Sparse coupled logistic regression to estimate co-activation and modulatory influences of brain regions

Thomas A. W. Bolton[1,2] **& Dimitri Van De Ville**[1,2]

[1] Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[2] Department of Radiology and Medical Informatics, University of Geneva (UNIGE), Geneva, Switzerland

E-mail: `thomas.bolton@epfl.ch`

November 2019

Abstract. Please magically appear =)

Keywords: dynamic functional connectivity, effective connectivity, logistic regression, ℓ_1 regularisation

Submitted to: *J. Neural Eng.*

1. Introduction

Understanding the structural wiring of the brain at its most global scale, and how information flows between remote processing centres, are essential questions to shed light on higher-order behaviours involving multi-modal integration and associated brain disorders. When it comes to functional magnetic resonance imaging (fMRI), the mapping of brain function is commonly performed from resting-state (RS) recordings through the computation of *functional connectivity* (FC), that is, the statistical interdependence between different time courses reflective of regional activity (Friston 1994), as can be assessed from an array of measures (Smith, Miller, Salimi-Khorshidi, Webster, Beckmann, Nichols, Ramsey & Woolrich 2010). This approach has revealed the presence of a set of RS networks (RSNs) (Damoiseaux, Rombouts, Barkhof, Scheltens, Stam, Smith & Beckmann 2006, Power, Fair, Schlaggar & Petersen 2010, Yeo, Krienen, Sepulcre, Sabuncu, Lashkari, Hollinshead, Roffman, Smoller, Zöllei, Polimeni et al. 2011), whose properties are critical landmarks of brain function and cognition (Bressler & Menon 2010, van den Heuvel & Hulshoff Pol 2010).

Over the past decade, it has become increasingly clear that quantifying FC between two brain regions as one scalar for a full scanning session is an overly simplistic approach that does not characterise the numerous reconfigurations that occur at the time scale of seconds (Chang & Glover 2010). Accordingly, many methodological pipelines have been developed to dig into time-resolved FC, and map brain function dynamically (see (Preti, Bolton & Van De Ville 2017, Lurie, Kessler, Bassett, Betzel, Breakspear, Keilholz, Kucyi, Liégeois, Lindquist & McIntosh 2018) for contemporary reviews).

The most notorious family of dynamic approaches simplifies the originally voxel-wise fMRI data into a state-level representation: first, FC is computed over successive temporal sub-windows, and the concatenated data across the full subject population at hand is subjected to hard clustering to yield a set of dynamic FC (dFC) states (Allen, Damaraju, Plis, Erhardt, Eichele & Calhoun 2014, Damaraju, Allen, Belger, Ford, McEwen, Mathalon, Mueller, Pearlson, Potkin, Preda, Turner, Vaidya, van Erp & Calhoun 2014). Because spatial Independent Component Analysis (ICA) is typically performed prior to clustering, each state stands for a set of RSNs showing specific correlational relationships.

In other analytical schemes, whole-brain voxelwise activity (Liu, Chang & Duyn 2013), or activity transients (Karahanoğlu & Van De Ville 2015), undergo clustering instead of FC patterns; in this case, each of the retrieved centroids directly stands for an RSN. If temporal ICA is applied after spatial ICA, temporally mutually independent RSNs are retrieved (Smith, Miller, Moeller, Xu, Auerbach, Woolrich, Beckmann, Jenkinson, Andersson, Glasser et al. 2012). Finally, the use of a hidden Markov model (HMM) also enables to derive RSNs, as represented under the form of (sparse) FC patterns (Eavani, Satterthwaite, Gur, Gur & Davatzikos 2013, Vidaurre, Smith & Woolrich 2017) or

vectors of activation (Chen, Langely, Chen & Hu 2016).

In all the above cases, there is the underlying assumption that the raw fMRI data can be downscaled to a set of RSNs, and that the dynamics of brain function should be understood from this simplified starting point. Recent results, however, question the validity of this assumption: for instance, some brain regions do not remain attached to the same network throughout a scanning session, but instead adjust their modular allegiance over time in a way that relates to cognitive performance (Chen, Cai, Ryali, Supekar & Menon 2016, Pedersen, Zalesky, Omidvarnia & Jackson 2018). In addition, brain regions or networks also morph spatially over time (Kiviniemi, Vire, Remes, Elseoud, Starck, Tervonen & Nikkinen 2011, Kottaram, Johnston, Ganella, Pantelis, Kotagiri & Zalesky 2018, Iraji, Fu, Damaraju, DeRamus, Lewis, Bustillo, Lenroot, Belger, Ford, McEwen et al. 2019).

To capture these spatially more subtle reconfigurations, novel methodologies have attempted to operate at the regional scale, and the assessment of *causal* relationships (*i.e.*, from time t to $t + 1$) between distinct areas showed particular merits as an alternative conceptualisation of RS functional brain dynamics, be it through autoregressive models (Liégeois, Laumann, Snyder, Zhou & Yeo 2017, Lennartz, Schiefer, Rotter, Hennig & LeVan 2018) or Ornstein-Uhlenbeck processes (Gilson, Moreno-Bote, Ponce-Alvarez, Ritter & Deco 2016).

At present, there are thus two conceptually discrepant ways to view RS dFC: on the one hand, expressing it as sets of simultaneously activating regions that make networks, and on the other hand, viewing it as effective connectivity between individual areas. It remains to be determined which of these two viewpoints offers the best representation of RS dynamics, and whether they describe overlapping or distinct facets of the data.

In this work, we have attempted to progress in answering these questions by developing a novel methodological framework that jointly estimate sets of co-activations, and causal couplings, between individual brain regions. A dedicated parameter also enables to modulate the trade-off in data fitting between these two viewpoints.

