Università
di Genova

PhD in Mathematics and Applications

XXXVIII Cycle

Advanced statistical techniques for brain functional connectivity: from network estimation to network comparisons

Candidate:
Laura Carini

Supervisor:
Dr. Sara Sommariva

Borsa di dottorato cofinanziata con risorse dellUnione europea-NextGeneration EU Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 Potenziamento dellofferta dei servizi di istruzione: dagli asili nido allUniversità.

Ma nei campus di tutto il mondo, studentesse e studenti manifestano per il popolo palestinese, ricordandoci che cosa sia (cosa dovrebbe essere), veramente, l'università: non un luogo di perpetuazione del mondo com'è, ma un laboratorio di insorgenza critica che forgi strumenti per cambiarlo, il mondo.

T. Montantari, Libera Università, 2025.

Abstract

Magneto- and electro-encephalography (M/EEG) are two non-invasive neuroimaging techniques with high temporal resolution capable of recording the magnetic field outside the head and the scalp potential generated by the electric currents within the brain. From M/EEG recordings, it is possible to derive brain connectivity networks. In this thesis, we examine two main aspects concerning the study of brain networks. In the first part, we investigate the estimation of brain connectivity networks at the cortical level from M/EEG measurements. In detail, we first present a procedure based on ℓ_1 regularisation that directly estimates the cross-power spectrum between brain regions while controlling the number of false positives. Then, we introduce an approach that exploits structural equation modelling (SEM) and statistical inference for estimating directed cortical networks aiming also to provide information on potential causal relantionships. In the second part, we investigate statistical tools to compare the estimated brain networks. In particular, we focus on the network-based statistic (NBS) tool for comparing groups of brain networks associated with different external factors. We provide a mathematical formulation of the NBS algorithm and present two alternatives to the F-test used when dealing with more than two groups: we compare the Tukey-Kramer test and a Bonferroni-corrected pairwise t-test, both capable of identifying which groups of networks differ significantly. Finally, we will apply the NBS to a dataset of subjects affected by Dementia with Lewy Bodies (DLB) to investigate the relationship between functional connectivity metrics and the core clinical feature characterising DLB patients.

The overall contribution of this PhD is based mainly on five works [1, 2, 3, 4, 5], whereas the present thesis primarily focuses on the first four.

Contents

List of Figures

List of Tables

Introduction

The human brain comprises around 100 billion neurons, linked by 100 trillion synapses. These connections are anatomically structured over multiple scales of space and their interactions are organized over multiple scales of time, representing one of the most complex networks known. The brain is the biological engine that generates our thoughts, feelings and behaviour. Dysfunctions in human brain networks lead to clinical disorders, such as dementia, Alzheimer's Disease and Parkinsonism, which are among the most intractable global health problems [6, 7]. For these reasons, a central goal of neuroscience in recent years has been the understanding and analysis of brain connectivity networks supported by mathematical and statistical tools. In fact, these networks are not directly observed and we can infer information on them mostly by indirect measurements. In mathematics, these are known as inverse problems and they are naturally applied to estimate brain connectivity networks from the given observations. Moreover, brain connectivity networks are influenced by many external factors, such as age, clinical conditions and neurological disorders. This motivates the increasing interest in developing statistical tools for the analysis and inference on networks. Inference might involve the comparison of groups of networks associated to subjects presenting, for example, different neurological disorders, or the correlation of the connectivity network with clinical indices to provide a characterisation of the subject condition.

Different modalities are available to measure brain activity: functional magnetic resonance imaging (fMRI) [8], positron emission tomography (PET)

[9], magnetoencephalography (MEG) [10] and electroencephalography (EEG) [11]. In this thesis we are interested in magneto- and electro-elecephalography (M/EEG). These two techniques measure the magnetic field outside the head and the electric potential on the scalp, respectively. Two main advantages of M/EEG compared to other imaging modalities are the non-invasiveness and the high temporal resolution, of the order of milliseconds. From a mathematical point of view, signals generated by brain activity or representing M/EEG recordings can be modelled, for example, either as multivariate stochastic processes or square-integrable functions. The former represent a finite discretisation over time, while the latter represent continuous models. Besides the type of modelling, our aim is to estimate the interactions between the brain activity $\boldsymbol{X}(t)$ arising from n brain sources. However, brain activity is not directly observed and needs to be derived from the measurements quantified by the m M/EEG sensors, denoted by $\boldsymbol{Y}(t)$. The unobserved brain activity and the observed M/EEG recordings are related by a mathematical model, which in this thesis is assumed to be linear and corrupted by additive noise $\boldsymbol{N}(t)$. The equation describing this framework is $\boldsymbol{Y}(t) = \boldsymbol{G}\boldsymbol{X}(t) + \boldsymbol{N}(t)$, which is used to solve the M/EEG inverse problem, that is, to estimate $\boldsymbol{X}(t)$ from $\boldsymbol{Y}(t)$. Another step is required to quantify the interactions between the estimated brain activity.

The estimation of functional brain connectivity usually involves a two-step procedure: first, the brain activity is estimated by solving the M/EEG inverse problem mentioned above; then, functional brain connectivity is computed in terms of statistical dependencies from the estimated brain activity. The M/EEG inverse problem is ill-posed, therefore regularisation techniques are required for the first step. A number of recent studies demonstrated that, when regularisation is performed via the Tikhonov method [12], the optimal regularisation parameter for brain activity estimation leads to a sub-optimal brain connectivity estimate [13, 14]. Moreover, this two-step procedure returns a large number of false positives, that is, it identifies the presence of statistical interactions between brain regions that are actually independent.

In the last years, several methods have been proposed to avoid such a two-step procedure [15, 16, 17]. In Chapter 3, we propose an alternative approach, presented in [1], that directly estimates functional brain connectivity from the observed M/EEG recordings. To promote sparsity on the solution and consequently reduce the number of false positives, we use the ℓ_1 regularisation through the fast iterative shrinkage-thresholding algorithm (FISTA) [18]. A computational strategy is implemented to carry out the FISTA update in an efficient way.

In many neuroscientific studies, directed brain connectivity is also of interest in order to investigate potential causal relationships among brain sources. Several methods have been proposed in the literature for this purpose, such as Granger causality [19], dynamic causal modelling (DCM) [20], directed transfer function (DTF) [21], partial directed coherence (PDC) and structural equation modelling (SEM) [22]. Among these, SEM models the relationships not only between brain activity and the M/EEG measurements, but also among brain sources. Moreover, it naturally leads to a mathematical formulation that can be used to derive the probability distributions of all the variables involved, and consequently exploit a statistical framework for inference, such as expectation-maximisation (EM) algorithm [23]. In Chapter 4, we introduce a novel approach [2] that uses SEM to estimate directed brain connectivity in the form of directed acyclic graphs (DAGs). We propose to implement the EM algorithm within the DAG with NO TEARS method [24], that estimates DAGs in an efficient way.

Statistical inference for comparing groups of brain networks influenced by different external factors, like neurological conditions, may be done in several ways. For example, it may involve the computation of a global topological measure to investigate between-group differences. Even if this approach provides a useful summary of the network properties, it lacks specificity because it does not provide any information on whether the effects are distributed throughout the brain or are confined to a specific subset of nodes or edges. On the contrary, measures on each element (node or edge) of the network, can

also be computed. Thus, the same hypothesis is tested over many different elements of a brain network. This procedure is known as mass univariate hypothesis testing and naturally gives rise to a multiple comparisons problem. Many standard statistical approaches can be applied to solve the problem of multiple comparisons [25]. Among these, the network-based statistic (NBS) [26] is a network-specific approach that identifies differential connected components across brain networks and simultaneously controls for the family-wise error rate (FWER), when performing mass univariate testing on all connections in a network. When comparing more than two groups of networks, NBS exploits an F-test in the univariate test. Chapter 5 describes the NBS and compares two alternatives to the F-test, namely a Tukey-Kramer test and a Bonferroni-corrected pairwise t-test. This work is reported in [3]. In Chapter 6, we present results reported in [4]. In particular, we apply the NBS toolbox on a EEG dataset of subjects affected by Dementia with Lewy Bodies (DLB), which represents the second most common cause of neurodegenerative dementia after Alzheimer's Disease. DLB patients are characterised by specific core clinical features, such as cognitive fluctuations, visual hallucinations, parkinsonism and REM sleep behaviour disorder (RBD). In this chapter, we explore the relationships between each core clinical feature of DLB and abnormalities in functional brain connectivity, by comparing the connectivity networks associated to DLB patients to those associated with healthy individuals.

Summarising, in Chapter 1 we provide an introduction to brain functional connectivity, ranging from brain functioning to mathematical models of the M/EEG inverse problem and brain connectivity metrics. Chapter 2 is dedicated to the theory of inverse problems, comprehensive of the most common numerical regularisation approaches and the expectation-maximisation algorithm. In Chapter 3 we describe the approach for direct estimation of functional brain connectivity presented in [1]. In Chapter 4, we present some preliminary results on directed brain connectivity estimation [2]. Chapter 5 gives a description of the NBS and we compare two alternative tests to the

use of an F-test in the NBS framework presented in [3]. Finally, in Chapter 6 we show that both the presence of visual hallucinations and RBD is associated with an increase in functional connectivity, as reported in [4].

Chapter 1

Fundamentals of brain networks

Our experience of life goes through the brain. It is the brain that makes us feel pain when we touch fire. Our joys and sorrows, our intelligence, creativity and passions, are all processed via the brain. Indeed, the brain continuously processes real-world inputs perceived through the senses, and actively generates outputs. It can be seen as an active agent which constructs reactions based on sensed data, memories and instincts.

As far as we know, the brain obeys laws of nature, and at the most fundamental level, of physics. The complexity of this powerful system relies on a precise molecular structure and atomic-level operations. Some atoms in the brain are incorporated into relatively stable structures, such as cell membranes and organelles. Their architecture provides the physical basis for information analysis and storage. Other atoms, such as neurotransmitter molecules, together with extracellular fluids, are responsible for real-time responses. Their activity produces our sensations, thoughts, and behaviour [27].

The complexity of the brain thus emerges from the interactions of many interconnected elements. The network structure naturally arising from these connected elements is known as brain connectivity. Understanding brain networks is essential for investigating cognition, neurological disorders, and the computational principles underlying brain function. For this reason, it has

been widely studied by scientists - from neuroscientists and medical doctors to data scientists. Moreover, considerable effort has been devoted to quantifying, visualising and analysing brain connectivity networks. Indeed, brain connections span multiple spacial scales, ranging from the microscopic level of neurons to the macroscopic level of cortical areas. Connections can be classified as chemical or electrical, as well as anatomical, functional or effective connections [27, 25].

In this chapter, we give an overview of brain function and present some techniques used to record it. We also introduce the concept of brain connectivity together with the mathematical tools to describe it. In particular, in Section 1.1, we present some basic notions on the brain structure and its function; in Section 1.2, we describe the techniques of magnetoencephalography (MEG) and electroencephalography (EEG), used to record brain activity. In Section 1.3, we describe the M/EEG forward model. We introduce the mathematical formalism used to describe brain connectivity in Section 1.4.

1.1 The brain

The nervous system has a complex organization that enables the regulation of internal bodily functions and the ability to respond to and interact with the external environment. It is divided into two primary components: the central nervous system, which includes the spinal cord and the encephalon, and the peripheral nervous system, which comprises the sensory and motor pathways. The encephalon is an organ entirely enclosed within the skull and weighs approximately 1400 grams in adults. It plays a crucial role in regulating physiological body functions and processes such as interaction with the external world, memory, thinking, and emotions.

The encephalon is subdivided into the brain, brainstem, and cerebellum. The brain itself is further divided into the diencephalon and the telencephalon, with the former located at the core of the brain and the latter forming its outer portion. The telencephalon accounts for about 80% of the total volume of the

central nervous system and it is divided into two cerebral hemispheres, right and left. These hemispheres are connected by the corpus callosum, which enables communication and coordination between them. The hemispheres are covered by a thin layer of grey matter characterized by numerous folds, the cerebral cortex, which is responsible for information processing.

The cortex of each cerebral hemisphere is divided into four lobesfrontal, parietal, occipital, and temporalas shown in Figure 1.1. Each lobe is associated with specific functions. For example, the frontal lobe includes the motor cortex, responsible for the planning, control, and execution of voluntary movements. On the parietal lobe the somatosensory cortex processes sensory input related to taste, touch, pain, and temperature.

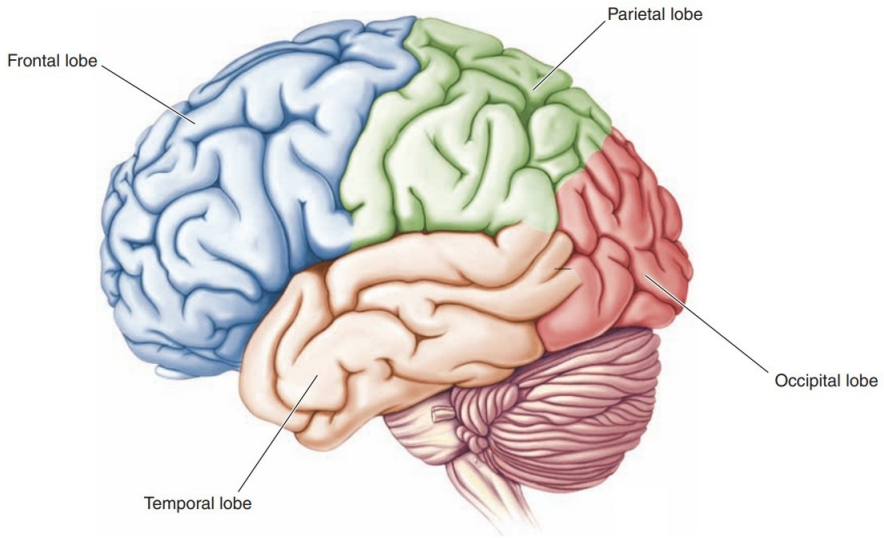

Figure 1.1: The cortex is divided into four lobes: frontal, parietal, occipital and temporal. Each lobe is further divided into smaller areas based on cellular composition and function. Courtesy of [28].

The fundamental units of the nervous system are neurons, cells that transmit nerve impulses. Neurons are composed of a cell body, containing the nucleus and most of the cell's metabolic machinery; dendrites, which are short, filament-like cytoplasmic extensions that, together with the cell body, receive input from other cells; and an axon, a long extension that rapidly conducts

electrochemical signals. The axon terminates at a synapse, which is the interconnection of two neurons and where signals are transmitted from one cell to the next one. Figure 1.2 shows the structure of a typical neuron. Neurons are specialized to receive signals from the internal and external environments and to transmit this information to other neurons or to structures throughout the body.

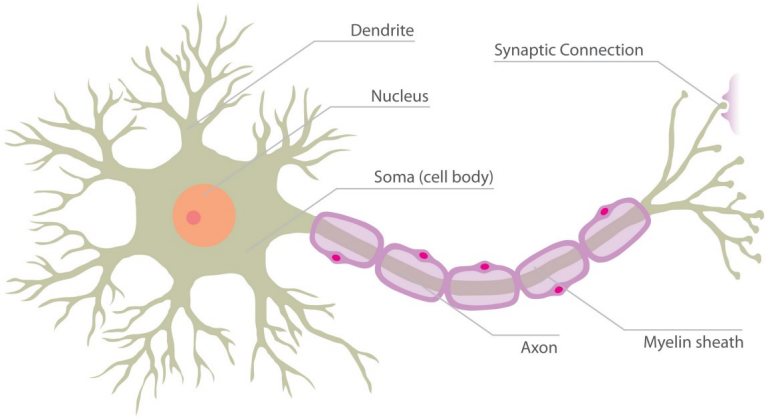

Figure 1.2: The neuron and its main components: the cell body, the dendrites and the axon. Neurons are specialized to receive signals and transmit it to other neurons. Courtesy of [29].

The transmission of information between neurons is linked to electrical phenomena. A difference in electrical potential, known as the membrane potential, exists between the interior and exterior of the cellular membrane, with the interior being negatively charged relative to the exterior at rest. When the axon is stimulated, a rapid and transient reversal of this polarity occurs, during which the interior becomes positively charged. This transient event is known as the action potential and underlies the transmission of information between neurons. Further details on brain function and neuronal communication can be found in [28].

While individual neurons communicate through action potentials, large-scale brain activity emerges from the coordinated firing of neuronal populations. The electrical activity of populations of neurons is often synchronised by a structure in the encephalon, namely the dorsal thalamus, which plays a

central role in the generation of specific brain rhythms. These rhythms usually fall in specific frequency bands and represent an important topic when addressing neurophysiological phenomena. Indeed, they may act as biomarkers used to characterise neurological disorders. The most commonly investigated frequency bands are: the delta band ($[1,4]$ Hz), the theta band ($[4,8]$ Hz), the alpha band ($[8,13]$ Hz) and the beta band ($[13,30]$ Hz) [30]. The dominant posterior alpha rhythm, for example, is an oscillatory rhythm that arises in the posterior cortex when the subject is awake with closed eyes and in a relaxed state. In healty subjects it usually falls within the frequencies of the alpha band. However, it is shifted to lower frequencies when dealing with subjects affected by neurodegenerative diseases such as Alzheimers disease, Parkinsons disease and dementia with Lewy bodies [31].

While electrical activity in individual cells can be assessed only through invasive techniques, on the macroscopic level, measurements of synchronized activity generated by aggregates of neurons can be performed non-invasively through the use of magneto- and electroencephalography (MEG and EEG) techniques. We discuss this further in the next sections.

1.2 The M/EEG devices

The magnetoencephalography (MEG) and electroencephalography (EEG) techniques allow noninvasive, real-time observation of neuronal signaling activity, offering insight into the activity of the brain.

A standard EEG device consists of a cap with multiple electrodes ranging from 19 to 256, placed directly on the head surface. The placement of the electrodes follows the international 10-20 system, though their actual positions may vary slightly between measurements. It is fundamental that the electrode locations are known so that their signals can be properly analysed. Figure 1.3 shows an example of an EEG cap with multiple electrodes.

A MEG system consists of a helmet containing up to 500 low-Tc SQUID (low-temperature Superconducting Quantum Interference Device) sensors, ar-

Figure 1.3: EEG cap with multiple electrodes.

ranged to cover the scalp. To detect extremely small variations in the magnetic field, these sensors must be kept at cryogenic temperatures, for example by immersion in liquid helium. Typically, two types of sensors are used in MEG systems: magnetometers and gradiometers. Magnetometers usually measure the component of the magnetic field perpendicular to the surface of the MEG helmet and are sensitive to fields originating from both nearby and distant sources. Gradiometers measure the spatial gradient of the magnetic field, resulting in a sensitivity that decreases more rapidly with distance [32]. An example of a MEG device and its electrodes is shown in Figure 1.4.

Both EEG and MEG provide excellent temporal resolution on the order of milliseconds, allowing them to track rapid changes in brain activity. MEG generally offers better spatial resolution (millimeter scale) compared to EEGs more limited spatial resolution (centimeter scale) [33]. A more detailed description on MEG and EEG can be found in [27].

1.3 The M/EEG forward problem

The M/EEG forward problem consists in computing the electric potential on the scalp and magnetic field outside the head, generated by neuronal activity within the brain.

The quasi-static approximation of Maxwell's equations provides a model

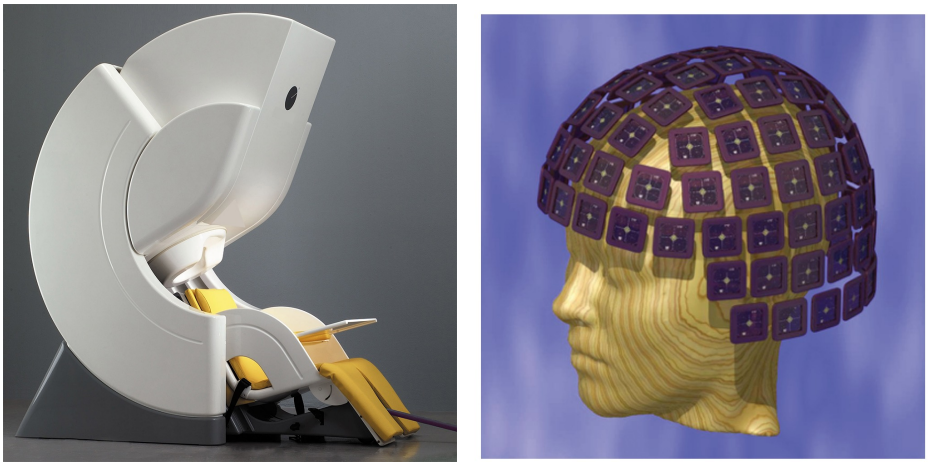

Figure 1.4: On the left: example of MEG device; on the right: examples of MEG sensors. Courtesy of [27].

for computing these two quantities. By assuming that the magnetic permeability of brain tissues is that of a vacuum, Maxwell's equations in the quasi-static case can be written as follows:

$$\nabla \cdot \mathbf{E}(\mathbf{r}, t) = \frac{\rho(\mathbf{r}, t)}{\varepsilon_0}, \tag{1.1}$$

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = 0, \tag{1.2}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r}, t) = 0, \tag{1.3}$$

$$\nabla \times \mathbf{B}(\mathbf{r}, t) = \mu_0 \mathbf{J}(\mathbf{r}, t), \tag{1.4}$$

where $\mathbf{E}(\mathbf{r}, t)$ denotes the electric field, $\mathbf{B}(\mathbf{r}, t)$ is the magnetic field, and $\mathbf{J}(\mathbf{r}, t)$ the electric current at location $\mathbf{r}$ and time t, respectively; ρ is the charge density, μ_0 is the vacuum magnetic permeability and ε_0 is the vacuum electric permittivity [34]. Assuming the quasi-static approximation, which means that the time derivatives are assumed to be zero, and given Equation (1.2), the electric field is irrotational and thus, according to the Poincaré Lemma, there exists an electric scalar potential V, such that

$$\mathbf{E}(\mathbf{r}, t) = -\nabla V(\mathbf{r}, t). \tag{1.5}$$

The brain activity generates an electric current which can be considered as the sum of two contributions: the primary current, $\mathbf{J}^p$, directly related to brain activity, and the volume current, $\mathbf{J}^v$, which is due to the non-null conductivity of the brain, i.e., $\mathbf{J}^v(\mathbf{r},t) = \sigma(\mathbf{r})\mathbf{E}(\mathbf{r},t)$ where $\sigma(\mathbf{r})$ is the conductivity at location $\mathbf{r}$. Thus, the total current is given by

$$\begin{aligned}\mathbf{J}(\mathbf{r},t) &= \mathbf{J}^p(\mathbf{r},t) + \sigma(\mathbf{r})\mathbf{E}(\mathbf{r},t),\\ &= \mathbf{J}^p(\mathbf{r},t) - \sigma(\mathbf{r})\nabla V(\mathbf{r},t).\end{aligned} \tag{1.6}$$

Now, by recalling that the divergence of a curl is zero, if we apply the divergence operator on (1.4) we get

$$0 = \nabla \cdot \mathbf{J}(\mathbf{r},t) = \nabla \cdot \mathbf{J}^p(\mathbf{r},t) - \nabla \cdot (\sigma(\mathbf{r})\nabla V(\mathbf{r},t)). \tag{1.7}$$

From Equations (1.3) and (1.4), the Biot-Savart equation [10] follows as

$$\mathbf{B}(\mathbf{r},t) = \frac{\mu_0}{4\pi}\int_\Omega \mathbf{J}(\mathbf{r}',t) \times \frac{\mathbf{R}}{R^3}\, d\mathbf{r}', \tag{1.8}$$

which, by exploiting Equation (1.6), becomes

$$\mathbf{B}(\mathbf{r},t) = \frac{\mu_0}{4\pi}\int_\Omega \left(\mathbf{J}^p(\mathbf{r}',t) - \sigma(\mathbf{r}')\nabla' V(\mathbf{r}',t)\right) \times \frac{\mathbf{R}}{R^3}\, d\mathbf{r}', \tag{1.9}$$

where Ω is the head volume, $\mathbf{R} = \mathbf{r} - \mathbf{r}'$ and $R = |\mathbf{R}|$.

As previously mentioned, in this section we are interested in the M/EEG forward problem, that is, in computing the electric potential on the scalp and the magnetic field outside the head from the brain activity. By definition of the forward problem, the primary current $\mathbf{J}^p(\mathbf{r},t)$ is assumed to be known, and the goal is to determine the resulting electromagnetic fields. More details regarding the modelling of $\mathbf{J}^p(\mathbf{r},t)$ can be found in Section 1.3.1 .

Given $\mathbf{J}^p(\mathbf{r},t)$, by solving Equation (1.7) we get V and then we compute $\mathbf{B}$ and $\mathbf{E}$ by solving Equations (1.9) and (1.5), thus we finally obtain a solution for the M/EEG forward problem.

A typical assumption is that the head is a piecewise homogeneous conductor, namely it consists of isotropic contiguous regions $\{\Omega_i\}_{i=1,\dots,J}$ of conductivity $\{\sigma_i\}_{i=1,\dots,J}$, such as scalp, skull, cerebrospinal fluid, grey matter and white matter. In this setting, it can be shown [10] that

$$\mathbf{B}(\mathbf{r},t) = \mathbf{B}_0(\mathbf{r},t) + \frac{\mu_0}{4\pi}\sum_{i,j}(\sigma_i - \sigma_j)\int_{\partial\Omega_{i,j}} V(\mathbf{r}',t)\frac{\mathbf{R}}{R^3}\times\mathbf{n}_{i,j}(\mathbf{r}')\,ds', \tag{1.10}$$

and

$$(\sigma_i+\sigma_j)V(\mathbf{r},t) = 2\sigma_0 V_0(\mathbf{r},t) - \frac{1}{2\pi}\sum_{i,j}(\sigma_i-\sigma_j)\int_{\partial\Omega_{i,j}} V(\mathbf{r}',t)\frac{\mathbf{R}}{R^3}\times\mathbf{n}_{i,j}(\mathbf{r}')\,ds', \tag{1.11}$$

where $\partial\Omega_{i,j}$ is the contact surface between regions Ω_i and Ω_j, $\mathbf{n}_{i,j}(\mathbf{r}')$ is the unit vector normal to the surface $\partial\Omega_{i,j}$ at $\mathbf{r}'$ pointing from region i to region j, and

$$\mathbf{B}_0(\mathbf{r},t) = \frac{\mu_0}{4\pi}\int_\Omega \mathbf{J}^p(\mathbf{r}',t)\times\frac{\mathbf{R}}{R^3}\,dv', \tag{1.12}$$

$$V_0(\mathbf{r},t) = \frac{1}{4\pi\sigma_0}\int_\Omega \mathbf{J}^p(\mathbf{r}',t)\times\frac{\mathbf{R}}{R^3}\,dv'. \tag{1.13}$$

The forward problem can be solved following two main steps.

1. First, from Equation (1.13), compute V_0 and find V by solving equation (1.11);

2. Then, from Equation (1.12), compute $\mathbf{B}_0$ and, exploiting V, compute $\mathbf{B}$ by solving Equation (1.10).

We note that both the scalp potential V and the magnetic field $\mathbf{B}$ depend linearly on the primary current $\mathbf{J}^p$. The electric field $\mathbf{E}$ can be compute from V by solving Equation (1.5).

Analytical solutions of these equations are feasible only for very simple head geometries, such as a set of nested concentric homogeneous spherical shells corresponding to the different tissues. In all other cases, we require numerical methods.

1.3.1 The distributed source model and the discretisation of the forward problem

When dealing with real-world applications, the continuous forward model needs to be discretised. Indeed, both the volume occupied by the brain and the volume outside the head undergo a discretisation process.

The so-called distributed source model describes the brain volume, which is uniformly divided into n small parcels. The activity in each brain parcel can be approximated by a point-like source when n is sufficiently large. In this case, the source is described by a dipole and the overall contribution of the dipoles returns the current $\mathbf{J}^p$. Mathematically speaking, each dipole is a vector whose strength and direction represent the intensity and orientation of the primary current in the corresponding brain area. Given this model, the primary current can be expressed as

$$\mathbf{J}^p(\mathbf{r}, t) = \sum_{k=1}^{n} \mathbf{q}_k(t)\, \delta(\mathbf{r} - \mathbf{r}_k), \tag{1.14}$$

where $\mathbf{q}_k(t)$ is the moment of dipole k at time t.

As for the volume outside the head, the sensors of the M/EEG instrument naturally discretise it. Let us consider m the number of sensors of the instrument and denote the measured magnetic field or scalp potential as $\mathbf{y}(t) = (y_1(t), \ldots, y_m(t))$. Since, as previously stated, there is a linear dependence of both the magnetic field $\mathbf{B}$ and the scalp potential V on the primary current $\mathbf{J}^p$, it holds

$$\mathbf{y}(t) = \sum_{k=1}^{n} \mathbf{G}(\mathbf{r}_k)\, \mathbf{q}_k(t) + \mathbf{e}(t), \tag{1.15}$$

where $\mathbf{G}(\mathbf{r}_k) \in \mathbb{R}^{m \times 3}$ represents the leadfield matrices and $\mathbf{e}(t)$ is the measurement noise. This is commonly modelled as a zero-mean Gaussian process, which can be white on the properties of the sensors and the environment.

The l-th column of $\mathbf{G}(\mathbf{r}_k)$ contains the measurement at the sensor level when a unit current dipole is placed at location $\mathbf{r}_k$ and oriented along the l-th

orthogonal direction. Here we assume the orientation of the dipoles to be normal to the brain surface [35]. In this case, the electric current intensities are scalars (we refer to them as $\{q_k\}_{k=1,\dots,n}$) and the leadfield matrices are column vectors (we refer to them as $\{\mathbf{G}_k\}_{k=1,\dots,n}$). Let us now define

$$\mathbf{x}(t) := (q_1(t), \dots, q_n(t)), \tag{1.16}$$

and

$$\mathbf{G} := [\mathbf{G}_1, \dots, \mathbf{G}_n] \in \mathbb{R}^{m \times n}. \tag{1.17}$$

Each column of the leadfield matrix $\mathbf{G}$ describes the pattern of the field generated by a unit dipole at a specific location and orientation. Its derivation strictly depends on the conductivity model of the head previously introduced, ensuring consistency between the continuous forward equations and the discretised operator.

Henceforth, we will refer to $\mathbf{G}$ as the leadfield matrix. Finally, reassembling Equations (1.16) and (1.17) into Equation (1.15), we get

$$\mathbf{y}(t) = \mathbf{G}\mathbf{x}(t) + \mathbf{e}(t). \tag{1.18}$$

This equation forms the basis of the M/EEG inverse problem, that we will discuss in detail in Chapter 2.

1.4 Brain connectivity

1.4.1 Brief history of connectomics

For more than a century, a central concept in neurosciene was that individual neurons are the basic structural and functional units of the nervous system. In the late 1800s, Ramòn y Cajal revealed a wide diversity of neuronal cell types and morphologies, arranged into intricate circuitry. His great intuition extended beyond the anatomical structure of neurons to include their functional *wiring diagrams*, which specify the direction and flow of neural signals.

Indeed, Ramon y Cajal and some others claimed that neurons were discrete cells, connected to one another through the presence of synaptic junctions. They were the first to introduce the idea of physiological mechanisms of neuronal communication - an idea that directly opposed the theory proposed by Camillo Golgi. Golgi was convinced that neurons formed a continuous syncytial connection. However, his theory ultimately proved to be wrong in the 1950s, when electron microscopy resolved the theoretical debate in favour of the discrete setting of Ramon y Cajal.

Around the same time as Ramon y Cajal's studies on microscopic neuronal connectivity, other scientists began exploring the organization of macroscopic networks of interconnected cortical areas. In particular, Theodor Meynert, Carl Wernicke, and Ludwig Lichtheim used network diagrams to summarize white matter connections between cortical areas. These diagrams also illustrated the relationships between symptoms of brain disorders and pathological lesions. Wernicke, for example, proposed an associative theory of brain function, suggesting that higher-order cognitive abilities emerge from the integration of distinct anatomical cortical areas, while psychoses and disorders stem from the pathological disintegration of these areas. These early macroscopic brain networks laid the conceptual foundation for large-scale network models of the nervous system, just as Ramón y Cajal's neuron theory did for cellular-level models.

It is important to note, however, that the quality of the data available at that time for the two different levels of networks, macroscopic and microscopic, was very different. Indeed, high-quality pictures of neurons could be obtained thanks to the technical developments in optics and tissue staining. On the contrary, macroscopic diagrams were derived from postmortem brain appearance and then correlated with the patient's clinical symptoms. For this reason, large-scale networks analysis of neurological and psychiatric disorder were neglected for the first half of the twentieth century. These early ideas anticipated the modern concept of brain networks, but their impact was limited by available technology. With better quality data available to link cortical

anatomy to psychological functions and clinical symptoms, the importance of large-scale networks for understanding brain function and brain disorders, first advocated in the nineteenth century, was securely reaffirmed around 100 years later. The reader may consult [25] for more historial details.

Beyond the evolution of technologies for measuring the nervous system, also the development of mathematical methods in networks science greatly contributed to the growth of studies in the fields of brain connectivity and connectomics.

In the following sections, we briefly introduce what is brain connectivity and the mathematical tools used to describe it.

1.4.2 A roadmap to brain connectivity

As discussed in the previous paragraph, speaking about brain connectivity networks is not trivial. For example, it may involve different brain scales, from the microscopic neuronal level to the macroscopic cortical areas. Moreover, anatomical connections are distinguished from functional connections due to their intrinsically different nature. Indeed, anatomical –or structural– connectivity is defined as an axonal projection from one area to another or, when considering larger scales, by the projection between tracts of axons. Functional connectivity is instead explained as the statistical dependencies between neurophysiological signals. According to the imaging modalities, signals can refer to single neurons or to the activity of neuronal populations, thus spanning different spatial scales. Another class of brain connectivity is given by the effective connectivity. This is defined at the neuronal level and describes the causal influences between neural activity. Effective connectivity often requires a model of how neuronal dynamics generate the measured signal. Directionality is an important aspect for all the three classes of connectivity, with some differences. Structural connectivity is inherently directed, even if some measurements can not resolve its directionality. Functional connectivity can be directed or undirected according to the metric used to compute it, while effective connectivity is always directed, as it is based on a model of

causal interactions between neural systems. For a more complete description on brain connectivity, we refer to [25].

Numerous techniques can be used to assess connectivity depending on the objective of the study. Our main interest lies in functional connectivity that is usually measured via M/EEG and fMRI techniques. Another level of complexity when speaking about functional connectivity consists in considering the measured signals at the sensor level or the reconstructed signals at the cortical brain level. Computing functional connectivity between the measured signals is more straightforward, as it does not require to reconstruct the an estimate of the true brain activity. However, functional brain connectivity at the sensor level is inherently affected by the mixing of contributions from different brain sources. This can be problematic because the measured sensor signals represent a mixture of activity from spatially distributed brain regions, making it difficult to discriminate whether observed dependencies arise from true functional interactions between neural sources. Instead, computing functional connectivity at the source level requires that brain activity must be first reconstructed from the measured signals. This involves solving the ill-posed M/EEG inverse problem. Then, connectivity metrics can be computed from the reconstructed activity of cortical brain sources. We discuss methods for the estimation of brain connectivity of cortical brain sources from M/EEG data in Chapter 3 and in Chapter 4.

1.4.3 Graph theory for brain connectivity networks

Graph theory is a branch of mathematics that studies systems with interacting elements. In particular, it defines how elements are linked together, how these links are organized and possibly evolve. It is an important tool to model, estimate and simulate the topology and the dynamics of brain networks. A graph models such systems by using a set of nodes linked by edges. This simple representation is very flexible and can be used to investigate many aspects of brain organization. Here we revise the formal definition of graph and of few related concept that will be useful in the remaining of the thesis.

We refer to [36] to for a more detailed description of this topic.

Definition 1. *A graph G consists of a collection V of vertices and a collection E of edges, for which we write $G = (V, E)$.*

Each edge $e \in E$ is said to join two vertices $u, v \in V$, called end points. We write $e = (u, v)$. The two vertices are said to be adjacent and the edge e is said to be incident with the two vertices. Vertices are equivalently called *nodes.* In brain network analysis, nodes can represent neurons or bigger cortical areas, according to the scale of interest, while edges can be classified according to whether they represent structural connections or functional and statistical dependencies. In this thesis we will focus on the latter. The graph is said to be weighted when a value, called weight, is assigned to each edge. This can be done by defining a function $w : E \to \mathbb{R}$, so that $\forall(u, v), w(u, v)$ represents the weight on (u, v). The graph is said to be undirected if the edges of the graph do not have any orientation. A directed graph is defined as follows.

Definition 2. *A directed graph G consists of a collection of vertices V and a collection of edges E which are ordered pairs of vertices.*

Given a directed graph, each edge is a directed connection represented by an arrow from a vertex $u \in V$ to another vertex $v \in V$. Vertex u is called the tail while v is the head.

In a directed graph as above, a sequence of vertices $\{v_1, v_2, \ldots, v_k, v_1\}$ such that given each pair (v_i, v_{i+1}) is connected by an edge and no vertex is repeated except for the first one, is said a *cycle*. A directed acyclic graph (DAG) is a directed graph with no cycles.

A simple way to represent graphs is the use of the adjacency matrix. Given a graph G with n nodes, the adjacency matrix A is an $n \times n$ matrix where each element $A_{u,v}$ indicates whether an edge exists between nodes u and v. The adjacency matrix can be binary or weighted. A binary adjacency matrix associated with a graph $G = (V, E)$ of n nodes can be written as

$$A_{u,v} = \begin{cases} 1 & \text{if } (u,v) \in E, \\ 0 & \text{if } (u,v) \notin E. \end{cases}$$

Similarly, a weighted adjacency matrix associated with a graph $G = (V, E)$ of n nodes $A \in \mathbb{R}^{n \times n}$ is defined by

$$A_{u,v} = \begin{cases} w(u,v) & \text{if } (u,v) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, each entry $A_{u,v}$ represents the weight of the edge connecting nodes u and v; if no such edge exists, the corresponding entry is zero.

The adjacency matrix for an undirected graph is symmetric, while for a directed graph it is not.

The total number of edges, $|E|$, referred to as the network degree is a metric of interest when dealing with brain networks. This quantity is calculated as the total number of unique nonzero elements in the adjacency matrix. Let us restrict ourselves to the framework of undirected networks and thus symmetric adjacency matrices. In this case, since connections do not have a direction the edge from node i to node j is identical to the edge from j to i. Therefore, summing all nonzero elements in an undirected, adjacency matrix counts each edge twice, once for (i, j) and once for (j, i). For an adjacency matrix, this can be expressed as:

$$\sum_{u,v} A_{u,v} = 2|E|.$$

Similarly, the degree of an individual node, denoted k_i, is defined as the number of edges connected to that node, which can be computed as the sum of the corresponding row (or column) of the adjacency matrix:

$$k_u = \sum_{v \neq u} A_{u,v}.$$

These definitions will be useful later on in the thesis.

In brain network analysis, a common procedure to construct a binary adjacency metric from a weighted one, is to apply a threshold τ to the weighted matrix, retaining only connections whose strength exceeds the value τ. Thresholding helps separate meaningful connections from noise, highlights more reliable network structure, and reduces computational cost by limiting the number of edges.

However, thresholding has notable drawbacks. Selecting an appropriate threshold is often arbitrary and depends on the researcher, as existing methods each have limitations. Additionally, applying a fixed threshold can lead to unequal network densities across individuals, since networks with higher overall connectivity retain more edges than those with lower connectivity. This variability is problematic because many network measures are sensitive to edge density. Although adaptive thresholds can enforce equal edge counts across networks, they introduce their own analytical challenges.

The method used to measure connectivity can significantly influence the resulting network density. In correlation-based networks, such as those derived from functional connectivity or morphometric covariance, every pair of nodes typically has a nonzero weight, making the network fully connected by definition. In these cases, applying a threshold is essential, as it determines the network density and is a critical preprocessing step.

Connectivity in nervous systems is not distributed uniformly across nodes. Instead, as in many real-world networks, these systems exhibit a skewed, long-tailed degree distribution, indicating the presence of highly connected hub nodes. This heterogeneity suggests that different nodes serve distinct topological roles, with hubs exerting a particularly strong influence on network function. While degree measures provide some insight into node importance, they offer only a partial view. A broader set of metrics, known as centrality measures, can quantify the influence of a node on overall network function by capturing various aspects of its topological position. Examples of such measures include betweenness centrality, closeness centrality, and eigenvec-

tor centrality, though we will not discuss these further here. The interested reader may refer to [25] on the topic.

1.4.4 Towards the definition of cross-power spectrum

Stochastic stationary processes

Definition 3. *Let I be a set and $(I, \mathcal{M})$ a measurable space. A random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is an $(\mathcal{F}, \mathcal{M})$-measurable mapping*

$$X : \Omega \longrightarrow I.$$

Definition 4. *A stochastic process $\{X(t)\}_{t\in S}$ is a time-dependent family of random variables. A multivariate stochastic process of dimension N,*

$$\{\mathbf{X}(t)\}_{t\in S} = \{(X_1(t), \ldots, X_N(t))\}_{t\in S},$$

is a collection of N stochastic processes.

Depending on the set S, the process may be continuous or discrete, and finite or infinite. In this section, only continuous and infinite stochastic processes are considered, that is, $S = \mathbb{R}$. From now on, we will consider the stochastic process $\{\mathbf{X}(t)\}_{t\in S}$ to be real-valued, although the theory can be readily extended to complex-valued processes.

Definition 5. *Let $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ be a multivariate stochastic process of dimension N. At each time instant t, the mean vector is defined as*

$$\boldsymbol{\mu}(t) = \mathbb{E}[\mathbf{X}(t)] = (\mathbb{E}[X_1(t)], \ldots, \mathbb{E}[X_N(t)]) \in \mathbb{R}^N,$$

where

$$\mathbb{E}[X_j(t)] = \int_{\mathbb{R}} x\, p_{X_j(t)}(x)\, dx,$$

and $p_{X_j(t)}$ denotes the probability density function of $X_j(t)$.

Remark 1. *In general, $\boldsymbol{\mu}(t_1) \neq \boldsymbol{\mu}(t_2)$ for $t_1 \neq t_2$.*

Covariance function

Definition 6. *Let $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ be a multivariate stochastic process of dimension N. Given two time instants t and $t+\tau$, with $t,\tau\in\mathbb{R}$, the* covariance function *returns the covariance between the variables $X_j(t)$ and $X_k(t+\tau)$, for any $j,k\in\{1,\dots,N\}$, namely*

$$\begin{aligned} C_{j,k}(t,\tau) &= \mathbb{E}\left[(X_j(t)-\mu_j(t))(X_k(t+\tau)-\mu_k(t+\tau))\right] \\ &= \int_{\mathbb{R}^2}(x_j-\mu_j(t))(x_k-\mu_k(t+\tau))\, p_{X_j(t),X_k(t+\tau)}(x_j,x_k)\,dx_j\,dx_k, \end{aligned}$$

where $p_{X_j(t),X_k(t+\tau)}(x_j,x_k)$ denotes the joint probability density function of $(X_j(t),X_k(t+\tau))$.

By considering all possible pairs of processes of the multivariate process, one obtains a family of covariance matrices

$$\mathbf{C}^{\mathbf{X}}(t,t+\tau)=\mathbb{E}[(\mathbf{X}(t)-\boldsymbol{\mu}(t))(\mathbf{X}(t+\tau)-\boldsymbol{\mu}(t+\tau))^T]$$

$$=\begin{bmatrix} C_{1,1}(t,t+\tau) & C_{1,2}(t,t+\tau) & \cdots & C_{1,N}(t,t+\tau) \\ C_{2,1}(t,t+\tau) & C_{2,2}(t,t+\tau) & \cdots & C_{2,N}(t,t+\tau) \\ \vdots & \vdots & \ddots & \vdots \\ C_{N,1}(t,t+\tau) & C_{N,2}(t,t+\tau) & \cdots & C_{N,N}(t,t+\tau) \end{bmatrix}.$$

The definition of the covariance function allows the introduction of the concepts of strongly and weakly stationary processes.

Definition 7. *A multivariate stochastic process $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ is said to be* strongly stationary *if, for any $t_1,\dots,t_n$, with $\{1,\dots,n\}\subseteq\mathbb{N}$, and for any $\tau\in\mathbb{R}$, the joint probability distributions of $(\mathbf{X}(t_1),\dots,\mathbf{X}(t_n))$ and $(\mathbf{X}(t_1+\tau),\dots,\mathbf{X}(t_n+\tau))$ are identical.*

Definition 8. *A multivariate stochastic process $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ is said to be* weakly stationary *if:*

i. the second-order moment of $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ *is finite for any* t*, that is*

$$\mathbb{E}[(\mathbf{X}(t)-\boldsymbol{\mu}(t))(\mathbf{X}(t)-\boldsymbol{\mu}(t))^T]<\infty,$$

ii. the mean vector does not depend on time, namely

$$\boldsymbol{\mu}(t)=\boldsymbol{\mu}, \qquad \forall\, t\in\mathbb{R},$$

iii. the covariance function depends on the time lag but not on time, that is

$$\mathbf{C}_{\mathbf{X}}(t_1,t_1+\tau)=\mathbf{C}_{\mathbf{X}}(t_2,t_2+\tau), \qquad \forall\, t_1,t_2,\tau\in\mathbb{R}.$$

From condition (iii), for ease of notation, the dependence of the covariance function on time can be dropped, that is,

$$\mathbf{C}_{\mathbf{X}}(\tau)=\mathbf{C}_{\mathbf{X}}(t_1,t_1+\tau).$$

From this point onward, only weakly stationary stochastic processes will be considered and, for simplicity, they will be referred to simply as stationary stochastic processes.

Correlation function

The correlation function quantifies the linear dependence between two variables as a function of the time lag. This quantity is of interest since it can be used, for instance, to identify causal relationships between variables.

Definition 9. *Let* $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ *be a stationary stochastic process. For any* $j,k\in\{1,\dots,N\}$*, the* correlation function *is defined as*

$$R_{j,k}(\tau)=\mathbb{E}\left[X_j(t)\,X_k(t+\tau)\right].$$

In particular, if $j=k$ the function is referred to as the *autocorrelation*, whereas for $j\neq k$ it is called the *cross-correlation*. Since $p_{X_j(t),X_k(t+\tau)}(x_j,x_k)$

denotes the joint probability density function of $(X_j(t), X_k(t+\tau))$, Equation (9) can be written as

$$R_{j,k}(\tau) = \int_{\mathbb{R}^2} x_j x_k \, p_{X_j(t),X_k(t+\tau)}(x_j, x_k) \, dx_j \, dx_k.$$

The family of correlation matrices is defined as

$$\mathbf{R}_{\mathbf{X}}(\tau) = \mathbb{E}\left[\mathbf{X}(t)\,\mathbf{X}(t+\tau)^T\right]$$

$$= \begin{bmatrix} R_{1,1}(\tau) & R_{1,2}(\tau) & \cdots & R_{1,N}(\tau) \\ R_{2,1}(\tau) & R_{2,2}(\tau) & \cdots & R_{2,N}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ R_{N,1}(\tau) & R_{N,2}(\tau) & \cdots & R_{N,N}(\tau) \end{bmatrix}.$$

Remark 2. *Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ be the mean vector of a stationary stochastic process. For any $j, k \in \{1, \ldots, N\}$, the following properties hold:*

i.
$$C_{k,j}(\tau) = R_{k,j}(\tau) - \mu_k \mu_j,$$

ii.
$$R_{k,j}(\tau) = R_{j,k}(-\tau).$$

Note that for zero-mean stochastic processes the covariance and correlation functions coincide.

Definition of cross-power spectrum via correlation

The cross-power spectrum can be defined through the correlation function, via the Fourier transform, or by means of filtering-squaring-averaging operations. In this section, the first two approaches are described. For the latter method, the reader is referred to [37].

Definition 10. *Let $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ be a stationary stochastic process of dimension N such*

that

$$\int_{\mathbb{R}} \left|R_{j,k}(\tau)\right| \, d\tau < +\infty, \qquad \forall\, j,k \in \{1,\dots,N\}.$$

The cross-power spectrum *is a one-parameter family of $N \times N$ matrices, $\mathbf{S}(f)$, whose (j,k)-th entry is defined as*

$$S_{j,k}(f) = \int_{\mathbb{R}} R_{j,k}(\tau)\, e^{-2\pi i \tau f}\, d\tau. \tag{1.19}$$

Note that Equation (1.19) corresponds to the Fourier transform of the correlation function.

Proposition 1. *For real-valued stochastic processes, the cross-power spectrum matrices are symmetric with respect to frequency and Hermitian, namely*

$$S_{j,k}(-f) = S^*_{j,k}(f) = S_{k,j}(f).$$

Proof. The first equality follows from

$$S_{j,k}(-f) = \int_{\mathbb{R}} R_{j,k}(\tau)\, e^{2\pi i \tau f}\, d\tau = \left(\int_{\mathbb{R}} R_{j,k}(\tau)\, e^{-2\pi i \tau f}\, d\tau\right)^* = S^*_{j,k}(f),$$

for all $j,k \in \{1,\dots,N\}$.

On the other hand,

$$\begin{aligned} S_{j,k}(-f) &= \int_{\mathbb{R}} R_{j,k}(\tau)\, e^{2\pi i \tau f}\, d\tau = \int_{\mathbb{R}} R_{j,k}(-u)\, e^{-2\pi i u f}\, du \\ &= \int_{\mathbb{R}} R_{k,j}(u)\, e^{-2\pi i u f}\, du = S_{k,j}(f), \end{aligned}$$

where the last equality follows from Remark 2. □

From Property 1 it follows that $S_{j,j}(f)$ is an even, real-valued function.

Definition of cross-power spectrum via Fourier transform

Definition 11. *Let $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ be a stationary stochastic process of dimension N. The* cross-power spectrum *is a one-parameter family of $N \times N$ matrices whose (j,k)-th entry is defined as*

$$S_{j,k}(f) = \lim_{T\to+\infty} S_{j,k}(f,T),$$

where

$$S_{j,k}(f,T) = \frac{1}{T}\,\mathbb{E}\left[\hat{X}_j(f)\,\hat{X}_k(f)^*\right].$$

Here, $\hat{X}_j(f)$ denotes the Fourier transform of $X_j(t)$ over the interval $[0,T]$, defined as

$$\hat{X}_j(f) = \int_0^T X_j(t)\,e^{-2\pi i f t}\,dt.$$

Proposition 2. *Definitions 10 and 11 are equivalent.*

We observe that both the Fourier transform $\hat{X}_j(f)$ of the real-valued process $X_j(t)$ and the cross-power spectrum $S_{j,k}(f)$ are complex-valued.

1.4.5 Connectivity measures

Coherence and its derivatives

Definition 12. *Let $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ be a stationary stochastic process of dimension N. For each pair $\{X_j(t), X_k(t)\}$, with $j,k \in \{1,\dots,N\}$, the* coherency function *at frequency f is defined as*

$$\mathrm{COH}_{j,k}(f) = \frac{S_{j,k}(f)}{\sqrt{S_{j,j}(f)\,S_{k,k}(f)}},$$

where $S_{j,k}(f)$ denotes the cross-power spectrum between $X_j(t)$ and $X_k(t)$.

