Learning and Evaluating Graphs From Spatially Structured Data : Application to Brain Functional Connectivity Networks

Hanâ Lbath

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques et Informatique
Unité de recherche : Laboratoire Jean Kuntzmann

Méthodes d'apprentissage et d'évaluation de graphes à partir de données structurées spatialement : application aux réseaux de connectivité fonctionnelle cérébrale

Learning and Evaluating Graphs From Spatially Structured Data: Application to Brain Functional Connectivity Networks

Présentée par :

Hanâ LBATH

Direction de thèse :

Sophie ACHARD DIRECTRICE DE RECHERCHE, Université Grenoble Alpes	Directrice de thèse

Rapporteurs :

NATHALIE PEYRARD
DIRECTEUR DE RECHERCHE, INRAE CENTRE OCCITANIE-TOULOUSE
DANIEL MARGULIES
DIRECTEUR DE RECHERCHE, CNRS ILE-DE-FRANCE VILLEJUIF

Thèse soutenue publiquement le **25 septembre 2023**, devant le jury composé de :

SOPHIE ACHARD DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES	Directrice de thèse
NATHALIE PEYRARD DIRECTEUR DE RECHERCHE, INRAE CENTRE OCCITANIE-TOULOUSE	Rapporteure
DANIEL MARGULIES DIRECTEUR DE RECHERCHE, CNRS ILE-DE-FRANCE VILLEJUIF	Rapporteur
CLEMENTINE PRIEUR PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Présidente
SARAH MORGAN ASSOCIATE PROFESSOR, UNIVERSITY OF CAMBRIDGE	Examinatrice

Invités :

ALEXANDER PETERSEN
ASSOCIATE PROFESSOR, BRIGHAM YOUNG UNIVERSITY
WENDY MEIRING
FULL PROFESSOR, UNIVERSITY OF CALIFORNIA SANTA BARBARA

Abstract

The main objective of this thesis is to develop methods for statistically consistent estimation and analysis of graphs from multivariate datasets such as those observed in neuroimaging. More precisely, we aim to better account for spatial dependencies of the signals and uncertainties from the data acquisition and models, which are known to greatly impede both network inference and downstream analyses. Most of the contributions presented in this thesis could be applied to generic multivariate grouped data. While these can be found in a wide range of fields, from econometry, to family studies, to meteorology, we are motivated in particular by a brain functional connectivity application. In this setting, networks are often constructed from functional Magnetic Resonance Imaging (fMRI) data. Nodes then correspond to predefined brain regions, containing voxels associated to time series, and edges routinely link together correlated regions.

We first propose a large-scale correlation-screening-based approach for binary network inference, and in particular thresholding, in the presence of arbitrary spatial dependence. In some contexts, weighted networks may be preferred over their binary counterparts. We hence contribute to a study that leverages both topological data analysis and spatial information to improve weighted network discriminability. We then tackle the challenge of consistent edge weight estimation and introduce a clustering-based inter-regional correlation estimator that simultaneously offsets the effects of noise and arbitrary spatial dependence. Instead of considering point estimates, as is usually performed in the literature, we subsequently construct densities of correlations as connectivity phenotypes for an individual scan, and consider distribution-weighted networks. We assess repeatability and performance on common machine learning tasks, and highlight the effects of high inter-subject variability. The proposed distribution-based paradigm introduces foundations for the definition of a framework that could evaluate and account for uncertainty and quality in connectivity network estimation.

Throughout this thesis we evaluate and validate our methods on both synthetic and real-world brain fMRI datasets.

Résumé

L'objectif principal de cette thèse est de développer des méthodes d'estimation statistiquement consistantes et d'analyse de graphes à partir de données multivariées telles que celles observées en neuroimagerie. Plus précisément, nous visons à mieux prendre en compte les dépendances spatiales des signaux ainsi que les incertitudes liées à l'acquisition des données et à la modélisation, qui sont connues pour entraver l'inférence de réseaux et les analyses en aval. La plupart des contributions présentées dans cette thèse pourraient être appliquées à des données groupées multivariées génériques. Bien que ces données soient présentes dans un large éventail de domaines, de l'économétrie aux études familiales, en passant par la météorologie, nous sommes motivés en particulier par une application à la connectivité fonctionnelle cérébrale. Dans ce contexte, les réseaux sont souvent construits à partir de données d'imagerie par résonance magnétique fonctionnelle (IRMf). Les nœuds correspondent alors à des régions cérébrales prédéfinies, contenant des voxels associés à des séries temporelles, et les arêtes relient couramment des régions corrélées.

Nous proposons tout d'abord une approche basée sur le criblage en grande dimension de corrélations pour l'inférence de graphes binaires, et en particulier l'étape de seuillage, en présence de dépendance spatiale arbitraire. Dans certains contextes, les réseaux pondérés peuvent être préférés à leurs équivalents binaires. Nous contribuons donc à une étude qui exploite à la fois l'analyse topologique de données et des informations spatiales pour améliorer la discriminabilité de graphes pondérés. Nous nous attaquons ensuite au défi de l'estimation consistante du poids des arêtes et introduisons un estimateur de corrélation inter-régionale basé sur des techniques de clustering. Ce dernier compense simultanément les effets du bruit et de dépendances spatiales arbitraires. Enfin, au lieu de considérer des estimations ponctuelles, comme c'est généralement le cas dans la littérature, nous proposons plutôt de considérer des densités de corrélations comme phénotypes de connectivité à l'échelle des individus et de construire des réseaux pondérés par des fonctions de distribution. Nous évaluons alors leur répétabilité et mesurons leurs performances sur des tâches courantes d'apprentissage automatique avant de mettre en évidence les effets d'une forte variabilité inter-sujets. Le paradigme proposé, basé sur

l'utilisation de distributions, introduit les bases pour la définition d'un cadre qui pourrait évaluer et prendre en compte l'incertitude dans l'estimation de réseaux de connectivité.

Tout au long de cette thèse, nous évaluons et validons nos méthodes sur des données synthétiques ainsi que des données d'IRMf cérébrales réelles.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Sophie Achard for this extremely enriching opportunity, and for always being available for questions and discussions, encouraging me, and providing me with guidance throughout this journey.

I also would like to address my warmest thanks to the entire Q-Func project team. In addition to our regular discussions, I am grateful to have had the opportunity to meet with you in person in different parts of the world. In particular, I am deeply thankful to Alex Petersen and Wendy Meiring for welcoming me in Provo and Santa Barbara, respectively, and for providing continuous guidance and feedback. I would also like to sincerely thank Jonas Richiardi for welcoming me in Lausanne and providing insightful comments and feedback throughout these years.

Then, I would like to thank Nathalie Peyrard and Daniel Margulies for taking the time to carefully review this thesis. I thank as well Clémentine Prieur and Sarah Morgan for accepting to be part of my thesis defense jury.

I would also like to thank my other collaborators, Lucrezia, Chao, Jean-François, as well as the Master's students who helped shape parts of this work: Paul, Youssef, Yassir, Francesco, and Simone.

Thank you as well to the Inria Statify team, led by Florence Forbes, with too many members, past and present, to cite here.

At last, I would like to extend my gratitude to my family and friends, in particular to my brothers and my parents who are, and have always been, there for me.

Table of contents

Chapter 1

Introduction

This PhD was conducted as part of an American-French research grant in the context of a multilateral collaboration, and included research visits at Brigham Young University and the University of California, Santa Barabara, in the United States, as well as the Lausanne University Hospital, in Switzerland. During the course of this project, two Master's student interns, a team of two Master's students working on an extracurricular research project, and a Master's thesis student were recruited and co-supervised.

The label "Research, Enterprise, Innovation" was also obtained in conjunction with this PhD project. It consists in an introduction to business organization, project management and entrepreneurship fundamentals.

1.1 General Introduction

Modeling spatially structured data using graphs, or complex networks, which help capture non-trivial relationships in an intuitive fashion, has been gaining traction over the last decades. Such data are inherent to a wide range of applications, ranging from neuroimaging, to meteorology, to bioinformatics. We are primarily interested in multivariate data that can be grouped according to one of their attributes, such as spatial structure. For instance, meteorological data, such as temperature or rainfall records, can be aggregated according to geographical location of measurement site.

This work is motivated in particular by an application to brain functional connectivity (cf. Figure 1.1). In this setting, networks are often constructed from functional Magnetic Resonance Imaging (fMRI) data. These provide 3D images of the brain, where each voxel—which is a 3D pixel—is associated with a blood-oxygen-level-dependent (BOLD) signal. The latter captures oxygen levels as a proxy for the underlying neural activity. Once signals are acquired, the brain may be divided into anatomical regions. In brain functional

connectivity networks, nodes correspond to these predefined brain regions, containing groups of voxels, and edges often link together correlated regions. Brain functional connectivity networks have been extensively used to try to improve our understanding of, not only the healthy, but also, the diseased or injured brain, which, to this day, remains quite mysterious in many ways. More down-to-earth, but still far-off, applications also include helping provide patients, who could be comatose for instance, with a more precise prognosis, and consequently help optimize their treatment and rehabilitation. In such scenarios, providing reliable methods to model the brain, along with a way to account for quality is hence paramount.

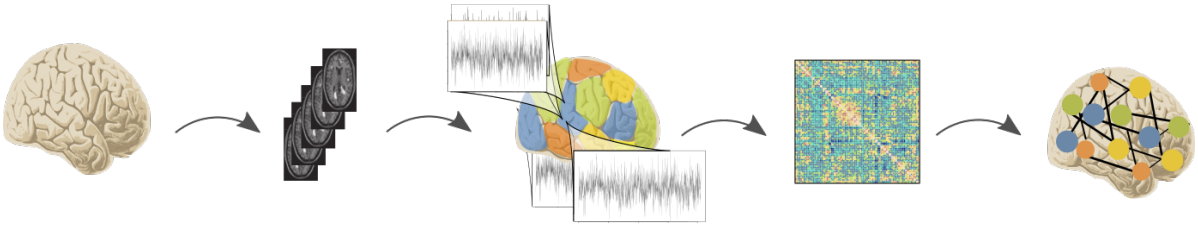


Fig. 1.1 Pipeline to learn a brain functional connectivity network from fMRI data. First fMRI scans of the brain are acquired. The brain is subsequently parcellated into anatomical regions that contain voxel-level BOLD time series. Inter-regional connectivity is then estimated to construct the weighted functional connectivity network. The latter can finally be thresholded to obtain a binary graph.

The overarching goal of this thesis is to develop dependable methods for learning and evaluation of subject-specific networks. More precisely, we aim to better account for spatial dependencies of the signals as well as uncertainties from the data acquisition and estimation models, which both considerably hinder network inference.

1.2 Contributions

We present in this section the outline of this thesis, and identify our main contributions.

- Chapter 2 presents the theoretical context of this thesis as well as notations, and provides an extensive literature review of different concepts utilized throughout this thesis, from correlation estimation to network inference and analysis. This chapter includes in particular excerpts from a recently published journal article. It compares various correlation estimators in the context of functional brain connectivity.

Achard, S., Coeurjolly, J.-F., de Micheaux, P. L., Lbath, H., and Richiardi, J. (2023). Inter-regional correlation estimators for functional magnetic resonance imaging. *NeuroImage*, 282:120388

- Chapter 3 specifies the different synthetic and real-world brain fMRI datasets used during the course of this thesis. We first detail how the simulated data was generated, before describing the different rat and human brain fMRI datasets, as well as the additional preprocessing we performed.

- In Chapter 4, we introduce a novel correlation-screening-based method for binary network estimation, and in particular thresholding. More precisely, we present a novel approach to infer connectivity networks when nodes represent groups of dependent variables, which is the case in our brain functional connectivity application. We first formally establish the importance of leveraging dependence structures to reliably detect inter-group correlations. Our method then consists in estimating, for each pair of groups, an inter-correlation distribution before deriving a tailored correlation threshold based on a correlation screening approach. In particular, we propose simplified expressions for the mean number of discoveries that allow for easier theoretical and empirical manipulation, while taking into account dependence within groups. We also apply our framework on both synthetic data and a real dataset of rat brain fMRI.

  This work was presented at the "Brain Connectivity Networks: Quality and Reproducibility" satellite symposium of the Conference on Complex Systems 2021 (October 2021), in Lyon, France, and at the Joint Statistical Meeting 2022 (August 2022), in Washington DC, USA. It is in the process of being submitted to a journal.

  Lbath, H., Petersen, A., and Achard, S. (2021). Brain functional connectivity estimation. In *Brain Connectivity Networks: Quality and Reproducibility - Satellite of the Conference on Complex Systems 2021*, Lyon, France

  Lbath, H., Petersen, A., and Achard, S. (2022a). Large-scale correlation screening under dependence for brain functional connectivity inference. In *JSM 2022 - Joint Statistical Meetings*, Washington, United States

- In some applications, handling weighted networks, and notably circumventing the thresholding step, may be preferred over manipulating their binary counterparts.

Chapter 5 hence leverages both topological data analysis and regional label information to propose a multi-scale comparison of weighted connectivity networks. The effectiveness of this approach is illustrated via a comparison of comatose and healthy subjects, and of real-world data against null models. The latter could also be seen as way to assess the quality of estimated networks.

This contribution is mainly the product of a Master's thesis student's work who was personally supervised as part of this PhD's thesis project. It is in the process of being submitted to a journal.

Sitoleux, P., Carboni, L., Lbath, H., and Achard, S. (in preparation 2023). Multiscale and multi-density comparison of functional brain networks through label-informed persistence diagrams

- Chapter 6 subsequently tackles the challenge of consistent edge weight estimation. To that end, we propose a novel non-parametric estimator of the correlation between groups of arbitrarily dependent variables in the presence of noise. The challenge resides in the fact that both noise and intra-group correlation lead to inconsistent correlation estimation. However, previous works handle either one or the other but fail to tackle both at the same time. To address this problem, we propose to fully harness the dependency structures of the data, and utilize hierarchical clustering to simultaneously offset the effects of both noise and intra-group correlation. We derive the limiting behavior of our estimator. We also empirically show our approach surpasses popular estimators in terms of quality, and provide illustrations on real-world datasets.

  This work was presented at the IMS International Conference on Statistics and Data Science 2022 (December 2022), in Florence, Italy, where it was rewarded with a Student Award. This contribution has also been published in a journal.

  Lbath, H., Petersen, A., Meiring, W., and Achard, S. (2022b). Clustering-based inter-group correlation estimation. In *ICSDS 2022 - IMS International Conference on Statistics and Data Science*, Florence, Italy

  Lbath, H., Petersen, A., Meiring, W., and Achard, S. (2023). Clustering-based inter-regional correlation estimation. *Computational Statistics & Data Analysis*, page 107876

- Chapter 7 leverages results from Chapter 6 to introduce distribution-based weighted networks, where distribution or density functions are assigned to edges, instead of

the traditional point estimates. This framework hence enables, to a certain extent, accounting for uncertainties in the connectivity estimation. This chapter then aims to validate the practicalities of such a framework on a real-world resting-state fMRI application. We hence evaluate test-retest repeatability and performance on common machine learning tasks, such as classification and regression. This chapter presents the first conclusions of on an ongoing project.

Lbath, H., Richiardi, J., Petersen, A., Meiring, W., and Achard, S. (working paper). Distribution-based weighted networks validation on rs-fMRI data

- In Chapter 8, we discuss possible future directions based on the different contributions of this thesis, and more specifically the proposed distribution-based connectivity framework. We highlight how it could reveal itself as the starting point for the development of functional data analysis-based methodologies that better account for dependencies in high-dimensional settings, as well as provide foundations for an uncertain functional connectivity graph paradigm.

Chapter 2

Literature Review

In this chapter, we introduce the theoretical context of this thesis, provide a literature review of various aspects covered in this manuscript, and highlight key challenges we tackled.

2.1 Preliminaries: Data Structure and Notation

2.1.1 Multivariate grouped data

Multivariate grouped data are inherent to a wide range of applications. Spatio-temporal data are a natural example, with spatially located variables that can be grouped for instance geographically, as in agriculture data (Desloires et al., 2023), or anatomically, as in neuroimaging studies (De Vico Fallani et al., 2014). Data points are typically collected over a period of time that could span minutes, such as via functional Magnetic Resonance Imaging (fMRI), to years, as in satellite imaging of fields for agricultural studies. In these settings, similarity metrics are then often computed between the spatially aggregated measurements. Other applications include familial studies, where traits, such as height, are often measured on n independent families, with the aim to compare for example parents with their children, e.g., (Rosner et al., 1977; Donner et al., 1998). One of the common types of data used in econometry consists of panels with p subjects, which frequently are firms in finance panel regression studies, observed across n time periods. Such data are often clustered across at least one dimension (Cameron and Miller, 2011). An example of multivariate grouped data is illustrated in Figure 2.1.

In this thesis, we are interested in particular by neuroimaging data, and more specifically resting-state fMRI data, where brain signals are often spatially grouped by brain regions, with the aim to estimate inter-regional functional connectivity.

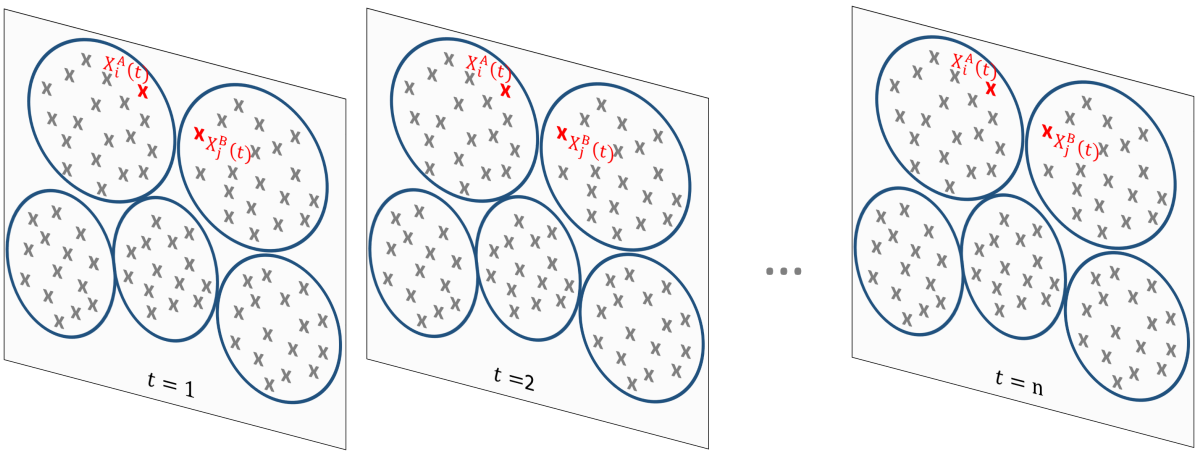


Fig. 2.1 Representation of multivariate grouped data consisting of n samples, e.g., time points, of p variables (gray crosses), e.g., voxels, divided into $J = 5$ groups (blue circles), e.g., anatomical brain regions. For instance, $X_i^A(t)$ is the signal associated to voxel i in region A at time t.

Vocabulary: Multivariate Grouped Data

We define **multivariate grouped data** as data that consist of measurements of a quantity that can be grouped along at least one dimension, such as space. An illustration of such kind of data is displayed in Figure 2.1.

Depending on the context, these groups may also be denominated clusters or classes. For instance, these groups may correspond to spatial regions (e.g., neuroimaging (De Vico Fallani et al., 2014), agricultural (Desloires et al., 2023) or meteorological data (Wigley et al., 1984)), siblings and parents (in familial data literature (Rosner et al., 1977)), or clusters of companies (e.g., econometry literature (Cameron and Miller, 2011) or organization studies (Ostroff, 1993)). Throughout this work, we will refer to *groups* and *regions* interchangeably.

Handling data that consists of measurements of a quantity that can be grouped along one or more dimensions is a challenging, but understudied, undertaking for several reasons. In this section we will highlight key challenges and briefly overview their repercussions. Some of them will be further detailed in the following sections.

Dependence. First, dependence between variables within a group can substantially hamper exploitation of the data. This is especially true when one is interested in exploring associations between groups, by computing for instance inter-group correlations. Effects of the intra-group dependence on inter-group correlation estimation have been demonstrated

in various contexts, such as physical activity assessments (Perisic and Rosner, 1999), or organization studies (Ostroff, 1993). Comparing dependent correlations is often the next step, and specific test statistics need to be derived, e.g., (Steiger, 1980). Intra-group correlation has in fact been shown to impact the effective degrees of freedom, and thus the variance of the sample inter-correlation estimators, which is used for significance testing of the inter-correlation, see, e.g. (Elston, 1975; Rosner et al., 1979; Konishi, 1982; Clifford et al., 1989; Donner et al., 1998). Econometry studies, such as financial panel data regression, are also faced with the issue of accounting for correlation within clusters, which affects estimation of the variance of the regression estimator. Moreover, data could be clustered across one (e.g., across individuals) or multiple dimensions (e.g., across both individuals and time) (Thompson, 2011; Cameron and Miller, 2011; Cameron et al., 2011). The latter can be also linked to autocorrelation variance estimation in spatio-temporal settings (Cameron and Miller, 2011). Note however that inter-cluster correlation is assumed to be zero in these settings, unlike in many other applications, such as neuroimaging.

Vocabulary: Intra- and Inter-Correlation

Dependence between variables both within and between groups, or regions, is a salient point that will be addressed throughout this thesis. We highlight here the definition of two terms that are central to much of this work:

- **Intra-correlation** is the correlation between any pair of random variables within the *same* region.
- **Inter-correlation** is the correlation between two random variables from two *distinct* regions.

High dimensionality. In some contexts, the number of variables, for instance, the number of genes in gene expression data, or voxels in brain imaging, can be much larger than the number of samples, such as the number of subjects or the number of time points. This has consequences for downstream analyses, such as correlation network inference or correlation screening. For example, increasing the number of variables impacts the sample correlation distribution, and increases the chance to estimate abnormally large correlations (Fan and Lv, 2008a). In the context of correlation screening or variable selection, this may lead to an increase in spuriously detected, or selected, correlations. This phenomenon is closely linked to the vast multiple testing literature, and is exacerbated when variables are dependent (Goeman et al., 2019). In the network inference setting, the latter has been

leveraged for example to recover partial-correlation-weighted networks while controlling for family-wise error rate (FWER) (Drton and Perlman, 2007). Regression of grouped variables in high dimensions is also challenging (Qiu and Ahn, 2020).

Noise. Measurement noise is another challenge. While it is an issue that is not necessarily restricted to multivariate grouped data, its effects are still noteworthy but often neglected in this setting. For instance, additive noise is known to attenuate correlation estimation between two variables, e.g., (Perisic and Rosner, 1999; Matzke et al., 2017). In order to alleviate the effect of noise, a prevalent and effective strategy is to aggregate the data, for example, by averaging across groups, e.g., (Ostroff, 1993; Dunlap et al., 1983). In fMRI data, a common preprocessing step is the application of spatial smoothing for the purpose of increasing the signal-to-noise ratio, e.g., (Liu et al., 2017a). This method consists in averaging neighboring voxels, weighted by a Gaussian kernel of user-defined width. However, aggregation can result in information loss and potentially lead to undesired side effects, such as inter-correlation overestimation (Liu et al., 2017a).

2.1.2 Notation

In this section, we define our notation together with the inter- and intra-correlation coefficients. As detailed previously, we are interested in multivariate data that are organized into groups. Let us consider p variables grouped into J regions. As an illustration, we will focus in this section on brain fMRI data where individual observed variables correspond to blood-oxygen-level-dependent (BOLD) signals that are assigned to *voxels* (which are 3D pixel), and are grouped into brain *regions*. Nonetheless, the following notations and definitions can be applied to any dataset of grouped measurements of a quantity.

Let us now consider one such region, which we denote A, and that contains $p_A < p$ voxels. Let $X_1^A, \ldots, X_i^A, \ldots, X_{p_A}^A$ denote the p_A spatially dependent random variables inside region A. Quantities of interest include the correlation coefficients between the variables. We define the **intra-correlation** as Pearson's correlation between any pair of random variables *within* a given region A. We denote by $\eta_{i,i'}^A$ the population intra-correlation of the variables $X_i^A, X_{i'}^A$. Similarly, we define the **inter-correlation** as Pearson's correlation between any pair of random variables from two *distinct* regions A and B. Let $\rho_{ij}^{A,B}$ denote the population inter-correlation coefficient between X_i^A and X_j^B. These concepts are illustrated in Figure 2.2.

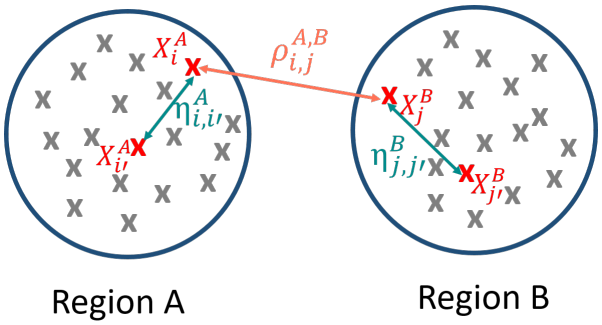


Fig. 2.2 Simplified representation of two regions A and B, as well as the voxel-level intra-correlation $\eta_{i,i'}^A$ and inter-correlation $\rho_{i,j}^{A,B}$ coefficients. Gray crosses represent possibly spatially located random variables, e.g., voxels.

We consider as well n samples of X_i^A, and define the corresponding vector of observations $\mathbf{X}_i^A = [X_i^A(1), \ldots, X_i^A(t), \ldots X_i^A(n)]$. In the context of brain functional connectivity, $\mathbf{X}_i^A$ corresponds to the fMRI BOLD signal time series associated with voxel i of brain region A. It has n time points. Details about the estimation of Pearson's correlation coefficient will be elaborated in Section 2.2.

Throughout this thesis $\bar{\cdot}$ represents an empirical average across either time points or voxels, and $\hat{\cdot}$ corresponds to an empirical estimate.

2.2 Correlation Estimation

2.2.1 Pearson's sample correlation coefficient

Let X_i^A and X_j^B be two random variables. Then, **Pearson's correlation coefficient**, also known as Pearson's product-moment correlation, between these two variables is defined as:

$$Cor(X_i^A, X_j^B) = \frac{Cov(X_i^A, X_j^B)}{\sqrt{Var(X_i^A) \cdot Var(X_j^B)}}, \tag{2.1}$$

where $Cov(\cdot, \cdot)$ is the population covariance of two random variables, and $Var(\cdot)$ is the population variance of a random variable. In saptio-temporal contexts, this corresponds to the zero-lag cross-correlation. In our multivariate grouped data setting, when $B \neq A$, $Cor(X_i^A, X_j^B)$ is equal to the inter-correlation $\rho_{i,j}^{A,B}$ between voxel i in region A and voxel j in region B, while when $B = A$, $Cor(X_i^A, X_j^A)$ corresponds to the intra-correlation $\eta_{i,j}^A$. We now assume n samples $X_i^A(t), X_j^B(t)$, $t = 1, \ldots, n$ of each of the two variables X_i^A, X_j^B are available. The **sample Pearson correlation coefficient** is then defined as follows:

$$R_{i,j}^{A,B} = \widehat{Cor}(X_i^A, X_j^B) = \frac{\sum_{t=1}^{n}\left(X_i^A(t) - \overline{X_i^A}\right)\left(X_j^B(t) - \overline{X_j^B}\right)}{\sqrt{\sum_{t=1}^{n}\left(X_i^A(t) - \overline{X_i^A}\right)^2 \sum_{t=1}^{n}\left(X_j^B(t) - \overline{X_j^B}\right)^2}}, \tag{2.2}$$

where $\overline{X_i^A} = \frac{1}{n}\sum_{t=1}^{n} X_i^A(t)$ and $\overline{X_j^B} = \frac{1}{n}\sum_{t=1}^{n} X_j^B(t)$ are the sample means.

Theoretical properties of the sample correlation coefficient have been extensively investigated under the assumption $\left(X_i^A(1), X_j^B(1)\right), \ldots, \left(X_i^A(n), X_j^B(n)\right)$ are independent and follow a bivariate normal distribution. For instance, exact closed-form expressions of the density of sample correlations can be derived under these conditions (Fisher, 1915; Hotelling, 1953; Muirhead, 2005), and several transformations of the sample correlation coefficient that exhibit simplified distributional properties have been proposed, e.g., (Fisher, 1921; Harley, 1957; Ruben, 1966). In addition, the sample correlation coefficient is known to be asymptotically normal, e.g., (Hotelling, 1953; Muirhead, 2005), and asymptotic expressions of its mean and variance have also been established, e.g., (Fisher, 1915; Hotelling, 1953).

However, in practice, the bivariate normality assumption is often violated. In such cases, while asymptotic results of varying complexity have been proposed under relaxed assumptions, such as bivariate elliptical distributions (Muirhead, 2005), or even non-normality, e.g., (Ogasawara, 2006), to the best of our knowledge, no exact analytical expression of the sample correlation density has been derived in the general case. This adds complexity to the study of the influence of various elements, such as noise or spatial dependence, on the estimation of the correlation.

2.2.2 Impact of noise and dependence on correlation estimation

Temporal dependence. If the temporal samples are dependent, the variance of the sample inter-correlation will be artificially increased. This phenomenon has been described in different contexts, such as meteorological data (Gunst, 1995) or fMRI data (Afyouni et al., 2019). Filtering, another common preprocessing step in fMRI data was shown to introduce temporal dependence, which exacerbates the inflation of the variance of the sample inter-correlation (Davey et al., 2013). On the other hand, wavelet transformation of the BOLD signals, an alternative to filtering, decreases this temporal dependence (Whitcher et al., 2000b).

Spatial dependence. As mentioned previously, spatial dependence also impacts inter-correlation estimation. Many works on the estimation of inter-correlations have mostly focused on aggregating variables within predefined regions (De Vico Fallani et al., 2014;

Dadi et al., 2019). In the context of brain functional connectivity network inference, some prefer techniques based on independent component analysis (ICA) (Calhoun et al., 2012), while most focus on summarizing all voxels within predefined brain regions by their average, before computing Pearson's correlation across time, possibly after wavelet or other filtering, e.g., (Achard et al., 2006; Bolt et al., 2017; Ogawa, 2021; Zhang et al., 2016; Malagurski et al., 2019). However, such approaches suffer from loss of relevant information and can lead to statistical inconsistency and incorrect correlation estimation (Ostroff, 1993). In particular, the average of weakly correlated time series, which corresponds to samples of a single variable in our data model, is difficult to reliably estimate (Wigley et al., 1984). Additionally, it was observed on small samples that the correlation of averages is different than the average of correlations (Dunlap et al., 1983). This phenomenon can also be easily checked with arbitrary large samples. Furthermore, correlation of averages are known to overestimate the true correlation, especially when intra-correlations are weak (Spearman, 1913; Hotelling, 1953; Achard et al., 2011, 2023). It was also empirically observed in fMRI data that the application of spatial smoothing, which is a common preprocessing step to reduce the effect of noise, causes the inter-regional correlations to be overestimated (Liu et al., 2017a). Overestimation is particularly problematic when the goal is to identify significant inter-correlations, such as when inferring binary connectivity networks, since it may lead to identifying spurious edges. Several methods tackling the impact of intra-correlation on the estimation of inter-correlation have been proposed in familial data literature, e.g., (Elston, 1975; Rosner et al., 1977; Srivastava and Keen, 1988; Wilson, 2010). These approaches nonetheless do not address the impact of noise. Moreover, they require normality assumptions on the samples.

Noise. It has been established in various contexts that correlation is underestimated in the presence of noise (Ostroff, 1993; Matzke et al., 2017; Saccenti et al., 2020). Bayesian inference methods have been proposed to offset the effect of measurement errors (Matzke et al., 2017). However they only handle pairs of variables, as opposed to groups of variables—which is what we are interested in. Robust correlation estimation has also been extensively investigated but mostly for specific distributions, such as contaminated normal distributions (Shevlyakov and Smirnov, 2016) or with heavy tails (Lindskog, 2000). Furthermore, groups of variables are not considered either in these contexts. As stated earlier, cluster-robust inference in the presence of both noise and within-group correlation has been studied in the econometric literature (Cameron and Miller, 2015). However, inter-correlation, which often is the quantity of interest, is assumed to be zero.

There hence seems to be a lack of non-parametric inter-correlation estimators that are simultaneously robust to noise and intra-correlation.

2.2.3 Review of inter-correlation estimators

This section is mostly based on portions of the following journal article where several inter-correlation estimators are compared.

> Achard, S., Coeurjolly, J.-F., de Micheaux, P. L., Lbath, H., and Richiardi, J. (2023). Inter-regional correlation estimators for functional magnetic resonance imaging. *NeuroImage*, 282:120388

Model. We first define a model of the data that will be needed to highlight theoretical properties of the various estimators. For a given region A, let $\epsilon_1^A, \ldots, \epsilon_i^A, \ldots, \epsilon_{p_A}^A$ represent random *local noise* variables, and e represent random *global noise* that is equally corrupting all voxels of all regions. We assume that the latent (unobserved) process X_i^A at each voxel i at time t is contaminated by both local noise $\epsilon_i^A(t)$, and global noise $e(t)$ so that the observed variables $Y_i^A(t)$ in region A are

$$Y_i^A(t) = X_i^A(t) + \varepsilon_i^A(t) + e(t), \quad i = 1, \ldots, p_A \quad \text{and} \quad t = 1, \ldots, n. \tag{2.3}$$

We now need to introduce some assumptions in order to facilitate the derivation of theoretical properties of the different inter-correlation estimators. We first assume that $X_i(\cdot)$, $\varepsilon_i^A(\cdot)$ and $e(\cdot)$ are mutually independent and independent in time. We assume as well within-region homoscedasticity of signal and global homoscedasticity for global noise, i.e.,

$$\sigma_A^2 = Var\left(X_i^A\right), \ \sigma_e^2 = Var\left(e\right), \quad i = 1, \ldots, p_A, \quad \text{with } \sigma_A > 0, \sigma_e \geq 0.$$

We assume that inside each region A, the signals of interest have positive intra-correlation $\eta_{ii'}^A$, and that for each time t and for each region both the latent signal and the local noise are stationary random field defined over the voxels within that region. We furthermore assume that the intra-correlations between latent signals (resp. between local noise variables) are stationary and isotropic with respect to the uniform distance, that is, they only depend on the uniform distance between the two voxels i and i'. Our a priori is that the intra-correlation is close to 1 for moderate distances δ, meaning that close neighbors are strongly (positively) correlated. We also assume that there exists d such that the local noise correlation is equal to 0 for any $\delta \geq d$. Without loss of generality, we

also intrinsically assume that for all $i = 1, \dots, p_A$ and $j = 1, \dots, p_B$, $\varepsilon_i^A(t)$ and $\varepsilon_j^B(t)$ are uncorrelated. The intra-regional covariance of the local noise can be defined as follows:

$$Cov\left(\varepsilon_i^A(t), \varepsilon_{i'}^A(t)\right) = \sigma^2_{\varepsilon^A}\eta^A_{\varepsilon,ii'},\ i, i' = 1, \dots, p_A, \text{ with } \sigma^2_{\varepsilon^A} = Var\left(\varepsilon_i^A\right) = \left(\varepsilon_{i'}^A\right) \geq 0. \quad (2.4)$$

For a given pair of distinct regions, A, B, the inter-correlation between any pair of latent random variables X_i^A, X_j^B is assumed to be constant across voxels, and is denoted as $\rho^{A,B}$. This is the quantity we aim to estimate.

Inter-correlation estimators. We now detail several estimators and their theoretical properties. Gaining a clear understanding of these aspects will prove crucial when discussing the main contributions of this thesis. The definition of these estimators, as well as their limit when the number of samples n tends to infinity, are reported in Table 2.1.

The *correlation of averages* (CA) estimator is the most popular in functional connectivity networks and was designed to reduce the impact of local noise. It consists in spatially averaging the signals within each region, before computing Pearson's correlation between these regional averages. CA is a strongly consistent estimator of (2.6). In the absence of global noise ($\sigma_e = 0$), (2.6) depends on both the intra-correlation and local noise, which both appear in the denominator. The local noise is nevertheless smoothed since the spatial average of the local noise intra-correlations $\overline{\eta}_\epsilon^A = \mathcal{O}(1/p_A)$, which can be small. However, even in the absence of noise, CA is still affected by the intra-correlation: the weaker the spatial intra-regional dependence, which corresponds to inhomogeneous regions, the larger the overestimation of the inter-correlation. This effect may also be compounded when regions are large, as was observed by Achard et al. (2011).

Instead of evaluating correlation of spatial averages, it is natural to perform the spatial *average of correlations* (AC). The AC estimator of the inter-correlation corresponds to the ensemble estimator from the familial data literature, e.g. (Rosner et al., 1977). AC is a strongly consistent estimator of (2.8), which does not depend on intra-correlation values. However, even in the absence of global noise, it remains biased by local noise, which is present in the denominator of (2.8). This implies that AC underestimates the true inter-correlation when the local noise is large, which is in accordance with results mentioned in Section 2.2.2.

In order to cancel out the effect of local noise, we can consider a slight adaptation of the replicate-based estimator introduced by Bergholm et al. (2010) in the context of

image analysis. The main idea is to average the correlations computed using voxels that are sufficiently far from each other to ensure local noises are uncorrelated. The *replicate* estimator (R) is defined by (2.11), where, for $s = 1, \dots, S$, the voxels $i_1^{(s)}, i_2^{(s)}$ in region A are such that $|i_2^{(s)} - i_1^{(s)}| = \delta \geq d$. In the same way, voxels $j_1^{(s)}, j_2^{(s)}$ in region B are such that $|j_2^{(s)} - j_1^{(s)}| = \delta \geq d$. R is a strongly consistent estimator of (2.12), which does not depend on the local noise. Nonetheless, the intra-correlation η_δ between voxels at distance δ is present in its denominator, although it can be close to 1 for small δ. In that case, and in the absence of global noise, R would be a consistent estimator of $\rho^{A,B}$.

All these estimators are contaminated by global noise, which appears in both the numerator and denominator of their limit. In order to counteract the effect of global noise, we assume there exists two regions K, K', which are uncorrelated between themselves and from all the other ones. In the context of fMRI data, the field of view is typically larger than the brain itself. The definition of additional, and uncorrelated, regions is hence possible, for instance using air voxels. The idea is then to subtract their signals from the that of the regions of interest in order to remove the global noise term. Denominated the *disconnected* estimator (D), it is defined by (2.15), where

$$\widetilde{Cor}(\mathbf{Y}^A_{i^{(s)}}, \mathbf{Y}^B_{j^{(s)}}; \mathbf{Y}^K_{k^{(s)}}, \mathbf{Y}^{K'}_{k'^{(s)}}) = \frac{\widehat{\mathrm{Cov}}(\mathbf{Y}^A_{i^{(s)}} - \mathbf{Y}^K_{k^{(s)}}, \mathbf{Y}^B_{j^{(s)}} - \mathbf{Y}^{K'}_{k'^{(s)}})}{\widehat{s}(\mathbf{Y}^A_{i^{(s)}}, \mathbf{Y}^K_{k^{(s)}}, \mathbf{Y}^{K'}_{k'^{(s)}})\ \widehat{s}(\mathbf{Y}^B_{j^{(s)}}, \mathbf{Y}^K_{k^{(s)}}, \mathbf{Y}^{K'}_{k'^{(s)}})},$$

and where for three vectors $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ with same length

$$\widehat{s}^2(\mathbf{U}, \mathbf{V}, \mathbf{W}) = \left(\widehat{Var}(\mathbf{U} - \mathbf{V}) + \widehat{Var}(\mathbf{U} - \mathbf{W}) - \widehat{Var}(\mathbf{V} - \mathbf{W}) \right) / 2.$$

D is a strongly consistent estimator of (2.16), which does not depend on the global noise. However, local noise is present in the denominator.

In order to eliminate the effects of both local and global noise, the R and D estimators can be combined into the RD estimator, which is defined by (2.19). It is a strongly consistent estimator of (2.20), which is free of noise terms. It remains nevertheless contaminated by the η_δ intra-correlation term.

Another approach to reduce the impact of noise, with the side-effect of decreasing the effect of intra-correlation, would be to go back to spatially averaging the signals. As mentioned previously, when noisy signals are averaged, the signal to noise ratio increases. However, as detailed above, averaging over a large area may introduce considerable

bias. Instead of averaging over entire regions, like in the CA estimator, or of considering individual voxels, as in the AC or R estimators, one can average over neighborhoods of voxels of interest. A *local* version of the previous estimators can thus be defined (cf. Table 2.1). Additional notation need to be introduced accordingly. For any region A, we define a ν-neighborhood as the subset of voxels within region A such that they are at a distance less than or equal to ν from the center of the neighborhood. For $s = 1, \ldots, S$, denote by $\mathcal{V}_A^{(s)}$ the ν-neighborhood of a voxel in region A. Let

$$\bar{Y}_{\mathcal{V}_A^{(s)}}(t) = \frac{1}{|\mathcal{V}_A^{(s)}|} \sum_{i \in \mathcal{V}_A^{(s)}} Y_i^A(t), \ t = 1, \ldots, n,$$

be the spatial average of the signals associated to the voxels within the neighborhood $\mathcal{V}_A^{(s)}$ with cardinality $|\mathcal{V}_A^{(s)}|$. Analogously to $\bar{\eta}^A$, let $\bar{\eta}^{\mathcal{V}_A}$ be the spatial average of correlations between voxels within ν-neighborhood $\mathcal{V}_A$. Define as well the average correlation between voxels within two distinct ν-neighborhoods $\mathcal{V}, \mathcal{V}'$ of voxels i_1, i_2 at distance $|i_2 - i_1| = \delta$:

$$\bar{\eta}_\delta^{\mathcal{V},\mathcal{V}'} = \frac{1}{|\mathcal{V}| \cdot |\mathcal{V}'|} \sum_{i \in \mathcal{V}, i' \in \mathcal{V}'} \eta_{i,i'}.$$

Looking at the limits of the local versions of the estimators, it is apparent that the local noise terms are still smoothed, although maybe not as much as in the non-local versions, since $\eta_\varepsilon^{\mathcal{V}_A} = \mathcal{O}(1/|\mathcal{V}_A|)$. Moreover, local averaging allows to modulate the impact of the intra-correlation. Indeed, within-neighborhood correlations are expected to be larger than within-region correlations.

Table 2.1 Definitions of inter-correlation estimators and their limit under model (2.3) when n tends to infinity. Adapted from Table 1 in (Achard et al., 2023)

Estimator	**Limit of $r^{\bullet}_{A,B}$**
Correlation of averages (CA) $r^{\mathrm{CA}}_{A,B} = \widehat{Cor}\left(\frac{1}{p_A}\sum_{i=1}^{p_A}\mathbf{Y}^A_i, \frac{1}{p_B}\sum_{j=1}^{p_B}\mathbf{Y}^B_j\right)$ (2.5)	$\frac{\rho^{A,B} + \sigma_e^2/\sigma_A\sigma_B}{\sqrt{(\bar{\eta}^A + \frac{\sigma_\varepsilon^2}{\sigma_A^2}\bar{\eta}^A_\varepsilon + \frac{\sigma_e^2}{\sigma_A^2})(\bar{\eta}^B + \frac{\sigma_\varepsilon^2}{\sigma_B^2}\bar{\eta}^B_\varepsilon + \frac{\sigma_e^2}{\sigma_B^2})}}$ (2.6)
Average of correlations (AC) $r^{\mathrm{AC}}_{A,B} = \frac{1}{p_A p_B}\sum_{i=1}^{p_A}\sum_{j=1}^{p_B}\widehat{Cor}(\mathbf{Y}^A_i, \mathbf{Y}^B_j)$ (2.7)	$\frac{\rho^{A,B} + \sigma_e^2/\sigma_A\sigma_B}{\sqrt{(1 + \frac{\sigma_\varepsilon^2}{\sigma_A^2} + \frac{\sigma_e^2}{\sigma_A^2})(1 + \frac{\sigma_\varepsilon^2}{\sigma_B^2} + \frac{\sigma_e^2}{\sigma_B^2})}}$ (2.8)
Local correlation of averages (ℓCA) $r^{\ell\mathrm{CA}}_{A,B} = \frac{1}{S}\sum_{s=1}^{S}\widehat{Cor}\left(\bar{\mathbf{Y}}_{\mathcal{V}_A^{(s)}}, \bar{\mathbf{Y}}_{\mathcal{V}_B^{(s)}}\right)$ (2.9)	$\frac{\rho^{A,B} + \sigma_e^2/\sigma_A\sigma_B}{\sqrt{(\bar{\eta}^{\mathcal{V}_A} + \frac{\sigma_\varepsilon^2}{\sigma_A^2}\bar{\eta}^{\mathcal{V}_A}_\varepsilon + \frac{\sigma_e^2}{\sigma_A^2})(\bar{\eta}^{\mathcal{V}_B} + \frac{\sigma_\varepsilon^2}{\sigma_B^2}\bar{\eta}^{\mathcal{V}_B}_\varepsilon + \frac{\sigma_e^2}{\sigma_B^2})}}$ (2.10)
Replicates (R) $r^{\mathrm{R}}_{A,B} = \frac{1}{S}\sum_{s=1}^{S}\frac{\frac{1}{4}\sum_{\alpha,\beta=1}^{2}\widehat{Cor}\left(\mathbf{Y}^A_{i^{(s)}_\alpha}, \mathbf{Y}^B_{j^{(s)}_\beta}\right)}{\sqrt{\left\lvert\widehat{Cor}\left(\mathbf{Y}^A_{i^{(s)}_1}, \mathbf{Y}^A_{i^{(s)}_2}\right)\widehat{Cor}\left(\mathbf{Y}^B_{j^{(s)}_1}, \mathbf{Y}^B_{j^{(s)}_2}\right)\right\rvert}}$ (2.11)	$\frac{\rho^{A,B} + \sigma_e^2/\sigma_A\sigma_B}{\sqrt{\lvert(\eta_\delta + \frac{\sigma_\varepsilon^2}{\sigma_A^2})(\eta_\delta + \frac{\sigma_\varepsilon^2}{\sigma_B^2})\rvert}}$ (2.12)
Local averages + Replicates (ℓR) $r^{\ell\mathrm{R}}_{A,B} = \frac{1}{S}\sum_{s=1}^{S}\frac{\frac{1}{4}\sum_{\alpha,\beta=1}^{2}\widehat{Cor}(\bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j_\alpha}}, \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j'_\beta}})}{\sqrt{\lvert\widehat{Cor}(\bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j_1}}, \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j_2}})\widehat{Cor}(\bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j'_1}}, \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j'_2}})\rvert}}$ (2.13)	$\frac{\rho^{A,B} + \sigma_e^2/\sigma_A\sigma_B}{\sqrt{\lvert(\bar{\eta}^{\mathcal{V},\mathcal{V}'}_\delta + \frac{\sigma_\varepsilon^2}{\sigma_A^2})(\bar{\eta}^{\mathcal{V},\mathcal{V}'}_\delta + \frac{\sigma_\varepsilon^2}{\sigma_B^2})\rvert}}$ (2.14)
Disconnected (D) $r^{\mathrm{D}}_{A,B} = \frac{1}{S}\sum_{s=1}^{S}\widetilde{Cor}(\mathbf{Y}^A_{i^{(s)}}, \mathbf{Y}^B_{j^{(s)}}; \mathbf{Y}^K_{k^{(s)}}, \mathbf{Y}^{K'}_{k'^{(s)}})$ (2.15)	$\frac{\rho^{A,B}}{\sqrt{\left(1 + \frac{\sigma_\varepsilon^2}{\sigma_A^2}\right)\left(1 + \frac{\sigma_\varepsilon^2}{\sigma_B^2}\right)}}$ (2.16)
Local averages + Disconnected (ℓD) $r^{\ell\mathrm{D}}_{A,B} = \frac{1}{S}\sum_{s=1}^{S}\widetilde{Cor}(\bar{\mathbf{Y}}_{\mathcal{V}_A^{(s)}}, \bar{\mathbf{Y}}_{\mathcal{V}_B^{(s)}}; \bar{\mathbf{Y}}_{\mathcal{V}_K^{(s)}}, \bar{\mathbf{Y}}_{\mathcal{V}_{K'}^{(s)}})$ (2.17)	$\frac{\rho^{A,B}}{\sqrt{(\bar{\eta}^{\mathcal{V}_A} + \frac{\sigma_\varepsilon^2}{\sigma_A^2}\bar{\eta}^{\mathcal{V}_A}_\varepsilon)(\bar{\eta}^{\mathcal{V}_B} + \frac{\sigma_\varepsilon^2}{\sigma_B^2}\bar{\eta}^{\mathcal{V}_B}_\varepsilon)}}$ (2.18)
Replicates + Disconnected (RD) $r^{\mathrm{RD}}_{A,B} = \frac{1}{S}\sum_{s=1}^{S}\frac{\frac{1}{4}\sum_{\alpha,\beta=1}^{2}\widetilde{Cor}(\mathbf{Y}_{i^{(s)}_\alpha}, \mathbf{Y}_{j^{(s)}_\beta}; \mathbf{Y}_{k^{(s)}}, \mathbf{Y}_{k'^{(s)}})}{\sqrt{\lvert\widetilde{Cor}(\mathbf{Y}_{i^{(s)}_1}, \mathbf{Y}_{i^{(s)}_2}; \mathbf{Y}_{k^{(s)}}, \mathbf{Y}_{k'^{(s)}})\widetilde{Cor}(\mathbf{Y}_{j^{(s)}_1}, \mathbf{Y}_{j^{(s)}_2}; \mathbf{Y}_{k^{(s)}}, \mathbf{Y}_{k'^{(s)}})\rvert}}$ (2.19)	$\frac{\rho^{A,B}}{\lvert\eta_\delta\rvert}$ (2.20)
Local averages + Replicates + Disconnected (ℓRD) $r^{\ell\mathrm{RD}}_{A,B} = \frac{1}{S}\sum_{s=1}^{S}\frac{\frac{1}{4}\sum_{\alpha,\beta=1}^{2}\widetilde{Cor}(\bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j_\alpha}}, \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j'_\beta}}; \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_k}, \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{k'}})}{\sqrt{\widetilde{Cor}(\bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j_1}}, \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j_2}}; \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_k}, \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{k'}})\widetilde{Cor}(\bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j'_1}}, \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{j'_2}}; \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_k}, \bar{\mathbf{Y}}_{\mathcal{V}^{(s)}_{k'}})}}$ (2.21)	$\frac{\rho^{A,B}}{\lvert\bar{\eta}^{\mathcal{V},\mathcal{V}'}_\delta\rvert}$ (2.22)

Properties. We now study the practical properties of these different estimators. Three datasets were used to that end: a synthetic dataset, a rat dataset including both dead and live animals, and a healthy human subject dataset from the Human Connectome Project (HCP). We refer to (Achard et al., 2023) and Chapter 3 for a full description of the datasets.

First, a simulation study (cf. Figure 2.3) on a pair of synthetic regions show that, as expected, the CA estimator tends to overestimate the true inter-correlation for all settings. The D-based estimators are more dispersed than the others for all settings. Even in the absence of local and global noise, ℓR and ℓRD overestimate the true inter-correlation. Furthermore, AC, ℓCA, D and D are affected by the presence of local noise, and exhibit a negative bias, while the replicates-based estimators display a slight positive bias. However, the presence of global noise does not seem to have much impact on the inter-correlation estimation in these simulation settings.

