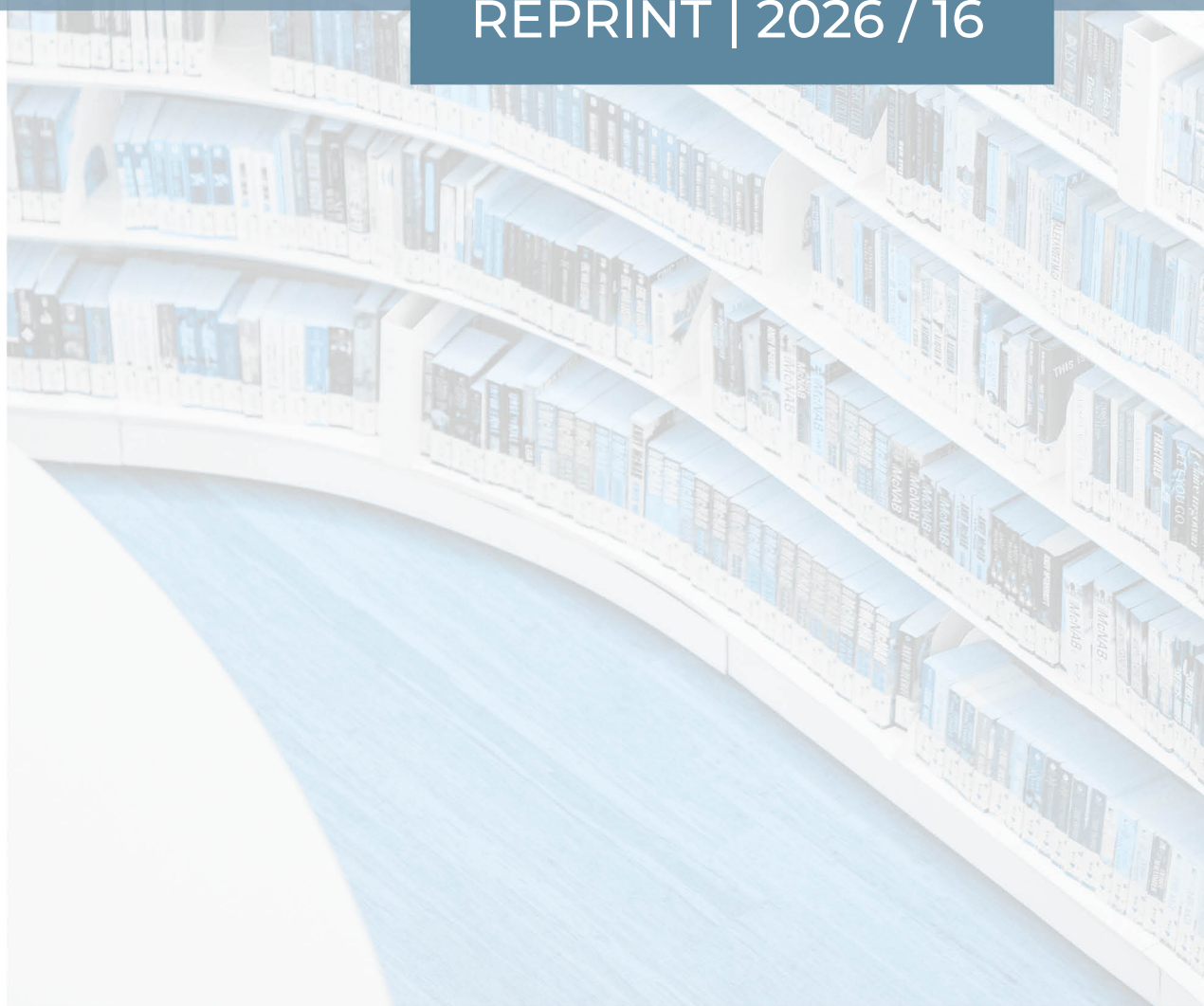
OVERLAPPING CLUSTERING OF TIME DEPENDENT VARIABLES FOR FMRI DATA

Eugen Pircalabelu, Xin Bing

REPRINT | 2026 / 16



UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication

Overlapping clustering of time dependent variables for fMRI data

Eugen Pircalabelu

UCLouvain, Institute of Statistics, Biostatistics and Actuarial Sciences, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium

E-mail: eugen.pircalabelu@uclouvain.be

Xin Bing

University of Toronto, Department of Statistical Sciences, Ontario, Canada

Summary. We propose in this paper a framework for performing overlapping clustering in the presence of time dependence, where the main goal is to identify overlapping coarse parcelations of the brain based on resting state fMRI time series. Our procedure is based on the Latent OVErlapping (LOVE) clustering method of Bing et. al (2020) which we extend to weakly dependent time series. Although the method is developed in an fMRI context, it is general enough to be directly applicable to other contexts, such as gene expression data, where one has at disposal multiple time series and is interested in identifying overlapping groups of similar elements.

Keywords: rs-fMRI, brain parcelation, overlapping clustering, reduced-rank regression, latent factor model

1. Introduction

The field of neuroscience has long been interested in characterizing and understanding functional brain connectivity and identifying homogeneous regions in the brain that are associated with certain tasks such as seeing, hearing, memory, etc. This has lead scientists in the last decade to investigate data-driven 'brain parcelations', where based on the measured brain activation signal, one desires to obtain via a clustering procedure, groups of voxels that are believed to act together to form a homogeneous region in the brain. We stress here that due to computational convenience classical, non-overlapping clustering (such as hierarchical, k-means or spectral clustering) is most often used. We refer to Arslan et al. (2018) for a systematic and thorough review on clustering methods applied for fMRI data and Jiang et al. (2004) and Wiwie et al. (2015) for large genetic datasets. However, groups of voxels need not belong to one single region. There is empirical evidence that suggests clusters are overlapping (Saxe et al., 2006; Downing et al., 2007) and as such, one is interested in quantifying the degree of association between a group of voxels and coarser functional regions. Other examples of overlapping clustering approaches in neurosciences include Cecchi et al. (2009), Yan et al. (2011), Ghebreab and Smeulders (2011), Jennings et al. (2016) and Li et al. (2023) among others. Relative to the cited work which constructs the clusters either by optimizing metrics that search for overlapping partitions or, the more popular route, using a form of dimensionality-reduction decomposition of the matrix of observed data, in the spirit of Independent

component analysis (ICA, Hyvärinen and Oja, 2000) or (functional) Principal component analysis, our approach proposes using latent factors to model the brain networks. Generally, the ICA-based approaches (see Calhoun et al., 2009; Varoquaux et al., 2010; Xu et al., 2013; Shi and Guo, 2016, among others) have shown good performance on real data, but relative to such an approach that tries to maximally separate the underlying sources, for the LOVE method the membership matrix dictating (i) which regions of interest (ROIs) belong to which cluster and, (ii) to which degree the clusters are overlapping, is a direct parameter of interest which will be explicitly estimated, as shown further in Sections 2 and 3. Moreover, our proposed estimator is sparse, thus by definition focusing on retaining the strong signal in the data and eliminating unimportant sources of variation. Recent work around sparse versions of ICA as in Wang et al. (2024) will thus form the most appropriate, direct comparison from an ICA perspective. Sections 4 and 5 will focus on such comparisons to sparse and dense versions of ICA.

To motivate the proposed methodology we use data from $n = 406$ healthy, unrelated subjects from the 'Young Adult' HCP study (available at `https://www.humanconnectome.org/study/hcp-young-adult`) measuring brain activity in a resting state fMRI study (see for more details Van Essen et al., 2013). In this manuscript, we have used the 3T preprocessed resting-state fMRI data from the 406 healthy, unrelated subjects from the S1200 Subjects release of the study. During the scan, the subjects were instructed to keep their eyes open and maintain fixation on a projected bright crosshair on a dark background and each scan ran for nearly 15 minutes. In the structural preprocessing pipeline, subject-level rs-fMRI data in their native space were co-registered using the Multimodal Surface Matching (MSM) algorithm (Robinson et al., 2014) and subsequently denoised using the ICA-FIX algorithm (Salimi-Khorshidi et al., 2014). Details on the MSM-All method employed by the MSM-All HCP pipeline are provided in Glasser et al. (2016b). Each scan was spatially smoothed using a surface-based Gaussian kernel with 6-mm FWHM and the smoothed grayordinate surface time series were resampled to FreeSurfer's fsaverage5 template and temporally standardized to zero mean and unit variance. We further performed global signal regression on these pre-processed data to improve the correspondence to anatomical connectivity (Fox et al., 2009). Finally, we use the multi-modal Glasser parcelation (Glasser et al., 2016a) to parcelate the brain recordings into ROI-level time series comprising $p = 360$ ROIs and $T = 1200$ time points. Additional pre-processing and the visuals have been performed in Python using a suite of hand-written functions, combined with standard `nilearn` functionalities, while the main body of functions for performing overlapping clustering was written in `R`.

Figure 1 illustrates what is understood by an 'overlapping' clustering solution. It is common in the literature to equate different clusters to distinct 'brain networks', each of which comprises multiple spatially non-overlapping/dissociated regions that perform a similar function and show correlated activity at rest (Biswal et al., 1995). Contrary to this non-overlapping paradigm, the proposed procedure allows some groups of voxels to belong to multiple brain networks at the same time. As such, the final configuration will contain groups of voxels that are associated to only *one* single network, which we graphically depict in Figure 1 using the color red, while other groups will be associated to *multiple* networks, which we graphically depict then using different intensities of yellow depending on the strength of association to multiple networks. Figure 1 illustrates certain



Fig. 1. Toy example illustrating clusters coloured by association intensity where yellow denotes non-pure regions of interest, while red denotes pure regions. The left lateral view is presented.

groups of voxels that are 'pure', in the sense of being associated with one single such coarse network, while other groups display a weaker form of association with multiple networks. In the right panel of the figure, the red regions indicate that the corresponding voxels are only associated with what is often labelled in neuroscience as the 'motor' network and are referred to as 'pure'. On the other hand, the yellow regions in the prefrontal cortex seem to be associated with both the 'visual' and the motor network and are hence referred to as 'non-pure'. A similar interpretation holds for the left panel as well, in the sense that some regions are solely associated with the visual network (red regions), while some are associated with multiple networks (yellow regions). The strength of association to the different networks is reflected by the different intensities, where lighter or darker shades reflect the magnitude of the estimated values of the $\boldsymbol{A}$ matrix from model (3), denoting a weaker or stronger form of association.

We propose in this manuscript a procedure that allows for parcelation of the brain where ROIs are allowed to belong simultaneously to multiple clusters. By 'parcelation', as in Eickhoff et al. (2018), we refer to the process of obtaining 'distinct partitions in the brain, be they areas or networks that comprise multiple discontinuous but closely interacting regions'. The goal is thus to identify parcels, i.e., spatially contiguous groups of voxels that have similar brain activity profiles, but can be spatially non-localized. In doing so, as opposed to the heuristic clustering approaches, we offer a principled, model-based framework that describes how the data generating process directly depends on the clustering via latent factors. The procedure is flexible enough to cover both low and high-dimensional settings by allowing the number of parameters to exceed the available sample size, and is underpinned by sound theoretical guarantees.

Notation: For any vector $v \in \mathbb{R}^m$ and $0 < q \leq \infty$, we use $\|v\|_q = (\sum_{a=1}^{m} |v_a|^q)^{1/q}$ to denote the ℓ_q norm of v. For any matrix $M \in \mathbb{R}^{m_1 \times m_2}$, we write $\|M\|_{p,q} = (\sum_{a=1}^{m_1} \|M_{a\bullet}\|_q^p)^{1/p}$. In particular, we write $\|M\|_F = \|M\|_{2,2}$ and $\|M\|_\infty = \|M\|_{\infty,\infty}$. We use M^+ to denote the Moore-Penrose inverse of M. For a subset $S \subseteq \{1, \ldots, m_1\}$, we write $M_{S\bullet}$ as the submatrix of M corresponding to the rows in S. For any $S \subseteq \{1, \ldots, m_2\}$, we similarly write $M_{\bullet S}$ as the submatrix of M corresponding to the columns in S.

2. Proposed model formulation

Assume that for a collection of n independent units, one measures a p-dimensional characteristic X at T separate occasions. The area of fMRI functional brain mapping fits

perfectly in this framework, as for such studies one measures the brain activity (X) of n independent subjects, where for each subject time series of length T are available for p different brain regions of interest.

For a sampling unit $i = 1, \ldots, n$ at a time point $t = 1, \ldots, T$, we assume that $X_t^{(i)} \in \mathbb{R}^p$ follows a Markov model of order 1 (i.e. current values are affected only by the immediate past, one lag apart) as

$$X_t^{(i)} = \boldsymbol{D} X_{t-1}^{(i)} + \epsilon_t^{(i)}, \tag{1}$$

where $X_t^{(i)}$ and $X_{t-1}^{(i)}$ are the p-dimensional measurements at time points t and $t-1$, respectively, and $\epsilon_t^{(i)} \in \mathbb{R}^p$ is a random vector with $\mathbb{E}(\epsilon_t^{(i)}) = 0$ and $\mathrm{Cov}(\epsilon_t^{(i)}) = \Sigma$. The transition matrix $\boldsymbol{D} \in \mathbb{R}^{p \times p}$ fully characterizes the dependency of two adjacent measurements and is assumed to be the same among all subjects. We treat it as fixed and unknown in this paper. We stress here that if desired, the model can directly be extended to (i) higher order Markov models or (ii) by allowing subject-specific transition matrices, at the cost of increased computational complexity. We have not fully pursued this avenue here, as for the HPC study we focus in this work, our analysis in Section 5 shows that the order 1 assumption seems to be satisfied, but Section 7 treats briefly these extensions and presents the obtained results on the HPC data.