2. Materials and Methods

2.1. Mathematical framework

Let us denote the activity of a region r (out of R in total) at time t as $h_t^{(r)}$. We hypothesise two possible states of activity: *baseline* ($h_t^{(r)} = 0$) or *active* ($h_t^{(r)} = +1$). Each region may interact with all the other areas $s \neq r$ in two ways: (1) showing simultaneous activity (that is, episodes of co-activation), or (2) being causally modulated. To jointly describe these two phenomena, we characterise the probability of a region r to switch between activity states as a logistic regression (Friedman, Hastie & Tibshirani 2010):

$$\begin{cases} \mathcal{P}(h_{t+1}^{(r)} = +1 | h_t^{(r)} = 0, \mathbf{h}_t^{(-\mathbf{r})}, \mathbf{h}_{t+1}^{(-\mathbf{r})}) = \frac{1}{1+e^{-(\alpha_A^{(r)} + \boldsymbol{\gamma}_A^{(r)\top} \mathbf{h}_{t+1}^{(-\mathbf{r})} + \boldsymbol{\beta}_A^{(r)\top} \mathbf{h}_t^{(-\mathbf{r})})}} \\ \mathcal{P}(h_{t+1}^{(r)} = 0 | h_t^{(r)} = +1, \mathbf{h}_t^{(-\mathbf{r})}, \mathbf{h}_{t+1}^{(-\mathbf{r})}) = \frac{1}{1+e^{-(\alpha_D^{(r)} + \boldsymbol{\gamma}_D^{(r)\top} \mathbf{h}_{t+1}^{(-\mathbf{r})} + \boldsymbol{\beta}_D^{(r)\top} \mathbf{h}_t^{(-\mathbf{r})})}} \end{cases} . \quad (1)$$

The baseline-to-active transition is modelled by the first equation, while the return to baseline from an active state is governed by the second. Associated coefficients are respectively written with the $\cdot_A$ and $\cdot_D$ subscripts. In what follows, for the sake of clarity, we will omit these subscripts and only show one set of equations, as the formulations are strictly equivalent for both types of transitions.

If all other regions are at a baseline level of activity at the start ($\mathbf{h}_t^{(-\mathbf{r})} = \mathbf{0}$) and end ($\mathbf{h}_{t+1}^{(-\mathbf{r})} = \mathbf{0}$) of the transition, only the scalar coefficient $\alpha^{(r)}$ plays a role in shaping the transition likelihood. The vector $\boldsymbol{\gamma}^{(r)} \in \mathbb{R}^{R-1}$ contains the co-activation coefficients for all regions $s \neq r$: positive-valued coefficients will enhance the likelihood of the transition of interest if $h_{t+1}^{(s)} = +1$ (that is, if regions r and s are co-active at time $t+1$). Negative-valued coefficients will, likewise, reduce the transition probability. The reasoning is similar for the vector $\boldsymbol{\beta}^{(r)} \in \mathbb{R}^{R-1}$, except that a modulatory effect is then exerted if $h_t^{(s)} = +1$ (*i.e.*, region s is active before the transition, resulting in a causal modulation instead of a co-activation).

If the above pair of equations is considered for each brain region, the resulting coefficients can be arranged in two types of matrices, where the r^{th} column contains the influences onto region r (diagonal elements are left empty): one type is reflective of co-activations, which we be termed $\boldsymbol{\Gamma}$, and one symbolises causal modulations, and will be referred to as $\mathbf{B}$. $\boldsymbol{\Gamma}$ and $\mathbf{B}$ can respectively be interpreted as equivalents of the functional connectome and effective connectome. An overview of our framework is provided in Figure 1.

The concomitant modelling of co-activations and causal modulations enables to jointly derive the two sets of coefficients. Given the fact that the resting brain is often described as a series of RSNs (Damoiseaux et al. 2006, Power et al. 2010, Yeo et al. 2011), we expect $\boldsymbol{\Gamma}$ to only contain a sparse subset of non-null entries. Similarly, only a restricted amount of areas or networks are expected to causally modulate each other (Christoff,