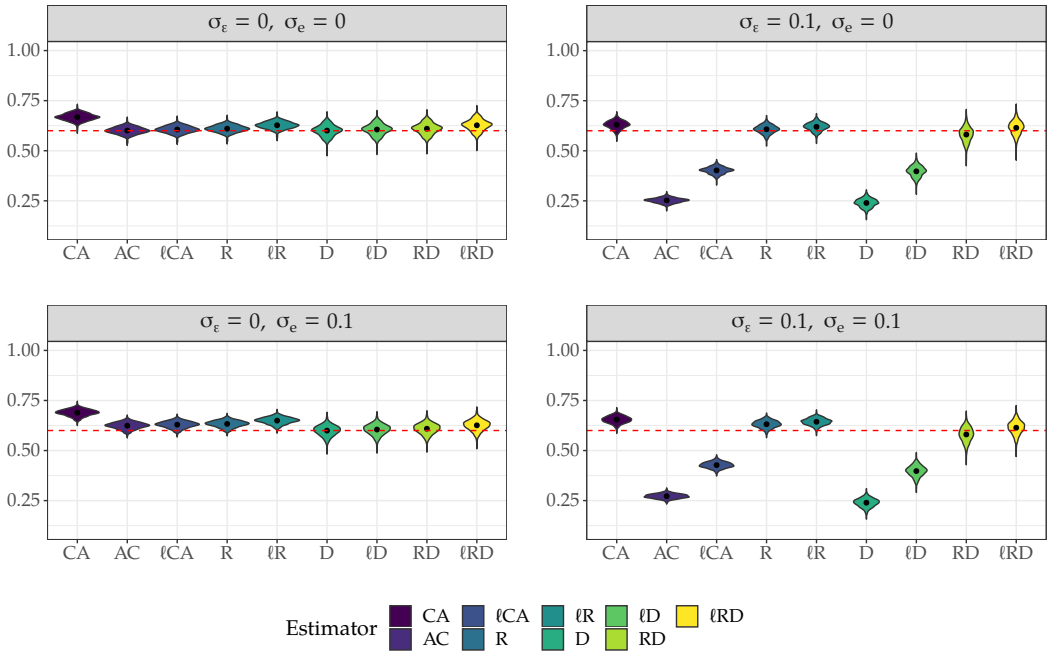


Fig. 2.3 Simulation results. Estimates of the inter-correlation between two regions, containing 20 and 40 voxels, respectively, each corresponding to a time series of length 1000, based on 500 simulation runs of the general model (2.3) with Toeplitz intra-correlation structure. Situations with no noise, local noise or global noise are considered. The true inter-correlation is depicted by the red dashed line. Adapted from Figure 2 in (Achard et al., 2023).

We next evaluate the different estimators on real fMRI data. First, using rat data, we perform a face validity analysis of the estimators, with the premise that dead rats should show no functional connectivity (cf. Figure 2.4). This implies the estimated inter-correlation coefficients should be close to zero. This is the case for the estimators AC, R, ℓCA, D and ℓD. However, the other four estimators showcase a bias towards positive values. Results on live rats show that AC and D provide inter-correlation values close to zero where non-zero correlations should be observed. Combining dead and live rats results, the estimators ℓCA, R and ℓD seem to be the most adequate. Indeed, they yield near-zero correlation values in the dead rat case, while displaying non-zero values in live rats. However, as shown in equation (2.17), ℓD is difficult to implement. Indeed, it requires the definition of two additional regions that are uncorrelated with the main brain regions of the parcellation and uncorrelated with themselves. Moreover, R cannot be estimated when regions are too small, which is often the case in rat data. It hence becomes apparent that the ℓ_{CA} estimator should be favored.

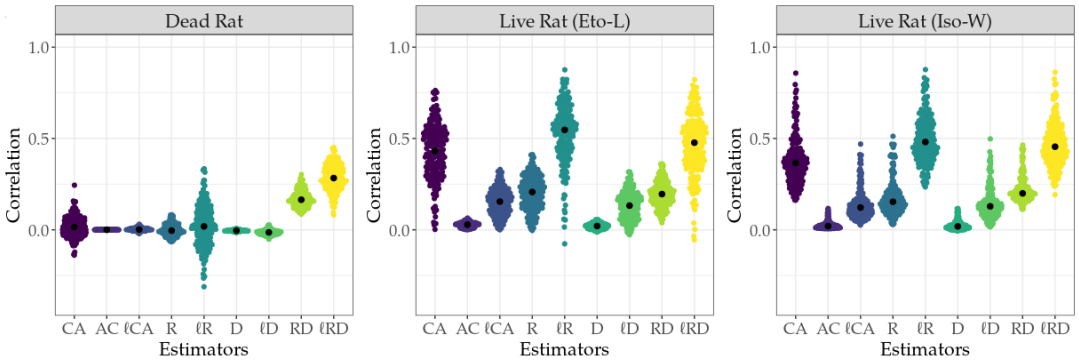


Fig. 2.4 Rat data results. Empirical distribution of the correlation estimators for all pairs of brain regions for a dead and two anesthetized rats, for all proposed estimators. Adapted from Figure 3 in (Achard et al., 2023).

These results are then corroborated on human data. Based on our findings on the rats datasets, we evaluate the performances of the three estimators CA (most common estimator in functional connectivity estimation), AC (familial data estimator, with high dead-live rat similarities) and ℓCA on 100 subjects of the HCP dataset. Figure 2.5 reports the correlation values among all pairs of regions for four randomly selected HCP subjects. Similarly to the rat results, the estimator CA yields the largest inter-correlation values, AC yields values close to zero, while ℓCA values are different from zero, but smaller than that of CA.

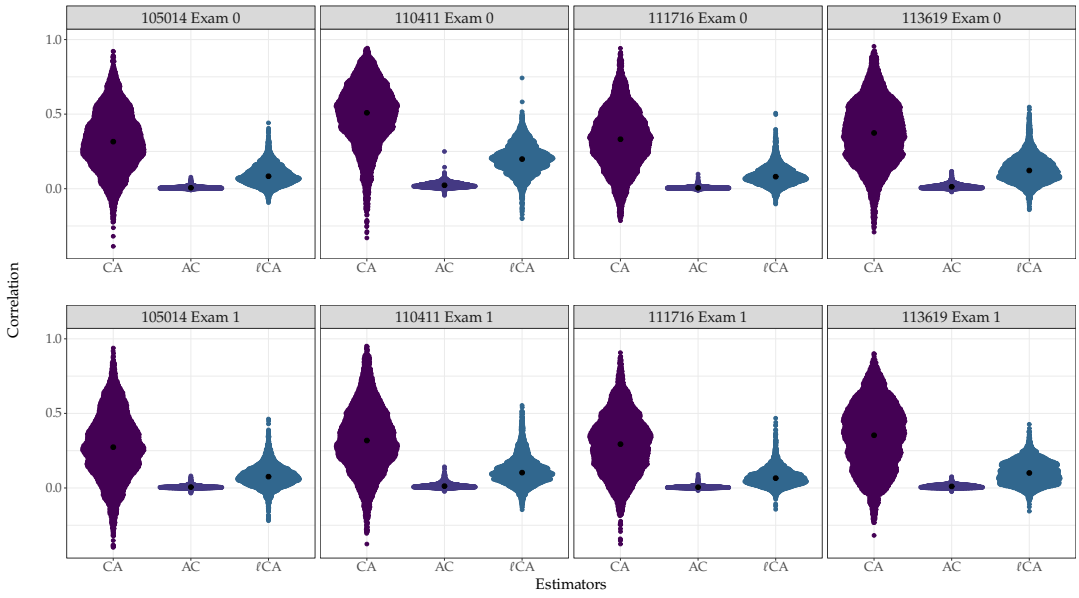


Fig. 2.5 Human data results. Empirical distribution of inter-regional correlations for three selected estimators for all pairs of brain regions for four human subjects. Each subject was scanned twice, on different days. Adapted from Figure 4 in (Achard et al., 2023).

More results on the repeatability and subject identifiability performance of the different estimators are available in the full article (Achard et al., 2023) and underscore the strengths of the ℓCA estimator.

Lastly, we highlight the impact of inter-correlation estimation choices on brain connectivity network configuration. Figure 2.6 shows median differences between the estimators CA and ℓCA in brain space across the HCP subjects. It brings to light systematic spatial variations between the two estimators, exhibiting dorsal posterior hyper-connectivity and corresponding ventral anterior hypo-connectivity for CA compared to ℓCA. The figure also suggests that the largest differences between the two estimators comes between regions that are the largest, highlighting the reduced effect of region size for the ℓCA estimator. The spatial distribution of these differences suggests that caution is in order when examining large-scale resting-state networks derived from the CA estimator, as some apparent topological properties of brain networks, such as modularity, could be driven in part by region size and intra-correlation. In fact, in our experiments, thresholded CA- and ℓCA-based graphs differed in a large proportion of edges, both in rats (around 30%-50% edge differences) and humans (around 30% edge differences).

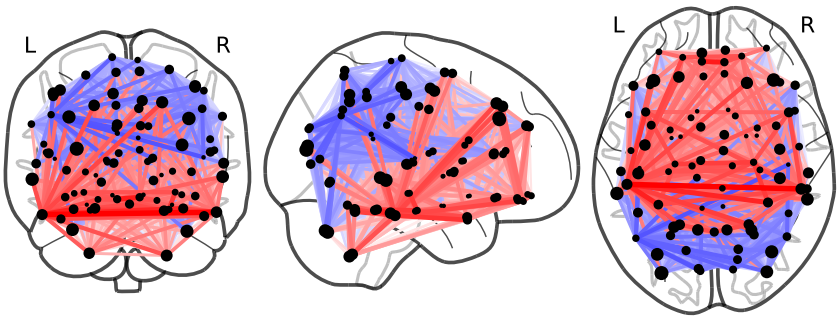


Fig. 2.6 Largest differences between the CA and ℓCA estimators, median over 100 HCP subjects. Only the top 20% differences are shown. Inter-regional correlations are taken in absolute value and rank-transformed prior to computing differences (rank 1 for the strongest correlation, rank 2 for the second-strongest, and so on). Red indicates absolute correlations that are higher for the ℓCA than the CA estimator, while blue indicates the reverse. Node size is proportional to region size in the atlas. Estimator CA on average shows hyperconnectivity in occipital and generally dorsal posterior regions, and hypoconnectivity in frontal, temporal, and general ventral anterior regions. Figure 5 in (Achard et al., 2023), courtesy of Jonas Richiardi.

2.2.4 Conclusion

It emerges from this section that the inter-correlation estimator must be simultaneously robust to intra-correlation and noise. We have seen how the choice of the estimator can have an impact on the topology of the corresponding connectivity networks, and hence on downstream analyses interpretation. It is thus imperative inter-correlation estimation be as reliable as possible. The ℓCA estimator seems to provide the best trade-off among the estimators analyzed in this section, while remaining fairly simple to implement. Nevertheless, it raises the question of neighborhood choice, which we will tackle in Chapter 6.

2.3 Network Inference from Spatio-Temporal Data

2.3.1 Definitions

A **network** or **graph** $G = (\nu, e)$ is defined by its set of *nodes* ν and of *edges* e, which link or connect together pairs of nodes. In *binary* networks, edges simply link together connected regions and an absence of edge implies an absence of connection. In *weighted* networks, edges are associated with a scalar value that often indicates the strength of the connection. The edges of an **undirected** graph define a symmetric relation between the corresponding nodes. On the other hand, edges in **directed** graphs correspond to a

connection that is only valid in one direction, that is, for a pair of distinct nodes $A, B \in \nu$, the edge (A, B) is an ordered pair.

Networks are often represented by their **adjacency matrix**, which encodes connections between nodes. Entries equal to 0 correspond to the absence of an edge between two nodes, while the presence of an edge is represented by a 1 in binary networks and the edge weight in weighted networks. The adjacency matrix of undirected graphs is symmetric.

In the context of spatio-temporal data, nodes often correspond to a variable attached to a spatial location. For instance, consider p voxels of a brain fMRI scan of an individual subject. Connectivity networks can then be constructed at the voxel-level, where nodes correspond to voxels, or at the regional-level, where nodes correspond to brain regions. In the latter case, the p voxels are grouped into J anatomical regions and their signals are typically aggregated to obtain one representative signal per region.

2.3.2 Connectivity Inference

In this section we provide a brief overview of common approaches to infer connectivity networks from spatio-temporal data. Several comprehensive reviews exist, see for instance (Dong et al., 2019; Brugere et al., 2018; De Vico Fallani et al., 2014).

Gaussian Graphical Models (GGM). Also known as **Gaussian Markov Random Fields** (MRF) (see, e.g., Dong et al. (2019); Koller and Friedman (2009)), these approaches infer undirected networks under the assumption the data matrix $\mathbf{X}_{n\times J}$, corresponding to the J region-specific time series of length n, follow a J-dimensional Gaussian distribution. The GGM structure encodes their joint probability distribution. The lack of graph edges between two nodes hence represents the *conditional independence* of the two corresponding random variables. Several techniques to learn such networks exist. They all aim to estimate a sparse *precision matrix*, which is the inverse of the covariance matrix and in fact encodes conditional independence (it can also be used to compute partial correlations). They typically fall into two categories: (i) local inference, where node neighborhoods are learned sequentially, often using *Lasso* (Tibshirani, 1996), or (ii) global inference. In the latter paradigm, the entire precision matrix is estimated at once, often using *graphical Lasso* (gLasso), which is built upon maximum-likelihood estimation and $l1$-regularization (Friedman et al., 2008). In order to control Type I errors, the Lasso requires stability assumptions (Meinshausen and Bühlmann, 2006), which may not

hold in practice. Convergence of gLasso may also be problematic, in addition to being computationally expensive (Mazumder and Hastie, 2012).

Graph Signal Processing. In this paradigm, nodes are formally associated with a signal. Most of these approaches are then based on estimating the graph Laplacian, which is the difference between the degree and adjacency matrices (Dong et al., 2019).

Correlation Networks. Another popular, and less computationally expensive, approach is to estimate correlations, or equivalently covariances, as opposed to the partial correlations obtained with GGMs. Similarly to GGMs, some methods estimate the entire covariance matrix at once, using for instance Lasso regression (Bien and Tibshirani, 2011). However, these approaches are only valid under normality assumption of the underlying variables. Pairwise estimation of the inter-regional correlation, often combined with a thresholding step, is another, and more standard, approach.

Under certain conditions, which may not be straightforward, including normality of the variables and assumptions on the structure of the underlying network, e.g., it must be sparse or acyclic, gLasso and thresholding of covariance matrices have been shown to be equivalent (Fattahi and Sojoudi, 2019).

Brain Functional Connectivity Networks. Typically, each region of the brain, defined by a structural or functional parcellation, is associated to a given set of voxels among the thousands for which a signal is recorded. The idea is then to extract a representative of the set of voxels to attach one time series to each region. When structural atlases are used, the most common approach is to take the average of the voxel time series at each time point. Indeed, almost 70% of papers on PubMed that used the Human Connectome Project dataset to conduct functional-connectivity-related studies in the last five years (e.g., Ogawa (2021); Figueroa-Jimenez et al. (2021); Bolt et al. (2017); Zhang et al. (2016)) use this method.

The literature review was conducted on PubMed using the keywords “brain connectivity graph resting state ‘human connectome project’ ” on September 30, 2021. It was included in (Achard et al., 2023). The search returned 32 papers written between 2014 and 2021. Out of those papers, 5 were not open access and 2 papers were literature reviews, and were not conisdered further. 3 papers were either using seed-based or voxel-to-voxel correlation. Out of the remaining 24 papers 71% (17/24) first averaged voxels before computing the inter-regional correlations and 88% (21/24) employed some kind of spatial aggregation method, including but not limited to averaging over voxels, ICA or dictionary learning.

There are numerous functional connectivity metrics other than Pearson's correlation, including coherence, Granger causality De Vico Fallani et al. (2014), or mutual information. However, it has been shown (Hlinka et al., 2011) that nonlinear connectivity metrics are not substantially more informative than basic linear correlation.

2.3.3 Network thresholding

Once edge weights are estimated, one may want to binarize the network in order to only keep the most connected edges. One may also be interested in only conserving significant edges. Various thresholding methods can be employed. They usually belong to either of these two paradigms: (i) *relative thresholding*, that is estimation of the binary network by extracting a fixed proportion of edges (van den Heuvel et al., 2017), and (ii) *absolute thresholding* where one chooses a fixed threshold that will be applied to all edges. In the context of functional connectivity, these two approaches are implemented in several popular packages, such as the Brain Connectivity toolbox (Rubinov and Sporns, 2010) and CONN (Whitfield-Gabrieli and Nieto-Castanon, 2012), with little guidance on the choice of threshold.

Relative thresholds will always detect the same proportion of edges regardless of the true inter-correlation value, which may be desired in some network comparison applications (van den Heuvel et al., 2017). However, it also means that non-significant edges may be retained if the threshold is too low, or, conversely that significant edges may be missed.

Absolute thresholds can be set more or less arbitrarily. Simple methods based on the distribution of the edge weights, but without theoretical guarantees, are sometimes employed, e.g., (Poli et al., 2015; Boschi et al., 2021). Multiple testing approaches, which aim control to some extent the FDR or FWER by using corrected p-values to determine the absolute thresholds, have also been used in functional connectivity contexts, see, e.g., (Becq et al., 2020b; Váša et al., 2018). Nevertheless, these approaches fail to account for intra-regional spatial dependencies, which as we have seen previously, impact both inter-correlation estimation and testing.

2.3.4 Conclusion

While there are many connectivity estimators available, we focus in this thesis on the inter-correlation. Indeed, it is widely used and is more straightforward to manipulate than many other connectivity metrics. This is all the more important that there is a lack

of non-parametric approaches that take into account both noise intra-regional spatial dependence. This will be explored in Chapter 4 and 6.

2.4 Machine Learning for Connectivity Networks

In this section we introduce various concepts that are needed to perform machine learning tasks on networks, and that will be used in this thesis, mainly in Chapter 7. In these contexts, both nodes and edges can be associated with a vector of features.

2.4.1 Network embedding

Some machine learning methods, such as Random Forests or Support Vector Machines (SVM), which we will describe afterwards, require inputs to be in a vector space. Network or graph embeddings are typically used to transform networks into a latent vector space. A basic approach, which we refer to as *direct* embedding in this thesis, is to stack all edge weights in a vector. Although simple, information from the network structure is lost. *Node2vec* (Grover and Leskovec, 2016) is a widespread network embedding technique that aims to remedy this issue. Based on random-walks, it embeds nodes in a low-dimensional feature space while aiming to preserve node neighborhoods. A more recent approach is the *Feather* embedding, which uses characteristic functions of node features (Rozemberczki and Sarkar, 2020). *Graph2vec* (Narayanan et al., 2017), an adaptation of Node2vec, aims to preserve subgraph patterns. These graph-based embeddings also operate as dimension reduction tools.

2.4.2 Common regression and classification methods

A wide range of methods performing classification and regression task are available. One of the most basic ones is linear regression, but it has a widely acknowledged tendency to overfit in high-dimensional settings. In this section, we will focus on two alternatives that will be used in Chapter 7.

Random Forests provide an easily interpretable approach that comes with guarantees on overfitting. Random Forests are an ensemble learning technique for either classification or regression tasks. The main idea is to combine, most often via averaging or majority vote, predictions from individual *trees* (Breiman, 2001). Each tree is in fact a classifier or regressor. Random forests have previously been used for classification and regression tasks from neuroimaging data such as fMRI (Kesler et al., 2017).

Support Vector Machines (SVM) are another category of common classification and regression method (Cortes and Vapnik, 1995; Platt, 1999), and have been extensively used in brain functional connectivity literature (Dadi et al., 2019). Effective in high dimensions, SVM aim to construct a hyperplane that optimally separate classes. Initially designed for linear separation, nonlinear extensions were proposed by mapping the data into a space where a hyperplane can be identified. SVM have also been extended to perform regression.

Chapter 3

Data Description

We describe in this chapter different ways to generate synthetic multivariate grouped data, as well as the real-world fMRI datasets used throughout this thesis.

3.1 Synthetic Data Generation

In order to evaluate connectivity inference methods, we need to simulate data with known ground-truth connectivity. Synthetic data can be generated either at the *network*- or the *voxel*-level. Synthetic networks are useful when studying empirical binary or weighted networks that have already been estimated. For instance, one could need to compare their properties with that of null networks. Several approaches to generate null networks exist (Váša et al., 2018), ranging from generating a random correlation matrix (Zalesky et al., 2012), to aiming to simulate a weighted connectivity network, to generating random binary networks using random graph models (e.g., Erdős-Rényi). Nevertheless, with network-level data, it is not possible to study the effect of voxel-specific activity, which is what we are most often interested in in this thesis. Alternatively, voxel-level synthetic data is particularly valuable when one is interested in evaluating connectivity inference methods. In fact, it allows to finely control diverse parameters impacting the estimation procedures, such as noise and intra-regional dependence. We will focus on voxel-level synthetic data generation in the remainder of this section. A few models have been developed and implemented to simulate brain activity and the corresponding BOLD signals, such as the SimTB MATLAB toolbox (Erhardt et al., 2012), the NeuRosim R package (Welvaert et al., 2011), or the more complex The Virtual Brain (TVB)[1] software (Ritter et al., 2013; Schirner et al., 2022; Sanz Leon et al., 2013), which accommodates

[1] https://docs.thevirtualbrain.org/

several types of neuroimaging modalities, including fMRI. However, while the overall spatial structure of the correlation can be defined to a certain extent, these methods do not provide direct and simultaneous control over inter- and intra-correlation. We will hence next present a simple and principled approach to generate voxel-level data with user-defined inter- and intra-correlation structure.

3.1.1 Generating a pair of regions

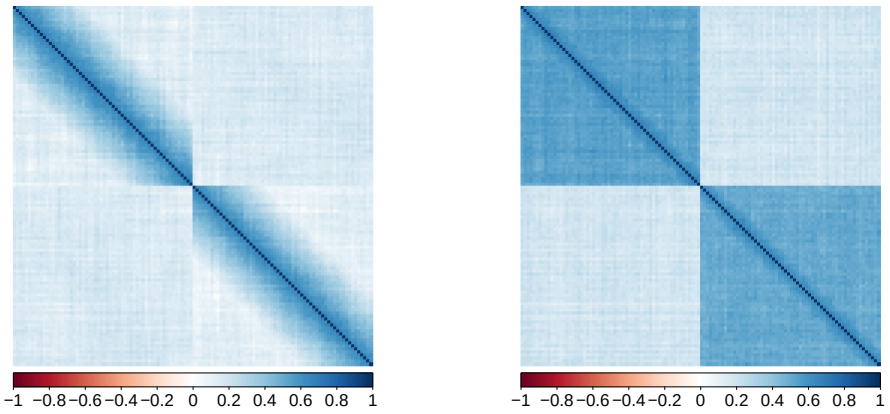
We start by describing how to generate voxel-level signals from a pair of regions.

For each simulation, we simultaneously generate n independent samples of a pair of inter-correlated regions, containing each p_a, p_b intra-correlated variables, respectively. These variables follow a multivariate normal distribution with a predefined covariance structure contaminated by Gaussian noise. Local or global noise can be used. The true inter-correlation is assumed to be constant across all pairs of voxels. The different parameters are chosen to ensure the population covariance matrix of the two regions is positive semidefinite. For instance, one cannot generate a covariance matrix where both intra- and inter-correlation values are low. It has been shown indeed that, when intra- and inter-correlation coefficients are constant within their corresponding region, or pair of regions, respectively, and that region A only contains one variable (such as in mother-siblings studies), $\rho^{A,B} < \eta^B$ is a necessary and sufficient condition for the pairwise population covariance matrix to be positive and semi-definite (Rosner et al., 1977). Different intra-regional covariance structures can be used, and we detail some them in this section.

Constant Covariance Structure. One of the easiest structure is to define the population intra-correlation as constant across all voxels within a given region.

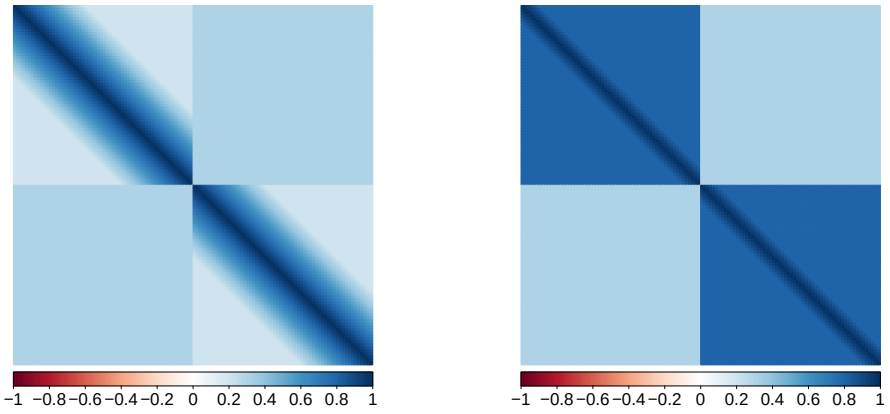
Toeplitz Covariance Structure. A more realistic setting would be to generate 1-dimensional data with a Toeplitz intra-regional covariance structure (later denoted 1D Toeplitz), see for instance (Achard et al., 2023). For each region, intra-correlation is defined such that it decreases as the distance between two variables increases: for any voxel i, i' in region A, $Cor(X_i^A, X_{i'}^A) = \max(1 - |i' - i|/30, \eta_{min}^A)$, where $|i' - i|$ is the uniform norm between voxels i and i', and η_{min}^A the minimal population intra-correlation of a region A. Several experimental settings can be considered by varying the population intra-correlation, inter-correlation and the variance of the noise. The sample pairwise correlation matrices of the observed signals are represented in Figure 3.1 for a low

intra-correlation and a high intra-correlation setting with high noise. The population version of these pairwise correlation matrices are displayed in Figure 3.2.



(a) $\eta^A_{min} = \eta^B_{min} = 0.2$ (b) $\eta^A_{min} = \eta^B_{min} = 0.8$

Fig. 3.1 Sample pairwise correlation matrices (from the 1D Toeplitz model) for different minimum intra-correlation values, with an inter-correlation $\rho^{A,B} = 0.3$ and noise standard deviation $\sigma^A_\varepsilon = \sigma^B_\varepsilon = \sqrt{0.5}$. The diagonal blocks correspond to the intra-correlation of each region.

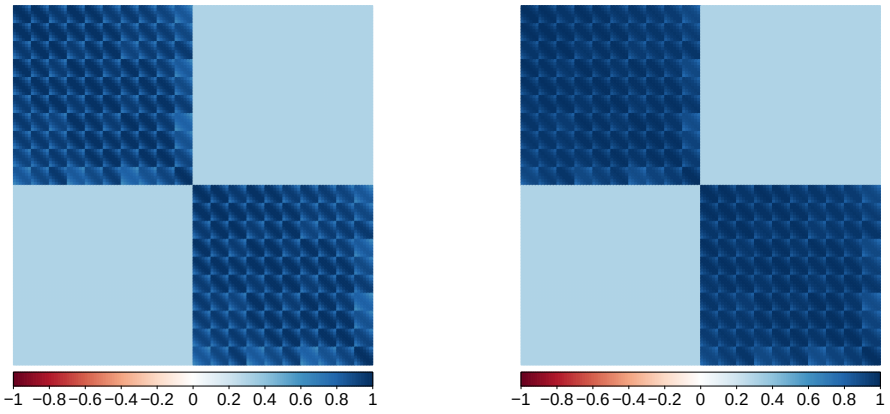


(a) $\eta^{min}_A = \eta^{min}_B = 0.2$ (b) $\eta^{min}_A = \eta^{min}_B = 0.8$

Fig. 3.2 Population pairwise correlation matrices (from the 1D Toeplitz model) for different minimum intra-correlation values, with an inter-correlation $\rho^{A,B} = 0.3$. The diagonal blocks correspond to the intra-correlation of the two regions.

Matérn Covariance Structure. Similarly we simulate 3-dimensional data with a Matérn intra-regional covariance structure that depends on the Euclidean distance (later denoted 3D Matérn) (Ribeiro and Diggle, 2001). In this thesis, we set the smoothness

parameter to $\kappa_A = \kappa_B = 70$ to maintain the positive-definiteness of the input covariance matrix. We then vary the range parameters ϕ_A, ϕ_B and the variance of the noise. The lower the range parameter, the lower the mean intra-correlation. The population pairwise correlation matrices of data generated using a Matérn structure under two settings are shown in Figure 3.3.



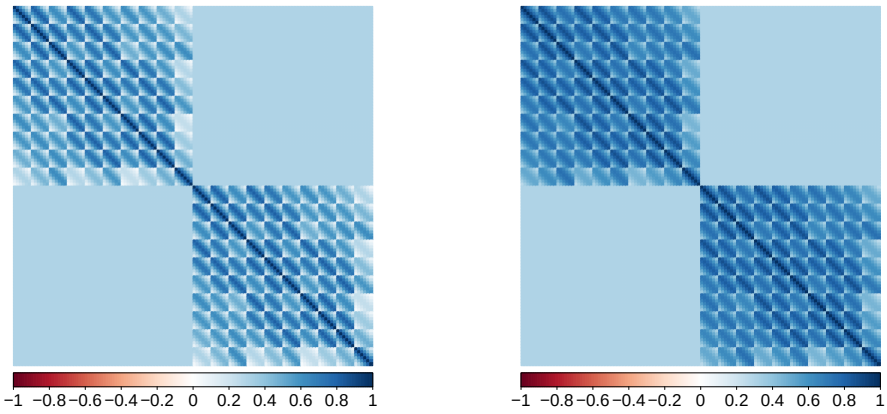
(a) $\phi_A = \phi_B = 0.6$ (b) $\phi_A = \phi_B = 0.8$

Fig. 3.3 Population pairwise correlation matrices (from the 3D Matérn model) for different range values, with an inter-correlation $\rho^{A,B} = 0.3$. The diagonal blocks correspond to the intra-correlation of the two regions.

Spherical Covariance Structure. We generate 3-dimensional data with a spherical intra-regional covariance structure that also depends on the Euclidean distance between voxels (later denoted 3D Spherical) (Ribeiro and Diggle, 2001). We vary the range parameters ϕ_A, ϕ_B and the variance of the noise. The lower the range parameter, the lower the mean intra-correlation. The population pairwise correlation matrices of data generated using a spherical structure under two settings are shown in Figure 3.4.

3.1.2 Generating an arbitrary number of regions

We can also generate synthetic datasets with several inter-connected regions with an underlying ground-truth network. For each dataset, J regions are simultaneously simulated, each region containing p intra-correlated variables following a multivariate normal distribution. n independent samples of each of these variables are obtained. For each region, any intra-correlation structure can be used. The true inter-correlation is assumed to be constant and is set to 0 for region pairs with no ground-truth connection.



(a) $\phi_A = \phi_B = 8$ (b) $\phi_A = \phi_B = 12$

Fig. 3.4 Population pairwise correlation matrices (from the 3D Spherical model) for different range values, with an inter-correlation $\rho^{A,B} = 0.3$. The diagonal blocks correspond to the intra-correlation of the two regions.

3.2 Real-World Data: Brain fMRI

3.2.1 Datasets

Throughout this thesis various resting-state fMRI datasets are used. They are described in this section.

Rat Brain fMRI Dataset. This dataset consists of fMRI scans acquired on both dead and live, anesthetized rats (Becq et al., 2020a,b). The following anesthetics are used: Etomidate (EtoL), Isoflurane (IsoW), Medetomidine (MedL) and Urethane (UreL). After quality control, 3 dead rats and 4 IsoW, 6 EtoL, 2 UreL, and 4 MedL rats are retained. The dataset is freely available at https://doi.org/10.5281/zenodo.2452871. The scanning duration is 30 min with a time repetition of 0.5 s. After preprocessing as described in (Becq et al., 2020b), 51 groups of time series, each associated with its BOLD signal with a number of time points in the order of thousands, were extracted for each rat. They correspond to rat brain regions defined by an anatomical atlas obtained from a fusion of the Tohoku and Waxholm atlases (Becq et al., 2020b).

Human Connectome Project. We also consider 100 healthy subjects from the human connectome project[2] (HCP), WU-Minn Consortium pre-processed (Glasser et al., 2013). Two resting-state fMRI acquisitions on different days are available for each

[2]http://www.humanconnectomeproject.org/

subject. The dataset is freely available at: https://www.humanconnectome.org/study/hcp-young-adult/data-releases. The scanning duration is 14 min and 24 s with a time repetition of 720 ms. A modified AAL template is used to parcellate the brain into 89 regions. The details of the preliminary preprocessing as well as the parcellation are available in (Termenon et al., 2016). Additional subject-specific variables are available, and contain for example age range and family information. Also included are psychometric variables such as personality scores from the *NEO-FFI* Big 5 personality questionnaire, fluid intelligence scores from the *Penn Matrix Test*, which is based on an abbreviated version of the *Raven's matrices*, as well as other cognitive function scores.

PIOP2. The PIOP2 dataset is another open-source resting-state fMRI dataset. It is part of the Amsterdam Open MRI Collection (AOMIC) and information about the preliminary preprocessing can be found in (Snoek et al., 2021). It is freely available at https://openneuro.org/datasets/ds002790/versions/2.0.0. Scans of 223 healthy human subjects are available, as well as subject-specific variables, including age, education category (applied vs academic) and psychometric variables such as the *NEO-FFI* personality questionnaire results and *Raven's matrices* scores. The provided fMRI *derivative fmriprep* data, which is minimally preprocessed, is used to extract voxel-level BOLD signal time series. The same modified AAL template is used to parcellate the brain into 89 regions.

3.2.2 Data preprocessing

Throughout this thesis, we use voxel-level data, unlike most functional connectivity studies—which most often use region-averaged signals. Therefore, open-source data are often not ready for use as is, and further preparation is needed. Starting from the partially preprocessed voxel-level BOLD signals that are provided, we applied the following additional preprocessing steps for each individual scan:

1. Assign each voxel to a unique brain region, based on a predefined anatomical parcellation.
2. Apply a gray-matter mask to each subject individually.
3. Remove voxels with a signal equal to zero at all time points.
4. Perform a wavelet decomposition of the voxel-level time series (Whitcher et al., 2000a). The target frequency band is $(0.06 - 0.13$ Hz$)$.

At the end of the preprocessing pipeline, the size of the rat dataset is over 6 GB, that of the HCP dataset over 160 GB, and that of the PIOP2 over 45 GB. Due to the size of the human datasets, they were uploaded and processed on a computer cluster, which came with its own set of hurdles, including finely tuning computing resource demands to reduce access waiting time. Several of the analyses presented in this thesis were also performed on a cluster.

Note that, in this thesis, we are interested in improving connectivity estimation and analysis once a parcellation has been defined, no matter its quality. Optimizing the choice of parcellation is out of the scope of this thesis. For ease of interpretability, we chose to exclusively use fixed predefined anatomical parcellations in our practical applications. The contributions presented in this thesis could however also be applied to data parcellated using data-driven approaches, such as ICA.

Chapter 4

Correlation-Screening-Based Binary Network Estimation

Large-Scale Correlation Screening under Dependence for Network Inference

We present in this chapter a novel framework to infer binary connectivity networks. Previous work has focused on aggregating data across voxels within predefined regions to infer inter-regional connectivity. However, the presence of intra-correlations has noticeable impacts on inter-correlation detection, and thus edge identification. To alleviate them, we propose to leverage techniques from the large-scale correlation screening literature, and derive simple and practical characterizations of the mean number of correlation discoveries that flexibly incorporate intra-regional dependence structures. A connectivity network inference framework is then presented. First, inter-correlation distributions are estimated. Then, correlation thresholds that can be tailored to one's application are constructed for each edge. Finally, the proposed framework is implemented on synthetic and real-world datasets. This novel approach for handling arbitrary intra-regional correlation is shown to limit false positives while improving true positive rates.

This chapter is based on the following contributions. It is in the process of being submitted to a journal.

Lbath, H., Petersen, A., and Achard, S. (2021). Brain functional connectivity estimation. In *Brain Connectivity Networks: Quality and Reproducibility - Satellite of the Conference on Complex Systems 2021*, Lyon, France Lbath, H., Petersen, A.,

and Achard, S. (2022a). Large-scale correlation screening under dependence for brain functional connectivity inference. In *JSM 2022 - Joint Statistical Meetings*, Washington, United States

4.1 Introduction

Large-scale network inference is a problem inherent to numerous fields, including gene regulatory networks, spatial data studies, and brain imaging. This work is motivated by an application to resting-state brain functional connectivity networks of single subjects. Such networks connect together correlated brain regions, which consist in groups of dependent voxels. These networks are key to providing insights into the diseased or injured brain (Achard et al., 2012a; Richiardi et al., 2013; Malagurski et al., 2019). In this chapter, the terms *region* and *group* will be used interchangeably, the former being associated with the motivating application, and the second with other data sources of similar structure to which the proposed methods also apply.

The goal of this work is to infer a binary network where nodes correspond to regions and edges are present only between nodes that are sufficiently highly correlated. The challenge is two-fold: not only does dependence between voxels within a region impact inter-regional correlation estimation, but it also affects inter-regional correlation threshold estimation, and thus edge detection. We propose a correlation screening approach, and tackle the problem of reliable large-scale correlation discovery between two groups of arbitrarily dependent variables.

In the context of brain functional connectivity, networks are often constructed from functional Magnetic Resonance Imaging (fMRI) data by spatially aggregating blood-oxygen-level-dependent (BOLD) time series within predefined brain regions, e.g. (De Vico Fallani et al., 2014). However, this may lead to overestimation of the inter-regional correlation, or inter-correlation for brevity, e.g., (Halliwell, 1962), and hence incorrect edge detection. We propose a novel network inference framework that leverages, for each pair of regions, inter-correlation distributions instead of aggregation. To obtain the associated binary network we then present a thresholding step based on correlation screening. Existing approaches typically assume variables are independent within their region. Yet, as detailed in this work, any violation of this assumption markedly impacts

inter-correlation discovery, i.e., when the sample inter-correlation coefficient is greater than a given threshold. As will be showcased later, high intra-regional correlation, or intra-correlation for brevity, which corresponds to settings with homogeneous regions, leads to lowered true positive rates (TPR). On the other hand, low intra-correlation, a characteristic of inhomogeneous regions, leads to increased false positive rates (FPR). In (Hero and Rajaratnam, 2011), a theoretical framework that accounts for arbitrary dependence is presented. Their approach is nevertheless very difficult to implement in practice and their empirical evaluation only covers the cases of independence or sparse dependence. We hence introduce simple and practical expressions to characterize the number of discoveries that flexibly incorporate dependence structures. These can then be employed to find a correlation threshold per pair of regions that improves true discovery rates under dependence, while limiting the number of false discoveries. The main steps of the proposed pipeline are depicted in Figure 4.1 and are presented in Sections 4.3, 4.4 and 4.5. We illustrate our work on synthetic data throughout this chapter and demonstrate the effectiveness of our framework on synthetic and real-world brain rat imaging datasets in Section 4.6.

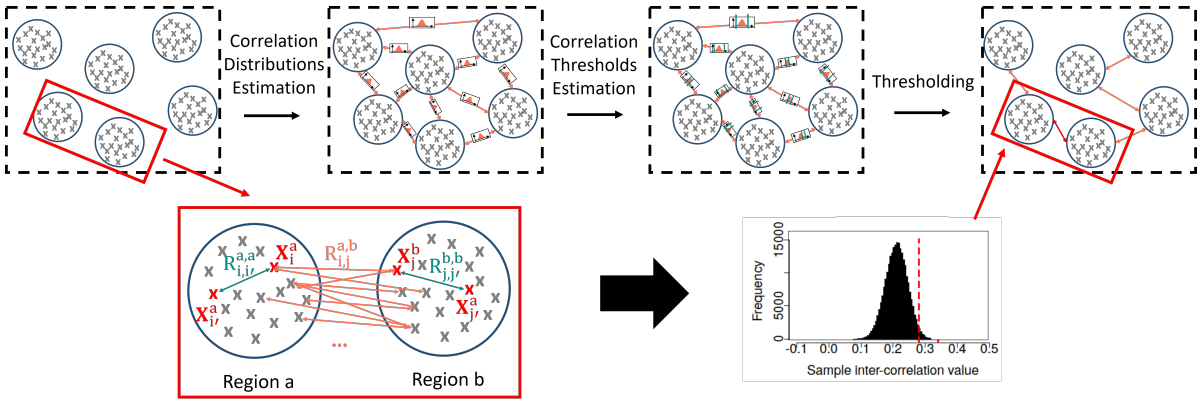


Fig. 4.1 Main steps of our proposed network inference pipeline. Each circle corresponds to a group of variables (represented by crosses). The sample inter-correlation estimation and thresholding steps are detailed for a pair of regions. Some edges were left out to improve readability.

4.2 Preliminaries

In this section we define the data model and its parameters that will be used in the rest of this chapter.

4.2.1 Correlation coefficients

Let A and B be indices of two regions or groups consisting of p_A and p_B random variables, respectively. Denote by $\mathcal{R}_A$ the set of variables in region A and X_i^A the ith random variable in $\mathcal{R}_A$. Assume n independent samples of X_i^A are available and define the corresponding vector $\mathbf{X}_i^A = [X_{i,1}^A, \ldots, X_{i,n}^A]^T$. $\mathcal{R}_B, X_j^B$ and $\mathbf{X}_j^B$ are similarly defined. As an illustration, in the context of brain functional connectivity, X_i^A corresponds to voxel i of brain region A, which is associated with an fMRI BOLD signal time series with n time points. We define *intra-correlation* as the Pearson correlation between each pair of random variables *within* a given region. *Inter-correlation* is the Pearson correlation between pairs of random variables from *two different* regions.

Let $\rho_{ij}^{A,B}$ denote the true population inter-correlation coefficient between X_i^A and X_j^B. For $A \neq B$, define the corresponding sample inter-correlation coefficient

$$R_{i,j}^{A,B} = \frac{\sum_{k=1}^n (X_{i,k}^A - \overline{X_i^A})(X_{j,k}^B - \overline{X_j^B})}{\sqrt{\sum_{k=1}^n (X_{i,k}^A - \overline{X_i^A})^2 \sum_{k=1}^n (X_{j,k}^B - \overline{X_j^B})^2}},$$

with $\overline{X_i^A}$, $\overline{X_j^B}$ the sample means. Denote the probability density, cumulative distribution, and quantile functions of $R_{i,j}^{A,B}$ by $f_{R_{i,j}^{A,B}}$, $F_{R_{i,j}^{A,B}}$, and $F^{-1}_{R_{i,j}^{A,B}}$, respectively. Sample intra-correlation coefficients and their distributions can analogously be defined by choosing $B = A$. Population intra-correlation between X_i^A and $X_{i'}^A$ is denoted by $\eta_{i,i'}^A$. Asymptotic closed-form expressions of the density of correlation can be obtained for Gaussian independently identically distributed (i.i.d.) variables X_i^A, X_j^B (Muirhead, 2005). Note however that this work aims to tackle arbitrary dependence between variables, and in this context, to the best of our knowledge, such explicit formulas have not been derived without defining a parametric dependence structure.

In most of this chapter, and for ease of calculation, we assume the joint distribution of pairs of voxels i, j from a fixed region pair A, B are identically distributed. In such cases, the sample inter-correlation coefficients $R_{i,j}^{A,B}$ are identically distributed, and the i, j subscripts will be dropped, though we emphasize that independence within regions is not assumed.

4.2.2 Synthetic data examples

We illustrate the different concepts introduced in this chapter with data simulated as follows. We consider two regions A and B, both containing p intra-correlated variables following a multivariate normal distribution with a predefined Toeplitz covariance structure.

n independent samples of each of these p variables are generated. We hence obtain data with a block diagonal covariance matrix of size $2p \times 2p$, where each block corresponds to each region. The off-diagonal blocks correspond to the inter-correlation coefficients, which are set to be constant across all pairs of voxels. The diagonal blocks correspond to the intra-correlation coefficients, which follow a Toeplitz dependence structure.

4.3 Inter-Correlation Estimation

4.3.1 Related work

As detailed in Chapter 2, previous works on the estimation of inter-correlations have mostly focused on aggregating variables within predefined regions (De Vico Fallani et al., 2014; Dadi et al., 2019). In the context of brain functional connectivity network inference, some prefer techniques based on independent component analysis (ICA) (Calhoun et al., 2012), while most focus on summarizing all voxels within predefined brain regions by their average, e.g., (Achard et al., 2012a, 2006; Di Martino et al., 2014; Malagurski et al., 2019). However, such approaches suffer from loss of relevant information and can lead to statistical inconsistency and incorrect correlation estimation (Ostroff, 1993). In particular, the estimate of the average of weakly correlated time series, which corresponds to samples of a single variable in our data model, is poor (Wigley et al., 1984). Additionally, it has been observed on small samples that the correlation of averages is different than the average of correlations (Dunlap et al., 1983). This phenomenon can also be easily checked with arbitrary large samples. Furthermore, correlation of averages were empirically observed to overestimate the true correlation (Halliwell, 1962; Achard et al., 2011). Therefore, when correlating regional averages for binary network inference, one will tend to identify spurious edges. Some previous works attempted to improve false positive rate control utilizing multiple testing approaches (Drton and Perlman, 2007). However, in the context of arbitrary dependence structures, such methods cannot be straightforwardly applied. One alternative to aggregation is to measure the similarity, such as the Wasserstein distance or covariance (Petersen and Müller, 2019) between intra-correlation densities. However, this approach is not equivalent to that of the Pearson correlation. Indeed, while the Wasserstein distance may provide a first intuition about how regions are connected, it does not capture as much information about the relationship between the two regions as inter-correlations do.

4.3.2 On the impact of intra-correlation on inter-correlation estimation and detection

We first illustrate how intra-correlation affects the variability of sample inter-correlations in a simplified scenario, before considering a more general case. It has been known for some time in familial data studies that intra-correlations impact inter-correlation estimation (Rosner et al., 1977; Donner and Eliasziw, 1991). In the multivariate normal case, and under the assumption of within-group homoscedasticity, the asymptotic variance of the maximum-likelihood estimator of the inter-correlation, denoted as $R^{A,B}_{MLE}$, was derived by Elston (1975). This estimator showcases similar behavior to the voxel-to-voxel sample inter-correlation coefficients $R^{A,B}_{i,j}$ and will help provide us with a first intuition about the impact of intra-correlation. We need to assume all variables X^A_i, X^B_j have the same true inter-correlation $\rho^{A,B}$ and intra-correlation η^A and η^B. This amounts to saying sample intra- and inter-correlation coefficients are identically distributed within their corresponding group, or pair of groups, respectively. Under these assumptions, and according to Elston (1975), the variance of the maximum-likelihood estimator is:

$$Var(R^{A,B}_{MLE}) = \frac{1}{n}\left[(\rho^{A,B})^2 - \frac{1}{p_A}[1+(p_A-1)\eta^A]\right] \times \left[(\rho^{A,B})^2 - \frac{1}{p_B}[1+(p_B-1)\eta^B]\right] + \frac{(\rho^{A,B})^2}{2n}\left[\frac{p_A-1}{p_A}(1-\eta^A)^2 + \frac{p_B-1}{p_B}(1-\eta^B)^2\right] \quad (4.1)$$

The expression in (4.1) shows that the variance of the sample inter-correlation coefficient explicitly depends on the true intra-correlation coefficients η^A and η^B. When the number of samples n is sufficiently large, the inter-correlation variance in the multivariate normal case hence increases when intra-correlation decreases. This observation implies that for a fixed threshold that does not depend on regional dependency structures, more false positive correlations are likely to be discovered.

This intuition is illustrated in the left hand-side of Figure 4.2 where the true inter-correlation is zero and no positive correlations are expected to be discovered. Conversely, for the same fixed threshold, when the true inter-correlation is positive (cf. right hand-side of Figure 4.2), increased intra-correlations, which lead to lower inter-correlation variance, may lead to decreased number of true positives. This phenomenon is observed regardless of the number of time points or variables (cf. supporting information).

In fact, the impact that intra-correlation distributions have on the spatial average of sample inter-correlations can be quantified even without any distributional assumptions. The following result shows that, when the intra-correlation densities of two regions A

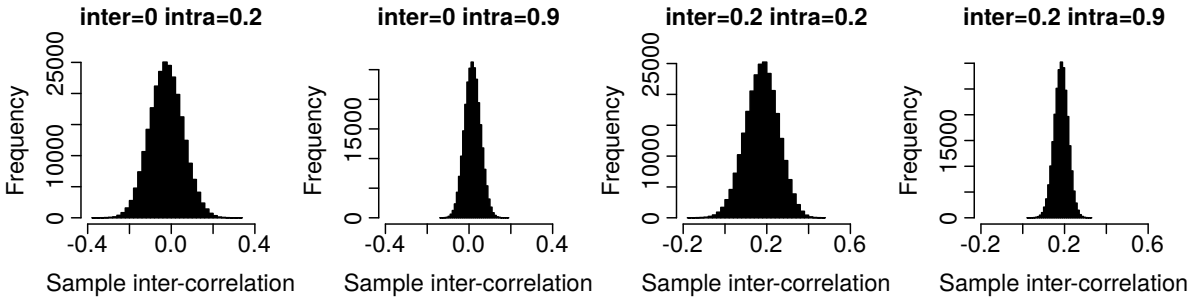


Fig. 4.2 Effect of intra-correlation on sample inter-correlation distribution, for different population inter-correlation values. The correlation samples were computed between all pairs of variables from two groups. Each group contains $n = 150$ samples of $p = 500$ intra-correlated random variables following a multivariate normal distribution with Toeplitz intra-correlation (cf. Section 4.2.2). We can note the higher the intra-correlation, the lower the variance of the inter-correlation distribution.

and B are highly dissimilar, as quantified by a large Wasserstein distance, the average inter-correlation is upper-bounded. In particular, this phenomenon is by no means limited to the Gaussian case. We recall here the definition of the Wasserstein distance between two correlation densities (Petersen and Müller, 2019; Panaretos and Zemel, 2019): $d_W^2(f_{R^{A,A}}, f_{R^{B,B}}) = \int_0^1 [F^{-1}_{R^{A,A}}(c) - F^{-1}_{R^{B,B}}(c)]^2 dc$. The full proof is available in Appendix 4.A.

Proposition 4.3.1. *For any regions A, B with p_A, p_B voxels, respectively, and $\mathcal{R}_A, \mathcal{R}_B$ the corresponding voxel sets, if there exists $d \in \mathbb{R}^+$ such that*

$$d_W^2(f_{R^{A,A}}, f_{R^{B,B}}) \geq \min_{c\in[0,1]} \left(F^{-1}_{R^{A,A}}(c) - F^{-1}_{R^{B,B}}(c)\right)^2 \geq d,$$

then, $\overline{R^{A,B}} = \frac{1}{p_A p_B} \sum_{i\in\mathcal{R}_A} \sum_{j\in\mathcal{R}_B} R^{A,B}_{i,j} \leq 1 - \frac{\sqrt{d}}{2}$.

4.3.3 Our proposed approach: inter-correlation distribution estimation

As previously discussed, aggregating variables within regions to estimate inter-correlation leads to loss of information and incorrect edge detection during the binary network inference step. We have also brought to light the importance of taking intra-correlation into account when manipulating inter-correlations. In addition, all the previously cited approaches that aim to infer a binary network where nodes are groups of variables only provide a single correlation threshold to be applied to all pairs of regions. In this chapter,

we propose to derive a correlation threshold specific to each pair of regions to better harness the particularities of the regional dependence structures. Instead of averaging variables within regions, we hence propose to estimate the distribution of correlations measured between all pairs of variables from two different regions. We then obtain an inter-correlation distribution per pair of region, which then needs to be thresholded. To that end, we propose to leverage correlation screening. In that paradigm, an edge is said to be detected in the associated binary graph if a sufficient number of sample inter-correlation coefficients of the corresponding pair of regions are large enough. In the following sections, we derive simplified expressions of the number of discoveries to propose a reliable method to threshold these inter-correlation distributions.