After taking out the dependence on the previous time point, we assume that the remaining part of $X_t^{(i)}$ follows a latent factor model, specified as

$$\epsilon_t^{(i)} = \boldsymbol{A} Z_t^{(i)} + u_t^{(i)}, \tag{2}$$

where $Z_t^{(i)} \in \mathbb{R}^K$ is a vector of (unobservable) latent factors with zero mean and $\mathrm{Cov}(Z_t^{(i)}) = \boldsymbol{C}$, and $u_t^{(i)} \in \mathbb{R}^p$ is a random noise vector with $\mathbb{E}(u_t^{(i)}) = 0$ and $\mathrm{Cov}(u_t^{(i)}) := \boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$. The number of latent factors, K, is unknown and is assumed to be smaller than p. We further assume that $X_{t-1}^{(i)}$, $Z_t^{(i)}$ and $u_t^{(i)}$ are mutually independent for $1 \leq t \leq T$ and $1 \leq i \leq n$. Combining (1) with (2) gives

$$X_t^{(i)} = \boldsymbol{D} X_{t-1}^{(i)} + \boldsymbol{A} Z_t^{(i)} + u_t^{(i)}. \tag{3}$$

By using the latent factors $Z_t^{(i)}$ to model the brain networks, the membership matrix $\boldsymbol{A} \in \mathbb{R}^{p \times K}$ in (2) is therefore the main parameter of our interest as it defines the overlapping clusters of the p measurements in $X_t^{(i)}$, after modelling out the time dependency, via their association with $Z_t^{(i)}$. Model (3) is seen as a generalization of the factor model investigated in Bing et al. (2020) where $X_t^{(i)}$ was assumed to be independent of $X_{t-1}^{(i)}$, that is, $\boldsymbol{D} = 0$.

Since the matrix $\boldsymbol{A}$ is not identifiable, as one can insert any invertible matrix $\boldsymbol{Q}$ such that $(\boldsymbol{AQ})(\boldsymbol{Q}^{-1} Z_t^{(i)}) = \boldsymbol{A} Z_t^{(i)}$ with $\boldsymbol{AQ} \neq \boldsymbol{A}$, one needs to impose additional identifiability conditions. We adopt the following identifiability conditions proposed in Bing et al. (2020).

Assumption 1.

(a) For each $1 \leq k \leq K$, there exists $j \neq j'$ such that $|\boldsymbol{A}_{j\bullet}| = |\boldsymbol{A}_{j'\bullet}| = \boldsymbol{e}_k$.

(b) For each $1 \leq j \leq p$, $\|\boldsymbol{A}_{j\bullet}\|_1 \leq 1$.
(c) $\boldsymbol{C}$ *is positive definite and* $\boldsymbol{C}_{kk} \geq |\boldsymbol{C}_{kk'}|$ *for all* $1 \leq k, k' \leq K$.

By $|\boldsymbol{A}_{j\bullet}|$ and $|\boldsymbol{A}_{j'\bullet}|$ we denote in Assumption 1 the vector of absolute values of the elements on the j-th and j'-th row of $\boldsymbol{A}$, while $\boldsymbol{e}_k$ is a vector containing the value 1 on position k and having all other elements equal to 0. Item (a) is called the *pure variable* assumption since it requires that for each latent factor, there exist at least two measurements that are solely associated with this latent factor. Under our fMRI data context, the group centres defined by the latent factors $Z_t^{(i)}$ correspond to different brain networks, hence (a) is a reasonable assumption because there are multiple brain regions (measurements in $X_t^{(i)}$) that are only associated with each network. On the other hand, since some brain regions are associated with multiple networks, the resulting clusters are expected to be overlapping which are enabled by the measurements that are not pure. Assumption $1(b)$ is a scaling condition so that $\boldsymbol{A}$ can be interpreted as a membership matrix. In terms of the covariance matrix of $Z_t^{(i)}$, (c) is a mild incoherence condition that implies any two factors are not essentially the same in the sense that $|[Z_t^{(i)}]_k| \neq |[Z_t^{(i)}]_{k'}|$ almost surely for any $k \neq k'$. This is a reasonable and realistic assumption for fMRI studies as well, as one wishes to identify multiple, distinct latent factors from the data. Without such an assumption it is impossible to disentangle the effect of different factors from each other.

Under Assumption 1, Theorem 2 of Bing et al. (2020) shows that the membership matrix $\boldsymbol{A}$ is identifiable, up to a signed permutation matrix, from $\Sigma = \mathrm{Cov}(\varepsilon_t^{(i)})$. Since Σ can be identified by regressing $X_t^{(i)}$ on $X_{t-1}^{(i)}$, we conclude that $\boldsymbol{A}$ is identifiable under model (3) coupled with Assumption 1. After identifying $\boldsymbol{A}$, the groups $G = \{G_1, \ldots, G_K\}$ of $X_t^{(i)}$ are defined via the support of $\boldsymbol{A}$ as

$$G_k = \{1 \leq j \leq p : \boldsymbol{A}_{jk} \neq 0\}, \quad \forall\ 1 \leq k \leq K. \tag{4}$$

As the columns of $\boldsymbol{A}$ can have common support, the groups could be overlapping.

3. Estimation

We discuss in this section the proposed estimation procedure for constructing the overlapping clusters defined in (4). To this end, we start by concatenating the information from all time points and all subjects in model (1) and write $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{D} + \boldsymbol{E}$, where

$$\boldsymbol{Y} \in \mathbb{R}^{nT\times p} = \begin{bmatrix} X_1^{(1)} \\ \vdots \\ X_T^{(1)} \\ \vdots \\ X_T^{(n)} \end{bmatrix}, \quad \boldsymbol{X} \in \mathbb{R}^{nT\times p} = \begin{bmatrix} X_0^{(1)} \\ \vdots \\ X_{T-1}^{(1)} \\ \vdots \\ X_{T-1}^{(n)} \end{bmatrix}, \quad \boldsymbol{E} \in \mathbb{R}^{nT\times p} = \begin{bmatrix} \varepsilon_1^{(1)} \\ \vdots \\ \varepsilon_T^{(1)} \\ \vdots \\ \varepsilon_T^{(n)} \end{bmatrix}.$$

The matrix $\boldsymbol{Y}$ is the response matrix containing the 'present' information and $\boldsymbol{X}$ is the predictor matrix containing the 'past' information. Without loss of generality, we assume that both $\boldsymbol{Y}$ and $\boldsymbol{X}$ are centred.

Our ultimate goal is to estimate the allocation matrix $\boldsymbol{A}$ from which one can obtain overlapping clusters of the ROIs via (4). Since the optimal estimation of $\boldsymbol{A}$ is provably guaranteed by the LOVE algorithm proposed in Bing et al. (2020), provided a good estimator of Σ with small entry-wise error, we first discuss below how to estimate Σ.

3.1. Estimation of the residual covariance matrix Σ

To estimate Σ, we distinguish two cases depending on the total sample size nT and the number of features p. When the total sample size is larger than p, we propose to use the following estimator of Σ

$$\widehat{\Sigma}_1 = \frac{1}{(nT-q)}\boldsymbol{Y}^\top \boldsymbol{P}^\perp \boldsymbol{Y}, \qquad \text{if} \quad nT > p, \tag{5}$$

where $\boldsymbol{P}^\perp = \boldsymbol{I} - \boldsymbol{P}$, $q = \text{rank}(\boldsymbol{X})$ and $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top$ denotes the projection matrix onto the space spanned by the columns of $\boldsymbol{X}$. The matrix $\widehat{\Sigma}_1$ can be viewed as the sample covariance matrix of the residuals after regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$ by using the ordinary least squares estimator $\widehat{\boldsymbol{D}}_{\text{ols}} = (\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{Y}$ to estimate $\boldsymbol{D}$.

However, when $nT < p$, it is well-known that $\widehat{\boldsymbol{D}}_{\text{ols}}$ performs poorly for estimating $\boldsymbol{D}$, hence (5) is not a good estimator of Σ. In this high-dimensional setting, one needs additional structure on $\boldsymbol{D}$ to ensure that it can be estimated consistently. We focus on the assumption that $\boldsymbol{D}$ is low-rank, that is, $r = \text{rank}(\boldsymbol{D}) < p$, since under the context of fMRI data, it has the implication that the current measurements $X_t^{(i)}$ depend on the previous $X_{t-1}^{(i)}$ only through their r linear combinations. We refer to Remark 2 for a discussion on other structures on $\boldsymbol{D}$. The following reduced-rank estimator (Izenman, 2008) has been widely used for estimating $\boldsymbol{D}$:

$$\widehat{\boldsymbol{D}}_d := \underset{\boldsymbol{D}:\ \text{rank}(\boldsymbol{D})\le d}{\arg\min} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{D}\|_F^2 \tag{6}$$

for any integer $d \in \{1, 2, \ldots, q\}$. A closed-form solution of $\widehat{\boldsymbol{D}_d}$ is known to be

$$\widehat{\boldsymbol{D}}_d = (\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{Y}\sum_{k=1}^{d} v_k v_k^\top \tag{7}$$

where $v_1, \ldots, v_p$ are eigenvectors of $\boldsymbol{Y}^\top\boldsymbol{P}\boldsymbol{Y}$ corresponding to non-increasing eigenvalues. By using $\widehat{\boldsymbol{D}}_d$, we propose to further estimate Σ by

$$\widehat{\Sigma}_2 = \frac{1}{nT}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{D}}_d\right)^\top\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{D}}_d\right), \qquad \text{if} \quad nT \le p. \tag{8}$$

Remark 1 (Selection of d). In practice an optimal value to be used for d is not a priori known. As in the spirit of (6), we propose to select a data-driven $\widehat{d}$ based on a validation set $(\boldsymbol{Y}^v, \boldsymbol{X}^v)$ independent of $(\boldsymbol{Y}, \boldsymbol{X})$ as

$$\widehat{d} = \underset{d\in\{1,2,\ldots,q\}}{\arg\min} \|\boldsymbol{Y}^v - \boldsymbol{X}^v\widehat{\boldsymbol{D}}_d\|_F^2. \tag{9}$$

Here $\widehat{\boldsymbol{D}}_d$ is defined in (7) by using $(\boldsymbol{Y}, \boldsymbol{X})$. Alternatively, we could select d by

$$\widehat{d} = \underset{d \in \{1,2,\ldots,q\}}{\arg\min} \|\boldsymbol{Y}^v - \boldsymbol{X}^v \widehat{\boldsymbol{D}}_d\|_{2,\infty}, \tag{10}$$

with $\|M\|_{2,\infty} = \max_a \|M_{a\bullet}\|_2$. The intuition of using (10) to select d comes from the fact that it provides a better control over $\|\widehat{\Sigma} - \Sigma\|_\infty$, which is the key requirement of the LOVE algorithm in Bing et al. (2020). Our finite sample simulation results indicate that these two approaches, namely using (9) or (10), have close performance in terms of yielding similar overlapping groups.

Remark 2 (Other structures on $\boldsymbol{D}$). We choose to impose the low-rank structure on $\boldsymbol{D}$. In the context of fMRI data, it has the implication that the current measurements $X_t^{(i)}$ depend on the previous $X_{t-1}^{(i)}$ only through their r linear combinations. This AR(1) time dependency scheme for fMRI data has been widely used in practice, see among others Arbabshirani et al. (2014); Liégeois et al. (2017). One of our starting points, most closely connected to the proposed order 1 model, is the work of Liégeois et al. (2017) which uses multivariate AR processes of order 1 for HCP S900 data. In Liégeois et al. (2021) the same AR(1) model is further proposed as a baseline model in a hypothesis testing strategy for fMRI. Liao et al. (2010) identified as well (based on the BIC information criterion) an order 1 for the autoregressive models they used to evaluate Granger causality relations for resting state fMRI, while Nagle et al. (2023) and Endemann et al. (2022) both use sparse high-dimensional multivariate AR models for resting data. In many other applications, it might be more appealing to consider different structures on $\boldsymbol{D}$, such as various sparsity assumptions on $\boldsymbol{D}$, see among others Han and Liu (2013); Davis et al. (2016). Our approach can be easily modified to accommodate these structures and an empirical evaluation of this strategy is presented in Section 4.