Definition 13. *Given the coherency function, the* coherence function *is defined as*

its magnitude, namely

$$\Gamma_{j,k}(f) = \frac{|S_{j,k}(f)|}{\sqrt{S_{j,j}(f)\, S_{k,k}(f)}}.$$

Coherence takes values in the interval $[0, 1]$. When $\Gamma_{j,k}(f) = 0$, the activities of the two signals $\{X_j(t), X_k(t)\}$ at frequency f are linearly independent. Conversely, $\Gamma_{j,k}(f) = 1$ indicates maximal correlation between the signals [38]. It is also worth noting that coherence is sensitive to both phase and amplitude variations in the signals.

One of the main issues that connectivity metrics must address is is the spatial spread of signals from a common source. In EEG, this effect is mainly due to the spread of electric currents through biological tissue, which is known as volume conduction. In MEG, instead, multiple sensors detect the magnetic field generated by the same neural source. In both cases, the activity of a single source can be recorded by multiple sensors, leading to spurious connectivity estimates that do not reflect true interactions between distinct neural sources.

On the other hand, signal spread, together with the intrinsic limitations of the inverse operators used to estimate brain activity, also gives rise to spatial blurring at the source level. In this case, the estimated activity at a given location spreads to other locations, producing artificial and spurious interactions among the estimated source activities.

To mitigate these effects, Nolte and colleagues [39] proposed considering only the imaginary part of the coherency function, which necessarily reflects genuine interactions.

Definition 14. *Let $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ be a stationary stochastic process of dimension N. For each pair $\{X_j(t), X_k(t)\}$, with $j, k \in \{1, \ldots, N\}$, the* imaginary part of coherency *at frequency f is defined as*

$$\text{imCOH}_{j,k}(f) = \frac{\text{Im}\left(S_{j,k}(f)\right)}{\sqrt{S_{j,j}(f)\, S_{k,k}(f)}}.$$

where $S_{j,k}(f)$ denotes the cross-power spectrum between $X_j(t)$ and $X_k(t)$.

Phase Lag Index and its derivatives

Definition 15. *Let $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ be a stationary stochastic process of dimension N. For each pair $\{X_j(t), X_k(t)\}$, with $j,k \in \{1,\dots,N\}$, the* Phase Lag Index *(PLI) at frequency f is defined as*

$$\mathrm{PLI}_{j,k}(f) = \left|\mathbb{E}\left[\mathrm{sign}\left(\mathrm{Im}\left(\hat{X}_j(f)\hat{X}_k^*(f)\right)\right)\right]\right|, \tag{1.20}$$

where sign *is the sign function,* $\mathrm{Im}(\cdot)$ *indicates the imaginary part and $\hat{X}_j(f)$ denotes the Fourier transform of $X_j(t)$.*

The PLI was introduced by Stam and colleagues [40] with the aim of providing a reliable estimate of phase synchronization that is invariant to volume conduction and source leakage. The PLI takes values in the interval $[0,1]$: a value of 0 indicates either the absence of coupling or a coupling centered around $k\pi$, whereas a value of 1 denotes perfect phase synchronization with a consistent nonzero phase lag.

Definition 16. *Let $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ be a stationary stochastic process of dimension N. For each pair $\{X_j(t), X_k(t)\}$, with $j,k \in \{1,\dots,N\}$, the* weighted Phase Lag Index *(wPLI) at frequency f is defined as*

$$\mathrm{wPLI}_{j,k}(f) = \frac{\left|\mathbb{E}\left[\mathrm{Im}\left(\hat{X}_j(f)\hat{X}_k^*(f)\right)\right]\right|}{\mathbb{E}\left[\left|\mathrm{Im}\left(\hat{X}_j(f)\hat{X}_k^*(f)\right)\right|\right]},$$

where $\mathrm{Im}(\cdot)$ *indicates the imaginary part and $\hat{X}_j(f)$ denotes the Fourier transform of $X_j(t)$.*

As a final remark, it should be noted that all functional connectivity measures can provide valuable information on brain function; however, it is good practice to complement such information with that obtained from other analysis approaches. Indeed, it has been shown that features derived from the

power spectrum of individual sources and those derived from functional connectivity metrics are not independent [41], and therefore should be considered jointly rather than separately.

Part I.

Towards the estimation of brain connectivity networks at the cortical level

Chapter 2

Fundamentals of inverse problem theory

This chapter provides an introduction to the mathematical theory of inverse problems, focusing on standard numerical approaches, such as regularisation methods, and iterative optimisation algorithms, like the expectation-maximisation (EM) one. The theory of inverse problems aims to investigate quantities of interest that can not be directly observed, but can instead be estimated from indirect measurements.

Inverse problems play a crucial role in neuroscience and in the study of brain dynamics. When considering MEG or EEG, the devices measure either the magnetic field outside the head or the scalp potentials produced by neuronal currents flowing within the brain. From this recorded data, one can gain information on unobserved brain activity. This is done by defining a model that relates the observed and unobserved variables and then solving the so-called inverse problem. Precisely, the M/EEG inverse problem aims at reconstructing an estimate of brain activity, including the intensity of the signals and the location of the neural sources, as well as functional brain connectivity networks.

We introduce the formal definition of inverse problem in Section 2.1. In Section 2.2 we describe the standard generalised inverse and in Section 2.3

the regularisation methods, such as the Tikhonov and the ℓ_1 approaches. Finally, in Section 2.4, we present the expectation-maximisation algorithm and its properties.

2.1 Inverse problems

Let $\mathbf{T} : X \to Y$ be a linear continuous operator between the Hilbert spaces X and Y, with bounded range $R(\mathbf{T})$, such that

$$\mathbf{T}\mathbf{x} = \mathbf{y}, \tag{2.1}$$

with $\mathbf{x} \in X$ and $\mathbf{y} \in Y$. To solve the inverse problem associated with (2.1) means estimating $\mathbf{x}$ given the data $\mathbf{y}$ and the operator $\mathbf{T}$. This problem is easy to solve when $\mathbf{T}$ is invertible, which is not the case in many real-world applications, as the one of interest related to the M/EEG framework. Let us understand when it is possible to solve the problem defined by (2.1).

Definition 17. *A problem of the form (2.1) is said to be well-posed in the sense of Hadamard [42] if the solution exists, it is unique and it changes continuously with the initial conditions.*

A problem is said to be ill-posed if one of the conditions presented in Definition 17 is not satisfied. We also note that the conditions stated above are strictly linked to the properties of the operator $\mathbf{T}$. Indeed, we have that the existence of the solution is equivalent to the surjectivity of $\mathbf{T}$ while the uniqueness of the solution is equivalent to the injectivity of $\mathbf{T}$. Thus, if $\mathbf{T}$ is invertible, the existence and uniqueness properties are guaranteed. The stability, namely the continuous dependence on the initial conditions, depends also on the the inverse of the operator $\mathbf{T}$: it is guaranteed if $\mathbf{T}^{-1}$ is continuous.

As mentioned before, the M/EEG inverse problem is ill-posed, thus we are interested in understanding how to deal with such a problem to obtain a meaningful solution.

2.2 Generalised inverse

Let us consider an ill-posed inverse problem as in Equation (2.1). Even when the conditions for well-posedness are not met, it is desirable to obtain a meaningful solution. In many practical situations, for example, the uniqueness condition in Definition 17 is violated. In such cases, we seek a solution that only approximately satisfies Equation (2.1), or, when multiple solutions exist, we introduce additional constraints or criteria to select a unique and meaningful one. These considerations lead to the notion of a generalised solution [43].

Definition 18. *Let $\mathbf{T} : X \to Y$ be a linear continuous operator between the Hilbert spaces X and Y with closed range, $R(\mathbf{T})$. The generalised solution of problem (2.1) is equivalently defined as:*

(a)

$$\mathbf{x}^\dagger \in S \text{ s.t. } \|\mathbf{x}^\dagger\| \leq \|\mathbf{u}\| \quad \forall \mathbf{u} \in S,$$

being $S = \{\mathbf{u} \in X : \|\mathbf{T}\mathbf{u} - \mathbf{y}\| \leq \|\mathbf{T}\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x} \in X\}$.

(b)

$$\mathbf{x}^\dagger = \mathbf{T}^\dagger \mathbf{y},$$

where $\mathbf{T}^\dagger$ can be defined as follows. Given an invertible linear continuous operator $\tilde{\mathbf{T}} : X \to Y$, such that $\tilde{\mathbf{T}}_{|N(\mathbf{T})^\perp} = \mathbf{T}_{|N(\mathbf{T})^\perp}$, where $N(\cdot)$ denotes the kernel operator and $\cdot^\perp$ the orthogonal space, $\mathbf{T}^\dagger$ is defined as

$$\mathbf{T}^\dagger = P_{N(\mathbf{T})^\perp} \tilde{\mathbf{T}}^{-1} : X \to Y,$$

being $P_{N(\mathbf{T})^\perp}$ the orthogonal projection on $N(\mathbf{T})^\perp$.

For clarity of notation, we recall that by $\|\cdot\|$ we mean the norm induced by the inner product defined in the Hilbert spaces X and Y, respectively, $\|\cdot\|_X$ or $\|\cdot\|_Y$ being the argument an element of X or Y.

Theorem 1. *Let $\mathbf{T} : X \to Y$ be a linear continuous operator between the Hilbert spaces X and Y, and let $\mathbf{u} \in X$. The following are equivalent:*

i. $\mathbf{Tu} = P_{R(\mathbf{T})}\mathbf{y}$;

ii. $\|\mathbf{Tu} - \mathbf{y}\| \leq \|\mathbf{Tx} - \mathbf{y}\| \quad \forall \mathbf{x} \in X$;

iii. $\mathbf{T}^*\mathbf{Tu} = \mathbf{T}^*\mathbf{y}$, *with* $\mathbf{T}^*$ *being the adjoint operator.*

2.3 Regularisation methods

Let us now consider the special case of a linear continuous operator $\mathbf{T} : X \to Y$, between the Hilbert spaces X and Y, with non-closed range, $R(\mathbf{T})$. In this case, there is no continuous dependence on the data since the non-closeness of $R(\mathbf{T})$ leads to a non-bounded operator $\mathbf{T}^\dagger$. Regularisation methods approximate an ill-posed problem, such as the one above, with a family of neighbouring well-posed problems, and provide a regularised solution.

Definition 19. *Let* $\mathbf{T} : X \to Y$ *be a linear continuous operator between the Hilbert spaces* X *and* Y. *The family of operators* $\{\mathbf{R}_\lambda\}_{\lambda>0}$, $\mathbf{R}_\lambda : Y \to X$, *is said a regularisation algorithm if:*

1. $\mathbf{R}_\lambda$ *is linear, continuous, and bounded;*
2. $\lim_{\lambda\to 0} \mathbf{R}_\lambda \mathbf{y} = \mathbf{T}^\dagger \mathbf{y} \quad \forall y \in R(\mathbf{T}) \oplus R(\mathbf{T})^\perp$.

The parameter λ *is called the regularisation parameter.*

Let us now consider the problem

$$\mathbf{T}^*\mathbf{Tx} = \mathbf{T}^*\mathbf{y},$$

which is equivalent to problem (2.1).

Definition 20. *Let* $\mathbf{T} : X \to Y$ *be an operator. We define the spectrum of* $\mathbf{T}$ *as*

$$\sigma(\mathbf{T}) = \{\omega \in \mathbb{C} \text{ s.t. } \omega\mathbf{I} - \mathbf{T} \text{ is not invertible}\},$$

where I is the identity operator.

We observe that if $\mathbf{X} = \mathbb{R}^N$, $N \in \mathbb{N}$, then

$$\sigma(\mathbf{T}) = \{\omega \in \mathbb{C} \text{ s.t. } \omega \text{ is an eigenvalue of } \mathbf{T}\}.$$

Definition 21. *Consider $\tilde{\mathbf{R}}_\lambda : \sigma(\mathbf{T}^*\mathbf{T}) \subseteq [0, \|\mathbf{T}\|^2] \subset \mathbb{R}$, a continuous operator such that it is an approximation of the function $f(t) = t^{-1}$, $t \in (0, +\infty)$, for $\lambda \to 0$, then*

$$\mathbf{x}_\lambda = \tilde{\mathbf{R}}_\lambda(\mathbf{T}^*\mathbf{T})\mathbf{T}^*\mathbf{y} \tag{2.2}$$

is said a regularised solution.

Theorem 2. *Let $\mathbf{T} : X \to Y$ be a linear continuous operator between the Hilbert spaces X and Y and consider problem (2.1). Let $\{\tilde{\mathbf{R}}_\lambda\}_{\lambda>0}$, $\tilde{\mathbf{R}}_\lambda : (0, \|T\|^2] \to \mathbb{R}$, be a family of real value functions. Then the regularised solution converges to the generalised solution, that is*

$$\mathbf{x}_\lambda = \tilde{\mathbf{R}}_\lambda(\mathbf{T}^*\mathbf{T})\mathbf{T}^*\mathbf{y} \xrightarrow{\lambda\to 0} \mathbf{x}^\dagger = \mathbf{T}^\dagger\mathbf{y},$$

with $\mathbf{y} \in R(\mathbf{T}^\dagger) \oplus R(\mathbf{T}^\dagger)^\perp$, if :

1. $\tilde{\mathbf{R}}_\lambda \xrightarrow{\lambda\to 0} t^{-1} \quad \forall t \in (0, \|\mathbf{T}\|^2]$,
2. *$t\tilde{\mathbf{R}}_\lambda(t)$ is uniformly bounded.*

In the case of noise-free data, a good approximation of the generalised solution can be achieved through regularisation methods. In addition, their role is crucial also in the presence of noise. Suppose our data are affected by noise with intensity δ, that is

$$\|\mathbf{y} - \mathbf{y}^\delta\| \leq \delta,$$

where $\mathbf{y}$ is the noise-free data. The problem now reads

$$\mathbf{y}^\delta = \mathbf{T}\mathbf{x} + \mathbf{n}_\delta.$$

In the case of a linear regularisation algorithm, it holds

$$\|\mathbf{R}_\lambda \mathbf{y}^\delta - \mathbf{x}^\dagger\| \leq \|\mathbf{R}_\lambda \mathbf{T}\mathbf{x}^\dagger - \mathbf{x}^\dagger\| + \delta\|\mathbf{R}_\lambda\|, \tag{2.3}$$

where the first term on the right-hand side is the approximation error due to the use of $\mathbf{R}_\lambda$ instead of the generalised inverse and it tends to zero when λ tends to zero. The second term quantifies the error on the regularised solution $\mathbf{R}_\lambda \mathbf{y}^\delta$ due to the presence of noise and it tends to infinity when λ tends to zero. This implies that for any δ there exists an optimal value of the regularisation parameter $\lambda_{\text{opt}}(\delta)$ such that the right-hand side of expression (2.3) is minimised. We can now define the following.

Definition 22. *A regularisation algorithm $\{\mathbf{R}_\lambda\}_{\lambda>0}$ is said to be regular if, for $\delta \to 0$,*

$$\lambda_{opt}(\delta) \to 0 \quad \textit{and} \quad \mathbf{R}_{\lambda_{opt}}\mathbf{y}^\delta \to \mathbf{x}^\dagger.$$

In the next section we introduce some regularisation algorithms that are of interest for later discussion.

2.3.1 Tikhonov regularisation

Tikhonov introduced the so-called Tikhonov regularisation method in 1943 [44]. It imposes two constraints on the solution of the ill-posed problem (2.1). One condition requires that the desired solution fits well the data and is a good approximation of the generalised solution, namely we ask that $\|\mathbf{T}\mathbf{x} - \mathbf{y}\| \leq \varepsilon$, where $\|\mathbf{T}\mathbf{x} - \mathbf{y}\|$ is the norm of the residual –or fidelity term – and ε can be interpreted as the entropy, which should be small. The second condition is used to find a regular solution by imposing a small energy E, namely $\|\mathbf{x}\| \leq E$. This is called a penalty term.

Definition 23. *Let $\mathbf{T} : X \to Y$ be a linear continuous operator between the Hilbert spaces X and Y and λ be a positive real value. The Tikhonov regularised solution is the minimum of the functional*

$$\Phi_\lambda : X \to \mathbb{R}, \quad \mathbf{x} \mapsto \Phi_\lambda := \|\mathbf{T}\mathbf{x} - \mathbf{y}\|^2 + \lambda\|\mathbf{x}\|^2.$$

Thus the solution reads

$$\mathbf{x}_\lambda = \arg\min_x \left\{ \|\mathbf{T}\mathbf{x} - \mathbf{y}\|^2 + \lambda \|\mathbf{x}\|^2 \right\}.$$

The regularisation parameter λ is a trade-off that weights between the fidelity and penalty terms. It is usually set a-priori but many methods have been proposed to set such a parameter, each one with its advantages and disadvantages [43].

Let us now return to the conditions

i. $\|\mathbf{T}\mathbf{x} - \mathbf{y}\| \leq \varepsilon$,

ii. $\|\mathbf{x}\| \leq E$.

A schematic representation is depicted in Figure 2.1. Two cases can be described. In the first case, among all the solutions that satisfy condition (*i.*), we seek the one with minimum energy, that is $\|\mathbf{x}\| = E_{\min}$. As shown in Figure 2.1(a), such a solution lies on the boundary of $\|\mathbf{T}\mathbf{x} - \mathbf{y}\| \leq \varepsilon$, therefore the smaller ε is, the higher $\|\mathbf{x}\|$ is. In the second case, among all the solutions that satisfy condition (*ii.*), we seek the one with smallest discrepancy, that is $\|\mathbf{T}\mathbf{x} - \mathbf{y}\| = \varepsilon_{\min}$. As shown in Figure 2.1(b), such a solution lies on the boundary of $\|\mathbf{x}\| = E$, therefore the smaller E is, the higher $\|\mathbf{T}\mathbf{x} - \mathbf{y}\|$ is. It is clear that the two conditions can not be satisfied at the same time for any value of ε and E.

Definition 24. *We call the compatibility region the set of solutions that satisfy both conditions (i.) and (ii.):*

$$\{\mathbf{x} \in X : \|\mathbf{x}\| \leq E \text{ and } \|\mathbf{T}\mathbf{x} - \mathbf{y}\| \leq \varepsilon\}.$$

A representation of the compatibility region just defined is given by Figure 2.2.

Theorem 3. *For any $\mathbf{y} \in Y$ and $\lambda > 0$, the Tikhonov functional $\Phi_\lambda(\mathbf{x})$ has a unique*

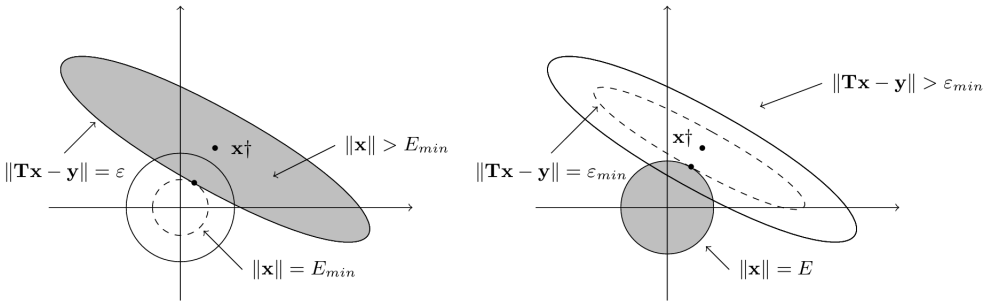

Figure 2.1: Schematic representation of the constraints (i) and (ii). (a) The solution, among all of those that satisfy condition (i), that has minimum energy lies on the boundary of $\|\mathbf{Tx}-\mathbf{y}\| \leq \varepsilon$. (b) Among all the solutions that satisfy condition (ii) the one with minimum discrepancy lies on the boundary of $\|\mathbf{x}\| \leq E$. Courtesy of [45].

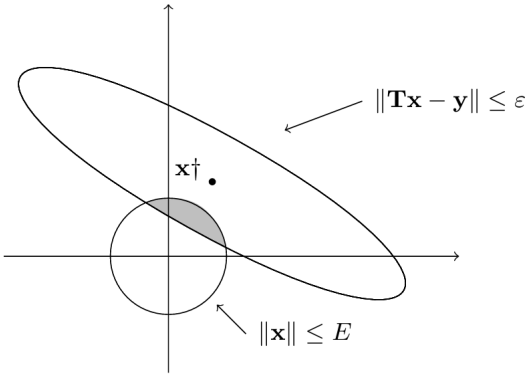

Figure 2.2: Schematic representation of the compatibility region. Courtesy of [45].

minimum point $\mathbf{x}_\lambda \in N(\mathbf{T})^\perp$. *Such a point is the solution of the Euler equation*

$$(\mathbf{T}^*\mathbf{T} + \lambda\mathbf{I})\mathbf{x}_\lambda = \mathbf{T}^*\mathbf{y},$$

where $\mathbf{I}$ *is the identity operator. In particular,*

$$\mathbf{x}_\lambda = (\mathbf{T}^*\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^*\mathbf{y}.$$

We note that the operator $(\mathbf{T}^*\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^*$ is bounded for every fixed $\lambda > 0$, which guarantees the regularised problem is well-posed.

Theorem 4. *The one-parameter family* $\{\mathbf{R}_\lambda\}_{\lambda>0}$ *defined by*

$$\mathbf{R}_\lambda = (\mathbf{T}^*\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}^*$$

is a regular regularisation algorithm.

2.3.2 ℓ_1 regularisation

The ℓ_1 regularisation method is another commonly used regularisation method. Differently from the Tikhonov one that seeks a solution with a small ℓ_2 norm, the ℓ_1 regularisation method promotes sparsity on the solution.

Let us assume X to be in a Banach space [46]. Sparsity is then obtained by considering the ℓ_1 norm instead of the ℓ_2 norm in the penalty term of the minimised functional in Definition 23 . We define the ℓ_1 regularised solution as follows.

Definition 25. *Let* $\mathbf{T} : X \to Y$ *be a linear continuous operator between the Banach space* X *and the Hilbert space* Y *and* λ *be a positive real value. The* ℓ_1 *regularised solution is the minimum of the functional*

$$\Phi_\lambda : \mathbb{R}^N \longrightarrow \mathbb{R}, \quad \mathbf{x} \mapsto \Phi_\lambda := \|\mathbf{T}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{2.4}$$

where we stress that $\|\cdot\|_2$ *is the* ℓ_2*-norm in the Hilbert space* Y*, and* $\|\cdot\|_1$ *is the* ℓ_1*-norm in the Banach space* X*. Thus*

$$\mathbf{x}_\lambda = \arg\min_{\mathbf{x}} \left(\|\mathbf{T}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1 \right).$$

In this case, an explicit computation of the solution is not possible and iterative methods have to be used. Such methods are typically defined for the discretised version of problem (2.1), which is always the case in experimental contexts, where $\mathbf{X} = \mathbb{R}^N$, $\mathbf{Y} = \mathbb{R}^M$, and $\boldsymbol{T} \in \mathbb{R}^{M\times N}$. Here we will focus on

the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [18], which is an improved version of the Shrinkage-Thresholding Algorithm (ISTA). The next section is dedicated to the definition of FISTA. To get to this definition we first need to describe the Proximal Gradient Method, together with its fast version, and ISTA.

As a further remark before introducing FISTA, we stress that, as mentioned above, ℓ_1 regularisation promotes sparsity on the solution, whereas Tikhonov regularisation promotes smoothness. The difference between the two methods is illustrated in Figure 2.4 for a bi-dimensional case. In the ℓ_1 case, among all the solutions with a given ℓ_1-norm, the one with smallest entropy lies on the corner of $\|\mathbf{x}\|_1 = E$, that is a solution with null x_1 component. Depending on the properties that are desirable for the solution one may choose to use Tikhonov regularisation or ℓ_1 regularisation, however there are some intrinsic properties that are worth mentioning as they may influence the choice. Tikhonov regularisation has the great advantage of having a closed-form solution, which is easy to compute. However, the matrix $\mathbf{T}\mathbf{T}^* + \lambda \mathbf{I}_M$ needs to be inverted in the computations. Such operation might not be computationally feasible for high dimensional problems. This difficulty can be overcome with the ℓ_1 regularisation method, though using an iterative method can be time consuming and needs to set parameters, such as the tolerance and the maximum number of iterations, which add subjectivity to the method.

Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)

We now provide a formal definition of FISTA; to this end, we need to introduce the Proximal Gradient Method [47] and its accelerated form, as well as ISTA itself.

Let us consider a function f, continuously differentiable and convex. We can apply the gradient method to find its minimum, namely to compute

$$\arg\min_{\mathbf{x}}\{f(\mathbf{x})\}.$$

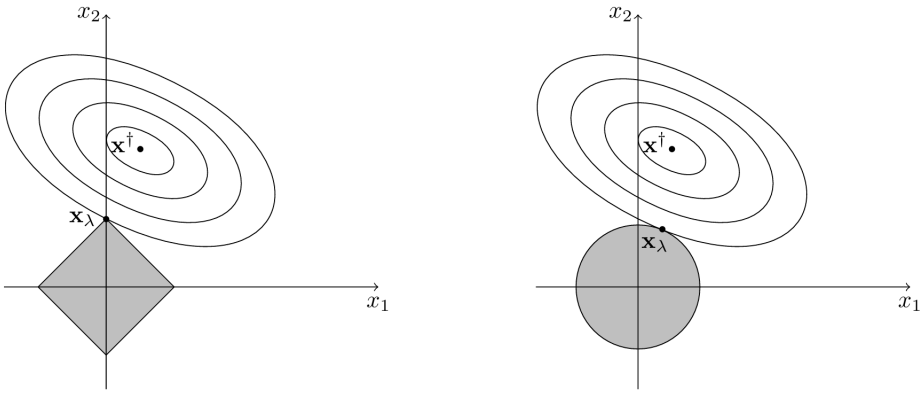

Figure 2.3: Schematic representation of the difference between ℓ_1 (a) and Tikhonov (b) regularisation methods. While ℓ_1 promotes sparsity, Tikhonov promotes smoothness. Courtesy of [45].

At each iteration of the gradient method, $\mathbf{x}_k$ moves into the direction of the steepest descent. Thus, given a suitable starting point $\mathbf{x}_0$ and stepsize $t_k > 0$, a generic point of the sequence $\{\mathbf{x}_k\}$ is given by

$$\mathbf{x}_k = \mathbf{x}_{k-1} - t_k \nabla f(\mathbf{x}_{k-1}). \tag{2.5}$$

Given the quadratic approximation of f, that is

$$q_t(\mathbf{x}, \mathbf{w}) := f(\mathbf{w}) + \langle \mathbf{x} - \mathbf{w}, \nabla f(\mathbf{w}) \rangle + \frac{1}{2t} \|\mathbf{x} - \mathbf{w}\|^2$$

and by observing that it holds

$$\|\mathbf{x} - \mathbf{w}\|^2 = \|\mathbf{x} - \mathbf{w} + t\nabla f(\mathbf{w})\|^2 - t^2 \|\nabla f(\mathbf{w})\|^2 - 2t\langle \mathbf{x} - \mathbf{w}, \nabla f(\mathbf{w}) \rangle,$$

the quadratic form can be rewritten as

$$q_t(\mathbf{x}, \mathbf{w}) = f(\mathbf{w}) + \frac{1}{2t} \|\mathbf{x} - \mathbf{w} + t\nabla f(\mathbf{w})\|^2 - \frac{t}{2} \|\nabla f(\mathbf{w})\|^2.$$

Since the gradient iteration (2.5) can be equivalently written as [48]

$$\begin{aligned}\mathbf{x}_k &= \arg\min_{\mathbf{x}}\{q_{t_k}(\mathbf{x}, \mathbf{x}_{k-1})\} \\ &= \arg\min_{\mathbf{x}}\Big\{f(\mathbf{x}_{k-1}) + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}_{k-1} + t_k\nabla f(\mathbf{x}_{k-1})\|^2 - \frac{t_k}{2}\|\nabla f(\mathbf{x}_{k-1})\|^2\Big\},\end{aligned}$$

by ignoring the constant terms, we finally obtain the following expression

$$\mathbf{x}_k = \arg\min_{\mathbf{x}}\Big\{\frac{1}{2t_k}\|\mathbf{x} - (\mathbf{x}_{k-1} - t_k\nabla f(\mathbf{x}_{k-1}))\|^2\Big\}. \tag{2.6}$$

Let us now introduce a general version of problem (2.4). Given $g : \mathbb{R}^N \to \mathbb{R}$ a non-smooth continuous convex function and $f : \mathbb{R}^N \to \mathbb{R}$ a continuously differentiable convex function, let us consider the functional Φ such that

$$\Phi : \mathbb{R}^N \longrightarrow \mathbb{R}, \quad \mathbf{x} \mapsto \Phi := f(\mathbf{x}) + g(\mathbf{x}),$$

Note that the ℓ_1 regularised functional can be obtained by considering $f(\mathbf{x}) = \|T\mathbf{x} - \mathbf{y}\|_2^2$ and $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$.

The optimisation model can be expressed as

$$\arg\min_{\mathbf{x}}\{\Phi(\mathbf{x})\}, \tag{2.7}$$

where Φ is non-smooth. For this reason, the gradient algorithm can not be applied to (2.7). However, the same iterative idea of (2.6) can be applied to (2.7), leading to the following iterative scheme

$$\mathbf{x}_k = \arg\min_{\mathbf{x}}\Big\{\frac{1}{2t_k}\|\mathbf{x} - (\mathbf{x}_{k-1} - t_k\nabla f(\mathbf{x}_{k-1}))\|^2 + g(\mathbf{x})\Big\}. \tag{2.8}$$

This scheme can be rewritten in terms of the proximal operator, which is defined as follows.

Definition 26. *For any $t > 0$ and for any convex function $g : \mathbb{R}^N \to \mathbb{R}$ the proximal*

operator associated with g is defined as

$$\text{prox}_t^g(\mathbf{w}) := \arg\min_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{1}{2t}\|\mathbf{x}-\mathbf{w}\|^2 \right\}, \quad \mathbf{x}, \mathbf{w} \in \mathbb{R}^N.$$

If we use the proximal operator definition, Equation (2.8) becomes

$$\mathbf{x}_k = \text{prox}_{t_k}^g \left(\mathbf{x}_{k-1} - t_k \nabla f(\mathbf{x}_{k-1}) \right).$$

Let us now further assume that f is a continuously differentiable convex function with Lipschitz continuous gradient with constant $L > 0$, i.e., for any $\mathbf{x}, \mathbf{w} \in \mathbb{R}^N$ it holds

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{w})\| \leq L\|\mathbf{x}-\mathbf{w}\|.$$

The Lipschitz constant L can be used to define the step size t_k. Thus, the step size turns out to be constant and we can refer to the Proximal Gradient Method with constant step size rule, which consists of the following step:

Input: L, a Lipschitz constant of ∇f;

Step 0. Take $\mathbf{x}_0 \in \mathbb{R}^N$, and set $t_k = \dfrac{1}{L}$;

Step k. (with $k \geq 1$) compute:

$$\mathbf{x}_k = \text{prox}_{\frac{1}{L}}^g \left(\mathbf{x}_{k-1} - \frac{1}{L}\nabla f(\mathbf{x}_{k-1}) \right). \tag{2.9}$$

The Proximal Gradient Method stops when a suitable stopping criterion is satisfied, such as the maximum number of iterations or a tolerance on the difference between two consecutive iterates, namely $\|\mathbf{x}_k - \mathbf{x}_{k-1}\| < \epsilon$. The following theorem provides the convergence rate for the function values $\Phi(\mathbf{x}_k)$.

Theorem 5. *Let $\Phi : \mathbb{R}^N \to \mathbb{R}$ be defined as $\Phi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, with $g : \mathbb{R}^N \to \mathbb{R}$ a continuous convex function and $f : \mathbb{R}^N \to \mathbb{R}$ a continuously differentiable convex*

function with Lipschitz continuous gradient with constant $L > 0$. For any $k \geq 1$ and for any minimiser $\mathbf{x}^* \in \mathbb{R}^N$ of $\Phi(\mathbf{x})$ it holds

$$\Phi(\mathbf{x}_k) - \Phi(\mathbf{x}^*) \leq \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2k},$$

being $\mathbf{x}_k$ defined by Equation (2.9) and $\mathbf{x}_0 \in \mathbb{R}^N$ a suitable starting point.

This theorem shows that the order of convergence of the Proximal Gradient Method is $O\left(\frac{1}{k}\right)$. The Fast Proximal Gradient Method improves this convergence rate by maintaining the same structure of the Proximal Gradient Method and introducing an auxiliary point $\mathbf{w}_k$ at which the proximal operator is computed. This additional point is defined by a linear combination of $\mathbf{x}_{k-1}$ and $\mathbf{x}_{k-2}$. The Fast Proximal Gradient Method algorithm is defined by the following steps:

Input: L, a Lipschitz constant of ∇f;

Step 0. Take $\mathbf{x}_0 \in \mathbb{R}^N$, and set $\mathbf{w}_1 = \mathbf{x}_0$, $t_1 = 1$;

Step k. (with $k \geq 1$) compute:

$$\mathbf{x}_k = \text{prox}^g_{\frac{1}{L}}\left(\mathbf{w}_k - \frac{1}{L}\nabla f(\mathbf{w}_k)\right), \tag{2.10}$$

$$t_{k+1} = \frac{1}{2} + \frac{1}{2}\sqrt{1 + 4t_k^2},$$

$$\mathbf{w}_{k+1} = \mathbf{x}_k + \frac{t_k - 1}{t_{k+1}}(\mathbf{x}_k - \mathbf{x}_{k-1}).$$

The stopping criterion for the Fast Proximal Gradient Method is the same as for the Proximal Gradient Method. In this case, the order of the convergence rate is $O\left(\frac{1}{k^2}\right)$. Indeed, the following theorem holds.

Theorem 6. *Let $\Phi : \mathbb{R}^N \to \mathbb{R}$ be defined as $\Phi(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, with $g : \mathbb{R}^N \to \mathbb{R}$ a continuous convex function and $f : \mathbb{R}^N \to \mathbb{R}$ a continuously differentiable convex function with Lipschitz continuous gradient with constant $L > 0$. For any $k \geq 1$*

and for any minimiser $\mathbf{x}^* \in \mathbb{R}^N$ *of* $\Phi(\mathbf{x})$ *it holds*

$$\Phi(\mathbf{x}_k) - \Phi(\mathbf{x}^*) \leq \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{(k+1)^2},$$

being $\mathbf{x}_k$ *defined by Equation (2.10) and* $\mathbf{x}_0 \in \mathbb{R}^N$ *a suitable starting point.*

Let us now get back to problem (2.4) and consider the particular case where $f(\mathbf{x}) = \|\mathbf{T}\mathbf{x} - \mathbf{y}\|_2^2$ and $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$. The convex function is continuously differentiable and its gradient is expressed as

$$\nabla f(\mathbf{x}) = 2\mathbf{T}^\top(\mathbf{T}\mathbf{x} - \mathbf{y}).$$

Moreover, $f(\mathbf{x})$ is Lipschitz continuous with smallest constant $L = 2\sigma_{\max}(\mathbf{T}^\top\mathbf{T})$, where $\sigma_{\max}(\mathbf{T}^\top\mathbf{T})$ is the maximum eigenvalue of $\mathbf{T}^\top\mathbf{T}$, and $g(\mathbf{x})$ is a continuous convex non-smooth function. The proximal operator associated to g can be analytically computed and it coincides with the shrinkage operator, that we define as follows.

Definition 27. *Given* $\alpha > 0$*, the shrinkage operator is defined as*

$$\mathcal{T}_\alpha(\mathbf{x})_i = (|x_i| - \alpha)_+ \operatorname{sgn}(x_i) = \begin{cases} x_i - \alpha & \text{if } x_i \geq \alpha, \\ x_i + \alpha & \text{if } x_i \leq -\alpha, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 7. *Given* $g : \mathbb{R}^N \to \mathbb{R}$ *defined as* $g(\mathbf{x}) := \lambda\|\mathbf{x}\|_1$*, with* $\lambda > 0$*, the proximal operator associated with* g *with* $t > 0$ *coincides with the shrinkage operator of threshold* $\alpha = \lambda t$*, i.e.,*

$$\operatorname{prox}_t^g(\mathbf{x}) = \mathcal{T}_{\lambda t}(\mathbf{x}).$$

We can now define ISTA as the Proximal Gradient Method applied to problem (2.4) by using the shrinkage operator and the Lipschitz constant $L = \sigma_{\max}(\mathbf{T}^\top\mathbf{T})$. The ISTA iterative scheme results in:

Input: $L = \sigma_{\max}(\mathbf{T}^\top \mathbf{T})$;

Step 0. Take $\mathbf{x}_0 \in \mathbb{R}^N$;

Step k. (with $k \geq 1$) compute

$$\mathbf{x}_k = \mathcal{T}_{\frac{\lambda}{L}} \left(\mathbf{x}_{k-1} - \frac{2}{L} \mathbf{T}^\top (\mathbf{T}\mathbf{x}_{k-1} - \mathbf{y}) \right).$$

Similarly, we can define FISTA as the Fast Proximal Gradient method applied to problem (2.4) by using the shrinkage operator and the Lipschitz constant $L = \sigma_{\max}(\mathbf{T}^\top \mathbf{T})$. The FISTA iterative scheme results in:

Input: $L = \sigma_{\max}(\mathbf{T}^\top \mathbf{T})$

Step 0. Take $\mathbf{x}_0 \in \mathbb{R}^N$, and set $\mathbf{w}_1 = \mathbf{x}_0$, $t_1 = 1$;

Step k. (with $k \geq 1$) compute:

$$\mathbf{x}_k = \mathcal{T}_{\frac{\lambda}{L}} \left(\mathbf{w}_k - \frac{2}{L} \mathbf{T}^\top (\mathbf{T}\mathbf{x}_{k-1} - \mathbf{y}) \right),$$

$$t_{k+1} = \frac{1}{2} + \frac{1}{2}\sqrt{1 + 4t_k^2},$$

$$\mathbf{w}_{k+1} = \mathbf{x}_k + \frac{t_k - 1}{t_{k+1}} (\mathbf{x}_k - \mathbf{x}_{k-1}).$$

The stopping criterion for both ISTA and FISTA is the same as for the Proximal Gradient Method, that is the maximum number of iterations or a tolerance on the difference between two consecutive iterates.

2.4 The expectation-maximisation algorithm

Inverse problems are naturally formulated within a statistical inference framework, since the observed data are typically affected by stochastic noise. While

regularisation methods provide a deterministic strategy to solve the inverse problem, a probabilistic approach models the data explicitly as random variables and exploits their statistical properties to estimate the underlying quantities of interest.

The maximum likelihood approach is among the most widely used statistical techniques for solving inverse problems. It consists in the maximisation of the likelihood function, which describes how the unknown parameters give rise to the observed data. Its maximisation consequently yields an estimate of the unknown variables, providing a statistical alternative to direct inversion. However, when the model depends on unobserved latent variables, the direct maximisation of the likelihood function is not possible and it is replaced by an iterative method to compute the maximum likelihood estimates. This method is known as the expectation-maximisation (EM) algorithm [23].

In the following sections, we describe the likelihood function in more detail and then we present the steps of the EM algorithm.

2.4.1 Likelihood function and its maximisation

Let us define the likelihood function, an important statistic used to summarize data.

Definition 28. *Let us consider a sample* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *and its observation* $\boldsymbol{X} = \mathbf{x}$*. The joint probability density function of the sample is denoted as* $f(\mathbf{x}|\boldsymbol{\theta})$*. Then, we call the likelihood function the function of* $\boldsymbol{\theta}$

$$L(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}).$$

The values assumed by the likelihood $L(\boldsymbol{\theta}|\mathbf{x})$ tell us how likely it is that the sample is actually observed under those very values of the parameters $\boldsymbol{\theta}$. Thus, when we compare the likelihood functions at two parameter values, namely $L(\boldsymbol{\theta}_1|\mathbf{x})$ and $L(\boldsymbol{\theta}_2|\mathbf{x})$, and we get

$$L(\boldsymbol{\theta}_1|\mathbf{x}) > L(\boldsymbol{\theta}_2|\mathbf{x}),$$

$\boldsymbol{\theta}_1$ can be interpreted as a more plausible value for the true value of $\boldsymbol{\theta}$ than $\boldsymbol{\theta}_2$. We also note that the only distinction between the likelihood function and the probability density function is which variable is considered fixed and which one is varying. When we consider the probability density function $f(\mathbf{x}|\boldsymbol{\theta})$, we are considering $\boldsymbol{\theta}$ as fixed and $\mathbf{x}$ as the variable; instead, when we consider the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$, we are considering $\mathbf{x}$ to be the fixed observed sample point and $\boldsymbol{\theta}$ to be varying over all possible parameter values.

There exists many methods for finding estimators of the parameters $\boldsymbol{\theta}$. In particular, we are interested in the method of maximum likelihood, which is by far the most popular technique for this purpose. Let us first give a definition of the maximum likelihood estimator.

Definition 29. *For each sample point* $\mathbf{x}$*, let* $\hat{\boldsymbol{\theta}}(\mathbf{x})$ *be a parameter value at which* $L(\boldsymbol{\theta}|\mathbf{x})$ *attains its maximum as a function of* $\boldsymbol{\theta}$*, with* $\mathbf{x}$ *fixed. Then* $\hat{\boldsymbol{\theta}}(\mathbf{x})$ *is a maximum likelihood estimate (MLE) of the parameter* $\boldsymbol{\theta}$ *based on a sample* $\boldsymbol{x}$*.*

The MLE thus represents the parameter point for which the observed sample is most likely. The MLE is considered to be a good point estimator, even if the maximum likelihood estimation inherently presents two main drawbacks. The first problem refers to finding the actual global maximum of the function, while the second problem is that of numerical sensitivity. Indeed, it is often the case that a slightly different sample will produce a vastly different MLE, showing that the estimate is sensitive to small changes in the data. A specific procedure to compute the MLE in the case of missing data, besides differentiation, is the EM algorithm that is guaranteed to converge to the MLE [49]. More details on the likelihood function and its maximisation approach can be found in [49].

2.4.2 The EM steps

The expectation-maximisation algorithm has been introduced independently both by Harteley [50] in 1958 and by Dempster [23] in 1977. The core idea of this method is to replace one difficult maximisation with a sequence of easier

maximisations whose limit tends to be the solution of the original problem. Dempster and collegues, in particular, proposed a general approach to iteratively compute maximum-likelihood estimates from incomplete data. Given two sample spaces Y and X, the incomplete data case verifies when the observed data lie in Y and provides an indirect measure of $\mathbf{x}$, where $\mathbf{x} \in X$ can not be observed directly. We can assume the existence of a mapping that relates $\mathbf{x} \in X$ to $\mathbf{y} \in Y$ and we denote by $X(\mathbf{y})$ the subset of X that is consistent with the observed data $\mathbf{y}$, that is $\mathbf{y} = \mathbf{y}(\mathbf{x})$. Given a family of sampling densities $f(\mathbf{x}|\boldsymbol{\theta})$ depending on parameter $\boldsymbol{\theta}$, we aim at deriving the corresponding family of sampling densities $g(\mathbf{y}|\boldsymbol{\theta})$.

The family $f(\mathbf{x}|\boldsymbol{\theta})$ is called complete-data density function, while $g(\mathbf{y}|\boldsymbol{\theta})$ is the incomplete-data function, since it depends only on the observed data $\mathbf{y}$ and not on the complete samples. Note that some formulations of the complete-data function make the dependence on the observed variable $\mathbf{y}$ explicit, namely $f(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})$. In this chapter we follow the notation used in Dempster's paper, that is $f(\mathbf{x}|\boldsymbol{\theta})$ for the complete-data function.

The two families are related by the equation

$$g(\mathbf{y}|\boldsymbol{\theta}) = \int_{X(\mathbf{y})} f(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}. \tag{2.11}$$

The EM algorithm seeks a value of $\boldsymbol{\theta}$ which maximises $g(\mathbf{y}|\boldsymbol{\theta})$ given an observed $\mathbf{y}$, by exploiting the associated family $f(\mathbf{x}|\boldsymbol{\theta})$. Indeed, maximising $f(\mathbf{x}|\boldsymbol{\theta})$ is preferred because of the closed form of the complete-data density, while direct maximisation of $g(\mathbf{y}|\boldsymbol{\theta})$ is often intractable. Also note that, given the incomplete-data specification $g(\mathbf{y}|\boldsymbol{\theta})$, there are many possible complete-data specifications $f(\mathbf{x}|\boldsymbol{\theta})$ that generate $g(\mathbf{y}|\boldsymbol{\theta})$. Thus, the choice of defining the associated $f(\mathbf{x}|\boldsymbol{\theta})$ may be done in several different ways.

The EM algorithms consists in two steps, namely the expectation step, also known as E–step, and the maximisation step, that is called the M–step.

We first consider the EM algorithm in a restricted framework, namely the case where $f(\mathbf{x}|\boldsymbol{\theta})$ assumes the form of a regular exponential-family and then we expand to a more general setting. Thus, let us consider a regular

exponential-family

$$f(\mathbf{x}|\boldsymbol{\theta}) = b(\mathbf{x}) \exp\left[(\boldsymbol{\theta} t(\mathbf{x})^T)/a(\boldsymbol{\theta})\right], \tag{2.12}$$

where $\boldsymbol{\theta}$ denotes the vector of parameters and $a(\boldsymbol{\theta})$ is a real-valued function depending only on the parameters $\boldsymbol{\theta}$, while the functions $b(\mathbf{x}) \geq 0$ and $t(\mathbf{x})$ depend on $\mathbf{x}$. The function $t(\mathbf{x})$ summarises all information of the data which are needed for the estimation of $\boldsymbol{\theta}$, assuming that the complete data were known. For this reason, it is known as the complete-data sufficient statistics. The regularity of the exponential family comes from the assumption that $\boldsymbol{\theta}$ is restricted only to a finite dimensional convex set Ω, such that expression (2.12) defines a density for all $\boldsymbol{\theta} \in \Omega$. Consequently, up to an arbitrary non-singular linear transformation, the parametrization $\boldsymbol{\theta}$ and the choice of $t(\mathbf{x})$ are unique. The set of parameters $\boldsymbol{\theta}$ are often called natural parameters, although in many examples the conventional parameters are often non-linear functions of $\boldsymbol{\theta}$.

Let us present the procedure for the EM algorithm when considering this restricted framework of regular exponential families. Given $\boldsymbol{\theta}^{(p)}$ the value of $\boldsymbol{\theta}$ after p iterations of the algorithms, the EM framework can be described in two steps:

- E-step: Estimate the complete-data sufficient statistics $t(\mathbf{x})$ by computing
$$t^{(p)} = \mathbb{E}(t(\mathbf{x})|\mathbf{y}, \boldsymbol{\theta}^{(p)}). \tag{2.13}$$
- M–step: Determine $\boldsymbol{\theta}^{(p+1)}$ by solving the equations
$$\mathbb{E}(t(\mathbf{x})|\boldsymbol{\theta}) = t^{(p)}. \tag{2.14}$$

Equations (2.14) define the maximum-likelihood estimator of $\boldsymbol{\theta}$, given that $t^{(p)}$ is the vector of sufficient statistics computed from an observed $\mathbf{x}$ drawn from the exponential family (2.12). Indeed, let us consider the expression of the

likelihood for an exponential family

$$\log f(\mathbf{x}|\boldsymbol{\theta}) = -\log(a(\boldsymbol{\theta})) + \log(b(\mathbf{x})) + \boldsymbol{\theta} t(\mathbf{x})^T.$$

Computing its derivative with respect to the parameters $\boldsymbol{\theta}$ and setting it to zero leads to

$$-\frac{\partial}{\partial \boldsymbol{\theta}} (\log a(\boldsymbol{\theta})) + t(\mathbf{x})^T = 0.$$

We obtain (2.14) considering that $\frac{\partial}{\partial \boldsymbol{\theta}} (\log a(\boldsymbol{\theta})) = \mathbb{E}(t(\mathbf{x})|\boldsymbol{\theta})$ for an exponential family.

Equations (2.14) are not always solvable for $\boldsymbol{\theta} \in \Omega$. For such cases, a more general definition must be used, because the maximising value of $\boldsymbol{\theta}$ lies on the boundary of Ω. Instead, for those cases that can be solved for $\boldsymbol{\theta} \in \Omega$, because of the well-known convexity property of the log-likelihood for regular exponential families, the solution is unique.

Let us now understand why alternating the E–step and the M–step leads to find the optimal value $\boldsymbol{\theta}^*$ that maximises

$$\ell(\boldsymbol{\theta}) = \log g(\mathbf{y}|\boldsymbol{\theta}), \tag{2.15}$$

where $g(\mathbf{y}|\boldsymbol{\theta})$ is the incomplete-data specification previously introduced in expression (2.11). We recall that the maximisation of (2.15) is obtained by working with the complete-data and conditional densities. Since the conditional density of $\mathbf{x}$ given $\mathbf{y}$ and $\boldsymbol{\theta}$ is given by

$$k(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{\theta}) / g(\mathbf{y}|\boldsymbol{\theta}), \tag{2.16}$$

the expression for the log-likelihood can be written as

$$\ell(\boldsymbol{\theta}) = \log f(\mathbf{x}|\boldsymbol{\theta}) - \log k(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}). \tag{2.17}$$

In the special case of exponential families, we observe that the conditional density takes the following form

$$k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}) = \frac{b(\mathbf{x})\, \exp\big(\boldsymbol{\theta} t(\mathbf{x})^T\big)}{a(\boldsymbol{\theta} \mid \mathbf{y})},$$

where the normalizing function $a(\boldsymbol{\theta}|\mathbf{y})$ is given by

$$a(\boldsymbol{\theta} \mid \mathbf{y}) = \int_{\mathbf{x} \in X(\mathbf{y})} b(\mathbf{x})\, \exp\Big(\boldsymbol{\theta} t(\mathbf{x})^T\Big)\, d\mathbf{x}. \tag{2.18}$$

Note that both functions $f(\mathbf{x} \mid \boldsymbol{\theta})$ and $k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})$ share the natural parameter $\boldsymbol{\theta}$ and the sufficient statistic $t(\mathbf{x})$, because they belong to the same exponential family. However, they are defined on different sample spaces, namely X and $X(\mathbf{y})$. Considering this, we may rewrite (2.17) as

$$\ell(\boldsymbol{\theta}) = -\log a(\boldsymbol{\theta}) + \log a(\boldsymbol{\theta} \mid \mathbf{y}),$$

where the counterpart to (2.18) is

$$a(\boldsymbol{\theta}) = \int b(\mathbf{x})\, \exp\Big(\boldsymbol{\theta}^T t(\mathbf{x})\Big)\, d\mathbf{x}. \tag{2.19}$$

By differentiating (2.19) and writing $t(\mathbf{x})$ simply as t, we obtain

$$D \log a(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log a(\boldsymbol{\theta}) = \mathbb{E}(\mathbf{t} \mid \boldsymbol{\theta}).$$

Similarly, by differentiating (2.18) we get

$$D \log a(\boldsymbol{\theta} \mid \mathbf{y}) = \mathbb{E}(\mathbf{t} \mid \mathbf{y}, \boldsymbol{\theta}).$$

Finally, we can write the derivative of the log-likelihood that comes into play when performing the maximisation. Indeed, it becomes

$$D\ell(\boldsymbol{\theta}) = -\mathbb{E}(\mathbf{t} \mid \boldsymbol{\theta}) + \mathbb{E}(\mathbf{t} \mid \mathbf{y}, \boldsymbol{\theta}), \tag{2.20}$$

namely it admits a convenient representation as the difference between an unconditional and a conditional expectation of the sufficient statistics.

