Politecnico di Torino - Collegio di Ingegneria Elettronica, delle Telecomunicazioni e Fisica (ETF)

Corso in Fisica dei Sistemi Complessi

Tesi di Laurea Magistrale

Robust inference of causal models: applications to spontaneous activity of zebrafish brain

Candidato:
Francesca Mastrogiuseppe

Relatore:
prof. Riccardo Zecchina

Luglio 2014

Abstract

Robust inference of causal models: applications to spontaneous activity of zebrafish brain

by Francesca MASTROGIUSEPPE

During my internship I focused on the problem of inferring causal models from finite observational data. The possibility of learning descriptive and intuitive models from raw and huge datasets has become more and more relevant in the current framework of amazing technological possibilities. The theoretical framework of causal modeling sets the possibility of learning graphical causality relationships among variables, by exploiting the presence of peculiar patterns, called *v-structures*.

If a correct list of conditional independencies among all the variables is provided, traditional constraint-based algorithms, like the well-known PC algorithm, are known to give back an exact result. In practice, however, conditional independencies are inferred from statistical tests on real data, and this kind of approach often results in not satisfying performances. An alternative procedure, proposed by our group, combines the constraint-based approach with a Maximal Likelihood environment, in which entropy evaluations are performed in order to quantitatively estimate the reliability of each single sub-structure. Because of the possibility to include at each iteration only the structures with a high score, our 3off2 algorithm is expected to be more robust against the noise arising from finite dataset.

In the report, this approach is used to analyze datasets recording the single-cell resolution activity of the zebra-fish brain. Data, recorded by the G. Debrégas group at UPMC, derive from fluorescence imaging performed through a genetically encoded calcium indicator. In our analysis, we first focus on limited neuronal structures, whose connectivity is object of specific investigations from the scientific community. In a second step, we exploit the same approach in dealing with full-brain data, on the two-dimensional scale dictated by the scanning laser sheet technology. Inference tests allow us to test in a rigorous way precise pattern of communication which have been hypothesized through anatomical knowledge or raw data analysis. Furthermore, it gives access to more general features of zebra-fish brain connectome, its main paths for information transmission and its set of topological parameters.

Contents

Chapter 1

Inference of causal graphs

1.1 Causal graphs

1.1.1 Introduction

Understanding cause-effect relationships between variables is the central aim of many fields in science, which deal with physical, biological, behavioral and social phenomena. The aim of many sciences, indeed, is to understand the mechanism by which variables take their own value, and predict the values they would assume if some manipulations are made on the natural system.

Usually, experimental intervention is used to find these relationships. In many settings, however, experimental interventions are infeasible because of time, cost or ethical constraints. More often, only observations, i.e. non-experimental, huge data sets are available. In the last 30 years, the typology of available data in science has changed a lot. Thanks to amazing experimental setups, and largely improved storage techniques, a huge amount of raw data is now available for many phenomena of science. Extrapolating simple phenomenological models from this new class of data, by using only traditional methodologies, could become a task of insurmountable difficulty.

It is not a coincidence, then, that during the same years, causal models have been tested and formalized. Causal relationships have the undeniable merit of being intuitive, since they provide a *descriptive* analysis of observational data. The theoretical framework which we are going to introduce in this chapter sets the possibility to learn simple models starting from some probability distributions, in which causal relationships can

be hidden. In the following we will consider, therefore, the problem of inferring causal information from observational data.

From a historical point of view, *functional models* are the first methods through which causal relationships have been encoded. In those models, events are represented by variables, whose value is determined, in an exact way, by functional equations, which encode some more intuitive causal relationships. An approach which results to be more handy from an algorithmic point of view suggests to focus on the inference of graphical structures (graphs), in which relationships between causes and effect are encoded by directed edges. In our applications, vertices will stand for the discrete variables of the problem, which can assume values in a fixed range. Figure 1.1 shows an over-simplified example of this kind of models.

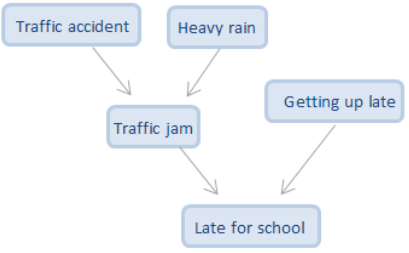


FIGURE 1.1: Graphical representation of the causal model for the phenomenon "being late for school".

Suppose that two variables are given; in this toy model we will consider "traffic jam" (X) and "heavy rain" (Y). In our simplified view, the variables can assume only the values *TRUE* or *FALSE*. An observational dataset provides us information about the weather and the traffic situation, which are recorded every morning. The only kind of information that we can extract from this framework is the *mutual information* between variables $I(X, Y)$. In other terms, we can statistically decide if X and Y show some kind of probabilistic dependency or not.
If we discover that X and Y show dependency, such that $I(X, Y) > 0$, we will know that some relationships can exist between the two phenomena, but we will never be able to decide whether the rain is the cause of traffic, or the other way round, or maybe X and Y stand for the effects or a common cause which was not directly observed. How to teach a machine, then, to recognize proper relationships of causality?

Ordinary causal reasoning in human beings often relies on temporal schemes: a cause is always thought to precede the effect. Temporal mechanisms, however, can lead to mistakes: swallows are known to reach cold regions a few days before the arrival of spring, but they do not cause the spring to come. Furthermore, time-series datasets are not always available. Very often, scientists deal with datasets which gather together many data points which have been registered during the same experimental session, or in a few different moments.

The kind of approach we will use in this study does not rely on temporal patterns. As a first step, we will focus on small systems constituted at least by three variables, which are considered to be the smallest subset about which some causal models can be assessed. We will exploit again tests of independency, which now can be simple (as before), or *conditioned* on some external group of variables:

Definition 1.1. Being A, B and C three subsets of the variables X_1,...,X_N, on which the joint probability distribution P is defined, we say that A and B are conditionally independent given C, or $(A \perp B|C)_P$, if and only if:

$$P(a|b,c) = P(a|c) \tag{1.1}$$

where a,b and c are respectively the values assumed by the subsets A, B and C.

In other words, once the value of C is fixed, the value of B does not influence the knowledge we had about A. When $C = \emptyset$, we recover the usual definition of independency.

The previous definition turns out to be exceptionally instructive for a particular class of causal sub-structures, as we will explain in the following. Consider the triplet of variables "heavy rain", "traffic jam" and "late for school". In our final model, they are connected through a causality flux which originates from the first variable and finally reaches the third one. We expect then to see those two variables to show dependency in the data.
Suppose now to fix the value of the intermediate one: suppose to access data and check whether there is traffic in the town. This knowledge makes rain and our phenomenon (being late for school) completely independent.
An opposite behavior can be underlined in the triplet "traffic accident", "heavy rain" and "traffic jam" where the last variable is recognized to be a *common effect* of the first two. Here, by fixing the common effect (conditioning on it), we are creating a dependency between the causes. Indeed knowing that an accident happened, in our naive and simplified example, does not change our knowledge about the weather. In other words, we expect the two causes to behave as independent variables. Suppose then to discover that a traffic jam formed in the town. The awareness of the accident will significantly change the probability estimation that we previously had about the rain. In poor words, variables have become dependent. This kind of triplet is called a *v-structure*, and is schematized in picture 1.2.

Definition 1.2. A v-structure (or *vee-structure* [1], or *unshielded collider* [2]) is a triple of nodes X_1, X_2 and X_3 such that in the graph G: $X_1 \to X_2$ and $X_3 \to X_2$, but X_1 and X_3 are not adjacent.

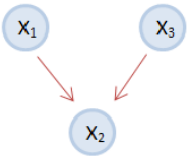


FIGURE 1.2: A v-structure.

As a second, more explicit example, we focus on the v-structure created in 1.1 by the variables "traffic jam", "getting up up late" and "late for school". Before conditioning on the value of the third variable, we would say that the first two are completely independent. Discovering that "late for school"$=FALSE$, however, would decrease the probability of having traffic jam and getting up late in a dependent way. In particular, if we ascertain the presence of traffic jam, we expect the probability of getting up late to become smaller and smaller. In the other way round, when "late for school"$=TRUE$, we expect the same two probabilities to increase. If we know that "traffic jam"$=TRUE$, then the probability of also getting up late will decrease with respect to the non-conditioned value: the appearance of one of the two causes tends to exclude the other one. Conditioning on being late is then creating a fictitious anti-correlated behavior among independent variables.

V-structures will represent the leading element of the inference process, since it turns out that when the appearance of such a conditional dependency arises, the causal sub-graph which can be inferred is unambiguously a v-structure. The same possibility does not hold for the first triplet of variables we analyzed. All the oriented, open triplets which are not v-structures are colled *non-v-structures*. They are characterized by one of the three shapes shown in 1.3, and by the vanishing of a dependency when a conditioning on its vertex is performed. As we will characterize in chapter 2 through entropy calculation, it is not possible to distinguish from independency tests which shape among the three is assumed by the graphical model which generated the data. This feature will limit considerably our possibility of inference. It is possible, indeed, that perfect performances of inference algorithms will be able to give back only partially oriented graphs, when the presence of v-structures is not enough to induce orientation on all the edges of the network.

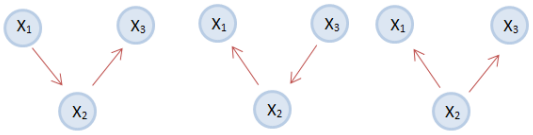


FIGURE 1.3: The three equivalent forms of non-v-structures.

1.1.2 Bayesian networks

In the following, we will introduce a few elements which help to formalize the starting intuition we presented so far. Because of the nature of causality, in our applications we will deal with directed graphs. If you imagine to remove all the arrowhead from the edges of the graph, then you will get the undirected graph which is called *skeleton*. A path in the graph is said to be *directed* if the orientations of all the involved edges point in the same direction. If the starting and the ending point of the route coincide, then the direct path is said to be a direct *cycle*.

When the directed graph contains no cycles, we will say that it is acyclic. In the following, we will focus on *DAGs*, that is, on Directed Acyclic Graphs. Indeed, this typology of structures lies at the heart of the definition of the versatile and efficient Bayesian networks.

A Bayesian Network can be seen as a convenient representation of a given probability distribution. Suppose you want to characterize N variables through their joint distribution, $P(x_1, ..., x_N)$. A graph representation of the distribution P can represent a huge advantage if each X_i depends just on a small number of variables. Those variables are called *Markovian parents* for the variable X_i.

Definition 1.3. Given $X_i \in V$, the variables in the set $PA_i \subset V$ are said to be Markovian parents of X_i if PA_i is the minimal set of predecessors of X_i that makes X_i independent of all its others predecessors in a defined arbitrary ordering.

In other words, thanks to Markovian parenthood, it is possible to get the following key simplification:

$$P(x_i|x_1, ..., x_{i-1}) = P(x_i|pa_i) \tag{1.2}$$

Notice that for each distribution P the *chain rule* of conditional probabilities holds:

$$P(x_1, ..., x_N) = \prod_i P(x_i|x_1, ..., x_{i-1}) \tag{1.3}$$

By exploiting the Markovian property in 1.2, it becomes possible to obtain the following decomposition for P:

$$P(x_1, ..., x_N) = \prod_i P(x_i|pa_i) \tag{1.4}$$

In 1988 Pearl shown that if P is always strictly positive for all the possible configurations, then the sets of Markovian parents for all the variables are unique [3].

In order to get from them the Bayesian graphical model, imagine to draw the N variables as nodes of a graph. Then you have to draw arrows between them in such a way that each Markovian parent is linked by an outgoing arrow with its son X_i. You will get a DAG G which is called *Bayesian network* for P.

The special relationship between P and G encoded by the graph will be called *Markov compatibility.*

Definition 1.4. Given a DAG G, for each distribution P which satisfies the decomposition 1.4, we will say that P and G are Markovian compatible.

The compatibility between the probability distribution and its Bayesian graph turns out to be a far richer relationship. It is possible to provide a completely equivalent characterization of compatibility, which is enriched of a more intuitive interpretation. Here we will use blindly the concept of *d-separation* between variables in a graph G. If a subset of variables A is said to be *d-separated* from B, then we will denote, in formulae: $(A \perp B|C)_G$. The following result holds [4]:

Theorem 1.5. *Consider the three subsets of V A, B and C. If A and B are d-separated by C in the DAG G, then A and B are conditionally independent, given C, in all the distributions P which are compatible with G. In formulae:* $(A \perp B|C)_G \Rightarrow (A \perp B|C)_P$.

We notice that the converse is not always true. That is, it is possible that there are some *accidental* independencies encoded in P which cannot be read in G. They are typically due to some fine tuning of the underlying parameters.

Now it is necessary to explicit the proper definition of d-separation, which makes the theorem 1.5 to hold. We will say that A and B are d-separated by C if and only if all the paths starting from a vertex $X \in A$ and ending in $Y \in B$ are d-separated by a vertex $Z \in C$. In particular:

Definition 1.6. A path from i to j is said to be *d-separated* by a set C if and only if it contains one of this two sequences:

1. $i \to v \to j$ or a fork $i \leftarrow v \to j$, such that v is in C, or
2. an inverted fork $i \to v \leftarrow j$, such that v, and all its descendants, are not in C.

We can try to have an insight into this property through intuition. Suppose that arrows are encoding a causal relationships between variables. By fixing the value of one of the vertices v in C, in case 1 we are blocking the flux the information which flows from i to j, and then from A to B. Instead, condition 2 requires that C, in G, does not contain any common effect of i and j. By fixing the common effect, indeed, we justified the creation of some dependencies between the two causes which were, in principle, completely independent. This is what happens, typically, in v-structures.

The role of the v-structures is of great importance in DAG recovering. They encode all the essential information which is inferred from the probability distribution, and they turn out to be the same in all the DAGs which are compatible with a given P (*Markov equivalent* graphs). Indeed, the following theorem holds:

Theorem 1.7. *Two DAGs are Markov equivalent if and only if they have the same skeleton and the same set of v-structures, that is, the same* pattern *[5].*

Because of the limitations in inference due to the presence of non-v-structures, it looks natural that the best an inference algorithm can do with observational data, is to recover the full equivalence class to which the true causal explanation belongs (a CPDAG: Completely Partially Directed Acyclic Graph).