3.2. Estimation of the membership matrix $\boldsymbol{A}$ and overlapping groups

Let $\widetilde{\Sigma}$ be our estimate of Σ as constructed either in (5) or (8). We apply the procedure of Bing et al. (2020) to $\widetilde{\Sigma}$ for estimating the membership matrix $\boldsymbol{A}$ and constructing the estimated overlapping clusters of the ROIs. We follow their notation and denote by I the set of indices of pure variables as specified in Assumption 1(a). Specifically, define

$$I = \bigcup_{k=1}^{K} I_k, \qquad I_k = \{1 \le j \le p : |\boldsymbol{A}_{j\bullet}| = \boldsymbol{e}_k\}.$$

The estimation of $\boldsymbol{A}$ mainly consists of three steps: (1) estimate I and its partition $\{I_1, \ldots, I_K\}$, (2) estimate the submatrix $\boldsymbol{A}_{I\bullet}$ of $\boldsymbol{A}$ and (3) estimate the submatrix $\boldsymbol{A}_{J\bullet}$ of $\boldsymbol{A}$, with $J = \{1, \ldots, p\} \setminus I$.

For step (1), the LOVE Algorithm given in Appendix A estimates the number of latent factors (K), the set of pure variables (I) and its partition by taking the input $\widetilde{\Sigma}$ and a tuning parameter δ. It is an optimization-free algorithm and has computational complexity of order p^2. In step (2), for each $k \in \{1, \ldots, \widehat{K}\}$ given the estimate $\widehat{I}$ of I,

we estimate $\boldsymbol{A}_{I\bullet}$ by

$$\begin{aligned}\widehat{\boldsymbol{A}_{a\bullet}} &= \boldsymbol{e}_k, \qquad && \text{pick any } a \in \widehat{I}_k,\\ \widehat{\boldsymbol{A}_{b\bullet}} &= \text{sign}(\widetilde{\Sigma_{ab}})\boldsymbol{e}_k, \qquad && \text{for any } b \in \widehat{I}_k \setminus \{a\}.\end{aligned}$$

Next, in step (3), the estimation of $\boldsymbol{A}_{J\bullet}$ is based on the key identity $\Sigma_{JI} = \boldsymbol{A}_{J\bullet}\boldsymbol{C}\boldsymbol{A}_{I\bullet}^{\top}$, implied by model (2) and the diagonal structure of $\text{Cov}(u_t^{(i)})$. Provided $\widehat{\boldsymbol{A}_{I\bullet}}$ and an estimator $\widehat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega} := \boldsymbol{C}^{-1}$, by letting $\widehat{J} = \{1, \ldots, p\} \setminus \widehat{I}$, it is natural to estimate $\boldsymbol{A}_{J\bullet}$ by

$$\widetilde{\boldsymbol{A}_{\widehat{J}\bullet}} = \widetilde{\Sigma_{\widehat{J}\widehat{I}}}\widehat{\boldsymbol{A}_{\widehat{I}\bullet}}\left(\widehat{\boldsymbol{A}_{\widehat{I}\bullet}^{\top}}\widehat{\boldsymbol{A}_{\widehat{I}\bullet}}\right)^{-1}\widehat{\boldsymbol{\Omega}}.$$

Recall that the groups are defined via the support of $\boldsymbol{A}$ as specified by (4). For support recovery, Bing et al. (2020) proposes to soft-threshold $\widetilde{A}_{\widehat{J}\bullet}$ to form the final estimator

$$\widehat{\boldsymbol{A}_{j\bullet}} = \arg\min_{\beta \in \mathbb{R}^K} \|\beta\|_1 \quad \text{s.t. } \|\beta - \widetilde{\boldsymbol{A}_{j\bullet}}\|_\infty \leq \mu, \qquad \forall\, j \in \widehat{J}.$$

Here μ is a tuning parameter. Combining $\widehat{\boldsymbol{A}_{\widehat{I}\bullet}}$ and $\widehat{\boldsymbol{A}_{\widehat{J}\bullet}}$ gives the estimate of $\boldsymbol{A}$ which, in view of (4), is used to form the groups of ROIs as

$$\widehat{G_k} = \left\{1 \leq j \leq p : \widehat{\boldsymbol{A}_{jk}} \neq 0\right\}, \quad \forall\, 1 \leq k \leq K.$$

It is important at this stage to note that the estimation of homogeneous clusters of ROIs for fMRI applications comes from the homogeneity/similarity of entries in the estimates of the covariance matrix Σ (see Algorithm 1 in Appendix A), whereas previous work using ICA-based methods focuses on directly estimating the mixing matrix via SVD decompositions. The consequence of this is that LOVE determines the grouping structure in a data-adaptive way and thus does not need any information on the number of potential latent groups, whereas ICA-based methods are constructed with an a priori specified number of known sources.

Finally, for estimating $\boldsymbol{\Omega} := \boldsymbol{C}^{-1}$, we solve the linear program

$$(\widehat{\boldsymbol{\Omega}}, \widehat{\zeta}) = \underset{\zeta \in \mathbb{R}^+, \boldsymbol{\Omega} \in \mathbb{R}^{K\times K}}{\arg\min} \zeta \qquad \text{subject to}$$

$$\boldsymbol{\Omega} = \boldsymbol{\Omega}^{\top}, \qquad \|\boldsymbol{\Omega}\widehat{\boldsymbol{C}} - \boldsymbol{I}\,\|_\infty \leq \lambda\zeta, \qquad \|\boldsymbol{\Omega}\|_{\infty,1} \leq \zeta$$

with a regularization parameter $\lambda > 0$.

The whole procedure of estimating $\boldsymbol{A}$ consists of three tuning parameters δ, λ and μ. When the random vectors $Z_t^{(i)}$ and $u_t^{(i)}$ are sub-Gaussian, the theoretical order of δ and λ is $\sqrt{\log(p)/(nT)}$ while $\mu = \mathcal{O}(\|\widehat{\boldsymbol{\Omega}}\|_{\infty,1}\delta)$. We refer to Section 5.1 of Bing et al. (2020) for the selection of leading constants of δ, λ and μ in a data-driven way.

4. Simulations

In this section we evaluate the performance of the estimators of Σ proposed in Section 3.1 in a controlled simulation study. We compare the proposed estimators with the sample covariance estimator (for which $\boldsymbol{D} = 0$) and sparse alternatives when estimating $\boldsymbol{D}$.

- **Sample covariance estimator.** The estimator of Σ is $\widehat{\Sigma}_S = \frac{1}{nT}\boldsymbol{Y}^\top\boldsymbol{Y}$, the sample covariance matrix of $\boldsymbol{Y}$. This estimator completely ignores the dependence structure in (1) hence it corresponds to using $\widehat{\boldsymbol{D}_S} = 0$ in (8) in lieu of $\widehat{\boldsymbol{D}_d}$.
- **Lasso-based sparse estimator.** We consider $\widehat{\Sigma}_{\text{lasso}} = \frac{1}{nT}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{D}}\right)^\top\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{D}}\right)$ where $\widehat{\boldsymbol{D}}$ is a sparse estimator obtained based on an ℓ_1 regularization procedure as in Basu and Michailidis (2015).
- **EN-based sparse estimator.** We consider $\widehat{\Sigma}_{\text{en}} = \frac{1}{nT}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{D}}\right)^\top\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{D}}\right)$ where $\widehat{\boldsymbol{D}}$ is a sparse estimator obtained with an elastic net penalty.

For the sparse estimators the implementation offered in the `sparsevar` (Vazzoler, 2021) package from the `R` software is used.

In the first part of the simulation study, data from the correctly specified model expressed in (1) and (2) have been generated for multivariate time series (containing p components) of length $T = 30$ for $n = 50$ independent samples and where $p = 200$. The true number of clusters was set to $K = 8$. The matrix $\boldsymbol{A}$ was generated by sampling randomly first a number (between $\lfloor p/(2K)\rfloor$ and $2K$) of indices corresponding to pure variables. Some of the pure variables have been set randomly to the value 1, while the remaining ones were set to -1. All non-pure variables were allowed to load on either 2 or 3 clusters, randomly chosen. The corresponding entries were set to either $\mathtt{s} \times 0.5$ (in case the variable loads on 2 clusters) or $\mathtt{s} \times 0.33$ (in case the variable loads on 3 clusters) where the value `s` was randomly sampled from the set $\{+1, -1\}$. Each of the p variables thus loads on at least 1 and at most 3 clusters. The matrix $\boldsymbol{C}$ was generated as $C_{ab} = (-1)^{a+b}\rho^{|a-b|}$ for each $a, b \in \{1, \ldots, K\}$, where $\rho \in \{.3, .6, .9\}$. In the last step, the matrix $\boldsymbol{\Gamma}$ was generated as a diagonal matrix with elements sampled from Unif[1, 2] and then Σ was set to $\Sigma = \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top + \boldsymbol{\Gamma}$.

The true autoregressive matrix was generated according to the following scenario. We first generate the entries of $\boldsymbol{D}$ i.i.d. from Unif[-0.2, 0.2] and then truncate its SVD decomposition by retaining the first 10 components to have $r = \text{rank}(\boldsymbol{D}) = 10$. Considering $nT = 1500$ as the effective sample size and $r(p + \min(nT, p) - r)$ as the number of free parameters for the reduced-rank regression problem in (6) (thus ignoring the number of parameters needed to parametrize Σ), one needs to estimate 3900 parameters. Once Σ and $\boldsymbol{D}$ have been constructed, the `R` package `tsDyn` (Di Narzo et al., 2023) has been used to simulate multivariate VAR models with the specified parameters.

In the second part of the study, data from misspecified models have been generated. Different forms of misspecification have been considered: (i) lag misspecification (where the true data came from a VAR(3) process with coefficient matrices $\boldsymbol{D}_1$, $\boldsymbol{D}_2/5$ and $\boldsymbol{D}_3/10$ where $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $\boldsymbol{D}_3$ are generated as above); (ii) subject-specific effects (where the true data were generated by using the matrices $\boldsymbol{D}_i$, $i = 1, \ldots, n$ in model 3) and (iii) long memory processes (where the true data were generated by using multivariate fractional white noise with correlation between the innovations set at $\rho = 0.7$ and memory parameters set at 0.4 for all series). Roughly, the long-memory DGP is constructed such that the simulated data manifest strong persistence of autocorrelations, which generally take much longer to decay to zero than classical ARMA-type models. For the long

memory misspecification the `R` package `LongMemoryTS` (Leschinski, 2019) has been used to simulate multivariate data with the specified parameters.

For all five estimators $\widehat{\Sigma}_1, \widehat{\Sigma}_2, \widehat{\Sigma}_S, \widehat{\Sigma}_{\text{lasso}}$ and $\widehat{\Sigma}_{\text{en}}$, we apply the LOVE algorithm (stated in Appendix A) and obtain the estimated overlapping clustering. For the selection of the tuning parameters (δ, λ), in order to avoid overfitting, we have used a data-splitting approach on the training set. The parameter δ was selected from a grid of 100 different values and once δ was selected, we proceeded and selected λ from a grid of 100 different values. For each parameter the training data were split randomly five times and the average performance across the five different splits was used to select the parameter. For μ, we have fixed the leading constant to 0.1. For the competitors based on sparse coefficient matrices the regularisation levels have been chosen based on a standard 3-fold cross-validation scheme.

All competitors are evaluated with respect to out-of-sample prediction error, model fit relative to the true quantities and discrepancy between the estimated and the true underlying number of communities. For all measures smaller values denote better performance. The criteria are defined as:

$$\begin{aligned} \text{PE}_{\text{test}} &= || \boldsymbol{Y}_{\text{test}} - \boldsymbol{X}_{\text{test}} \widehat{\boldsymbol{D}}_r ||_F & \text{Fit}_3 &= || \widehat{\boldsymbol{A}}_r \widehat{\boldsymbol{A}}_r^\top - \boldsymbol{A}\boldsymbol{A}^\top ||_F \\ \text{Fit}_1 &= || \widehat{\Sigma}_r - \Sigma ||_\infty & \text{Fit}_4 &= || \widehat{\boldsymbol{C}}_r - \boldsymbol{C} ||_F \\ \text{Fit}_2 &= || \widehat{\boldsymbol{A}}_r \widehat{\boldsymbol{C}}_r \widehat{\boldsymbol{A}}_r^\top - \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top ||_F & \text{Cl.Diff.} &= | \widehat{K}_r - K | \end{aligned}$$

where $\boldsymbol{Y}_{\text{test}}$, $\boldsymbol{X}_{\text{test}}$ represent the out-of-sample data for 50 new, unseen subjects for which one desires a low prediction error; $\widehat{\Sigma}_r$ with $r \in \{1, 2, S, \text{lasso}, \text{en}\}$ generically denotes the estimator obtained using the training set, $\widehat{\boldsymbol{D}}_r$ generically denotes the transition matrix and $\widehat{\boldsymbol{A}}_r$, $\widehat{\boldsymbol{C}}_r$ and $\widehat{K}_r$ represent the estimated quantities for the parameters in model (2). The quantities $\Sigma, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{C}$ and K represent the true parameters. The norms denote the usual ℓ_∞ and Frobenius norms.