4.4 Characterization of the Number of Discoveries Under Dependence

Correlation screening (Hero and Rajaratnam, 2011), or independence screening (Fan and Lv, 2008b), is often used in variable or feature selection problems. In such approaches, the goal is to discover sufficiently highly correlated variables. A practical method consists in defining a correlation threshold, above which correlation coefficients, and their associated variables, are said to be *detected* or *discovered*. Nonetheless, in high dimension, such approaches may suffer from a high number of false discoveries. In (Hero and Rajaratnam, 2011), the authors aim to mitigate this issue in the following way. They first propose the following maximum-based definition of the number of discoveries, pertaining to inter-correlation coefficients, with $\phi_{ij}^{AB}(\rho) = \mathbb{1}(|R_{i,j}^{A,B}| > \rho)$ for all voxels i, j in regions A, B and correlation threshold $\rho \in [0, 1]$:

$$N^{AB}(\rho) = \sum_{i=1}^{p_A} \max_{j=1,\dots,p_B} \phi_{ij}^{AB}(\rho). \tag{4.2}$$

The authors provide as well an approximation of the expected number of discoveries $E[N^{AB}]$ that depends on the number of variables p, the number of samples n and a function of the joint distribution of a transformation of the variables. Then they employ the derived formula to compute critical threshold values based on a phase transition approach. Furthermore, the expected number of discoveries is used to control the number of false discoveries. It is then all the more essential to have an expression of the number of discoveries that is both interpretable and can easily be theoretically and empirically utilized. However, the expression for $E[N^{AB}]$ derived in (Hero and Rajaratnam, 2011),

which depends on joint distributions, is difficult to compute, especially for single subject analysis where we have access to only a single sample of each signal. We hence provide simplified explicit expressions of the mean number of discoveries that still harness information contained in the intra-correlation distributions.

4.4.1 Maximum-based expression: N^{AB}

Empirically, intra-correlation has an impact on N^{AB} and its average (cf. Figure 4.3, left). Indeed, for a given inter-correlation threshold, the smaller the intra-correlations, the larger the number of discoveries. This is in accordance with the observations from Figure 4.2. Additionally, we can remark that in this example the true inter-correlation is zero. Thus any discovery is a false positive. As the intra-correlation increases, a lower correlation threshold is then sufficient to maintain similar levels of false discoveries.

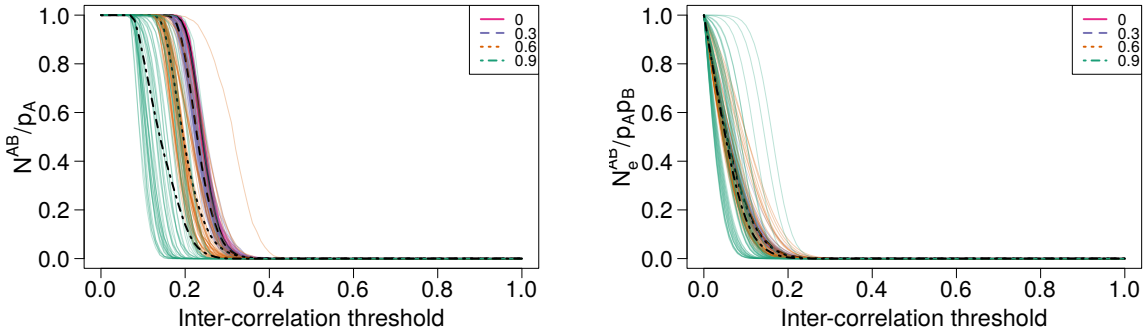


Fig. 4.3 Normalized number of discoveries, N^{AB}/p_A (**Left**) and N_e^{AB}/p_Ap_B (**Right**) as a function of the inter-correlation threshold for data simulated as described in Section 4.2.2, with $p_A = p_B = 500, n = 150$, true inter-correlation $\rho^{A,B} = 0$ and Toeplitz intra-correlation with varying minimal intra-correlations. For each of the four intra-correlation values, 50 datasets were simulated and used to compute the number of discoveries (the colored curves), and their average (the black dotted curves). N^{AB}/p_A decreases as the intra-correlation increases, while N_e^{AB}/p_Ap_B does not seem to be much impacted on average, even though its variability seems to increase with the intra-correlation value.

From (4.2), we can also conclude that for any $\rho \in [0, 1]$,

$$E[N^{AB}(\rho)] = \sum_{i=1}^{p_A} \left(1 - F_{|R_{i,1}^{A,B}|,\dots,|R_{i,p_B}^{A,B}|}(\rho, \dots, \rho)\right), \tag{4.3}$$

with $F_{|R_{i,1}^{A,B}|,\dots,|R_{i,p_B}^{A,B}|}$ the joint distribution of the absolute values of the corresponding correlation coefficients, which will inherently take into account dependence structures between the inter-correlation coefficients. However, joint distributions are complicated to estimate and manipulate. Let $\tilde{\nu}^{AB} = E[N^{AB}]/p_A$. We then propose an approximate

expression of $\tilde{\nu}^{AB}$ that depends on the distribution of inter-correlations, and that is exact under some particular assumptions (cf. Appendix 4.C):

$$\nu^{AB}(\rho) = 1 - F_{|R^{A,B}|}(\rho)^{p_B}, \quad \rho \in [0,1]. \tag{4.4}$$

We can also derive the following inequality.

Proposition 4.4.1. *Consider two regions A and B and a correlation threshold $\rho \in [0,1]$. If all variables X_i^A and X_i^B in both regions follow a normal distribution and their sample inter-correlation coefficients are identically distributed, then for sufficiently large n,*

$$\nu^{AB}(\rho) \geq \tilde{\nu}^{AB}(\rho) \tag{4.5}$$

Proof. As defined in (Lehmann, 1966), the random variables $T_1, T_2, \ldots, T_p$ are Positively Quadrant Dependent (PQD) if for any positive number $t_1, t_2, \ldots, t_p$,

$$P\left(\bigcap_{k=1}^{p} T_k \leq t_k\right) \geq \prod_{k=1}^{p} P\left(T_k \leq t_k\right). \tag{4.6}$$

Under the assumption X_i^A, X_j^B are normal for all i, j, the distribution of their sample correlation coefficients $R_{i,j}^{A,B}$ is asymptotically normal, e.g., (Ruben, 1966; Hotelling, 1953). Hence, according to Theorem 1 in (Šidák, 1967), when n is large enough, $|R_{i,j}^{A,B}|$ are PQD. We can then note that equation (4.6) is equivalent to $E\left[\prod_{k=1}^{p} \mathbb{1}(|T_k| \leq t_k)\right] \geq \prod_{k=1}^{p} E[\mathbb{1}(|T_k| \leq t_k)]$. Under the assumption the sample correlation coefficients are identically distributed, and setting $t_k = \rho$ and $T_k = R^{A,B}$ for all variables, we can thus write:

$$\begin{aligned} p_A \cdot \nu^{AB}(\rho) &= \sum_{i=1}^{p_A} \left(1 - \prod_{j=1}^{p_B} E[\mathbb{1}(|R^{A,B}| \leq \rho)]\right) \\ &\geq \sum_{i=1}^{p_A} \left(1 - E\left[\prod_{j=1}^{p_B} \mathbb{1}(|R^{A,B}| \leq \rho)\right]\right) = E[N^{AB}(\rho)] = p_A \cdot \tilde{\nu}^{AB}(\rho). \end{aligned}$$

This concludes the proof. □

This result ensures ν^{AB} will provide thresholds that are at least as conservative as that of $\tilde{\nu}^{AB}$. Moreover, the assumptions needed in Proposition 4.4.1 are often reasonable in practice—and notably in functional brain connectivity applications where the signals associated with each voxel can appropriately be transformed (Whitcher

et al., 2000b). In addition, ν^{AB} can be estimated by $\widehat{\nu}^{AB}(\rho) = 1 - \widehat{F}_{|R^{A,B}|}(\rho)^{p_B}$, where $\widehat{F}_{|R^{A,B}|}(\rho) = \frac{1}{p_B p_A} \sum_{i=1}^{p_A} \sum_{j=1}^{p_B} \mathbb{1}(|R_{i,j}^{A,B}| \leq \rho)$ is the empirical cumulative distribution function (ecdf) of $|R^{A,B}|$. The result stated above can then be empirically observed in Figure 4.4. The curve of $\widehat{\nu}^{AB}$ as a function of thresholds is also particularly close to that of the empirical values of $\tilde{\nu}^{AB}$ as long as both the inter-correlation and the intra-correlation of region B are not too high. $\widehat{\nu}^{AB}$ provides hence an approximation for the normalized expected number of discoveries that is easier to compute, while still accounting for the inter-correlation distribution. Moreover, in practice, the ecdf of the inter-correlation coefficients can be shown to depend on the intra-correlation structure (Azriel and Schwartzman, 2014) and hence allows us to account for it.

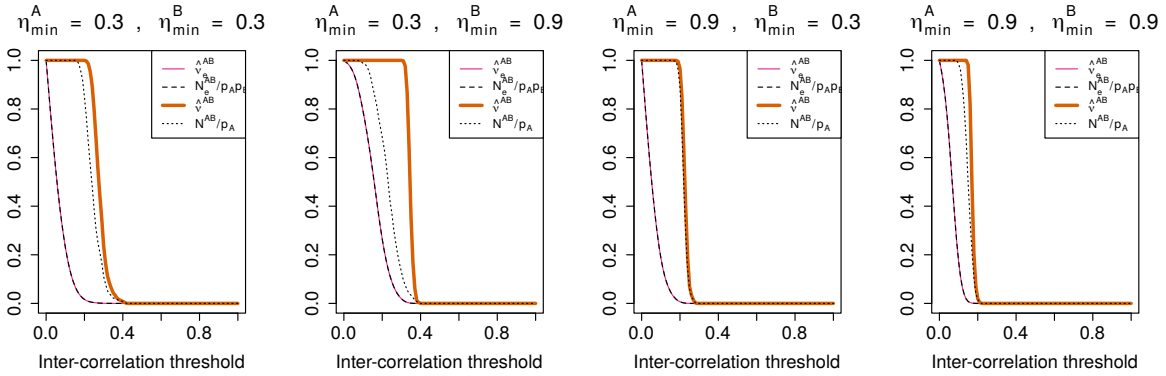


Fig. 4.4 Normalized number of discoveries $\widehat{\nu}^{AB}$, $\widehat{\nu}_e^{AB}$, N^{AB}/p_A and $N_e^{AB}/p_A p_B$ as a function of the correlation threshold for different minimal intra-correlation values $\eta_{min}^A, \rho_{min}^{BB}$ and a zero inter-correlation for data simulated as described in Section 4.2.2, with $p_A = p_B = 500, n = 150$. We can note $\widehat{\nu}_e^{AB} \leq \widehat{\nu}^{AB}$. We can also observe the asymmetric behavior of N^{AB}/p_A and how it is close to $\widehat{\nu}^{AB}$ when $\eta_{min}^A \gg \eta_{min}^B$, unlike when $\eta_{min}^B \gg \eta_{min}^A$.

4.4.2 Sum-based expression: N_e^{AB}

We now present another, and more intuitive, way to characterize the number of discoveries N_e^{AB} (Hero and Rajaratnam, 2011). It represents the total number of discoveries and will enable us to propose less conservative thresholds:

$$N_e^{AB}(\rho) = \sum_{i=1}^{p_A} \sum_{j=1}^{p_B} \phi_{ij}^{AB}(\rho), \quad \rho \in [0,1]. \tag{4.7}$$

However, it is not straightforward to derive a critical correlation threshold from this expression. We propose the simplified expression below:

$$\widehat{\nu}_e^{AB}(\rho) = 1 - \widehat{F}_{|R^{A,B}|}(\rho), \quad \rho \in [0,1]. \tag{4.8}$$

We can also remark in Figure 4.4 that $\widehat{\nu}_e^{AB}$ and $N_e^{AB}/p_A p_B$ look indistinguishable.

4.4.3 Link between N_e^{AB} and N^{AB}

We have presented so far two ways to characterize the number of discoveries. We will now discuss how they relate to one another. We can remark the following inequality.

Proposition 4.4.2. *For all* $\rho \in [0,1]$,

$$\widehat{\nu}_e^{AB}(\rho) \leq \widehat{\nu}^{AB}(\rho). \tag{4.9}$$

The proof is straightforward and can be found in Appendix 4.B. This result can notably be observed in Figure 4.4. Thus $\widehat{\nu}^{AB}$ is more conservative than $\widehat{\nu}_e^{AB}$, which in some circumstances may be desirable. Nonetheless, when the inter-correlation is zero we expect no discoveries. In this case, a critical correlation threshold can hence be defined as the minimum correlation such that the number of discoveries is zero. In such cases, using $\widehat{\nu}_e^{AB}$ then seems to be preferable, since it provides a lower correlation threshold for a similar number of false discoveries.

4.5 Correlation Threshold Definition

Now we have better characterized the number of discoveries, we can use it to construct correlation thresholds tailored to one's data and that ensure, to a certain extent, a restricted number of false discoveries and improved number of true discoveries. We present in this section two possible correlation threshold definition approaches. The idea behind correlation threshold definition is to ensure that it is very unlikely for any discovery to correspond to a correlation value that could have happened at random, which amounts to a setting where the true inter-correlation is zero, which may not be true in practice. Surrogate data defined such that the population inter-correlation is zero can hence be utilized to estimate the correlation thresholds. We denote $\widehat{F}^{-1}_{0,|R^{A,B}|}$ the corresponding quantile function.

FWER-based threshold. Correlation thresholds with family-wise error rate (FWER) theoretical control can be derived for specific dependence structures. These approaches control the probability of making at least one false discovery. Analogously to Proposition 2 in (Hero and Rajaratnam, 2011), it can be shown, under a weak dependence condition, that N_e^{AB} converges to a Poisson random variable when $p_A, p_B \to \infty$ and $P(N_e^{AB} > 0) \to 1 - \exp\left(-E[N_e^{AB}]\right)$. Our proposed expression $\widehat{\nu}_e^{AB}$ can then be used to compute correlation thresholds ρ_α^{AB} that guarantee a FWER at level α. Nevertheless, the weak dependence assumption upon which this approach hinges is often not reasonable in practice.

Quantile-based threshold. The correlation threshold can also be defined such that the False Positive Rate (FPR) is guaranteed to be less than a given level α. The FPR is the ratio between the number of false positives (FP) and the total number of ground truth negatives, that is $p_A \cdot p_B$ in the $\rho^{A,B} = 0$ case. Controlling the FPR at level α is thus equivalent to ensuring the number of discoveries is less than $FP = \alpha \cdot p_A \cdot p_B$. Since in our setting ($\rho^{A,B} = 0$) any discovery is a false positive, we can set $\widehat{\nu}_e^{AB} \cdot p_A \cdot p_B = \alpha \cdot p_A \cdot p_B$, which leads to the threshold $\rho_{q,\alpha}^{A,B} = \widehat{F}^{-1}_{0,|R^{A,B}|}(1-\alpha)$. We can remark that when $\alpha = 0$, the chosen threshold is larger than any of the observed absolute correlations, ensuring there will be no discoveries. Additionally, this threshold will depend on the intra-correlation, as does the ecdf (Azriel and Schwartzman, 2014). We can also remark this threshold guarantees a FWER at level $\alpha = 0$ under the previous weak dependence assumption. $\widehat{\nu}^{AB}$ can also be used to similarly derive a critical threshold, although stringent conditions on the region sizes would then need to be verified when $\alpha \neq 0$.

Numerical results. We compare in Figure 4.5 the two correlation thresholds defined above with two other approaches: the critical thresholds ρ_{hero} obtained in (Hero and Rajaratnam, 2011), and a simple method where the threshold is set to $\rho_{poli}^{AB} = \widehat{\mu}_0^{AB} + \widehat{\sigma}_0^{AB}$ where $\widehat{\mu}_0^{AB}$ and $\widehat{\sigma}_0^{AB}$ are the sample mean and standard deviation of the sample inter-correlation of the surrogate data (Poli et al., 2015). We observe that, when intra-correlation is lower than 0.5, both our proposed approaches provide similar thresholds to Hero and Rajaratnam (2011). Nonetheless, our FWER- and quantile-based methods provide less conservative thresholds when the intra-correlation is high. The lower thresholds imply an increase in the true positive rate. We can note as well that ρ_{poli}^{AB} is lower than all three other thresholds for all intra-correlation values, which could lead to a large number of false discoveries, as will be shown in the next section. It is also decreasing when intra-correlation increases in accordance with the observations about

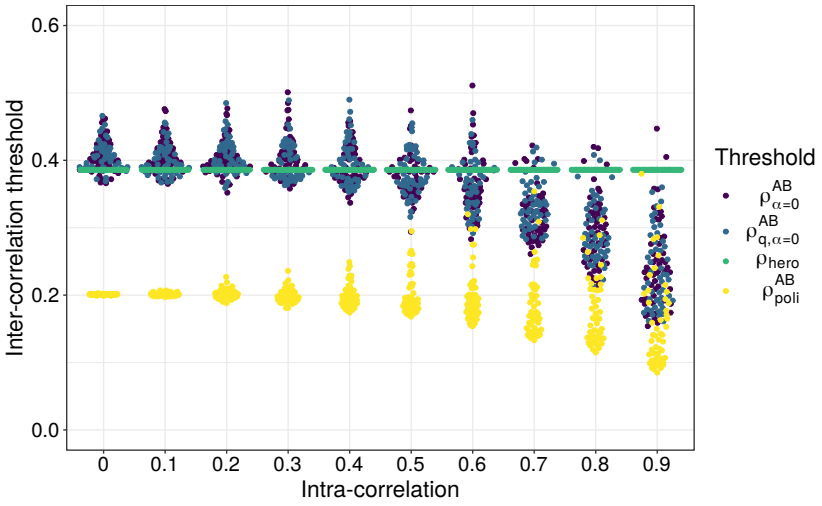


Fig. 4.5 Comparison of different critical thresholds for 50 replicates of data simulated as described in Section 4.2.2 with $p = 150$, $n = 100$, $\rho^{A,B} = 0$ and varying constant intra-correlation values. The FWER- and quantile-based thresholds ρ^{AB}_{α} were computed for $\alpha = 0$.

the effect of intra-correlation on the distribution of sample inter-correlation in Section 4.3.

4.6 Network Inference Results

In this section we provide an illustration of our network inference approach on synthetic and real-world data and compare it to several methods.

Comparison to other methods. As mentioned in Section 4.3.1 and Chapter 2, most single-subject fMRI studies that use anatomical parcellations estimate the inter-correlation by computing the correlation coefficient between spatial regional averages of the signals. We will refer to the correlation of averages approach by CA and our proposed correlation screening method by CS. As detailed in Chapter 2, various methods can then be employed to define the correlation thresholds. They usually belong to either of these two paradigms: (i) relative thresholding, that is, estimation of the binary network by extracting a fixed proportion of edges (van den Heuvel et al., 2017), and (ii) absolute thresholding where one chooses, more or less arbitrarily, a fixed threshold that will be applied to all edges—as opposed to our proposed edge-specific thresholds. These two approaches are implemented in several popular packages, such as the Brain Connectivity toolbox (Rubinov and Sporns, 2010) and CONN (Whitfield-Gabrieli and Nieto-Castanon,

2012), with little guidance on the choice of threshold. Relative thresholds will always detect the same proportion of edges regardless of the true inter-correlation value, and as such can be disregarded in this work. Both the thresholds proposed by Hero and Rajaratnam (2011) and Poli et al. (2015) are absolute thresholds. The latter was used in (Boschi et al., 2021) to threshold CA-based functional connectivity. In (Becq et al., 2020b), the authors apply to all edges of a CA-based functional connectivity network a fixed threshold ρ_{becq} determined according to a multiple testing approach (see Appendix G of their work for more details). In practice, the values of ρ_{becq} are close to that of ρ_{hero}. The thresholds ρ^{AB}_{poli}, ρ^{AB}_{α} and $\rho^{AB}_{q,\alpha}$ are estimated using surrogate data where the true inter-correlation is zero and the intra-correlation is constant and equal to the average sample intra-correlation.

Synthetic data results. We generated synthetic datasets with ten inter-connected regions. For each dataset, 10 regions are simultaneously simulated, each region containing $p = 150$ intra-correlated variables following a multivariate normal distribution. 100 independent samples of each of these variables are obtained. For each region, a Toeplitz intra-correlation is used with the same minimal intra-correlation value across all ten regions. Further details about data generation are presented in Chapter 3.

There are 41 true positive (constant true inter-correlation $\rho^{A,B} = 0.2$) and 4 true negative edges (constant inter-correlation $\rho^{A,B} = 0$) in the ground-truth network. The different simulation parameters are chosen to ensure the population covariance matrix of the ten regions is positive semidefinite. To identify the edges of the binary network, the pairwise thresholds are applied to the corresponding distributions of the absolute value of the sample inter-correlation. In particular, pairs of regions where the inter-correlation is larger than the threshold with a probability at most 0.05 are not identified as edges.

Table 4.1 displays the false positive and true positive rates (FPR and TPR, respectively) of the different methods, for varying minimal intra-correlation values. The FPR and TPR are defined as follows: FPR = FP/(FP+TN) and TPR = TP/(TP+FN). FN stands for false negatives (i.e., an edge is undetected when it actually exists). FPR is expected to be close to 0 and TPR to 1. Results in Table 4.1 showcase that, as expected from the previous section, using ρ^{AB}_{poli} leads to high FPRs, while ρ_{hero} and ρ_{becq} lead to decreasing TPRs as intra-correlation increases. Additionally, while the FPR is slightly increased, correlation screening methods with FWER- or quantile-based thresholds markedly improve the TPR when the intra-correlation is high, and should be preferred in that case. Indeed, when intra-correlation is 0.9 all other methods (except CS + ρ^{AB}_{poli}) have a TPR close to zero, while the FWER- and quantile-based thresholds have

a TPR close to 0.7. The CS + ρ^{AB}_{poli} method displays a FPR of 1 for all intra-correlation values and should thus to be avoided. Since the FWER- and quantile-based thresholds are empirically equivalent, from now on we will be using $\rho^{AB}_{q,\alpha=0}$, which, unlike $\rho^{AB}_{\alpha=0}$, has a theoretical control over false positives that is valid for any dependence structure.

Table 4.1 Comparison of the mean (standard deviation) TPR and FPR of several network inference methods on synthetic data across 100 repetitions. The ground-truth networks consist of 10 regions, with 4 real negative edges (true inter-correlation $\rho^{A,B} = 0$), 41 real positive edges (true inter-correlation $\rho^{A,B} = 0.2$), five minimal intra-correlation values η^A_{min}, $n = 100$, and $p = 150$.

	$\eta^A_{min} = 0.5$		$\eta^A_{min} = 0.6$		$\eta^A_{min} = 0.7$		$\eta^A_{min} = 0.8$		$\eta^A_{min} = 0.9$	
Method	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR
CA + ρ^{AB}_{poli}	0.51 (0.23)	0.97 (0.02)	0.46 (0.22)	0.90 (0.07)	0.45 (0.25)	0.69 (0.12)	0.43 (0.26)	0.33 (0.14)	0.30 (0.24)	0.04 (0.03)
CA + ρ_{becq}	0.13 (0.13)	0.47 (0.20)	0.04 (0.09)	0.22 (0.13)	0.03 (0.08)	0.11 (0.09)	0.01 (0.04)	0.05 (0.06)	0.01 (0.04)	0.04 (0.04)
CS + ρ^{AB}_{poli}	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
CS + ρ_{hero}	0.03 (0.08)	0.13 (0.12)	0.03 (0.08)	0.13 (0.11)	0.04 (0.09)	0.12 (0.09)	0.02 (0.07)	0.11 (0.09)	0.01 (0.06)	0.09 (0.07)
CS + $\rho^{AB}_{\alpha=0}$	0.06 (0.10)	0.23 (0.17)	0.08 (0.12)	0.32 (0.17)	0.13 (0.13)	0.42 (0.17)	0.15 (0.16)	0.52 (0.16)	0.23 (0.17)	0.68 (0.13)
CS + $\rho^{AB}_{q,\alpha=0}$	0.06 (0.11)	0.23 (0.18)	0.07 (0.11)	0.32 (0.17)	0.14 (0.13)	0.42 (0.16)	0.15 (0.16)	0.53 (0.16)	0.25 (0.18)	0.67 (0.13)

Real-world data results. We applied our framework on functional Magnetic Resonance Imaging (fMRI) data acquired on both dead and live rats, anesthetized using Isoflurane (Becq et al., 2020a,b). The scanning duration was 30 minutes with a time repetition of 0.5 second so that 3600 time points were acquired. Additional information about this dataset are documented in Chapter 3. After preprocessing as explained in (Becq et al., 2020b), based on an anatomical atlas, 51 groups of time series, corresponding to the rat brain regions, were extracted for each rat. Due to insufficient signal, four regions were excluded. Each time series captures the functioning of a given voxel. The dead rats provide experimental data where the ground-truth network is empty. Indeed, no legitimate functional activity should be detected, whereas for the live rat under anesthetic, we expect non-empty graphs as brain activity keeps on during anesthesia. We can note that no ground-truth is available for the live rat networks. As expected, the networks of the dead rats estimated using our proposed correlation screening method are empty, i.e. it does not detect any false positive edges, with the exception of one edge in one rat. However, the CA + ρ^{AB}_{poli} approach Boschi et al. (2021); Poli et al. (2015) detects over 300 false positive edges, and Becq et al. (2020b) (later denoted B2020) detects between one and four false positive edges (cf. Table 4.2). While our approach is more conservative than the other two, important edges are still detected in the live rats, mainly in motor regions (M1 and M2) and somatosensory regions (S1 and S2), as shown for instance in Figure 4.6.

Table 4.2 Comparison of the number of edges in the networks obtained via our proposed network inference approach and two methods from the literature (B2020 and CA + ρ^{AB}_{poli}) for dead (**Left**) and live (**Right**) rat brain fMRI data. In the dead rat brain networks, any detected edge is a false positive.

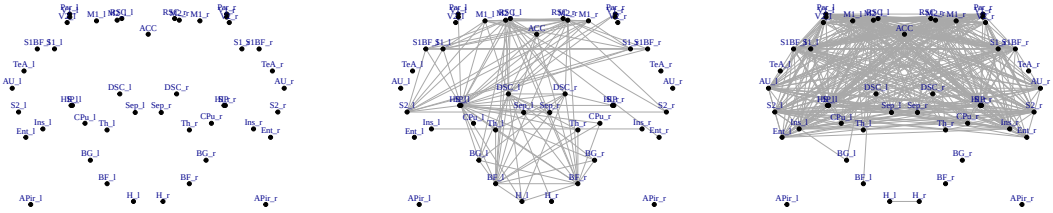
	NUMBER OF EDGES		
DEAD RATS ID	**CS + $\rho^{\mathbf{AB}}_{\mathbf{q},\alpha=0}$**	B2020	CA + ρ^{AB}_{poli}
20160524_153000	**1**	4	316
20160609_161917	**0**	4	317
20160610_121044	**0**	1	325

	NUMBER OF EDGES		
LIVE RATS ID	**CS + $\rho^{\mathbf{AB}}_{\mathbf{q},\alpha=0}$**	B2020	CA + ρ^{AB}_{poli}
20160615_103000	**25**	647	820
20160614_095825	**411**	847	910
20160615_121820	**116**	477	692
20160421_133725	**83**	591	910

4.7 Discussion

We have presented a novel approach to infer connectivity networks when nodes represent groups of correlated variables. We have formally established the importance of leveraging dependence structures to reliably discover inter-correlations. Our method consists in estimating, for each pair of groups, an inter-correlation distribution before deriving a tailored threshold based on a correlation screening approach. In particular, we proposed simplified expressions for the mean number of discoveries that allow for easier theoretical and empirical manipulation, and flexibly take into account dependence within groups. Motivated by a real-world application, we have demonstrated the feasibility of our approach on a real dataset of rat brain images.

This work has several possible theoretical extensions. First, while we provide a method with theoretical FPR control for any setting, we provide theoretical FWER control guarantees only under a weak dependence assumption. Nevertheless, this assumption is often unrealistic and relaxing it is difficult and would be an interesting direction to explore. Additionally, FWER approaches may sometimes be too conservative. On the



(a) Dead rat (20160609_161917) (b) Live rat (20160615_121820) (c) Live rat (20160614_095825)

Fig. 4.6 Brain functional connectivity network of a dead and two live rats (anesthetized with Isofluorane) inferred using our proposed correlation screening framework with the quantile-based threshold (CS + $\rho^{AB}_{q,\alpha=0}$).

other hand, the false discovery rate (FDR) enables a control of the average number of FPs, which is often sufficient. A procedure to define correlation thresholds was proposed in (Cai and Liu, 2016) that leverages a quantity linked to the sum-based mean number of discoveries $E[N_e^{AB}]$. While they provide FDR control, it is only valid under some particular dependence conditions. It would nonetheless be interesting to extend their work to arbitrary dependence.

In this chapter, the aim was to reliably detect one edge at a time. It would then be interesting to build upon the proposed edge-centric correlation thresholds to develop a multiple testing framework so as to provide theoretical control over the estimation of the connectivity of all pairs of region, perhaps by leveraging existing bootstrapping techniques (Cai and Liu, 2016).

Finally, it would be valuable to provide practitioners with a way to quantify edge detection uncertainty. For instance, confidence intervals for each edge of the whole inferred network could be defined. Some work has been done to determine confidence intervals for correlation coefficients in the bivariate case, both for underlying normality, e.g., (Ruben, 1966; Muirhead, 2005), and unknown distributions (Hu et al., 2020). It would be worth exploring how these methods could build upon our approach to extend them to a more general case in order to account for dependence.

Appendix

4.A Proof of Proposition 4.3.1

4.A.1 U-scores

Before proving Proposition 4.3.1, we need to introduce U-scores. *U-scores* are an orthogonal projection of the Z-scores of random variables. They are confined to an $(n-2)$-sphere centered around 0 and with radius 1, denoted S_{n-2}, with n the number of samples. We refer to (Hero and Rajaratnam, 2011) for a full definition. U-scores namely provide a practical expression of the correlation coefficient as an inner product of U-scores: $R^{AB}_{i,j} = (U^A_i)^T U^B_j = 1 - \|U^A_i - U^B_j\|^2/2$, where U^A_i, U^B_j are the random variables of the U-scores of voxels i and j in regions a and b, respectively, and $\|.\|^2$ is the squared Euclidean distance. Consequently, when U-scores are close to one another on S_{n-2}, they are associated with a high correlation.

Intuition behind Proposition 4.3.1. Roughly speaking, Proposition 4.3.1 means that if the Wasserstein distance between the densities of sample intra-correlation coefficients is sufficiently large, which means that the two distributions are highly different, then the average Euclidean distance between U-scores from the two regions is large too. Hence the average of inter-correlations is quite low. This phenomenon is illustrated in Figure 4.A.1 where two regions with different intra-correlation densities are depicted when $n = 3$.

4.A.2 Proof of Proposition 4.3.1

Let us first remark:

$$\inf_{c\in[0,1]} \left(F^{-1}_{R^{A,A}}(c) - F^{-1}_{R^{b,b}}(c)\right)^2 \leq d^2_W(f_{R^{A,A}}, f_{R^{b,b}}) \leq \sup_{c\in[0,1]} \left(F^{-1}_{R^{A,A}}(c) - F^{-1}_{R^{b,b}}(c)\right)^2 \quad (4.10)$$

and as $\left(F^{-1}_{R^{A,A}} - F^{-1}_{R^{b,b}}\right)^2$ is continuous on $[0,1]$, it attains its supremum and infimum.

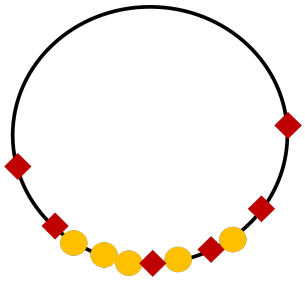


Fig. 4.A.1 S_{n-2} with $n = 3$ and U-scores from two regions (red diamonds and orange discs) that have a high intra-correlation density Wasserstein distance. Recalling that a high Euclidean distance between U-scores implies a low correlation, we can intuitively observe that the average inter-correlation is upper-bounded.

Additionally, we can notice that for each $c \in [0,1]$, there exist two points $U, V \in S_{n-2}$ such that $F^{-1}_{R^{A,A}}(c) = 1 - \|U - V\|^2/2$, and similarly for region b. Moreover, for all $x, y \in \mathcal{R}_A$, there exists $c \in [0,1]$ such that, with their corresponding U-scores denoted U_x, U_y (which are in S_{n-2}), $R^{A,A}_{x,y} = 1 - \|U_x - U_y\|^2/2 = F^{-1}_{R^{A,A}}(c)$, and analogously for region B.

Therefore, under the assumption $\min_{c\in[0,1]} \left(F^{-1}_{R^{A,A}}(c) - F^{-1}_{R^{b,b}}(c)\right)^2 \geq d$, there exist U_{x_A}, U_{y_A}, $U_{x_B}, U_{y_B} \in S_{n-2}$ such that

$$\min_{c\in[0,1]} \left(F^{-1}_{R^{A,A}}(c) - F^{-1}_{R^{b,b}}(c)\right)^2 = \frac{1}{4}\left(\|U_{x_B} - U_{y_B}\|^2 - \|U_{x_A} - U_{y_A}\|^2\right)^2,$$

and it follows for all $v_A, w_A \in \mathcal{R}_A$, $v_B, w_B \in \mathcal{R}_B$,

$$d \leq \frac{1}{4}\left(\|U_{x_B} - U_{y_B}\|^2 - \|U_{x_A} - U_{y_A}\|^2\right)^2 \leq \frac{1}{4}\left(\|U_{v_B} - U_{w_B}\|^2 - \|U_{v_A} - U_{w_A}\|^2\right)^2.$$

Thus, expanding the term on the right and applying the triangle inequality, followed by the reverse triangle inequality,

$$\begin{aligned}
2\sqrt{d} &\leq (\|U_{v_B} - U_{w_B}\| + \|U_{v_A} - U_{w_A}\|) \cdot \Big|\|U_{v_B} - U_{w_B}\| - \|U_{v_A} - U_{w_A}\|\Big| \\
&\leq (\|U_{v_B} - U_{v_A}\| + \|U_{v_A} - U_{w_B}\| + \|U_{v_A} - U_{w_B}\| + \|U_{w_B} - U_{w_A}\|) \cdot \|U_{v_B} - U_{w_B} - (U_{v_A} - U_{w_A})\| \\
&\leq (\|U_{v_B} - U_{v_A}\| + \|U_{v_A} - U_{w_B}\| + \|U_{v_A} - U_{w_B}\| + \|U_{w_B} - U_{w_A}\|) \cdot (\|U_{v_B} - U_{v_A}\| + \|U_{w_B} - U_{w_A}\|) \\
&\leq (\|U_{v_B} - U_{v_A}\|^2 + \|U_{v_B} - U_{v_A}\| \cdot \|U_{w_B} - U_{v_A}\| + \|U_{w_B} - U_{v_A}\| \cdot \|U_{w_B} - U_{w_A}\| \\
&\quad + \|U_{v_B} - U_{v_A}\| \cdot \|U_{w_B} - U_{w_A}\|) \\
&\quad + (\|U_{w_B} - U_{w_A}\|^2 + \|U_{w_B} - U_{v_A}\| \cdot \|U_{v_B} - U_{v_A}\| + \|U_{w_B} - U_{w_A}\| \cdot \|U_{w_B} - U_{v_A}\| \\
&\quad + \|U_{v_B} - U_{v_A}\| \cdot \|U_{w_B} - U_{w_A}\|).
\end{aligned}$$

We can then notice

$$\begin{aligned}
\overline{\|U^B - U^A\|}^2 &= \left(\frac{1}{p_A p_B}\sum_{v_A\in\mathcal{R}_A}\sum_{v_B\in\mathcal{R}_B}\|U_{v_B} - U_{v_A}\|\right)^2 \\
&= \frac{1}{(p_A p_B)^2}\sum_{h_A\in\mathcal{R}_A}\sum_{h_B\in\mathcal{R}_B}\|U_{h_B} - U_{h_A}\|^2 + \\
&\frac{1}{(p_A p_B)^2}\sum_{h_A\in\mathcal{R}_A}\sum_{h_B\in\mathcal{R}_B}\sum_{k_B\in\mathcal{R}_B-\{h_B\}}\|U_{h_B} - U_{h_A}\|\cdot\|U_{k_B} - U_{h_A}\| + \\
&\frac{1}{(p_A p_B)^2}\sum_{h_A\in\mathcal{R}_A}\sum_{h_B\in\mathcal{R}_B}\sum_{k_A\in\mathcal{R}_A-\{h_A\}}\|U_{h_B} - U_{h_A}\|\cdot\|U_{h_B} - U_{k_A}\| + \\
&\frac{1}{(p_A p_B)^2}\sum_{h_A\in\mathcal{R}_A}\sum_{h_B\in\mathcal{R}_B}\sum_{k_A\in\mathcal{R}_A-\{h_A\}}\sum_{k_B\in\mathcal{R}_B-\{h_B\}}\|U_{h_B} - U_{h_A}\|\cdot\|U_{k_B} - U_{k_A}\|.
\end{aligned}$$

Thus $\overline{\|U^B - U^A\|}^2 \geq \frac{1}{(p_A p_B)^2}\cdot\frac{(p_A p_B)^2}{2}\cdot 2\sqrt{d} = \sqrt{d}$. From the Cauchy-Schwarz inequality,

$$\overline{R^{A,B}} \leq 1 - \frac{\overline{\|U^B - U^A\|}^2}{2},$$

which completes the proof.

4.B Proof of Proposition 4.4.2

Let us recall Proposition 4.4.2.

Proposition 4.4.2. *For all $\rho \in [0,1]$,*

$$\widehat{\nu}_e^{AB}(\rho) \leq \widehat{\nu}^{AB}(\rho). \tag{4.11}$$

Proof. Since for all $\rho \in [0,1]$, $0 \leq \widehat{F}_{|R^{A,B}|}(\rho) \leq 1$, then $\widehat{F}_{|R^{A,B}|}(\rho)^{p_B} \leq \widehat{F}_{|R^{A,B}|}(\rho)$. Thus, $\widehat{\nu}^{AB}(\rho) = 1 - \widehat{F}_{|R^{A,B}|}(\rho)^{p_B} \geq 1 - \widehat{F}_{|R^{A,B}|}(\rho) = \widehat{\nu}_e^{AB}$ □

4.C Additional insights on ν^{AB}

We can derive the following proposition.

Proposition 4.C.1. *If, for a fixed $i = 1, \ldots, p_A$, all sample inter-correlation coefficients $R_{i,j}^{A,B}$ are i.i.d., $\nu^{AB} = \tilde{\nu}^{AB}$.*

Proof. We can first remark $\max_{j \in \mathcal{R}_B} \phi_{ij}^{AB} = 1 - \prod_{j=1}^{p_B} (1 - \phi_{ij}^{AB})$. Thus,

$$\begin{aligned} E[N^{AB}] &= \sum_{i=1}^{p_A} E[1 - \prod_{j=1}^{p_B} (1 - \phi_{ij}^{AB})] \\ &= \sum_{i=1}^{p_A} (1 - \prod_{j=1}^{p_B} E[(1 - \phi_{ij}^{AB})]) \text{ under the assumption of independence.} \end{aligned}$$

For a fixed $i = 1, \ldots, p_A$, under the assumption all $|R_{i,j}^{a,b}|$ are identically distributed, then, for all $j, l = 1, \ldots, p_B$, $F_{|R_{i,j}^{A,B}|} = F_{|R_{i,l}^{A,B}|}$. Denote, $F_{|R^{A,B}|}$ the distribution function such that $F_{|R_{i,j}^{A,B}|} = F_{|R^{A,B}|}$ for all $i = 1, \ldots, p_A$, $j = 1, \ldots, p_B$. Thus, $p_A \cdot \tilde{\nu}^{AB} = E[N^{AB}] = p_A \cdot \left[1 - F_{|R^{A,B}|}^{p_B}\right] = p_A \cdot \nu^{AB}$.

□

4.D Inter-correlation distributions

The effect of intra-correlation on the sample inter-correlation distribution is explored for additional sets of parameters: either decreasing the number of variables p or increasing the number of samples n with respect to Figure 4.2. Figures 4.D.1 and 4.D.2 highlight the fact that the variance decreases for increased intra-correlation is observed no matter the region or sample size.

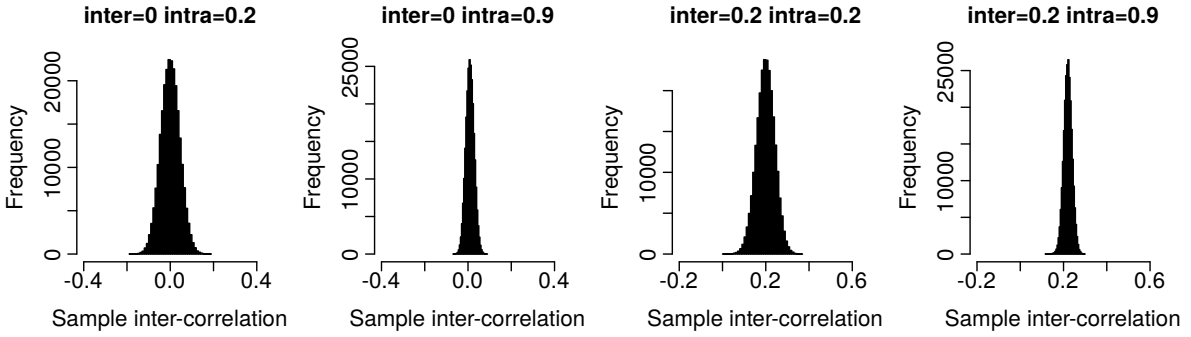


Fig. 4.D.1 Effect of intra-correlation on sample inter-correlation distribution, for different population inter-correlation values. The correlation samples were computed between all pairs of variables from two groups. Each group contains $n = 500$ samples of $p = 500$ intra-correlated random variables following a multivariate normal distribution with Toeplitz intra-correlation (cf. Section 2.2). We can note the higher the intra-correlation, the lower the variance of the inter-correlation.

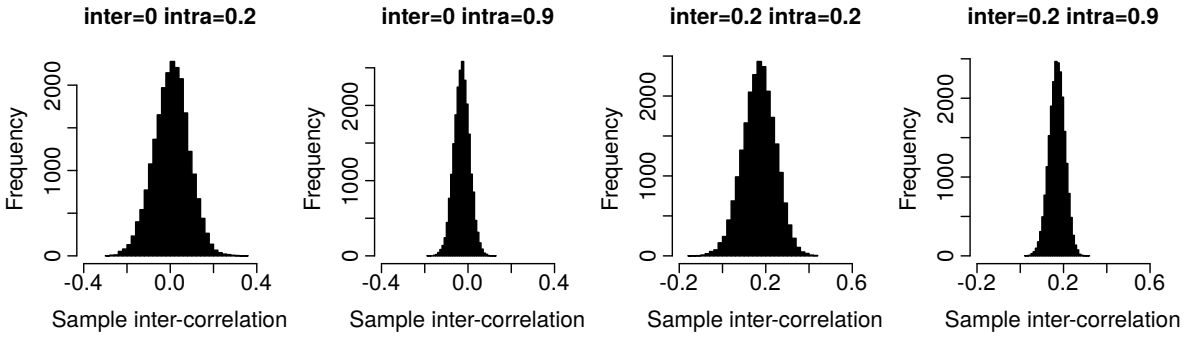


Fig. 4.D.2 Effect of intra-correlation on sample inter-correlation distribution, for different population inter-correlation values. The correlation samples were computed between all pairs of variables from two groups. Each group contains $n = 150$ samples of $p = 150$ intra-correlated random variables following a multivariate normal distribution with Toeplitz intra-correlation (cf. Section 2.2). We can note the higher the intra-correlation, the lower the variance of the inter-correlation.

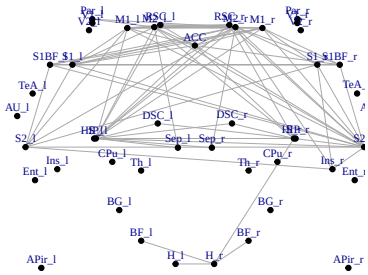
4.E Additional intuitions about correlations and U-scores

One can turn to the approximate expression of $E[N^{AB}]$ derived in (Hero and Rajaratnam, 2011), which we denote ν_h^{AB}. It employs the Bhattacharyya affinity, which quantifies the overlap between two probability densities and was originally interpreted as a cosine of the angle between distributions (Bhattacharyya, 1946). In (Hero and Rajaratnam, 2011) the authors leverage it as a dependency measure between U-scores. It can be reframed as well, using the law of total probability, as quantifying dependency between pairs of U-scores from each region of interest, which can be associated with inter-correlation coefficients. ν_h^{AB} is thence maximized when inter-correlation coefficients are independent. Intuitively, a high intra-correlation when the average inter-correlation is mild will increase the dependence between inter-correlations and decrease the number of discoveries, which is consistent with Figure 4.3. Furthermore, their proposed approximation ν_h^{AB} is proportional to the number of variables p_A and p_B. Therefore this phenomenon can also be explained by the decrease in the effective number of variables observed in the presence of dependence (Nyholt, 2004). In (Hero and Rajaratnam, 2011) the authors leverage it as a dependency measure between U-scores. It can be reframed as well, using the law of total probability, as quantifying dependency between pairs of U-scores from each region of interest, which can be associated with inter-correlation coefficients. ν_h^{AB} is thence maximized when inter-correlation coefficients are independent. Intuitively, a high

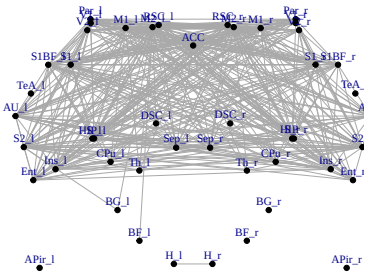
intra-correlation when the average inter-correlation is mild will increase the dependence between inter-correlations and decrease the number of discoveries, which is consistent with Figure 4.3. Furthermore, their proposed approximation ν_h^{AB} is proportional to the number of variables p_A and p_B. Therefore this phenomenon can also be explained by the decrease in the effective number of variables observed in the presence of dependence (Nyholt, 2004). These remarks further highlight the need to take into account the dependency structures of the variables to reliably detect inter-correlations.

4.F Rat data networks

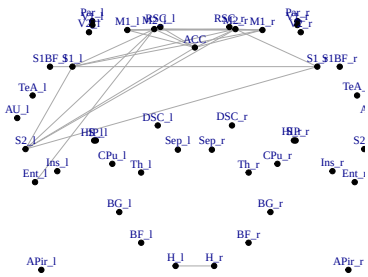
Figures 4.F.1 and 4.F.2 display the live and dead rat brain networks estimated in Section 4.6 and presented in Table 4.2.



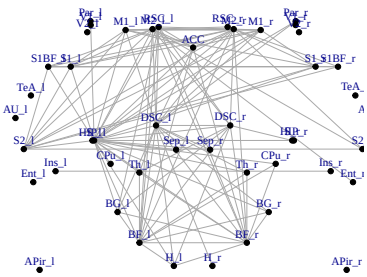
(a) 20160421_133725



(b) 20160614_095825

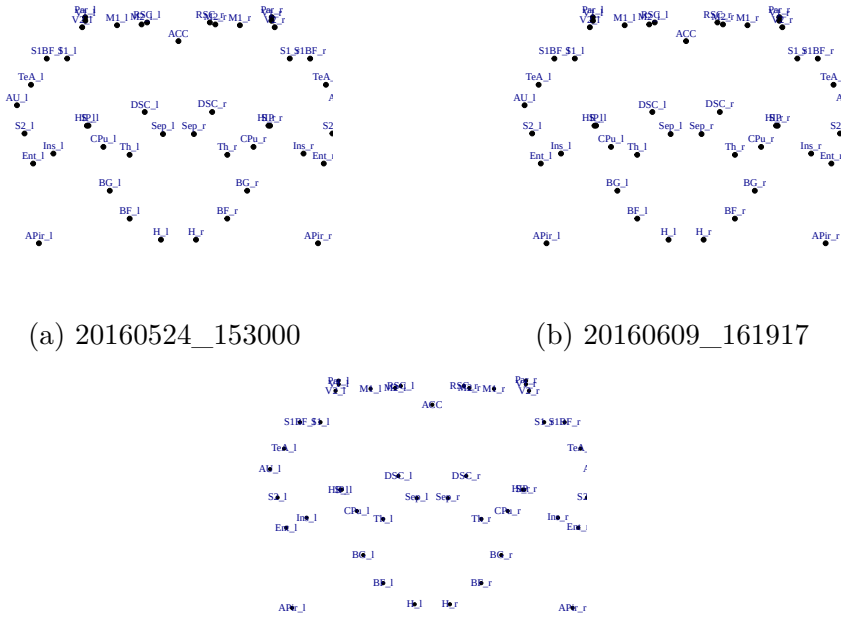


(c) 20160615_103000



(d) 20160615_121820

Fig. 4.F.1 Brain functional connectivity networks of four live rats, anesthetized using Isofluorane, inferred applying our proposed method with quantile-based thresholds. Nodes represent brain regions and edges connect nodes with a sufficiently high inter-correlation.



(a) 20160524_153000 (b) 20160609_161917

(c) 20160610_121044

Fig. 4.F.2 Brain functional connectivity networks of three dead rats inferred applying our proposed method with quantile-based thresholds. Nodes represent brain regions and edges connect nodes with a sufficiently high inter-correlation.

4.G Additional empirical results about $\hat{\nu}^{AB}$ and $\hat{\nu}_e^{AB}$

In this section, we present additional illustrations of our proposed approximations for the number of discoveries on datasets generated as presented in Section 4.2.2.

In Figure 4.G.1, the inter-correlation is set to 0.3, and can be compared to Figure 4.4. We can remark how, for a given pair of average intra-correlation values, the number of discoveries reaches zero at a larger correlation threshold value when the inter-correlation is larger, which is as expected. To assess the impact of dataset generation variability, $\hat{\nu}^{AB}$ and $\hat{\nu}_e^{AB}$ are plotted as a function of the correlation threshold for ten different simulations (cf. Figures 4.G.2 and 4.G.3). We can observe an increase in the variability of $\hat{\nu}_e^{AB}$ when the intra-correlation increases, which is as expected with regards to Figure 4.2.