1.1.3 Inference of causal networks

Suppose now you do not have any pre-printed causal model, but you want to find a probable one by just looking at observational data, which encode the state of each variable during the experiment. That is, you have access only to an empirical joint probability distribution.
You would like to extract the true *causal model*, which consists of the CPDAG G, whose links express causality relations.

In practice, we will require the final model to be *stable* with respect to data. This corresponds to require the causality graph G to be compatible with the distribution P, but also the distribution P to be compatible with G. In raw words, we will assume that the inverse of theorem 1.5 is valid also. Suppose that you can extract from data a list M of conditional independency statements ranging over all the variables. We will adopt the following definition:

Definition 1.8. The inferred graph G is said to be the *stable* if and only if the set of conditional independencies entailed by G is exactly the same set M which can be be extracted from data.

A stable model is also called a *complete causal explanation* [2], or a *faithfull* one [6]. The existence of a stable model is used as hypothesis in many algorithm whose task is to reconstruct the graph compatible with the causal model. In facts, it was proved that the distributions which admit a stable graph are indeed the wide majority [7].

1.2 PC algorithm

The PC algorithm relies on a solid theoretical background. It is possible indeed to prove that, given an ideal oracle of structural independencies which is not sensitive to the disturbances of the experimental noise, this algorithm gives as output an exact result, provided that a stable underlying DAG exists.

The main idea is that, since the final DAG model should encode the right conditional independencies, then by the latter we can try to infer some aspects of the former. Indeed, the algorithm first infers a complete list of conditional independencies between the variables. In practice, the oracle is replaced by a statistical test, which needs to be provided with a significance level α. The arbitrariness in the typology of the test and in the used value of α are at the origin of the major limitations of PC.

The PC algorithm takes as input the inferred structural independencies structure and provides as output the *maximally oriented* graph (or CPDAG), which encodes the highest level of inference you can reach about G. The *maximally oriented* newtork contains all the oriented edges which are present in all the complete causal explanations of M. It is possible, indeed, that there exist more than one causal explanation for M. However, as we saw, they all share the same set of v-structures, and then they belong to the same class of equivalence.

The PC algorithm acts in three steps:

S1:

Start from a complete, undirected graph with nodes V. Get an undirected graph G by applying the following rule: A is adjacent to B if and only if there does not exist a set

$S \subseteq V \setminus \{A, B\}$ such that, in M, $(A \perp B|S)_M$. If such S exists, then set $Sep(A, B) = S$.

In facts, what the authors Spirtes and Glymour suggested to do [8], is to test the existence of the set S by starting from the empty set and then progressively increasing its cardinality by including more and more neighbours of the vertices A and B. The cardinality of this set is then bounded by the sum of the two degrees of the nodes with higher number of edges in the graph. This trick makes possible to run the algorithm in polynomial time on sparse graphs.

Step 1 gives back the skeleton of the inferred graph. In the hypothesis that our distribution is stable, it is possible to prove that the skeleton found is exact. Indeed, it is possible to prove the following lemma [1]:

Lemma 1.9. *Let M be an exact independency list. A DAG G is compatible with M if and only if:*

1. *the edge $\overline{AB}$ is in G if and only if, $\forall S$, $(A \not\perp B|S)_M$;*
2. *the v-structure $\overrightarrow{AB}\overleftarrow{BC}$ is in G if and only if $\overline{AB}$ and $\overline{BC}$ are in G but not $\overline{AC}$, and $\forall S$, if $(A \perp C|S)_M$, then $B \notin S$.*

The second point of lemma 1.9 suggests moreover a second step which turns out to be useful in the orientation procedure, and involves v-structures.

S2:
Start from the skeleton of G. For each pair of non-adjacent nodes A and C see if there is a node B which is not in $Sep(A, C)$. In this case, orient the edges $\overrightarrow{AB}$ and $\overleftarrow{BC}$.

In facts, because of noise, it can happen that PC tries to reconstruct v-structures which contradict each other about the orientation of one edge. In this case, the version of the R package we used (pcalg), simply takes into account the last inferred orientation. This is the maximum effort of which PC is capable, having no kind of score which can test the statistical faithfulness of the two reconstructed structures. The third step, finally, extends the orientations of the pattern by applying a few orientation rules. It is possible to prove, indeed, that those laws encode the only choice that we can make, because orienting in the opposite direction would lead to the creation of loops or new v-structures, which were not inferred by the algorithm [2].

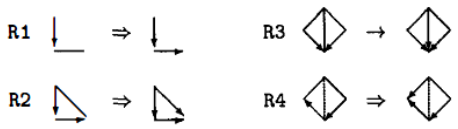


FIGURE 1.4: The four rules used by PC algorithm in order to propagate orientations [2].

S3:
Start from the pattern of G. While G has some undirected edges, orient as many edges as you can by applying the rules encoded in picture 1.4.

A theorem [2] ensures that, in the case of an ideal oracle of independencies, the result given back by steps 1,2 and 3 is the true maximally oriented graph.
Moreover, PC is a quite efficient algorithm. For fixed graph connectivity, its complexity increases polynomially in the number of vertices N [8].

1.3 Constraint based and Bayesian methods

Probably PC is the most well-known example of *constraint based* search algorithms. Typically, the constraint based methods, taking as input an oracle of conditional independency which acts in the population of variables, return a representation of the Markov equivalence class of the causal graph. However, as it is possible to discuss, they present some advantages and some disadvantages.

One main advantage of constraint-based algorithms are their very fast performances; moreover, they can be easily generalized to the case when some latent variables are assumed to exist outside from the recorded dataset. Their output is a unique, clear graph, which codifies all the equivalence class of the true model. Moreover, if the algorithm is provided of an ideal independencies estimator, it can rely on strong theoretical arguments which ensure the correctness of the output.

Their main disadvantage is that constraint-based algorithms, in fact, show bad accuracy and robustness in dealing with real world data. At small sample size, indeed, tests of conditional independency are characterized by low precision, especially when conditioning on many variables is needed. In the case of huge dataset, instead, statistical evaluation can be extremely slow.

Structural independencies, moreover, are identified according to an arbitrary statistical significance level. A very problematic issue is that results from contraint-based approaches depend on the arbitrary ordering of the variables in the dataset. Furthermore, small mistakes made in the early part of the algorithm can propagate and lead to later mistakes. Constraint-based methods, eventually, do not provide a quantitative measure of reliability about the result, neither any indication of how much reliable the output model is compared to the next best one.

The more reasonable alternative to this approach is to develop an efficient way of implementing *score-based* methods, which can assess in a quantitative way the likelihood that our data sample is generated by a given graph. In this case, algorithms associate a score to each possible model, measuring the closeness between graph and data and the level of essentiality (e.g. the number of free parameters) of the model.
In the limit of large data sample, it is expected that the DAG with the highest score G belongs to the equivalence class of the model which underlies the data.
The practical definition of the score to be used has some arbitrariness. In the Bayesian approach, for example, the quality measure is essentially the probability of having a given network, by knowing the measured database. Let D indicate the dataset, in order to compare the probabilities of two DAG structures G_1 and G_2, you can calculate:

$$\frac{P(G_1|D)}{P(G_2|D)} = \frac{\frac{P(G_1,D)}{P(D)}}{\frac{P(G_2,D)}{P(D)}} = \frac{P(G_1,D)}{P(G_2,D)} \tag{1.5}$$

Then what you need is just to compute the joint probabilities $P(G,D)$.

Since the number of possible DAGs is super-exponential in the number of vertices, even with a modest N it is not possible to examine each graph and test its compatibility with the probability distribution. It is possible to prove, indeed, that if each vertex has more than 2 parents in G, then the problem results to be NP hard [9]. Some heuristic search procedures have then been developed, for selecting at most a polynomial number of different structures; one example consists of the algorithms based on the hill-climbing method [10] [11].

Chapter 2

The 3off2 algorithm

2.1 Theoretical background

The purpose of our algorithm is to combine the two strategies we discussed so far, keeping the more advantageous features of each approach. By computing a likelihood score for all the local structures, the algorithm performs a process of optimization on small scale, eventually constructing a unique output model. The Maximum Likelihood framework, in which the algorithm is built, circumvents the necessity to set an arbitrary significance threshold in independencies identification.

The conception and the implementation of the 3off2 algorithm involves all the group of prof. Isambert (UMR 168, Institut Curie); the project started before my arrival but is still developing. A first paper, which concerns only the skeleton reconstruction part (see further), is currently under review [12].

2.1.1 A Maximum Likelihood approach

The maximum likelihood method is a well-known criterion exploited in the process of statistical inference of the free parameters in a given model. In order to briefly recall this methodology, suppose we have a sample of x_1, x_2,... x_n repeated observations of a variable X. We suppose that they fit a given family of probability distribution f and we want to find the best choice possible for the parameter (or the vector of parameters) θ.

Then we suppose to fix θ and construct the probability of observing the outcome sample, by multiplying all the terms coming from the independent trials: $p_\theta(x_1, x_2, ...x_n) = \prod_{i=1}^{n} p_\theta(x_i)$. If you look at the same quantity as a function of the parameter θ, being the outcome of the trials fixed, you get a function $\mathcal{L}(\theta|x_1, x_2, ...x_n)$ which is called *likelihood* function. For practical purposes, it can be convenient to deal with its logarithm, which is called *log-likelihood* function:

$$\ell(\theta|x_1, x_2, ...x_n) = \ln \mathcal{L}(\theta|x_1, x_2, ...x_n) = \sum_{i=1}^{n} \log p_\theta(x_i) \tag{2.1}$$

The Maximum Likelihood (ML) principle assumes that the most reliable value for the parameter θ is the one which maximizes the quantity $\ell(\theta|x_1, x_2, ...x_n)$.

The likelihood $\mathcal{L}(G)$ for a model G can also be expressed in terms of *cross entropy* between the probability distribution emerging from data and the one encoded in G.

Definition 2.1. In information theory, if X is a discrete variable, if two distributions of probability $e(x)$ and $p(x)$ are given, their cross entropy H is defined as: $H(e, p) = -\sum_x e(x) \log p(x)$.

By introducing the single-distribution entropy function $H(e) = \sum_x e(x) \log e(x)$, and the *Kullback-Leibler divergence* (or Relative Entropy) $D_{KL}(e||p)$, it is straightforward to see that the following simple relation holds: $H(e, p) = H(e) + D_{KL}(e||p)$.

In the following, we will assume that e is the experimental distribution, emerging from data, while p is the one encoded in the model to be tested. It is clear that in this case, D_{KL} is the only non constant term with respect to the distribution p which has to be tested. In particular, it is possible to show that the minimization of this single term provides the same result for θ which is given by the ML criterion.

Theorem 2.2. *Let $e(x)$ be the experimental probability distribution extracted by data, and $p(x)$ a generic one. Then the distribution $\tilde{p}$ is the ML estimator if and only if:*

$$\tilde{p} = argmin_p\{D(e||p)\} \tag{2.2}$$

In the following we will exploit the alternative form of the likelihood function we introduced in the previous reasonings:

$$\mathcal{L}(G) = e^{-nH(e,p)} \tag{2.3}$$

where n is the number of independent observational data points. At this stage we extend our analysis to N variables $\{X_i\} \equiv \{X_1, X_2, ..., X_N\}$, which will become the nodes of the graph we want to infer. We suppose that each variable takes value x_i. If the model we are assuming is compatible with a DAG G, then for p it must hold: $p(\{x_i\}) = \prod_{i=1}^{N} e(x_i|pa_i)$. This factorization on the single node propagate to the likelihood function too, since:

$$H(e,p) = -\sum_{i=1}^{N}\sum_{x_i} e(\{x_i\}) \log e(x_i|pa_i) = \sum_{i=1}^{N} H(x_i|pa_i) \tag{2.4}$$

In 2.4 we used the notion of *conditional entropy* for two variables X, Y: $H(X|Y) = H(X,Y) - H(Y) =$
$\sum_{x,y} p_{XY}(x,y) \log p_{X|Y}(x|y)$. Eq. 2.3 will allow us to compare two alternative graphical models G and G' by taking the ratio of their likelihood functions. The two models differ for the parental set predicted for each variable X_i:

$$\frac{\mathcal{L}(G')}{\mathcal{L}(G)} = e^{-n\sum_{i=1}^{N}[H(X_i|PA_i') - H(X_i|PA_i)]} \tag{2.5}$$

By applying this result to open triplets of nodes X, Y and Z, one can get an interesting characterization of v-structures in terms of entropies.

Suppose your graph consists of an isolated v-structure $\mathcal{V}$ with basis XY. Then by applying the definition of *mutual information* $I(X;Y) = H(X) + H(Y) - H(X,Y)$, since X and Y have no parents in $\mathcal{V}_{XY}$, you get:

$$\mathcal{L}(\mathcal{V}_{XY}) = e^{-n[H(Z|X,Y)+H(X)+H(Y)]} = e^{-n[H(X,Y,Z)+I(X;Y)]} \tag{2.6}$$

A non-v-structure $N\mathcal{V}_{XY}$ can assume one of the three forms in picture . For all of them, by using the following equality for conditional entropy: $H(XY) = H(X|Y) + H(Y)$ you get:

$$\begin{aligned}\mathcal{L}(N\mathcal{V}_{XY}) &= e^{-n[H(X)+H(Z|X)+H(Y|Z)]} = e^{-n[H(X|Z)+H(Y|Z)+H(Z)]} \\ &= ... = e^{-n[H(X,Y,Z)+I(X;Y|Z)]}\end{aligned} \tag{2.7}$$

In the last step we recognized the *conditional mutual information* between X and Y given Z: $I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$. By computing 2.5 for a v and a non-v model you end up with a *multi-variate mutual information*, and in particular,

with the three-point one. Indeed:

$$\frac{\mathcal{L}(\mathcal{V}_{XY})}{\mathcal{L}(N\mathcal{V}_{XY})} = e^{-n[I(X;Y)-I(X;Y|Z)]} = e^{-nI(X;Y;Z)} \tag{2.8}$$

where $I(X;Y;Z) = H(X)+H(Y)+H(Z)-H(X,Y)-H(Y,Z)-H(Z,X)+H(X,Y,Z)$. In general, *k-variate mutual information*, with $k \geq 3$, are interesting objects. Indeed, while the two-points $I(X,Y)$ is always larger than 0, they can be either positive or negative.
In this case, the sign of $I(X,Y,Z)$ provides an order of magnitude of the relative likelihood of the v-structure versus non-structures.
When $I(X;Y;Z) < 0$ v-structures are characterized by a larger probability of being present in the true model.

