Clustering-Based Inter-Regional Correlation Estimation

Hanâ Lbath, Alexander Petersen, Wendy Meiring, Sophie Achard

Clustering-Based Inter-Regional Correlation Estimation

Hanâ Lbath[a,*], Alexander Petersen[b], Wendy Meiring[c], Sophie Achard[a]

[a]Univ. Grenoble Alpes, CNRS, Inria, Grenoble-INP, LJK,
38000 Grenoble, France

[b]Department of Statistics, Brigham Young University,
Provo, UT 84602, USA

[c]Department of Statistics and Applied Probability,
University of California Santa Barbara, Santa Barbara, CA 93106, USA

November 3, 2023

Abstract

A novel non-parametric estimator of the correlation between grouped measurements of a quantity is proposed in the presence of noise. This work is primarily motivated by functional brain network construction from fMRI data, where brain regions correspond to groups of spatial units, and correlation between region pairs defines the network. 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 a trade-off between two procedures: correlating regional averages, which is not robust to intra-regional correlation; and averaging pairwise inter-regional correlations, which is not robust to noise. To that end, we project the data onto a space where Euclidean distance is used as a proxy for sample correlation. We then 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.

Keywords: correlation estimation, hierarchical clustering, Ward's linkage, spatio-temporal data, brain functional connectivity

*Corresponding author: hana.lbath@inria.fr

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 diffi-

cult, 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 prove our estimator is consistent 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.

2 Related Work

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., 2017).

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.

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 paper 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

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

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 (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.

4 Proposed Inter-Correlation Estimator

After defining the sample correlation coefficient in Section 4.1, we highlight in Section 4.2 the impact of the combined presence of noise and intra-correlation, when using popular estimators of inter-correlation. In Section 4.3 we then propose an inter-correlation estimator that limits these effects. Consistency of our estimator is proved in Section 4.4.

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

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

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

$$r_{i,j}^{A,B} = \widehat{Cor}(\mathbf{Y}_i^A, \mathbf{Y}_j^B). \quad (3)$$

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

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 (1) and Achard et al. (2020), $r_{i,j}^{A,B}$ converges almost surely to

$$\frac{Cov(Y_i^A, Y_j^B)}{\sqrt{(\sigma_A^2 + \gamma_A^2) \cdot (\sigma_B^2 + \gamma_B^2)}} = \frac{Cov(X_i^A, X_j^B)}{\sqrt{(\sigma_A^2 + \gamma_A^2) \cdot (\sigma_B^2 + \gamma_B^2)}}. \tag{4}$$

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 (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 $r_{A,B}^{AC}$, it corresponds to the ensemble estimator in familial data literature (Rosner et al., 1977):

$$r_{A,B}^{AC} = \frac{1}{N_A \cdot N_B} \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} r_{i,j}^{A,B}. \tag{5}$$

However, the latter is similarly impacted by noise.

As mentioned in Section 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 $\overline{\mathbf{Y}}^A = N_A^{-1} \sum_{i=1}^{N_A} \mathbf{Y}_i^A$ and $\overline{\mathbf{Y}}^B = N_B^{-1} \sum_{j=1}^{N_B} \mathbf{Y}_j^B$. Then this estimator is

$$r_{A,B}^{CA} = \widehat{Cor}(\,\overline{\mathbf{Y}}^A, \overline{\mathbf{Y}}^B\,). \tag{6}$$

Under model (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, $r_{A,B}^{CA}$ converges almost surely to:

$$\frac{\rho^{A,B}}{\sqrt{\left[\frac{1}{N_A^2} \cdot \sum_{i,i'=1}^{N_A} \eta_{i,i'}^A + \frac{\gamma_A^2}{N_A \cdot \sigma_A^2}\right] \left[\frac{1}{N_B^2} \cdot \sum_{j,j'=1}^{N_B} \eta_{j,j'}^B + \frac{\gamma_B^2}{N_B \cdot \sigma_B^2}\right]}}, \tag{7}$$

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

It follows from (7) that intra-correlation and noise both contribute to inconsistency of the inter-correlation estimator (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 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.

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).

4.3.1 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.

4.3.2 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. A comparison of different clustering methods, which empirically validates the use of Ward's linkage in our context, is presented in Section 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 that 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 1). Proofs can be found in the appendix.