The criterion PE_{test}, has the purpose to assess how well the transition matrices $\widehat{\boldsymbol{D}}_r$ fit the mean structure and extrapolate to new, unseen data. Fit_1 has the purpose to assess the entry-wise accuracy of the proposed estimators relative to the true covariance matrix used in the data generating process. We remind the reader that to ensure an optimal recovery of the allocation matrix $\boldsymbol{A}$ with LOVE, one necessarily needs an accurate entry-wise estimator for Σ. Fit_2 assesses how closely a 'reconstructed' Σ based on the estimated values coming from the LOVE algorithm, comes to the true covariance. Fit_3 and Fit_4 focus on how accurately the allocation matrix $\boldsymbol{A}$ and the matrix $\boldsymbol{C}$ are estimated, while Cl.Diff. assesses how far off the estimated number of clusters is from the true number.

Table 1 shows the obtained results averaged over 1000 different simulation runs. We conclude that if the data are generated from the correctly specified model in (1) and (2), then using the estimators $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$ can result in better performance in terms of fit to new, unseen data and accuracy of recovering the unknown quantities over the naive sample covariance estimator and the sparsity-based alternatives. Applying LOVE with $\widehat{\Sigma}_2$ provides as expected a good performance with respect to all performance measures (inspecting other norms such as the element-wise ℓ_1 norm or the supremum norm led to similar conclusions, results not shown but available from the authors) and regardless of the misspecification type, thus showing empirically that using LOVE on the residual

Table 1. Simulated data with $T = 30, n = 50$ and $p = 200$. Mean performance metrics over 1000 different simulation runs are presented. Lower values denote better performance.

	ρ	$\mathrm{PE}_{\mathrm{test}}$					Fit_1					Fit_2				
		$\widehat{\Sigma}_2$	$\widehat{\Sigma}_1$	$\widehat{\Sigma}_S$	$\widehat{\Sigma}_{\mathrm{Lasso}}$	$\widehat{\Sigma}_{\mathrm{EN}}$	$\widehat{\Sigma}_2$	$\widehat{\Sigma}_1$	$\widehat{\Sigma}_S$	$\widehat{\Sigma}_{\mathrm{Lasso}}$	$\widehat{\Sigma}_{\mathrm{EN}}$	$\widehat{\Sigma}_2$	$\widehat{\Sigma}_1$	$\widehat{\Sigma}_S$	$\widehat{\Sigma}_{\mathrm{Lasso}}$	$\widehat{\Sigma}_{\mathrm{EN}}$
Correct	.3	809	864	1096	1154	1161	12	14	148	172	175	18	23	166	174	178
	.6	810	864	1095	1163	1169	13	15	149	180	183	20	22	167	180	183
	.9	819	863	1090	1186	1193	18	19	152	209	215	28	24	172	204	210
VAR(3)	.3	823	877	1120	1182	1189	13	14	160	186	189	20	23	177	187	191
	.6	824	876	1117	1189	1196	14	16	157	191	195	21	22	176	191	194
	.9	832	875	1115	1212	1220	18	20	163	221	226	28	24	183	212	218
D_i	.3	1099	1145	1200	1229	1231	40	29	1852	1911	1921	44	41	1858	1913	1925
	.6	1100	1143	1212	1245	1248	40	30	1634	1680	1695	47	40	1641	1679	1693
	.9	1106	1141	1240	1290	1294	42	33	1275	1328	1336	38	30	1273	1334	1339
Memo.	.3	970	970	1057	1104	1110	15	15	76	105	111	27	28	70	95	101
	.6	970	970	1057	1117	1128	16	16	85	138	152	28	28	87	128	143
	.9	970	970	1056	1149	1175	18	18	104	217	256	25	19	105	211	252

	ρ	Fit_3					Fit_4					Cl.Diff.				
		$\widehat{\Sigma}_2$	$\widehat{\Sigma}_1$	$\widehat{\Sigma}_S$	$\widehat{\Sigma}_{\mathrm{Lasso}}$	$\widehat{\Sigma}_{\mathrm{EN}}$	$\widehat{\Sigma}_2$	$\widehat{\Sigma}_1$	$\widehat{\Sigma}_S$	$\widehat{\Sigma}_{\mathrm{Lasso}}$	$\widehat{\Sigma}_{\mathrm{EN}}$	$\widehat{\Sigma}_2$	$\widehat{\Sigma}_1$	$\widehat{\Sigma}_S$	$\widehat{\Sigma}_{\mathrm{Lasso}}$	$\widehat{\Sigma}_{\mathrm{EN}}$
Correct	.3	27	37	633	615	598	1	1	15	15	15	1	2	11	6	6
	.6	44	51	779	631	600	1	2	14	15	15	3	3	9	5	5
	.9	90	102	1090	420	392	5	5	10	15	15	6	6	4	4	4
VAR(3)	.3	30	37	716	527	496	1	1	15	17	16	2	2	10	6	6
	.6	46	50	814	604	533	1	2	14	16	16	3	3	8	5	5
	.9	95	102	1211	453	382	5	5	11	16	16	6	7	4	4	4
D_i	.3	76	71	109	119	110	2	2	68	71	71	4	4	5	5	5
	.6	87	77	138	158	183	3	3	61	66	67	6	5	6	5	5
	.9	144	138	346	338	285	5	5	50	58	58	7	7	6	4	4
Memo.	.3	36	39	142	114	137	1	1	3	9	10	1	1	1	2	2
	.6	64	70	334	86	86	1	1	4	14	15	2	2	2	2	2
	.9	98	100	633	105	170	2	2	3	27	32	3	3	3	3	3

covariance matrix obtained after multivariate mean removal with a rank-reduced matrix is a justified choice. The results for misspecified DGPs also point to a degradation of performance for the rank-reduced procedure, where having subject-specific matrices produces, consistently over all metrics of evaluation, the largest impact on the deterioration, followed by using long-memory processes.

Overall, estimating sequentially the autoregressive matrix of coefficients based on rank-reduced methods and the latent factorization, produces comparable or better results than those provided by sparsity-based competitors.

4.1. Clustering accuracy

We have further investigated the performance of the clustering solution based on the LOVE estimator with $\widehat{\Sigma}_2$, relative to fuzzy clustering solutions, namely fuzzy k-means (see Bezdek, 1981), fuzzy k-means with entropy regularization (Li and Mukaidono, 1995) and fuzzy k-medoids (Krishnapuram et al., 2001). We have used the implementation offered in the `R` package `fclust` (Ferraro et al., 2019) and all estimators use as input the covariance matrix $\widehat{\Sigma}_2$. Moreover, as the proposed procedure estimates a mixing matrix $\boldsymbol{A} \in \mathbb{R}^{p\times K}$, one can favour using Independent component analysis on the data matrix $\boldsymbol{X}$ to obtain a mixing matrix similar in spirit to $\boldsymbol{A}$. As such, we also compare to the overlapping clustering solution estimated by (i) sparse ICA and sparse group ICA (Wang et al., 2024) as implemented in the `R` package `SparseICA`, (ii) an ICA analysis performed on pooled data across subjects as implemented in the `R` package `iTensor`, (iii) an ICA analysis performed on pooled data across subjects after first reducing the dimensionality by using SVD and retaining the singular vectors for the analysis as suggested in Varoquaux et al. (2010) and (iv) a group ICA analysis followed by hierarchical clustering as in Esposito et al. (2005), using code from the `R` package `fmri`.

To evaluate the methods, we have defined based on pairs of features (j,k) with $1 \leq j < k \leq p$ the indicators corresponding to the true positive (TP), true negative (TN), false positives (FP) and false negative (FN) discoveries, as

$$\begin{aligned}
TP_{jk} &= \mathbf{1}\{\text{if } j,k \in G_a \text{ and } j,k \in \hat{G}_b \text{ for some } 1 \leq a \leq K \text{ and some } 1 \leq b \leq \hat{K}\}\\
TN_{jk} &= \mathbf{1}\{\text{if } j,k \notin G_a \text{ and } j,k \notin \hat{G}_b \text{ for some } 1 \leq a \leq K \text{ and some } 1 \leq b \leq \hat{K}\}\\
FP_{jk} &= \mathbf{1}\{\text{if } j,k \notin G_a \text{ and } j,k \in \hat{G}_b \text{ for some } 1 \leq a \leq K \text{ and some } 1 \leq b \leq \hat{K}\}\\
FN_{jk} &= \mathbf{1}\{\text{if } j,k \in G_a \text{ and } j,k \notin \hat{G}_b \text{ for some } 1 \leq a \leq K \text{ and some } 1 \leq b \leq \hat{K}\}
\end{aligned}$$

where G_a and K represent the true group membership and true number of clusters, while $\hat{G}_b$ and $\hat{K}$ represent the estimated counterparts.

The global metric of evaluation is the F_1 score (higher means better) defined as

$$F_1 = \frac{2 \times (\sum_{jk} TP_{jk})}{2 \times (\sum_{jk} TP_{jk}) + (\sum_{jk} FP_{jk}) + (\sum_{jk} FN_{jk})}.$$

In order to create an assignment of a feature j to a cluster b for the fuzzy competitors (i) the number of clusters is required and (ii) a cut-off value ν has to be used for the entry M_{jb}, where we generically denote the estimated fuzzy membership matrix as $M_{p\times K}$. As

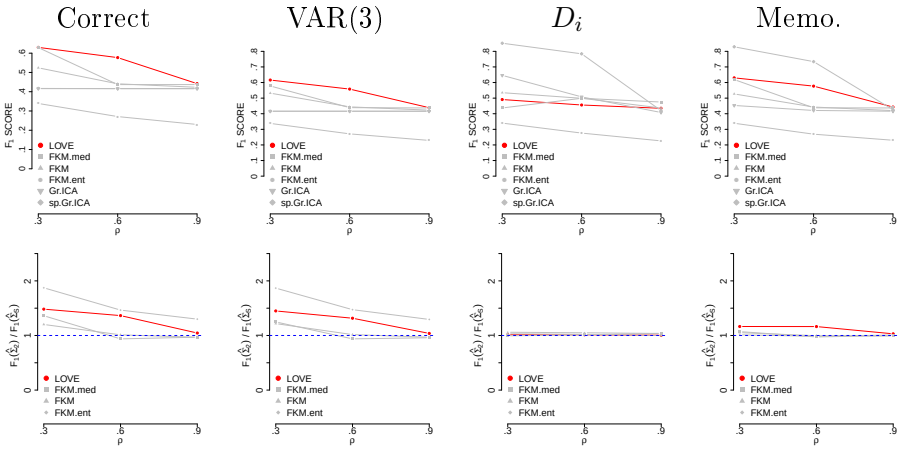


Fig. 2. Top row: F_1 scores of the LOVE-based estimator and competitors. Bottom row: ratios of the F_1 scores when using $\widehat{\Sigma}_2$ relative to using the sample covariance matrix $\widehat{\Sigma}_S$.

such, in the comparison, for the fuzzy methods we indicate the true number of clusters $K = 8$ and then search for an optimal cut-off value $\nu \in \{.001, .005, 01, \ldots, 0.3\}$ such that the F_1 score is maximal. Clearly, these two choices put the fuzzy methods at an advantage and are used as benchmark as in general, they are unknown in practice.

Figure 2 shows on the top row the averages over 1000 independent repetitions of the $F_1(\widehat{\Sigma}_2)$ score performance and indicates that for the correctly specified model and the 'slightly misspecified' VAR(3) model, both with low to moderate ρ values, the LOVE-based estimator using $\widehat{\Sigma}_2$ is at an advantage relative to the fuzzy methods and the ICA-based competitors. We show here the performance of the sparse, pooled-data ICA and the sparse group version of ICA as they were either comparable or slightly better than the other ICA methods. For high correlation i.e. $\rho = 0.9$ the performance of the three methods is generally on par. Unsurprisingly, and reinforcing the conclusion of Table 1, the misspecified DGP having subject-specific matrices proves to be the most challenging one for the LOVE method, followed by the long-memory DGP. For these two DGPs the ICA delivers superior performance.

The bottom row of Figure 2 inspects the ratio of the F_1 scores when using $\widehat{\Sigma}_2$ relative to the same F_1 score performance when using the sample covariance matrix $\widehat{\Sigma}_S$. This ratio has the purpose to evaluate if mean-removal using a low-rank approximation is effective relative to not performing mean-removal. Higher values speak in favour of mean-removal with a low-rank approximation. Similar conclusions hold: if the DGP is not severely misspecified then using $\widehat{\Sigma}_2$ proves advantageous, but for DGPs far away from the correct one, mean-removal with a rank-reduced matrix might not offer a large clustering accuracy advantage.

4.2. Mimicking real data

To further validate the use of the LOVE method on fMRI data, we have performed a last simulation study where the DGP we use, mimics the features of the observed, real data

used in Section 5. Similarly, we use here series of length $T = 1200$, focusing on brain activation of $p = 360$ ROIs grouped in $K = 6$ communities.

We describe below how data were generated. First, the matrix $\mathbf{\Gamma}$ was generated as a diagonal matrix with elements sampled from Unif[1, 2] and then Σ was set to $\Sigma = \boldsymbol{ACA}^\top + \mathbf{\Gamma}$, where to be as close as possible to the features of the real data, the model parameters $\boldsymbol{A}$ and $\boldsymbol{C}$ have both been set to the values estimated on the real dataset. Next, the true model parameter $\boldsymbol{D}$ from equation (3) has been set as well to the values estimated on the real dataset, where the rank $d = 235$ was used.