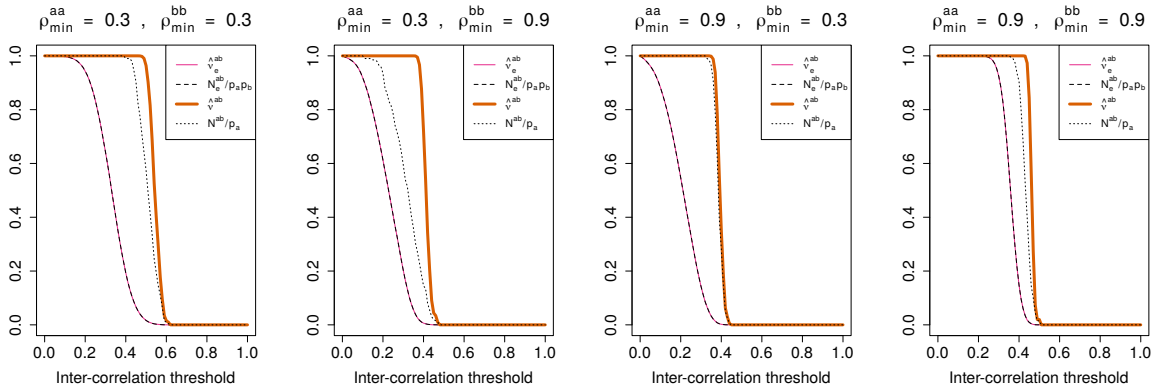


Fig. 4.G.1 Normalized number of discoveries $\hat{\nu}^{AB}$, $\hat{\nu}_e^{AB}$, N^{AB}/p_A and $N_e^{AB}/p_A p_B$ as a function of the correlation threshold for different intra-correlation values ρ^{aa}, ρ^{bb} and a 0.3 inter-correlation, with $p_A = p_B = 200, n = 100$. The dataset was generated as described in Section 4.2.2.

4.H Further information about our implementation and code availability

Our implementation is based on R 4.2.3. All experiments were performed on a laptop running on Ubuntu 18.04 with eight 1.8GHz 64-bits Intel Core i7-10610U CPUs, 32 GB of memory and a 1 TB hard drive. Our source code can be found at https://gitlab.inria.fr/q-func/csinference.

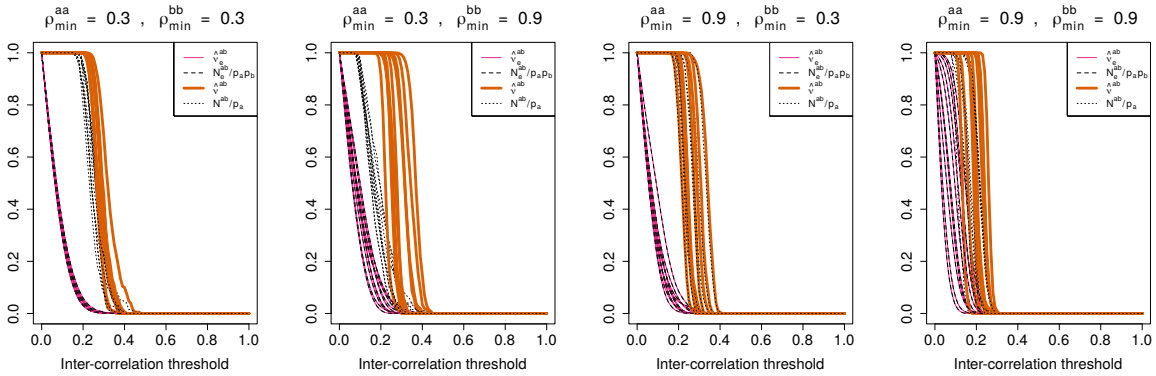


Fig. 4.G.2 Normalized number of discoveries $\widehat{\nu}^{AB}$, $\widehat{\nu}_e^{AB}$, N^{AB}/p_A and N_e^{AB}/p_Ap_B as a function of the correlation threshold for different intra-correlation values ρ^{aa}, ρ^{bb} and 10 different simulations. The inter-correlation is 0, with $p_A = p_B = 200, n = 100$. The ten datasets were generated as described in Section 4.2.2.

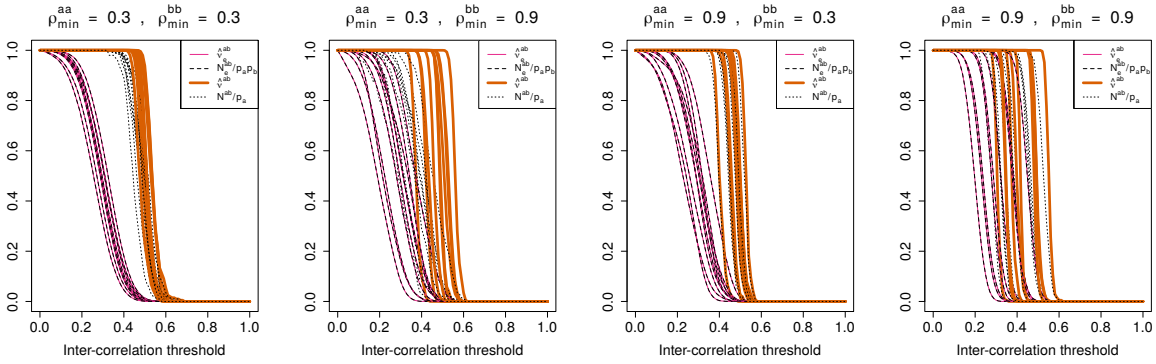


Fig. 4.G.3 Normalized number of discoveries $\widehat{\nu}^{AB}$, $\widehat{\nu}_e^{AB}$, N^{AB}/p_A and N_e^{AB}/p_Ap_B as a function of the correlation threshold for different intra-correlation values ρ^{aa}, ρ^{bb} and 10 different simulations. The inter-correlation is 0.3, with $p_A = p_B = 200, n = 100$. The ten datasets were generated as described in Section 4.2.2.

Chapter 5

Topological Data Analysis for Spatially-Informed Weighted Network Comparison

Multiscale and Multi-Density Comparison of Functional Brain Networks Through Label-Informed Persistence Diagrams

In Chapter 4 we introduced a novel binary network inference pipeline that accounted for dependency structures within nodes. Nonetheless, in some applications, handling weighted networks, and notably circumventing thresholding steps, may be preferred over manipulating their binary counterparts. Chapter 5 hence leverages both topological data analysis and regional label information to propose a multi-scale comparison of weighted connectivity networks. The effectiveness of this approach is illustrated via a comparison of comatose and healthy subjects, and of real-world data against null models. The latter could also be seen as way to assess the quality of estimated networks by ensuring they capture meaningful information beyond mere noise.

This contribution is based on the following contribution, which is in the process of being submitted to a journal.

Sitoleux, P., Carboni, L., Lbath, H., and Achard, S. (in preparation 2023). Multiscale and multi-density comparison of functional brain networks through label-informed persistence diagrams

5.1 Introduction

Networks are widely used to model brain functional connectivity and have been characterized via the various tools and concepts from graph theory and network sciences (Sporns, 2022), such as small-worldness (Achard and Bullmore, 2007), degree distribution, or rich-club coefficient. As previously seen, the correlation of averages (CA) approach is widely used to define functional brain networks. It requires parcellating the rs-fMRI signal with a predefined anatomical atlas, giving us the N nodes of the graph (Stanley et al., 2013). For each node, the rs-fMRI time series is estimated to be the average of voxel activity in the corresponding parcellated region. Each pair of nodes is then correlated to yield a dense adjacency matrix. One then generally applies a threshold, in order to obtain a sparse binary graph (Bassett and Bullmore, 2006; Bordier et al., 2017; Theis et al., 2021; Achard and Bullmore, 2007). As explored in Chapter 4, an important issue with this approach is the choice of a threshold. The latter can be based on correlation distribution properties, as we did in the previous chapter, average degree, connectivity (Couto et al., 2017), graph regime, multiple testing, or to optimize a metric such as discrimination accuracy (Zanin et al., 2012), etc. To avoid this, analyses can be conducted at multiple density levels (Kartun-Giles and Bianconi, 2019). In this chapter, density is understood as the proportion of edges present in the graph compared to the maximum possible number of edges. However, there is a lack of attention to multi-density approaches exploring and summarizing information in functional brain networks.

Moreover, purely graph-based methods are often limited to global or local statistics, only presenting information at the scale of the whole graph or of a single node neighborhood (Sizemore et al., 2019). Persistent homology is a topological data analysis (TDA) approach that produces multiscale summaries from a point cloud or a distance matrix. Persistent homology tracks homological features, in short, the number of connected components and the number of topological cavities (equivalent to a circle or a hollow sphere in dimensions 1 and 2, or to higher dimensional cavities), by building simplices. This is a way to look at structures in the graph that are orders higher than the dyadic relationships the edges of the graph represent (Torres et al., 2021). The interest in higher-order interactions has taken off in recent years, whether in the broader context of

complex systems modeling (Battiston et al., 2020, 2021) or in structural (Andjelković et al., 2020) or functional brain networks (Gatica et al., 2021; Herzog et al., 2022; Gatica et al., 2022), with notable findings in aging and neurodegenerative disorders. Another argument that has been advanced in favor of persistent homology is that it is easily applicable to biological data, which is often highly dimensional and lacks a natural concept of distance (Carlsson, 2009). For instance, in neuroscience networks, it has been applied to differentiate healthy subjects and subjects with neural disorders (Lee et al., 2012; Chung et al., 2015; Caputi et al., 2021), study the influence of psychotropic substances and sedatives (Petri et al., 2014; Varley et al., 2021) or to analyze neuronal network simulations (Reimann et al., 2017; Bardin et al., 2019).

One of the main goals of functional connectivity network studies is detecting and recognizing brain dysfunctions related to pathological conditions. Indeed, functional networks offer a unique way of extracting new noninvasive biomarkers (Hallett et al., 2020) that provide powerful understanding of physio-pathological mechanisms. Recognizing a given pathology by analysis of the functional connectivity requires graph comparison distances, similarities functions, or statistical tools, such as efficiency, small-worldness, global or nodal statistics (Carboni et al., 2023), spectral graph analysis, or edit distances (Mheich et al., 2020). For this purpose, it is fundamental to be able to precisely quantify whole-brain variations that can arise in different cohorts, probing the influence of age difference, state of consciousness, or neurological disorders. To show the benefit of our proposed framework in real data applications, we consider functional networks of comatose patients. Discriminating between subjects within different states of consciousness using functional networks is a difficult task: previous work did not detect significant modifications in graph metrics between comatose patients and healthy controls, but what appears to be a reorganization of network hubs (Achard et al., 2012b). Here, we propose a new distance among brain networks that combines homological features and regional information.

We are particularly interested in applications of TDA in functional brain networks. To that end, we consider null models, which generate surrogate inter-correlation matrices as benchmarks of real data. This permits the identification of relevant features captured by a persistent homology approach in brain connectivity. Null models are ubiquitous in network neuroscience (Váša and Mišić, 2022). For instance, they allow one to test the statistical significance of graph features of empirical networks against a null hypothesis, given by the choice of a null model. Furthermore, null models which can be tuned to reproduce a specific graph layout offer a tool for comparing graph distances and understanding which properties they capture.

In this work, we present applications of persistent homology that allow insights into functional brain connectivity networks without requiring an arbitrary threshold choice to obtain binary graphs. Persistent homology is newly applied to a density-adjacency matrix to probe its ability to discriminate between real data and surrogate data generated from null models, as well as between a set of healthy subjects and comatose patients. However, while persistent homology summaries are multiscale, they forego critical information: the identity of nodes and edges. Indeed, these summaries, such as the persistence diagram or Betti curves,do not depend on the labeling of nodes. We hence introduce distances taking into account the node labeling, and explore them in the above-mentioned tasks.

This chapter explores four main directions: (i) a novel density-based filtration, (ii) label-informed distance on persistence diagrams, useful in non-permutation invariant settings, (iii) application to null-models and comatose-control discrimination, and (iv) gain insights on the construction of null models from real brain data.

5.2 Materials and methods

5.2.1 Data

We consider a set of resting-state functional networks from the Human Connectome Project (HCP) (Essen et al., 2012) to investigate the general properties of topological features of brain graphs. To investigate how significant these features are, attempts will be made to cluster a second dataset of 20 healthy controls and 17 patients in a comatose state (Achard et al., 2012b) (8 patients were dropped due to excessive head motion). The patients were between 21 and 82 years old, had an initial Wessex Head Injury Matrix (WHIM) score between 1 and 37 (on a scale that goes from 0, meaning deep coma, to 62, meaning full recovery), and were scanned between 3 and 32 days after the acute medical event. Twelve of them fell into a coma following a cardiac and respiratory arrest, two after going into hypoglycemia, two after a gaseous embolism and one after an extracranial arterial dissection. Six months following their fall into a coma, 9 patients had died, 5 remained in a vegetative state and 3 made a recovery.

Data were corrected for head motion and co-registered with the subject's structural MRI images. This allowed mapping fMRI images to a customized parcellation template. The data were not spatially smoothed and, unless specified, did not undergo any cerebrospinal fluid, white matter, or gray matter regression. The process needed to go from raw blood-oxygen-level-dependent (BOLD) image sequences to a meaningful functional connectivity graph comprises multiple steps, each with its set of assumptions and conse-

quences on the final network. For example, in the case of global signal regression, that is, regressing over average of the whole image, it has been established that one has to make a choice between discarding some neural signals and keeping non-neural nuisances (Liu et al., 2017b).

5.2.2 Functional connectivity network construction

The measurements were parcellated into $N = 90$ regions using the Automated Anatomical labeling (AAL) template (Tzourio-Mazoyer et al., 2002). For each region, the time series were aggregated by averaging over all voxels, weighted by the proportion of gray matter in each voxel (estimated through structural MRI) and corrected for head motion. Inter-correlations were then estimated between all pairs of region-averaged time series.

5.2.3 Functional connectivity network representation

A graph $G = (V, E)$ is a collection of nodes or vertices V and edges linking those vertices, i.e., each element in E is an element of $V \times V$. There is a range of ways to represent a graph, here it will be represented by its adjacency matrix, usually denoted A. To avoid any possible confusion, pairs of nodes, i.e., brain regions, will be denoted i, j in this chapter, as opposed to A, B in all the other chapters of this thesis. An adjacency matrix is a symmetric matrix where each off-diagonal element a_{ij} is a weight, representing a chosen attribute of the relation between nodes i and j.

Using the inter-correlation matrix between the inter-regional correlation coefficients C, we construct the graph adjacency matrix A:

$$A = 1 - |C|. \tag{5.1}$$

To neutralize the effect of inter-subject fluctuations in the values of the inter-correlations, an alternative matrix can be defined by considering the density-based adjacency matrix $\tilde{A}$, computed using the level of edge-appearance as follows:

$$\tilde{a}_{ij} = \frac{2u_{ij}}{N(N-1)}, \quad i \neq j, \tag{5.2}$$

where $\tilde{a}_{ij}$ is the normalized index of the $u_{ij}{}^{\text{th}}$ off-diagonal element of matrix A, ordered according to the absolute value of the inter-correlations. In this new matrix, the order between each off-diagonal element is respected, but these new values are spaced at a constant interval $\frac{2}{N(N-1)}$.

5.2.4 Null models for functional connectivity

Null models are an ubiquitous tool in network neuroscience (Váša and Mišić, 2022). By offering a way to generate networks according to a simplified model, they allow benchmarking of the properties of empirical networks against the null hypothesis provided by a given model.

Comparing null models and empirical functional connectivity allows for the appreciation of null model properties and improved understanding of the information captured by various distances. Here, four null models for correlation networks are presented, each conserving some features from an initial functional connectivity matrix (Váša and Mišić, 2022). First, we consider the Zalesky matching algorithm, which generates correlation matrices with a given average correlation and variance between correlations (Zalesky et al., 2012). This is achieved through generating an $N+1$ normally distributed random vectors of length T $x_i, y \sim \mathcal{N}(0, \mathbf{I})$ $(i \in \{1, \dots N\})$, with $\mathbf{I}$ the T-dimensional identity matrix. Then, the values of x_i are repeatedly adjusted as $x_i \leftarrow x_i + ay$ until the desired average correlation between the vectors is obtained. The process is repeated with a different number of time points T until the correlation variances also match.

As a second model, we consider a spatio-temporal approach that generates time series imitating spatial and temporal autocorrelation from the regional time series of a given scan of a subject (Shinn et al., 2023).

The last two null models are a phase randomization model, generating new time series where each regional time series has the same power spectrum, and an Erdös-Rényi model, where each correlation value is randomly distributed.

5.2.5 Persistent homology

Persistent homology is a mathematical formalism from the larger field of topological data analysis, that allows the production of multiscale summaries of point cloud or graph distance matrices (Chazal and Michel, 2021), such as Betti numbers and persistent diagrams.

Starting from a distance matrix, simplicial complexes are built, using a rule denominated *filtration*, at each possible scale in the matrix. At each step, the appearance and disappearance of topological features are tracked. These features correspond to the dimension of p^{th} homology group (Croom, 1978), with p a non-negative integer. Essentially, the number of 0-dimensional features is the number of connected components, and higher dimensional ones represent *topological cavities*: 1-dimensional features are circle-like, 2-dimensional features are hollow sphere-like, and so on.

For a set of $k+1$ vertices $\{v_0, v_1, \ldots, v_k\}$, the *k--dimensional abstract simplex* or *k-simplex* is the set containing all nonempty subsets of vertices. A simplex σ^k is a *face* of a simplex σ^n, $k < n$ means that each vertex of σ^k is a vertex of σ^n. For example, 0-simplex is a single point, a 1-simplex is an edge linking two points, a 2-simplex is a triangle with its three edges and three points as 1 and 0-faces, a 3-simplex is a tetrahedron, etc.

A *simplicial complex* K is a finite set of properly joined simplices, i.e., each intersection of two simplices is either a face of both or empty, and where each face of a member of K is also a member of K. The *dimension* of K is the largest positive integer r such that K has an r-simplex. 5.2.1 presents a 3-dimensional simplicial complex

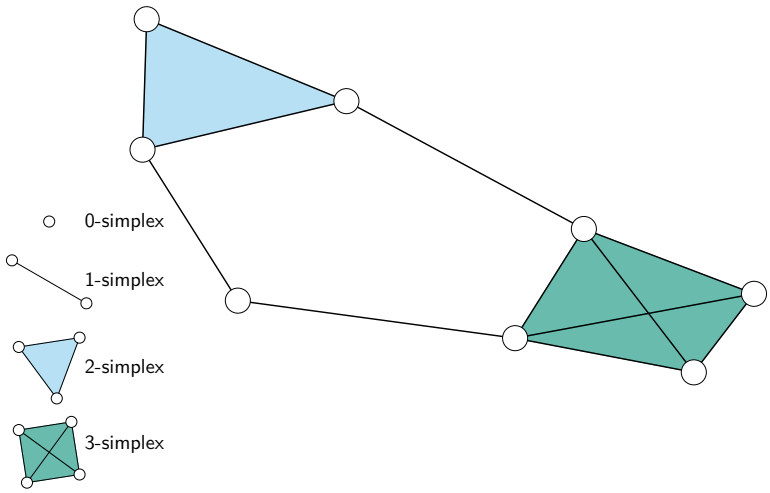


Fig. 5.2.1 A simplicial complex of dimension 3.

For a given simplicial complex, the number of k-dimensional *holes* or *features* is called the k^{th} *Betti number*, denoted β_k. Formally, β_k is the dimension of the k^{th} homology group H_k, which is itself the quotient of the k^{th} cycle group Z_k with the k^{th} boundary group of the given simplicial complex:

$$\beta_k = \dim(H_k) = \dim\left(\frac{Z_k}{B_k}\right). \tag{5.3}$$

In short, this means that β_k counts the number of $k+1$-dimensional volumes that are enclosed by, at least, a k-dimensional cycle that does not correspond to the boundary of a simplex of the given simplicial complex, with β_0 counting the number of connected components of the simplicial complex. Therefore, in 5.2.1 the simplicial complex has $\beta_0 = 1$, since it has a unique connected component, $\beta_1 = 1$ with the non-filled central cycle 1 and $\beta_k = 0$ for $k \geq 2$.

In persistent homology, there is some freedom in how simplices are built. For its lower computational cost relative to the other options, in this work, and in most persistent homology applications, the *Vietoris-Rips* filtration is used. In this setting, a simplex is in the relevant simplicial complex if all its 1-faces or edges are. Edges are successively added to the simplicial complex when the corresponding value in the adjacency matrix is equal to or lower than the given filtration parameter. In this work, instead of using inter-correlation values, the homology groups are computed at each scale of the density-based adjacency matrix.

A way to present this topological information is to plot the *Betti curves* of the Betti numbers against the filtration values, which can be either inter-correlations or density levels. Another way is to plot the *death* values of features against their *birth* values. This yields a *persistence diagram*.

Let us consider the following example. The first row of 5.2.2 presents simplicial complexes with $\beta_0 = 1$ and $\beta_1 = 0$. In all those three cases, there exists at least one cycle, but each corresponds to the boundary of a simplex. For the second row, each simplicial complex has $\beta_1 = 1$. The first two have a unique representative: ABCD and ABCDEFGHIJ, while the last one has multiples, with two of them disjoint: ACEGI and BDFHJ. This illustrates both the robustness of persistent homology with regard to the general structure and the limitations of its discriminative power.

From the examples given here, it also seems that $\beta_1 > 0$ is a signature of a relatively low clustering coefficient, i.e., the proportion of possible triangles. This shows how these features could be significant. Taking the analogy of friendship networks, it means that there is a cycle of length 4 with some gap with A and C both being friends with B and D, but neither A and C or B and D are. This might indicate either a pure coincidence or some hidden organizational principle.

5.2.6 Comparing persistence diagrams

We first recall various distances between persistence diagrams.

The *p-Wasserstein distance* between measures μ, ν, $p \in [1, \infty[$ is

$$W_p(\mu, \nu) = \left(\int_{\mathcal{X} \times \mathcal{X}} c(x, y)^p \, \mathrm{d}\pi(x, y) \right)^{1/p} \tag{5.4}$$

where $\pi(x, y)$ is the optimal coupling between μ and ν under cost $c(\cdot, \cdot)^p$. In these works, the focus is mainly on the 1-Wasserstein distance.

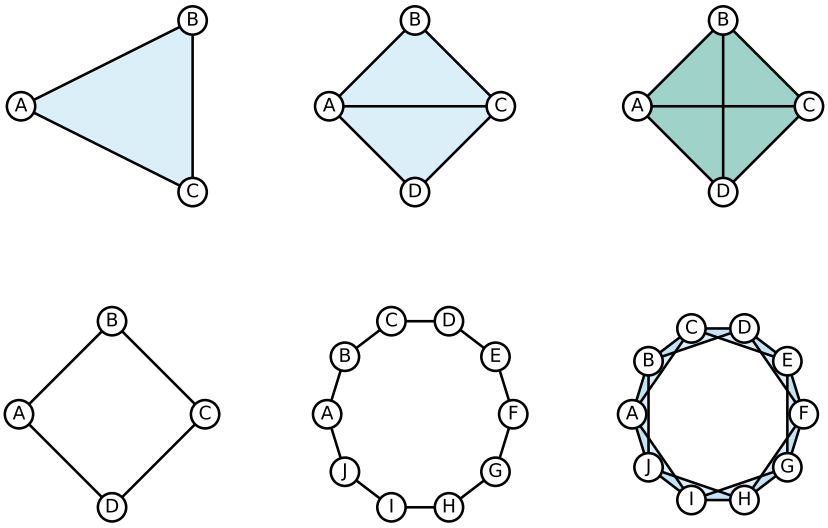


Fig. 5.2.2 Examples of graphs with different Betti numbers. On the top row, they have $\beta_1 = 0$ while those on the bottom have $\beta_1 = 1$.

Another common option is the *bottleneck distance*, which is the $p \to \infty$ limit of the p-Wasserstein distance:

$$\mathrm{Bo}(\mu, \nu) = \lim_{p \to +\infty} W_p(\mu, \nu). \tag{5.5}$$

For a, b two discrete distributions, the *optimal transport distance* can be expressed as

$$d_{\mathrm{OT}}(a, b) = \gamma^*_{ab} = \underset{\gamma \in \mathbb{R}^{m \times n}}{\arg\min} \sum_{i,j} \gamma_{ij} M_{ij} \tag{5.6}$$

such that $\gamma \mathbf{I} = a, \gamma^T \mathbf{I} = b, \gamma \geq 0$ and with M the cost matrix defining the cost of moving the histogram bin a_i to bin b_j, $\mathbb{I}$ the identity matrix of the relevant dimension.

We now introduce a label-informed distance on persistence diagrams, by designing a novel cost matrix. First, we list edges $\{\{e_k[i]\}_{i \in \{1,\ldots,n_k\}}\}$ which correspond to the birth or death of a the k^{th} homological feature in the persistence diagrams, where n_k is the number features. Next we choose a simple binary cost matrix defined as follows:

$$M^*_{ij} = \begin{cases} 0 & \text{if } e[i] = e[j] \\ 1 & \text{otherwise} \end{cases} \tag{5.7}$$

We define the *edge comparison optimal transport* distance as the optimal transport distance d_{OT} where $M = M^*$:

$$d_{e,k\cdots}(a,b) = \arg\min_{\gamma \in \mathbb{R}^{m\times n}} \sum_{i,j} \gamma_{ij} M^*_{ij}, \tag{5.8}$$

where $k\cdots$ denotes the included homological features. It should be noted that this is an approach that is similar to the fused Gromov-Wasserstein distance (Titouan et al., 2019; Vayer et al., 2020), applied to a labeled persistence diagram instead of a labeled graph. Fused Gromov-Wasserstein takes into account both labels and graph structure by taking a weighted sum of two terms representing each aspect. The label term is equivalent to the distance introduced here for a general cost matrix M.

In the following, the persistence diagrams and the Wasserstein distances between them are computed using giotto-tda, with an error tolerance of 0.01 (Tauzin et al., 2020). The optimal transport distance d_{OT} between the persistence diagram is computed with the POT package (Flamary et al., 2021).

5.2.7 Comparing edges of labeled graphs

A simple distance between labeled graphs would be to look at the *Frobenius* norm of the difference of adjacency matrices P, Q.

$$d_{\mathrm{F}}(P,Q) = \frac{1}{2}\frac{||P-Q||_{\mathrm{F}}}{||P||_{\mathrm{F}} + ||Q||_{\mathrm{F}}} \tag{5.9}$$

with $||P||_{\mathrm{F}} = \left[\sum_{i,j} p_{ij}^2\right]^{1/2}$ the Frobenius norm of matrix P (with a normalization factor to facilitate its interpretation).

In the context of thresholded graphs, node labels can be used to compare graphs. For $\mathcal{A}, \mathcal{B}$ two sets, the *overlap* similarity is

$$\mathrm{Overlap}(\mathcal{A},\mathcal{B}) = \frac{|\mathcal{A}\cap\mathcal{B}|}{\min(|\mathcal{A}|,|\mathcal{B}|)} \tag{5.10}$$

and the overlap distance

$$d_{\mathrm{O}}(\mathcal{A},\mathcal{B}) = 1 - \mathrm{Overlap}(\mathcal{A},\mathcal{B}). \tag{5.11}$$

In our setting, $\mathcal{A}, \mathcal{B}$ correspond to two sets of node label pairs, associated to edges in the thresholded graph.

5.2.8 Quantitative class comparison

In order to have a more quantitative understanding of the various distances presented here we evaluate them in a class comparison task. To that end, we use common clustering algorithms. The distance matrix structure makes it complicated to find groups if one only considers pairwise distances. To counteract this issue, the distance matrix are treated as if it were a matrix of observations. While debatable, this approach allows exploring this kind of class structure using simple and proven clustering algorithms: KMeans, hierarchical clustering with complete linkage, and spectral clustering on the nearest-neighbors graph (Pedregosa et al., 2011).

The performance of the class comparisons are evaluated using the *adjusted mutual information* (AMI):

$$\mathrm{AMI}(X;Y) = \frac{I(X;Y) - \mathbb{E}\left[I(X;Y)\right]}{\max(H(X), H(Y)) - \mathbb{E}\left[I(X;Y)\right]} \tag{5.12}$$

where $H(X)$ is the Shannon entropy of the random variable X and $I(X;Y)$ the mutual information between random variables X and Y. The AMI takes values 1 for a perfect identification and has an expected value of 0 for a random assignment of labels (Vinh et al., 2010).

5.3 Results

5.3.1 Density-based adjacency matrix

In this section, we first compare the correlation- and density-based adjacency matrices on real-world data. The plot of the inter-subject Frobenius distance for the coma cohort is depicted in 5.3.1. In the upper diagonal, which corresponds to the density-based matrices, a clear separation between the comatose and healthy subjects is apparent. However, the two groups cannot be distinguished when using the correlation-based adjacency matrices.

By varying the density of the graphs, one can also determine an optimal density threshold value at which groups can be clustered. To that end, we apply clustering algorithms on the overlap distance. Moreover, comparing different densities provide insights on the underlying structure of different graph groups. For healthy subjects and comatose patients, the results are presented in 5.3.2a, and in 5.3.2b for random graphs compared with healthy controls and comatose patients. Random graph adjacency

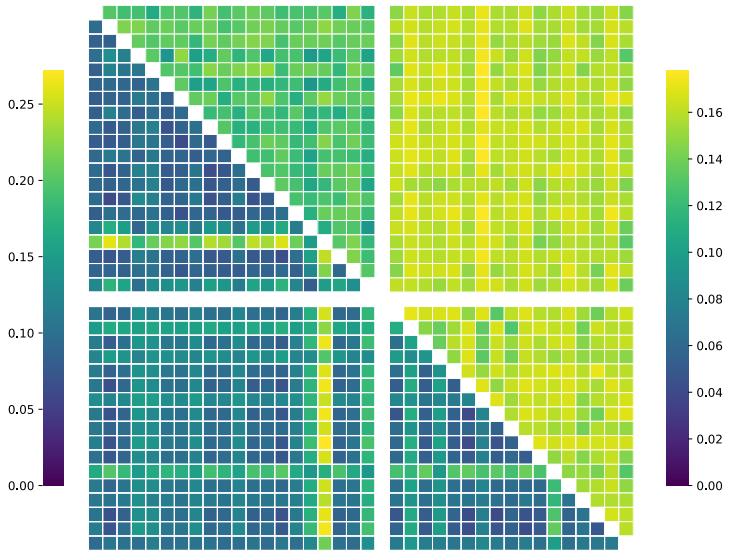


Fig. 5.3.1 Frobenius distances on the coma dataset. Lower diagonal: normalized Frobenius norm for correlation adjacency matrices A. Upper diagonal: normalized Frobenius norm for density-based adjacency matrices $\tilde{A}$.

matrices are generated as symmetric matrices, with normal independent and identically distributed (i.i.d.) off-diagonal elements.

At lower densities, it is obviously impossible to cluster between groups. Starting from a density of $\sim 10^{-3}$—which corresponds to a four-edge graph—the different chosen algorithms improve and, eventually, cluster groups perfectly for all problems with a relatively low density. Here, the fMRI measurements of the comatose patients stray markedly away from the healthy controls and appear collectively closer to each other than to the random graphs (see 5.3.1). But when individually comparing these two groups with random graphs, more edges are required to discriminate comatose patients from random graphs than healthy subjects from random graphs.

When comparing functional brain graphs to random graphs, this gives an estimate of the scale at which the structure starts to fade away. For non-random graphs, it gives the scales at which the variations get drowned in the noise.

5.3.2 Betti numbers

Figure 5.3.3 presents the Betti curves of the different rs-fMRI inter-correlation matrices and their null models using the density-based filtration, which does not require any thresholding step. The latter are presented in the first two rows.

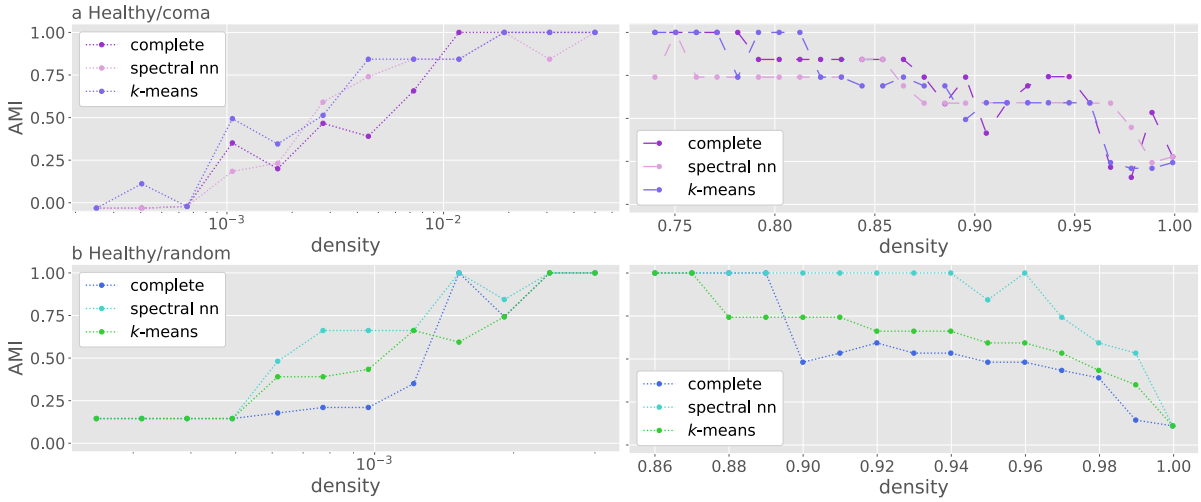


Fig. 5.3.2 Adjusted mutual information score (AMI) for density edge overlaps: (a) 20 healthy controls and 17 patients in a comatose state (b) 20 healthy controls and 17 random matrices (symmetric, with normal i.i.d. off-diagonal elements).

Null models. Both Erdös-Rényi and phase randomization models exhibit fast percolation, reaching a single connected component at low density (Figure 5.3.3a). The β_1 and β_2 curves also have similar behavior, with a localized peak where the maximal value is slightly lower for the phase randomization model (Figure 5.3.3b and Figure 5.3.3c). The Zalesky and spatio-temporal models exhibit slightly different behavior, with a slower decrease in the β_0 curve (Figure 5.3.3d), particularly in the case of the spatio-temporal model. The β_1 and β_2 features are on average more present in the Zalesky graphs than in the spatio-temporal graphs, and tend to appear at lower densities in the former than in the latter (Figure 5.3.3e and Figure 5.3.3f).

Betti numbers hence seem to capture some characteristics that are specific to each null model.

Real data. Betti curves highlight the importance of the preprocessing of brain graphs, and synthesize how it influences the dependencies of the large number of random variables at play. Here, we compare three preprocessing strategies. The first (HCP) corresponds to only taking the wavelet correlation between time series, while the other ones include a regression on white matter and cerebrospinal fluid (WM & CSF), with the last adding global signal regression (GSR), which are common preprocessing step. For β_0, additional signal regressions make the curve fall faster, meaning that fewer regions are poorly connected to the main connected component (Figure 5.3.3g). For β_1, β_2, the curves follow each other at low and high densities but reach higher maxima (Figure 5.3.3h and Figure

Table 5.3.1 Graph density thresholds at which each clustering algorithm starts and stops to cluster the relevant groups perfectly. coma+/coma- correspond to the smallest/highest density threshold which perfectly identifies healthy and comatose subjects. randh+/randh- are analogously defined for random graphs and healthy subjects. randc+/randc- similarly corresponds to random graphs and comatose subjects.

	k-means	hierarchical	spectral
coma+	$(1.3 \pm 0.1) \cdot 10^{-2}$	$(8.8 \pm 0.3) \cdot 10^{-3}$	$(1.68 \pm 0.01) \cdot 10^{-2}$
coma-	0.81	0.79	0.75 ± 0.02
randh+	$(2.3 \pm 0.3) \cdot 10^{-3}$	$(1.7 \pm 0.3) \cdot 10^{-3}$	$(1.4 \pm 0.3) \cdot 10^{-3}$
randh-	0.87	0.90	0.97
randc+	$(1.3 \pm 0.1) \cdot 10^{-2}$	$(1.0 \pm 0.1) \cdot 10^{-2}$	$(4.9 \pm 0.1) \cdot 10^{-2}$
randc-	0.79	0.70	0.75

5.3.3i). However, it seems that there is an important variability in the number of features, making it hard to reach conclusions that could be generalized over the entire cohort.

For the healthy subject/comatose patient cohort, there are limited deviations between the mean Betti curve and they are restricted to some density intervals (Figure 5.3.3j, Figure 5.3.3k, Figure 5.3.3l). However, the shape of the confidence interval for the coma class indicates that there are some patients who present Betti numbers noticeably higher than that of the class average and are probably driving it up.

Finally, it seems Betti numbers mostly capture preprocessing differences, and are not sufficient to clearly identify the different groups.

5.3.3 Persistence diagrams

With the mean Betti curves having shown their ability to capture some variability between the different classes of empirical and simulated networks, this section explores the possibilities offered by standard optimal transport distances between persistence diagrams.

Examples of persistence diagrams are presented in Figure 5.3.4. The Erdös-Rényi model and the phase randomization model show a large number of β_1, β_2 features, clear separation between the different feature scales where the β_1 or β_2 features dominate, and important variations in the lifetime of the different features. The example of an HCP subject, comatose patient, spatio-temporal and Zalesky model all exhibit similar behavior: fewer features with most of them having a short lifetime, and fuzzier separation between β_1 and β_2-dominated scales.

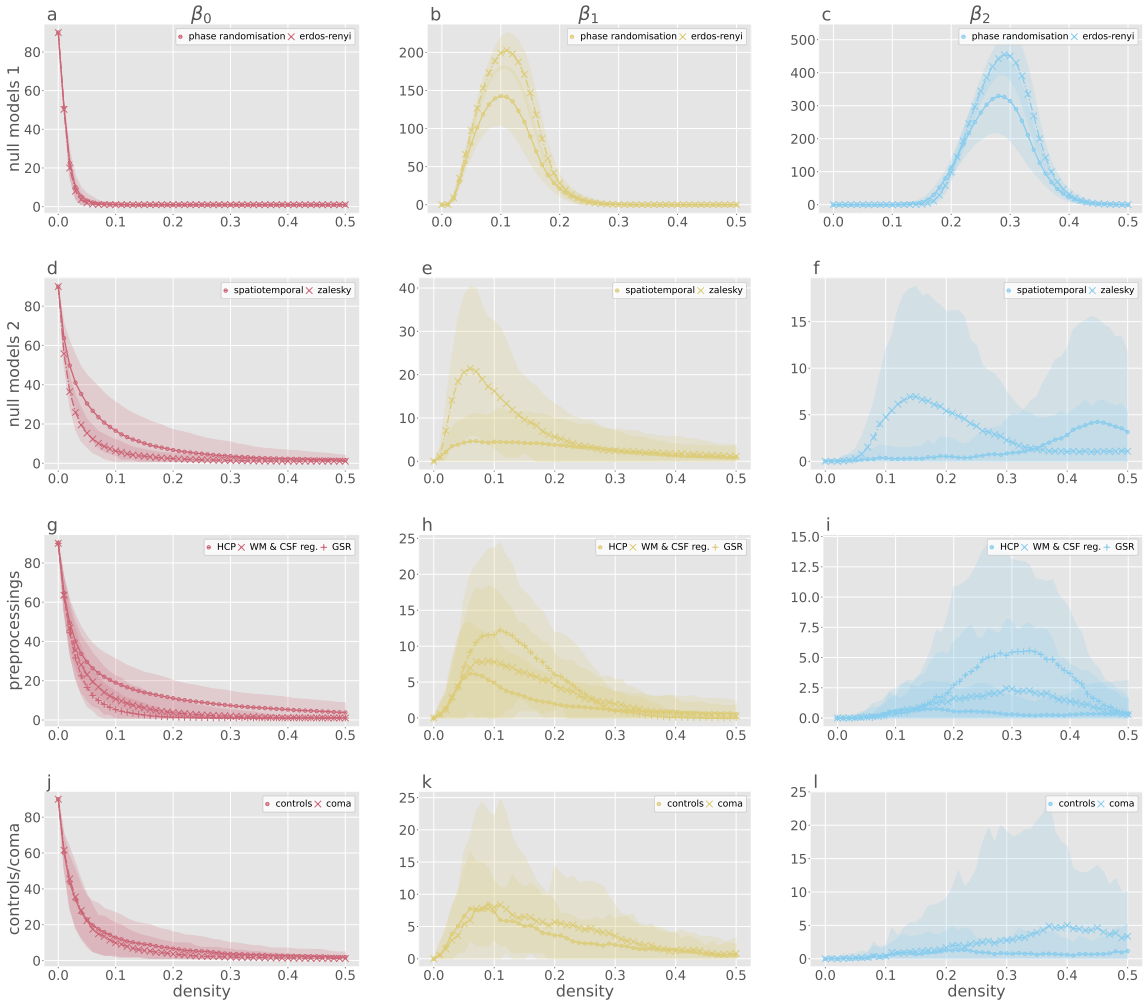


Fig. 5.3.3 From left to right: Betti curves for β_0, β_1 and β_2. (a) (b) (c) Null model Betti curves: phase randomization and Erdös-Rényi (d) (e) (f) Null model Betti curves: spatio-temporal and Zalesky (g) (h) (i) Empirical functional connectivity in HCP subjects with different preprocessings, (j) (k) (l) healthy subjects and comatose patients. Shaded areas give a 95% confidence interval around the mean β_i value at the given density.

Furthermore, we consider the distance between persistence diagrams without label information (cf. Figure 5.3.5). For the healthy subject and comatose patient cohort, the 1-Wasserstein distance does not capture the difference between the two groups. Most subjects and patients seem equidistant to each other, with a handful of outliers being a greater distance away from the main group. The bottleneck distance presents a similar behavior.

Meanwhile, for the null models, the Wasserstein distance clearly allows to discriminate between three groups: phase randomization, Erdös-Rényi and a third group including

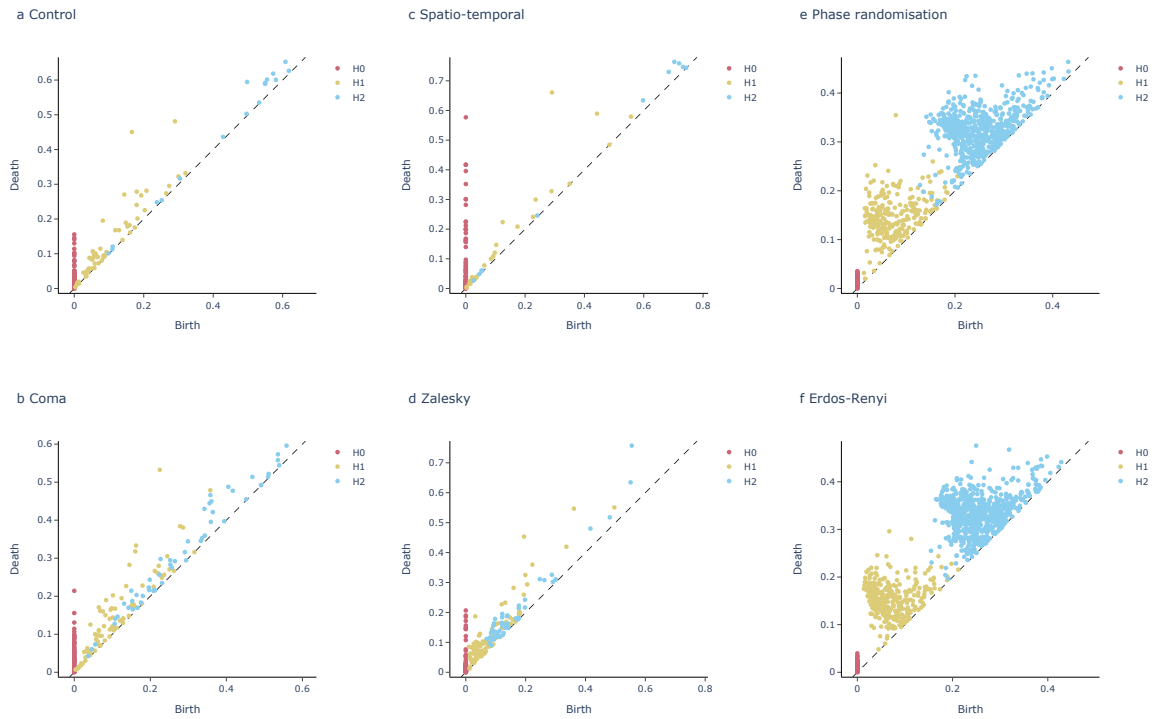


Fig. 5.3.4 persistence diagrams (a) of functional connectivity of a healthy subject (b) of a comatose patient (c) of a spatio-temporal null model (d) of a Zalesky null model (e) of a phase randomization null model and (f) of an Erdös-Rényi model.

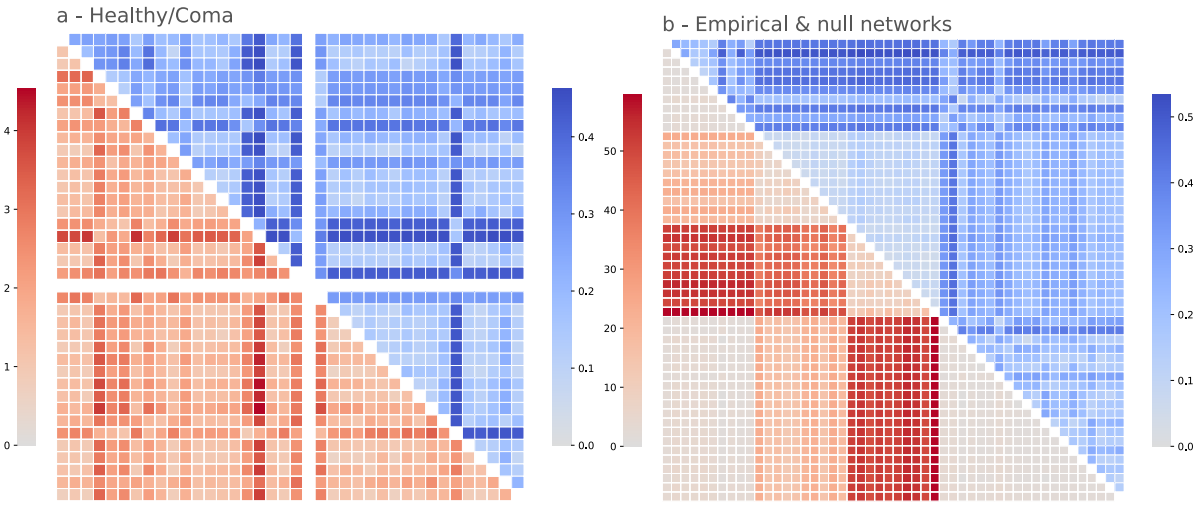


Fig. 5.3.5 **Upper triangle:** Bottleneck distance. **Lower triangle:** 1-Wasserstein distance. (a) Distance matrices for 20 healthy subjects and 17 comatose patients and (b) 10 healthy subjects with 10 realizations of each of the four null models (phase randomization, Erdös-Rényi , spatio-temporal, Zalesky).

healthy subjects, spatio-temporal, and Zalesky models. The bottleneck distance only separates the latter from a single group containing both phase randomization and Erdös-Rényi models.

5.3.4 Node-label information

It appears that the persistence diagram alone is not enough to identify fine differences in true functional network data, unlike a simple overlap approach combined with density-based thresholding. Hence, taking into account the node ordering is essential to effectively compare functional networks. Thus, we evaluate our proposed node-label informed distance on the persistence diagrams.

First, we apply the overlap distance to the coma dataset and to null models at a density of 0.05 (approximately the percolation density in random graphs, but this remains an arbitrary choice). Results are reported in the upper diagonal of the matrices in Figure 5.3.6. A simple inspection of the distance matrix reveals a high separation between comatose and healthy subjects. Similarly, in the comparison of null models and real data, the overlap distance clearly groups apart healthy subjects from the null models. Nevertheless, this distance does not differentiate between the various null models.

Then, we compute the edge comparison optimal transport distance which is reported in the lower diagonal of matrices in Figure 5.3.6. The distance is obtained by considering H_0, H_1, and H_2 homological features. A clear separation between the healthy subjects and the comatose patients is apparent. Similarly to the overlap distance, the optimal transport edge comparison distance separates real data from null models. Moreover, contrary to the overlap distance, our proposal clusters together the phase randomization and Erdös-Rényi models.

Finally, we quantitatively compare all the considered distances by their reached adjusted mutual information score (AMI) for the class comparison task in the coma dataset (See Table 5.3.2).

The overlap, the Frobenius and our proposed edge comparison optimal transport $d_{\mathrm{e},012}$ $d_{\mathrm{e},01}$ distances all succeed in realizing a perfect separation of the healthy-coma cohort. However, the standard Wasserstein and Bottleneck distances do not manage to cluster the data, with an AMI in the order of 0.05. It should also be noted that, even if some distances are far from being able to separate the two groups, they can still be noticeably better than random assignments, see Table 5.A.1.

Node label information is hence crucial to high-quality persistence diagram comparison.

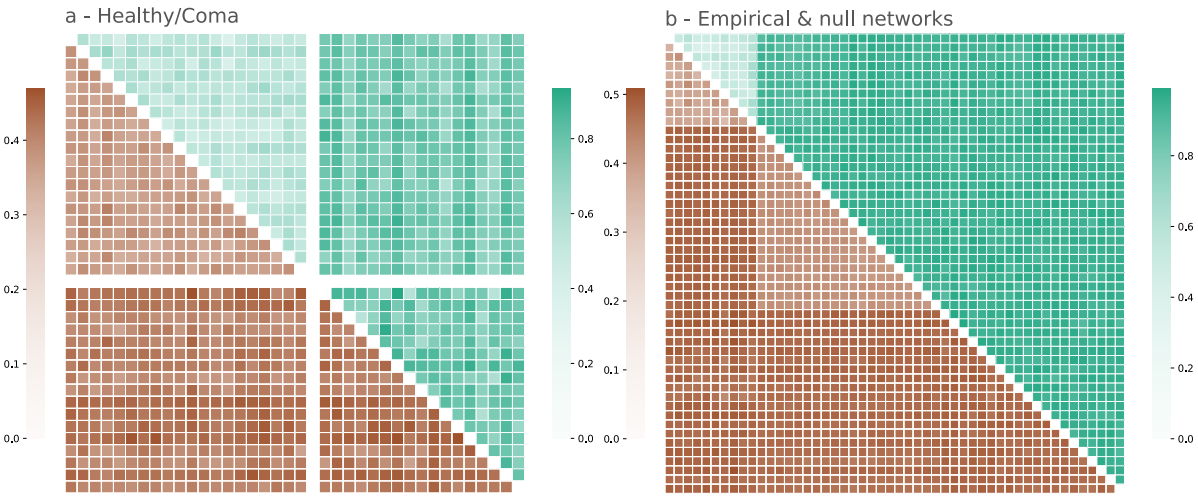


Fig. 5.3.6 Optimal transport $d_{e,012}$ between edges of the persistence diagrams (with $\beta_0, \beta_1, \beta_2$ features, lower triangle) and overlap distances between edges at density 0.05 (upper triangle), (a) of 20 healthy subjects and 17 comatose patients (left) and (b) 10 healthy subjects with 10 realizations of each of the 4 null models (right, phase randomization, Erdös-Rényi, spatio-temporal, Zalesky).

