{PS}}{\partial h_i} = \frac{1}{B} \sum_{t=1}^B \frac{e^{h_i + \sum_j J_{ij}\sigma_j(t)}}{1 + e^{h_i + \sum_j J_{ij}\sigma_j(t)}} - p_i
\]

(2.30a)

\[
\frac{\partial S_i^{PS}}{\partial J_{ij}} = \frac{1}{B} \sum_{t=1}^B \sigma_j(t) \frac{e^{h_i + \sum_j J_{ij}\sigma_j(t)}}{1 + e^{h_i + \sum_j J_{ij}\sigma_j(t)}} - p_{ij}
\]

(2.30b)

from which it can be noticed that the Pseudo-likelihood approximation leads to the Callen’s identities for the Ising model:

\[
\langle \sigma_i \rangle_{\text{Ising}} \approx \frac{1}{B} \sum_{t=1}^B \frac{e^{h_i + \sum_j J_{ij}\sigma_j(t)}}{1 + e^{h_i + \sum_j J_{ij}\sigma_j(t)}}
\]

\[
\langle \sigma_i \sigma_j \rangle_{\text{Ising}} \approx \frac{1}{B} \sum_{t=1}^B \sigma_j(t) \frac{e^{h_i + \sum_j J_{ij}\sigma_j(t)}}{1 + e^{h_i + \sum_j J_{ij}\sigma_j(t)}}
\]

Differently from the BM learning algorithm, the Pseudo-likelihood approach requires knowledge of the full empirical configurations and not only of the neuron firing rates and pairwise correlations. Inferred couplings \( J_{ij} \) and \( J_{ji} \) are in general not equal and, in some applications [162], their average is simply taken as an estimate of the model symmetric \( J_{ij} \). Applications of the Pseudo-likelihood algorithm are more natural in community detection problems, where an L1 regularization on \( S^{PS} \) is typically added to enforce sparsity and what matters is not the precise value of the couplings, but which ones are different from zero, in order to separate neighborhoods on the interaction graph (i.e. communities of...
strongly interacting variables). It has been shown [163] that neighborhoods are correctly
recovered with high probability when data are generated by an Ising model (which is likely
to be false in real neuronal systems) with a sparse matrix of strong enough couplings
-scaling as \( \sqrt{d \log N/B} \), \( d \) being the maximum community size); the computational
time is polynomial in \( N \).

**Mean Field methods**

Mean Field (MF) approximations consist in deriving a simplified expression for
\( \log Z \), from which it is possible to compute \( m_i = \langle \sigma_i \rangle_{MF} \) (called magnetizations) and
\( C_{ij} = \langle \sigma_i \sigma_j \rangle_{MF} - m_i m_j \) (\( \langle \cdot \rangle_{MF} \) average with respect to the Ising distribution in the MF
approximation); \( \log Z_{MF} \), \( m_i \) and \( C_{ij} \) are good approximations of the exact \( \log Z_{Ising} \),
\( \langle \sigma_i \rangle_{Ising} \) and \( (C_{Ising})_{ij} = \langle \sigma_i \sigma_j \rangle_{Ising} - \langle \sigma_i \rangle_{Ising} \langle \sigma_j \rangle_{Ising} \) only when the coupling matrix of
the Ising model is of a particular form, as briefly discussed below.

The simplest MF approximation, also called naïve MF, is obtained expanding the
Ising Hamiltonian in fluctuations \( \delta \sigma_i \) and neglecting the second order terms \( \delta \sigma_i \delta \sigma_j \) where
\( \delta \sigma_i = \sigma_i - m_i \):

\[
\mathcal{H} = - \sum_{i<j} J_{ij} \sigma_i \sigma_j - \sum_i h_i \sigma_i = \\
= - \sum_{i<j} J_{ij} (m_i + \delta \sigma_i)(m_j + \delta \sigma_j) - \sum_i h_i \sigma_i \approx \\
\approx - \sum_{i<j} J_{ij} (m_i m_j + m_i \delta \sigma_j + m_j \delta \sigma_i) - \sum_i \sigma_i = \\
= \frac{1}{2} \sum_{ij} J_{ij} m_i m_j - \sum_i h_{i}^{eff} \sigma_i = \\
= \mathcal{H}_{MF}
\]

with:

\( h_{i}^{eff} = h_i + \sum_j J_{ij} m_j \)

(couplings \( J_{ii} \) are set to zero). In this way, the N-body system is replaced by N one-body
systems (similarly to the Pseudo-likelihood approach), subject to an effective mean field
\( h_{i}^{eff} \), representing the average effect of all components \( j \neq i \) on the individual component \( i \).
This approximation holds for dense networks of weak interactions [164, 165] (for example
in the Curie-Weiss model of a ferromagnet, where all interactions are equal and of the
order of \( 1/N \), or in the Sherrington-Kirkpatrick model, where all neurons interact with
one another through couplings of the order of \( \sim N^{-1/2} \), after addition of the Onsager term,
see below), but is poor for dilute graphs with strong interactions.

Within this approximation, the partition function is easily obtained:

\[
Z_{MF} = e^{-\frac{1}{2} \sum_{ij} J_{ij} m_i m_j} \prod_i \left( 1 + e^{h_i^{eff}} \right)
\]

Substituting \( \log Z_{MF} \) into the exact relation \( \langle \sigma_i \rangle_{Ising} = \frac{\partial \log Z_{Ising}}{\partial h_i} \), one obtains the
self-consistency equations for the magnetizations:

\[
m_i = \frac{\exp(h_i + \sum_j J_{ij} m_j)}{1 + \exp(h_i + \sum_j J_{ij} m_j)} \quad (2.32)
\]
Inverting 2.32 and calculating the linear response $\frac{\partial h_i}{\partial m_j} = (C_{\text{Ising}}^{-1})_{ij}$, it is straightforward to obtain the couplings and the fields as functions of the magnetizations and the connected correlations:

$$J_{ij}^{\text{MF}} = -(C^{-1})_{ij} \quad (2.33a)$$

$$h_i^{\text{MF}} = \log \left( \frac{m_i}{1 - m_i} \right) + \sum_{j \neq i} (C^{-1})_{ij} m_j \quad (2.33b)$$

Equating $m_i$ and $C_{ij}$ to the empirical firing rates and connected correlations allows us to infer the parameter values from the data in a time polynomial in $N$.

Several improvements to the naïve MF approximation have been proposed, starting from the so-called TAP approximation [166], in which the effective field $h_i^{\text{eff}}$ is corrected with the addition of a term, called the Onsager reaction term:

$$h_i^{\text{eff}} = h_i + \sum_j J_{ij}(m_j - J_{ij}(1 - m_j^2)m_i)$$

Intuitively, the Onsager reaction term is aimed at removing the effect of variable $\sigma_i$ on the marginal probabilities $m_j$ of all other variables when computing $m_i$. The TAP self-consistency equations for the magnetizations $m_i$ can be obtained with the variational method, using the Plefka expansion of the Gibbs free energy. The purpose is to find a tractable distribution $Q(\sigma)$ (within a certain family) which approximates the intractable distribution $P(\sigma) = \frac{e^{-\mathcal{H}}}{Z}$, with $\mathcal{H} = -\sum_i h_i \sigma_i - \sum_{i<j} J_{ij} \sigma_i \sigma_j$ Ising Hamiltonian. The best distribution $Q(\sigma)$ in the family is the one which minimizes the Kullback-Leibler divergence

$$D(Q||P) = \sum_{\sigma} Q(\sigma) \log \frac{Q(\sigma)}{P(\sigma)} = \log Z + E(Q) - S(Q) = \log Z + F(Q)$$

with $S(Q)$ entropy of the distribution $Q$, $E(Q) = \langle \mathcal{H} \rangle_Q$ variational energy and $F(Q) = E(Q) - S(Q)$ variational free energy. When $Q = P$, $D(Q||P) = 0$ and the variational free energy equals the exact free energy $F(P) = -\log Z$. Since $Z$ does not depend on $Q$, minimizing $D(Q||P)$ is equivalent to minimizing $F(Q)$. $F(Q)$ can be minimized in two steps [167]. In the first step, minimization is performed in the generic family of distributions $Q$, with the constraint $\langle \sigma \rangle_Q = m$. By introducing the Lagrange multipliers $\theta_i$ to enforce the constraint on the magnetizations, this minimization leads to:

$$Q(\sigma) = \frac{e^{-\mathcal{H}(\sigma) + \sum_i \theta_i \sigma_i}}{Z(\mathcal{H})}$$

In the second step, the Gibbs free energy $G(m) = \min_Q \{ F(Q)|\langle \sigma \rangle_Q = m \}$ should be minimized with respect to $m$. Since no approximation has been introduced so far, the solution to this last minimization problem can not be found analytically, unless an approximation of $G(m)$ is introduced. An approximation that enables to solve the problem analytically is obtained through the Plefka expansion of $G(m)$ in small couplings [168], which consists in multiplying $\mathcal{H}$ by a small parameter $\alpha$ and expanding $G(m)$ in powers of $\alpha$ at fixed $m$ and $\{ J_{ij} \}$ (but with fields $\{ h_i + \theta_i = f_i(\alpha) \}$). Truncating the expansion at the second order, and then restoring the value $\alpha = 1$, the following expression for $G(m)$ is obtained (in 0,1 variables):

$$G(m) = \sum_i \left( m_i \log m_i + (1 - m_i) \log(1 - m_i) \right) - \sum_i h_i m_i - \sum_{i<j} J_{ij} m_i m_j$$

$$- \frac{1}{2} \sum_{i<j} J_{ij}^2 (1 - m_i^2)(1 - m_j^2) \quad (2.34)$$
Minimizing this approximation of the Gibbs free energy with respect to $m$ leads to the TAP self-consistency equations

$$m_i = \frac{\exp(h_i + \sum_j J_{ij} m_j - \sum_j J_{ij}^2 (1 - m_j^2) m_i)}{1 + \exp(h_i + \sum_j J_{ij} m_j - \sum_j J_{ij}^2 (1 - m_j^2) m_i)}$$  \hspace{1cm} (2.35)

from which, computing the linear response, one obtains that each coupling is solution of this simple equation:

$$- J_{ij}^{TAP} - 2 (J_{ij}^{TAP})^2 m_i m_j = (C^{-1})_{ij}$$  \hspace{1cm} (2.36)

The Plefka expansion up to the first order gives back the naïve MF approximation, which can also be obtained minimizing, directly with respect to $m$, the variational free energy computed for the product distribution $Q = \prod_i (\sigma_i m_i + (1 - \sigma_i)(1 - m_i))$. The derivation by Plefka expansion clearly shows why these approximations are good only in the limit of weak couplings (becoming exact for $N \to \infty$ in the SK model).

Recently, Sessak and Monasson [169] derived another approximation for the inverse Ising problem, which outperforms the TAP approximation [170], performing a power expansion in small correlations around the non-interacting case: this expansion is obtained expressing $S(Q)$ in terms of magnetizations and connected correlations, multiplying the connected correlations by a small parameter $\beta$ and expanding $S(Q)$ in powers of $\beta$ at fixed $m$ and $C$ (while $\{J_{ij} = J_{ij}(\beta)\}$, $\{h_i = h_i(\beta)\}$). Expressions for $\{J_{ij}\}$ and $\{h_i\}$ in terms of magnetizations and connected correlations are finally obtained restoring $\beta = 1$ at the end of the calculation.

Another well known approximation, which holds not only for small couplings but also for couplings of $O(1)$, is the Bethe approximation (BA): this one, however, is exact only when the interaction network is a tree. Indeed, in the BA, the free energy is obtained as the difference between the energy at given magnetizations and neighbouring correlations and the entropy of the following approximation (exact for trees) of the joint probability distribution of the $N$ variables:

$$P(\sigma) \approx \prod_{\langle ij \rangle} \frac{p_i(\sigma_i) p_j(\sigma_j)}{p_i(\sigma_i) p_j(\sigma_j)} \prod_i p_i(\sigma_i)$$

where the first product runs over all pairs of neighboring variables. The solution to the problem of minimizing the Bethe free energy is obtained with an iterative algorithm known as Belief Propagation (BP) [171]. The idea of BP is to recursively compute marginals $P_i(\sigma_i)$ exploiting the tree-like structure of the graph. The starting point is the exact expression:

$$P_i(\sigma_i) = \sum_{\sigma\setminus\sigma_i} P(\sigma) \propto \sum_{\sigma\setminus\sigma_i} e^{\sigma_i(h_i + \sum_j J_{ij} \sigma_j)} P(\sigma \setminus \sigma_i)$$

where $P(\sigma \setminus \sigma_i)$ is the joint distribution for the system $\sigma \setminus \sigma_i$ containing all variables $\sigma_j$ with $j \neq i$ (and all couplings except those connecting $j$ to $i$). If the graph is a tree, $P(\sigma \setminus \sigma_i)$ factorizes into the product of individual marginals $P_{j\setminus i}(\sigma_j)$; therefore $P_i(\sigma_i)$ can be rewritten as:

$$P_i(\sigma_i) \propto \prod_j \left( \sum_{\sigma_j} e^{J_{ij} \sigma_i \sigma_j} P_{j\setminus i}(\sigma_j) \right)$$

This reasoning can be applied recursively to express each marginal $P_{j\setminus i}(\sigma_j)$ in terms of its neighbors (excluding $i$), in order to obtain a close set of equations: this procedure leads to the BP algorithm. BP and other similar message passing algorithms [172] are guaranteed to
2.2. MODEL-BASED METHODS

converge on graphs that are trees or have long-range loops, but do not work for networks with strongly interacting clusters of neurons. Message passing algorithms are developed and used to solve optimization problems in many different fields beyond neuroscience, like the satisfiability of Boolean formulas [173–175], the reconstruction of the patient zero of an epidemic outbreak [176] (e.g. in social and computer networks), or the reconstruction of discrete images from tomographic measurements [177].

The Adaptive Cluster Expansion algorithm

Cocco and Monasson [178] have recently proposed a new method for the inverse Ising problem: the Adaptive Cluster Expansion algorithm (ACE). In summary, this algorithm (see also [157, 179]) is based on an expansion of the entropy \( S \) where, for 1-neuron clusters, \( \Delta S \) with numerical methods (gradient descent or Newton’s method, or a combination of both) with \( i, j \) only \( \Delta S \) in time growing polynomially with \( \sim N \) partition function can be calculated exactly in time of clusters \( \Gamma \) (defined as the minimum of the regularized cross-entropy between the model and the data \( S^* \) coming from all clusters of neurons contained in the system. Many clusters in this expansion can (and should) be discarded because their entropies reflect the sampling noise: only clusters whose entropies reflect the signal present in the data are kept in the expansion. Minimization of the regularized cross-entropy \( S^*_\Gamma \) for each relevant subset \( \Gamma \) allows one to obtain the relevant cluster contributions \( \Delta S^*_\Gamma \) to the Ising entropy \( S \); minimization of \( S^*_\Gamma \) is computationally tractable because the relevant clusters turn out to be the smaller ones. The solution for the whole system is therefore built up recursively from the solutions found for the small relevant clusters.

More in details, for each subset of \( k \) neurons \( \Gamma(k) \) \( (k = 1, \ldots, N) \), two entropies are defined:

- the subset entropy \( S^*_\Gamma(p) \), which is the entropy of the subset \( \Gamma(k) \) given the data \( (p = \{ p_i, p_{ij} \}) \), that is the minimum, with respect to \( J \), of the regularized cross-entropy \( S^*_\Gamma(p) \) between the model and the data, restricted to the \( k \) variables in \( \Gamma(k) \) (when \( \Gamma = \Gamma(N) \), \( S^*_\Gamma(p) = S(p) \), entropy of the whole system);
- the cluster entropy \( \Delta S^*_\Gamma(p) \), which is the remaining contribution to the subset entropy, once the contributions coming from all clusters of \( \Gamma' \) contained in \( \Gamma \) have been subtracted: \( \Delta S^*_\Gamma(p) = S^*_\Gamma(p) - \sum_{\Gamma' \subset \Gamma} \Delta S^*_\Gamma' \).

Therefore, the total entropy \( S(p) \) can be expressed as a sum of the \( 2^N - 1 \) cluster entropies of all clusters with \( 1, 2, \ldots, N \) neurons, each of which depends only on the firing rates \( \{p_i\} \) and pairwise correlations \( \{p_{ij}\} \) of the neurons within the cluster:

\[
S(p) = \sum_i \Delta S_i(p_i) + \sum_{i < j} \Delta S_{ij}(p_i, p_j, p_{ij}) + \sum_{i < j < k} \Delta S_{ijk}(p_i, p_j, p_k, p_{ij}, p_{ik}, p_{jk}) + \ldots \quad (2.37)
\]

where, for 1-neuron clusters, \( \Delta S_i(p_i) = S_i(p_i) \).

Cluster entropies can be computed recursively: at the \( k \)th step cluster entropies \( \Delta S^*_\Gamma(p) \) of clusters \( \Gamma(k) \) can in principle be computed minimizing \( S^*_\Gamma(p) \) with respect to \( \{h_i, J_{ij}\} \) with \( i, j \in \Gamma \) (to get \( S^*_\Gamma(p) \)) and subtracting all cluster entropies calculated at the step \( k - 1 \). For small enough clusters \( (k \lesssim 20) \), \( S^*_\Gamma(p) \) can be obtained quite easily: the partition function can be calculated exactly in time \( \sim 2^k \), and \( S^*_\Gamma(p) \) can be minimized with numerical methods (gradient descent or Newton’s method, or a combination of both) in time growing polynomially with \( k \). Minimization of \( S^*_\Gamma(p) \) allows one to calculate not only \( \Delta S^*_\Gamma(p) \), but also the contributions \( \Delta J^*_\Gamma \) of each cluster \( \Gamma \) to the couplings to be
inferred. Indeed, a decomposition in cluster contributions similar to that performed for the entropy holds for the model parameters:

$$J = -\frac{\partial S}{\partial p} = \sum \Delta J_{\Gamma}$$

(2.38)

with $\Delta J_{\Gamma} = J_{\Gamma} - \sum_{\Gamma' \subset \Gamma} \Delta J_{\Gamma'}$ and $\Delta J_{\Gamma} = J_{\Gamma}$ for the 1-neuron clusters. Minimizing $S^*_{\Gamma}(p)$ gives $J_{\Gamma}$ from which $\Delta J_{\Gamma}$ can be obtained removing the contributions computed at the previous step.

In conclusion, if the cluster expansion can be truncated at $k^*$ (thus removing all clusters with size $k > k^*$) and $k^*$ is not too large (to be able to compute $Z$ exactly), the procedure just described enables to infer the parameters $J$ for the whole system.

Expansion 2.37 (and 2.38) can be really truncated, including only small clusters and without missing any relevant information from the data. This is the consequence of two facts (summarized here and illustrated better in [157, 178]): first, small cluster entropies have a universal distribution, not specific to the interaction network and reflecting the sampling noise of the empirical correlations; on the contrary, large cluster entropies are network specific and reflect the signal present in the data; secondly, large cluster entropies decay exponentially in the length of the shortest closed interaction path between the neurons in the cluster. Thus, all clusters with an entropy smaller than a certain threshold $\theta^*$ can be discarded in expansion 2.37 and these “noisy” clusters are those with large closed interaction paths: in practice, beyond a certain size $k^*$, all clusters will be discarded (together with some other clusters with smaller sizes).

The threshold $\theta^*$ on the minimum cluster entropy to be considered in the expansion is chosen such that the model inferred from all clusters with $\Delta S > \theta^*$ reproduces the data without overfitting. This is achieved in the following way: one starts by choosing a high $\theta^*$; cluster entropies for all 1-neuron clusters are computed and those which are larger than $\theta^*$ (relevant clusters) are retained (together with the $\Delta J$ inferred from them); clusters with 2 neurons are formed combining pairs of relevant 1-neuron clusters, and again retained (together with their $\Delta J$) if their $\Delta S > \theta^*$; at step $k+1$, new clusters with $k + 1$ neurons are formed combining all relevant clusters with $k$ neurons, differing by 1 neuron only, that have been selected at the previous step. When none of the new clusters has $\Delta S > \theta^*$, all contributions $\Delta J$ coming from the collected relevant clusters are summed up, and a Monte Carlo simulation is performed for this inferred Ising model to obtain the reconstruction errors on the firing rates $\{p_i\}$ and on the connected correlations $\{c_{ij} = p_{ij} - p_ip_j\}$. The Adaptive Cluster Expansion algorithm halts if the reconstruction errors are of the order of the fluctuations due to noise, that is when the model reconstructs the empirical $\{p_i\}$ and $\{c_{ij}\}$ within their sampling errors:

$$\epsilon_p = \sqrt{\frac{1}{N} \sum_i \frac{(p_i^{rec} - p_i)^2}{(\delta p_i)^2}} \simeq 1$$

(2.39a)

$$\epsilon_c = \sqrt{\frac{2}{N(N-1)} \sum_{i<j} \frac{(c_{ij}^{rec} - c_{ij})^2}{(\delta c_{ij})^2}} \simeq 1$$

(2.39b)

where $\delta p_i = \sqrt{\frac{p_i(1-p_i)}{B}}$ is the sampling error on $p_i$ and $\delta c_{ij}$ is simply calculated from $\delta p_i$, $\delta p_j$ and $\delta p_{ij} = \sqrt{\frac{p_{ij}(1-p_{ij})}{B}}$. If the reconstruction errors are larger, the threshold on cluster
2.2. MODEL-BASED METHODS

entropies $\theta^*$ is decreased, new clusters are added to the expansion and new parameters are inferred. The procedure is iterated until criterion 2.39 is reached. In many practical applications to neuronal data, all clusters with size larger than a few neurons ($k^* \sim 5$) turn out to be a consequence of sampling noise and are discarded.

ACE has multiple advantages compared to the other available algorithms for the inverse Ising problem. These advantages are a consequence of the careful selection of the relevant clusters described above. These advantages can be summarized as follows:

- elimination of many clusters from the expansion, and in particular of all clusters larger than a certain size, makes the problem computationally solvable in a much shorter time than that required by other algorithms like Boltzmann Machine learning;

- the selection procedure allows one to carefully fit the model on all and only the clusters that contain information on the true interactions: overfitting is therefore avoided in a much more controlled way than with a simple regularization of the cross-entropy (which is the standard approach for the other inference methods);

- the very accurate inference procedure makes the inferred model a generative model of the activity: in several applications to neuronal data, it has been shown (see [179] and next chapter) that not only the empirical firing rates and pairwise correlations are reconstructed, but higher order statistics is also well predicted by the Ising model inferred via ACE, for example triplet firing probabilities $\{p_{ijk}\}$ and the probability $p(k)$ that $k$ neurons are active in a time-bin, even for large $k$ ($k \sim 10$); with other inference methods (e.g. Pseudo-likelihood and Mean Field) not even the $\{p_i, p_{ij}\}$ are usually well reproduced;

- differently from message passing algorithms, ACE converges, by construction, even in the presence of strongly interacting groups of neurons (loops), provided that the size of these clusters does not exceed the limit beyond which exact calculation of $Z$ becomes impracticable;

- ACE builds up the solution $J$ such that the inverse Ising problem is well-conditioned, that is a small change $\delta p_{ij}$ in some empirical frequencies (e.g. between neurons 1 and 2) produces localized changes in the interaction parameters (in other words, only couplings in the “neighborhood” of neurons 1-2 will change). This property is very important for the result of the inference to be meaningful: indeed, in typical experiments, only a small region of the brain is recorded and a model can only be inferred from very partial information; for the result to be significative, the inferred interaction network should not depend dramatically on the part of the brain that is not recorded, which means that the response $\frac{\partial J}{\partial p}$ should be localized. Clusters that are selected in ACE are precisely the neighborhoods in the interaction graph: indeed, the contribution $\Delta J_\Gamma$ to the reconstruction of the network coming from the cluster entropy $\Delta S_\Gamma$ represents the variation (response) $\delta J$ induced by correlations between neurons of different sub-clusters $\Gamma'$ contained in $\Gamma$; if sub-clasters $\Gamma'$ are not independent but interact significantly with one another, the response or correction $\delta J$ to the inferred couplings will be different from zero. Since many clusters only represent noise and are not selected, many entries in the matrix $\frac{\partial J}{\partial p}$ turn out to be zero, and the neighborhoods of the reconstructed interaction graph are spatially localized.
2.2.3 Generalized linear models

Generalized linear models (GLM) express the spiking probability of a neuron $i$ at time $t$ in terms of a conditional intensity function $\lambda_{it}$, which represents the expected number of spikes of neuron $i$ in time-bin $t$ and is related to the model parameters and to some relevant covariates, such as the neuron’s own past activity, the spiking history of the other neurons and extrinsic covariates, like stimuli or behavior. This conditional intensity is related to a linear combination of the model parameters through an invertible function, which can be non-linear, called link function (hence the name of GLM). I will show a few representative examples of this very large class of models, which can differ by the choice of the covariates, of the specific functions used to model these covariates, of the link function, or of the stochastic process used to model spiking activity. Importantly, all GLMs can catch the simultaneous effects of multiple covariates and allow for the assessment of their relative importance, as discussed below; moreover, GLMs can capture the dependence on all past history and not only on the spikes occurring in a short time window before time $t$.

Poisson- and Bernoulli-GLMs

The classical GLM is the Poisson-GLM with exponential link function (also called natural link function):

\[
P(n_{it}) = \frac{e^{-\lambda_{it}} \lambda_{it}^{n_{it}}}{n_{it}!}
\]

where $n_{it}$ is the number of spikes emitted by neuron $i$ in time-bin $t$, the total input to neuron $i$ at time-bin $t$, $h_{tot}^{it}$, is a linear function of model parameters $\{\theta_q\}$, and $g_q$ is a generic function of covariate $v_q$. The exponential non-linearity in (2.40b) implements a ‘soft threshold’, converting a neuron total input at time-bin $t$ into spike probability within the same time-bin.

A typical expression for $h_{tot}^{it}$ is:

\[
h_{tot}^{it} = h_i + \sum_{j \neq i} J_{ij} \sum_{\tau} e^{-\tau/c_1} \sigma_{j,t-\tau} + J_{ii} \sum_{\tau} e^{-\tau/c_2} \sigma_{i,t-\tau}
\]

with parameters $\{h_i, J_{ij}, J_{ii}\}$: $h_i$ represents an external input to neuron $i$, setting its baseline firing rate; $J_{ij}$ represents the directional interaction from $j$ to $i$ (or in other words how much neuron $i$ firing is affected by neuron $j$ firing); $J_{ii}$ captures the dependence of the spiking activity of neuron $i$ on the past activity of the neuron itself, due to mechanisms like refractoriness, bursting and adaptation. The dependence on past activity of both the neuron itself and the other neurons is assumed to decay exponentially over time, and $\lambda_{it}$ is a history-dependent generalization of the inhomogeneous Poisson rate function.

It is possible to add a dependence on other covariates in eq. (2.41). For example, to model activity of neurons in motor cortex, it can be useful to add a velocity vector component, like $|V_t| (\alpha_1 \cos(\phi_t) + \alpha_2 \sin(\phi_t))$, as shown in [180], with parameters $\alpha_1$ and $\alpha_2$.

The generic Poisson-GLM likelihood of the whole set of neuron spike trains is given by:
\[ P(\{n_{it}\}|\{\theta_q\}) = \prod_{it} \frac{e^{-\lambda_{it}} \lambda_{it}^{n_{it}}}{n_{it}!} \]

When the time-bin is small enough, a neuron either does not spike or it spikes just once in a time-bin; therefore \( n_{it} \) can be replaced with a binary variable \( \sigma_{it} \) and eq. 2.40a becomes:

\[ P(\sigma_{it}) = e^{-\lambda_{it}} \lambda_{it}^{\sigma_{it}} \]  

(2.42)

However, this approximation is not reasonable when \( \lambda_{it} \) is not \( \ll 1 \) \( \forall i, t \) (indeed, according to eq. 2.42, \( <\sigma_{it}>= e^{-\lambda_{it}} \lambda_{it} \), which is a decreasing function of \( \lambda_{it} \) for \( \lambda_{it} > 1 \) and is \( \sim \lambda_{it} \) only when \( \lambda_{it} << 1 \)).

It is possible to correctly model a neuron spiking in each time-bin as a binary (Bernoulli) process within the GLM framework, independently on how large is the total input a neuron receives. Such a model is the \emph{Bernoulli-GLM with logistic link function}, defined by the following equations:

\[ P(\sigma_{it}) = \lambda_{it}^{\sigma_{it}}(1 - \lambda_{it})^{1-\sigma_{it}} \]  

(2.43a)

\[ \lambda_{it}' = \frac{e^{h_{tot}^{it}}}{1 + e^{h_{tot}^{it}}} = \frac{\lambda_{it}}{1 + \lambda_{it}} \]  

(2.43b)

\[ h_{tot}^{it} = \sum_q \theta_q g_q(v_q(t)) \]  

(2.43c)

where \( \lambda_{it}' \) is the spiking probability of neuron \( i \) in time-bin \( t \). The complete likelihood of the population of neurons is expressed by:

\[ P(\{\sigma_{it}\}|\{\theta_q\}) = \prod_{it} \lambda_{it}^{\sigma_{it}}(1 - \lambda_{it})^{1-\sigma_{it}} \]

When \( \lambda_{it} << 1 \), the Poisson and Bernoulli-GLMs are equivalent [180]. Indeed:

\[ \lambda_{it}^{\sigma_{it}}(1 - \lambda_{it})^{1-\sigma_{it}} = \left( \frac{\lambda_{it}'}{1 - \lambda_{it}} \right)^{\sigma_{it}} (1 - \lambda_{it}') \sim \lambda_{it}^{\sigma_{it}} e^{-\lambda_{it}'} \]

and

\[ \lambda_{it}' = \frac{\lambda_{it}}{1 + \lambda_{it}} \sim \lambda_{it} \]

The likelihood of the Poisson and Bernoulli-GLMs is a concave function of the parameters \( \{\theta_q\} \) [181]; therefore inference can be easily carried out by maximizing the log-likelihood with the gradient ascent algorithm or more efficiently with Newton’s method.

The Poisson-GLMs are part of the class of Linear-Nonlinear-Poisson (LNP) models. These last are defined on three levels. The first level consists of a linear filter, which represents the neuron spatio-temporal integration of external stimuli and, in principle, internal inputs, e.g. those coming from the other neurons, like in GLMs; however LNP models are generally used to characterize the neuronal responses in early stages of sensory processing [182] and the linear filter often represents a neuron receptive field. The second level consists of a non-linear function, which takes as an input the output of the linear filter \((h_{tot}^{it})\) and returns the neuron’s instantaneous spike rate (its conditional intensity \( \lambda_{it} \)); finally, the instantaneous spike rate is converted into a series of spike times, under the hypothesis that spikes are generated according to an inhomogenous Poisson process. Some differences
between GLMs and LNP models are that in LNP models the non-linearity is allowed to be a non invertible function, while in GLMs it is constrained to be invertible. Moreover, in GLMs the spike generator process can be non Poissonian, like in Bernoulli-GLMs.

**Some goodness-of-fit techniques for the evaluation of different covariates**

The Poisson or the Bernoulli-GLMs are often used to assess the distinct effects of the covariates of interest. Even the simple study of the parameters of GLMs, inferred from the data, can be useful to improve knowledge about neuron properties, compared to what is known from descriptive statistics: in the GLM with neuron interactions derived in [183], for instance, the inferred stimulus filters indicate that the receptive fields of parasol ganglion cells have smaller surrounds than those obtained by the spike-triggered average of the stimulus, meaning that a portion of the classical surround can be explained by interactions between nearby cells.

Different goodness-of-fit analysis, applied to GLMs, can provide more refined and complementary information. Some typical methods are the Akaike’s standard information criterion, cross-validation analysis, point process residual analysis, and Bayesian neural spike train decoding.

**Akaike’s standard information criterion** (AIC) [184] provides a rank between different candidate models, based on their relative goodness-of-fit; if each one of these models includes a different set of covariates, this measure can be interpreted as a measure of the relative importance of these covariates. For a model with \( q \) parameters:

\[
AIC(q) = -2 \log P \left( \{\sigma_i\} | \hat{\theta}_q \right) + 2q
\]

where \( \hat{\theta}_q \) is the maximum-likelihood estimate of the parameters. The best model (the one with the smallest estimated relative Kullback-Leibler distance from the true distribution) is the model with the smallest AIC. This measure represents the trade-off between how well the model fits the data and the number of parameters required to achieve that fit. For example, in [180], the authors show that, according to AIC, the spiking activity of neurons in the arm region of the primary motor cortex of monkeys is represented better by a GLM model containing an autoregressive spiking history component plus a velocity component than a GLM model which only contains the velocity component, and adding the ensemble activity covariate further improves the goodness-of-fit. This is not trivial and captures a true dependence on spike-history and ensemble covariates, since AIC penalizes models with many parameters.

Another simple method which is often used to assess a model goodness-of-fit, taking into account the issue of overfitting, is *cross-validation*: this technique consists in fitting the model on one part of the data (called training set) and in quantifying the predictive power of the model by measuring the error with which the other part of the data (the test set) is reproduced by model simulations. For example in [18], the authors show by cross-validation that a GLM containing an ensemble activity covariate reproduces better the spike times of hippocampal pyramidal cells, than a GLM model containing only the spatial location and the theta phase covariates: this result means that some residual synchronization is left between these neurons, once the timing effects induced by location in space and theta phase have been subtracted; this kind of analysis also reveals the time-scale of this synchronization (as already mentioned in chapter 1).

