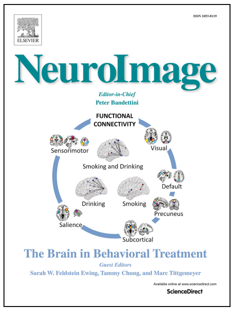


Sparse DCM for whole-brain effective connectivity from resting-state fMRI data

Giulia Prando, Mattia Zorzi, Alessandra Bertoldo, Maurizio Corbetta, Marco Zorzi, Alessandro Chiuso

PII: S1053-8119(19)30958-9

DOI: https://doi.org/10.1016/j.neuroimage.2019.116367

Reference: YNIMG 116367

To appear in: *NeuroImage*

Received Date: 25 June 2019

Revised Date: 12 November 2019

Accepted Date: 13 November 2019

Please cite this article as: Prando, G., Zorzi, M., Bertoldo, A., Corbetta, M., Zorzi, M., Chiuso, A., Sparse DCM for whole-brain effective connectivity from resting-state fMRI data, *NeuroImage* (2020), doi: https://doi.org/10.1016/j.neuroimage.2019.116367.

Author Contribution Statement

G. Prando, M. Zorzi, A. Bertoldo, M. Corbetta, M. Zorzi, A. Chiuso

November 12, 2019

Giulia Prando: Methodology, Software, Investigation, Writing - Original draft, Visualization

Mattia Zorzi: Methodology, Formal analysis, Writing - Review and editing

Alessandra Bertoldo: Methodology, Formal analysis, Writing - Review and editing

Maurizio Corbetta: Conceptualization

Marco Zorzi: Conceptualization, Writing - Review and editing

Alessandro Chiuso: Methodology, Formal analysis, Writing - Review and editing, Supervision

Sparse DCM for whole-brain effective connectivity from resting-state fMRI data

Giulia Prando[a], Mattia Zorzi[a], Alessandra Bertoldo[a], Maurizio Corbetta[b], Marco Zorzi[c], Alessandro Chiuso[a]

[a] *Department of Information Engineering, University of Padova, Padova (Italy)*
[b] *Department of Neuroscience, University of Padova, Padova (Italy)*
[c] *Department of General Psychology, University of Padova, Padova (Italy) and IRCCS San Camillo Hospital, Venice-Lido (Italy)*

Abstract

Contemporary neuroscience has embraced network science and dynamical systems to study the complex and self-organized structure of the human brain. Despite the developments in non-invasive neuroimaging techniques, a full understanding of the directed interactions in whole brain networks, referred to as *effective connectivity*, as well as their role in the emergent brain dynamics is still lacking. The main reason is that estimating brain connectivity requires solving a formidable large-scale inverse problem from indirect and noisy measurements. Building on the dynamic causal modeling framework, the present study offers a novel method for estimating whole-brain effective connectivity from resting-state functional magnetic resonance data. To this purpose sparse estimation methods are adapted to infer the parameters of our novel model, which is based on a linearized, region-specific haemodynamic response function. The resulting algorithm is shown to compare favorably with state-of-the art methods when tested on both synthetic and real data. We also provide a graph-theoretical analysis on the whole-brain effective connectivity estimated using data from a cohort of healthy individuals, which reveals properties such as asymmetry in the connectivity structure as well as the different roles of brain areas in favoring segregation or integration.

Keywords: effective connectivity, fMRI, sparsity, Dynamic Causal Modeling, resting-state

1. Introduction

The study of the human brain as a complex network plays a central role in contemporary neuroscience. It is now widely believed that cognitive processes are not localized to a specific brain region but arise from the interplay of several areas [70]. The study and

Email addresses: prandogi@dei.unipd.it (Giulia Prando), zorzimat@dei.unipd.it (Mattia Zorzi), bertoldo@dei.unipd.it (Alessandra Bertoldo), maurizio.corbetta@unipd.it (Maurizio Corbetta), marco.zorzi@unipd.it (Marco Zorzi), chiuso@dei.unipd.it (Alessandro Chiuso)

validation of this concept, known as functional integration, critically relies on the analysis of the anatomical and functional relations between brain regions, which is defined in terms of brain connectivity. The development of non-invasive neuroimaging techniques has allowed to identify different types of brain connectivity, ranging from anatomical (structural) links, to statistical (functional) and directed (effective) connections (see [71, 24, 43] for a review). Their joint analysis appears to be crucial to understand the complex organization of the human brain, which in turn plays a key role in predicting the effect of brain lesions [7, 67] as well as in studying and designing brain stimulation treatments [41, 38, 36, 10]. While whole-brain characterizations exist for structural (SC) and functional connectivity (FC) [33, 83, 89], a thorough understanding of whole-brain directed interactions (as described by effective connectivity) remains elusive. The main reason for this gap lies in the fact that effective connectivity (EC) is often defined in terms of a generative model for the blood oxygen level dependent (BOLD) signal. The latter is measured with functional Magnetic Resonance Imaging (fMRI). Inferring EC requires estimating a large number of parameters from a relatively small dataset [81], which turns out being an ill-posed inverse problem.

Accordingly, inference of whole-brain effective connectivity appears as a key open challenge for the neuroscience community [52, 18, 3]. Besides the complexity of the estimation problem, validation of the estimated effective connectivity networks is still an open issue.

A classical approach for effective connectivity estimation relies on a nonlinear dynamical model. The latter accounts for both the directed dependencies among neural populations and the mapping from neural activity to observations. This framework is known in the neuroimaging community as Dynamic Causal Modelling[1] (DCM); it was originally developed to deal with fMRI data [25, 11] and later extended to handle EEG and MEG data [12, 37]. The original deterministic formulation [25] only accounted for task-dependent fMRI data, where neural activity is driven by known external stimuli. A stochastic DCM, driven by endogenous random fluctuations, was later developed to deal with resting state [26]. DCM inversion is commonly performed assuming a prior for the model parameters and using the Variational Bayes approach to compute an approximation of their posterior [25, 21]. This procedure is particularly challenging for stochastic DCMs, because it requires to infer not only the model parameters but also the latent neural activity [29, 23]. This latter issue was solved in [26, 53] by postulating a linear model for the haemodynamic response, allowing to reformulate the DCM in the frequency domain and simplifying the model inversion.

Within the DCM framework, effective connectivity estimation typically starts by postulating a family of candidate network topologies and proceeds by inverting a DCM for each topology; finally, the best hypothesis is chosen using Bayesian model selection (BMS) [22, 75, 74]. However, the number of possible network topologies is combinatorial

[1] There is some debate in the literature on the use of the terms *directed* and *causal*, see e.g. [82]. We prefer to avoid entering in this debate and therefore we shall always use the term *directed* connections when talking about EC.

in the number of network nodes (i.e., brain regions). This poses severe challenges due to (i) the need to invert a huge number of competing DCMs and (ii) the need to compare a combinatorial number of alternatives which leaves very low statistical significance to the final selected network topology (EC). These issues have been partially overcome by resorting to techniques known as *post-hoc model selection* [22] or *Bayesian model reduction* [27], which allow to invert one fully connected model and to subsequently perform a greedy selection over the nested models. Despite the availability of these approaches, the inversion of a classical DCM remains ill-posed and computationally intensive for large brain networks, thus limiting its applicability to networks including about ten nodes [14, 53, 80]. More recently, the introduction of sparsity inducing priors on the connectivity matrix has extended the use of resting-state DCM to graphs composed of tens of nodes [63, 54]. Another approach, known as regression DCM [19], was recently applied to infer task-dependent effective connectivity among 104 brain regions [18]; the price to be paid in regression DCMs is that a *linear and known* haemodynamics model needs to be postulated.

Outside the DCM framework, models which attempt to establish Granger-type causality directly on observed BOLD signals have been developed. For instance, effective connectivity was recently treated as a parameter of the model describing brain resting-state dynamics as an Ornstein-Uhlenbeck process. Note that this approach neglects the effect of the haemodynamic response. Under this modelling assumption, a fast procedure was proposed to estimate brain directed dependencies [32] and applied to whole-brain networks. The estimated effective connectivity profiles proved to be reliable signatures for subject identification as well as for task/rest condition detection [31, 42].

While the aforementioned approaches for effective connectivity estimation rely on the specification of a generative model of the available measurements, Bayesian nets provide an alternative model-free framework. Under the assumption that brain effective connections form a directed acyclic graph (DAG), these methods typically evaluate conditional probabilities to assess network adjacencies [48, 50, 57]. Among them, the Fast Greedy Equivalence Search (FGES) was recently applied to a voxel-wise whole-brain network [47]. According to the validation study performed in [68], and more recently confirmed in [61], Bayesian nets successfully detect existing connections, but are much less powerful in estimating link directionality.