5.4 Discussion

5.4.1 Threshold

As far as the authors know, this work is the first attempt to use graph-density weighted persistence diagrams for the analysis of dense weighted graphs. This original approach is grounded in our results presented in Figure 5.3.1 where the Frobenius distance of inter-correlation matrices fails to detect any difference between healthy subjects and patients. It seems that in functional brain connectivity, inter-correlation rank is more informative than its value. Commonly, functional connectivity matrices are thresholded before performing downstream analysis. A common task in neuroimaging research is to differentiate multiple classes. Thus, the threshold can be defined to optimize the class separation as it is done in Table 5.3.1. The inter-subject comparison appears easier when simply defining a distance through the intersection of the edge sets of thresholded, binary graphs. Yet, this exclusively requires a supervised approach. Indeed, in these cases, the density threshold needs to be learned on data with known class assignments. The benefit of this approach is also highlighted by the perfect AMI score of the overlap distance-based class comparison reported in Table 5.3.2. This is the only distance to have a perfect score in all three considered algorithms. However, it should be noted that the distance d_{O} was applied to a thresholded matrix, where the density was chosen to

Table 5.3.2 Adjusted mutual information score of different distances for the comparison of healthy subjects and comatose patients. The distances are $d_{\mathrm{O},\tau=0.05}$ the edge-overlap distance for thresholded matrices at density level 0.05, d_{F} the Frobenius-distance, W_p the p-Wasserstein distance between persistence diagrams, Bo the Bottleneck distance between persistence diagrams, $d_{\mathrm{e},k_1k_2k_3}$ the edge-optimal transport distance on the persistence diagram with homology features $H_{k_1}, H_{k_2}, H_{k_3}$.

	k-means	hierarchical	spectral
$d_{\mathrm{O},\tau=0.05}$	1	1	1
d_{F}	1	0.59	1
W_1	0.04	0.04	0.03
W_2	0.06	0.06	0.06
Bo	0.06	0.06	0.12
$d_{\mathrm{e},012}$	1	0.32	1
$d_{\mathrm{e},0}$	0.74	0.32	0.74
$d_{\mathrm{e},1}$	0.16	0.06	0.07
$d_{\mathrm{e},2}$	0.05	-0.02	0.05
$d_{\mathrm{e},01}$	0.84	0.41	1
$d_{\mathrm{e},12}$	0.00	0.00	0.25

maximize the AMI. This means these results might not hold in general settings where the threshold choice is not optimized for the AMI, or in unsupervised cases.

5.4.2 Graph comparison and label information

Recent developments in the study of complex systems and the increase of data that can be naturally modeled as networks have led to an explosion of graph-centric problems. From this arises the need to develop approaches to compare graphs, for example, in order to establish the properties of a given molecule, proteins, etc. Despite the great interest in graph distances, relatively few methods are engineered to take into account edge and node labels, compared with the proliferation of permutation invariant distances (Tantardini et al., 2019). While introducing labeled and weighted edges brings an additional layer to the task, it allows leveraging alternative approaches in order to quantify graph similarity.

The traditional persistent homology approach allows to compare brain functional connectivity networks without requiring a threshold choice or supervised data. Indeed, one of the main arguments for the use of persistent homology is its multiscale approach. However, classical persistent distances fail to capture the variability between the comatose patients and healthy subjects (Figure 5.3.5). This might be due to the permutation invariant representation which does not take into account node labeling. This is especially

precarious for functional brain networks, where nodes are brain regions and are obviously distinct from each other.

Moreover, standard methods might be limited to discriminating groups with radically different underlying graph structures, as it happens for null models in Figure 5.3.5. This issue can be partially solved by including in the persistence diagram distance the label information of the edges associated with the appearance and disappearance of each feature.

This allows a considerable increase in the AMI score, reaching a perfect separation of the coma and control groups. While it offers an appealing baseline, it remains somewhat limited from an interpretation standpoint. Nevertheless, the H_0 persistence might be interpreted as an implicit node-wise threshold, where one keeps roughly only the most significant edge for each node.

5.4.3 Real-world data

Our reported persistent homology results in real data with different preprocessing reveal the effect of preprocessing on the Betti numbers, and thus on high-order graph structure (Fig. 5.3.3). This highlights the importance of documenting the applied framework for inferring brain networks from fMRI time series. This also suggests that studies on the effect of preprocessing on connectivity require more attention. Our application to the coma cohort is highly valuable since discriminating between subjects within different states of consciousness is a difficult task. With the exception of the Frobenius and traditional persistence diagram distances, all the considered distances manage to perfectly cluster patients and controls (Figure 5.3.5, Table 5.3.2). Interestingly, the pairwise distance between patients and controls is approximately the same as the distance between patients. It would seem that healthy controls are organized along a similar pattern while comatose patients are not.

5.4.4 Null models

A qualitative comparison of persistence diagrams shows that the Zalesky and the spatio-temporal model can roughly reproduce the persistent homology of empirical functional brain networks (Figure 5.3.4). The persistence diagrams of Phase randomization and Erdös-Rényi models are similar one to another, but markedly different from the others.

This visual inspection is reinforced by the results on the considered distances (Figures 5.3.5,5.3.6). Interestingly, when considering the W_1 distance the real data are grouped together with the Zalesky and spatio-temporal model. This means that real data birth

and death feature distribution can be approximated by considering a dominant Gaussian signal and adjusting the noise of individual regions in order to match the inter-correlation distribution or by capturing limited spatial and temporal auto-correlations from the fMRI data. Meanwhile, the amplitude of the persistence diagrams appears to vary across real data and these two models. This leads to high bottleneck distances and prevents grouping them all together. The opposite is observed for phase randomization and Erdös-Rényi models: they are grouped together by bottleneck distance and not by W_1. Moreover, they are also gathered by the edge comparison optimal transport, suggesting they might produce similar label feature distribution.

Both considered label-based distances differentiate the spatio-temporal and Zalesky null models from empirical data, but not from each other (Figure 5.3.6). This might be expected for the latter, since it does not take into account any label information to generate surrogate matrices, but is more surprising for the former. Hence, these label-dependent distances demonstrate that even null models that input some kind of spatial information do not reproduce label-dependent behavior.

None of the null models we consider manage to reproduce the real data label organization. From a quality control perspective, this helps ensure that the inferred networks carry meaningful information and are not overcome by noise.

5.4.5 Limitations and perspectives

For somewhat large networks, persistent homology stays limited to the lower dimensions, as the computational cost for computing higher-order features increases exponentially as it requires finding an arrangement of cliques. Although in this case, as the dependencies seem strong, finding high-order features would be surprising, it cannot yet be ruled out. Furthermore, in functional brain networks, first and second-order homology features seem to be short-lived, limiting the interest in persistent homology, where longer-lived features are the signature of particular topological invariants and are the main attribute that is targeted by persistent homology. In short, persistent homology appears to capture some of the texture of functional brain networks but does not uncover larger-scale organization.

The approach of building simplicial complexes using inter-correlation matrices has been criticized as a heavily limited framework for the exploration of higher-order interactions (Rosas et al., 2022). An alternative would be to consider information measures able to capture synergistic behavior and, either use them to investigate general properties of the system (Rosas et al., 2019) or to build simplicial complexes using the tricorrelation and upwards. The second option is strongly limited by the high combinatorial price to pay, but could be implemented on a restricted set of nodes in the graph. Additionally,

approaches based on density could solve one of the big issues in studying higher-order interactions: the lack of an equivalent to correlations. Indeed, there is no equivalent to the Cauchy-Schwartz inequality for trilinear and higher dimensional forms, and hence, for example, the co-skewness of 3 random variables:

$$S(X,Y,Z) = \frac{\mathbb{E}(X-\mathbb{E}(X))\mathbb{E}(Y-\mathbb{E}(Y))\mathbb{E}(Z-\mathbb{E}(Z))}{\sigma_X\sigma_Y\sigma_Z} \tag{5.13}$$

is not bounded and its usefulness for building a filtration is not straightforward. If instead, one looks at the triangle density (or, 2-simplex density), for triangles ordered using $|S|$, one would be able to obtain two-dimensional homology that truly takes into account triadic dynamics (and this stays true for higher-order interactions). In practice, this could prove beneficial, as it should yield a finer understanding of the dynamics. Nevertheless, it stays limited by the exponential growth of the number k-simplices when k increases.

It should also be noticed that these works have been done on a restricted cohort. While this has not been investigated in this setting, it is becoming more and more apparent that small sample sizes can have calamitous consequences on the confidence in results of functional MRI studies (Marek et al., 2022). This presents multiple challenges, notably the need for analysis procedures that scale well and for an increase in data availability. One can easily see how problematic this is in the case of comatose patients, as their ability to give informed consent is at best uncertain (Bruni et al., 2019; White et al., 2020).

5.5 Conclusion

To the best of our knowledge, our proposed graph-density-based filtration is the first attempt to include density levels in weighted graph persistent filtration instead of inter-correlation values. Moreover, we propose to include label information when comparing non-permutation invariant applications. Particularly, this is required in functional brain networks where nodes are associated with brain regions and are not perfectly exchangeable. We evaluate our approach on both real data and null models.

While standard persistent homology fails in discriminating healthy controls from comatose patients, our label-informed distance addresses this weakness. In this task, our distance is comparable to simple overlap or matrix distances. Furthermore, our proposal does not require the choice of an arbitrary threshold and is more informative.

Finally, these label-dependent distances show that node information included in some null models does not constrain the model enough to be close to the real data. This suggests new objectives in the design of null models for brain connectivity.

Although the application of persistent homology for the comparison of functional brain networks remains limited, it offers some understanding of the role of multiscale approaches and of higher-order interactions. It also helps to bring new insights about the importance of integrating node label information for quantitative functional brain network comparison.

Appendix

5.A AMI Scores for random labels

Figure 5.A.1 presents AMI quantiles of a set of $1.5 \cdot 10^6$ random binary labels of length 37 (matching the number of points in the healthy/coma dataset) with the true data set labels. Since $1.5 \cdot 10^6 / 2^{37} \simeq 10^{-5}$, roughly one hundred thousandths of the possible labelings have been explored. The mean AMI value is $\sim 10^{-5}$, providing a useful sanity check. The maximal value is ~ 0.66. Figure 5.A.1 a shows that the clustering score stays close to zero and is inferior to 0.1 before the top 1% assignments.

Leveraging this allows for a finer understanding of AMI values in Table 5.3.2. This table is reproduced in Table 5.A.1, with instead of the AMI values, the (estimated from $1.5 \cdot 10^6$ samples) probability that a random assignment is better than the predictions with the given distance matrix and clustering algorithm. Although this shows that it is relatively rare for those predictions to be completely meaningless (most of them outperform at least 90 % or 95% of random labelings), they remain inoperative for practical predictions. This also demonstrates that, for distances that can perfectly cluster the two groups for one or two algorithms, the other predictions still significantly outperform random labels.

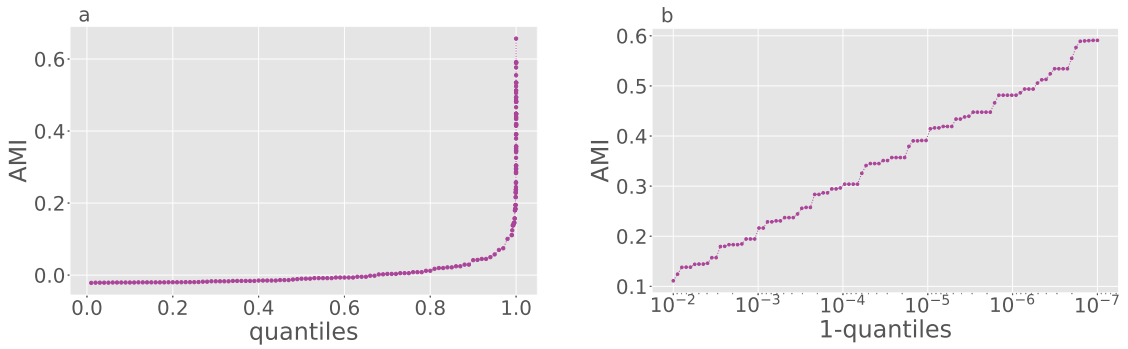


Fig. 5.A.1 (a) Plot of the AMI quantiles (b) Plot of the AMI quantiles for quantiles in the range $[1 - 10^{-2}, 1 - 10^{-7}]$.

Table 5.A.1 Probability that a randomly sampled label assignment performs better, with regards to the adjusted mutual information score, on labels predicted by a given algorithm on the different distance matrices.

	k-means	hierarchical	spectral
d_{O}	0	0	0
d_{F}	0	$6.7 \cdot 10^{-7}$	0
1-WA	0.10	0.10	0.11
2-WA	$5.0 \cdot 10^{-2}$	$5.0 \cdot 10^{-2}$	$5.0 \cdot 10^{-2}$
Bo	$5.0 \cdot 10^{-2}$	$5.0 \cdot 10^{-2}$	$9.4 \cdot 10^{-3}$
$d_{\mathrm{e},012}$	0	$5.8 \cdot 10^{-5}$	0
$d_{\mathrm{e},0}$	0	$5.8 \cdot 10^{-5}$	0
$d_{\mathrm{e},1}$	$2.9 \cdot 10^{-3}$	$5.0 \cdot 10^{-2}$	$3.6 \cdot 10^{-2}$
$d_{\mathrm{e},2}$	$5.5 \cdot 10^{-2}$	0.82	$5.5 \cdot 10^{-2}$
$d_{\mathrm{e},01}$	0	$1.0 \cdot 10^{-5}$	0
$d_{\mathrm{e},12}$	0.32	0.32	$3.5 \cdot 10^{-4}$

Chapter 6

Clustering-Based Edge Weight Estimation in Connectivity Networks

Clustering-Based Inter-Regional Correlation Estimation

We highlighted in Chapter 5 some possible applications of weighted connectivity networks. Yet, in Chapter 4 we had also started drawing attention to the negative impact of intra-correlation on the estimation of connectivity network weights, that is, inter-correlation. It is then all the more essential we use consistent inter-correlation estimators when learning connectivity networks. Therefore, we present in this chapter a novel non-parametric estimator of the correlation between grouped measurements of a quantity in the presence of noise. The challenge resides in the fact that both noise and intra-regional correlation lead to inconsistent inter-regional correlation estimation using classical approaches. While some existing methods handle either one of these issues, no non-parametric approaches tackle both simultaneously. To address this problem, we propose to leverage hierarchical clustering to gather together highly correlated variables within each region prior to inter-regional correlation estimation. We provide consistency results, and empirically show our approach surpasses several other popular methods in terms of quality. We also provide illustrations on real-world datasets that further demonstrate its effectiveness.

This chapter is based on the following contribution, which was presented at an international conference and published in a journal.

Lbath, H., Petersen, A., Meiring, W., and Achard, S. (2022b). Clustering-based inter-group correlation estimation. In *ICSDS 2022 - IMS International Conference on Statistics and Data Science*, Florence, Italy

Lbath, H., Petersen, A., Meiring, W., and Achard, S. (2023). Clustering-based inter-regional correlation estimation. *Computational Statistics & Data Analysis*, page 107876

6.1 Introduction

Correlation estimation is integral to a wide range of applications, and is often the starting point of further analyses. However, data are often contaminated by noise. If data are additionally inherently divided into separate, and study-relevant groups, inter-group correlation estimation becomes all the more challenging. Such datasets are often encountered in spatio-temporal studies, such as single-subject brain functional connectivity network estimation, where voxel-level signals acquired via functional Magnetic Resonance Imaging (fMRI) are grouped into predefined spatial brain regions (De Vico Fallani et al., 2014). This work is relevant as well to other fields, such as organizational studies, where individuals are grouped by organization (Ostroff, 1993). As such, we will be using the words group, region, and parcellation interchangeably. In these contexts, measurement replicates of each individual element, most often collected across time, are available and used to compute the sample correlation between different regions. These elements are grouped according to a parcellation which is fixed and corresponds to a practical reality, like anatomical brain regions in fMRI studies. As a result, regions could themselves be inhomogeneous. This work hence aims to estimate inter-regional correlation, later shortened to inter-correlation, no matter the quality of the parcellation.

However, both noise and arbitrary within-region correlation, later called intra-correlation, lead to inconsistent inter-correlation estimation by Pearson's correlation coefficient (Ostroff, 1993; Saccenti et al., 2020). Indeed, it has been established in various contexts that correlation is underestimated in the presence of noise (Ostroff, 1993; Matzke et al., 2017; Saccenti et al., 2020). Furthermore, data are often high dimensional, which presents a challenge of its own. In practice, including many fMRI studies, variables hence are commonly spatially averaged by regions prior to inter-correlation estimation (Achard et al., 2006; De Vico Fallani et al., 2014). Yet, intra-correlation may be weak, which would lead to overestimation of inter-correlations (Wigley et al., 1984). This phenomenon may also be compounded by unequal region sizes (Achard et al., 2011). Thus, standard correlation estimators are not well-suited for the setting of grouped variables under noise

contamination. Nonetheless, simultaneously tackling noise and intra-group dependence structures can be quite difficult, especially in a non-parametric setting. Failing to do so can be especially problematic for downstream analyses. For instance, in functional connectivity network estimation, a threshold is often applied to sample inter-correlation coefficients in order to identify edges between brain regions. Under- or over-estimation of the inter-correlation would then lead to missing or falsely detecting edges.

To address these problems, we present a data-driven, and non-parametric, approach with an astute intermediate aggregation. First, we propose to gather together highly correlated variables within each region. To this end, variables are projected onto a space where Euclidean distance can serve as a substitute to sample correlation, with lower values of the former corresponding to higher correlations. Hierarchical clustering with Ward's linkage (Ward, 1963; Murtagh and Legendre, 2014) is then applied to the projected variables within each region, resulting in intra-regional clusters of highly correlated variables. Within each intra-regional cluster, these variables are next spatially averaged. For each pair of regions, a sample correlation is then computed for each pair of cluster-averages from different regions. Our approach hence provides a distribution of the sample inter-correlations between each pair of regions, containing as many sample correlations as there are pairs of clusters from the two regions. For a point estimate of the inter-regional correlation for a given pair of regions, the average of the sample inter-correlation coefficients can then be considered. We summarize our main contributions as follows:

- We propose a novel non-parametric estimator of inter-regional correlation that offsets the combined effect of noise and arbitrary intra-correlation by leveraging hierarchical clustering.
- Based on the properties of hierarchical clustering with Ward's linkage, we derive the limiting behavior of our estimator for an appropriate choice of the cut-off height of the dendrograms thus obtained.
- We then empirically corroborate our results about the impact of the cut-off height on the quality of the estimation. We also show our proposed inter-correlation estimator outperforms popular estimators in terms of quality, and illustrate its effectiveness on real brain imaging datasets.

6.2 Related Work

We recall here relevant related work that were first detailed in Chapter 2. In the context of functional connectivity, the vast majority of papers that build correlation networks first average signals within each brain region for each time point, before computing Pearson's correlation across time, possibly after wavelet or other filtering, e.g., (Achard et al., 2006; Bolt et al., 2017; Ogawa, 2021; Zhang et al., 2016). Nevertheless, and as mentioned in the previous section, the correlation of averages overestimates the true correlation when intra-regional correlations are weak, while high noise may lead to underestimation. It was also empirically observed in fMRI data that the application of spatial smoothing, which is a common preprocessing step to reduce the effect of noise, causes the inter-regional correlations to be overestimated (Liu et al., 2017a).

Several methods tackling the impact of intra-correlation on the estimation of inter-correlation have been proposed in familial data literature, e.g., (Elston, 1975; Rosner et al., 1977; Srivastava and Keen, 1988; Wilson, 2010). These approaches nonetheless do not address the impact of noise. Moreover, they require normality assumptions on the samples, while we provide consistency guarantees for our proposed estimator that do not require parametric assumptions on the signal distribution. Bayesian inference methods have been proposed to offset the effect of measurement errors (Matzke et al., 2017). However they require a careful choice of priors, in addition to only handling pairs of variables, as opposed to groups of variables—which is what we are interested in. Robust correlation estimation has also been extensively investigated but mostly for specific distributions, such as contaminated normal distributions (Shevlyakov and Smirnov, 2016) or with heavy tails (Lindskog, 2000), whereas we are interested in robustness to noise and weak intra-group dependence. Furthermore, groups of variables are not considered either. Cluster-robust inference in the presence of both noise and within-group correlation has been studied in the econometric literature (Cameron and Miller, 2015). However, inter-correlation, which is the quantity we aim to estimate in this work, is assumed to be zero. To the best of our knowledge, we are the first to propose a method to simultaneously tackle the impact of noise and within-group inhomogeneity to estimate inter-correlation in a non-parametric fashion.

6.3 Preliminaries

From this point forward, and without loss of generality, we will focus on spatio-temporal contexts. In particular, we are motivated by an application to brain fMRI data where

individual observed variables correspond to blood-oxygen-level-dependent (BOLD) signals that are assigned to *voxels*, and are grouped by *regions*. Nonetheless, the following results can be applied to any dataset of grouped measurements of a quantity. In this section we define our notation and model, together with the inter- and intra-correlation coefficients.

Throughout this chapter we consider two regions, generically denoted A and B. In reality, datasets will involve a potentially large number of regions but, for the purpose of correlation network construction, the correlations can be estimated in a pairwise fashion at the regional-level. Let $X_1^A, \ldots, X_i^A, \ldots, X_{N_A}^A$ denote N_A spatially dependent latent (unobserved) random variables in region A, each variable corresponding to an individual voxel in that region. Let $\epsilon_1^A, \ldots, \epsilon_i^A, \ldots, \epsilon_{N_A}^A$ represent random noise variables. We assume that the latent process X_i^A at each voxel i is contaminated by noise ϵ_i^A, so that the observed variables Y_i^A in region A are

$$Y_i^A = X_i^A + \epsilon_i^A, \quad i = 1, \ldots, N_A. \tag{6.1}$$

We assume within-region homoscedasticity of both signal and noise, i.e.,

$$\sigma_A^2 = \mathrm{Var}\left(X_i^A\right), \ \gamma_A^2 = \mathrm{Var}\left(\epsilon_i^A\right), \qquad i = 1, \ldots, N_A.$$

Analogously we define N_B, X_j^B, ϵ_j^B, Y_j^B, σ_B^2 and γ_B^2, for region B and voxels $j = 1, \ldots, N_B$. We assume the noise variables are spatially uncorrelated both within and across regions, and that they are also uncorrelated to the latent state both within and between regions.

A critical reality of the observed data is the *intra-correlation* or Pearson's correlation between any pair of random variables *within* a given region A. We denote by $\eta_{i,i'}^A$ the intra-correlation of the latent variables $X_i^A, X_{i'}^A$. We place no further constraints on the intra-correlation structure. Similarly, we define the *inter-correlation* as Pearson's correlation between any pair of random variables from two *distinct* regions. For a given pair of distinct regions, A, B, the inter-correlation between any pair of latent random variables X_i^A, X_j^B is assumed to be constant across voxels, and is denoted as $\rho^{A,B}$.

Consider now n temporally independent and identically distributed (i.i.d.) samples of all observed signals. That is, for each region A and voxel $i = 1, \ldots, N_A$, we have n i.i.d. observations $Y_i^A(t)$, $t = 1, \ldots, n$, each distributed as in (6.1) with the same intra- and inter-correlation properties as those outlined previously. In particular, for any time point $t = 1, \ldots, n$, and voxels i and j from distinct regions A and B, respectively, $Cov(Y_i^A(t), Y_j^B(t)) = \rho^{A,B}\sigma_A\sigma_B$. Denote by $\mathbf{Y}_i^A = [Y_i^A(1), \ldots, Y_i^A(t), \ldots Y_i^A(n)]$ the vector of observations for the i-th voxel of region A.

6.4 Proposed Inter-Correlation Estimator

After defining the sample correlation coefficient in Section 6.4.1, we highlight in Section 6.4.2 the impact of the combined presence of noise and intra-correlation, when using popular estimators of inter-correlation. In Section 6.4.3 we then propose an inter-correlation estimator that limits these effects. Consistency of our estimator is proved in Section 6.4.4.

6.4.1 Computing sample correlations

We denote by $\widehat{Cor}(\cdot,\cdot)$ the sample (Pearson's) correlation between any two equal-length vectors of samples. This corresponds to the zero-lag empirical cross-correlation in spatio-temporal studies. To be specific, suppose $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ are any vectors of the same length, and let $\overline{a} = n^{-1}\sum_{t=1}^{n} a_t$ and $\overline{b} = n^{-1}\sum_{t=1}^{n} b_t$ be the averages of their elements, respectively. Let $\mathbf{1}_n$ be the n-vector of ones, $\mathbf{a}^c = \mathbf{a} - \overline{a}\mathbf{1}_n$, and $\mathbf{b}^c = \mathbf{b} - \overline{b}\mathbf{1}_n$ their centered versions. With $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ being the Euclidean inner product and norm, respectively, we define

$$\widehat{Cov}(\mathbf{a},\mathbf{b}) = n^{-1}\langle \mathbf{a}^c, \mathbf{b}^c\rangle,\ \widehat{Var}(\mathbf{a}) = n^{-1}\left\|\mathbf{a}^c\right\|^2,\ \widehat{Cor}(\mathbf{a},\mathbf{b}) = \frac{\widehat{Cov}(\mathbf{a},\mathbf{b})}{\sqrt{\widehat{Var}(\mathbf{a})\widehat{Var}(\mathbf{b})}}. \quad (6.2)$$

Using this notation, the sample correlation between any two voxels i and j in regions A and B is

$$R_{i,j}^{A,B} = \widehat{Cor}(\mathbf{Y}_i^A, \mathbf{Y}_j^B). \quad (6.3)$$

Observe that this definition applies equally to sample inter-correlations ($A \neq B$) as well as intra-correlations ($A = B$).

6.4.2 Impact of noise and intra-correlation

Previously, Matzke et al. (2017) showed that the presence of noise attenuates the observed correlation. Indeed, this phenomenon is captured in the following result: from model (6.1) and Achard et al. (2020), $R_{i,j}^{A,B}$ converges almost surely to

$$\frac{Cov(Y_i^A, Y_j^B)}{\sqrt{(\sigma_A^2+\gamma_A^2)\cdot(\sigma_B^2+\gamma_B^2)}} = \frac{Cov(X_i^A, X_j^B)}{\sqrt{(\sigma_A^2+\gamma_A^2)\cdot(\sigma_B^2+\gamma_B^2)}}. \quad (6.4)$$

Therefore, if distinct regions A,B with latent signals observed contaminated by noise, $R_{i,j}^{A,B}$ is not a consistent estimator of true inter-correlation $\rho^{A,B}$ due to the presence of the noise variances in the denominator of (6.4). Furthermore, in settings where a

single point estimate of the inter-correlation of the unobserved latent signal between two regions is needed, the corresponding pairwise sample inter-correlation coefficients can be averaged to provide an estimator. Denoted $r^{AC}_{A,B}$, it corresponds to the ensemble estimator in familial data literature (Rosner et al., 1977):

$$r^{AC}_{A,B} = \frac{1}{N_A \cdot N_B} \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} R^{A,B}_{i,j}. \tag{6.5}$$

However, the latter is similarly impacted by noise.

As mentioned in Section 6.2, one of the most popular estimators in neuroimaging studies consists of spatially averaging the observation random variables within each distinct region for each time t, before computing the sample correlation between these averages. Specifically, define regional (spatial) averages $\overline{\mathbf{Y}}^A = N_A^{-1} \sum_{i=1}^{N_A} \mathbf{Y}^A_i$ and $\overline{\mathbf{Y}}^B = N_B^{-1} \sum_{j=1}^{N_B} \mathbf{Y}^B_j$.

Then this estimator is

$$r^{CA}_{A,B} = \widehat{Cor}(\overline{\mathbf{Y}}^A, \overline{\mathbf{Y}}^B). \tag{6.6}$$

Under model (6.1), and according to results from (Achard et al., 2020), together with intra-regional uncorrelatedness between latent and noise random variables, as well as inter-regional uncorrelatedness of noise, $r^{CA}_{A,B}$ converges almost surely to:

$$\frac{\rho^{A,B}}{\sqrt{\left[\frac{1}{N_A^2} \cdot \sum_{i,i'=1}^{N_A} \eta^A_{i,i'} + \frac{\gamma_A^2}{N_A \cdot \sigma_A^2}\right] \left[\frac{1}{N_B^2} \cdot \sum_{j,j'=1}^{N_B} \eta^B_{j,j'} + \frac{\gamma_B^2}{N_B \cdot \sigma_B^2}\right]}}, \tag{6.7}$$

where $N_A^{-2} \cdot \sum_{i,i'=1}^{N_A} \eta^A_{i,i'}$ is the spatial average of the pairwise latent intra-correlation coefficients within region A.

It follows from (6.7) that intra-correlation and noise both contribute to inconsistency of the inter-correlation estimator (6.6). Indeed, both quantities appear in the denominator. It is then apparent that the smaller the regions (smaller N_A), the higher the impact of noise on the correlation estimation. Additionally, the weaker the spatial intra-regional dependence, the larger the overestimation of the true inter-correlation. This effect may also be compounded when regions are large, as was observed by Achard et al. (2011). One would then need to have regions as large as possible, while having an average intra-correlation as close to 1 as possible in order to offset these biases. However, large regions tend to be inhomogeneous in practical scenarios, and thus tend to have low intra-correlation.

6.4.3 A clustering-based inter-correlation estimator

Based on these findings, we propose an inter-correlation estimator specifically designed to limit the combined effects of noise and intra-correlation. Instead of aggregating over entire regions, we propose to aggregate over small groups of highly intra-correlated variables (cf. Steps 1 and 2), before computing the correlation of the corresponding local averages (cf. Step 3).

Step 1: U-Scores Computation

To facilitate the grouping of the variables within each region, we can leverage U-scores to project the sample vectors $\mathbf{Y}_i^A$ onto a space where the Euclidean distance can be used as a proxy for the sample correlations. We could then apply any clustering algorithm in the U-score space. *U-scores* are an orthogonal projection of the Z-scores of random variables onto a unit $(n-2)$-sphere centered around 0. The U-score $\mathbf{U}_i^A$ of $\mathbf{Y}_i^A$ is defined by $\mathbf{U}_i^A = \mathbf{H}_{2:n}^T \mathbf{Z}_i^A$, where $\mathbf{H}_{2:n}^T$ is a $(n-1) \times (n-1)$ matrix obtained by Gramm-Schmidt orthogonalization, and $\mathbf{Z}_i^A$ the Z-score of $\mathbf{Y}_i^A$. We refer to (Hero and Rajaratnam, 2011) for a full definition. Sample correlations can then be expressed as an inner product of U-scores: $R_{i,j}^{A,B} = (\mathbf{U}_i^A)^T \mathbf{U}_j^B = 1 - \|\mathbf{U}_i^A - \mathbf{U}_j^B\|^2/2$, where $\mathbf{U}_i^A$, $\mathbf{U}_j^B$ are the U-scores of the ith and jth voxels in regions A and B, respectively, and $\|\cdot\|^2$ is the squared Euclidean distance.

Step 2: Clustering

Once the U-scores are calculated, any standard clustering algorithm can be applied to obtain homogeneous groups of variables within each region. Agglomerative hierarchical clustering with Ward's linkage (Ward, 1963; Murtagh and Legendre, 2014), which is closely related to the k-means algorithm (Hartigan and Wong, 1979), aims to minimize the intra-cluster variance, which implies a maximization of the intra-cluster correlation. More specif-ically, agglomerative hierarchical clustering starts by assigning each element, e.g., a voxel in our setting, to its own cluster. Then, clusters are iteratively merged according to a pre-defined rule. Ward's linkage specifies that, at each step, the pair of clusters to be merged is chosen to minimize the increase in the combined error sum of squares. We used the `hclust` function from the `stats` R package, with the `ward.D2` method and default parameters (Murtagh and Legendre, 2014). A comparison of different clustering methods, which empirically validates the use of Ward's linkage in our context, is presented in Section 6.5.3. In practice, the number of clusters generally needs to be specified. However, such a strategy, while often satisfactory in common clustering tasks,

such as exploratory analyses, does not provide any obvious theoretical guarantees on the homogeneity of the clusters, which is what we are interested in. Nevertheless, hierarchical clustering outputs a dendrogram,which indicates the between-cluster distance at which clusters are merged. It can then be cut off at a designated height to produce a clustering. Therefore, instead of setting a number of clusters, we propose to specify a cut-off height through which cluster radii, and by proxy intra-correlations, can be controlled to a certain extent (cf Theorem 6.4.1). Proofs can be found in the appendix.

Theorem 6.4.1. *For a region A, a fixed cut-off height h_A, and all clusters ν_A thus obtained, the spatial average of the sample intra-cluster correlation is bounded as follows:*

$$1 - \frac{h_A^2}{2} \;\leq\; \frac{1}{|\nu_A|^2} \sum_{i,i'=1}^{|\nu_A|} R_{i,i'}^{A,A} \;\leq 1, \tag{6.8}$$

where $|\nu_A|$ is the size of cluster ν_A.

Theorem 6.4.1 shows that through careful choice of the cut-off heights, clusters of highly correlated variables can be selected within each region. This choice can be guided by the ensuing observations about the maximum distance between U-scores within a given region, denoted by $h_A^{\max}$, which follow immediately from Theorem 6.4.1 and the fact that $1 - (h_A^{\max})^2/2 \;= \min_{i,i'=1,\ldots,N_A} R_{i,i'}^{A,A}$:

- if $h_A \geq h_A^{\max}$,

$$1 - \frac{h_A^2}{2} \;\leq\; \min_{i,i'=1,\ldots,N_A} R_{i,i'}^{A,A} \;\leq\; \frac{1}{|\nu_A|^2} \sum_{i,i'=1}^{|\nu_A|} R_{i,i'}^{A,A} \tag{6.9}$$

- and if $h_A \leq h_A^{\max}$,

$$\min_{i,i'=1,\ldots,N_A} R_{i,i'}^{A,A} \;\leq\; 1 - \frac{h_A^2}{2} \;\leq\; \frac{1}{|\nu_A|^2} \sum_{i,i'=1}^{|\nu_A|} R_{i,i'}^{A,A}. \tag{6.10}$$

Therefore, to ensure all clusters contain more than one voxel, the maximum distance between any two clusters of the region (i.e., the cut-off height) would need to be larger than the maximum distance between any two voxels within the region (i.e., $h_A^{\max}$). Thus, setting the cut-off height to $h_A^{\max}$ would ensure to obtain the smallest possible clusters guaranteed to contain at least two variables. Moreover, computing $h_A^{\max}$ is computationally inexpensive. It also does not depend on any ground-truth, which remains unknown in practice. Empirical comparisons of this data-driven choice with the optimal cut-off heights are made on simulated data in Section 6.5.2. As the optimal cut-off heights are

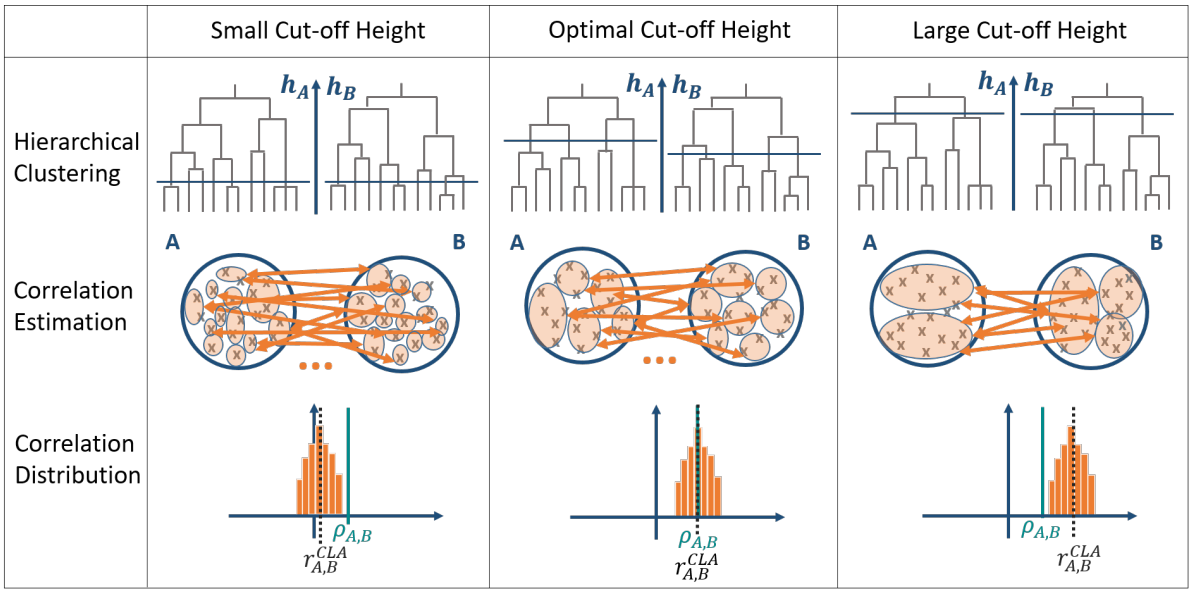


Fig. 6.4.1 Illustration of the inter-correlation estimation of a pair of regions for different cut-off heights. The top panel shows the dendrograms of the hierarchical clustering applied to each region. The horizontal line over each dendrogram indicates the cut-off heights h_A, h_B. The grey crosses in the middle panel correspond to the random variables inside each regions, and are grouped into the resulting clusters (orange ellipses). The arrows represent the sample inter-correlation between the average of the variables inside each cluster (some arrows were left out to improve readability). The bottom panel displays the distribution of the pairwise sample inter-correlation. The true inter-correlation $\rho^{A,B}$ (solid line) is best approximated by the sample inter-correlation $r^{CLA}_{A,B}$ (dotted line) when the cut-off heights are neither too small nor too large.

not known in practice and cannot be computed from the data, these results demonstrate the practical effectiveness of setting the cut-off height to $h_A^{\max}$.

Step 3: Clustered Correlation Estimation

Once clusters are obtained within each region, the inter-correlation is estimated as follows. For two distinct regions A and B, for fixed cut-off heights h_A, h_B, and any two pairs of clusters ν_A, ν_B within each of these regions, we define the following cluster-level inter-correlation estimator:

$$r^{CLA}_{\nu_A,\nu_B} = \widehat{Cor}(\,\overline{\mathbf{Y}}^{\nu_A},\ \overline{\mathbf{Y}}^{\nu_B}\,), \tag{6.11}$$

where $\overline{\mathbf{Y}}^{\nu_A} = |\nu_A|^{-1}\sum_{i\in\nu_A}\mathbf{Y}^A_i$, and $\overline{\mathbf{Y}}^{\nu_B}$ is defined similarly. A distribution of sample inter-correlation coefficients is hence obtained for this pair of regions, as seen in Figure 6.4.1. As mentioned earlier, if a point estimate is needed, one can then simply average

the cluster-level estimates to derive the following regional-level estimator:

$$r_{A,B}^{CLA} = \frac{1}{N_A^{clust} \cdot N_B^{clust}} \sum_{\nu_A,\nu_B} r_{\nu_A,\nu_B}^{CLA}, \tag{6.12}$$

where N_A^{clust} is the number of clusters within region A. We refer to Algorithm 1 for a detailed description of our proposed clustering-based correlation estimation procedure for J regions.

Algorithm 1: Clustering-Based Correlation Estimation

input : N variables grouped in J regions with n samples each
output : Cluster-level and regional-level inter-correlation estimates

▷ Clustering
for *each region* A **do**
  Apply hierarchical clustering to A;
  Choose the cut-off height h_A;
  Cut the dendrogram at height h_A;
  for *each cluster* ν_A *in* A **do**
    $\overline{\mathbf{Y}}^{\nu_A} \leftarrow \sum_{i=1}^{|\nu_A|} \mathbf{Y}_i^A / |\nu_A|$;

▷ Correlation of local averages estimation
for *each pair of regions* A, B **do**
  for *each pair of clusters* ν_A, ν_B **do**
    $r_{\nu_A,\nu_B}^{CLA} \leftarrow \widehat{Cor}(\overline{\mathbf{Y}}^{\nu_A}, \overline{\mathbf{Y}}^{\nu_B})$
  $r_{A,B}^{CLA} \leftarrow \sum_{\nu_A,\nu_B} r_{\nu_A,\nu_B}^{CLA} / N_A^{clust} \cdot N_B^{clust}$

6.4.4 Consistency of the proposed estimator

The clusters derived in Algorithm 1 are data-driven, and thus random from a probabilistic perspective. To simplify analysis and allow us to demonstrate the expected behavior of the proposed estimator as the number of time points n grows, let us assume that clusters ν_A and ν_B are fixed. Then define the following quantity, which will be used in several of the subsequent results:

$$\rho_{\nu_A,\nu_B}^{CLA} = \frac{\rho^{A,B}}{\sqrt{\left[\frac{1}{|\nu_A|^2} \cdot \sum_{i,i'=1}^{|\nu_A|} \eta_{i,i'}^A + \frac{\gamma_A^2}{|\nu_A| \cdot \sigma_A^2}\right] \cdot \left[\frac{1}{|\nu_B|^2} \cdot \sum_{j,j'=1}^{|\nu_B|} \eta_{j,j'}^B + \frac{\gamma_B^2}{|\nu_B| \cdot \sigma_B^2}\right]}}. \tag{6.13}$$

Theorem 6.4.2. *Under the assumptions of model (6.1), for a fixed pair of clusters ν_A, ν_B, as n tends towards infinity,*

$$r^{CLA}_{\nu_A,\nu_B} \xrightarrow{a.s.} \rho^{CLA}_{\nu_A,\nu_B}. \tag{6.14}$$

The proof is detailed in the appendix. We obtain similar results for the regional-level point estimate $r^{CLA}_{A,B}$.

Corollary 6.4.1. *Under the same assumptions as Theorem 6.4.2, for two regions A, B, as n tends towards infinity,*

$$r^{CLA}_{A,B} \xrightarrow{a.s.} \frac{1}{N^A_{clust} N^B_{clust}} \sum_{\nu_A,\nu_B} \rho^{CLA}_{\nu_A,\nu_B}. \tag{6.15}$$

Corollary 6.4.1 is a direct consequence of Theorem 6.4.2.

Theorem 6.4.2 and Corollary 6.4.1 emphasize the fact that controlling the denominator of $\rho^{CLA}_{\nu_A,\nu_B}$ is key to obtaining a consistent estimator of $\rho^{A,B}$. This brings to light the influence of the cut-off height, and thereby the cluster size and intra-cluster correlation, on the consistency of the inter-correlation estimate, both at the cluster- and regional-level.

For a pair of regions A, B, as the cut-off heights h_A, h_B become larger, the impact of noise diminishes. Moreover, the clusters increase in size until there is only a single cluster left that corresponds to the entire region. Thus, for h_A, h_B sufficiently large, our proposed estimator $r^{CLA}_{\nu_A,\nu_B}$, and the corresponding point estimate $r^{CLA}_{A,B}$ are equal to the correlation of averages $r^{CA}_{A,B}$ mentioned earlier. Conversely, as h_A, h_B become smaller the maximum distance between U-scores within a cluster decreases, hence the minimal intra-cluster correlation increases (cf. Theorem 6.4.1). There are also gradually less variables within each cluster, until they eventually contain only a single variable. It follows that when $h_A, h_B = 0$, $r^{CLA}_{A,B}$ corresponds to a correlation estimate with no aggregation $r^{AC}_{A,B}$. This can be visualized in Figure 6.4.1, where sample correlation distributions are depicted for different cut-off heights.

Therefore, to simultaneously lessen the impact of noise and intra-correlation a trade-off is necessary between a sufficiently high cut-off height (to decrease the impact of noise), and a low enough height (to decrease the impact of intra-cluster correlation). Thus for suitable cut-off heights, we expect the limits of both $r^{CLA}_{\nu_A,\nu_B}$ and $r^{CLA}_{A,B}$ to be closer to the population inter-correlation $\rho^{A,B}$ than that of $r^{CA}_{A,B}$ and $r^{AC}_{A,B}$. We will empirically compare these three estimators in Section 6.5.3, where the results suggest that the data-driven cut-off height does indeed lead to improvement.

6.5 Experimental Results

In this section we empirically determine the optimal cut-off height, evaluate our proposed inter-correlation estimator on synthetic data, and illustrate our approach on real-world datasets.

6.5.1 Datasets

We first present the different datasets used in this chapter. Additional dataset information and preprocessing details are documented in Chapter 3.

Real-world datasets

A rat and a human brain fMRI dataset is used in this chapter.

Rat brain fMRI dataset. We apply our estimator on fMRI data acquired on both dead and anesthetized rats (Becq et al., 2020a,b). In this chapter we consider the following anesthetics: Etomidate (EtoL), Isoflurane (IsoW) and Urethane (UreL). The scanning duration is 30 min with a time repetition of 0.5 s. After preprocessing (Becq et al., 2020b), 25 groups of voxels, each associated with its BOLD signal with a number of time points in the order of thousands, were extracted for each rat. They correspond to rat brain regions defined by an anatomical atlas obtained from a fusion of the Tohoku and Waxholm atlases (Becq et al., 2020b). Region sizes vary from about 40 up to approximately 200 voxels.

Human Connectome Project. We also consider 35 subjects from the human connectome project (HCP), WU-Minn Consortium pre-processed (Glasser et al., 2013). Subjects were pseudonymized. Two fMRI acquisitions on different days are available for each subject. The scanning duration is 14 min and 24 s with a time repetition of 720 ms. A modified AAL template is used to parcellate the brain into 89 regions. The details of the pre-processing are available in (Termenon et al., 2016). Region sizes are in the order of thousands of voxels, and number of time points are in the order of thousands.

Synthetic datasets

We consider several synthetic datasets to evaluate our estimator. For each simulation, we simultaneously generate 800 independent samples of a pair of inter-correlated regions, containing each 60 intra-correlated variables that follow a multivariate normal distribution

with a predefined covariance structure contaminated by Gaussian noise. The inter-correlation is constant across all pairs of voxels. The different parameters are chosen to ensure the population covariance matrix of the two regions is positive semidefinite. For instance, one cannot generate a covariance matrix where both intra- and inter-correlation values are low.

Toeplitz covariance structure. We first generate 1-dimensional data with a Toeplitz intra-regional covariance structure (later denoted 1D Toeplitz). For each region, intra-correlation is defined such that it decreases as the distance between two variables increases: for any voxel i, i' in region A, $Cor(X_i^A, X_{i'}^A) = \max(1-|i'-i|/30, \eta_{min}^A)$, where $|i'-i|$ is the uniform norm between voxels i and i', and η_{min}^A the minimal population intra-correlation of a region A. In this chapter we consider several experimental settings by varying the population intra-correlation, inter-correlation and the variance of the noise. The sample pairwise correlation matrices of the observed signals are represented in Figure 3.1 for a low intra-correlation and a high intra-correlation setting with high noise.

Matérn covariance structure. Similarly we then simulate 3-dimensional data with a Matérn intra-regional covariance structure that depends on the Euclidean distance (later denoted 3D Matérn) (Ribeiro and Diggle, 2001). In this chapter, we set the smoothness parameter to $\kappa_A = \kappa_B = 70$ to maintain the positive-definiteness of the input covariance matrix. We then vary the range parameters ϕ_A, ϕ_B and the variance of the noise. The lower the range parameter, the lower the mean intra-correlation.

Spherical covariance structure. We then generate 3-dimensional data with a spherical intra-regional covariance structure that also depends on the Euclidean distance between voxels (later denoted 3D Spherical) (Ribeiro and Diggle, 2001). We vary the range parameters ϕ_A, ϕ_B and the variance of the noise. The lower the range parameter, the lower the mean intra-correlation.

6.5.2 Choice of the cut-off heights

In this section we empirically evaluate on the 1D-Toeplitz dataset the impact of the cut-off heights h_A, h_B on the proposed clustering-based correlation estimator. We also propose a heuristic to choose optimal cut-off heights.

We consider different scenarios, including one that loosely matches live rat data settings, where the noise is high and the intra-correlation low. For each simulated pair

of regions, and for various cut-off heights h_A, h_B, the squared error of the cluster-level estimators are computed and then averaged across the different clusters:

$$\text{ERROR} = \frac{1}{N^A_{clust} N^B_{clust}} \sum_{\nu_A, \nu_B} (r^{CLA}_{\nu_A,\nu_B} - \rho^{A,B})^2. \tag{6.16}$$

The resulting surfaces are displayed in Figure 6.5.1. The lower the error, the better the quality of the estimator. As expected from Theorems 6.4.1 and 6.4.2, the error is lowest (refer to the orange points in Figure 6.5.1) for cut-off heights that are neither too small nor too large. Moreover, when both the intra-correlation and the variance of the noise are low, the error is low, even for low cut-off heights, as there is no need to aggregate the data to obtain a consistent estimator. However, the error is high for large cut-off heights regardless of the scenario. Indeed, even in the high noise settings, intra-correlation still influences the inter-correlation, and this effect is compounded by that of the cluster size.

In Section 6.4.3, we proposed a computationally cheap heuristic to determine a suitable cut-off height. Empirically, it seems the maximum distance between U-scores within a given region A, $h^{\max}_A$, could indeed be an optimal cut-off height. It is represented by a yellow diamond in Figure 6.5.1. In fact, it seems to be located at the bottom of a valley and quite close to the minimal error for all settings.

We then compare our proposed optimal cut-off height, in terms of Mean Squared Error (MSE), to that obtained using a more standard criterion from the clustering literature: the maximum silhouette score. The Squared Error (SE) of a simulation-specific correlation estimate $r^{CLA}_{A,B}$ can be defined as

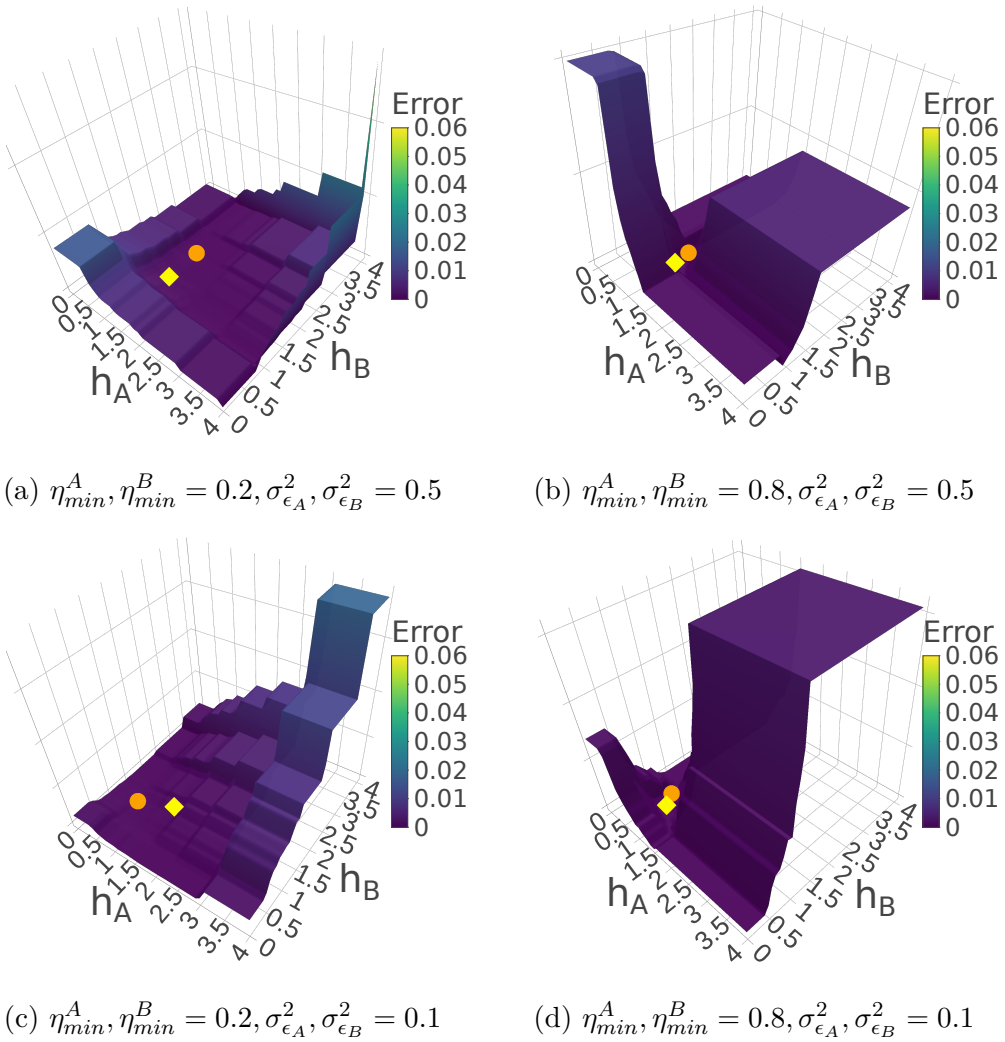
$$\text{SE} = (r^{CLA}_{A,B} - \rho^{A,B})^2. \tag{6.17}$$

In this section, the MSE is computed by averaging the SEs across 50 replicates. The MSE for varying intra- and inter-correlation values and a fixed high noise variance are depicted in Figures 6.5.2 and 6.5.3. The MSE is lower when using our proposed cut-off heights in all the considered scenarios.