The *point process residual analysis* is a typical method to assess the contribution of an extrinsic covariate to a single neuron’s spiking activity and the precise form in which a
neuron activity depends on this covariate. The point process residual for neuron $i$ at time $t_k$ is the difference between the true and the predicted number of spikes of neuron $i$ in a time window of size $T$ ending at time-bin $t_k$:

$$R_i(t_k) = \sum_{t-k-T}^{k} n_{it} - \sum_{t-k-T}^{k} \lambda_{it}$$

The presence of a positive or negative correlation between $R_i$ and the extrinsic covariate indicates the presence of some structure in the spiking activity of neuron $i$ which is induced by the covariate and is not captured by the model. In [180], the point process residual analysis shows that introducing the velocity covariate in a Bernoulli-GLM model in the form $|V_t| (\alpha_1 \cos(\phi_t) + \alpha_2 \sin(\phi_t))$ accounts for a significative amount (compared to other models not containing a velocity component) of statistical structure related to hand velocity in the activity of single neurons in primary motor cortex; however, it also reveals that this velocity model does not completely capture the single-neuron spiking structure related to hand velocity since, for some neuron, $R_i$ is correlated to some aspects of the velocity.

Differently from the point process residual analysis, neural spike train decoding can give an estimate of the contribution of an extrinsic covariate not only to a single neuron’s spiking activity but also to the network activity at the ensemble level. A Bayesian decoding algorithm for Poisson-GLMs with $\lambda \ll 1$, based on state estimation with point process observations, is derived in [185], [186] and is briefly illustrated below. The decoded extrinsic covariate is obtained as the covariate mean relative to an approximation of the posterior probability distribution. The algorithm computes recursively the posterior mean and the posterior covariance at successive time-bins, exploiting the point process observations (the neuron spike times until the time-bin of the current estimate) and previously estimated quantities. The idea is to first model the temporal evolution of a multidimensional extrinsic covariate $x_k$ as a stochastic process with a Gaussian component (which is in many cases a reasonable assumption):

$$x_k = \mu + F_k x_{k-1} + \eta_k$$

(2.44)

where $\eta_k$ is a Gaussian noise with zero mean and covariance matrix $Q_k$; $\mu$, $\lambda$ and $F_k$ are fitted to the true covariate. Then the mean $x_{k|k-1}$ and covariance $W_{k|k-1}$ for the one-step prediction density (i.e. the a priori mean and covariance estimates) are simply obtained from 2.44:

$$x_{k|k-1} = \mu + F_k x_{k-1|k-1}$$

(2.45a)

$$W_{k|k-1} = F_k^T W_{k-1|k-1} F_k + Q_k$$

(2.45b)

where $x_{k-1|k-1}$ and $W_{k-1|k-1}$ are the posterior mean and covariance at time $k - 1$. These one-step (or a priori) estimates are used for the computation of the posterior covariance and mean at the successive time $k$:

$$W_{k|k}^{-1} = W_{k|k-1}^{-1} + \sum_i \left( \frac{\partial \log \lambda_{ik}}{\partial x_k} \lambda_{ik} \left( \frac{\partial \log \lambda_{ik}}{\partial x_k} \right)' - (\sigma_{ik} - \lambda_{ik}) \frac{\partial^2 \log \lambda_{ik}}{\partial x_k \partial x_{i'}} x_{k|k-1} \right)$$

(2.46a)

$$x_{k|k} = x_{k|k-1} + W_{k|k}^{-1} \sum_i \left( \frac{\partial \log \lambda_{ik}}{\partial x_k} ' (\sigma_{ik} - \lambda_{ik}) \right) x_{k|k-1}$$

(2.46b)
Eqs. 2.46 are derived from the posterior probability, obtained with Bayes’ rule, after a Gaussian approximation

\[
P(x_k|\{\sigma_{i1}\}, \ldots, \{\sigma_{ik}\}, \{\hat{\theta}\}) \propto \prod_i e^{-\frac{1}{2}(x_k-x_{k|k-1})'W_{k|k-1}^{-1}(x_k-x_{k|k-1})}
\]

and performing a quadratic expansion of the log of both sides of eq. 2.47. Starting from an initial condition \(\{x_{1|1}, W_{1|1}\}\), the algorithm can compute the a priori estimates for the mean and the covariance, and then the posterior covariance and the posterior mean. These latter allow for the computation of confidence regions and coverage probability, that is the probability that the true covariate is within the confidence regions during a certain time interval, which gives an estimate of how well the covariate is decoded by the model.

This decoding algorithm has been used in [180] to show the success of the previously mentioned GLM model with velocity covariate, fitted to motor cortex neuronal activity, in decoding the direction of hand movement and the slower velocity fluctuations. A similar Bayesian decoding algorithm has been applied in [183] to compare decoding performance of a visual stimulus with a Poisson-GLM model containing pairwise neuron interactions and a similar model without interactions, both fitted to the activity of parasol retinal ganglion cells of monkeys: comparison shows that the stimulus is reconstructed better under the coupled model than under the uncoupled model, indicating that the interaction structure contributes to the representation of sensory information even at very early stages of visual processing.

Some studies suggest that Bayesian decoding is not just a mathematical formalism, but may be implemented by neuron populations [187]. Bayesian algorithms known as Kalman filters, similar to the algorithm described above, are thought to be biologically plausible mechanisms by which the brain would decode extrinsic variables in real-time: Kalman filters combine in an optimal way (i.e. minimizing the mean square error of the estimate) a priori information obtained from an internal model of the process to be decoded with the sensory feedback, representing a noisy estimate of the variable. In [188], the authors propose a neural implementation of Kalman filters in recurrent cortical circuits as a plausible mechanism for optimal sensorimotor integration, for example to track the position of one’s own arm during motion in darkness, from the proprioceptive sensory feedback and from the prediction generated by an internal model of arm dynamics, predicting the current arm position given the preceding position and motor command; or to locate a static visual stimulus while the eyes are moving, from the response of neurons in the retina and from the internal knowledge of average eye velocity.

The Kinetic Ising model

The Kinetic Ising model can be seen as a particular type of Bernoulli-GLM, where the dependence on the spike history of the other neurons is restricted to the previous time-bin, \(\{J_{ii} = 0\}\) and the local fields \(h_i\) are time-dependent, that is:

\[
h_{it}^{tot} = h_{i,t-1} + \sum_{j \neq i} J_{ij}\sigma_{j,t-1}
\]
A Kinetic Ising model with 0,1 variables is therefore defined by the following equation:

\[ p(\sigma_{it}) = \frac{\exp(\sigma_{it}(h_{i,t-1} + \sum_{j \neq i} J_{ij}\sigma_{j,t-1}))}{1 + \exp(h_{i,t-1} + \sum_{j \neq i} J_{ij}\sigma_{j,t-1})} \]

This model, and some variations of it, have been studied by Roudi and collaborators [189–193] in the form with −1,1 variables (where \( \sigma_{it} = -1 \) indicates that neuron \( i \) is silent in time-bin \( t \), while \( \sigma_{it} = 1 \) that neuron \( i \) is active in time-bin \( t \)). In this form, the model equation reads:

\[ p(\sigma_{it}) = \frac{\exp(\sigma_{it}(h_{i,t-1} + \sum_{j \neq i} J_{ij}\sigma_{j,t-1}))}{2 \cosh(h_{i,t-1} + \sum_{j \neq i} J_{ij}\sigma_{j,t-1})} \quad (2.49) \]

and the likelihood is given by:

\[ p(\{\sigma_{it}\}|\{h_{it}, J_{ij}\}) = \prod_{it} p(\sigma_{it}) \quad (2.50) \]

Parameters \( \{h_{it}, J_{ij}\} \) can be inferred when the recording consists of several repeats (necessary to estimate the time-dependent local fields \( h_{it} \)). One possibility is to infer these parameters by an iterative algorithm, analogous to Boltzmann learning for the equilibrium Ising model. Starting from some initial values of the couplings and fields, the algorithm, at each step, updates these values with the rule:

\[ h_{it} \rightarrow h_{it} + \alpha_h \left( \langle \sigma_{i,t+1} \rangle_r - \langle \tanh(h_{it}^{tot}) \rangle_r \right) \quad (2.51a) \]

\[ J_{ij} \rightarrow J_{ij} + \alpha_J \left( \langle \sigma_{i,t+1}\sigma_{j,t} \rangle_{r,t} - \langle \tanh(h_{it}^{tot})\sigma_{j,t} \rangle_{r,t} \right) \quad (2.51b) \]

where \( \alpha_h \) and \( \alpha_J \) are learning rates, \( \langle \cdot \rangle_r \) denotes averages over repeats and \( \langle \cdot \rangle_{r,t} \) averages over repeats and time-bins within each repeat; \( \delta h_{it} \) and \( \delta J_{ij} \) are proportional to \( \frac{\partial \mathcal{L}}{\partial h_{it}} \) and \( \frac{\partial \mathcal{L}}{\partial J_{ij}} \), respectively (\( \mathcal{L} = \log p(\{\sigma_{it}\}|\{h_{it}, J_{ij}\}) \)) being the log-likelihood). The algorithm converges to the maximum of \( \mathcal{L} \), without the need to carry out simulations: indeed, inference with this model does not have the problem of the calculation of the partition function because the log-likelihood factorizes over the neurons, and the computation of the gradient of \( \mathcal{L} \) reduces to the computation of empirical averages. However, the algorithm can still be slow because new averages have to be computed after each parameter update. The need for faster solutions has motivated the development of a mean-field theory for this out-of-equilibrium model. In [189], the authors derive dynamical mean-field and TAP equations: writing \( \sigma_{it} = m_{it} + \delta \sigma_{it} \), expanding the tanh to first order in \( \delta \sigma_{it} \) in eq. 2.51b and setting \( \delta J_{ij} = 0 \), finally exploiting the naïve MF relation \( m_{i,t+1} = \tanh(h_{it} + \sum J_{ij}m_{j,t}) \) to eliminate the zero order term, they obtain:

\[ \langle D_{ij,t} \rangle_t = \sum_k J_{ik}^{nMF} \langle (1 - m_{i,t+1}^2) C_{kj,t} \rangle_t \quad (2.52) \]

with \( D_{ij,t} = \langle \delta \sigma_{i,t+1}\delta \sigma_{j,t} \rangle_r \) and \( C_{ij,t} = \langle \delta \sigma_{it}\delta \sigma_{jt} \rangle_r \). Inverting this formula, they finally obtain the couplings in terms of the time-dependent correlations:

\[ J_{ij}^{nMF} = \sum_k \langle D_{ik,t} \rangle_t \left[ (B_{kj}^{(i)})^{-1} \right]_{kj} \quad (2.53) \]

with \( B_{kj}^{(i)} = \langle (1 - m_{i,t+1}^2) C_{kj,t} \rangle_t \). TAP expressions are obtained similarly, but taking into account the Onsager reaction term.
The dynamical TAP approximation performs better than the equilibrium TAP approximation in reconstructing the connections of simulated dilute networks with asymmetric connections: for instance, in [189], the authors show that couplings corresponding to inhibitory, excitatory and zero synapses, inferred with dynamical TAP, have well separated distributions, which is not true for couplings inferred with equilibrium TAP. Moreover, in cases where correlations are stimulus-induced (for instance in retinal cells subject to a time-varying visual stimulus), a Kinetic Ising model with zero couplings can give a better representation of the data (e.g. in terms of AIC test) than an Ising model with non-zero couplings [191] (when the dataset includes a sufficiently large number of repetitions), since in the Ising model local fields are constant over time and the stimulus-induced correlations are reproduced through the couplings. Therefore, the Kinetic Ising model with time-dependent local fields proves useful to analyse data characterized by the presence of a time-varying external input, repeated with the same temporal profile over many experimental trials. Alternatively, it is sometimes possible to model the time-dependent input of a neuron as a function with constant parameters: in this case less data are usually required to carry out inference. For example, place fields of place cells or grid cells can be seen as time-dependent external inputs, which can be modeled by linear combinations of Gaussian basis functions, centered on a lattice covering the recording environment. This is the approach proposed in [193] to infer a Kinetic Ising model from a population of grid cells and to get an estimate of the functional connectivity not biased by overlapping place fields (as well as other factors like theta rhythm and head directional inputs): the authors prove in this way that connectivity between grid cells decays with increasing phase difference, becoming negative for large differences; moreover they show that connections are on average stronger for cells belonging to the same module (that is having similar spacing and orientation [194]).

2.2.4 State-space models

A state-space model is defined by two equations: a state equation and an observation equation. The first one describes an unobservable state process on which neuronal activity depends; the second one sets the relation between the observed spiking activity and the unobservable state process. Two representative examples of this class are state-space GLMs and integrate and fire models (I&F).

State-space GLMs

State-space GLMs (SS-GLMs) are a special class of state-space models in which the observation equation is a GLM. An interesting example is shown in [195], where a SS-GLM is used to account for between-trial activity modulations of single hippocampal neurons in a monkey performing a location-scene association task. The observation equation of this SS-GLM is a Poisson-GLM with conditional intensity function:

\[ \lambda_t = e^{\sum \theta_q,kg_q(t)} e^{\sum \gamma_{\tau_\text{tot}} \tau} \]

where \( \{g_q(t)\} \) are functions that model the within-trial stimulus or task-specific effects on spiking activity, \( \theta_k = \{\theta_{1,k}, \ldots, \theta_{q,k}, \ldots, \theta_{\text{tot},k}\} \) are parameter vectors which depend on each trial \( k \), and \( \gamma = \{\gamma_1, \ldots, \gamma_{\tau}, \ldots, \gamma_{\text{tot}}\} \) is a parameter vector that accounts for the dependence on the neuron past spiking history. The state equation expresses the evolution of parameter vectors \( \theta_k \) from one trial to the next one:

\[ \theta_k = \theta_{k-1} + \epsilon_k \]
with $\epsilon_k$ Gaussian random vector with zero mean and unknown covariance $\Sigma$. Modeling the evolution of $\theta_k$ as a random walk represents a reasonable continuity constraint on $\theta_k$ and limits the number of free parameters to infer to the set $\psi = \{\theta_0, \Sigma, \gamma\}$.

The authors also show, by AIC analysis, that this model explains the data better than simple statistics like a PSTH, than a state-space PSTH (defined by eq. 2.54 without the spike-history dependent part and by eq. 2.55), and it also outperforms the corresponding GLM with constant $\theta$ across trials (which models the stimulus and spike-history effects but not inter-trial dynamics). The SS-GLM provides a rich structure to quantify in a well-defined framework both spike history dependence and behavioural and learning effects on the neuronal activity. As pointed out in [195], not only the SS-PSTH and the classical GLM are special cases of the SS-GLM, but also the simple PSTH. A PSTH is indeed obtained by replacing $\theta_k = \theta$ for all trials, $\gamma = 0$ and unit pulse functions of width $B/q_{tot}$, centered at times $(q-1)B/q_{tot} + \frac{1+B/q_{tot}}{2}$, to $g_q(t)$, with $B$ number of time-bins in the recording.

Inference of a SS-GLM, as well as of other models that depend on an unobserved latent process, is typically done with the Expectation-Maximization algorithm (EM). Each iteration of the EM algorithm consists of two steps: an expectation step (E-step) and a maximization step (M-step). In the E-step, at iteration $i+1$, the expected value of the complete log-likelihood function is computed with respect to the conditional probability distribution of $\theta$, given the observed spike times and the set of parameters estimated at the previous step $\psi^i$. The complete log-likelihood is defined as $\log P(\sigma \mid \theta, \psi) = \log P(\sigma \mid \theta) + \log P(\theta \mid \psi)$, where $\log P(\sigma \mid \theta)$ is the model observation equation and $\log P(\theta \mid \psi)$ is derived from the state equation (in the example above, it is a Gaussian centered in $\theta_{k-1}$ with covariance $\Sigma$). In the M-step, the expected value of the complete log-likelihood is maximized (e.g. with Newton’s method) with respect to the parameters $\psi$, obtaining a new estimate $\psi^{i+1}$, which will be used in the E-step of the subsequent iteration.

### Integrate and Fire models

In Integrate and Fire models (I&F) the state equations give the evolution of a neuron membrane potential $V_i(t)$ and the observation equation expresses the relation between the neuron firing times $t_{ik}$ and the membrane potential. The equations of the simplest I&F model (the leaky I&F) are:

\begin{align}
C \frac{dV_i(t)}{dt} &= - \frac{C}{\tau_m} V_i(t) + I_i(t) \\
V_i(t) &= V_{th} \iff t = t_{ik} \\
V_i(t_{ik}^+) &= V_r
\end{align}

where $C$ is the membrane capacitance, $\tau_m = RC$ is the membrane leaking time, $V_{th}$ is the spike threshold and $V_r$ is the resting potential. Eq. 2.56b tells that the spike times coincide with the times at which the potential reaches the threshold value $V_{th}$; eq. 2.56c tells that, soon after a spike is emitted, the membrane potential is reset to its resting value $V_r$; the dynamics then restarts following eq. 2.56a with the new initial condition $V_r$ (hence, eqs. 2.56a and 2.56c are the state equations, and eq. 2.56b is the implicit observation equation). The model has a memory that extends back to the time of the preceding spike for each neuron, that is the evolution of $V_i(t)$ in a given ISI is independent from the evolution in the other ISIs. The equation for the dynamics of the membrane
potential 2.56a is derived representing a neuron as an electrical circuit, composed by a membrane capacitance $C$ in parallel with a membrane resistance $R$, driven by a current $I_i(t)$: in such a circuit, $I_i(t)$ is the sum of the resistive current $I_{i}^{s}(t) = V_{i}(t)/R$ and of the capacitive current $I_{i}^{c}(t) = C \frac{dV_{i}(t)}{dt}$. The total current $I_i(t)$ is usually modeled as the sum of a constant term $I_{i}^{c}(t)$ (representing an external input to the neuron) and a synaptic current $I_{i}^{syn}(t)$ (input coming from the other neurons):

$$ I_i(t) = I_{i}^{c}(t) + I_{i}^{syn}(t) \tag{2.57a} $$

$$ I_{i}^{syn}(t) = \sum_{j} J_{ij} \sum_{k} f(t - t_{jk}) \tag{2.57b} $$

with $J_{ij}$ strength of the interaction from neuron $j$ to neuron $i$ and $f(t - t_{jk})$ postsynaptic current induced at time $t$ by a spike of neuron $j$ at time $t_{jk}$. The simplest choice for $f(s)$ is the Dirac function $f(s) = \delta(s)$ (instantaneous integration of synaptic inputs); a more realistic choice is $f(s) = H(s)e^{-s/\tau_s}$ (exponentially decaying integration of synaptic inputs), where $H(s)$ is the Heaviside function and $\tau_s$ the synaptic time constant.

The model just described is deterministic. Its stochastic counterpart includes an additional fluctuating current, modeled as a Gaussian variable $\eta_i(t)$ with $\langle \eta_i(t) \rangle = 0$, $\langle \eta_i(t)\eta_j(t') \rangle = \sigma^2 \delta_{ij} \delta(t - t')$. The stochastic I&F model is more popular than the deterministic one because it takes into account the noise always present in neuronal recordings and can be studied with likelihood-based methods. In a stochastic I&F model, the observation equation is the likelihood of all spike times given parameters $\{I_{i}^{c}, J_{ij}\}$, which can be expressed, as pointed out in [196], as the product over all neurons, and all spike times of each neuron, of the probabilities that $V_i$ crosses $V_{th}$ for the first time at $t_{i,k+1}$, starting from $V_{i}$ at $t_{ik}$ and conditioned on the spike times of the other neurons $\{t_{jl}\}$ in the interval $[t_{ik}, t_{i,k+1}]$, that is the product of the first-passage time (FPT) probabilities:

$$ p(\{t_{ik}\}|\{I_{i}^{c}, J_{ij}\}) = \prod_{ik} p_{FPT}(t_{i,k+1}|t_{ik}, \{t_{jl}\}, \{J_{ij}\}, I_{i}^{c}) \tag{2.58} $$

Inference of the parameters of I&F models is difficult and most studies focus on the inverse problem for a single I&F neuron.

For a single stochastic I&F neuron, subject to an external stimulus $s(t)$, the total current is typically given by:

$$ I(t) = I_{stim}(t) + I_{hist}(t) + \eta(t) \tag{2.59a} $$

$$ I_{stim}(t) = k \cdot s(t) \tag{2.59b} $$

$$ I_{hist}(t) = h \cdot r(t) \tag{2.59c} $$

where $k$ is the stimulus filter, representing the spatio-temporal receptive field of the neuron, $h$ is a linear filter convolved with the recent spike train history $r$ to account for refractoriness, burstiness or adaptation effects, and $\eta(t)$ is a Gaussian white noise with standard deviation $\sigma$. Parameters of this model are $\{k, h, \sigma, \tau_m, V_{r}\}$. An elegant solution to this inference problem is shown in [197], where the authors derive a method to compute the log-likelihood by solving the Fokker Planck equation for the subthreshold voltage probability during each inter-spike-interval: indeed, eq. 2.56a has exactly the same form of the first cardinal equation for a Brownian particle (in this analogy, $V_{i}(t)$ corresponds to the particle velocity, $C$ to its mass, $\tau_m$ to the mass divided by the viscous friction coefficient, $\eta_i(t)$ to the force resultant from the impacts with the molecules of the fluid and $I_i(t) - \eta_i(t)$...
to an external drive). Hence \( V_i(t) \) satisfies a Fokker Planck equation in each ISI, which can be solved with standard techniques. The explicit expression for \( V_i(t) \) is then used to solve the following equation for \( p(t) \) (probability of a spike at time \( t \)):

\[
\int p(V(t)) dV = 1 - \int_{t_{k-1}}^t p(s) ds
\]

Finally the likelihood is simply the product of \( p(t) \) over all spike times of the neuron. This likelihood is log-concave and can be maximized with standard gradient ascent techniques.

This kind of single neuron I&F model can be used, for instance, to characterize the responses of sensory neurons to external stimuli more precisely than with simpler models which do not take into account spike history effects, as shown in [198].

One of the first attempts to infer a small network of I&F neurons is illustrated in [199] for the deterministic model described by eqs. 2.56 and 2.57 with \( f(s) \) defined by the exponential function. The authors infer the parameters of a 5-neuron network, by minimizing a cost-function defined as the sum of the squared differences between the experimental spike times and the spike times predicted by the model. However, this approach is not extensible to larger populations of neurons.

Two examples of inference of larger networks of stochastic I&F neurons are shown in [196, 200], where inference is performed in two steps: first, the most probable dynamical path of the membrane potential in each ISI is computed given the neuron spike times and the model parameters (maximum a posteriori path); then the log-likelihood of the spike times given the model parameters is approximated, in the limit of weak noise variance, by the contribution of the MAP path in [196], or with a Laplace approximation centered by the MAP path in [200]; in both cases, the approximated log-likelihood is concave and can be easily maximized. The method proposed in [196] is able to infer, in a reasonable time, currents and couplings of several tens of neurons (~40) recorded for about one hour.

### 2.3 Methods to identify cell assemblies and replay

In this paragraph, I will review the main available methods to detect and characterize cell assemblies and replay in neuronal recordings. I will first show some examples in which cell assemblies can be identified in a relatively simple way, focusing in particular on sensory and hippocampal cell assemblies; finally I will discuss the problem in more complex cases, like when cell assemblies do not have evident sensory correlates or when these correlates are unknown.

#### 2.3.1 Some introductory examples

When studying very small populations of neurons, it is possible to search for highly synchronous patterns by simply computing the empirical probabilities of all possible neuron configurations, binned in narrow time windows (e.g. 100 ms or smaller), and comparing their values with the cumulative Poisson probability that the same or a larger number of neurons are active in the time-bin, when the parameter \( \lambda \) is the average number of independently active neurons. A pattern is considered to be synchronous if the cumulative probability is very low (e.g. less than 5%). This approach, used in [82, 83] to detect synchronicity in populations of \( \leq 7 \) motor neurons, is clearly not applicable to larger populations.
In some cases (e.g., in the visual cortex [72–76]), cell assemblies are evident already at a mesoscopic level and their presence is signalled by LFP fast oscillations, which can only be observed when many neurons fire in synchrony; the high frequency of these oscillations (often in the gamma range) indicates that synchrony is generated with high temporal precision. Neurons composing these cell assemblies are usually identified at a microscopic scale through simple correlation analysis. However, correlation-based approaches are not always successful, especially when the signal is low and cell assemblies are very hidden in the data: the study illustrated in the next chapter, for instance, shows that correlation-based analysis in prefrontal cortex recordings may be limiting in terms of extraction of information about cell assemblies.

Bathellier, Ushakova, and Rumpel [79] have given an example of cell assembly characterization in the auditory cortex: in this study, neurons were recorded using two-photon calcium imaging, a technique which enables simultaneous recording of the activity of large neuronal populations in vivo, with single cell precision but low temporal resolution. Cell assembly activations, in this case, are local collective events, with a large number of locally clustered neurons firing synchronously. Hence, detecting these collective activations is not too problematic, while their characterization with respect to the external sensory stimulus is not trivial because responses are very noisy (repeated presentation of the same sound never gives rise to exactly the same response). Cell assemblies, in this case, are assimilated to response modes: a response mode is an activity pattern which reliably represents the population response to a sound or a group of sounds. Each mode is identified in this way: the overlap between population vectors of pairs of trials is computed (a population vector for a trial being the vector of all neuron firing rates in response to a sound); the average overlap over all pairs of trials in which the same sound is presented represents the reliability of the population response to a sound and is compared with the reliability value obtained for randomized data. Reliability, despite low because of stochasticity of responses, is significantly higher than for randomized data in some populations, meaning that those populations respond reliably to a sound. The same average overlap is computed between trials in which different sounds are presented and if this quantity (called similarity) is comparable with the reliability values for the single sounds it means that the population responds indistinguishably to the two sounds. In practice, once the similarity matrix is constructed (each entry representing the similarity for a pair of sounds), a hierarchical clustering algorithm applied to this matrix groups together sounds eliciting the most similar responses. The response pattern to each group of sounds is a response mode. Correlation between neurons firing in each response mode is much stronger than correlation between neurons firing in different modes: each mode can therefore be considered as a cell assembly from a functional point of view. It is relatively simple to study cell assemblies in sensory systems: in the example just described, for instance, it is clear that cell assemblies should encode sound features, and it is possible to characterize them by studying their response properties to sounds, easily reproducible in an experimental setting. However, this approach is not applicable to all regions of the brain. In the prefrontal cortex, for example, it is usually not known what is, if any, the external sensory input eliciting cell assembly activation; very likely, cell assemblies in prefrontal cortex, as well as in other high level centers of the brain, respond to internal cognitive states, which are very difficult to determine and control experimentally; therefore, cell assembly identification and characterization is very challenging.
Template matching for hippocampal cell assemblies

Activation and reactivation of cell assemblies of place cells are also identified quite easily since it is known that place cells respond to location in space and place cell assemblies represent place fields or, at a larger scale, sequences of adjacent place fields (trajectories), as discussed in chapter 1. A neuron place field is simply determined as the position of peak firing, calculated over the entire recording, and it often depends on the direction of motion.

It is possible to determine all reactivation events of significative trajectories (replay) with template matching techniques. Template matching consists in quantifying the correlation between the ordered activation sequences of place cells observed during salient moments, candidate for replay (e.g. SWR events), either during sleep or during wakefulness, and the sequences of place cells determined by the behaviourally relevant trajectories, observed during locomotion [21]. These behaviourally significant sequences are precisely the “templates”, to be matched with the “candidate” replay events. When place fields change with the direction of motion, different templates are defined for the same spatial trajectory traversed in opposite directions. To quantify the significance of the match between the template and the candidate event, the standard approach is to create many surrogate sequences, by shuffling the identity of the neurons in the candidate event, and to compute their correlations (overlap) with the template. If the true candidate sequence is more correlated with the template than are 95% of the surrogate sequences, the candidate is considered to be a significant replay event ([201]). This method can also be applied to detect preplay and reverse replay [27], and allows one to find both sequences that are positively correlated with the template (replay) and sequences that are negatively correlated (when the match is much smaller than chance level): this last case is less frequent, but it can occur for example when the animal knows that a certain path does not lead to reward and should not be followed, as shown in [26]. Since place cells fire multiple times within their place field and, at the overlap between adjacent place fields, spikes of different place cells are intermingled, candidate events are usually defined considering only the first spike fired by each place cell [35, 201] and converting the resulting place cell sequence into a sequence of place fields. A more refined estimate of the series of positions corresponding to the ensemble activity during candidate events is obtained with Bayesian decoding algorithms [25] of the kind illustrated in 2.2.3. A representative example is illustrated in [202], where place cell activity is modeled as an inhomogeneous Poisson process whose instantaneous rate is a function of position and theta phase, and the dynamical model for the position is a Gaussian random walk. The algorithm returns both position predictions and confidence regions for these predictions. An extension of this algorithm has been developed in [46].

PCA-based methods and community detection techniques

The logic of template matching can be extended to detect the activation of cell assemblies in other brain regions, even when neurons are not activated in a well-defined, decodable temporal order and the extrinsic covariates represented by the assembly are unknown. In such a case, templates can be approximated using the correlational structure of data through Principal Component Analysis (PCA).

PCA finds the orthogonal directions of largest variances of multidimensional data: the first principal component (PC1) identifies the direction of largest variance; the second principal component (PC2) identifies the direction, orthogonal to PC1, which explains the largest portion of the remaining variance, and so on. Principal components coincide
CHAPTER 2. METHODS FOR THE STUDY OF NEURONAL NETWORKS

with the eigenvectors of the Pearson correlation matrix of the data and the variance they account for with the corresponding eigenvalues. The Pearson correlation matrix $\Gamma$ is the covariance matrix of the $z$-scored spike counts of the $N$ neurons:

$$\Gamma = \frac{ZZ^T}{B}$$

$$Z_{it} = \frac{\sigma_{it} - \langle \sigma_{it} \rangle_t}{\sqrt{\langle \sigma_{it} \rangle_t(1 - \langle \sigma_{it} \rangle_t)}}$$

with $B$ number of time-bins in the recording. Since $\Gamma$ is real and symmetric, it can be decomposed into eigenvector contributions:

$$\Gamma = \sum_{\mu} \lambda_{\mu} v_{\mu} v_{\mu}^T$$

where $v_{\mu}$ is the $\mu$th eigenvector (or $\mu$th principal component) and $\lambda_{\mu}$ the $\mu$th eigenvalue.

Intuitively, when the data are correlated in some direction in the variable space, variance in that direction increases; therefore, the top principal components can be interpreted as correlation modes: variables that have large entries with the same sign in one top PC are correlated with each other and they are anti-correlated with variables having large entries with opposite sign in the same PC. In applications to neuronal data, neurons having large entries (with the same sign) in the top PCs can be considered as members of a cell assembly. Two questions naturally arise: a) what is the threshold on the eigenvalues which separates signal from non-signal components and b) what is the threshold on the variable entries $v_{\mu}(i)$ in a signal PC to be used to circumscribe neurons belonging to the cell assembly?

As for the first point, the number of signal PCs, indicating the number of correlation modes or cell assemblies in this framework, is usually determined from the Marčenko-Pastur distribution [203], which represents the distribution of eigenvalues of the covariance matrix of random, independent $N$-dimensional datasets. In the limit $B \to \infty$, $N \to \infty$, with $q = \frac{B}{N} \geq 1$, the distribution is:

$$\rho(\lambda) = \frac{q}{2\pi\sigma^2} \sqrt{(\lambda_{\text{max}} - \lambda)(\lambda - \lambda_{\text{min}})}$$

which has finite support in the interval $[\lambda_{\text{min}}, \lambda_{\text{max}}]$, with $\lambda_{\text{max}} = \sigma^2(1 \pm \sqrt{1/q})^2$, $\sigma^2$ variance of the elements of the random matrix (in this case the variance is set to 1 to compare the covariance matrix of the random variables with that of the $Z_{it}$, which have variance equal to 1). Therefore, in the approximation of infinite dimensional datasets, eigenvalues greater than $\lambda_{\text{max}}$ indicate the presence of correlations and are associated to signal components. A correction, based on the Tracy-Widom distribution [204], can be added to approximate the upper bound of the eigenvalues of finite matrices, as proposed in [2]:

$$\lambda_{\text{TW}}^{\text{max}} = \lambda_{\text{max}} + N^{-2/3}$$

where $N^{-2/3}$ is the standard deviation of the fluctuations around $\lambda_{\text{max}}$ for finite sampling. This threshold for separating the signal from the non-signal components is reasonable as a first approximation. However, while it is guaranteed that non-signal components have eigenvalues $< \lambda_{\text{TW}}^{\text{max}}$, it is not excluded that some signal components may have $\lambda < \lambda_{\text{max}}$ [205] (in particular the lowest-eigenvalue components of the spectrum have been shown to
be important for reconstruction of structural information in proteins \cite{206, 207}) and they are discarded with this approximation.