Theorem 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:*

$$1 - \frac{h_A^2}{2} \;\leq\; \frac{1}{|\nu_A|^2} \sum_{i,i'=1}^{|\nu_A|} r_{i,i'}^{A,A} \;\leq 1, \tag{8}$$

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

Theorem 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 1 and the fact that $1 - (h_A^{\max})^2/2 \;= \min_{i,i'=1,\ldots,N_A} r_{i,i'}^{A,A}$:

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

$$1 - \frac{h_A^2}{2} \;\leq\; \min_{i,i'=1,\ldots,N_A} r_{i,i'}^{A,A} \;\leq\; \frac{1}{|\nu_A|^2} \sum_{i,i'=1}^{|\nu_A|} r_{i,i'}^{A,A} \tag{9}$$

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

$$\min_{i,i'=1,\ldots,N_A} r_{i,i'}^{A,A} \;\leq\; 1 - \frac{h_A^2}{2} \;\leq\; \frac{1}{|\nu_A|^2} \sum_{i,i'=1}^{|\nu_A|} r_{i,i'}^{A,A}. \tag{10}$$

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 explorations of an optimal choice are presented in Section 5.2, and demonstrate the practical effectiveness of setting the cut-off height to $h_A^{\max}$.

4.3.3 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:

$$r_{\nu_A,\nu_B}^{CLA} = \widehat{Cor}(\ \overline{\mathbf{Y}}^{\nu_A},\ \overline{\mathbf{Y}}^{\nu_B}\ ), \tag{11}$$

where $\overline{\mathbf{Y}}^{\nu_A} = |\nu_A|^{-1} \sum_{i \in \nu_A} \mathbf{Y}_i^A$, 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 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:

$$r_{A,B}^{CLA} = \frac{1}{N_A^{clust} \cdot N_B^{clust}} \sum_{\nu_A,\nu_B} r_{\nu_A,\nu_B}^{CLA}, \tag{12}$$

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.

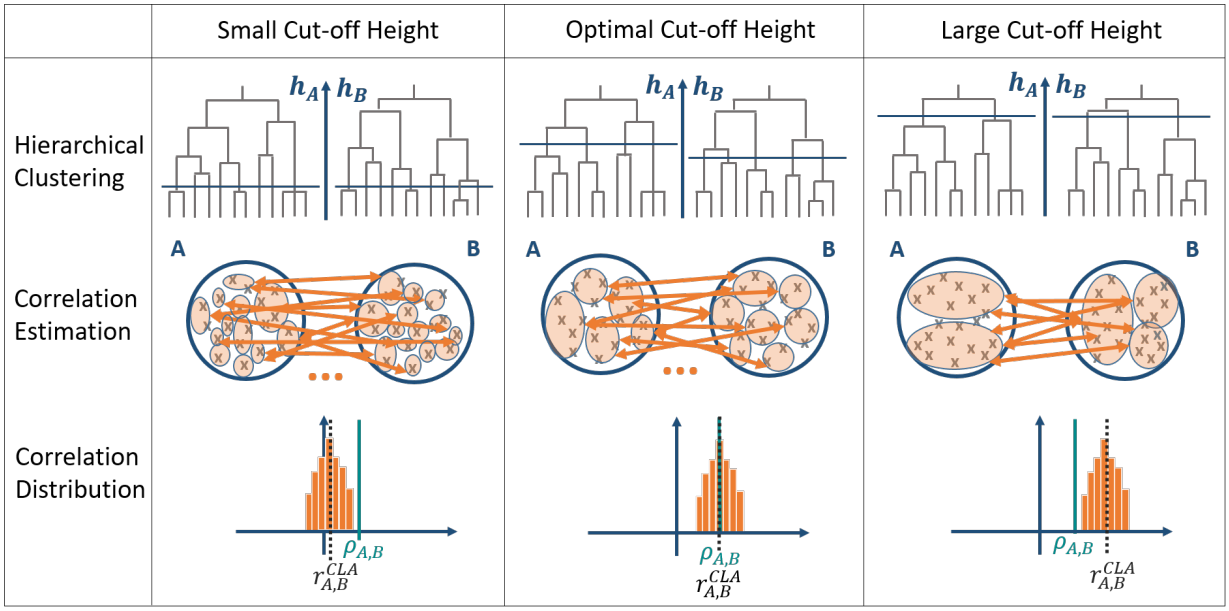


Figure 1: Illustration of the inter-correlation estimation of a pair of regions for different cut-off heights. The top panel shows the dendrograms of the hierarchical clustering applied to each region. The horizontal line over each dendrogram indicates the cut-off heights h_A, h_B. The grey crosses in 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 $r_{A,B}^{CLA}$ (dotted line) when the cut-off heights are neither too small nor too large.