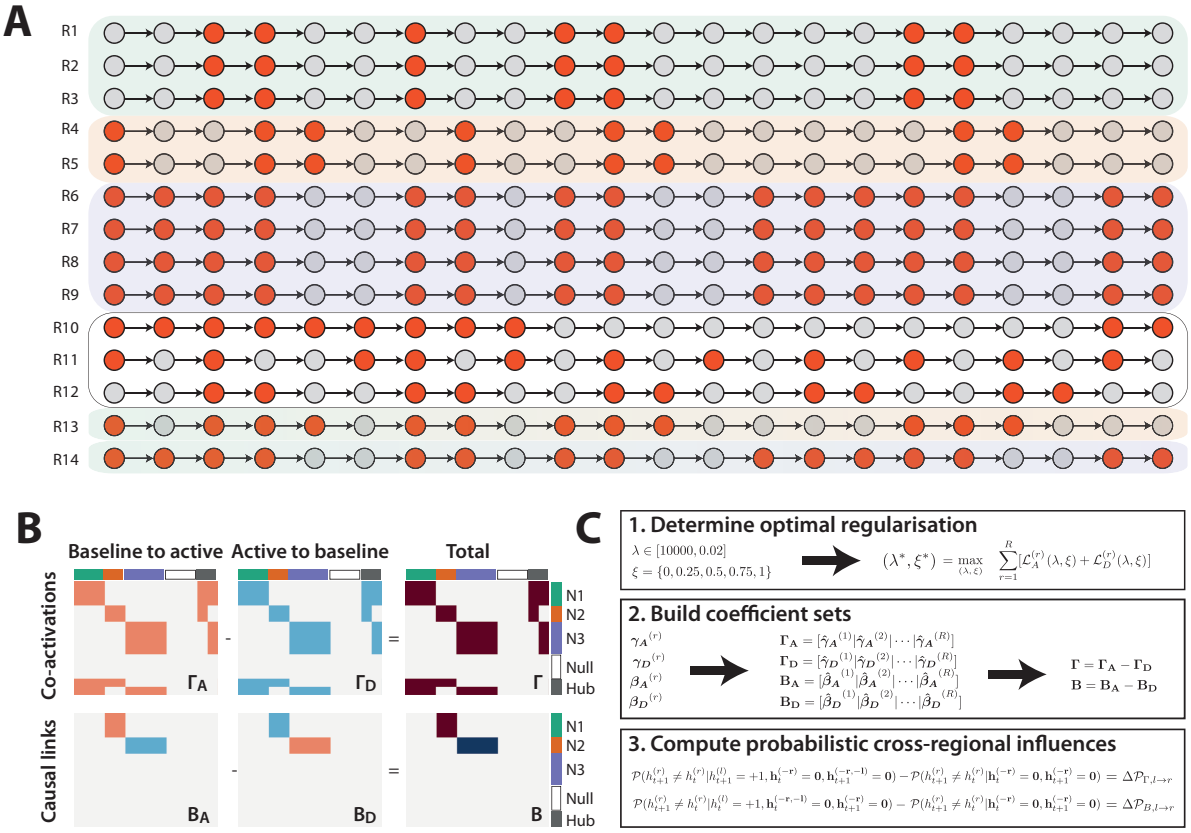

Figure 1. Overview of the framework. (A) Example activity time courses for a set of 14 regions; each can transit between a baseline state of activity (symbolised by a grey circle) and an active state (red circle). The green, salmon and blue underlays highlight the regions that belong to the same RSN, and thus exhibit a similar transitory dynamics. Regions 10 to 12 evolve according to their own dynamics, which are independent from all the others. As for regions 13 and 14, they are *hubs* that belong to two networks at a time (as rendered by the mixed colour underlay), and thus turn active as soon as one of their affiliated networks does so. **(B)** Coefficient matrices associated to the example presented in (A) for co-activations (top row) and causal modulations (bottom row). The left column pertains to the transition from the baseline to the active state: a positive-valued coefficient at l, r means that when region l is active, it enhances the likelihood of a transition for region r at the same time point (for co-activations) or one time point later (for causal modulations). The middle column similarly characterises transitions from the active to the baseline state; thus, modulations that enhance the overall activity of an area are here reflected by negative-valued coefficients (*i.e.*, the probability to go down in activity is lowered). The right column yields total influences summed across both transition types. **(C)** To solve the framework, optimal regularisation parameters λ and ξ are first determined by extracting local maxima of the full likelihood across regions and transition types (top box). Then, co-activation and causal coefficients are computed for each region r and transition type (middle box). Finally, the likelihood to switch activity state can be compared with and without an external region's influence, to compute a pair-wise probabilistic modulation coefficient (bottom box). R: region. N: network.

Irving, Fox, Spreng & Andrews-Hanna 2016, Bolton, Tarun, Sterpenich, Schwartz & Van De Ville 2017). To fit these neurobiological priors, while also enabling convergence of the framework with fewer data points, we appended an ℓ_1 regularisation term:

$$\xi||\boldsymbol{\gamma}^{(r)}||_1 + (1-\xi)||\boldsymbol{\beta}^{(r)}||_1 < \rho \quad \forall \quad r = 1, ..., R. \tag{2}$$

In the above, the parameter ρ controls the extent of regularisation casted on all coefficients (it is associated to an inversely proportional parameter λ in the optimisation equation detailed below). The parameter ξ enables to balance the extent with which the

co-activation and causal sets are regularised: if $\xi = 0$, regularisation only operates on causal coefficients, while if $\xi = 1$, only co-activation coefficients are made sparse. This respectively amounts to a description of regional brain dynamics where co-activations, or causal influences, dominate.

2.2. Implementation

Solving the above set of coupled logistic regression equations requires that the activity levels of all regions be known. To binarise the input time courses, we individually z-score each, and set to 1/0 the time points with a value above/below 0. While binarisation may remove part of the insightful information from the original data, it has been used in recently developed methodological pipelines (Kang, Pae & Park 2019). In the discussion, we touch upon possibilities to make the framework amenable to a case with more than 2 states of activity.

After defining the activation states, initial parameter estimates can be computed. Co-activation and modulatory coefficients are all set to 0, and intrinsic transition probabilities are estimated by a standard HMM approach (Rabiner 1989).

Following (Friedman et al. 2010), in a regularised logistic regression, one attempts to solve the following:

$$\min_{\alpha^{(r)},\boldsymbol{\gamma}^{(r)},\boldsymbol{\beta}^{(r)}} -\mathcal{L}^{(r)}(\alpha^{(r)},\boldsymbol{\gamma}^{(r)},\boldsymbol{\beta}^{(r)}) + \lambda(\xi||\boldsymbol{\gamma}^{(r)}||_1 + (1-\xi)||\boldsymbol{\beta}^{(r)}||_1), \tag{3}$$

where r is the assessed region, and the log-likelihood is approximated as:

$$\mathcal{L}^{(r)}(\alpha^{(r)},\boldsymbol{\gamma}^{(r)},\boldsymbol{\beta}^{(r)}) = -\frac{1}{2|\mathcal{T}|}\sum_{t\in\mathcal{T}}\omega_t(z_t - \alpha^{(r)} - \boldsymbol{\gamma}^{(r)\top}\mathbf{h}_{t+1}^{(-\mathbf{r})} - \boldsymbol{\beta}^{(r)\top}\mathbf{h}_t^{(-\mathbf{r})}) + C. \tag{4}$$