As regards the second question, the threshold on the entries of each signal PC for delimiting a cell assembly is often chosen arbitrarily. A possibility to reduce the arbitrariness of this choice is to represent each neuron as a point in the space of signal components, with coordinates \(\sqrt{\lambda_\mu} v_\mu(i), \mu = 1, \ldots, N_s\) (\(N_s\) number of signal components), and to apply clustering techniques (e.g. k-means) to identify groups of correlated neurons in this space: indeed, when data are clearly structured into distinct cell assemblies, groups of closely correlated neurons appear as points far from the origin and close to each other in this dimensionally-reduced space. Another possibility, suggested in \cite{208}, is to approximate the number of assembly neurons by the number of eigenvalues outside the support of the Marchenko-Pastur distribution (above or below the theoretical bounds): the rational is that when two neurons are correlated, the variance increases in one direction (the cell assembly axis) and decreases in the orthogonal direction, leading to one eigenvalue above \(\lambda_{\text{max}}\) and another one below \(\lambda_{\text{min}}\); Santos et al. \cite{208} generalize this reasoning for groups of \(k\) correlated neurons, arguing that one eigenvalue, corresponding to the assembly, will have \(\lambda > \lambda_{\text{max}}\), and other \(k-1\) will have \(\lambda < \lambda_{\text{min}}\).

PCA can also be used to track cell assembly activity in the data, through the quantity:

\[
A_t = \sum_{i=1}^{N} v(i)Z_{it} = v^T Z_t
\]

with \(v\) principal component which identifies the cell assembly and \(Z_t\) vector of all neuron \(z\)-scores in time bin \(t\). \(A_t\) coincides with the length of the projection of the configuration vector \(Z_t\) onto the principal component \(v\). \(A_t\) is large and positive when several neurons with large, positive entries in the PC are active in time-bin \(t\), and it is large and negative when several neurons with large, negative entries are active; so \(A_t\) allows detection of both the times of cell assembly activation and the times of activation of neurons negatively correlated with the cell assembly. \(A_t\) has been used in \cite{1} to localize, along the experimental maze, the activation of the first three PCs of the activity recorded in rat prefrontal cortex neurons during a task. In \cite{209}, the same quantity has allowed the authors to discover sensory-behavioural correlates of cell assemblies in the rat vibrissal sensory area of the thalamus, which represents the activity of whiskers: their analysis shows that activation of the PC1 cell assembly reflects global functions, such as the overall magnitude of sensory stimuli eliciting whiskers, regardless of whisker identity and direction, and is also associated with global \(8 – 12\) Hz oscillations during states of attention anticipating whisker movement \cite{210}; on the contrary, PC2 and PC3 cell assemblies represent specific directions and speeds of whisker movement. In \cite{1, 2}, a similar quantity \(R_t\) has been defined to track the reactivation (replay) during sleep following the task of the first PC extracted from the task recording \((v^{\text{Task}})\):

\[
R_t = \sum_{i<j} v^{\text{Task}}(i)v^{\text{Task}}(j)Z_{it}^{\text{Sleep}}Z_{jt}^{\text{Sleep}}
\]

\(R_t\) coincides with \(\frac{1}{2}(A^2_t - v_i^2Z_{ii}^2)\). Subtraction of the terms with \(i = j\) guarantees that activation of single members of the cell assembly does not contribute to the replay estimation. \(R_t\) is large when the length of the projection is large regardless of its sign: it is possible that a signal component has both a group of large positive entries and another group of large negative entries, corresponding to two mutually inhibiting cell assemblies;
in this case \( R_t \) does not distinguish between activation of one or the other. In [2], \( R_t \) in the prefrontal cortex peaks 40 ms after SWR occurrences during sleep following the task, in agreement with the hypothesis that hippocampal SWRs trigger prefrontal cortex replay. Moreover, Benchenane et al. [3] have shown that, during the task, \( R_t \) of the first principal component computed from periods of high theta coherence between prefrontal cortex and hippocampus peaks at the choice point of the maze.

Peyrache et al. [1] have also derived an expression to quantify the contribution of each principal component to the total amount of reactivation during sleep of task-related activity patterns. They define the correlation similarity between task and sleep (representing an estimate of the epoch-wide replay) through the quantity \( M \):

\[
M = \sum_{i<j} \Gamma_{ij}^{Sleep} \Gamma_{ij}^{Task}
\]

Substituting \( \Gamma_{ij}^{Sleep} = \frac{1}{B} \sum_t Z_{it}^{Sleep} Z_{jt}^{Sleep} \) and \( \Gamma_{ij}^{Task} = \sum_\mu \lambda_\mu v_\mu^{Task}(i) v_\mu^{Task}(j) \) into 2.63 they obtain:

\[
M = \frac{1}{B} \sum_t \sum_\mu \lambda_\mu \sum_{i<j} v_\mu^{Task}(i) v_\mu^{Task}(j) Z_{it}^{Sleep} Z_{jt}^{Sleep} =
\]

\[
= \frac{1}{B} \sum_t \sum_\mu \lambda_\mu R_t^{(\mu)}
\]

where \( R_t^{(\mu)} \) is an estimate of the reactivation at time \( t \) of sleep of the \( \mu \)th PC extracted from the task and its contribution to the total amount of replay of task-related coactivation patterns is weighted by its eigenvalue. In the next chapter, we will see that this estimation can be improved giving an interpretation of the replay that is natural in the context of statistical physics.

Despite PCA has been applied successfully in several cases, it has some drawbacks. For example, it cannot detect overlapping cell assemblies because principal components are orthogonal by definition. To overcome this problem, Santos et al. [208] have proposed the Assembly Vector (AV) estimation method. AV consists in searching for possibly overlapping cell assemblies in the subspace, called assembly space, spanned by the signal PCs. First, neuron vectors are defined by their coordinates in the assembly space:

\[
a_k = (v_1(k), \ldots, v_\mu(k), \ldots, v_L(k))
\]

where \( L \) is the number of signal PCs (defined as the number of eigenvalues \( \lambda > \lambda_{max} \)). Clustering is not applied directly to these neuron vectors, but to a binary adjacency matrix, calculated as follows. The significative neuron vectors are extrapolated as the \( q \) neuron vectors with largest norm, with \( q \) number of eigenvalues outside the support of the Marchenko-Pastur distribution. Then, each element of an “interaction matrix” is computed as

\[
M_{ij} = \frac{a_i \cdot a_j}{a_j \cdot a_j}
\]

\( M_{ij} \) is the projection of the significative neuron vector \( i \) onto the direction of the significative neuron vector \( j \). The interaction matrix is finally transformed into an adjacency matrix, i.e. it is digitalized, choosing the best threshold which separates low and high \( M_{ij} \) entries and assigning value 1 to high entries and value 0 to low entries. Finally, a
clustering algorithm is applied to find a number of (possibly overlapping) clusters in the adjacency matrix corresponding to the number of signal PCs. These clusters identify the cell assemblies. The Assembly Vector for an assembly $A$ is the (normalized) mean of the neuron vectors exclusive to $A$ and therefore it can be used (in place of $v$ in eq. 2.62) to track the activation of that cell assembly only (regardless of its overlap with other cell assemblies).

As shown in the next chapter, a potential problem in applying this method is that the interaction matrix is often not bimodal; in this case, the choice of the threshold to convert it into an adjacency matrix is totally arbitrary.

Another example of community detection technique for cell assembly identification, exploiting the Markov Stability method, has been recently presented in [211]. The method consists in a) defining the interaction graph from the neuron spike trains, b) finding a series of partitions on the graph with varying coarse-grain level, c) identifying the most robust partition. Interactions are derived only considering the effects of the last spike of a neuron immediately preceding each spike of another neuron and do not disentangle direct from indirect correlations. To identify a series of partitions at different coarse-grain levels on this interaction network, the authors define a continuous diffusion process on the graph; then, given a partition $H (H_{ij} = 1$ if node $i$ belongs to community $j$), the Markov Stability of the partition is defined as the probability that a random walker at stationarity starts in community $i$ and is found again in the same community after time $t_M$, minus the probability of such an event happening in an unstructured graph, summed over all communities. The Markov Stability is locally maximized in the space of all possible graph partitions for a given $t_M$, using the locally greedy Louvain algorithm, which gives an optimized partition for each initial condition and each Markov time $t_M$. By increasing $t_M$ the size of the communities tends to increase and their number to decrease. Once several possible partitions have been identified, the most robust ones are found based on two criteria: robustness with respect to variations of $t_M$ (plateau in the curve of the number of communities as a function of $t_M$) and robustness with respect to the choice of the initial condition in the Louvain algorithm (distance between two partitions obtained from two different initial conditions is estimated using an information-theoretic notion of distance called “variation of information”).

This method can in principle find groups of highly correlated neurons without assuming a priori the size and number of these groups; however the strong simplifications in the definition of the interaction graph can prevent its applicability to dataset in which cell assemblies are not elicited by sensory stimuli but arise from internal processes relying on a complex interaction network between the neurons, like the avalanche processes described in chapters 3 and 4.

In the next two chapters, an alternative method for detecting cell assemblies and replay will be presented, based on the inference and simulation of a microscopic generative model of the neuronal activity, the Ising model, relying on a network of functional connectivity between the neurons. In chapter 5, it will be shown that inference and sampling of a non-stationary model, a binary-GLM, can provide complementary information on the identified cell assemblies.
Chapter 3

A new statistical physics inspired method to study cell assemblies

As anticipated in the Foreword, it starts here, and continues in chapters 4, 5 and 6, the work that I developed under the supervision of and in collaboration with Simona Cocco and Rémi Monasson, in the framework of the European FP7-ICT project ENLIGHTENMENT. The aspects of this chapter relative to the measure of the effective coupling potentiation and to the formula for the quantification of the replay have been developed with the collaboration of Ulisse Ferrari. The data studied in this and in the following two chapters have been collected by Francesco Battaglia’s group, and have been previously analysed in [1–3].

We propose a statistical physics inspired method to identify cell assemblies from multi-electrode recordings of neuronal activity in vivo, different from all methods presented in par. 2.3, and we apply this method to study the prefrontal cortex activity of behaving rats, during performance of a decision-making task and during preceding and following sleep epochs.

We first infer models for the distribution of the neuronal activity in each experimental phase, based on an estimate of the effective coupling network between the neurons (Fig. 3.1, left). Among the large variety of graphical models illustrated in the previous chapter, we choose the Ising model to represent and study this data, since it is the simplest (maximum entropy) model able to reproduce low order statistics of the data and to reliably predict higher order statistics; more complicated, non-stationary models are generally more difficult to infer and they often do not share this generative property.

Comparison of the interaction networks inferred from the three epochs (Sleep Pre Task, Task and Sleep Post Task) reveals the presence of task-related potentiation or depression of the effective couplings.

The effective coupling-based model also allows us to characterize the firing probability of any neuron conditional to the activity of the other cells in the population, and to search for self-sustaining activity patterns, in which each neuron ‘reads’ the activities of the other recorded cells and, in turn, participates as an input to those cells in a coherent way. This concept of self-sustaining pattern is closely related to Hebb’s classical definition of a cell assembly [4] (see chapter 1). Self-sustaining patterns are encoded in the effective coupling network, arise spontaneously whenever favorable network or cellular excitability are met, and may be detected by a downstream ‘reader’ neuron [7]. A computationally efficient way to search through the combinatorial number of putative self-sustaining patterns consists in adding a driving input into the model distribution of the neuronal activity, favoring
CHAPTER 3. A NEW METHOD TO STUDY CELL ASSEMBLIES

Figure 3.1: Model for the neuronal activity: definition and simulation. Spiking times are binned into time-bins of width $\Delta t$; each neuron $i$ is assigned the variable $\sigma_i = 1$ or 0, if it is active in the time-bin or not (top, left). A model of the neuronal activity distribution ($P$ in eq. 3.1) is inferred to reproduce the 1- and 2-cell firing frequencies of this binned data (bottom, left); red and blue links correspond, respectively, to positive and negative effective couplings $J_{ij}$ in the inferred network. The model distribution is then simulated with the addition of an increasing drive, which favors configurations with more and more active neurons; values of the activity variables $\sigma_i$ in the most likely configuration given the drive are shown (right). As the drive increases a group of neurons (comprised in the dashed contour) may abruptly coactivate, defining a cell assembly.

high-activity configurations, that is mimicking transient increases in network excitability, as may be induced by e.g. slow oscillations of sleep. As the drive gets stronger self-sustaining activity patterns with more and more active neurons appear (Fig. 3.1, right) and groups of strongly coactivating neurons are revealed. These groups define our cell assemblies.

In surprising agreement with Hebbian plasticity, the cell assemblies supporting the effectively potentiated couplings are shown to strongly coactivate in the behavioral epoch but not in the preceding sleep epoch, and to be replayed in the subsequent sleep epoch. A wide-scale study of about 100 experimental sessions shows a variety of possible scenarios for the cell assemblies across the epochs, which allows us to formulate empirical rules for their formation and replay.
3.1 Effective network model for the neuronal activity

To model the distribution of activity of the \(N\) recorded neurons, the spiking times are binned within small time-bins of duration \(\Delta t = 10\) ms; the activity configurations \((\sigma_1, \sigma_2, \ldots, \sigma_N)\) are snapshots of the neuronal activity, where \(\sigma_i\) takes values one or zero depending on whether the \(i\)-th neuron is, respectively, active or inactive in the time-bin.

We model the probability distribution of activity configurations as

\[
P(\sigma_1, \sigma_2, \ldots, \sigma_N) = \frac{1}{Z} \exp \left( \sum_{i<j} J_{ij} \sigma_i \sigma_j + \sum_i h_i \sigma_i \right)
\]

where \(Z\) ensures normalization of the distribution. The \(\frac{1}{2}N(N+1)\) parameters \(h_i\) and \(J_{ij}\) are fitted to reproduce the \(N\) individual spiking frequencies \(f_i\) and the \(\frac{1}{2}N(N-1)\) pairwise spiking frequencies \(f_{ij}\) (within a time-bin \(\Delta t\)) estimated from the recording data as the average values over time-bins of, respectively, the variables \(\sigma_i(t)\) and \(\sigma_i(t)\sigma_j(t)\).

As discussed in the previous chapter, \(P\) in eq. 3.1 defines an Ising model, which is the least constrained (with maximum entropy), default probability distribution reproducing this low-order spiking statistics. Parameters \(J_{ij}\) define the effective pairwise couplings between the cells (Fig. 3.1): \(J_{ij}\) different from zero expresses the presence of a conditional dependence between neurons \(i\) and \(j\), not mediated by other neurons in the recorded population. From eq. 3.1, it is straightforward to derive the conditional average activity of neuron \(i\) given the other neuron activities \(\{\sigma_j\}\), with \(j \neq i\):

\[
\langle \sigma_i \rangle = \frac{P(\sigma_1, \ldots, \sigma_i=1, \ldots, \sigma_N)}{P(\sigma_1, \ldots, \sigma_i=0, \ldots, \sigma_N) + P(\sigma_1, \ldots, \sigma_i=1, \ldots, \sigma_N)} = \frac{e^{V_i}}{1 + e^{V_i}}, \quad \text{with} \quad V_i = \sum_{j(i \neq i)} J_{ij} \sigma_j + h_i
\]

\(\langle \sigma_i \rangle\) is a logistic function of the neuron total input \(V_i\), equal to the sum of the other neuron activities \(\sigma_j\) weighted by the couplings \(J_{ij}\), and of the local input \(h_i\).

3.2 Inference and validation of the model

As anticipated, the model is inferred from recordings of the activity of tens of neurons in the medial prefrontal cortex (mPFC) of five behaving rats [1], obtained from up to six tetrodes, implanted in both superficial and deep layers. Each recording session is divided in three \(\approx 30\)-minute epochs: a Task epoch in which the rat had to learn a rule (go left, right, where the light were on, or off, in a Y-shaped maze), which was changed as soon as the rat had learned it, and two Sleep epochs, one before (Sleep Pre) and one after (Sleep Post) the Task epoch. The same neurons are recorded in the three epochs of each session, but are generally different in different sessions, due to inevitable electrode movements.

With the Adaptive Cluster Expansion (ACE) algorithm of [178], recalled in section 2.2.2, we infer the parameters \(h_i\) and \(J_{ij}\) for the three epochs of the 97 recorded sessions, together with their statistical error bars due to the finite recording time, \(\Delta h_i = \sqrt{\frac{1}{B} (\chi^{-1})_{i,i}}\) and \(\Delta J_{ij} = \sqrt{\frac{1}{B} (\chi^{-1})_{ij,ij}}\) with \(\chi\) Hessian of the regularized cross-entropy (see section 2.2.2). To regularize the inverse problem, we first remove from the datasets neurons spiking only a few times, i.e. with \(f_i < 10/B\) (where \(B\) is the number of time-bins) in any recorded epoch; then we add an L2 regularization, which sets the maximal absolute value of the couplings to about \(J_{\text{Max}} \simeq 8\). As a result the couplings \(J_{ij}\) associated to vanishing pair
probabilities $f_{ij} = 0$ are inferred to be $-J_{\text{Max}} \simeq -8$, with error bar $\Delta J_{\text{Max}} = \sqrt{10}$. Note that the value of $J_{\text{Max}}$ (and of those couplings) can be tuned arbitrarily as long as it stays larger than the couplings associated to co-firing neurons, i.e. pairs of neurons $i,j$ with $f_{ij} > 0$. Throughout this study we find convenient to set the couplings corresponding to $f_{ij} = 0$ to $-\Delta J_{\text{Max}}$.

To test the model, we compute the average values of observables with the inferred distribution $P$, through Monte Carlo simulations, and we compare them with the empirical values. The inferred model distribution reproduces the single-neuron and pairwise spiking probabilities in a time-bin with great accuracy (Fig. 3.2 a). In addition, it also predicts the value of higher-order moments such as triplet firing probabilities, and the probability of multiple neuron firing in a time-bin, in excellent agreement with the data (Fig. 3.2 b). Our model approach also successfully complies with standard criteria for statistical inference, such as cross-validation. To cross-validate the model we divide the data set of the Task epoch of one representative session, called session A, in two halves. We extract the spiking frequencies $f_i$ and $f_{ij}$ from the first half of the recording, and we infer the Ising model able to reproduce these data within their statistical errors. We then compare in Fig. 3.2 c the connected correlations $c_{ij} = f_{ij} - f_i f_j$ obtained with the model to their experimental counterparts computed from the first half of the data (used for the inference, left panel), and from the second half (independent of the inference, right panel). We choose to represent the connected correlations $c_{ij}$ rather than the pairwise probabilities $f_{ij}$ as the former are more sensitive to errors in the inference than the latter. The excellent agreement confirms the absence of overfitting in our inference.

We also test the robustness of the inference with respect to the choice of the threshold $\Theta$ in the ACE (that is the threshold that discriminates the significant clusters to be used to fit the model from the “noisy” clusters), or in other words how the couplings resulting from inference vary when the threshold $\Theta$ is made slightly larger or smaller than the optimal value $\Theta^*$. As already recalled in chapter 2, the optimal $\Theta^*$ is the value at which the inferred Ising model reproduces the experimental low-order statistics within the expected sampling accuracy. For large threshold values with respect to $\Theta^*$ a small number of clusters are selected, and the approximation to the entropy is poor. For $\Theta$ small, the approximation is much better, but requires to treat an exceedingly large number of clusters and is likely to overfit the data. For the values of $\Theta^*$ corresponding to our data, many 2-neuron clusters are discarded, and the corresponding couplings vanish. As a result the coupling networks are quite sparse. By lowering the threshold $\Theta$, more and more couplings acquire a value different from zero. In Fig. 3.3 we compare the couplings obtained for a high ($\Theta^+ \simeq 4.10 \times 10^{-5}$, 130 clusters selected) and a low ($\Theta^- = 8.63 \times 10^{-6}$, 325 clusters selected) threshold with those reproducing optimally the experimental correlations ($\Theta^* \simeq 1.63 \times 10^{-5}$, 228 clusters selected) for session A, which has 37 neurons. At high thresholds many couplings are equal to zero, which is not compatible with the error bars calculated at the optimal $\Theta^*$. At low thresholds many couplings, which are zero in the correct set, are set to a non-zero value; the difference are however small and statistically compatible with zero. Due to the absence of a sharp definition for overfitting, there is some arbitrariness in the best $\Theta^*$ value. We have extensively checked for various sessions and epochs that small $\Theta$ variations (within 5%) around $\Theta^*$ do not affect our analysis and results.

The structure of the inferred interaction network is found to be largely sparse, with an average of about 60% of zero couplings across epochs and sessions, while about 40% of pairwise correlations are compatible with zero within one standard deviation. This
3.2. INFERENCE AND VALIDATION OF THE MODEL

Figure 3.2: Quality and validation of the inferred model. a. Reproduction of the low-order statistics of the spiking data. Scatter plots of the single-neuron ($f_i$, left panel) and of the pairwise ($f_{ij}$, right panel) frequencies. Values of the frequencies computed from the spiking data are shown along the $x$-axis, while their counterparts computed with Monte Carlo simulations from the inferred model distribution $P$, eq. 3.1, are shown along the $y$-axis. b. Predictions for higher-order statistics. Left panel: scatter plot of the triplet frequencies $f_{ijk}$ computed from the data ($x$-axis) and from the inferred model distribution ($y$-axis). Right panel: probability $p(k)$ that $k$ neurons are active in a time-bin (of duration $\Delta t = 10$ ms), computed from the data and from the model distribution. The agreement is excellent for $k$ such that $p(k)$ times the number of time-bins is larger than or equal to one, that is, provided the recording time is sufficient to sample those rare configurations of multiple neuron firing. c. Cross-validation of the model distribution $P$, inferred from the spikes emitted in the first half of the recording of the Task epoch in session A. The correlations $c_{ij} = f_{ij} - f_i f_j$ are shown along the $x$-axis (left panel: first half of the recording, right panel: second half), and compared to the values computed from the model distribution ($y$-axis). In both panels the points lie close to the diagonal line, within one or two error bars corresponding to the statistical standard deviation due to the finite sampling. Moreover the offsets from the diagonal are of the same order of magnitude in the two panels, confirming the absence of overfitting in our method.
CHAPTER 3. A NEW METHOD TO STUDY CELL ASSEMBLIES

Figure 3.3: **Comparison of couplings obtained at different thresholds $\Theta$.** We plot the differences between the couplings $J_{ij}$ obtained for a large (top, $\Theta^+ \simeq 4.10 \times 10^{-5}$) and a low (bottom, $\Theta^- = 8.63 \times 10^{-6}$) threshold with those corresponding to $\Theta^*(\simeq 1.63 \times 10^{-5})$ as a function of the pair index, running from 1 to $N(N-1)/2$. Green (respectively, blue) points show pairs $i,j$ whose coupling differences are more than two (resp., three) error bars $\Delta J_{ij}$ away from zero. With a too large threshold $\Theta$, too many couplings are set to zero and the model is not able to reproduce the dataset statistics. For too small $\Theta$ values many couplings become nonzero; the network is not sparse any longer, and overfits the data.

is not a result of the regularization (since the chosen L2 regularization does not enforce sparsity), but of the truncation of the cluster expansion of the entropy, which prevents overfitting. Therefore, the Ising model offers an accurate and compressed representation for the empirical distribution of activity snapshots, over different dynamical regimes, like active behaviour and sleep.

### 3.3 Choice of the time-bin width $\Delta t$

The choice of the value of the time-bin $\Delta t$ is related to the characteristic correlation time in the spiking data and to sampling. The spiking probability in a time-bin should be much smaller than unity; otherwise, for cells which spike more than once within $\Delta t$, the activity binary variable $\sigma_i$ would be set to one, and one would lose spiking events in the sampling. Moreover the time-bin should not be too small to avoid missing correlated spiking events in the inference. Average correlation in the firing of a pair of neurons is described by the cross-correlation histogram of spiking delays between the two cells, as illustrated in paragraph 2.1. To avoid missing correlated spiking events, the bin window $\Delta t$ should be chosen to capture the central peak of delays $\tau$ in $H$. The cross-correlation histograms of four representative pairs of cells in session A (Task epoch) are shown in...
3.3. CHOICE OF THE TIME-BIN WIDTH $\Delta T$

Figure 3.4: Cross-correlation histograms $H$ for four pairs of cells during the Task epoch of session A vs. time delay $\tau$ (in seconds). The binning time of the histograms is $\Delta T = 10$ ms. Note the differences of scales along the vertical and horizontal axis between the panels.

Figure 3.5: Scatter plots of the inferred fields (left) and couplings (right) for time-bin widths $\Delta t^{(1)} = 10$ ms and $\Delta t^{(2)} = 30$ ms, for the Task epoch of session A. Left: The five red circles correspond to five strongly coactivating neurons (those of the Replay group, see the following part of the chapter). The blue and red lines correspond to, respectively, $h_{30} = h_{10} + \log (\Delta t^{(2)}/\Delta t^{(1)})$ and $h_{30} = h_{10}$. Right: Blue points identify reliable couplings, i.e. whose absolute value is larger than three times their statistical standard deviation. Black points are unreliable couplings, statistically compatible with zero. The red line corresponds to $J_{30} = J_{10}$. Note that there are more zero couplings for $\Delta t = 30$ ms since the number of neuronal configurations (time-bins) is 3 times smaller than for $\Delta t = 10$ ms: correlation values are known with less accuracy, and can be reproduced within their error bars with a sparser interaction network ($\Theta^* = 1.96 \times 10^{-4}$; 152 clusters).
Fig. 3.4. For some pairs, e.g., cells 8-14, the histogram is essentially flat, showing no correlation at all. Negative and narrow peaks, extending over short delays \( \sim 10 \text{ ms} \) are found for other pairs, e.g., 35-37 or 14-28. Cross-correlation histograms may also vary smoothly over longer time scales \( \tau \), ranging between 10 and 100 ms, see for instance cells 1-9. We choose \( \Delta t = 10 \text{ ms} \), the smallest time-scale which guarantees to capture the cross-correlation peaks; this choice gives spiking probabilities \( \leq 0.3 \) for the most active cells (firing frequencies of about 30 Hz), and allows us to avoid double spiking events in a time-bin.

We further check the robustness of the inferred couplings with respect to the choice of the time-bin width. Fig. 3.5, right panel, shows couplings \( \{J_{ij}\} \) inferred for two different time-bin widths, \( \Delta t^{(1)} \) and \( \Delta t^{(2)} \): the couplings have no systematic dependence on the time-bin width. As long as \( \Delta t \geq 10 \text{ ms} \) and it is not too large (\( \Delta t \ll 50 \text{ ms} \)), couplings statistically different from zero are, up to small differences, independent of the time-bin width.

On the contrary, we observe in Fig. 3.5, left panel, that most fields \( \{h_i\} \) depend logarithmically on the time-bin width: \( h_i^{(2)} - h_i^{(1)} \approx \log \left( \Delta t^{(2)}/\Delta t^{(1)} \right) \), as expected for a system of independent neurons. Indeed, in the Ising model of a single neuron \( \sigma \), with firing rate \( f \), the value of the local field \( h \) is related to the probability \( p(\Delta t) \) that the neuron is active in a time-bin of width \( \Delta t \), computed from the spike recordings, through

\[
p(\Delta t) \sim f \cdot \Delta t = \frac{e^h}{1 + e^h}
\]

Hence, \( h \approx \log(f \Delta t) \) for small \( \Delta t \). In a system of more than one neuron with non-zero couplings this calculation does not hold anymore, but offers a simple approximation for the dependence of most fields upon the time-bin width. This simple rule breaks down for strongly coactivating (coupled) neurons, see Fig. 3.5, left panel.

### 3.4 Comparison of the coupling networks across the epochs

I will first describe our findings for one typical experimental session, called A, with \( N = 37 \) neurons, and I will summarize the results obtained on other six representative sessions, labelled B to G; results on sessions B to G (together with further details on session A) will then be illustrated extensively in paragraph 3.11. In this section, I will also present a global measure which allows us to rapidly scan all 97 sessions and retain the most significative ones for the characterization of cell assemblies and the study of experience-related replay.

We first compare the networks inferred from the three recorded epochs of session A. The scatter plot of the inferred couplings is shown in Fig. 3.6, giving information about how couplings change between the epochs. Most couplings do not vary much between Sleep Pre, Task, and Sleep Post, while some are strong in the Sleep epochs only. Interestingly, some couplings are weak or negative in Sleep Pre, and become stronger and positive in Task and in Sleep Post. In session A those effectively potentiated couplings are mostly supported by a group of five neurons (1-9-20-21-26 in Fig. 3.7, top panels), which are strongly and positively interconnected in Task and in Sleep Post, but not in Sleep Pre.

Similar scatter plots can be drawn and studied for all available sessions. Session B, for instance, with \( N = 10 \) recorded cells only, displays a behavior similar to session A.
3.4. **Comparison of the Coupling Networks Across the Epochs**

![Diagram](image)

**Figure 3.6:** Comparison of the inferred couplings in the different epochs of session A. Scatter plot of the Ising couplings inferred for the Task vs. the Sleep Post epochs in session A. Positive and negative (or null) couplings in the Sleep Pre epoch are shown with, respectively, + and − symbols. A group of five neurons (1-9-20-21-26) supports most of the potentiated couplings, shown by red circles; five of the pairs are shown with the corresponding neuron numbers.
CHAPTER 3. A NEW METHOD TO STUDY CELL ASSEMBLIES

Figure 3.7: Relevant subnetworks of couplings in sessions A, B, C, D. Relevant subnetworks of couplings in the three epochs of sessions A to D (red: $J > 0$, blue: $J < 0$; line thickness is proportional to $|J|$). Pyramidal cells are shown with triangles, undetermined cells with circles, and interneurons with squares. In Session A to C, we show the subnetworks of couplings with most changes between Sleep Pre and Post; the subnetworks in Sleep Post show large similarities with the ones in Task. In Session D, where no significant change in the inferred couplings is observed between Sleep Pre and Post, we show the most interconnected subgroup of (3) neurons. Session A: the potentiated group is composed of neurons 1-9-20-21-26 identified in Fig. 3.6. Session B: the potentiated group is composed of neurons 3-4-6-10. Session C: neurons in the potentiated subgroup (9-29-45) have inhibitory connections with the group (12-40-42-44). Session D: one 3-cell, largely connected group (28-31-32) is conserved across all three epochs.
3.5. NULL MODEL FOR THE COUPLING ADJUSTMENT

(par. 3.11.2, Fig. 3.27), with an effectively potentiated group of four cells (Fig. 3.7). Session C exhibits a more complex network reconfiguration between the Sleep Pre and Post epochs, with the appearance of new positive (as in sessions A and B) and new negative couplings (Fig. 3.31). As shown in Fig. 3.7, the effectively potentiated group is made of three neurons (9-29-45), with couplings vanishing in Sleep Pre but large and positive in Task and Sleep Post. In addition, many couplings between this group and another group of four neurons (12-40-42-44) are depressed, decreasing from zero values in Sleep Pre to negative values in Task and in Sleep Post. In session D no signature of task-related change in the couplings is found (Fig. 3.36); the three largest positive couplings define the network shown in Fig. 3.7, which is conserved across the three epochs. Session E includes two strongly potentiated couplings between two unrelated pairs of neurons (Fig. 3.40). Sessions F and G show very similar behaviors to, respectively, A and B, see Figs. 3.42 & 3.47.

The examples above show that Task-related changes of the couplings between the Sleep epochs greatly vary across the sessions. To quantify this effect in a way allowing us to scan efficiently the 97 sessions we introduce the following session-wide estimator:

\[
Adj = \sum_{\text{pairs } i,j \text{ with nonzero couplings in Task and Sleep Post}} \text{sign} \left( J_{ij}^{\text{Task}} - J_{ij}^{\text{Sleep Pre}} \right) \times \left( J_{ij}^{\text{Sleep Post}} - J_{ij}^{\text{Sleep Pre}} \right) \quad (3.4)
\]

\(Adj\) is measure of the task-related adjustment of the inferred couplings between the Sleep Pre and Post epochs. The presence of the sign function allows us to sum constructively contributions corresponding to effective potentiation (as in sessions A, B, C) and depression (as in session C) of the couplings. The summation is restricted to pairs \(i, j\) of neurons whose couplings are significantly different from zero in Task and Sleep Post. In practice we require that \(|J_{ij}|/\Delta J_{ij} > 3\), where \(\Delta J_{ij}\) is the error bar on the inferred coupling \(J_{ij}\), though the value of \(Adj\) varies little upon relaxing the criterion to \(|J_{ij}|/\Delta J_{ij} > 2\).

Fig. 3.8, left panel, shows the values of \(Adj\) vs. the numbers \(N\) of recorded cells for the 97 sessions. Some sessions (including, but not restricted to, A, B, C, E, F, G) have large and positive \(Adj\), more than one standard deviation above the average for a null model, where the correspondence between pairs of neurons across the epochs is removed by reshuffling the neuron indices, see next paragraph. On the other hand, there is basically no session with large and negative \(Adj\), more than one standard deviation below the average for the null model, meaning that potentiation and depression of the couplings are never anticorrelated to \(J_{ij}^{\text{Task}} - J_{ij}^{\text{Sleep Pre}}\); either a session does not undergo any significative mean potentiation or depression (dots lying within the red dashed lines in Fig. 3.8), or the change is related to the Task (dots above the upper red dashed line in Fig. 3.8). The outcome of a control calculation, where, for each session, we exchange the Sleep Pre and Sleep Post couplings in eq. 3.4 is shown in Fig. 3.8, right panel. As expected no large-\(Adj\) session is found. This simple control provides a clear evidence for the fact that \(Adj\) captures experience-related changes in the couplings. We then focus mainly (but not solely) on sessions with high coupling adjustment for the identification and characterization of cell assemblies and replay.
Figure 3.8: **Coupling adjustment across all 97 recorded sessions.** Coupling adjustment $Adj$, see eq. 3.4, is shown for the 97 sessions in the left panel. The right panel shows a control calculation, where we have exchanged the Sleep Pre and Post inferred couplings. Red lines show the predictions of the null model (average: full lines, ±1 standard deviation: dashed lines), see par. 3.5. Colors identify the five recorded rats; circles locate sessions A to G.