Equation (2.20) is fundamental for understanding the E- and M–steps of

the EM algorithm. When the algorithm converges to $\boldsymbol{\theta}^*$, meaning that

$$\mathbb{E}(\mathbf{t} \mid \boldsymbol{\theta}^*) = \mathbb{E}(\mathbf{t} \mid \mathbf{y}, \boldsymbol{\theta}^*),$$

which is equivalent to having a null derivative, namely $DL(\boldsymbol{\theta}^*) = 0$.

The formulation of Equation (2.20) is commonly used in the context of EM. Its general form was given by Sundberg in 1974 [51], who traced it back to unpublished 1966 lecture notes of Martin-Löf.

We now extend our treatment of the EM algorithm to a more general setting. Our second level of generality assumes that the completedata model is not a regular exponential family, but rather a *curved* exponential family. In this situation, representation (2.12) for the exponential family remains valid, but the parameters $\boldsymbol{\theta}$ must lie in a curved submanifold Ω_0 of the dimensional convex region Ω. The E-step of the EM algorithm is unchanged, but the assumptions for Sundbergs identities are no longer valid. Consequently, the M-step must be reformulated as follows:

M-step: Determine $\boldsymbol{\theta}^{(p+1)}$ as a value of $\boldsymbol{\theta} \in \Omega_0$ that maximises

$$-\log a(\boldsymbol{\theta}) + \boldsymbol{\theta}(\mathbf{t}^{(p)})^T.$$

In other words, the M-step now consists of maximising the likelihood under the assumption that the sufficient statistics are given by $\mathbf{t}^{(p)}$. We note that this extended definition of the M-step, with Ω substituted for Ω_0, is appropriate in those regular exponentialfamily cases for which equations (2.14) can not be explicitly solved for $\boldsymbol{\theta}$ within Ω mentioned before.

Our third and most general setting removes any reference to exponential-family structure. Here we introduce the function

$$Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = \mathbb{E}\left[\log f(\mathbf{x} \mid \boldsymbol{\theta}') \,\middle|\, \mathbf{y}, \boldsymbol{\theta}\right], \tag{2.21}$$

which we assume exists for all pairs $(\boldsymbol{\theta}', \boldsymbol{\theta})$. We note that the expectation in expression (2.21) is taken under the parameter guess $\boldsymbol{\theta}$. The guiding idea is

that we would ideally maximise $\log f(\mathbf{x} \mid \boldsymbol{\theta})$. Since this quantity is unknown, we maximise instead its conditional expectation given $\mathbf{y}$ and the parameter value $\boldsymbol{\theta}$.

We assume $f(\mathbf{t} \mid \boldsymbol{\theta}) > 0$ almost everywhere on X for all $\boldsymbol{\theta} \in \Omega$. The full EM update $\boldsymbol{\theta}^{(p)} \to \boldsymbol{\theta}^{(p+1)}$ is now defined as:

E-step: Compute $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(p)})$;

M-step: choose $\boldsymbol{\theta}^{(p+1)} \in \Omega$ to maximise $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(p)})$.

For exponential families, we have

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(p)}) = -\log a(\boldsymbol{\theta}) + \mathbb{E}[\, b(\mathbf{x}) \mid y, \boldsymbol{\theta}^{(p)} \,] + \boldsymbol{\theta}\mathbf{t}^{(p)T},$$

so maximising $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(p)})$ is equivalent to maximising $-\log a(\boldsymbol{\theta}) + \boldsymbol{\theta}\mathbf{t}^{(p)T}$, matching the more specialised M-step defined earlier. In exponentialfamily models, the E-step of (2.13) is typically simpler than the general E-step, since only the expected sufficientstatistic components must be computed.

2.4.3 General properties of the EM algorithm

We now collect some basic results for the EM algorithm. Throughout, the observable data $\mathbf{y}$ are treated as fixed.

For convenience, define

$$H(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = \mathbb{E}\big[\log k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}') \mid \mathbf{y}, \boldsymbol{\theta}\big], \tag{2.22}$$

where the expectation in expression (2.22) is taken again under the current parameter guess $\boldsymbol{\theta}$. Thus, using (2.15), (2.16), and (2.21),

$$Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = \ell(\boldsymbol{\theta}') + H(\boldsymbol{\theta}' \mid \boldsymbol{\theta}).$$

Lemma 1. *For any pair* $(\boldsymbol{\theta}', \boldsymbol{\theta}) \in \Omega \times \Omega$,

$$H(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) \leq H(\boldsymbol{\theta} \mid \boldsymbol{\theta}),$$

with equality if and only if $k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}') = k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})$ *almost everywhere.*

This is a standard consequence of Jensens inequality [52].

To specify an iterative algorithm, we only need to list the sequence

$$\boldsymbol{\theta}^{(0)} \to \boldsymbol{\theta}^{(1)} \to \boldsymbol{\theta}^{(2)} \to \cdots$$

starting from a given initial value $\boldsymbol{\theta}^{(0)}$. More generally, an iterative algorithm is a mapping $M(\boldsymbol{\theta}) : \Omega \to \Omega$ such that

$$\boldsymbol{\theta}^{(p+1)} = M(\boldsymbol{\theta}^{(p)}).$$

Definition 30. *An iterative algorithm with mapping* $M(\boldsymbol{\theta})$ *is called a* generalized EM algorithm (GEM) *if, for every* $\boldsymbol{\theta} \in \Omega$*,*

$$Q(M(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) \geq Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}).$$

The definitions of the full EM algorithm given above require

$$Q(M(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) \geq Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$$

for all $(\boldsymbol{\theta}', \boldsymbol{\theta}) \in \Omega \times \Omega$, meaning that $\boldsymbol{\theta}' = M(\boldsymbol{\theta})$ maximises $Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$.

Theorem 8. *For every GEM algorithm,*

$$\ell(M(\boldsymbol{\theta})) \geq \ell(\boldsymbol{\theta}) \quad \forall\, \boldsymbol{\theta} \in \Omega,$$

where equality holds if and only if both

$$Q(M(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}),$$

and

$$k(\mathbf{x} \mid \mathbf{y}, M(\boldsymbol{\theta})) = k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}) \quad \textit{almost everywhere.}$$

Proof. We have

$$\ell(M(\boldsymbol{\theta})) - \ell(\boldsymbol{\theta}) = \big[\, Q(M(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) - Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}) \,\big] + \big[\, H(\boldsymbol{\theta} \mid \boldsymbol{\theta}) - H(M(\boldsymbol{\theta}) \mid \boldsymbol{\theta}) \,\big]. \tag{2.23}$$

For every GEM algorithm, the difference in Q functions in (2.23) is strictly positive. By Lemma 1, the difference in H functions is nonnegative, with equality if and only if

$$k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}) = k(\mathbf{x} \mid \mathbf{y}, M(\boldsymbol{\theta})) \quad \text{almost everywhere.}$$

□

Corollary 1. *Suppose that for some $\boldsymbol{\theta}^* \in \Omega$,*

$$\ell(\boldsymbol{\theta}^*) \geq \ell(\boldsymbol{\theta}) \quad \forall\, \boldsymbol{\theta} \in \Omega.$$

Then for every GEM algorithm:

(a) $\ell(M(\boldsymbol{\theta}^*)) = \ell(\boldsymbol{\theta}^*)$,

(b) $Q(M(\boldsymbol{\theta}^*) \mid \boldsymbol{\theta}^*) = Q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^*)$,

(c) $k(\mathbf{x} \mid \mathbf{y}, M(\boldsymbol{\theta}^*)) = k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}^*)$ *almost everywhere.*

Corollary 2. *If for some $\boldsymbol{\theta}^* \in \Omega$,*

$$\ell(\boldsymbol{\theta}^*) > \ell(\boldsymbol{\theta}) \quad \forall\, \boldsymbol{\theta} \in \Omega,\ \boldsymbol{\theta} \neq \boldsymbol{\theta}^*,$$

then for every GEM algorithm,

$$M(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^*.$$

Theorem 9. *Suppose that $\{\boldsymbol{\theta}^{(p)}\}_{p=0}^{\infty}$ is an instance of a GEM algorithm such that:*

1. *the sequence $\ell(\boldsymbol{\theta}^{(p)})$ is bounded, and*

2.

$$Q(\boldsymbol{\theta}^{(p+1)} \mid \boldsymbol{\theta}^{(p)}) - Q(\boldsymbol{\theta}^{(p)} \mid \boldsymbol{\theta}^{(p)}) \;\geq\; \lambda\,(\boldsymbol{\theta}^{(p+1)} - \boldsymbol{\theta}^{(p)})(\boldsymbol{\theta}^{(p+1)} - \boldsymbol{\theta}^{(p)})^T$$

for some $\lambda > 0$ *for all* p.

Then the sequence $\boldsymbol{\theta}^{(p)}$ *converges to some* $\boldsymbol{\theta}^*$ *in the closure of* Ω.

Proof. From assumption (1) and Theorem 8, the sequence $\ell(\boldsymbol{\theta}^{(p)})$ converges to some finite limit $\ell^* < \infty$. Hence, for any $\varepsilon > 0$, there exists $p(\varepsilon)$ such that for all $p \geq p(\varepsilon)$ and all $r \geq 1$,

$$\sum_{j=1}^{r} \left[\, \ell(\boldsymbol{\theta}^{(p+j)}) - \ell(\boldsymbol{\theta}^{(p+j-1)}) \,\right] = \ell(\boldsymbol{\theta}^{(p+r)}) - \ell(\boldsymbol{\theta}^{(p)}) < \epsilon.$$

□

Theorem 8 implies that $\ell(\boldsymbol{\theta})$ is nondecreasing on each iteration of a GEM algorithm, and is strictly increasing on any iteration such that

$$Q(\boldsymbol{\theta}^{(p+1)} \mid \boldsymbol{\theta}^{(p)}) > Q(\boldsymbol{\theta}^{(p)} \mid \boldsymbol{\theta}^{(p)}).$$

The corollaries imply that a maximum-likelihood estimate is a fixed point of a GEM algorithm. Theorem 9 provides the conditions under which an instance of a GEM algorithm converges. However, these results do not imply convergence to a maximum-likelihood estimator. Usually, the convergence to maximum likelihood holds under the assumption of the existence and continuity of a sufficient number of derivatives of the functions $Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$, $\ell(\boldsymbol{\theta})$, $H(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$.

In almost all applications, the limiting value $\boldsymbol{\theta}^*$ specified in Theorem 9 will occur at a local—if not global—maximum of $\ell(\boldsymbol{\theta})$. An exception could occur if $DM(\boldsymbol{\theta}^*)$ has eigenvalues exceeding unity. In that case $\boldsymbol{\theta}^*$ could be a saddle point of $\ell(\boldsymbol{\theta})$, for certain convergent sequences $\boldsymbol{\theta}^{(p)} \to \boldsymbol{\theta}^*$ that are asymptotically orthogonal to the eigenvectors of $DM(\boldsymbol{\theta}^*)$ associated with the large eigenvalues. Note that if $\boldsymbol{\theta}$ were given a small random perturbation

away from a saddle point $\boldsymbol{\theta}^*$, then the EM algorithm would diverge from that saddle point. When all eigenvalues of $DM(\boldsymbol{\theta}^*)$ are less than 1, the largest such eigenvalue determines the rate of convergence of the algorithm. We do not investigate the convergence results any further. For more details the reader can refer to the seminal papers [50, 23] and to [53].

Chapter 3

Sparse optimisation of cross-power spectra in M/EEG inverse problem

In this chapter, we tackle the problem of functional brain connectivity estimation in the brain source domain. To this aim, common methods usually involve a two-step procedure. First, the brain activity is estimated from M/EEG data via the solution of an ill-posed inverse problem. Then, brain connectivity is computed by means of connectivity metrics, such as the cross-power spectrum discussed in Chapter 1, from the estimated brain activity.

However, a number of recent studies demonstrated that this two–step procedure is inherently sub-optimal. Indeed, when the initial optimisation step is achieved through Tikhonov regularization [12], a regularization parameter has to be set, and this is typically done in order to obtain the best possible estimate of the brain activity. On the other hand, the cross-power spectrum computed from the time–series estimated with this value of the regularization parameter is usually sub-optimal. It was shown that a better estimate can be obtained from the compued time–series using a different value of the regularization parameter [13, 14] depending on features of the neural time-courses themselves, such as their spectral complexity [54].

Moreover, the estimates of the cortical cross-power spectrum obtained through this two–step procedure are usually affected by a large number of

false positives, that is, identifying statistical interactions between pairs of brain regions that are actually independent. One of the main causes of these spurious interactions is the residual mixing (or source–leakage) of the hidden process describing brain activity [55]. To overcome this problem, existing methods, described in Chapter 1, have been proposed, with the obvious drawback that they also ignore instantaneous (or nearly instantaneous) true interactions [15].

In the last years, few optimisation techniques have been proposed to avoid such two–step procedure. In [16], the authors propose a new method of MEG source reconstruction that simultaneously estimates the source amplitudes and interactions across the whole brain by a variational Bayesian algorithm; in particular, they use a multivariate autoregressive (AR) model to represent directed interaction between sources together with a prior knowledge on structural brain connectivity inferred from diffusion magnetic resonance imaging (dMRI). In [17, 56], a Kalman filter is used to achieve a simultaneous estimation of source activities and their dynamic functional connectivity. However, also in this case AR models are used for representing source interactions. Ossadtchi and colleagues [15] introduce a novel projection matrix that operates on the cross–power spectrum of the observable process in order to mitigate the contribution of source–leakage to its real part. Once applied the projector, an optimisation step is still required to estimate the cortical cross–power spectrum from the modified sensor–level data.

In this chapter, we present the results of our work [1]. In particular, we suggest an alternative approach where the cross–power spectrum of the hidden process describing the cortical activity is directly estimated from that of observed M/EEG time–series by setting up a linear optimisation problem. In order to reduce the number of false positive, ℓ_1 regularization [57] is used so as to promote sparsity in the resulting estimator of the cortical cross–power spectrum. The Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [18] is a classical approach for solving linear optimisation problems with ℓ_1 penalty. However, its application to our problem is not trivial due to

the high dimension of the matrix describing the forward operator, which typically includes few hundreds of rows and few thousands of columns. Thus, we propose a computational strategy for efficiently carrying on the FISTA update by exploiting structural properties of the forward operator.

In Section 3.1 we derive the forward model relating the cross–power spectrum of the hidden process to that of the observable one. In Section 3.2 we present the novel proposed approach for estimating the cross–power spectrum of the hidden process. We also recall a classical two-step approach used as benchmark. Section 3.3 focuses on the strategy we develop to make our approach computationally affordable. Finally, numerical validations of the presented approach are shown in Section 3.4 and our conclusions are discussed in Section 3.5.

3.1 Problem formulation and forward modelling

Let us recall, first of all, the definition of the cross–power spectrum given in Chapter 1. We assume without loss of generality that the mean values of all the considered stochastic processes are zero throughout this chapter.

Definition 31. *Let $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ and $\{\mathbf{E}(t)\}_{t\in\mathbb{R}}$ be two real-valued, multivariate, stationary stochastic processes of dimension n and m, respectively. Denoted with $\left\{\Gamma^{\mathbf{XE}}(\tau)\right\}_{t\in\mathbb{R}}$ the corresponding covariance function, that is $\Gamma^{\mathbf{XE}}(\tau) = \mathbb{E}\left[\mathbf{X}(t)\mathbf{E}(t+\tau)^\top\right]$, we assume $\Gamma^{\mathbf{XE}}_{j,k}(\tau)$ to be absolutely integrable for all $j = 1,\dots,n$ and $k = 1,\dots,m$.*
Then, the cross-power spectrum between $\mathbf{X}(t)$ and $\mathbf{E}(t)$ is a one–parameter family of complex–valued matrices $\mathbf{S}^{\mathbf{XE}}(f) \in \mathbb{C}^{n\times m}$ defined as [37]

$$S^{\mathbf{XE}}(f) = \int_{-\infty}^{+\infty} \Gamma^{\mathbf{XE}}(\tau) e^{-2\pi i f\tau} \mathrm{d}\tau,$$

where the integral is computed component-wise.

Remark 3. *Following Definition 31, we define the cross-power spectrum of the process $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ as a one–parameter family of Hermitian matrices of size $n \times n$ denoted as $\mathbf{S}^{\mathbf{X}}(f) := \mathbf{S}^{\mathbf{XX}}(f)$. Analogously, we denote $\Gamma^{\mathbf{X}}(\tau) := \Gamma^{\mathbf{XX}}(\tau)$.*

The ext theorem introduces the linear relationship between the cross-power spectrum of the observable process $\{\mathbf{Y}(t)\}_{t\in\mathbb{R}}$ and that of the unknown process $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$.

Theorem 10. *Let $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ and $\{\mathbf{Y}(t)\}_{t\in\mathbb{R}}$ be two real-valued, multivariate, stationary stochastic processes of dimension n and m, respectively. We further assume that $\mathbf{Y}(t)$ is a linear mixture of the components of $\mathbf{X}(t)$ corrupted by independent additive noise, that is*

$$\mathbf{Y}(t) = \mathbf{G}\mathbf{X}(t) + \mathbf{E}(t)\,, \tag{3.1}$$

where $\mathbf{G} \in \mathbb{R}^{m\times n}$ and $\{\mathbf{E}(t)\}_{t\in\mathbb{R}}$ is a multivariate, stationary stochastic process independent from $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$. Then,

$$\mathbf{S}^{\mathbf{Y}}(f) = \mathbf{G}\mathbf{S}^{\mathbf{X}}(f)\mathbf{G}^{\top} + \mathbf{S}^{\mathbf{E}}(f)\,. \tag{3.2}$$

Proof. We observe that from the linearity of the expectation and the hypothesis on $\mathbf{Y}(t)$ it follows

$$\begin{aligned}\Gamma^{\mathbf{Y}}(\tau) &= \mathbb{E}\left[(\mathbf{G}\mathbf{X}(t) + \mathbf{E}(t))(\mathbf{G}\mathbf{X}(t+\tau) + \mathbf{E}(t+\tau))^{\top}\right] \\ &= \mathbf{G}\Gamma^{\mathbf{X}}(\tau)\mathbf{G}^{\top} + \mathbf{G}\Gamma^{\mathbf{X}\mathbf{E}}(\tau) + \left(\mathbf{G}\Gamma^{\mathbf{X}\mathbf{E}}(\tau)\right)^{\top} + \Gamma^{\mathbf{E}}(\tau) \\ &= \mathbf{G}\Gamma^{\mathbf{X}}(\tau)\mathbf{G}^{\top} + \Gamma^{\mathbf{E}}(\tau),\end{aligned}$$

where the last equality holds due to the fact that $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$ and $\{\mathbf{E}(t)\}_{t\in\mathbb{R}}$ are independent and thus $\Gamma^{\mathbf{X}\mathbf{E}}(\tau) = 0$ for all τ. The thesis follows from Definition 31 by eploiting the linearity of the Fourier transform. □

By denoting with $\mathcal{S}^{\mathbf{X}}(f) \in \mathbb{C}^{n^2}$, and $\mathcal{S}^{\mathbf{Y}}(f), \mathcal{S}^{\mathbf{E}}(f) \in \mathbb{C}^{m^2}$ the vector obtained by stacking the columns (vec$(\cdot)$ operator) of matrices $\mathbf{S}^{\mathbf{X}}(f)$, $\mathbf{S}^{\mathbf{Y}}(f)$, and $\mathbf{S}^{\mathbf{E}}(f)$, respectively, it follows from Equation (3.2) that

$$\mathcal{S}^{\mathbf{Y}}(f) = (\mathbf{G}\otimes\mathbf{G})\,\mathcal{S}^{\mathbf{X}}(f) + \mathcal{S}^{\mathbf{E}}(f), \tag{3.3}$$

where $\otimes$ is the Kronecker product.

Then, splitting real and imaginary parts, Equation (3.3) can be rewritten so to include only real-valued quantities:

$$\begin{pmatrix} \mathrm{Re}(\mathcal{S}^{\mathbf{Y}}(f)) \\ \mathrm{Im}(\mathcal{S}^{\mathbf{Y}}(f)) \end{pmatrix} = \begin{pmatrix} \mathbf{G} \otimes \mathbf{G} & 0 \\ 0 & \mathbf{G} \otimes \mathbf{G} \end{pmatrix} \begin{pmatrix} \mathrm{Re}(\mathcal{S}^{\mathbf{X}}(f)) \\ \mathrm{Im}(\mathcal{S}^{\mathbf{X}}(f)) \end{pmatrix} + \begin{pmatrix} \mathrm{Re}(\mathcal{S}^{\mathbf{E}}(f)) \\ \mathrm{Im}(\mathcal{S}^{\mathbf{E}}(f)) \end{pmatrix}. \tag{3.4}$$

We finally remark that, since $\mathbf{S}^{\mathbf{X}}(f)$ is a Hermitian matrix, $\mathrm{Re}(\mathcal{S}^{\mathbf{X}}(f))$ and $\mathrm{Im}(\mathcal{S}^{\mathbf{X}}(f))$ are the vectorization of a symmetric and antisymmetric matrix, respectively.

3.2 Inverse modelling

Given a realization $\{\mathbf{y}(t)\}_{t\in\mathbb{R}}$ of the observable process $\{\mathbf{Y}(t)\}_{t\in\mathbb{R}}$, our approach aims at providing a sparse estimation of the cross-power spectrum $\mathbf{S}^{\mathbf{X}}(f)$ by exploiting the model described by Equation (3.4). When studying cortical functional connectivity from M/EEG data, this gives rise to a high-dimensional optimisation problem. In fact the forward matrix

$$\boldsymbol{\mathcal{G}} := \begin{pmatrix} \mathbf{G} \otimes \mathbf{G} & 0 \\ 0 & \mathbf{G} \otimes \mathbf{G} \end{pmatrix} \tag{3.5}$$

has size $2m^2 \times 2n^2$, where, in connectivity studies, $m \propto 10^2$ is the number of M/EEG sensors and $n \propto 10^3$ is the number of cortical locations where the brain activity and connectivity are estimated. Hence the matrix $\boldsymbol{\mathcal{G}}$ has size proportional to $(2 \cdot 10^4) \times (2 \cdot 10^6)$ making it necessary to develop proper computational strategies, described in Section 3.3, to carry out any algorithmic step involving such a matrix.

Throughout this chapter our approach is compared to a classical two-step approach summarized in the next subsection.

3.2.1 Benchmark: classical two–step approach

A widely used approach for estimating the cross-power spectrum $\mathbf{S}^{\mathbf{X}}(f)$ consists in the following two steps [55, 54, 14].

Step 1. A regularised estimate, $\{\mathbf{x}_\lambda(t)\}_{t\in\mathbb{R}}$, of the unknown process $\{\mathbf{X}(t)\}_{t\in\mathbb{R}}$, is obtained by solving the inverse problem associated to Equation (3.1). Here we consider the Tikhonov estimator presented in the previous chapter, defined as

$$\mathbf{x}_\lambda(t) = \arg\min_{\mathbf{x}(t)} \left\{ \|\mathbf{G}\mathbf{x}(t) - \mathbf{y}(t)\|_2^2 + \lambda \|\mathbf{x}(t)\|_2^2 \right\} \quad \forall t \in T\,, \tag{3.6}$$

where λ is a proper regularisation parameter, $\|\cdot\|_2$ is the ℓ_2-norm, and $T \subset \mathbb{N}$ is the discrete set of time points where the data were collected.

Step 2. The Welch's estimator, $\mathbf{S}^{\mathbf{x}_\lambda}(f)$, of the time–series $\{\mathbf{x}_\lambda(t)\}_{t\in T}$ is computed as follows [58]. First the time–series $\{\mathbf{x}_\lambda(t)\}_{t\in T}$ is partitioned in P overlapping segments of length L, denoted as $\left\{\mathbf{x}_\lambda^{(p)}(\tau)\right\}_{\tau=0,\dots,L-1}$ with $p \in \{1,\dots,P\}$, and the discrete Fourier transform

$$\hat{\mathbf{x}}_\lambda^{(p)}(f) = \frac{1}{L}\sum_{\tau=0}^{L-1} \mathbf{x}_\lambda^{(p)}(\tau) w(\tau) e^{\frac{-2\pi i \tau f}{L}}$$

is computed for each segment, $\{w(\tau)\}_{\tau=0,\dots,L-1}$ being the Hamming window [59]. Then we define

$$\mathbf{S}^{\mathbf{x}_\lambda}(f) = \frac{L}{PW}\sum_{p=1}^{P} \hat{\mathbf{x}}_\lambda^p(f)\hat{\mathbf{x}}_\lambda^p(f)^H,$$

where $W = \frac{1}{L}\sum_{t=0}^{L-1} w(t)^2$.

3.2.2 Direct estimation of the cross–power spectrum of the hidden process

Fixed a frequency f, we propose to estimate the cross-power spectrum $\mathbf{S}^{\mathbf{x}}(f)$ through a least-squares approach with ℓ_1 regularization applied to the model in Equation (3.4). Hence, after computing the Welch's estimator $\mathbf{S}^{\mathbf{y}}(f)$ of the cross–power spectrum of the observed data, we define [60, 57]

$$\begin{pmatrix} \widehat{\mathrm{Re}(\mathcal{S}^{\mathbf{x}})} \\ \widehat{\mathrm{Im}(\mathcal{S}^{\mathbf{x}})} \end{pmatrix} = \underset{\mathbf{s}}{\operatorname{argmin}} \left\{ \left\| \begin{pmatrix} \mathbf{G} \otimes \mathbf{G} & 0 \\ 0 & \mathbf{G} \otimes \mathbf{G} \end{pmatrix} \mathbf{s} - \begin{pmatrix} \mathrm{Re}(\mathcal{S}^{\mathbf{y}}) \\ \mathrm{Im}(\mathcal{S}^{\mathbf{y}}) \end{pmatrix} \right\|_2^2 + \lambda \|\mathbf{s}\|_1 \right\}, \quad (3.7)$$

where, for the sake of readability, we set $\mathbf{s} = \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{pmatrix}$ and we omit the specification of the argument f.

In order to solve the optimisation problem in (3.7) we used the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [18] introduced in Chapter 2. Algorithm 1 describes its implementation. In particular we observe that in line 4 and 6 of the algorithm we have exploited the block-diagonal structure of the forward matrix defined in Equation (3.5).

Following the original paper by Beck and Teboulle [18], in Algorithm 1 we used the shrinkage operator $\mathcal{T}_{\frac{\lambda}{L}} : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ defined as, for all $i = 1, \ldots, n^2$,

$$\mathcal{T}_{\frac{\lambda}{L}}(\mathbf{x})_i := \left(|x_i| - \frac{\lambda}{L} \right)^+ \mathrm{sign}(x_i) = \begin{cases} x_i - \frac{\lambda}{L} & \text{if } x_i \geq \frac{\lambda}{L} \\ 0 & \text{if } |x_i| \leq \frac{\lambda}{L} \\ x_i + \frac{\lambda}{L} & \text{if } x_i \leq -\frac{\lambda}{L} \end{cases}, \quad (3.8)$$

where for all $z \in \mathbb{R}$, $(z)^+ := \max\{z, 0\}$ and $\mathrm{sign}(z)$ denote the ramp function (or positive part) and the sign function, respectively. The Lipschitz constant L

is computed as

$$L = 2\lambda_{\max}(\boldsymbol{\mathcal{G}}^\top \boldsymbol{\mathcal{G}}), \tag{3.9}$$

with $\lambda_{\max}(\boldsymbol{\mathcal{G}}^\top \boldsymbol{\mathcal{G}})$ being the highest eigenvalue of the squared matrix $\boldsymbol{\mathcal{G}}^\top \boldsymbol{\mathcal{G}}$.

Remark 4. *The value in Equation (3.9) is a common choice for the constant L in optimisation problems of the form (3.7) [18] . In fact, it corresponds to the Lipschitz constant of the gradient of the function* $\mathbf{f} : \mathbb{R}^{2n^2} \to \mathbb{R}$*, defined as*

$$\mathbf{f}\begin{pmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{pmatrix} = \left\| \boldsymbol{\mathcal{G}}\begin{pmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{pmatrix} - \begin{pmatrix}\mathrm{Re}(\mathcal{S}^{\mathbf{y}})\\ \mathrm{Im}(\mathcal{S}^{\mathbf{y}})\end{pmatrix} \right\|_2^2 .$$

In our case, computing the Lipschitz constant L by means of Equation (3.9) requires determing the eigenvalues of a matrix of size $2n^2 \times 2n^2$*. However, a more efficient computation may be carried on by exploiting the block-diagonal structure of* $\boldsymbol{\mathcal{G}}$ *and the properties of the Kronecker product:*

$$\begin{aligned} L &= 2\lambda_{\max}\left(\begin{pmatrix} \mathbf{G}\otimes\mathbf{G} & 0 \\ 0 & \mathbf{G}\otimes\mathbf{G} \end{pmatrix}^\top \begin{pmatrix} \mathbf{G}\otimes\mathbf{G} & 0 \\ 0 & \mathbf{G}\otimes\mathbf{G} \end{pmatrix} \right) \\ &= 2\lambda_{\max}\begin{pmatrix} \mathbf{G}^\top\mathbf{G}\otimes\mathbf{G}^\top\mathbf{G} & 0 \\ 0 & \mathbf{G}^\top\mathbf{G}\otimes\mathbf{G}^\top\mathbf{G} \end{pmatrix} \\ &= 2\left[\lambda_{\max}(\mathbf{G}^\top\mathbf{G})\right]^2 . \end{aligned}$$

Hence, from a practical point of view, L can be computed as

$$L = 2\sigma^4_{\max}(\mathbf{G}), \tag{3.10}$$

with $\sigma_{\max}(\mathbf{G})$ *being the highest singular value of* $\mathbf{G}$*.*

Since we seek a solution to the optimisation problem (3.7) to be interpreted as the cross-power spectrum of a stochastic process, $\mathbf{s}_1$ and $\mathbf{s}_2$ need to be the vectorization of a symmetric and antisymmetric matrix, respectively. The following results prove that this can be achieved by initialising FISTA through a random Hermitian matrix, as done in Algorithm 1.

Algorithm 1 FISTA for a direct estimation of the cross–power spectrum $\mathbf{S}^{\mathbf{x}}$ in the frequency domain.

Input: $\mathbf{G} \in \mathbb{R}^{m \times n}$; λ, L, K, $\varepsilon > 0$; $\mathbf{S}^{\mathbf{y}} \in \mathbb{C}^{m \times m}$ and $\mathbf{S}_0 \in \mathbb{C}^{n \times n}$ Hermitian matrices and corresponding vectorization $\mathcal{S}^{\mathbf{y}} := \mathrm{vec}(\mathbf{S}^{\mathbf{y}}) \in \mathbb{C}^{m^2}$ and $\mathcal{S}_0 := \mathrm{vec}(\mathbf{S}_0) \in \mathbb{C}^{n^2}$.

1: **Initialize:** $k = 0$, $t_0 = 0$, $\begin{pmatrix} \mathbf{w}_{0,1} \\ \mathbf{w}_{0,2} \end{pmatrix} = \begin{pmatrix} \mathbf{s}_{0,1} \\ \mathbf{s}_{0,2} \end{pmatrix} = \begin{pmatrix} \mathrm{Re}(\mathcal{S}_0) \\ \mathrm{Im}(\mathcal{S}_0) \end{pmatrix}$

2: **while** $k \leq K$ **and** $e \leq \varepsilon$ **do**

3: $\quad k = k + 1$

4: $\quad \begin{pmatrix} \mathbf{s}_{k,1} \\ \mathbf{s}_{k,2} \end{pmatrix} = \begin{pmatrix} \mathcal{T}_{\frac{\lambda}{L}} \left(\mathbf{w}_{k-1,1} - \frac{2}{L} (\mathbf{G} \otimes \mathbf{G})^\top ((\mathbf{G} \otimes \mathbf{G}) \mathbf{w}_{k-1,1} - \mathrm{Re}(\mathcal{S}^{\mathbf{y}})) \right) \\ \mathcal{T}_{\frac{\lambda}{L}} \left(\mathbf{w}_{k-1,2} - \frac{2}{L} (\mathbf{G} \otimes \mathbf{G})^\top ((\mathbf{G} \otimes \mathbf{G}) \mathbf{w}_{k-1,2} - \mathrm{Im}(\mathcal{S}^{\mathbf{y}})) \right) \end{pmatrix}$

5: $\quad t_k = \frac{1 + \sqrt{1 + 4 t_{k-1}^2}}{2}$

6: $\quad \begin{pmatrix} \mathbf{w}_{k,1} \\ \mathbf{w}_{k,2} \end{pmatrix} = \begin{pmatrix} \mathbf{s}_{k,1} \\ \mathbf{s}_{k,2} \end{pmatrix} + \frac{t_{k-1} - 1}{t_k} \begin{pmatrix} \mathbf{s}_{k,1} - \mathbf{s}_{k-1,1} \\ \mathbf{s}_{k,2} - \mathbf{s}_{k-1,2} \end{pmatrix}$

7: $\quad e = \frac{||\mathbf{s}_{k,1} - \mathbf{s}_{k-1,1}||_1 + ||\mathbf{s}_{k,2} - \mathbf{s}_{k-1,2}||_1}{||\mathbf{s}_{k,1}||_1 + ||\mathbf{s}_{k,2}||_1}$

8: **end while**

Output: $\begin{pmatrix} \mathbf{s}_{k,1} \\ \mathbf{s}_{k,2} \end{pmatrix}$

Proposition 3. *Let $\mathcal{A} := \mathrm{vec}(\mathbf{A}) \in \mathbb{R}^{m^2}$ be the vectorization of a matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ and let $\mathcal{T}_\alpha$ be the shrinkage operator defined as in Equation (3.8) by replacing $\frac{\lambda}{L}$ with a generic parameter $\alpha > 0$. The following properties hold:*

(a) if $\mathbf{A}$ is symmetric, then $\mathcal{T}_\alpha(\mathcal{A})$ is the vectorization of a symmetric matrix;

(b) if $\mathbf{A}$ is antisymmetric, then $\mathcal{T}_\alpha(\mathcal{A})$ is the vectorization of an antisymmetric matrix.

Proof. Since $\mathbf{A}_{ij} = \mathcal{A}_{m(j-1)+i}$ for all $i, j = 1, \dots, m$, it can be shown that

$$\mathbf{A} \text{ is symmetric} \iff \mathcal{A}_{m(j-1)+i} = \mathcal{A}_{m(i-1)+j}, \text{ for all } i, j = 1, \dots, m,$$

and

$$\mathbf{A} \text{ is antisymmetric} \iff \mathcal{A}_{m(j-1)+i} = -\mathcal{A}_{m(i-1)+j}, \text{ for all } i, j = 1, \dots, m.$$

Hence, the thesis follows by the definition of $\mathcal{T}_\alpha$ observing that if $\mathbf{A}$ is symmetric then

$$\begin{aligned}\mathcal{T}_\alpha(\mathcal{A})_{m(j-1)+i} &= \Big(|\mathcal{A}_{m(j-1)+i}| - \alpha\Big)^+ \operatorname{sign}(\mathcal{A}_{m(j-1)+i}) \\ &= \Big(|\mathcal{A}_{m(i-1)+j}| - \alpha\Big)^+ \operatorname{sign}(\mathcal{A}_{m(i-1)+j}) = \mathcal{T}_\alpha(\mathcal{A})_{m(i-1)+j}.\end{aligned}$$

Similarly, if $\mathbf{A}$ is antisymmetric, then

$$\begin{aligned}\mathcal{T}_\alpha(\mathcal{A})_{m(j-1)+i} &= \Big(|\mathcal{A}_{m(j-1)+i}| - \alpha\Big)^+ \operatorname{sign}(\mathcal{A}_{m(j-1)+i}) \\ &= \Big(|-\mathcal{A}_{m(i-1)+j}| - \alpha\Big)^+ \operatorname{sign}(-\mathcal{A}_{m(i-1)+j}) = -\mathcal{T}_\alpha(\mathcal{A})_{m(i-1)+j}.\end{aligned}$$

□ □

Theorem 11. *Let $\left\{\begin{pmatrix}\mathbf{s}_{k,1}\\ \mathbf{s}_{k,2}\end{pmatrix}\right\}_{k\geq 0}$ and $\left\{\begin{pmatrix}\mathbf{w}_{k,1}\\ \mathbf{w}_{k,2}\end{pmatrix}\right\}_{k\geq 0}$ the sequences generated with Algorithm 1 with input the cross–power spectrum of the observed data $\mathbf{S}^{\mathbf{y}} \in \mathbb{C}^{m\times m}$ and a random Hermitian matrix $\mathbf{S}_0 \in \mathbb{C}^{n\times n}$ as initial point. Then $\{\mathbf{s}_{k,1}\}_{k\geq 0}$ and*

$\{\mathbf{w}_{k,1}\}_{k\geq 0}$ are vectorizations of symmetric matrices and $\{\mathbf{s}_{k,2}\}_{k\geq 0}$ and $\{\mathbf{w}_{k,2}\}_{k\geq 0}$ are vectorizations of antisymmetric matrices.

Proof. To prove the theorem we proceed by induction.

For $k = 0$, $\begin{pmatrix}\mathbf{w}_{0,1}\\ \mathbf{w}_{0,2}\end{pmatrix} = \begin{pmatrix}\mathbf{s}_{0,1}\\ \mathbf{s}_{0,2}\end{pmatrix} = \begin{pmatrix}\mathrm{Re}(\mathcal{S}_0)\\ \mathrm{Im}(\mathcal{S}_0)\end{pmatrix}$. Hence the thesis follows from the fact that $\mathbf{S}_0$ is Hermitian.

Fixed $k > 0$, we assume that $\mathbf{s}_{k-1,1}$ and $\mathbf{w}_{k-1,1}$ are vectorizations of two symmetric matrices denoted with $\mathbf{S}_{k-1,1}$ and $\mathbf{W}_{k-1,1}$, respectively. Analogously, we assume that $\mathbf{s}_{k-1,2}$ and $\mathbf{w}_{k-1,2}$ are vectorizations of two antisymmetric matrices denoted with $\mathbf{S}_{k-1,2}$ and $\mathbf{W}_{k-1,2}$, respectively.

By exploiting the linearity of the $\mathrm{vec}(\cdot)$ operator and the properties of the Kronecker product, from line 4 of Algorithm 1 it follows:

$$\begin{aligned}\mathbf{s}_{k,1} &= \mathcal{T}_{\frac{\lambda}{L}}\left(\mathbf{w}_{k-1,1} - \frac{2}{L}\left(\mathbf{G}\otimes\mathbf{G}\right)^\top((\mathbf{G}\otimes\mathbf{G})\mathbf{w}_{k-1,1} - \mathrm{Re}(\mathcal{S}^{\mathbf{y}})\right)\right)\\ &= \mathcal{T}_{\frac{\lambda}{L}}\left(\mathrm{vec}(\mathbf{W}_{k-1,1}) - \frac{2}{L}\left(\mathbf{G}\otimes\mathbf{G}\right)^\top\mathrm{vec}\left(\mathbf{G}\mathbf{W}_{k-1,1}\mathbf{G}^\top - \mathrm{Re}(\mathbf{S}^{\mathbf{y}})\right)\right) \qquad (3.11)\\ &= \mathcal{T}_{\frac{\lambda}{L}}\left(\mathrm{vec}\left(\mathbf{W}_{k-1,1} - \frac{2}{L}\mathbf{G}^\top\mathbf{G}\mathbf{W}_{k-1,1}\mathbf{G}^\top\mathbf{G} + \frac{2}{L}\mathbf{G}^\top\mathrm{Re}(\mathbf{S}^{\mathbf{y}})\mathbf{G}\right)\right).\end{aligned}$$

Since $\mathbf{W}_{k-1,1}$ and $\mathrm{Re}(\mathbf{S}^{\mathbf{y}})$ are symmetric for the inductive hypothesis and the properties of the cross-power spectrum, respectively, the argument on the right side of the last equation in (3.11) results to be a symmetric matrix. Hence Proposition 3 ensures $\mathbf{s}_{k,1}$ is the vectorization of a symmetric matrix, $\mathbf{S}_{k,1}$.

Similarly from line 4 of Algorithm 1 it follows:

$$\begin{aligned}\mathbf{s}_{k,2} &= \mathcal{T}_{\frac{\lambda}{L}}\left(\mathbf{w}_{k-1,2} - \frac{2}{L}(\mathbf{G}\otimes\mathbf{G})^\top((\mathbf{G}\otimes\mathbf{G})\mathbf{w}_{k-1,2} - \mathrm{Im}(\mathcal{S}^{\mathbf{y}}))\right)\\ &= \mathcal{T}_{\frac{\lambda}{L}}\left(\mathrm{vec}\left(\mathbf{W}_{k-1,2} - \frac{2}{L}\mathbf{G}^\top\mathbf{G}\mathbf{W}_{k-1,2}\mathbf{G}^\top\mathbf{G} + \frac{2}{L}\mathbf{G}^\top\mathrm{Im}(\mathbf{S}^{\mathbf{y}})\mathbf{G}\right)\right),\end{aligned}$$

where the argument of the right side of the last equation is antisymmetric because $\mathbf{W}_{k-1,2}$ and $\mathrm{Im}(\mathbf{S}^{\mathbf{y}})$ are antisymmetric. Hence Proposition 3 ensures

$\mathbf{s}_{k,2}$ is the vectorization of a antisymmetric matrix, $\mathbf{S}_{k,2}$.

The fact that $\mathbf{w}_{k,1}$ and $\mathbf{w}_{k,2}$ are the vectorizations of a symmetric and an antisymmetric matrix then follows from line 6 of Algorithm 1, that can be rewritten as follows

$$\begin{pmatrix} \mathbf{w}_{k,1} \\ \mathbf{w}_{k,2} \end{pmatrix} = \begin{pmatrix} \operatorname{vec}\left(\mathbf{S}_{k,1} + \frac{t_{k-1}-1}{t_k}(\mathbf{S}_{k,1} - \mathbf{S}_{k-1,1})\right) \\ \operatorname{vec}\left(\mathbf{S}_{k,2} + \frac{t_{k-1}-1}{t_k}(\mathbf{S}_{k,2} - \mathbf{S}_{k-1,2})\right) \end{pmatrix}.$$

□

3.3 Smart product for an efficient computation of the FISTA update

The FISTA update at line 4 of Algorithm 1 requires several matrix–vector multiplications involving the matrix $\mathbf{G} \otimes \mathbf{G}$ and its transpose $(\mathbf{G} \otimes \mathbf{G})^\top = \mathbf{G}^\top \otimes \mathbf{G}^\top$. Since $\mathbf{G} \in \mathbb{R}^{m \times n}$, $\mathbf{G} \otimes \mathbf{G}$ and $\mathbf{G}^\top \otimes \mathbf{G}^\top$ have size $m^2 \times n^2$ and $n^2 \times m^2$, respectively. Hence, the product of a vector by each one of these matrices has a cost proportional to $O(m^2 n^2)$. In this section we show how the properties of the Kronecker product can be exploited to reduce both the computational cost and the memory requirements of such product. More specifically, our approach has two main advantages: on the one hand it reduces the computational cost to a value proportional to $O(\max(m, n)\, mn)$, on the other hand it avoids explicitly assembling the matrix $\mathbf{G} \otimes \mathbf{G}$, by directly employing the matrix $\mathbf{G}$.

First, we recall that the Kronecker product enjoys the mixed-product property [61], therefore it holds

$$\underbrace{\mathbf{G} \otimes \mathbf{G}}_{m^2 \times n^2} = (\mathbf{G}\, \mathbf{I}_n) \otimes (\mathbf{I}_m\, \mathbf{G}) = \underbrace{(\mathbf{G} \otimes \mathbf{I}_m)}_{m^2 \times nm}\, \underbrace{(\mathbf{I}_n \otimes \mathbf{G})}_{nm \times n^2}. \tag{3.12}$$

Now we present a result that establishes a connection between the elements of $\mathbf{G} \otimes \mathbf{I}_m$ and those of $\mathbf{I}_m \otimes \mathbf{G}$ through matrices that permute rows

and columns. This relationship plays a crucial role in the procedure for efficiently computing the matrix-vector product with $(\mathbf{G} \otimes \mathbf{G})$ (or its transpose, $(\mathbf{G} \otimes \mathbf{G})^\top$).

Proposition 4. *Let $\mathbf{G} \in \mathbb{R}^{m \times n}$ be a rectangular matrix. Then there exist two matrices $\mathbf{P}_{m^2} \in \mathbb{R}^{m^2 \times m^2}$ and $\mathbf{P}_{nm} \in \mathbb{R}^{nm \times nm}$ such that*

$$\mathbf{G} \otimes \mathbf{I}_m = \mathbf{P}_{m^2}(\mathbf{I}_m \otimes \mathbf{G})\mathbf{P}_{nm}. \tag{3.13}$$

Proof. The thesis is equivalent to identifying two permutation functions for the row and column indices that rearrange the elements of $\mathbf{I}_m \otimes \mathbf{G}$, making it identical to $\mathbf{G} \otimes \mathbf{I}_m$. Relation (3.13) is then established by defining $\mathbf{P}_{m^2} \in \mathbb{R}^{m^2 \times m^2}$ and $\mathbf{P}_{nm} \in \mathbb{R}^{nm \times nm}$ as identity matrices of the corresponding sizes, with the specified permutations applied to their rows and columns [61].

Then we first highlight that the structures of $\mathbf{G} \otimes \mathbf{I}_m$ and $\mathbf{I}_m \otimes \mathbf{G}$ are given by

$$\mathbf{G} \otimes \mathbf{I}_m = \underbrace{\begin{bmatrix} g_{1,1}\mathbf{I}_m & g_{1,2}\mathbf{I}_m & \cdots & g_{1,n}\mathbf{I}_m \\ g_{2,1}\mathbf{I}_m & g_{2,2}\mathbf{I}_m & \cdots & g_{2,n}\mathbf{I}_m \\ \vdots & \vdots & \ddots & \vdots \\ g_{m,1}\mathbf{I}_m & g_{m,2}\mathbf{I}_m & \cdots & g_{m,n}\mathbf{I}_m \end{bmatrix}}_{nm \text{ columns}}, \quad \mathbf{I}_m \otimes \mathbf{G} = \left.\begin{bmatrix} \mathbf{G} & O_{m,n} & \cdots & O_{m,n} \\ O_{m,n} & \mathbf{G} & \cdots & O_{m,n} \\ \vdots & \vdots & \ddots & \vdots \\ O_{m,n} & O_{m,n} & \cdots & \mathbf{G} \end{bmatrix}\right\} m^2 \text{ rows}$$

Consequently, we obtain the elements of $\mathbf{G} \otimes \mathbf{I}_m$ from those of $\mathbf{I}_m \otimes \mathbf{G}$ with a permutation that brings the $i-$th row in position

$$\mathrm{mod}(i-1, m)\, m + \left\lfloor \frac{i-1}{m} \right\rfloor + 1, \quad \text{for } i = 1 \dots m^2; \tag{3.14}$$

and the $j-$th column in position

$$\mathrm{mod}(j-1, n)\, m + \left\lfloor \frac{j-1}{n} \right\rfloor + 1, \quad \text{for } j = 1 \dots nm, \tag{3.15}$$

where, for all $a, b \in \mathbb{N}$ with $\mathrm{mod}(a, b)$ we denote the remainder after the

division of a by b and with $\lfloor a \rfloor$ the floor of a.

Indeed, note that, $\mathrm{mod}(i-1, m)$ (resp. $\mathrm{mod}(j-1, n)$) is the quantity that serves to group the row indices (resp. column) into blocks of m rows (resp. column), determining the relative position within each block. The value $\left\lfloor \frac{i-1}{m} \right\rfloor$ (resp. $\left\lfloor \frac{j-1}{n} \right\rfloor$) determines the block number we are considering. See [62] and [63] for further visualizations and definitions on the block rectangular permutations.

Let $\mathbf{e}_\ell$ be the $\ell-$th canonical vector with 1 in position ℓ and zeros elsewhere, that is $(\mathbf{e}_\ell)_j = \delta_{j\ell}$. We conclude the proof choosing $\mathbf{P}_{m^2}$ and $\mathbf{P}_{nm}$ the two permutation matrices where the one entry equal to 1 is given exploiting the two permutation functions (3.14) and (3.15). In details,

$$\mathbf{P}_{m^2} = \left[e_{\xi_r(1)} \;\middle|\; e_{\xi_r(2)} \;\middle|\; \cdots \;\middle|\; e_{\xi_r(m^2)} \right], \quad \mathbf{P}_{nm} = \left[e_{\xi_c(1)} \;\middle|\; e_{\xi_c(2)} \;\middle|\; \cdots \;\middle|\; e_{\xi_c(nm)} \right],$$

where

$$\begin{aligned} \xi_r(i) &= \mathrm{mod}(i-1, m)\, m + \left\lfloor \frac{i-1}{m} \right\rfloor + 1, \quad i = 1 \dots m^2, \\ \xi_c(j) &= \mathrm{mod}(j-1, n)\, m + \left\lfloor \frac{j-1}{n} \right\rfloor + 1, \quad j = 1 \dots nm. \end{aligned} \tag{3.16}$$

□

Theorem 12. *Let $\mathbf{G} \in \mathbb{R}^{m \times n}$ be a rectangular matrix and $\mathbf{x} \in \mathbb{R}^{n^2}$. The matrix–vector product $(\mathbf{G} \otimes \mathbf{G})\mathbf{x}$ can be performed in $O(\max(m, n)\, mn)$ operations with a matrix–less approach (denoted by Algorithm 2).*

Proof. Equation (3.12) and Proposition 4 imply

$$(\mathbf{G} \otimes \mathbf{G})\mathbf{x} = (\mathbf{G} \otimes \mathbf{I}_m)(\mathbf{I}_n \otimes \mathbf{G})\mathbf{x} = \mathbf{P}_{m^2}(\mathbf{I}_m \otimes \mathbf{G})\mathbf{P}_{nm}(\mathbf{I}_n \otimes \mathbf{G})\mathbf{x},$$

where $\mathbf{P}_{m^2}$ and $\mathbf{P}_{nm}$ are the permutation matrices defined by functions ξ_r and ξ_c in Equation (3.16) in Proposition 4. Hence the product $(\mathbf{G} \otimes \mathbf{G})\mathbf{x}$ can be decomposed in few steps that form Algorithm 2. We now analyse the proposed procedure and demonstrate that it does not require assembling any of

the matrix tensor products. Moreover, we prove that the total computational cost of each step is proportional to $\max(m,n)\cdot mn$.