From now on, and unless stated otherwise, we will hence estimate the inter-correlation using this optimal cut-off height.

6.5.3 Comparison with other methods

We then empirically evaluate our choice of clustering method and compare our proposed approach with other estimators in terms of MSE.



(a) $\eta^A_{min}, \eta^B_{min} = 0.2, \sigma^2_{\epsilon_A}, \sigma^2_{\epsilon_B} = 0.5$

(b) $\eta^A_{min}, \eta^B_{min} = 0.8, \sigma^2_{\epsilon_A}, \sigma^2_{\epsilon_B} = 0.5$

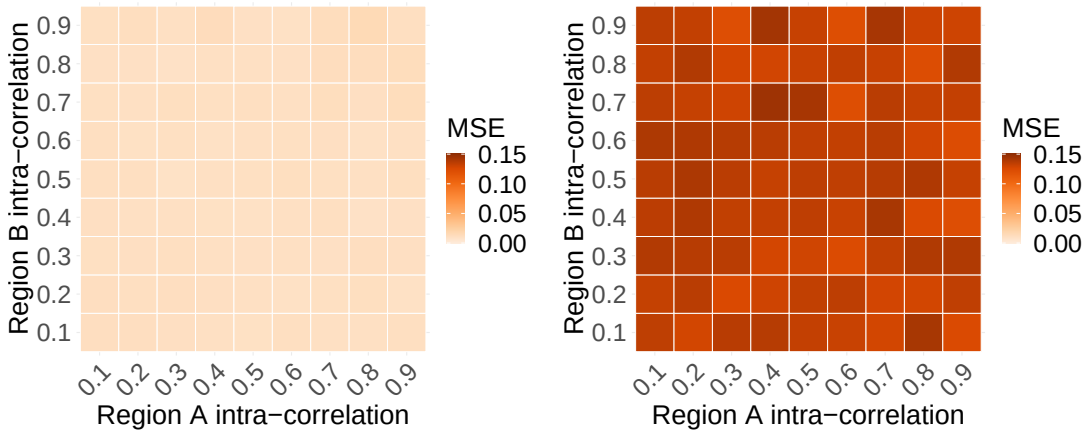
(c) $\eta^A_{min}, \eta^B_{min} = 0.2, \sigma^2_{\epsilon_A}, \sigma^2_{\epsilon_B} = 0.1$

(d) $\eta^A_{min}, \eta^B_{min} = 0.8, \sigma^2_{\epsilon_A}, \sigma^2_{\epsilon_B} = 0.1$

Fig. 6.5.1 Error as a function of the cut-off heights h_A, h_B for a pair of simulated regions for four simulation scenarios, with a true inter-correlation $\rho^{A,B} = 0.3$. The yellow diamond represents the error for cut-off heights equal to the maximum distance between U-scores within each region. The orange point corresponds to the minimal error.

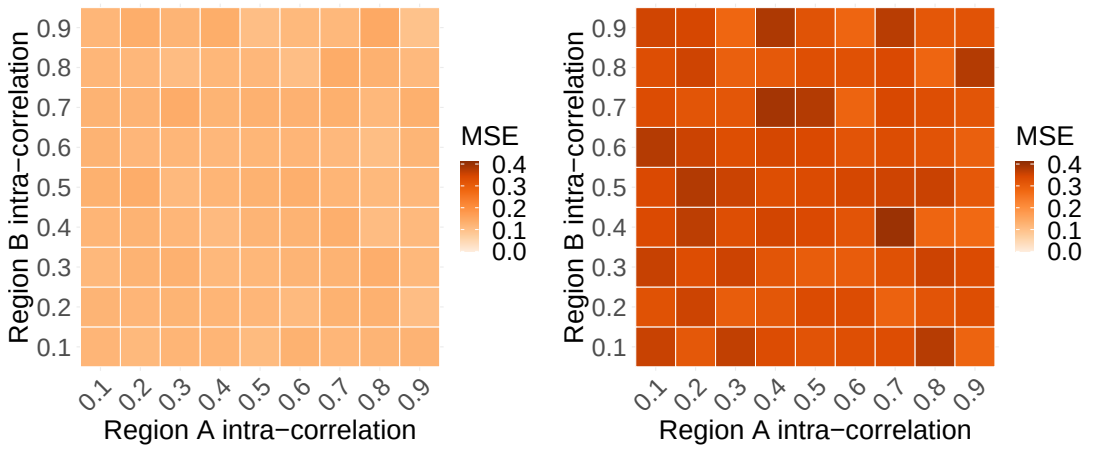
We first compare the performance of hierarchical clustering with Ward's linkage (our proposed choice and later denoted WardMaxU) with that of k-means (Hartigan and Wong, 1979) and ClustOfVar (CoV) (Chavent et al., 2012). ClustofVar is a hierarchical clustering method which is based on a principal component analysis approach, and closely related to works from Dhillon et al. (2003) and Vigneau et al. (2015). DBSCAN (Ester et al., 1996), which allows to directly control the cluster radii, was also considered. However, it fails to produce any clustering on the type of data we handle, which is high-dimensional. We also compare these clustering methods with a random assignment

of the voxels into clusters (Random). We choose the cut-off heights required by Ward's method according to the heuristic validated in the previous section (that is the maximum distance between U-scores). ClustOfVar, k-means and Random all require a choice of the number of clusters (and not of the cut-off heights). We hence choose the former as that obtained with our proposed method. We also evaluate ClustOfVar with the number of clusters chosen according to the maximum rand index (randCoV), which is the proposed criterion in (Chavent et al., 2012). Results are presented in Table 6.5.1. All methods with the same number of clusters are similar, with the exception of the random assignment. As expected, the latter displays MSEs an order of magnitude higher than that of the other clustering techniques, except when both minimal intra-correlations are high. Indeed, in such cases, the intra-correlation is high enough that the intra-cluster correlation will be high regardless of the choice of clusters. This demonstrates the importance of constructing clusters with high intra-cluster correlation to correctly estimate the inter-correlation. The method randCoV showcases the second highest MSE in all scenarios, except when both intra-correlation and noise are high, in which case its MSE is similar to that of the k-means and CoV. Moreover, the computation of the rand index requires a bootstrapping step and is thus very computationally expensive. Indeed, the average CPU time of clustering two regions using the method randCov is in the order of 10 min, while average CPU time is approximately 5 s when using CoV,



(a) Maximum distance between U-scores. (b) Maximum silhouette score

Fig. 6.5.2 MSE ($\times 10$), averaged over 50 replicates, for varying intra-correlation values for regions A and B. The true inter-correlation $\rho^{A,B}$ is 0.3 and the noise variance $\gamma_A^2 = \gamma_B^2 = 0.5$.



(a) Maximum distance between U-scores. (b) Maximum silhouette score.

Fig. 6.5.3 MSE (×100), averaged over 50 replicates, for varying intra-correlation values for regions A and B. The true inter-correlation $\rho^{A,B}$ is 0.1 and the noise variance $\gamma_A^2 = \gamma_B^2 = 0.5$.

300 ms using kmeans, and 30 ms using WardMaxU. Additionally, neither k-means nor CoV provide any obvious theoretical guarantees on the intra-correlation values within each cluster. Furthermore, they require to compute the U-scores, unlike our method. Indeed, our approach only depends on the distance between U-scores, which can be obtained directly from the sample voxel-to-voxel inter-correlation coefficients, without transforming the signals into U-scores. This step has a CPU time of about 15 s per region. These methods are thus much more computationally heavy. This confirms the choice of hierarchical clustering with Ward's linkage for our purposes, and will be used in all subsequent results.

We then compare our proposed estimator with the standard correlation of averages estimator $r_{A,B}^{CA}$, and the average of correlations $r_{A,B}^{AC}$ (Rosner et al., 1977). We also conduct comparisons with another inter-correlation estimator from the familial data literature, which is specifically designed for groups of dependent variables but fails to take into account noise (Elston, 1975). Its quality is similar to that of $r_{A,B}^{AC}$, and these results are hence included in the supplementary materials. Comparison with other correlation estimators from the literature would not be fair as they either only consider pairs of variables or do not handle arbitrary inter-correlation. To proceed we compute the regional-level point estimator $r_{A,B}^{CLA}$. We then calculate the MSE across 50 simulations. The results obtained for several simulation scenarios are recorded in Table 6.5.2. As expected from Theorem 6.4.2 and its corollary, our proposed estimator $r_{A,B}^{CLA}$ outperforms

Table 6.5.1 Mean ($\times 10^{-3}$) and standard deviation in parenthesis ($\times 10^{-3}$) of the squared errors over 50 replicates for different clustering methods and different simulation scenarios from the 1D Toeplitz model. The inter-correlation $\rho^{A,B}$ is set to 0.3.

Scenarios			Clustering Methods				
η^A_{min}	η^B_{min}	$\gamma^2_A = \gamma^2_B$	K-means	CoV	randCoV	Random	**WardMaxU**
0.2	0.2	0.5	2.0 (1.4)	2.0 (1.4)	4.8 (7.8)	15 (5.2)	2.0 (1.4)
0.8	0.8	0.5	1.2 (1.5)	1.2 (1.5)	1.1 (1.3)	1.0 (1.0)	1.2 (1.5)
0.2	0.8	0.5	1.1 (1.2)	1.1 (1.2)	2.9 (4.2)	5.0 (3.1)	1.1 (1.2)
0.2	0.2	0.1	1.0 (0.9)	1.0 (0.9)	4.6 (10)	26 (8.1)	1.0 (0.9)
0.8	0.8	0.1	0.6 (1.0)	0.6 (1.1)	1.0 (1.4)	1.4 (1.6)	0.6 (1.1)
0.2	0.8	0.1	0.4 (0.6)	0.4 (0.5)	2.7 (4.4)	10 (4.5)	0.4 (0.5)

the other estimators for all settings, except the low noise scenarios with 3D Spherical intra-correlation, where the MSE for $r^{AC}_{A,B}$ is slightly lower. Even in this case, the MSE for $r^{AC}_{A,B}$ and $r^{CLA}_{A,B}$ are in the same order of magnitude. More generally, we can note that in all scenarios where the intra-correlation is quite high and the noise variance is low, the MSE for these two estimators are also in the same order of magnitude. Indeed, according to equation (6.4), Theorem 6.4.1, and Corollary 6.4.1 $r^{AC}_{A,B}$ and $r^{CLA}_{A,B}$ would be very similar. Therefore, not only is the quality of the estimation greatly improved in the presence of noise and low intra-correlation, but it is also not deteriorated when intra-correlation is high and the noise is low. Furthermore, in practice, data are expected to be quite noisy with a low intra-correlation.

We can remark here that we did not include in Table 6.5.2 scenarios where the intra-correlation is close to zero. Indeed, in such cases no clusters of highly correlated variables can be found. In practical situations, this could be due to either high regional inhomogeneity or high noise, and could indicate an issue with the parcellation or data acquisition. Our clustering approach can hence help identify problematic datasets and thus provide information on the quality of the data.

6.5.4 Illustration on real-world data

We now apply our proposed estimator on real-world fMRI datasets, with the goal of estimating functional connectivity. At first, the sample cluster-level inter-correlation and voxel-level intra-correlation of different subjects can be visually inspected. The correlation estimates of three rats, including a dead one, are displayed in Figure 6.5.4, and that of three healthy human subjects (from the HCP dataset) are shown in Figure 6.5.6.

Table 6.5.2 Mean and standard deviation (in parenthesis) of the squared error over 50 replicates for different simulation scenarios and different estimators. The inter-correlation $\rho^{A,B}$ is set to 0.3.

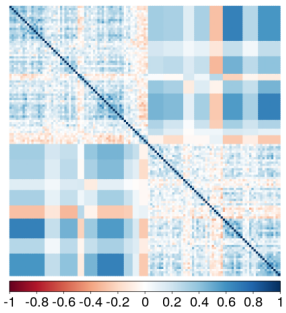
	Scenarios			Estimators		
	η^A_{min}	η^B_{min}	γ^2_A, γ^2_B	$r^{AC}_{A,B}$	$r^{CLA}_{A,B}$	$r^{CA}_{A,B}$
1D Toeplitz	0.2	0.2	0.5	1.8×10^{-2} (2.8×10^{-3})	$\mathbf{2.0 \times 10^{-3}}$ (1.4×10^{-3})	1.5×10^{-1} (1.8×10^{-1})
	0.8	0.8	0.5	1.2×10^{-2} (3.7×10^{-3})	$\mathbf{1.2 \times 10^{-3}}$ (1.5×10^{-3})	1.0×10^{-1} (1.0×10^{-1})
	0.2	0.8	0.5	1.4×10^{-2} (3.0×10^{-3})	$\mathbf{1.1 \times 10^{-3}}$ (1.2×10^{-3})	1.0×10^{-1} (1.0×10^{-1})
	0.2	0.2	0.1	5.4×10^{-3} (2.0×10^{-3})	$\mathbf{1.0 \times 10^{-3}}$ (9.1×10^{-4})	2.3×10^{-1} (2.7×10^{-1})
	0.8	0.8	0.1	1.9×10^{-3} (2.0×10^{-3})	$\mathbf{6.4 \times 10^{-4}}$ (1.0×10^{-3})	1.2×10^{-1} (1.2×10^{-1})
	0.2	0.8	0.1	2.7×10^{-3} (1.7×10^{-3})	$\mathbf{4.3 \times 10^{-4}}$ (5.5×10^{-4})	1.4×10^{-1} (1.6×10^{-1})
	$\phi_{A,A}$	$\phi_{B,B}$	γ^2_A, γ^2_B	$r^{AC}_{A,B}$	$r^{CLA}_{A,B}$	$r^{CA}_{A,B}$
3D Matérn	0.6	0.6	0.5	1.0×10^{-2} (3.8×10^{-3})	$\mathbf{7.0 \times 10^{-4}}$ (1.1×10^{-3})	1.6×10^{-3} (1.9×10^{-3})
	0.8	0.8	0.5	1.0×10^{-2} (4.0×10^{-3})	$\mathbf{7.9 \times 10^{-4}}$ (1.2×10^{-3})	1.0×10^{-3} (1.4×10^{-3})
	0.6	0.8	0.5	1.0×10^{-2} (3.9×10^{-3})	$\mathbf{7.2 \times 10^{-4}}$ (1.1×10^{-3})	1.0×10^{-3} (1.6×10^{-3})
	0.6	0.6	0.1	1.3×10^{-3} (1.5×10^{-3})	$\mathbf{7.7 \times 10^{-4}}$ (1.0×10^{-3})	1.7×10^{-3} (2.0×10^{-3})
	0.8	0.8	0.1	1.4×10^{-3} (1.6×10^{-3})	$\mathbf{7.5 \times 10^{-4}}$ (1.0×10^{-3})	1.1×10^{-3} (1.4×10^{-3})
	0.6	0.8	0.1	1.3×10^{-3} (1.6×10^{-3})	$\mathbf{7.7 \times 10^{-4}}$ (1.0×10^{-3})	1.3×10^{-3} (1.7×10^{-3})
	$\phi_{A,A}$	$\phi_{B,B}$	γ^2_A, γ^2_B	$r^{AC}_{A,B}$	$r^{CLA}_{A,B}$	$r^{CA}_{A,B}$
3D Spherical	8	8	0.5	1.0×10^{-2} (2.3×10^{-3})	$\mathbf{4.6 \times 10^{-3}}$ (2.4×10^{-3})	8.8×10^{-2} (1.4×10^{-2})
	12	12	0.5	1.0×10^{-2} (2.8×10^{-3})	$\mathbf{2.4 \times 10^{-3}}$ (1.9×10^{-3})	2.5×10^{-2} (8.2×10^{-3})
	8	12	0.5	9.4×10^{-3} (2.5×10^{-3})	$\mathbf{4.2 \times 10^{-3}}$ (2.3×10^{-3})	5.3×10^{-2} (1.1×10^{-2})
	8	8	0.1	$\mathbf{9.1 \times 10^{-4}}$ (7.9×10^{-4})	8.9×10^{-3} (3.8×10^{-3})	9.3×10^{-2} (1.3×10^{-2})
	12	12	0.1	$\mathbf{1.0 \times 10^{-3}}$ (1.0×10^{-3})	4.5×10^{-3} (2.8×10^{-3})	2.6×10^{-2} (8.4×10^{-3})
	8	12	0.1	$\mathbf{7.3 \times 10^{-4}}$ (7.8×10^{-4})	7.7×10^{-3} (3.3×10^{-3})	5.6×10^{-2} (1.1×10^{-2})

In brain functional connectivity studies, point estimates for each pair of regions are needed to construct a correlation matrix. A thresholding step is then applied to obtain a binary connectivity network where only the edges corresponding to the highest correlation values remain. In this section, we will therefore mostly focus on evaluating the regional-level entries of these correlation matrices.

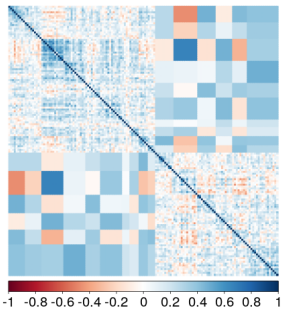
Rat data

We first examine rat data results.

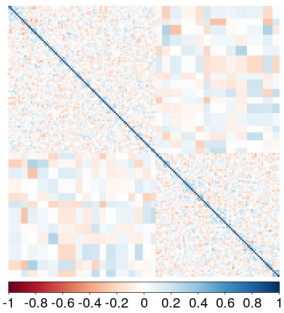
Dead rats. No functional activity should be detected in dead rats, unlike in live rats. Dead rats hence provide experimental data where the ground-truth inter-correlation is zero. We can therefore compute the MSE across all pairs of regions (each region pair is a replicate). We expect as well that the intra-correlation is zero within all regions. In fact, no discernible structure of the dead rat's intra-correlation can be noted in Figure 6.5.4, where motor (M1_l, M1_r) and sensory (S1_l, S1_r) regions are represented. We find



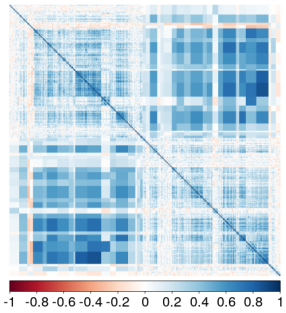
(a) M1_l, M1_r-rat 24 (IsoW)



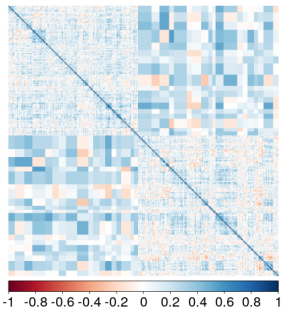
(b) M1_l, M1_r–rat 31 (EtoL)



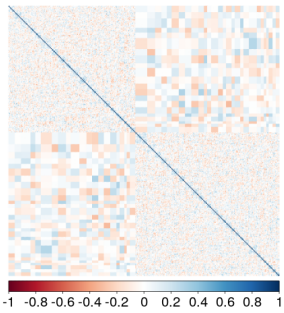
(c) M1_l, M1_r–rat 9 (dead)



(d) S1_l, S1_r–rat 24 (IsoW)



(e) S1_l, S1_r–rat 31 (EtoL)



(f) S1_l, S1_r–rat 9 (dead)

Fig. 6.5.4 Sample pairwise correlation matrices for different rats and brain region pairs. Voxels are ordered by clusters. The diagonal blocks correspond to the voxel-to-voxel sample intra-correlation $r_{i,i'}^{A,A}$, while the off-diagonal blocks correspond to the sample inter-correlation between clusters r_{ν_A,ν_B}^{CLA}.

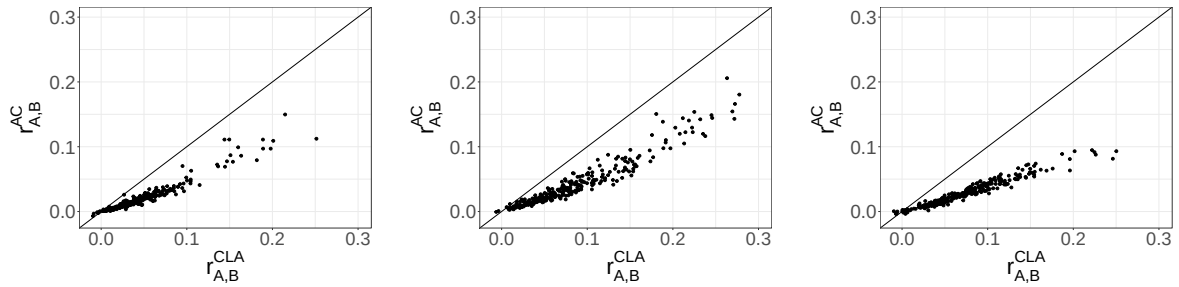
the MSE of $r_{A,B}^{CLA}$ is slightly higher than that of $r_{A,B}^{AC}$ (cf. Table 6.5.3). Nonetheless, they are both very low and several orders of magnitude lower than the MSE of $r_{A,B}^{CA}$. This indicates that for dead rat data, $r_{A,B}^{CLA}$ displays similar quality to $r_{A,B}^{AC}$, and a considerable improvement over the standard $r_{A,B}^{CA}$.

Table 6.5.3 MSE across all pairs of regions for different dead rats and different estimators.

Dead Rat ID	$r_{A,B}^{AC}$	$r_{A,B}^{CLA}$	$r_{A,B}^{CA}$
16	5.2×10^{-6}	5.6×10^{-5}	1.3×10^{-2}
18	4.7×10^{-6}	5.4×10^{-5}	1.3×10^{-2}
9	5.7×10^{-6}	6.0×10^{-5}	1.3×10^{-2}

Live rats. To further illustrate the advantages of our proposed approach, we consider three live rats under different anesthetics. Unlike for dead rats, no ground-truth inter-

correlation is available. We thus inspect directly the values of the estimated inter-correlations. We can first remark correlation values are visually very different in live and dead rats. Indeed, both intra- and inter-correlations are higher, in addition to displaying an apparent structure (cf. Figure 6.5.4). While we could not clearly demarcate $r^{AC}_{A,B}$ from $r^{CLA}_{A,B}$ using solely the dead rat data, we can note in Figure 6.5.5 that for any pair of regions, $r^{CLA}_{A,B}$ is both larger than $r^{AC}_{A,B}$ and further away from zero, which corresponds to dead rat connectivity. In the context of functional connectivity, this implies that, when applying a thresholding step, $r^{CLA}_{A,B}$ may allow us to increase the number of rightfully detected edges in the corresponding connectivity network.



(a) Rat 24 (IsoW) (b) Rat 4 (UreL) (c) Rat 31 (EtoL)

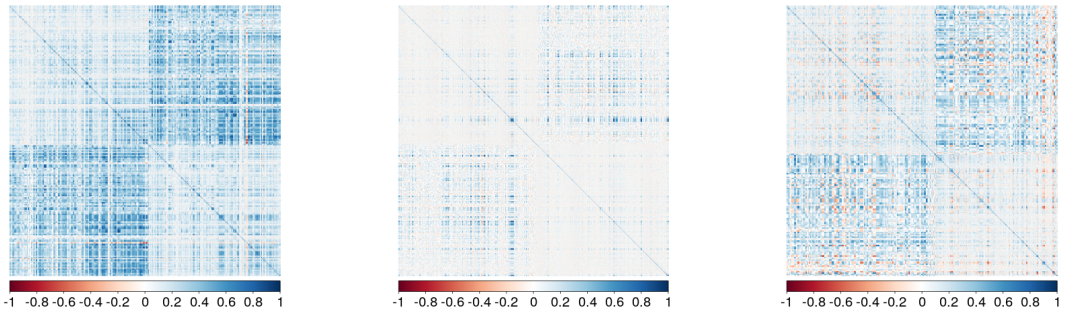
Fig. 6.5.5 Sample inter-correlation coefficients estimated using $r^{AC}_{A,B}$ against our proposed estimator $r^{CLA}_{A,B}$ for three live rats under different anesthetics. Each point represents a pair of brain regions.

HCP data

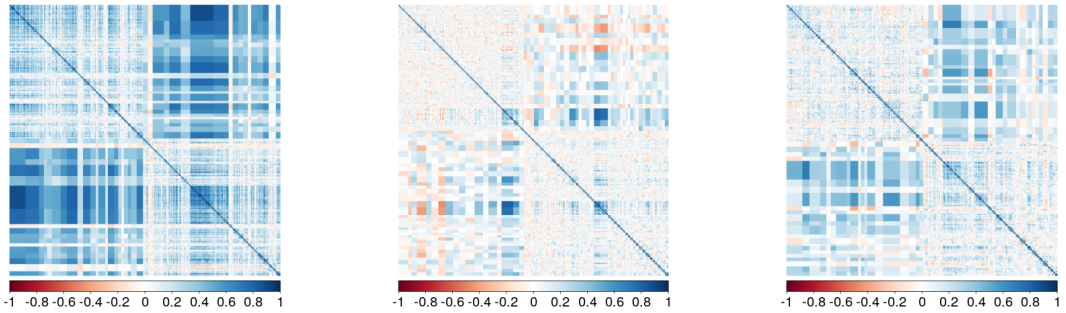
We then illustrate our proposed approach on human data from healthy live subjects. No ground-truth is available.

Figure 6.5.6 showcases sample correlations of the Precentral regions (Pr_l, Pr_r), which are large regions containing about 1700 voxels, and Heschl's gyri (H_l, H_r), which are ten times smaller. We can first note that the intra-correlation displays some structure, as in the live rats. Nonetheless, overall, subject 2 seems to have both lower sample intra- and inter-correlation values, compared to most other subjects (including subjects 1 and 3). Subject 2 has in fact a benign anatomical brain anomaly. Our proposed approach hence allowed us to single out an unusual subject just by visually inspecting its sample intra- and inter-correlation values.

We can then compare the sample distribution of our proposed estimator $r^{CLA}_{A,B}$ with that of the standard estimator $r^{CA}_{A,B}$ (cf. Figure 6.5.7) and of $r^{AC}_{A,B}$ (cf. Figure 6.5.8). Overall, and as expected from equations (6.4) and (6.7) and Corollary 6.4.1, the correlation of



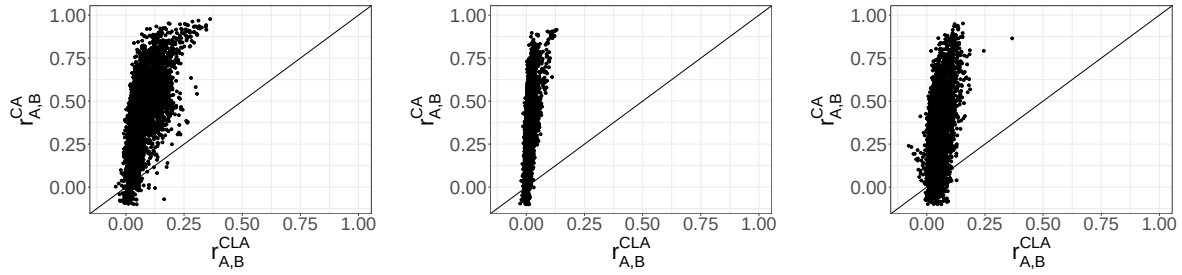
(a) Pr_l, Pr_r – subject 1 (b) Pr_l, Pr_r – subject 2 (c) Pr_l, Pr_r – subject 3



(d) H_l, H_r – subject 1 (e) H_l, H_r – subject 2 (f) H_l, H_r – subject 3

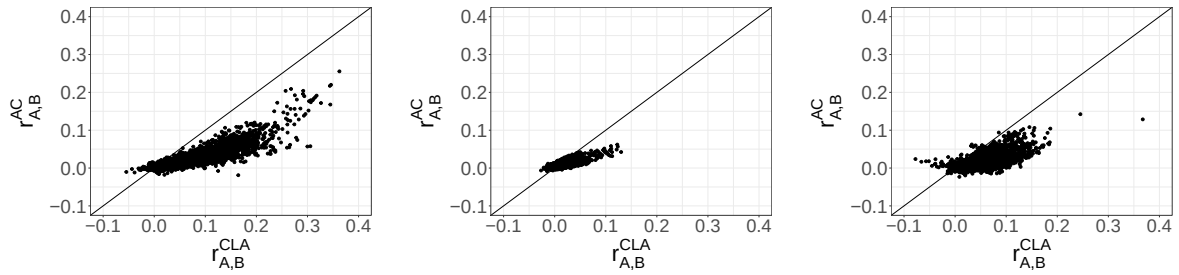
Fig. 6.5.6 Sample pairwise correlation matrices for different HCP subjects and brain region pairs. Voxels are ordered by clusters. The diagonal blocks correspond to the voxel-to-voxel sample intra-correlation $r_{i,i'}^{A,A}$, while the off-diagonal blocks correspond to the sample inter-correlation between clusters r_{ν_A,ν_B}^{CLA}.

averages $r_{A,B}^{CA}$ values are higher than that of $r_{A,B}^{CLA}$, while the sample values of the average of correlations estimator $r_{A,B}^{AC}$ are lower. In terms of functional connectivity, this means using the $r_{A,B}^{CA}$ estimator could lead to falsely detecting edges, while using $r_{A,B}^{AC}$ could lead to missing edges. These results are in accordance with what was observed in the rat data.



(a) Subject 1 – session 0 (b) Subject 2 – session 0 (c) Subject 3 – session 0

Fig. 6.5.7 Inter-correlation coefficients estimated using $r_{A,B}^{CA}$ against our proposed estimator $r_{A,B}^{CLA}$ for three HCP subjects. Each point represents a pair of brain regions.



(a) Subject 1 – session 0 (b) Subject 2 – session 0 (c) Subject 3 – session 0

Fig. 6.5.8 Inter-correlation coefficients estimated using $r^{AC}_{A,B}$ against our proposed estimator $r^{CLA}_{A,B}$ for three HCP subjects. Each point represents a pair of brain regions.

Since we have access to two separate sessions for each subject, we then evaluate the reproducibility of our estimator. To do so, for each subject, we calculate the Concordance Correlation Coefficient (CCC) (Lin, 1989) between the inter-correlations estimates from their two sessions. The CCC is scaled between -1 and 1, with 1 corresponding to a perfect concordance. This means that the higher the CCC, the more reproducible the estimator. The estimator $r^{CLA}_{A,B}$ exhibits the highest CCC, with an average (variance) across the 35 subjects of 0.69 (0.03), while that of $r^{CA}_{A,B}$ is 0.63 (0.02) and $r^{AC}_{A,B}$ is 0.67 (0.04). Our proposed estimator hence improves reproducibility over existing estimators.

6.6 Conclusion

In this chapter, we proposed a novel and non-parametric estimator of the correlation between groups of arbitrarily dependent variables in the presence of noise. We devised a clustering-based approach that simultaneously reduces the impact of noise and intra-correlation through judicious aggregation. We analyzed the convergence of our proposed estimator, and provided a heuristic selection of cut-off heights of the dendrograms. Moreover, our method yields both point estimates and a corresponding empirical distribution that could be used, for instance, for uncertainty quantification. We conducted experiments on synthetic data that showed our proposed estimator surpasses popular existing methods in terms of quality, and demonstrated the effectiveness and reproducibility of our approach on real-world datasets.

Appendix

6.A Proof of Theorem 6.4.1

The proof follows from the properties of hierarchical clustering. In the context of Ward's linkage, the distance between two clusters ν_1 and ν_2 is defined according to Kaufman and Rousseeuw (2005, p. 230) as:

$$D(\nu_1, \nu_2) = \sqrt{\frac{2 \cdot |\nu_1| \cdot |\nu_2|}{|\nu_1| + |\nu_2|} \cdot \left\| \overline{\mathbf{U}}^{\nu_1} - \overline{\mathbf{U}}^{\nu_2} \right\|^2}, \tag{6.18}$$

where $\overline{\mathbf{U}}^{\nu_1}$ is the centroid and $|\nu_1|$ the cardinality of cluster ν_1. Consider a region A and fix a cut-off height h_A. Then, from properties of agglomerative clustering, for any cluster ν_A, and for all pairs of U-scores $\mathbf{U}_i^A, \mathbf{U}_{i'}^A$ inside ν_A, $D(\{\mathbf{U}_i^A\}, \{\mathbf{U}_{i'}^A\}) \leq h_A$. Therefore, by combining this inequality with properties of the U-scores (Hero and Rajaratnam, 2011), the sample intra-correlation can be lower-bounded by a function of h_A:

$$1 - \frac{h_A^2}{2} \;\leq\; 1 - \frac{\|\mathbf{U}_i^A - \mathbf{U}_{i'}^A\|^2}{2} \;= r_{i,i'}^{A,A}, \tag{6.19}$$

which implies the left-hand side of (6.8). The right-hand side follows from properties of correlation coefficients. This concludes the proof.

6.B Proof of Theorem 6.4.2

For two clusters ν_A, ν_B in regions A, B, from (6.11),

$$r_{\nu_A,\nu_B}^{CLA} = \frac{\widehat{Cov}(\overline{\mathbf{Y}}^{\nu_A}, \overline{\mathbf{Y}}^{\nu_B})}{\sqrt{\widehat{Var}(\overline{\mathbf{Y}}^{\nu_A}) \cdot \widehat{Var}(\overline{\mathbf{Y}}^{\nu_B})}}. \tag{6.20}$$

Since we have assumed variables are temporally i.i.d., and according to the model definition (cf. Section 4.2), as n tends towards infinity,

$$\widehat{Cov}(\overline{\mathbf{Y}}^{\nu_A}, \overline{\mathbf{Y}}^{\nu_B}) \overset{a.s.}{\to} Cov(\overline{Y}^{\nu_A}(t), \overline{Y}^{\nu_B}(t)), \tag{6.21}$$

for any time point t and where

$$\begin{aligned} Cov(\overline{Y}^{\nu_A}(t), \overline{Y}^{\nu_B}(t)) &= \frac{1}{|\nu_A| \cdot |\nu_B|} \sum_{i=1}^{|\nu_A|} \sum_{j=1}^{|\nu_B|} Cov(Y_i^A(t), Y_j^B(t)) \\ &= \frac{1}{|\nu_A| \cdot |\nu_B|} \sum_{i=1}^{|\nu_A|} \sum_{j=1}^{|\nu_B|} \sigma_A \sigma_B \rho^{A,B} \\ &= \sigma_A \sigma_B \rho^{A,B}, \end{aligned} \tag{6.22}$$

and, from equation (6.1),

$$\widehat{Var}(\overline{\mathbf{Y}}^{\nu_A}) \overset{a.s.}{\to} Var(\overline{Y}^{\nu_A}(t)) \;=\; \sigma_A^2 \cdot \frac{1}{|\nu_A|^2} \cdot \sum_{i,i'=1}^{|\nu_A|} \eta_{i,i'}^A + \frac{\gamma_A^2}{|\nu_A|}, \tag{6.23}$$

which gives (6.14), and concludes the proof.

6.C Relaxing assumptions about the noise

In the preliminaries section of the main paper we assumed noise variables $\boldsymbol{\epsilon}_i^A$ were spatially uncorrelated both within and across regions. However, in practice these assumptions are often violated. Nevertheless, they can easily be relaxed to the presence of spatial correlation within regions and uncorrelatedness between distinct regions. We can then extend the consistency results presented at the end of Section 4 of the proposed estimator r_{ν_A,ν_B}^{CLA}accordingly.

First, ρ_{ν_A,ν_B}^{CLA} is now equal to:

$$\frac{\rho^{A,B}}{\sqrt{\left[\frac{1}{|\nu_A|^2} \sum\limits_{i,i'=1}^{|\nu_A|} \eta_{i,i'}^A + \frac{\gamma_A^2}{\sigma_A^2 \cdot |\nu_A|^2} \sum\limits_{i,i'=1}^{|\nu_A|} Cor(\boldsymbol{\epsilon}_i^A, \boldsymbol{\epsilon}_{i'}^A)\right] \cdot \left[\frac{1}{|\nu_B|^2} \sum\limits_{i,i'=1}^{|\nu_B|} \eta_{i,i'}^B + \frac{\gamma_B^2}{\sigma_B^2 \cdot |\nu_B|^2} \sum\limits_{i,i'=1}^{|\nu_B|} Cor(\boldsymbol{\epsilon}_i^B, \boldsymbol{\epsilon}_{i'}^B)\right]}}.$$

It follows immediately that r_{ν_A,ν_B}^{CLA} is a consistent estimator of ρ_{ν_A,ν_B}^{CLA}. Interpretation about the impact of the cluster size on lessening the effects of the noise is however slightly

less straightforward. We can note that for a given cluster ν_A of region A,

$$\sum_{i,i'=1}^{|\nu_A|} Cor(\boldsymbol{\epsilon}_i^A, \boldsymbol{\epsilon}_{i'}^A) = \frac{1}{|\nu_A|} + \frac{1}{|\nu_A|^2} \sum_{i \neq i'=1}^{\nu_A} Cor(\boldsymbol{\epsilon}_i^A, \boldsymbol{\epsilon}_{i'}^A).$$

Therefore, as long as the correlation between noise variables within a cluster remain quite small, interpretation of the effect of the cut-off height on the noise remains similar to the one described in the main paper—that is, the larger, the cut-off height, the larger the clusters, and the smaller the impact of the noise. All results presented in the main paper then remain valid.

6.D Comparison with an additional estimator

In this section we compare our proposed estimator $r_{A,B}^{CLA}$ with the one used by Elston (1975) in the setting of groups of dependent variables.

Table 6.D.1 Mean and standard deviation (in parenthesis) of the squared error over 50 replicates for different simulation scenarios and different estimators. The inter-correlation $\rho^{A,B}$ is set to 0.3.

Scenarios				Estimators		
	η_A^-	η_B^-	$\sigma^2_{\epsilon^A}, \sigma^2_{\epsilon^B}$	$r_{A,B}^{AC}$	$r_{A,B}^{CLA}$	Elston (1975)
1D Toeplitz	0.2	0.2	0.5	1.8×10^{-2} (2.8×10^{-3})	$\mathbf{2.0 \times 10^{-3}}$ (1.4×10^{-3})	1.8×10^{-2} (2.8×10^{-3})
	0.8	0.8	0.5	1.2×10^{-2} (3.7×10^{-3})	$\mathbf{1.2 \times 10^{-3}}$ (1.5×10^{-3})	1.2×10^{-2} (3.8×10^{-3})
	0.2	0.8	0.5	1.4×10^{-2} (3.0×10^{-3})	$\mathbf{1.1 \times 10^{-3}}$ (1.2×10^{-3})	3.1×10^{-3} (3.1×10^{-3})
	0.2	0.2	0.1	5.4×10^{-3} (2.0×10^{-3})	$\mathbf{1.0 \times 10^{-3}}$ (9.1×10^{-4})	5.5×10^{-3} (2.0×10^{-3})
	0.8	0.8	0.1	1.9×10^{-3} (2.0×10^{-3})	$\mathbf{6.4 \times 10^{-4}}$ (1.0×10^{-3})	1.9×10^{-3} (2.0×10^{-3})
	0.2	0.8	0.1	2.7×10^{-3} (1.7×10^{-3})	$\mathbf{4.3 \times 10^{-4}}$ (5.5×10^{-4})	2.7×10^{-3} (1.7×10^{-3})

Suppose we have two fixed clusters ν_A and ν_B within regions A and B, respectively. We place ourselves in the context of model 1 defined in Section 3 of the main paper. Let Σ_{ν_A} be the intra-cluster covariance matrix of the variables within cluster ν_A, and Σ_{ν_A,ν_B} be the inter-cluster covariance matrix between clusters ν_A and ν_B. It follows from Section 3 that the inter-cluster covariance elements are constant across all pairs of variables within theses two clusters. We also assume the intra-cluster variance and covariance are constant. We now suppose we have n independent samples of the vector $\left[Y_1^A, \ldots, Y_{|\nu_A|}^A, Y_1^B, \ldots, Y_{|\nu_B|}^B\right]$ from a multivariate normal distribution with mean μ and

covariance matrix with the following structure:

$$\Sigma = \begin{pmatrix} \Sigma_{\nu_A} & \Sigma_{\nu_A,\nu_B} \\ \Sigma_{\nu_A,\nu_B} & \Sigma_{\nu_B} \end{pmatrix} \tag{6.24}$$

In the context of fMRI data, this would correspond to the BOLD signal of the voxels within clusters ν_A and ν_B.

Let us consider the following estimators of the elements in Σ. The intra-cluster variance can be estimated by $v_{\nu_A} = \frac{1}{|\nu_A|} \sum_{i=1}^{|\nu_A|} \widehat{Var}(\mathbf{Y}_i^A)$, and the inter-cluster covariance by $v_{\nu_A,\nu_B} = \frac{1}{|\nu_A|\cdot|\nu_B|} \sum_{i=1}^{\nu_A} \sum_{j=1}^{\nu_B} \widehat{Cov}(\mathbf{Y}_i^A, \mathbf{Y}_j^B)$. Then, according to Elston (1975), $v_{\nu_A,\nu_B}/\sqrt{v_{\nu_A} v_{\nu_B}}$ is the maximum likelihood estimator of $\rho^{A,B}$. This estimator is closely related to the average of correlations estimator $r_{A,B}^{AC}$ (Rosner et al., 1977). However, instead of spatially averaging the correlations, Elston averages the covariance and variance terms separately. While this estimator is specifically designed for groups of dependent variables, it is impacted by noise in the same manner as $r_{A,B}^{AC}$. It turns out the MSE of Elston's estimator is also very similar to that of $r_{A,B}^{AC}$, and consequently higher than that of our estimator in most scenarios (cf. Table 6.D.1).

6.E Details about the implementation and code availability

Our implementation is based on R 4.2. All experiments were performed on a laptop running on Ubuntu 18.04 with eight 1.8GHz 64-bits Intel Core i7-10610U CPUs, 32 GB of memory and a 1 TB hard drive.

Source code, including a notebook detailing how to reproduce the figures of this chapter, is available at: https://gitlab.inria.fr/q-func/clustcorr.

Chapter 7

Distribution-Based Weighted Networks Validation on rs-fMRI Data

In this chapter we build upon the inter-correlation estimator presented in Chapter 6 to introduce distribution-based weighted networks, hence fully leveraging inter-correlation distributions. This chapter aims to validate their use in a practical scenario. We namely demonstrate our proposed framework improves repeatability, regression and classification performances compared to that of the standard correlation of averages approach.

This chapter presents results from an on-going project:

Lbath, H., Richiardi, J., Petersen, A., Meiring, W., and Achard, S. (working paper). Distribution-based weighted networks validation on rs-fMRI data

7.1 Introduction

Resting-state functional Magnetic Resonance Imaging (rs-fMRI) is widely used in studies aiming to predict personality traits (Girn et al., 2023), cognitive functions (van den Heuvel et al., 2009; Cui and Gong, 2018), or either diagnose or improve understanding of neurological or psychiatric disorders, such as Alzheimer's or schizophrenia (Dadi et al., 2019; Sarica et al., 2017). In these contexts, resting-state functional connectivity (RSFC) networks, constructed from rs-fMRI data, are prevalent. As presented in Chapter 2, RSFC networks are most commonly built as follows: nodes represent brain regions, and edge weights are equal to the correlation between region-averaged signals. We previously

referred to this inter-correlation estimator as the correlation of averages (CA). However, scalar-weighted networks only offer limited opportunities to account for uncertainties when inferring networks.

We hence introduce distribution-based weighted networks, where edges are associated to a distribution or density function, instead of a scalar value. To build these new types of networks we leverage the clustering-based inter-correlation estimator introduced in Chapter 6 (Lbath et al., 2023). Indeed, the latter provides an empirical inter-correlation distribution in addition to a point estimate.

Our goal in this chapter is to evaluate the use of distribution-based weighted networks in terms of both repeatability and performance in machine learning pipelines. To this end, we explored classification and regression tasks on psychometric scores. Indeed, classification performance has been previously used as a validation criterion in selection of connectivity network construction models (Zanin et al., 2012). Moreover, studying associations between RSFC and variables of interests, such as personality traits or medical diagnosis, is in itself highly valuable. Yet, classification of RSFC data is challenging, and performances are often moderate at best (Dadi et al., 2019; Cui and Gong, 2018; Sarica et al., 2017). Furthermore, reliably identifying univariate associations in these contexts is tremendously difficult (Marek et al., 2022). Exploring multivariate associations is similarly challenging (Marek et al., 2022; Girn et al., 2023). Improving performances on these tasks is then a practical and effective way to demonstrate the strengths of our proposed framework.

It has also been suggested that small populations and high subject heterogeneity may hinder generalization performances of machine learning models (Vabalas et al., 2019; Schnack and Kahn, 2016; Marek et al., 2022; Cui and Gong, 2018). We hence highlight the effect of heterogeneity on univariate effect sizes, and show our proposed distribution-based approach decreases subject variance in terms of edge weights.

We can summarize the contributions presented in this chapter as follows. We first introduce a novel distribution-based weighted network paradigm. We then proceed to validate its effectiveness on two real-world human datasets. To that end, we evaluate its test-retest reliability in terms of both RSFC network edge weights and graph metric values, as well as its performance in various machine learning tasks. These include classification and regression tasks on cognitive and personality scores at both brain- and edge-level, and the latter encompassing different types of graph embeddings.

7.2 Materials and Methods

7.2.1 Data

We present in this section the datasets used in this chapter. Additional dataset information and preprocessing details are documented in Chapter 3.

Human Connectome Project (HCP). The work presented in this chapter uses the test-retest fMRI data from 100 healthy subjects of the HCP dataset, WU-Minn Consortium preprocessed (Glasser et al., 2013). Two resting-state fMRI acquisitions on different days are available for each subject, and are later denoted HCP 0 and HCP 1. A modified AAL template is used to parcellate the brain into 89 regions (Termenon et al., 2016). We also use subject-specific psychometric variables, such as NEO-FFI Big 5 personality questionnaire scores and fluid intelligence scores from the Penn matrix test, which is based on an abbreviated version of Raven's matrices.

PIOP2. Some of the analyses presented in this work are repeated on the PIOP2 dataset from the Amsterdam Open MRI Collection (AOMIC) (Snoek et al., 2021). The latter includes several subject variables that can be used for regression or classification tasks, including psychometric variables such as NEO-FFI and Raven's matrices scores. We exclude subjects with missing fMRI data, missing Raven's matrices scores, or with excessive head motion. The motion exclusion criteria pipeline used by Thiele et al. (2022) is applied. The final sample consists of 204 healthy subjects.

The brain is parcellated into 89 regions with the modified AAL template used to parcellated the HCP data. Regions with less than 10 voxels with non-zero BOLD signal in at least one subject are discarded. The same 76 out of 89 regions from the initial parcellation are retained in all subjects.

7.2.2 Distribution-based weighted connectivity networks

We introduced in Chapter 6 the clustering-based estimator (CLA) (Lbath et al., 2023), which, in addition to being more consistent than the standard CA estimator, also provides an empirical inter-correlation distribution by design. In this chapter we build on that to introduce distribution-based weighted connectivity networks. To this end, we construct fully connected subject-specific functional connectivity networks where nodes correspond to brain regions and edge weights are inter-regional correlation density functions (cf. Figure 7.2.1).

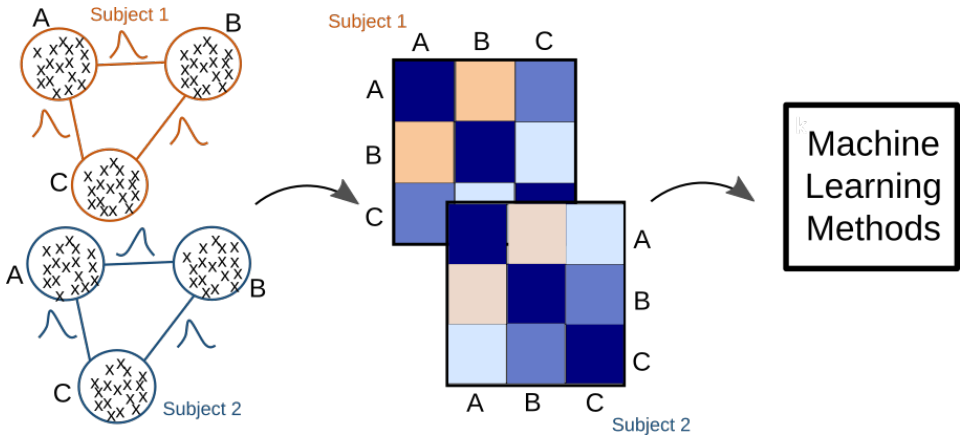


Fig. 7.2.1 Simplified machine learning pipeline for distribution-based networks of two subjects. Networks are weighed by inter-correlation density functions. Quantile values are then extracted to obtain the corresponding connectivity matrices to be input into machine learning methods.

For each subject, and for each pair of regions A, B, the clustering-based correlation estimator is used to obtain the empirical inter-correlation distribution between regions A and B. The corresponding inter-correlation densities and quantile functions are then obtained using default kernel density estimates from the frechet (version 0.2.0) and fdadensity (version 0.1.2) R (version 4.2.0) packages.

7.2.3 Network representation

We now present how our proposed distribution-based networks can be leveraged in practice. In order to incorporate them into standard network evaluation and machine learning pipelines, the edge weights need to be summarized so as to be equal to a single scalar value (cf. Figure 4.1). A straightforward approach is to only keep the correlation value corresponding to a given quantile value, later denoted by Qxx where xx is the corresponding percentile (e.g., Q10 for 10th percentile or 0.1 quantile).

We first utilize *brain-level* summaries where scalar edge weights are averaged across all edges of a given subjects. These were used in a recent fMRI reproducibility study (Marek et al., 2022).

We additionally perform *edge-level* analyses. These require inputs to lay in a vector space. In our setting, once edges are summarized by a scalar, quantile-based, value, various graph embedding techniques can be considered. The most basic approach is a direct embedding where all edge weights are stacked in a vector. We also employ more sophisticated approaches, such as the random-walk-based embedding Graph2vec

(Narayanan et al., 2017) or the Feather embedding, which uses characteristic functions of node features (Rozemberczki and Sarkar, 2020). Note that embeddings must be applied on training and testing data separately.

7.2.4 Validation methods

We describe in this section the different methods we use to evaluate the practicality of employing distribution-based networks in brain functional connectivity studies. Due to the nature of these evaluation methods, this assessment is focused more specifically on quantile-based connectivity estimators. We first assess test-retest reliability, before evaluating classification and regression performances. Results are compared to that of the popular CA estimator.