Furthermore, we notice that $I(X;Y;Z)$ is a symmetric function with respect to its variables. Then, whatever basis is chosen for the open structure, the same result for the likelihood ratio holds. It follows that we cannot conclude the inference process by simply looking at $I(X;Y;Z)$.
To this end, however, it is possible to easily show that the most probable basis is the one characterized by the lower value of mutual information between its nodes, since:

$$\frac{\mathcal{L}(\mathcal{V}_{XY})}{\mathcal{L}(\mathcal{V}_{YZ})} = \frac{e^{-nI(X;Y)}}{e^{-nI(Y;Z)}} \tag{2.9}$$

It is straightforward to show that the same holds for non-v-structures. Two-point and three-point information, then, are enough to infer the true structure of the triplet according to the ML principle.

2.1.2 Generalized v-structures

The issue of inferring a real, large graph requires some generalizations. In this section we will extend the results for simple triplets to larger structures which will be called *generalized*.

As a first stage, we fix a couple of non-adjacent nodes X and Y. We consider then the set $\{U_i\}_{XY}$ of all the *upstream nodes*. Each upstream node has at least one direct and oriented connection to X, Y or another upstream node. In the following, we will not specify, if not strictly necessary, the pedex XY.

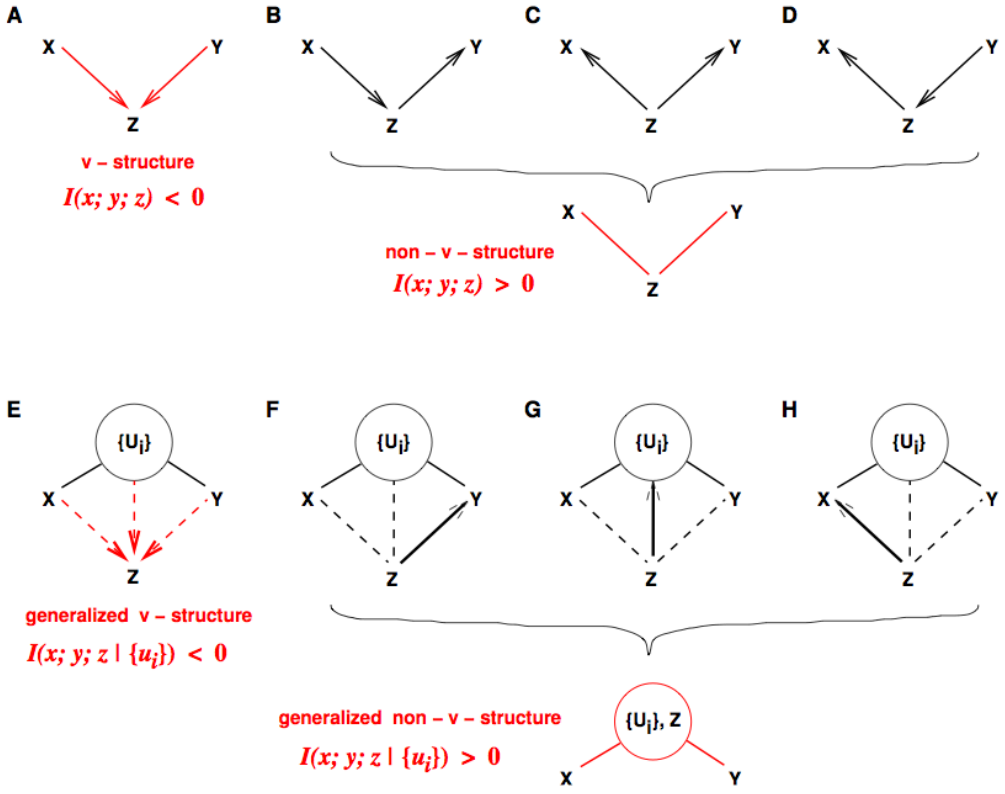


FIGURE 2.1: Brief resume of the found results. Pictures A, B, C and D refer to isolated structures; E, F, G and H to the generalized ones.

Suppose to have a third vertex Z which does not belong to the $\{U_i\}$ set. This is possible only if all the connections which exist between Z and X, Y and $\{U_i\}$ point towards Z. We will say that Z is the apex of a *generalized* v-structures. In all the other cases, Z has at least one edge oriented in the opposite direction, than it is part of the upstream nodes. In this case, we will say that nodes form a *generalized* non-v-structures. If you want to check, like the PC algorithm does, the conditional independency between X and Y, all the Z belonging to generalized non-v-structures should be considered. Indeed, it is likely that part of the information traveling from X to Y passes through them.

It is quite straightforward to compute the likelihood functions also in this case, by applying 2.3 to the most general case. Here we will denote with $\mathcal{V}$ and $N\mathcal{V}$ the generalized

structures.

$$\mathcal{L}(\mathcal{V}_{XY}) = e^{-n[H(Z|X,Y,\{U_i\})+H(X|\{U_i\})+H(Y|\{U_i\})]} = e^{-n[H(X,Y,Z,\{U_i\})+I(X;Y|\{U_i\})]} \tag{2.10}$$

$$\mathcal{L}(N\mathcal{V}_{XY}) = e^{-n[H(X|Z,\{U_i\})+H(Y|Z,\{U_i\})+H(Z|\{U_i\})]} = e^{-n[H(X,Y,Z,\{U_i\})+I(X;Y|Z,\{U_i\})]} \tag{2.11}$$

In the same way you get the likelihood ratio:

$$\frac{\mathcal{L}(N\mathcal{V}_{XY})}{\mathcal{L}(N\mathcal{V}_{YZ})} = e^{-nI(X;Y;Z|\{U_i\})} \tag{2.12}$$

where we introduced, in analogy with 2.8, the three-point conditional mutual information: $I(X;Y;Z|\{u_i\}) = I(X;Y|\{U_i\}) - I(X;Y|Z,\{U_i\})$. A significantly positive $I(X;Y;Z|\{U_i\})$ means, then, that a generalized non-v-structure is more likely that a v one. Also in this case, eventually, we end up with a symmetric quantity, which cannot tell us how to set the basis of the inferred structure. Again, we have to look for the couple with lower mutual information, in this case conditioned on $\{u_i\}$:

$$\frac{\mathcal{L}(\mathcal{V}_{XY})}{\mathcal{L}(\mathcal{V}_{YZ})} = \frac{\mathcal{L}(N\mathcal{V}_{XY})}{\mathcal{L}(N\mathcal{V}_{YZ})} = \frac{e^{-nI(X;Y|\{U_i\}_{XY})}}{e^{-nI(Y;Z|\{U_i\}_{YZ})}} \tag{2.13}$$

2.2 The main idea

The 3off2 algorithm starts from a fully connected graph and, like PC, tries to remove one edge at time by looking at the conditioning independencies between X and Y.
In this paragraph we will characterize the same process in terms of mutual information. This new perspective will suggest a proper way of choosing, at each step, a conditioning set S. We suppose to start by computing $I(X;Y)$. This quantity is always non-negative, and can be typically larger than zero even if X and Y are not in direct causal relationship. Information can be indeed mediated by all the nodes which are present on the pathways existing between the two variables.

As in PC, we are looking for some mutual conditional independencies, that is, we are looking for the set S such that $I(X;Y|S) \simeq 0$ up to a certain significance level. Here the task is to build this set in a proper way, by including the more robust patterns: the ones which are more likely to contribute to communication between X and Y.

Then, once X and Y have been fixed, you peek up iteratively all the possible third nodes Z_i and you measure the reliability of these subgraph to be generalized v or non-v-structures, by assigning them a *score* which is based on their likelihood. Then you select only the reliable non-v-structures and you test the mutual information between the two original nodes, by conditioning on the new one. If Z_i has been chosen properly, you have in facts reduced the mutual information between X and Y. By including at each step a new Z_i in S, you can find that conditioning on it it is sufficient to get $I(X;Y|Z) \simeq 0$, or not. In the first case, you include Z_i in the upstream nodes of the edge XY, and you include a good candidate in S. If X and Y are causally non-connected in the true model, at a certain point you expect to have a set S large enough to block all the information pathways between the two nodes.

More formally, we imagine to start from the definition of three-point mutual information, and then apply it recursively:

$$\begin{aligned} I(X;Y) &= I(X;Y;Z_1) + I(X;Y|Z_1) \\ &= I(X;Y;Z_1) + I(X;Y;Z_2|Z_1) + ... + I(X;Y;Z_N|\{Z_i\}_{N-1}) + I(X;Y|\{Z_i\}_N) \end{aligned} \tag{2.14}$$

Inverting the formula we get:

$$I(X;Y|\{Z_i\}_N) = I(X;Y) - I(X;Y;Z_1) - I(X;Y;Z_2|Z_1) - ... - I(X;Y;Z_N|\{Z_i\}_{N-1}) \tag{2.15}$$

The quantity on the l.h.s. is the one we want compare with 0. By constructing the set $\{Z_i\}_N$, ideally you are, starting from the two-point mutual information, *taking off* some contributions from $I(X,Y)$, on the r.h.s.. This is why the name we propose for the algorithm is 3off2. In our perspective, then, we would like to collect all the Z_i with positive three-points terms, that is, the one which are likely to form a non-v-structure which can participate to information transmission.

2.2.1 The score attribution

In order to get a proper score for each triplet of nodes, we will combine two quantities: the normalized probability of forming a generalized non-v-structures, and the one that the base of the more probable pattern is exactly XY. We use here the likelihood

functions computed before:

$$\begin{aligned} P_{NV}(X,Y,Z) &= \frac{\mathcal{L}(N\mathcal{V}_{XY})}{\mathcal{L}(N\mathcal{V}_{XY}) + \mathcal{L}(\mathcal{V}_{XY})} \\ &= \frac{1}{1+\exp\{-nI(X;Y;Z|\{U_i\})\}} \end{aligned} \tag{2.16}$$

$$\begin{aligned} P_{B=XY}(X,Y,Z) &= \frac{\mathcal{L}(N\mathcal{V}_{XY})}{\mathcal{L}(N\mathcal{V}_{XY}) + \mathcal{L}(N\mathcal{V}_{YZ}) + \mathcal{L}(N\mathcal{V}_{ZX})} \\ &= \frac{1}{1+\frac{\exp\{-nI(X;Z|\{U_i\})\}}{\exp\{-nI(X;Y|\{U_i\})\}} + \frac{\exp\{-nI(Y;Z|\{U_i\})\}}{\exp\{-nI(X;Y|\{U_i\})\}}} \end{aligned} \tag{2.17}$$

Score computation requires the knowledge of the upstream nodes of X and Y. This set is built from scratch during the execution of the algorithm, such that the ranking of each triplet evolves continuously in time.
The score should reflect the likelihood that the triplet X,Y,Z is a non-v structure of basis XY. This is possible if and only if both the conditions expressed for 2.16 and 2.17 are realized. Then we will define the rank as:

$$r(Z;XY) = \min[P_{NV}(X,Y,Z), P_{B=XY}(X,Y,Z)] \tag{2.18}$$

In order to remove each edge in a robust way, we define the *score* of XY by taking the more informative Z:

$$R(XY) = \max_{Z}[r(Z;XY)] \tag{2.19}$$

2.3 Corrections: the Minimum Description Length principle

When dealing with small databases, it is known that the purely ML framework adopted up to now to rank structures can be inaccurate [13].
Coding theory suggest a new typology of measure which is coherent with the *Minimum Description Length* (MDL) principle. The MDL principle reflects the Occam's Razor or some principle of *parsimony* in choosing the best model; it was firstly formulated by Rissanen in 1978 [14]. In his view, the best model is the one which can give the *shortest description* of data.

A natural possibility would be then to assume the Description Length (DL) equal to the *Kolmogorov descriptive complexity*, which measures the quantity of computability

resources needed to specify a given object. Unfortunately, in algorithmic information theory, it can be shown that the descriptive complexity is not a *computable* quantity. Rissanen proposed then to enrich the DL approach of a new interpretation based on *coding theory*, such that the new measure can be thought as the number of binary digits needed to code the data, in order to transmit them.
The coding interpretation relies on the Shannon's *Source Coding Theorem* [15].

In the simplest case, when a family of parametrized distribution is given and the best set of parameters has to be chosen, it is easy to show that the MDL measure coincides in facts with the ML one. In more complex cases, MDL introduces a cost for each degree of freedom, which are finally included in the model only if their introduction significantly improve the fitting with data; otherwise, they are considered to be redundant. As a result, a simpler model is preferred.
In order to introduce the measure which naturally derives from the MDL principle, let us set a few definitions. We will consider the usual set of N variables X_i, which can assume one among r_{Xi} possible values. Let D be the used dataset, consisting of n independent points. Let $\{U_i\}$ be the set of upstream nodes of X and Y, and let $\prod_i r_{U_i}$ be the number of their possible instantiations.

Definition 2.3. The *description length* $L(G, D)$ of the graphical model G given the database D is given by:

$$L(G, D) = \log P(G) - nH(G, D) - \frac{1}{2}k \cdot \log n \tag{2.20}$$

where k is the number of degrees of freedom of the model G and H is the cross entropy between the true distribution and the one encoded by G.

The first term of 2.20 encodes the prior knowledge we can impose on the inference process.
Since it is a measure of the uncertainty in the model, the entropy term (the one we used up to now) tends to decrease by adding nodes and degrees of freedom in the graph. The third term, instead, introduces the cost for the complexity of the model. What we expect is that a model involving many degrees of freedom results to be preferred on a simpler one only if its cross entropy is much smaller than the one of the other model. It is possible to prove that the DL measure has the same property of the Bayesian one in the limit of infinite data samples [16].