Figure 3.9: **Histograms of coupling adjustments $a_{ij}$.** Upper panel: straight data (28790 pairs $i,j$). Lower panel: shuffled data (28790^3 pairs).
3.5 Null model for the coupling adjustment

We define the adjustment $a_{ij}$ of the coupling between two cells $i$ and $j$ through

$$a_{ij} = \text{sign}(J_{ij}^{\text{Task}} - J_{ij}^{\text{Sleep Pre}}) \times (J_{ij}^{\text{Sleep Post}} - J_{ij}^{\text{Sleep Pre}}) \text{ if } \frac{|J_{ij}^{\text{Sleep Post}}|}{\Delta J_{ij}^{\text{Sleep Post}}} > 3 \text{ and } \frac{|J_{ij}^{\text{Task}}|}{\Delta J_{ij}^{\text{Task}}} > 3$$

$$a_{ij} = 0 \text{ otherwise}$$

(3.5)

where $\Delta J$ denotes the statistical error on the couplings. As can be seen in Fig. 3.9, the distribution $P(a)$ is asymmetric, with a larger density on positive weights. The total coupling adjustment of an experimental session, $\text{Adj}$, is defined as the sum of $a_{ij}$ over all pairs $i < j$, see eq. 3.4.

To derive the null model for the coupling adjustment shown in Fig. 3.8, we first estimate the histogram $P(a^{SH})$ of the reshuffled coupling adjustment $a^{SH}$, where the reshuffling consists in randomly choosing the pairs $ij, kl, mn$ of the three couplings for the three epochs over the set of $N_{\text{tot}}$ couplings obtained from all sessions of all rats:

$$a^{SH}_{ij,kl,mn} = \text{sign}(J_{ij}^{\text{Task}} - J_{kl}^{\text{Sleep Pre}}) \times (J_{mn}^{\text{Sleep Post}} - J_{kl}^{\text{Sleep Pre}}) \text{ if } \frac{|J_{mn}^{\text{Sleep Post}}|}{\Delta J_{mn}^{\text{Sleep Post}}} > 3 \text{ and } \frac{|J_{ij}^{\text{Task}}|}{\Delta J_{ij}^{\text{Task}}} > 3$$

$$a^{SH}_{ij,kl,mn} = 0 \text{ otherwise.}$$

(3.6)

The histogram of the reshuffled coupling adjustment $P(a^{SH})$ is shown in Fig. 3.9, bottom panel. The means $\mu$, standard deviations $\Delta a$, and statistical errors $\Delta \mu$ on the means are (first two rows):

<table>
<thead>
<tr>
<th>Data</th>
<th>$\mu = \langle a \rangle$</th>
<th>$\Delta a = \sqrt{\langle a^2 \rangle - \langle a \rangle^2}$</th>
<th>$\Delta \mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Task-Post</td>
<td>$8.92 \times 10^{-3}$</td>
<td>$0.1129$</td>
<td>$6.6 \times 10^{-4}$</td>
</tr>
<tr>
<td>Reshuffled</td>
<td>$4.39 \times 10^{-3}$</td>
<td>$0.1283$</td>
<td>$2.6 \times 10^{-8}$</td>
</tr>
<tr>
<td>Post-Task-Pre</td>
<td>$4.08 \times 10^{-3}$</td>
<td>$0.1297$</td>
<td>$4.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>Reshuffled</td>
<td>$4.45 \times 10^{-3}$</td>
<td>$0.1298$</td>
<td>$2.7 \times 10^{-8}$</td>
</tr>
</tbody>
</table>

where $\Delta \mu$ is computed as the standard deviation $\Delta a$ divided by the square root of the number of pairs $ij$ (straight data), or triplets of pairs $ij, kl, mn$ (reshuffled data) used to estimate the mean. We also present in the Table (last two rows) the values obtained in the control calculation, upon exchange of the Sleep Pre and Post coupling sets. Interestingly, the value of the mean in this control case is not compatible with the mean obtained for the real data, but is close (one standard deviation) to the value obtained after reshuffling.

The values of $\mu^{SH}$ and $\Delta a^{SH}$ listed on the second and fourth lines of the Table above allow us to construct a null model for the total coupling adjustment $\text{Adj}$ for the true and the control data. For a session with $N$ recorded neurons the null model distribution $P(\text{Adj}|N)$ for $\text{Adj}$ is obtained by summing $N(N-1)/2$ coupling adjustments $a^{SH}$, drawn randomly from the reshuffled histogram. It is easy to compute the mean and the standard deviation of $P(\text{Adj}|N)$:

$$\langle \text{Adj} \rangle = \frac{N(N-1)}{2} \mu^{SH}, \quad \sqrt{\langle \text{Adj}^2 \rangle - \langle \text{Adj} \rangle^2} = \sqrt{\frac{N(N-1)}{2}} \Delta a^{SH},$$

(3.7)

which define the red lines of Fig. 3.8. The list of the sessions whose coupling adjustment $\text{Adj}$ exceeds the null-model average by one standard deviation or more ($\text{StN} = (\text{Adj} - \langle \text{Adj} \rangle)/\sqrt{\langle \text{Adj}^2 \rangle - \langle \text{Adj} \rangle^2} \geq 1$) is (session D is not included in the list):
3.6 Simulations of the inferred model

We now simulate the model distribution $P$ derived above to identify groups of neurons, or ‘putative cell assemblies’ most likely to coactivate. We call **self-sustaining** a configuration of activity $(\sigma_1, \sigma_2, ..., \sigma_N)$ such that for each cell $i$, the total input $V_i$ in eq. 3.2 is positive and the neuron is active ($\sigma_i = 1$), or $V_i$ is negative and the neuron is silent ($\sigma_i = 0$). In other words, in a self-sustaining configuration, each neuron activity is consistent with the inputs coming from the other neurons (and from $h_i$). It is easy to show that self-sustaining configurations are local maxima of the distribution $P$, eq. 3.1, and are therefore prototypical patterns of the neuronal activity. Indeed, when $\sigma_i = 1$ and $V_i > 0$, changing the activity of neuron $i$ to $\sigma_i = 0$ changes the log-probability log $P$ by $-V_i < 0$; similarly, when $\sigma_i = 0$ and $V_i < 0$, changing the activity of neuron $i$ to $\sigma_i = 1$ changes the log-probability log $P$ by $V_i < 0$. In both cases inverting the value of the neuron activity leads to a decrease of log $P$.

The search procedure for configurations in which all neurons are self-sustaining is the following: we start with the all-silent neuron configuration ($\sigma_i = 0$ for $i = 1, ..., N$). If the configuration is self-sustaining, the algorithm has found a maximum of $P$ and halts. If one or more neurons are not self-sustaining, i.e. their values $\sigma_i$ do not agree with the signs of their total inputs $V_i$, we pick up uniformly at random one of them, say, $i$, and flip its value $\sigma_i$ (from silent to active, or vice-versa). The asynchronous updating is iterated until the configuration is self-sustaining. This algorithm is a Glauber dynamics at zero temperature and it is guaranteed to converge as the log-probability of the configuration increases after each updating step. Re-running the dynamics may, however, produce different maxima, due to the stochasticity in the choice of the (non self-sustaining) neuron to flip at each step.

As the effective couplings $J_{ij}$ and local inputs $h_i$ reflect the neuronal activity distribution over the entire epoch, which is quite sparse, a simulation of the model distribution with the parameters as inferred will not generate any self-sustaining state but the all-silent
neuron configuration, \( \sigma_i = 0 \) for all \( i = 1, \ldots, N \). In neuronal activity, cell assemblies are not sustained for a long period, but occur transiently during bouts of high neuronal excitability. To simulate that, we introduce an extra parameter \( H \) which increases the likelihood of configurations with many active neurons (in a uniform way \([212]\)) and allows us to reproduce assembly-generating transients. In practice, for each epoch of a chosen recorded session, we identify the self-sustaining configurations upon changing the total inputs \( V_i \) into \( V_i + H \) in eq. 3.2 for all neurons \( i \). For each value of the drive \( H \) we determine the self-sustaining configurations \( \sigma \) (to search for all self-sustaining patterns we run the Glauber dynamics one million time for each value of \( H \)), their numbers of active neurons, and the total inputs \( V_i(\sigma) + H \) to those active neurons.

Figure 3.10 shows the number of active neurons in the self-sustaining configurations as a function of \( H \) for the three epochs of session A. As \( H \) increases from zero, neurons start to activate one after the other, in decreasing order of their local inputs \( h_i \). As more and more neurons \( i \) get activated, contributions to the total inputs of the other neurons \( j \) build up, facilitating \( (J_{ij} > 0) \) or hindering \( (J_{ij} < 0) \) their activations.

Discontinuous ‘jumps’ in the number of active neurons are found at special values of \( H \), and are indicated with arrows in Fig. 3.10. A jump signals the coexistence of two self-sustaining configurations \( \sigma^{(1)} \) and \( \sigma^{(2)} \), with respectively, a low and a high number of active neurons. We take such co-existence of a high and a low-activity configuration as the hallmark of a cell assembly, as it highlights how transiting to a new self-sustaining state requires the concomitant activation of a group of neurons. We show in the insets of Fig. 3.10 the variations \( \delta \langle \sigma_i \rangle \) in the conditional average activities, \( \langle \sigma_i \rangle \) in eq. 3.2, between the ‘low’ and the ‘high’ self-sustaining configurations. As the local inputs \( h_i \) and the drive \( H \) are constant across the jump, large variations \( \delta \langle \sigma_i \rangle \) (in absolute value) may come only from the collective activation of a group of neurons \( i \) (Fig. 3.1), which we call putative cell-assembly (As). More precisely, to determine the cell assembly attached to the jump we rank all the cells \( i \) according to the variation \( \delta \langle \sigma_i \rangle = \langle \sigma_i^{(2)} \rangle - \langle \sigma_i^{(1)} \rangle \) of their conditional average values. Neurons with the largest \( \delta \langle \sigma_i \rangle \) are included in the assembly. The cut-off value over \( \delta \langle \sigma_i \rangle \) is chosen to be \( \simeq 0.2 \) for session A, but may depend on the session, see paragraph 3.11.

In session A, in either Task or Sleep Pre, only one out of the two jumps defines a putative cell assembly, while the other jump does not show strong variations \( \delta \langle \sigma_i \rangle \), see Insets in Figs. 3.10 a&b (and Fig. 3.22 in 3.11.1). In Sleep Post the single jump defines a large putative cell assembly (Fig. 3.10 c).

Comparing the cell assemblies of session A across the different epochs we find common subgroups of neurons (Fig. 3.10 and caption). The ‘Sleep’ group (neurons 6-7-11-12) is shared by As. 2 in Sleep Pre and the large assembly As. 3 in Sleep Post (Figs. 3.10 b&c); the ‘Replay’ group (neurons 1-9-20-21-26) is shared by As. 1 in Task and As. 3 in Sleep Post (Figs. 3.10 a&c). The Replay group coincides with the group of strongly potentiated couplings in Figs. 3.6 & 3.7. The subnetworks of couplings, inferred in the three epochs of session A, between the neurons in the Sleep and Replay groups are shown in Fig. 3.11. Neurons in the Sleep group are the only ones to be interconnected by large and positive couplings in the Sleep Pre epoch, and the same statement holds for the cells in the Replay group in the Task epoch. In the Sleep Post epoch, however, the two groups become largely interconnected. The merging of the Sleep and Task cell assemblies in a large, interacting group of cells in Sleep Post, corresponding to the large coactivation jump in Fig. 3.10 c, is a general finding for all sessions with a large coupling adjustment and a fairly large number of recorded neurons, see description of scenarios below and par. 3.11.
CHAPTER 3. A NEW METHOD TO STUDY CELL ASSEMBLIES

Figure 3.10: Identification of putative cell assemblies in session A. Number of active neurons in the self-sustaining patterns of the Ising model as a function of the drive $H$ for the Task (a), Sleep Pre (b), Sleep Post (c) epochs. Distinct self-sustaining patterns may coexist (at a given $H$); they are indicated by the same colored dots and define jumps indicated by arrows. The changes in the conditional averages $\langle \sigma_i \rangle$ corresponding to the self-sustaining patterns shown by colored dots are given in insets. While the first two jumps in Sleep Pre and Task correspond to small changes in $\langle \sigma_i \rangle$, the second two jumps in those epochs and the jump in Sleep Post include a group of neurons with substantial changes in $\langle \sigma_i \rangle$, which define our cell assemblies. As. 1 includes neurons 1-9-10-16-18-20-21-26-29, As. 2 includes 2-6-7-8-9-11-12-13, and As. 3 includes 1-5-6-7-8-9-11-12-15-20-21-26-29-34-35. Two groups of neurons, common to the cell-assemblies of different epochs, are the Replay (1-9-20-21-26) and the Sleep (6-7-11-12) groups.
3.7. COACTIVATION OF THE ‘PUTATIVE CELL ASSEMBLIES’

Figure 3.11: Association of the Replay- and the Sleep-group subnetworks in the Sleep Post epoch of session A. Subnetwork of strong couplings within the Replay (top five neurons 1-9-20-21-26, black symbols) and the Sleep (bottom four neurons 6-7-11-12, violet symbols) groups identified in Fig. 3.10. Red: $J > 0$, Blue: $J < 0$. Line width is proportional to $|J|$. Pyramidal cells are shown with triangles.

3.7 Coactivation of the ‘putative cell assemblies’

We now check that the putative assemblies found in the model simulations correspond to real coactivations of the associated neurons in the spiking data. To this aim we define the coactivation ratio (CoA) of a group $G$ of neurons over the time scale $\tau$ through

$$\text{CoA}(G, \tau) = \frac{f(G)}{\prod_{i \in G} f_i}$$

where $f(G)$ is the probability that all the neurons in the group are active within the time scale $\tau$, and the denominator is the product of the individual spiking probabilities. For a group of independent cells the CoA is on average equal to unity. To assess the statistical validity of the CoA for a group $G$ of neurons we also compute the error bar on CoA. Assuming a Poisson distribution for the coactivation events, the standard deviation of the CoA is estimated to be $\text{CoA}(\tau) / \sqrt{N_G(\tau)}$, where $N_G(\tau)$ is the number of coactivation events for the cells in $G$ over the time scale $\tau$. As a relevant coactivation event contributing to $f(G)$ should correspond to a sequence of ‘readings’ of spikes, triggering in turn the next spike [7], we expect $\tau$ to be not larger than $n \times \Delta t$, where $n$ is the number of neurons in $G$ and $\Delta t = 10$ ms is the time-bin duration used for the inference.

We compute the CoA of the cell assembly As. 1 from the spiking data, with the results shown in Fig. 3.12 a, left (and extensively in Fig. 3.23 of section 3.11.1). As. 1 is found to strongly coactivate in Task and in Sleep Post (on much longer time scales), but only during Slow-Wave-Sleep periods (SWS), in which hippocampal sharp waves are known to be important for memory consolidation. In Sleep Pre As. 1 does not coactivate, which is
Figure 3.12: Coactivation ratios (CoA) in session A. a. CoA of As. 1 in Task ($\delta \log P \simeq 11.9$), of the group 2-6-9-11-12-13 in As. 2 of Sleep Pre ($\delta \log P \simeq 8.9$) and of the group 7-8-11-12-20-21-26-35 in As. 3 of Sleep Post ($\delta \log P \simeq 12.3$). b. CoA of the Replay and Sleep groups; see caption of Fig. 3.10 for the lists of corresponding neurons. CoAs are shown for time scales $\tau$ ranging from 5 ms to $n \times 20$ ms, where $n$ is the number of neurons in each group considered. Note the variations in the CoA and temporal scales along the $y$- and $x$-axis between the panels. See text for the computation of error bars. CoA equal to zero are compatible with the independent-cell hypothesis due to the low-firing rates of the neurons (see text).
compatible with the independent-cell hypothesis due to the low firing frequencies: indeed, assuming independence between cells, simultaneous firing events (contributiong to $f(G)$) are unlikely to be found, and the CoA is likely to be zero, if the duration of the recording is small compared to $T_{\text{min}} = \tau / \prod_{i \in G} f_i(\tau)$.

The ‘Sleep’ group (neurons 6-7-11-12), shared by As. 2 in Sleep Pre and the large assembly As. 3 in Sleep Post, coactivates in Sleep Pre and Sleep Post, both in SWS and non-SWS periods over $\tau \simeq 30 - 40$ ms (Fig. 3.12 b). We find essentially no coactivation in Task (CoA close to 1). The ‘Replay’ group (neurons 1-9-20-21-26), shared by As. 1 in Task and As. 3 in Sleep Post, and coinciding with the five-neuron group which supports the strongly potentiated couplings (Fig. 3.7), strongly coactivates in Task and in SWS-Sleep Post, on similar time scales, respectively, $\tau \simeq 20 - 30$ ms and $\tau \simeq 30 - 40$ ms, and does not coactivate in Sleep Pre nor in non-SWS periods of Sleep Post (Fig. 3.12 b, note that in the independent-cell hypothesis, the CoA of the Replay group should be zero for time scales $\tau \leq 55$ ms in the Task epoch, and for all the values of $\tau$ considered in Sleep Pre and Post). In addition, the large CoAs of the Replay group found in Task and in SWS-Sleep Post are significantly higher than CoAs for random groups of five neurons, as illustrated in Fig. 3.25 of par. 3.11.1. These findings support the hypothesis that the five-cell Replay group is (part of) a cell assembly involved in memory consolidation.

It is interesting that the model, inferred from spiking frequencies and pairwise correlations computed on a 10 ms time-scale, is able to predict the coactivation of multi-neuron firing patterns, which occurs on much longer time-scales, spanning several tens of ms. This is possible because multi-neuron patterns are built up from pairwise interactions, which concatenate neurons in twos: catching the pairwise interactions, the model is able to reconstruct the whole coactivation chain upon stimulation with the external drive. This concept will be illustrated in more detail in the next chapter.

The coactivation of subgroups of neurons in the putative cell assemblies of each epoch can be studied further. Looking at subgroups rather than the whole assembly allows us to investigate the internal structure of the putative cell assemblies, e.g. the alternative activation of subgroups resulting from the presence of negative couplings, or simply from undersampling. As an example, in As. 2 of Sleep Pre, large CoA are found for the subgroups 2-6-9-11-12-13 (shown in Fig. 3.12 a) and 2-6-7-11-12-13 (peak CoA value $\simeq 500$), but their simultaneous activation is not observed in the data.

It is possible to quantitatively understand and predict which group of cells corresponds to strong or to weak coactivations based on the model distribution $P$. To this purpose we introduce the log–likelihood variation ($\delta \log P$) of a cell assembly, which measures the difference in the log-likelihoods of the high and the low activity configurations ($\sigma^{(2)}$ and $\sigma^{(1)}$ respectively) due to the interactions between the neurons in the assembly:

$$\delta \log P = \sum_{i<j} J_{ij} \left( \sigma_i^{(2)} \sigma_j^{(2)} - \sigma_i^{(1)} \sigma_j^{(1)} \right)$$

(3.9)

The log-likelihood variation of a subgroup of neurons in $\sigma^{(2)}$ is defined as in eq. 3.9, with the double sum restricted to the neurons in the subgroup.

In session A, the only three five-cell groups found to have CoA comparable to the Replay group are obtained by replacing one of the five cells in the group with another neuron in As. 1; these three variants are the groups with the largest $\delta \log P$ values (see section 3.11.1). Similarly, variants of the Sleep group or subgroups of the Sleep Post assembly As. 3 with large $\delta \log P$ and CoAs are also identified, see Figs. 3.12 a & 3.23.
3.8 Possible scenarios for cell assemblies across sessions

The approach described above for session A has been applied to the other available sessions; results for other six representative sessions (B to G), are summarized here (see paragraph 3.11 for more details). Given the strong and random undersampling of the neuronal activity, it is not surprising that we find different scenarios for cell assemblies.

A prototypical scenario, encompassing session A, is that a Task-related group of coactivating neurons is found in Sleep Post and it was not present in Sleep Pre. An example is provided by session B, which consists of 10 recorded cells only. The number of active neurons in the self-sustaining configurations of the inferred Ising model of the three epochs of session B is shown in Fig. 3.28. No cell assembly coactivates in Sleep Pre. However, a 4-cell assembly is found in Task and is almost perfectly reproduced in Sleep Post (Fig. 3.28 and effective networks in Fig. 3.7); this cell assembly strongly coactivates in both epochs (Fig. 3.30). The same scenario is encountered in session G (Figs. 3.48 & 3.50).

Other sessions are found to be even more similar to A in that they have a cell assembly in Sleep Pre, a different one in Task, and both merge in Sleep Post. These sessions are those with large coupling adjustment $Adj$ and a fairly large number of recorded neurons. An example is given by session F, see Figs. 3.43, 3.45 & 3.46. In session C, the Replay group is composed by three neurons, which coincide with those that undergo strong coupling potentiation and inhibit the 4-cell group of Fig. 3.7. Remarkably, the complex structure of the Replay group is apparent through a decrease of the CoA upon addition of inhibited neurons to the potentiated 3-cell group, compare Figs. 3.34 & 3.35.

In some sessions, the same cell assembly is encountered across all three epochs. This scenario is illustrated by session D, see Fig. 3.37, where the ‘conserved’ cell assembly is supported by three neurons (effective network in Fig. 3.7). We find similar values for $\delta \log P (= 4.4, 4.4, 3.8)$ and for the maximal CoA $(= 40, 25, 40)$ in, respectively, Sleep Pre, Task and Sleep Post, see Fig. 3.39.

Many sessions show very small values of the adjustment $Adj$, see Fig. 3.8, and no coactivation at all or a conserved cell assembly as in session D above. Interestingly, a few sessions with large or intermediate values of $Adj$ do not exhibit any cell assembly; for those sessions, the effectively potentiated couplings do not interconnect a small set of recorded neurons, but are scattered over non-overlapping pairs of neurons. An example is provided by session E, see Figs. 3.40 and 3.41.

3.9 Comparison with PCA–based methods

In this paragraph, I will compare our method to identify cell-assemblies in data with techniques based on Principal Component Analysis (PCA), illustrated in section 2.3.3. We consider again session A, for which Peyrache et al. [1] have identified the presence of reactivation during Sleep Post of the first principal component of the Pearson correlation matrix of the activity in Task. Despite this result, the identification of cell assemblies in each epoch is in general difficult with PCA, as shown by the following analysis.

We bin the neuron spike trains into time windows of 10 ms (as in our model inference) and 100 ms (as used by Peyrache et al. in their analysis) and compute the Pearson correlation matrix of the activity for these two choices of the time-bin. For $\Delta t = 10$ ms
3.9. COMPARISON WITH PCA–BASED METHODS

Figure 3.13: First and second principal components of the Pearson correlation matrices of the three epochs with $\Delta t = 10$ ms. Similar results are obtained for $\Delta t = 100$ ms. Orange: neurons of the Replay group; violet: neurons of the Sleep group.

Figure 3.14: CoA ratio vs. time scale $\tau$ for the groups composed by the neurons with the largest entries in absolute value ($> 0.2$) in the first (first row) and second (second row) principal component of each epoch for 10 ms time-bins. Top left: group 14,18,20,21,33,37; bottom left: group 2,8,11,12,13,24; top middle: group 1,9,20,21,26; bottom middle: group 9,10,24,29,33; top right: group 7,8,11,21,26,35; bottom right: group 18,20,21,33,37.
we identify six signal (principal) components, and five for $\Delta t = 100$ ms. We use the standard criterion for identification of the signal eigenvectors, i.e. we select the modes whose corresponding eigenvalues are larger than the upper bound of the Marchenko-Pastur distribution, $\lambda_+ = \left(1 + \sqrt{\frac{N}{B}}\right)^2$, where $B =$ number of time-bins in the recording, $N =$ number of recorded neurons. In Fig. 3.13 we show the two principal components for $\Delta t = 10$ ms. The largest entries of the first component in the Task epoch correspond to the Replay group 1-9-20-21-26, which is also represented (at least partially) in the two principal components of Sleep Post. This result explains the agreement between the measure of reactivation in [1] and our method.

However it is, in general, not easy to separate the statistically significative components from the background noise and to identify cell assemblies in each epoch with PCA. Choosing an arbitrary threshold of 0.2, the largest entries of the second component in Task are 9-10-24-29-33, among which 9-10-29 are also present in our As. 1. The CoA of the group 9-10-24-29-33 is relatively large (see Fig. 3.14, bottom middle panel), but significantly smaller than the CoA of As. 1 identified by our method (see Fig. 3.12 a of par. 3.7). With the same threshold (0.2), the largest entries of the first component in Sleep Pre correspond to neurons 14-18-20-21-33-37, which are not present in our Sleep Pre cell assembly As. 2. This group has, indeed, a very weak CoA compared to the group 2-6-9-11-12-13 identified with our method in As. 2, and strongly coactivating (Fig. 3.12 a of 3.7). The largest entries of the second component in Sleep Pre correspond to the group 2-8-11-12-13, which has a larger overlap with 2-6-9-11-12-13, and is found to coactivate in the spiking data (even though with a smaller CoA than our group and with very large error bars). The signal components in Sleep Post give similar results to those in Sleep Pre, but with an inversion between the first and the second component: the largest entries ($> 0.2$) of the first component in Sleep Post correspond to 7-8-11-21-26-35, contained in our Sleep Post cell assembly As. 3; the CoA of this group is significant, but much smaller than the CoA of the groups we have identified in this epoch (see Fig. 3.12 a of 3.7 and Fig. 3.23 of 3.11.1). The second component is peaked on neurons 18-20-21-33-37, which have a very small CoA.

In summary, the largest entries of the first two principal components of the Pearson correlation matrix have some correspondence with the cell assemblies we have identified with our method, but the groups found in each epoch with our analysis show much stronger coactivation properties than those found with PCA. Moreover this correspondence becomes less and less obvious as we move to the next (third, fourth, ...) components.

We then try to identify neuronal groups refining the previous PCA approach with clustering procedures. We represent each neuron as a point in the space of the signal components, with coordinate $v_\mu(i)\sqrt{\lambda_\mu}$ (where $v_\mu(i)$ is the entry corresponding to neuron $i$ in the $\mu$th signal eigenvector, and $\lambda_\mu$ is the $\mu$th eigenvalue). We then apply the classical $k$-means clustering algorithm to these $N$ points (where the number $k$ of clusters is arbitrarily chosen), as we expect groups of closely correlated neurons to be represented by points far from the origin and close to each other in this dimensionally-reduced space. Unfortunately, this method applied to our data does not seem to be able to identify significant clusters, well separated from noisy clusters, as shown in Fig. 3.15, where the identified clusters are projected onto the bi-dimensional space of the first two signal components. Each panel shows the clustering (optimized over $10^4$ random initial conditions) in each epoch, for a particular choice of $k = 2, 3, 4$. Neurons in the same cluster are represented by the same symbol (regardless of its color); full symbols show the farthest clusters from the origin (distance $d_c > 0.3$), empty symbols correspond to the closest ones ($d_c < 0.3$). With a few exceptions, e.g. the cluster of upward full triangles in Task, which is rather robust
with respect to the choice of \( k \), signal clusters are in general not clearly separated from each other and from the noisy clusters. In addition, no obvious choice for the value of \( k \) seems to be optimal to extract the groups of maximally coactivated neurons in each epoch. Finally, this method assigns each neuron to one cluster, and does not allow us to identify overlapping cell assemblies.

Despite these difficulties, we try to find cell assemblies in this framework and to compare them to those extracted with our method. To this purpose, we first compute the CoA ratios for the signal clusters, defined as those with centroid distance \( d_c \) from the origin greater than \( d = 0.3 \). We obtain the following results: in Sleep Pre, for \( k = 2 \), we find the signal cluster 14-17-18-20-21-25-29-33-36-37 with CoA=16±9 (peak value); for \( k = 3 \), cluster 18-20-21-23-36-37 with CoA=3.5±0.3; for \( k = 4 \), clusters 9-14-17-20-21-25-29-33 (CoA=58±34) and 2-3-4-5-6-7-8-10-11-12-13-15 (CoA=0 at all relevant time scales); in Task, we find for \( k = 2 \) the signal cluster 1-9-10-20-21-26 (CoA=213±151); for \( k = 3 \), clusters 1-20-21-26 (CoA=8.8±0.5) and 9-10-24-29-32-33-37 (CoA=81±31); for \( k = 4 \), cluster 1-20-21-26 again, and 9-10-29-33 (CoA=30±11); in Sleep Post for \( k = 2 \) the cluster 9-14-18-20-21-22-23-24-25-29-33-36-37 (CoA=3±1); for \( k = 3 \), clusters 14-18-23-33-36-37 (CoA=1.9±0.1) and 1-7-9-20-21-24-26-28-29-32 (CoA=290±50); for \( k = 4 \), clusters 14-18-33-36-37 (CoA=1.53±0.02) and 1-20-21-26-32 (CoA=280±90). All these groups have CoA smaller than those identified by our method. Since these clusters are mixed with the noisy ones, lowering the threshold on the centroid distance \( d_c \) would not help to detect cell assemblies.

We also try to identify Replay and Sleep groups by comparing the clusters with \( d_c > 0.3 \) in the different epochs: we identify two putative Replay groups (1-20-21-26 and 9-24-29), common to clusters of Task and Sleep Post, and two putative Sleep groups (18-23-36-37 and 14-18-33-36-37), shared by clusters of the two Sleep epochs, see Fig. 3.15. Neurons in these groups are very weakly coactivated (if not even independent), except those in the Replay group 1-20-21-26, which is in good agreement with our findings, but misses neuron 9 and its CoA is smaller than that of 1-9-20-21-26.

In conclusion, our method allows us to unveil cell-assemblies in a more rigorous way and to extract much more signal from the data. Through the study of the variations in the interaction part of the log-likelihood and in the neuron conditional averages between two self-sustaining patterns at different levels of activity, it also allows us both to estimate the group coactivation (eq. 3.9) and to rank the neurons according to their coactivation with the rest of the group. Finally it does not impose \textit{a priori} the number of clusters to be found in the data and allows a neuron to belong to more than one cell-assembly.

We also compare our method with the approach used by Santos et al. [208], who developed a procedure to identify groups of strongly coactivated, possibly overlapping neurons, described in section 2.3.3. The “interaction matrices” we find for the three epochs of session A are highly non-bimodal, see Fig. 3.16. The absence of a gap in the element distribution prevents us from constructing a robust and non-arbitrary adjacency matrix.

Trying, nevertheless, to define a threshold (red line in Fig. 3.16) we obtain three graphs consisting in only one connected component. Consequently the clusterization procedure results difficult and arbitrary. Moreover the selected group in Sleep Post contains neither neurons 1, 9 nor 10, 29, 35, which constitute our Replay group and its significant variants, see Table 3.2.

The same PCA-based approaches applied to other sessions lead to similar or worse results. For example, in session C, the Replay group 9-29-45 has large entries in the first principal component of Task, but is represented neither in the first nor in the second
Figure 3.15: Clustering of neurons in the space of principal components. Projection onto the first two principal components is shown in each panel, for a particular choice of the number of clusters $k$, from $k = 2$ (first row) to $k = 4$ (third row). Different clusters are identified by different symbols, with full symbols for signal clusters ($d_c > 0.3$) and empty symbols for fully noisy clusters ($d_c < 0.3$). Colors indicate putative cell assemblies identified within the signal clusters: orange for Replay groups, violet for Sleep groups. Putative Replay groups are: $1, 20, 21, 26$ (orange upward triangles in Task for all $k$ and in Sleep Post for $k = 3$ and $k = 4$), $9, 24, 29$ (orange squares in Task for $k = 3$ and orange upward triangles in Sleep Post for $k = 3$, in the same cluster as $1, 20, 21, 26$). Putative Sleep groups: $18, 23, 36, 37$ (purple squares for $k = 3$ both in Sleep Pre and Post); $14, 18, 33, 36, 37$ (purple squares for $k = 2$ in Sleep Pre and for $k = 3, 4$ in Sleep Post).
3.10 Quantitative estimates of the replay

In this paragraph, I will present the method we have developed to quantify the replay in neuronal data and its relation with the formulation given by Peyrache et al. [2]. I will then show an application of this new method to the same dataset of the previous analysis.