The cost of step 2 and 4 of Algorithm 2 is 0 since they only require the index exchange suggested by ξ_r and ξ_c in Equation (3.16). The vector $\mathbf{x}\in\mathbb{R}^{n^2}$ can be split into n vectors $\hat{x}_\ell$ of size n. In particular,

$$\hat{x}_\ell = \left[x_{(\ell-1)n+k}\right]_{k=1}^{n}, \quad \ell = 1,\dots,n.$$

Similarly, the vector $\widehat{\mathbf{y}}$ can be split into m vectors of size n.

The latter and the block diagonal structures of $\mathbf{I}_n\otimes\mathbf{G}$ and $\mathbf{I}_m\otimes\mathbf{G}$, imply that the multiplications in steps 1 and 3 of Algorithm 2 avoid assembling the matrix $\mathbf{G}\otimes\mathbf{G}$ (or any other matrix tensor product). Moreover, they consist of n and m matrix-vector products with the matrix $\mathbf{G}$, respectively. Therefore the total computational cost is proportional to $n\cdot nm + m\cdot nm$, that is $O(\max(n,m)\cdot mn)$. □

Algorithm 2 Efficient computation of the matrix–vector multiplication $(\mathbf{G}\otimes\mathbf{G})\mathbf{x}$

Input: $\mathbf{G}\in\mathbb{R}^{m\times n}$; $\mathbf{x}\in\mathbb{R}^{n^2}$; $\mathbf{P}_{nm}$ and $\mathbf{P}_{m^2}$ as in Proposition 4.
1: Compute $\mathbf{y} := (\mathbf{I}_n\otimes\mathbf{G})\,\mathbf{x}\in\mathbb{R}^{nm}$
2: Permute the element of $\mathbf{y}$ to define $\widehat{\mathbf{y}} := \mathbf{P}_{nm}\mathbf{y}$
3: Compute $\mathbf{z} := (\mathbf{I}_m\otimes\mathbf{G})\,\widehat{\mathbf{y}}\in\mathbb{R}^{m^2}$
4: Permute the element of $\mathbf{z}$ to define $\widehat{\mathbf{z}} := \mathbf{P}_{m^2}\mathbf{z}$
Output: $\widehat{\mathbf{z}}$

Remark 5. *Since the results presented in this section hold true for any generic rectangular matrix, Algorithm 2 can be straightforwardly applied for computing also the matrix–vector product* $\left(\mathbf{G}^\top\otimes\mathbf{G}^\top\right)\mathbf{x}$. *Moreover, in connectivity studies,* $m\propto 10^2$ *is the number of M/EEG sensors and* $n\propto 10^3$ *is the number of cortical locations and in general* $n >> m$. *Then, the estimation cost of the Theorem 12 becomes* $O(m^2)$.

Remark 6. *The previous result allows us to handle the natural increase in the complexity of the inverse problem and of the associated optimisation problem, for which an*

iterative scheme such as FISTA is required. We stress that, in terms of computational efficiency, in the classical Tikhonov regularization framework, the inverse problem reduces instead to the solution of a smooth least-squares problem, which can be handled efficiently. The latter is not a computational bottleneck that our method is designed to address. However, in the context of M/EEG connectivity studies, efficiency should rather be understood in terms of the trade-off between computational effort and reconstruction accuracy. The method we propose attains substantially higher accuracy in regimes where Tikhonov regularization fails, at the price of solving a more involved linear inverse problem. The results of Theorem 12 and Algorithm 2, together with the complexity estimate discussed above, ensure that this increased algebraic complexity does not translate into prohibitive per-iteration costs, so that the FISTA-based scheme remains computationally viable in realistic large-scale scenarios.

3.4 Numerical validation

This section is organised as follows. In Subsections 3.4.1 and 3.4.2 we describe how MEG synthetic data were simulated an analysed using both the classical two–step approach and the proposed one, referred as one-step approach. The results and comparison of the two methods are summarized in Subsection 3.4.3. The Python codes implementing the proposed approach are freely available at the GitHub repository `https://github.com/theMIDAgroup/fista_cps_conn`. We highlight that our approach may still be regarded as a two-step procedure as it requires the estimation of the cross-power spectrum of the recorded time–series before the optimisation step. However, since the optimisation is carried out in the frequency domain and the regularization is applied directly to the variable to be estimated, the use of the term one-step is justified for the sake of conciseness in the current numerical section.

3.4.1 Brain connectivity configurations

To test the performance of the proposed approach under different experimental conditions, we considered the following two brain configurations.

Configuration 1: three active sources with unidirectional coupling from source 1 to source 2. Inspired by previous works [64, 14, 65], the time–courses of the active sources were simulated by filtering in the band $[8, 12]$Hz (α band) a multivariate autoregressive (MVAR) process of order $P = 5$ defined as

$$\begin{pmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \end{pmatrix} = \sum_{k=1}^{P} \begin{pmatrix} a_{1,1}(k) & 0 & 0 \\ a_{2,1}(k) & a_{2,2}(k) & 0 \\ 0 & 0 & a_{3,3}(k) \end{pmatrix} \begin{pmatrix} z_1(t-k) \\ z_2(t-k) \\ z_3(t-k) \end{pmatrix} + \begin{pmatrix} \epsilon_1(t) \\ \epsilon_2(t) \\ \epsilon_3(t) \end{pmatrix} . \quad (3.17)$$

The non-zero elements $a_{i,j}(k)$ of the coefficient matrix were drawn from a normal distribution of zero mean and standard deviation 0.9. We retained only coefficients resulting in (i) a stable MVAR process [66] and (ii) triplets of signals, $(z_1(t), z_2(t), z_3(t))^T$, such that the ℓ_2 norm of the strongest one is less than 3 times the ℓ_2 norm of the weakest one and such that the average over the range $[8, 12]$Hz of the sum of their power spectra was at least 1.2 times the average over the entire frequency range [14].

Configuration 2: three sources with unidirectional coupling from source 1 to source 2 and 3. The time–courses of the active sources were generating as for Configuration 1, but substituting the model in Equation (3.17) with

$$\begin{pmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \end{pmatrix} = \sum_{k=1}^{P} \begin{pmatrix} a_{1,1}(k) & 0 & 0 \\ a_{2,1}(k) & a_{2,2}(k) & 0 \\ a_{3,1}(k) & 0 & a_{3,3}(k) \end{pmatrix} \begin{pmatrix} z_1(t-k) \\ z_2(t-k) \\ z_3(t-k) \end{pmatrix} + \begin{pmatrix} \epsilon_1(t) \\ \epsilon_2(t) \\ \epsilon_3(t) \end{pmatrix} . \quad (3.18)$$

3.4.2 Simulation and analysis of the observed MEG time–series

We exploited the model in Equation (3.1) for generating 50 realisations, $\{\mathbf{y}(t)\}_{t=1}^{T}$, of the observable process for each configuration defined in the previous section. Specifically, for each realisation, we fixed $T = 10,000$, and for all $t = 1, \ldots, T$, we set

$$\mathbf{y}(t) = \mathbf{G}\mathbf{x}(t) + \mathbf{e}(t), \quad (3.19)$$

where

- we extracted the forward operator $\mathbf{G}$ from the sample dataset within the MNE Python package [67] by considering only magnetometers, and by downsampling the available source space to $n = 6940$ points. Hence, in our numerical experiment $\mathbf{G}$ has size 102×6940 and each column $\mathbf{g}_i$, $i = 1, \dots, n$, represents the magnetic field generated by a point-like unit source placed at the i–th point of the source space with orientation normal to the local cortical surface;

- we defined $\mathbf{x}(t)$ by randomly selecting three points of the source space so that the pairwise source distances were greater than 4 cm and the pairwise ratios of the ℓ_2 norms of the corresponding columns of $\mathbf{G}$ were close to one. The components of $\mathbf{x}(t)$ corresponding to the drawn location were set equal to the time–courses defined by Equation (3.17) or (3.18) while the value of the remaining $n - 3$ components was kept equal to 0;

- we sampled $\mathbf{e}(t)$ from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$ where we chose σ^2 so as to obtain a signal-to-noise ratio (SNR) equal to 5 dB.

To test the proposed method, for each one of the simulated data, $\{\mathbf{y}(t)\}_{t=1}^{T}$, we then computed the Welch's estimator, $\mathbf{S}^{\mathbf{y}}(f)$, and we applied Algorithm 1 by setting f equal to the frequency in the range [8, 12]Hz where the component $S_{12}^{\mathbf{y}}(f)$ peaks. In our experiments, we set the maximum number of iterations K equal to 5,000, the tolerance ε equal to 10^{-5}, and we computed the Lipshitz constant L as in Equation (3.10). To avoid inverse crime and mimic real-life scenarios where the active brain sources seldom match points of the source space, the matrix $\mathbf{G}$ used within Algorithm 1 is obtained from that employed in Equation (3.19) for simulating the MEG data by further reducing the source space to $n = 644$ points. Finally, we tested four different values of the regularization parameters, by choosing four different scaling factor κ evenly spaced in log-space in the range $[10^{-2}, 10^{-1}]$. Then we set $\lambda = \kappa \lambda^*$,

where $\lambda^* = 2 \left\| \boldsymbol{\mathcal{G}} \begin{pmatrix} \mathrm{Re}(\mathcal{S}^{\mathbf{y}}) \\ \mathrm{Im}(\mathcal{S}^{\mathbf{y}}) \end{pmatrix} \right\|_{\infty}^{2}$, being $\mathcal{S}^{\mathbf{y}} = \mathrm{vec}(\mathbf{S}^{\mathbf{y}}(f))$. The value λ^* has been shown to be an upper bound for the optimisation problem (3.7) to admit a non-null solution [68].

The value λ^* has been shown to be an upper bound for the optimisation problem (3.7) to admit a non-null solution [68], while the lower bound $10^{-2}\lambda^*$ has been chosen to guarantee enough sparsity in the obtained estimates.

The performance of the proposed approach were compared to that of the classical two–step approach described in Section 3.2.1. In detail, we first computed the Tikhonov estimator $\{\mathbf{x}_\lambda(t)\}_{t=1}^{T}$ of the neural sources as in Equation (3.6) by using the coarse operator $\mathbf{G}$ including $n = 644$ source locations used also within Algorithm 1. Four different regularization parameters were tested, namely $\lambda = \xi\ 10^{-SNR/10}$, with $\xi \in \{0.1, 1, 10, 100\}$. This range has been chosen based on previous literature showing that the value $10^{-SNR/10}$ corresponds to the regularization parameter that provides the best possible estimate of the brain activity in case of uncorrelated Gaussian signals [13]. For each value of the parameter, we then computed the Welchs estimator, $\mathbf{S}^{\mathbf{x}_\lambda}(f)$, of the time–series $\{\mathbf{x}_\lambda(t)\}_{t \in T}$.

Let $\widehat{\mathbf{S}}$ be an estimate of the cross-power spectrum obtained with either the proposed method or the classical two–step approach. In order to quantitatively compare the two approaches, we separated the real and the imaginary part of $\widehat{\mathbf{S}}$. For both of them, we then computed a weighted sum over all the estimated pairwise interaction exceeding a given threshold of the Euclidean distance between the interacting-source locations and the closest pair of sources truly connected in sense of the Wasserstein 2-distance. More formally, denoted with $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{6940}\}$ and $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_{644}\}$ the source spaces associated to the forward operators used for simulating the data and within the inverse procedures, respectively, we defined

$$\mathrm{Err}^{\mathrm{Re}} = \sum_{(\mathbf{w}_i, \mathbf{w}_j) \in \mathcal{E}^{rec}} \frac{\left| \mathrm{Re}\left(\widehat{S}_{ij}\right) \right|}{\max\limits_{i<j} \left| \mathrm{Re}\left(\widehat{S}_{ij}\right) \right|} \min_{(\mathbf{v}_p, \mathbf{v}_q) \in \mathcal{E}^{true}} d\left((\mathbf{w}_i, \mathbf{w}_j), (\mathbf{v}_p, \mathbf{v}_q)\right), \qquad (3.20)$$

where $\mathcal{E}^{rec} = \left\{(\mathbf{w}_i, \mathbf{w}_j) \in \mathcal{W} \times \mathcal{W} \mid i < j \text{ and } \left|\text{Re}\left(\widehat{S}_{ij}\right)\right| \geq \tau\right\}$ collects the pairs of source location between which the estimated cross spectrum exceeds a given threshold, namely $\tau = 0.5 \max_{i<j} \left\{\left|\text{Re}\left(\widehat{S}_{ij}\right)\right|\right\}$; $\mathcal{E}^{true} \subset \mathcal{V} \times \mathcal{V}$ is the set of truly connected sources, and

$$d\left((\mathbf{w}_i, \mathbf{w}_j), (\mathbf{v}_p, \mathbf{v}_q)\right) = \sqrt{\frac{1}{2}\min\left\{||(\mathbf{w}_i, \mathbf{w}_j) - (\mathbf{v}_p, \mathbf{v}_q)||_2^2, ||(\mathbf{w}_i, \mathbf{w}_j) - (\mathbf{v}_q, \mathbf{v}_p)||_2^2\right\}}$$

is the Wasserstein 2-distance [69].

Similarly, Err^{Im} was defined as in (3.20) by replacing $\text{Re}\left(\widehat{S}_{ij}\right)$ with $\text{Im}\left(\widehat{S}_{ij}\right)$. For both the proposed method and the classical two-step approach we chose the value of the regularization parameter that provides the lowest total error $\text{Err}^{\text{Re}} + \text{Err}^{\text{Im}}$. The corresponding estimated cross-power spectra were compared in order to asses the advantages of the proposed method.

3.4.3 Results

Table 3.1 illustrated the behaviour of the proposed one–step approach when varying the amount of regularization. As expected, the higher the value of the regularization parameter the sparser the resulting estimation of the cross–power spectrum. In detail, for the highest value of the parameter, namely $\lambda_4 := 0.1\,\lambda^*$, the estimated cross-power spectrum exhibits no non-zero interactions in many simulated datasets. On the other hand, for the lowest value of the parameter, namely $\lambda_1 := 0.01\,\lambda^*$, only in two datasets for both the configurations the imaginary part of the estimated cross-power spectrum was equal to zero, but the number of supra-threshold connections increased up to few hundreds for Configuration 1 and few thousands for Configuration 2. We recall that the threshold τ was set equal to half the value of the strongest connection. Figures 3.1 and 3.2 confirm that these are reasonable ranges for both the methods by illustrating the level of sparsity reached by the obtained solutions and the behaviour of the total estimation error as function of the tested regularization parameters. Figures 3.1 and 3.2 also confirm that the total estimation error of the proposed method is systematically lower than

that of the classical two-step approach. As extensively discussed in previous literature (see [70, 13, 14]) , selecting the optimal regularization parameter for either the proposed or the classical approach is a cumbersome problem that lies outside the scope of this work.

Table 3.1: Level of sparsity in the solution provided by the proposed one-step approach for decreasing values of the regularization parameters. For each configuration, each cell of the table shows in the first row the percentage of simulated data where the real (first column) or the imaginary (second column) part of the estimated cross-power spectrum show at least one non-null interaction; in the second row the minimum and maximum number of supra-threshold connections across these data; and in the third row the mean number of supra-threshold interactions.

	Conf 1		**Conf 2**	
	Real part	**Imag. part**	**Real part**	**Imag. part**
λ_4	98.0% (1, 87) 4	58.0% (1, 41) 4	100.0% (1, 292) 17	78.0% (1, 284) 17
λ_3	98.0% (1, 156) 6	76.0% (1, 318) 18	100.0% (1, 401) 19	88.0% (1, 221) 14
λ_2	100.0% (1, 281) 20	86.0% (1, 462) 31	100.0% (1, 1013) 44	92.0% (1, 526) 30
λ_1	100.0% (1, 260) 20	96.0% (1, 325) 42	100.0% (1, 4303) 144	96.0% (1, 2942) 120

Figure 3.3 shows the main advantages of the proposed method over the classical two-step approach. In this example, both method identified connected sources nearby the truly interacting ones, however the number of false positive is much higher for the two–step approach. Further, for the two–step approach the value of the cross-power spectrum is underestimated of several orders of magnitude. As expected, this improved estimation accuracy comes at the price of a higher computational cost: as shown by Figure 3.4 for this simulation more than 3000 iterations of the FISTA algorithm were required before the objective function in Equation (3.7) and the total estimation er-

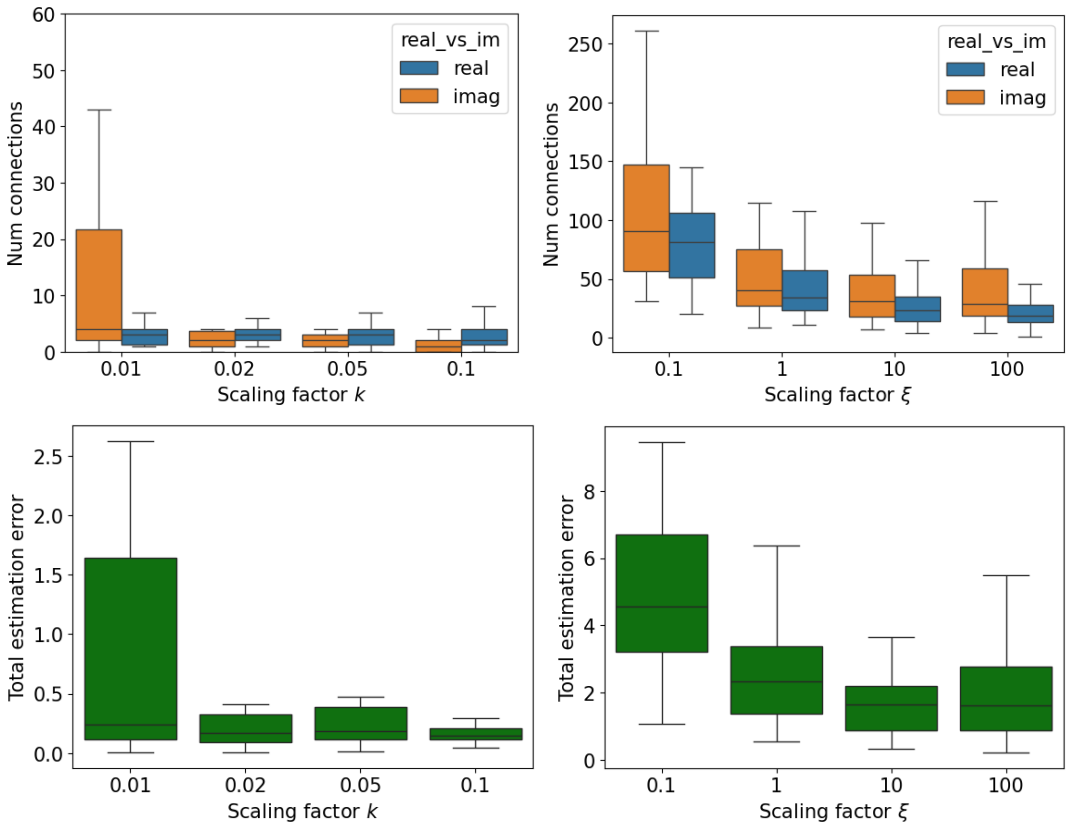

Figure 3.1: Number of supra-threshold connections and total estimation error $\mathrm{Err}^{\mathrm{Re}} + \mathrm{Err}^{\mathrm{Im}}$ as function of the tested scaling factors κ and ξ for Configuration 1 of the proposed approach (left column) and the classical two step approach (right column). Please notice the different scales on the y axis.

ror $\mathrm{Err}^{\mathrm{Re}} + \mathrm{Err}^{\mathrm{Im}}$ reached a plateau. As a consequence, the running time for analysing each one of the simulated MEG data with the proposed one-step approach is about 30 minutes, while the classical two-step method takes around 2 minutes.

More in general, in our experiments, the one-step approach outperformed the classical two-step method as demonstrated by Table 3.2 and Figure 3.5. More specifically, for both the approaches we selected for each simulated data the best regularization parameter among those tested as the one minimising the sum $\mathrm{Err}^{\mathrm{Re}} + \mathrm{Err}^{\mathrm{Im}}$. When using these optimal parameters, the average error of the one–step approach is systematically lower than that of the two–

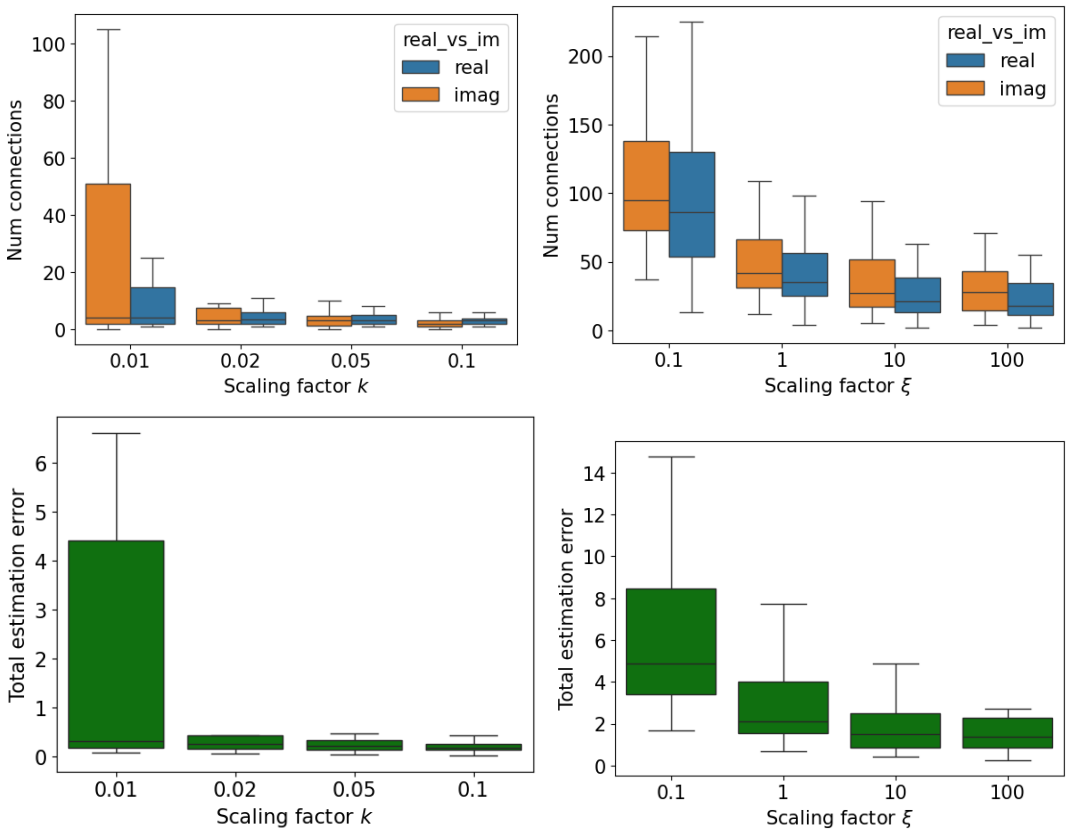

Figure 3.2: Number of supra-threshold connections and total estimation error $\mathrm{Err}^{\mathrm{Re}} + \mathrm{Err}^{\mathrm{Im}}$ as function of the tested scaling factors κ and ξ for Configuration 2 of the proposed approach (left column) and the classical two step approach (right column). Please notice the different scales on the y axis.

step approach for both the real and the imaginary part of the cross–power spectrum in both the configurations. More quantitatively, Table 3.2 shows that on average the total error $\mathrm{Err}^{\mathrm{Re}} + \mathrm{Err}^{\mathrm{Im}}$ of the one-step approach is almost one order of magnitude lower than the total error of the two-step approach. This is due to the fact that the number of spurious interactions is much higher for the two–step approach than for the proposed method, as shown in the second row of Figure 3.5.

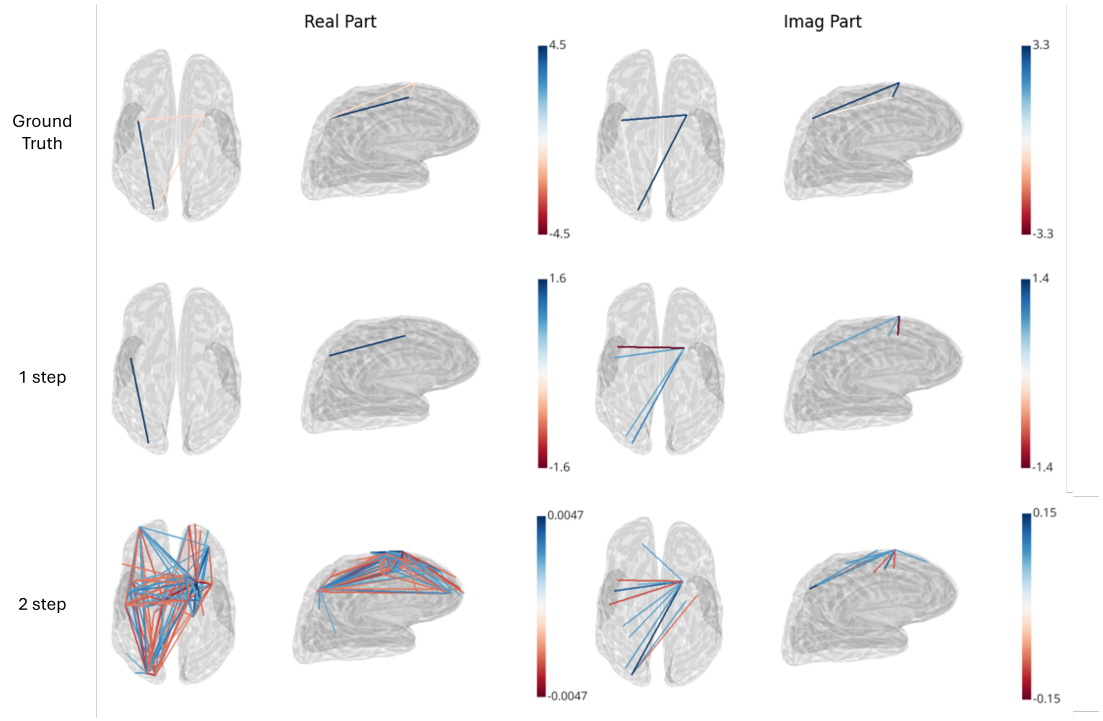

Figure 3.3: Original and estimated cross-power spectrum for one MEG data simulated by using Configuration 2. In this simulation, $\mathrm{Err}^{\mathrm{Re}}$ is equal to 0.077 for the one–step approach and to 0.521 for the two–step, while $\mathrm{Err}^{\mathrm{Im}}$ is equal to 0.019 and 0.086 respectively.

Table 3.2: Comparison of the total error $\mathrm{Err}^{\mathrm{Re}} + \mathrm{Err}^{\mathrm{Im}}$ for the best estimates of the one-step and the two-step approach. Each cell shows mean (minimum, maximum) value across the 50 simulated data.

	Conf 1	**Conf 2**
1-step	0.28 (0.006, 2.63)	0.39 (0.020, 3.04)
2-step	1.50 (0.23, 6.93)	1.65 (0.26, 8.86)

3.5 Discussion

The estimation of the brain cortical cross-power spectrum from M/EEG data is typically achieved in two steps: first the Tikhonov's least square estimator $\{\mathbf{x}_\lambda(t)\}_{t\in\mathbb{R}}$ of the brain cortical activity is computed, and then the cross-power spectrum of $\{\mathbf{x}_\lambda(t)\}_{t\in\mathbb{R}}$ is estimated through e.g. Welch's method. However this approach often results in a large number of false positives. This issue is

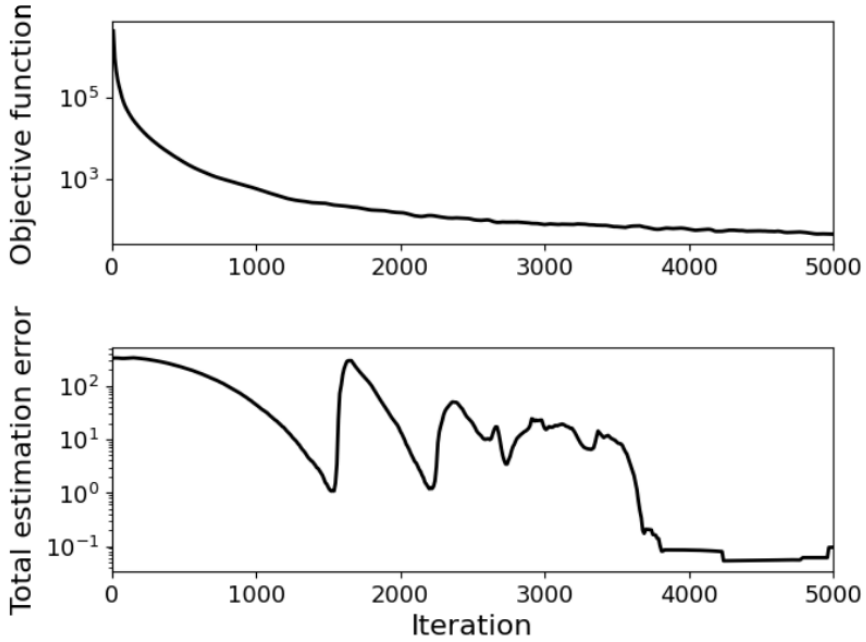

Figure 3.4: Objective function and total estimation error, $\mathrm{Err}^{\mathrm{Re}} + \mathrm{Err}^{\mathrm{Im}}$, with respect to the iteration number when the one-step is used for analysing the same simulated MEG data considered in Figure 3.3

partially overcome by deriving from the cross-power spectrum metrics, such as the imaginary part of coherency, that are insensitive to linear mixing of the truly interacting brain source. However, such metrics fail in identifying instantaneously correlated sources. In this work we overcome this issue by suggesting a novel approach that directly estimates the cross-power spectrum of the cortical sources from that of the recorded M/EEG time–series. The number of false positives is controlled by constructing an ℓ_1–regularized least square estimator of the cross-power spectrum computed through FISTA.

The proposed method leverages the tensorial structure of the global coefficient matrix. Therefore, we developed an efficient algorithm where the computational cost is primarily driven by operations involving the forward matrix $\mathbf{G}$. The advantages of our method over the classical two-step approach have been demonstrated on a large number of synthetic data mimicking different brain configurations.

From a methodological point of view, we are planning two main developments. The first one concerns the implementation of an automatic procedure to choose the regularization parameter. Moreover, advanced variants of FISTA, such as schemes with restarts or adaptive step sizes, possibly employing back-tracking strategies [71] or adaptive estimation of the Lipschitz constant L [72], could be employed to further improve the computational efficiency of the proposed method, and to enable a more systematic investigation of the trade-off between accuracy and computational cost.

The second one concerns implementing different approaches for solving the inverse problem in Equation (3.3), such as Bayesian Monte Carlo approaches [73]. The latter would have the advantage of also quantifying the uncertainty of the provided estimation even though at the price of a higher computational cost.

Finally, we observe that the current implementation of the proposed approach provides estimates of the cross-power spectrum at fixed frequencies. Future work will be devoted to investigate how to combine the information from multiple frequencies and/or frequency ranges.

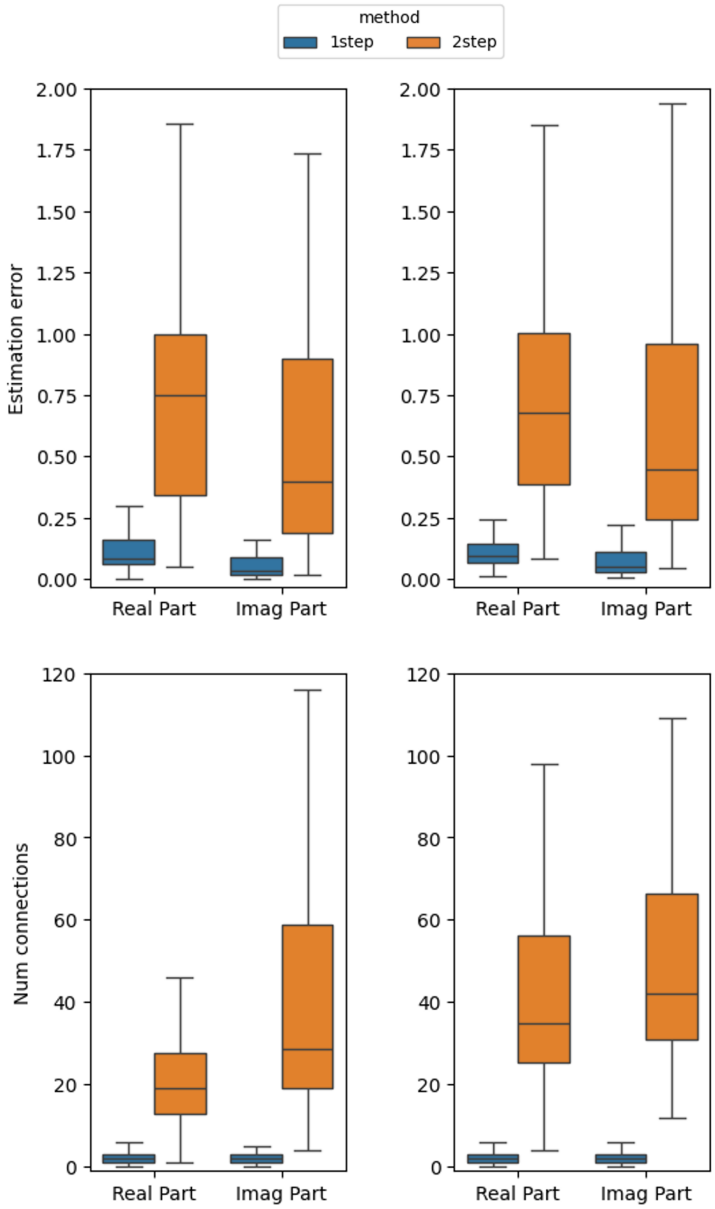

Figure 3.5: Quantitative comparison of the performance of the one-step and the two-step approach. First row: error in estimating the real and the imaginary part of the cortical cross-power spectrum. Second row: number of supra-threshold connections in the real and imaginary part of the estimated cross-power spectrum. In each row the left-side panel refers to the first simulated brain configuration where only 1 pair of truly connected sources are present, while the right-side panel refers to the second brain configuration involving two pairs of interacting sources. Boxplots summarize results across 50 different simulated data. For the ease of visualization outliers were omitted.

Chapter 4

Estimating latent directed brain connectivity via structural equation modelling

In this chapter, we introduce a novel method for estimating latent directed brain connectivity of cortical brain activity from EEG functional data. In the previous chapter, we proposed a method for estimating the cross-power spectrum of cortical brain activity, an undirected measure of functional connectivity defined in the frequency domain. In contrast, the present chapter focuses on the estimation of directed connectivity in the time domain, presented in [2]. To this end, we adopt a functional data analysis framework to model EEG recordings and the underlying cortical brain signals. In many neuroscientific studies, estimation of directed networks is of interest to investigate potential causal relationships among cortical brain sources [74]. Several works focused on estimating directed networks from EEG recordings, some examples are [75, 76, 77, 78]. Moreover, methods such as the Granger causality approach [79], dynamic causal modelling (DCM) [19], directed transfer function (DTF) [20] and partial directed coherence (PDC) [80], have been proposed to model directed networks. Among these techniques, structural equation modelling (SEM) allows to simultaneously model the relationships between the observed

EEG data and the latent cortical activity, while encoding directional interactions. It consists of a set of linear structural equations describing the variables of interest. Through these equations, SEM specifies the causal relationships arising from the underlying structure among the variables and models the variance and covariance structure. It is a common method to estimate effective connectivity in neuroscience [75, 81, 82], even though it is often applied on the reconstructed brain activity thus requiring a previous step for the estimation of cortical brain activity. For example, Astolfi et al. [21] propose a two-step approach that first reconstructs the cortical brain activity by performing a specific form of Tikhonov regularisation; then, they estimate directed brain connectivity from the reconstructed activity via SEM modelling.

In this study, we exploit the SEM framework to model the interactions occurring among the cortical brain sources directly, without the need to reconstruct their activity. This is done by considering in the SEM framework, both the activity of the cortical brain sources and the EEG recordings. To solve the SEM-based model, we adopt an expectation-maximisation (EM) algorithm, which iteratively estimates the latent directed brain connectivity, using the DAG with NO TEARS method proposed by Zheng et al. [24]. The latter is an efficient approach for estimating directed acyclic graphs (DAGs) which are commonly used to describe causal and directed interactions [83]. Up to our knowledge the DAG with NO TEARS method has been used to estimate directed connectivity [84, 85] from functional magnetic resonance imaginig (fMRI) and diffusion tensor imaging (DTI) data. Instead, we exploit it for directed brain connectivity from M/EEG data.

The chapter is organised as follows. In Section 4.1 we present the functional SEM that describes the interactions between the cortical brain activity and the EEG data. In Section 4.2 we show the inference procedure for the latent directed brain connectivity estimation, through the derivation of the likelihood function, the two steps of the EM algorithm and the explicit minimisation problem derived from the DAG with NO TEARS approach. In Section 4.3 we describe the numerical simulation developed to explore the

proposed approach and we present preliminary results. Finally, we discuss the proposed method and its limitations in Section 4.4.

4.1 Functional structural equation model

Let us consider the space of square integrable function $L^2(\mathcal{T}) = \{f : \mathcal{T} \to \mathbb{R} : \int_{\mathcal{T}} |f(t)|^2 dt < \infty\}$. Let $\boldsymbol{Y}(t) = [Y_1(t), \dots, Y_m(t)]^T$ and $\boldsymbol{X}(t) = [X_1(t), \dots, X_n(t)]^T$ be two functional variables, respectively of dimensions m and n, defined on the interval $\mathcal{T} \subseteq \mathbb{R}$ and such that $Y_i(t), X_j(t) \in L^2(\mathcal{T})$. In our application, the variables $\boldsymbol{Y}$ describe the signals recorded by the m EEG sensors and $\boldsymbol{X}$ denote the latent activity of the n cortical brain sources. Given $i \in \{1, \dots, m\}$ and $j \neq h, j, h \in \{1, \dots, n\}$, the functional structural equation model assumes the following relations

$$\begin{cases} Y_i(t) = \int_{\mathcal{T}} \sum_j \beta_{X_j Y_i}(t, \tau) X_j(\tau) d\tau + \epsilon_i^{\boldsymbol{Y}}(t) \\ X_h(t) = \int_{\mathcal{T}} \sum_j \beta_{X_j X_h}(t, \tau) X_j(\tau) d\tau + \epsilon_h^{\boldsymbol{X}}(t), \end{cases} \tag{4.1}$$

where the coefficients $\beta_{X_j Y_i}$ and $\beta_{X_j X_h}$ are functions in $L^2(\mathcal{T} \times \mathcal{T})$ that describe the structural effect of the variable X_j on the variable Y_i and on X_h respectively. The terms $\epsilon_i^{\boldsymbol{Y}}(t)$ and $\epsilon_h^{\boldsymbol{X}}(t)$ represent mean-zero stochastic processes accounting for observation noise. In particular, $\boldsymbol{\epsilon}^{\boldsymbol{Y}}(t) = \left[\epsilon_1^{\boldsymbol{Y}}(t), \dots, \epsilon_m^{\boldsymbol{Y}}(t)\right]^\top$ and $\boldsymbol{\epsilon}^{\boldsymbol{X}}(t) = \left[\epsilon_1^{\boldsymbol{X}}(t), \dots, \epsilon_n^{\boldsymbol{X}}(t)\right]^\top$ represent, respectively, the residual variability of the EEG recordings and the cortical brain sources. The two noise terms $\boldsymbol{\epsilon}^{\boldsymbol{Y}}$ and $\boldsymbol{\epsilon}^{\boldsymbol{X}}$ are assumed to be independent. Moreover, $\forall i, j \in 1, \dots, m$, $\epsilon_i^{\boldsymbol{Y}}$ and $\epsilon_j^{\boldsymbol{Y}}$ are mutually independent, with an analogous property holding for the components of $\boldsymbol{\epsilon}^{\boldsymbol{X}}$.

In our model we assume there are no self-effects, namely $\beta_{X_j X_j}(t, \tau) = 0 \quad \forall t, \forall \tau$. Furthermore, we assume there are no structural effects of the EEG-sensor activity on cortical brain activity, nor any influence of the activity recorded at one sensor on that of another, namely $\forall i \in \{1, \dots, m\}, \forall j \in \{1, \dots, n\}$, $\beta_{Y_i X_j}(t, \tau) = 0$ and $\forall i, j \in \{1, \dots, m\}, \beta_{Y_i Y_j}(t, \tau) = 0$, $\forall t, \forall \tau$.

The coefficients $\beta_{X_iY_j}$ in model (4.1) are assumed to be known. Indeed, the effect of the cortical brain activity $\boldsymbol{X}$ on the EEG-recordings $\boldsymbol{Y}$ is determined by the topography and the conductivity of the head, which can be modelled as discussed in Chapter 1. We discuss this further in Section 4.1.1.

Our goal is to estimate the interactions among the cortical brain sources $\boldsymbol{X}$, namely $\beta_{X_jX_h}$ for $j \neq h, j, h = 1, \ldots, n$. We provide now an example for a clearer understanding.

Example 1. *Let us consider $\boldsymbol{Y}(t) = [Y_1(t), Y_2(t)]$ the signal recorded by $m = 2$ EEG sensors and $\boldsymbol{X}(t) = [X_1(t), X_2(t), X_3(t)]$ the latent brain activity of $n = 3$ cortical sources. Figure 4.1 shows an illustrative example. The edges linking the $\boldsymbol{X}$ variables to the $\boldsymbol{Y}$ variables, such as $\beta_{X_3Y_2}$, reflect the effect of the latent cortical activity on the EEG sensors. Assuming that $\forall i \in \{1,2\}, \forall j \in \{1,2,3\}$, $\beta_{Y_iX_j}(t,\tau) = 0$ and $\forall i, j \in \{1,2\}, \beta_{Y_iY_j}(t,\tau) = 0$, $\forall t, \forall \tau$, as mentioned before, reflect the absence of edges going from $\boldsymbol{Y}$ variables to $\boldsymbol{X}$ variables and from $\boldsymbol{Y}$ to themselves. The coefficients $\beta_{X_3X_1}$ and $\beta_{X_2X_1}$ describe the effect of X_3 and of X_2 on X_1 respectively. These connections represent the latent connectivity between cortical brain sources that we aim to estimate.*

The complete functional structural equation model in equation (4.1) can be written in a compact form as

$$\begin{bmatrix} \boldsymbol{Y}(t) \\ \boldsymbol{X}(t) \end{bmatrix} = \int_{\mathcal{T}} \begin{bmatrix} \boldsymbol{0}_{m\times m} & \boldsymbol{\beta}_{XY}(t,\tau) \\ \boldsymbol{0}_{n\times m} & \boldsymbol{\beta}_{XX}(t,\tau) \end{bmatrix} \begin{bmatrix} \boldsymbol{Y}(\tau) \\ \boldsymbol{X}(\tau) \end{bmatrix} d\tau + \begin{bmatrix} \boldsymbol{\epsilon}^{\boldsymbol{Y}}(t) \\ \boldsymbol{\epsilon}^{\boldsymbol{X}}(t) \end{bmatrix}, \tag{4.2}$$

where $\boldsymbol{\beta}_{XY} \in L^2(\mathcal{T}\times\mathcal{T})^{m\times n}$ contains the structural effects of the $\boldsymbol{X}$ variables on the $\boldsymbol{Y}$ variable, namely $(\beta_{XY})_{ij}(t,\tau) = \beta_{X_jY_i}(t,\tau)$. Similarly, $\boldsymbol{\beta}_{XX} \in L^2(\mathcal{T}\times\mathcal{T})^{n\times n}$ contains the structural effects of the $\boldsymbol{X}$ variables on themselves and $(\beta_{XX})_{ij}(t,\tau) = \beta_{X_jX_i}(t,\tau)$.

We reduce the infinite-dimensional problem to a finite-dimensional representation by expanding the functional variables into a finite sum of orthonormal basis functions. Let us consider a general orthonormal (ON) basis set $\{\phi_k(t)\}_{k=1}^{\infty} \subset L^2(\mathcal{T})$. We assume that the system of variables $\boldsymbol{Y}$ and $\boldsymbol{X}$ can be

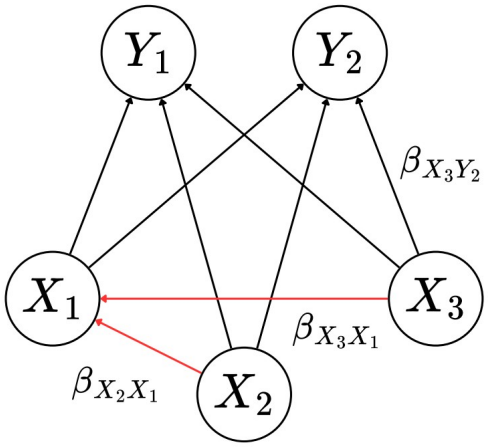

Figure 4.1: Illustration of the relationships between the latent brain activity $\mathbf{X}$ of three cortical sources and the EEG recordings $\mathbf{Y}$ of two observed sensors. Directed edges among the sources represent brain connectivity of interest, while edges from sources to sensors describe the effect of brain activity to EEG measurements.

expressed in terms of the basis functions and their corresponding scores. In particular, we have that

$$Y_i(t) = \sum_{k=1}^{\infty} \zeta_{i,k}\phi_k(t) \simeq \sum_{k=1}^{K} \zeta_{i,k}\phi_k(t) \tag{4.3}$$

and

$$X_j(t) = \sum_{k=1}^{\infty} \xi_{j,k}\phi_k(t) \simeq \sum_{k=1}^{K} \xi_{j,k}\phi_k(t), \tag{4.4}$$

where we approximate the functions with the projection onto the first $K > 0$ elements of the basis. The projection scores can be computed as $\zeta_{i,k} = \langle Y_i, \phi_k \rangle$ and $\xi_{j,k} = \langle X_j, \phi_k \rangle$.

Similarly, we can write $\epsilon_i^{\mathbf{Y}}(t) \simeq \sum_{k=1}^{K} \eta_{i,k}^{Y}\phi_k(t)$ and $\epsilon_j^{\mathbf{X}}(t) \simeq \sum_{k=1}^{K} \eta_{j,k}^{\mathbf{X}}\phi_k(t)$. Let us denote the vector score as $\boldsymbol{\zeta} = [\zeta_{1,1}, \ldots, \zeta_{m,K}]^T$. The vectors $\boldsymbol{\xi}$, $\boldsymbol{\eta}^{\mathbf{Y}}$, $\boldsymbol{\eta}^{\mathbf{X}}$ can be analogously defined.

By exploiting Equations (4.3) and (4.4), the structural equation model (4.2) becomes

$$\begin{bmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\xi} \end{bmatrix} = \boldsymbol{B} \otimes \boldsymbol{I}_K \begin{bmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\xi} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\eta}^{\boldsymbol{Y}} \\ \boldsymbol{\eta}^{\boldsymbol{X}} \end{bmatrix}, \tag{4.5}$$

where

$$\boldsymbol{B} = \begin{bmatrix} \mathbf{0}_{m\times m} & \boldsymbol{B}_{XY} \\ \mathbf{0}_{n\times m} & \boldsymbol{B}_{XX} \end{bmatrix}.$$

The matrices $\boldsymbol{B}_{XY} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B}_{XX} \in \mathbb{R}^{n\times n}$ are intrinsically related to the functional coefficients $\boldsymbol{\beta}_{XY}$ and $\boldsymbol{\beta}_{XX}$ of model (4.2) [86]. For example, it can be shown that

$$\beta_{X_jY_i}(t,\tau) = \sum_{k=1}^{K} b_{i,j}^{XY} \phi_k(t)\phi_k(t).$$

We define the overall matrix $\tilde{\boldsymbol{B}} = \boldsymbol{B} \otimes \boldsymbol{I}_K \in \mathbb{R}^{(m+n)K\times(m+n)K}$. The Kronecker product between $\boldsymbol{B}$ and the identity matrix encodes homogeneous effects across basis functions, meaning that the same structural relationships apply at each basis component.

The noise vector of basis coefficients $\boldsymbol{\eta} = [\boldsymbol{\eta}^{\boldsymbol{Y}}, \boldsymbol{\eta}^{\boldsymbol{X}}]^T \in \mathbb{R}^{(n+m)K}$ is modelled as a multivariate normal distribution, namely $\boldsymbol{\eta} \sim \mathcal{N}(0, \boldsymbol{D} \otimes \boldsymbol{I}_K)$ where $\otimes$ denotes the Kronecker product, which replicates the covariance structure across the K basis coefficients. The block-diagonal covariance matrix

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_Y & 0 \\ 0 & \boldsymbol{D}_X \end{bmatrix}$$

has blocks $\boldsymbol{D}_Y \in \mathbb{R}^{m\times m}$ and $\boldsymbol{D}_X \in \mathbb{R}^{n\times n}$, representing the covariance of the noise components for $\boldsymbol{Y}$ and $\boldsymbol{X}$, respectively.

Finally, Equation (4.5) can be rewritten as

$$\begin{bmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\xi} \end{bmatrix} = (\boldsymbol{I} - \tilde{\boldsymbol{B}})^{-1}\boldsymbol{\eta}, \tag{4.6}$$

assuming the matrix $\boldsymbol{I} - \tilde{\boldsymbol{B}}$ is invertible. In practice, the cortical brain activity $\boldsymbol{\xi}$ is not directly observed, while the scores $\boldsymbol{\zeta}$ are estimated from the EEG recordings. This motivates the probabilistic formulation of Section 4.2.1, where $\boldsymbol{\xi}$ is

treated as a latent variable and $\boldsymbol{\zeta}$ as the observed data. Before moving to the inference framework, we discuss in more details the structural relationship between the cortical activity $\boldsymbol{X}$ and the EEG recorded data $\boldsymbol{Y}$.

4.1.1 Derivation of the structural relation between cortical brain activity and EEG signals

In this section we show how the structural effects of the variables $\boldsymbol{X}$ on $\boldsymbol{Y}$ are linked to the topology and conductivity of the brain in consideration. In EEG forward model, the cortical brain activity is related to the EEG data through a linear model, namely

$$\boldsymbol{Y}(t) = \boldsymbol{G}\boldsymbol{X}(t) + \boldsymbol{E}(t), \tag{4.7}$$

where $\boldsymbol{E}(t) \in \mathbb{R}^m$ represents additive noise independent from $\boldsymbol{Y}$ and $\boldsymbol{X}$. The matrix $\boldsymbol{G} \in \mathbb{R}^{m \times n}$ describes the topology and conductivity of the brain and it is often derived from magnetic resonance imaging (MRI). In the M/EEG context, this matrix is called the leadfiel matrix and maps the activity of cortical brain sources to the recorded data. More details regarding its construction can be found in Chapter 1.

Theorem 13. *Consider $\boldsymbol{Y}$ and $\boldsymbol{X}$ and their structual relation* (4.2)*. Assume instantaneous linear mixing between these two variables, that is for t it holds*

$$\boldsymbol{Y}(t) = \boldsymbol{G}\boldsymbol{X}(t) + \boldsymbol{E}(t).$$

Moroever, given a set of orthonormal basis function $\{\phi_k(t)\}_{k=1}^{\infty} \subset L^2(\mathcal{T})$ and the decomposition of $\boldsymbol{Y}$ and $\boldsymbol{X}$ described in (4.3) *and* (4.4)*, respectively, assume that the corresponding scores follows*

$$\begin{bmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\xi} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{m \times m} & \boldsymbol{B}_{XY} \\ \mathbf{0}_{n \times m} & \boldsymbol{B}_{XX} \end{bmatrix} \begin{bmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\xi} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\eta}^{Y} \\ \boldsymbol{\eta}^{X} \end{bmatrix}.$$

Then, it holds that $\boldsymbol{B}_{XY} = \boldsymbol{G}$.