Once Σ and $\boldsymbol{D}$ have been constructed to reproduce the features of the real data, the `R` package `tsDyn` was been used to simulate independent, multivariate VAR(1) models with autoregressive matrix $\boldsymbol{D}$ and covariance matrix Σ. Data were generated for $n \in \{10, 25, 50, 100\}$ different subjects.

Figure 3 shows the empirical averages, over 100 independent replications from the DGP, for the F_1 score and it clearly sets apart LOVE from the competitors. As expected, for LOVE as n the number of subjects increases, the F_1 improves as well, however this seems not to be the case for the ICA-based competitors. We again show just the sparse ICA alternatives, as their performance was similar or better to that of the other ICA-based methods. The explanation for less than optimal performance of the sparse ICA methods is probably due to proposed selection of the regularization parameter, directly linked to the amount of sparsity in the solution. As we have used the default BIC, suggested method it seems that the final estimated solution overestimates the number of communities to which an ROI should belong, whereas the LOVE estimates come closer to the true distribution. Table 2 presents the average number of ROIs in the estimated solution that do not load on any cluster, that load on one single cluster (hence are pure ROIs as they are cluster determinants), or load on a higher number of clusters. Clearly the true distribution, determined based on the true $\boldsymbol{A}$ used in the DGP, is best-matched by the LOVE estimator using $\widehat{\Sigma}_2$ or $\widehat{\Sigma}_1$ as estimators for Σ. Performing mean-removal based on a sparse $\boldsymbol{D}$ estimated with Elastic net (with regularization levels selected by cross-validation), rather than reduced-rank methods as proposed here, also seems sub-optimal as it seems to favour larger levels of sparsity. At the opposite spectrum, the sparse ICA alternatives (with regularization levels selected by BIC) seem to favour quite dense solutions, resulting in completely shifted empirical distributions of the loading pattern and hence sub-optimal F_1 score performance.

All in all, the simulation exercise performed in this section, validates empirically that the proposed LOVE procedure has the capacity to accurately recover the overlapping structure of the clusters on data similar to the real HCP dataset, while proposing estimates that are striking a balance between being too sparse or too dense.

5. Real data: rs-fMRI analysis

We turn in this section to the HCP dataset introduced in Section 1 containing brain activation time series using the multi-modal Glasser atlas comprising $p = 360$ ROIs for $n = 406$ subjects. Out of the total number of subjects, only 395 subjects had complete observations on two separate scan sessions. For replication reasons, we consider in this section only the subjects with complete 2-sessions scans. We focus on the data acquired

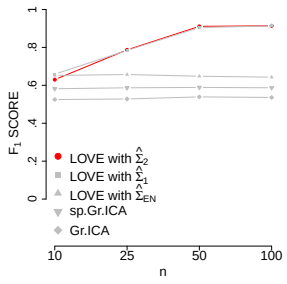


Fig. 3. Average F_1 scores of the LOVE estimator and competitors for increasing sample sizes.

Table 2. Simulated rs-fMRI data mimicking the real data. Average number of ROIs that load on 0, 1 (denoting pure nodes), 2, . . . , 6+ clusters.

Clusters	True	$\widehat{\Sigma}_2$	$\widehat{\Sigma}_1$	$\widehat{\Sigma}_{\text{EN}}$	sp.Gr.ICA	Gr.ICA
0	64	71	71	146	1	3
1	202	210	209	131	5	9
2	71	60	61	48	12	23
3	22	18	18	53	37	50
4	1	2	2	39	81	86
5	0	0	0	22	129	108
6+	0	0	0	28	94	81

at the first scanning session and consider the second session as replication data. For each of the 395 subjects, BOLD (short for 'blood oxygen level-dependent') time series are obtained where each time series contains $T = 1200$ time points. After normalizing each time series to have mean 0 and variance 1, a total of 300 subjects have been used for training, while 50 subjects are used for validation purposes and 45 subjects are used for evaluating the out-of-sample performance on new, unseen data.

To evaluate the three estimators of Σ used in Section 4, $\widehat{\Sigma}_1$, $\widehat{\Sigma}_2$ and $\widehat{\Sigma}_S$, by recalling that they all correspond to using different estimators of $\boldsymbol{D}$, we use the prediction error criterion $\text{PE}_{\text{test}} = || \boldsymbol{Y}_{\text{test}} - \boldsymbol{X}_{\text{test}}\widehat{\boldsymbol{D}}_{\text{train}} ||_F$ where $\boldsymbol{Y}_{\text{test}}$, $\boldsymbol{X}_{\text{test}}$ represent the out-of-sample

Table 3. rs-fMRI data: Out-of-sample prediction error performance and estimated number of clusters for $\widehat{\Sigma}_1$, $\widehat{\Sigma}_2$ (under Markov assumptions of order 1 and 2) and $\widehat{\Sigma}_S$. Empirical averages and standard deviations (in parentheses) across 100 random splits of the data are presented.

		PE_{test}	$\widehat{K}$	$\widehat{d}$
	$\widehat{\Sigma}_S$	4405 (0.1)	4.3 (1.0)	/
Or.1	$\widehat{\Sigma}_1$	3696 (22.4)	5.9 (1.7)	/
Or.1	$\widehat{\Sigma}_2$	3705 (30.0)	5.9 (1.7)	235
Or.2	$\widehat{\Sigma}_1$	3674 (26.1)	6.1 (1.7)	/
Or.2	$\widehat{\Sigma}_2$	3686 (31.9)	6.0 (1.7)	193

data corresponding to new, unseen subjects for which one desires a low prediction error. For $\widehat{\Sigma}_2$, the unknown rank is selected by using (10). We repeat randomly splitting the data into training/validation/test sets 100 times and on each such random split the above prediction errors have been recorded for each $\widehat{\boldsymbol{D}}_{\text{train}}$ (corresponding to $\widehat{\Sigma}_1$, $\widehat{\Sigma}_2$ and $\widehat{\Sigma}_S$). Table 3 contains the average prediction errors of all competitors across all 100 random splits of the data. It suggests that as expected the estimates provided by $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$ offered better results than the competitor since they take into account the time dependency in the data. As expected, due to the large effective sample size, $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$ have close performance, as the estimators agreed to a large extent. Moreover, for this particular dataset, using a second order Markov assumption did not provide a substantial gain in reducing the prediction errors.

The main conclusion of this section, points to the fact that if one is interested in performing brain parcelation based on covariance matrices, eliminating first the time dependency inherently linked to the mean structure, would be more beneficial in discovering the hidden clustering structure and should be advocated as a first step.

5.1. Overlapping clusters solution

In this section, we apply LOVE with $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$, as suggested by the results in the last section, using the entire data corresponding to all 395 subjects. A preliminary step was conducted to select the optimal tuning parameters and the optimal rank.

Tuning parameters (δ, λ) are selected as in Section 4, while for the rank values used in $\widehat{\Sigma}_2$, we used the average value identified in Table 3, i.e. $\widehat{d} = 235$. Since very similar results have been observed when LOVE is applied with either $\widehat{\Sigma}_1$ or $\widehat{\Sigma}_2$, while using the same set of values for the tuning parameters, we only present the results obtained from using $\widehat{\Sigma}_2$. As using a Markov assumption of order 2 did not substantially improve the prediction error metric used in Table 3, we decided to use further the first order Markov assumption and have first checked the stability of the estimated process by inspecting the modulus of the eigenvalues of $\widehat{\boldsymbol{D}_{235}}$. All values were less than 1, hence the processes is stable, therefore stationary as well, following Proposition 2.1 from Lütkepohl (2005). A follow-up residual analysis also suggested that the order 1 assumption is reasonable.

Applying LOVE to $\widehat{\Sigma}_2$ finds $\widehat{K} = 6$ overlapping clusters of ROIs. In Figure 4 we plot all ROIs from each estimated cluster (on the columns) for LOVE. The visuals were obtained using Python's `nilearn` package (see Abraham et al., 2014) by plotting ROI intensity (measured by the estimated loadings) on an fsaverage5 surface mesh (inflated version) with sulcal depth maps for both the left and right hemispheres. For each hemisphere the lateral and medial views are presented. Comparing the obtained solution with the well-known data-driven, 7-networks parcelation from Yeo et al. (2011), we observe that LOVE identifies accurately part of the Auditory network, parts of the Sylvian fissure and of the Entorhinal Cortex. Interestingly, it also identifies a 'Super' cluster formed out of large numbers of pure and non-pure variables. In this cluster, the visual and motor networks seem to be characterized by pure regions, whereas all other regions from the cluster load on multiple networks. It is unclear why this cluster is dominant in size relative to the other ones, but we conjecture it might point to a baseline status of highly interconnected ROIs in the brain. As the procedure lets the data speak for themselves and since each

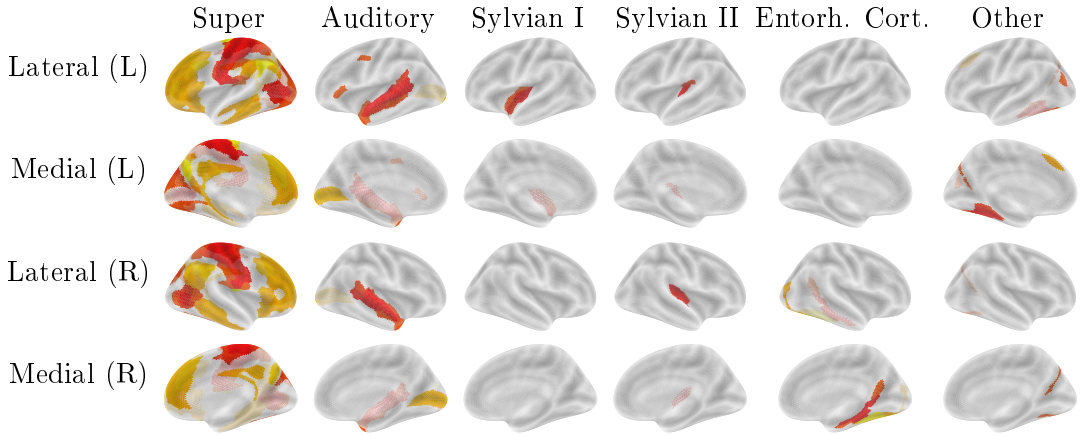


Fig. 4. LOVE with $\widehat{\Sigma}_2$ when $K = 6$ clusters (on the columns) coloured by loading intensity (yellow denotes non-pure ROIs, while red denotes pure ROIs). Lateral and medial views (on the rows) for the left (L) and correspondingly, right (R) hemisphere are presented.

ROI can load on multiple factors, it might not be surprising that the maps are more intricate and pointing to a more complex structure.

To disentangle information, we zoom-in by performing LOVE again when focusing only on the regions contained in the Super cluster. The resulting groups are plotted in Figure 5 and we identify nine clusters that accurately recover the Visual, Somatomotor and Default mode networks. The estimated clusters suggest a new finer split of these three classical networks. Moreover, the benefits of the overlapping clusters approach are well illustrated by these three networks. Both DMN II and the Motor networks identify different ROIs that are common. As well, the Visual I and DMN II/III clusters identify cluster-specific ROIs, but also point to a number of common regions. More concretely, the Visual I group in Figure 5 contains ROIs that are included in the 'Early visual' and the 'Dorsal stream visual' cortical areas identified by Glasser et al. (2016a), while Visual II and III focus almost exclusively on the 'MT+Complex' and its neighbouring area, supporting the hypothesis of more specialized, smaller networks. Moreover, the non-pure ROIs shared by Visual I and the DMN network focus mostly in the dorsolateral prefrontal and marginally in the inferior parietal cortical areas. One of the possible explanations why the dorsolateral prefrontal ROIs are non-pure and are identified both by the DMN and the Visual network, might rest on the evidence in the literature which links the dorsolateral prefrontal area to visual search for conjunctions (Kalla et al., 2009), as well as being anti-correlated with the DMN network (Bauer et al., 2020).

We have next compared the overlapping solution to estimated cortical hubs which were obtained by thresholding the absolute values of the empirical correlation matrix estimated after concatenating all subjects, at 100 values equally spaced in the interval [.01,.3] to eliminate low temporal correlation attributable to noise. Given that roughly 98% of the entries in absolute value in the empirical correlation matrix are lower than 0.35, we deem this interval appropriate for eliminating low, noise-attributed correlation.

Fig. 5. LOVE with $\widehat{\Sigma}_2$ when zooming-in on the 'Super' cluster with $K = 9$ smaller clusters (on the columns) coloured by loading intensity (yellow denotes non-pure ROIs, while red denotes pure ROIs). Lateral and medial views (on the rows) for the left (L) and correspondingly, right (R) hemisphere are presented.