Since we did not assume any prior knowledge, by using the DL measure the likelihood function becomes:

$$\mathcal{L}'(G) = e^{-n\sum_{i=1}^{N} H(X_i|PA_i) - \frac{1}{2}k\log n} \tag{2.21}$$

Here we will derive the new threshold, predicted by the MDL principle, that will be used in order to quantitatively assert if $I(X;Y|\{u_i\}) \simeq 0$. This step will circumvent the need for an arbitrary statistical value as in the PC case.
The discovery of the mutual conditional independency would allow us to remove the edge between vertices X and Y. Then we will assume that G is the complete graphical model, while G' is the one in which XY has been removed. It is immediate to show that:

$$\frac{\mathcal{L}'(G)}{\mathcal{L}'(G')} = e^{nI(X;Y|\{U_i\}) - \frac{1}{2}(k_G - k_{G'})\log n} \tag{2.22}$$

Indeed, $I(X;Y|\{U_i\})$ is the price you have to pay for removing the edge. Here if we set $k_G = k_{max}$, then $k_{G'} = k_{max} - (r_X - 1)(r_Y - 1)\prod_i r_{U_i}$ [16]. Then the new graph is more likely than the second one if:

$$I(X;Y|\{U_i\}) < \frac{1}{2n}(r_X - 1)(r_Y - 1)\prod_i r_{U_i}\log n \tag{2.23}$$

which is the condition we will assess during the skeleton reconstruction.

The orientation procedure is led by the presence of v-structures. At the beginning, 3off2 collects all the open and closed triplets of the inferred graph. Among the latter, it is possible to distinguish the v and the non-v-structures through a significantly positive or negative value of their three-point mutual information.

Since for them you can for sure infer the orientation, you look for the most reliable among the v-structures in the list and you orient its edges. Then you look at all the triplets (open and closed ones) which are involved in the procedure. If possible, you orient all the involved triplets, taking care that the process is not creating any oriented loop in the graph. If it is found that the propagation eventually contradicts itself or some previous structures with a higher score in the list, the original v-structure is considered to be *unfaithful* and it is discarded.

2.4 Performances

A brief summary of the performances of the skeleton reconstruction of the 3off2 algorithm can be found in [12]. Several quality parameters have been evaluated: the precision $Prec = TP/(TP + FP)$, the recall $Rec = TP/(TP + FN)$, and the F-score $F_{sc} = (1 + \beta^2)Prec \times Rec/(\beta^2 Prec + Rec)$, for the values of the parameter β equal to 1 or 1/2 [1].

The tests have been performed by the PhD student S. Affeldt on benchmark causal graphs containing from 20 to 70 nodes, as a function of the dimension of the sample n. Typically 3off2 reaches very good levels of precision for a small value of n, if compared with other methods like PC, ARACNE, or Bayesian search. The value of the recall, instead, seems to grow slower; in complexity, however, the F-scores result very often to be the highest one among all the compared algorithms.

2.5 Pseudo-code

[1] TP (True Positives) stands for the number of edges which are present in the true causal model and are correctly inferred by the algorithm; FP (False Positives) is the number of reconstructed edges which are not present in the underlying graph; FN (False Negatives) is the number of links which should be inferred, being part of the true model, but they are missing in the inferred graph.

Algorithm 1: 3off2 Skeleton Reconstruction

In: observational data of finite size N

Out: skeleton of causal graph $\mathcal{G}$

1. Initiation Start with complete undirected graph **forall the** *links* xy **do**
  if $I(x;y) < (r_x - 1)(r_y - 1)\log N/2N$ **then**
    xy **link is** non-essential and **removed separation set** of xy: $\text{Sep}_{xy} = \emptyset$
  else
    find the **most contributing node** z neighbor of x or y and **compute** 3off2 **rank**, $R(xy;z|\emptyset)$
  end
end

2. Iteration while $\exists$ *xy link with* $R(xy;z|\{u_i\}) > 1/2$ **do**
  for *top link xy with highest rank* $R(xy;z|\{u_i\})$ **do**
    expand contributing set $\{u_i\} \leftarrow \{u_i\} + z$
    update contributing nodes and **ranks** of links xz & yz: $R(xz;z'|\{u_i'\})$ & $R(yz;z''|\{u_i''\})$
    if $I(x;y|\{u_i\}) < (r_y - 1)(r_x - 1)\prod_i r_{u_i}\log N/2N$ **then**
      xy **link is** non-essential and **removed separation set** of xy: $\text{Sep}_{xy} = \{u_i\}$
    else
      find **next most contributing node** z neighbor of x or y and **compute new** 3off2 **rank** of xy: $R(xy;z|\{u_i\})$
    end
    sort the 3off2 **rank list** $R(xy;z|\{u_i\})$
  end
end

Algorithm 2: 3off2 Orientation Reconstruction

In: skeleton of causal graph $\mathcal{G}$

Out: oriented graph $\mathcal{G}'$

1. Identify the triplets (x_i, x_k, x_j) that are *unshielded* $(x_i not - x_j, x_i - x_k, x_k - x_j)$ or *closed* $(x_i - x_j, x_i - x_k, x_k - x_j)$

2. For each *unshielded* triplet, estimate its probability of being a *generalized v-structure* or *non v-structure*, and assign a score corresponding to this probability

3. Order the unshielded triplet in decreasing order of their score

4. For each non oriented or partially oriented *unshielded* triplet, apply the following orientation rules:

- *v-structure* $\Rightarrow (x_i \rightarrow x_k)$ and $(x_j \rightarrow x_k)$
- *non v-structure* $\Rightarrow if(x_i \rightarrow x_k - x_j$ (*resp.*, $x_i - x_k \leftarrow x_j)$), do $x_i \rightarrow x_k \rightarrow x_j$ (*resp.* $x_i \leftarrow x_k \leftarrow x_j$)

5. For each *closed* triplets, if the triplet has 2 non converging oriented edges, then orient the 3^{rd} edge to avoid directed cycle.

Chapter 3

Zebra-fish brain imaging

3.1 Introduction

Here we introduce the causal analysis performed on a biological dataset, extrapolated by imaging the whole brain of zebra-fish at larval stage.
Experimental data are provided by the research group of G. Debregeas and R. Candelier, working at the CNRS/UPMC Jean Perrin Laboratory (Paris, France) [17].

Traditional methodologies for neural activity recording are based on *in vivo* single electrode measures. These techniques give access, with high temporal resolution, to the electrical signal emitted by a single neuron. Using micro-arrays of electrodes, moreover, it is possible to record the neural activity of some hundreds of cells at the same time. However, since the number of neurons composing the brain of an adult animal is of several order of magnitudes larger, this methodology cannot give access to interactions between cells in different brain areas; moreover, it does not allow the detection of ensembles of neurons which are strongly functionally related.
In order to study the activity of a larger number of neurons at the same time, one possibility is to optically monitor the fluctuations of Ca^{++} quantity in the cells [18]. This kind of signal is said to be *intrinsic*, since it does not reflects the simple electrical one, and it derives from secondary dynamics which affects the neuronal cell during the spiking process.

The action potential (AP) is mainly due to the sudden depolarization of the cellular membrane. When the neuron is at rest, the channel proteins on the membrane are closed. In this condition, the density of some charged ions, like Ca and Na, is much

larger outside than inside the cell. The electrical ΔV between the membrane and the environment is stable, while it is suddenly lowered during the short duration of a spike ($\sim$ 1 ms). Depolarization causes a large variation in the cytoplasmic free calcium.

If some particular dyes are present inside the cell, it is possible to record with a camera the fluorescence activity linked to the Ca^{++} presence. Those special molecules, called *calcium indicator*, can bind the Ca^{++} ions, and they are provided with a fluorescent protein which makes possible the *in vivo* optical recording.

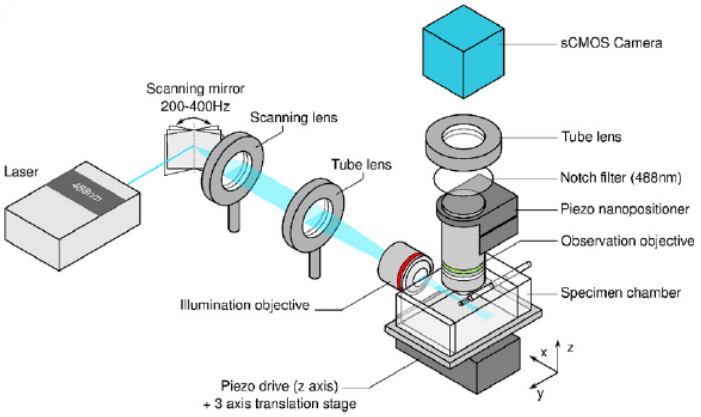
3.2 Zebra-fish brain imaging

Because of its transparency, one animal which results to be suitable for this kind of imaging approach is the zebra-fish in its larval stage. Since in this stage the fish is very small, moreover, its brain (typically $200 \times 500 \times 1000\ \mu m$) consists only of few thousands of neural cells. Nevertheless, the brain in the larval stage is sufficiently developed, and it is able to respond to simple stimuli, like contact, light and movement.
A transgenic line of fishes, whose genome codes spontaneously for the presence of a calcium indicator in cells, was engineered. The genetically encoded dye chosen in this study is the GCaMP3, which has been developed in Hughes Medical Institute, Ashburn (USA) [19]. During the imaging experiment, the larvae used as sample are completely paralyzed, allowing for a single neuron resolution over 30 min or longer.

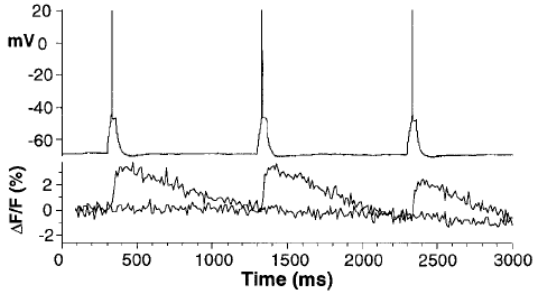
The optical apparatus used to detect the fluorescence, based on Selective-Plane Illumination Microscopy (SPIM), allows to superate the limits in recording imposed by point-scanning imaging techniques, whose low acquisition speed sets, in fact, a limitation on the number on neurons which can be simultaneously observed. Recording is performed through the lateral illumination of the larva with a thin laser-sheet. Different parts of the larva, then, are illuminated and they are observed at the same time. Fluorescence recording can access to neural activity with a single-cell resolution, at least in the non-marginal areas of brain. In the data we will deal with, neurons are monitored with a frequency of 10 Hz, and the total imaging time is of 40 minutes.

3.2.1 Data processing

The main disadvantage of the SPIM method is given by the fact that it deals with fluorescence, which is a more complex signal with respect to the electrical one. Neural



(a)



(b)

FIGURE 3.1: (a) The experimental setup at the Jean Perrin Laboratory [17]. (b) The relationship between AP (top) and the fluorescence signal (down). Those measurements were conducted with one of the earlier calcium indicator which was available in 1999 [20]. Depolarizations are inducted by an external current and the fluorescence signal is detected.

spikes (or action potentials: AP), indeed, have a fixed amplitude of ∼100 mV. The fluorescence signal, instead, can have a varying amplitude, and has a slow rising and decreasing behavior, which makes very difficult the reconstruction of single spikes in low and high frequency trains.

Here we briefly resume the refinement procedure that scientists of Jean Perrin laboratory applied on their data. First, a segmentation algorithm identifies the different regions of interest corresponding to individual somata. Given $F(t)$ the fluorescence signal for a single neuron, the second step is to extract the relative output, $dF/F = (F(t) - b_{slow})/b_{slow}$. b_{slow} is the value of the basement line of each neuron, when it shows no activity. Picture 3.1 shows the expected shape for a fluorescence signal caused by a single neural spike. The spike lasts about 1 ms, while the fluorescence

activity up to $\sim$ 1s. The shape is the one of a double exponential $f(t) \propto e^{\lambda_r} + e^{-\lambda_d}$, with two characteristic times which depends on the performances of the calcium indicator that you are using. In our case, λ_r= 100 ms and λ_d= 600 ms [21].

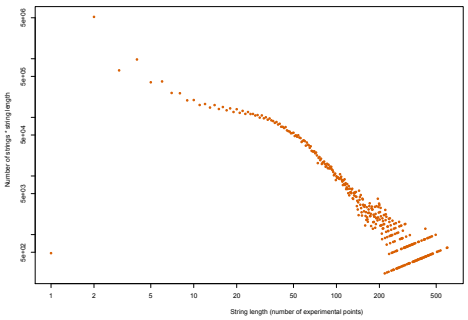
In order to remove the long-tail effect given by the calcium indicator, the signal is processed with a Wiener filter (linear deconvolution) which takes into account also a Gaussian noise. After the filter process, the signal becomes more symmetric and more pronounced, but it is not possible to reduce further the width of the peaks. The temporal resolution of our data τ, then, can be measured as the full width half maximum of peaks. With our experimental setting, $\tau \sim$ 600 ms [22].

3.3 The binary fluorescence signal

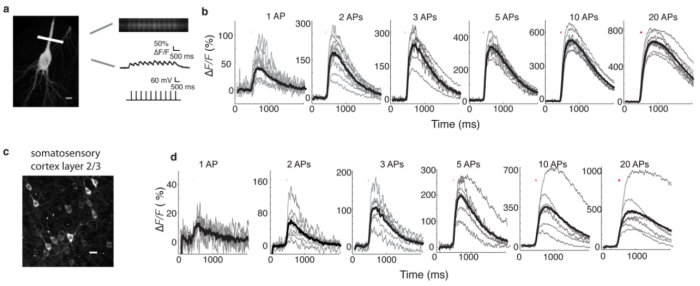
Together with the analog signal obtained with this procedure, the group of the Jean Perrier Laboratory provided us also the relative binary dataset. In the binary case, for each experimental point, neurons are said to be switched on, if their signal is equal to 1, or switched off, if their signal is equal to 0.

The group provided us six databases, which correspond to the *cutting* of the analogical signal with different values of threshold, ranging from 3 to 7. Data with low threshold presents a lot of small strings of 1, consisting of 2-5 experimental points. These very short signals disappear if they apply an higher threshold, and they are assumed to derive also from the background noise. This assumption is consistent with the temporal resolution dictated by the calcium dye, which is estimated to be around 600 ms. In facts, by looking at the length distribution of strings of activity, it is possible to notice a large part of the signal whose duration ranges from 10 to hundreds of experimental points (see figure 3.2). In some cases, like in the hindbrain oscillating area of brain, that we will analyze deeper in the following, neural signals last even longer.