Proof. Consider the basis function decomposition of $Y_i(t) \simeq \sum_{k=1}^{K} \zeta_{i,k}\phi_k(t)$. We can write:

$$\begin{aligned} Y_i(t) &= \sum_{k=1}^{K} \zeta_{i,k}\phi_k(t) \\ &= \sum_{k=1}^{K}\sum_{j=1}^{n} [b_{i,j}\xi_{j,k} + \eta_{i,k}^{\boldsymbol{Y}}]\phi_k(t) \\ &= \sum_{j=1}^{n} b_{i,j}X_j(t) + \epsilon_i^{\boldsymbol{Y}}(t), \end{aligned} \tag{4.8}$$

where we use $b_{i,j} \in \boldsymbol{B}_{XY}$ to denote the effect of the j-th $\boldsymbol{X}$-variable on Y_i. By recalling (4.7), namely $Y_i(t) = G_{i,.}\boldsymbol{X}(t) + E_i(t) \quad \forall i$, we can conclude that $\boldsymbol{B}_{XY} = \boldsymbol{G}$.

□

4.2 Inference method

4.2.1 Derivation of the likelihood function

In this section we describe the distribution of the observed data $\boldsymbol{\zeta}$ and of the latent variable $\boldsymbol{\xi}$ defined in the previous section. We then derive the likelihood function. As discussed in Chapter 2, the likelihood function measures how well a statistical model explains the observed data under values of the model parameters. In our specific case, it thus describes how likely it is that we observe the joint variable $[\boldsymbol{\zeta}, \boldsymbol{\xi}]$ under the estimated parameters $\boldsymbol{B}_{XX}$ describing the cortical brain connectivity network. Moreover, the likelihood function naturally provides a framework for parameter estimation and inference.

In our specific application, the scores $\boldsymbol{\xi}$ are not directly observed and their estimation is conditioned to the observed scores $\boldsymbol{\zeta}$. For this reason, our setting lies within the incomplete-data framework and can be addressed by implementing the expectation-maximisation (EM) algorithm.

We now present the probabilistic model for the complete data, namely

$[\boldsymbol{\zeta}, \boldsymbol{\xi}]$, and then show the EM derivation in the Section 4.2.2.

Proposition 5. *Let us assume that the joint distribution of* $[\boldsymbol{\zeta}, \boldsymbol{\xi}]$ *follows a normal distribution of zero mean and covariance matrix* $\boldsymbol{\Lambda}$*, that is*

$$[\boldsymbol{\zeta}, \boldsymbol{\xi}] \sim \mathcal{N}(0, \boldsymbol{\Lambda}).$$

The covariance matrix is computed as

$$\boldsymbol{\Lambda} = (\boldsymbol{I} - \tilde{\boldsymbol{B}})^{-1} \tilde{\boldsymbol{D}} (\boldsymbol{I} - \tilde{\boldsymbol{B}})^{-T}, \tag{4.9}$$

where $\tilde{\boldsymbol{D}} = \boldsymbol{D} \otimes \boldsymbol{I}_K$ *and* $\tilde{\boldsymbol{B}} = \boldsymbol{B} \otimes \boldsymbol{I}_K$*. The covariance matrix* $\boldsymbol{\Lambda}$ *can be partitioned in four blocks, each one depending on* $\boldsymbol{B}_{XX}$ *so that:*

$$\boldsymbol{\Lambda}(\boldsymbol{B}_{XX}) = \begin{bmatrix} \boldsymbol{\Lambda}_{\zeta}(\boldsymbol{B}_{XX}) & \boldsymbol{\Lambda}_{\zeta\xi}(\boldsymbol{B}_{XX}) \\ \boldsymbol{\Lambda}_{\xi\zeta}(\boldsymbol{B}_{XX}) & \boldsymbol{\Lambda}_{\xi}(\boldsymbol{B}_{XX}) \end{bmatrix},$$

where

$$\begin{aligned} \boldsymbol{\Lambda}_{\xi}(\boldsymbol{B}_{XX}) &= (\boldsymbol{I} - \boldsymbol{B}_{XX} \otimes \boldsymbol{I}_K)^{-1} \tilde{\boldsymbol{D}}_X (\boldsymbol{I} - \boldsymbol{B}_{XX} \otimes \boldsymbol{I}_K)^{-T}, \\ \boldsymbol{\Lambda}_{\zeta}(\boldsymbol{B}_{XX}) &= \boldsymbol{B}_{XY} \boldsymbol{\Lambda}_{\xi}(\boldsymbol{B}_{XX}) \boldsymbol{B}_{XY}^T + \tilde{\boldsymbol{D}}_Y, \\ \boldsymbol{\Lambda}_{\zeta\xi}(\boldsymbol{B}_{XX}) &= \boldsymbol{B}_{XY} \boldsymbol{\Lambda}_{\xi}(\boldsymbol{B}_{XX}) = \boldsymbol{\Lambda}_{\xi\zeta}^T(\boldsymbol{B}_{XX}). \end{aligned} \tag{4.10}$$

Let us note that in (4.10), we distinguish between the identity matrix involved in the Kronecker product, namely $\boldsymbol{I}_K$, and the identity matrix in the difference $\boldsymbol{I}$, that is $\boldsymbol{I}_{nK}$. Moreover, in Equation (4.10) we stressed the dependence of each block on the matrix $\boldsymbol{B}_{XX}$. For the ease of notation in most cases we will omit this dependence. We now provide a proof of the above proposition.

Proof. The covariance structure is induced by the structural equation model (4.6), that is

$$\begin{bmatrix} \zeta \\ \xi \end{bmatrix} = (\boldsymbol{I} - \tilde{\boldsymbol{B}})^{-1}\boldsymbol{\eta}.$$

By taking into account the distribution of the noise term $\boldsymbol{\eta} \sim \mathcal{N}(0, \tilde{\boldsymbol{D}})$, we obtain:

$$\begin{aligned}\boldsymbol{\Lambda} &= \mathbb{V}\Big(\begin{bmatrix} \zeta \\ \xi \end{bmatrix} \Big) = \mathbb{E}\Big(\begin{bmatrix} \zeta \\ \xi \end{bmatrix} \begin{bmatrix} \zeta & \xi \end{bmatrix} \Big) = \mathbb{E}\Big((\boldsymbol{I} - \tilde{\boldsymbol{B}})^{-1}\boldsymbol{\eta}\boldsymbol{\eta}^T(\boldsymbol{I} - \tilde{\boldsymbol{B}})^{-T} \Big) \\ &= (\boldsymbol{I} - \tilde{\boldsymbol{B}})^{-1}\mathbb{E}\Big(\boldsymbol{\eta}\boldsymbol{\eta}^T\Big)(\boldsymbol{I} - \tilde{\boldsymbol{B}})^{-T} = (\boldsymbol{I} - \tilde{\boldsymbol{B}})^{-1}\tilde{\boldsymbol{D}}(\boldsymbol{I} - \tilde{\boldsymbol{B}})^{-T},\end{aligned}$$

where $\tilde{\boldsymbol{B}} = \boldsymbol{B}_{XX} \otimes \boldsymbol{I}_K$. Exploiting the formulation in model (4.5), we can derive the covariance matrix blocks in (4.10), namely $\boldsymbol{\Lambda}_\xi, \boldsymbol{\Lambda}_\zeta, \boldsymbol{\Lambda}_{\zeta\xi}$. □

Remark 7. *From the joint distribution of $[\boldsymbol{\zeta}, \boldsymbol{\xi}]$, it follows that the marginal distributions of $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ are normal distributions with zero mean and covariance given by the corresponding block of $\boldsymbol{\Lambda}$, namely $\boldsymbol{\zeta} \sim \mathcal{N}(0, \boldsymbol{\Lambda}_\zeta)$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{\Lambda}_\xi)$, respectively.*

Remark 8. *From the structural equation model in* (4.6)*, the conditional distribution of $\boldsymbol{\zeta}$ given $\boldsymbol{\xi}$ is given by*

$$\boldsymbol{\zeta}|\boldsymbol{\xi} \sim \mathcal{N}(\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}, \tilde{\boldsymbol{D}}_Y),$$

where we denote with $\tilde{\boldsymbol{B}}_{XY} = \boldsymbol{B}_{XY} \otimes \boldsymbol{I}_K$.

Remark 9. *Finally, conditioning the joint distribution of $[\boldsymbol{\zeta}, \boldsymbol{\xi}]$ on $\boldsymbol{\zeta}$ yields the conditional distribution*

$$\boldsymbol{\xi}|\boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_\zeta^{-1}\zeta, \boldsymbol{\Lambda}_\xi - \boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_\zeta^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T).$$

Following the distribution defined in the previous remarks, let us denote with $p(\boldsymbol{\zeta}, \boldsymbol{\xi})$ the probability density function of the complete data. As discussed in Chapter 2, we exploit the factorization of the complete-data probability density function, namely $p(\boldsymbol{\zeta}, \boldsymbol{\xi}) = p(\boldsymbol{\xi})p(\boldsymbol{\zeta}|\boldsymbol{\xi})$. This decomposition is

particularly advantageous for the EM framework, as it separates the latent variable $p(\xi)$ from the observed one $p(\zeta|\xi)$, allowing us to compute the expected log-likelihood conditional on the observed data $\boldsymbol{\zeta}$. When considering a population of N independent and identically distributed individuals, the complete-data log-likelihood expression for the joint distributions is given by

$$\begin{aligned}\ell_{\xi\zeta}(\boldsymbol{B}_{XX}) &= \sum_{i=1}^{N} \log\left[p_{0,\boldsymbol{\Lambda}_\xi}(\boldsymbol{\xi}^{(i)}) p_{\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)},\tilde{\boldsymbol{D}}_Y}(\boldsymbol{\zeta}^{(i)})\right] \\ &= \sum_{i=1}^{N} \log p_{\mathbf{0},\boldsymbol{\Lambda}_\xi}(\boldsymbol{\xi}^{(i)}) + \sum_{i=1}^{N} \log p_{\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)},\tilde{\boldsymbol{D}}_Y}(\boldsymbol{\zeta}^{(i)}),\end{aligned} \tag{4.11}$$

where $\boldsymbol{\xi}^{(i)}$ and $\boldsymbol{\zeta}^{(i)}$ are respectively the scores related to the brain activity and the observed EEG recording for the i-th subject.

Proposition 6. *Let us consider the setting presented in Proposition 5 and in Remarks 7, 8, and 9. The expected log-likelihood of the complete data, conditional on the observed data $\boldsymbol{\zeta}$, is given by*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\xi}\sim\boldsymbol{B}_{XX}}[\ell_{\xi,\zeta}(\boldsymbol{B}_{XX})|\boldsymbol{\zeta}] = \\ &-\frac{N}{2}(\log|\boldsymbol{\Lambda}_\xi| + nK\log(2\pi)) \\ &-\frac{1}{2}\sum_i \mathrm{Tr}\left[(\boldsymbol{\Lambda}_\xi)^{-1}(\boldsymbol{\Lambda}_\xi - \boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_\zeta^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T + \boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_\zeta^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_\zeta^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T)\right] \\ &-\frac{N}{2}(\log|\tilde{\boldsymbol{D}}_Y| + mK\log(2\pi)) \\ &-\frac{1}{2}\sum_i \mathrm{Tr}[(\tilde{\boldsymbol{D}}_Y)^{-1}(\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T} - 2\boldsymbol{\zeta}^{(i)}(\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_\zeta^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T)\tilde{\boldsymbol{B}}_{XY}^T \\ &+\tilde{\boldsymbol{B}}_{XY}(\boldsymbol{\Lambda}_\xi - \boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_\zeta^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T + \boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_\zeta^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_\zeta^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T)\tilde{\boldsymbol{B}}_{XY}^T)].\end{aligned} \tag{4.12}$$

Proof. From equation (4.11), the expected complete-data log-likelihood conditional on the observed data $\boldsymbol{\zeta}$ follows as

$$\mathbb{E}_{\boldsymbol{\xi}}[\ell_{\xi,\zeta}(\boldsymbol{B}_{XX})|\boldsymbol{\zeta}] = \sum_i \mathbb{E}_{\boldsymbol{\xi}^{(i)}}[\log p_{\mathbf{0},\boldsymbol{\Lambda}_\xi}(\boldsymbol{\xi}^{(i)})] + \sum_i \mathbb{E}_{\boldsymbol{\xi}^{(i)}}[\log p_{\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi},\tilde{\boldsymbol{D}}_Y}(\boldsymbol{\zeta}^{(i)})].$$

Let us now develop first term as

$$\begin{aligned}
&\mathbb{E}_{\xi^{(i)}}[\log p_{\mathbf{0},\mathbf{\Lambda}_{\tilde{\xi}}}(\boldsymbol{\xi}^{(i)})] = \mathbb{E}_{\xi^{(i)}}\left[-\frac{1}{2}(\log|\mathbf{\Lambda}_{\tilde{\xi}}| + kn\log(2\pi) + (\boldsymbol{\xi}^{(i)})^T\mathbf{\Lambda}_{\tilde{\xi}}^{-1}\boldsymbol{\xi}^{(i)})|\boldsymbol{\zeta}^{(i)}\right] \\
&= -\frac{1}{2}\left[\log|\mathbf{\Lambda}_{\tilde{\xi}}| + kn\log(2\pi) + \mathrm{Tr}(\mathbf{\Lambda}_{\tilde{\xi}}^{-1}\mathbb{E}_{\xi^{(i)}}[\boldsymbol{\xi}^{(i)}(\boldsymbol{\xi}^{(i)})^T|\boldsymbol{\zeta}^{(i)}])\right] \\
&= -\frac{1}{2}\Big[\log|\mathbf{\Lambda}_{\tilde{\xi}}| + kn\log(2\pi) \\
&\qquad + \mathrm{Tr}(\mathbf{\Lambda}_{\tilde{\xi}}^{-1}(\mathbf{\Lambda}_{\tilde{\xi}} - \mathbf{\Lambda}_{\tilde{\xi}\zeta}\mathbf{\Lambda}_{\zeta}^{-1}\mathbf{\Lambda}_{\tilde{\xi}\zeta}^T + \mathbf{\Lambda}_{\tilde{\xi}\zeta}\mathbf{\Lambda}_{\zeta}^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\mathbf{\Lambda}_{\zeta}^{-1}\mathbf{\Lambda}_{\tilde{\xi}\zeta}^T))\Big].
\end{aligned} \tag{4.13}$$

Similarly, the second term can be computed as

$$\begin{aligned}
&\mathbb{E}_{\xi^{(i)}}[\log p_{\tilde{\boldsymbol{B}}_{XY}\xi,\tilde{\boldsymbol{D}}_Y}(\boldsymbol{\zeta}^{(i)})] \\
&= \mathbb{E}_{\xi^{(i)}}\Big[-\frac{1}{2}\Big(\log|\tilde{\boldsymbol{D}_Y}| + km\log(2\pi) \\
&\qquad + (\boldsymbol{\zeta}^{(i)} - \tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)})^T\tilde{\boldsymbol{D}_Y}^{-1}(\boldsymbol{\zeta}^{(i)} - \tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)})\Big)\,\Big|\,\boldsymbol{\zeta}^{(i)}\Big] \\
&= \mathbb{E}_{\xi^{(i)}}\Big[-\frac{1}{2}\Big(\log|\tilde{\boldsymbol{D}_Y}| + km\log(2\pi) \\
&\qquad + \mathrm{Tr}(\tilde{\boldsymbol{D}_Y}^{-1}(\boldsymbol{\zeta}^{(i)} - \tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)})(\boldsymbol{\zeta}^{(i)} - \tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)})^T)\Big)\,\Big|\,\boldsymbol{\zeta}^{(i)}\Big] \\
&= -\frac{1}{2}\Big[\log|\tilde{\boldsymbol{D}_Y}| + km\log(2\pi) \\
&\qquad + \mathrm{Tr}(\tilde{\boldsymbol{D}_Y}^{-1}\mathbb{E}_{\xi^{(i)}}\Big[(\boldsymbol{\zeta}^{(i)} - \tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)})(\boldsymbol{\zeta}^{(i)} - \tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)})^T\,\Big|\,\boldsymbol{\zeta}^{(i)}\Big])\Big].
\end{aligned} \tag{4.14}$$

If focusing on the expectation term, we obtain

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}^{(i)}}\left[(\boldsymbol{\zeta}^{(i)}-\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)})(\boldsymbol{\zeta}^{(i)}-\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)})^T|\boldsymbol{\zeta}^{(i)}\right]\\
&=\mathbb{E}_{\boldsymbol{\xi}^{(i)}}\left[\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}-2\boldsymbol{\zeta}^{(i)}\boldsymbol{\xi}^{(i)T}\tilde{\boldsymbol{B}}_{XY}^T+\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi}^{(i)}\boldsymbol{\xi}^{(i)T}\tilde{\boldsymbol{B}}_{XY}^T|\boldsymbol{\zeta}^{(i)}\right]\\
&=\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}-2\boldsymbol{\zeta}^{(i)}\mathbb{E}_{\boldsymbol{\xi}^{(i)}}[\boldsymbol{\xi}^{(i)T}|\boldsymbol{\zeta}^{(i)}]\tilde{\boldsymbol{B}}_{XY}^T+\tilde{\boldsymbol{B}}_{XY}\mathbb{E}_{\boldsymbol{\xi}^{(i)}}[\boldsymbol{\xi}^{(i)}\boldsymbol{\xi}^{(i)T}|\boldsymbol{\zeta}^{(i)}]\tilde{\boldsymbol{B}}_{XY}^T\\
&=\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}-2\boldsymbol{\zeta}^{(i)}(\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\zeta}^{(i)})^T\tilde{\boldsymbol{B}}_{XY}^T\\
&\qquad+\tilde{\boldsymbol{B}}_{XY}(\boldsymbol{\Lambda}_{\xi}-\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T+\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T)\tilde{\boldsymbol{B}}_{XY}^T\\
&=\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}-2\boldsymbol{\zeta}^{(i)}(\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T)\tilde{\boldsymbol{B}}_{XY}^T\\
&\qquad+\tilde{\boldsymbol{B}}_{XY}(\boldsymbol{\Lambda}_{\xi}-\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T+\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T)\tilde{\boldsymbol{B}}_{XY}^T.
\end{aligned}
$$

Expression (4.14) can be rewritten as

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}^{(i)}}[\log p_{\tilde{\boldsymbol{B}}_{XY}\xi,\tilde{\boldsymbol{D}}_Y}(\boldsymbol{\zeta}^{(i)})]=\\
&-\frac{1}{2}[\log|\tilde{\boldsymbol{D}}_Y|+km\log(2\pi)+\\
&+\mathrm{Tr}(\tilde{\boldsymbol{D}}_Y{}^{-1}(\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}-2\boldsymbol{\zeta}^{(i)}(\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T)\tilde{\boldsymbol{B}}_{XY}^T\\
&+\tilde{\boldsymbol{B}}_{XY}(\boldsymbol{\Lambda}_{\xi}-\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T+\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^T)\tilde{\boldsymbol{B}}_{XY}^T)).
\end{aligned}
\tag{4.15}
$$

By summing together expressions (4.13) and (4.15), we obtain expression (4.12) for the expected log-likelihood. □

We stress that each term $\boldsymbol{\Lambda}_{\xi},\boldsymbol{\Lambda}_{\zeta},\boldsymbol{\Lambda}_{\xi\zeta}$ in the expected log-likelihood of the complete data depends on the parameter of interest $\boldsymbol{B}_{XX}$ through the covariance structure defined in Proposition 5.

4.2.2 The EM framework

Given the expected log-likelihood described in the previous section, we can now explicit the two steps of the EM algorithm. We recall that the E-step and the M-step iteratively alternate until convergence or the stopping criterion is satisfied. Let us consider the p-th iteration and the corresponding set of

parameters $\boldsymbol{B}_{XX}^{(p)}$.

The E-step consists in computing the expectation under the current parameter guess $\boldsymbol{B}_{XX}^{(p)}$ of the complete-data likelihood, given the observed data $\boldsymbol{\zeta}$, that is

$$\begin{aligned}
Q(\boldsymbol{B}_{XX}|\boldsymbol{B}_{XX}^{(p)}) =& \mathbb{E}_{\boldsymbol{\xi}\sim\boldsymbol{B}_{XX}^{(p)}}[\ell_{\tilde{\xi},\tilde{\zeta}}(\boldsymbol{B}_{XX})|\boldsymbol{\zeta}] \\
=& -\frac{N}{2}(\log|\boldsymbol{\Lambda}_{\tilde{\xi}}^{(p)}| + nK\log(2\pi)) \\
& -\frac{1}{2}\sum_i \mathrm{Tr}\left[(\boldsymbol{\Lambda}_{\tilde{\xi}}^{(p)})^{-1}(\boldsymbol{\Lambda}_{\tilde{\xi}} - \boldsymbol{\Lambda}_{\tilde{\xi}\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\tilde{\xi}\zeta}^T + \boldsymbol{\Lambda}_{\tilde{\xi}\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\tilde{\xi}\zeta}^T)\right] \\
& -\frac{N}{2}(\log|\tilde{\boldsymbol{D}}_Y| + mK\log(2\pi)) \\
& -\frac{1}{2}\sum_i \mathrm{Tr}[(\tilde{\boldsymbol{D}}_Y)^{-1}(\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T} - 2\boldsymbol{\zeta}^{(i)}(\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\tilde{\xi}\zeta}^T)\tilde{\boldsymbol{B}}_{XY}^T \\
& + \tilde{\boldsymbol{B}}_{XY}(\boldsymbol{\Lambda}_{\tilde{\xi}} - \boldsymbol{\Lambda}_{\tilde{\xi}\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\tilde{\xi}\zeta}^T + \boldsymbol{\Lambda}_{\tilde{\xi}\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\tilde{\xi}\zeta}^T)\tilde{\boldsymbol{B}}_{XY}^T)].
\end{aligned} \tag{4.16}$$

We denote with $\boldsymbol{\Lambda}_{\tilde{\xi}}^{(p)}$ the covariance matrix of $\tilde{\boldsymbol{\xi}}$ under the current parameter guess $\boldsymbol{B}_{XX}^{(p)}$, namely $\boldsymbol{\Lambda}_{\tilde{\xi}}^{(p)} = \boldsymbol{\Lambda}_{\tilde{\xi}}(\boldsymbol{B}_{XX}^{(p)}) = (I - \boldsymbol{B}_{XX}^{(p)} \otimes \boldsymbol{I}_K)^{-1}\tilde{\boldsymbol{D}}_X(I - \boldsymbol{B}_{XX}^{(p)} \otimes \boldsymbol{I}_K)^{-T}$.

The M-step aims at choosing the new parameter $\boldsymbol{B}_{XX}^{(p+1)}$ by maximising $Q(\boldsymbol{B}_{XX}|\boldsymbol{B}_{XX}^{(p)})$. For the maximisation procedure, we compute the derivative of $Q(\boldsymbol{B}_{XX}|\boldsymbol{B}_{XX}^{(p)})$ with respect to the matrix $\boldsymbol{B}_{XX}$. Since the computation of the derivative of $Q(\boldsymbol{B}_{XX}|\boldsymbol{B}_{XX}^{(p)})$ is not trivial, we first present all the steps of the derivation for one of the addends in (4.16), and then we provide the final expression for the complete derivatative of $Q(\boldsymbol{B}_{XX}|\boldsymbol{B}_{XX}^{(p)})$. We refer to [87] for standard rules of matrix-form derivation.

Let us consider the term $\mathrm{Tr}((\boldsymbol{\Lambda}_{\tilde{\xi}}^{(p)})^{-1}\boldsymbol{\Lambda}_{\tilde{\xi}})$. We proceed by computing the derivative of $\mathrm{Tr}((\boldsymbol{\Lambda}_{\tilde{\xi}}^{(p)})^{-1}\boldsymbol{\Lambda}_{\tilde{\xi}})$ with respect to $\tilde{\boldsymbol{B}}_{XX}$, and then reconstruct the derivative with respect to $\boldsymbol{B}_{XX}$ by exploiting the block structure of the computed matrix derivative.

Let us note that the derivative of a scalar function with respect to a matrix

variable is a matrix of the same dimensions of the variable of differentiation. Therefore, since $\tilde{\boldsymbol{B}}_{XX} \in \mathbb{R}^{nK \times nK}$, also the derivative $\frac{\partial \operatorname{Tr}((\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\tilde{\xi}})}{\partial \tilde{\boldsymbol{B}}_{XX}} \in \mathbb{R}^{nK \times nK}$. Moreover, we have that

$$\left[\frac{\partial \operatorname{Tr}((\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\tilde{\xi}})}{\partial \tilde{\boldsymbol{B}}_{XX}}\right]_{i,j} = \frac{\partial \operatorname{Tr}((\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\tilde{\xi}})}{\partial (\tilde{\boldsymbol{B}}_{XX})_{i,j}},$$

for some $i, j \in \{1, \dots, nK\}$. We now proceed to compute the derivative with respect to the (i,j) generic element $(\tilde{\boldsymbol{B}}_{XX})_{i,j}$ of the matrix $\tilde{\boldsymbol{B}}_{XX}$, and then we reconstruct the full matrix derivative by collecting all the computed elements. Let us focus on the derivative of $\boldsymbol{\Lambda}_{\tilde{\xi}} = (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T}$, which is computed by applying the standard rules for the derivative of inverse and transpose functions [87], leading to the following expression:

$$\begin{aligned}
&\frac{\partial \boldsymbol{\Lambda}_{\tilde{\xi}}}{\partial (\tilde{\boldsymbol{B}}_{XX})_{i,j}} = \frac{\partial}{\partial (\tilde{\boldsymbol{B}}_{XX})_{i,j}} \left((\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} \right) \\
=&\frac{\partial (I - \tilde{\boldsymbol{B}}_{XX})^{-1}}{\partial (\tilde{\boldsymbol{B}}_{XX})_{i,j}}\tilde{\boldsymbol{D}}_X(\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} + (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X\frac{\partial (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T}}{\partial (\tilde{\boldsymbol{B}}_{XX})_{i,j}} \\
=&(\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1}_{\cdot,i}(\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1}_{j,\cdot}\tilde{\boldsymbol{D}}_X(\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} \\
&\quad + \quad (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T}_{\cdot,i}(\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T}_{j,\cdot}.
\end{aligned} \tag{4.17}$$

We denote by $(\cdot, i)$ the i-th column of a given matrix and by $(j, \cdot)$ its j-th row. By applying the chain rule and using the expression (4.17), the derivative

of $\mathrm{Tr}((\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi})$ with respect to $(\tilde{\boldsymbol{B}}_{XX})_{i,j}$ is given by

$$\begin{aligned}
&\frac{\partial}{\partial(\tilde{\boldsymbol{B}}_{XX})_{i,j}}\left[\mathrm{Tr}\Big((\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi}\Big)\right] = \mathrm{Tr}\left[\frac{\partial}{\partial(\tilde{\boldsymbol{B}}_{XX})_{i,j}}\Big((\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi}\Big)\right] \\
&= \mathrm{Tr}\Big[(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}_{\cdot,i}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}_{j,\cdot}\tilde{\boldsymbol{D}}_X(I-\tilde{\boldsymbol{B}}_{XX})^{-T} \\
&\qquad + (\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}_{\cdot,i}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}_{j,\cdot}\Big] \\
&= \Big[(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}_{j,\cdot}\tilde{\boldsymbol{D}}_X(I-\tilde{\boldsymbol{B}}_{XX})^{-T}(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}_{\cdot,i} \\
&\qquad + (\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}_{j,\cdot}(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}_{\cdot,i}\Big].
\end{aligned} \tag{4.18}$$

We exploit the linearity of the trace operator in the first equality and the cyclic property of the trace in the last equation. We remind the reader that $(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}$ does not depend on the variable $\boldsymbol{B}_{XX}$ but on the previous parameter guess $\boldsymbol{B}_{XX}^{(p)}$, and thus it can be treated as a constant in the differentiation procedure.

Overall, we obtain

$$\begin{aligned}
\frac{\partial\,\mathrm{Tr}((\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi})}{\partial\tilde{\boldsymbol{B}}_{XX}} &= (\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\
&\quad + (\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}
\end{aligned} \tag{4.19}$$

The resulting matrix is a block matrix with $n \times n$ blocks of dimension $K \times K$. Given $u, v \in \{1, \ldots, n\}$, because $\tilde{\boldsymbol{B}}_{XX}$ depends the Kronecker product, computing the trace of (u, v)-block returns the value of the derivative with respect to the (u, v)-th element of $\boldsymbol{B}_{XX}$, that is

$$\frac{\partial\,\mathrm{Tr}((\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi})}{\partial\boldsymbol{B}_{XX,uv}} = \mathrm{Tr}\left(\frac{\partial\,\mathrm{Tr}((\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi})}{\partial\tilde{\boldsymbol{B}}_{XX}}\Big|_{u:u+K,v:v+K}\right), \tag{4.20}$$

where $(\cdot)_{u:u+K,v:v+K}$ denotes the (u, v)-block of dimensions $K \times K$.

The derivative of the overall function $Q(\boldsymbol{B}_{XX}|\boldsymbol{B}_{XX}^{(p)})$ with respect to $\tilde{\boldsymbol{B}}_{XX}$ is

given by summing together the derivatives of each additive term, namely

$$\begin{aligned}\frac{\partial Q(\boldsymbol{B}_{XX}|\boldsymbol{B}_{XX}^{(p)})}{\partial \tilde{\boldsymbol{B}}_{XX}} = &-\frac{N}{2}\frac{\partial \operatorname{Tr}((\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi})}{\partial \tilde{\boldsymbol{B}}_{XX}} \\ &-\frac{N}{2}\frac{\partial \operatorname{Tr}(-(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^{T})}{\partial \tilde{\boldsymbol{B}}_{XX}} \\ &-\frac{1}{2}\frac{\partial \operatorname{Tr}(-(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^{T})}{\partial \tilde{\boldsymbol{B}}_{XX}} \\ &-\frac{1}{2}\frac{\partial \operatorname{Tr}((\boldsymbol{D}_{Y}^{-1})(-2\boldsymbol{\zeta}^{(i)}(\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^{T})\tilde{\boldsymbol{B}}_{XY}^{T}))}{\partial \tilde{\boldsymbol{B}}_{XX}} \\ &-\frac{N}{2}\frac{\partial \operatorname{Tr}(\boldsymbol{D}_{Y}^{-1}\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\Lambda}_{\xi}\tilde{\boldsymbol{B}}_{XY}^{T})}{\partial \tilde{\boldsymbol{B}}_{XX}} \\ &-\frac{N}{2}\frac{\partial \operatorname{Tr}(\boldsymbol{D}_{Y}^{-1}\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^{T}\tilde{\boldsymbol{B}}_{XY}^{T})}{\partial \tilde{\boldsymbol{B}}_{XX}} \\ &-\frac{1}{2}\frac{\partial \operatorname{Tr}(\boldsymbol{D}_{Y}^{-1}\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^{T}\tilde{\boldsymbol{B}}_{XY})}{\partial \tilde{\boldsymbol{B}}_{XX}}.\end{aligned} \tag{4.21}$$

To finally obtain the derivative with respect to $\boldsymbol{B}_{XX}$, we compute the trace of each block of the resulting matrix, as described in Equation (4.20), that is

$$\frac{\partial Q(\boldsymbol{B}_{XX}|\boldsymbol{B}_{XX}^{(p)})}{\partial (\boldsymbol{B}_{XX})_{u,v}} = \frac{\partial Q(\boldsymbol{B}_{XX}|\boldsymbol{B}_{XX}^{(p)})}{\partial \tilde{\boldsymbol{B}}_{XX}}\Big|_{u:u+K,v:v+K}. \tag{4.22}$$

We present the explit derivation of each term in (4.21) in the next section. Maximising the expected log-likelihood with respect to $\boldsymbol{B}_{XX}$ yields the connectivity structure among cortical sources that best explains the observed EEG-derived coefficients $\boldsymbol{\zeta}$ under the assumed Gaussian model.

Explicit computation of the derivative components

By following the procedure described above, we determine the explicit expression for the derivatives of all the terms in Q. We now present their derivation, for which we need to define some intermediate terms to ease the notation.

Denoting

$$a_1 = \tilde{\boldsymbol{B}}_{XY}\, \boldsymbol{\Lambda}_{\zeta}^{-1}\, \boldsymbol{\Lambda}_{\xi\zeta}^{\top}\, (\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}$$

and

$$a_2 = (\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\, \boldsymbol{\Lambda}_{\xi\zeta}\, \boldsymbol{\Lambda}_{\zeta}^{-1}\, \tilde{\boldsymbol{B}}_{XY},$$

we write

$$\begin{aligned}
&\frac{\partial \operatorname{Tr}(-(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^{T})}{\partial \tilde{\boldsymbol{B}}_{XX}} = \\
&-((\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}a_1(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\
&+(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}a_1(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T} \\
&-(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}a_1\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\tilde{\boldsymbol{B}}_{XY}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\
&-(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}a_1\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\tilde{\boldsymbol{B}}_{XY}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T} \\
&+(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}a_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\
&+(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}a_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}).
\end{aligned} \tag{4.23}$$

For the third term of $Q(\boldsymbol{B}_{XX}|\tilde{\boldsymbol{B}}_{XX}^{(p)})$, we define the following intermediate terms

$$b_1 = \tilde{\boldsymbol{B}}_{XY}^{\top}\, \boldsymbol{\Lambda}_{\zeta}^{-1}\, \boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\, \boldsymbol{\Lambda}_{\zeta}^{-1}\, \boldsymbol{\Lambda}_{\xi\zeta}^{\top}\, (\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1},$$

$$b_2 = \boldsymbol{\Lambda}_{\xi\zeta}\, \boldsymbol{\Lambda}_{\zeta}^{-1}\, \tilde{\boldsymbol{B}}_{XY},$$

$$b_3 = \tilde{\boldsymbol{B}}_{XY}^{\top}\, \boldsymbol{\Lambda}_{\zeta}^{-1}\, \boldsymbol{\Lambda}_{\xi\zeta}^{\top},$$

and finally

$$b_4 = (\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\, \boldsymbol{\Lambda}_{\xi\zeta}\, \boldsymbol{\Lambda}_{\zeta}^{-1}\, \boldsymbol{\zeta}^{(i)}\, \boldsymbol{\Lambda}_{\zeta}^{-1}\, \tilde{\boldsymbol{B}}_{XY}.$$

Thus, the derivative of the third term of $Q(\boldsymbol{B}_{XX}|\tilde{\boldsymbol{B}}_{XX}^{(p)})$ can be written as

$$
\begin{aligned}
&\frac{\partial \operatorname{Tr}(-(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\zeta}^{(i)}\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^{T})}{\partial \tilde{\boldsymbol{B}}_{XX}} = \\
&(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}b_1(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\
&+(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}b_1(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T} \\
&-(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}b_1b_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\
&-(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}b_1b_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T} \\
&-(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}b_3b_4(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\
&-(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}b_3b_4(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T} \\
&+(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}b_4(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\
&+(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}b_4(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}.
\end{aligned}
\tag{4.24}
$$

The fourth term of $Q(\boldsymbol{B}_{XX}|\tilde{\boldsymbol{B}}_{XX}^{(p)})$ can be rewritten as

$$
\begin{aligned}
&\frac{\partial \operatorname{Tr}((\boldsymbol{D}_Y^{-1})(-2\boldsymbol{\zeta}^{(i)}(\boldsymbol{\zeta}^{(i)T}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^{T})\tilde{\boldsymbol{B}}_{XY}^{T}))}{\partial \tilde{\boldsymbol{B}}_{XX}} = \\
&-2(-\left((\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}c_1c_2(I-\tilde{\boldsymbol{B}}_{XX})^{-1}\right)^{\top} \\
&-\left((\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}c_1c_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}\right)^{\top} \\
&+\left((\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}c_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\right)^{\top} \\
&+\left((\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}c_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_X(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}\right)^{\top}),
\end{aligned}
\tag{4.25}
$$

where we defined

$$c_1 = \tilde{\boldsymbol{B}}_{XY}^{\top}\,\boldsymbol{\Lambda}_{\zeta}^{-1}\,\boldsymbol{\Lambda}_{\xi\zeta}^{\top}$$

and

$$c_2 = \tilde{\boldsymbol{B}}_{XY}^{\top}\,(\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1}\,\boldsymbol{\zeta}^{(i)}\,\boldsymbol{\Lambda}_{\zeta}^{-1}\,\tilde{\boldsymbol{B}}_{XY}.$$

By defining

$$d = \tilde{\boldsymbol{B}}_{XY}^{\top} \left(\boldsymbol{\Lambda}_{\xi}^{(p)}\right)^{-1} \tilde{\boldsymbol{B}}_{XY},$$

the derivative of the fifth term of $Q(\boldsymbol{B}_{XX}|\tilde{\boldsymbol{B}}_{XX}^{(p)})$ becomes

$$\begin{aligned}\frac{\partial \operatorname{Tr}(\boldsymbol{D}_{Y}^{-1}\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\Lambda}_{\xi}\tilde{\boldsymbol{B}}_{XY}^{T})}{\partial \tilde{\boldsymbol{B}}_{XX}} &= (\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_{X}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}d(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\ &\quad + (\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}d(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_{X}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}.\end{aligned} \tag{4.26}$$

We define the following intermediate terms for the derivative of the sixth term of $Q(\boldsymbol{B}_{XX}|\tilde{\boldsymbol{B}}_{XX}^{(p)})$:

$$e_1 = \tilde{\boldsymbol{B}}_{XY}^{\top}\,\boldsymbol{\Lambda}_{\zeta}^{-1}\,\boldsymbol{\Lambda}_{\xi\zeta}^{\top}\,\tilde{\boldsymbol{B}}_{XY}^{\top}\left(\boldsymbol{\Lambda}_{\xi}^{(p)}\right)^{-1}\tilde{\boldsymbol{B}}_{XY}$$

$$e_2 = \boldsymbol{\Lambda}_{\xi\zeta}\,\boldsymbol{\Lambda}_{\zeta}^{-1}\,\tilde{\boldsymbol{B}}_{XY}$$

$$e_3 = \tilde{\boldsymbol{B}}_{XY}^{\top}\left(\boldsymbol{\Lambda}_{\xi}^{(p)}\right)^{-1}\tilde{\boldsymbol{B}}_{XY}$$

so that

$$\begin{aligned}&\frac{\partial \operatorname{Tr}(\boldsymbol{D}_{Y}^{-1}\tilde{\boldsymbol{B}}_{XY}\boldsymbol{\Lambda}_{\xi\zeta}\boldsymbol{\Lambda}_{\zeta}^{-1}\boldsymbol{\Lambda}_{\xi\zeta}^{T}\tilde{\boldsymbol{B}}_{XY}^{T})}{\partial \tilde{\boldsymbol{B}}_{XX}} = \\ &-\Big((\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_{X}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}e_1(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\ &\quad + (\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}e_1(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_{X}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T} \\ &\quad - (\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_{X}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}e_1e_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\ &\quad - (\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}e_1e_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_{X}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T} \\ &\quad + (\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_{X}(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}e_3e_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1} \\ &\quad + (\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-T}e_3e_2(\boldsymbol{I}-\tilde{\boldsymbol{B}}_{XX})^{-1}\tilde{\boldsymbol{D}}_{X}(I-\tilde{\boldsymbol{B}}_{XX})^{-T}\Big).\end{aligned} \tag{4.27}$$

Finally, considering

$$f_1 = \tilde{\boldsymbol{B}}_{XY}^{\top} \boldsymbol{\Lambda}_{\zeta}^{-1} \zeta_{\tilde{\xi}}^2 \boldsymbol{\Lambda}_{\zeta}^{-1} \boldsymbol{\Lambda}_{\xi\zeta}^{\top} \tilde{\boldsymbol{B}}_{XY}^{\top} (\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1} \tilde{\boldsymbol{B}}_{XY}$$

$$f_2 = \boldsymbol{\Lambda}_{\xi\zeta} \boldsymbol{\Lambda}_{\zeta}^{-1} \tilde{\boldsymbol{B}}_{XY}$$

$$f_3 = \tilde{\boldsymbol{B}}_{XY}^{\top} \boldsymbol{\Lambda}_{\zeta}^{-1} \boldsymbol{\Lambda}_{\xi\zeta}^{\top}$$

$$f_4 = \tilde{\boldsymbol{B}}_{XY}^{\top} (\boldsymbol{\Lambda}_{\xi}^{(p)})^{-1} \tilde{\boldsymbol{B}}_{XY} \boldsymbol{\Lambda}_{\xi\zeta} \boldsymbol{\Lambda}_{\zeta}^{-1} \boldsymbol{\zeta}^{(i)} \boldsymbol{\zeta}^{(i)T} \boldsymbol{\Lambda}_{\zeta}^{-1} \tilde{\boldsymbol{B}}_{XY}$$

we obtain that

$$\begin{aligned}
&\frac{\partial \operatorname{Tr}(\boldsymbol{D}_Y^{-1} \tilde{\boldsymbol{B}}_{XY} \boldsymbol{\Lambda}_{\xi\zeta} \boldsymbol{\Lambda}_{\zeta}^{-1} \boldsymbol{\zeta}^{(i)} \boldsymbol{\zeta}^{(i)T} \boldsymbol{\Lambda}_{\zeta}^{-1} \boldsymbol{\Lambda}_{\xi\zeta}^{T} \tilde{\boldsymbol{B}}_{XY})}{\partial \tilde{\boldsymbol{B}}_{XX}} = \\
&(\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \tilde{\boldsymbol{D}}_X (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} f_1 (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \\
&+ (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} f_1 (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \tilde{\boldsymbol{D}}_X (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} \\
&- (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \tilde{\boldsymbol{D}}_X (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} f_1 f_2 (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \\
&- (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} f_1 f_2 (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \tilde{\boldsymbol{D}}_X (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} \\
&- (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \tilde{\boldsymbol{D}}_X (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} f_3 f_4 (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \\
&- (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} f_3 f_4 (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \tilde{\boldsymbol{D}}_X (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} \\
&+ (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \tilde{\boldsymbol{D}}_X (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} f_4 (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \\
&+ (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T} f_4 (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1} \tilde{\boldsymbol{D}}_X (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-T}.
\end{aligned} \tag{4.28}$$

4.2.3 Optimisation procedure

Since we are interested in estimating causal interactions among the activity of cortical brain sources, we ask the estimated graph to be a directed acyclic graph (DAG), namely a graph with directed edges and no cycles. We briefly discussed this definition in Chapter 1. To this aim, we include the EM framework described in the previous section, in the approach implemented by Zheng and collegues [24], named DAGs with NO TEARS. This method uses a characterisation of acyclicity to convert the traditional combinatorial optimi-

sation problem arising when dealing with DAG constraints into a continous program.

To include our EM framework in the DAG with NO TEARS approach, we first need to consider the minimisation of the function $-Q(\boldsymbol{B}|\boldsymbol{B}_{XX}^{(p)})$, rather than the equivalent maximisation of $Q(\boldsymbol{B}|\boldsymbol{B}_{XX}^{(p)})$. Moreover, we consider the regularised score function $F(\boldsymbol{B}) : \mathbb{R}^{n\times n} \to \mathbb{R}$, defined as

$$F(\boldsymbol{B}) = -Q(\boldsymbol{B}|\boldsymbol{B}_{XX}^{(p)}) + \lambda\|\text{vec}(\boldsymbol{B})\|_1,$$

where λ is a regularisation parameter and $\|\text{vec}(\boldsymbol{B})\|_1$ is the ℓ_1 norm of the vectorized form of $\boldsymbol{B}$ used to induce sparsity in the solution. Consequently, the minimisation problem we are interested in solving is

$$\begin{aligned} &\min_{\boldsymbol{B}\in\mathbb{R}^{nK}} F(\boldsymbol{B}) \\ &\text{subject to } G(\boldsymbol{B}) \in \text{DAGs}, \end{aligned} \tag{4.29}$$

considering $G(\boldsymbol{B})$ the graph of n nodes induced by the matrix $\boldsymbol{B}$. Zheng et al. prove that the DAG constraint is equivalent to the smooth equality constraint $h(\boldsymbol{B}) = \text{Tr}(e^{\boldsymbol{B}\circ\boldsymbol{B}}) - n = 0$. This equality-constrained problem is solved by using an augmented Lagrangian method which (1) introduces a quadratic penalty in the loss function to be minimised; (2) approximates the solution of a constrained problem by the solution of an unconstrained problem. The reader may consult [88] for an accurate description of the Lagrangian method.

The optimisation (4.29) reduces to

$$\min_{\boldsymbol{B}\in\mathbb{R}^{n\times n}} L^{\rho}(\boldsymbol{B}, \alpha)$$

where

$$L^{\rho}(\boldsymbol{B}, \alpha) = F(\boldsymbol{B}) + \frac{\rho}{2}|h(\boldsymbol{B})|^2 + \alpha h(\boldsymbol{B}).$$

We refer to [24] for more details regarding the DAG with NO TEARS implementation. In Algorithm 3 we present the complete optimisation proce-

dure.

Algorithm 3 DAG with NO TEARS for the estimation of latent brain connectivity

Require: $w_{\text{est}}, \rho = 1, \alpha = 0, h = \text{inf}$

for $t = 1, \ldots, T_{\max}$ **do**
  while $\rho < \rho_{\max}$ **do**
    Compute $w_{\text{new}} = \arg\min_w(L^\rho(\boldsymbol{w}, \alpha))$.
    $h_{\text{new}} = h(w_{\text{new}})$
    Update $Q(w|w_{\text{new}})$
    if $h_{\text{new}} > 0.25\, h$ **then**
      $\rho = 10\rho$
    else
      break
    end if
  end while
  Update primal and dual variables: $w_{\text{est}} = w_{\text{new}}, h = h_{\text{new}}, \alpha = \alpha + \rho h$.
  if $h \leq h_{\text{tol}}$ **or** $\rho \geq \rho_{\max}$ **then**
    break
  end if
end for
return w_{est}.

4.3 Numerical validation

4.3.1 Simulation settings

In this section we present a simple simulation to evaluate the proposed method. We generate directly the scores vectors $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ by means of the noise terms $\boldsymbol{\eta}^{\boldsymbol{X}}$ and $\boldsymbol{\eta}^{\boldsymbol{Y}}$, without simulating the functional data $\boldsymbol{X}$ and $\boldsymbol{Y}$. We choose to use two basis functions, namely $K = 2$ and simulate $n = 3$ cortical brain sources and $m = 2$ EEG sensors. We consider $N = 100$ subjects in our sample.

We define $\boldsymbol{D}_X = \boldsymbol{D}_Y = \begin{bmatrix} 1 & 0 \\ 0 & 0.8 \end{bmatrix}$ so that $\tilde{\boldsymbol{D}}_X = \boldsymbol{D}_X \otimes \boldsymbol{I}_K$ and $\tilde{\boldsymbol{D}}_Y = \boldsymbol{D}_Y \otimes \boldsymbol{I}_K$. For each subject in the dataset, the noise terms are then sampled as $\boldsymbol{\eta}^{\boldsymbol{X}} \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{D}}_X)$ and $\boldsymbol{\eta}^{\boldsymbol{Y}} \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{D}}_Y)$.

The ground truth matrix $\boldsymbol{B}_{XX}$ representing the latent brain connectivity of interest, is exploited to generate the latent brain scores $\boldsymbol{\xi}$. In particular, we simulated $\boldsymbol{B}_{XX}$ as

$$\boldsymbol{B}_{XX} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{4.30}$$

and $\tilde{\boldsymbol{B}}_{XX} = \boldsymbol{B}_{XX} \otimes \boldsymbol{I}_K$. We also simulate the leadfield matrix $\boldsymbol{B}_{XY}$ describing the effect of the cortical brain activity on the EEG recordings as

$$\boldsymbol{B}_{XY} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

and $\tilde{\boldsymbol{B}}_{XY} = \boldsymbol{B}_{XY} \otimes \boldsymbol{I}_K$.

Finally, the scores vectors are computed as $\boldsymbol{\xi} = (\boldsymbol{I} - \tilde{\boldsymbol{B}}_{XX})^{-1}\boldsymbol{\eta}^{\boldsymbol{X}}$ and $\boldsymbol{\zeta} = \tilde{\boldsymbol{B}}_{XY}\boldsymbol{\xi} + \boldsymbol{\eta}^{\boldsymbol{Y}}$.

The initial point of the optimisation algorithm is a matrix $\boldsymbol{B}_0$ of dimensions $n \times n$, obtained by perturbing the ground truth $\boldsymbol{B}_{XX}$ in the following way: first, we set $\boldsymbol{B}_0 = \boldsymbol{B}_{XX}$; then, we choose a random set of elements in the matrix $\boldsymbol{B}_0$ and we sample those from a normal distribution of zero mean and variance 0.5, namely $\mathcal{N}(0, 0.5)$. For this simulation, we do not consider any regularisation, which means we set the regularisation parameter $\lambda = 0$.

4.3.2 Result

Figure 4.2 illustrates the binary ground truth and the estimated latent directed brain connectivity obtained when simulating the data as described in the previous section. In particular, Figure 4.2(a) shows the ground truth $\boldsymbol{B}_{XX}$ generated according to (4.30). Figure 4.2(b) and 4.2(c) show the estimated weighted and thresholded binary graphs, respectively. The latter is obtained by retaining the 80% higher connections.

The width of edges in Figure 4.2(b) is proportional to the magnitude of the

corresponding estimated coefficients. The edges (X_2, X_1) and (X_3, X_1) exhibit substantially larger weights than (X_2, X_3) and (X_3, X_2). The true directed connections (X_2, X_1) and (X_3, X_1) are therefore correctly recovered in this toy simulation, while the remaining estimated connections are negligible in magnitude.

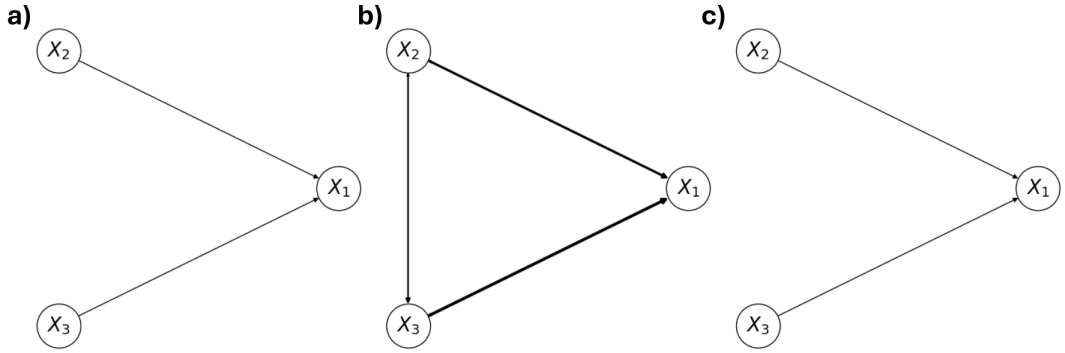

Figure 4.2: True and estimated latent brain connectivity in the toy simulation: (a) ground-truth binary network; (b) estimated weighted network; and (c) binary network, thresholded by retaining the strongest 80% of connections.