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

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**

    $r_{\nu_A,\nu_B}^{CLA} \leftarrow \widehat{Cor}(\overline{\mathbf{Y}}^{\nu_A}, \overline{\mathbf{Y}}^{\nu_B})$

  $r_{A,B}^{CLA} \leftarrow \sum_{\nu_A,\nu_B} r_{\nu_A,\nu_B}^{CLA} / N_A^{clust} \cdot N_B^{clust}$

subsequent results:

$$\rho_{\nu_A,\nu_B}^{CLA} = \frac{\rho^{A,B}}{\sqrt{\left[\frac{1}{|\nu_A|^2} \cdot \sum_{i,i'=1}^{|\nu_A|} \eta_{i,i'}^A + \frac{\gamma_A^2}{|\nu_A| \cdot \sigma_A^2}\right] \cdot \left[\frac{1}{|\nu_B|^2} \cdot \sum_{j,j'=1}^{|\nu_B|} \eta_{j,j'}^B + \frac{\gamma_B^2}{|\nu_B| \cdot \sigma_B^2}\right]}}. \tag{13}$$

Theorem 2 *Under the assumptions of model* (1)*, for a fixed pair of clusters* ν_A, ν_B*, as* n *tends towards infinity,*

$$r_{\nu_A,\nu_B}^{CLA} \xrightarrow{a.s.} \rho_{\nu_A,\nu_B}^{CLA}. \tag{14}$$

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

Corollary 1 *Under the same assumptions as Theorem 2, for two regions* A, B*, as* n *tends*

towards infinity,

$$r_{A,B}^{CLA} \overset{a.s.}{\to} \frac{1}{N_{clust}^{A} N_{clust}^{B}} \sum_{\nu_A,\nu_B} \rho_{\nu_A,\nu_B}^{CLA}. \tag{15}$$

Corollary 1 is a direct consequence of Theorem 2.

Theorem 2 and Corollary 1 emphasize the fact that controlling the denominator of ρ_{ν_A,ν_B}^{CLA} 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_{ν_A,ν_B}^{CLA}, and the corresponding point estimate $r_{A,B}^{CLA}$ are equal to the correlation of averages $r_{A,B}^{CA}$ 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 1). There are also gradually less variables within each cluster, until they eventually contain only a single variable. It follows that when $h_A, h_B = 0$, $r_{A,B}^{CLA}$ corresponds to a correlation estimate with no aggregation $r_{A,B}^{AC}$. This can be visualized in Figure 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). In such cases, both r_{ν_A,ν_B}^{CLA} and $r_{A,B}^{CLA}$ are consistent estimators of the population inter-correlation.

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.

5.1 Datasets

We first present the different datasets used in this paper.

5.1.1 Real-World Datasets

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 paper we consider the following anesthetics: Etomidate (EtoL), Isoflurane (IsoW) and Urethane (UreL). The dataset is freely available at `https://dx.doi.org/10.5281/zenodo.7254133`. 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.

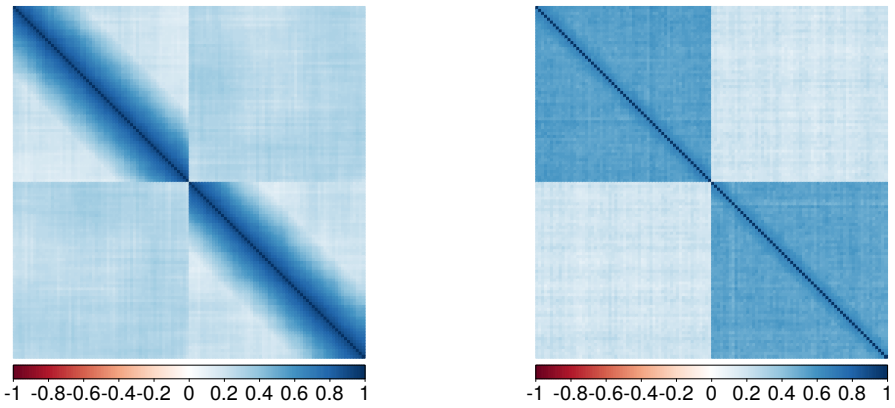
Human Connectome Project We also consider 35 subjects from the human connectome project (HCP), WU-Minn Consortium pre-processed (Glasser et al., 2013). Subjects were pseudonymized. Two fMRI acquisitions on different days are available for each subject. The scanning duration is 14 min and 24 s with a 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.

5.1.2 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, $Cor(X_i^A, X_{i'}^A) = \max(1 - |i' - i|/30, \eta_A^-)$, where $|i' - i|$ is the uniform norm between voxels i and i', and η_A^- the minimal population intra-correlation of a region A. In this paper 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 2 for a low intra-correlation and a high intra-correlation setting with high noise.



(a) $\eta_A^- = \eta_B^- = 0.2$ (b) $\eta_A^- = \eta_B^- = 0.8$

Figure 2: 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 variance $\gamma_A^2 = \gamma_B^2 = 0.5$. The diagonal blocks correspond to the intra-correlation of the two regions.