The main contribution of the present work is to offer a novel effective connectivity estimation procedure for resting-state fMRI data, hereafter named *sparse DCM*. Our method is based on a simplification of the standard resting-state DCM [26] and can be applied to whole-brain data. The main differences with respect to standard DCMs are the following:

1. DCMs and the Ornstein-Uhlenbeck model adopted in [32] are formulated in continuous-time; our model is converted in discrete-time while keeping a continuous-time physical parametrization (effective connectivity). In this way we better exploit the low temporal resolution of fMRI scanners so as to simplify (from the computational point of view) the burden of model inversion without loosing in statistical performance.

2. We propose a statistical linearisation of the haemodynamic response function (HRF), thus obtaining a linear stochastic generative model of resting-state fMRI data. This allows to translate the priors on the physiological parameters describing the haemodynamic model [6] into a prior on the HRF that can be exploited when performing model inversion. A preliminary version of this procedure was proposed in [45], and it is generalized in the present study to account for the haemodynamic variability across brain areas.
3. Following the Sparse Bayesian Learning (SBL) approach [78], a sparsity-inducing prior is formulated on the matrix describing the effective connectivity network. In addition, the iterative reweighted procedure introduced in [85] is adapted to our framework.
4. An expectation-maximization (EM) algorithm [13, 66] is used to invert our simplified (linear) DCM. Insights on the algorithm initialization are provided in terms of (i) a procedure for automatic initialization and (ii) analyses on the role of prior knowledge about effective connectivity patterns on initialization, which might be important for clinical applications.

The second contribution of the present work is to provide a thorough comparison of state-of-the-art methods for estimating effective connectivity models, ranging from DCM-type [26] (including our sparse DCM) to Bayesian nets [69, 49, 47, 64, 56, 61] and Granger causality [2].

The third contribution of the current study is to offer an extensive study on empirical fMRI data for a whole-brain parcellation [34]. In this real scenario, the effective connectivity pattern inferred by sparse DCM was validated by measuring its ability to reproduce subject-specific functional connectivity on new data. Building on these results, we also provide a graph-theoretical analysis on the whole-brain effective connectivity networks estimated for a cohort of subjects, computing metrics such as nodes strength, clustering coefficient and path lengths [58]. This large-scale analysis is typically performed on functional or structural networks [5, 17], while only few results are available for effective connectivity graphs, see e.g. [32, 55, 18].

Note that the generative model adopted here is related to that proposed in [60] and further developed in [59]. There are however some key differences, most notably: (i) a different linearization strategy for the HRF and (ii) the use of the EM algorithm combined with an iterative-reweighted procedure [85] to invert the specified generative model and to obtain a sparse connectivity pattern.

Alternative, and possibly richer, modelling frameworks could of course be considered. For instance, in the control and system identification community several dynamical models with an underlying network structure have been studied (see [87, 8, 84, 90] and references therein). The sparse DCM model provides a good trade-off between model complexity and the need to account for physiological insights and computational issues, all of which should be considered when estimating models for high dimensional data (fMRI recordings) from relatively small (i.e., measured for short time intervals) datasets.

This article is structured to reflect its three main contributions. The first part reviews the classical DCM framework and introduces our sparse DCM model. The second part establishes the face, construct and predictive validity of the scheme. The third part illustrates a further application of sparse DCM by addressing a generic issue in functional integration from the perspective of graph theory.

2. Methods and Materials

2.1. Dynamic Causal Modeling

A Dynamic Causal Model (DCM), as proposed by Friston et al. [25], is a nonlinear multiple input multiple output (MIMO) dynamical system. It is driven by experimentally designed inputs (task) and by random neural fluctuations (resting-state). It outputs the BOLD fMRI response $y(t)$ for each of the monitored brain areas. The DCM consists of two components: a differential equation describing the coupling among neuronal populations, and a dynamic map from the neuronal activity to the measured BOLD signal $y(t)$, the so-called haemodynamic response. Let $x(t) = [x_1(t) \ \cdots \ x_n(t)]^\top$ denote the hidden neural activity of n brain regions at time t. The DCM takes the following form:

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t); \theta_f) + v(t) \\ y(t) &= h(x(t); \theta_h) + e(t), \quad e(t) \sim \mathcal{N}(0, R), \end{aligned} \tag{1}$$

where $u(t)$ denotes experimental manipulations (such as external stimuli, task demands), $v(t)$ is a stochastic term representing intrinsic brain fluctuations, and $e(t)$ accounts for observation noise with covariance matrix R. The parameters θ_f describe the model at the neuronal level, including effective connectivity, while θ_h are biophysical parameters defining the haemodynamic response. The original DCM formulation [25] assumes that the neural activity is elicited only by external stimuli $u(t)$, thus neglecting the stochastic source $v(t)$, and postulates a bilinear form for f:

$$\dot{x}(t) = \left(A + \sum_{j=1}^{m} u_j(t) B_j \right) x(t) + Cu(t). \tag{2}$$

In this case $\theta_f := \{A, B_1, \cdots, B_m, C\}$ encode couplings among neural activity and external inputs. Specifically, A represents the network connectivity (effective connectivity) in the absence of external excitations, B_j accounts for the change in the neuronal coupling due to the j-th input; finally, C models the direct influence of experimental manipulations on the neuronal activity.

A variant of the original DCM was introduced by [26] in order to deal with resting-state fMRI (rs-fMRI) data. In this setting, external stimuli are absent, that is $u(t) = 0$, and the random fluctuations $v(t)$ are responsible for driving the neural activity. Function f in (1) becomes linear:

$$\dot{x}(t) = Ax(t) + v(t) \tag{3}$$

with A representing effective connectivity.

The second component of a DCM, that is the haemodynamic response h appearing in Eq. (1), is modeled through a nonlinear, biophysically inspired, dynamical system. It takes the neural activity $x_i(t)$ as input and outputs the corresponding BOLD signal $b_i(t)$ [6, 28, 76] :

$$\dot{r}_i(t) = x_i(t) - \kappa_i r_i(t) - \eta_i(f_i(t) - 1), \qquad i = 1, ..., n \tag{4}$$

$$\dot{f}_i(t) = r_i(t) \tag{5}$$

$$\tau_i \dot{v}_i(t) = f_i(t) - v_i^{1/\xi_i}(t) \tag{6}$$

$$\tau_i \dot{q}_i(t) = (f_i(t)/\rho_i)\left[1 - (1-\rho_i)^{1/f_i(t)}\right] - v_i^{1/\xi_i - 1}(t) q_i(t) \tag{7}$$

$$b_i(t) = V_0 k_1 (1 - q_i(t)) + V_0 k_2 (1 - q_i(t)/v_i(t)) + V_0 k_3 (1 - v_i(t)). \tag{8}$$

The haemodynamic states $\{r_i, f_i, v_i, q_i\}$ are biophysical quantities: r_i denotes the vasodilatatory signal, f_i is the blood inflow, v_i and q_i are respectively the blood volume and the deoxyhemoglobin content. The output equation (8) depends on the resting blood volume fraction V_0 (typically $V_0 = 0.02$) and on the constants k_1, k_2 and k_3. These have found different characterizations in the literature, as reviewed in [76]. Also the parameters $\theta_h = \{\kappa_i, \eta_i, \tau_i, \xi_i, \rho_i;\ i = 1, .., n\}$ have a biological meaning, see [25]. In the latter study, a prior distribution for θ_h has been specified. When adopting the DCM framework, effective connectivity is estimated by inverting the DCM using measured fMRI data. In a Bayesian framework this inversion cannot be computed in closed form. Most often Variational Bayes techniques under the Laplace approximation (VBL) [21, 11] are exploited. When resting-state fMRI data are considered and neural dynamics is assumed to be described by Eq. (3), the DCM inversion becomes more challenging than in the task-dependent domain (that is, when Eq. (2) is used). While in the latter case, only the parameters θ_f and θ_h have to be inferred, in the first situation, also the neural states $x(t)$ have to be estimated. Two procedures are commonly used to this end: *Dynamic Expectation Maximization* (DEM) [29] and *Generalized Filtering* (GF) [23, 39]; even if both adopt the Variational Bayes procedure, DEM uses the mean-field and the Laplace approximations, while GF only exploits the latter. However, a different approach, known as *spectral DCM* (spDCM) [26, 52] has been proven superior to these methods, both in terms of face validity and of computational complexity [53]. Differently from DEM and GF, which operate in time domain, spDCM replaces the stochastic generative model (3)-(8) with a deterministic model producing the cross-spectra of the original fMRI time-series. In this way, endogenous neural states $x(t)$ are no longer estimated, but only the time-invariant parameters describing their cross-spectra have to be inferred.