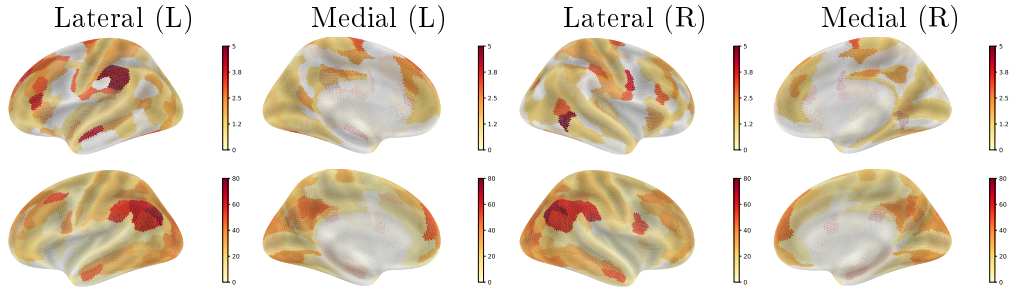


Fig. 6. Contrasting LOVE (clusters coloured by the row-wise ℓ_0 norm) - top row, with hub ROIs (coloured by node degree intensity) - bottom row. The first two and the last two columns show the lateral and medial views for the left (L) and correspondingly, right (R) hemisphere.

The values are then mapped to an adjacency matrix (containing only $0/1$ values which denote that the pairwise correlation is smaller/larger than the threshold) based on which the node degree is calculated. The larger the degree, the more 'hub'-like an ROI is. We refer to Achard et al. (2006) and Buckner et al. (2009), amongst others, for details and applications centred around hub ROIs and functional connectivity. For illustration, in Figure 6 the overlapping membership of ROIs to the latent clusters (top row) is plotted against the hub configuration with the threshold level $\nu = .25$ (bottom row). It suggests that the brain maps produced by the overlapping clustering and the network hubs are to a certain extent complementary. This complementary aspect suggests that it is not the nodes that tend to be pure which connect highly to other nodes to act as hubs, but rather the non-pure nodes exhibit the hub-like tendency to connect to other nodes. Spearman's rank correlation coefficient between the two maps was estimated at $\hat{\rho} = .3$ (p.val$< .001$). This observation contradicts the importance that is usually attributed to cortical hubs in the fMRI literature, in the sense that fuzzier ROIs (that cannot be attributed as pure nodes to a cluster) rather than cluster determinants (ie, pure ROIs) connect heavily to other nodes.

Additionally, since our approach focuses on modelling also the mean structure by estimating the transition matrix $\boldsymbol{D}$ using a low-rank estimator, one could envision a competing strategy to the approach of Achard et al. (2006) by thresholding $\hat{\boldsymbol{D}}_d$ rather than the empirical correlations between the series. Figure 7 shows the estimated Spearman's rank correlation coefficient as a function of a 100×100 grid of thresholding parameters ν_1 (used to threshold the absolute values of the entries of $\hat{\boldsymbol{D}}_d$) and ν_2 (used to threshold the absolute values of the entries of the empirical correlation matrix) where $\nu_1 \in [.01, .2]$ and $\nu_2 \in [.01, .3]$. The figure clearly suggests the existence of couples of thresholding parameters (ν_1, ν_2) for which the cortical hubs obtained by using the transition matrix match to a large extent those obtained from the empirical correlation matrix.

As ICA is routinely used in fMRI studies to unravel overlapping brain networks, we have next benchmarked our results relative to the results obtained with ICA methods. For a relevant comparison, since our method produces sparse estimates, we focused on using a sparse version of ICA proposed for fMRI analysis in Wang et al. (2024), based on the proposed implementation in the `sparseICA` package accompanying the paper. We

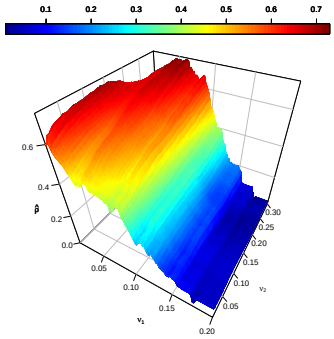


Fig. 7. Estimated Spearman's rank correlation coefficient between the maps of the cortical hubs obtained by thresholding $\hat{\boldsymbol{D}}_d$ at ν_1 and by thresholding the empirical correlation matrix at ν_2.

varied the number of components $K \in \{6, 7, 8, 9\}$ to match our solution with LOVE. Our comparison suggests that with the optimal tuning parameters selected by the BIC and the numerical convergence tolerance levels set to default values, the sparse ICA solution seems to (i) merge more heavily the distinct networks into large super-networks and to (ii) suggest much more overlapping between the networks to the point where the voxels load on much larger numbers of networks at the same time. A visualization of the solution with $K = 6$ is offered in the Appendix in Figures 14 and 15 and although the solution is a lot murkier, elements of the Visual, Motor, Auditory and DMN networks are still discernible. Similar conclusions hold also when using larger K values.

Additional comparisons to non-sparse, group or pooled-data ICA solutions, reveal similar conclusions: the competitors produce much murkier solutions where everything is overlapping, since no sparsity in imposed, but based on the magnitude of the estimates some elements of the Visual, Motor, Auditory and DMN networks are sometimes discernible, however to a far lesser degree as clear as when using LOVE. For space constraints reasons we do not present here the visuals, but they are available from the authors.

To conclude, the novelty of our findings in this section, relative to the data-driven atlas of Yeo et al. (2011), points to the fact that our proposed procedure not only recovers certain large networks such as DMN, Visual, Motor and Auditory, but also suggests a finer split of some networks than that proposed by Yeo et al. (2011). Future research might explore deeper to investigate if parts of the large brain networks are functionally distinct. Our result also indicates that such a clear separation between distinct, coarse networks is an oversimplification, as in fact some networks are in large part overlapping. Furthermore, our analysis reveals that the 'hub' behaviour of ROIs is linked to their non-pure character as opposed to being cluster determinants. Furthermore, it is noteworthy that our procedure does not impose any spatial continuity of the clusters, yet the data speak in large agreement with this characteristic for fMRI clustering. Relative to the ICA approach, the LOVE solution seems much sharper, more parsimonious and thus easier to interpret.

6. Replication on the second scanning session

Using the second scanning session we have performed a replication analysis, where using the same tuning parameters identified in Section 5.1, we fitted the LOVE algorithm to obtain overlapping clusters. See Figures 10-12 in Appendix B. Two 'Super clusters' have been obtained which matched to a large degree the Super cluster identified in Figure 4 and when zooming-in on the two clusters, we have observed a very similar amalgamation of the DMN, Visual and Motor networks, albeit a slightly different split than that of Figure 5. Despite a slightly larger fragmentation on the replication data, we conclude that the two solutions agree to a large extent. Comparing to the cortical hubs solution, similar conclusions are reached: fuzzier ROIs identified by LOVE tend to be connected to other nodes in the cortical hubs network (see Figure 13 in Appendix B); with Spearman's ρ between the two maps estimated at $\hat{\rho} = .21$ (p.val$< .001$) using the same value $\nu = .25$ as in Section 5.1. All in all, the analysis in this section points to replicability of results on a separate, second scanning session, thus giving more credit to the conclusions of Section 5.1.

7. Extensions

In this section, we propose a slight extension of model (3) by considering subject-specific transition matrices $\boldsymbol{D}^{(i)}$ rather than a single, common effects matrix. The model we consider is formulated as

$$X_t^{(i)} = \boldsymbol{D}^{(i)} X_{t-1}^{(i)} + \boldsymbol{A} Z_t^{(i)} + u_t^{(i)}, \tag{11}$$

for which the estimator $\widehat{\Sigma}_3 = \frac{1}{nT} \sum_{i=1}^{n} \left(\boldsymbol{Y}^{(i)} - \boldsymbol{X}^{(i)} \widehat{\boldsymbol{D}}_d^{(i)} \right)^\top \left(\boldsymbol{Y}^{(i)} - \boldsymbol{X}^{(i)} \widehat{\boldsymbol{D}}_d^{(i)} \right)$ is considered. Here we use the notation

$$\boldsymbol{Y}^{(i)} \in \mathbb{R}^{T \times p} = \begin{bmatrix} X_1^{(1)} \\ \vdots \\ X_T^{(1)} \end{bmatrix}, \quad \boldsymbol{X}^{(i)} \in \mathbb{R}^{T \times p} = \begin{bmatrix} X_0^{(1)} \\ \vdots \\ X_{T-1}^{(1)} \end{bmatrix}.$$

Tuning parameters (δ, λ) are selected as described in Section 4, the rank is set to $\widehat{d} = 235$ for all estimated $\boldsymbol{D}^{(i)}$ matrices, and applying LOVE to $\widehat{\Sigma}_3$ finds $\widehat{K} = 11$ overlapping clusters of ROIs. In Figure 8 we plot all ROIs from each estimated cluster and the obtained solution agrees to a large extent on the same major cortical networks as obtained in Section 5.1.

A second order Markov model has also considered as a possible extension and applying LOVE to $\widehat{\Sigma}_2$ (after mean removal) finds $\widehat{K} = 8$ overlapping clusters of ROIs. Based on the results presented in Table 3 the estimated mean rank $\hat{d} = 193$ has been used with the same search for tuning parameters as in Section 4. The results are illustrated in Figure 9 and suggest to a large degree a similar grouping of regions as the first order solution with a slightly murkier Motor network than in the previous analysis and quite close to the results obtained on the replication data. Nonetheless, the two solutions (first and second order Markov models) are judged to be qualitatively similar, indicative of a stable grouping structure.

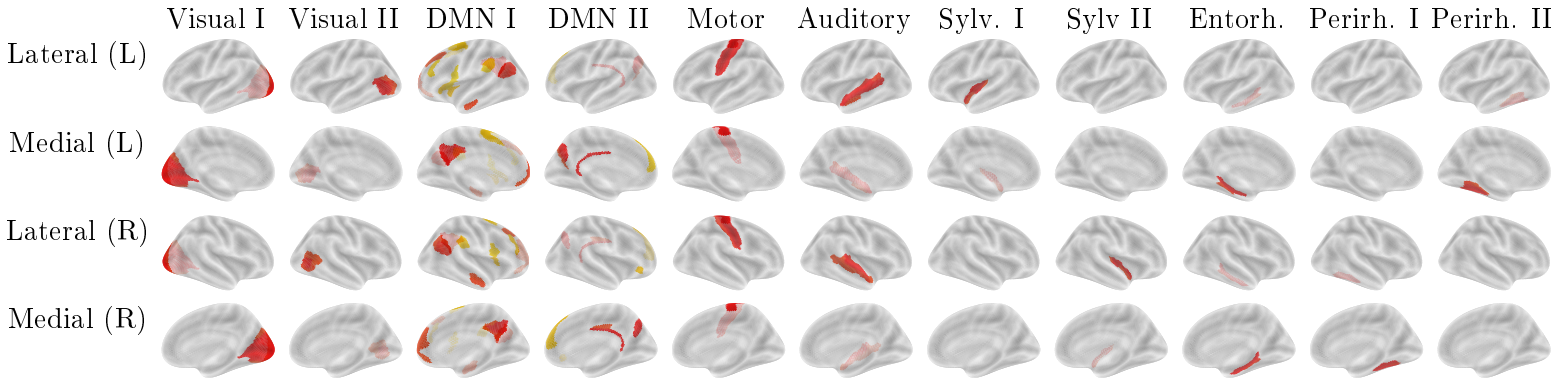


Fig. 8. LOVE with $\widehat{\Sigma}_3$ and $K = 11$ clusters (on the columns) coloured by loading intensity (yellow denotes non-pure ROIs, while red denotes pure ROIs). Lateral and medial views (on the rows) for the left (L) and correspondingly, right (R) hemisphere are presented.

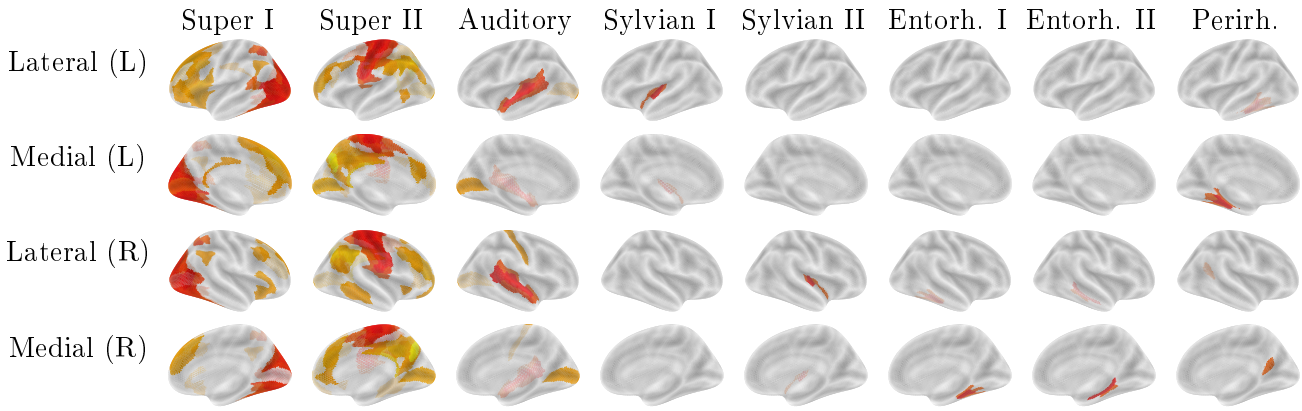


Fig. 9. Second order Markov model. LOVE with $\widehat{\Sigma}_2$ and $K = 8$ clusters (on the columns) coloured by loading intensity (yellow denotes non-pure ROIs, while red denotes pure ROIs). Lateral and medial views (on the rows) for the left (L) and correspondingly, right (R) hemisphere are presented.