Test-retest reliability

We use the Concordance Correlation Coefficient (CCC) Lin (1989) to measure the repeatability of the inter-correlation estimators. For each subject, fMRI scans from two separate examinations, which are available in the HCP dataset but not PIOP2, are required to compute the CCC. For each subject, the latter is defined as follows for two sets of measurements x, y, corresponding to the two scans:

$$\mathrm{CCC} = \frac{2s_{xy}}{s_x^2 + s_y^2 + \left(\frac{1}{K}\sum_{j=1}^{K} x_j - \frac{1}{K}\sum_{j=1}^{K} y_j\right)^2}, \tag{7.1}$$

where s_{xy} and s_x are the empirical covariance and variance across all edges, respectively, and K is the number of observations, i.e., the number of edges in our case. CCC values are comprised between -1 and 1, with higher values indicating higher repeatability. For each subject, the CCC of both edge weights and a classical graph centrality metric, the betweenness, is computed for various clustering-based inter-correlation quantile values and the CA estimator. The CCC method from the R (version 4.3.0) DescTools (version 0.99.48) package is used.

Betweenness quantifies for any given node the extent to which it lies on the shortest path between other nodes. We compute the betweenness for each node of each of the 100 HCP subjects for a range of network sparsity threshold values. The latter are defined as a percentage of the total number of edges, keeping only edges with the highest correlation. Edges are thus binarized. In order to calculate the betweenness, which is not well-defined for disconnected graphs, we then force the network to be connected by

applying a minimum spanning tree (Alexander-Bloch et al., 2010). The R igraph (version 1.4.2) package is used.

Univariate brain-wide association

In order to preliminarily assess effect size, univariate associations are obtained. Pearson's r correlation between individual psychometric scores and brain-level RSFC is hence computed for various clustering-based inter-correlation quantiles and the CA estimator, similarly to univariate brain-wide association computation in (Marek et al., 2022). A linear least square fit is also obtained using the Python (version 3.9.16) numpy (version 1.23.5) polyfit method. In addition, for each edge, Pearson's r between the edge weights and individual psychometric scores are computed for the different inter-correlation estimators. Brain- and edge-level brain-wide association results are obtained for both the HCP and PIOP2 datasets.

Classification

We use RSFC networks obtained from the HCP and PIOIP2 dataset to classify individuals into high and low fluid intelligence categories. To that end, fluid intelligence scores, corresponding to the Penn matrix test scores (PMAT24_A_CR) and Raven's scores in the HCP and PIOP2 datasets, respectively, are binarized in the same manner as in (Dadi et al., 2019): subjects are split into thirds according to score quantiles 0.33 and 0.66, and individuals in the middle class are excluded in order to facilitate the binary classification task. In order to fully leverage our connectivity data, we also examine multivariate associations. Classification is hence performed both at the brain- and edge-level for varying clustering-based inter-correlation quantile values and the CA inter-correlation estimator. We use Support Vector Classification (SVC) (Platt, 1999), which is among recommended classifiers for rs-fMRI functional connectivity studies (Dadi et al., 2019), with default rbf kernel from the Python scikit-learn library (version 1.2.2). A 10-fold cross-validation pipeline is used to measure accuracy and the Area Under the Receiver Operating Characteristics Curve (AUC). Evaluation scores are averaged across the 10 cross-validation folds. Accuracy is equal to the proportion of correct classifications, while the AUC measures how well a classifier separates classes. In balanced cases, such as our own, uninformative classifiers yield 0.5 accuracy and AUC, while both metrics are equal to 1 for perfect classifiers. When sample size is small, Combrisson and Jerbi (2015) showed that chance level accuracy is higher. In fact, for the HCP data, with 100 subjects, accuracy would need to be at least 0.58 for the classification to be deemed significant

with a p-value < 0.05. For the PIOP2 data, with 204 subjects, accuracy larger than 0.56 would be required.

Regression

We then use Random Forest regression to predict NEO-FFI personality scores from brain- or edge-level RSFC. Random Forest has been previously shown to perform adequately for small effect sizes in neuroimaging studies (Jollans et al., 2019). A Random Forest regressor with 501 trees and a squared error criterion from the scikit-learn Python library is employed. Similarly to the classification task, 10-fold cross-validation is used. Pearson's r correlation between actual and predicted values, and the Root Mean Squared Error (RMSE) are computed to evaluate models. RMSE is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(c_m - \hat{c}_m)^2}, \tag{7.2}$$

where c_m are actual scores, $\hat{c}_m$ predicted scores, and M is the number of subjects in the testing dataset. Pearson's r is expected to be close to 1, and the RMSE close to 0 for perfect predictions. The RMSE is particularly interesting as it can be compared to the standard deviation of the variables to be predicted.

Subject heterogeneity

In order to study subject variability within the HCP and PIOP2 dataset, we first compute edge weight standard deviation across subjects for each edge and for different inter-correlation quantile values and the CA estimator. Furthermore, associations between psychometric scores and brain-level connectivity is obtained for two subsets of subjects. The latter are constructed according to a selected cross-validation fold. The same linear fit as the one used in the univariate association study is obtained from each of the two subsets of subjects for fixed quantile values. Edge-level connectivity is also obtained for individual edges. For a fixed edge, linear fits are similarly computed from each of two subsets of subjects. Finally, we construct edge variability networks for different inter-correlation estimators. In such networks, edge weights are set to the standard deviation of the inter-correlation values of the corresponding edge, calculated across all subjects of the dataset.

7.3 Results

7.3.1 Inter-correlation edge summarization distribution

To begin with, we visually inspect the distribution of the absolute value of the inter-correlation edge summaries for the two examinations of four HCP subjects (cf. Figure 7.3.1). This figure can be compared to Figure 2.5 in Chapter 2, which displayed inter-correlation distribution of various estimators for the same four subjects. We can first note that, for a fixed subject and session, CA is more variable across edges than quantiles of the clustering-based inter-correlation. Moreover, CA, which is known to overestimate true inter-correlation (Achard et al., 2023; Lbath et al., 2023), reaches correlation values up to almost 1. While the central quantiles (Q50, Q60) showcase correlation values centered close to zero, they attain values of up to $0.3 - 0.5$. The tails of the inter-correlation distribution (Q10, Q90) provide absolute values of correlations that are higher. They are nevertheless lower than the most extreme CA values, which is in accord with consistency results in Chapter 6.

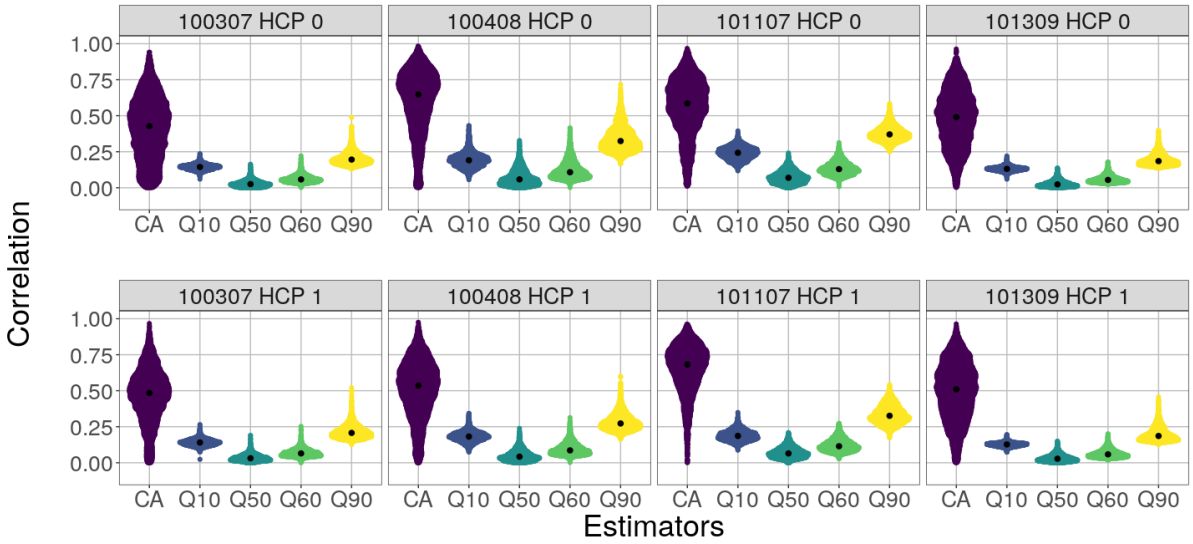


Fig. 7.3.1 Edge weight distribution. Empirical distribution across all pairs of brain regions of the absolute value of inter-correlations for the CA estimator and selected quantile values of the clustering-based inter-correlation estimator for four subjects of the HCP dataset. Each subject was scanned twice, on different days. The black dot corresponds to the median. The CA estimator is much more variable than the quantile-based estimators.

7.3.2 Test-retest repeatability

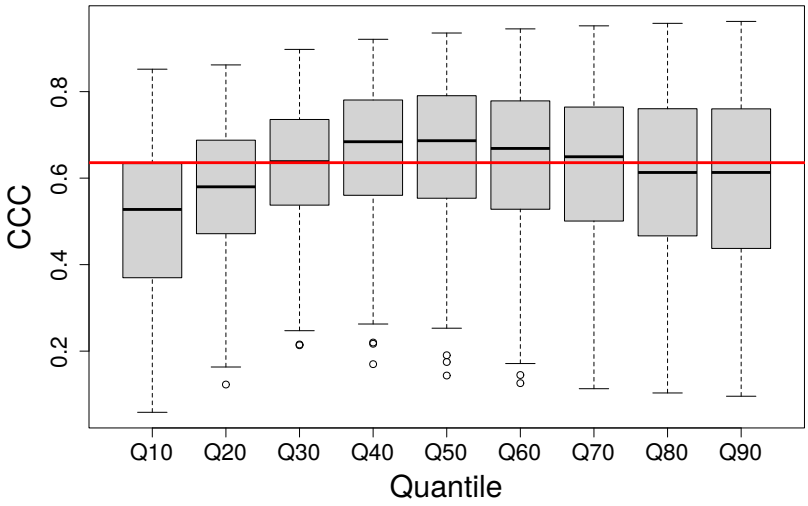
We now evaluate repeatability of the RSFC networks. Results are showcased in Figure 7.3.2. The inter-correlation CCC distribution across the 100 HCP subjects varies with the quantile value and tend to be highest for quantiles ranging from 0.4 to 0.6, i.e. Q40 to Q60 (cf. Figure 7.3.2a). This implies that the tails of the inter-correlation distributions are less repeatable. Additionally, the average CCCs across subjects for quantiles Q40 to Q60 range from 0.64 to 0.66, which are slightly higher than that of the CA estimator (0.63). Graph metric betweenness repeatability results are reported in Figure 7.3.2b for varying network threshold values. It shows all distribution-based estimators yield higher betweenness CCC than CA, with Q10 and Q90 the most repeatable for all network sparsity thresholds. Indeed average betweenness CCC across all 100 subjects peaks at 0.85 for Q90 and 0.35 for CA. Furthermore, while the betweenness CCC of CA decays to below 0.2 as the percentage of edges included in the RSFC networks increases, that of Q10 and Q90 remain high at 0.85, initially reached for slightly less than a 10% threshold, and only slightly decrease to reach approximately 0.70 for a 50% threshold.

7.3.3 Machine learning for psychometric variables prediction and classification

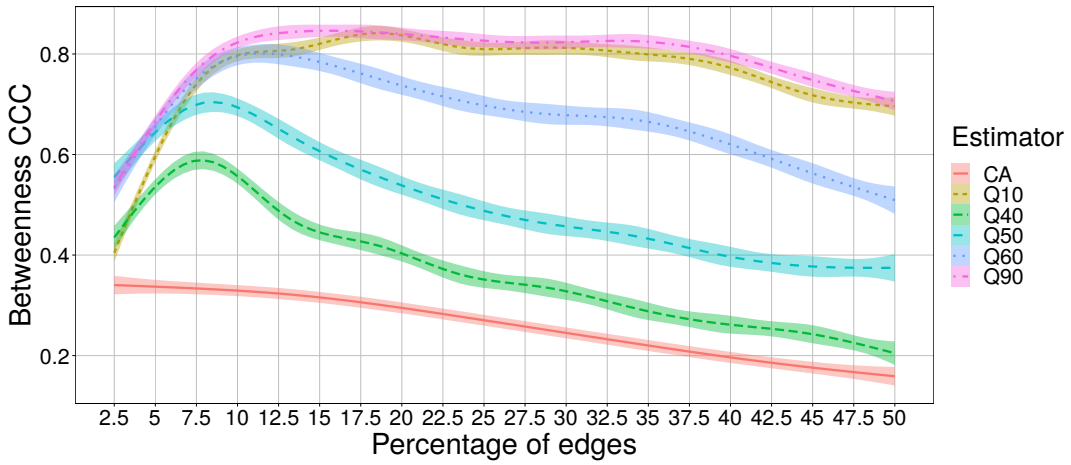
We now evaluate our proposed framework in terms of machine learning task performance.

Brain-level analysis

We first report results on brain-level summaries. Figure 7.3.3 depicts univariate associations between brain-level functional connectivity and either conscientiousness (NEO C) scores or fluid intelligence (Raven) for the HCP data. Pearson's r with Raven's scores is in the order of 0.1, which is quite low, for all quantile values as well as the CA estimator. It is however in the same order of magnitude as that obtained by Marek et al. (2022), where authors considered three datasets, including $1,200$ subjects from the HCP dataset, and the CA estimator is used on a parcellation with 394 regions. We note as well in Figure 7.3.3 that effect size from association with NEO C scores are slightly larger for the 0.5 and 0.9 quantiles (Q50, Q90) and the CA estimator, the largest being obtained by Q90 ($|r| = 0.18$). Nevertheless, correlation between the 0.1 (Q10) quantile at the brain-level and NEO C is 0.01. Association with other NEO-FFI personality variables are conducted, but not reported here as they show similar results, with NEO C yielding the highest effect size. Brain-level association results are similar as well for the PIOP2 data (cf. Figure 7.A.1).



(a) Edge weight repeatability.



(b) Graph metric repeatability.

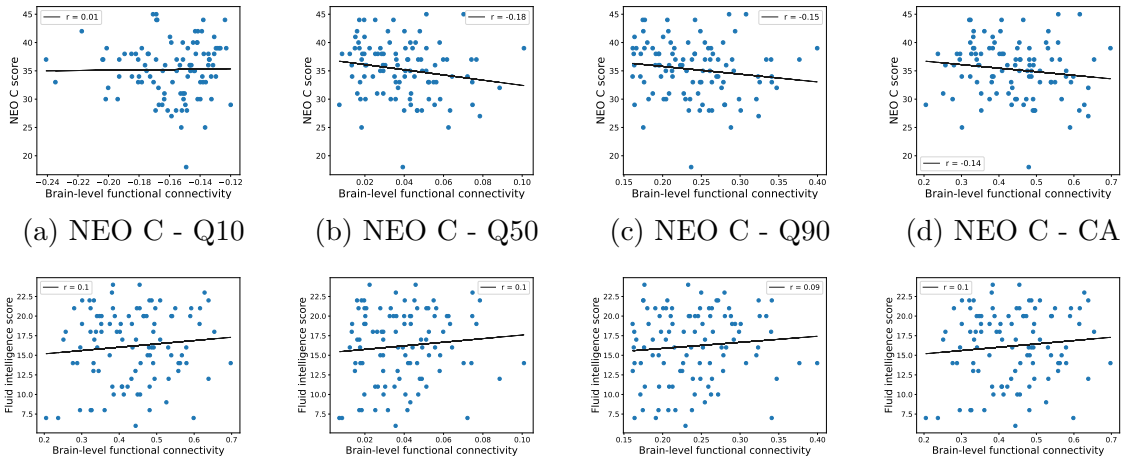
Fig. 7.3.2 Repeatability of the HCP data. **(a)** Distribution across the 100 HCP subjects of the Concordance correlation Coefficient (CCC) for the test-retest reliability of edge weights for varying quantiles of the inter-correlation distributions, ranging from the 10th to the 90th percentile. The CCC is computed between the two examinations for each subject. The mean CCC across the 100 subjects for the CA estimator is represented by a solid red line. Higher CCC indicates a more repeatable estimator. the center of the inter-correlation distribution is the most repeatable. **(b)** CCC of a topological graph metric (betwenness) according to varying choices of network threshold and for different quantiles of the inter-correlation distributions and the CA estimator. The tails of the inter-correlation distribution provide the most repeatable network organization in terms of betweenness.

Table 7.3.1 Raven's score classification results using SVC. Accuracy and AUC are obtained using 10-fold cross-validation on the HCP 0, HCP 1 and PIOP2 datasets, and compared across different edge weight estimates for various graph representations.

		Q10			Q40			Q50			Q60			Q90			CA		
	Metric	HCP 0	HCP 1	PIOP2	HCP 0	HCP 1	PIOP2	HCP 0	HCP 1	PIOP2	HCP 0	HCP 1	PIOP2	HCP 0	HCP 1	PIOP2	HCP 0	HCP 1	PIOP2
Brain	Accuracy	0.44	0.58	0.54	0.34	**0.67**	0.55	0.44	0.60	0.55	0.52	0.56	**0.56**	**0.54**	0.45	0.53	0.44	0.47	**0.56**
	AUC	0.32	0.55	0.53	0.33	**0.63**	0.52	0.44	**0.63**	0.52	0.50	0.59	0.49	**0.55**	0.41	**0.54**	0.27	0.46	0.53
Edge	Accuracy	0.51	0.61	0.51	0.41	0.64	**0.60**	0.43	**0.67**	0.57	0.48	0.61	0.53	**0.53**	0.50	0.46	0.42	0.44	0.57
	AUC	0.45	0.55	0.54	0.26	**0.69**	0.52	0.37	0.66	**0.57**	0.49	0.65	0.56	**0.55**	0.52	0.42	0.38	0.54	0.53
Edge + Feather	Accuracy	0.33	0.37	0.42	0.31	**0.65**	**0.47**	**0.41**	0.34	0.44	0.36	0.37	0.40	0.35	0.34	0.37	0.35	0.31	0.45
	AUC	0.48	0.50	0.50	0.44	**0.65**	**0.54**	**0.50**	50	0.51	0.48	0.51	0.50	0.48	0.49	0.49	0.48	0.44	0.52
Edge + g2vec	Accuracy	**0.52**	**0.52**	0.46	0.46	**0.52**	0.42	**0.52**	**0.52**	0.46	**0.52**	**0.52**	0.46	0.51	0.51	**0.47**	**0.52**	**0.52**	0.46
	AUC	0.50	**0.50**	**0.50**	0.48	**0.50**	0.43	0.50	**0.50**	0.48	0.50	**0.50**	0.49	**0.51**	**0.50**	0.48	0.50	**0.50**	0.48

Table 7.3.2 Psychometric scores regression performance using Random Forest. RMSE and Pearson's r are obtained using 10-fold cross-validation for NEO C on the HCP 0 and HCP 1 scans, and NEO N on the PIOP2 datasets, and compared across different edge weight estimates for various graph representations. * and ** correspond to positive correlation values with a p-value < 0.05, < 0.001, respectively (zero correlation null hypothesis).

		Q10			Q40			Q50			Q60			Q90			CA		
	Metric	HCP 0	HCP 1	PIOP2	HCP 0	HCP 1	PIOP2	HCP 0	HCP 1	PIOP2	HCP 0	HCP 1	PIOP2	HCP 0	HCP 1	PIOP2	HCP 0	HCP 1	PIOP2
Brain	Pearson's r	**0.27****	−0.20	−0.10	−0.09	**0.16**	−0.22	0.06	0.11	**0.12**	0.08	0.11	0.04	−0.18	0.04	0.02	0.20*	0.11	−0.12
	RMSE	**5.3**	6.2	9.7	6.0	**5.5**	10	5.7	5.7	**9.0**	5.7	5.7	9.3	6.3	5.6	9.3	5.4	5.6	10
Edge	Pearson's r	0.07	0.08	−.08	0.09	0.16	−0.15	**0.20***	**0.26****	−0.08	0.15	0.24*	−0.02	−0.04	0.19	**0.04**	0.20	0.15	0.02
	RMSE	4.9	4.9	8.2	4.9	4.8	8.2	**4.7**	**4.7**	8.2	4.8	**4.7**	**8.1**	5.1	4.8	**8.1**	4.8	4.8	**8.1**
Edge + Feather	Pearson's r	**0.23***	−0.23	−0.17	0.01	**0.25***	−0.21	0.01	0.12	**0.08**	0.05	0.17	0.01	−0.23	0.06	−0.02	0.17	0.20*	−0.19
	RMSE	**5.2**	6.1	9.9	5.5	**5.0**	9.7	5.6	5.4	**8.9**	5.7	5.3	9.4	6.3	5.5	9.6	5.3	5.2	10
Edge + g2vec	Pearson's r	−0.01	0.05	−0.13	0.07	0.04	−0.03	**0.16**	0.03	**0.01**	−0.03	**0.09**	−0.09	0.11	0.03	0.00	−0.02	−0.04	−0.02
	RMSE	5.3	5.1	8.5	5.0	**5.0**	8.2	**4.8**	5.1	8.4	5.2	5.1	8.4	5.1	5.3	**8.1**	5.1	5.4	8.6



(a) NEO C - Q10 (b) NEO C - Q50 (c) NEO C - Q90 (d) NEO C - CA

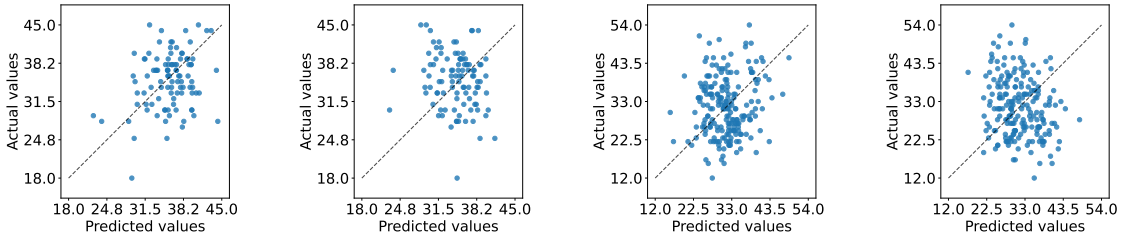
(e) Raven - Q10 (f) Raven - Q50 (g) Raven - Q90 (h) Raven - CA

Fig. 7.3.3 Brain-level connectivity association with psychometric scores. **(a-d)** Conscientiousness (NEO C) and **(e-h)** fluid intelligence (Raven) scores are considered for three quantile values (Q10, Q50, Q90) and the CA estimator. The black solid lines represent the linear regression fits, and Pearson's correlation value r are reported. Data from the first examination (HCP 0) of 100 HCP subjects is used and reveal low effect size (|r| in the order of 0.1).

Brain-level classification results are recorded in the first two rows of Table 7.3.1. They show at least one quantile-based estimator always outperforms the CA estimator on the HCP data, in terms of both AUC and accuracy. The latter reaches 0.67 for Q40 on the second scan (HCP 1), which is in the same order of magnitude as the median accuracy of the optimal fluid intelligence classification pipeline on HCP data in the following benchmarking study (Dadi et al., 2019). Brain-level classification of the PIOP2 data using quantile-based edge summaries is better, or at least as good as that using the CA estimator in terms of both accuracy and AUC.

Similarly, quantile-based RSFC estimators achieve improved psychometric score regression results compared to the CA estimator. Table 7.3.2 reports regression results for NEO C scores for the HCP dataset and neuroticism (NEO N) scores for the PIOP2 dataset. We recall that good predictions yield high Pearson's r and low RMSE. Results for other NEO-FFI personality scores are not recorded as they yield lower performance. It can first be noted that Pearson's r between predicted and actual values is the highest across all three sets of scans and estimators for Q10 of the first HCP session (HCP 0), and is significantly different from 0 ($r = 0.27$, p-value< 0.01). Nonetheless, the lowest RMSEs, which are obtained by quantile-based RSFC, are quite high, and equal to slightly more than the standard deviation of the scores for all three sets of scans. Figure 7.3.4

showcases actual values plotted against predicted values for the best (highest Pearson's r) and worst (highest RMSE) performance among quantile-based estimators for the HCP 0 and PIOP2 data. Points tend to be scattered along the identity line in the best regression scenarios, while in the worst regression case, actual and predicted values are visibly negatively associated. In fact, Q90 yields a Pearson's r of -0.18 for the HCP 0 data, and Q40 yields a Pearson's r of -0.22 for the PIOP2 data. However, in both scenarios, predictions seem to be concentrated around their respective mean psychometric scores.



(a) NEO C - Q10 - HCP 0 (b) NEO C - Q90 - HCP 0 (c) NEO N - Q50 - PIOP2 (d) NEO N - Q40 - PIOP2

Fig. 7.3.4 Brain-level regression results. Actual psychometric score values versus values predicted by a random forest regressor for the **(a,c)** best case (highest Pearson's r across quantile values) and **(e-h)** the worst case (highest RMSE) using the first examination of the 100 HCP subjects (HCP 0) and the 204 PIOP2 subjects. **(a,b)** Conscientiousness (NEO C) and **(c,d)** neuroticism (NEO N) scores are considered. Points tend to be scattered along the identity line (black dashed line) in the best regression scenarios, while in the worst regression case, actual and predicted values are visibly negatively associated. In both scenarios, predictions seem to be concentrated around mean scores.

Edge-level analysis

We now examine univariate edge-level association results. Figure 7.3.5 depicts the distribution across edges of Pearson's r correlation between edge weights and fluid intelligence scores for different clustering-based inter-correlation quantiles and the CA estimator. We note edge weight associations are quite weak for all inter-correlation estimators. They are however higher than brain-level associations, with Peason's r values ranging from -0.3 to 0.3. Effect sizes are slightly different for each of the two sessions from the HCP datasets as well as the PIOP2 dataset. Distributions of Pearson's r between edge weights and additional psychometric scores can be found in the appendix and are similarly weak (cf. Figure 7.C.1). Conscientiousness is the score with the highest correlations with HCP edge connectivity, with maximal absolute values reaching 0.4.

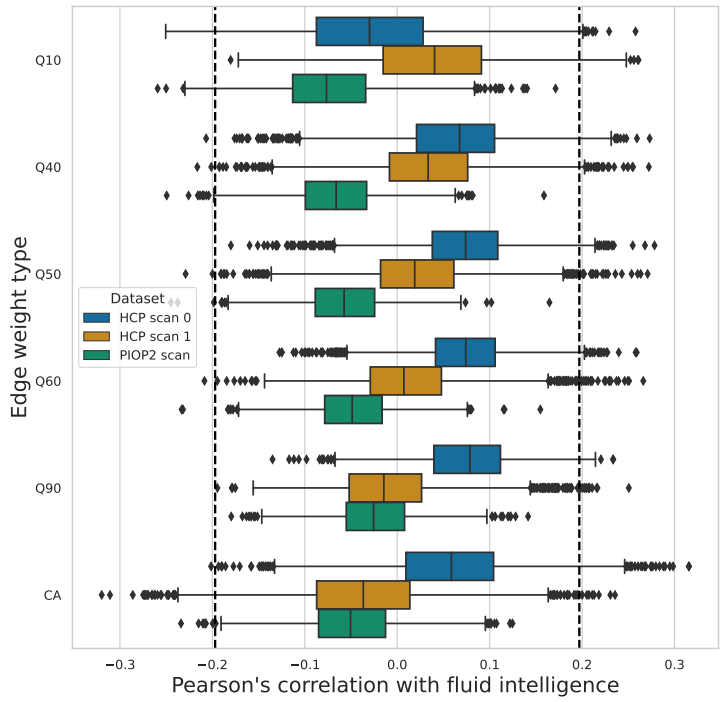
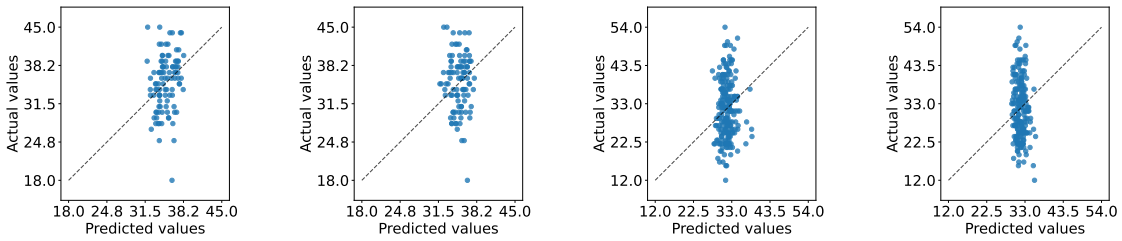


Fig. 7.3.5 Distribution across all edges of Pearson's correlation between edges weights and fluid intelligence scores. The correlation values are computed across all subjects for different quantiles of the edge inter-correlation distributions and the CA estimator. Distributions for PIOP2 dataset and both scans of the HCP dataset are depicted. The dashed vertical lines represent the critical correlation values corresponding to a 5% nominal level.

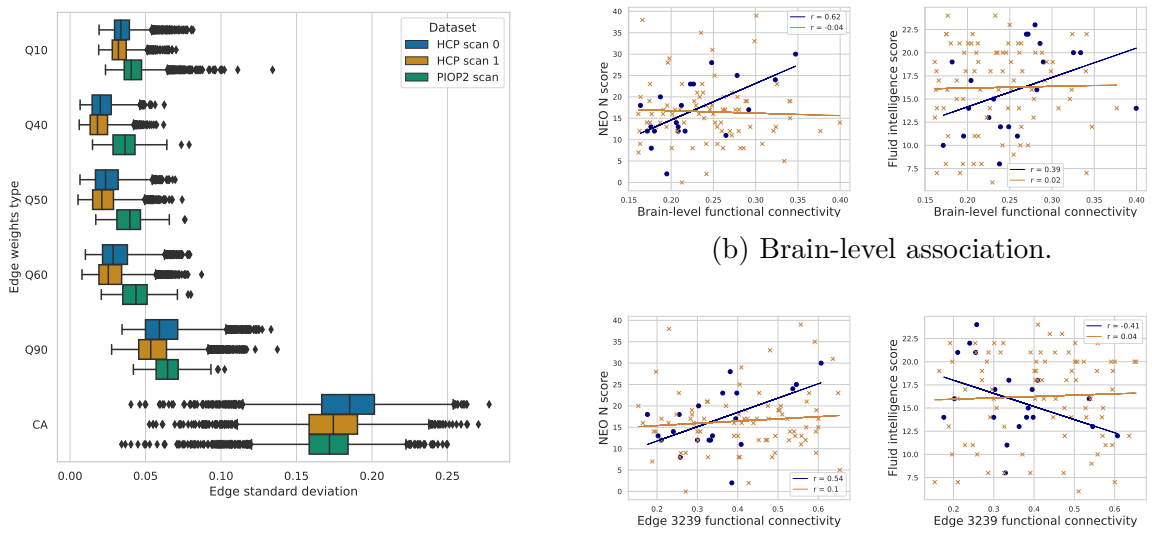
We then explore multivariate edge-level associations. First, classification results of the edge-level connectivity are similar to that of brain-level, with quantile-based connectivity performing better than CA and the highest accuracy at 0.67 and AUC at 0.69 for the HCP data (cf. Table 7.3.1). Feather embedding yields similar results to the direct embedding for the HCP 1 data, but deteriorates HCP 1 and PIOP2 results, with a maximum accuracy at 0.47 for Q40 of PIOP2. Furthermore, quality of the graph2vec embedding classification is worsened compared to the brain- and edge-level network representations. In fact, graph2vec accuracy and AUC are close to 0.50 for all settings. Nonetheless, our proposed quantile-based RSFC improve classification results for both embeddings.



(a) NEO C - Q50 - HCP 0 (b) NEO C - Q90 - HCP 0 (c) NEO N - Q90 - PIOP2 (d) NEO N - Q40 - PIOP2

Fig. 7.3.6 Edge-level regression results. Actual psychometric score values versus values predicted by a random forest regressor for the **(a,c)** best case (highest Pearson's r across quantile values) and **(e-h)** the worst case (highest RMSE) using the first examination of the 100 HCP subjects (HCP 0) and the 204 PIOP2 subjects. **(a,b)** Conscientiousness (NEO C) and **(c,d)** neuroticism (NEO N) scores are considered. Points tend to be more scattered along the identity line (black dashed line) in the best HCP data edge-level regression scenario. In both scenarios, predictions seem to be concentrated around mean scores, even more so in the PIOP2 data.

Evaluation of the regression of psychometric scores on multivariate edge-level connectivity for various inter-correlation quantile values and the CA estimator are reported in Table 7.3.2. Once again, our proposed quantile-based estimators outperform CA. Edge-level regression somewhat improves RMSE, compared to brain-level regression, and is now slightly less than one standard deviation of the psychometric scores to be predicted. Nevertheless, edge-level regression provides Pearson's r that are highly dissimilar to that of brain-level regression. Indeed, the highest r on the HCP 0 data is obtained for Q50 ($r = 0.20$, p-value < 0.05), while it is obtained for Q10 in the brain-level regression. Strongest performances are similarly mismatched between edge- and brain-level for HCP 1 and PIOP2. However, Feather embedding is similar to brain-level RSFC for all three datasets in terms of both RMSE and Pearson's r, with the exception of HCP 1, which yields $r = 0.25$, p-value< 0.05 for Q40, which is much higher than r at brain-level and closer to that of the direct embedding at Q50. The graph2vec embedding is worse for all three datasets in terms of Pearson's r, but is similar to the direct embedding in terms of RMSE. Plots of actual values against predicted values showcase predictions that tend to be more concentrated around the mean scores than that of brain-level classification (cf. Figure 7.3.6).



(a) Edge standard deviation.

(b) Brain-level association.

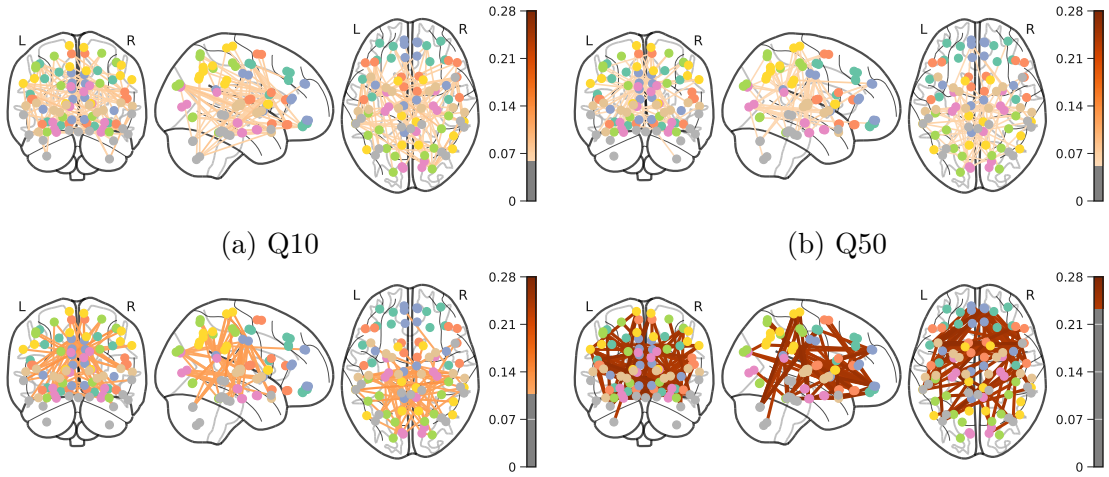
(c) Edge-level association.

Fig. 7.3.7 Subject heterogeneity. **(a)** Distribution of edge standard deviation calculated across all subjects for each edge for various inter-correlation quantiles and the CA estimates. **(b-c)** Brain- and edge-level connectivity association with neuroticism (NEO N) and fluid intelligence (Raven's score) scores, for the 0.9 quantile. The navy solid line represents the linear regression fit of the smallest of the two subsets (20 subjects represented by navy dots) and the golden solid line corresponds to the linear fit of the remaining 80 subjects (golden crosses). Subsets are different for different figures. Pearson's correlation value r are reported. Data from the first examination (scan 0) of 100 HCP subjects is used and reveal that subsets of subjects can have opposing associations.

Subject heterogeneity

We now investigate subject heterogeneity in the HCP and PIOP2 datasets. To begin with, the distribution across edges of the standard deviation of edge weights calculated across subjects for both HCP scans and the PIOP2 data are shown in Figure 7.3.7a. The CA estimator displays the most variability across subjects, with median edge standard deviation of about 0.17 for all datasets. All quantile-based estimators enable a marked decrease in variability, with edge standard deviation in the order of 0.01.

Nonetheless, Figure 7.3.7 highlights the fact that independent subsets of the subjects can display highly differing association effects, e.g., $r = 0.62$ and $r = 0.04$ for the relationship between NEO N and brain-level RSFC on the HCP data. This phenomenon is also observed for univariate edge-level associations, and are corroborated by the PIOP2 data.



(a) Q10 (b) Q50

(c) Q90 (d) CA

Fig. 7.3.8 Edge variability networks. Data from the first examination (scan 0) of 100 HCP subjects. Brain edge standard deviation networks for **(a-c)** different inter-correlation quantiles and **(d)** the CA estimate. Top 2% of most variable edges. Color legend represents the standard deviation value associated to the edge.

To examine the spatial location of the most variable edges, edge standard deviation networks are represented in Figure 7.3.8 for the first scans of the HCP dataset. Edge variability networks of the second HCP scans are qualitatively similar, and hence are not reported here. PIOP2 edge variability networks can be found in the appendix (cf. Figure 7.A.2). HCP edge variability networks indicate that the most variable CA-based edges are located in frontal brain regions, unlike quantile-based edges. Indeed, the latter display the highest variability in the occipital regions. PIOP2 edge variability networks indicate similar spatial location of primary edge variability, with the exception of Q10, which showcases some additional variability in frontal regions, and of CA, whose most variable edges are now distributed across the entire brain. These figures highlight as well the fact that quantile-based estimates are much less variable than CA, as previously shown in Figure 7.3.7a.

7.4 Discussion

7.4.1 Test-retest repeatability

Our first validation step involved assessing test-retest reliability. We showed that the center of the clustering-based inter-correlation distribution was the most repeatable and that it was also more reliable on average than the CA estimator (cf. Figure 7.3.2). We

additionally remark that the most repeatable quantiles were also the ones where edge weight variability is the lowest (cf. Figure 7.3.7a).

Nevertheless, edge weight CCC values were highly variable across subjects, ranging from less than 0.2 to over 0.8 for all quantile values. In a previous study, nearly half of the healthy adult participants were found to have fallen asleep during resting-state fMRI scan (Soehner et al., 2019). This could help explain why some subject showcase particularly low repeatability. Regardless of cause, it might be prudent to remove less repeatable individuals in future studies.

As opposed to edge weight CCC, betweenness CCC was highest for the tails of the inter-correlation distribution (cf. Figure 7.3.2b). This might be explained by the fact betweenness is sensitive to region disconnection, since changing a single edge in the graph may imply alterations in the shortest paths of the graph. Indeed, central quantiles correspond to low inter-correlation value (cf. Figure 7.3.1), which might be more sensitive to network thresholding, and thus lead to fluctuating network topologies. It may be interesting to explore as well the repeatability of additional graph metrics, such as efficiency. In addition to that, the betweenness CCC of CA sharply declined for denser networks, which contain many edges with weights not significantly different from zero, unlike that of quantile-based estimators. This indicates that the latter are much more reliable estimators of lower-valued edges than CA.

7.4.2 Classification and regression methods choice

We then evaluated performance on machine learning tasks. Before discussing the results presented above, we briefly address the choice of classification and regression method. In addition to SVC, Ridge regression and l_2 logistic regression were both recommended for classification tasks in fMRI studies (Dadi et al., 2019). These methods were applied to both brain- and edge-level classification of HCP datasets. However, Ridge regression tended to assign all subjects to a single class, and performance of the l_2 logistic regression was worst than that of SVC. A random forest classifier, which is also sometimes used in neuroimaging data (Sarica et al., 2017), was also evaluated. Overall, its performance was similar to that of SVC on the HCP data, but it was much worse on the PIOP2 data. SVC is also the most commonly used classifier in autism-related neuroimaging studies according to a survey conducted by Vabalas et al. (2019), and reinforced our decision to use SVC in this chapter. Regarding the regression task, random forest regression, SVR and ridge regression were explored. However, performance of the latter two were much worse, in terms of both Pearson's r and RMSE.

7.4.3 Brain-level analysis

We now turn to the performance of machine learning tasks on brain-level RSFC. We showed our proposed distribution-based approach improved both classification and regression results compared to CA, although performance remained modest. Nevertheless, the highest accuracy level (0.67) obtained on the HCP 1 data (using Q40) was on par with state-of-the-art psychometry-related classification of RSFC fMRI data. Indeed, accuracy of classifiers in the context of neuroimaging studies is expected to be in the order of $60 - 70\%$, e.g., (Kassraian-Fard et al., 2016; Dadi et al., 2019). Regression results were similarly moderate. Nevertheless, Pearson'r was almost always significantly greater than 0 for at least one quantile. We remark however that classification and regression performances were not quite replicable across the three sets of scans. Indeed, while performances were improved by our proposed quantile-based estimates, the specific quantiles that worked best were different for each set of scans. Furthermore, univariate associations unexpectedly did not seem to be linked to classification and regression performance of the various inter-correlation quantile values. Precise interpretation of association between personality and cognitive scores and brain-level RSFC may be difficult, as brain-level aggregation may be too coarse. It might however be linked to a baseline physiological state, analogously to the known relationship with the amplitude of BOLD signals (Hall et al., 2016).

7.4.4 Edge-level analysis

On the other end of the spectrum, we considered edge-level RSFC. Our clustering-based correlation estimator once again performed better than the standard CA estimator in the classification task (cf. Table 7.3.1). Moreover, evaluation of the classification and regression tasks indicate moderate, and sometimes non-significant, results, with regression RMSE roughly equal to the standard deviation of the psychometric scores to be predicted, and the highest accuracy across all estimators, network embedding techniques, and datasets, equal to 0.67. However, the latter is a state-of-the-art result (Dadi et al., 2019), as mentioned previously. Similarly to the brain-level analysis, edge-level classification and regression results for both quantile-based and CA RSFC, highly differ across sets of scans. Regression results also vary across network representation levels, highlighting the complexity of the task.

Furthermore, with $3,916$ and $2,850$ edges, and 100 and 204 subjects in HCP and PIOP2 RSFC networks, respectively, we find ourselves in a high-dimensional setting. That information, combined with the moderate results of the edge-level analysis indicate

a need for dimension reduction. Additionally, crucial information about connectivity networks of comatose patients was previously found at a higher scale level (Achard et al., 2012a), further suggesting edge-level RSFC may be too fine-grained. Nevertheless, brain-level classification results were worsened for the PIOP2 data, but similar for the HCP data compared to that of the edge-level. They were however improved for the brain-level regression on PIOP2. Yet, while brain-level RSFC may be too coarse, edge-level Feather and graph2vec embeddings yielded the overall worst classification results. Further inspection revealed these could be explained by their tendency to assign all subjects to a single class. Therefore, it appears the two embeddings failed to capture the structure that is the most relevant to associations with psychometric scores. Further work is hence required to identify effective dimension reduction approaches in this context.

7.4.5 Subject heterogeneity

Previous works suggested subject heterogeneity may help explain modest performances in RSFC network machine learning tasks (Schnack and Kahn, 2016; Marek et al., 2022).

We first noted the effect size of the association between psychometric scores and both brain- and edge-level RSFC varied across subsets of subjects (cf. Figure 7.3.7). This corroborated previous observations by Marek et al. (2022), and indicated high subject heterogeneity. Subsequently, qualitative inspection of the distribution across edges for individual subjects of inter-correlation estimates showed the CA estimator yielded exceedingly high inter-correlation values compared to our proposed quantile-based estimators (cf. Figure 7.3.7a). It is worth noting CA is also the estimator that produced the worst classification and regression results.

The impact of subject variability may be heightened by the sample size. This hypothesis is supported by a recent work that advocated for the use of thousands of individuals in rs-fMRI connectivity association studies (Marek et al., 2022). Previously, Cui and Gong (2018) had also advised a minimum of 200 subjects for RSFC regression tasks. Moreover, it has been suggested small sample sizes may yield artificially larger effect sizes in RSFC settings because of their particularly high homogeneity (Schnack and Kahn, 2016; Marek et al., 2022). This brings us back to our initial observation on subsets of subjects, some of which displayed particularly high effect sizes.

It is therefore particularly relevant that our proposed quantile-based estimator enable a reduction in edge-specific heterogeneity compared to CA. We additionally showed that this phenomenon is mostly localized in frontal brain regions, which are known to be more prone to the effect of breathing and motion artifacts (Xifra-Porxas et al., 2021).

7.4.6 Distribution-based weighted networks

In this work, we introduced distribution-based connectivity networks. Assigning distributions or density functions to edges, instead of point estimates such as CA, enables more flexibility and creates new opportunities for an expanded utilization of data. As a first step, different quantile values can be chosen for different applications, and we showed their relevance throughout this chapter. These distributions could be further exploited, for instance for data augmentation purposes or for ensemble learning, by obtaining several scalar-weighted graphs per individual subject via approaches akin to bootstrapping. Manipulating density functions also opens up to the possibility of using functional data analysis tools. For example, functional PCA scores of the inter-correlation densities could be used as input features for machine learning tasks.

7.4.7 Conclusion and perspectives

To conclude, our proposed distribution-based network framework improved repeatability, classification and regression performance over the standard CA estimator on both the HCP and PIOP2 datasets. However, while our proposed method emerged as the most superior option, and classification results were similar to state-of-the-art performances (Dadi et al., 2019), the latter remained moderate. Furthermore, our exploration of several datasets emphasized the arduous nature of performing reliable classification or regression using resting-state fMRI to predict psychometric scores. Several factors may help account for these modest performance, which still surpass that from CA-based networks.

First, as previously detailed, subject heterogeneity may impair performance in RSFC settings (Schnack and Kahn, 2016). Nevertheless, we showed our proposed quantile-based edge weights are much less variable across subjects than the CA weights.

However, our sample sizes, of approximately 100 and 200, may prove too small to offset the effects of subject heterogeneity. Indeed, recent work recommended the use of thousands of subjects (Marek et al., 2022) in brain-wide RSFC association studies. Furthermore, classification accuracy of RSFC was also shown to artificially increase when sample size decreases (Vabalas et al., 2019), further underlining the need for larger sample size. This may however be difficult to achieve in practice, and even more so when investigating certain pathologies.

Additional preprocessing may also help increase performance. For instance, removing physiological and motion nuisance has been shown to improve subject identifiability (Xifra-Porxas et al., 2021). However, previous works have stressed the importance of

exercising caution when choosing preprocessing techniques as valuable signals may be lost (Murphy and Fox, 2017; Liu et al., 2017b).

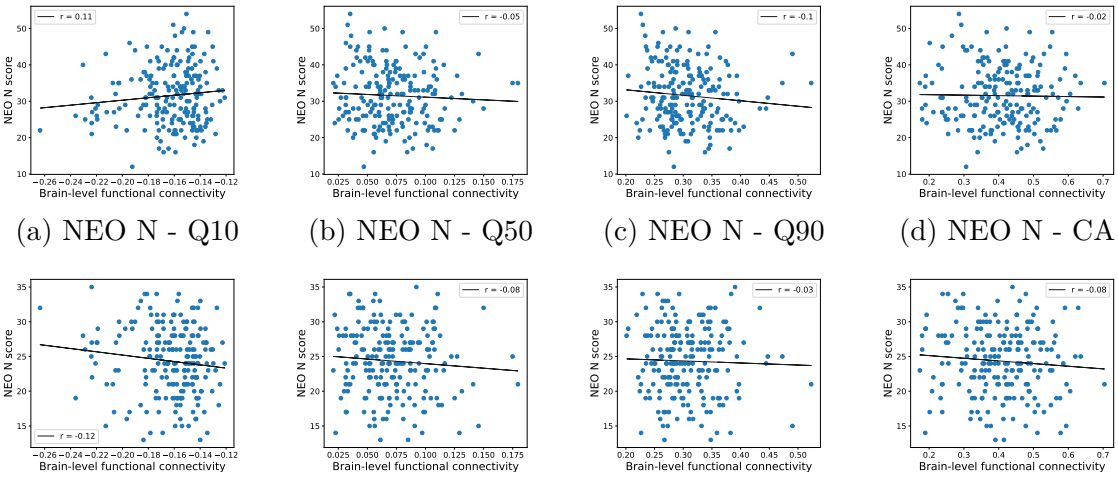
Yet, it could simply be links between rs-fMRI functional connectivity and psychometric information are tenuous. It would hence be interesting to validate the proposed framework on datasets where the informativeness of resting-state imaging would be a more apparent, such as in comatose patient studies, where resting-state is the only available fMRI modality. Multi-modality may also help improve performances (Sarica et al., 2017).

It remains that we are in a high-dimensional setting. In order to counteract the lack of subjects, which is compounded by the subject heterogeneity and high number of edges more thoughtful dimension reduction methods need to be applied. This can be translated in terms of improved aggregation of edges. For instance, a node-level analysis that take into account edge information could be considered. Functional connectivity strength (FCS) is a common summary metric in brain connectivity studies. It consists in averaging each row of the RSFC matrix, which in other words amounts to averaging weights of all edges connected to each node. However, preliminary classification and regression results were not promising. Furthermore, Cui and Gong (2018) have shown that FCS performs worse than RSFC in regression tasks. Another, and more insightful, approach could be to group nodes into equivalence-based node clusters (Carboni et al., 2023) and average inter-correlations accordingly, instead of aggregating over the entire brain.

Nevertheless, Random Forests and SVM require input graphs to be embedded into vector space, and preclude the use of multidimensional node and edge features. Graph Neural Networks (GNN) could be seen through the lens of node and edge aggregation. Indeed, they traditionally contain message-passing layers, which combine neighboring node or edge information. For instance, Graph Attention Network layers perform weighted aggregation based on the importance of node and edge neighbors. Furthermore, GNNs perform end-to-end learning and take networks as an input, circumventing the network embedding step. GNNs were hence recently used for classification of major depressive disorder from RSFC (Gallo et al., 2023). However, edge features were not taken into account and accuracy results were similarly modest to that of more common classifiers. We are expecting the use of possibly multivariate edge weights would significantly improve performances. Preliminary explorations were conducted but application of GNNs to distribution-based networks is not all that straightforward.