3.10.1 A statistical interpretation of the replay

It is possible to use the model inferred from the Task to quantify the replay, that is, to estimate how compatible the Sleep coactivation patterns are with the Task model, and if this compatibility improves in Sleep Post over Sleep Pre. In other words, we interpret the replay in a probabilistic framework, as the interaction part of the log-likelihood of the Sleep configurations with the model inferred from the Task activity. To obtain an estimate independent of the firing rates, which differ between the Task and the Sleep epochs, we introduce the Z-score variables, or rescaled Ising variables, which have zero mean and unit variance in all epochs:

$$Z_i = \frac{\sigma_i - f_i}{\sqrt{f_i(1 - f_i)}}$$

(3.10)

The probability distribution $\hat{P}$ of the configuration $\{Z_i\}$ has the same functional form as the Ising distribution $P$ (3.1), with rescaled couplings

$$\hat{J}_{ij} = \sqrt{f_i(1 - f_i)} J_{ij} \sqrt{f_j(1 - f_j)}$$

(3.11)

We can then estimate the replay $R(t)$ as the interaction part of the log-probability of the Sleep configuration in time-bin $t$ with the rescaled Ising model inferred from the Task recording:

$$R(t) = \sum_{i<j} \hat{J}_{ij}^{\text{Task}} Z_i^{\text{Sleep}}(t) Z_j^{\text{Sleep}}(t)$$

(3.12)

$R(t)$ measures how well the Sleep activity agrees with the coupling structure inferred from the Task epoch.
The total amount of replay in a session (session-wide replay) can be estimated as the difference between the average replay during the SWS periods of Sleep Post and the average ‘preplay’ during the SWS periods of Sleep Pre (which can be seen as a baseline value for the replay):

$$R = \langle R(t) \rangle_{SWSPost} - \langle R(t) \rangle_{SWSPre} =$$

$$= \sum_{i<j} J_{ij}^{\text{Task}} \left( \langle Z_i(t) Z_j(t) \rangle_{SWSPost} - \langle Z_i(t) Z_j(t) \rangle_{SWSPre} \right) =$$

$$= \sum_{i<j} J_{ij}^{\text{Task}} \left( \Gamma_{ij}^{SWSPost} - \Gamma_{ij}^{SWSPre} \right)$$

where $\Gamma_{ij}^{SWSPre}$ and $\Gamma_{ij}^{SWSPost}$ are the Pearson correlation matrices of the activity during SWS Pre and SWS Post, respectively:

$$\Gamma_{ij} = f_{ij} - f_i f_j \sqrt{f_i(1-f_i) f_j(1-f_j)}$$

### 3.10.2 Comparison with previous interpretations

It is interesting to compare this definition of replay and session-wide replay with the estimates given in previous works [1, 2].

As recalled in section 2.3.3, Peyrache et al. have quantified the replay as the reactivation in the Sleep Post epoch of the first principal component of the Pearson correlation matrix computed from the Task recording:

$$R(t) = \sum_{i<j} v_{Task}(i)v_{Task}(j)Z_i^{Sleep}(t)Z_j^{Sleep}(t)$$

where $v$ denotes the first PC. Moreover, they derive a formula which allows one to estimate the total amount of replay in a session:

$$M = \sum_{i<j} C_{ij}^{Sleep} C_{ij}^{Task} = \frac{1}{B} \sum_{t} \sum_{\mu} \lambda_{\mu} \sum_{i<j} v_{\mu}^{Task}(i)v_{\mu}^{Task}(j)Z_i^{Sleep}(t)Z_j^{Sleep}(t) =$$

$$= \frac{1}{B} \sum_{t} \sum_{\mu} \lambda_{\mu} R^{(\mu)}(t)$$

$M$ represents the correlation similarity (‘match’) between Task and Sleep and takes into account the contributions of all eigenvectors with a weight equal to their eigenvalues: contributions will therefore be very small for the lowest-eigenvalue components. However, it has been shown [206, 207] that the signal part of the spectrum comprises not only the high-eigenvalue components ($\lambda \gg 1$) but also the low-eigenvalue ones ($\lambda \ll 1$).

Our definition of replay (3.12) is similar, however it includes contributions from both high- and low-eigenvalue components, as illustrated below.

If we approximate the non-rescaled Ising couplings $J_{ij}$ in eq. 3.11 with their naïve Mean Field values, we derive the following equivalence:

$$\hat{J}_{ij} \simeq \sqrt{f_i(1-f_i)} J_{ij}^{MF} \sqrt{f_j(1-f_j)} = \sqrt{f_i(1-f_i)(-C^{-1})_{ij} f_j(1-f_j)} =$$

$$= - (\Gamma^{-1})_{ij}$$

$$\sum_{i<j} \hat{J}_{ij} \left( \langle Z_i(t) Z_j(t) \rangle_{SWSPost} - \langle Z_i(t) Z_j(t) \rangle_{SWSPre} \right) =$$

$$= \sum_{i<j} \hat{J}_{ij} \left( \Gamma_{ij}^{SWSPost} - \Gamma_{ij}^{SWSPre} \right)$$
3.10. QUANTITATIVE ESTIMATES OF THE REPLAY

where $C$ is the matrix of the connected correlations between the non-rescaled Ising variables $\sigma_i$ and $\Gamma$ is the Pearson correlation matrix (3.14). Decomposing $\Gamma^{-1}$ into its eigenvectors, we obtain from eq. 3.12:

$$R(t) = \sum_{i<j} (-\Gamma^{-1})_{ij}^{\text{Task}} Z_i^{\text{Sleep}}(t) Z_j^{\text{Sleep}}(t) = \sum_{i<j} \sum_{\mu} \left(1 - \frac{1}{\lambda_\mu}\right) v_\mu^{\text{Task}}(i)v_\mu^{\text{Task}}(j) Z_i^{\text{Sleep}}(t) Z_j^{\text{Sleep}}(t) = \sum_{\mu} \left(1 - \frac{1}{\lambda_\mu}\right) R^{(\mu)}(t)$$

(3.18)

where $R^{(\mu)}(t)$ has the same functional form as the reactivation (eq. 3.15) of the $\mu$th principal component defined by Peyrache et al. However, eq. 3.18 expresses the total amount of reactivation at time $t$ as the sum of contributions coming from all the components $\mu = 1, \ldots, N$, each one with a proper weight: the noisy components, which lie within the support of the Marchenko-Pastur distribution, are concentrated around $\lambda = 1$ and will give contributions $\approx 0$ to $R(t)$ (due to the prefactor $1 - 1/\lambda_\mu$); on the contrary, the signal part of the spectrum, comprising both the high-eigenvalue components ($\lambda \gg 1$) and the low-eigenvalue ones ($\lambda \ll 1$), will contribute significantly to the total reactivation $R(t)$ [206, 207]. In particular, large positive contributions will come from neurons having large entries with the same sign in $\lambda \gg 1$ components and from neurons having large entries with opposite sign in $\lambda \ll 1$ components.

It is worth noting that the rescaled Ising couplings $\hat{J}_{ij}$ in the naive Mean Field approximation coincide with the couplings of a Gaussian model [214], in which the Z-score variables are assumed to be drawn from a Gaussian distribution, with zero mean and unit variance:

$$\hat{P}^G(Z_{it}) = \sqrt{\det(\text{Id} - J^G)} \frac{(2\pi)^{N/2}}{\exp\left(\frac{1}{2} \sum_{i,j} J_{ij}^G Z_i Z_j - \frac{1}{2} \sum_i Z_i^2\right)}$$

(3.19)

(Id denoting the $N$-dimensional identity matrix). Indeed, according to a well-known property of the multi-variate Gaussian measure, the covariance matrix of the Z-scores, coinciding with the Pearson correlation matrix of the non-rescaled activity variables $\sigma_i(t)$, is given by

$$\Gamma_{ij} = [\text{Id} - J^G]^{-1}$$

(3.20)

Hence, the coupling matrix $J^G$ of the Gaussian model is

$$J^G = \text{Id} - \Gamma^{-1}$$

(3.21)

which, for the non-diagonal elements, reduces to (3.17).

3.10.3 Replay as a function of time in session A

We now apply our definition, eq. 3.12, to estimate the replay $R(t)$ in each time-bin $t$ (of the Sleep epoch) in session A. Note that $Z_i^{\text{Sleep}}(t)$ and $\hat{J}_{ij}^{\text{Task}}$ depend on the time-bins chosen for the Sleep and the Task epochs, through the spiking frequencies $f_i$ in eqs. 3.10 and 3.11. Since the coactivation of multi-neuron patterns occurs on time-scales larger than the time-scale of pairwise correlations, we bin the Sleep epochs with time windows $\Delta t^{\text{Sleep}}$ of several tens of milliseconds in order to search for Task-related coactivation.
CHAPTER 3. A NEW METHOD TO STUDY CELL ASSEMBLIES

Figure 3.17: Replay $R(t)$ vs. time $t$ in the Sleep epochs (left) and histograms over the SWS periods (right) in session A. $R(t)$ is strong in SWS of Sleep Post (dark blue periods) compared to non-SWS of Sleep Post (light blue) and to Sleep Pre (top). Bin widths are $\Delta t = 10$ ms for the Task Ising model, 40 ms for the Sleep configurations.

patterns: in particular, we choose $\Delta t^{\text{Sleep}} = 40$ ms since the Replay group identified by our method has its coactivation peak on this time-scale in the Sleep Post epoch (see Fig. 3.12 b). As for the Task model, we always consider couplings $J_{ij}$ inferred from the pairwise correlations in the $\Delta t^{\text{Task}} = 10$ ms time-bin. This prescription agrees with our observation that model distributions based on short time-scale pairwise couplings are capable of predicting coactivation patterns on longer time-scales. Moreover, as pointed out in 3.3, coupling values are essentially unchanged when the time-bin is increased up to $\sim 50$ ms.

It should be noted that $\tilde{J}_{ij}^{\text{Task}}$ has still a dependence on the time-bin, which for small firing rates, is approximately linear in $\Delta t^{\text{Task}}$ (since $\tilde{J}_{ij}^{\text{Task}} \sim \sqrt{f_i f_j} J_{ij} \sim \Delta t^{\text{Task}} J_{ij}$ for small $f_i, f_j$, see eq. 3.11). To get an estimate of the replay which does not depend on the choice of the time-bin for the Task epoch, a possibility would be to study the tail of the distribution $P(t)_I = \sum_{i<j} J_{ij}^{\text{Task}} \sigma_i^{\text{Sleep}} \sigma_j^{\text{Sleep}}$, which is the interaction part of the log-likelihood of the non-rescaled Ising model: $J_{ij}^{\text{Task}}$ is robust with respect to the choice of the time-bin; however, $P(t)_I$ has a noise bulk, due to the spurious contributions coming from the non-correlated neurons with high firing rates in the Sleep epoch; nevertheless, the rare events of top replay could be extracted from the tail of the distribution $P(t)_I$. We leave this study at a future work, and we focus here on definition (3.12).

The replay $R(t)$ for session A is shown as a function of the time-bin index $t$ of Sleep Pre and Sleep Post in Fig. 3.17. Due to the huge amount of time-bins ($\sim 10^6$), which would make difficult to appreciate the full plot of $(t, R(t))$, we select randomly
3.10. QUANTITATIVE ESTIMATES OF THE REPLAY

Figure 3.18: Scatter plot of probabilities that cells are active in Sleep Post vs. their counterparts across top replay events (top 0.05%) for session A. The five neurons (red) in the potentiated-Replay group are particularly active during top replay events, compared to the other periods of Sleep Post. Time-bin width $\Delta t^{\text{Sleep}} = 40$ ms.

3.10.4 Session-wide replay

Session-wide replay (3.13) is plotted in Fig. 3.19 as a function of the number of recorded cells, for each one of the 97 sessions in the data set. At least for some sessions, like A and B, $\mathcal{R}$ is significantly larger than a null model estimate (red dashed lines), where replay is uniformly distributed on all pairs of cells, as explained in the next section. Strong $\mathcal{R}$ values signal the existence of strongly potentiated and densely connected cell assemblies.
Figure 3.19: Session-wide replay $R$ vs. number of recorded cells for all 97 sessions. Colors correspond to the five rats involved in the experiment. Red dashed lines show the average replay ± one standard deviation in the null model (see 3.10.5).

It is possible to assess the contribution of each neuron to the total amount of replay by studying the impact on $R$ produced by removal of the neurons one by one from the data set. The replay upon removal of neuron $i$, i.e. the replay $R^{(i)}$ we would observe if cell $i$ was not recorded, is:

$$R^{(i)} = \sum_{k<j}^N (J^{(i)}_{kj})_{\text{Task}} \left( (\Gamma^{(i)}_{kj})_{\text{SWSPost}} - (\Gamma^{(i)}_{kj})_{\text{SWSPre}} \right)$$

where $\Gamma^{(i)}$ is the reduced Pearson correlation matrix after removal of row and column $i$ and $J^{(i)}$ are the couplings re-inferred after removal of neuron $i$ from the dataset. As shown in Fig. 3.20 for sessions A, B, D (see corresponding dots in Fig. 3.19), the dependence of $R^{(i)}$ on the neuron-index $i$ may be highly non trivial and inhomogeneous. For sessions A and B, which have larger replays than the null-model estimate, we observe the presence of a few cells with large contributions to the replay. Those cells coincide with the potentiated-Replay group identified for session A by the previous analysis and with neuron 6 for session B, which is the cell supporting most of the potentiated couplings of the Replay group (see Fig. 3.7 of par. 3.4). In session D, the three neurons of the conserved group, 28, 31, 32, give respectively a positive, ~null and negative contribution to $R$, in agreement with the coactivation of the group in both Sleep Pre and Sleep Post.

3.10.5 Null model for the replay

We interpret the session-wide replay $R^{(s)}$ of a session $s$ as a sum of random variables $r_{ij}$, representing the contributions given by each pair $i, j$:

$$R^{(s)} = \sum_{i<j} r_{ij}^{(s)}$$
Figure 3.20: Session-wide replay $\mathcal{R}$ after removal of one cell in sessions A, B, D. The largest decrease of the replay for session A is obtained by removing one cell among the potentiated-Replay group; the largest decrease for session B is obtained upon removal of neuron 6; in session D, neurons 28 and 32 give opposite contributions to $\mathcal{R}$.

where the neuron indices $i, j$ run between 1 and $N^{(s)}$, the number of recorded cells in session $s$. To define a null probabilistic model of the replay, we assume that the pairwise $r_{ij}$ are independent and identically distributed Gaussian random variables, with

$$\langle r_{ij}^{(s)} \rangle = r$$

$$\langle r_{ij}^{(s)} r_{kl}^{(s')} \rangle - r^2 = \sigma^2 \delta_{s,s'} \delta_{ij,kl}$$

where $\delta$ is the Kronecker delta function. Under these assumptions, replay $\mathcal{R}^{(s)}$ is Gaussianly distributed, with mean $N_2^{(s)} r$ and variance $N_2^{(s)} \sigma^2$ where $N_2 = N(N - 1)/2$ is the number of neuron pairs. Hence, the log-likelihood of the $S = 97$ experimental sessions according to the null model is:

$$\log P = - \sum_s \left( \frac{(\mathcal{R}^{(s)} - r N_2^{(s)})^2}{2 N_2^{(s)} \sigma^2} + \frac{1}{2} \log(2\pi N_2^{(s)} \sigma^2) \right)$$

Maximizing eq. 3.26 with respect to $r$ and $\sigma^2$ we obtain:

$$r = \frac{\sum_{s=1}^S \mathcal{R}^{(s)}}{\sum_{s=1}^S N_2^{(s)}}, \quad \sigma^2 = \frac{1}{S} \sum_{s=1}^S \frac{(\mathcal{R}^{(s)} - N_2^{(s)} r)^2}{N_2^{(s)}} \cdot$$

Red dashed lines in Fig. 3.19 correspond to the mean $r N_2$ and to the mean $\pm$ one standard deviation $r N_2 \pm \sigma \sqrt{N_2}$. 
3.11 Insights on sessions A, B, C, D, E, F, G

As previously mentioned, a 6-digit number, RRMMDD, is attributed to each session, where RR is the rat number, MM and DD are, respectively, the month and the day in which the experiment was carried out [1].

For each session, I will show the scatter plot of the couplings in Task and Sleep Post (as in Fig. 3.6 above); the number of active neurons in the self-sustaining configurations vs. the drive $H$ for each epoch (as in Fig. 3.10); the variations $\delta \langle \sigma_i \rangle$ in the conditional averages between coexisting self-sustaining configurations at the same drive $H$; a table summarizing information about cell assemblies and groups coactivating in each epoch; and finally the CoA ratios for some of these groups. Additional information is also shown for session A.

3.11.1 Session A (181014)

When high activity and low activity self-sustaining patterns coexist for a range of values of the drive $H$, as in the Task epoch of session A for $2 \leq H \leq 2.25$ (first jump) and $2.15 \leq H \leq 2.65$ (second jump), one may wonder if the cell assembly signaled by the jump changes across the values of $H$ at which the jump is defined. The answer is that it does not change because neurons of self-sustaining patterns on the same activity level have very similar conditional averages $\langle \sigma_i \rangle$, as shown in Fig. 3.21, top panel, and therefore $\delta \langle \sigma_i \rangle$ between two activity levels remains fundamentally constant across the different $H$. So we can choose arbitrarily a value $H$ for each jump and define a cell assembly as the group of neurons with highest $\delta \langle \sigma_i \rangle$ (larger than $\simeq 0.2$ for this session) in the jump, as previously mentioned. As shown in Fig. 3.10 a & b above, two jumps in the number of active neurons in the self-sustaining configurations are found in Sleep Pre and Task, but only one cell assembly is statistically significant in each epoch. The variations $\delta \langle \sigma_i \rangle$ are indeed small ($\lesssim 0.2$) for all neurons in the first jump of Sleep Pre and Task, as is clear from Fig. 3.22, where we rank the $\delta \langle \sigma_i \rangle$ in decreasing order for each jump. Neurons belonging to both Task and Sleep Post cell assemblies (1-9-20-21-26-29, composing the ‘Replay’ group and its variation) are shown in orange, while neurons belonging to the Sleep Pre and Sleep Post cell assemblies (6-7-9-11-12, composing the ‘Sleep’ group and its variation) are in violet. The merging of the Replay and Sleep groups in the large cell assembly of Sleep Post reflects the emergence of positive couplings connecting those two groups (Fig. 3.11 above).

Neurons in cell assemblies have large variations $\delta \langle \sigma_i \rangle$ of their conditional averages between the two coexisting self-sustaining patterns. Those neurons are either newly activated across the jump, or already active in the self-sustaining configuration with lower activity and in strong interactions with other newly activated neurons. This is illustrated in Fig. 3.21, bottom panel. The bottom, right panel shows that newly activated neurons in the second jump of the Task epoch (at $H = 2.3$) coactivate with three neurons (20-21-18) already active in the low activity configuration. Those neurons, indeed, belong to As. 1 (and 20-21 are also part of the potentiated-Replay group). The bottom, left panel of Fig. 3.21 shows instead the CoA ratios of the self-sustaining patterns delimiting the first jump of the Task, indicated by the purple and green dots in Fig. 3.21, top panel, and of the newly activated neurons in that jump: neurons in the purple configuration are those with largest spiking frequencies; therefore they are activated for the smallest input strength, but they are independent (CoA$\sim 1$, purple curve); the green configuration and the group of newly activated neurons have the same weak CoA (CoA$\sim 2.5$, green and black curves.
3.11. INSIGHTS ON SESSIONS A, B, C, D, E, F, G

Figure 3.21: Session A, Task: Top: neurons of self-sustaining patterns on the same activity level have similar conditional averages $\langle \sigma_i \rangle$. Neuron conditional averages $\langle \sigma_i \rangle$ in the self-sustaining patterns indicated by colored dots are shown with the same color in the inset: conditional averages $\langle \sigma_i \rangle$ change little between configurations on the same activity level, like the blue and purple configurations (bottom panel in the inset), the three green configurations (middle panel), and the two orange ones (top panel).

Bottom: CoA ratios for the self-sustaining patterns defining the two jumps and for the newly activated neurons. Colors refer to the self-sustaining patterns shown in the top panel (values of $H$ are given in the panel titles). Left panel: the purple configuration (14-24-28-37) is made of four independent cells (CoA $\sim 1$), while neurons in the green configuration (14-24-28-37-8-20-21-23-35) are weakly coactivated (CoA $\sim 2.5$). The 'difference' between the two configurations (8-20-21-23-35) has the same weak CoA (black curve) as the green configuration. Right panel: the self-sustaining pattern with higher activity (dark orange, with 15 neurons) has a CoA (top value 1300) much larger than the CoA of neurons 1-9-10-11-16-26-29 activated in the jump (black curve). The latter cells interact with those (20 and 21 in particular) in the self-sustaining pattern with lower activity at the same $H$ (dark green in the top panel): remarkably the CoA of the high-activity pattern is larger than the product of the CoAs of the group coactivated in the jump and of the low activity pattern (CoA $\sim 4$ at peak value, not shown).
Figure 3.22: Session A: Variation $\delta \langle \sigma_i \rangle$ of the conditional averages, sorted in decreasing order, between the coexisting self-sustaining configurations indicated by colored dots in Fig. 3.10. Neurons composing a cell assembly are represented by top dots in these curves ($\delta \langle \sigma_i \rangle$ larger than $\sim 0.2$). For coexisting configurations at $H=2.7$ in Sleep Pre and $H=2.15$ in Task, all $\delta \langle \sigma_i \rangle$ are small, meaning that no cell assembly is present. Colored dots indicate the conditional average variation for the neurons composing the Replay (orange) and the Sleep (violet) groups.

In Table 3.1 we list the neurons of the cell assemblies found in session A and some relevant information about the coactivating groups of neurons. The value of $\delta \log P$ (eq. 3.9 above) for each group is indicative of the order of magnitude of the group coactivation in the data, see CoA in Fig. 3.23.

Within the Sleep Pre and Sleep Post cell assemblies we identify the Sleep group 6-7-11-12 and its variant, 6-9-11-12. Similarly, among the subgroups of the Task and Sleep Post cell assemblies, we identify the potentiated-Replay group 1-9-20-21-26 and a variant, 1-9-20-26-29.

We have tested the CoA of all 160 possible groups obtained from the Replay group upon substitution of one of the five neurons by another neuron among the 37 recorded cells. One group only is found to have a CoA as large as 1-9-20-21-26, both in Task and in SWS-Sleep Post (Fig. 3.24, cyan curve) and coincides with the variant 1-9-20-26-29, in perfect agreement with the model prediction. Other two groups among the 160 are found to have large CoA values, comparable to the one of 1-9-20-21-26: these variants are obtained by replacing 26 $\rightarrow$ 10 (stronger CoA in Task, see Fig. 3.24, blue curve), and 21 $\rightarrow$ 35 (stronger CoA in SWS-Sleep Post, see Fig. 3.24, pink curve). The group with neuron 10 is a subgroup of the Task cell assembly extracted by our model (see Table 3.1), and the group with neuron 35 is a subgroup of the Sleep Post cell assembly. The log-likelihood variations of these three groups (collected in Table 3.2) are large and close to that of the Replay group 1-9-20-21-26 ($\delta \log P=6.47$ in Task and $\delta \log P=9.29$ in Sleep Post). As a conclusion, these three variants of the Replay group are found in the spiking data and are predicted with a remarkable accuracy by the model distribution. It is worth noticing that, contrary to the two variants of the Sleep group, these three variants of the Replay group are not mutually exclusive: in particular, the CoA of 1-9-20-21-26-29-35 is very large in
### Table 3.1: Session A: List of neurons in the cell assemblies (As. 1,2,3) and in some coactivated groups therein.

Neurons in cell assemblies are listed in order of decreasing values of $\delta \langle \sigma_i \rangle$. For each group we indicate the value of the drive $H$ necessary to unveil it; for all coactivated groups, we also give the values of $\delta \log P$ and CoA (time scale and maximum value, distinguishing between SWS and nSWS=nonSWS periods). Groups in each cell assembly are listed in order of decreasing $\delta \log P$ to facilitate comparison with the CoA values.
Figure 3.23: **Session A: CoA vs. time scale of some groups identified in the cell assemblies of the three epochs.** The CoA ratios shown correspond to the groups: 6-7-11-12-13 in the Sleep Pre cell assembly As. 2, As. 1 of Task (full cell assembly), 1-9-10-16-20-21-26-29 in the Task cell assembly As. 1, 7-11-12-20-21-26 in the Sleep Post cell assembly As. 3 (see Table 3.1). All these subgroups are variants of those shown in Fig. 3.12 of par. 3.7. CoAs are shown in red, dark blue, and light blue for, respectively, Task and the SWS and non-SWS periods of Sleep Pre and Post.

Figure 3.24: **Session A: CoA of the Replay group and of its possible fluctuations.** Top panels: CoA of the Replay group 1-9-20-21-26 and of the three groups with largest CoA obtained upon substitution of one neuron (indicated by stars). Note the large error bars on the CoA of the cyan and magenta group at small time scales. Bottom panels: CoA of the 7-neuron group, which extends the Replay group 1-9-20-21-26 by adding neurons 29 and 35.
Table 3.2: Session A: Three one-neuron substitutions from the Replay group 1-9-20-21-26, leading to significantly coactivated groups.

<table>
<thead>
<tr>
<th>Substitution</th>
<th>$\delta \log P \text{ Task}$</th>
<th>$\delta \log P \text{ Sleep Post}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$21 \rightarrow 29$</td>
<td>5.48</td>
<td>7.35</td>
</tr>
<tr>
<td>$26 \rightarrow 10$</td>
<td>5.33</td>
<td>4.67</td>
</tr>
<tr>
<td>$21 \rightarrow 35$</td>
<td>5.08</td>
<td>7.77</td>
</tr>
</tbody>
</table>

Figure 3.25: Session A: CoA ratios for groups of five randomly chosen neurons, compared to the CoA for the 5-cell Replay group. Black curves: average of 1000 random samples (full), ± average over the standard deviations (dashed-dotted); red curves: CoA ratios of the Replay group, corresponding to the results shown in Fig. 3.12).
CHAPTER 3. A NEW METHOD TO STUDY CELL ASSEMBLIES

Figure 3.26: Session A: Temporal trend of the coactivation of the Replay group in the Task and Sleep Post epochs. Left: we devide the Task epoch in three equally long periods and we show the CoA ratios vs. time scale $\tau$ in the first (brown curve), second (red curve) and third (yellow curve) interval; in the last two intervals, the group coactivates at smaller time-scales ($\tau \sim 20$ ms). Right panel: CoA ratios vs. time-scale $\tau$ in the first (brown), second (red) and third (yellow) SWS periods of Sleep Post Task; in the third period, coactivation of the Replay group is significantly weaker than in the first two periods.

Further tests of the statistical significance of the CoA of the Replay group are given in Fig. 3.25. We compare the CoA of the potentiated-Replay group 1-9-20-21-26 with the CoA of groups of 5 randomly chosen neurons (among the $N = 37$ cells of session A). For each random group, we compute the CoA at various time scales $\tau$, and the standard deviation within the Poisson hypothesis explained in 3.7. The outcome is averaged over 1000 random groups, and shown with black curves in Fig. 3.25. Note that, as the standard deviation in the case of zero CoA is infinite, we discard those samples in the calculation of the standard deviations; we therefore consider only samples with non-zero CoA, and then multiply the outcome by the fraction of samples with non-zero CoA.

We observe that the CoA of the 5-cell potentiated group is much larger than the CoA for random groups of 5 neurons in the Task and Sleep Post epochs, even when statistical uncertainties are taken into account. This provides further evidence for the statistical significance of the strong coactivation of this group. In Sleep Pre the average CoA is about 2, which shows the existence of a weak positive correlation between randomly chosen cells in the data. The CoA for the 5-cell Replay group remains equal to zero. This value does not mean that the 5 cells are anti-correlated, and is, indeed, statistically compatible with the independent-cell hypothesis: the product of the spiking frequencies of the 5 cells, multiplied by the duration of the recording is much smaller than one.

In session A, we also observe that the coactivation of the Replay group becomes faster toward the end of the Task phase (red and yellow curves in the left panel of Fig. 3.26), which is in perfect agreement with Hebbian learning and seems to indicate that the real synaptic connections between the neurons in the group get stronger and stronger during learning of the rule. Interestingly, coactivation of the same Replay group during Sleep Post, which is large in the first and second SWS periods (brown and red curves, respectively, in the right panel of Fig. 3.26), becomes weaker in the last SWS period (yellow curve in the same panel), in agreement with studies showing that replay is more frequent immediately
3.11. INSIGHTS ON SESSIONS A, B, C, D, E, F, G

Figure 3.27: Session B: Scatter plot of the couplings in Sleep Post and Task. As in Fig. 3.6 of par. 3.4, blue symbols represent couplings with 95% confidence in Sleep Post and Task, whereas brown symbols are confident also in Sleep Pre. A coupling is shown by a + if it positive in Sleep Pre, by a − if it is negative or null in Sleep Pre. Couplings between cells of the potentiated group (3-4-6-10) are shown within red circles.

after an experience and decays with time [22]. Unfortunately, we do not find systematically this temporal trend for other sessions: however, session A is the only session of the dataset in which a Replay group is sampled and the rule to be learned has just been changed (that is the rat is starting learning a new rule). It would be interesting to test this prediction on a larger dataset.

3.11.2 Session B (200208)

Session B has only 10 recorded neurons, among which 6 pyramidal cells and 2 interneurons. Despite the small number of neurons, it is characterized by a large coupling adjustment (Fig. 3.8 in par. 3.4 and table in par. 3.5). From the scatter plot of the couplings (Fig. 3.27), we identify an effectively potentiated group 3-4-6-10. The same neurons also show large $\delta\langle\sigma_i\rangle$ in correspondence of the jump between the pattern with low activity and the pattern with high activity in Task and in Sleep Post (Fig. 3.28 and 3.29), but not in Sleep Pre, where no cell assembly is found. Indeed the group 3-4-6-10 is found to coactivate in Task and in Sleep Post, with larger CoA in SWS (Fig. 3.30), thus representing a Replay group for this session. See Table 3.3 for additional information.
Figure 3.28: **Session B: Simulations of the model.** Number of active neurons in the self-sustaining configurations of the model activity distribution as a function of the drive $H$, for all epochs. Coexistence of distinct self-sustaining configurations (at a given $H$) with different levels of activity (colored dots) defines each cell assembly (As.) indicated by an arrow. The changes in the conditional averages $\delta \langle \sigma_i \rangle$ corresponding to the appearance of the assemblies are given in insets. Neurons of the potentiated, Replay group (3-4-6-10) (indicated in orange) show a large $\delta \langle \sigma_i \rangle$ in both Task and Sleep Post.
3.11. INSIGHTS ON SESSIONS A, B, C, D, E, F, G

Figure 3.29: Session B: Variation $\delta \langle \sigma_i \rangle$ of the conditional averages for all neurons (in decreasing order of $\delta \langle \sigma_i \rangle$) at the values of $H$ corresponding to the appearance of the cell assemblies (As. 1 and 2) in Fig. 3.28. Neurons composing each cell assembly are represented by the top dots of each curve ($\delta \langle \sigma_i \rangle$ larger than $\sim 0.1$). Orange dots indicate the conditional average variation for the neurons composing the potentiated-Replay group, coinciding with As. 1 and 2 for this session.

**Table 3.3:** Session B: List of neurons in the cell assemblies of Fig. 3.28 (As. 1 and 2) and in the coactivated group found within As. 2. The coactivated subgroup of As. 2 coincides with As. 1 and with the potentiated-Replay group for this session. Values of the drive $H$, $\delta \log P$ and CoA (time scale and max) are indicated. See caption of Table 3.1 for more details.

Figure 3.30: Session B: CoA of the potentiated, Replay group 3-4-6-10 vs. time scale in Task (red), SWS (blue) and non-SWS (light blue) periods of Sleep Pre and Post. This group coincides with As. 1 in Task, and with a coactivated group found in the Sleep Post cell assembly, As. 2.
Figure 3.31: **Session C: Scatter plot of couplings \( J_{ij} \) in the Task and Sleep Post epochs.** Same color code as in Fig. 3.27. Red circles indicate couplings between the neurons of the potentiated group (9-29-45), whereas red squares points to couplings between those neurons and the ones of the inhibited group (12-40-42-44). Note that the minus signs in the squares usually corresponds to zero couplings in Sleep Pre. Moreover, several other couplings are potentiated, but they correspond to isolated pairs.

### 3.11.3 Session C (181021)

Session C has 45 recorded neurons, among which 28 are pyramidal cells and 6 are interneurons. Its coupling adjustment \( \text{Adj} \) is 1.6 standard deviation larger than the null model average (Fig. 3.8 in par. 3.4 and table in par. 3.5). As shown in Fig. 3.31 (red circles) and Fig. 3.7 in 3.4, the couplings between neurons 9-29-45 are negative or null in Sleep Pre and become positive in Task and in Sleep Post. Moreover, many couplings between the subgroups 9-29-45 and 12-40-42-44 are depressed from Sleep Pre to Sleep Post (note that cell 29 is an interneuron): the large \( \text{Adj} \) is therefore a result of both potentiation and depression.

A substantial reconfiguration of cell assemblies is observed in session C from Sleep Pre to Sleep Post: As. 1 in Sleep Pre contains a group that strongly coactivates in non-SWS Pre (Fig. 3.34) but has weaker CoA in Task and is not coactivated in Sleep Post, in agreement with the model simulations. A Sleep group, common to As. 1 in Sleep Pre and As. 4 in Sleep Post, coactivates in the two Sleep epochs but not significantly in Task. The potentiated group 9-29-45 is a Replay group: it has high \( \delta(\sigma_i) \) in Task and in Sleep Post.
3.11. INSIGHTS ON SESSIONS A, B, C, D, E, F, G

Figure 3.32: **Session C: Simulations of the model.** Number of active neurons in the self-sustaining configurations of the model activity distribution as a function of the drive $H$, for all epochs. Coexistence of distinct self-sustaining configurations (at a given $H$) with different levels of activity (colored dots) defines each cell assembly (As.) indicated by an arrow. The changes in the conditional averages $\delta \langle \sigma_i \rangle$ corresponding to the appearance of the assemblies are given in insets. Neurons of the potentiated, Replay group 9-29-45 (indicated in orange) show a large $\delta \langle \sigma_i \rangle$ in both Task and Sleep Post; neurons in the Sleep group 2-5-7-20 (in violet) show a large $\delta \langle \sigma_i \rangle$ in both sleep epochs; neurons of the inhibited group (12-40-42-44) have negative $\delta \langle \sigma_i \rangle$ both in Task and in Sleep Post.
Chapter 3. A New Method to Study Cell Assemblies

Figure 3.33: Session C: Variation $\delta \langle \sigma_i \rangle$ of the conditional averages for all neurons (in decreasing order of $\delta \langle \sigma_i \rangle$) at the values of $H$ corresponding to the appearance of the cell assemblies (As. 1 to 4) in Fig. 3.32. Neurons composing each cell assembly are represented by the first dots along each curve ($\delta \langle \sigma_i \rangle$ larger than $\sim 0.45$). Orange and violet dots indicate the conditional average variations for the neurons composing the Replay group (9-29-45) and the Sleep group (2-5-7-20) respectively.