Despite the widespread use of these approaches in computational neuroscience, their applicability is limited to small brain networks, in the order of ten nodes. Increasing the number of regions leads to a relevant rise in the number of parameters to be estimated and in turn to an exponential growth of the computational time required to invert these models. These limitations particularly affect DEM and GF, which have to estimate both the hidden neural states trajectories $x(t)$ and the parameters. On the other hand, the computational efficiency of spDCM was recently exploited to invert large-scale DCMs, comprising up to 36 brain regions [54]. To improve the robustness and further reduce the computational burden, we introduce below a simplification of the original DCM framework and a simplified (Expectation-Maximization) procedure for its inversion.

2.2. Linear DCM

The proposed reformulation of the classical DCM for rs-fMRI [26] involves both a discretization and a linearization of the original non-linear continuous-time model. The former is dictated by the low temporal resolution of fMRI scanners: since they indirectly measure the neuronal activity at time intervals of length T_R (typically ranging from 0.7 to 3 seconds), it is reasonable to adopt a discrete-time version[2] of Eq. (3). This is derived by simply observing that

$$\begin{aligned} x(kT_R + T_R) &= e^{A(kT_R+T_R-kT_R)}x(kT_R) + \int_{kT_R}^{kT_R+T_R} e^{A(kT_R+T_R-s)}v(s)\,ds \\ &= e^{AT_R}\;x(kT_R) + \int_0^{T_R} e^{A\tau}v(\tau)\,d\tau. \end{aligned}$$

Using the simplified notation $x(k) := x(kT_R)$ and defining $w(k) := \int_0^{T_R} e^{A\tau}v(\tau)\,d\tau$, the sampled version of Eq. (3) becomes

$$x(k+1) = e^{AT_R}x(k) + w(k). \tag{9}$$

Furthermore, we assume that $v(t)$, $t \in \mathbb{R}$, in (3) is white Gaussian noise with intensity $\sigma^2 I_n$ where I_n denotes the identity matrix of size n ; consequently, $w(k)$ is white Gaussian with variance [30]

$$Q = \sigma^2 \int_0^{T_R} e^{A\tau}e^{A^\top\tau}\mathrm{d}\tau. \tag{10}$$

The haemodynamic response (4)-(8) is linearised following a statistical approach as follows: we consider a Finite Impulse Response (FIR) model which takes as input a neuronal state $x_i(k)$ and outputs the BOLD signal $b_i(k) := b_i(kT_R)$:

$$b_i(k) = \sum_{l=0}^{s-1} h_{i,l}\; x_i(k-l), \qquad i = 1, ..., n. \tag{11}$$

The length s of the impulse response $h_i := [h_{i,0}\;\cdots\;h_{i,s-1}]^\top$ is chosen large enough to retain the relevant temporal dependencies. The finite impulse responses h_i are assigned a Gaussian prior distribution $h_i \sim \mathcal{N}(\mu_h, \Sigma_h)$, by exploiting the empirical priors for the parameters θ_h appearing in the non-linear model (4)-(8) of the haemodynamic

[2] Issues related to estimation of sparse continuous time models from low-rate (i.e. large T_R) measurements have been also recently discussed in [88].

response. The exact procedure we followed rests on statistical linearization techniques and is reported in Appendix A.

Having replaced the non-linear component of the DCM for rs-fMRI with a linear map, we can formulate the proposed DCM variant as a stochastic linear state-space model. In particular, defining

$$\begin{aligned}\mathbf{x}(k) &:= \begin{bmatrix} x^\top(k) & x^\top(k-1) & \cdots & x^\top(k-s+1)\end{bmatrix}^\top \in \mathbb{R}^{ns}\\ \mathbf{w}(k) &:= [w^\top(k)\ \mathbf{0}]^\top \in \mathbb{R}^{ns},\end{aligned}$$

model (1) with the linearization (11) can be written in the form

$$\begin{cases} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{w}(k) \\ y(k) &= \mathbf{H}\mathbf{x}(k) + e(k). \end{cases} \tag{12}$$

Matrices **A** and **H** in (12) are defined as

$$\mathbf{A} := \begin{bmatrix} e^{AT_R} & \mathbf{0} \\ I_{n(s-1)} & \mathbf{0} \end{bmatrix} \tag{13}$$

$$\mathbf{H} := \left[\begin{array}{cccc|cccc|c|cccc} h_{1,0} & 0 & \cdots & 0 & h_{1,1} & 0 & \cdots & 0 & \cdots & h_{1,s-1} & 0 & \cdots & 0 \\ 0 & h_{2,0} & \ddots & \vdots & 0 & h_{2,1} & \ddots & \vdots & \cdots & 0 & h_{2,s-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \ddots & 0 & \cdots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & h_{n,0} & 0 & \cdots & 0 & h_{n,1} & \cdots & 0 & \cdots & 0 & h_{n,s-1} \end{array}\right].$$

To complete the model specification, in line with Eqs. (10) and (1), we further assume:

$$\mathbf{w}(k) \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \qquad \mathbf{Q} := \text{blkdiag}(Q, \varsigma I_{n(s-1)}) \tag{14}$$

$$e(k) \sim \mathcal{N}(\mathbf{0}, R), \qquad R := \text{diag}(\lambda_1^2, ..., \lambda_n^2) \tag{15}$$

where blkdiag(·) and diag(·) respectively denote the block-diagonal and the diagonal operators, while $\varsigma = 10^{-15}$ is a scalar positive constant chosen small enough to guarantee that **Q** is invertible. The bold notation has been used to represent extended quantities. Section 2.3 will describe how the parameters

$$\theta := \{A, \sigma, h_1, ..., h_n, \lambda_1, ..., \lambda_n\} \tag{16}$$

which specify the linear model in (12), are estimated using an Expectation-Maximization (EM) algorithm.

For a fixed value of θ, we define the model Functional Connectivity (FC) as

$$[\widehat{\text{FC}}]_{ij} = \frac{[\Sigma_y]_{ij}}{\sqrt{[\Sigma_y]_{ii}[\Sigma_y]_{jj}}}. \tag{17}$$

where Σ_y is the stationary output covariance matrix $\Sigma_y = \mathbf{H}\Sigma_\mathbf{x}\mathbf{H}^\top + R$ and the stationary state covariance $\Sigma_\mathbf{x}$ is the solution of the Lyapunov equation $\Sigma_\mathbf{x} = \mathbf{A}\Sigma_\mathbf{x}\mathbf{A}^\top + \mathbf{Q}$. The model FC (17) can be computed using the parameters $\hat{\theta}$ estimated from a run of rs-fMRI data (according to the procedure detailed in Section 2.3). Note that the "empirical" FC, defined as the correlation between the empirical BOLD time-series, can be seen as a sample estimate of the model FC in (17). A comparison between the two, e.g. based on the Pearson Correlation Coefficient (PearsonCC) between the two matrices, will serve as a validation step for the estimation routine when a ground-truth is not available, that is when dealing with empirical fMRI data. A more reliable validation can be obtained by comparing the model FC in (17) computed from the estimated $\hat{\theta}$ with the empirical FC obtained in a different run from the same subject, i.e. from new data that have not been used for parameters inference: a good agreement would be a reasonable indicator of the generalization capabilities of the estimated model.

2.3. Sparse estimation algorithm

We now describe a procedure to estimate the parameter vector θ of model (12) from measurements $\{y(k)\}_{k=1}^{N}$ of the BOLD signal.

Following a Bayesian perspective, we first assign a prior $p_\gamma(\theta)$ so as to reflect either prior knowledge (e.g., on typical haemodynamic responses as in [25]) or to favor reconstruction of a sparse effective connectivity matrix A. The parameters γ, known as hyperparameters in the Bayesian learning framework, define the prior and are also estimated from data as discussed below. Ideally, one would like to find θ and γ that maximize the marginal posterior

$$p_\gamma(\theta|Y) = \int p_\gamma(\mathbf{X}, \theta|Y) \; \mathrm{d}\mathbf{X} \tag{18}$$

where $Y := [y^\top(1) \; \cdots \; y^\top(N)]^\top$ and $\mathbf{X} := [\mathbf{x}^\top(0) \; \cdots \; \mathbf{x}^\top(N)]^\top$, playing the role of measured and latent variables, respectively. However, the computation of such a high-dimensional integral is typically avoided by exploiting the decomposition $p_\gamma(\theta|Y) \propto p(Y|\theta)p_\gamma(\theta)$ and a tractable lower bound of the likelihood

$$p(Y|\theta) = \int p(\mathbf{X}, Y|\theta) \; \mathrm{d}\mathbf{X}. \tag{19}$$

An appropriate bound can be found e.g. resorting to the EM algorithm [13].