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 paper, 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.

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:

$$\text{ERROR} = \frac{1}{N^A_{clust} N^B_{clust}} \sum_{\nu_A, \nu_B} (r^{CLA}_{\nu_A,\nu_B} - \rho^{A,B})^2. \tag{16}$$

The resulting surfaces are displayed in Figure 3. The lower the error, the better the quality of the estimator. As expected from Theorems 1 and 2, the error is lowest (refer to the orange points in Figure 3) 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 4.3.2, 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 3. In fact, it seems to be located at the bottom of a valley and quite close to the minimal error for all settings.

(a) $\eta_A^-, \eta_B^- = 0.2, \sigma^2_{\epsilon_A}, \sigma^2_{\epsilon_B} = 0.5$ (b) $\eta_A^-, \eta_B^- = 0.8, \sigma^2_{\epsilon_A}, \sigma^2_{\epsilon_B} = 0.5$

(c) $\eta_A^-, \eta_B^- = 0.2, \sigma^2_{\epsilon_A}, \sigma^2_{\epsilon_B} = 0.1$ (d) $\eta_A^-, \eta_B^- = 0.8, \sigma^2_{\epsilon_A}, \sigma^2_{\epsilon_B} = 0.1$

Figure 3: 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 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

$$\text{SE} = (r^{CLA}_{A,B} - \rho^{A,B})^2. \tag{17}$$

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 4 and 5. The MSE is lower when using our proposed cut-off heights in all the considered scenarios.

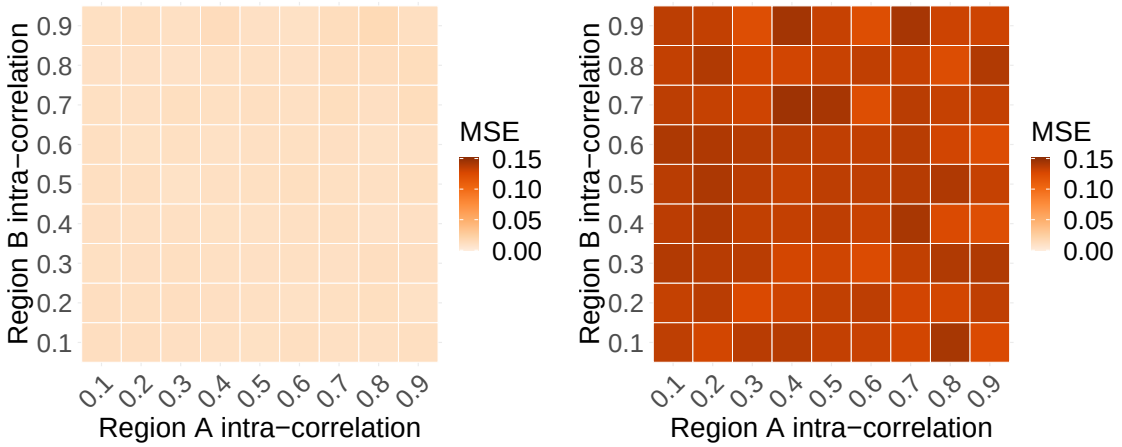
From now on, and unless stated otherwise, we will hence estimate the inter-correlation using this optimal cut-off height.

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.

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

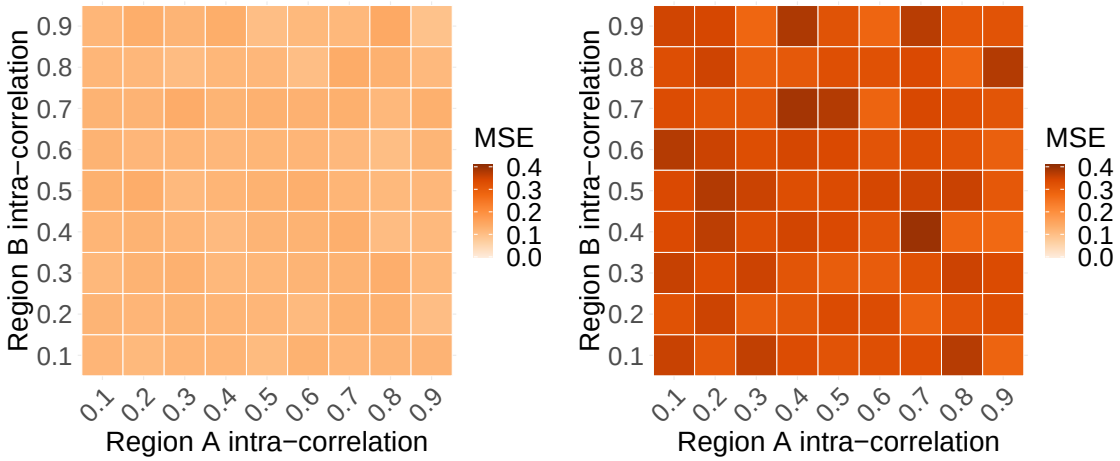
method which is based on a principal component analysis approach, and closely related to works from Dhillon et al. (2003) and Vigneau et al. (2015). DBSCAN (Ester et al., 1996), which allows to directly control the cluster radii, was also considered. However, it fails to produce any clustering on the 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 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