Table 3.4: Session C: List of neurons in the cell assemblies of Fig. 3.32 (As. 1 to 4) and in the coactivated groups found within them. Values of the drive $H$, $\delta \log P$ and CoA (time scale and max) are indicated for all coactivated groups. See caption of Table 3.1 for more details.
3.11. INSIGHTS ON SESSIONS A, B, C, D, E, F, G

Figure 3.34: Session C: CoA of some groups identified in the assemblies of the three epochs vs. time scale. CoA ratio of the coactivated group found within As. 1 of Sleep Pre (2-6-8-10-11), of the full As. 2 of Task (1-6-8-9-10-29-31-45) and of the coactivated group in As. 4 of Sleep Post (9-10-11-18-20-29); of the potentiated, Replay group (9-29-45) and of the Sleep group (2-5-7-20), in all epochs (Task: red; SWS Pre and Post: blue; non-SWS Pre and Post: light blue).

(Fig. 3.32&3.33), is coactivated in Task and in SWS-Sleep Post but not in Sleep Pre. As in session A, the Replay and Sleep groups merge in the assembly of Sleep Post. Coherently with the depression of couplings, neurons 12-40-42-44 have negative $\delta \langle \sigma_i \rangle$ coinciding with the coactivation of 9-29-45 in Task and Sleep Post; Fig. 3.35 shows this inhibition in the data.

The two assemblies found in Task (Fig. 3.32) are very similar (see also Fig. 3.33). More information about assemblies and coactivated groups found for this session is given in Table 3.4 and Fig. 3.34.
Figure 3.35: Session C: Ratio of the CoA of 9-29-45-12 (first row), 9-29-45-40 (second row), 9-29-45-42 (third row), 9-29-45-44 (fourth row) over the CoA of 9-29-45 in all three epochs vs. time scale. As neurons 12, 40, 42, 44 are inhibited by the group 9-29-45, the observed ratios are smaller than 1 in Task and $\lesssim 1$ in Sleep Post (taking into account the error bars).
3.11. INSIGHTS ON SESSIONS A, B, C, D, E, F, G

3.11.4 Session D (150720)

Session D has 34 recorded neurons, among which 22 pyramidal cells and 2 interneurons have been identified. As pointed out in Fig. 3.8, this session shows no effective potentiation or depression \( (\text{Adj} \sim \text{null-model average}) \). In agreement with this observation, the interaction network is characterized by three conserved and positive couplings between three neurons (28-31-32), see Fig. 3.36, dark '+' in the upper-right part of the plot, and Fig. 3.7. The same group emerges from the model simulation in the presence of drive \( H \) (Figs. 3.37, 3.38 and Table 3.5) and is found to coactivate in all epochs, with comparable values in the non-SWS and in the SWS periods of the Sleep epochs (Fig. 3.39).
Figure 3.37: Session D: Simulations of the model. Number of active neurons in the self-sustaining configurations of the model activity distribution as a function of the drive $H$, for all epochs. Insets: changes in the conditional averages $\delta \langle \sigma_i \rangle$ corresponding to the jumps indicated by the arrows. Only three assemblies out of 5 jumps are found to be significant on the basis of $\delta \langle \sigma_i \rangle$, and coincide with the group 28-31-32 (indicated in green) in every epoch. No Replay or Sleep group has been found in this session.
3.11. INSIGHTS ON SESSIONS A, B, C, D, E, F, G

Figure 3.38: Session D: Variation $\langle \sigma_i \rangle$ of the conditional averages for all neurons (in decreasing order of $\langle \sigma_i \rangle$) at the values of $H$ corresponding to the jumps of Fig. 3.37. Neurons composing each cell assembly are represented by top dots in these curves ($\langle \sigma_i \rangle$ larger than ~ 0.4). The three significative assemblies unveiled at $H = 2.6$ in Sleep Pre, $H = 2.7$ in Task and $H = 3.1$ in Sleep Post of Fig. 3.32 coincide with the conserved group 28-31-32 (green dots), while the jumps found at $H = 2.35$ in Sleep Pre and $H = 2.85$ in Sleep Post have very small $\langle \sigma_i \rangle$ values for all neurons and do not represent cell assemblies.

<table>
<thead>
<tr>
<th>Epoch</th>
<th>Drive H</th>
<th>Name</th>
<th>Neurons</th>
<th>$\delta \log P$</th>
<th>CoA $\tau$ (ms)</th>
<th>CoA Max. value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep Pre</td>
<td>2.6</td>
<td>As. 1 (conserved group)</td>
<td>31-32-28</td>
<td>4.41</td>
<td>10(SWS)</td>
<td>42±13 (SWS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20(nonSWS)</td>
<td>35±3.5(nonSWS)</td>
</tr>
<tr>
<td>Task</td>
<td>2.7</td>
<td>As. 2 (conserved group)</td>
<td>31-32-28</td>
<td>4.38</td>
<td>10</td>
<td>25±3</td>
</tr>
<tr>
<td>Sleep Post</td>
<td>3.1</td>
<td>As. 3 (conserved group)</td>
<td>31-32-28</td>
<td>3.82</td>
<td>10(SWS)</td>
<td>39±22(SWS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30(nonSWS)</td>
<td>8±1(nonSWS)</td>
</tr>
</tbody>
</table>

Table 3.5: Session D: List of neurons (in order of decreasing $\langle \sigma_i \rangle$) in the significant cell assemblies of Fig. 3.37, As. 1 to 3, which represent a unique conserved coactivated group. Values of the drive $H$, $\delta \log P$ and CoA (time scale and max) are indicated for each epoch.

Figure 3.39: Session D: CoA of the conserved group 28-31-32 (coinciding with As. 1, 2 and 3) in all three epochs vs. time scale.
Figure 3.40: **Session E: Scatter plot of the couplings.** Same color code as in Fig. 3.27. Coupling $J_{14,18}$, equal to -8 in the Sleep Pre epoch, is very strongly potentiated. Coupling $J_{7,11}$ is also potentiated, as its value in Sleep Pre is weakly positive, $J \approx 0.3$, and much smaller than its value in Sleep Post, $J \approx 2.9$. Although there is not a Replay group, the presence of two isolated potentiated couplings increases the coupling adjustment $Adj$ of this session.

### 3.11.5 Session E (190228)

This session has 27 recorded neurons, among which 15 are pyramidal cells, and 3 are interneurons. Session E is characterized by a large coupling adjustment, see Fig. 3.8. As is evident in Fig. 3.40 there are two largely potentiated couplings, corresponding to pairs 14-18 and 7-11. However, those couplings are isolated and do not form an interconnected, potentiated group. No significant jump in the number of active neurons composing the self-sustaining patterns is observed in any experimental epoch. As shown in Fig. 3.41, neurons progressively activate as the drive is increased, at values of $H$ depending on their local input $h_i$. Therefore, no Replay group emerges from the model simulations in the presence of the drive.
Figure 3.41: Session E: Simulations of the model. Number of active neurons in the self-sustaining configurations of the model activity distribution vs. drive $H$ in the three epochs.
Figure 3.42: Session F: Scatter plot of couplings $J_{ij}$ in the three epochs. Same color code as in Fig. 3.27.

3.11.6 Session F (181012)

Among the 44 neurons of session F, three cells (23, 30, 43) were identified as interneurons. Session F bears a strong similarity with session A, though a conserved group (8-9-15) is also observed here, together with a Replay and a Sleep groups (Fig. 3.43): the jump associated with the conserved group takes place at a large value of $H$ in Sleep Post (4.35, not shown). As shown in Fig. 3.42 many couplings are potentiated in Sleep Post with respect to Sleep Pre, and several effectively potentiated couplings connect neurons in the group 2-17-21-24-26. This large potentiated group coincides with the Replay group found within the Task and Sleep Post cell assemblies (Figs. 3.43, 3.44 and Table 3.6). The Replay group and the Sleep group merge in the same cell assembly in Sleep Post (As. 4), as already found for session A and C. Neurons of the Replay group are coactivated in the data during Task and SWS Sleep Post, but never coactivate in Sleep Pre and in non-SWS Sleep Post (Fig. 3.46). In agreement with predictions, neurons in the Sleep group have basically the same CoA in Sleep Pre and Sleep Post, while they are almost independent in Task. As in session A, the Replay group has some possible variants, such as 4-12-21-24 (see Fig. 3.45). All four neurons have large $\delta\langle \sigma_i \rangle$ in Task (and are contained in As. 2); three of them have large $\delta\langle \sigma_i \rangle$ in Sleep Post too (and are contained in As. 4), see Table 3.6.
Figure 3.43: **Session F: Simulations of the model.** Number of active neurons in the self-sustaining configurations of the model activity distribution vs. drive $H$ in the three epochs. Insets: changes in the conditional averages $\delta \langle \sigma_i \rangle$ corresponding to the jumps (arrows) between low activity and high activity patterns. The first jump in Sleep Pre is not significant due to very low $\delta \langle \sigma_i \rangle$ values; the other jumps define cell assemblies As. 1 to 4. Green labels: neurons of the conserved group 8-9-15; orange labels: neurons of the Replay group 2-17-21-24-26; violet labels: neurons of the Sleep group 16-20-24-33; note that neuron 24 belongs to both the Replay and the Sleep groups.
Figure 3.44: Session F: Variation $\delta \langle \sigma_i \rangle$ of the conditional averages for all neurons (in decreasing order of $\delta \langle \sigma_i \rangle$) at the values of $H$ corresponding to the jumps of Fig. 3.43. Neurons composing each cell assembly are represented by the top dots of each curve ($\delta \langle \sigma_i \rangle$ larger than $\sim 0.2$). The jump at $H = 3.15$ in Sleep Pre is not significant as all $\delta \langle \sigma_i \rangle$ values are small. $\delta \langle \sigma_i \rangle$ for neurons of the Replay, Sleep and conserved groups are given by the orange, violet and green dots, respectively.

Figure 3.45: Session F: CoA vs. time scale of some coactivated groups found in the cell assemblies of the three epochs. CoA of a group within As. 1 of Sleep Pre (3-8-15-21), a group within As. 2 (4-12-21-24) and another one within As. 3 (15-9-42) of Task, and of a group in As. 4 of Sleep Post (2-16-17-21-26).
Table 3.6: Session F: List of neurons in the significant cell assemblies of Fig. 3.43 (As. 1 to 4) and in the coactivated groups found within them. Values of the drive $H$, $\delta \log P$ and CoA (time scale and max) are indicated for all coactivated groups. See caption of Table 3.1 for more details.

Figure 3.46: Session F: CoA vs. time scale for the Replay, Sleep and conserved groups.
Figure 3.47: Session G: Scatter plot of couplings $J_{ij}$ in the three epochs. Same color code as in Fig. 3.27.

### 3.11.7 Session G (200209)

Session G (200209) has only 12 recorded neurons, all of which are interneurons. The behavior of session G is similar to that of session B, with the presence of coupling potentiation (high $Adj$), despite the little number of neurons, and basically the same coactivated group in Task and in Sleep Post, representing a Replay group and coinciding with the potentiated one: see Figs. 3.47, 3.48, 3.49, 3.50 and Table 3.7.
Figure 3.48: **Session G: Simulations of the model.** Number of active neurons in the self-sustaining configurations of the model distribution vs. drive $H$ in the three epochs. Insets: changes in the conditional averages $\delta \langle \sigma_i \rangle$ corresponding to the jumps (orange labels: neurons of the Replay group).
Figure 3.49: Session G: Variation $\delta \langle \sigma_i \rangle$ of the conditional averages for all neurons (in decreasing order of $\delta \langle \sigma_i \rangle$) at the values of $H$ corresponding to the jumps of Fig. 3.48. Neurons composing cell assemblies are represented by top dots in these curves ($\delta \langle \sigma_i \rangle$ larger than $\sim 0.1$). The jump at $H=4$ in Sleep Pre and the one at $H=3.25$ in Sleep Post are not significant as all $\delta \langle \sigma_i \rangle$ values are small. $\delta \langle \sigma_i \rangle$ for neurons of the Replay group are located by the orange dots.

<table>
<thead>
<tr>
<th>Epoch</th>
<th>Drive $H$</th>
<th>Name</th>
<th>Neurons</th>
<th>$\delta \log P$</th>
<th>CoA $\tau$ (ms)</th>
<th>CoA Max. value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>2.55</td>
<td>As. 1</td>
<td>3-12-7-8-1</td>
<td>3.45</td>
<td>25</td>
<td>8.1±0.5</td>
</tr>
<tr>
<td>Task</td>
<td>2.55</td>
<td>Replay group</td>
<td>3-7-12</td>
<td>2.24</td>
<td>15</td>
<td>4.6±0.1</td>
</tr>
<tr>
<td>Sleep Post</td>
<td>3.95</td>
<td>As. 2</td>
<td>3-12-7</td>
<td>2.69</td>
<td>10(SWS)</td>
<td>21±6(SWS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Replay group)</td>
<td></td>
<td></td>
<td>25(nonSWS)</td>
<td>7±1(nonSWS)</td>
</tr>
</tbody>
</table>

Table 3.7: Session G: List of neurons in the significant cell assemblies shown in Fig. 3.48 (As. 1 and 2). Values of the drive $H$, $\delta \log P$ and CoA (time scale and max) are indicated.

Figure 3.50: Session G: CoA vs. time scale of the Task cell assembly As. 1 (1-3-7-8-12) and of the Replay group (3-7-12). The Replay group coincides with the Sleep Post cell assembly As. 2 for this session.
Chapter 4

Simulations of the models in the presence of noise

Results of the previous chapter show that cell assemblies can be unveiled through the study of the locally most probable configurations of neuronal activity, or self sustaining patterns, which are uncovered through a gradient ascent (Glauber) dynamics in the log-likelihood landscape at $T = 0$, that is in the absence of noise.

In this chapter, we include noise into the analysis, by taking into account the thermal fluctuations around those local maxima. We perform Markov Chain Monte Carlo simulations at $T = 1$ of the inferred neural networks and we study the response of these virtual networks upon increasing the external drive: synchronously activated (inhibited) neurons emerge as the neurons with a large increase (decrease) in simulated activity at close values of the drive, when the system is at equilibrium. This is a more rigorous approach than that illustrated in chapter 3, since real neuronal networks are noisy: stochasticity might arise from e.g. neuron ion channels, synaptic transmission and inputs from the surrounding environment (including non recorded neurons). We compare the results obtained in the presence of such fluctuations with those extracted at $T = 0$ for three representative sessions, called A, B and D in the previous chapter, recorded from three different rats. To test the stability of our method we slightly change the definition of the start and end times of the epochs, which is somewhat arbitrary, and we compare the results obtained at $T = 0$ for these weakly modified datasets with those of chapter 3. We use the models inferred from these slightly different datasets for all the analysis of this chapter. We also treat the case in which the drive is not global and uniform but localized on specific pairs of neurons, favoring or disfavoring cell assembly activation according to the targeted pair. Finally, we discuss in detail the significance of the external drive in the simulations of the neural networks, and more generally the meaning and potentiality of this method.

4.1 Results at $T = 0$ on slightly modified datasets

Results of the analysis at $T = 0$ on the slightly modified datasets are summarized in Figs. 4.1 4.2 and Table 4.1 for session A and in Tables 4.2, 4.3 for sessions B and D, respectively.

Cell assemblies and coactivated groups are very similar to those obtained in chapter 3 with a slightly different set of inferred couplings and fields: the few differences in the identified cell assemblies are represented mainly by neurons that have $\delta \langle \sigma_i \rangle > \text{th.}$ in one set and $\delta \langle \sigma_i \rangle \lesssim \text{th.}$ in the other set and which could be included in the cell assembly
Figure 4.1: Session A: Simulations of the models inferred for the slightly modified datasets. Number of active neurons in the self-sustaining patterns of the distribution of activity configurations as a function of the drive $H$. Variations of the neuron conditional average activities between the low and the high activity configurations corresponding to each jump are shown in insets. Red labels indicate neurons of the Replay group.
### 4.1. Results at $T = 0$ on slightly modified datasets

<table>
<thead>
<tr>
<th>Epoch</th>
<th>Drive $H$</th>
<th>Name</th>
<th>Neurons</th>
<th>$\delta \log P$</th>
<th>CoA $\tau$ (ms)</th>
<th>CoA Max. value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>3</td>
<td>As.1</td>
<td>12-2-6-11-12-13</td>
<td>8.8</td>
<td>105 (nonSWS)</td>
<td>(5.1 ± 1.5)·$10^2$ (nonSWS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sleep group</td>
<td>6-7-11-12</td>
<td>3.5</td>
<td>30 (SWS)</td>
<td>27 ± 13 (SWS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sleep group</td>
<td>2-6-7-11-12-13</td>
<td>7.5</td>
<td>60 (4 ± 0.9)·$10^2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sleep group</td>
<td>2-5-6-11-12-13</td>
<td>7.5</td>
<td>105 (1.4 ± 0.7)·$10^2$</td>
<td></td>
</tr>
<tr>
<td>Task</td>
<td>2.4</td>
<td>As. 2</td>
<td>26-9-1-10-11-12-13</td>
<td>11.6</td>
<td>40</td>
<td>(8 ± 4)·$10^2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inhib. by As. 2</td>
<td>26-9-10-21-26</td>
<td>6.5</td>
<td>20</td>
<td>73 ± 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>potentiated-Replay group</td>
<td>1-9-20-21-26</td>
<td>5.5</td>
<td>20</td>
<td>63 ± 20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replay group</td>
<td>1-9-10-20-21-26</td>
<td>5.4</td>
<td>25</td>
<td>77 ± 19</td>
</tr>
<tr>
<td>Post</td>
<td>2.6</td>
<td>As. 3</td>
<td>26-10-12-9-6-11-12-13</td>
<td>15.5</td>
<td>60</td>
<td>(9.8 ± 4.4)·$10^4$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>group in As. 3</td>
<td>7-8-10-11-12-13-20-21-26-35</td>
<td>12.3</td>
<td>120</td>
<td>96 ± 48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>group in As. 3</td>
<td>7-8-10-11-12-13-20-21-26-35</td>
<td>9.2</td>
<td>35 (SWS)</td>
<td>(1.6 ± 0.6)·$10^3$(SWS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replay group</td>
<td>1-9-20-21-26</td>
<td>7.5</td>
<td>25 (SWS)</td>
<td>(2.2 ± 1.1)·$10^3$(SWS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replay group</td>
<td>1-9-20-21-26</td>
<td>7.3</td>
<td>25 (SWS)</td>
<td>(3.6 ± 1.6)·$10^3$(SWS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replay group</td>
<td>1-9-20-21-26</td>
<td>5.3</td>
<td>30</td>
<td>(5 ± 2)·$10^4$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>group in As. 3</td>
<td>5-6-7-8-11-12-13-20-21-26</td>
<td>5.2</td>
<td>60 (SWS)</td>
<td>(1.1 ± 0.6)·$10^3$(SWS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replay group</td>
<td>1-9-10-20-21</td>
<td>4.2</td>
<td>55 (SWS)</td>
<td>(5 ± 2)·$10^3$(SWS)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sleep group</td>
<td>6-7-11-12</td>
<td>2.2</td>
<td>40 (SWS)</td>
<td>32 ± 8 (SWS)</td>
</tr>
</tbody>
</table>

Table 4.1: Session A: Summary table on the analysis at $T = 0$ (slightly modified datasets). We list neurons in the cell assemblies (in order of decreasing $\delta \langle \sigma_i \rangle$) and in some coactivated groups therein, and neurons inhibited by those cell assemblies (in order of increasing $\delta \langle \sigma_i \rangle$). For each group we indicate the value of the drive $H$ necessary to unveil it; for all coactivated groups, we also give the values of $\delta \log P$ and CoA (time scale and maximum value) in the respective epochs (for the replay and sleep groups we also distinguish between SWS and nonSWS periods). Groups in each cell assembly are listed in order of decreasing $\delta \log P$ to facilitate comparison with the CoA values.
Figure 4.2: Session A: Networks of couplings $J_{ij}$ between neurons of the Replay group and neurons of the Sleep group for each epoch. Red (blue) lines indicate positive (negative) couplings and the line thickness is proportional to $|J_{ij}|$. Couplings within the Replay group undergo a strong potentiation from Sleep Pre to Sleep Post, while several couplings within the Sleep group are strong in the two Sleep epochs and absent in the Task.

Table 4.2: Session B: Summary table on the analysis at $T = 0$ (slightly modified datasets). See caption of Table 4.1 for more details.

Table 4.3: Session D: Summary table on the analysis at $T = 0$ (slightly modified datasets). See caption of Table 4.1 for more details.
4.1. RESULTS AT $T = 0$ ON SLIGHTLY MODIFIED DATASETS

4.1.1 Non-uniform inputs in the model of Sleep Post activity

Fig. 4.3 shows that specific and partial stimulation can facilitate the activation of the 5-cell Replay group in Sleep Post of session A. The procedure we follow is equivalent to adding a non-uniform drive, very strong on a pair of cells within the Replay group, and homogeneous (equal to $H$) on the other neurons. Not surprisingly, with pairs of neurons already on, the drive $H$ to be applied on the other neurons to activate the Replay group is lowered with respect to the value found with a homogeneous stimulation, $H \simeq 2.6$, see Fig. 4.1, right. However, we observe that this value varies considerably with the specific pair of neurons which is targeted. For instance, $H \simeq 1.5$ is sufficient to induce coactivation of the Replay group when pair 9-26 is already on, while $H \simeq 2.5$ is necessary with pair 20-21. In general, stimulation of pairs chosen among 1-9-26 has the largest effect, with pairs 9-26 and 1-26 producing a quite synchronous response of the other three neurons in the Replay group at the lowest $H$ values (1.5 and 1.75 respectively) and pair 1-9 producing the most synchronous response. This result is coherent with the fact that neurons 1, 9 and 26 support the strongest and positive couplings in the group (see Fig. 4.2, right), and are slightly lowering th. (th. threshold on $\delta(\sigma_i)$ defining the cell assembly). Neither the Replay groups nor the conserved group of session D change in the two sets; the Sleep group of session A is slightly enlarged, including neurons 2 and 13.

In section 4.2.2 these results will be compared with the analysis at $T = 1$. 

Figure 4.3: Number of active neurons in the self-sustaining patterns of the Sleep Post model of session A with specific stimulation. Two targeted cells of the Replay group (indicated in panel titles) are forced to be active (very strong specific input), while all the other neurons receive an increasing uniform drive $H$ (x axis). The non-targeted neurons of the Replay group are listed in each panel at the point of their activation.
therefore the optimal target cells to propagate the signal. On the other hand, stimulation of pair 20-21 does not favor group coactivation (which occurs at $H \simeq 2.6$, very close to the value of $H$ required in the absence of specific stimulation). Neurons 20 and 21, which have very similar firing rates (and $h_i$) and weaker connections in the group, respond the same way when other two neurons are kept active (see 9-26, 1-26 and 1-9 in Fig. 4.3) and induce very similar responses in the other neurons when one of them is in the targeted pair (compare 1-20 and 1-21, 9-20 and 9-21, 20-26 and 21-26 in Fig. 4.3).

In section 4.2.3, we will confirm these results through a slightly different analysis at $T = 1$.

We think that exploration of the effects produced by localized stimulations of specific neurons through simulations of the inferred neural network is a potentially useful approach which, together with the new optogenetic techniques [215–218] could help in the manipulation of cortical cell assemblies in a controlled way.

## 4.2 Simulations at $T = 1$

### 4.2.1 Description of the method

We use a Monte Carlo Metropolis algorithm with asynchronous updating to sample the Ising distribution $P(\sigma_1, \sigma_2, ..., \sigma_N)$ (eq. 3.1 of the previous chapter) inferred from each experimental epoch. We define the energy of a configuration $\sigma$ as minus its log-likelihood (up to a configuration-independent additive constant):

$$E(\sigma) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i$$

(4.1)

The algorithm starts from a randomly chosen neuronal configuration, $\sigma^0$. At each step $t$, a neuron $i$ is chosen uniformly at random among the $N$ neurons, and its state $\sigma_i^t$ in the neuronal configuration $\sigma^t$ is flipped into $\sigma_i^{t+1} = 1 - \sigma_i^t$ with probability

$$p(\text{flip}) = \begin{cases} 1 & \text{if } \Delta E < 0, \\ e^{-\Delta E} & \text{if } \Delta E > 0 \end{cases}$$

(4.2)

where $\Delta E = E(\sigma^{t+1}) - E(\sigma^t)$ is the difference between the energies of the neuronal configurations after and before the flip. As this dynamics satisfies detailed balance, the probability distribution over the neuronal configurations will eventually converge to $P(\sigma)$.

To decide when convergence has been reached we study the evolution of observables $Q(\sigma)$ (examples given below) as the algorithm runs. We define a round as a set of $N$ (number of neurons) Monte Carlo steps. The round numbers are divided into windows $[2^n, 2^{n+1}]$, where $n$ is an integer. The empirical averages $Q^{(n)}$ of the observable $Q(\sigma)$ over the sampled configurations $\sigma$ are computed in each window (one every $2^4$ Monte Carlo rounds is kept in the average in order to make the system decorrelate). The algorithm stops when $n \geq 17$ (that is more than 4000 samples are used to compute the averages) and the relative fluctuations of the averages from one window to the next one are smaller than a chosen accuracy:

$$\left| \frac{Q^{(n)} - Q^{(n-1)}}{Q^{(n-1)}} \right| < \text{accuracy}$$

(4.3)
for all the observables $Q$ under consideration. In particular we consider the observables related to:

- the population activity: $A = \frac{1}{N} \sum_i \sigma_i \quad (4.4a)$
- the single-neuron activities: $a_i = \sigma_i \quad (4.4b)$
- the pairwise activities: $a_{ij} = \sigma_i \sigma_j \quad (4.4c)$

In a reasonable computational time ($n \lesssim 22$) we can reach an accuracy of $\sim 0.001$ on the computation of $A^{(n)}$ and of $\sim 0.05$ on $a_i^{(n)}$ and $a_{ij}^{(n)}$. The global susceptibility, defined as $\sum_{ij} a_{ij}^{(n)} - a_i^{(n)} a_j^{(n)}$ (see 4.2.2) is also computed with very good accuracy ($\lesssim 0.001$).

### 4.2.2 Study of the susceptibility maxima and minima

Through Monte Carlo simulations we compute the local (single-neuron) susceptibilities $\chi_i(H)$ and the global susceptibilities $\chi(H)$ for each experimental epoch. Local susceptibilities $\chi_i(H)$ are defined as the change in the probability of firing (in a time-bin of width $\Delta t$) of neuron $i$ resulting from a small increase of the external drive $H$:

$$\chi_i(H) = \frac{\partial \langle \sigma_i \rangle}{\partial H}(H) \quad (4.5)$$

where $\langle \cdot \rangle(H)$ denotes the average over the distribution $P(\sigma)$, with an external drive $H$ added to all local inputs: $h_i \rightarrow h_i + H$. According to the fluctuation-dissipation theorem [219], responses to perturbations are equal to connected correlations:

$$\chi_i(H) = \sum_{j=1}^{N} \frac{\partial \langle \sigma_i \rangle}{\partial h_j}(H) = \sum_{j=1}^{N} \left[ \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle \right](H) \quad (4.6)$$

Formula (4.6) is useful compared to definition (4.5) as it requires to compute the values of connected correlations, which is easy to do numerically, rather than to estimate a derivative, generally much harder to do in an accurate way. Formula (4.6) can be extended to define the global susceptibility,

$$\chi(H) = \sum_{i=1}^{N} \chi_i(H) = \frac{\partial}{\partial H} \sum_{i=1}^{N} \langle \sigma_i \rangle(H) \quad (4.7)$$

which is the derivative of the average number of active neurons in the time-bin with respect to the global input strength.

Peaks in the global susceptibilities in Fig. 4.4 indicate the presence of specific values of $H$ for which the average activity abruptly increases. Those events correspond to strong coactivations, and are reminiscent of the jumps in the number of active neurons in the self-sustaining patterns at $T = 0$. Maxima of $\chi(H)$ are found for values of the drive similar to, or slightly smaller than, the range of values of $H$ at which cell assemblies are unveiled at $T = 0$ (vertical lines in Fig. 4.4): in fact thermal fluctuations can switch on a neuron at a value of $H$ slightly smaller than the value expected in the absence of fluctuations, and can eventually anticipate the coactivation event.
Figure 4.4: **Global susceptibility** \( \chi \) as a function of drive \( H \) for sessions A, B and D in all epochs. Note differences in the y-scale for the different sessions. Vertical dashed lines indicate the range of \( H \) values for which there is coexistence of two self-sustaining patterns, one with low activity and the other one with high activity, defining the cell assemblies at \( T = 0 \).

Figure 4.5: **Single-neuron susceptibilities** \( \chi_i \) as a function of drive \( H \) for sessions A, B and D in all epochs. The number of curves is equal to the number of recorded neurons for each session (respectively \( N = 37, 10, \) and 34). Red curves: susceptibilities of neurons in the cell assemblies identified at \( T = 0 \); green curves: susceptibilities of the inhibited neurons identified at \( T = 0 \) (see Tables 4.1, 4.2, 4.3).
4.2. SIMULATIONS AT $T = 1$

The global susceptibilities of session A show large peaks in all three epochs. Peak height is particularly large in Sleep Post (in agreement with the merging of different sub-groups coactivated in the previous epochs, found in the analysis at $T = 0$). In session B no significative peak is observed in the susceptibility of Sleep Pre (in agreement with the absence of jumps at $T = 0$), while susceptibilities of Task and Sleep Post have a clear maximum.

The large global peaks of sessions A and B are due to high and overlapping peaks in the single-neuron susceptibility curves of Fig. 4.5: single-neuron susceptibility maxima give detailed information on the composition of coactivated groups, while susceptibility minima indicate inhibition. Local susceptibility maxima and minima are reminiscent of the variations of the neuron conditional average activities in the jumps at $T = 0$: neurons with the highest susceptibility maxima coincide with neurons having the largest $\delta \langle \sigma_i \rangle$ in the jumps at $T = 0$ (red in Fig. 4.5), and neurons with susceptibility minima largely coincide with those having the most negative $\delta \langle \sigma_i \rangle$ in the same jumps at $T = 0$ (green in Fig. 4.5).

In session D the global susceptibility shows a bump in the three epochs - slighter than in session A despite the similar number of recorded neurons - due to the fact that few neurons (28-31-32) have overlapping susceptibility peaks (red curves in Fig. 4.5, bottom panels).

We now study single-neuron susceptibilities in more detail.

The five main types of neuron susceptibilities found in the dataset are shown in Fig. 4.6: the majority of neurons have a bell-shaped susceptibility with a maximum at a certain value of the drive (a), other neurons have a convex susceptibility with a minimum (b), or a more complex susceptibility with both a maximum and a minimum (c) or two maxima and a minimum (d), and finally some neurons have an increasing susceptibility which does not reach a stationary point for the values of $H$ considered here (e).

For each neuron $i$ with a susceptibility of any type but e, we can define

$$H_i^+ = \operatorname{argmax} \chi_i(H) \quad \text{or} \quad H_i^- = \operatorname{argmin} \chi_i(H)$$

(4.8)

corresponding to, respectively,

$$\chi_i^+ = \max \chi_i(H) \quad \text{or} \quad \chi_i^- = \min \chi_i(H)$$

(4.9)
for each maximum and (or) minimum of the susceptibility $\chi_i(H)$. When several neurons have susceptibilities with maxima (or minima) at the same (or at a similar) value of $H$, the value of each maximum is indicative of how likely it is that the corresponding neuron takes part in the coactivation event, while the value of each minimum tells how strong the inhibition of the corresponding neuron is due to the coactivation event.

More precisely, for an independent neuron with local field $h$, the average activity is

$$\langle \sigma \rangle(H) = \frac{e^{h+H}}{1 + e^{h+H}}$$

(4.10)

and the susceptibility is given by

$$\chi(H) = \frac{\partial \langle \sigma \rangle(H)}{\partial H} = \frac{e^{h+H}}{(1 + e^{h+H})^2}$$

(4.11)

the maximal value of which is 0.25 (for $H = -h$). The baseline $\chi = 0.25$ therefore represents a null value for the maximal susceptibility, corresponding to neurons essentially independent from any other neuron and coactivated group. Maxima higher than 0.25 indicate coactivating neurons; maxima lower than 0.25 have a less clear-cut interpretation, as they may indicate neurons that are weakly activated or weakly inhibited by the other neurons activating at the same $H$, depending on whether their $|h_i|$ is different or $\sim H_i^+$. Minima of single-neuron susceptibilities correspond to neurons whose average activity minimally increases (positive minima) or maximally decreases (negative minima) when $H$ increases around $H_i^-$, that is neurons that are weakly or strongly inhibited, respectively.