By adopting a higher threshold to cut data, together with small strings of noise, also some longer strings of activity, of duration $\sim$ 1 - 2 s, disappear. It seems possible, then, that together with noise, we are throwing aways some short signals.
This problem could be due to the amplitude of the fluorescence signal generated by the dye, which is not constant in time, neither before nor after the application of the data processing filters. For the calcium indicator used in this analysis, peaks amplitude has been seen to depend on the recording condition (single cell in vitro, or layer of packed cortical cells...) and on the neuron species (see figure 3.2). Furthermore, the



(a)

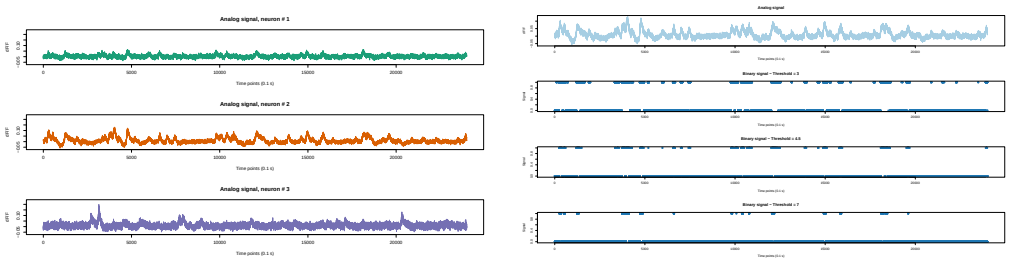


(b)

FIGURE 3.2: (a) Length distribution for all the strings of activity in the dataset. The result is then multiplied for the string length, in order to get a measure of the probability of having a given length by considering one single active point. (b) Performances of the GCamp3 calcium indicator in detecting single AP and small trains of activity [19]. The b trial refers to the isolated neuron in a, while the d one refers to a neuron belonging to the layer of cortical neurons shown in c. The thick line underlines the average signal.

signal amplitude is strongly related to the number of action potentials they correspond to, as the study in picture 3.2 shows. One important consequences is that there can be, in data, some signals of small amplitude (probably related to trains of short time duration) which in principle we want to detect. The risk to avoid is to lose them by cutting the analog signal by using a high threshold (see figure 3.3).

Then it looks reasonable to deal with the dataset corresponding to the lowest values of threshold, and then try to get rid of the very short trains (shorter than the temporal resolution), which are thought to be noise. This task has been performed by using a smoothing filter which relies on a sliding window of width equal to 7 experimental points (700 ms, a bit larger than the temporal resolution). Inside the window, a majority rule is performed: if the number of *active* points is larger than 3, you save a 1 in another string, in the place corresponding to the middle of the window. Otherwise, you save a 0. This results in a non-drastic changing in data, which nevertheless can be



(a) (b)

FIGURE 3.3: (a) Analog signal for three neurons picked at random in the hind-brain oscillator zone (see further). The y scale has been chosen to be the same for three of them. The amplitude of the signal varies for different neurons, and also within the signal of a single cell. (b) Analog signal of the neuron # 2 of 3.3a compared with the relative binary one obtained by setting the threshold to 3, 4.5 and 7. Already with an intermediate value of threshold it is possible to notice that some short peaks of the analog signal are not recorded in the binary one.

necessary, since clustering of neurons will requires logical OR operations of the single neuron signals, which could result in a rapidly propagating noise.

3.4 Inferring brain connectivity

It is widely believed that brain activity relies on a fine balance between functional segregation and integration of information [23]. The first feature to be historically underlined was segregation. This concept was proposed in a first, raw fashion by phrenologists; during the XIX century, later experiments in dogs and monkeys revealed a crucial dependency between brain areas and their specific functions.
However, within few years, it became clear that, given the presence of anatomical communication channels between segregated zones, it was not possible to map uniquely each area of cortex to a specific task. Today, we have widely accepted that each cortical structure, which often codes for a single function, may in facts involve many specialized areas, whose interaction is mediated by signal integration.

Functional segregation have been firmly proved by brain imaging; evidences of integration are harder to collect. Two main approaches have been developed in the recent years, which aim at looking at available data from different perspectives.
The analysis of *functional connectivity* aims at reconstructing the structure of statistical dependencies (mainly correlations) between remote neural activities.
A more powerful tool is given by *effective connectivity*, which refers to the influences

that some neural structures impose on some others, creating dynamical relationships of coupling and causality. While functional connectivity can then be extracted directly from data, functional connectivity has to be inferred and it requires the assumption of a specific model.

3.4.1 Causality inference over time

Within this framework, a widely accepted stream of causal modeling is the one based on time relationships, in opposition to the one based on Bayesian dependency graphs we analyzed so far. This alternative kind of approach is called *dynamical causal modeling* (DCM) [24]. In this dynamical framework, the first step is to set the relationship between the observed response $y(t)$, an exogenous input $u(t)$ and some random fluctuations v. The behavior of y reflects the hidden dynamics of the true physiological state $x(t)$:

$$\begin{aligned} \dot{x} &= f(x,u) + w \\ y &= g(x,u) + v \end{aligned} \tag{3.1}$$

The approach of DCM consists in inverting or fitting this set of equations given some experimental data.
The first possibility is to linearize the equations in 3.1 by using Taylor approximations. You obtain a new family of models whose parameters can be found through autoregression procedures. This approximation lies at the heart of the *Granger causality* approach, which has been widely used in treating fMRI imaging data [25]. A signal X is said to be the *Granger-cause* of Y if the past values of X contain information which helps predicting the value of Y.

The graphical modeling we will adopt, instead, deals with static functional relationships between variables, and then it can be used also when time series are not available. Of course, since our methodology ignores time, it is limited to discovering conditional independencies in DAGs, and it cannot deal with causal feedback loops. This could in principle create problems in dealing with data from brain, since it is widely accepted that brain behaves like a recursive, cyclic structures.

Chapter 4

Inference results

4.1 Neuron clustering

The aim of our analysis is to apply the causality inference procedure to large numbers of neurons, while typically the 3off2 algorithm can run in a reasonable amount of time only when the number of variables is not larger than 100 - 200. Then it has been fundamental to develop a clustering method which allows to deal with small brain regions as a unique variable. This procedure will be applied, during the analysis, on different scales. Since we want to derive local structures which capture the mean features of fluorescence activity, we based the cluster procedure on spatial closeness and neural affinity.

With this purpose, we exploited a hierarchical clustering algorithm (hclust, provided in the R environment), by providing it a modified distance matrix. It consists of a matrix of size $N \times N$, where N is the number of neurons which had to be clustered. We imposed a arbitrarily maximal spatial distance d_{max} and we analyzed each couple of neurons. If they are found not to be neighbours, according to distance threshold which was set, their cell in the matrix is set to -1, otherwise the value of correlation among their signal is computed.
The output of the hclust algorithm is the hierarchical tree (a *dendrogram*) which encodes affinity between groups of neurons. According to the dendrogram it is possible to get the single clusters by cutting it through the setting of a reasonable value for the number of clusters.
While using the algorithm, the agglomeration rule to be used has to be decided. Indeed, hclust starts by considering each point as an isolated leaf; iteratively it associates

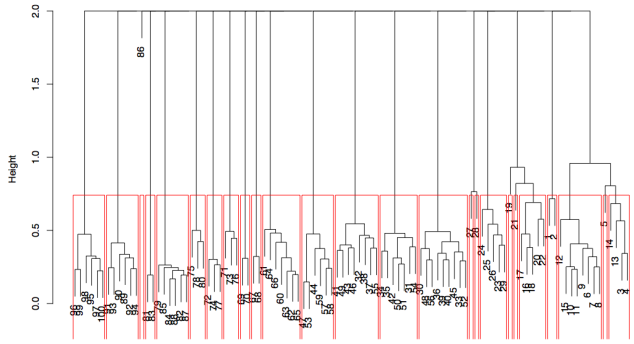


FIGURE 4.1: Example of the dendrograms provided as output from the hclust algorithm (here: complete option). On the y axis distance between elements is reported. By cutting the structure at a given fixed distance you build the single clusters. This dendrogram refers to the hindbrain oscillator area (see next paragraph).

each leaf with the *closest* cluster which has been formed in the previous steps. By setting the agglomeration rule, we chose the measure of distance between groups of variables.
If you choose the complete option, distances will be computed as the maximum distance between the selected object and all the objects inside the cluster. The single option, on the other side, will pick up the nearest neighbour inside the cluster. Its distance from the fixed object is then compute. Other well-known methods are the average one, which measures an average distance between the fixed point and all the variables in the cluster, and the Ward one, which tries to give back clusters of similar dimensions.

The clustering procedure is based on the data smoothed by the filter we described in the previous chapter. When the clusters are ready, their signal is collapsed together in a unique one, through some logical operation among the binary signals.

4.1.1 Clustering rule for signals

In principle we would define a cluster of neurons to be active if a significant fraction of its cells is emitting a fluorescence signal. In practice, to set an exact and reasonable value for this fraction can lead to some difficulties.

Our purpose would be to define a final signal which fully characterizes the cluster as single variable. By adopting a small fraction, as threshold, the risk is to end up with the most active clusters completely switched on through all the duration of the experiment.
On the contrary, by requiring the activity of a large part of cells, it is not difficult to obtain a complete absence of signal after the clustering. Neural activity inside each cluster, indeed, is not homogeneous, and is characterized by a certain number of neurons with lower activity with respect to the other.

The second problem arises when dealing with large cluster, which can be the case, for example, in treating the inference problem from the full brain. It is quite clear, then, that the chosen threshold should change according to the dimension of the cluster and the kind of behavior we expected from the analyzed cells.

4.2 Number of independent data points

In this analysis, we will test the inference algorithm 3off2 on the zebra-fish brain data, and we will compare the obtained results with the graph inferred by PC. One of the main advantages of 3off2 algorithm is that it does not rely on arbitrary thresholds for statistical significance, since the cutoff value is provided naturally within the MDL environment. It is important, then, to have a good estimate for n, which stands for the number of independent data points recorded in the experiment.

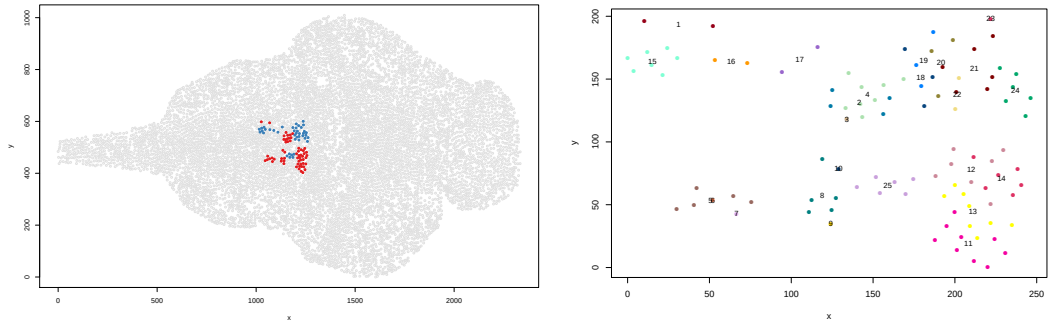
Our dataset contains 24000 experimental points of which, due to the time decay of the fluorescence dye, just a fraction of them are completely independent. If we assume to have one independent point each $\tau \simeq 600$ ms (the temporal resolution), then we get $n = 4000$ effectively independent data points. Since the value for τ is nothing more than an estimation, in our analysis we will allow n to fluctuate a bit around its estimated value, in order to test its robustness.

4.3 The hindbrain oscillator

In the first part of our analysis we focused our attention on restricted groups of neurons. We selected the brain areas about which a particular behavior has been observed during the zebra-fish brain imaging.

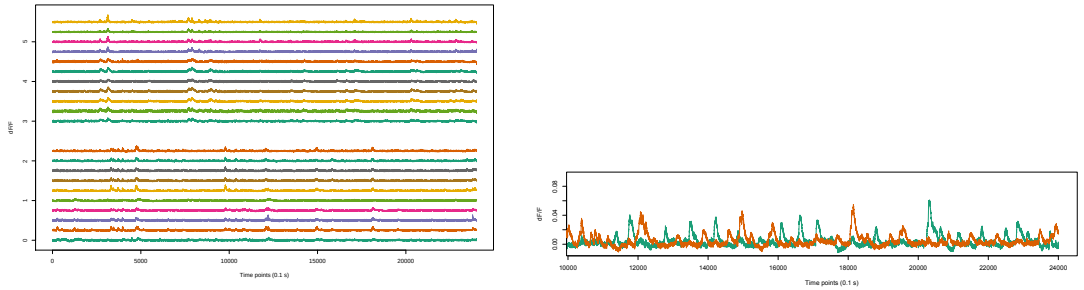
The first set of neurons we looked at is located in the hindbrain of the fish, that is, in the section of the brain which is closer to the spinal cord. It consists of two groups of cells which seem to oscillate in counter-phase, with long period (around $20 \sim 30$ s). This behavior has been founded to be robust and well-marked in different experiments [26], but the function of the so called *hindbrain oscillator* has still to be clarified.

The same structure has been individuated by the experimentalists at the Jean Perrin laboratory (see figure 4.2). The synchronized oscillating behavior is underlined in pictures 4.2.



(a) (b)

FIGURE 4.2: (a) Localization of the hindbrain oscillator in the full brain of the zebrafish larva utilized in the experiment. We will indicate as area 1 the group of neurons plotted in red; the blue group will be indicated with 2. The x and y scale of all the spatial plot we will present have the number of pixels as unity of measure. (b) Hindbrain oscillator: result of the clustering procedure.



(a) (b)

FIGURE 4.3: (a) Analog signal of 10 neurons randomly picked up from area 1 (bottom) and 10 from area 2 (up). (b) Average signal of area 1(red) and area 2 (green). Detail from the last part of the full signal.

After isolating the neurons belonging to the oscillator, we clustered them in a reduced number of points (25). In this case, given the small number or neurons involved in the analysis, it was possible to use the stricter measure for distance between groups provided by hclust (complete one). In picture 4.2 we show the spatial distribution of the obtained clusters. Even if we treated together neurons coming from the two different oscillating areas, the clustering procedure returns small structures which are compatible with the division in the two distinct zones which are oscillating in counterphase.