Before delving into algorithmic details, the prior $p_\gamma(\theta)$ will be specified. It will be assumed that $p_\gamma(\theta) \propto p_\gamma(A)p(\sigma)\prod_{i=1}^{n} p(h_i)p(\lambda_i)$, where $p(\sigma)$ and $p(\lambda_i)$, $i = 1, ..., n$, are uninformative priors while the h_i's are i.i.d. Gaussian $h_i \sim \mathcal{N}(\mu_h, \Sigma_h)$. A key role is played by the sparsity inducing prior $p_\gamma(A)$ for the connectivity matrix A. Following the Sparse Bayesian Learning (SBL) perspective [78], the elements $[A]_{ij}$ of matrix A are postulated to be independent zero mean Gaussian with variances γ_k, i.e. $p_\gamma(a) \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\gamma_1, \cdots, \gamma_{n^2}))$ with $a := \mathrm{vec}(A^\top)$ denoting the vectorization of $A^\top$. SBL was originally proposed to deal with classical regression problems where no hidden variables are present and where observations are corrupted by white Gaussian noise. Under this setting, the hyperparameters $\gamma := \{\gamma_k\}_{k=1}^{n^2}$ will be estimated through marginal likelihood maximization (also known as type-II maximum likelihood method). As a consequence, under generic conditions, the maximum likelihood ML estimates of certain γ_i's will be zero and the Gaussian posterior distribution of the corresponding element a_i of matrix A will concentrate around zero, leading to a zero MAP estimate. To compute the ML estimate, the hyperparameters $\{\gamma_i\}_{i=1}^{n^2}$ are updated, as an inner step of the EM-type algorithm described below, following the reweighted ℓ_1 approach proposed in [85]. This procedure provides an automatic selection of a sparsity pattern in the estimated effective connectivity matrix A, thus avoiding the combinatorial search over candidate network structures, which becomes practically infeasible in reasonably sized networks (tens to hundreds of nodes). Further details will be provided in Appendix C. Though the introduction of *post-hoc* model selection [22] and Bayesian Model Reduction [27] have extremely simplified the search over candidate DCM models, classical DCM approaches [21, 29, 23, 26, 53] remain affected by the issue of combinatorial model search. During the last decade, several studies have tried to alleviate this drawback by specifying different sparsity priors for the connectivity matrix A [60, 63, 59, 54, 18], in line with the approach we propose here.

We now provide the details regarding how the MAP estimate

$$\hat{\theta} = \arg\max_{\theta} \ \ \ln p(Y|\theta) + \ln p_\gamma(\theta) \tag{20}$$

is obtained using an EM procedure that iteratively optimizes a lower bound of the (log)posterior. Classically, EM maximizes $\ln p(Y|\theta)$ by iteratively maximizing its lower bound

$$\mathcal{L}(q(\mathbf{X}),\theta) = \int q(\mathbf{X})\left(\ln p(\mathbf{X},Y|\theta) - \ln q(\mathbf{X})\right)\ \mathrm{d}\mathbf{X} \tag{21}$$

with respect to an arbitrary distribution $q(\mathbf{X})$ and θ. In the statistical learning literature $\mathcal{L}(q(\mathbf{X}),\theta)$ is also known as (negative) *free-energy*. At the l-th iteration of the algorithm, $\mathcal{L}(q(\mathbf{X}),\theta^{(l)})$ is maximized by $q^{(l+1)}(\mathbf{X}) = p(\mathbf{X}|Y,\theta^{(l)})$. Plugging this into (21), one obtains

$$\mathcal{L}(q^{(l+1)}(\mathbf{X}),\theta) = \int p(\mathbf{X}|Y,\theta^{(l)})\ln p(\mathbf{X},Y|\theta)\mathrm{d}\mathbf{X} - \int p(\mathbf{X}|Y,\theta^{(l)})\ln p(\mathbf{X}|Y,\theta^{(l)})\mathrm{d}\mathbf{X}. \tag{22}$$

In our MAP setting (20) the a-priori information on θ needs to be included. Neglecting the terms that do not depend on θ and γ, a lower bound of the posterior is given by

$$\mathcal{Q}(\theta,\theta^{(l)}) = \int p(\mathbf{X}|Y,\theta^{(l)})\ln p(\mathbf{X},Y|\theta)\ \mathrm{d}\mathbf{X} + \ln p_\gamma(\theta). \tag{23}$$

Using the Markovian property of system (12), $\mathcal{Q}(\theta,\theta^{(l)})$ can be rewritten as [62, Ch.12]

$$\begin{aligned}\mathcal{Q}(\theta,\theta^{(l)}) = &\sum_{k=1}^{N}\int p(\mathbf{x}(k),\mathbf{x}(k-1)|Y,\theta^{(l)})\ \ln p(\mathbf{x}(k)|\mathbf{x}(k-1),\theta)\ \mathrm{d}\mathbf{x}(k)\mathrm{d}\mathbf{x}(k-1) \\ &+\sum_{k=1}^{N}\int p(\mathbf{x}(k)|Y,\theta^{(l)})\ \ln p(y(k)|\mathbf{x}(k),\theta)\ \mathrm{d}\mathbf{x}(k) + \ln p_\gamma(\theta)\end{aligned} \tag{24}$$

where the smoothing distributions

$$p(\mathbf{x}(k)|Y,\theta^{(l)}) = \mathcal{N}(\hat{\mathbf{x}}^s(k), \mathbf{P}^s(k)) \tag{25}$$

$$p(\mathbf{x}(k),\mathbf{x}(k-1)|Y,\theta^{(l)}) = \mathcal{N}\left(\begin{bmatrix}\hat{\mathbf{x}}^s(k)\\ \hat{\mathbf{x}}^s(k-1)\end{bmatrix}, \begin{bmatrix}\mathbf{P}^s(k) & \mathbf{P}^s(k)\mathbf{G}^\top(k-1)\\ \mathbf{G}(k-1)\mathbf{P}^s(k) & \mathbf{P}^s(k-1)\end{bmatrix}\right) \tag{26}$$

can be computed by means of the Rauch-Tung-Striebel smoother (RTSS) [51]. Its implementation is summarized in Appendix B (Algorithm 2). Plugging (25) and (26) into (24) we get

$$\begin{aligned}\mathcal{Q}(\theta,\theta^{(l)}) &= \ln p_\gamma(\theta) - \frac{N}{2}\ln|2\pi\mathbf{Q}| - \frac{N}{2}\ln|2\pi R| \\ &\quad - \frac{N}{2}\mathrm{tr}\left[\mathbf{Q}^{-1}\left(\Lambda - \Psi\mathbf{A}^\top - \mathbf{A}\Psi^\top + \mathbf{A}\Upsilon\mathbf{A}^\top\right)\right] \\ &\quad - \frac{N}{2}\mathrm{tr}\left[R^{-1}\left(\Delta - \Xi\mathbf{H}^\top - \mathbf{H}\Xi^\top + \mathbf{H}\Lambda\mathbf{H}^\top\right)\right]\end{aligned} \tag{27}$$

where

$$\Lambda = \frac{1}{N}\sum_{k=1}^{N}\mathbf{P}^s(k) + \hat{\mathbf{x}}^s(k)\left[\hat{\mathbf{x}}^s(k)\right]^\top,$$

$$\Psi = \frac{1}{N}\sum_{k=1}^{N}\mathbf{P}^s(k)\mathbf{G}(k-1) + \hat{\mathbf{x}}^s(k)\left[\hat{\mathbf{x}}^s(k-1)\right]^\top,$$

$$\Upsilon = \frac{1}{N}\sum_{k=1}^{N}\mathbf{P}^s(k-1) + \hat{\mathbf{x}}^s(k-1)\left[\hat{\mathbf{x}}^s(k-1)\right]^\top,$$

$$\Xi = \frac{1}{N}\sum_{k=1}^{N}y(k)\left[\hat{\mathbf{x}}^s(k)\right]^\top, \qquad \Delta = \frac{1}{N}\sum_{k=1}^{N}y(k)y^\top(k).$$

In summary, our algorithm alternates between an RTS smoother, which computes the distributions (25)-(26) for a fixed θ, and the maximization of function $\mathcal{Q}(\theta,\theta^{(l)})$ in Eq. (27) to update θ. At each iteration also the hyper-parameters $\{\gamma_i\}_{i=1}^{n^2}$ are updated; this is the key step for inducing sparsity on A. The complete routine is reported in Appendix B (Algorithm 1).