The ensemble $\mathcal{T}$ contains all the data points for which the probed region is in the start state of interest at time t (*e.g.*, baseline for the baseline-to-active transitions), and C is a constant. If we denote the probability of the transition of interest as $p(\alpha^{(r)},\boldsymbol{\gamma}^{(r)},\boldsymbol{\beta}^{(r)},\mathbf{h}_t^{(-\mathbf{r})},\mathbf{h}_{t+1}^{(-\mathbf{r})})$, the parameters ω_t and z_t depend on the current estimates of the coefficients—which we denote with a tilda—as:

$$\begin{cases} \omega_t = p(\tilde{\alpha}^{(r)},\tilde{\boldsymbol{\gamma}}^{(r)},\tilde{\boldsymbol{\beta}}^{(r)},\mathbf{h}_t^{(-\mathbf{r})},\mathbf{h}_{t+1}^{(-\mathbf{r})}) - p(\tilde{\alpha}^{(r)},\tilde{\boldsymbol{\gamma}}^{(r)},\tilde{\boldsymbol{\beta}}^{(r)},\mathbf{h}_t^{(-\mathbf{r})},\mathbf{h}_{t+1}^{(-\mathbf{r})})^2 \\ z_t = \tilde{\alpha}^{(r)} + \tilde{\boldsymbol{\gamma}}^{(r)\top}\mathbf{h}_{t+1}^{(-\mathbf{r})} + \tilde{\boldsymbol{\beta}}^{(r)\top}\mathbf{h}_t^{(-\mathbf{r})} + \frac{y_t - p(\tilde{\alpha}^{(r)},\tilde{\boldsymbol{\gamma}}^{(r)},\tilde{\boldsymbol{\beta}}^{(r)},\mathbf{h}_t^{(-\mathbf{r})},\mathbf{h}_{t+1}^{(-\mathbf{r})})}{\omega_t} \end{cases}. \tag{5}$$

y_t defines whether there was a change in activity level from time t to $t+1$ or not (respectively, $y_t = 1$ or $y_t = 0$). Coefficients are iteratively estimated by a coordinate-wise descent algorithm, following (Friedman, Hastie, Höfling, Tibshirani et al. 2007): the initial estimates outlined above are used at the maximal regularisation level λ_{MAX}, and individual coefficients are successively re-estimated in random order (note that for $\alpha^{(r)}$ coefficients, which do not enter the ℓ_1 regularisation term, soft shrinkage is not required). The process continues until the change across two iterations becomes lower

than a defined tolerance threshold ϵ. The next regularisation level is then considered, using warm restarts to speed up computations (*i.e.*, the estimates obtained at the end of a regularisation cycle are used as initial values for the following one).

In all the analyses performed in this work, we considered a regularisation path with $\lambda \in [10000, 0.02]$ (206 logarithmically distributed values), compared five levels of trade-off between co-activation and causal coefficients ($\xi = \{0, 0.25, 0.5, 0.75, 1\}$), and used a tolerance $\epsilon = 10^{-40}$.

2.3. Validation of the framework on simulated data

We first sought to validate our pipeline on simulated data containing cross-regional causal modulations as well as co-activations. To do so, we considered parameters resembling those of the assessed experimental data (see the following section) as much as possible. We simulated activity time courses for $R = 45$ regions, for a total of $S = 135$ subjects and $T = 1190$ time points per subject.

To design our simulations in accordance with the RS literature (Yeo et al. 2011), we considered the presence of $N = 7$ separate RSNs, each of which could contain between 4 and 7 areas. Time courses for all regions belonging to the same network were similar (prior to the addition of noise). In addition, we also included a set of areas evolving according to their own, independent dynamics; since in such a setting, no co-activation or causal coefficients should be retrieved, these regions can be regarded as a negative control. Furthermore, a few regions were also set as *hubs* that jointly belong to two networks, and activate as soon as one of the networks turns on. Figure 2C (top left matrix) shows the ground truth co-activation relationships between the set of simulated regions.

Each simulated dynamics was associated to a probability to switch from the baseline to the active state, selected uniformly in the [0.2,0.5] interval. Similarly, the probability to transit from the active to the baseline state was uniformly selected in the [0.7,0.9] interval. Causal modulations were introduced between a subset of networks, as summarised in Figure 2C (top right matrix): when a modulating network turned active, it could enhance the activity of the modulated network (both by enhancing the likelihood of a 0 to +1 transition, and reducing that of a +1 to 0 one), as symbolised by a positive-valued causal coefficient, or decrease that activity, as reflected by a negative-valued element. We used a shift in transition probability $\Delta P = 0.6$.

Eventually, all time courses were corrupted with Gaussian noise, at standard deviation $\sigma = 2$; indicative time courses for a simulated subject are presented in Figure 2A, where noise is sufficient not to be able to infer any cross-regional relationships by mere eyesight.

To assess the ability of the framework to recover the ground truth, we computed Pearson's spatial correlation coefficient between ground truth and estimated coefficients,

separately for the co-activation and causal sets, and contrasted these similarity measures to the evolution of the log-likelihood of the data. In addition, we examined whether the information contained in the co-activation coefficients was sufficient to re-order the regions into their underlying networks, by computing Ward's linkage on probabilistic co-activation values (see Figure 1C, bottom box).

2.4. Application of the framework to experimental fMRI data

We applied our framework to experimental RS fMRI data from the *Human Connectome Project* (Van Essen, Smith, Barch, Behrens, Yacoub & Ugurbil 2013). We considered one scanning session long of $T = 1190$ time points for $S = 135$ subjects. The data was acquired at a fast TR of 720 ms, at a spatial resolution of $2 \times 2 \times 2$ mm^3; additional acquisition details can be found elsewhere (Smith, Beckmann, Andersson, Auerbach, Bijsterbosch, Douaud, Duff, Feinberg, Griffanti, Harms et al. 2013).