Appendix

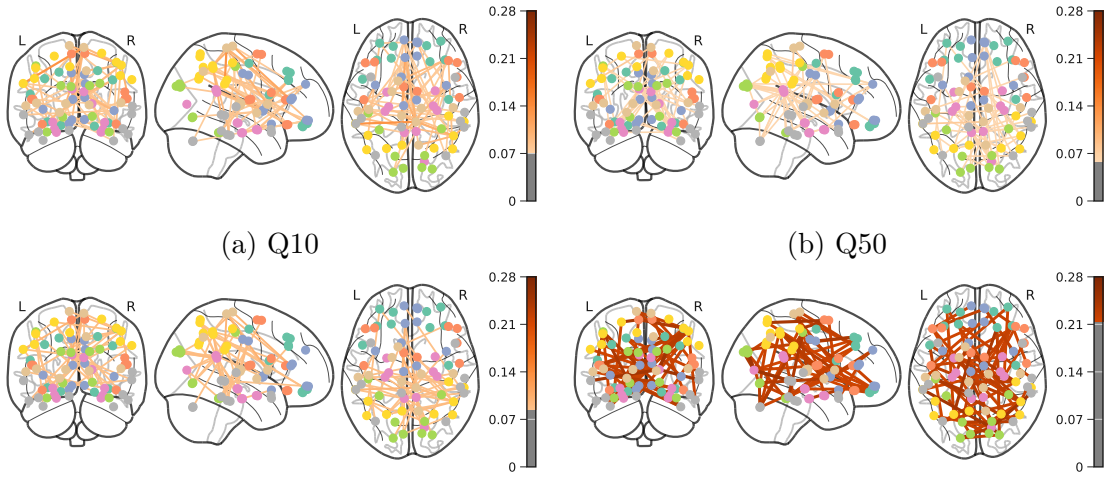
7.A Additional PIOP2 data figures



(a) NEO N - Q10 (b) NEO N - Q50 (c) NEO N - Q90 (d) NEO N - CA

(e) Raven - Q10 (f) Raven - Q50 (g) Raven - Q90 (h) Raven - CA

Fig. 7.A.1 Brain-level connectivity association with psychometric scores. **(a-d)** Neuroticism (NEO N) and **(e-h)** fluid intelligence (Raven) scores were considered for three quantile values and the CA estimator. The black solid lines represent the linear regression fits, and Pearson's correlation value r are reported. Data from the 204 PIOP2 subjects was used and revealed low effect size (|r| in the order of 0.1).

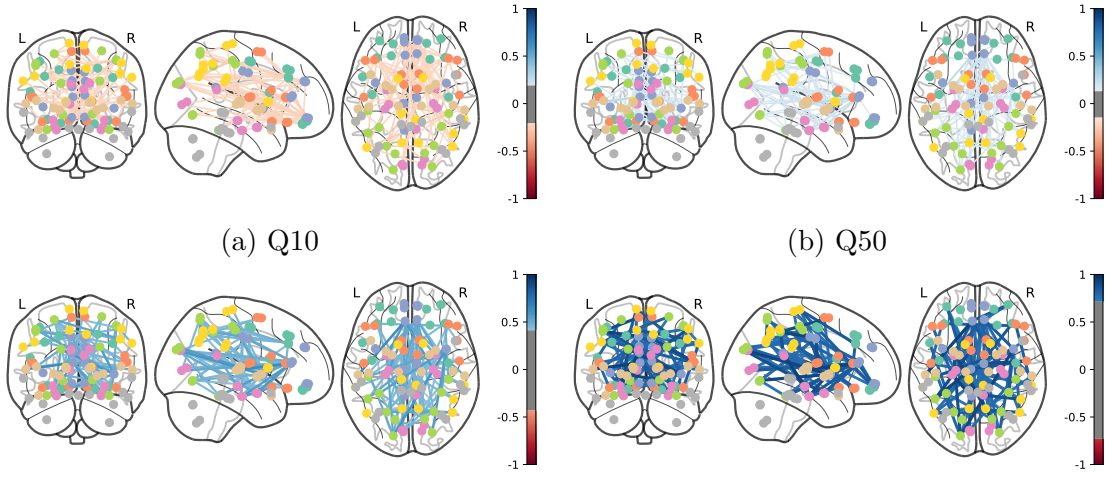


(a) Q10 (b) Q50

(c) Q90 (d) CA

Fig. 7.A.2 Edge variability networks. Data from the 204 PIOP2 subjects. Brain edge standard deviation networks for **(a-c)** different inter-correlation quantiles and **(d)** the CA estimate. Top 2% of most variable edges. Color legend represents the standard deviation value associated to each edge.

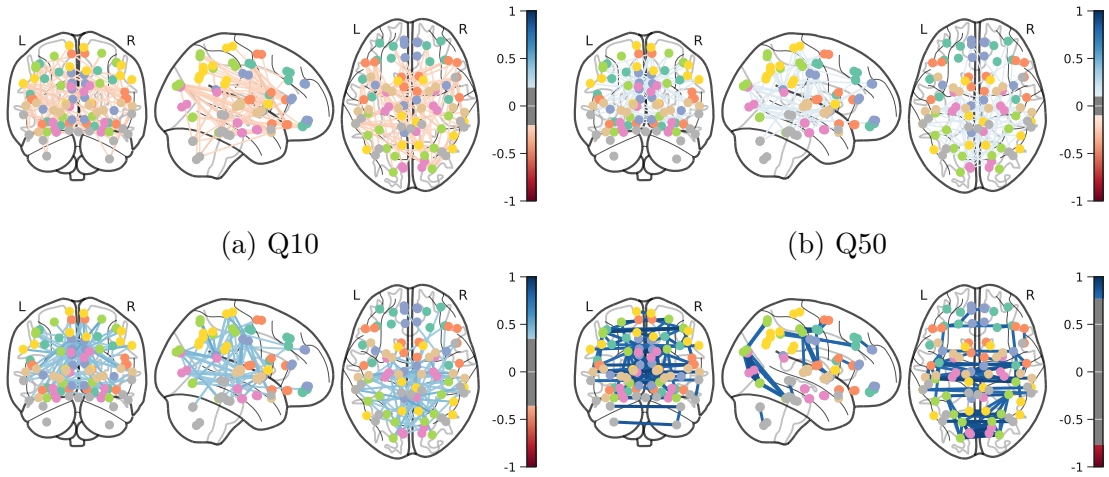
7.B Edge connectivity mean brain networks



(a) Q10 (b) Q50

(c) Q90 (d) CA

Fig. 7.B.1 Edge mean networks. Data from the 204 PIOP2 subjects. Brain edge mean networks for **(a-c)** different inter-correlation quantiles and **(d)** the CA estimate. Top 2% of edges with the highest connectivity. Color legend represents the average across subjects of the inter-correlation value associated to each edge.

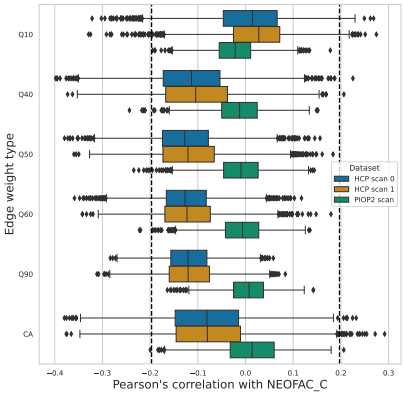


(a) Q10 (b) Q50

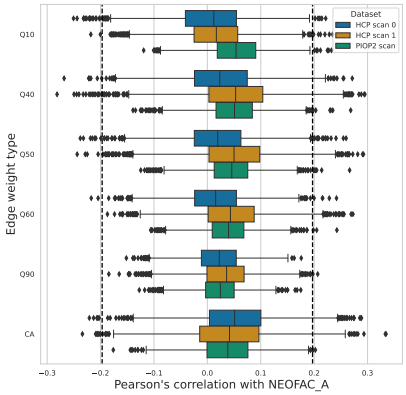
(c) Q90 (d) CA

Fig. 7.B.2 Edge mean networks. Data from the first scan of the 100 HCP subjects. Brain edge mean networks for **(a-c)** different inter-correlation quantiles and **(d)** the CA estimate. Top 2% of edges with the highest connectivity. Color legend represents the average across subjects of the inter-correlation value associated to each edge.

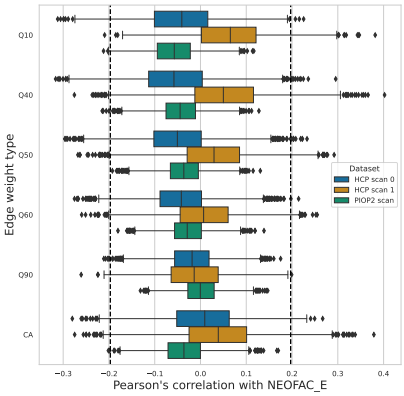
7.C Edge-level brain-wide association for additional scores



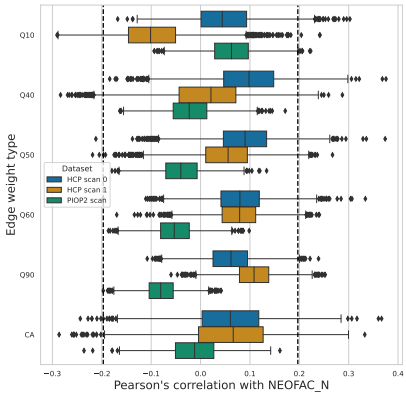
(a) NEO-FFI Conscientiousness



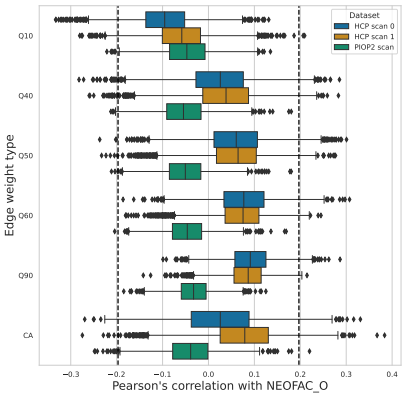
(b) NEO-FFI Agreeableness



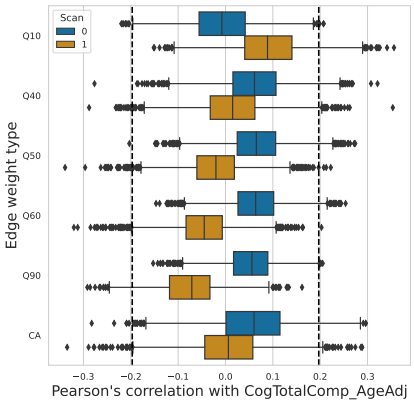
(c) NEO-FFI Extraversion



(d) NEO-FFI Neuroticism



(e) NEO-FFI Openness to Experience



(f) Cognitive function

Fig. 7.C.1 Distribution across edges of Pearson's correlation between edges weights and psychometric scores computed across all subjects for different quantiles of the inter-correlation distribution and the CA estimator.Distributions for both scans of the HCP dataset are depicted. The vertical dashed lines represent the critical correlation values corresponding to a 5% nominal level. Age-adjusted cognitive function composite scores are only available in the HCP dataset. Conscientiousness seems to be the score with the highest correlations with HCP edge connectivity.

Chapter 8

Conclusion and Perspectives

8.1 Summary

Manipulating multivariate grouped data is a daunting task, particularly when the objective is to learn, and evaluate, associated connectivity networks.

In Chapter 4, we introduced a novel binary network learning pipeline that leverages correlation screening and accounts for dependency structures. Once connectivity values are inferred, it is sometimes preferable to circumvent the thresholding step, and to namely consider weighted networks instead of their binary counterparts. Chapter 5 hence combined topological data analysis and regional label information to propose a multi-scale comparison of weighted connectivity networks.

However, intra-correlation and noise are known to negatively impact the estimation of connectivity network weights, that is, inter-correlation, consequently affecting downstream analyses. It is therefore crucial to use consistent inter-correlation estimators to infer connectivity networks. We then proposed in Chapter 6 a novel non-parametric estimator of the correlation between groups of measurements of a quantity that simultaneously tackles the presence of both intra-correlation and noise. We derived consistency results, and empirically established the superiority of our proposed estimator. Leveraging hierarchical clustering, our approach provides both point estimates and matching empirical distributions, which inherently quantify some uncertainties. In order to fully utilize this information, we introduced in Chapter 7 the concept of distribution-based weighted connectivity networks, where edge weights are inter-correlation distribution or density functions. We then proceeded to demonstrate its relevance and improved performance compared to a traditional estimator on real-world resting-state fMRI data in terms of repeatability and performance in machine learning tasks.

Distribution-based weighted networks unlock a vast array of possibilities and introduce the foundations for the definition of a powerful framework that could rigorously quantify and account for uncertainty and quality in connectivity networks.

8.2 Future Directions

We end this manuscript by detailing some possible future directions arising from this thesis.

8.2.1 Functional data analysis for connectivity networks

We introduced in Chapter 7 distribution-based weighted networks. These can be reframed as multivariate functional data, with a density function associated to each edge of the individual networks. This shift in perspective enables us to utilize tools from functional data analysis (FDA), such as functional principal component analysis (fPCA), or functional analysis of variance (fANOVA) and its multivariate counterpart (fMANOVA), for the purpose of exploring mean and variance effects.

To provide a more tangible perspective, we can consider an application to anesthetic effects on rat brains. We can recall for instance the rat brain fMRI dataset introduced in Chapter 3 where rats were administered one of four different anesthetics. In this context, we can construct brain functional connectivity networks, where each node is a brain region and each edge is assigned its corresponding inter-correlation density.

Mean effects: high-dimensional fMANOVA. Anesthetic mean effects could then be examined. Simultaneous comparison of several groups of subjects, i.e., anesthetic categories in our scenario, is typically performed using ANOVA. It emerges the latter has been previously successfully used to compare traditional scalar-weighted brain networks (Fraiman and Fraiman, 2018). This approach even manages to identify edges and nodes with differences, but only supports scalar weights. fANOVA, which can only handle one distribution-weighted edge at a time in our context, has also been extensively used in a wide range of applications, including spatio-temporal climate data, e.g., (Johny, 2021; Cuevas et al., 2004). Multivariate versions of fANOVA (fMANOVA) could hence be applied to compare anesthetics while simultaneously accounting for all inter-correlation densities of a given subject. The fdanova R package provides an implementation of functional multivariate ANOVA (fMANOVA) (Gorecki and Smaga, 2019, 2017). The

corresponding null hypothesis is then:

$$H_0 : \mathbf{m}_1 = \cdots = \mathbf{m}_k = \cdots = \mathbf{m}_4, \tag{8.1}$$

where $\mathbf{m}_k$ denotes the vector containing the mean inter-correlation density functions of all edges, computed across the anesthetic group k. After projecting the density functions into a Hilbert space (Petersen and Müller, 2016), as required per (Gorecki and Smaga, 2017), preliminary results on the rat datasets are promising and allow to differentiate anesthetic groups. However, we have too few subjects for those results to truly be reliable. We are indeed in a high-dimensional setting, with at most 6 rats per group and a large number of variables, that is, the number of edges, which are in the thousands. High dimensional approaches could offer a possible solution. Some high-dimensional MANOVA methods have already been developed (Lin et al., 2021) and could be extended to functional data. The approach presented by Lin et al. (2021) relies on maxima among variables, which would need to be adapted in a functional data setting, perhaps by leveraging quantiles of the inter-correlation densities.

Variance effects. Anesthetic variance effects could also be of interest. For instance, distribution-based connectivity networks could be decomposed into a network representing joint variations across all rats, networks representing anesthetic-specific perturbations, and individual networks representing rat-specific perturbations. To that end, low-rank models that could plausibly capture a latent part of the inter-correlation densities could be used. These would need to be chosen carefully. Preliminary explorations of the Joint and Individual Variation explained (JIVE) method (Lock et al., 2013) highlighted the challenges of the task. Indeed, while we were able to obtain joint and individual edge variation components across all rats, the JIVE framework proved too restrictive to recover anesthetic-specific variation components. Other approaches, such as fPCA could also be investigated. More work hence would need to be done and is out of the scope of this thesis.

8.2.2 Uncertain graphs

We are convinced providing theoretically sound and practical methods to model brain connectivity only partially meets expectations, and that quantifying their quality is necessary for physicians to adopt these tools.

Incidentally, the proposed distribution-based paradigm could also serve as a foundation for uncertainty quantification of connectivity networks. Inter-correlation distributions

already constitute in and of themselves a characterization of uncertainty, and could potentially be linked to the Bayesian literature, possibly by way of ensemble learning (Pearce et al., 2018). Caution is however warranted due to the presence of dependence between the correlation estimates used to derive individual inter-correlation densities.

A related and pertinent extension would be to consider *uncertain graphs*, e.g. (Hu et al., 2017; Liu et al., 2023), that is, graphs whose edge weights are equal to their own probability of existence. Hailing from the database community, and leveraging the possible world paradigm, they have been used in various undertakings, from graph embedding (Hu et al., 2017), to identifying protein complexes (Zhao et al., 2014). In the context of functional connectivity, such graphs have been introduced as *fuzzy networks* (Raimondo and De Domenico, 2021). In the setting where networks edges are associated with an inter-correlation distribution, the probability of existence of the edge between any regions A and B could potentially be defined as follows:

$$\pi_{AB} = P(\rho^{A,B} \geq \rho_q^{A,B}), \tag{8.2}$$

where $\rho_q^{A,B}$ is the inter-correlation threshold used to determine whether the edge between regions A and B is sufficiently highly correlated to be deemed to exist. It could be for instance equal to the quantile-based threshold introduced in Chapter 4. We could then propose to define the empirical probability of existence of the edge between any regions A and B as follows:

$$\widehat{\pi}_{AB} = P(|R^{A,B}| \geq \rho_q^{A,B}) = 1 - \widehat{F}_{|R^{A,B}|}(\rho_q^{A,B}). \tag{8.3}$$

Other definitions of the probability of existence could be alternatively be proposed. For instance, Raimondo and De Domenico (2021) leverage the hypothesis testing literature as well as the Bayes factor to derive it.

Once the probability of existence of edges are defined and computed, uncertain graphs could be exploited in a wide range of situations highly relevant to brain functional connectivity.

For instance, various graph theory metrics, such as node degree, the clustering coefficient or the closeness centrality have been extended to uncertain graphs (Raimondo and De Domenico, 2021; Liu et al., 2023). Since each edge is now assigned a probability of existence, the probability distribution of these graph metrics can be derived from individual networks. Graph metrics have recently been used to determine node equivalence classes via a nodal-statistics-based equivalence relation (Carboni et al., 2023). The latter

could be extended to the uncertain graph paradigm by taking into account node metric probability distributions when defining the equivalence classes.

Uncertainty could also be taken into account in subsequent analyses, such as in machine learning pipelines, possibly using uncertain graph embeddings (Hu et al., 2017).

Finally, uncertainty visualization could be more easily produced. This is all the more compelling in functional connectivity settings, where brain maps depicting graph metric information, such as node degrees, could be combined with connectivity uncertainty for each region. These representations could be specifically designed to help meet the needs of physicians, for example in terms of prognosis assessment.

8.2.3 Graph neural networks for distribution-based networks

We discuss in this section one last possible future direction. Graph neural networks (GNN) perform end-to-end learning, taking graphs as an input, and thus circumventing the graph embedding step. In this paradigm, nodes and edges can furthermore both have vector features. GNNs have recently been used to classify major depressive disorders using scalar-weighted functional connectivity networks (Gallo et al., 2023). However, edge features were not taken into account.

In our proposed distribution-based framework, density functions are attached to each edge and could easily be discretized to be appointed as input edge features. Node features could analogously be obtained from intra-correlation distributions. The latter could be derived from voxel-to-voxel intra-correlations in fMRI contexts.

A typical GNN model contains at least an input and output layer, and one or more pooling steps that aggregate node or edge features so that the output prediction is a scalar. Pooling function could be the mean, minimum, maximum, sum, or more sophisticated aggregation function (e.g., linear regression). A popular type of layers are message-passing layers, which take into account the input graph structure by aggregating neighboring node and edge information. One such layer is the Edge Graph Attention Network (EGAT) layer (Kamiński et al., 2021), which performs weighted aggregation based on the importance of node and edge neighbors and can handle vector edge features. The EGAT layer would hence be particularly appropriate for our setting. It is implemented in the dgl Python library and has recently been used to classify multi-channel electroencephalography data (Lin et al., 2023). We are expecting the use of multivariate edge features derived from inter-correlation densities to improve GNN performances over univariate, or even non-existent, edge features.

References

Achard, S. and Bullmore, E. (2007). Efficiency and cost of economical brain functional networks. *PLOS Computational Biology*, 3(2):1–10.

Achard, S., Coeurjolly, J., Marcillaud, R., and Richiardi, J. (2011). fMRI functional connectivity estimators robust to region size bias. *Proceedings of IEEE Workshop on Statistical Signal Processing (SSP)*, pages 813–816.

Achard, S., Coeurjolly, J.-F., de Micheaux, P. L., Lbath, H., and Richiardi, J. (2023). Inter-regional correlation estimators for functional magnetic resonance imaging. *NeuroImage*, 282:120388.

Achard, S., Coeurjolly, J.-F., Lafaye de Micheaux, P., and Richiardi, J. (2020). Robust correlation for aggregated data with spatial characteristics. *arXiv:2011.08269*.

Achard, S., Delon-Martin, C., Vértes, P., Renard, F., Schenck, M., Schneider, F., Heinrich, C., Kremer, S., and Bullmore, E. T. (2012a). Hubs of brain functional networks are radically reorganized in comatose patients. *Proc. Natl. Acad. Sci.*, 109(50):20608–20613.

Achard, S., Delon-Martin, C., Vértes, P. E., Renard, F., Schenck, M., Schneider, F., Heinrich, C., Kremer, S., and Bullmore, E. T. (2012b). Hubs of brain functional networks are radically reorganized in comatose patients. *Proceedings of the National Academy of Sciences*, 109(50):20608–20613.

Achard, S., Salvador, R., Whitcher, B., Suckling, J., and Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *The Journal of Neuroscience*, 26(1):63–72.

Afyouni, S., Smith, S. M., and Nichols, T. E. (2019). Effective degrees of freedom of the pearson's correlation coefficient under autocorrelation. *NeuroImage*, 199:609–625.

Alexander-Bloch, A. F., Gogtay, N., Meunier, D., Birn, R., Clasen, L., Lalonde, F., Lenroot, R., Giedd, J., and Bullmore, E. T. (2010). Disrupted Modularity and Local Connectivity of Brain Functional Networks in Childhood-Onset Schizophrenia. *Frontiers in Systems Neuroscience*, 4.

Andjelković, M., Tadić, B., and Melnik, R. (2020). The topology of higher-order complexes associated with brain hubs in human connectomes. *Scientific Reports*, 10(1).

Azriel, D. and Schwartzman, A. (2014). The empirical distribution of a large number of correlated normal variables. *Journal of the American Statistical Association*, 110:00–00.

Bardin, J.-B., Spreemann, G., and Hess, K. (2019). Topological exploration of artificial neuronal network dynamics. *Network Neuroscience*, 3(3):725–743.

Bassett, D. S. and Bullmore, E. (2006). Small-world brain networks. *The neuroscientist*, 12(6):512–523.

Battiston, F., Amico, E., Barrat, A., Bianconi, G., de Arruda, G. F., Franceschiello, B., Iacopini, I., Kéfi, S., Latora, V., Moreno, Y., Murray, M. M., Peixoto, T. P., Vaccarino, F., and Petri, G. (2021). The physics of higher-order interactions in complex systems. *Nature Physics*, 17(10):1093–1098.

Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.-G., and Petri, G. (2020). Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports*, 874:1–92.

Becq, G. J.-P. C., Barbier, E., and Achard, S. (2020a). Brain networks of rats under anesthesia using resting-state fmri: comparison with dead rats, random noise and generative models of networks. *Journal of Neural Engineering*, 17:045012.

Becq, G. J.-P. C., Habet, T., Collomb, N., Faucher, M., Delon-Martin, C., Coizet, V., Achard, S., and Barbier, E. L. (2020b). Functional connectivity is preserved but reorganized across several anesthetic regimes. *NeuroImage*, 219:116945.

Bergholm, F., Adler, J., and Parmryd, I. (2010). Analysis of bias in the apparent correlation coefficient between image pairs corrupted by severe noise. *J Math Imaging Vis*, 37:204–219.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, 7(4):401–406.

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820.

Bolt, T., Nomi, J. S., Rubinov, M., and Uddin, L. Q. (2017). Correspondence between evoked and intrinsic functional brain network configurations. *Human Brain Mapping*, 38(4):1992–2007.

Bordier, C., Nicolini, C., and Bifone, A. (2017). Graph analysis and modularity of brain functional connectivity networks: searching for the optimal threshold. *Frontiers in neuroscience*, 11:441.

Boschi, A., Brofiga, M., and Massobrio, P. (2021). Thresholding functional connectivity matrices to recover the topological properties of large-scale neuronal networks. *Frontiers in Neuroscience*, 15.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Brugere, I., Gallagher, B., and Berger-Wolf, T. Y. (2018). Network structure inference, a survey: Motivations, methods, and applications. *ACM Comput. Surv.*, 51(2).

Bruni, T., Graham, M., Norton, L., Gofton, T., Owen, A. M., and Weijer, C. (2019). Informed consent for functional mri research on comatose patients following severe brain injury: balancing the social benefits of research against patient autonomy. *Journal of medical ethics*, 45(5):299–303.

Cai, T. T. and Liu, W. (2016). Large-scale multiple testing of correlations. *Journal of the American Statistical Association*, 111(513):229–240. PMID: 27284211.

Calhoun, V. D., Sui, J., Kiehl, K., Turner, J. A., Allen, E. A., and Pearlson, G. (2012). Exploring the psychosis functional connectome: aberrant intrinsic networks in schizophrenia and bipolar disorder. *Frontiers in psychiatry*, 2:75.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Cameron, A. C. and Miller, D. (2011). Robust inference with clustered data. In Ullah, A. and Giles, D. E., editors, *Handbook of Empirical Economics and Finance*, pages 1–28. CRC Press.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *The Journal of Human Resources*, 50:317 – 372.

Caputi, L., Pidnebesna, A., and Hlinka, J. (2021). Promises and pitfalls of topological data analysis for brain connectivity analysis. *NeuroImage*, 238:118245.

Carboni, L., Dojat, M., and Achard, S. (2023). Nodal-statistics-based equivalence relation for graph collections. *Phys. Rev. E*, 107:014302.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Chavent, M., Kuentz-Simonet, V., Liquet, B., and Saracco, J. (2012). **ClustOfVar** : An *R* Package for the Clustering of Variables. *Journal of Statistical Software*, 50(13).

Chazal, F. and Michel, B. (2021). An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4.

Chung, M. K., Hanson, J. L., Ye, J., Davidson, R. J., and Pollak, S. D. (2015). Persistent homology in sparse regression and its application to brain morphometry. *IEEE Transactions on Medical Imaging*, 34(9):1928–1939.

Clifford, P., Richardson, S., and Hemon, D. (1989). Assessing the significance of the correlation between two spatial processes. *Biometrics*, 45(1):123–134.

Combrisson, E. and Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, 250:126–136. Cutting-edge EEG Methods.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Couto, C. M. V., Comin, C. H., and Costa, L. d. F. (2017). Effects of threshold on the topology of gene co-expression networks. *Mol. Biosyst.*, 13(10):2024–2035.

Croom, F. H. (1978). *Basic Concepts of Algebraic Topology*. Springer New York.

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational Statistics & Data Analysis*, 47(1):111–122.

Cui, Z. and Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *NeuroImage*, 178:622–637.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., Varoquaux, G., Initiative, A. D. N., et al. (2019). Benchmarking functional connectome-based predictive models for resting-state fmri. *NeuroImage*, 192:115–134.

Davey, C. E., Grayden, D. B., Egan, G. F., and Johnston, L. A. (2013). Filtering induces correlation in fmri resting state data. *NeuroImage*, 64:728–740.

De Vico Fallani, F., Richiardi, J., Chavez, M., and Achard, S. (2014). Graph analysis of functional brain networks: practical issues in translational neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1653):20130521.

Desloires, J., Ienco, D., and Botrel, A. (2023). Out-of-year corn yield prediction at field-scale using sentinel-2 satellite imagery and machine learning methods. *Computers and Electronics in Agriculture*, 209:107807.

Dhillon, I. S., Marcotte, E. M., and Roshan, U. (2003). Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, 19(13):1612–1619.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659.

Dong, X., Thanou, D., Rabbat, M., and Frossard, P. (2019). Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3):44–63.

Donner, A. and Eliasziw, M. (1991). Methodology for inferences concerning familial correlations: a review. *J Clin Epidemio*, 44(4/5):449–455.

Donner, A., Eliasziw, M., and Shoukri, M. M. (1998). Review of inference procedures for the interclass correlation coefficient with emphasis on applications to family studies. *Genetic Epidemiology*, 15.

Drton, M. and Perlman, M. D. (2007). Multiple Testing and Error Control in Gaussian Graphical Model Selection. *Statistical Science*, 22(3):430 – 449.

Dunlap, W., Jones, M., and Bittner, A. (1983). Average correlations vs. correlated averages. *Bulletin of the Psychonomic Society*, 21:213–216.

Elston, R. C. (1975). On the correlation between correlations. *Biometrika*, 62(1):133–140.

Erhardt, E. B., Allen, E. A., Wei, Y., Eichele, T., and Calhoun, V. D. (2012). Simtb, a simulation toolbox for fmri data under a model of spatiotemporal separability. *NeuroImage*, 59(4):4160–4167.

Essen, D. V., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S., Penna, S. D., Feinberg, D., Glasser, M., Harel, N., Heath, A., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S., Prior, F., Schlaggar, B., Smith, S., Snyder, A., Xu, J., and Yacoub, E. (2012). The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4):2222–2231.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226–231.

Fan, J. and Lv, J. (2008a). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Lv, J. (2008b). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fattahi, S. and Sojoudi, S. (2019). Graphical lasso and thresholding: Equivalence and closed-form solutions. *Journal of Machine Learning Research*, 20(10):1–44.

Figueroa-Jimenez, M. D., Cañete-Massé, C., Carbó-Carreté, M., Zarabozo-Hurtado, D., Peró-Cebollero, M., Salazar-Estrada, J. G., and Guàrdia-Olmos, J. (2021). Resting-state default mode network connectivity in young individuals with Down syndrome. *Brain and Behavior*, 11(1):e01905.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Fraiman, D. and Fraiman, R. (2018). An anova approach for statistical comparisons of brain networks. *Scientific Reports*, 8.

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

Gallo, S., El-Gazzar, A., Zhutovsky, P., Thomas, R. M., Javaheripour, N., Li, M., Bartova, L., Bathula, D., Dannlowski, U., Davey, C., Frodl, T., Gotlib, I., Grimm, S., Grotegerd, D., Hahn, T., Hamilton, P. J., Harrison, B. J., Jansen, A., Kircher, T., Meyer, B., Nenadić, I., Olbrich, S., Paul, E., Pezawas, L., Sacchet, M. D., Sämann, P., Wagner, G., Walter, H., Walter, M., PsyMRI, and van Wingen, G. (2023). Functional connectivity signatures of major depressive disorder: machine learning analysis of two multicenter neuroimaging studies. *Mol. Psychiatry*.

Gatica, M., Cofré, R., Mediano, P. A., Rosas, F. E., Orio, P., Diez, I., Swinnen, S. P., and Cortes, J. M. (2021). High-order interdependencies in the aging brain. *Brain Connectivity*, 11(9):734–744.

Gatica, M., Rosas, F. E., Mediano, P. A. M., Diez, I., Swinnen, S. P., Orio, P., Cofré, R., and Cortes, J. M. (2022). High-order functional redundancy in ageing explained via alterations in the connectome in a whole-brain model. *PLOS Computational Biology*, 18(9):e1010431.

Girn, M., Spreng, R. N., Margulies, D. S., Van Elk, M., and Lifshitz, M. (2023). Trait absorption is not reliably associated with brain structure or resting-state functional connectivity. *Neuroimage: Reports*, 3(2):100171.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., and Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124.

Goeman, J. J., Meijer, R. J., Krebs, T. J. P., and Solari, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106(4):841–856.

Gorecki, T. and Smaga, L. (2017). Multivariate analysis of variance for functional data. *Journal of Applied Statistics*, 44(12):2172–2189.

Gorecki, T. and Smaga, L. (2019). fdanova: an r software package for analysis of variance for univariate and multivariate functional data. *Computational Statistics*, 34.

Grover, A. and Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864, New York, NY, USA. Association for Computing Machinery.

Gunst, R. F. (1995). Estimating spatial correlations from spatial-temporal meteorological data. *Journal of Climate*, 8(10):2454 – 2470.

Hall, C. N., Howarth, C., Kurth-Nelson, Z., and Mishra, A. (2016). Interpreting bold: towards a dialogue between cognitive and cellular neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1705):20150348.

Hallett, M., de Haan, W., Deco, G., Dengler, R., Di Iorio, R., Gallea, C., Gerloff, C., Grefkes, C., Helmich, R. C., Kringelbach, M. L., et al. (2020). Human brain connectivity: Clinical applications for clinical neurophysiology. *Clinical Neurophysiology*, 131(7):1621–1651.

Halliwell, J. W. (1962). Dangers inherent in correlating averages. *The Journal of Educational Research*, 55(7):327–329.

Harley, B. I. (1957). Relation Between the Distributions of Non-central t and of a Transformed Correlation Coefficient. *Biometrika*, 44(1-2):219–224.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hero, A. and Rajaratnam, B. (2011). Large scale correlation screening. *Journal of the American Statistical Association*, 106:1540–1552.

Herzog, R., Rosas, F. E., Whelan, R., Fittipaldi, S., Santamaria-Garcia, H., Cruzat, J., Birba, A., Moguilner, S., Tagliazucchi, E., Prado, P., and Ibanez, A. (2022). Genuine high-order interactions in brain networks and neurodegeneration. *Neurobiology of Disease*, 175:105918.

Hlinka, J., Paluš, M., Vejmelka, M., Mantini, D., and Corbetta, M. (2011). Functional connectivity in resting-state fmri: Is linear correlation sufficient? *NeuroImage*, 54(3):2218–2225.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):193–232.

Hu, J., Cheng, R., Huang, Z., Fang, Y., and Luo, S. (2017). *On Embedding Uncertain Graphs*, page 157–166. Association for Computing Machinery, New York, NY, USA.

Hu, X., Jung, A., and Qin, G. (2020). Interval estimation for the correlation coefficient. *The American Statistician*, 74(1):29–36.

Johny, M. M. (2021). *Functional ANOVA-Type Methods with Interpretable Visualization for Comparisons among Groups of Time Series*. PhD thesis, Iowa State University.

Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivières, S., Grigis, A., Martinot, J.-L., Paus, T., Smolka, M. N., Walter, H., Schumann, G., Garavan, H., and Whelan, R. (2019). Quantifying performance of machine learning methods for neuroimaging data. *NeuroImage*, 199:351–365.

Kamiński, K., Ludwiczak, J., Jasiński, M., Bukala, A., Madaj, R., Szczepaniak, K., and Dunin-Horkawicz, S. (2021). Rossmann-toolbox: a deep learning-based protocol for the prediction and design of cofactor specificity in Rossmann fold proteins. *Briefings in Bioinformatics*, 23(1). bbab371.

Kartun-Giles, A. P. and Bianconi, G. (2019). Beyond the clustering coefficient: A topological analysis of node neighbourhoods in complex networks. *Chaos, Solitons & Fractals: X*, 1:100004.

Kassraian-Fard, P., Matthis, C., Balsters, J. H., Maathuis, M. H., and Wenderoth, N. (2016). Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Frontiers in Psychiatry*, 7.

Kaufman, L. and Rousseeuw, P. J. (2005). *Finding groups in data: an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Wiley.

Kesler, S. R., Rao, A., Blayney, D. W., Oakley-Girvan, I. A., Karuturi, M., and Palesh, O. (2017). Predicting long-term cognitive outcome following breast cancer with pre-treatment resting state fmri and random forest machine learning. *Frontiers in Human Neuroscience*, 11.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Konishi, S. (1982). Asymptotic properties of estimators of interclass correlation from familial data. *Annals of the Institute of Statistical Mathematics*, 34(3):505–515.

Lbath, H., Petersen, A., and Achard, S. (2021). Brain functional connectivity estimation. In *Brain Connectivity Networks: Quality and Reproducibility - Satellite of the Conference on Complex Systems 2021*, Lyon, France.

Lbath, H., Petersen, A., and Achard, S. (2022a). Large-scale correlation screening under dependence for brain functional connectivity inference. In *JSM 2022 - Joint Statistical Meetings*, Washington, United States.

Lbath, H., Petersen, A., Meiring, W., and Achard, S. (2022b). Clustering-based inter-group correlation estimation. In *ICSDS 2022 - IMS International Conference on Statistics and Data Science*, Florence, Italy.

Lbath, H., Petersen, A., Meiring, W., and Achard, S. (2023). Clustering-based inter-regional correlation estimation. *Computational Statistics & Data Analysis*, page 107876.

Lbath, H., Richiardi, J., Petersen, A., Meiring, W., and Achard, S. (working paper). Distribution-based weighted networks validation on rs-fMRI data.

Lee, H., Kang, H., Chung, M. K., Kim, B.-N., and Lee, D. S. (2012). Persistent brain network homology from the perspective of dendrogram. *IEEE Transactions on Medical Imaging*, 31(12):2267–2277.

Lehmann, E. L. (1966). Some Concepts of Dependence. *The Annals of Mathematical Statistics*, 37(5):1137 – 1153.

Lin, L. I.-K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1):255–268. Publisher: [Wiley, International Biometric Society].

Lin, X., Chen, J., Ma, W., Tang, W., and Wang, Y. (2023). Eeg emotion recognition using improved graph neural network with channel selection. *Computer Methods and Programs in Biomedicine*, 231:107380.

Lin, Z., Lopes, M. E., and Müller, H.-G. (2021). High-dimensional manova via bootstrapping and its application to functional and sparse count data. *Journal of the American Statistical Association*, 0(0):1–15.

Lindskog, F. (2000). Linear correlation estimation. *Risklab Research Paper, ETH-Zentrum, Zürich*.

Liu, P., Calhoun, V., and Chen, Z. (2017a). Functional overestimation due to spatial smoothing of fmri data. *Journal of Neuroscience Methods*, 291:1–12.

Liu, T. T., Nalci, A., and Falahpour, M. (2017b). The global signal in fMRI: Nuisance or information? *NeuroImage*, 150:213–229.

Liu, Z., Ye, J., and Zou, Z. (2023). Closeness centrality on uncertain graphs. *ACM Trans. Web*. Just Accepted.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523 – 542.

Malagurski, B., Péran, P., Sarton, B., Vinour, H., Naboulsi, E., Riu, B., Bounes, F., Seguin, T., Lotterie, J. A., Fourcade, O., Minville, V., Ferré, F., Achard, S., and Silva, S. (2019). Topological disintegration of resting state functional connectomes in coma. *NeuroImage*, 195:354–361.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., Moore, L. A., Conan, G. M., Uriarte, J., Snider, K., Lynch, B. J., Wilgenbusch, J. C., Pengo, T., Tam, A., Chen, J., Newbold, D. J., Zheng, A., Seider, N. A., Van, A. N., Metoki, A., Chauvin, R. J., Laumann, T. O., Greene, D. J., Petersen, S. E., Garavan, H., Thompson, W. K., Nichols, T. E., Yeo, B. T. T., Barch, D. M., Luna, B., Fair, D. A., and Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660.

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., and Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychology*, 3(1). 25.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125 – 2149.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Mheich, A., Wendling, F., and Hassan, M. (2020). Brain network similarity: methods and applications. *Network Neuroscience*, 4(3):507–527.

Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory*. Wiley-Interscience.

Murphy, K. and Fox, M. D. (2017). Towards a consensus regarding global signal regression for resting state functional connectivity mri. *NeuroImage*, 154:169–173. Cleaning up the fMRI time series: Mitigating noise with advanced acquisition and correction strategies.

Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3):274–295.

Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., and Jaiswal, S. (2017). graph2vec: Learning distributed representations of graphs. *CoRR*, abs/1707.05005.

Nyholt, D. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American journal of human genetics*, 74:765–9.

Ogasawara, H. (2006). Asymptotic expansion of the sample correlation coefficient under nonnormality. *Computational Statistics & Data Analysis*, 50(4):891–910.

Ogawa, A. (2021). Time-varying measures of cerebral network centrality correlate with visual saliency during movie watching. *Brain and Behavior*, 11(9):e2334.

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology*, 78:569–582.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431.

Pearce, T., Leibfried, F., Brintrup, A., Zaki, M. H., and Neely, A. D. (2018). Uncertainty in neural networks: Approximately bayesian ensembling. In *International Conference on Artificial Intelligence and Statistics*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perisic, I. and Rosner, B. (1999). Comparisons of measures of interclass correlations: the general case of unequal group size. *Statistics in Medicine*, 18(12):1451–1466.

Petersen, A. and Müller, H.-G. (2016). Functional data analysis for density functions by transformation to a hilbert space. *The Annals of Statistics*, 44(1):183–218.

Petersen, A. and Müller, H.-G. (2019). Wasserstein covariance for multiple random densities. *Biometrika*, 106(2):339–351.

Petri, G., Expert, P., Turkheimer, F., Carhart-Harris, R., Nutt, D., Hellyer, P. J., and Vaccarino, F. (2014). Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface*, 11(101):20140873.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Poli, D., Pastore, V. P., and Massobrio, P. (2015). Functional connectivity in in vitro neuronal assemblies. *Frontiers in Neural Circuits*, 9.

Qiu, D. and Ahn, J. (2020). Grouped variable screening for ultra-high dimensional data for linear model. *Computational Statistics & Data Analysis*, 144:106894.

Raimondo, S. and De Domenico, M. (2021). Measuring topological descriptors of complex networks under uncertainty. *Phys. Rev. E*, 103:022311.

Reimann, M. W., Nolte, M., Scolamiero, M., Turner, K., Perin, R., Chindemi, G., Dłotko, P., Levi, R., Hess, K., and Markram, H. (2017). Cliques of neurons bound into cavities provide a missing link between structure and function. *Frontiers in Computational Neuroscience*, 11.

Ribeiro, P. J. and Diggle, P. J. (2001). geoR: a package for geostatistical analysis. *R-NEWS*, 1(2):14–18.

Richiardi, J., Achard, S., Bunke, H., and Van De Ville, D. (2013). Machine learning with brain graphs: Predictive modeling approaches for functional imaging in systems neuroscience. *IEEE Signal Process. Mag.*, 30(3):58–70.

Ritter, P., Schirner, M., McIntosh, A. R., and Jirsa, V. K. (2013). The virtual brain integrates computational modeling and multimodal neuroimaging. *Brain Connectivity*, 3(2):121–145. PMID: 23442172.

Rosas, F. E., Mediano, P. A. M., Gastpar, M., and Jensen, H. J. (2019). Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100(3).

Rosas, F. E., Mediano, P. A. M., Luppi, A. I., Varley, T. F., Lizier, J. T., Stramaglia, S., Jensen, H. J., and Marinazzo, D. (2022). Disentangling high-order mechanisms and high-order behaviours in complex systems. *Nature Physics*.

Rosner, B., Donner, A., and Hennekens, C. H. (1977). Estimation of interclass correlation from familial data. *Applied Statistics*, 26:179–187.

Rosner, B., Donner, A., and Hennekens, C. H. (1979). Significance testing of interclass correlations from familial data. *Biometrics*, 35(2):461–471.

Rozemberczki, B. and Sarkar, R. (2020). Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1325–1334, New York, NY, USA. Association for Computing Machinery.

Ruben, H. (1966). Some new results on the distribution of the sample correlation coefficient. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(3):513–525.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069. Computational Models of the Brain.

Saccenti, E., Hendriks, M. M. W. B., and Smilde, A. K. (2020). Corruption of the pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models. *Scientific Reports*, 10:438.

Sanz Leon, P., Knock, S., Woodman, M., Domide, L., Mersmann, J., McIntosh, A., and Jirsa, V. (2013). The virtual brain: a simulator of primate brain network dynamics. *Frontiers in Neuroinformatics*, 7.

Sarica, A., Cerasa, A., and Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: A systematic review. *Frontiers in Aging Neuroscience*, 9.

Schirner, M., Domide, L., Perdikis, D., Triebkorn, P., Stefanovski, L., Pai, R., Prodan, P., Valean, B., Palmer, J., Langford, C., Blickensdörfer, A., van der Vlag, M., Diaz-Pier, S., Peyser, A., Klijn, W., Pleiter, D., Nahm, A., Schmid, O., Woodman, M., Zehl, L., Fousek, J., Petkoski, S., Kusch, L., Hashemi, M., Marinazzo, D., Mangin, J.-F., Flöel, A., Akintoye, S., Stahl, B. C., Cepic, M., Johnson, E., Deco, G., McIntosh, A. R., Hilgetag, C. C., Morgan, M., Schuller, B., Upton, A., McMurtrie, C., Dickscheid, T., Bjaalie, J. G., Amunts, K., Mersmann, J., Jirsa, V., and Ritter, P. (2022). Brain simulation as a cloud service: The virtual brain on ebrains. *NeuroImage*, 251:118973.

Schnack, H. G. and Kahn, R. S. (2016). Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Frontiers in Psychiatry*, 7.

Shevlyakov, G. and Smirnov, P. (2016). Robust estimation of the correlation coefficient: An attempt of survey. *Austrian Journal of Statistics*, 40:147–156.

Shinn, M., Hu, A., Turner, L., Noble, S., Preller, K. H., Ji, J. L., Moujaes, F., Achard, S., Scheinost, D., Constable, R. T., Krystal, J. H., Vollenweider, F. X., Lee, D., Anticevic, A., Bullmore, E. T., and Murray, J. D. (2023). Functional brain networks reflect spatial and temporal autocorrelation. *Nat Neurosci*, 26(5):867–878.

Sitoleux, P., Carboni, L., Lbath, H., and Achard, S. (in preparation 2023). Multiscale and multi-density comparison of functional brain networks through label-informed persistence diagrams.

Sizemore, A. E., Phillips-Cremins, J. E., Ghrist, R., and Bassett, D. S. (2019). The importance of the whole: Topological data analysis for the network neuroscientist. *Network Neuroscience*, 3(3):656–673.

Snoek, L., van der Miesen, M., Beemsterboer, T., Leij, A., Eigenhuis, A., and Scholte, H. (2021). The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. *Scientific Data*, 8:85.

Soehner, A. M., Chase, H. W., Bertocci, M. A., Greenberg, T., Stiffler, R., Lockovich, J. C., Aslam, H. A., Graur, S., Bebko, G., and Phillips, M. L. (2019). Unstable wakefulness during resting-state fmri and its associations with network connectivity and affective psychopathology in young adults. *Journal of Affective Disorders*, 258:125–132.

Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology, 1904-1920*, 5(4):417–426.

Sporns, O. (2022). Graph theory methods: applications in brain networks. *Dialogues in clinical neuroscience*.

Srivastava, M. S. and Keen, K. J. (1988). Estimation of the interclass correlation coefficient. *Biometrika*, 75(4):731–739.

Stanley, M., Moussa, M., Paolini, B., Lyday, R., Burdette, J., and Laurienti, P. (2013). Defining nodes in complex brain networks. *Frontiers in computational neuroscience*, 7:169.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87:245–251.

Tantardini, M., Ieva, F., Tajoli, L., and Piccardi, C. (2019). Comparing methods for comparing networks. *Scientific Reports*, 9(1).

Tauzin, G., Lupo, U., Tunstall, L., Pérez, J. B., Caorsi, M., Medina-Mardones, A., Dassatti, A., and Hess, K. (2020). giotto-tda: A topological data analysis toolkit for machine learning and data exploration.

Termenon, M., Jaillard, A., Delon-Martin, C., and Achard, S. (2016). Reliability of graph analysis of resting state fMRI using test-retest dataset from the human connectome project. *NeuroImage*, 142:172–187.

Theis, N., Rubin, J., Cape, J., Iyengar, S., Gur, R. E., Gur, R. C., Roalf, D. R., Pogue-Geile, M. F., Almasy, L., Nimgaonkar, V. L., et al. (2021). Evaluating network threshold selection for structural and functional brain connectomes. *bioRxiv*, pages 2021–10.

Thiele, J. A., Faskowitz, J., Sporns, O., and Hilger, K. (2022). Multitask brain network reconfiguration is inversely associated with human intelligence. *Cerebral Cortex*, 32(19):4172–4182.

Thompson, S. B. (2011). Simple formulas for standard errors that cluster by both firm and time. *Journal of Financial Economics*, 99(1):1–10.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Titouan, V., Courty, N., Tavenard, R., Laetitia, C., and Flamary, R. (2019). Optimal transport for structured data with application on graphs. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6275–6284. PMLR.

Torres, L., Blevins, A. S., Bassett, D., and Eliassi-Rad, T. (2021). The why, how, and when of representations for complex systems. *SIAM Review*, 63(3):435–485.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1):273–289.

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS One*, 14(11):e0224365.

van den Heuvel, M. P., de Lange, S. C., Zalesky, A., Seguin, C., Yeo, B. T., and Schmidt, R. (2017). Proportional thresholding in resting-state fmri functional connectivity networks and consequences for patient-control connectome studies: Issues and recommendations. *NeuroImage*, 152:437–449.

van den Heuvel, M. P., Stam, C. J., Kahn, R. S., and Pol, H. E. H. (2009). Efficiency of functional brain networks and intellectual performance. *Journal of Neuroscience*, 29(23):7619–7624.

Varley, T. F., Denny, V., Sporns, O., and Patania, A. (2021). Topological analysis of differential effects of ketamine and propofol anaesthesia on brain dynamics. *Royal Society Open Science*, 8(6):201971.

Váša, F. and Mišić, B. (2022). Null models in network neuroscience. *Nature Reviews Neuroscience*, 23(8):493–504.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2020). Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212.

Vigneau, E., Chen, M., and Qannari, E. M. (2015). ClustVarLV: An R Package for the Clustering of Variables Around Latent Variables. *The R Journal*, 7(2):134.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Váša, F., Bullmore, E. T., and Patel, A. X. (2018). Probabilistic thresholding of functional connectomes: Application to schizophrenia. *NeuroImage*, 172:326–340.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Welvaert, M., Durnez, J., Moerkerke, B., Berdoolaege, G., and Rosseel, Y. (2011). neurosim: An r package for generating fmri data. *Journal of Statistical Software*, 44(10):1–18.

Whitcher, B., Guttorp, P., and Percival, D. (2000a). Wavelet analysis of covariance with application to atmospheric time series. *Journal of Geophysical Research. D, Atmospheres*, 105(D11):14941–14962.

Whitcher, B., Guttorp, P., and Percival, D. B. (2000b). Wavelet analysis of covariance with application to atmospheric time series. *Journal of Geophysical Research*, 105(D11)(14):941–962.

White, T., Blok, E., and Calhoun, V. D. (2020). Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed. *Human Brain Mapping*, 43(1):278–291.

Whitfield-Gabrieli, S. and Nieto-Castanon, A. (2012). Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectivity*, 2(3):125–141. PMID: 22642651.

Wigley, T. M. L., Briffa, K. R., and Jones, P. D. (1984). On the average value of correlated time series, with applications in dendroclimatology and hydrometeorology. *Journal of Applied Meteorology and Climatology*, 23(2):201–213.

Wilson, C. (2010). *A study of relationships between family members using familial correlations*. PhD thesis, Old Dominion University Libraries.

Xifra-Porxas, A., Kassinopoulos, M., and Mitsis, G. D. (2021). Physiological and motion signatures in static and time-varying functional connectivity and their subject identifiability. *eLife*, 10:e62324.

Zalesky, A., Fornito, A., and Bullmore, E. (2012). On the use of correlation as a measure of network connectivity. *NeuroImage*, 60(4):2096–2106.

Zanin, M., Sousa, P., Papo, D., Bajo, R., García-Prieto, J., del Pozo, F., Menasalvas, E., and Boccaletti, S. (2012). Optimizing functional network representation of multivariate time series. *Sci. Rep.*, 2(1):630.

Zhang, C., Cahill, N., Arbabshirani, M., White, T., Baum, S., and Michael, A. (2016). Sex and age effects of functional connectivity in early adulthood. *Brain Connectivity*, 6:700–713.

Zhao, B., Wang, J., Li, M., Wu, F.-X., and Pan, Y. (2014). Detecting protein complexes based on uncertain graph model. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 11(3):486–497.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.