In terms of computational cost, each iteration has complexity $O\left((ns)^3\right)$ due to matrix inversions in the RTS smoother, see step 13 of Algorithm 1. Thus, for N iterations, the computational cost scales as $O((ns)^3N)$. In our experiments the number N of EM iterations ranged in the interval $[20, 400]$ depending on sampling time and on the number of monitored regions. Results on average execution times on a specific hardware can be found in Sections 3.1 and 3.2.

A further issue that calls for attention is the non-convexity of problem (20), which might have many local minima. As such the initialization $\theta^{(0)}$ and $\{\gamma_i^{(0)}\}_{i=1}^{n^2}$ plays a crucial role, especially when dealing with large DCMs. We experimentally investigated the impact of this stage on the estimated DCM and we found that the values of $\{\gamma_i^{(0)}\}_{i=1}^{n^2}$ do not strongly affect the final outcome of the EM algorithm, while the initialization of the effective connectivity matrix A seems to be more critical. Some of these results will be reported in Section 3.3.5 where large-scale DCMs are considered. As an outcome of the latter investigation further hints on initialization are found in Appendix B (Algorithm 1).

A last warning concerns sparsity of A: due to numerical issues, some of the parameters $\hat{\gamma}_k$'s become small but not zero. Thus, we obtain quasi-sparse solutions $\hat{A}$, i.e. with many entries having very small absolute values. These entries are irrelevant to any practical purpose and, to facilitate interpretation of the results, are thresholded to zero to make $\hat{A}$ rigorously sparse. An automated thresholding criterion based on functional connectivity notions will be proposed and discussed (see Sec. 3.1 and also Sec. 3.3.2).

2.4. Synthetic data

Some Monte-Carlo studies on synthetic datasets were conducted. The synthetic rs-fMRI data were generated using SPM12 routines `spm_int_J`, `spm_fx_fmri` and `spm_gx_fmri` (`http://www.fil.ion.ucl.ac.uk/spm/`). Routine `spm_gx_fmri` was modified to generate different haemodynamic responses, each generated randomly drawing θ_h from the empirical distributions reported in [25].

Two generative models, with respectively 7 and 66 brain regions, were used. The former setup resembles a local brain network, while the latter simulates a whole-brain network. A fixed sparsity pattern was assigned to matrix A in both setups. In the 7-regions network the non-zero entries were fixed in order to resemble the connectivity of a local brain network. The 66-nodes network was obtained using the human connectome derived from diffusion-weighting imaging in [34]: following [18], a structural connection between two brain areas was assumed to be present only if the average inter-regional fiber density was larger than 0.06. This thresholding favored the stability of the DCM constructed starting from the connectome matrix. In both setups, the absolute values of the non-zero (off-diagonal) entries of A were sampled from a normal distribution with mean 0.2 and variance 0.0025, while their signs were drawn from a Bernoulli distribution with parameter $p = 0.5$. The diagonal entries of A were fixed to -0.5 to prevent instability issues. The endogenous fluctuations $v(t)$ were modelled using Gaussian white noise with intensity $\sigma^2 I_n = 0.01 \cdot I_n$. Using these generative models, 20 Monte-Carlo sets with $N = 300$ samples of BOLD signal time-series were generated, randomizing both the driving noise and the effective connectivity matrix A. Several sampling times $T_R = \{0.5\text{s}, 1\text{s}, 2\text{s}\}$ were tested in the 7-regions setting, while only $T_R = 2s$ was considered for the 66-nodes DCM. The signal-to-noise ratio (SNR) was fixed to 3 in all datasets. Figures S1 and S2[3] in the Supplementary Material show a sample of the generated data.

In addition, to study the impact of data SNR on the performance of our algorithm, we conducted an extensive simulation study using a fixed connectivity matrix A, defined

[3] All the tables, sections and figures referenced with the prefix "S" (e.g. S1(a)) are found in the Supplementary Material.

as

$$A = \begin{bmatrix} -0.5 & 0 & 0 & 0 & -0.2 & 0 & 0 \\ 0 & -0.5 & 0 & -0.45 & -0.3 & 0 & 0 \\ 0 & 0 & -0.5 & 0.8 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & -0.5 & -0.1 & 0.6 & 0 \\ 0.3 & 0 & -0.55 & 0 & -0.5 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0.3 & -0.5 & 0.45 \\ 0.15 & 0 & 0.2 & 0 & 0 & 0 & -0.5 \end{bmatrix}. \tag{28}$$

We generated 9 datasets with 20 Monte-Carlo runs each, by varying the data SNR in the set $\{1, 5, 10\}$ and the sampling time T_R in the set $\{1, 5, 10\}$. The generated BOLD time-series contained again $N = 300$ samples and the endogenous fluctuations $v(t)$ were modelled as described above.

Remark 1. *Note that in this paper synthetic data were always generated using a sparse directed connectivity matrix A. This reflects the belief that brain networks are organized as small world (and thus sparse). In future work we shall also consider more general conditions where the model might be quasi-sparse or not sparse at all, in order to test how different approaches perform in terms of approximating a generic model with a sparse one.*

We tested our method against several state-of-the-art algorithms in terms of its ability to retrieve the true underlying directed connectivity, namely:

- *spDCM with post-hoc selection* [22]. The SPM12 routines `spm_dcm_fmri_csd` and `spm_dcm_post_hoc` were used with driving inputs a-priori switched off.

- *Multivariate Granger Causality* (MVGC). The order of the estimated VAR model was chosen through Bayesian Information Criterion (BIC) for the data coming from the 7-regions DCM and by means of Akaike Information Criterion (AIC) when dealing with the whole-brain DCM. In both cases, the Geweke's χ^2 test with FDR correction and significance level equal to 0.2 was used to detect the connectivity structure. Routine `tsdata_to_var` of the MVGC Matlab Toolbox [2] was used to estimate effective connectivity, while routines `var_to_autocov`, `autocov_to_pwcgc` and `mvgc_pval` were used to assess the connectivity structure.

- Some causal search algorithms included in the suite Tetrad (`http://www.phil.cmu.edu/tetrad/`):

  - *Peter and Clark* (PC) algorithm equipped with Fisher-Z test [69];
  - *Peter and Clark* method using Fast Adjacency Search stable algorithm [9] for the adjacency estimation (PCstable), also equipped with Fisher-Z test;
  - *Fast Greedy Equivalence Search* (FGES) adopting Fisher-Z score [49, 47];
  - *Linear Non-Gaussian Acyclic Modelling* (Lingam) [64];
  - an optimized version of the *CCD algorithm* (CCDmax) using Fisher-Z test [56];

- *Fast Adjacency Skewness* (FASK) algorithm equipped with Gaussian BIC score as a conditional independence test [61].

For all these algorithms we used classes `LoadContinuousDataAndSingleGraph` and `Simulations` to import the synthetic BOLD data and the true effective connectivity graph. We used classes `Statistics` and `Comparison` (in particular the routine `compareFromSimulations`) to evaluate the performance of the various algorithms. Finally, for all of them we used the default parameters settings.

The performance was measured both in terms of Root Mean Squared Error (RMSE) on the estimated connectivity matrix A, as well as in terms of accuracy, precision, sensitivity and specificity in retrieving the effective connectivity network (presence/absence of directed links). These are defined as:

$$\begin{aligned} \text{accuracy} &= \frac{\text{TP}+\text{TN}}{\text{P}+\text{N}}, & \text{precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\ \text{sensitivity} &= \frac{\text{TP}}{\text{P}}, & \text{specificity} &= \frac{\text{TN}}{\text{N}} \end{aligned} \tag{29}$$

where P and N respectively denote the number of non-zero (existing edges) and zero entries in the true connectivity matrix, while TP and TN are respectively the number of non-zero and zero entries that are correctly retrieved by the estimation algorithm; finally, FP is the number of connections that exist in the estimated connectivity, but do not exist in the true effective network.