We started from the publicly available minimally preprocessed data. Each voxel-wise time course was detrended, and constant, linear and quadratic trends were regressed out at the same time as a Discrete Cosine Transform basis (cutoff frequency: 0.01 Hz). We chose not to perform global signal regression, since it remains a debated preprocessing step (Murphy & Fox 2017), and in light of recent results showing extensive relationships between spatio-temporal motion patterns and human behaviour (Bolton, Zöller, Caballero-Gaudes, Kebets, Glerean & Van De Ville 2019), also decided not to include individual motion time course regressors (note that motion is handled by conservative scrubbing at a later stage of the pipeline—see below).

Voxel-wise time courses were averaged into 90 regions of interest defined from the AAL atlas (Tzourio-Mazoyer, Landeau, Papathanassiou, Crivello, Etard, Delcroix, Mazoyer & Loliot 2002); although more accurate parcellation schemes have been introduced (Glasser, Coalson, Robinson, Hacker, Harwell, Yacoub, Ugurbil, Andersson, Beckmann, Jenkinson et al. 2016, Schaefer, Kong, Gordon, Laumann, Zuo, Holmes, Eickhoff & Yeo 2017), they involve a larger amount of brain regions and would thus require an amount of data larger than the available one for accurate estimation. As the main goal of the present report is the introduction of our framework, rather than its application to neurobiologically relevant questions, we opted to operate at the smaller AAL scale.

As a final preprocessing step, scrubbing was performed at a framewise displacement threshold (Power, Barnes, Snyder, Schlaggar & Petersen 2012) of 0.3 mm, and discarded frames were re-estimated by cubic spline interpolation.

To assess the reproducibility of our findings, we separately applied our framework to each hemisphere of the brain; in each case, co-activations and causal modulations were thus estimated between $R = 45$ separate areas.

3. Results

3.1. Validation of the framework on simulated data

Figure 2 displays the results of our simulations. Around the largest regularisation extents ($\lambda_1 = 9000$), the log-likelihood was low regardless of the balance between the regularisation of co-activation and causal coefficients, and this was associated to overall low similarity to the ground truth transition probability modulation patterns (Figure 2B), an unsurprising feature given that probabilistic modulation coefficients were then extremely sparse, or (for the less regularised set) randomly distributed (see λ_1 cases in Figure 2C).

When regularisation decreased (*i.e.*, going to the left in Figure 2B plots), the log-likelihood remained low when regularisation was principally casted on causal modulations (see the orange and purple curves in the top plot); as seen in the associated coefficient matrices from Figure 2C, this is because many erroneous coefficients still populated the co-activation set, which is the dominating factor in the simulated data. Log-likelihood was more elevated for the schemes that favoured sparsity of co-activation coefficients (red and blue curves), or enabled an equal regularisation between both sets (green curve). At the global log-likelihood optimum ($\lambda_2 = 190, \xi = 0.5$), co-activation probabilistic modulations were accurately retrieved in a majority (but not all) of cases, as well as for a still limited subset of causal relationships. This resulted in intermediate similarity to the ground truth.

When regularisation was further lowered, regardless of the ξ parameter value, all curves converged towards a common, almost full representation that captured ground truth co-activation and causal influences with high fidelity: all regional similarity values exceeded 0.8 for co-activations, and for causal modulations, the majority exceeded 0.6. Only hub regions (for which underlying patterns are by construction more complex) and areas from network 3 (linked to a negative-valued modulation from network 7) showed slightly lower similarity values around 0.5, but the related patterns could still be captured in the associated coefficient matrices from Figure 2C.

Log-likelihood reached a local optimum at $\lambda_3 = 8.4$, which was very close to the global one. The slightly lower likelihood value despite the closer match to the ground truth is explained by the presence of a wide array of small noisy coefficients, seen as small negative-valued entries in the λ_3 matrices of Figure 2C.

To summarise, although the arrangement of regions into networks and their relationships could not be determined from inspecting the time courses (Figure 2A), they could be retrieved following the application of our framework. In addition, all regions could be correctly assigned to their associated network from co-activation probabilistic couplings (Figure 2D): following hierarchical clustering, 8 distinct groups could indeed be determined, including the 7 networks of interest and an extra cluster for independent