Fig. 4.7 shows the pairs $(H_i^+, \chi_i^+)$, upward full triangles, and $(H_i^-, \chi_i^-)$, downward empty triangles, for all recorded neurons in each epoch of sessions A, B and D (horizontal line at 0.25). In session A (top panels), orange and violet labels indicate neurons with close susceptibility maxima (above 0.25) or minima in Task and in Sleep Post, or in the two sleep epochs, respectively. Orange (violet) maxima coincide with neurons of the Replay (Sleep) group identified at $T = 0$. The orange and violet minima may be interpreted as the recorded part of cell-assemblies inhibited in, respectively, Task and Sleep Post or Sleep Pre and Sleep Post. This picture shows that the activation-inhibition effects observed in Sleep Pre and Task partially sum up in the Sleep Post epoch (as signalled by the merging of the orange and violet maxima and minima): this reflects the fact that the Replay and Sleep groups, which are negatively connected or not significantly interacting with one another in Sleep Pre and in Task, become positively connected in Sleep Post, synchronously inhibiting both the orange and violet groups in Fig. 4.7 (minima). While couplings within the Replay group are potentiated from Sleep Pre to Sleep Post, couplings between the Replay group and the group of neurons with susceptibility minima in Task and Sleep Post are not depotentiated on average: indeed they are negative also in Sleep Pre, but inhibition between the two groups is not observed in this epoch as replay neurons are not coactivated. In agreement with the analysis at $T = 0$, in session B neurons 3-4-6-10 have large and synchronized susceptibility peaks both in Task and in Sleep Post (Figs. 4.7, middle), and in session D neurons 28-31-32 have the largest susceptibility peaks in all three epochs (Figs. 4.7, bottom). In this session, some negative interactions are formed during the Task and retained in Sleep Post, producing inhibition of neurons 8-14-17-30 in these two epochs.
Figure 4.7: Maxima (upward full triangles) and minima (downward empty triangles) of single-neuron susceptibilities vs. drive $H$, for all epochs of session A (top row), B (middle row) and D (bottom row). The horizontal line at $\chi = 0.25$ shows the maximal susceptibility for independent neurons (see text). Orange (violet) labels attached to susceptibility maxima indicate neurons of the Replay (Sleep) group; orange (violet) labels attached to susceptibility minima indicate neurons inhibited in Task and Sleep Post (in Sleep Pre and Sleep Post).
4.2.3 Localized stimulation in the model of Sleep Post activity

In section 4.1.1 I have shown that adding to the uniform global drive \( H \) a strong input on a specific pair of neurons in the Replay group may facilitate the coactivation of the rest of the group if the two neurons forced to be active are sufficiently connected to the others. Here, we show that a similar result is obtained at \( T = 1 \) if we put an increasing drive \( H \) on specific pairs and no input at all on all the other neurons. This specific drive may mimic sharp waves during Slow-Wave-Sleep, when synchronous reactivation of replay neurons takes place. In the presence of this localized drive, single-neuron susceptibilities become:

\[
\chi_i(H) = \sum_{j \in \text{pair}} \frac{\partial \langle \sigma_i \rangle}{\partial h_j}(H) = \sum_{j \in \text{pair}} [\langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle] (H) \tag{4.12}
\]

where the sum runs over the two neurons in the targeted pair only.

The similarity between Fig. 4.8 and Fig. 4.3 is evident: synchronized responses in Fig. 4.3 typically correspond to synchronized responses in Fig. 4.8; moreover, the smaller the value of \( H \) needed to activate non-targeted neurons of the Replay group in Fig. 4.3, the...
4.2. SIMULATIONS AT $T = 1$

<table>
<thead>
<tr>
<th>Targeted pair</th>
<th>9, 26</th>
<th>1, 26</th>
<th>1, 9</th>
<th>21, 26</th>
<th>20, 26</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^+_{t_i}$</td>
<td>4.85, 4.8</td>
<td>4.85, 4.65</td>
<td>4.9, 4.75</td>
<td>3.3, 4.45</td>
<td>3.25, 3.95</td>
</tr>
<tr>
<td>$\chi^+_i$</td>
<td>0.3107, 0.3104</td>
<td>0.3457, 0.3453</td>
<td>0.3423, 0.3421</td>
<td>0.2809, 0.2808</td>
<td>0.3089, 0.3086</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Targeted pair</th>
<th>1, 20</th>
<th>1, 21</th>
<th>9, 20</th>
<th>9, 21</th>
<th>20, 21</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^+_{t_i}$</td>
<td>5.2, 3.35</td>
<td>5.45, 3.35</td>
<td>4.65, 3.35</td>
<td>4.65, 3.3</td>
<td>2.95, 2.9</td>
</tr>
<tr>
<td>$\chi^+_i$</td>
<td>0.267, 0.2674</td>
<td>0.2611, 0.261</td>
<td>0.2744, 0.2741</td>
<td>0.2716, 0.2716</td>
<td>0.3058, 0.3059</td>
</tr>
</tbody>
</table>

Table 4.4: Coordinates $(H^+_{t_i}, \chi^+_i)$ of susceptibility maxima for the targeted neurons in each panel of Fig. 4.8.

Figure 4.9: Crosscorrelation histograms between neurons of the Replay group in session A (1-9-20-21-26), restricted to the SWS phase of Sleep Post. Time windows of histograms: 50 ms. Range: −1 s, +1 s.

higher are the susceptibility peaks of the same neurons in Fig. 4.8. In Fig. 4.8 we see that if $H$ is applied to a pair of neurons among 26, 9, 1 a relatively strong response is elicited in all other neurons of the Replay group. Indeed, neurons 26, 9 and 1 are well connected to the others of the Replay group in Sleep Post (see Fig. 4.2, right), as mentioned in par. 4.1.1, and they also respond synchronously to drive $H$: their $H^+_{t_i}$ values are very close, see Table 4.4, which potentiates the effect of their activation on the response of the neurons interacting with them. Notice that interaction between the two neurons of the stimulated pair is not necessary to elicit the coactivation of the Replay group (e.g. 26 and 9 are not correlated in the Slow-Wave-Sleep phase of Sleep Post Task, Fig 4.9). On the other hand, stimulation of pair 20-21 does not produce any significative coactivation of the group (like in Fig. 4.3). Like in Fig. 4.3, neurons 20 and 21 behave in a very similar way both when they respond to the targeted pair and when they are in the targeted pair.
4.3 From $T = 0$ to $T = 1$

In this section I will show how neuron susceptibilities progressively change when temperature $T$ in Monte Carlo simulations increases from 0.1 to 1 and I will compare the neuron susceptibility peaks at $T = 1$ with the variations of neuron conditional averages $\delta \langle \sigma_i \rangle$ obtained at $T = 0$.

4.3.1 Neuron susceptibilities on varying $T$

Neuron susceptibilities for ten different values of $T$, ranging from $T = 0.1$ to $T = 1$ (step of 0.1) are shown in Fig. 4.10: each panel represents one neuron and each curve the susceptibility for that neuron at a fixed temperature as the global drive $H$ increases; color of the curve indicates the value of $T$, with blue denoting $T = 0.1$, red $T = 1$ and intermediate progressive colors all intermediate temperatures. Neuron susceptibilities change in a progressive way as $T$ increases. Temperature increase may produce two different effects. The first one is that new energy minima may be explored which are not reached at very low values of $T$, due to barriers separating different minima in the energy landscape; exploration of several energy minima in the Monte Carlo simulations may lead to the emergence of several susceptibility maxima and/or minima at different $H$ values: this is the case of neurons 32 and 25, for which a second susceptibility maximum (absent at very low $T$) appears at intermediate $T$ values; of neurons 14 and 24, which only at intermediate $T$ have a very pronounced susceptibility minimum; of neurons 30 and 27, which do not have any susceptibility peak at low $T$ values, but show a non-flat
4.3. FROM $T = 0$ TO $T = 1$

Figure 4.11: Scatter plot of maxima $\chi_i^+$ (upward full triangles) and minima $\chi_i^-$ (downward empty triangles) of the local susceptibilities ($T = 1$) vs. variations $\delta\langle \sigma_i \rangle$ of the neuron conditional average activities ($T = 0$). Results refer to the cell assembly activation in the Task epoch of session A. Horizontal (vertical) line is the reference for independent neurons in the analysis at $T = 1$ ($T = 0$).

behaviour vs. $H$ at higher $T$. The second effect, typically present at $T = 1$ or closer values, is the averaging over different minima in the energy landscape (when barriers between one minimum and the following one are overcome frequently in both directions); this leads to the smoothing of susceptibility peaks, like in neurons 14 and 24, and in some cases (see neurons 32 and 25) just one peak is left at $T \sim 1$. For some neurons (for example 31) no susceptibility peak is found at any $T$ at the reasonable values of $H$ shown in the figures, and for many neurons a unique susceptibility maximum is present for all $T$ values, which gets smoother and smoother as $T$ increases and the neuron activity is averaged over the noise (see neurons 9 and 21).

4.3.2 Local susceptibility peaks vs. conditional average variations

As mentioned in the previous paragraph, single-neuron susceptibilities tell how much a neuron is likely to be activated or inhibited by the synchronous activation of other neurons, in a similar but more precise way compared to the variations in the conditional average activities at $T = 0$. In Fig. 4.11 we compare values $\chi_i^+$ and $\chi_i^-$ of the susceptibility maxima and minima with the variations $\delta\langle \sigma_i \rangle$ between the self-sustaining patterns with low and high activity defining cell assemblies at $T = 0$. There is approximately a monotonic relationship between the two quantities, showing that the two analysis lead to qualitatively equivalent results. It is worth noting that the point at which the susceptibility peaks crosses the line at 0.25 (representing the reference for independent neurons) corresponds to a zero value of $\delta\langle \sigma_i \rangle$, which indicates independent neurons in the analysis at $T = 0$. Inhibition effects are weaker than coactivation effects and more subject to noise, which gives larger differences $\chi_i^- - \delta\langle \sigma_i \rangle$ compared to $\chi_i^+ - \delta\langle \sigma_i \rangle$. 
4.4 Discussion on the meaning of the method

4.4.1 Is inference of the Ising model necessary to predict rare coactivation events?

One can wonder if inference of the neuron effective interactions is really necessary to predict the relatively rare events in which a cell-assembly is activated, or if the same coactivation events can be extracted more simply from the observed distribution of activity in snapshots of 10 ms. To assess this issue we compute neuron susceptibilities (eq. 4.6) from the following distribution, obtained directly by rescaling the data distribution in snapshots of 10 ms according to the global drive \( H \), without inference of the Ising model:

\[
P_{\text{data}}(\sigma|H) = \frac{N(\sigma(10 \, \text{ms})) \cdot e^{H \sum_i \sigma_i}}{Q},
\]

(4.13)

where \( N(\sigma(10 \, \text{ms})) \) is the number of times configuration \( \sigma \) appears in the data and \( Q \) ensures normalization of the distribution. We compute neuron susceptibilities for several values of \( H \) and compare them to those obtained from Monte Carlo sampling of the Ising distribution in the presence of the same drive \( H \). Comparison is shown in Fig. 4.12 for neurons of the Replay group (top row) and other representative neurons (bottom row) recorded from session A, Sleep Post epoch: susceptibilities obtained from the rescaled data distribution (red) overlap with those obtained from the rescaled Ising distribution (black) only for very low values of the rescaling parameter \( H (H \lesssim 1.5) \). For larger values not only there is no overlap between the two curves, but their shapes often become very different: in the Replay group only neuron 20 has a bell-shaped susceptibility obtained from the data, and among the other neurons only few (like 14 and 35) have similar shaped curves. This...
Figure 4.13: Sketch of the efficacy of the method in unveiling rare configurations important for coding. Differently from the data rescaled distribution, eq. 4.13 (schematically represented by the cyan curve), the Ising distribution (schematically represented by the black curve), reconstructs accurately the tail of the ‘true’ distribution of all activity patterns potentially generated in such network. Rare (high-activity) patterns are uncovered by adding into the Ising model a global drive \( H \), which triggers, in the virtual network, avalanche processes resulting in the collective activation of multiple neurons. Inset: mean global activity as a function of drive \( H \), computed from Monte Carlo sampling of the Ising distribution (black curve), and from the data rescaled distribution (cyan curve); curves refer to the Task phase of session A. The cyan curve saturates at \( \sim 0.2 \), corresponding to the maximal activity configuration observed in the data, while the black curve keeps growing as \( H \) increases; note that at the value of \( H \) necessary to uncover coactivation of replay neurons (\( H \sim 2 \)) the two curves are already separated.
comparison highlights that coactivation events, relatively rare but important for coding, which are unveiled by adding a drive into the inferred Ising model, are not deducible at all by adding the same drive into the distribution of snapshots of recorded activity. In [220], the authors notice that the Ising model, relying on accurate sampling of pairwise correlations only (which can be reliably estimated from a relatively short data set), gives a better prediction of network activity patterns than what can be achieved by empirical sampling. Here we push this consideration further: inferring the interaction network of the data not only allows us to get an accurate estimate of the distribution of activity patterns in the conditions at which the network is inferred, but also to reconstruct activity patterns generated by this effective network at different time-scales and in the presence of a real input. Drive $H$ pushes the virtual network into a state of increased neuronal excitability and triggers avalanche processes of neuron activations that can occur in the real system: when a neuron belonging to a cell assembly increases its firing rate (e.g. in the presence of sharp waves) neurons which are strongly connected to it will also undergo a strong increase in their activity, and the signal will propagate to all neurons of the cell assembly in an avalanche process. The potentiality of the Ising model in catching these network effects is schematically illustrated in Fig. 4.13: the inferred model reconstructs the tail of the distribution of all configurations potentially observable in the system; rare, high activity configurations are accessible by adding a uniform drive $H$. In this sense, a cell-assembly activation is indeed shown to emerge from the structural properties of a pairwise interaction network.

4.4.2 What does drive $H$ represent?

As just observed, adding a global drive $H$ into the model allows us to reconstruct the tail of the configuration distribution and eventually to unveil cell-assembly activations that are confirmed by direct computation of the CoA ratios. But what can account for the drive $H$ in the real system? From CoA estimations we observe that cell-assembly activations take place at time-scales larger than 10 ms (that is the time-scale of the firing rates and pairwise correlations reproduced by the inferred Ising model). Therefore drive $H$ should represent, at least partly, a time-scale.

As mentioned in par. 3.3, for a system of independent neurons with firing rates $f_i$, we can compute with very good approximation the relation between $H$ and the time-scale: values of the fields $h_i$ in an Ising model with zero couplings are related to the empirical probabilities $p_i(\Delta t)$ that each neuron $i$ is active in a time-bin of width $\Delta t$ through

$$p_i(\Delta t) \sim f_i \cdot \Delta t = \frac{e^{h_i}}{1 + e^{h_i}} \sim e^{h_i}$$  \hspace{1cm} (4.14)$$

from which we obtain:

$$h_i \simeq \log (f_i \cdot \Delta t)$$  \hspace{1cm} (4.15)$$

for small $\Delta t$: local inputs $h_i$ vary logarithmically with $\Delta t$. Increasing the time-bin in the inference from $\Delta t^{(1)}$ to $\Delta t^{(2)}$ produces a uniform increase in all $h_i$:

$$h_i^{(2)} - h_i^{(1)} \sim \log \left( \frac{\Delta t^{(2)}}{\Delta t^{(1)}} \right)$$  \hspace{1cm} (4.16)$$

Therefore, in a system of independent neurons, adding a uniform drive $H$ to each field $h_i$ (inferred with time-bins of width $\Delta t^{(1)}$) would be equivalent to rescale the time-bin width in the inference by factor $e^H$, at least for small values of $H$. 
However, this logarithmic correspondence between $H$ and the time-scale at which the system is observed only holds for independent neurons and small values of $H$ (or $\Delta t$). In our system, where no neuron is completely independent, we observe (Fig. 4.14) more and more important deviations from the uniform logarithmic increase of the inferred fields $h_i$ with the time-bin ratio, as this ratio increases. We compare fields inferred with the 10 ms time-bin with those inferred with three larger time-bins (30 ms, 50 ms and 100 ms) in the Sleep Post of session A. For $\Delta t^{(2)} = 30$ ms, the independent neuron approximation is quite good for some neurons but already poor for the coactivated ones (red circles): fields $h_i$ of coactivated neurons increase less because the activity of these neurons is mainly driven by couplings with other cells. In particular, in the mean-field approximation, eq. 4.14 becomes:

$$p_i(\Delta t) \sim f_i \cdot \Delta t \sim \frac{e^{h_i + \sum_j J_{ij} \langle \sigma_j \rangle}}{1 + e^{h_i + \sum_j J_{ij} \langle \sigma_j \rangle}}$$

from which we obtain:

$$h_i^{(2)} - h_i^{(1)} \sim \log \left( \frac{\Delta t^{(2)}}{\Delta t^{(1)}} \right) - \sum_j J_{ij} \left( \langle \sigma_j \rangle^{(2)} - \langle \sigma_j \rangle^{(1)} \right)$$

where $\langle \sigma_j \rangle^{(2)} - \langle \sigma_j \rangle^{(1)}$ is non negative (mean activities are larger on larger time-scales) and the increase of the fields tends to be smaller than the logarithm of the time-bin ratio. For $\Delta t^{(2)} = 50$ ms and $\Delta t^{(2)} = 100$ ms, the increase of $h_i$ is even less uniform. In general, the largest deviations from the logarithmic behaviour concern the most negative $h_i$, which tend to increase less than $\log(\Delta t^{(2)}/\Delta t^{(1)})$: this suggests that on average neurons with low firing rates are more interacting than neurons with high firing rates, consideration that will be further confirmed afterwards.

To better assess the differences in the effects produced by the drive $H$ and the time-scale we also compute neuron susceptibilities directly from pairwise correlations in the data (eq. 4.6) at different time-scales (plotted in green as a function of $\log(\Delta t/10 \text{ ms})$ in Fig. 4.15) and we compare them to the susceptibilities obtained through Monte Carlo sampling of the Ising distribution with parameters inferred on a 10 ms time-scale (black in Fig. 4.15) as $H$ increases. Only for almost independent neurons (susceptibility peak $\sim 0.25$), like 14, 25 and 18, the agreement between the two curves is good and the effect induced by $H$ reduces to a simple increase of the time-scale. On the contrary, curves are very different for replay neurons, which do not show high synchronized peaks by just increasing the time-scale, and significantly different for many other interacting neurons, such as 8 and 13. This suggests that coactivation events are not the typical (most frequent) events that can be found in the recordings, not even at their characteristic time-scale, but they are probably transient events triggered by a real input feeding the system at particular times.

We estimate the entity of the real input necessary to elicit activation of the Replay group (1-9-20-21-26) of session A, in the following way: we compute the susceptibilities of replay neurons through Monte Carlo sampling of the Ising distribution (in the presence of drive $H$) inferred with time-bins equal to the time-scale at which coactivation of those neurons is observed in the data. Neuron susceptibilities should be peaked at $H = 0$ if neurons are on average coactivated at that time-scale. As already observed, this is not the case and indeed susceptibility peaks are shifted with respect to the zero both in Task and in Sleep Post, see Fig. 4.16. Drive $H$ corresponding to those susceptibility peaks gives an estimate of the real input necessary to trigger transient coactivations of replay
Figure 4.14: Scatter plots of fields $h_i$ inferred with 10 ms time-bins vs. fields inferred with larger time-bin widths: 30 ms (left), 50 ms (middle), 100 ms (right). Parameters refer to the Sleep Post epoch of session A. Red circles locate neurons of the Sleep Post cell assembly (see Table 4.1) and the other labels locate the nearly independent neurons 14, 25, 18 represented in Fig. 4.15, second row. Cyan lines are the bisectors, while blue lines are obtained by shifting the bisectors of $\log(\Delta t^{(2)}/\Delta t^{(1)})$.

Figure 4.15: Neuron susceptibilities computed from Monte Carlo sampling of the Ising distribution in the presence of drive $H$ (black), compared to susceptibilities computed directly from correlations in the data at different time-scales $\Delta t$ (green). Time-bin for model inference: 10 ms. Only susceptibilities of ten representative neurons are shown and refer to session A, Sleep Post epoch.
4.4. DISCUSSION ON THE MEANING OF THE METHOD

Figure 4.16: Susceptibilities vs. $H$ of replay neurons of session A, computed through Monte Carlo sampling of the Ising distribution with parameters inferred on a 10 ms time-scale (right curves) and on the time-scale of neuron coactivation in the data (left curves). Time-scales of coactivation in the data are 40 ms for the Sleep Post and 30 ms for the Task. Each color identifies a neuron (see legend). Vertical bars indicate the mean value of $H_i^+$ of the 5 neuron susceptibility peaks relative to the two time-bin widths. Difference in the time-bins results in susceptibility shifts towards smaller values of $H$, but a global drive of $\sim 0.9$ in Sleep Post and $\sim 1.1$ in Task still has to be applied to see susceptibility peaks, even if inference of Ising parameters is done at the time-scale of coactivation in the data.
Figure 4.17: Susceptibility vs. $H$ of neuron 14 of session A in Sleep Post, computed through Monte Carlo sampling of the Ising distribution with parameters inferred with 10 ms (right curve), 40 ms (middle curve) and 60 ms (left curve) time-bins. 60 ms is the average time-scale of neuron activation in the data ($1/f$). Susceptibility peaks are just above 0.25 and their $H_i^+$ (full vertical bars) are close to the opposite of the fields $h_i$ (indicated by a red star for each time-bin width), meaning that the neuron is almost independent. Difference in the time-bins results in susceptibility shifts of $\sim \log(\Delta t^{(2)}/\Delta t^{(1)})$ and no global drive $H$ is necessary to see neuron activation at the time-scale of 60 ms (susceptibility peak at $H = 0$).
neurons. As a control, we plot (Fig. 4.17) susceptibilities vs. $H$ of neuron 14 (almost independent) in Sleep Post, for three different values of the time-bin used in the inference: 10 ms, 40 ms and 60 ms, the last one being the typical time-scale of neuron activation computed from the data $(1/f)$: susceptibility peak is progressively shifted leftwards of $\sim \log(\Delta t^{(2)}/\Delta t^{(1)})$ and at 60 ms the maximum value is for $H = 0$, confirming that $H$ has simply the meaning of a time-scale for independent neurons. On the contrary, values $H_i^+$ for replay neurons, obtained from the Ising model inferred with 10 ms time-bins, account for both a time-scale increase from 10 ms to 40 ms in Sleep Post (or 30 ms in Task), that is the characteristic time-scale of group coactivation, and for a real input which transiently turns that coactivation into the most likely event.

In Sleep Post this input ($H \sim 0.9$) is very likely to be represented by sharp waves coming from the hippocampus: coactivation of replay neurons is observed during SWS periods, when sharp waves reach the prefrontal cortex.

As for the Task phase, the necessary input ($H \sim 1.1$) is likely to be related to the rat position in the maze. Indeed the probability of coactivation $P_{CoA}$ of the Replay group in session A is peaked at specific positions of the maze, as shown by top panels of Fig. 4.18, which represents the heat maps of this probability over the maze, during the forward path of successful trials (first column) and failed trials (second column) and during the backward path of successful trials (third column) and failed trials (fourth column). $P_{CoA}$ is computed as the number of coactivation events in each maze patch devided by the number of times the rat crosses that patch. $P_{CoA}$ for the Replay group of session A is peaked at the beginning of the maze (starting point of the forward path) and, even more, close to the end of the maze (at the beginning of the backward path), where it is stronger in successful trials and may be related to reward. Localization of coactivation events at the end of the maze is also observed for the Replay group of session B (second row of panels in Fig. 4.18): interestingly, this group is found to coactivate only at the final point and in all 14 successful trials, while it does not coactivate in any of the 9 failed trials. This result also holds for session G of chapter 3, which was recorded the day after session B: Replay groups of these two sessions (which probably have very close recorded neurons) are clearly activated by reward. The input necessary to activate them may still be represented by sharp waves, as the rat stops for a while at the end of the maze. On the contrary, the probability distribution of coactivation events for the conserved group of session D is spread all over the maze and in both successful and unsuccessful trials.

4.4.3 Properties of coactivating neurons

The divergence from independence of a neuron $i$ is well represented by the shift between $H_i^+$ (input corresponding to the peak of the neuron susceptibility obtained by MC sampling of the distribution inferred with 10 ms time-bins) and $\log\left(\frac{1}{f_i \cdot 10\text{ms}}\right)$ (with $f_i$ firing rate of the neuron), which is approximately the value of $H_i^+$ for an independent neuron (see eqs. 4.11 and 4.15). Indeed, there is approximately a monotonic relationship between this shift and both the variation of the conditional average activity $\delta\langle \sigma_i \rangle$ in the jumps between the self-sustaining patterns at $T = 0$ (see full dots in the left panels of Fig. 4.19) and the height of the susceptibility peak $\chi_i^+$ (colors of the full dots in the same panels). The more a neuron is positively interacting with the others, the more the shift $H_i^+ - \log\left(\frac{1}{f_i \cdot 10\text{ms}}\right)$ is
Figure 4.18: Heat map of the probability $P_{\text{CoA}}$ that the Replay group in sessions A and B (top and middle panels) and the conserved group in session D (bottom panels) are coactivated across the maze. $P_{\text{CoA}}$ is computed in every area (black rectangles) of the maze crossed by the rat. We consider coactivation events on a $\lesssim 100$ ms time-scale for the 5 neuron Replay group of session A, and on a $\lesssim 70$ ms time-scale for the 4 neuron Replay group of session B and the 3 neuron conserved group of session D. Results are shown for the forward path (first and second columns, corresponding to successful and failed trials respectively) and for the backward path (third and fourth columns, corresponding to successful and failed trials respectively). The probability distribution of coactivation events for the Replay groups is much more localized than for the conserved group. Probability peaks are mainly localized at the final points of the maze in successful trials.
negative, since positive interactions reduce the value of the input necessary to activate the neuron.

This shift, or divergence from independence, can be related to other neural properties, namely the neuron firing rates and the time variability of the neuron firing rates. Fig. 4.19 shows that the shift decreases in absolute value with the neuron firing rate (middle panels), while it increases with the neuron firing variability (right panels), in agreement with the Poisson nature of most neuron activities, according to which firing variability should scale as $1/\sqrt{f_i}$. Therefore, in most cases, the smaller the neuron mean firing rate, the greater the interaction with the rest of the network, that is the neuron encoding efficacy.

Our method is particularly powerful in detecting those rarely active neurons or, in other words, in catching the deviations from the average of the neuron activity, caused by network effects. Drive $H$ can be seen as a global forcing on the system which allows the underlying interaction structure of the network to manifest itself, in the form of transient coactivation events that would not occur otherwise, due to neuron low firing rates.

A confirmation of the fact that cell assembly activations are rare at all time-scales (also that of the CoA peak) is obtained comparing the shift $H_i^+ - \log(\frac{1}{f_i \cdot 10\text{ms}})$ with the difference between $H_i^+$ and the value $H_{i,\text{ts}}^+ = \log(\frac{\Delta t}{10\text{ms}})$ corresponding to the peak of the susceptibilities computed directly from the data on different time-scales (green curves in Fig. 4.15). Shifts $H_i^+ - H_{i,\text{ts}}^+$ (empty dots in Fig. 4.19) are very close to $H_i^+ - \log(\frac{1}{f_i \cdot 10\text{ms}})$; therefore susceptibilities computed as a function of the time-scale are peaked at $\Delta t \approx \frac{1}{f_i}$. This means that these susceptibilities only catch the neuron mean spiking frequencies and not the activity modulations induced by coactivation events because these events are atypical on all time-scales.

A careful comparison of the properties of the replay neurons in Sleep Post and Task unveils a difference between the two epochs: in Sleep Post replay neurons are characterized by larger shifts $H_i^+ - \log(\frac{1}{f_i \cdot 10\text{ms}})$ and higher firing rates compared to the Task epoch, particularly neurons 20 and 21. This is indicative of the fact that in Task coactivation events are less rare than in Sleep Post, which has a biological reason: in Task several coactivation events are needed to form new couplings important for encoding information, whereas in Sleep Post few offline pattern reactivations are necessary to consolidate the newly formed couplings.
Figure 4.19: **Y-axis:** full dots: difference between $H_i^+ \ (\text{input corresponding to the peak of the Montecarlo susceptibility, computed with 10 ms time-bins in the inference})$ and $\log(\frac{1}{f_i \cdot 10ms}) \ (\text{value of the input at the susceptibility peak of an independent neuron})$; empty dots: difference between $H_i^+$ and the value $H_i^{+ts} = \log(\frac{\Delta t}{10ms}) \ (\text{corresponding to the peak of the susceptibilities in Fig. 4.15 (green curves)})$. Notice that $H_i^+ - H_i^{+ts}$ is similar to $H_i^+ - \log(\frac{1}{f_i \cdot 10ms})$. **X-axis:** variation of the neuron conditional average activity in the jump at $T = 0$ (left); neuron mean firing rate (middle); neuron time variability (right), defined as the ratio between the standard deviation and the mean of the instantaneous firing rate $r_t$; $r_t$ is computed from sliding windows of 100 ms, progressively shifted by 50 ms. All dots refer to neurons with a Montecarlo susceptibility having a peak value $> 0.25$; height of the peak is indicated by the color (from blue for the lowest to red for the highest). Top panels are relative to the Sleep Post epoch of session 1, bottom panels to the Task epoch of the same session.
Chapter 5

Inference and sampling of a Bernoulli-GLM

The Ising model gives a time-independent probability distribution of all possible neuronal configurations in a time-bin, and it does not allow to study spatio-temporal patterns of the activity; moreover the Ising couplings are symmetric by definition and if there are asymmetries in the neuron interactions, the model can not capture them. In order to understand if neuron interactions are indeed symmetric or not, and what is the effect produced by a localized input not only on the coactivation properties of the replay neurons, but also on their activation order, we have inferred a Bernoulli-generalized-linear model from the same dataset of chapter 3. Results are illustrated for session A.

5.1 Model inference and goodness-of-fit

The class of Bernoulli-GLMs has already been presented in 2.2.3. The specific set of equations defining our Bernoulli-GLM is:

\[ P(\sigma_{it}) = \frac{\lambda_{it}^{\sigma_{it}}}{1 + \lambda_{it}} \]  

\[ \log \lambda_{it} = h_i + \sum_{j \neq i} J_{ij} \sum_{\tau=0}^{T_1} e^{-\tau/c_1} \sigma_{jt-\tau} + J_{ii} \sum_{\tau=1}^{T_2} e^{-\tau/c_2} \sigma_{it-\tau} \]

with parameters \( \{h_i, J_{ij}, J_{ii}\} \). Eq. 5.1a is equivalent to eqs. 2.43a and 2.43b. We choose time-bins of width \( dt = 5 \) ms and \( \sigma_{it} = 1 \) if neuron \( i \) spikes at least once in time-bin \( t \), \( \sigma_{it} = 0 \) otherwise. We set the synaptic integration time constant \( c_1 = 2 \) (in time-bins), which is a biophysically plausible value [221], and the integration time constant of a neuron past history \( c_2 = 8 \) (in time-bins). This value is chosen by doing the histogram of the inter-spike-intervals for each neuron: \( c_2 = 8 \) \( dt = 40 \) ms is the typical −slope−1 of the log of these histograms. We put a cut-off at \( e^{-4} \) in the second and third addend of eq. 5.1b, that is we choose \( T_1 = 8 \) and \( T_2 = 32 \).

The log-likelihood of the data under this model is given by:

\[ \log P = \sum_{it} \left( \sigma_{it} \log \lambda_{it} - \log(1 + \lambda_{it}) \right) \]

This \( \log P \) is not normalized because it includes instant terms (with \( \tau = 0 \)) in the second addend of 5.1b; hence the joint likelihood of the variable configuration at each time \( t \) is
replaced by the product of the marginal (single-variable) probabilities and $\log P$ should be interpreted as a pseudo-log-likelihood rather than a log-likelihood (see 2.2.2).

$\log P$ is a concave function of parameters $\psi = \{h_i, J_{ij}, J_{ji}\}$; therefore inference can be done with Newton’s method, computing the Gradient and the Hessian of $\log P$ and recursively updating $\psi$

$$
\psi^{i+1} = \psi^i - \left[ H(\log P(\psi^i)) \right]^{-1} \nabla \log P(\psi^i)
$$

until $||\psi^{i+1} - \psi^i|| < \epsilon$ (we choose $\epsilon = 10^{-3}$).

To test the goodness-of-fit of this model we compare the log-likelihood per time-bin of the data, given the inferred parameters of the model, with the distribution of log-likelihoods (per time-bin) of spatio-temporal configurations generated by the inferred model with Monte Carlo simulations. At each iteration of the simulation, we update a variable chosen randomly in the full spatio-temporal pattern $\{\sigma_{it}\}$ with $i = 1, \ldots, N_{\text{tot}}$ ($N_{\text{tot}} = 37 = \text{number of recorded neurons}$) and $t = 1, \ldots, T_{\text{tot}}$ ($T_{\text{tot}} = 640$, in time-bins $dt$ of 5 ms), using the Metropolis acceptance rule: if the variable flip causes a positive variation of $\log P$, the flip is accepted; if the variable flip causes a negative variation of $\log P$, the flip is accepted with probability $e^{\Delta \log P}$. We choose periodic boundary conditions for each $N_{\text{tot}} \times T_{\text{tot}}$-dimensional configuration (sampled configurations are substantially independent from the boundary condition, since time dependence in the model extends back for $T_2 = 32$ time-bins, and we consider spatio-temporal patterns which are 20 times longer). We call a set of $N_{\text{tot}}T_{\text{tot}}$ consecutive flip trials a Monte Carlo step, and every $2^{25}$ Monte Carlo steps (decorrelation time), we compute the $\log P$ per time-bin of the sampled pattern. We stop the simulation at the $2^{15}$th Monte Carlo step, collecting a total of $(2^{15}/2^2) = 1024$ values. We then plot the histogram of these values of $\log P$/bin. Fig. 5.1 shows the histograms obtained for the three epochs of session A, compared to the $\log P$/bin of the real data (green dot): $\log P$/bin computed from the data is strikingly close (well within one standard deviation) to the mean $\log P$/bin of configurations generated by the model. This indicates that the model is a good representation of these data, which corroborates the analysis illustrated in the following paragraph.