Remark 2. *We warn the reader that all these measures can be criticized to some extent, as they compare the estimated model with the "true" model. Of course in practice a true model does not exist and, most importantly, several models of different complexity may explain the observed data, thus calling for methods that, as ours and Bayesian model reduction, trade complexity with fit. We stress that one of the final goals of whole brain modelling is to find an interconnection structure that can be interpreted and used for clinical and translational purposes. Thus, we regard as a plus the ability to recover a model which is close to some ground truth for "typical" sparse network topologies.*

2.5. Empirical data: 66 regions

We applied our algorithm to the dataset used in [44] consisting of 48 BOLD time-series measured in 24 right-handed healthy young volunteers (15 females, age range 20-31 years).[4] Two scanning sessions of 10 minutes, sampling time $T_R = 2$ sec, are available for each subject. Participants were asked to relax and maintain fixation on a

[4] We report here the Ethics statement included in [44]: "This research was conducted in agreement with the Code of Ethics of the World Medical Association (Declaration of Helsinki) and informed consent was obtained from all subjects before performing the study, in accordance with institutional guidelines. The study design was approved by the local Ethics Committee of Chieti University and the local Ethics Committee of Lausanne."

red point of 0.3 visual degrees positioned in the center of a black screen during scanning. Data were acquired on a 3T MR scanner (Achieva; Philips Medical Systems) using a T2-weighted echo-planar-imaging (EPI) sequence ($T_R = 2000$ ms, $T_E = 35$ ms, 32 axial slices, voxel size 3×3×3.5 mm^3). Data pre-processing was performed using the SPM5 software package (Wellcome Department of Cognitive Neurology, London, UK) with the following steps: (1) correction for slice-timing differences; (2) correction of head-motion; (3) co-registration of the anatomical image and the mean functional image; (4) spatial normalization of all images to the MNI space with a voxel size of 3×3×3 mm^3; (5) spatial Independent Component Analysis (ICA) of the BOLD time-series in MNI space for the removal of artifacts due to blood pulsation, head movement and instrumental spikes. Finally, for each recording session, the mean BOLD time-series were extracted from the $n = 66$ brain regions of the Hagmann atlas [34]. Further details on the acquisition and processing of these data can be found in Section "Methods" of [44], while the list of ROIs and their abbreviations is reported in Table S11 of the Supplementary Material.

To validate the estimated models, we compared the PearsonCC between the empirical FC matrix of a given data run and the model FC inferred using the same data (see Eq. (17)). In addition, we also compared the latter model FC with the empirical FC estimated from the *s*econd data run for the same subject. This was done to evaluate, on the one hand, the dependence of the estimated DCM on the specific data run and, on the other hand, to what extent the estimated effective connectivity is able to capture subject-specific features.

Next, a one-sample t-test was performed to assess which effective connections are stable across subjects in the population. In addition, we exploited graph theory measures to characterize the estimated effective connectivity networks. We used the Brain Connectivity Toolbox (BCT, `https://sites.google.com/site/bctnet/`) [58] to compute centrality measures such as *strength*, *betweenness centrality*, *within-module degree z-score* and *participation coefficient* of the network nodes, or segregation measures such as the *clustering coefficient*. The purpose of the latter analyses is to understand which brain regions play a role in favouring network segregation (*provincial hubs*) and which instead are crucial for network integration (*connector hubs*). Since these two properties have been widely studied in undirected brain networks arising from structural or functional connectivity [73, 5, 89], we conducted the same analysis on both effective and functional graphs in order to assess the role of directionality in brain connectivity.

Finally, we investigated the impact of EM initialization on the estimated effective connectivity A by comparing two initialization strategies for A.

3. Results

3.1. Synthetic data: 7 regions DCM

We start our experimental validation by suggesting a criterion for the thresholding of the estimated effective connectivity matrix. The top plot in Fig. 1(a) reports the

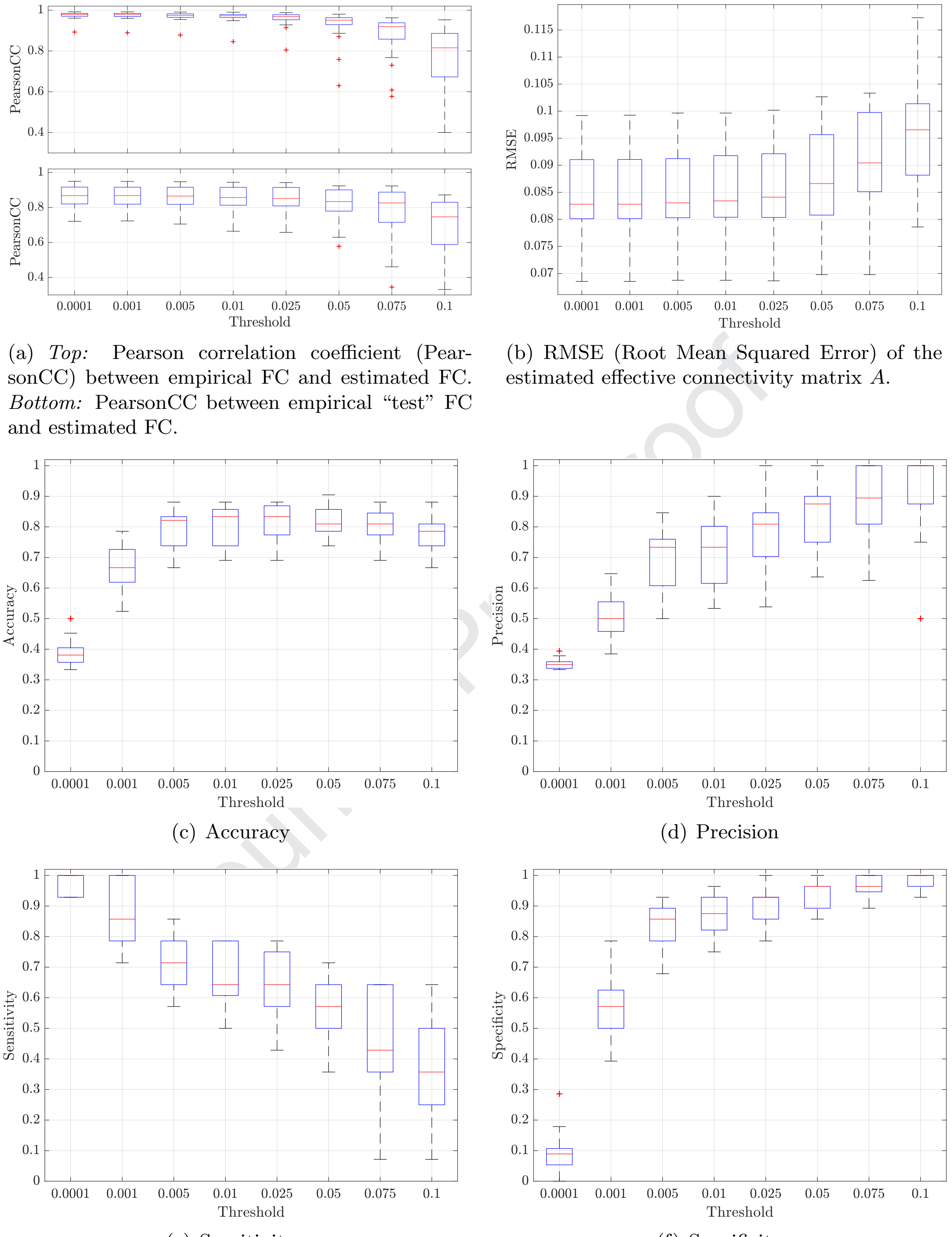
PearsonCC [5] between the empirical FC matrix directly computed from the BOLD time-series and the estimated model FC (calculated as in Eq. (17)) as a function of the threshold applied to the estimated effective connectivity matrix A. It is apparent that increasing the threshold value leads to a deterioration of the agreement between the two FC matrices. We suggest to fix the threshold to the largest value that leads to a degradation of at most 3% in the correlation between the empirical and the estimated FC. As a result the thresholds ranges between 0.025 and 0.075 in the 20 Monte-Carlo runs. This choice of the threshold provides good generalization capabilities in terms of predicting empirical FC on a new dataset (Fig. 1(a)-bottom). This choice also leads to a good estimate of A. The RMSE is essentially not affected by the thresholding (Fig. 1(b)). Accuracy, precision and specificity (Figs. 1(c)-1(d)-1(f), respectively) improve if a larger threshold is adopted. However, increasing the threshold leads to a worse sensitivity (Fig. 1(e)), since the connectivity matrix becomes too sparse and many links are not detected.

Adopting this threshold strategy, we compared the performance of our sparse DCM with the other methods listed in Sec. 2.4. The results are reported in Fig. 2 for $T_R = 2$s. Our sparse DCM approach always appears within the two best-performing methods. In particular, the performance in terms of RMSE are comparable with those achieved by MVGC and are superior to those obtained by the algorithms included in the suite Tetrad (see Fig. 2(a)). Concerning the reconstruction of the true effective connectivity structure, we observe that sparse DCM provides very good results in terms of accuracy and sensitivity. The performance related to sensitivity is superior if compared to the algorithms of the suite Tetrad. The performance of spDCM in terms of sensitivity is very poor. Indeed, it tends to overestimate the degree of sparsity in the effective connectivity matrix. Tables S1-S5 show the comparison for different sampling times T_R. Notably, sometimes spDCM estimated completely disconnected networks, thus making it impossible to compute precision (see Table S5). MVGC may incur in a similar behavior if the significance level of the Geweke's χ^2-test is not properly set. This test is used by MVGC to select the significant connections. We observed that a larger significance level may prevent an excessive sparsity in the estimated connectivity matrix. We set it to 0.2 in the reported simulations.