(a) Maximum distance between U-scores.

(b) Maximum silhouette score

Figure 4: 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 $\sigma^2_{\epsilon^A} = \sigma^2_{\epsilon^B} = 0.5$.



(a) Maximum distance between U-scores.

(b) Maximum silhouette score.

Figure 5: MSE ($\times 100$), 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 $\sigma^2_{\epsilon^A} = \sigma^2_{\epsilon^B} = 0.5$.

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, 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 meth-

ods 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.

Table 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^-	η_B^-	$\gamma_A^2 = \gamma_B^2$	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)

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 2. As expected from Theorem 2 and its corollary, our proposed estimator $r_{A,B}^{CLA}$ outperforms the other estimators for all settings, except the low noise scenarios with 3D Spherical intra-correlation, where the MSE for $r_{A,B}^{AC}$ is slightly lower. Even in this case, the MSE for $r_{A,B}^{AC}$ and $r_{A,B}^{CLA}$ 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 (4), Theorem 1, and Corollary 1 $r_{A,B}^{AC}$ and $r_{A,B}^{CLA}$ 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 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.

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, and that of three healthy human subjects (from the HCP dataset) are shown in Figure 8.

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.

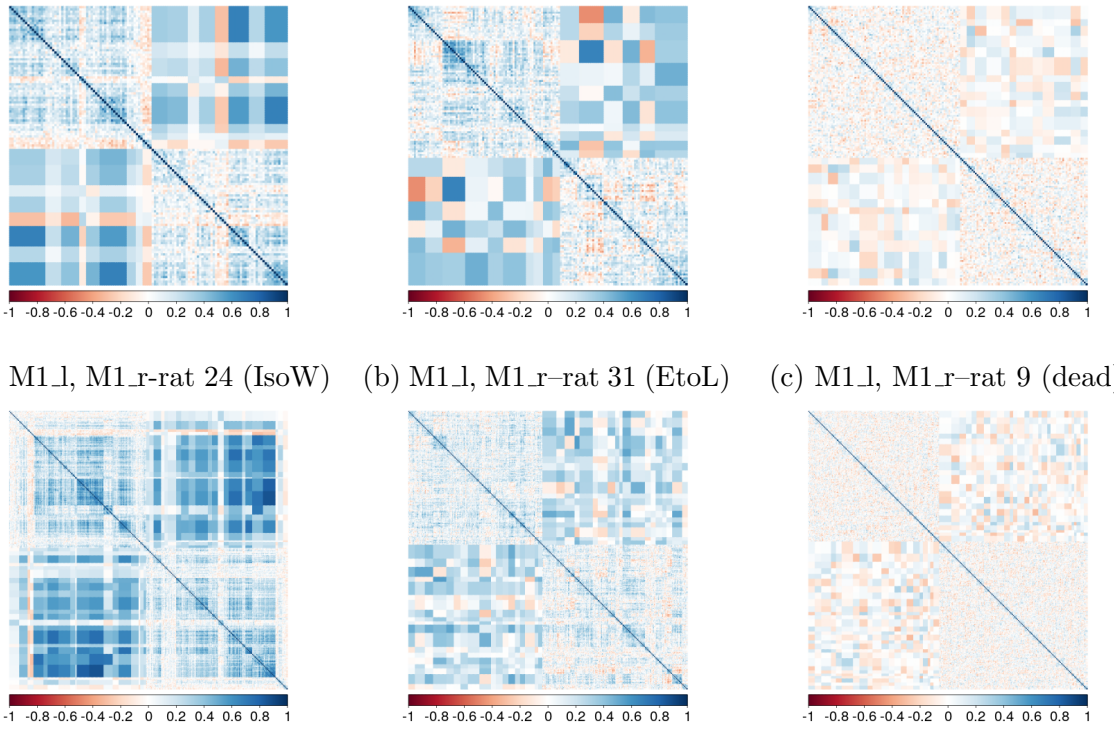
5.4.1 Rat Data

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,

Table 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^-	η_B^-	γ_A^2, γ_B^2	$r_{A,B}^{AC}$	$r_{A,B}^{CLA}$	$r_{A,B}^{CA}$
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_A^2, γ_B^2	$r_{A,B}^{AC}$	$r_{A,B}^{CLA}$	$r_{A,B}^{CA}$
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_A^2, γ_B^2	$r_{A,B}^{AC}$	$r_{A,B}^{CLA}$	$r_{A,B}^{CA}$
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})