One of the most important parameter to set in the analysis is the choice of the majority rule needed in order to construct the signal of each cluster. In this first case, clusters

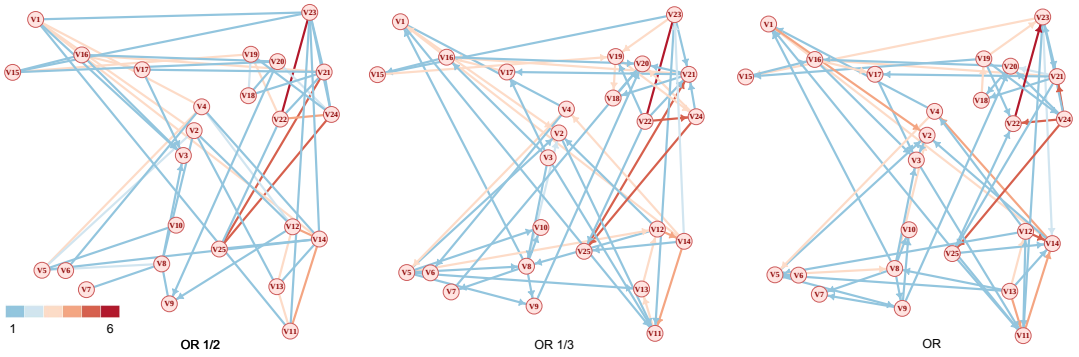


FIGURE 4.4: Hindbrain oscillator: 3off2 inference results ($n = 4000$) for different logical rules. Coloring scale: linear; blue stands for the weak edges, red for the strong ones.

contain a small number of neurons (on average, 4), so the neuronal activity inside each group is expected to be quite synchronized. We tried to investigate the effect of this choice on the final result by selecting quite different rules:

- by OR rule we mean that the logical OR operation has been performed on the signals of all the neurons inside each cluster;
- by OR 1/N rule (here N=2,3) we mean that a 1 is recorded in the final signal only if at the same time at least a fraction of 1/N neurons is active inside the cluster.

In picture 4.4 a comparing plot of the final results obtained by 3off2 is shown.

The coloring code of edges in the graphs reflects the *weight* that the inference procedure assigns to the single link. The exact quantity we used is the conditional mutual information corrected for the MDL threshold ($c = nI(X;Y|\{U_i\}) - (r_X - 1)(r_Y - 1)\prod_i r_{U_i} \log n$), which can be seen as the logarithm of the confidence that the algorithm has in the single edge. The color scale is linear in c and is shown in the figure. Blue colors indicate very weak strength, while the red edges should be considered as the one with higher weight. The position of the vertices resembles the spatial distribution of the clusters which are used as variables for the reconstruction procedure.

The three graphs show many similar features, although a small fraction of edges is reconstructed only in some cases and some orientations are not stable. Edges with higher weight are always inferred, meaning that for them the causality pattern inside signals is quite robust with respect to the clustering rule. The algorithm is not able

to infer many of the orientations in the case of the OR 1/2 rule. Since clusters are very small, it is possible that this strict logical rule could imply the loss the part of information which is hidden in the single neuron activity (which in this case can be determinant). Graph coloring confirms a robust communication within each area (1 and 2), while some links which cross the two regions are present, but they are characterized by low weight. The latter would be the responsibles of the synchronization of the oscillations which has been found in the imaging. In this area of brain, indeed, it is reasonable to expect a loopy pattern of communication between the two regions, which in facts is observed in our results. Graphs are characterized by the presence of many oriented loops which involve neurons of different areas. From this perspective, since our theoretical approach does not assume the presence of this kind of structures, one possibility for further investigations would be the analysis of hindbrain oscillator through a DCM approach.

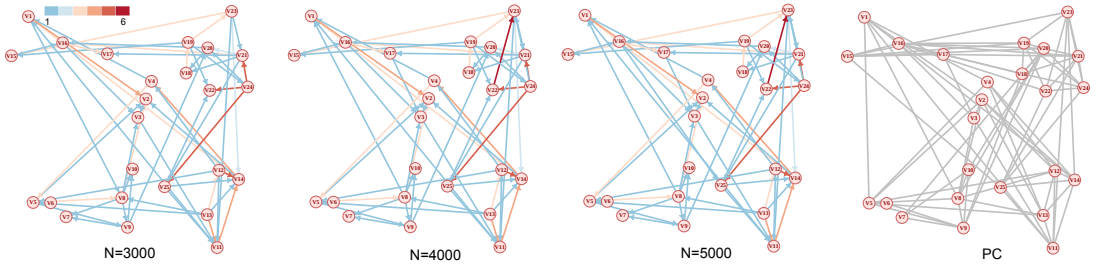


FIGURE 4.5: Hindbrain oscillator: 3off2 inference results (OR rule) for different values of n (first three graphs). Coloring scale: linear; blue stands for the weak edges, red for the strong ones. In the last graph, we show the result provided by the PC algorithm ($\alpha = 10^{-10}$; results were found to be stable in the range $10^{-4} - 10^{-10}$).

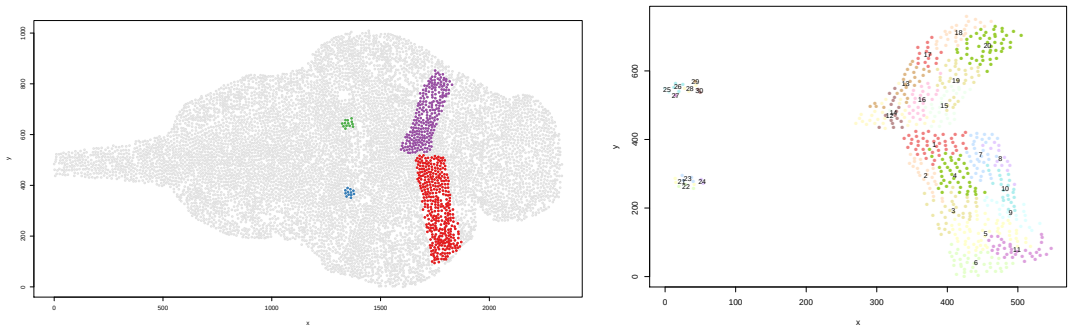
In a second time we tested the behavior of our algorithm by varying the number of independent data points n, which has been estimated to be around 4000. In 4.5 the results are shown; they turn out to be very stable.

As last comparison, we decided to test the behavior of the hindbrain oscillator through the PC algorithm. With this purpose, we used the conservative version implemented in the pcalgo package [6]. This more recent algorithm, which is expected to be more accurate, attributes the orientation of each edge only after performing again all the conditional independency test [27]. In the last element of figure 4.5 the result is shown. PC seems to reconstruct successfully the same kind of general structures inferred by 3off2. However, in this case, PC seems not to be able to infer many orientations on the edges. Our analysis, which propagates orientations starting from the stablest v-structures, seems to perform better in this case.

4.4 The tectum-cerebellum four areas

Another sub-group of neurons which shows an interesting behavior is displaced among two areas of the zebra-fish brain: the hindbrain and the midbrain. It has been first identificated at the Jean Perrin laboratory and consists of 630 cells divided in four small areas, two on the left and two in the right hemisphere of the brain. The first two are located in the optic tectum and they reveal a fairly elongated shape, the second two instead are characterized by a well-rounded shape and they are located in the cerebellum, on the edge of the hindbrain (see figure 4.6).

The experimental team noticed that each of these four areas is characterized by high values of correlation among its neurons. Moreover, they found significant average correlation between the areas 1 and 2, 2 and 3, 3 and 4, but only a weak one between areas 1 and 4 [22].
Then they decided to measure the probability of simultaneous activity between areas, by extracting it from data 4.7. Some preliminary results suggested a causal model in which the two small regions could be the apex of two v-structures: $1 \rightarrow 2 \leftarrow 3$ and $2 \rightarrow 3 \leftarrow 4$. This hypothesis motivated the causality inference analysis we performed.



(a) (b)

FIGURE 4.6: (a) Localization of the four areas of interest in the full brain of the zebra-fish larva utilized in the experiment. The area in red will be noted as 1, the blue as 2, the green as 3 and the violet as 4. (b) Four areas: result of the clustering procedure.

In order to apply the clustering procedure, we kept the neurons belonging to large areas (1 and 4) and to small ones (2 and 3) segregated. In this way, it is possible to set autonomously the number of clusters we desire in the regions 2 and 3; this method revealed to be necessary in order to capture the complexity of communication at smaller scale.

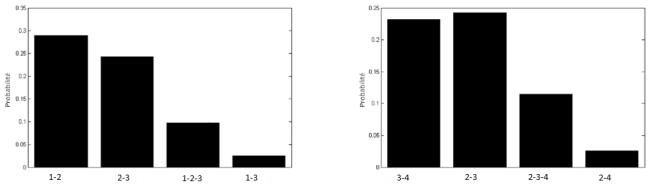


FIGURE 4.7: Probability that the specified areas are emitting fluorescence signals at the same time, for the triplets 1-2-3 (down) and 2-3-4 (up).

The total number of clusters chosen in this analysis is 35. Also in this case, given the high granularity of clusters, it has been possible to use the complete method for the clustering algorithm.

In figure 4.8 we report the results given by 3off2 and PC by using the OR 1/3 rule. In this case, clusters are typically larger than before (18 neurons for cluster, on average), and it looks reasonable that, by simply taking the logical OR between all the signals, a noisy activity is created in the final temporal string.
Also in this case, the strongest links look to be confined inside segregated areas (here into the two large ones, 1 and 4). The weakest edges inferred by 3off2 are the ones which cross the four areas in the diagonal direction. Indeed they are not inferred by PC and they are cut in 3off2 when a lower threshold is set with $n = 3000$. A very weak communication can be noticed also between areas 1 and 4. This discoveries strengthen the hypothesis that the causality pattern in those areas is led by the 1-2, 2-3 and 3-4 pathways.
In the reported results, moreover, it is possible to recognize several simple and generalized v-structures which have as vertices the areas 2 and 3.
By focusing on areas 1, 2 and 3, we underline the presence of simple v-structures of vertices $V23$ and $V22$ ($2 \to 23 \leftarrow 30$, $7 \to 22 \leftarrow 28$ ...) and generalized v-structures which involve more than three vertices ($2 \to 23 \to 21 \leftarrow 28$, $7 \to 22 \leftarrow 23 \leftarrow 30$ ...).
By focusing on areas 2, 3 and 4, it is possible to notice, as example, the structures $16 \to 26 \leftarrow 21$ and $16 \to 26 \leftarrow 28 \leftarrow 23$.

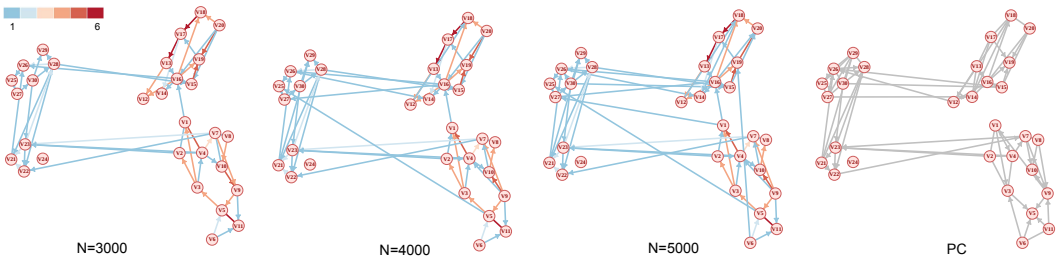


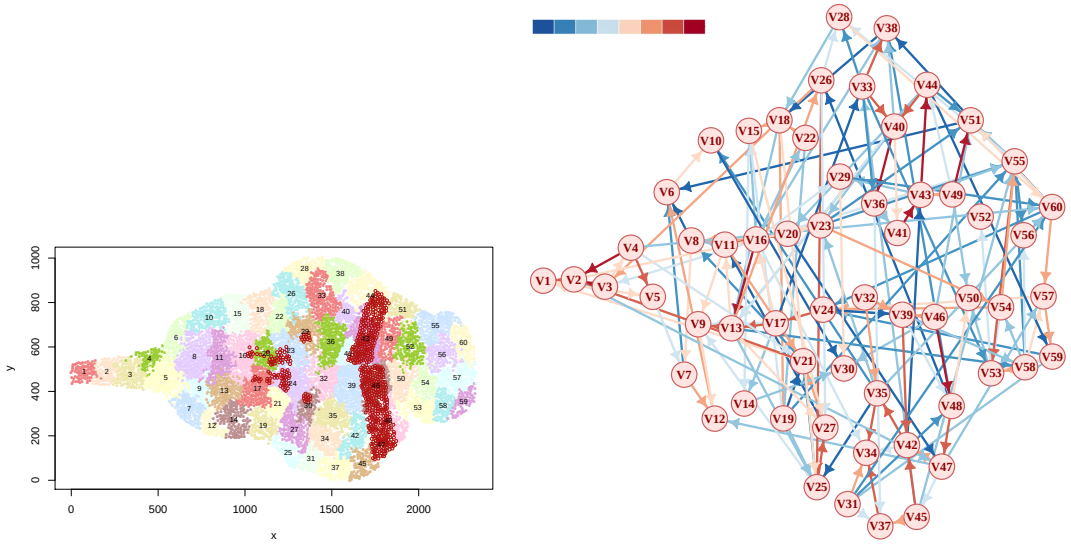
FIGURE 4.8: Four zones: 3off2 inference results (OR 1/3 rule) for different values of n (first three graphs). Coloring scale: linear; blue stands for the weak edges, red for the strong ones. In the last graph, the result provided by the PC algorithm ($\alpha = 10^{-10}$, results were found to be stable in the range $10^{-4} - 10^{-10}$).

We tested the presence of the same structures by using the other logical rules which have been exploited also for the hindbrain oscillator (OR, OR 1/2). In those cases results appear to be less stable by changing the parameter n. In some cases it is possible to reconstruct some v-structures only in the 1-2-3 or only in the 2-3-4 areas. A further investigation about the robustness of the v-structures hypothesis will be conducted on a larger scale, by seeing if it is possible or not to evidence the same local behavior by performing the network reconstruction for the whole brain.