Overall, we can conclude that our method outperforms the competitors in detecting "true" effective connections (in terms of sensitivity); remarkably, this is achieved while maintaining a good specificity.

Fig. 3 shows the performance of our approach as a function of sampling time and data SNR. In this case, the synthetic BOLD time-series were generated with the fixed connectivity matrix A in Eq. (28). The plots highlight how our approach significantly benefits from larger SNRs. Somewhat surprisingly, performance moderately improves when T_R increases. This behavior may be explained by the fact that low T_R data are gathered in a shorter time-horizon (since all the designed datasets always contain 300

[5] The Pearson Correlation Coefficient is computed only between the upper diagonal parts of the two FC matrices, due to their symmetry.



(a) *Top:* Pearson correlation coefficient (PearsonCC) between empirical FC and estimated FC. *Bottom:* PearsonCC between empirical "test" FC and estimated FC.

(b) RMSE (Root Mean Squared Error) of the estimated effective connectivity matrix A.

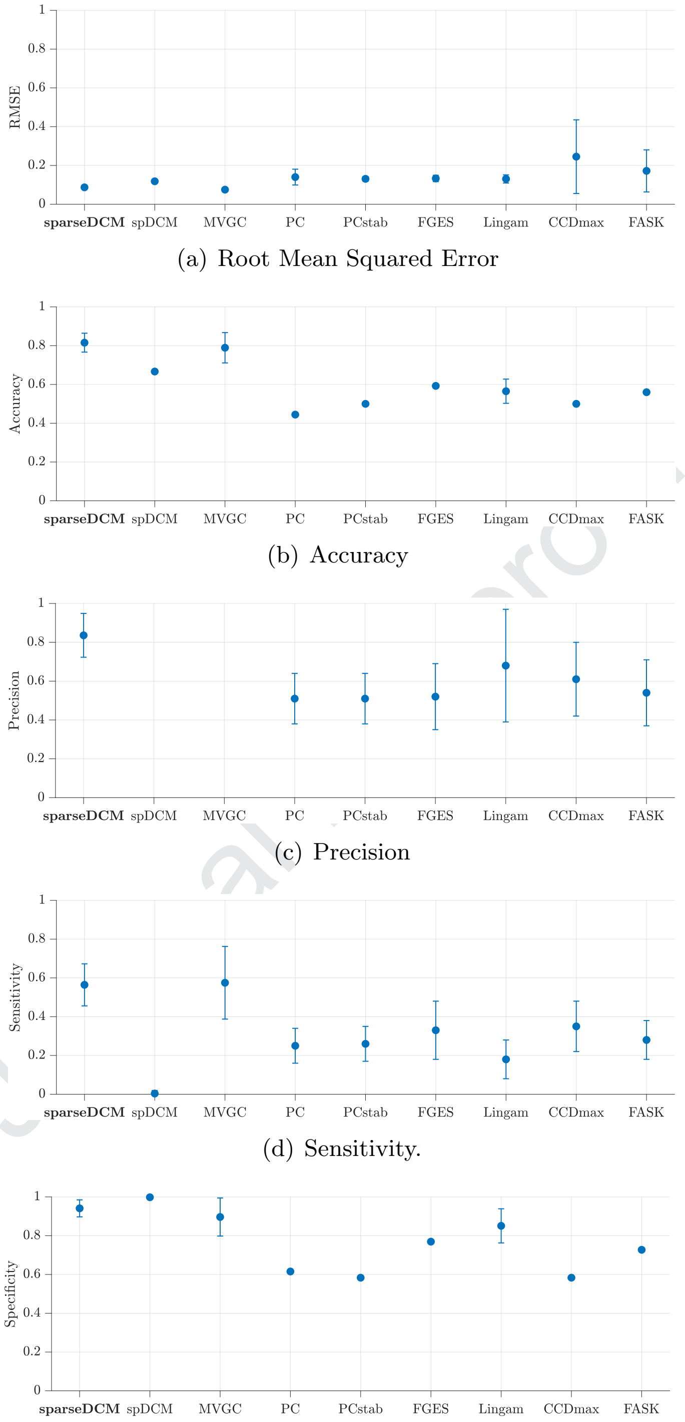
(c) Accuracy

(d) Precision

(e) Sensitivity

(f) Specificity

Figure 1: Synthetic data with 7 brain regions (nodes) and randomly drawn connectivity matrix (SNR=3, T_R=2s). Performance metrics as function of the thresholding applied on the estimated ECs.



(a) Root Mean Squared Error

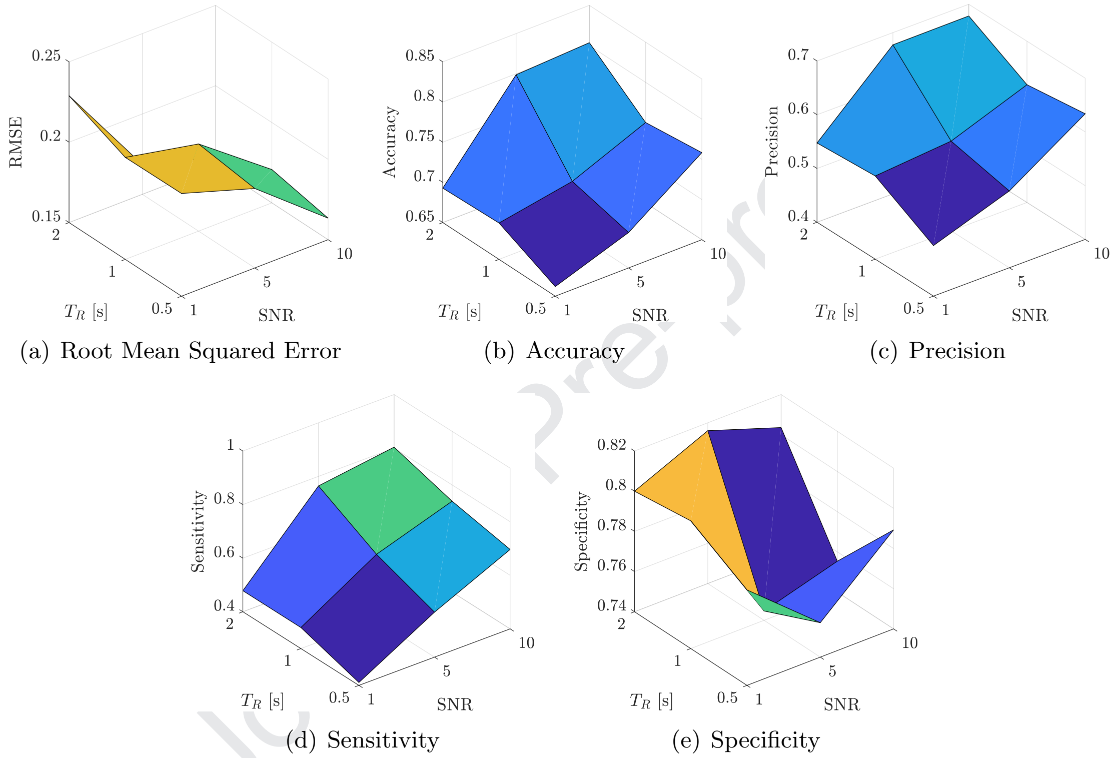
(b) Accuracy

(c) Precision

(d) Sensitivity.

(e) Specificity

Figure 2: Synthetic data with 7 brain regions (nodes) and randomly drawn connectivity matrix (SNR=3, T_R=2s). Performance metrics over 20 MC runs (mean ± standard deviation) are shown for our sparse DCM as well as for the compared methods.



(a) Root Mean Squared Error

(b) Accuracy

(c) Precision

(d) Sensitivity

(e) Specificity

Figure 3: Synthetic data with 7 brain regions and fixed connectivity matrix. Average performance metrics over 20 MC runs achieved by the proposed algorithm as function of data SNR and sampling time T_R.

samples), thus they might not be enough informative about the underlying dynamics. Indeed, a similar behavior is observed for spDCM in Tables S6-S10, which compare the performance of the tested algorithms on the same datasets. This trend is further confirmed by the results achieved on the data generated with randomly sampled true connectivity matrices: by inspecting Tables S1-S5 it can be noticed that not only our approach but also the method relying on Granger causality (MVGC) and FGES achieve better performance on data having higher sampling time.