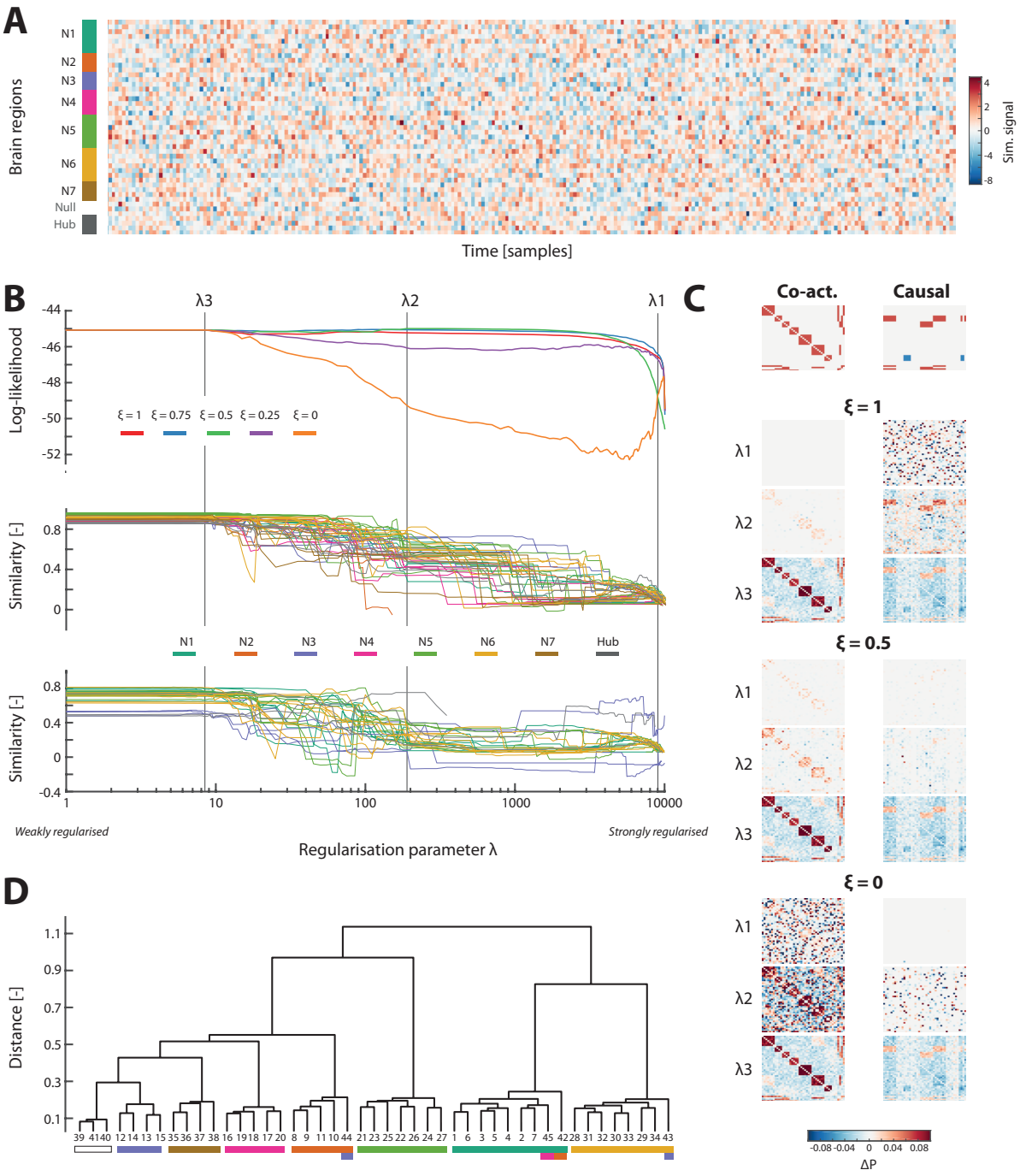

Figure 2. Results on simulated data. (A) Example simulated time courses for $R = 45$ regions, each displayed as one row for 250 samples. Colour coding denotes the network attribution of the regions (N_1 to N_7), as well as independent areas (in white) and hubs (in dark grey). **(B)** For the whole path of regularisation (strong to weak regularisation from the right to the left), whole log-likelihood of the data across brain regions and transition types (top plot), and associated similarity to the ground truth co-activation (middle plot) and causal (bottom plot) coefficients. The colour coding of the time courses respectively stands for the trade-off between the co-activation and causal set regularisations (top plot), and the network assignment of the regions (middle and bottom plot). The λ_1, λ_2 and λ_3 vertical lines highlight three indicative regularisation levels further detailed in **(C)**. **(C)** For co-activation (left column) and causal (right column) coefficients, ground truth values (top row), and probabilistic cross-regional influences for three co-activation/causal trade-off values ($\xi = 1, 0.5, 0$) and overall regularisation levels ($\lambda_1 = 9000, \lambda_2 = 190, \lambda_3 = 8.4$). **(D)** Dendrogram for regional clustering from co-activation probabilistic influences, with the same regional colour coding as in **(A)**.

regions. Note that hub areas were all assigned to one of the networks that they were linked to.

3.2. Application of the framework to experimental fMRI data

4. Discussion

In this work, we introduced a novel mathematical framework enabling to jointly derive the patterns of co-activation between brain regions, reflective of the brain's functional organisation as a set of RSNs (Damoiseaux et al. 2006, Yeo et al. 2011), and additional cross-regional causal modulations that enable to go beyond this network-level characterisation and also model more subtle cross-regional interplays. One can conceive our strategy as a joint recovery of FC (embedded in the $\mathbf{\Gamma}$ co-activation coefficients) and effective connectivity (in $\mathbf{B}$).

Our strategy is an improvement over previous work that also used a logistic regression characterisation to describe causal interactions between functional brain networks (Bolton et al. 2017): in this former methodology, however, network maps had to be computed in a separate analytical step, prior to the establishment of their causal interplays. As such, and much like the majority of other prominent dynamic FC approaches—see for instance (**?**, Allen et al. 2014, Karahanoğlu & Van De Ville 2015, Vidaurre et al. 2017), more subtle relationships at a smaller spatial scale than that of RSNs are then lost.

On simulated data, both co-activation and causal coefficient sets could accurately be retrieved by our framework despite marked noise. The optimal log-likelihood of the data was achieved in a weak regularisation setting, as we considered enough data points for accurate estimation of the full model: in total, we analysed 160650 time points for the estimation of $2(R + (R-1)R + (R-1)R) = 8010$ coefficients (two sets of coefficients—one per type of transition—for individual regional dynamics, co-activation and causal links), resulting in 20 data points available per estimate. Regularisation is expected to become more handy when dimensionally larger problems are addressed at a similar dataset size: for example, it will be interesting to derive coefficients on an extended set of brain regions obtained with finer parcellations that do not only operate from structural brain markers (Glasser et al. 2016, Schaefer et al. 2017).

As our simulations primarily included positive-valued coefficients, noisy coefficient estimates accompanying ground truth values were biased towards negative values (see the λ_3 settings in Figure 2C). This is why the simulated negative causal relationship between networks 7 and 3 was the least accurately captured one. At stronger regularisation levels, noisy coefficients disappeared, and a restricted subset of ground truth entries were recovered, owing to the ℓ_1 norm properties (**?**).