no discernible structure of the dead rat's intra-correlation can be noted in Figure 6, where motor (M1_l, M1_r) and sensory (S1_l, S1_r) regions are represented. We find the MSE of $r_{A,B}^{CLA}$ is slightly higher than that of $r_{A,B}^{AC}$ (cf. Table 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



(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)

Figure 6: 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 $r^{A,A}_{i,i'}$, while the off-diagonal blocks correspond to the sample inter-correlation between clusters $r^{CLA}_{\nu_A,\nu_B}$.

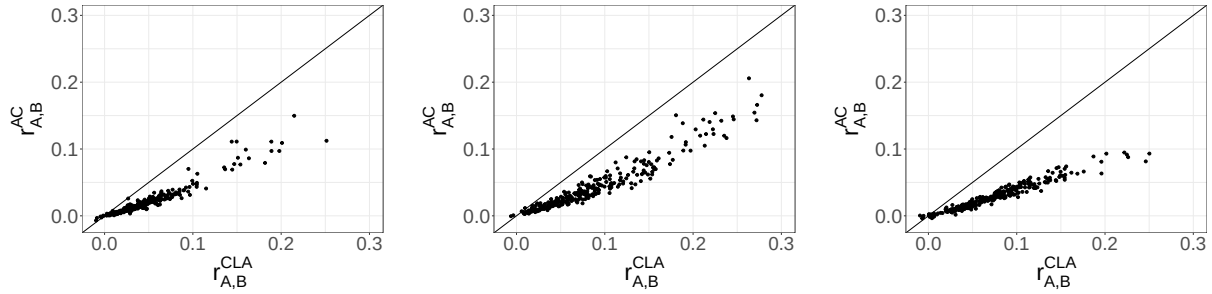
standard $r^{CA}_{A,B}$.

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). 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 7 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

Table 3: MSE across all pairs of regions for different dead rats and different estimators.

Dead Rat ID	$r^{AC}_{A,B}$	$r^{CLA}_{A,B}$	$r^{CA}_{A,B}$
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}

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.



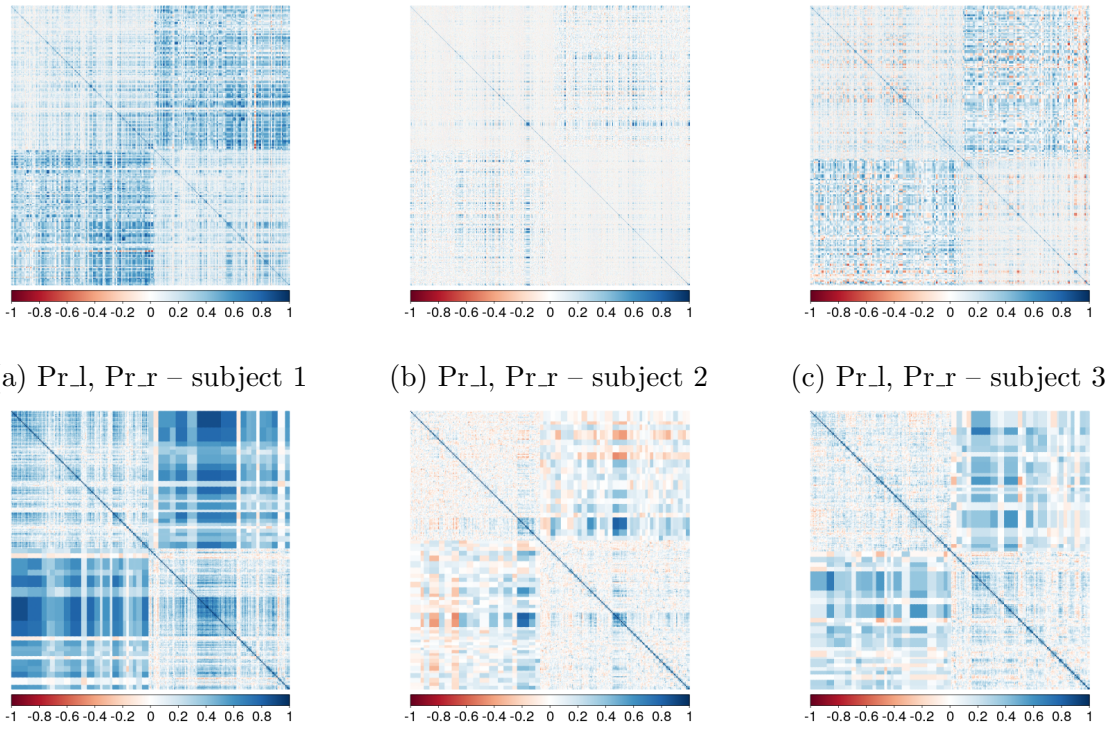
(a) Rat 24 (IsoW) (b) Rat 4 (UreL) (c) Rat 31 (EtoL)

Figure 7: 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.