4.5 Full brain results

Once the standard procedure was tested on restricted areas of brain, we decided to apply the same analysis to the totality of the neurons observed in the experimental trial (8082 cells).
By dealing with such a large number, it was necessary to play smartly with the number of clusters to impose and the logical rule to be used. When the number of clusters is around 100, the number of neurons inside each group is large enough to make the activity inside single clusters quite dis-homogeneous. In this framework, therefore, the complete option of hclust gives back a flat dendrogram which cannot be cut in a practical way; we decided then to exploit the less strict average method, which gives back, in this case, a useful result.

In a first step, we dealt with 60, 80 and 100 clusters. In figure 4.9 we show the result obtained for 60 clusters, by using the OR 1/6 rule. For the full brain results, we preferred to set a logarithmic scale for the strength coloring in the graphs. This choice, due to large variations in the confidence level for each edge, allow a better visualization. It is possible to notice, among the weak links (the pink - blue ones), the majority of the links on large scale. They create a series of lateral pathways which connect the right and the left hemisphere of the brain. Moreover, it is possible to notice some diagonal edges of communication, which start from the posterior part of the brain, which is closer to the spinal cord, and spread to the lateral edges of the anterior brain. It looks reasonable that long-scale links correspond, in reality, to chains of short-mediated pathways which flow in the rest of the brain, below or above the plane selected for imaging with the microscope. The strongest (red) links, instead, seem to refer to short scale interactions, and in most cases they describe fluxes of information which from the central areas of brain go to the periphery. In all the results that we will show, it is possible to notice that the largest fraction of strongest links are localized



(a) (b)

FIGURE 4.9: (a) Full brain: result of the clustering procedure. With the small red circles we underline the position of the hindbrain oscillator and of the four areas of interest. (b) Full brain: 3off2 inference results (OR 1/6 rule, $n = 4000$). Coloring scale: logarithmic; blue stands for the weak edges, red for the strong ones.

in lateral sides of the midbrain, in the areas involving the optical tectum. As we will show, neurons in those regions are characterized by high fluorescence activity.

Also in this case, we looked at the results obtained by setting three different values for n (see figure 4.10). Although some very small variations can be underlined, each single edge seem to be very stable. In the PC result, only a smaller fraction of the edges can be oriented (around one half). The general features of the pattern, instead, can be recognized also in this last result.

One interesting feature of 4.9 is the presence of structures which are compatible with the two generalized v-structures we investigated in the four areas discusses in the previous paragraph. In the graph it is possible to recognize a link from area 1 to 2 ($46 \rightarrow 30$) and from 4 to 3 ($43 \rightarrow 29$). Since the small areas 2 and 3 are included in single clusters, it is possible to observe only one edge between them. In this case the link is directed from 29 towards 30, and it is reasonable to think that it would correspond to the strongest direction of communication. This result is not observed for rules which are stricter than the OR 1/6 one. The same behavior has been observed in the same analysis conducted on 80 clusters. In that case, because of the huge number of variables, a graphical interpretation of the network becomes more difficult. By increasing the number of

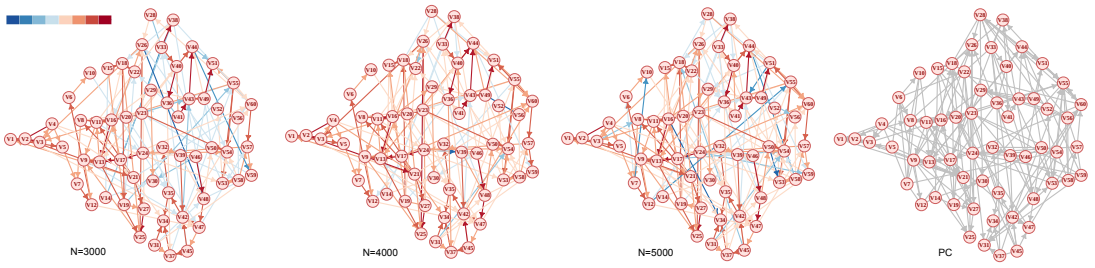


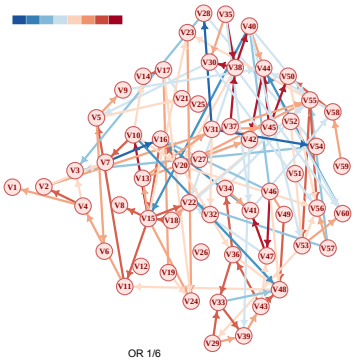
FIGURE 4.10: Full brain: 3off2 inference results (OR 1/5 rule) for different values of n (first three graphs). Coloring scale: logarithmic; blue stands for the weak edges, red for the strong ones. In the last graph, the result provided by the PC algorithm ($\alpha = 10^{-06}$; results are characterized by a slow increase in the number of edges with increasing α).

clusters up to 100, the threshold rule decreases to the OR 1/5. It is reasonable to think that, as the number of clusters increase, their inner activity becomes more and more synchronized. In that condition, in order to keep a fair amount of information, it is sufficient to use a stricter rule for the clustering of signals. Moreover, in picture 4.9, it is possible to notice that clusters which contain the cells of the hindbrain oscillator (here $V16$, $V20$, $V23$, $V17$, $V24$) look to be causally connected also at this larger scale.

At this stage we decided to test the choice we made about the threshold used for the binarization. Then we performed the same kind of analysis (starting from the application of the smoothing filter, up to the inference algorithm) on data obtained with threshold 4.5 (see figure 4.11). In this case it seems that many of the weakest edges inferred in the previous case disappear. This phenomenon could be due to the small loss of information which results from adopting a fairly higher threshold, as we suggested in the previous paragraphs. A less strict OR rule makes possible to infer a bit larger number of weak edges.
By using the new threshold it is not possible to recognize structures which are fully compatible with the ones we found before in the four zones, since some links between them are missing. However, a well assessed behavior seems to be the communication pathway from area 3 to 2, in this case included in clusters 31 and 32.

4.5.1 Most active neurons analysis

In order to better understand the consequences of the clustering procedure, we decided to filter out the neurons which are characterized by a lower activity. It looks reasonable, indeed, that during this procedure the silent neurons are forced to join one of the closest



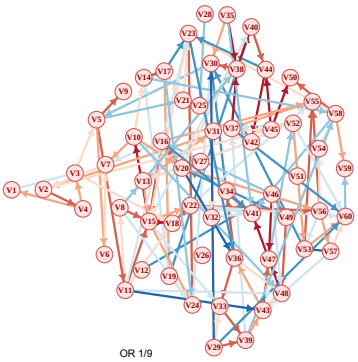
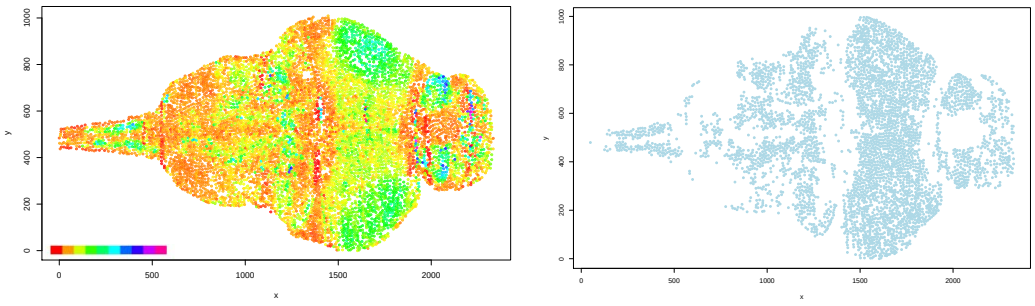


FIGURE 4.11: Full brain, with threshold 4.5: 3off2 inference results ($n = 5000$). Coloring scale: logarithmic; blue stands for the weak edges, red for the strong ones.

clusters. If a proper logical OR rule is not settled, their very low signal could hide the fluorescence dynamics of single active neurons inside each group. In map 4.12 we show the spatial distribution of the active and the silent neurons. Active neurons seem to be located mostly in the spinal cord, in the middle hindbrain and on the lateral edges of the midbrain.

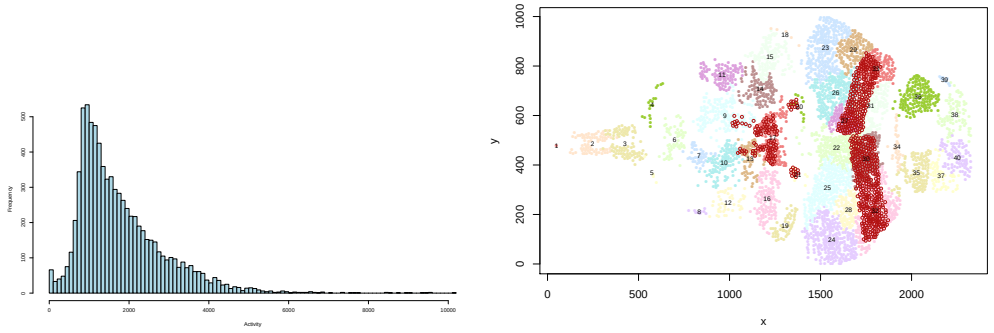
The activity distribution shows a single peak which turns out to be located not far from the median of the distribution. At this value we imagined to cut the distribution, and we kept for a further analysis only the signals belonging to the most active half of the 8082 neurons. On this new neuronal population we performed the same procedure of inference. In this frame it was possible to focus on a smaller number of clusters, which enables to an easier graphical interpretation.



(a) (b)

FIGURE 4.12: (a) Spatial distribution of the most active neurons in the full brain (indicated in green, blue, violet). Neurons with very weak or absent activity are plotted in yellow and red. (b) The half of neuronal population with higher activity, which has been considered for the further analysis.

In this case the results turn out to be a bit more robust by changing the logical rule adopted to cluster signals (see figure 4.14). This could be due to a more homogeneous dynamics inside each clusters, which contains now only the right tail of distribution in



(a) (b)

FIGURE 4.13: (a) Distribution of the activity of all the neurons. On the x axis, the number of time points during which the neuron is active is reported (total number of points: 24000). (b) Full brain, most active neurons: result of the clustering procedure. With the small red circles we underline the position of the hindbrain oscillator and of the four areas of interest.

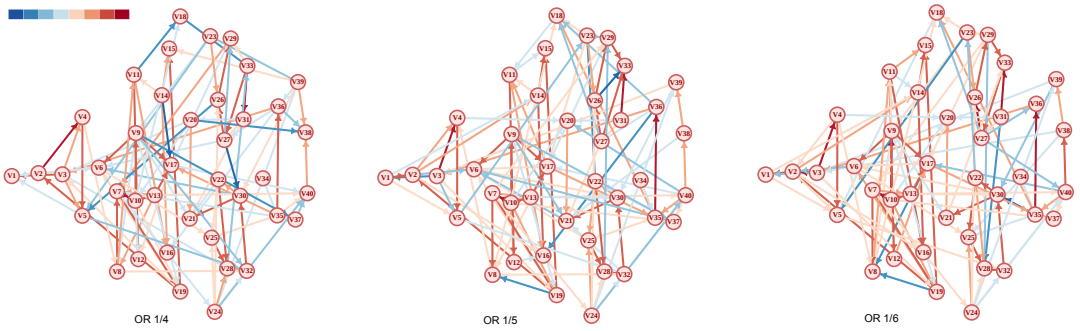


FIGURE 4.14: Full brain, most active neurons: 3off2 inference results ($n = 4000$). Coloring scale: logarithmic; blue stands for the weak edges, red for the strong ones.

4.13.

Neurons belonging to the areas of interest (hindbrain oscillator, four areas) belong to the fraction of active neurons. It is possible, then, to test their behavior in this new framework where a possible source of noise has been removed. In all the results shown in 4.14 the generalized v-structures of the cerebellum-hind brain can be recognized. If we focus on the OR 1/5 rule, it is possible to see the edge between areas 2 and 3 ($20 \rightarrow 21$), the two edges from 1 to 2 ($30/32 \rightarrow 21$) and the two from 4 to 3 ($31/27 \rightarrow 20$).

In a second time we focused on the common features of the communication patterns emerging in the results in 4.14. By analyzing the information pathways connecting the anterior to the posterior part of the brain (from right to left in the graph), it is possible

to notice the presence of a causal pathways which originates from the central regions of the posterior hindbrain and flows thought the root of the spinal cord (clusters 1, 2, 3, 6). Vertices 3 and 6 are, furthermore, the target of many edges originating in the very peripheral areas of the hindbrain and the midbrain. It is not excluded that those longer edges could be mediated by further connections in the hindbrain above or below the imaging plane. It is known that the posterior part of the zebra-fish hindbrain contains the *descending neurons*, which constitute the main pathway linking the brain to the spinal cord, whose cells control the motor reactions of the fish. The role of the hindbrain as an important area of reception and elaboration of information is nowadays well assessed [28]. One interesting feature, which can be underlined in the results obtained with the OR rules 1/5 and 1/6, is the presence of some hindbrain clusters characterized by high values of betweenness centrality ($V6$, $V17$, $V12$, $V11$)[1]. One of the robust features of the communication pathways from the posterior to the anterior areas of brain is instead the tendency to spread immediately, towards the peripheral edges of the hindbrain (clusters 4 and 5), the information coming from the vertices of the spinal cord. Information is then projected to the frontal regions of brain (clusters 32, 37, 40, 38) through long edges whose complex activity can be understood only through a 3D analysis.

4.6 Topological analysis of networks

The possibility to access to full-data brain, given by functional MRI and fluorescence imaging, has produced, over the last year, many experimental results which account for brain functional and effective connectivity. They can be obtained by affinity analysis, causal modeling or simulation.
An emerging behavior, which results to be quite robust among all the trials, is related to some particular topological features which can be recognized in functional graphs. The knowledge of these parameters would allow to deal with brain networks as models of a complex network, of which it is possible to test efficiency, robustness, vulnerability... In practice, graph topology can be quantitatively quantified by a large variety of measures; however, it is still not clear which are the most appropriate ones for the analysis of brain connectivity.

The simplest quantity to access is the *degree distribution* of the network. The degree of a node is the number of edges which are entering or exiting from it.

[1] The *betweenness centrality* of a node measures how many of the shortest paths between all other node pairs in the network pass through it.