8. Discussion

We have introduced in this manuscript an extension of the LOVE procedure that steps outside of the independent and identically distributed framework, to allow for high-dimensional, weakly dependent series to be jointly modelled. The goal was to identify in a resting state fMRI context, brain regions that belong to multiple networks at the same time. Traditional procedures, map one region to one brain network, whereas our framework allows more freedom as regions belong to multiple networks. Our procedure offers more support to the hypothesis that functional areas are based on overlapping regions containing neurons engaged in a common process. Clearly it introduces more flexibility in the modelling of brain activation. It also has general implications with respect to interpretations of rs-fMRI studies since for such purposes, one must take into account the presence of nearby, overlapping ROIs that are involved in multiple networks. Data driven atlases, such as MSDL (Varoquaux et al., 2011), the Yeo2011 (Yeo et al., 2011) or others alike, that associate groups of contiguous voxels to one single large, coarse structure, might be restrictive and overlooking the complexity and interdependence of overlapping fMRI activation. Our proposed procedure is model driven, thus at a distance relative to unsupervised techniques, but requires extra reflection on the problem at hand from the part of the researcher on determining the optimal level of overlap between ROIs.

Using the 'Young Adult' study from the Human Connectome Project we show that parts of well known brain networks are also identified with this methodology with the added knowledge on what ROIs play active roles across multiple networks. For example, the fact that groups of ROIs share associations with both the DMN and Motor networks, while other, distinct groups of ROIs share associations with the DMN and Visual networks, is a novel result that is worth further investigation. Controlled, simulation results are also presented and indicate that the two-step proposed procedure is better equipped at identifying latent groups of brain regions in the presence of time dependency, relative to the originally proposed method of Bing et al. (2020).

Publicly available software for the core routines of performing latent overlapping clustering can be found on the Github of X.B. (see `https://github.com/bingx1990/LOVE`) while wrapper routines for the fMRI example will be available on the webpage of E.P.

Acknowledgments

The authors thank F. Bunea, M. Wegkamp and M. R. Sabuncu for fruitful discussions. We thank as well M. Khosla for providing the pre-processed data.

Conflicts of interest

The authors declare not having any known competing interests related to the work presented in the paper.

Funding

No external funding was provided for this work.

Data availability

The data used in the paper are publicly available and can be accessed via: `https://www.humanconnectome.org/study/hcp-young-adult`.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8.

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.

Arbabshirani, M. R., Damaraju, E., Phlypo, R., Plis, S., Allen, E., Ma, S., Mathalon, D., Preda, A., Vaidya, J. G., Adali, T., and Calhoun, V. D. (2014). Impact of autocorrelation on functional connectivity. *NeuroImage*, 102:294–308.

Arslan, S., Ktena, S. I., Makropoulos, A., Robinson, E. C., Rueckert, D., and Parisot, S. (2018). Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. *NeuroImage*, 170:5–30.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Bauer, C., Rozenkrantz, L., Caballero, C., Nieto-Castanon, A., Scherer, E., West, M., Mrazek, M., Phillips, D., Gabrieli, J., and Whitfield-Gabrieli, S. (2020). Mindfulness training preserves sustained attention and resting state anticorrelation between default-mode network and dorsolateral prefrontal cortex: A randomized controlled trial. *Humman Brain Mapping*, 41(18):5356–5369.

Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press.

Bing, X., Bunea, F., Ning, Y., and Wegkamp, M. H. (2020). Adaptive estimation in structured factor models with applications to overlapping clustering. *The Annals of Statistics*, 48(4):2055–2081.

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34(4):537–541.

Buckner, R., Sepulcre, J., Talukdar, T., Krienen, F., Liu, H., Hedden, T., Andrews-Hanna, J., Sperling, R., and Johnson, K. (2009). Cortical hubs revealed by intrinsic functional connectivity: Mapping, assessment of stability, and relation to Alzheimer's disease. *The Journal of Neuroscience*, 29(6):1860–1873.

Calhoun, V. D., Liu, J., and Adalı, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage*, 45(1, Supplement 1):S163–S172.

Cecchi, G., Garg, R., and Rao, A. (2009). A cluster overlap measure for comparison of activations in fMRI studies. In Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., and Taylor, C., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, pages 1018–1025. Springer.

Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096.

Di Narzo, F. A., Aznarte, J. L., and Stigler, M. (2023). *tsDyn: Nonlinear Time Series Models with Regime Switching*. R package v. 11.0.4.1.

Downing, P. E., Wiggett, A. J., and Peelen, M. V. (2007). Functional magnetic resonance imaging investigation of overlapping lateral occipitotemporal activations using multi-

voxel pattern analysis. *Journal of Neuroscience*, 27(1):226–233.

Eickhoff, S. B., Yeo, B. T. T., Haughton, V. M., and Genon, S. (2018). Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19:672–541.

Endemann, C. M., Krause, B. M., Nourski, K. V., Banks, M. I., and Van Veen, B. (2022). Multivariate autoregressive model estimation for high-dimensional intracranial electrophysiological data. *NeuroImage*, 254:119057.

Esposito, F., Scarabino, T., Hyvarinen, A., Himberg, J., Formisano, E., Comani, S., Tedeschi, G., Goebel, R., Seifritz, E., and Di Salle, F. (2005). Independent component analysis of fMRI group studies by self-organizing clustering. *Neuroimage*, 25(1):193–205.

Ferraro, M., Giordani, P., and Serafini, A. (2019). fclust: An R package for fuzzy clustering. *The R Journal*, 11.

Fox, M. D., Zhang, D., Snyder, A. Z., and Raichle, M. E. (2009). The global signal and observed anticorrelated resting state brain networks. *Journal of Neurophysiology*, 101(6):3270–3283.

Ghebreab, S. and Smeulders, A. W. M. (2011). Identifying distributed and overlapping clusters of hemodynamic synchrony in fMRI data sets. *Pattern Analysis and Applications*, 14(2):175–192.

Glasser, M., Coalson, T., Robinson, E., Hacker, C., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C., Jenkinson, M., Smith, S., and Van Essen, D. (2016a). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178.

Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L. R., Auerbach, E. J., Behrens, T. E. J., Coalson, T. S., Harms, M. P., Jenkinson, M., Moeller, S., Robinson, E. C., Sotiropoulos, S. N., Xu, J., Yacoub, E., Ugurbil, K., and Van Essen, D. C. (2016b). The Human Connectome Project's neuroimaging approach. *Nature neuroscience*, 19(9):1175–1187.

Han, F. and Liu, H. (2013). Transition matrix estimation in high dimensional time series. *Proceedings of the 30th International Conference on Machine Learning*, 28(2):172–180.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer.

Jennings, J. R., Sheu, L. K., Kuan, D. C.-H., Manuck, S. B., and Gianaros, P. J. (2016). Resting state connectivity of the medial prefrontal cortex covaries with individual differences in high-frequency heart rate variability. *Psychophysiology*, 53(4):444–454.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on knowledge and data engineering*, 16:1370–1386.

Kalla, R., Muggleton, N., Cowey, A., and Walsh, V. (2009). Human dorsolateral prefrontal cortex is involved in visual search for conjunctions but not features: a theta TMS study. *Cortex*, 45(9):1085–1090.

Krishnapuram, R., Joshi, A., Nasraoui, O., and Yi, L. (2001). Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE Transactions on Fuzzy Systems*, 9(4):595–607.

Leschinski, C. (2019). *LongMemoryTS: Long Memory Time Series*. R package v. 0.1.0.

Li, J., Liu, Y., Wisnowski, J. L., and Leahy, R. M. (2023). Identification of overlapping and interacting networks reveals intrinsic spatiotemporal organization of the human

brain. *NeuroImage*, 270:119944.

Li, R.-P. and Mukaidono, M. (1995). A maximum-entropy approach to fuzzy clustering. *Proceedings of 1995 IEEE International Conference on Fuzzy Systems*, 4:2227–2232.

Liao, W., Mantini, D., Zhang, Z., Pan, Z., Ding, J., Gong, Q., Yang, Y., and Chen, H. (2010). Evaluating the effective connectivity of resting state networks using conditional Granger causality. *Biological Cybernetics*, 102:57–69.

Liégeois, R., Laumann, T. O., Snyder, A. Z., Zhou, J., and Yeo, B. (2017). Interpreting temporal fluctuations in resting-state functional connectivity MRI. *NeuroImage*, 163:437–455.

Liégeois, R., Yeo, B., and Van De Ville, D. (2021). Interpreting null models of resting-state functional MRI dynamics: not throwing the model out with the hypothesis. *NeuroImage*, 243:118518.

Lütkepohl, S. (2005). *New introduction to multiple time series analysis*. Springer.

Nagle, A., Gerrelts, J. P., Krause, B. M., Boes, A. D., Bruss, J. E., Nourski, K. V., Banks, M. I., and Van Veen, B. (2023). High-dimensional multivariate autoregressive model estimation of human electrophysiological data using fMRI priors. *NeuroImage*, 277:120211.

Robinson, E. C., Jbabdi, S., Glasser, M., Andersson, J., Burgess, G. C., Harms, M. P., Smith, S. M., Van Essen, D. C., and Jenkinson, M. (2014). MSM: A new flexible framework for multimodal surface matching. *NeuroImage*, 100:414–426.

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., and Smith, S. M. (2014). Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*, 90:449–468.

Saxe, R., Moran, J. M., Scholz, J., and Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *SCAN*, 1:229–234.

Shi, R. and Guo, Y. (2016). Investigating differences in brain functional networks using hierarchical covariate-adjusted independent component analysis. *The Annals of Applied Statistics*, 10(4):1930 – 1957.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn human connectome project: An overview. *NeuroImage*, 80:62–79.

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B. (2011). Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In Székely, G. and Hahn, H. K., editors, *Information Processing in Medical Imaging*, pages 562–573. Springer.

Varoquaux, G., Sadaghiani, S., Pinel, P., Kleinschmidt, A., Poline, J., and Thirion, B. (2010). A group model for stable multi-subject ICA on fMRI datasets. *NeuroImage*, 51(1):288–299.

Vazzoler, S. (2021). *sparsevar: Sparse VAR/VECM Models Estimation*. R package v. 0.1.0.

Wang, Z., Gaynanova, I., Aravkin, A., and Risk, B. (2024). Sparse independent component analysis with an application to cortical surface fMRI data in autism. *Journal of the American Statistical Association*, 119(548):2508–2520.

Wiwie, C., Baumbach, J., and Röttger, R. (2015). Comparing the performance of biomed-

ical clustering methods. *Nature Methods*, 12:1033–1038.

Xu, J., Zhang, S., Calhoun, V. D., Monterosso, J., Li, C.-S. R., Worhunsky, P. D., Stevens, M., Pearlson, G. D., and Potenza, M. N. (2013). Task-related concurrent but opposite modulations of overlapping functional networks as revealed by spatial ICA. *NeuroImage*, 79:62–71.

Yan, X., Kelley, S., Goldberg, M., and Biswal, B. (2011). Detecting overlapped functional clusters in resting state fMRI with connected iterative scan: A graph theory based clustering algorithm. *Journal of Neuroscience Methods*, 199(1):108–118.

Yeo, B., Krienen, F., Sepulcre, J., Sabuncu, M., Lashkari, D., Hollinshead, M., Roffman, J., Smoller, J., Zollei, L., Polimeni, J., Fischl, B., Liu, H., and Buckner, R. (2011). The organization of the human cerebral cortex estimated by functional correlation. *Journal of neurophysiology*, 106:1125–1165.

Appendix A: The LOVE algorithm

Algorithm 1 : Estimate the partition of the pure variables $\mathcal{I}$ by $\widehat{\mathcal{I}}$

procedure PUREVAR($\widehat{\Sigma}$, δ)
  $\widehat{\mathcal{I}} \leftarrow \varnothing$.
  for all $i \in [p]$ **do**
    $\widehat{I}^{(i)} \leftarrow \left\{l \in [p] \setminus \{i\} : \max_{j \in [p]\setminus\{i\}} |\widehat{\Sigma}_{ij}| \leq |\widehat{\Sigma}_{il}| + 2\delta\right\}$
    $Pure(i) \leftarrow True$.
    for all $j \in \widehat{I}^{(i)}$ **do**
      if $\left||\widehat{\Sigma}_{ij}| - \max_{k \in [p]\setminus\{j\}} |\widehat{\Sigma}_{jk}|\right| > 2\delta$ **then**
        $Pure(i) \leftarrow False$,
        break
    if $Pure(i)$ **then**
      $\widehat{I}^{(i)} \leftarrow \widehat{I}^{(i)} \cup \{i\}$
      $\widehat{\mathcal{I}} \leftarrow$ MERGE($\widehat{I}^{(i)}$, $\widehat{\mathcal{I}}$)
  return $\widehat{\mathcal{I}}$ and $\widehat{K}$ as the number of sets in $\widehat{\mathcal{I}}$

function MERGE($\widehat{I}^{(i)}$, $\widehat{\mathcal{I}}$)
  for all $G \in \widehat{\mathcal{I}}$ **do** ▷ $\widehat{\mathcal{I}}$ is a collection of sets
    if $G \cap \widehat{I}^{(i)} \neq \varnothing$ **then**
      $G \leftarrow G \cap \widehat{I}^{(i)}$ ▷ Replace $G \in \widehat{\mathcal{I}}$ by $G \cap \widehat{I}^{(i)}$
      return $\widehat{\mathcal{I}}$
  $\widehat{I}^{(i)} \in \widehat{\mathcal{I}}$ ▷ add $\widehat{I}^{(i)}$ in $\widehat{\mathcal{I}}$
  return $\widehat{\mathcal{I}}$

Algorithm 2 : The LOVE procedure for overlapping clustering.