5.4.2 HCP Data

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

Figure 8 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).



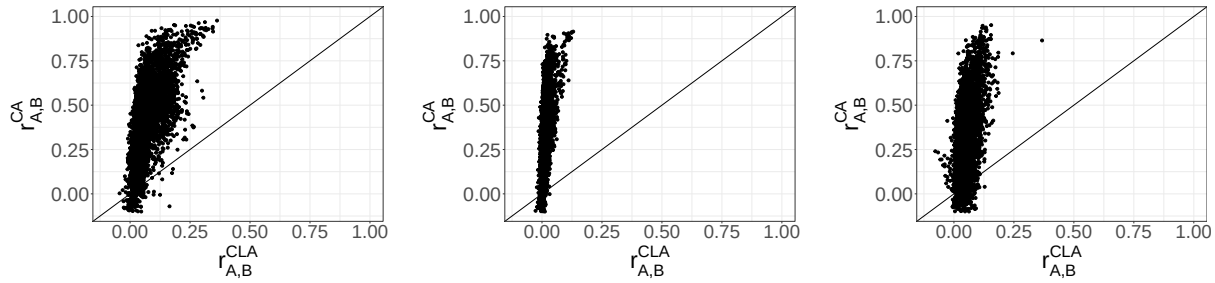
(a) Pr_l, Pr_r – subject 1 (b) Pr_l, Pr_r – subject 2 (c) Pr_l, Pr_r – subject 3

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

Figure 8: 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}.

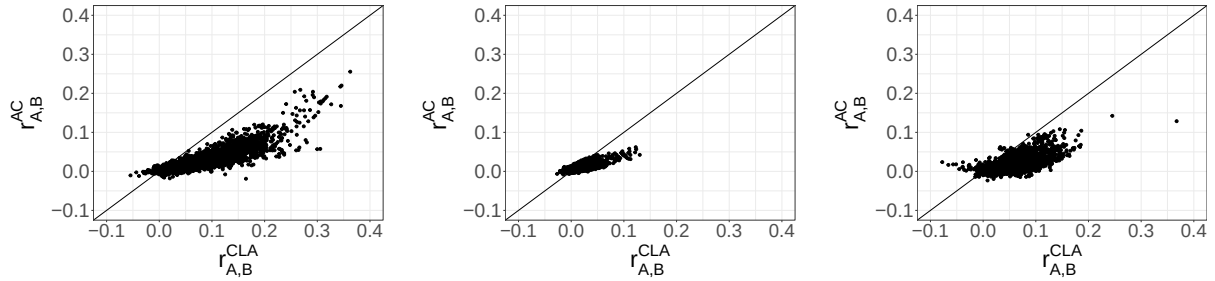
Subject 2 has in fact a benign anatomical brain anomaly. Our proposed approach hence allowed us to single out an unusual subject just by visually inspecting its sample intra- and inter-correlation values.

We can then compare the sample distribution of our proposed estimator $r_{A,B}^{CLA}$ with that of the standard estimator $r_{A,B}^{CA}$ (cf. Figure 9) and of $r_{A,B}^{AC}$ (cf. Figure 10). Overall, and as expected from equations (4) and (7) and Corollary 1, the correlation of 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

Figure 9: Inter-correlation coefficients estimated using $r^{CA}_{A,B}$ against our proposed estimator $r^{CLA}_{A,B}$ 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

Figure 10: 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 Conclusion

In this paper, 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 then proved that for an appropriate choice of cut-off heights of the dendrograms thus generated, our proposed estimator is a consistent estimator of the population inter-correlation. Moreover, our method yields both point estimates and a corresponding sample 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.

Supplementary Materials Proofs of the Theorems are available in the appendix. Discussion about the relaxation of assumptions on the noise, as well as additional details and results on the synthetic datasets are available in the supplementary materials. Source code, including a notebook detailing how to reproduce the figures of this paper, is available at: `https://gitlab.inria.fr/q-func/clustcorr`.