### 5.2 Interactions and spatio-temporal patterns

For each pair of neurons $ij$, we infer a coupling $J_{ij}$ and a coupling $J_{ji}$, together with their error bars, and we compare them in each epoch. Results are shown, for session A, in Fig. 5.2. Almost all couplings are statistically equivalent in the two directions. However, in the Sleep Post, we identify two couplings of the replay group which are significantly asymmetric: couplings 1-9 and 1-21. In the Task, 1-9 is also slightly asymmetric, but with opposite sign, while 1-21 is basically symmetric. We conclude that, differently from what reported in the hippocampus and in other areas, like motor, somatosensory and parietal cortices (reviewed in chapter 1), there is usually no dominant neuron activation order in mPFC cell assemblies. Even when a weak asymmetry is present in some interaction during the Task, it is not maintained in the Sleep Post with the same sign, meaning that replay in this region does not conserve the activation order, and should be understood as the highly synchronous re-activation of strongly interacting neurons, carrying non-sequential information.

However, since slight asymmetries are present in some couplings, we further studied the effect of a localized stimulation on the activation order of neurons of the replay group
5.2. INTERACTIONS AND SPATIO-TEMPORAL PATTERNS

Figure 5.1: Comparison between the log $P$/bin values of the real data and of configurations generated by the inferred models. Histograms represent the distributions of log $P$/bin for the configurations generated with Monte Carlo simulations by the models inferred from the Sleep Pre (left), the Task (middle) and the Sleep Post (right) epochs. Green full dots represent the log $P$/bin of the real data with the respective models; red empty dots (red crosses) indicate the mean (mean ± one standard deviation) of the log $P$/bin values generated by the models. Since the log $P$ we compute is not normalized, these histograms are defined up to a horizontal shift (which is the same for the log $P$/bin of the simulated and the real configurations of each epoch, but is different for the different epochs).
CHAPTER 5. INFERENCE AND SAMPLING OF A BERNOULLI-GLM

Figure 5.2: Scatter plots of couplings $J_{ij}$ vs. couplings $J_{ji}$ in the three epochs of session A. Red dots represent couplings between the neurons of the replay group (1-9-20-21-26). For clarity of the figure, error bars are shown only for the replay group couplings. Almost all couplings have $J_{ij} = J_{ji}$ compatible with zero. In the Sleep Post, couplings 1-9 and 1-21 of the replay group are significantly asymmetric; in the Task, 1-9 is also slightly asymmetric, but with opposite sign, while 1-21 is basically symmetric.

in the Sleep Post Task phase, by simulating the inferred Bernoulli-GLM in the presence of an external oscillatory drive acting on a single neuron of the replay group. More precisely, we apply to a neuron at a time of the replay group a strong constant input $H$ (added to 5.1b) for 8 dt= 40 ms, and we remove the input for the subsequent 40 dt=200 ms; this stimulation is repeated 13 times. A Monte Carlo simulation is run for this Bernoulli-GLM with periodic external drive. At each iteration, we update, with the Metropolis acceptance rule, a variable chosen randomly in the full spatio-temporal configuration $\{\sigma_{it}\}$ with $i = 1, \ldots, N_{tot}$ ($N_{tot} = 37$) and $t = 1, \ldots, T_{tot}$ ($T_{tot} = 624$, in time-bins dt, corresponding to 13 stimulation cycles of period 48 dt). Like in the previous case, we choose periodic boundary conditions for each $N_{tot} \times T_{tot}$-dimensional configuration, and we call the set of $N_{tot}T_{tot}$ flip trials a Monte Carlo step. Every $2^5$ Monte Carlo steps (decorrelation time), we store the 13 patterns (corresponding to the 13 stimulation cycles) of the values taken by $\{\sigma_{it}\}$ with $i \in$ replay group and $t = 1, \ldots, 48$. We stop the simulation at the $2^{15}$th Monte Carlo step, collecting a total of $13(2^{15}/2^5) = 13312$ spatio-temporal patterns of dimension $5 \times 48$. In the first 8 time-bins of each pattern, the stimulated neuron is always active, as expected, since drive $H$ is strong. We study the activation properties of the remaining 4 neurons of the replay group in the 48 time-bins of each pattern, by doing the histograms of all possible ordered sequences of activation of the non-stimulated neurons. In particular, we get a histogram for all patterns in which only 1 neuron out of 4 activates in the 48 time-bins (in this case no sequence is present and the histogram represents the number of patterns in which each neuron is activated); another histogram for all patterns in which 2 neurons activate (which rank the 12 possible ordered sequences of 2 activated neurons out of 4) and, similarly, other two histograms for the patterns with 3 and 4 activated neurons.
Figure 5.3: Histograms of the ordered sequences of activation of the neurons in the replay group (1-9-20-21-26), upon stimulation of one of them at a time, in the Sleep Post epoch of session A. Histograms on the 5 rows correspond to the stimulation of neurons 1, 9, 20, 21, 26 respectively. Histograms on the 4 columns rank the patterns with 1, 2, 3 and 4 active neurons, respectively (in addition to the stimulated neuron), into all possible ordered sequences (x-axis): for example, when neuron 1 is stimulated, all patterns with 2 active neurons are ranked into the 12 sequences 9-20, 9-21, 9-26, 20-9, 20-21, 20-26, 21-9, 21-20, 21-26, 26-9, 26-20, 26-21. The blue part of the histograms is drawn from patterns with all neurons activating in different time-bins ($dt = 5$ ms); the cyan part from patterns with 2 neurons $k, l$ activating in the same time-bin (each of these patterns gives a contribution $1/2$ to the two sequences in which $k, l$ appear in this order and in the opposite one); the yellow part is drawn from patterns with 3 neurons $k, l, m$ activating in the same time-bin (which give contributions $1/6$ to the 6 sequences resulting from all permutations of $k, l, m$). Error bars are computed from all contributions together.
When a neuron spikes more than once in the same pattern, the time of the first spike is considered. Fig. 5.3 shows these 4 histograms for each stimulated neuron in each one of the 5 rows. If 2 neurons $k, l$ fire at the same time-bin, we add $1/2$ to the two sequences containing $k, l$ and $l, k$ and we represent these contributions in cyan; if 3 neurons $k, l, m$ fire in the same time-bin, we add $1/6$ to the six sequences resulting from all possible permutations of $k, l, m$ and we represent these contributions in yellow. No configuration is found with all 4 neurons activating in the same time-bin. Contributions to the histograms coming from patterns with all neurons activating in different time-bins (representing true fully ordered sequences) are in blue. The probability of the different sequences is more and more uniform as more neurons of the replay group are activated. When the replay group is only partially activated, different sequences may have significantly different probabilities, but these differences do not reflect asymmetries in the couplings: the most numerous configurations are simply those in which the first neurons to be active are the neurons with the highest firing rates (21, 20, 26 in this order), meaning that coactivation events are triggered by the neurons that are independently more excitable. When the entire replay group is activated (histograms in the fourth column of Fig. 5.3), the probabilities of the different sequences are not statistically different, given the error bars and the contributions coming from non-fully sequential activations (cyan and yellow). Moreover, synchronicity between the neurons increases as more and more neurons of the replay group are activated (cyan and yellow contributions are relatively larger). We conclude that, in agreement with the substantial symmetry of the inferred couplings, a localized stimulation does not produce a predominant order of activation of the neurons in the replay group, other than that induced by the differences in firing rates.
Chapter 6

Correlation analysis of optogenetic in vitro cultures

In this brief chapter I will move away from the model-based methods and I will show an application of descriptive statistics to the study of in vitro cultures of rat cortical neurons, recorded through Micro-Electrode Arrays (MEAs) composed of 60 electrodes. The data analysed here have been collected by Michele Giugliano’s research group.

The first data set is composed of cultures that have been genetically transduced to express Channelrhodopsin (ChR2), a protein sensitive to blue light, and of control cultures, with no genetic transduction. Cultures have been recorded at different stages of maturation: at the 28th, 33rd, 39-40th and 43rd day in vitro. Wide-field light stimulation, driven using blue light (470 nm) pulses of duration from 1 to 20 ms, has been applied to the transduced cultures at the 33rd and 39-40th day in vitro to characterize their optogenetic response, and to the non-transduced cultures for control.

The second data set is composed of transduced cultures recorded between the 25th and the 27th day in vitro and stimulated at different frequencies: 10, 20, 40 Hz (10 s pause per 1 s of stimulation, repeated 54 times) and 1 Hz (10 s pause per 10 s of stimulation, repeated 54 times).

Signals have not been spike sorted, so we consider every electrode as a unit for our analysis. Note that I will use here the term ‘neuron’ in a broad sense to indicate all neighboring neurons recorded by the same electrode in the MEA.

We study the response to light of the cultures in the first data set by plotting the cross-correlation histograms of neuron pairs before and after light stimulation, at the 33rd and 39-40th day in vitro. Results are shown in Fig. 6.1 for three representative pairs recorded from a control culture (top panels) and a transduced culture (middle and bottom panels). While the cross-correlation histogram of the pair in the control culture does not vary significantly after light stimulation, those of the pairs in the transduced culture are strongly modified by the light at the 33rd day in vitro: the cross-correlation of pair 22-52 has a peak which becomes higher and narrower after application of light, while the cross-correlation peak of pair 29-50 is reset by light. Moreover, these cross-correlations are considerably reduced (at least in the transduced cultures) in more advanced stages of maturation (40th day in vitro), due to aging, and light stimulation is not effective anymore. These correlation profiles are observed for several other pairs.

To get an idea of the variations of the cross-correlation peaks for all neuron pairs, we calculate the correlation indices, defined as $CI_{ij} = p_{ij}/(p_i p_j)$, which represent the values of the cross-correlation histograms within 1 time-bin delay, coinciding with peak values in our
Figure 6.1: Cross-correlation histograms before and after light stimulation for three representative neuron pairs: 25-53 taken from a control culture (top row), 22-52 and 29-50 taken from a transduced culture (middle and bottom rows, respectively). The time-bin width of each histogram is $\Delta t = 1$ ms and the total range $[-1 s, +1 s]$.

The scatter plots of $CI_{ij}$ before vs. after light stimulation at the 33rd day in vitro is shown in Fig. 6.2 for the same control and transduced cultures of Fig. 6.1: light does not produce systematic changes in correlation indices of the control culture; on the contrary it produces significative changes in the transduced culture, where some $CI_{ij}$ decrease due to light stimulation and several $CI_{ij}$ (usually those that are large before the application of light) further increase.

Correlation indices are also the simplest approximation of the pairwise Ising couplings: indeed, Barton and Cocco [179] observe that when network effects are not taken into account, that is the entropy of the Ising model is approximated by one and two-neuron clusters only, couplings are approximated by:

$$J_{ij}^{(2)} = -\frac{\partial (S_{ij} - S_i - S_j)}{\partial p_{ij}} = \log p_{ij} - \log(p_i - p_{ij}) - \log(p_j - p_{ij}) + \log(1 - p_i - p_j + p_{ij})$$

where $S_{ij}$ is the entropy of the system composed by neurons $i$, $j$ and $S_i$, $S_j$ are the entropies of the one-neuron systems. For $p_{ij} << p_i << 1$, this equation gives

$$J_{ij}^{(2)} \approx \log \frac{p_{ij}}{p_i p_j} = \log CI_{ij}$$
We do not infer the Ising couplings because the recording periods are too short (5 minutes per epoch) and error bars on the couplings would be in any case large. We therefore use the correlation indices to also study other properties: in particular, the dependence of correlations on distance between cells and on culture maturation, and spike-timing-dependent plasticity.

The average correlation index at each distance in the MEA is shown in Fig. 6.3 for the control and the transduced culture of Figs. 6.1 and 6.2, at different levels of maturation: correlation indices decrease both with distance between neurons (independently of light stimulation) and with aging.

Spike-timing-dependent plasticity (STDP) is the phenomenon by which synapses between neurons are strengthened or weakened according to the relative timing of the pre- and post-synaptic spikes, and it represents the molecular mechanism of memory. STDP has been studied mainly in cortical and hippocampal slices [222, 223] with patch clamping techniques [224], but not with optogenetic stimulations. Here we investigate the frequency of light stimulation required to induce long-term potentiation (LTP) or long-term depression (LTD) in genetically transduced cultures of the second data set (see above). While patch clamp studies show that a stimulation frequency $>10$ Hz (typically a tetanic stimulation at 100 Hz) is required to induce LTP, we find that a lower frequency is required with optogenetic stimulation of transduced cultures. Among the frequencies of 1, 10, 20 and 40 Hz, only 1 Hz is observed to induce LTP. Our conclusion is based on Fig. 6.4. The top left panel shows that the increase from the Pre to the Post stimulation period of the correlation indices (which we use as approximate estimates of the synaptic strengths) in a transduced culture at the 25th day in vitro, stimulated at 1 Hz, depends on the timing of spikes during the stimulation interval, and this dependence has a typical LTP profile: the shorter the average delay between neuron spikes during stimulation, the stronger the increase in correlations from Pre to Post stimulation (however notice that several pairs do not undergo any change in their correlation). The top right panel shows that $C_{ij}$ between neurons far apart in the array (at a distance $d \sim 500 \, \mu m$) are not potentiated, as expected given the short-range nature of synaptic connections, though the modulation of
the average variation of the correlation indices with distance is weak, probably due to the large amount of neuron pairs at any distance not undergoing potentiation. The second and third rows of panels in Fig. 6.4 show the same plots for stimulation frequencies of 10 Hz and 20 Hz respectively: for these two frequencies, $CI_{ij}^{Post} - CI_{ij}^{Pre}$ tends to be negative for pairs of neurons spiking at short delays during light stimulation and $\langle CI_{ij}^{Post} - CI_{ij}^{Pre} \rangle$ tends to decrease in absolute value with distance between the electrodes, suggesting that LTD is optogenetically induced at these frequencies. Bottom panels show results obtained at a higher frequency (40 Hz): no increase of $CI_{ij}^{Post} - CI_{ij}^{Pre}$ at short delays during light stimulation and no modulation of $\langle CI_{ij}^{Post} - CI_{ij}^{Pre} \rangle$ with distance in the array is observed in this case. We hope to validate these results with the analysis of other cultures in the same conditions and to determine the precise range of light frequencies and culture maturation levels within which these results hold.
Figure 6.4: First row: evidence of LTP in a transduced culture at the 25th day in vitro stimulated at 1 Hz; second and third rows: evidence of LTD in the same culture stimulated at 10 Hz and 20 Hz; fourth row: neither LTP, nor LTD is observed when the culture is stimulated at 40 Hz. Left: Variation of the correlation indices from the Pre to the Post stimulation epochs vs. average delay between the neuron spikes during the stimulation interval; right: average variation of the correlation indices vs. distance between the neurons. $CI_{ij}$ are computed with 10 ms time-bins.
Conclusions

We have applied statistical physics techniques to the analysis of neuronal recordings in vivo from the prefrontal cortex of behaving rats, proposing a novel method to identify cell assemblies and to quantify the replay.

In our method, we map biological, observed neurons on an abstract graphical (Ising) model by inferring the values of connections that are most likely to reproduce the experimental data. Ising inference has been previously applied to neuronal data (e.g. in [143, 146]). In this work, we move considerably forward, by simulating the dynamics of the resulting abstract model, finding interesting parallels with the activity of the real system. In particular, we can identify which functional connections are most affected by learning, and we can detect cell assemblies that form as the rat learns during experience and strongly coactivate during behavior, as well as in the subsequent sleep. We can also identify cell assemblies that are not related to task learning, such as those that are specific to the Sleep epochs. By using the Adaptive Cluster Expansion algorithm of [178] to perform model inference, and with our novel simulation approach, we provide a conceptually sound, measurable definition of cell assemblies, even for the short data samples afforded by in vivo recordings during behavioral tasks, and for systems with highly complex cognitive functions such as the prefrontal cortex. Our Ising-model approach offers a natural way to detect and study cell assemblies in terms of the (co)activation properties of a virtual neural network and can detect replay events even when ‘templates’, as provided for example by the sequential activation of hippocampal place cells, are not available. The method we developed could enable cell assembly identification and characterization in other high-level centers of the brain, where cell assemblies are likely to activate in response to complex combinations of external inputs and internal cognitive states; in such cases our method may prove particularly useful, since those cell assemblies cannot be treated as response modes of the neuronal network to clearly defined external inputs (which is instead possible e.g. in sensory areas) and cannot be studied by direct exploration of the neuron responses to precise, experimentally controllable stimuli. Our analysis significantly extends the principal component analysis (PCA) of [1, 2], as it identifies the neurons participating to cell assemblies in a detailed way in all epochs and is able to predict, through the log–likelihood variation \( \delta \log P \) (3.9), which groups of cells correspond to strong or to weak coactivations.

The largest entries of the top principal components of the Pearson correlation matrix have some correspondence with the neurons in the coactivated groups identified with our method, especially in the Task epoch. However, disentangling coactivated groups in each epoch with PCA is difficult, even with the use of clustering procedures [213] and may lead to arbitrary results: some of the groups identified with PCA-based methods are not observed to coactivate in the data and even the ‘best’ groups extracted by PCA have a weaker coactivation than those extracted by our method. Moreover, differently from standard clustering techniques, our method is able to identify possibly overlapping cell assemblies, and without assuming their number a priori.
The approach we use consists in three steps: (1) the inference of couplings for each epoch allows us to compute the coupling adjustment $\text{Adj}$ (3.4) for each session, and retain large–$\text{Adj}$ sessions for the study of experience-related cell assemblies and replay; (2) the simulation of the inferred models under an external drive $H$ permits us to determine coactivating groups of neurons, and to assess their robustness through their log-likelihood variation $\delta \log P$; (3) the coactivation properties of these putative cell assemblies are then directly estimated from the spiking data through their CoA. Steps 1 & 2 complement and validate each other. Large coupling adjustments can be found without experience-related cell assembly and replay, e.g. when potentiated couplings do not sufficiently interconnect the recorded neurons, as in session E. Similarly, cell assemblies mixing coactivation and inhibition (as session C) would be hard to characterize from the ‘jump’ analysis only. The model-based statistical framework also allows us to quantitatively estimate the amount of replay $R(t)$ in each time-bin of sleep (eq. 3.12) and the total amount of replay $\mathcal{R}$ in each session (eq. 3.13).

The analysis of the self-sustaining patterns and of the ‘jumps’ between them is a zero temperature analysis, meaning that it does not take into account the noise present in the system. Results have been validated through the study of the neuron susceptibilities as a function of drive $H$, obtained with Monte Carlo simulations of the inferred models at $T = 1$, that is in the presence of activity fluctuations: cell assemblies extracted at $T = 1$ are indeed the same as those extracted with the analysis at $T = 0$. Moreover, the heights of the local susceptibility peaks computed at $T = 1$ allow us to assess which neurons of the cell assemblies most strongly take part in the coactivation events, in a more precise way compared to the variations of the conditional average activities computed at $T = 0$. Susceptibilities obtained at intermediate temperatures, between 0 and 1, could in principle be used to characterize the response of the network in the presence of different levels of noise, that is to explore the energy landscape at different coarse-grain levels: very low temperatures give a detailed but partial picture of the energy profile, restricted to the basin of attraction corresponding to the initial condition; intermediate temperatures potentially reveal other minima, not accessible at the lowest temperatures due to energy barriers; high temperatures give the most coarse-grained picture of the energy landscape. Our results show that cell assemblies are robust against (moderate) levels of noise. Moreover, changing the initial condition in the Glauber dynamics at zero temperature does not basically change the curve of the self-sustaining patterns as a function of the drive.

The effectiveness of our Ising model approach relies on several facts.

First, during wakeful experience and during sleep, neuronal activity shows very different regimes. The Ising models can bridge the gap between these states by clearly separating coupling and input effects. As the firing rates are, on average, very similar in Sleep Pre and Sleep Post, the creation of task-related cell assemblies mainly results from a change in the effective couplings $J_{ij}$, which we quantify with the coupling adjustment $\text{Adj}$. Whether $\text{Adj}$ is related to physiological synaptic plasticity or not is a fascinating and totally open question. The Adaptive Cluster Expansion Algorithm we use to perform inference considerably improves estimation of connectivity compared to other inference methods which do not accurately control overfitting. However, the inferred couplings are still “effective” or “functional” if electrodes miss to record a large portion of neurons locally, since the model cannot disentangle true interactions from those mediated by non-recorded neurons; moreover, spike sorting procedures may also introduce errors on the precise discrimination of different neuron spike trains, and some distortion on the estimation of the true connectivity is inevitable. These drawbacks, intrinsic to multi-electrode recordings,
can skew our estimate of experience-related changes to the couplings and may partially explain the presence of session-to-session fluctuations in $\text{Adj}$ (Fig. 3.8). Despite these difficulties, $\text{Adj}$ proves effective in rapidly scanning the whole data set and in providing a first separation between the sessions which are most likely to show replay and the other ones.

Secondly, a crucial ingredient of our approach is the addition of a global drive $H$ to the local inputs $h_i$ in order to detect the cell assemblies (Fig. 3.10). Our exploration of the energy landscape of the inferred Ising models and search for self-sustaining patterns reminds studies on the retina [145], uncovering features of the neural code for visual signals. However, in those studies parameter $H$ is inferred from the data together with the $N(N+1)/2$ parameters $h_i$ and $J_{ij}$, and it captures the ‘global synchrony’ or ‘population firing rate’ in the observed activity of the retina in response to a real visual stimulus. On the contrary, in our study we do not know which is the real stimulus (if any) feeding the system, and the global drive $H$ is a control parameter in our model, which allows us to detect high-activity self-sustaining patterns containing cell assemblies. The set of $h_i$ estimated from the data allows us to fit the diversity of session-averaged firing rates for different neurons, but cannot reproduce transient fluctuations in activity, which are important for cell assembly recruitment. It is a remarkable fact that relatively rare coactivation events can be evoked through an extra-stimulation of the model inferred from the session-average activity, while they are not uncovered by rescaling with the same parameter $H$ the empirical distribution of activity configurations. Indeed, the Ising model gives access to the neuron effective interaction network and drive $H$ triggers an avalanche process on this network: addition of a uniform drive in a system of independent neurons would produce a uniform increase in all firing rates, but in a network of interacting neurons, at some critical values of $H$, activity will propagate or ‘reverberate’ maximally within the cell assemblies, due to the mutual excitatory interactions between their neurons, with the result that increase of activity of cell assembly neurons will be much stronger compared to independent neurons. This is evident from the high susceptibility peaks in the analysis at $T = 1$ and from the large increase in the conditional average activities at $T = 0$ of neurons belonging to cell assemblies. Cell assembly activations are therefore shown to emerge from the microscopic structural properties of the inferred pairwise interaction networks. In physiological terms, $H$ may translate to increased neuronal excitability, synaptic facilitation states, or large transient inputs, such as those arising from hippocampal sharp waves [1]. $H$ also bridges the gap between the time-scale of cell assembly activation (peak of the CoA ratio (3.8)) and the time-scale of model inference (time-bin of the empirical $p_i$ and $p_{ij}$ reproduced by the model). The model, with forcing $H$, is indeed able to predict coactivation events which occur on time-scales of several tens of ms by reconstructing the chain of short-time-scale ($\sim 10$ ms) pairwise coactivations which compose the cell assembly (Fig. 3.1, top left): reconstruction of this activation chain takes place precisely through the avalanche process described above. The study of the neuron susceptibilities at $T = 1$ has allowed us to estimate the fraction of $H$ which represents a time-scale and the residual fraction representing the real input that should be present in the system to activate a cell assembly. The choice of a homogeneous drive $H$ to identify the groups of coactivating neurons ensures that the coactivating neuronal groups found are induced by the parameters fit on the data (set of $J_{ij}$ and $h_i$) only. In addition, non-homogeneous inputs targeting one or more neurons eventually reveal the same neuronal groups in case of strong coactivation.

The Ising model approach to identify cell assemblies is based on the notion of si-
multaneous coactivation (on a certain time-scale), irrespectively of temporal ordering aspects: as the Ising model gives the distribution of snapshots of the activity in each time-bin, couplings are symmetric and no ordering can be predicted by the model. In order to understand if neuron interactions are indeed symmetric or not, we have inferred a Bernoulli-GLM from the same recordings of prefrontal cortex neurons. Almost all inferred GLM-couplings are statistically symmetric; only few weak asymmetries are present and in general they do not maintain the same direction in Task and in Sleep Post. The distribution of spatio-temporal patterns generated by the inferred GLM with localized stimulations is also statistically symmetric over all possible orderings of the neurons in the Replay group. We conclude that there is usually no dominant neuron activation order in mPFC cell assemblies. A possible explanation is that information encoded in prefrontal cortex has aspects that are not inherently sequential, e.g. the current rule used to solve the task, and which may simply be encoded in the synchronous activation of groups of neurons in narrow time-windows. After all, even in the hippocampus (CA1), Hebbian cell assemblies have been observed [18, 29], with synchronously discharging neurons on a ∼30 ms time-scale and no apparent precise activation order. The time-scales of coactivation during the Task of the Replay groups we find in the different sessions are similar and approximately equal to the period of gamma oscillations (∼25 ms) which have been shown to synchronize the activity of neurons in other areas of the brain, like visual cortex (as exemplified in the first chapter) and to underlie the formation of coherent perceptions. It would be interesting to check if the coactivation peaks we observe are indeed time-locked to gamma oscillations, and more generally to study the temporal profile of group coactivations and reactivations over the different trials of the Task and during the subsequent Sleep. In session A, where the task rule has just been changed and learning is at an early stage, we find a non trivial shortening of the coactivation time-scales of the Replay group across trials (suggesting that synaptic plasticity processes may have progressively strengthened the synapses of those neurons in relation to task learning) and a gradual reduction of the replay over the SWS periods of Sleep Post (in agreement with previous studies on the hippocampus, illustrated in the first chapter, showing that replay of a past experience tends to decay with time). A large scale study, including more sessions with a new rule and more sampled neurons, could confirm these results and extend our knowledge of the relations between activation and replay of prefrontal cortex cell assemblies and different aspects of learning (e.g. the rule to be learned, the strategy pursued by the rat, the level of learning achieved, the continuous vs. discrete temporal profile of the learning process).

Our GLM-based method to sample spatio-temporal activity patterns in different conditions would probably show a higher potential if applied to recordings of neurons in brain areas where the sequential order of activation is known to be an important aspect of coding.

Study of the cell assemblies identified with our Ising model-based method in the various experimental sessions has revealed some simple empirical rules for cell assembly modification across the epochs. If a cell assembly is found in Task, while no coactivation is seen in Sleep Pre, then this cell assembly is found also in Sleep Post, e.g. in session B (Fig. 3.28). If cell assemblies are found both in Task and in Sleep Pre, but those assemblies differ in their constituting neurons, then they become associated in Sleep Post, as observed from the merging of jumps e.g. in session A (Figs. 3.10 & 3.11). Merging of cell-assemblies in the Sleep Post epoch suggests that an association phenomenon may take place between old and newly formed memories: it may be indeed a biological indication in favor of the “information overlap to abstract” mechanism (see section 1.4.5) proposed from a theoretical
point of view in [42]. This association can be accompanied by a reshaping of cell assemblies in the presence of effective inhibitory couplings, e.g. in session C. Last of all, if the same cell assembly is encountered in Task and in Sleep Pre, then it is conserved in Sleep Post, see session D in Fig. 3.37. Despite the variability of the possible scenarios, Task cell assemblies (when present) are always reactivated in the Sleep Post epoch: this is in agreement with the finding that a session either does not show any significative mean potentiation or depression of the couplings from Sleep Pre to Sleep Post (Adj within one standard deviation from the average for the null model) or it shows a Task-related change of the couplings (Adj above one standard deviation from the average for the null model); potentiation or depression are never anticorrelated with the Task, as reflected by the absence of sessions with large and negative Adj. Some task-related cell assemblies, which are replayed during subsequent sleep, like the Replay groups of sessions A, B and G, are maximally or solely coactivated at the end of the experimental maze in successful trials, and seem therefore to be elicited by reward. This result does not seem to depend on whether the rat has learned the rule or not, since in session A the rule has just been changed and the rat does not perform better than chance; in session B (recorded from a different rat) the animal is progressively learning and, at the end of the session, it starts to perform better than chance; in session G (recorded from the same rat the day after session B) the rule is finally acquired and changed. The sporadic coactivations observed at the end of the maze in unsuccessful trials suggest that these Replay cell assemblies may respond not only to the presence of reward, but also (even if less strongly) to the expectation of reward. These empirical rules on cell assembly modification across the epochs and this intuition about a possible (certainly not the unique) meaning of the Replay cell assemblies in information coding in prefrontal cortex could be further validated and extended by the analysis of data sets with more (hundreds of) recorded neurons, which are beginning to be collected with the most recent multi-electrode techniques.

In regards to this empirical rules, the large body of knowledge on Ising models with non-homogeneous couplings accumulated over the last decades [225] could prove useful to improve our theoretical understanding of how cell assemblies can be created, modified, suppressed, or combined with each other [7]. Simulations of inferred network models are also a potentially useful approach which, together with the new optogenetic techniques [216], could open exciting perspectives in the manipulation of cortical cell assemblies in a controlled way and in the exploration of the neural code across the brain. Optogenetic stimulations with different target neurons produce a large variety of neuronal responses, which are supported by the same interaction structure. A test of the stability of our inference method would be to compare the networks inferred under different optogenetic stimulations and to assess their similarity. The presence of a significative portion of couplings that are conserved (see scatter plots of the $J_{ij}$ in chapter 3) in the inferred Ising models of the three experimental epochs of [1], despite the very different activity regimes (e.g. firing rates) of the sleep and task phases, leaves high expectations about the stability of the inference method and its capability of reconstructing microscopic, structural properties of the real system in other (e.g. optogenetic) conditions. It would be interesting to directly assess the extent to which reconstructed networks in those conditions are generative models of the neuronal activity, that is the extent to which Montecarlo simulations of those models in the presence of a drive can predict the precise spatio-temporal patterns elicited by real optogenetic stimulations. The use of non-stationary models with potentially asymmetric couplings, like GLMs, could prove useful in reconstructing such patterns when the real synapses are not bidirectional. If the method we developed showed a high predictive
power in these tests, it could be used to dynamically investigate, in a virtual closed-loop platform, the responses of neuron populations to localized stimulations and how these responses would change upon dynamical, network state dependent, changes of the stimulus. Advances in multi-electrode recording techniques and spike-sorting methods, reducing errors deriving from imperfect sampling and imperfect separation of neuron signals, could further improve the predictive power of our method and make this perspective achievable. Zero temperature Glauber simulations of the model in the absence of drive $H$ could also be combined with optogenetic tools to explore the neural code in non-peripheral brain regions, not directly responding to sensory stimuli but accessible through optogenetic stimulations: in the retina, it has been shown indeed that the basins of attraction of the local maxima of the configuration distribution represent the external visual stimulus, since repetitions of the same stimulus may generate different activity configurations which lie in the basin of attraction of the same maximum; a similar dynamical search for the self-sustaining patterns corresponding to each observed activity configuration could in principle be used to study the properties of the neural code, like the degree of redundancy, across the brain. Simulations of inferred models in the presence of a driving input could also prove useful outside the context of optogenetics, for example to study the effects produced at the microscopic neuronal level by large-scale brain waves, which could be simulated by non-stationary, oscillatory forcings $H(t)$.

Finally, our study of in vitro cultures under optogenetic stimulations could be extended in two directions. The first one is at the inference level: with longer recordings, inference of Ising couplings would produce a more faithful estimate of the synaptic connectivity compared to the correlation indices. This more precise estimate could be used, for instance, to study with greater accuracy the dependence of the synaptic strength on distance between the neurons: our prediction is that inferred couplings would be more short-range than correlations, since they would better disentangle direct correlations from the indirect ones. Inference should also be adapted to these in vitro recordings, by modeling adequately the strong transient discharges (bursts) observed in cultures, due to the recurrent connections that neurons develop during maturation in vitro. A way to model these bursts would be to add an input parameter to each neuron, dependent on the global network activity, and to infer these parameters together with the $\{h_i\}$ and $\{J_{ij}\}$, in order to disentangle neuron interactions from the synchronization effect produced by activity bursts. The second direction in which the work could be extended is the investigation of the conditions under which LTP is induced with optogenetic stimulations, in particular the precise range of light frequencies and culture maturation levels. Our study suggests that in cultures transduced with ChR2, LTP is induced with a stimulation frequency of about 1 Hz at the $\sim$25th day in vitro. Determining the requirements to induce LTP and other STDP profiles with optogenetics would be a complementary advancement, with respect to those discussed above, in the context of the controlled manipulation of cell assemblies.
Bibliography


174 Bibliography


[91] O. Eschenko et al. “Noradrenergic neurons of the locus coeruleus are phase locked to cortical up-down states during sleep”. In: Cerebral Cortex 22.2 (2012), pp. 426–435.


J. L. McClelland, B. L. McNaughton, and R. C. O’Reilly. “Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory”. In: *Psychological review* 102.3 (1995), p. 419.