Several strategies may be envisioned to further improve the accuracy of the results obtained with our framework. First, the purely ℓ_1 regularisation strategy could be turned into an *elastic net* mix between ℓ_1 and ℓ_2 norms (**?**), but it would then come at the cost of an extra free parameter to specify. Second, additional assumptions could be explicitly introduced to the model formulation, such as the symmetric and non-negative nature of $\mathbf{\Gamma}$. Third, as noise operates to counterbalance strong positive-valued coefficients along

a given column of $\mathbf{\Gamma}$ or $\mathbf{B}$ (recall that coefficient estimates are obtained separately for each region r standing as one matrix column), the framework could be extended to successively run through a column-wise (as presently) and a row-wise solving step, where in the latter case, we would instead be estimating all the modulations emanating from a given region r (instead of impinging on it). Each of these three options has merits, but comes at the expense of a greater computational complexity and less streamlined modelling.

In real data, BLABLA.mI should mention the relevance of the ξ parameter, and the fact that both types of coefficients are captured.

Future work can model innovations instead of changes between baseline and active states: we would then just consider the probability to undergo an innovation. The advantage is that then, we can also more readily bridge the results from different datasets together, even if acquired at different TRs: indeed, we could consider whether an innovation occurred at time $t-1$, $t-2$. $t-3$...

Here, we make the assumption that the 0 to +1 and +1 to 0 transitions are mirroring each other (that is, if a region modulates another, it will boost one and decrease the other). Our framework already enables to also look for more subtle interplays by simply not combining the $\cdot_A$ and $\cdot_D$ cases anymore, but keeping them separate. For instance, maybe a given network only modulates another when the other is at baseline.

The hope is that in follow-up work, this approach enables to address brain disorders; for this purpose, bootstrapping could be conducted on both subject populations and coefficient distributions (or probabilistic modulations) compared statistically. The assessment of behavioural differences is also interesting, but harder to achieve: tailored ways to derive subject-level estimates despite too low data amount should then be developed.

More far-fetched ideas for future work: (1) apply this framework to hyperscanning to probe co-activation and causal links between two interacting subjects, and (2) apply this framework between concomitantly acquired fMRI and EEG data (that's the advantage of hidden states as a representational approach): of course, there would however be the need to define an equivalent temporal resolution between modalities.

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T. & Calhoun, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state, *Cerebral Cortex* **24**(3): 663–676.

Bolton, T. A. W., Tarun, A., Sterpenich, V., Schwartz, S. & Van De Ville, D. (2017). Interactions between large-scale functional brain networks are captured by sparse coupled hmms, *IEEE Transactions on Medical Imaging* **37**(1): 230–240.

Bolton, T. A. W., Zöller, D., Caballero-Gaudes, C., Kebets, V., Glerean, E. & Van De Ville, D. (2019). Agito ergo sum: correlates of spatiotemporal motion characteristics during fmri, *ArXiv (DOI: 1906.06445)* .

Bressler, S. L. & Menon, V. (2010). Large-scale brain networks in cognition: emerging methods and principles, *Trends in Cognitive Sciences* **14**(6): 277–290.

Chang, C. & Glover, G. H. (2010). Time-frequency dynamics of resting-state brain connectivity measured with fMRI, *Neuroimage* **50**(1): 81–98.
URL: *http://dx.doi.org/10.1016/j.neuroimage.2009.12.011*

Chen, S., Langely, J., Chen, X. & Hu, X. (2016). Spatiotemporal modeling of brain dynamics using resting-state functional magnetic resonance imaging with Gaussian hidden Markov model, *Brain Topography* **6**(4): 326–334.

Chen, T., Cai, W., Ryali, S., Supekar, K. & Menon, V. (2016). Distinct global brain dynamics and spatiotemporal organization of the salience network, *PLOS Biology* **14**(6): 1–21.
URL: *https://journals.plos.org/plosbiology/article/file?id=10.1371/journal.pbio.1002469&type=printable*

Christoff, K., Irving, Z. C., Fox, K. C. R., Spreng, R. N. & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: a dynamic framework, *Nature Reviews Neuroscience* **17**(11): 718–731.
URL: *http://www.nature.com/doifinder/10.1038/nrn.2016.113*

Damaraju, E., Allen, E. A., Belger, A., Ford, J. M., McEwen, S., Mathalon, D. H., Mueller, B. A., Pearlson, G. D., Potkin, S. G., Preda, A., Turner, J. A., Vaidya, J. G., van Erp, T. G. & Calhoun, V. D. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia, *Neuroimage: Clinical* **5**: 298–308.
URL: *http://dx.doi.org/10.1016/j.nicl.2014.07.003 http://linkinghub.elsevier.com/retrieve/pii/S2213158214000953*

Damoiseaux, J. S., Rombouts, S. A. R., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M. & Beckmann, C. F. (2006). Consistent resting-state networks across healthy subjects, *Proceedings of the National Academy of Sciences* **103**(37): 13848–13853.
URL: *http://www.pnas.org/content/103/37/13848.short*

Eavani, H., Satterthwaite, T. D., Gur, R. E., Gur, R. C. & Davatzikos, C. (2013). Unsupervised learning of functional network dynamics in resting state fMRI, *Lecture Notes in Computer Science* **7917**: 426–437.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3974209/pdf/nihms-504470.pdf*

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. et al. (2007). Pathwise coordinate optimization, *The Annals of Applied Statistics* **1**(2): 302–332.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**(1): 1.

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis, *Human Brain Mapping* **2**(1): 56–78.