Funding This work was supported by the project Q-FunC from Agence Nationale de la Recherche under grant number ANR-20-NEUC-0003-02 and the NSF grant IIS-2135859.

A Proof of Theorem 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:

$$D(\nu_1, \nu_2) = \sqrt{\frac{2 \cdot |\nu_1| \cdot |\nu_2|}{|\nu_1| + |\nu_2|} \cdot \left\| \overline{\mathbf{U}}^{\nu_1} - \overline{\mathbf{U}}^{\nu_2} \right\|^2}, \tag{18}$$

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:

$$1 - \frac{h_A^2}{2} \;\leq\; 1 - \frac{\|\mathbf{U}_i^A - \mathbf{U}_{i'}^A\|^2}{2} \;= r_{i,i'}^{A,A}, \tag{19}$$

which implies the left-hand side of (8). The right-hand side follows from properties of correlation coefficients. This concludes the proof.

B Proof of Theorem 2

For two clusters ν_A, ν_B in regions A, B, from (11),

$$r_{\nu_A,\nu_B}^{CLA} = \frac{\widehat{Cov}(\overline{\mathbf{Y}}^{\nu_A}, \overline{\mathbf{Y}}^{\nu_B})}{\sqrt{\widehat{Var}(\overline{\mathbf{Y}}^{\nu_A}) \cdot \widehat{Var}(\overline{\mathbf{Y}}^{\nu_B})}}. \tag{20}$$

Since we have assumed variables are temporally i.i.d., and according to the model definition (cf. Section 3), as n tends towards infinity,

$$\widehat{Cov}(\overline{\mathbf{Y}}^{\nu_A}, \overline{\mathbf{Y}}^{\nu_B}) \overset{a.s.}{\to} Cov(\overline{Y}^{\nu_A}(t), \overline{Y}^{\nu_B}(t)), \tag{21}$$

for any time point t and where

$$\begin{aligned} Cov(\overline{Y}^{\nu_A}(t), \overline{Y}^{\nu_B}(t)) &= \frac{1}{|\nu_A| \cdot |\nu_B|} \sum_{i=1}^{|\nu_A|} \sum_{j=1}^{|\nu_B|} Cov(Y_i^A(t), Y_j^B(t)) \\ &= \frac{1}{|\nu_A| \cdot |\nu_B|} \sum_{i=1}^{|\nu_A|} \sum_{j=1}^{|\nu_B|} \sigma_A \sigma_B \rho^{A,B} \\ &= \sigma_A \sigma_B \rho^{A,B}, \end{aligned} \tag{22}$$

and, from equation (1),

$$\widehat{Var}(\overline{\mathbf{Y}}^{\nu_A}) \overset{a.s.}{\to} Var(\overline{Y}^{\nu_A}(t)) \;=\; \sigma_A^2 \cdot \frac{1}{|\nu_A|^2} \cdot \sum_{i,i'=1}^{|\nu_A|} \eta_{i,i'}^A + \frac{\gamma_A^2}{|\nu_A|}, \tag{23}$$

which gives (14), and concludes the proof.

References

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

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., 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.

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 networksa," *Journal of Neural Engineering*a, 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 regimesb," *NeuroImage*b, 219, 116945.

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.

Cameron, A. C. and Miller, D. L. (2015), "A Practitioner's Guide to Cluster-Robust Inference," *The Journal of Human Resources*, 50, 317 – 372.

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.

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.

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

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

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*, 226–231.

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.

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.

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

Lin, Lawrence I-Kuei (1989), "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, 45, 1, 255–268.

Lindskog, F. (2000), "Linear Correlation Estimation," *Risklab Research Paper, ETH-Zentrum, Zürich*.

Liu, P., Calhoun, V., and Chen, Z. (2017), "Functional overestimation due to spatial smoothing of fMRI data," *Journal of Neuroscience Methods*, 291, 1–12.

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.

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.

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.

Ribeiro, P. J. and Diggle, P. J. (2001), "geoR: a package for geostatistical analysis," *R-NEWS*, 1, 2, 14–18.

Rosner, B., Donner, A., and Hennekens, C. H. (1977), "Estimation of interclass correlation from familial data," *Applied Statistics*, 26, 179–187.

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.

Shevlyakov, G. and Smirnov, P. (2016), "Robust Estimation of the Correlation Coefficient: An Attempt of Survey," *Austrian Journal of Statistics*, 40, 147–156.

Srivastava, M. S. and Keen, K. J. (1988), "Estimation of the Interclass Correlation Coefficient," *Biometrika*, 75, 4, 731–739.

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.

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.

Ward, J. H. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 301, 236–244.

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*, D.Phil. thesis, Old Dominion University Libraries.

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.