4.4 Discussion

In this chapter we introduced a novel method to estimate latent brain connectivity of cortical brain sources from EEG data. The proposed method exploits SEM to model the structural relationships between cortical brain activity and EEG recordings. This allows us to estimate the interactions occurring between the cortical brain sources directly, without the need to reconstruct their activity. Indeed, from the SEM formulation we can derive the distribution of the variables of interest and thus exploit maximum-likelihood estimation. Consequently, the EM algorithm is used to estimate the latent brain connectivity. Since SEM provides a model for directed connectivity, we propose to use the DAG with NO TEARS approach for estimating the directed acyclic network of interest. However, our inference method leads to the estimation of directed connectivity at the group level. This may represent a limitation of the

proposed approach since we do not have any information on the individual networks.

The preliminary simulation shows that the proposed approach can recover the main connectivity patterns, though further validation is needed. First of all, there is a need to evaluate the performance of the proposed method in a more complex simulation setting. Specifically, in the data generation process the number of sensors and cortical brain sources will be increased to construct a more realistic simulation of EEG real-world applications. Consequently, a leadfield matrix derived from a realistic conductivity model will be considered. A further step would be dedicated to the simulation of the EEG recordings and cortical brain signals, so that we can then test the decomposition in basis functions and scores. A systematic analysis is required for testing the optimisation procedure. In particular, optimisation is very sensitive to the choice of the initial point and further analysis is necessary for understanding its choice. Finally, validation on a real EEG dataset will be performed.

From a theoretical point of view, a further analysis of the maximum-likelihood estimation might be necessary. Indeed, the covariance structure of our model depends on a Kronecker product and the existence of the estimator of such matrix depends on the dimensionality of the data, thus it is not always granted [89, 90]. Future work will be devoted to address this topic.

In conclusion, this chapter presented a novel theoretical framework for the estimation of directed brain connectivity. Future work will be devoted to gain a more comprehensive empirical evaluation and practical implementation.

Part II.

Statistical analysis for the comparison of brain connectivity networks

Chapter 5

Fundamentals of the network-based statistics

In this chapter, we address the problem of comparing multiple groups of brain connectivity networks. While in the previous chapters our discussion focused on different approaches for the estimation of brain connectivity from magneto- or electro-encephalographic (M/EEG) data, we are now interested to perform analyses and statistical inference on brain connectivity networks. Indeed, a large number of variables, such as age, clinical conditions, neurodegenerative disorders, may influence the connections between the activity of different brain sources, both within and across individuals. Quantifying differences in functional connectivity across different conditions may allow to identify markers of such disorders, thus helping physicians in the discrimination procedure of subjects characterised by those different conditions. Literature on statistical methods to perform inference on brain networks and compare groups of graphs comprises statistical testing, multivariate and machine learning approaches. A comprehensive summary can be found in [25, 91, 92]. Recently, methods based on functional graphical models were developed to directly estimate differences among graphs [5, 93, 94]. Here, we focus on the approach proposed by Zalensky and colleagues [26], known as network-based statistics (NBS). This method seeks to identify differential connected compo-

nents among groups of graphs, by first performing an univariate test for each edge of the graph; then, the family-wise error rate (FWER) is controlled by grouping the significant edges in connected components and by assigning a p-value to each connected component according to the null distribution of the maximal component size approximated through a permutation test. When comparing three or more groups, the NBS implementation generally uses an F-test in the univariate analysis, thus testing for an overall group effect. Therefore, post-hoc analysis is required to identify which pairs of groups actually contribute to the overall effect. In this chapter, we present some preliminary results [3]. In particular, we discuss and compare two alternatives to the F-test, namely a Tukey–Kramer test and a Bonferroni-corrected pairwise t-test, to allow pairwise comparisons between groups while controlling this additional source of multiple comparisons.

The chapter is organized as follows. In Section 5.1 we introduce the basic statistical concepts that will be used in the following sections. In Section 5.2 we describe the multiple comparison problem that usually arises when testing for group differences in brain networks and some existing approaches to deal with it. In Section 5.3 we present the permutation test framework and in Section 5.4 we provide a formal description of the NBS algorithm. Finally, in Section 5.5 we provide a preliminary simulation setting to investigate the alternatives to the F-test, offering our evaluation and discussion.

5.1 Statistical hyphotesis testing

In the following section we introduce some concepts of basic statistics that will be used later in this chapter.

A *statistical hypothesis* is a statement concerning the parameters of a probability distribution or, more generally, the parameters of a statistical model. Such a hypothesis represents a conjecture about the underlying mechanism generating the observed data, and represents a formal method for inference by separating scientific claims from statistical noise. In particular, we call

the null hyphotesis H_0 a baseline assumption that we assess throught statistical tests. It typically states the absence of an effect or a difference. If H_0 is not rejected, any experimentally observed effect is due to chance alone. The statement that is being tested against the null hypothesis is instead called the alternative hypothesis, denoted by H_1.

Statistical hyphotesis testing usually consists in computing an appropriate test statistic based on the observed data, and deciding whether to reject or fail to reject the null hypothesis H_0 based on the observed value of the test statistic. An essential component of this procedure is the specification of a set of values of the test statistic for which the null hypothesis is rejected. This set is known as the *critical region* or *rejection region* of the test.

In the context of hypothesis testing, two types of errors may occur. A *type I error* (or false positive) occurs when the null hypothesis is rejected even though it is true. A *type II error* (or false negative) when the null hypothesis is not rejected despite being false. The probabilities of these errors are denoted by

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true}),$$

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \mid H_0 \text{ is false}).$$

It is often convenient to consider the *power* of a test, defined as

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 \mid H_0 \text{ is false}).$$

The probability of a type I error is denoted by α and is usuallly called the *significance level* of the test. The general approach to hypothesis testing consists of specifying an acceptable significance level and then designs the testing procedure so that the probability of a type II error β is as small as possible.

In the following we describe the statistical tests that will be used in this chapter, namely the t-test, the F-test and the Tukey–Kramer test.

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution $X \sim \mathcal{N}(\mu_X, \sigma^2)$, and $\bar{X}$ and S^2 be the sample mean and the sample variance of X, respectively. When the null and alternative hypothesis are respectively

$$H_0 : \mu_X = \mu_0, \quad H_1 : \mu_X > \mu_0,$$

the quantity

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \tag{5.1}$$

follows the Student's t-distribution with $n - 1$ degrees of freedom. The null hypothesis H_0 is rejected if and only if

$$\hat{t} > t_{\alpha;n-1},$$

where $\hat{t}$ is the sample value of (5.1) and $t_{\alpha;n-1}$ is the upper α percentage point of a Student's distribution with $n - 1$ degree of freedom.

Let us now consider a second random sample $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$ of sample size m. When comparing means of two groups sampled from distribution with equal variances, the null hypothesis is typically $H_0 : \mu_X = \mu_Y$, or equivalently, $H_0 : \mu_X - \mu_Y = 0$, while the alternative hypothesis reads $H_1 : \mu_X > \mu_Y$ -or, similarly $H_1 : \mu_X < \mu_Y$ or $H_1 : \mu_X \neq \mu_Y$. A pooled estimate of the common variance can be computed as

$$S^2_{pool} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n + m - 2}.$$

Under the null hypothesis H_0 the test statistic then becomes

$$T = \frac{\bar{X} - \bar{Y}}{S_{pool}\sqrt{\frac{1}{n} + \frac{1}{m}}}, \tag{5.2}$$

that follows the Student's t-distribution with $n + m - 2$ degrees of freedom. In this case H_0 is rejected if and only if

$$\hat{t} > t_{\alpha;n+m-2}$$

where $\hat{t}$ is the sample value of (5.2) and $t_{\alpha;n+m-2}$ is the upper α percentage

point of the Student's distribution with $n + m - 2$ degree of freedom.

When comparing variability across more than two groups, the F-test is commonly used to assess whether $Q > 2$ population means are equal. In this setting, we consider $X^1, \dots, X^Q$ random samples of dimension $n_1, \dots, n_Q$ respectively. The null hypothesis states

$$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_Q,$$

while the alternative hypothesis states that at least one group mean differs. The mean square between groups is defined as

$$\mathrm{MS_b} = \frac{1}{Q-1} \sum_{k=1}^{Q} n_k \left(\bar{X}^k - \bar{X} \right)^2,$$

where $\bar{X}^k$ is the sample mean of group k and $\bar{X}$ is the overall sample mean.

Moreover, the mean square within groups is given by

$$\mathrm{MS_w} = \frac{1}{n-Q} \sum_{k=1}^{Q} \sum_{i=1}^{n_k} \left(X_i^k - \bar{X}^k \right)^2,$$

where $n = \sum_{k=1}^{Q} n_k$ is the total sample size.

The F-test statistic is defined as the ratio of the variability between groups to the variability within groups, that is

$$F = \frac{\mathrm{MS_b}}{\mathrm{MS_w}}. \tag{5.3}$$

Denoted with $f_{\alpha;Q-1,n-Q}$ the upper α percentage point of a Fisher distribution with degree of freedom $Q-1$ and $n-Q$, the null hypothesis H_0 is rejected if and only if

$$\hat{f} > f_{\alpha;Q-1,n-Q}$$

where $\hat{f}$ is the sample value of (5.3).

The F-test allows to identify an overall group effect, but does not provide

any information on which group means actually differ. The Tukey–Kramer's test overcome this limitation by performing a statistical test for each pair of subgroups. In particular, under the null hypothesis of equal mean between any pair of groups, that is $H_0 : \mu_k = \mu_{k'}$ for $k, k' = 1, \dots, Q$, the Tukey–Kramer test statistics is

$$T_{KR} = \frac{|\bar{X}^k - \bar{X}^{k'}|}{\sqrt{0.5\mathrm{MS_w}(\frac{1}{n_k} + \frac{1}{n_{k'}})}}. \tag{5.4}$$

Given $\hat{t}_{KR}$ the sample value of (5.4), the null hypothesis is rejected if and only if

$$\hat{t}_{KR} > q_{\alpha;Q,n-Q}$$

with $q_{\alpha;Q,(n-Q)}$ is the upper α percentage point of a studentized range distribution for Q groups and $n - Q$ degree of freedoms. The Tukey–Kramer test also controls the multiple comparison problem arising from the $Q(Q-1)/2$ tests. We will discuss this further in Section 5.5.

More details regarding these basic statistical concepts can be found in [95, 49]. We now return to the problem of comparing brain connectivity networks associated with different clinical conditions.

5.2 Multiple comparison problem in brain networks analysis

When exploring differences among brain networks, it is a common practice to perform a specific statistical test independently to a large number of nodes or edges within the brain networks. This approach is known as mass univariate testing and provides a localized effect to individual nodes or edges.

Let us denote with *family* the collection of tested elements, namely nodes or edges; its size F can range from hundreds to millions depending on the size of the network [25]. Since mass univariate testing yields a test statistic and corresponding p-values for each element within the family, it gives rise to a

multiple comparison problem for which it is not appropriate to simply reject the null hypothesis based on the nominal significance for each test without further correction. Specifically, if we denote with α the individual test significance, the probability of rejecting at least one true null hypothesis becomes unacceptably large, approximately equal to αF. Therefore, a more severe significance level needs to be chosen to control the probability of making one or more false positives across the entire family. This probability is known as the family-wise error rate (FWER).

Numerous procedures have been proposed to deal with the multiple comparison problem. The simplest approaches, such as the Bonferroni correction [96], the Sidak procedure [97] and the Holm-Bonferroni step procedure [98], control the FWER in a strong sense. However, they are generally too conservative and often lead to high false negative rates. Alternatively, less conservative approaches have been proposed to control the FWER in a weak sense. Under the null hypothesis, weak FWER control methods behave in a similar way to the strong control ones. However, if at least one true effect exists, the probability of making even one false positive may exceed the desired threshold. Among the alternative methods to the strong FWER control, several focus on controlling the false discovery rate (FDR), which was introduced by Benjamini and Hochberg in 1995 [99]. Controlling the FDR provides greater power at the cost of an increased false positive rate. On the other side, the overly conservative nature of strong FWER control methods often lead to high false negative rates. So usually, researchers have turned to the alternative FDR approach, which offers a more balanced trade-off between sensitivity and specificity [25].

The analysis becomes more complex when measures computed across a family of nodes or connections, are not independent of each other. This is particularly evident in correlation-based networks, such as the brain functional networks under investigation. Indeed, even though family-wise errors are rarer under positive dependence, the consequences of such errors are more severe. This is because statistical dependencies among tests may cause an effect on a node or connection to influence neighboring elements, leading to

clusters of significant multiple test statistics, which are in reality false positives. In brain connectivity literature, this dependence issue is often ignored. Yet, resampling methods and permutation-based approachs can be used to address the problem of multple comparisons for correlated tests. Even if permutation tests can be computationally heavy, they ensure that the dependencies between tests is transferred to the test statistic null distribution, making it a practical choice for many applications in statistical connectomics [25].

5.3 Permutation approach for general linear model

In this section, we provide an overview of the permutation tests framework for general linear model (GLM). We discuss the exchangeability assumption and the permutation procedure, also considering the presence of nuisance variables; finally we present the Freedman–Lane procedure and its implementation. The main reference for this section is [100].

5.3.1 Introduction to linear modelling

Let us consider a general linear model [101] expressed as

$$\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{\psi} + \boldsymbol{\epsilon}, \tag{5.5}$$

where $\boldsymbol{Y} = (Y_1, \dots, Y_n) \in \mathbb{R}^{n\times 1}$ is the vector of n observations of a response variable, $\boldsymbol{M} = \begin{bmatrix} 1 & m_{1,1} & \dots & m_{1,r} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & m_{n,1} & \dots & m_{n,r} \end{bmatrix} \in \mathbb{R}^{n\times(r+1)}$ is the design matrix collecting the observed values of r explanatory variables (or regressors), $\boldsymbol{\psi} = (\psi_0, \dots, \psi_r) \in \mathbb{R}^{(r+1)\times 1}$ is the vector of regression coefficients and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n) \in \mathbb{R}^{n\times 1}$ is the vector of random errors. We are interested in testing the null hypothesis $H_0 : \boldsymbol{C}^T\boldsymbol{\psi} = 0$, where $\boldsymbol{C} \in \mathbb{R}^{r\times s}$ is a full-rank matrix of s contrasts, with $1 \leq s \leq r$. The contrast matrix specifies an arbitrary combination of some or all of the regressor coefficients that is tested for being equal to zero. This

univariate testing is repeated for all connections in the brain network, leading to the multiple comparison problem discussed in the previous section.

Example 2. *Let us consider n subjects and assume that for each subject we compute a connectivity matrix* $\boldsymbol{A}^{(i)}, i = 1, \dots, n$ *by using one of the connectivity metrics introduced in Chapter 1. The subjects are divided in two groups ,* $\mathcal{G}_1$ *containing the first* n_1 *subjects and* $\mathcal{G}_2$ *containing the remaining ones.*

Model (5.5) *can be exploited to identify differences for a brain edge* (u, v) *between the two groups. In this case* $\boldsymbol{Y}$ *collects the entry* $a_{u,v}$ *of each subject across the two groups, that is* $Y_1 = a^{(1)}_{u,v}, \dots, Y_n = a^{(n)}_{u,v}$*, and the design matrix can be written as*

$$\boldsymbol{M} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ \vdots & 1 & 0 \\ \vdots & 0 & 1 \\ \vdots & \vdots & \vdots \\ \vdots & 0 & 1 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1 \\ \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2 \end{matrix} .$$

If we are interested in testing for the equality of means between the two groups, the null and alternative hyphothesis to be tested are

$$H_0 : \psi_1 = \psi_2, \quad H_1 : \psi_1 \neq \psi_2.$$

The null hypothesis can be rewritten in terms of contrast vector $\boldsymbol{C\psi}$ *by defining the contrast vector as* $\boldsymbol{C} = [0, 1, -1]$.

In our discussion, we consider an equivalent model to (5.5) by partitioning the explanatory variables into those of interest, such as the group membership in the previous example, and those of nuisance, that is variables that encode possible confounding effects. The later variables can act as confounders in the model and introduce bias if they are correlated with the independent variables, so it is fundamental to take them into account. The model becomes

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{5.6}$$

where $\boldsymbol{X}$ is the design matrix for the regressors of interest, $\boldsymbol{Z}$ is the design matrix for the nuisance regressors, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the vectors of regression coefficients. This partitioning is not unique, but it can be defined in such a way that inference on $\boldsymbol{\beta}$ is equivalent to inference on $\boldsymbol{C}^T\boldsymbol{\psi}$.

Residuals in model (5.5) and (5.6) are equivalent and can be obtained as $\hat{\boldsymbol{\epsilon}} = \boldsymbol{R}_M\boldsymbol{Y}$, where $\boldsymbol{R}_M = \boldsymbol{I} - \boldsymbol{H}_M$ is the residual-forming matrix, $\boldsymbol{H}_M = \boldsymbol{M}\boldsymbol{M}^\dagger$ is the projection matrix obtained by multipling the design matrix $\boldsymbol{M}$ with its pseudoinverse $\boldsymbol{M}^\dagger$ and $\boldsymbol{I}_{n\times n}$ is the identity matrix. Similarly, the residuals due to the nuisance alone are $\hat{\boldsymbol{\epsilon}}_Z = \boldsymbol{R}_Z\boldsymbol{Y}$, with $\boldsymbol{R}_Z = \boldsymbol{I} - \boldsymbol{H}_Z$ and $\boldsymbol{H}_Z = \boldsymbol{Z}\boldsymbol{Z}^\dagger$. In the context of permutation methods, residuals assume an important role. In particular, a key aspect of the linear model is that residuals are not independent, even when the errors $\boldsymbol{\epsilon}$ are independent and homoscedastic a property that helps make these methods approximately exact. We further discuss this in Section 5.3.3.

In the next section, we discuss what statistic can be used to test the null hypothesis aforementioned. Moreover, we provide a definition of the p-value in the framework of permutation tests.

5.3.2 P-values in non-parametric hypothesis testing

In non-parametric settings, any statistic where large values reflect evidence against the null hypothesis could be used. It means that statistics that are independent of any unknown parameters can be chosen - these are known as pivotal statistics. Examples of such pivotal statistics include Student's t, the F ratio, the Pearson's correlation coefficient, the coefficient of determination, and more others used to construct confidence intervals and to compute p-values in parametric tests.

Regardless of the test statistic, p-values are usually used as a common measure against the null hypothesis. Thus, considered a certain test statistic T

and the value of the test statistic referred to the observed data, namely t^*, the p-value is defined as the probability, under the null hypothesis, of observing an equal or larger value of the test statistic than t^*, namely $P(T \geq t^* | \mathcal{H}_0)$. This definition refers only to one-sided tests, where the evidence against $\mathcal{H}_0$ corresponds to value larger than the observed one. Two-sided or negative-valued tests can also be considered and the corresponding p-values can be similarly defined.

Differently from parametric settings, the theoretical distribution of a specific test statistic can not be derived under some assumptions in non-parametric settings. Thus, instead of relying on parametric assumptions, the data are shuffled in a random way multiple times, maintaining it consistent with the null hypothesis. The model is then fitted again for each permutation j, and a new realization of the test statistic t^j is computed. This process results in an empirical distribution of T under the null hypothesis. The overall p-value is then calculated as:

$$\text{p-value} = \frac{1}{J} \sum_j I(t^j \geq t^*),$$

where J is the total number of permutations performed, and $I(\cdot)$ is the indicator function.

This formulation implies that non-parametric p-values are discrete, with each possible value being a multiple of $\frac{1}{J}$. Importantly, the permutation distribution must include the observed (unpermuted) test statistic [102, 103], which means that the smallest achievable p-value is $\frac{1}{J}$, not zero.

5.3.3 Permutations and exchangeability

The way data are shuffled under the null hypothesis is a central aspect of permutation tests. Indeed, the null hypothesis, together with the assumption of exchangeability, determines the appropriate permutation strategy. Depending on the null hypothesis, shuffling can be implemented using a permutation matrix or a sign-flipping matrix, as follows. Given a permutation matrix $\boldsymbol{P}_j \in \mathbb{R}^{n \times n}$, left-multiplying a data matrix by $\boldsymbol{P}_j$ permutes its rows. Similarly,

let $\boldsymbol{S}_j$ be a diagonal sign-flipping matrix of dimension $n \times n$; left-multiplication by $\boldsymbol{S}_j$ flips the signs of selected rows. Sign flips and permutations can be combined, depending on the structure of the null hypothesis. In such cases, terms like *shuffling* or *rearrangement* are often used to describe the process more generally. We now provide an example of a permutation matrix.

Example 3. *Let us consider a data vector* $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4) \in \mathbb{R}^4$. *Then, applying the permutation matrix*

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

to the data vector Y, *we obtain the permutation of the first and second rows, namely:*

$$\begin{bmatrix} Y_2 \\ Y_1 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix}.$$

More details and examples on permutation and sign-flipping matrices can be found in [100].

A fundamental assumption underlying permutation methods is that the joint distribution of a given set of variables remains invariant under rearrangment. This can be formalized using the concepts of exchangeable errors (EE) or, in more restrictive settings, independent and symmetric errors (ISE). The former is the traditional permutation requirement introduced by [104] and can be defined as follows.

Definition 32. *Let* $\mathcal{P}$ *be the set of permutations in* $\mathbb{R}^n$, *with n being the sample size of the data. For any permutation* $\boldsymbol{P}_j \in \mathcal{P}$ *an n-dimensional vector of errors* $\boldsymbol{\epsilon}$ *is said to be* exchangeable *if* $\boldsymbol{\epsilon} \overset{\delta}{=} \boldsymbol{P}_j\boldsymbol{\epsilon}$, *where* $\overset{\delta}{=}$ *denotes equality of distributions.*

Definition 32 means that errors are considered to be exchangeable if their joint distribution is invariant with respect to permutations. Exchangeability is

a weaker assumption than independence, as it allows for certain dependencies among variables while preserving invariance under permutations. Moreover, with respect to the common assumption of independent, normally and identically distributed errors, we note that normality and independence are no longer required. As long as the joint distribution remains invariant under permutations, this general condition applies to any distribution. As an example, a multivariate normal distribution is exchangeable when its covariance matrix exhibits compound symmetryi.e., all diagonal elements are equal and all off-diagonal elements are equal. The following definition refers to independent and symmetric errors:

Definition 33. *Let $\mathcal{S}$ be the set of sign flipping matrices, in $\mathbb{R}^n$, with n being the sample size of the data. For any sign flipping matrix $\boldsymbol{S}_j \in \mathcal{S}$ an n-dimensional vector of errors $\boldsymbol{\epsilon}$ is said to be independent and symmetric if $\boldsymbol{\epsilon} \overset{\delta}{=} \boldsymbol{S}_j\boldsymbol{\epsilon}$, where $\overset{\delta}{=}$ denotes equality of distributions.*

Compared to the assumption of independent, normally and identically distributed errors, ISE relaxes the requirment of normality, although symmetry (i.e., non-skewness) and independence of distributions are required. Indeed, independence is still necessary to ensure that sign flipping of one observation does not perturbes others. Such errors can arise when the variances from differences between two groups are not assumed to be the same.

The decision to use EE or ISE depends on the available knowledge or assumptions regarding the error terms. While ISE requires the error distributions to be symmetric, EE assumes that the variances and covariances are equal. However, when both EE and ISE are reasonable for a given model, combining permutations with sign flipping expands the set of possible rearrangementsa particularly beneficial property in studies with small sample sizes.

5.3.4 Freedman and Lane procedure

A number of approaches have been proposed to perform permutation tests under exchangeability in the presence of nuisance variables. We will focus on the Freedman–Lane procedure [105], because this is the one implemented in the network-based statistic algorithm from Zalesky et al [26]. Moreover, the paper from Winkler and collegues [100] shows that this method, together with the Smith method [106, 107], produces the best results in terms of control over error rates and power. Let us first recall the full model (5.6) presented in Section 5.3.1, namely

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

and define the reduced model as

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_{\boldsymbol{Z}}. \tag{5.7}$$

The core idea of the Freedman–Lane procedure is that, under the null hypothesis, the residuals of this full model, $\hat{\boldsymbol{\epsilon}}$, are equivalent to the residuals of the reduced model, $\hat{\boldsymbol{\epsilon}}_Z$. Therefore the residuals $\hat{\boldsymbol{\epsilon}}_Z$ can be used to build the reference distribution of the test statistic from which p-values can be obtained.

The Freedman–Lane procedure can be summarised as follows. First of all, the data $\boldsymbol{Y}$ are regressed against $\boldsymbol{X}$ and $\boldsymbol{Z}$ following the full model (5.6). The statistic of interest t^* can then be computed using the estimated parameters $\hat{\boldsymbol{\beta}}$. The estimated parameters $\hat{\gamma}$ and the estimated residuals $\hat{\boldsymbol{\epsilon}}_Z$ of the reduced models 5.7 are then computed. The permuted data $\boldsymbol{Y}^{(j)}$ are then computed by pre-multiplying the residual $\hat{\boldsymbol{\epsilon}}_Z$ by a permutation matrix $\boldsymbol{P}_j$ and adding back the estimated nuisance effects. Finally, $\boldsymbol{Y}^{(j)}$ is used as response variable in a full model, including both the variables of interest and the nuisance ones, namely: $\boldsymbol{Y}^{(j)} = P_j\hat{\boldsymbol{\epsilon}}_Z + \boldsymbol{Z}\hat{\boldsymbol{\gamma}}$, to estimate $\hat{\boldsymbol{\beta}}^{(j)}$, which is used to compute the statistic of interest for the j-th permutation. Repeating these steps for all permutations of the algorithm allows one to create a reference distribution under the null hypothesis, from which a p-value is ascribed by counting how

many times a statistic $t^{(j)}$ is equal to or greater than t^* and dividing this number by J. Algorithm 4 outlines the Freedman–Lane procedure.

Algorithm 4 Permutation test for effects of interest (Freedman–Lane)

Require: Response vector $\boldsymbol{Y}$, design matrix $\boldsymbol{X}$ (effects of interest) and $\boldsymbol{Z}$ (nuisance variables), number of permutations J

Fit full model: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$

Compute observed test statistic t^* from $\hat{\boldsymbol{\beta}}$

Fit reduced model: $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_Z$

for $j = 1$ to J **do**

  Generate permutation $\boldsymbol{P}_j$ of $\hat{\boldsymbol{\epsilon}}_Z$

  Construct permuted response:

$$\boldsymbol{Y}^{(j)} = \boldsymbol{P}_j\hat{\boldsymbol{\epsilon}}_Z + \boldsymbol{Z}\hat{\boldsymbol{\gamma}}$$

  Fit full model to permuted data:

$$\boldsymbol{Y}^{(j)} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

  Compute permuted test statistic $t^{(j)}$ from $\hat{\boldsymbol{\beta}}^{(j)}$

end for

Compute empirical p-value:

$$p = \frac{1}{J}\sum_{j=1}^{J}\mathbb{I}(t^{(j)} \geq t^*)$$

return p

5.3.5 Permutation-based approach for controlling the FWER

In the previous paragraph, we have described the permutation approach and the Freedman and Lane procedure for analysing a single element, such as node or edge, in the brain networks under investigation. However, when performing NBS, the described procedure needs to be repeated for every element of the considered family. Indeed, we introduced permutation methods specifically to address the multiple comparison problem arising from the numerous elements within a family.

Unlike parametric methods, permutation-based correction for multiple testing does not require additional assumptions. Holmes and colleagues [108] introduced this correction approach to control the FWER. It consists of retaining for each permutation j only the maximum value of the test statistic over all the elements within the family. This maximum value, denoted as $t^{(j)}_{\max}$, is then used to build the empirical distribution across permutations, resulting in an empirical distribution of maxima. For each test in the family, a FWER-corrected p-value can then be obtained by calculating the proportion of $t^{(j)}_{\max}$ values that exceed the observed statistic t^*. By looking at the maximum value of the test statistic under each permutation, one simulates the worst-case false positive that could occur by chance. Alternatively, a single FWER threshold can be applied to the entire statistical map of t^* values using the distribution of $t^{(j)}_{\max}$.

Because p-values are uniformly distributed in the interval $[0, 1]$ under the null hypothesis, they are also pivotal quantities and could, in principle, be used for multiple testing correction using a similar approach. Specifically, the distribution of the minimum p-value, $p^{(j)}_{\min}$, can be used instead of $t^{(j)}_{\max}$. However, due to the discrete nature of p-values, this method may introduce computational challenges and can result in a notable loss of statistical power.

False discovery rate (FDR) correction can also be applied once uncorrected p-values are obtained for each voxel or test. A single FDR threshold may be applied to the p-value map, or FDR-adjusted p-values can be calculated at each location.

5.4 Network-based statistics

The network-based statistic (NBS) is a network specific approach to control the FWER in the weak sense when performing mass univariate testing on correlated connections of groups of graphs. It is based on the fact that effects on the brain, such as changes in brain activity due to pathological conditions or specific cognitive tasks, are usually distributed throughout the brain ac-

cording to anatomical brain topology [109]. Thus, effects on brain networks likely involve multiple connections and nodes, which form interconnected subnetworks rather than being confined to single edges. The goal of NBS is to identify those subnetworks showing a significant effect while controlling the FWER. NBS can be summarised as follows: first, a the test statistic is computed for each pairwise connection in the network; the values of the test statistic are then thresholded to construct a set of suprathreshold links. Any connected component that may be present in the set of suprathreshold links is then identified using a breadth-first search algorithm (BFS) and their size is stored. Finally, permutation testing is used to controll for the FWER.

We do not discuss the BFS algorithm in this thesis. The reader can refer to [110] for its comprehensive characterisation. We now provide a detailed description of the NBS and its implementation.

Let us consider a collection of n undirected graphs $\{G_i\}_{i=1}^n$, which describe the brain connectivity networks of n subjects. We denote the graph associated to the i-th subject with $G_i = (V, E_i)$, where V is a fixed set of nodes and E_i is the set of edges for the corresponding graph. While V is assumed to be the same for every subject, the set of edges naturally changes according to the connectivity that characterises the brain of each subject.

Let $\{\boldsymbol{A}_i\}_{i=1}^n$ be the set of adjacency matrices associated to $\{G_i\}_{i=1}^n$, where $\boldsymbol{A}_i \in \mathbb{R}^{|V|\times|V|} \quad \forall i = 1, \dots, n$. NBS assumes the adjacency matrices to be weighted and symmetric. For this reason, in the following we will consider pair of edges (u, v) such that $u, v = 1, \dots, |V|$ and $u < v$.

Let us now recall the general linear model (5.6) presented in Section 5.3.1. Let $\mathbf{a}_{u,v} = (a_{u,v}^{(1)}, \dots, a_{u,v}^{(n)})^T \in \mathbb{R}^{n\times 1}$ denote the vector of edge weights across subjects and $a_{u,v}^{(i)}$ the element (u, v) of the adjcency matrix $(\boldsymbol{A}_i)_{u,v}$. Then, the full model can be written as

$$\mathbf{a}_{u,v} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{5.8}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times q}$ is the design matrix for the q explanatory variables of interest

and $\boldsymbol{Z} \in \mathbb{R}^{n \times q_z}$ is the design matrix for the q_z nuisance variables. The covariates of interest encoded in $\boldsymbol{X}$ often reflect the grouping of subjects according to clinical status or other factors of interest, while the nuisance variables in $\boldsymbol{Z}$ account for potential confounding effects, such as age or gender, without being the target of statistical inference. The vectors $\boldsymbol{\beta} \in \mathbb{R}^{q \times 1}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{q_z \times 1}$ are the corresponding regression coefficients, and $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times 1}$ is the error term.

The first step of NBS consists in fitting the full model (5.8) and estimating the regression coefficients $\hat{\boldsymbol{\beta}}$. Given $\boldsymbol{C} \in \mathbb{R}^{q \times s}$, $1 \leq s \leq q$, the contrast matrix introduced in Section 5.3.1, the NBS tests the null hypothesis $H_0 : \boldsymbol{C}^T \boldsymbol{\beta} = 0$. To this aim, for each edge (u, v) a test statistic, such as t- or F-statistic, is computed.

Since this procedure is repeated for every edge pair (u, v), it eventually leads to two triangular matrices: a matrix $\boldsymbol{T}$ containing the values of the computed test statistic, and a matrix $\boldsymbol{P}$ with the associated p-values. We note that NBS employs a one-tail test statistic. The matrix $\boldsymbol{T}$ is then thresholded to retain only connections greater than a pre-defined threshold level t_s, namely a new matrix $\tilde{\boldsymbol{T}}$ is defined as

$$\tilde{\boldsymbol{T}}_{u,v} = \begin{cases} \boldsymbol{T}_{u,v} & \text{if } \boldsymbol{T}_{u,v} > t_s \\ 0 & \text{otherwise} \end{cases}. \tag{5.9}$$

A breadth-first search algorithm (BFS) [110] is applied on the graph defined by $\tilde{\boldsymbol{T}}$ to detect connected components, namely $\{C_1, \ldots, C_K\}$. Each component is a subgraph of the original collection of graphs $\{G_i\}_{i=1}^n$ and it is denoted as $C_k = (V_k, E_k^{\text{cc}})$ with $V_k \subseteq V$ and $E_k^{\text{cc}} \subseteq \cup_{i=1}^n E_i$. The size of each component is then computed following one of two different definitions: either the *extent* size, namely the number of edges it contains $\sigma_k = |E_k^{\text{cc}}|$, or the *intensity* size, which is the weighted sum of the edges, $\sigma_k = \sum_{(u,v) \in E_k^{\text{cc}}} \tilde{\boldsymbol{T}}_{u,v}$.

Finally, a permutation test is performed to provide an empirical distribution of the maximal component size under the null hypothesis of exchangeability, discussed in Section 5.3.3. This empirical distribution is used to compute a p-value for each observed component while controlling for the FWER.

As mentioned in Section 5.3.5, exploiting the distribution of the maximal component size effectively simulates the worst-case false positive that could arise by chance.

When nuisance variables are not present in the model and the design matrix $\boldsymbol{X}$ reflects the subject's group membership, we can interpret the exchangeability assumption as the idea that the group label can be shuffled among subjects without affecting the analysis. Keeping this idea in mind can help approaching the overall rationale of the permutation procedure. However, this interpretation is only heuristic and the permutation procedure is implemented through the FreedmanLane procedure rather than by direct permutation of group labels.

Let us consider a permutation j, for $j = 1, \dots, J$. We now describe how the Freedman–Lane procedure is integrated in the NBS approach. First, the reduced model

$$\boldsymbol{a}_{u,v} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_Z$$

is fitted to estimate γ and $\boldsymbol{\epsilon}_Z$. The estimated residuals $\hat{\boldsymbol{\epsilon}}_Z$ are permuted to compute a permuted response vector $\mathbf{a}_{u,v}^{(j)} = \boldsymbol{P}_j\hat{\boldsymbol{\epsilon}}_Z + \boldsymbol{Z}\hat{\boldsymbol{\gamma}}$. The full model

$$\boldsymbol{a}_{u,v}^{(j)} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

is then fitted using the permuted responses $\mathbf{a}_{u,v}^{(j)}$ returning an estimate $\hat{\boldsymbol{\beta}}^{(j)}$ of the coefficients $\boldsymbol{\beta}$. At this point, the statistic is computed leading to new matrices $\boldsymbol{T}^{(j)}$ and $\boldsymbol{P}^{(j)}$. The thresholding step and the consequent identification of connected component is repeated, resulting in a new set of connected components $\{C_1^{(j)}, \dots, C_{K_j}^{(j)}\}$ for the j-th permutation.

The next step consists of retaining the size of the largest connected component, defined as $\sigma_{\max}^{(j)} = \max_{1,\dots,K}\{\sigma_1^{(j)}, \dots, \sigma_K^{(j)}\}$. This contributes to the construction of an empirical null distribution of the maximum component size.

Finally, the corrected p-value for an observed component C_k is computed

as

$$p_k = \frac{1}{J} \sum_{j=1}^{J} \mathbf{1}_{\{\sigma_{\max}^{(j)} \geq \sigma_k\}},$$

namely the proportion of permutations for which the maximum component size is equal to or greater than the observed component size.

We note that a somewhat arbitrary choice must be made to select the value of the test statistic threshold t_s and it is difficult to provide definitive rules guiding how to choose the set of suprathreshold links. If the threshold is chosen too low, large components can arise in the permuted data as a matter of chance and thereby reduce power. In contrast, if the threshold is set too high, connections comprising the effect of interest may not be admitted to the set of suprathreshold links. However, we note that the threshold t_s is mathematically related to the p-value of the univariate test statistic at the edge level, prior to any family-wise error control. Thus, we can interpret the threshold t_s in terms of statistical significance at the edge level, which can be fixed a priori.

Let us provide an example for the case of a one-tailed t-test. The relationship between the threshold and the p-values is expressed by $p_s = 1 - F_t(t_s; \nu)$, where ν are the degrees of freedom of the sample under consideration and F_t is the cumulative distribution function of the t-distribution. The choice of the threshold t_s can be guided by this expression. If we consider $\nu = 40$ and the significance level $p_s = 0.001$, the threshold is equal to $t_s = 3.31$.

In the NBS toolbox, users may choose either a t-test or an F-test as the test statistic within the general linear model. When dealing with more complex experimental designs for example, those including covariates of interest or nuisance variablesthese statistical tests can still be employed.

In practice, covariates of interest often define distinct groups of subjects, for which one may want to test the null hypothesis of equal connectivity. This happens, for example, when the covariates refer to presence or absence of a neurological disorder. For this reason, in the following section and in Chapter 6 we focus primarily on the special case of group comparisons.

Algorithm 5 Network-based statistic (NBS) algorithm

Input: Graphs $\{G_i\}_{i=1}^n$, adjacency matrices $\{\boldsymbol{A}_i\}_{i=1}^n$, design matrices $\boldsymbol{X}, \boldsymbol{Z}$, contrast $\boldsymbol{C}$, threshold t_s, permutations J

1. Edge-wise model fitting:

$$\mathbf{a}_{u,v} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}.$$

2. Testing null hyphotesis

$$H_0 : \boldsymbol{C}^T\boldsymbol{\beta} = 0$$

by computing chosen test statistic. Assemble $\boldsymbol{T}$ and $\boldsymbol{P}$.

3. Thresholding:

$$\tilde{\boldsymbol{T}}_{u,v} = \begin{cases} \boldsymbol{T}_{u,v} & \boldsymbol{T}_{u,v} > t_s \\ 0 & \text{otherwise.} \end{cases}$$

4. Detect connected components $\{C_1, \ldots, C_K\}$ **via BFS applied on** $\tilde{\boldsymbol{T}}$**.**

5. Compute component size $\sigma_k \quad \forall k = 1, \ldots, K$**.**

6. Permutation testing:

For $j = 1, \ldots, J$:

1. Estimate $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\epsilon}}_Z$ from $\boldsymbol{a}_{u,v} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_Z$.
2. Permute residuals of reduced model and compute $\mathbf{a}_{u,v}^{(j)} = \boldsymbol{P}_j\hat{\boldsymbol{\epsilon}}_Z + \boldsymbol{Z}\hat{\boldsymbol{\gamma}}$.
3. Estimate $\hat{\boldsymbol{\beta}}$ from the full model $\boldsymbol{a}_{u,v}^{(j)} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$.
4. Test H_0 and compute $\boldsymbol{T}^{(j)}$.
5. Compute thresholded matrix $\tilde{\boldsymbol{T}}^{(j)}$.
6. Detect $\{C_1^{(j)}, \ldots, C_{K_j}^{(j)}\}$.
7. Record $\sigma_{\max}^{(j)} = \max_{k=1,\ldots,K}\{\sigma_1^{(j)}, \ldots, \sigma_K^{(j)}\}$.

7. Compute corrected p-values:

$$p_k = \frac{1}{J}\sum_{j=1}^{J} \mathbf{1}_{\{\sigma_{\max}^{(j)} \geq \sigma_k\}} \quad \forall k = 1, \ldots K$$

Output: Connected components $\{C_k\}$ and corrected p-values $\{p_k\}$.

5.5 Network-based statistics for hypothesis testing in multiple groups of graphs

When coming to address the comparison of three or more groups, the NBS implementation uses an F-test in the univariate analysis. In particular, the F-test tests the null hypothesis that the groups are drawn from populations with equal mean values, thus testing for an overall group effect. If the null hypothesis is rejected, the test does not return any additional information on the groups for which there was an effect. Therefore, post-hoc analysis is required to identify which pairs of groups actually contribute to the overall effect. Here we discuss and compare two alternatives to the F-test, namely a Tukey–Kramer test and a Bonferroni-corrected pairwise t-test, to allow pairwise comparisons between groups while controlling this additional source of multiple comparisons. We introduced the definition of these statistical tests in Section 5.1.

In the following, we describe a preliminary simulation together with our evaluation and results.

5.5.1 Numerical simulations

Inspired by Normand and collegues [111], we propose the following data generation process to evaluate the Tukey–Kramer test and a Bonferroni-corrected pairwise t-test within the NBS framework. Our aim is to simulate brain connectivity networks for n subjects divided into Q groups.

As previously described, the NBS identifies differential connected components showing a significant effect, while controlling the FWER arising from the comparison of multiple edges. In order to detect significant differential components from simulated data, the generated graphs need to present at least a differential subnetwork. Thus, let us consider n undirected networks $\{G_i\}_{i=1}^{n}$, where $G_i = (V, E_i)$ as we modelled in the previous section, and their corresponding symmetric adjacency matrices $\{\boldsymbol{A}_i\}_{i=1}^{n}$. In these networks we distinguish among edges that belong, or not, to the differential subnetworks

and consequently sample their weights from different distributions.

Let us first present the procedure used to define each subnetwork S_k, $k = 1, \dots, K$, where K is the number of differential components that we want to simulate. We will later present the procedure to assign a weight to each edge of such subnetworks.

Definition of the differential network

To ensure the correct construction of a subnetwork, the following constraints are adopted. Since the graph under consideration is undirected, each pair of nodes defining an edge is indistinguishable from its reversed counterpart. Consequently, pairs (u, v) and (v, u), with $u, v \in \{1, \dots, |V|\}$, are considered equivalent and treated as a single edge. Furthermore, self-loops, represented by pairs of the form (u, u), are excluded from the construction process. Let us denote with N the desired number of edges in the subnetwork. The algorithm proceeds as follows:

1. **Initialization.** All nodes are initially assumed to be disconnected. An empty set of edges is represented by a matrix $\boldsymbol{F} \in \mathbb{R}^{2\times N}$, where each column corresponds to one of the two endpoints of an edge.

2. **Generation of the first edge.** Two distinct nodes $u, v \in \{1, \dots, |V|\}$ are selected at random and stored as the first column of $\boldsymbol{F}$.

3. **Generation of the second edge.** One node from the first edge is chosen at random, while the second node is drawn from the remaining nodes, subject to the previously defined constraints.

4. **Generation of the remaining edges.** The remaining $N - 2$ edges are generated iteratively following the same principle. At each step, a node already belonging to the partially constructed subnetwork is selected, and a candidate node from the graph is chosen according to the preliminary constraints. As more edges are added, the set of eligible nodes

expands. The process terminates once N edges have been generated, yielding a subnetwork of the original one.

Random sampling of the edge weights

Let us assume that the n networks are divided in two groups, that is, $Q = 2$. Suppose that the algorithm described in the previous section is used to generate K differential subnetworks between the two groups, namely $S_k = (V_k, E_k^{\mathrm{S}})$ where $V_k \subseteq V$ is the set of nodes and E_k^{S} is the set of edges, for $k = 1, \ldots, K$. The weights of the edges belonging to S_k are sampled from a normal distribution with subnetwork-specific mean μ_k^S and unit variance, that is: given $i \in \{1, \ldots, n\}$ if $a_{u,v}^{(i)} \in E_k^{\mathrm{S}}$ then

$$a_{u,v}^{(i)} \sim \mathcal{N}(\mu_k^S, 1).$$

Instead, the weights of those edges that do not belong to any subnetwork are drawn from a normal distribution with unit variance: given $a_{u,v}^{(i)} \notin E_k^{\mathrm{S}}$, $a_{u,v}^{(i)} \sim \mathcal{N}(\mu_0, 1)$. The mean μ_0 is itself a realization of a normal distribution with zero mean and unit variance, $\mu_0 \sim \mathcal{N}(0, 1)$, such as to allow for inter-subject variability. This setting works well when each subnetwork is differential only between two groups of subjects. However, when more than two groups are present, things become more complicated.

Let us suppose to have $Q = 3$ groups of subjects, namely $\mathcal{G}_1, \mathcal{G}_2$ and $\mathcal{G}_3$ and $K = 1$ differential subnetwork. This subnetwork may exhibit differential behaviour between any pair of subject groups. It could be differential between $\mathcal{G}_1$ and $\mathcal{G}_2$, $\mathcal{G}_1$ and $\mathcal{G}_3$, and $\mathcal{G}_2$ and $\mathcal{G}_3$. Alternatively, it may only be differential between a specific pair of groups, such as $\mathcal{G}_1$ and $\mathcal{G}_2$.

Thus, it seems necessary to model the edge weights not only according to the subnetworks they belong to, but also according to the groups of subjects they refer to. Let us consider a specific subject i, for $i = 1, \ldots, n$, and define the weight for an edge (u, v) as

$$a_{u,v}^{(i)} = \sum_{k=0}^{K} \sum_{c=1}^{Q} X_{i,c} w_c^k \mathbb{1}_{E_k^S}((u,v)), \tag{5.10}$$

where $\forall c = 1, \ldots, Q$,

$$w_c^0 \sim \mathcal{N}(\mu_0, 1) \quad \text{and} \quad w_c^k \sim \mathcal{N}(\mu_{k,c}^S, 1).$$

The element $X_{i,c}$ simply encodes whether the network of subject i belongs to the c-th group, allowing to modulate edge weights according to the group of interest. The indicator function $\mathbb{1}_{E_k^S}((u,v))$ is defined as $\mathbb{1}_{E_k^S}((u,v)) = 1$ if $(u,v) \in E_k^S$ or zero otherwise and describes whether the edge (u,v) belongs to the k-th differential subnetwork. The edges that do not belong to any subnetwork, namely for $k = 0$, are considered baseline edges. Regardless of the group c, they are always sampled from the same distribution. This general expression allows for an edge to be in more than one differential subnetworks and for a subject to reflect the influence of more than one covariate of interest. However, for simplicity in our simulation study we consider K non-overlapping subnetworks and Q distinct groups of subjects. In particular, non-overlapping subnetworks imply that for any edge (u,v), $\mathbb{1}_{E_k^S}((u,v)) = 1$ for at most one k, so that the sum over k reduces to a single active term. Similarly, distinct groups imply that for each subject i, $X_{i,c} = 1$ for exactly one c.

We now describe two different configurations used in our simulation study to evaluate the Tukey–Kramer test and a third one to compare the classical NBS pipeline to the Tukey–Kramer and the Bonferroni-corrected t-test. Before that, let us recall that a description of the Tukey–Kramer and the Bonferroni-corrected t-test was given in Section 5.1. Moreover, we note that when performing the Bonferroni-corrected t-test, the correction is performed at the edge level, so that the null hypothesis of equal connectivity between a pair of groups $H_0 : \mu_k = \mu_{k'}$ for $k \neq k', k, k' = 1, \ldots, Q$ with Q the number of groups, is rejected if and only if

$$\hat{t} > t_{\frac{\alpha}{Q(Q-1)}; n_k + n_{k'} - 2} \tag{5.11}$$

where n_k and $n_{k'}$ are the sample size for the k and k' groups, respectively, and $\hat{t}$ is the sample value of the t-test statistic. We denote with $t_{\frac{\alpha}{Q(Q-1)};n_k+n_{k'}-2}$ the upper $\frac{\alpha}{Q(Q-1)}$ percentage point of a Student's t distribution with $n_k + n_{k'} - 2$ degrees of freedom comprehensive of the Bonferroni correction.

For all the experiments, we set the overall level of significance of the permutation test equal to $\alpha = 0.05$ and the level of significance of the single univariate test equal to $\beta = 0.001$. We specify case by case the value of the threshold t_s related to the level of significance of the single univariate test β later on. The number of permutations in the NBS framework is set to $J = 5000$ and we used the extent size to compute the size of the connected components.

Configuration 1

In this first simulation setting, we mimic a three-groups scenario with a single differential subnetwork S, namely $Q = 3$ and $K = 1$. The three groups are denoted as $\mathcal{G}_1, \mathcal{G}_2$ and $\mathcal{G}_3$. We consider an overall population of $n = 63$, equally divided among the three groups, namely $n_q = 21$ for $q = 1, 2, 3$. As mentioned earlier, each subject belongs only to one group.

For this simulation, we assume the subnetwork to be differential between the first and the second groups, and between the first and the third groups. The first group thus operates as the control group: for each subject in the group and for all the edges, the edge weights are sampled from a normal distribution of zero mean and unit variance. Indeed, given a subject $i \in \mathcal{G}_1$, $\forall(u, v)$, the weight of the edge (u, v) reduces to $a_{u,v}^{(i)} \sim \mathcal{N}(\mu_0, 1)$ and $\mu_0 \sim \mathcal{N}(0, 1)$.

The other two groups present a common differential subnetwork S of 25 edges, namely $|E^S| = 25$. All the edges outside the subnetwork are sampled as for the first group:

$$\forall i \in \mathcal{G}_2 \cup \mathcal{G}_3, \forall(u, v) \notin E^S, a_{u,v}^{(i)} \sim \mathcal{N}(\mu_0, 1).$$

Instead, the edges within the subnetwork are sampled as follows:

1. $\forall i \in \mathcal{G}_2$ and $\forall (u,v) \in E^S$, $a_{u,v}^{(i)} \sim \mathcal{N}(\mu_2^S, 1)$, where $\mu_2^S = 3$;
2. $\forall i \in \mathcal{G}_3$ and $\forall (u,v) \in E^S$, $a_{u,v}^{(i)} \sim \mathcal{N}(\mu_3^S, 1)$, where $\mu_3^S = 3.2$;

Figure 5.1 shows the generated subnetwork S.

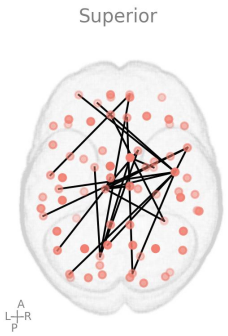

Figure 5.1: The generated subnetwork S for Configuration 1.

Configuration 2

For this simulation study, we consider $Q = 4$ groups and $K = 3$ differential subnetworks, namely S_1, S_2, S_3, each associated to a specific group. We consider an overall population of $n = 84$, equally divided among the four groups, namely $n_q = 21$ for $q = 1, \ldots, 4$.

As for the first simulation case, the first group operates as the control group, so that $\forall i \in \mathcal{G}_1$, $\forall (u,v)$, the weight of the edge (u,v) reduces to $a_{u,v}^{(i)} \sim \mathcal{N}(\mu_0, 1)$ and $\mu_0 \sim \mathcal{N}(0,1)$.