Require: $\widetilde{\Sigma}$ constructed in (5) or (8), the tuning parameters δ, λ and μ.

1: Apply Algorithm 1 to obtain the number of clusters $\widehat{K}$, the estimated set of pure variables $\widehat{I}$ and its partition of $\widehat{\mathcal{I}}$.
2: Estimate A_I by $\widehat{A}_{\widehat{I}}$.
3: Estimate C^{-1} by $\widehat{\Omega}$ and $\bar{\beta}^j$ for each $j \in \widehat{J}$.
4: Estimate A_J by $\widehat{A}_{\widehat{J}}$. Combine $\widehat{A}_{\widehat{I}}$ with $\widehat{A}_{\widehat{J}}$ to obtain $\widehat{A}$.
5: Estimate overlapping groups $\widehat{\mathcal{G}} = \{\widehat{G}_1, ..., \widehat{G}_{\widehat{K}}\}$ by using $\widehat{A}$ and output $\widehat{A}$, $\widehat{\mathcal{G}}$.

Appendix B: Extra figures

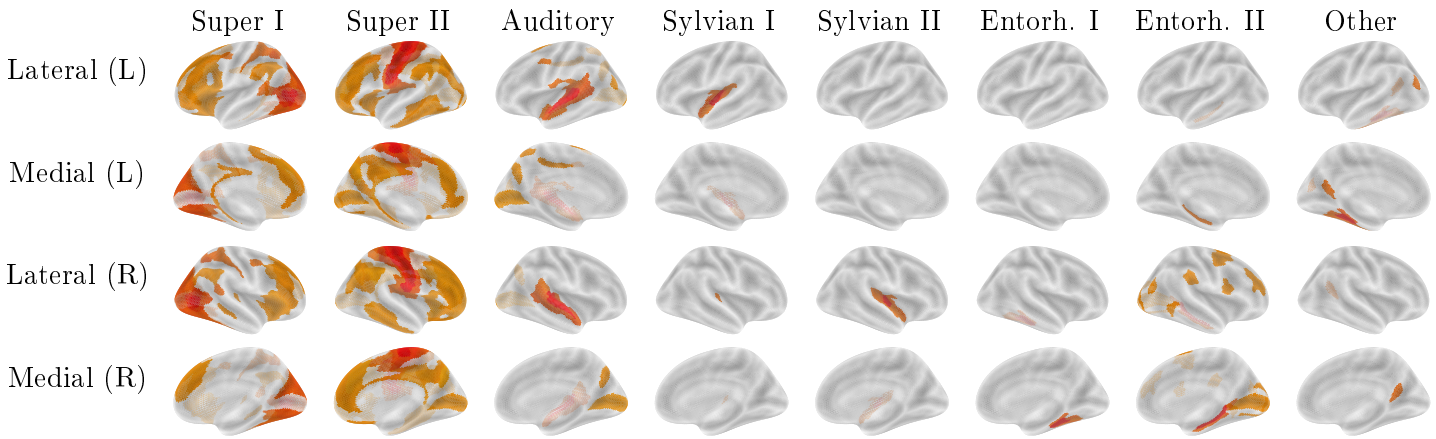


Fig. 10. LOVE on replication data with $\widehat{\Sigma}_2$ and $K = 8$ clusters coloured by loading intensity (yellow denotes non-pure ROIs, while red denotes pure ROIs). Lateral and medial views are presented.

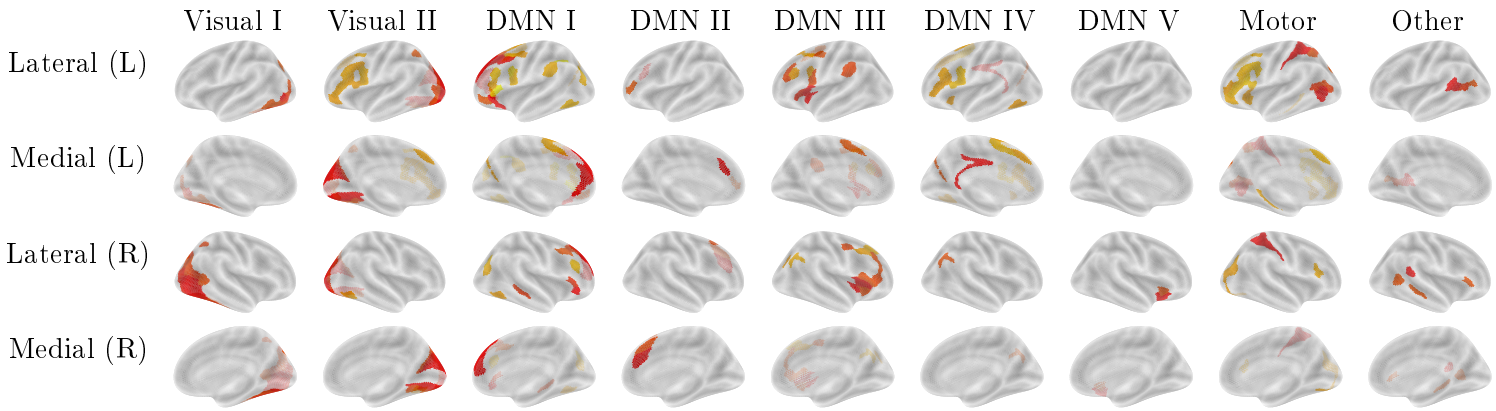


Fig. 11. LOVE on replication data with $\widehat{\Sigma}_2$ when zooming-in on the 'Super' cluster I with $K = 9$ smaller clusters coloured by loading intensity (yellow denotes non-pure ROIs, while red denotes pure ROIs). Lateral and medial views are presented.

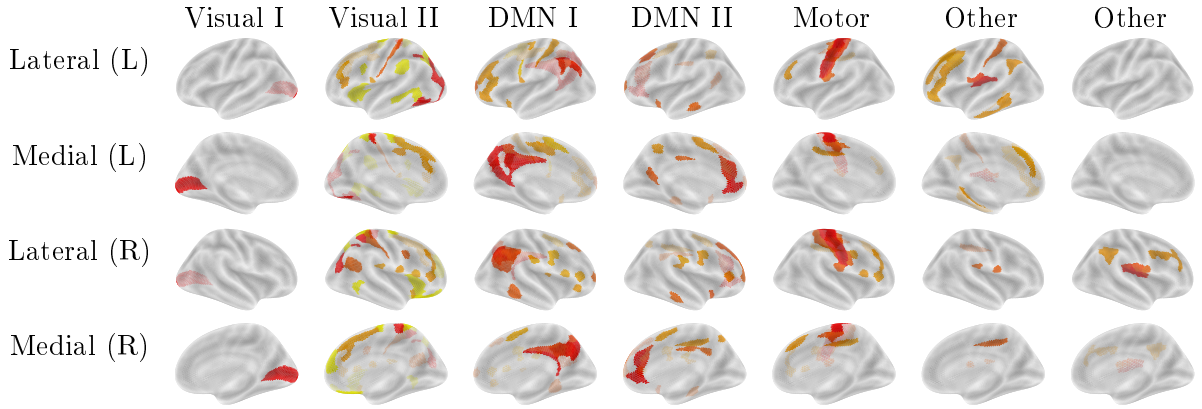


Fig. 12. LOVE on replication data with $\widehat{\Sigma}_2$ when zooming-in on the 'Super' cluster II with $K = 7$ smaller clusters coloured by loading intensity (yellow denotes non-pure ROIs, while red denotes pure ROIs). Lateral and medial views are presented.

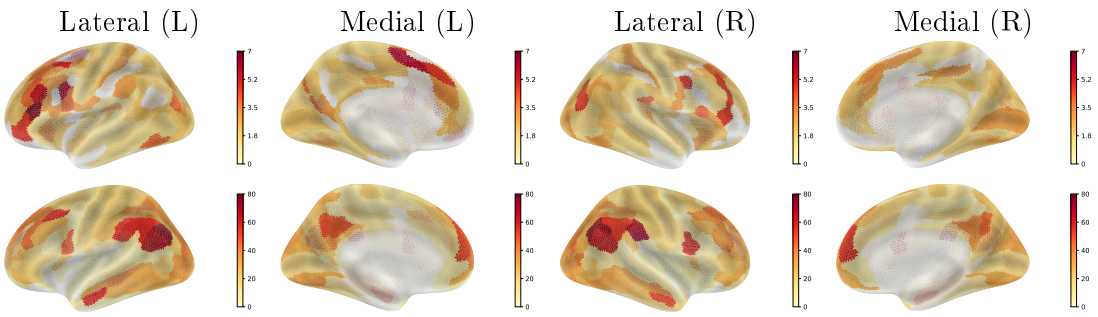


Fig. 13. Contrasting LOVE (clusters coloured by the row-wise ℓ_0 norm) - top row, with hub ROIs (coloured by node degree intensity) - bottom row, on replication data. Lateral and medial views for the left (L) and right (R) hemispheres are shown.

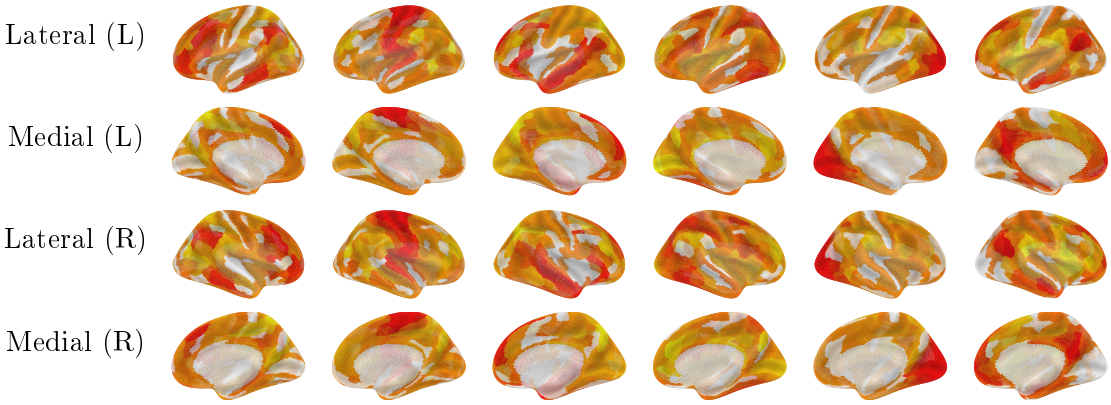


Fig. 14. Sparse ICA on pooled data when $K = 6$ clusters (on the columns) coloured by mixing intensity (yellow denotes low values, while red denotes high values). Lateral and medial views (on the rows) for the left (L) and correspondingly, right (R) hemisphere are presented.

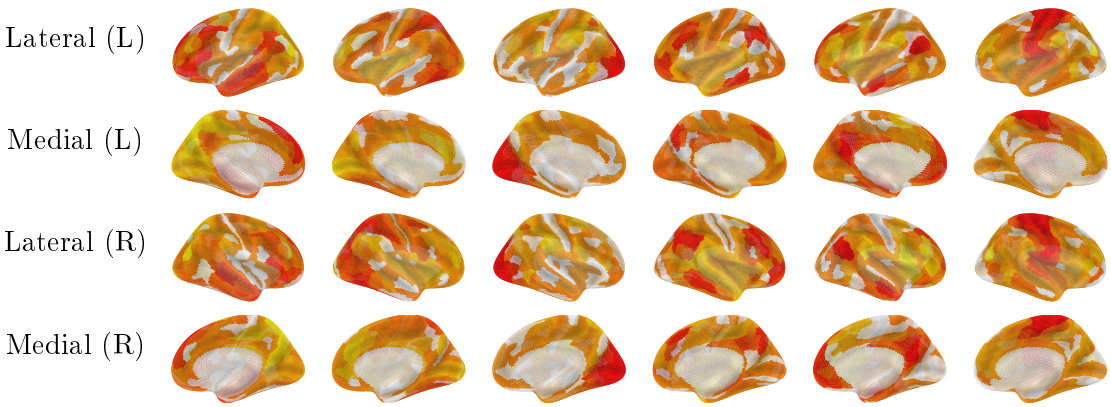


Fig. 15. Sparse group ICA when $K = 6$ clusters (on the columns) coloured by mixing intensity (yellow denotes low values, while red denotes high values). Lateral and medial views (on the rows) for the left (L) and correspondingly, right (R) hemisphere are presented.