Gilson, M., Moreno-Bote, R., Ponce-Alvarez, A., Ritter, P. & Deco, G. (2016). Estimation of directed effective connectivity from fmri functional connectivity hints at asymmetries of cortical connectome, *PLOS Computational Biology* **12**(3): e1004762.

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

Iraji, A., Fu, Z., Damaraju, E., DeRamus, T. P., Lewis, N., Bustillo, J. R., Lenroot, R. K., Belger, A., Ford, J. M., McEwen, S. et al. (2019). Spatial dynamics within and between brain functional domains: A hierarchical approach to study time-varying brain function, *Human Brain Mapping*

40(6): 1969–1986.

Kang, J., Pae, C. & Park, H. (2019). Graph-theoretical analysis for energy landscape reveals the organization of state transitions in the resting-state human cerebral cortex, *PLOS ONE* **14**(9): 0222161.

Karahanoğlu, F. I. & Van De Ville, D. (2015). Transient brain activity disentangles fMRI resting-state dynamics in terms of spatially and temporally overlapping networks, *Nature Communications* **6**: 7751.
URL: *http://www.nature.com/doifinder/10.1038/ncomms8751*

Kiviniemi, V., Vire, T., Remes, J., Elseoud, A. A., Starck, T., Tervonen, O. & Nikkinen, J. (2011). A sliding time-window ICA reveals spatial variability of the default mode network in time, *Brain Connectivity* **1**(4): 339–347.
URL: *http://www.liebertonline.com/doi/abs/10.1089/brain.2011.0036*

Kottaram, A., Johnston, L., Ganella, E., Pantelis, C., Kotagiri, R. & Zalesky, A. (2018). Spatio-temporal dynamics of resting-state brain networks improve single-subject prediction of schizophrenia diagnosis, *Human Brain Mapping* **39**(9): 3663–3681.

Lennartz, C., Schiefer, J., Rotter, S., Hennig, J. & LeVan, P. (2018). Sparse estimation of resting-state effective connectivity from fmri cross-spectra, *Frontiers in Neuroscience* **12**: 287.

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

Liu, X., Chang, C. & Duyn, J. H. (2013). Decomposition of spontaneous brain activity into distinct fMRI co-activation patterns, *Frontiers in Systems Neuroscience* **7**: 1–11.
URL: *http://journal.frontiersin.org/article/10.3389/fnsys.2013.00101/abstract*

Lurie, D., Kessler, D., Bassett, D., Betzel, R. F., Breakspear, M., Keilholz, S., Kucyi, A., Liégeois, R., Lindquist, M. A. & McIntosh, A. R. (2018). On the nature of resting fmri and time-varying functional connectivity, *PsyArXiv* .

Murphy, K. & Fox, M. D. (2017). Towards a consensus regarding global signal regression for resting state functional connectivity mri, *Neuroimage* **154**: 169–173.

Pedersen, M., Zalesky, A., Omidvarnia, A. & Jackson, G. D. (2018). Multilayer network switching rate predicts brain performance, *Proceedings of the National Academy of Sciences* **115**(52): 13376–13381.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion, *Neuroimage* **59**(3): 2142–2154.

Power, J. D., Fair, D. A., Schlaggar, B. L. & Petersen, S. E. (2010). The Development of Human Functional Brain Networks, *Neuron* **67**(5): 735–748.
URL: *http://dx.doi.org/10.1016/j.neuron.2010.08.017 http://ac.els-cdn.com/S0896627310006276/1-s2.0-S0896627310006276-main.pdf?_tid=530b28c0-358b-11e7-8056-00000aacb35d&acdnat=1494425996_7e5fc23be3ab984395a36c2a6fa42bfb*

Preti, M. G., Bolton, T. A. W. & Van De Ville, D. (2017). The dynamic functional connectome: State-of-the-art and perspectives, *Neuroimage* **160**: 41–54.
URL: *https://ac.els-cdn.com/S1053811916307881/1-s2.0-S1053811916307881-main.pdf?_tid=740fef8f-835b-4993-86d3-a6f60b841679&acdnat=1523360042_29ea611d8892cbd9e44c8ac92ff16e62*

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE* **77**(2): 257–286.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X., Holmes, A. J., Eickhoff, S. B. & Yeo, B. T. T. (2017). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri, *Cerebral Cortex* **28**(9): 3095–3114.

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P. et al. (2013). Resting-state fmri in the human connectome project, *Neuroimage* **80**: 144–168.

Smith, S. M., Miller, K. L., Moeller, S., Xu, J., Auerbach, E. J., Woolrich, M. W., Beckmann,

C. F., Jenkinson, M., Andersson, J., Glasser, M. F. et al. (2012). Temporally-independent functional modes of spontaneous brain activity, *Proceedings of the National Academy of Sciences* **109**(8): 3131–3136.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D. & Woolrich, M. W. (2010). Network modelling methods for fMRI, *NeuroImage* .

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. & Loliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, *NeuroImage* **15**: 273–289.

van den Heuvel, M. P. & Hulshoff Pol, H. E. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity, *European Neuropsychopharmacology* **20**(8): 519–534.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E. & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview, *Neuroimage* **80**: 62–79.
URL: *http://dx.doi.org/10.1016/j.neuroimage.2013.05.041 https://ac.els-cdn.com/S1053811913005351/1-s2.0-S1053811913005351-main.pdf?_tid=dd390d26-22c1-4bf3-8679-dd6749128504&acdnat=1543411897_b64e72580a4e4b27528bb4c99a53f9b8*

Vidaurre, D., Smith, S. M. & Woolrich, M. W. (2017). Brain network dynamics are hierarchically organized in time, *Proceedings of the National Academy of Sciences* **114**(48): 201705120.
URL: *http://www.pnas.org/lookup/doi/10.1073/pnas.1705120114*

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R. et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity, *Journal of Neurophysiology* **106**(3): 1125–1165.