All the differential subnetworks comprise 25 edges, namely $|E_k^S| = 25, k = 1,2,3$. All the edges outside the subnetwork are sampled as for the first group. The edges of the subnetworks are sampled as follows:

1. $\forall i \in \mathcal{G}_2$ and $\forall (u,v) \in E_1^S$, $a_{u,v}^{(i)} \sim \mathcal{N}(\mu_{1,2}^S, 1)$, where $\mu_{1,2}^S = 3$;
2. $\forall i \in \mathcal{G}_3$ and $\forall (u,v) \in E_2^S$, $a_{u,v}^{(i)} \sim \mathcal{N}(\mu_{2,3}^S, 1)$, where $\mu_{2,3}^S = 3.2$;
3. $\forall i \in \mathcal{G}_4$ and $\forall (u,v) \in E_3^S$, $a_{u,v}^{(i)} \sim \mathcal{N}(\mu_{3,4}^S, 1)$, where $\mu_{3,4}^S = 3.4$.

Figure 5.2 shows the generated subnetworks S_1, S_2 and S_3.

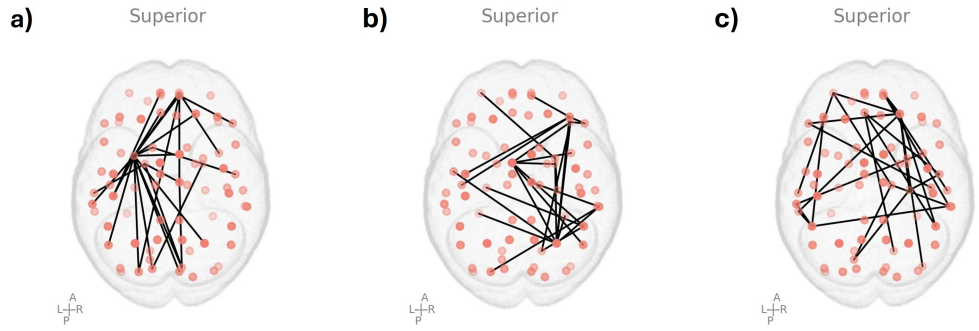

Figure 5.2: Generated subnetworks for Configuration 2: a) shows the differential subnetwork S_1; b) shows the differential subnetwork S_2 and c) shows the differential subnetwork S_3.

Comparison between Tukey–Kramer and Bonferroni-corrected pairwise t-test

For evaluating the proposed alternatives to the classical NBS pipeline, we used simulated data mimicking functional brain networks estimated across the recordings of a standard EEG helmet comprising $|V| = 61$ sensors. We assume $Q = 4$ balanced groups exist, each one including $n_q = 20$ graphs, $q = 1, \ldots, 4$. We simulated a scenario where the four groups present differences on a single subnetwork S of 50 edges, namely $|E^S| = 50$. For all $(u, v) \in S$, for all $i \in \mathcal{G}_q$, with $q = 1, \ldots, 4$, $a^{(i)}_{u,v}$ was randomly drawn from a folded normal distribution with mean μ^S_q and variance 1, that is $a^{(i)}_{u,v} \sim \mathcal{N}(\mu^S_q, 1)$. The remaining elements $(u, v) \notin S$ were drawn from a folded normal distribution of variance 1 and mean sampled from a standard normal distribution, $a^{(i)}_{u,v} \sim \mathcal{N}(\mu_0, 1)$ with $\mu_0 \sim \mathcal{N}(0, 1)$, thereby serving as sample-specific noise.

In a first experiment, we set $\mu^S_1 = 2.5$, $\mu^S_2 = 0$, $\mu^S_3 = \mu^S_4 = 1.1$. Hence $\mathcal{G}_1$ shows an increased connectivity in S with respect to all the other groups. Furthermore, $\mathcal{G}_2$ shows an impaired connectivity with respect to $\mathcal{G}_3$ and $\mathcal{G}_4$.

In a second experiment, we kept μ^S_1, μ^S_3 and μ^S_4 unchanged while we set μ^S_2 to 26 values equally spaced between 0 and 5. For each mean value, the test was repeated 20 times. Finally, the average values of true positive rate (TPR) was computed as $TPR = \frac{TP}{P}$, where P is the number of edges belonging to the

true differential subnetwork S and TP is the number of edges in S detected as significant.

5.5.2 Results

In this section, we present results for the Configurations 1 and 2 described in Section 5.5.1. For both configurations, we first present results for the F-test, which gives information on the presence of differential connected components beetween groups. We then show results obtained when applying the post-hoc Tukey–Kramer test. This pipeline represents a standard procedure when testing differences among multiple groups [95].

We then compare the results from the Tukey–Kramer test with those obtained from the t-test with Bonferroni-correction.

Configuration 1

To test the null hypothesis of equal connectivity among the three groups, we applied the F-test within the NBS framework. We set the threshold $t_s = 7.76$. Figure 5.3(a) shows the identified significant connected component ($p = 0.006$), comprising 25 edges. This result indicates the presence of a significantly differential connected component between at least two groups.

We then applied the Tukey–Kramer test to determine which group pairs drove this effect. We set the threshold $t_s = 4.89$. The result showed a significant component between the first and the second groups ($p = 0.007$), and between the first and the third groups ($p = 0.008$). Results are shown in Figures 5.3(b) and 5.3(c), respectively. In both cases, the identified differential connected component comprised 25 edges.

We compared the component identified from the F-test and the Tukey–Kramer tests with the generated subnetwork, confirming that in all three cases all the reconstructed edges were present in the original subnetworks, indicating that they all are true positives. As expected, no significant difference was observed between the second and the third groups.

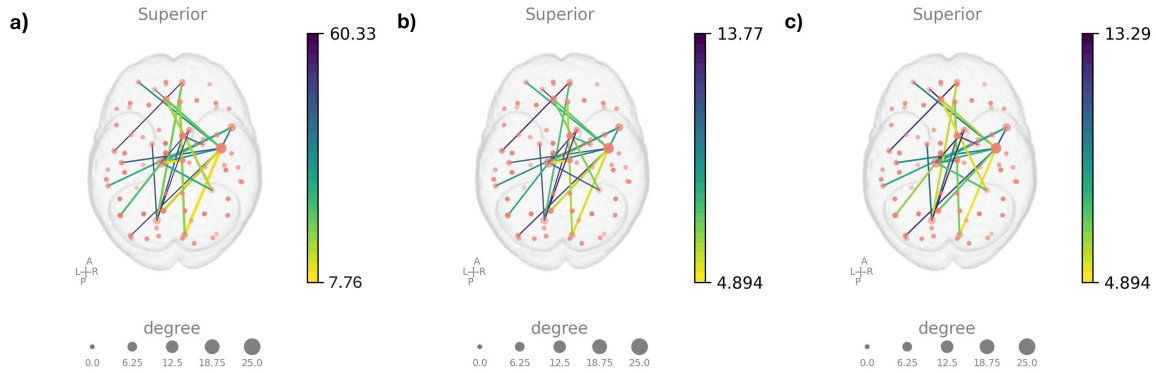

Figure 5.3: Differential connected component identified using network-based statistics (NBS). (a) Significant connected component detected by the F-test across the three groups. (b) Significant connected component identified by the Tukey–Kramer test comparing the first and second groups. (c) Significant connected component identified by the Tukey–Kramer test comparing the first and third groups.

Configuration 2

As for Configuration 2, we first tested the null hypothesis of equal connectivity among the four groups using an F-test. We set the threshold $t_s = 5.97$. This analysis returned a single significant connected component ($p = 0.0006$), comprising 76 edges. We compared this component with the three generated differential subnetworks. The identified connected component corresponds to their union, with one additional edge that represents a false positive.

We then applied the Tukey–Kramer test setting the threshold $t_s = 5.36$. The Tukey–Kramer test returned six significant differential connected components: the comparison between the first and the second groups, between the first and the third groups, between the first and the fourth ones; also the comparisons between the second and third groups, between the second and the fourth groups, and finally between the third and the fourth groups returned significant results. Table 5.1 summarises the significant tests, along with the corresponding p-values and the number of edges identified in the differential connected components. The number of edges identified in each pairwise comparison reflects the underlying simulated subnetworks: comparisons involv-

ing $\mathcal{G}_1$ identify 25 edges corresponding to a single differential subnetwork, while comparisons among $\mathcal{G}_2$, $\mathcal{G}_3$, and $\mathcal{G}_4$ identify 50 edges due to overlapping differential subnetworks in the simulation design. If compared with the edges of the generated differential subnetworks, all the identified edges were present in the original subnetworks. No false positives were identified.

Table 5.1: Significant Tukey–Kramer tests for Configuration 2.

Group comparisons	p-values	n. of edges identified
$\mathcal{G}_1$ vs $\mathcal{G}_2$	0.0004	25
$\mathcal{G}_1$ vs $\mathcal{G}_3$	0.0006	25
$\mathcal{G}_1$ vs $\mathcal{G}_4$	0.0002	25
$\mathcal{G}_2$ vs $\mathcal{G}_3$	0.0004	50
$\mathcal{G}_2$ vs $\mathcal{G}_4$	0.0002	50
$\mathcal{G}_3$ vs $\mathcal{G}_4$	0.0000	50

Figure 5.4 shows the identified differential connected components.

Comparison between Tukey–Kramer and Bonferroni-corrected pirwise t-test

In this section, we compare the results obtained from the Tukey–Kramer post-hoc analysis with the standard Bonferroni-corrected t-test. We applied a threshold $t_s = 3.56$ for the t-test, derived from the one-tail t-test tables when considering $df = 40$ degrees of freedom.

In the first experiment, the classical F-test-based implementation of NBS failed to reject the null hypothesis of equal mean across groups for all the edges. Instead, Table 5.2 summarises the results of the proposed procedures. When Tukey–Kramer is used, six two-sided tests were performed: a significant connected component was correctly identified when testing for differences between $\mathcal{G}_1$ and all other groups, while the approach failed to reject the null hypothesis when $\mathcal{G}_2$ is compared with $\mathcal{G}_3$ and $\mathcal{G}_4$. Conversely, the Bonferroni-corrected pairwise t-test performed 12 one-sided tests, unveiling

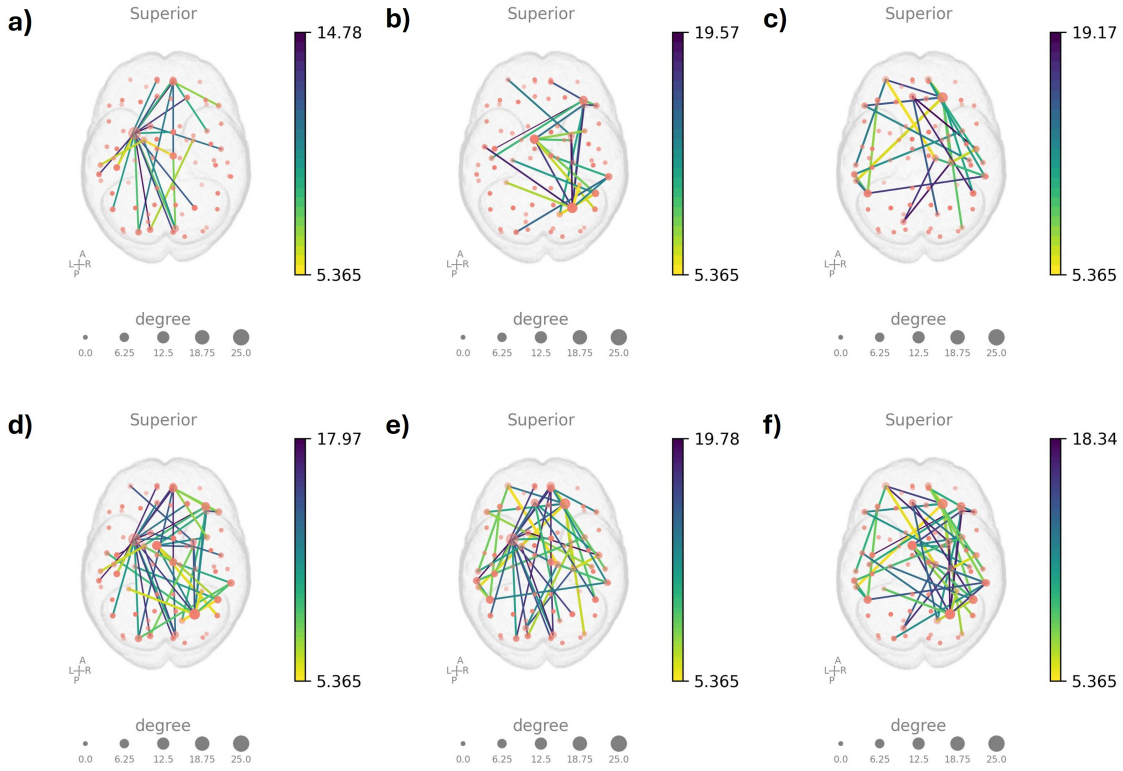

Figure 5.4: Differential connected components identified by the Tukey–Kramer test for Configuration 2. Each panel shows a significant differential connected component corresponding to one of the six pairwise group comparisons: a) comparison between $\mathcal{G}_1$ and $\mathcal{G}_2$; b) comparison between $\mathcal{G}_1$ and $\mathcal{G}_3$; c) comparison between $\mathcal{G}_1$ and $\mathcal{G}_4$; d) comparison between $\mathcal{G}_2$ and $\mathcal{G}_3$; e) comparison between $\mathcal{G}_2$ and $\mathcal{G}_4$; f) comparison between $\mathcal{G}_3$ and $\mathcal{G}_4$.

that the edge weights in $\mathcal{G}_1$ are on average higher than those in the other groups. However, such a procedure was slightly more conservative than the one based on Tukey–Kramer as the extent of the identified connected components was lower. Interestingly, the t-test-based NBS also recognized some differences between $\mathcal{G}_2$ and the other groups although in a single edge and with a p-value higher than 0.02.

These differences between the two proposed procedures were confirmed by the second experiment, where we recall we kept μ_1^S, μ_3^S and μ_4^S unchanged while we set μ_2^S to 26 values equally spaced between 0 and 5. The NBS with

Table 5.2: Results provided by NBS using a TukeyKramer test (left) and a Bonferroni-corrected pairwise t-test (right). For each pairwise test, no cc indicates that no edge survived the thresholding step, consequently, no connected component was identified. Otherwise, the p-value of the identified connected component is reported, together with its number of edges (in parentheses).

Tukey–Kramer					Bonferroni corrected t-test				
$\neq$	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$>$
$\mathcal{G}_1$	-	<0.001 (49)	<0.001 (39)	<0.001 (39)	-	no cc	no cc	no cc	$\mathcal{G}_1$
$\mathcal{G}_2$	-	-	no cc	no cc	<0.001 (48)	-	0.0460 (1)	0.0208 (1)	$\mathcal{G}_2$
$\mathcal{G}_3$	-	-	-	no cc	<0.001 (14)	no cc	-	no cc	$\mathcal{G}_3$
$\mathcal{G}_4$	-	-	-	-	<0.001 (27)	no cc	no cc	-	$\mathcal{G}_4$

the Bonferroni-corrected t-test correctly rejected the null hypothesis of no-difference between $\mathcal{G}_1$ and $\mathcal{G}_2$ in a larger number of cases than the NBS with Tukey–Kramer. However, the estimated connected components were usually smaller, resulting in a lower true positive rate, especially when $\mu_2^S - \mu_1^S \in (-2, 2)$.

5.5.3 Discussion

Given a set of three or more groups of weighted graphs, in this section we investigated the use of the post-hoc Tukey–Kramer test and the Bonferroni-corrected t-test in the NBS framework. After the F-test, both the Tukey–Kramer test and the Bonferroni-corrected t-test can be used to identify which specific groups show significant differences in connectivity. We showed that the Tukey–Kramer test is able to correctly identify the differential subnetworks in the two simulated configurations.

Moreover, we compared the Tukey–Kramer test and the Bonferroni-corrected pairwise t-test. These tests were chosen as they allow to control the family-

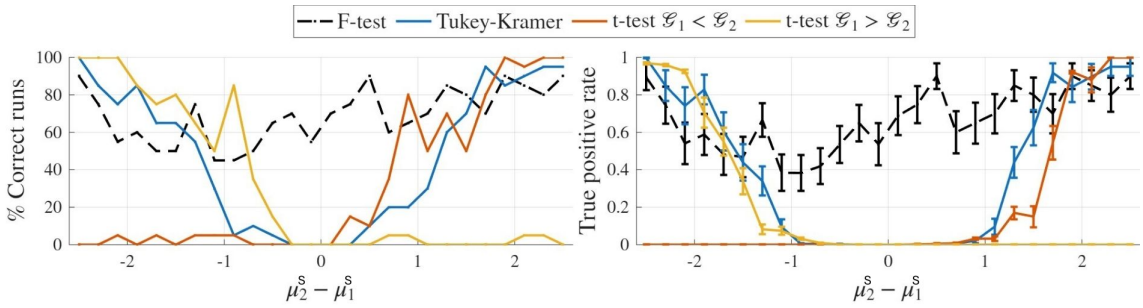

Figure 5.5: Sensitivity of the proposed approach as a function of the difference between the mean μ_1^S and μ_2^S. Left: percentage over 20 repetitions in which the methods identified a differential network; Right: portion of edges in S correctly identified (mean and standard error of the mean over the 20 repetitions). Results from the standard NBS implementation are shown as reference.

wise error of the edge-level univariate tests when comparing multiple pairs of groups. By allowing for one-sided alternative hypothesis, the t-test provides in principle more information. However, it sometimes yields more conservative results and thus shows lower sensitivity when the difference between the group-specific mean weight is small.

In the future, we plan to analyse the impact of input parameters, such as the significance level α within the mass-univariate test in NBS, will be studied. Furthermore, future effort will be devoted to a deeper comparison between the two approaches using both simulated and real EEG data.

Chapter 6

Network-based statistics in Lewy Body Dementia: effects of RBD and visual hallucinations

In this chapter, we present an application of the network-based statistics (NBS) method described in Chapter 5. In particular, we investigate how functional brain connectivity differs in subjects affected by Lewy Body Dementia (DLB) according to the presence of specific core clinical features. Dementia with Lewy Bodies represents the second most common cause of neurodegenerative dementia after Alzheimer's Disease (AD). It is an age-related form of dementia presenting the formation of Lewy bodies in both brain-stem nuclei, and in paralimbic and neocortical regions [112]. Indeed, DLB patients are characterised by specific core features reflecting the involvement of both cortical and sub-cortical structures, such as cognitive fluctuations, visual hallucinations, parkinsonism and REM sleep behaviour disorder (RBD) [112]. Some of these features, particularly RBD, can emerge years before the onset of dementia [113, 114]. Indeed, RBD is now considered a prodromal stage of alpha-synucleinopathies, including DLB [115, 116].

The early cortical dysfunction characterising DLB patients can be efficiently assessed using EEG, especially when advanced functional connectivity

metrics are applied to high-density EEG (hdEEG) data. Overall, network connectivity is typically reduced in DLB compared to healthy controls, with some possible increase specifically in the frontal-temporal network [117]. However, characterising differences in connectivity between DLB patients with different clinical features as well as between DLB and other neurodegenerative diseases is still an open problem [118, 119]. Most literature data focused on the relationship between functional connectivity metrics and cognitive functions and/or visual hallucinations [117]. Recent neurophysiological studies have provided initial empirical evidence for the possibility of modelling functional networks related to visual hallucinations in DLB [120, 121]. Nevertheless, other core features of DLB may also be associated with cortical dysfunctions that manifest as altered functional connectivity metrics.

For example, modifications in alpha phase synchronisation and delta amplitude correlation may represent an electrophysiological signature of cortical compensatory mechanisms in idiopathic-RBD patients. This suggests that large-scale functional networks could serve as early biomarkers for alpha-synucleinopathies [122]. Moreover, EEG-based machine-learning models have successfully identified idiopathic-RBD patients at high risk of short-term phenoconversion and tracked their longitudinal trajectories. Indeed, patients exhibiting more altered connectivity metrics are more likely developing DLB [123]. However, there is a lack of comprehensive literature assessing the relationship between functional brain connectivity dysfunctions and all core features of DLB.

To fill these gaps of the currently available literature on DLB, in this study [4] we explore the relationship between each core clinical feature of DLB (visual hallucinations, cognitive fluctuations, RBD, and parkinsonism) and abnormalities in functional brain connectivity estimated through hdEEG. We propose to compute functional brain connectivity metrics in the individual frequency bands, which can be more accurate than standard frequency bands in describing brain rhythms, as we describe in the next section. The final aim of the present study is to provide new insights on the brain mechanisms

underlying the clinical heterogeneity of DLB.

6.1 Definition of individual frequency bands

In EEG signal analysis, standard frequency bands, such as delta ([1,4]) Hz), theta ([4,8] Hz), alpha ([8,13] Hz) and beta ([13, 30] Hz) bands, are commonly used to investigate brain activity in both healthy and pathological conditions. When dealing with healthy subjects, these bands typically succeed in capturing the corresponding brain rhythms although they may vary considerably according to factors such as age, memory performance, brain volume and task demands [124, 125]. Instead, these standard frequency ranges may be inaccurately applied to subjects affected by neurodegenerative disorders. For instance, the dominant posterior rhythm, which typically falls within the standard alpha band for healthy subjects, may shift to lower frequencies in subjects with DLB [126], risking misclassification within the standard theta band. To address this issue, Klimesh et al. [127] introduced a datadriven approach for computing the individual theta-to-alpha transition frequency (TF), which represents the frequency that discerns the end of the theta band from the beginning of the alpha band. Different approaches have been developed to improve Klimeshs method by leveraging different physiological criteria [127, 128, 129].

In this study, we applied transfreq [129] (source code available at `https://github.com/elisabettavallarino/transfreq`), a Python package that allows for the computation of the theta-to-alpha transition frequency from resting-state data by clustering the spectral profiles associated with EEG channels based on their content in the alpha and theta bands. This method requires only one EEG time-series recorded at rest, while all other available approaches also require an additional EEG time-series recorded while the subject is performing a task inducing alpha desynchronisation.

The algorithm proceeds in an iterative way. First of all, the alpha and theta coefficients are computed by taking the average of the power spectrum over

the corresponding frequency band. Then, a clustering algorithm is used to identify two groups of channels, G_α and G_θ, characterised by low alpha and high theta activity, and by high alpha and low theta activity, respectively. Two spectral profiles are then obtained by averaging the power spectra over the two groups, namely S_α and S_θ. The TF is defined as the intersection of these two profiles. This procedure is repeated until convergence.

After computing the TF, the individual theta and alpha bands were defined as the frequency ranges [TF-2, TF] Hz and [TF, 13] Hz, respectively.

6.2 Participants and experimental protocol

The original data set included patients affected by DLB who underwent hdEEG. Participant recruitment took place at the IRCCS Ospedale Policlinico San Martino in Genova, Italy. Diagnoses were performed by following international guidelines [112], and the mean (sd) time interval between diagnosis and EEG recording was 1.5 month (4 months). All patients are considered to be in the early stage of DLB, as symptom onset was estimated within 1 year before diagnosis. All patients had at least one positive indicative biomarker, specifically presynaptic dopaminergic imaging and/or brain glucose imaging. Furthermore, Mini-mental State Examination (MMSE) and the movement disorder society Unified Parkinsons Disease Rating Scale motor examination (MDS-UPDRS-III) were used to assess global cognition and motor signs, respectively. The presence of visual hallucinations and cognitive fluctuations was inferred from clinical evaluation conducted by an experienced clinician. RBD was diagnosed using overnight polysomnography, serving as an additional indicative biomarker in these cases. Parkinsonism was diagnosed according to current criteria [112] and was defined by the presence of bradykinesia with resting tremor and/or rigidity. In our sample, all subjects with parkinsonism had an MDS-UPDRS-III equal or greater than 6.

For comparison, subjects without neurological or psychiatric disorders (healthy controls, HC) were selected from the lab internal database matched

for age to the DLB patients.

All participants signed an informed consent form in compliance with the Helsinki Declaration of 1975. The study was approved by the local institutional board of IRCCS Ospedale Policlinico San Martino, ethics committee num. N827 of the 24th of May 2023.

6.3 EEG acquisition and pre-processing

Recording sessions were performed at the IRCCS Ospedale Policlinico San Martino using a standard high-density EEG cap with 64 electrodes. For both DLB patients and HC, recordings lasted around $15 - 20$ minutes and took place in a dimly lit and silent room where participants sit, awake, at rest with eyes closed. From each EEG recording, an expert EEG technician visually inspected the EEG and extracted a 56 min time-series by combining artifact-free segments that were not necessarily consecutive. Recordings were digitised at 512 Hz and the EEG cap presented a reference electrode between FZ and AFZ.

Preprocessing of the EEG data was performed using the MNE-Python analysis package [67]. First, a Hamming-windowed finite impulse response (FIR) filter was applied to each EEG recording with a passband from 1 to 40 Hz. Then, time-series were visually inspected, and bad channels were manually removed and then interpolated (number of removed channels: 4 ± 3). Data were re-referenced using average reference [130] and independent component analysis (ICA) [131] was performed to remove noise and artefacts such as eye blinks, heartbeats, and muscular contractions (number of removed components: 9 ± 3).

6.4 Connectivity measures and network-based statistics implementation

To assess functional brain connectivity between pairs of EEG sensor recordings over time, we considered the weighted phase lag index (WPLI) [132] and imaginary part of coherency (IMCOH) [39] metrics, described in Chapter 1.

We recall that the WPLI ranges between 0 and 1, corresponding to lack of connectivity and full connectivity, respectively.

In this work, we use the absolute value of IMCOH, which ranges from 0 (lack of connectivity) to 1 (full connectivity). Hereafter, |IMCOH| denotes the absolute value of IMCOH. For each pair of signals $i, j = 1, \ldots, S$, where S is the number of EEG sensors, the averaged values of $\text{WPLI}_{i,j}(f)$ and $|\text{IMCOH}_{i,j}(f)|$ evaluated at frequency f within the individual theta and alpha bands were computed using routines from the MNE-Connectivity Python package (`https://github.com/mne-tools/mne-connectivity`). In particular, spectrum estimation was first performed using the multitaper method [133] with digital prolate spheroidal sequence (DPSS) windows [134]. Subsequently, each connectivity measure was computed and averaged across frequencies within a given individual frequency band, yielding one connectivity matrix of size $S \times S$ for each subject and each frequency band. Furthermore, for each connectivity matrix, the node degree was defined as the number of edges connected to that node. A proper definition of node degree can be found in Chapter 1. We used the NBS toolbox for statistical group-level analysis, as we describe in Chapter 5. In our work, the number of permutations J was kept fixed across all tests, specifically $J = 5000$, and the selected statistical threshold was set to $ts_{th} = 2.5$. Subnetworks were considered statistically significant if $p < 0.05$ for both WPLI and |IMCOH|. Statistically significant subnetworks were visualized using the `NetPlotBrain` Python package [135].

6.5 Data description and clinical heterogeneity

Our cohort included 33 patients affected by early DLB, and 21 healthy controls (HC), whose main demographic data and clinical scores are summarised in 6.1. All DLB patients had a positive presynaptic dopaminergic imaging as an indicative biomarker and/or a positive brain glucose imaging as a supportive biomarker [112].

Table 6.1: Demographic data and clinical scores for DLB and HC. Continuous variables are shown as mean $\pm$ standard deviation.

	DLB (n=33)	**HC (n=21)**
Age (years)	78 ± 6	71 ± 9
Male / Female	23 / 10	10 / 11
MMSE	22 ± 6	–
MDS-UPDRS-III	18 ± 13	–
Education (years)	9 ± 5	13 ± 4
Dominant Hand (R/L)	28 / 5	–

The ranges of age and education were similar between HC and DLB patients, although a statistically significant difference was detected (Mann-Whitney U test, $p = 0.002$ for age and $p = 0.007$ for education). No significant difference was found for Sex (chi-squared test, $p = 0.18$).

To limit possible cofounding effects due to ongoing medications, we checked whether the subjects included in our analysis were undergoing therapies known to affect the EEG. Within our cohort, only two patients were using low dose benzodiazepines at the time of EEG recording. Due to the limited sample size of our cohort, we decided to also keep these two subjects in our analysis. This choice is justified by the fact that when we reran all the analyses after excluding those two patients the results confirmed the trends discussed throughout the thesis. The UpSet Plot [136] in Figure 6.1 describes the distribution of the core clinical features within the DLB sample. In detail, cognitive fluctuations were present in 9 subjects, RBD in 19 subjects, visual hallucinations in 22 subjects, and parkinsonism was found in 28 subjects. One subject did not present any of the considered core clinical features. However, we de-

cided to keep this subject in our analysis because diagnosticated with DLB due to a polysomnography-confirmed RBD. This core clinical feature was not clinically evident anymore when the subject underwent the neuropsychology and the EEG recording, lightly due to the symptomatic treatment of RBD. As further confirmation that this patient is not an outlier, we repeated all the analysis after removing it and our results were confirmed.

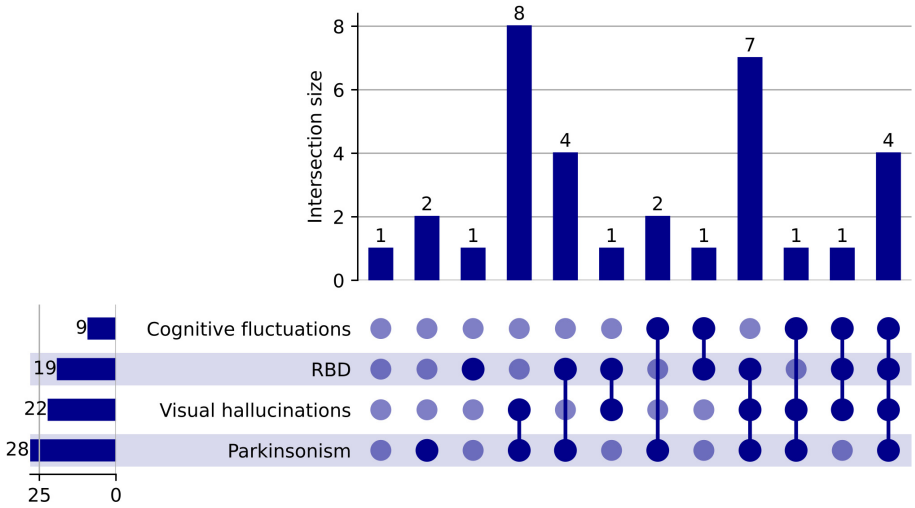

Figure 6.1: Distribution of the core clinical features and of their intersection within the cohort of DLB patients. The bar chart on the left expresses the number of subjects presenting a clinical feature; the bar chart on the top shows the size of each intersection set. Each row corresponds to a possible intersection: the filled-in cells show which set is part of an intersection.

6.6 Results

6.6.1 Transition frequency shows slower dominant posterior rhythm in DLB patients

For each DLB patient and each healthy control we used a modified version of the 2-dimensional *k*-means algorithm, introduced by [129], to group the EEG channels based on the corresponding averaged spectrum in the theta and al-

pha bands. For each subject, we defined two groups of channels, denoted as G_θ and G_α, showing a high value of the averaged spectrum in the theta and in the alpha band, respectively. Topographical maps in Figure 6.2 compare the results obtained by carrying out this clustering procedure using the standard frequency bands with the results obtained by using the individual frequency bands. As we mentioned in Section 6.1, the standard theta and alpha bands are defined as $[4, 8]$ Hz and $[8, 13]$ Hz, respectively, while the corresponding individual bands are defined as $[TF - 2, TF]$ Hz and $[TF, 13]$ Hz, where TF is the theta-to-alpha transition frequency computed through the publicly available Python package transfreq as described in Section 6.1. For DLB subjects, posterior-occipital channels, expected to more accurately reflect the dominant posterior rhythm, were clustered in the G_α group when using the individual band definition and in the G_θ group when applying the standard band definition. This confirms the shift to lower frequencies of the dominant posterior rhythm. In contrast, in the case of healthy controls, both the standard and individual analyses successfully clustered the posterior-occipital channels in the G_α group.

The distribution of the final estimated values of the theta-to-alpha transition TF is shown in Figure 6.3. DBL patients showed a statistically significantly lower TF than HC ($p < 0.001$, Mann–Whitney U test). A statistically significant difference was found also when comparing HC with the subgroups of DLB patients defined by considering the four core clinical features one at the time ($p < 0.005$, Mann–Whitney U test, for all the subgroups). Conversely, no significant differences were observed when comparing, for each clinical feature, the subgroup of patients presenting the feature to those who did not.

By exploiting the individual frequency band defined through the estimated theta-to-alpha TF, we then compared the mean (over EEG sensors) theta and alpha power. In accordance with results from previous studies [137, 138, 139], DLB patients tend to show an increased mean theta power than HC, see Figure 6.4. However, this difference is not statistically significant when the whole cohort of DLB patients is compared to HC ($p = 0.1361$, Mann–Whitney

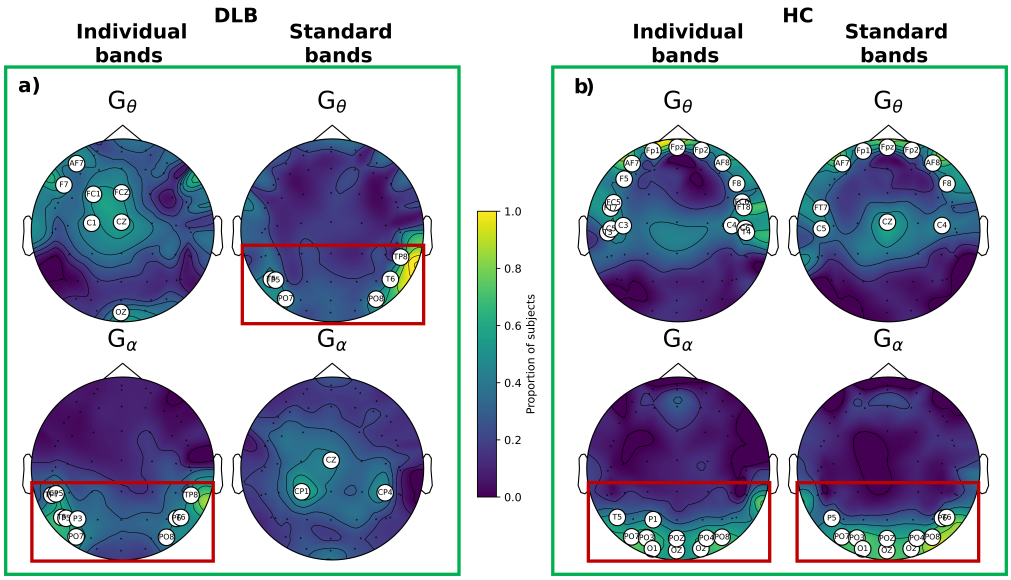

Figure 6.2: Topographical maps visualising, for each EEG channel, the number of subjects for which that channel is grouped into G_θ, as it shows a high content in the theta band (upper row), or into G_α, as it shows a high content in the alpha band (lower row). Panel a) shows the results for DLB patients, while panel b) concerns healthy controls. In each panel, the theta and alpha bands are defined using the individual analysis (left column) and the standard definition (right column). In each map we highlighted the channels for which the number of subjects is greater than 40%.

U test), while a statistically significant difference was found when comparing HC to specific subgroups of DLB patients, namely DLB patients with visual hallucinations ($p = 0.0178$, Mann–Whitney U test) and DLB patients without RBD ($p = 0.0178$, Mann–Whitney U test). When comparing, for each clinical feature, the subgroup of patients presenting that feature with the subgroup of patients who did not, the only statistically significant result was observed when comparing the individual mean theta power in DLB patients with visual hallucinations with respect to DLB patients without visual hallucinations ($p = 0.0410$, Mann–Whitney U test).

No statistically significant differences were found in the mean alpha power when comparing the DLB and HC groups, nor among the different DLB subgroups.

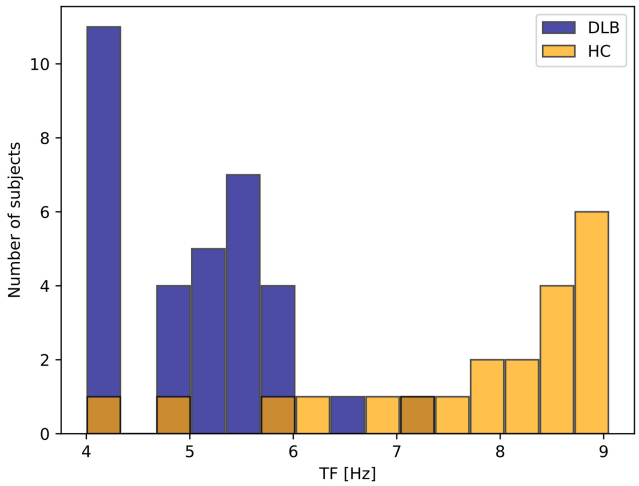

Figure 6.3: Comparison of the distribution of the theta-to-alpha transition frequency (TF) in the DLB patients (blue bars) and in the healthy control (HC, orange bars).

These results underline the importance of using individual frequency bands to study brain rhythms, particularly in subjects affected by neurodegenerative diseases. For this reason, in the remaining of the chapter we investigate connectivity by averaging the values of weighted phase lag index [132] (WPLI) and the absolute value of imaginary part of coherency [39] ($|\mathrm{IMCOH}|$) within the individual theta and alpha bands.

6.6.2 Clinical heterogeneity in DLB patients reflects in hdEEG functional dysconnectivity when compared to HC

We applied the Network Based Statistic (NBS) toolbox [26], described in Chapter 5, to identify functional connectivity associations that are significantly different between subgroups of our cohort. As described in Chapter 5, given two subgroups NBS performs a standard statistical test, such as a t-test, for each pair of EEG sensors in order to test the null hypothesis of equal connectivity between the two subgroups. Then a permutation test is performed to control the family-wise error rate due to multiple comparison.

To infer differences in hdEEG functional connectivity induced by DLB, we

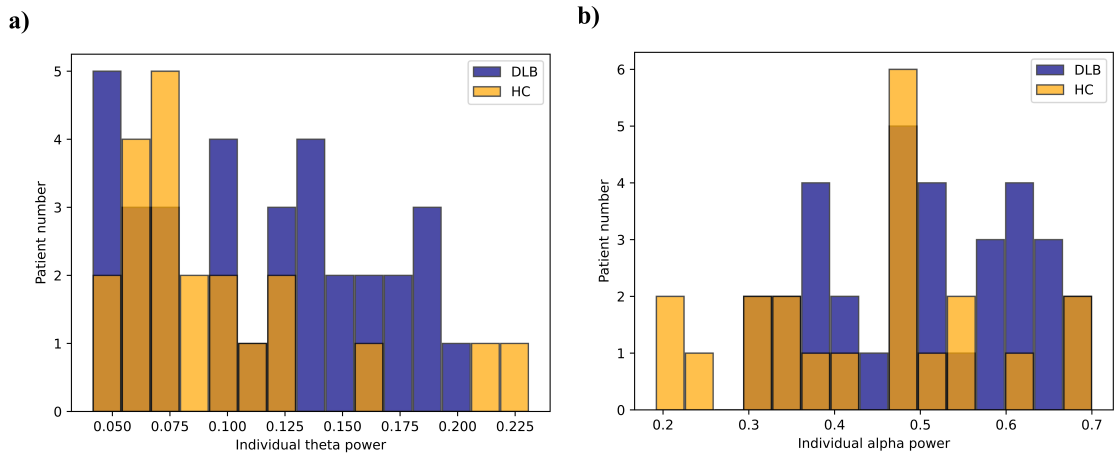

Figure 6.4: Comparison of the distribution of the mean (over EEG sensors) relative power for DLB patients and healthy controls (HC). a) The relative power is computed in the individual theta band. b) The relative power is computed in the individual alpha band. In both panels blue and orange bars represent DLB and HC, respectively.

used NBS to test the null hypothesis of equal connectivity between the cohort of DLB patients and the HC. Furthermore, for each of the four core clinical features assessed here, we divided the DLB patients in two groups: those presenting the feature and those who did not. As above, we used NBS for group-level comparison of the connectivity in each subgroup of DLB patients and the HC. Two groups of tests were performed by running NBS either without nuisance covariates (Unadjusted analysis) and by accounting for age and education (Adjusted analysis).

Table 6.2 shows the tests that returned a statistically significant ($p < 0.05$) differential network for at least one metric between WPLI and |IMCOH|.

We observe that, when no nuisance covariates are considered, the whole cohort of DLB patients as well as the subgroup of patients without visual hallucinations (DLB(Visual hallucinations-)) showed less connectivity compared to the HC cohort when using the averaged value of WPLI in the individual alpha band. Figures 6.5(a) and 6.5(b) show that the corresponding differential networks spread across the whole brain suggesting a general impairment of functional connectivity in DLB patients. However, this result was not confirmed when using |IMCOH| as a connectivity metric.

Table 6.2: Network-based statistics comparing DLB patients with healthy controls (HC).

		Unadjusted analysis		Adjusted analysis	
Band	Comparison	WPLI	|IMCOH|	WPLI	|IMCOH|
Individual α band	DLB $<$ HC	$p = 0.021$	n.s.	0.003	0.016
	DLB (Visual hallucinations$-$) $<$ HC	$p = 0.020$	n.s.	0.009	0.030
	DLB (RBD+) $>$ HC	$p = 0.049$	$p = 0.015$	n.s.	n.s.
	DLB (Visual hallucinations+) $>$ HC	n.s.	$p = 0.036$	n.s.	n.s.
Individual θ band	DLB (Cognitive fluctuations+) $>$ HC	$p = 0.017$	n.s.	0.022	n.s.

On the other hand, DLB patients with RBD (DLB(RBD+)) and with visual hallucinations (DLB(Visual hallucinations+)) separately showed increased values of connectivity in the individual alpha band while DLB patients with cognitive fluctuation (DLB(Cognitive fluctuations+)) showed an enhanced connectivity in the individual theta-band. Notably, the only result confirmed by both the considered connectivity metrics is the one concerning DLB patients with RBD. The corresponding differential networks are shown in Figure 6.5(c) for WPLI and in Figure 6.5(d) for |IMCOH|. Both networks showed connectivity from the right posterior regions to the contralateral fronto-temporal areas. However, WPLI seems to be more conservative than |IMCOH| as a lower number of nodes were involved in the corresponding network.

For the sake of conciseness, the differential networks for DLB patients with visual hallucination and with cognitive fluctuations, which are statistically significant for only one connectivity metric, are shown in Figure 6.6(a) and 6.6(b), respectively. The presence/absence of parkinsonism did not significantly impact the functional connectivity metrics in our population.

To test the effects on our findings of the difference in age between HC and DLB patients, we repeated the tests showed in Table 6.2 after adding the variable age as nuisance covariate within NBS. As shown in the last two columns of Table 6.2, when accounting for age and education related effects, the reduction in alpha-band connectivity for the whole cohort of DLB patients and for the subgroup that do not have visual hallucinations was confirmed, and a

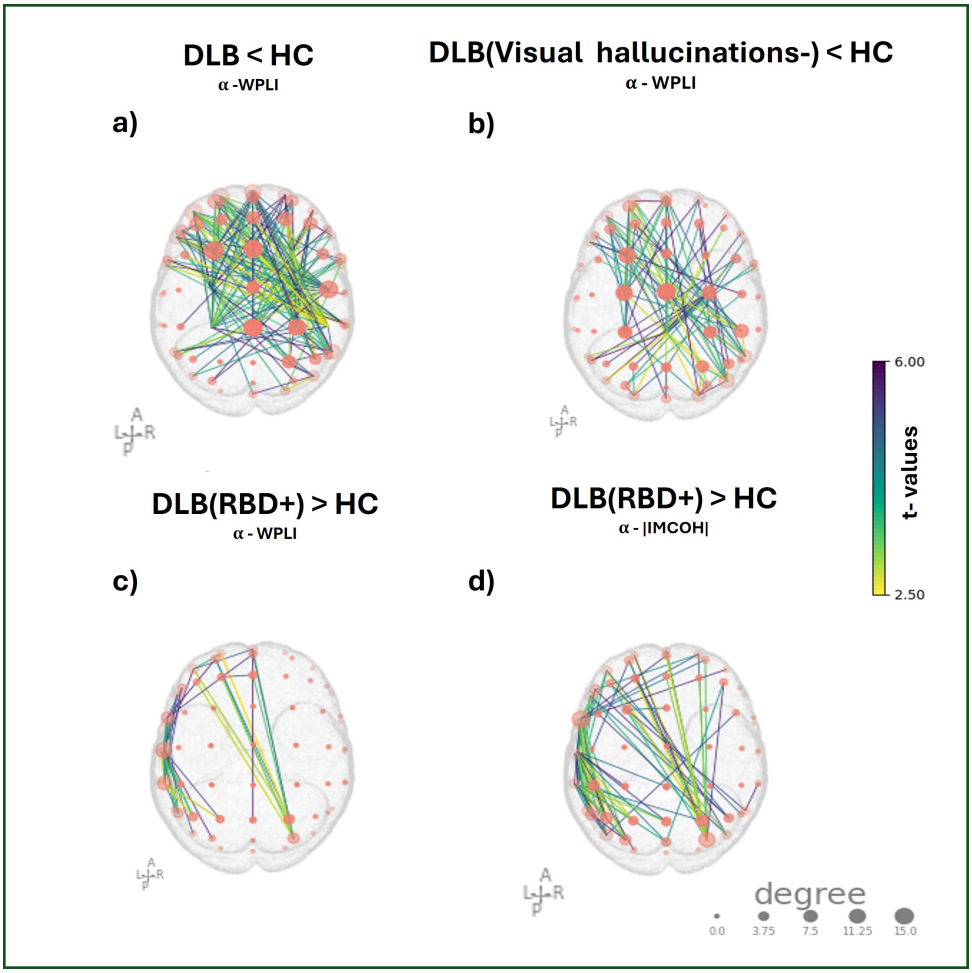

Figure 6.5: Statistically significant differential networks provided by NBS when comparing connectivity matrices of DLB patients and of HC. a) Results of testing for a diminished value of WPLI within the individual alpha band in the whole cohort of DLB patients than in HC. b) Results of testing for a diminished value of WPLI in the individual alpha band of DLB patients without visual hallucinations (DLB(Visual hallucinations-)) than in HC. c)–d) Results of testing for increased value of WPLI and $|\mathrm{IMCOH}|$, respectively, in the individual alpha band from DLB patients with RBD (DLB(RBD+)) than in HC. In each panel, the edge colours represent the t-values provided by the first step of NBS while each node size is proportional to its degree.

statistically significant differential network was found using both WPLI and $|\mathrm{IMCOH}|$. Conversely, the evidence of an increased alpha-band connectivity in DLB patients with RBD and with visual hallucinations were not confirmed.

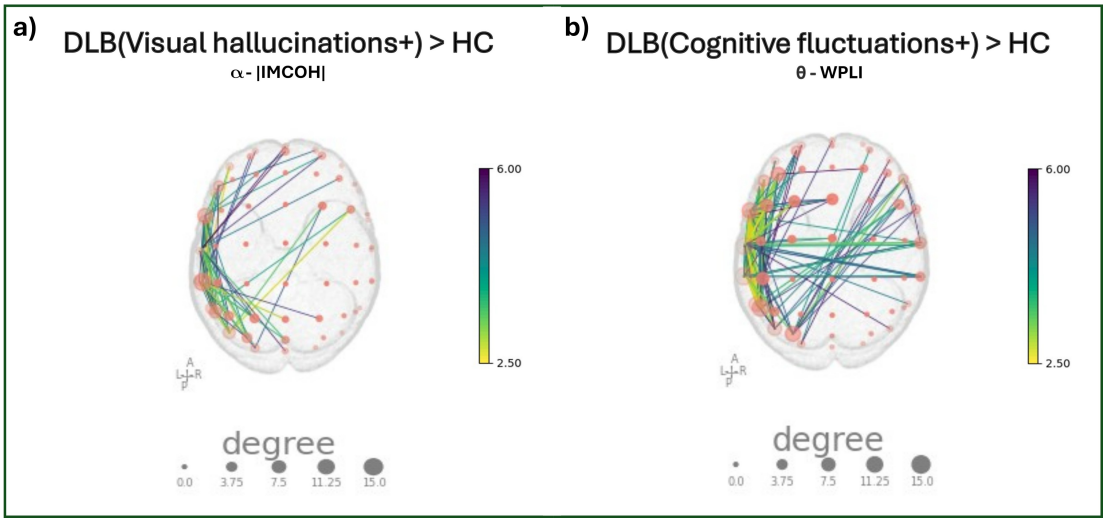

Figure 6.6: a) Statistically significant differential network provided by NBS when testing for an increased value of |IMCOH| within the individual alpha-band from DLB patients with visual hallucinations (DLB(Visual hallucinations+)) than in HC; b) Statistically significant differential network provided by NBS when testing for an increased value of WPLI within the individual theta-band from DLB patients with cognitive fluctuations (DLB(Cognitive fluctuations+)) than in HC. As in Figure 3, edge colours represent the t-values provided by the first step of NBS while each node size is proportional to its degree.

DLB patients with cognitive fluctuations presented higher levels of WPLI in the individual theta band than HC also when age and education are considered as nuisance covariates.

Table 6.3: Network-based statistics comparing DLB patients with healthy controls (HC) when adjusting for age covariate.

Band	Comparison	WPLI	|IMCOH|
$\boldsymbol{\alpha}$	DLB < HC	$p = 0.003$	$p = 0.017$
	DLB (Visual hallucinations−) < HC	$p = 0.011$	$p = 0.028$
	DLB (RBD+) > HC	n.s.	n.s.
	DLB (Visual hallucinations+) > HC	n.s.	n.s.
$\boldsymbol{\theta}$	DLB (Cognitive fluctuations+) > HC	$p = 0.021$	n.s.

6.6.3 RBD and visual hallucinations increase functional connectivity in DLB patients

We then investigated the effect of the four core clinical features (RBD, visual hallucinations, cognitive fluctuations, and parkinsonism) on the functional connectivity metrics of DLB patients. For all clinical features but RBD there were no statistically significant differences in age, sex, and education between the subgroup of patients presenting the feature and the subgroup of all other patients. For this reason, for all clinical features, we used NBS without nuisance variables to test the null hypothesis of equal connectivity between these two subgroups i.e., we performed one test for each core feature using the average value of WPLI and |IMCOH| in the individual theta and alpha bands. For the clinical feature RBD we also ran NBS with age as nuisance covariate and results did not change. NBS did not yield any statistically significant differential network when testing for greater connectivity in patients without the symptom. Instead, Table 6.4 summarises the results obtained when the alternative hypothesis was that connectivity is greater in patients with the symptom compared to those without. Patients with RBD presented a greater connectivity in the individual alpha-band than patients without RBD, while enhanced theta-band connectivity was found in patients with visual hallucinations as compared with those without visual hallucinations. Both results were confirmed using either WPLI or |IMCOH|.

Table 6.4: Network-based statistics comparing DLB subgroups presenting different clinical features.