Random networks are characterized by a symmetrically centered, Gaussian distribution, indicating that most nodes are connected through an average number of links. Other more complex networks, instead, show asymmetric distributions, often marked by fat tails towards high degrees [29].
In particular, some high-definition fMRI studies have suggested that brain networks (with voxel resolution) could show evidences of *scale-free* organization [30], which is reflected in degree distribution which resembles a power law [31] [32]. In other studies, instead, the power law tails are found to be exponentially truncated [33]. Studies about the very small ($\sim$ 300 neurons) *Caenorhabditis elegans* brain have revealed, instead, purely exponential decays [29].

A further well-assessed behavior in brain networks seems to be the *small-world* architecture (see, e.g. [34], [35]). This term was used for the first time by Watts and Strogatz [36], and since that moment small-worldnesss features have been recognized in a variety of complex networks belonging to different fields.
A small-world network is characterized by a highly clustered small scale structure which allows, nevertheless, the presence of short pathways between all the couple of nodes in the graph. The distance between two randomly picked vertices scales indeed like $\sim \log N$, where N is the total number of nodes. Clustering occurs when neighbor nodes which are linked to the same target have a high probability to be connected to each other. Short pathways, instead, are possible because of the presence of highly connected vertices (hubs), which create fast connections among different clusterss.

Small-worldness is often measured through the method indicated by Walsh (1999). In this framework, we define the small-worldness index as the ratio $S = \gamma/\lambda$, where γ is the ratio between the clustering coefficient [2] of the given graph and the one computed on a random network with the same number of edges and nodes, while λ is the ratio between the average path length in the target graph divided by the average path length of the random one.

For the zebra-fish brain, one known result is the set of parameters emerging from the functional graphical model designed by Stobb and Peterson [37]. This model, acting at a cellular level, combines together part of the existent anatomical knowledge and some stochastic rules.
For such a network, authors found a broad degree distribution, characterized by long

[2] The *clustering coefficient* (or transitivity) is a measure of the probability that two adjacent vertices of a given node are connected. It can been computed t/T, where t is the number of triangles connected to the node and T the number of triplets centered on the vertex.

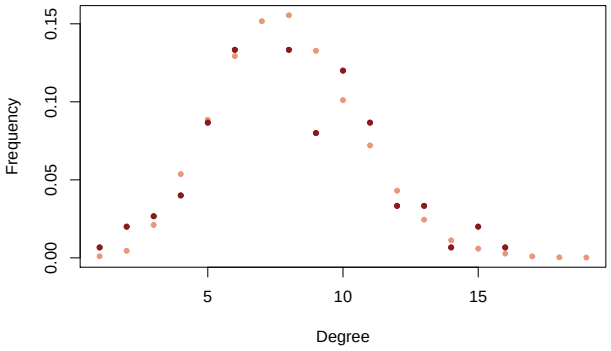


FIGURE 4.15: Degree distribution for 150 Clusters - OR 1/5 (dark red dots). The pink dots show the average degree distribution for the 100 generated random graphs.

tails but distant from the power law behavior. Moreover, they found a clustering coefficient significantly larger than one: (4.13 ± 0.01).

4.6.1 Results

We aimed at extracting the same kind of parameters from the graphs we inferred in our analysis (full brain results). For all the calculations, we used the tools implemented in the igraph package of R. For all the measures concerning random graphs, we generated 100 networks with the same procedure, and then we obtained an average value. Since orientations were seen not to be completely stable by tuning the analysis parameters, and since some undirected edges are present in the results, we decided to extract topological parameters from the skeleton of the inferred graphs.

In table 4.1 some topological parameters are reported. They refer to results obtaining by setting $n = 4000$. In facts, a weak dependency on n has been observed: the small-worldness index has been seen to decrease slowly as n increases.
The inferred graphs seem to have a small-worldness index which is, in all the cases, significantly larger than one. This result is mainly due to a large value for the clustering coefficient (which would be the same in the directed version of the graph). A quite robust behavior seem to be the increase of the small-worldness coefficient together with the number of clusters, as variables become closer and closer to single neurons. This increasing behavior seems to be compatible, in principle, with the value found in [37], which refers to isolated cells.
It is important to recall, however, that our results refer to a 2D network, in which vertical connections with other areas of brain are not included.

The degree distribution, instead, does not show any particular features, and it seems to be compatible with the Gaussian behavior (see figure 4.15).

Parameters	**Small-worldness index**	**Clustering coefficient**	**Average path length**
60 Clusters - OR 1/5	1.32 ± 0.01	0.123 (0.090)	2.709 (2.612)
60 Clusters - OR 1/6	1.69 ± 0.01	0.160 (0.091)	2.617 (2.652)
80 Clusters - OR 1/5	1.99 ± 0.01	0.143 (0.071)	2.723 (2.673)
80 Clusters - OR 1/6	1.84 ± 0.01	0.143 (0.075)	2.669 (2.599)
100 Clusters - OR 1/5	2.14 ± 0.01	0.128 (0.058)	2.878 (2.808)
150 Clusters - OR 1/5	2.71 ± 0.01	0.122 (0.043)	3.010 (2.877)
200 Clusters - OR 1/5	3.11 ± 0.01	0.103 (0.033)	3.034 (2.992)

TABLE 4.1: Topological parameters of the inferred networks. The value between parentheses refers to the average value for random graphs generated with the same number of edges and vertices. For the small-worldness index, also an estimation of the uncertainty is provided.

4.7 Conclusions

Fluorescence imaging of neurons, together with light-sheet microscopy, could be a precious tool in investigating the connectome structure in vertebrates. Technological progressions in this field produce a huge amount of data which is now available for scientists. A rigorous inference environment, such as the one offered by graphical causal models, can offer reliable results of intuitive interpretation.

In this report we introduced a new algorithm, which aims at making the traditional approaches more robust against the noise which exists in real data from the biological world. The 3off2 algorithm was tested on the fluorescence data deriving from imaging the brain of the zebra-fish brain, at the Jean Perrin laboratory.

Our approach allowed us to assess simple causality models for small scale structures, like the hindbrain oscillator and the peculiar four areas which are dislocated among the midbrain and the cerebellum. The robustness of the results was tested in a second stage, through a full-brain analysis. In order to fully understand the results obtained at this larger scale, it would be very useful to extend our analysis also to a 3D scale.

During the experimental trial, scientists of the Jean Perrin laboratory have the possibility to vary the height at which the fish brain is scanned. By scanning with high frequency a discrete number of brain layers, it is possible to record the simultaneous activity of neurons at different depth. Unfortunately, a suitable 3D database is still

not available, and we could not perform an extended analysis during the duration of my internship.

One interesting possibility for further investigations is to test the stereotypy of the results obtained for different individuals of zebra-fish. This kind of approach, suggested in [38], seems to give access to weak structures which can be hidden in single-fish analysis. One more tricky task could be to try to test the presence in other species of vertebrates of the areas of interest we analyzed in this report.

In this study, we analyzed the spontaneous brain activity of zebra-fish. Many more complex experimental setups, in which the sample fish is stimulated with specific visual patterns, have been built in the last years [39] [26]. In those cases, connectivity models for vision have been proposed, but a rigorous inference procedure could allow to test those hypothesis.

Finally, it is important to underline the discovery of many recursive structures, at small scale (in the hindbrain oscillator), and at a larger one (in the full-brain results). From this perspective, it could be useful to test, by applying them to the same dataset presented here, causality inference methods which exploit time series (DCM approaches).

Bibliography

[1] T. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. 1992. URL http://arxiv.org/ftp/arxiv/papers/1303/1303.5435.pdf.

[2] C. Meek. Causal inference and causal explanatiom with background knowledge. 1995. URL http://arxiv.org/ftp/arxiv/papers/1302/1302.4972.pdf.

[3] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 0-934613-73-7.

[4] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-77362-8.

[5] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pages 255–270, New York, NY, USA, 1991. Elsevier Science Inc. ISBN 0-444-89264-8. URL http://dl.acm.org/citation.cfm?id=647233.719736.

[6] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012. URL http://www.jstatsoft.org/v47/i11/.

[7] Christopher Meek. Strong completeness and faithfulness in bayesian networks. pages 411–418, 1995.

[8] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. August 1990. URL http://www.hss.cmu.edu/philosophy/techreports/15_Spirtes.pdf.

[9] David Maxwell Chickering, David Heckerman, and Christopher Meek. Large-sample learning of bayesian networks is np-hard. *J. Mach. Learn. Res.*, 5:1287–1330, December 2004. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1005332.1044703`.

[10] G. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[11] David Maxwell Chickering and Craig Boutilier. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[12] S. Affeldt, H. Isambert. Robust inference of conditional independencies from finite observational data. 2014, under review.

[13] Mark H. Hansen and Bin Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96:746–774, 1998.

[14] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[15] J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14, 1986.

[16] Remco R. Bouckaert. Probabilistic network construction using the minimum description length principle. 1994.

[17] Thomas Panier, Sebastian Romano, Raphael Olive, Thomas Pietri, German Sumbre, Raphael Candelier, and Georges Debregeas. Fast functional imaging of multiple brain regions in intact zebrafish larvae using selective plane illumination microscopy. *Frontiers in Neural Circuits*, 7(65), 2013. ISSN 1662-5110. doi: 10.3389/fncir.2013.00065. URL `http://www.frontiersin.org/neural_circuits/10.3389/fncir.2013.00065/abstract`.

[18] Rafael Yuste and Lawrence C. Katz. Control of postsynaptic ca2+ influx in developing neocortex by excitatory and inhibitory neurotransmitters. *Neuron*, 6(3):333 – 344, 1991. ISSN 0896-6273. doi: http://dx.doi.org/10.1016/0896-6273(91)90243-S. URL `http://www.sciencedirect.com/science/article/pii/089662739190243S`.

[19] L. Tian, S. A. Hires, T. Mao, D. Huber, M. E. Chiappe, S. H. Chalasani, L. Petreanu, J. Akerboom, S. A. McKinney, E. R. Schreiter, C. I. Bargmann, V. Jayaraman, K. Svoboda, and L. L. Looger. Imaging neural activity in worms,

flies and mice with improved gcamp calcium indicators. 6:875–81+, 2009. doi: 10.1038/nmeth.1398. URL http://www.ncbi.nlm.nih.gov/pubmed/19898485.

[20] Diana Smetters, Ania Majewska, and Rafael Yuste. Detecting action potentials in neuronal populations with calcium imaging. *Methods*, 18(2):215 – 221, 1999. ISSN 1046-2023. doi: http://dx.doi.org/10.1006/meth.1999.0774. URL http://www.sciencedirect.com/science/article/pii/S1046202399907740.

[21] Friedrich RW. Yaksi E. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved ca2+ imaging. *Nature Methods*, 2006.

[22] S. Wolf. M2 stage report: Imagerie calcique du cervau du poisson zebre.

[23] Karl J. Friston. Functional and effective connectivity: A review. *Brain Connectivity 1(1):13-36*, 2011.

[24] K.J. Friston. Dynamic causal models. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.

[25] C. W. J. Granger. Essays in econometrics. chapter Investigating Causal Relations by Econometric Models and Cross-spectral Methods, pages 31–47. Harvard University Press, Cambridge, MA, USA, 2001. ISBN 0-521-79697-0. URL http://dl.acm.org/citation.cfm?id=781840.781842.

[26] Misha B. Ahrens, Michael B. Orger, Drew N. Robson, Jennifer M. Li, and Philipp J. Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods*, 10(5):413–420, March 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2434. URL http://dx.doi.org/10.1038/nmeth.2434.

[27] Spirtes P Ramsey J., Zhang J. Adjacency-faithfullness and conservative causal inference. *Proceedings of the Twenty-Second Annual Conference on Uncertainty in Artificial Intelligence (UAI-06),*, pages 402–408, 2006.

[28] V. E. Prince C. B. Moens. Constructing the hindbrain: insights from the zebrafish. *Developmental Dynamics 224:1-17*, 2002.

[29] L. A. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proc Natl Acad Sci U S A*, 97(21):11149–11152, October 2000. ISSN 0027-8424. doi: 10.1073/pnas.200327197. URL http://dx.doi.org/10.1073/pnas.200327197.

[30] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. doi: 10.1126/science.286.5439.509. URL `http://www.sciencemag.org/content/286/5439/509.abstract`.

[31] M.P. van den Heuvel, C.J. Stam, M. Boersma, and H.E. Hulshoff Pol. Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain. *NeuroImage*, 43(3):528 – 539, 2008. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2008.08.010. URL `http://www.sciencedirect.com/science/article/pii/S1053811908009130`.

[32] Victor M. Eguiluz, Dante R. Chialvo, Guillermo A. Cecchi, Marwan Baliki, and A. Vania Apkarian. Scale-free brain functional networks. *Phys. Rev. Lett.*, 94: 018102, Jan 2005. doi: 10.1103/PhysRevLett.94.018102. URL `http://link.aps.org/doi/10.1103/PhysRevLett.94.018102`.

[33] Gaolang Gong, Yong He, Luis Concha, Catherine Lebel, Donald W Gross, Alan C Evans, and Christian Beaulieu.

[34] M.D Humphries, K Gurney, and T.J Prescott. The brainstem reticular formation is a small-world, not scale-free, network. *Proceedings of the Royal Society B: Biological Sciences*, 273(1585):503–511, 2006. doi: 10.1098/rspb.2005.3354. URL `http://rspb.royalsocietypublishing.org/content/273/1585/503.abstract`.

[35] Aaron Nagiel, Daniel Andor-Ardo, and A J Hudspeth.

[36] D. J. Watts and S. H. Strogatz. Collective dynamics of'small-world'networks. *Nature*, 393(6684):409–10, 1998.

[37] B. Mozzag E. Gahtan M. Stobb, J. M. Peterson. Graph theoretical model of a sensorimotor connectome in zebrafish. *Nature*, PLoS ONE 7(5): e37292, 2012. doi: 10.1372/journal.pone.0037292.

[38] Whole-brain activity maps reveal stereotyped, distributed networks for visuomotor behavior. *Neuron*, 81(6):1328–1343, March 2014. doi: 10.1016/j.neuron.2014.01.019. URL `http://dx.doi.org/10.1016/j.neuron.2014.01.019`.

[39] Fumi Kubo, Bastian Hablitzel, Marco Dal Maschio, Wolfgang Driever, Herwig Baier, and Aristides B Arrenberg.