Comparison	**θ Band**		**α Band**	
	WPLI	**|IMCOH|**	**WPLI**	**|IMCOH|**
DLB (RBD+) > DLB (RBD−)	n.s.	n.s.	$p = 0.042$	$p = 0.020$
DLB (VH+) > DLB (VH−)	$p = 0.017$	$p = 0.026$	n.s.	n.s.

The connected graphs differentiating DLB patients with RBD from those without RBD are shown in Figure 6.7(a) and consisted of 71 connections be-

tween 30 nodes when WPLI is used, while |IMCOH| showed 81 connections between 39 nodes. Furthermore, in both cases the node degree fell in a range between 0 and 15, with higher values for fronto-temporal and posterior nodes. The violin plots in Figure 6.7(b) quantitatively compare the connectivity for the two groups: the value of both WPLI and |IMCOH| was, on average, higher for the DLB patients with RBD (WPLI mean $= 0.39 \pm 0.05$; |IMCOH| mean = 0.09 ± 0.02) than for patients without RBD (WPLI mean = 0.27 ± 0.03; |IMCOH| mean = 0.04 ± 0.01). Figure 6.8(a) shows the connected graphs differentiating DLB patients with visual hallucinations from the other. When WPLI was used, the graphs consisted of 108 connections between 47 nodes, while when |IMCOH| is used, the graph involved 32 connections between 25 nodes. Furthermore, the violin plots in Figure 6.8(b) confirmed that the value of WPLI and |IMCOH| is, on average, higher for patients with visual hallucinations (WPLI mean $= 0.31 \pm 0.02$; |IMCOH| mean $= 0.06 \pm 0.01$) than for patients without hallucinations (WPLI mean $= 0.24 \pm 0.01$; |IMCOH| mean $= 0.03 \pm 0.006$).

The presence/absence of parkinsonism and cognitive fluctuations did not significantly affect functional connectivity metrics in our population.

6.7 Discussion

For the first time, in this study, we comprehensively investigated the relationship between functional connectivity metrics, assessed through network-based statistics in hdEEG data, and all the core clinical features characterising DLB patients, namely visual hallucinations, RBD, cognitive fluctuations and parkinsonism. We consistently demonstrated, using various metrics and approaches, that the presence of both visual hallucinations and RBD is associated with increased connectivity in early DLB patients, mostly in the left hemisphere. Conversely, DLB patients showed overall reduced functional connectivity compared to healthy controls. DLB patients with RBD and those with visual hallucinations showed an increased connectivity also with respect to HC,

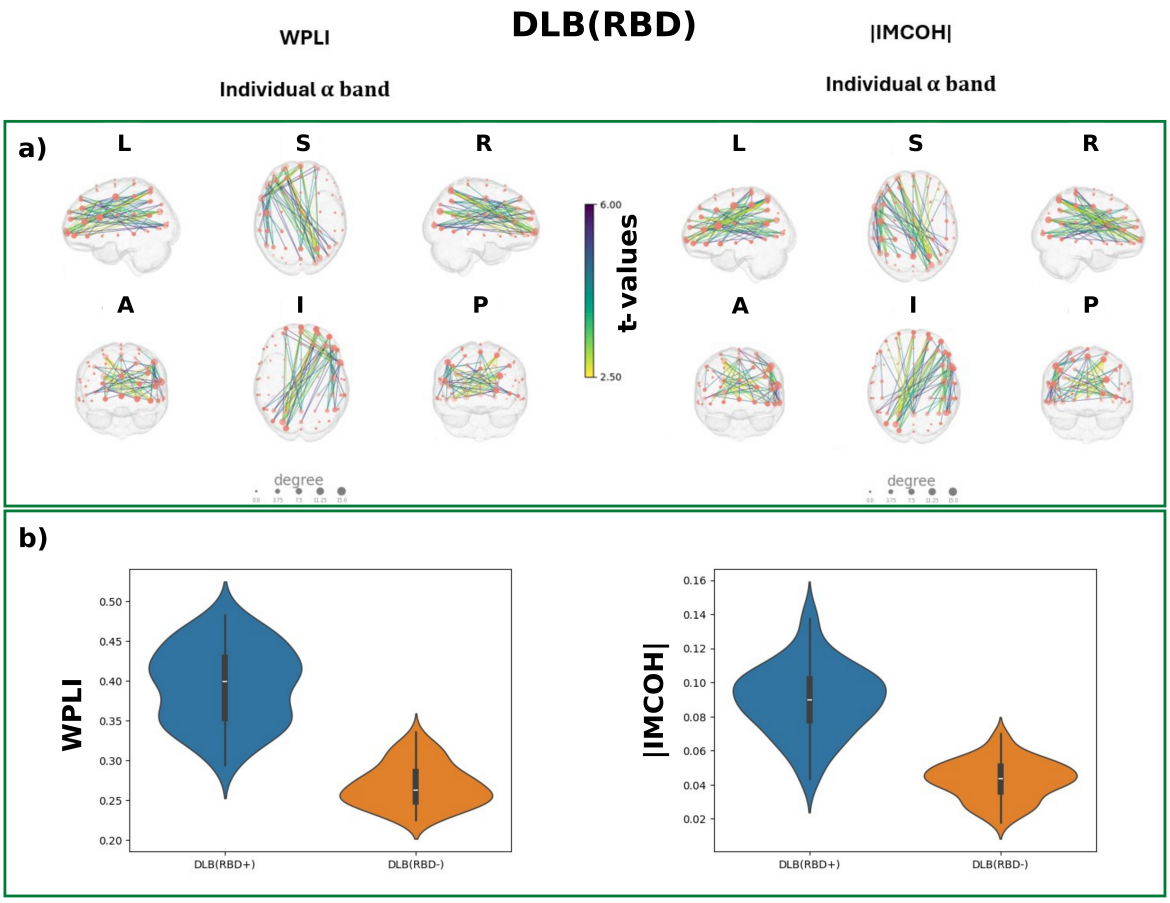

Figure 6.7: Differences in functional connectivity in the individual alpha band between DLB patients with and without RBD. a) Outcome of NBS when testing the alternative hypothesis of a greater connectivity in DLB patients with RBD (DLB(RBD+)) than in DLB patients without RBD (DLB(RBD-)). The colour of the edges represents the t-values obtained from the first univariate test in NBS while the size of each node is proportional to its degree. Six different brain views are shown denoted as L = Left, S = Superior, R = Right, A = Anterior, I = Inferior, P = Posterior. b) Distribution of the connectivity metricss values across subgroups at the edges of the graphs in a). In each panel the first column refers to the results obtained using WPLI while the second column concerns |IMCOH|.

although these results were not confirmed when considering the variables age and education as nuisance covariates within NBS. Future effort should be devoted in discerning the effects attributable to age and education from

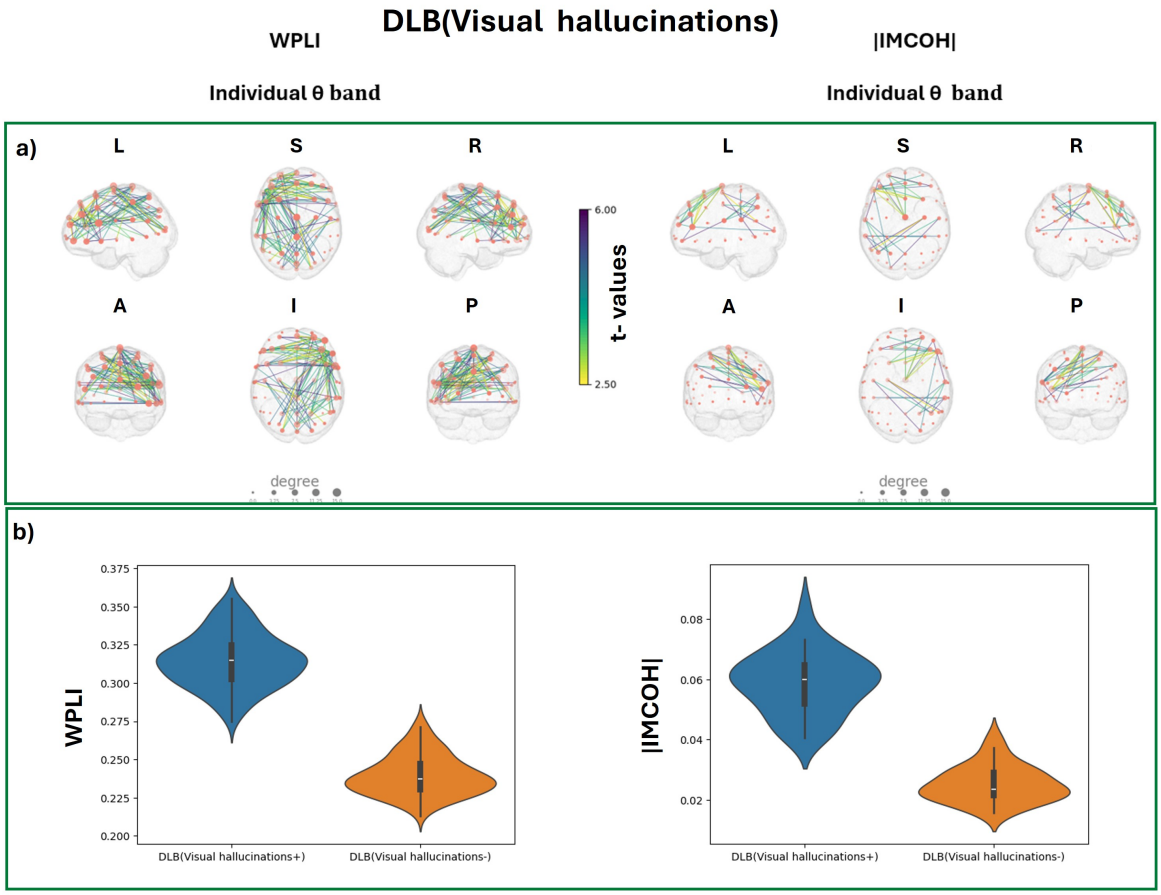

Figure 6.8: Differences in functional connectivity in the individual theta band between DLB patients with and without visual hallucinations. a) Outcome of NBS when testing the alternative hypothesis of a greater connectivity in DLB patients affected by visual hallucinations (DLB(Visual hallucinations+)) than in the other DLB patients (DLB(Visual hallucinations-)). The colour of the edges represents the t-values obtained from the first univariate test in NBS while the size of each node is proportional to its degree. Six different brain views are shown denoted as L = Left, S = Superior, R = Right, A = Anterior, I = Inferior, P = Posterior. b) Distribution of the connectivity metricss values across subgroups at the edges of the graphs in a). In each panel the first column refers to results obtained using WPLI while the second column concerns |IMCOH|.

those related to different stages of disease progression. Cognitive fluctuations and parkinsonism seem to have a non-significant impact on functional con-

nectivity metrics in our cohort, except from the fact that DLB patients with cognitive fluctuations exhibit increased connectivity compared to HC when the averaged value of WPLI over the individual theta band is used as connectivity metric. However, this result may require some additional confirmations due to the limited sample size of the subgroup of DLB patients with cognitive fluctuations, that includes 9 of the 33 subjects, and of the subgroup of DLB patients without parkinsonism, that includes 5 subjects.

Several studies have investigated functional connectivity in DLB patients, mainly using functional MRI and EEG, as well as DTI and FDG-PET [117, 140]. All studies agree that there is a global reduction in connectivity metrics in DLB patients compared to healthy subjects [117, 140] , and we confirm this finding also in our sample. However, literature shows conflicting results regarding the direction of connectivity when different hypothesis-driven approaches are employed, as well as in patients at different stages of the disease [117, 140].

A meaningful drawback of most previous literature is the reliance on standard frequency bands. However, it is known that neurodegenerative diseases affecting the central nervous system, especially DLB, cause a shift toward lower frequencies of both the dominant posterior rhythm and the theta-to-alpha transition frequency [31]. We confirm this finding also in our sample of DLB patients, and therefore we applied the network-based statistic to individual frequency bands, instead of standard ones, likely increasing the reliability of our findings. In fact, the use of standard frequency bands in DLB patients is critical, as it poses the risk of combining subjects with non-homogeneous individual EEG characteristics. For instance, the dominant posterior rhythm may fall within either the standard alpha or theta range, potentially leading to inconsistent results due to the specific characteristics of the study cohort. To note, in the present cohort, the transition frequency consistently fell within the standard theta frequency band, thus in line with existing literature [141]. This finding suggests that using the standard alpha and theta frequency bands may not have captured the individual characteristics of the evaluated subjects.

Furthermore, most literature focuses on specific, a-priori defined networks, such as the default mode network. While this approach is valuable when a strong a-priori hypothesis is made, it may not be the best choice for wider and more innovative aims. In the present study, we chose a whole-brain approach to investigate the unexplored relationship between all four DLB core features and functional connectivity metrics.

Additionally, it is important to emphasise the importance of using individual frequency bands rather than standard ones, also for topographic reasons. Indeed, as can be seen in the Figure 6.2, in DLB patients, the standard alpha frequency band is not expressed in the typical parieto-occipital regions, where instead standard theta is highly represented. This discrepancy has relevant implications when interpreting both network-based and region-based analyses applied on standard frequency bands, as typically shown in the current literature.

Interestingly, in the present study we found that the presence of both visual hallucinations and RBD was associated with increased functional connectivity in the individual theta and alpha bands, respectively. Notably, those findings were confirmed by applying two different functional connectivity measures, namely the weighted phase lag index (WPLI) and the absolute value of imaginary part of coherency (|IMCOH|). The main difference between these two connectivity metrics is that while WPLI is a measure of phase synchronization, the value of |IMCOH| depends on both phase and amplitude correlation [142]. From an electrophysiological point of view, phase synchronisation reflects the communication between coherently oscillating spatially distant neuronal groups [143]. Instead, amplitude correlation in EEG signals has been related to the concurrent activation of different neuronal populations triggered by a sensory stimulus [122].

It has been postulated that brain networks work in a balanced state to maintain efficient information integration as well as their functional state [144, 145] . Any deviation from this balanced state, in either direction, implies an irregularity. Thus, the increased connectivity associated with visual hallucina-

tions and RBD may represent an initial compensatory mechanism in response to the underlying neurodegeneration process. To note, all enrolled DLB patients were in an early disease stage, thus likely reflecting an initial neurodegeneration phase. Supporting this hypothesis, increased functional connectivity has also been found in patients with isolated/idiopathic RBD [122], thus indicating an even earlier alpha-synucleinopathy stage. Conversely, several studies have found reduced connectivity metrics associated with visual hallucinations. However, these studies evaluated more heterogeneous cohorts that included not only DLB patients, but also PD patients with dementia (PDD), who are likely in a different, more advanced stage of the disease. Additionally, these studies used standard frequency bands instead of individual ones [120, 146]. Finally, the presence of increased connectivity as a compensatory mechanism in the early neurodegeneration stages has been suggested also in the field of Alzheimer's disease [147, 148]. Another possible physiopathological interpretation of the increased connectivity is that the repeated presence of abnormal clinical manifestations, such as visual hallucinations and RBD, determine stressful effects on brain networks that, in response, react with increased activity, phenomenon that has been speculated to occur in psychiatric disorders characterised by hallucinations [149].

In the present study, both core features (visual hallucinations and RBD) were associated with increased connectivity across a broad network involving brain regions included in the fronto-parietal network, as well as additional temporal and central brain areas. Interestingly, this increase primarily affected the left (dominant) hemisphere. This is in agreement with existing literature that demonstrates widespread brain network dysfunctions when a whole-brain approach is applied, rather than relying on predefined networks [150, 151, 152], particularly when considering individual rather than standard frequency bands [153]. Furthermore, a previous, large, multicentric brain $[18F]$-FDG-PET study showed that the presence of visual hallucinations and RBD was associated with a relative increase in brain glucose metabolism, especially in the temporal lobe [154], speculating that this may be the result of

a compensatory brain mechanism to the neurodegeneration progression process. Similarly, the increase of the network connectivity metrics found in the present study may reflect as well a compensatory mechanism.

Notably, most previous literature on functional connectivity studies has focused on visual hallucinations, while we found that the presence of RBD significantly affects brain functional connectivity as well. This is in line with existing literature showing that presence of RBD from the prodromal to overt alpha-synucleinopathy continuum is associated with an early cortical involvement [122, 123, 155, 156]. Conversely, neither cognitive fluctuations nor parkinsonism significantly affected the functional connectivity metrics in our sample.

This study has some limitations. The main limitation concerns the relatively small sample size which results in a reduced statistical power, particularly for some of the subgroups analysed. In details, only about 27% of our DLB patients (i.e. 9 over 33 subjects) have cognitive fluctuations, while only about 15% (i.e. 5 over 33 subjects) do not have parkinsonism. Future effort should be devoted to enlarging the cohort of DLB patients in order to confirm the results obtained for the under-represented subgroups. Furthermore, a larger cohort will allow to perform a finer stratification of the patients and hence to investigate whether the associations we found between changes in functional connectivity and the presence of RBD and visual hallucinations are solely due to those features or may be modulated by the concurrent presence of other core clinical features.

Another limitation of this study is its single-modality nature, which means that except for hdEEG no other biomarkers of brain function, such as $[18F]$ FDG-PET or fMRI, were used.

A methodological limitation is the lack of correction for the multiple comparisons problem, arising from the numerous statistical tests performed. We report uncorrected p-values due to the exploratory nature of this study. However, we are aware this approach increases the risk of type I error. Therefore, the findings should be interpreted as preliminary results, warranting confir-

mation in future studies.

Despite these limitations, in this study we carefully characterised our sample based on the biological definition of the disease, by including patients with an abnormal presynaptic dopaminergic imaging. Moreover, all patients underwent overnight polysomnography, which is considered the gold standard for diagnosis of RBD. Furthermore, we used individual frequency bands, computed using a validated semi-automatic tool [129], likely enhancing the consistency and reliability of our results. Finally, this study is the first to investigate the relationship between functional connectivity and all four core clinical features of DLB.

Conclusions

In this thesis, we investigated the study of brain functional connectivity derived from magneto- and electro-electroencephalography (M/EEG) measurements. In particular, we examined two aspects: the estimation of brain functional connectivity networks from M/EEG recordings and statistical inference tools devoted to perform network comparisons.

Motivated by recent results on the standard two-steps procedure for cross-power spectrum estimation from M/EEG data [13, 14], in Chapter 3 we proposed an alternative procedure that directly estimates the cross-power spectrum and controls the number of false positives by exploiting the ℓ_1 regularisation through FISTA. The proposed method leverages the tensorial structure of the global coefficient matrix for efficient estimation. The advantages of our method over the classical two-step approach have been demonstrated on a large number of synthetic data that mimic different brain configurations. From a methodological point of view, we plan to implement an automatic procedure to choose the regularisation parameters. Moreover, advanced variants of FISTA could be employed to further improve the computational efficiency of the proposed method, and thus enable a more systematic investigation of the trade-off between accuracy and computational cost. While the current implementation of the proposed approach provides estimates of the cross-power spectrum at fixed frequencies, future work will be devoted to investigate how to combine the information from multiple frequencies and/or frequencies ranges.

In the context of brain connectivity estimation, another interesting ques-

tion aims to understand the directionality of the connections, which gives information on the potential causal relationships among the cortical brain sources. In Chapter 4, we introduced a novel method to estimate directed brain connectivity from M/EEG data. Our approach exploits SEM to model the relationships between the brain activity and the M/EEG measurements. It naturally leads to a mathematical formulation that can be used to derive the probability distributions of all the variables involved, and consequently exploit a statistical framework for inference. We implemented an EM algorithm within the DAG with NO TEARS method, which directly estimates acyclic graphs in an efficient way. The results showed that this approach is promising even though we need to investigate it further. From a methodological point of view, we plan to implement more complex simulation settings. Specifically, we plan to increase the number of EEG sensors and the number of cortical brain sources, and consequently use a leadfield matrix derived from a realistic conductivity model. A further step will be devoted to the simulation of the EEG recordings and the evaluation of the method from signals rather than from score basis functions. A systematic analysis is required for testing the optimisation procedure, with a specific focus on the choice of the initial point which is critical in EM frameworks. Finally, validation on a real EEG dataset is needed to validate the approach in real-world scenarios.

Because brain connectivity networks are influenced by many external factors, such as age, clinical conditions and neurological disorders, there is an increasing interest in developing statistical tools for comparisons of groups of networks associated with different external factors. The network-based statistics (NBS) is a network specific approach that identifies differential connected components across brain networks, controlling for the family-wise error rate (FWER). The task is achieved by first performing a univariate test for each edge in the network so as to identify differences at edge level. Then, connected components are detected from the obtained significant edges. For each component, a p-value is computed according to the null distribution of the maximal component size approximated through a permutation test. NBS

exploits an F-test for the univariate test, when dealing with more than two groups of networks. In Chapter 5, we discussed and compared two alternatives to the F-test, namely a Tukey–Kramer test and a Bonferroni-corrected pairwise t-test. In Chapter 6, we applied the NBS toolbox on a dataset of subjects affected by Dementia with Lewy Bodies (DLB). We comprehensively investigated the relationship between functional connectivity metrics and the core clinical features characterising DLB patients, namely visual hallucinations, RBD, cognitive fluctuations and parkinsonism. We consistently demonstrated that the presence of both visual hallucinations and RBD is associated with an increased connectivity in early DLB patients, mostly in the left hemisphere. Notably, most previous literature on functional connectivity studies has focused on visual hallucinations, while we found that the presence of RBD significantly affects brain functional connectivity as well. Furthermore, most literature focuses on specific, a-priori defined networks, while in the proposed study, we choose a whole-brain approach to investigate the unexplored relationship between all four DLB core features and functional connectivity metrics. Finally, it is important to emphasize the importance of using individual frequency bands, computed via the transfreq algorithm [129], rather than standard ones. The main limitation of the discussed study relates to the small sample size, that resulting in reduced statistical power. Future effort should be devoted to enlarge the cohort of DLB patients, also allowing finer stratification of the patients. Another limitation of this study is its single-modality nature, which means that except for EEG, no other biomarkers of brain functions, such as FDG-PET or fMRI, were used.

Bibliography

[1] L. Carini, I. Furci, and S. Sommariva. "Sparse optimization for estimating the cross-power spectrum in linear inverse models: from theory to the application in brain connectivity". *Accepted by Journal of Optimization Theory and Applications (JOTA)* (2024).

[2] L. Carini, S. Sommariva, and E. Wit. "Estimating latent directed brain connectivity via structural equation modelling". In preparation. 2026.

[3] L. Carini, M. Lai, and S. Sommariva. "Network-based statistics for hypothesis testing in multiple groups of brain connectivity graphs." Short paper submitted to 53rd SIS Scientific Meeting. 2026.

[4] L. Carini, S. Sommariva, F. Famà, L. Giorgetti, et al. "Network based statistics shows that rem sleep behavior disorder and visual hallucinations increase functional connectivity in early dementia with Lewy bodies". *Submitted to npj Parkinson's Diesease. Under review.* (2025).

[5] A. Mapelli, L. Carini, F. Ieva, and S. Sommariva. "A neighbour selection approach for identifying differential networks in conditional functional graphical models". *Submitted to Biometrical Journal* (2026).

[6] O. Sporns. "Towards network substrates of brain disorders". *Brain* 137.8 (July 2014), 2117–2118.

[7] J. D. Steinmetz et al. "Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021". *The Lancet Neurology* 23.4 (2024), 344–381.

[8] A. Howseman, S. Zeki, A. M. Howseman, and R. W. Bowtel. "Functional magnetic resonance imaging: imaging techniques and contrast mechanisms". *Philosophical Transactions of the Royal Society B: Biological Sciences* 354.1387 (July 1999), 1179–1194.

[9] S. I. Ziegler. "Positron Emission Tomography: Principles, Technology, and Recent Developments". *Nuclear Physics A* 752 (2005), 679–687.

[10] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, et al. "Magnetoencephalography - theory, instrumentation, and applications to noninvasive studies of the working human brain". *Rev. Mod. Phys.* 65 (2 Apr. 1993), 413–497.

[11] E. Niedermeyer and F. L. da Silva. *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. 5th. Lippincott Williams & Wilkins, 2005.

[12] M. S. Hämäläinen and R. J. Ilmoniemi. "Interpreting magnetic fields of the brain: minimum norm estimates". *Med. Biol. Eng. Comput.* 32.1 (1994), 35–42.

[13] E. Vallarino, S. Sommariva, M. Piana, and A. Sorrentino. "On the two-step estimation of the cross-power spectrum for dynamical linear inverse problems". *Inverse Probl.* 36.4 (2020), 045010, 18.

[14] E. Vallarino, A. S. Hincapié, K. Jerbi, R. M. Leahy, et al. "Tuning Minimum-Norm regularization parameters for optimal MEG connectivity estimation". *NeuroImage* 281 (2023), 120356.

[15] A. Ossadtchi, D. Altukhov, and K. Jerbi. "Phase shift invariant imaging of coherent sources (PSIICOS) from MEG data". *NeuroImage* 183 (2018), 950–971.

[16] M. Fukushima, O. Yamashita, T. R. Knösche, and M. Sato. "MEG source reconstruction based on identification of directed source interactions on whole-brain anatomical networks". *NeuroImage* 105 (2015), 408–427.

[17] F. Tronarp, N. P. Subramaniyam, S. Särkkä, and L. Parkkonen. "Tracking of dynamic functional connectivity from MEG data with Kalman filtering". *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018. ISBN: 978-1-5386-3647-3.

[18] A. Beck and M. Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". *SIAM J. Imaging Sci.* 2.1 (2009), 183–202.

[19] S. J. Kiebel, M. I. Garrido, R. J. Moran, and K. J. Friston. "Dynamic causal modelling for EEG and MEG". *Cognitive Neurodynamics* 2.2 (2008), 121–136.

[20] M. J. Kaminski and K. J. Blinowska. "A new method of the description of the information flow in the brain structures". *Biological Cybernetics* 65.3 (1991), 203–210.

[21] L. Astolfi, F. Cincotti, D. Mattia, M. G. Marciani, et al. "Comparison of different cortical connectivity estimators for high-resolution EEG recordings". *Human Brain Mapping* 28.2 (2007), 143–157.

[22] L. Astolfi, F. Cincotti, C. Babiloni, F. Carducci, et al. "Estimation of the cortical connectivity by high-resolution EEG and structural equation modeling: Simulations and application to finger tapping data". *IEEE Transactions on Biomedical Engineering* 52.5 (2005), 757–768.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (2018), 1–22.

[24] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing. "DAGs with NO TEARS: continuous optimization for structure learning". NIPS'18. Curran Associates Inc., 2018, pp. 9492–9503.

[25] A. Fornito, A. Zalesky, and E. T. Bullmore. *Fundamentals of Brain Network Analysis*. 1st ed. Academic Press, 2016.

[26] A. Zalesky, A. Fornito, and E. T. Bullmore. "Network-based statistic: Identifying differences in brain networks". *NeuroImage* 53.4 (2010), 1197–1207.

[27] R. J. Ilmoniemi and J. Sarvas. *Brain Signals: Physics and Mathematics of MEG and EEG*. The Mit Press, 2019.

[28] M. Bear, B. Connors, and M. Paradiso. *Neuroscience: Exploring the Brain*. Lippincott Williams & Wilkins, 2007.

[29] Z. Fountas. *Spiking Neural Networks for Human-like Avatar Control in a Simulated Environment*. 2011.

[30] O. Mecarelli. *Clinical Electroencephalography*. Ed. by O. Mecarelli. Springer Cham, 2019.

[31] C. Babiloni, C. Del Percio, M. Pascarelli, R. Lizio, et al. "Abnormalities of functional cortical source connectivity of resting-state electroencephalographic alpha rhythms are similar in patients with mild cognitive impairment due to Alzheimer's and Lewy body diseases". *Neurobiology of Aging* 77 (2019), 112–127.

[32] P. Garcés, D. López-Sanz, F. Maestú, and E. Pereda. "Choice of Magnetometers and Gradiometers after Signal Space Separation". *Sensors* 17.12 (2017).

[33] C. Babiloni, V. Pizzella, C. D. Gratta, A. Ferretti, et al. "Chapter 5 Fundamentals of Electroencefalography, Magnetoencefalography, and Functional Magnetic Resonance Imaging". Vol. 86. International Review of Neurobiology. Academic Press, 2009, pp. 67–80.

[34] J. Malmivuo. "Introduction". *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields*. Oxford University Press, Oct. 1995.

[35] F.-H. Lin, J. W. Belliveau, A. M. Dale, and M. S. Hämäläinen. "Distributed current estimates using cortical orientation constraints". *Human Brain Mapping* 27.1 (2006), 1–13.

[36] D. B. West. *Introduction to Graph Theory*. 2nd. Vol. 2. Upper Saddle River: Prentice Hall, 2001. ISBN: 0130144002.

[37] J. S. Bendat and A. G. Piersol. *Random data: analysis and measurement procedures*. Vol. 729. John Wiley & Sons, New York, 2011.

[38] E. Pereda, R. Q. Quiroga, and J. Bhattacharya. "Nonlinear multivariate analysis of neurophysiological signals". *Prog. Neurobiol.* 77.1 (2005), 1–37.

[39] G. Nolte, O. Bai, L. Wheaton, Z. Mari, et al. "Identifying true brain interaction from EEG data using the imaginary part of coherency". *Clinical Neurophysiology* 115.10 (2004), 2292–2307.

[40] C. J. Stam, G. Nolte, and A. Daffertshofer. "Phase lag index: Assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources". *Hum. Brain Mapp.* 28.11 (2007), 1178–1193.

[41] M. Demuru, S. M. La Cava, S. M. Pani, and M. Fraschini. "A comparison between power spectral density and network metrics: An EEG study". *Biomedical Signal Processing and Control* 57 (2020), 101760.

[42] J. Hadamard. "Sur les problèmes aux dérivés partielles et leur signification physique". *Princeton University Bulletin* 13 (1902), 49–52.

[43] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.

[44] A. N. Tikhonov. "On the stability of inverse problems". *Proceedings of the USSR Academy of Sciences* 39 (1943), 195–198.

[45] E. Vallarino. "Estimation and interpretation of the cross-power spectrum to quantify brain connectivity from electrophysiological data". PhD thesis. Genoa, Italy: University of Geno, 2022.

[46] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. S. Kazimierski. *Regularization Methods in Banach Spaces*. Vol. 10. Radon Series on Computational and Applied Mathematics. de Gruyter, 2012.

[47] P. L. Combettes and J.-C. Pesquet. "Proximal Splitting Methods in Signal Processing". *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Ed. by H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, et al. Springer New York, 2011, pp. 185–212.

[48] B. Polyak. *Introduction to Optimization*. Translations series in mathematics and engineering. Optimization Software, Publications Division, 1987.

[49] G. Casella and R. L. Berger. *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA, 2002.

[50] H. O. Hartley. "Maximum Likelihood Estimation from Incomplete Data". *Biometrics* 14.2 (1958), 174–194.

[51] R. Sundberg. "Maximum Likelihood Theory for Incomplete Data from an Exponential Family". *Scandinavian Journal of Statistics* 1.2 (1974), 49–58.

[52] C. Rao. *Linear Statistical Inference and Its Applications*. WILEY SERIES in PROBABILITY and STATISTICS: PROBABILITY and STATISTICS SECTION Series. Wiley, 1965.

[53] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. 2. ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley, 2008.

[54] E. Vallarino, A. Sorrentino, M. Piana, and S. Sommariva. "The role of spectral complexity in connectivity estimation". *Axioms* 10.1 (2021), 35.

[55] J.-M. Schoffelen and J. Gross. *Studying dynamic neural interactions with MEG*. In: Magnetoencephalography: From Signals to Dynamic Cortical Networks: Second Edition. Springer International Publishing, 2019, pp. 519–541.

[56] N. Puthanmadam Subramaniyam, F. Tronarp, S. Särkkä, and L. Parkkonen. "Joint estimation of source dynamics and interactions from MEG data". *Network Neuroscience* 9.3 (July 2025), 842–868.

[57] M. A. Figueiredo, R. D. Nowak, and S. J. Wright. "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems". *IEEE J. Sel. Top. Signal Process.* 1.4 (2007), 586–597.

[58] P. Welch. "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms". *IEEE Trans. Audio Electroacoust.* 15.2 (1967), 70–73.

[59] R. W. Hamming. *Digital filters, 2nd ed.* Englewood Cliffs, New Jersey: Prentice-Hall, 1983. ISBN: 0132128126.

[60] R. Tibshirani. "Regression shrinkage and selection via the lasso". *J. Roy. Statist. Soc. Ser. B* 58.1 (1996), 267–288.

[61] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1985, pp. xiii+561. ISBN: 0-521-30586-1.

[62] G. Barbarino, C. Garoni, M. Mazza, and S. Serra-Capizzano. "Rectangular GLT Sequences". *Electronic Transactions on Numerical Analysis* 55 (2022), 585–617.

[63] M. Bolten, M. Donatelli, P. Ferrari, and I. Furci. "Symbol based convergence analysis in block multigrid methods with applications for Stokes problems". *Appl. Numer. Math.* 193.C (2023), 109–130.

[64] S. Sommariva, A. Sorrentino, M. Piana, V. Pizzella, et al. "A comparative study of the robustness of frequency-domain connectivity measures to finite data length". *Brain Topogr.* 32.4 (2019), 675–695.

[65] S. Haufe and A. Ewald. "A simulation framework for benchmarking EEG-based brain connectivity estimation methodologies". *Brain Topogr.* 32.4 (2019), 625–642.

[66] H. Lütkepohl. *New introduction to multiple time series analysis*. Springer-Verlag, Berlin, 2005, pp. xxii+764. ISBN: 3-540-40172-5.

[67] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, et al. "MEG and EEG data analysis with MNE-Python". *Frontiers in Neuroscience* 7 (2013).

[68] P. Gerstoft, A. Xenaki, and C. F. Mecklenbräuker. "Multiple and single snapshot compressive beamforming". *J. Acoust. Soc. Am.* 138.4 (2015), 2003–2014.

[69] J. R. Hoffman and R. P. Mahler. "Multitarget miss distance via optimal assignment". *IEEE Trans. Syst. Man Cybern. A: Syst. Hum.* 34.3 (2004), 327–336.

[70] A. S. Hincapié, J. Kujala, J. Mattout, S. Daligault, et al. "MEG Connectivity and Power Detections with Minimum Norm Estimates Require Different Regularization Parameters". *Comput. Intell. Neurosci.* 2016 (2016).

[71] K. Scheinberg, D. Goldfarb, and X. Bai. "Fast first-order methods for composite convex optimization with backtracking". *Found. Comput. Math.* 14.3 (2014), 389–417.

[72] J.-F. Aujol, L. Calatroni, C. Dossal, H. Labarrière, et al. "Parameter-Free FISTA by Adaptive Restart and Backtracking". *SIAM Journal on Optimization* 34.4 (2024), 3259–3285.

[73] S. Sommariva and A. Sorrentino. "Sequential Monte Carlo samplers for semilinear inverse problems and application to magnetocephalography". *Inverse Probl.* 30.11 (2014), 114020, 23.

[74] S. H. Siddiqi, K. P. Kording, J. Parvizi, and M. D. Fox. "Causal mapping of human brain function". *Nature Reviews Neuroscience* 23.6 (2022), 361–375.

[75] L. Astolfi, F. Cincotti, D. Mattia, S. Salinari, et al. "Estimation of the effective and functional human cortical connectivity with structural equation modeling and directed transfer function applied to high-resolution EEG". *Magnetic Resonance Imaging* 22.10 (2004), 1457–1470.

[76] L. Astolfi, F. De Vico Fallani, F. Cincotti, D. Mattia, et al. "Estimation of Effective and Functional Cortical Connectivity From Neuroelectric and Hemodynamic Recordings". *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 17.3 (2009), 224–233.

[77] M. Papadopoulou, K. Friston, and D. Marinazzo. "Estimating Directed Connectivity from Cortical Recordings and Reconstructed Sources". *Brain Topography* 32.4 (2019), 741–752.

[78] A. Anzolin, P. Presti, F. Van De Steen, L. Astolfi, et al. "Quantifying the Effect of Demixing Approaches on Directed Connectivity Estimated Between Reconstructed EEG Sources". *Brain Topography* 32.4 (2019), 655–674.

[79] P. Manomaisaowapak, A. Nartkulpat, and J. Songsiri. "Granger Causality Inference in EEG Source Connectivity Analysis: A State-Space Approach". *IEEE Transactions on Neural Networks and Learning Systems* 33.7 (2022), 3146–3156.

[80] K. Sameshima and L. A. Baccalá. "Using partial directed coherence to describe neuronal ensemble interactions". *Journal of Neuroscience Methods* 94.1 (1999), 93–103.

[81] G. de Marco, P. Vrignaud, C. Destrieux, D. De Marco, et al. "Principle of structural equation modeling for exploring functional interactivity within a putative network of interconnected brain areas". *Magnetic resonance imaging* 27.1 (2009), 1–12.

[82] J. Cao, Y. Zhao, X. Shan, H.-l. Wei, et al. "Brain functional and effective connectivity based on electroencephalography recordings: A review". *Human Brain Mapping* 43.2 (2022), 860–879.

[83] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[84] G. Zhang, B. Cai, A. Zhang, Z. Tu, et al. "Detecting abnormal connectivity in schizophrenia via a joint directed acyclic graph estimation model". *NeuroImage* 260 (2022), 119451.

[85] A. Bagheri, M. Dehshiri, Y. Bagheri, A. Akhondi-Asl, et al. "Brain effective connectome based on fMRI and DTI data: Bayesian causal learning and assessment". *PLOS ONE* 18.8 (Aug. 2023), 1–22.

[86] G. He, H.-G. Müller, J.-L. Wang, and W. Yang. "Functional linear regression via canonical analysis". *Bernoulli* 16.3 (2010), 705–729.

[87] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Technical University of Denmark, 2008.

[88] A. Nemirovski. *Optimization II: Standard Numerical Methods for Nonlinear Continuous Optimization*. Lecture notes. 1999.

[89] B. Ro, F. Bijma, J. C. de Munck, and M. C. de Gunst. "Existence and uniqueness of the maximum likelihood estimator for models with a Kronecker product covariance structure". *Journal of Multivariate Analysis* 143 (2016), 345–361.

[90] I. Soloveychik and D. Trushin. "Gaussian and robust Kronecker product covariance estimation: Existence and uniqueness". *Journal of Multivariate Analysis* 149 (2016), 92–113.

[91] A. Mheich, F. Wendling, and M. Hassan. "Brain network similarity: methods and applications". *Network Neuroscience* 4.3 (July 2020), 507–527.

[92] D. E. Meskaldji, E. Fischi-Gomez, A. Griffa, P. Hagmann, et al. "Comparing connectomes across subjects and populations at different scales". *NeuroImage* 80 (2013), 416–425.

[93] S. D. Zhao, T. T. Cai, and H. Li. "Direct estimation of differential networks". *Biometrika* 101.2 (2014), 253–268.

[94] B. Zhao, Y. S. Wang, and M. Kolar. "FuDGE: A method to estimate a functional differential graph in a high-dimensional setting". *Journal of Machine Learning Research* 23.82 (2022), 1–82.

[95] D. Montgomery. *Design and Analysis of Experiments*. Wiley, 2017.

[96] C. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, 1936.

[97] Z. idák. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions". *Journal of the American Statistical Association* 62.318 (1967), 626–633.

[98] S. Holm. "A Simple Sequentially Rejective Multiple Test Procedure". *Scandinavian Journal of Statistics* 6.2 (1979), 65–70.

[99] Y. Benjamini and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (Dec. 1995), 289–300.

[100] A. Winkler, G. Ridgway, M. Webster, S. Smith, et al. "Permutation inference for the general linear model". *NeuroImage* 92 (2014), 381–397.

[101] S. Searle. *Linear models*. John Wiley & Sons, 1997.

[102] E. Edgington. "Approximate Randomization Tests". *The Journal of Psychology* 72.2 (1969), 143–149.

[103] B. Phipson and G. Smyth. "Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn". *Statistical Applications in Genetics and Molecular Biology* 9.1 (2010).

[104] P. Good. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. 3rd ed. Springer New York, 2005, pp. 33–65.

[105] D. Freedman and D. Lane. "A Nonstochastic Interpretation of Reported Significance Levels". *Journal of Business & Economic Statistics* 1.4 (1983), 292–298.

[106] T. O'Gorman. "The Performance of Randomization Tests that Use Permutations of Independent Variables". *Communications in Statistics - Simulation and Computation* 34.4 (2005), 895–908.

[107] T. Nichols, G. Ridgway, M. Webster, and S. Smith. "GLM permutation: nonparametric inference for arbitrary general linear models". *NeuroImage* 41.S1 (2008), S72.

[108] A. Holmes, R. Blair, J. Watson, and I. Ford. "Nonparametric analysis of statistic images from functional mapping experiments". *Journal of Cerebral Blood Flow and Metabolism* 16.1 (1996), 7–22.

[109] A. Fornito, A. Zalesky, and M. Breakspear. "The connectomics of brain disorders". *Nature Reviews Neuroscience* 16.3 (2015), 159–172.

[110] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms, fourth edition*. MIT Press, 2022.

[111] F. Normand, M. Gajwani, D. C. Côté, and A. Allard. "NBS-SNI, an extension of the network-based statistic: Abnormal functional connections between important structural actors". *Network Neuroscience* 8.1 (2024), 44–80.

[112] I. G. McKeith, B. F. Boeve, D. W. Dickson, G. Halliday, et al. "Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium." *Neurology* 89.1 (2017), 88–100.

[113] S. Joza, M. T. Hu, K.-Y. Jung, D. Kunz, et al. "Progression of clinical markers in prodromal Parkinson's disease and dementia with Lewy bodies: a multicentre study". *Brain* 146.8 (2023), 3258–3272.

[114] D. Arnaldi, P. Mattioli, S. Raffa, M. Pardini, et al. "Presynaptic Dopaminergic Imaging Characterizes Patients with REM Sleep Behavior Disorder Due to Synucleinopathy". *Annals of Neurology* 95.6 (2024), 1178–1192.

[115] T. Simuni, L. M. Chahine, K. Poston, M. Brumm, et al. "A biological definition of neuronal alpha-synuclein disease: towards an integrated staging system for research". *The Lancet Neurology* 23.2 (2024), 178–190.

[116] G. U. Höglinger, C. H. Adler, D. Berg, C. Klein, et al. "A biological classification of Parkinson's disease: the SynNeurGe research diagnostic criteria". *The Lancet Neurology* 23.2 (2024), 191–204.

[117] L. Kucikova, H. Kalabizadeh, K. Motsi, S. Rashid, et al. "A systematic literature review of fMRI and EEG resting-state functional connectivity in Dementia with Lewy Bodies: Underlying mechanisms, clinical manifestation, and methodological considerations". *Ageing Research Reviews* 93 (2024), 102159.

[118] J. Schumacher, L. R. Peraza, M. Firbank, A. J. Thomas, et al. "Functional connectivity in dementia with Lewy bodies: A within- and between-network analysis". *Human Brain Mapping* 39.3 (2018), 1118–1129.

[119] A. Wada, K. Tsuruta, R. Irie, K. Kamagata, et al. "Differentiating Alzheimer's Disease from Dementia with Lewy Bodies Using a Deep Learning Technique Based on Structural Brain Connectivity". *Magnetic Resonance in Medical Sciences* 18.3 (2019), 219–224.

[120] R. Mehraram, L. Peraza, N. R. E. Murphy, R. A. Cromarty, et al. "Functional and structural brain network correlates of visual hallucinations in Lewy body dementia". *Brain* 145.6 (2022), 2190–2205.

[121] A. Vallesi, C. Porcaro, A. Visalli, D. Fasolato, et al. "Resting-state EEG spectral and fractal features in dementia with Lewy bodies with and without visual hallucinations". *Clinical Neurophysiology* 168 (2024), 43–51.

[122] M. Roascio, A. Canessa, R. Trò, P. Mattioli, et al. "Phase and amplitude electroencephalography correlations change with disease progression in people with idiopathic rapid eye-movement sleep behavior disorder". *Sleep* 45.1 (2021), zsab232.

[123] E. Jeong, Y. Woo Shin, J.-I. Byun, J.-S. Sunwoo, et al. "EEG-based machine learning models for the prediction of phenoconversion time and subtype in isolated rapid eye movement sleep behavior disorder". *Sleep* 47.5 (2024), zsae031.

[124] W. Klimesch, H. Schimke, and G. Pfurtscheller. "Alpha frequency, cognitive load and memory performance". *Brain Topography* 5.3 (1993), 241–251.

[125] P. L. Nunez, L. Reid, and R. G. Bickford. "The relationship of head size to alpha frequency with implications to a brain wave model". *Electroencephalography and Clinical Neurophysiology* 44.3 (1978), 344–352.

[126] M. Stylianou, N. Murphy, L. R. Peraza, S. Graziadio, et al. "Quantitative electroencephalography as a marker of cognitive fluctuations in dementia with Lewy bodies and an aid to differential diagnosis". *Clinical Neurophysiology* 129.6 (2018), 1209–1220.

[127] W. Klimesch. "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis". *Brain Research Reviews* 29.2 (1999), 169–195.

[128] D. V. Moretti, C. Babiloni, G. Binetti, E. Cassetta, et al. "Individual analysis of EEG frequency and band power in mild Alzheimer's disease". *Clinical Neurophysiology* 115.2 (2004), 299–308.

[129] E. Vallarino, S. Sommariva, F. Famà, M. Piana, et al. "Transfreq: A Python package for computing the theta-to-alpha transition frequency from resting state electroencephalographic data". *Human Brain Mapping* 43.17 (2022), 5095–5110.

[130] F. Offner. "The EEG as potential mapping: The value of the average monopolar reference". *Electroencephalography and Clinical Neurophysiology* 2.1 (1950), 213–214.

[131] C. Jutten and J. Herault. "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture". *Signal Processing* 24.1 (1991), 1–10.

[132] M. Vinck, R. Oostenveld, M. Van Wingerden, F. Battaglia, et al. "An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias". *NeuroImage* 55.4 (2011), 1548–1565.

[133] D. Thomson. "Spectrum estimation and harmonic analysis". *Proceedings of the IEEE* 70.9 (1982), 1055–1096.

[134] D. Slepian. "Prolate spheroidal wave functions, fourier analysis, and uncertainty V: the discrete case". *The Bell System Technical Journal* 57.5 (1978), 1371–1430.

[135] S. Fanton and W. H. Thompson. "NetPlotBrain: A Python package for visualizing networks and brains". *Network Neuroscience* 7.2 (June 2023), 461–477.

[136] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, et al. "UpSet: Visualization of Intersecting Sets". *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), 1983–1992.

[137] J. J. van der Zande, A. A. Gouw, I. van Steenoven, P. Scheltens, et al. "EEG Characteristics of Dementia With Lewy Bodies, Alzheimer's Disease and Mixed Pathology". *Frontiers in Aging Neuroscience* 10 (2018).

[138] C. S. Musaeus, K. Engedal, P. Høgh, V. Jelic, et al. "EEG Theta Power Is an Early Marker of Cognitive Decline in Dementia due to Alzheimers Disease". *Journal of Alzheimer's Disease* 64.4 (2018), 1359–1371.

[139] T. Kai, Y. Asai, K. Sakuma, T. Koeda, et al. "Quantitative electroencephalogram analysis in dementia with Lewy bodies and Alzheimer's disease". *Journal of the Neurological Sciences* 237.1 (2005), 89–95.

[140] A. Habich, L. Wahlund, E. Westman, T. Dierks, et al. "(Dis-)Connected Dots in Dementia with Lewy BodiesA Systematic Review of Connectivity Studies". *Movement Disorders* 38.1 (2023), 4–15.

[141] L. Bonanni, R. Franciotti, F. Nobili, M. Kramberger, et al. "EEG Markers of Dementia with Lewy Bodies: A Multicenter Cohort Study". *Journal of Alzheimers Disease* 54.4 (2016), 1649–1657.

[142] J. Lachaux, E. Rodriguez, J. Martinerie, and F. Varela. "Measuring phase synchrony in brain signals". *Human Brain Mapping* 8.4 (1999), 194–208.

[143] P. Fries. "A mechanism for cognitive dynamics: neuronal communication through neuronal coherence". *Trends in Cognitive Sciences* 9.10 (2005), 474–480.

[144] D. Chialvo. "Emergent complex neural dynamics". *Nature Physics* 6.10 (2010), 744–750.

[145] D. Bassett and E. Bullmore. "Small-World Brain Networks Revisited". *The Neuroscientist* 23.5 (2017), 499–516.

[146] G. Leodori, A. Fabbrini, A. Suppa, M. Mancuso, et al. "Effective connectivity abnormalities in Lewy body disease with visual hallucinations". *Clinical Neurophysiology* 156 (2023), 156–165.

[147] S. Pusil, M. López, P. Cuesta, R. Bruña, et al. "Hypersynchronization in mild cognitive impairment: the 'X' model". *Brain* 142.12 (2019), 3936–3950.

[148] C. Frantzidis, A. Vivas, A. Tsolaki, M. Klados, et al. "Functional disorganization of small-world brain networks in mild Alzheimer's Disease and amnestic Mild Cognitive Impairment: an EEG study using Relative Wavelet Entropy (RWE)". *Frontiers in Aging Neuroscience* 6 (2014).

[149] M. Schutte, M. Bohlken, G. Collin, L. Abramovic, et al. "Functional connectome differences in individuals with hallucinations across the psychosis continuum". *Scientific Reports* 11.1 (2021), 1108.

[150] E. Kenny, J. O'Brien, M. Firbank, and A. Blamire. "Subcortical connectivity in dementia with Lewy bodies and Alzheimer's disease". *British Journal of Psychiatry* 203.3 (2013), 209–214.

[151] E. Kenny, A. Blamire, M. Firbank, and J. O'Brien. "Functional connectivity in cortical regions in dementia with Lewy bodies and Alzheimer's disease". *Brain* 135.2 (2011), 569–581.

[152] J. Schumacher, A. J. Thomas, L. R. Peraza, M. Firbank, et al. "Functional connectivity of the nucleus basalis of Meynert in Lewy body dementia and Alzheimer's disease". *International Psychogeriatrics* 33.1 (2021), 89–94.

[153] C. Babiloni, C. Del Percio, R. Lizio, G. Noce, et al. "Abnormalities of resting-state functional cortical connectivity in patients with dementia due to Alzheimer's and Lewy body diseases: an EEG study". *Neurobiology of Aging* 65 (2018), 18–40.

[154] S. Morbelli, A. Chincarini, M. Brendel, A. Rominger, et al. "Metabolic patterns across core features in dementia with lewy bodies". *Annals of Neurology* 85.5 (2019), 715–725.

[155] B. Orso, P. Mattioli, E.-J. Yoon, Y. K. Kim, et al. "Progression trajectories from prodromal to overt synucleinopathies: a longitudinal, multicentric brain [18F]FDG-PET study". *npj Parkinson's Disease* 10.1 (2024), 200.

[156] P. Mattioli, M. Pardini, F. Famà, N. Girtler, et al. "Cuneus/precuneus as a central hub for brain functional connectivity of mild cognitive impairment in idiopathic REM sleep behavior patients". *European Journal of Nuclear Medicine and Molecular Imaging* 48.9 (2021), 2834–2845.