As a final comparison, we report in Table 1 average execution times (in seconds) for each method over the 20 MC runs. The value reported under Tetrad is the sum of the execution times of all the algorithms included in the suite (i.e. PC, PCstab, FGES, Lingam, CCDmax and FASK). Simulations were conducted with a Macbook Pro 2017 (2.5 GHz Intel Core i7 processor, 16 GB RAM). As expected, the computational effort is low for the correlation based algorithms (Tetrad and MVGC), which simply compare correlations between the variables included in the model. Inverse methods, such as our sparse DCM and the spectral DCM, require the inversion of the specified model, thus being more expensive from the computational point of view.

	sparseDCM	spDCM	MVGC	Tetrad
Avg execution time [s]	133	62	0.10	0.42

Table 1: Synthetic data with 7 brain regions (nodes) and randomly drawn connectivity matrix (SNR=3, T_R=2s). Average execution time per run (computed over 20 MC runs).

3.2. Synthetic data: whole-brain-scale network

We now analyse the performance of sparse DCM in a more realistic whole brain setting (66 regions), still using synthetic data. The thresholding procedure was the same used in Sec. 3.1, leading to selected thresholds in the range [0.01, 0.025].

We compared sparse DCM with the methods listed in Sec. 2.4 (see Fig. 4) but excluding spDCM and Lingam due to their high computational load. Also CCDmax had to be dropped because it did not converge on most of the Monte-Carlo datasets. Overall, the results in Fig. 4 are in favor of sparse DCM, showing that its performance scales well with network size. The perfect score achieved by MVGC in terms of specificity is due to the fact that it provides, in most runs, a completely disconnected network. As a result its performance in terms of sensitivity is very poor. Differently, the high specificity performance obtained by our method is also accompanied by satisfying sensitivity and precision scores. The latter are the highest among the compared approaches.

	sparseDCM	spDCM	MVGC	Tetrad
Avg execution time [s]	19840	–	85	130

Table 2: Synthetic data with 66 brain regions and randomly drawn connectivity matrix (SNR=3, T_R=2s). Average execution time per run (computed over 20 MC runs).

Table 2 contains the average execution times per MC run of the compared algorithms. The simulations were conducted with the same hardware described in Section 3.1. Despite sparse DCM is significantly more expensive than the correlation-based approaches, it scales better than MVGC and Tetrad when the number of monitored brain regions increases.

3.3. Empirical data: 66 regions

We now consider the empirical fMRI data described in Sec. 2.5. Subjects 12 and 18 of the dataset have been excluded from the analyses reported below due to convergence problems in one of the two runs.

3.3.1. Effective and functional connectivity

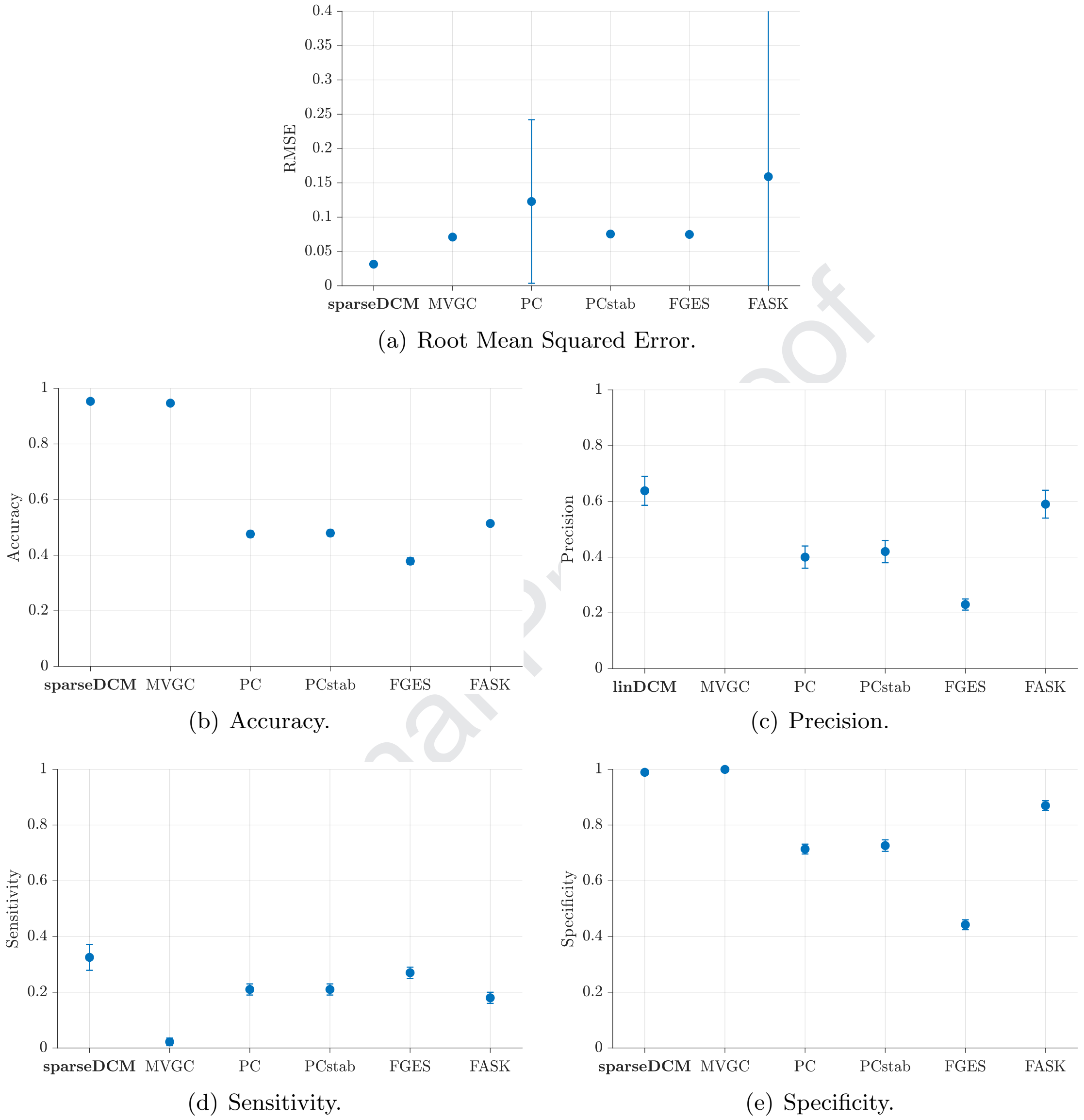
We first consider the data of a single subject to illustrate the sparse DCM outputs. Fig. 5(a) shows the estimated effective connectivity before thresholding. The matrix is actually (almost) sparse, with many entries very close to zero, even if not exactly zero. Fig. 5(b) illustrates the linear haemodynamic responses estimated for each of the 66 brain regions of the Hagmann atlas [34]. Their average is reported in black. It is interesting to observe that our algorithm indeed captures a significant variability of the haemodynamic responses for different brain areas. Finally, the agreement between empirical FC and model FC reconstructed using the estimated DCM can be appreciated by comparing Figs. 5(d) and 5(c).

The agreement between empirical and estimated FC is confirmed for the entire sample of subjects in terms of Pearson correlation coefficient (see the blue dots in Fig. 6). Most notably, Fig. 6 also reports the PearsonCC between the model FC coming from the DCM estimated using data from Run 2 and the empirical FC computed from Run 1 (red diamonds): this comparison can be viewed as a “model validation” stage, which aims at assessing the generalization capabilities of the estimated models. For completeness, the PearsonCCs between the empirical FCs from Run 1 and Run 2, which may be regarded as ceiling level for the corresponding red diamonds, are also shown (black squares).

3.3.2. Effective Connectivity Thresholding

Similarly to the synthetic scenario, the threshold value was fixed to 0.01 following the same selection approach we adopted with synthetic data. This threshold also guarantees a large agreement between the model FC and the empirical one, when computed on a different data run (Run 1 in this case), as shown by Fig. 7(b).

To further validate the threshold selection criterion, we exploited the availability of two scanning runs for each subject. Since we expect (a priori) that the ECs estimated from each of the two runs should be similar, we can evaluate if the chosen threshold guarantees such an agreement. To this purpose, we computed the so-called *Within Subject Similarity* (WSS), that is, the PearsonCC between the ECs inferred from the two data runs of the same subject. This quantity is shown in Fig. 7(c) as function of the thresholding. The results support our choice (i.e., 0.01), because larger values lead to a reduced similarity between the ECs inferred for the same subject.



(a) Root Mean Squared Error.

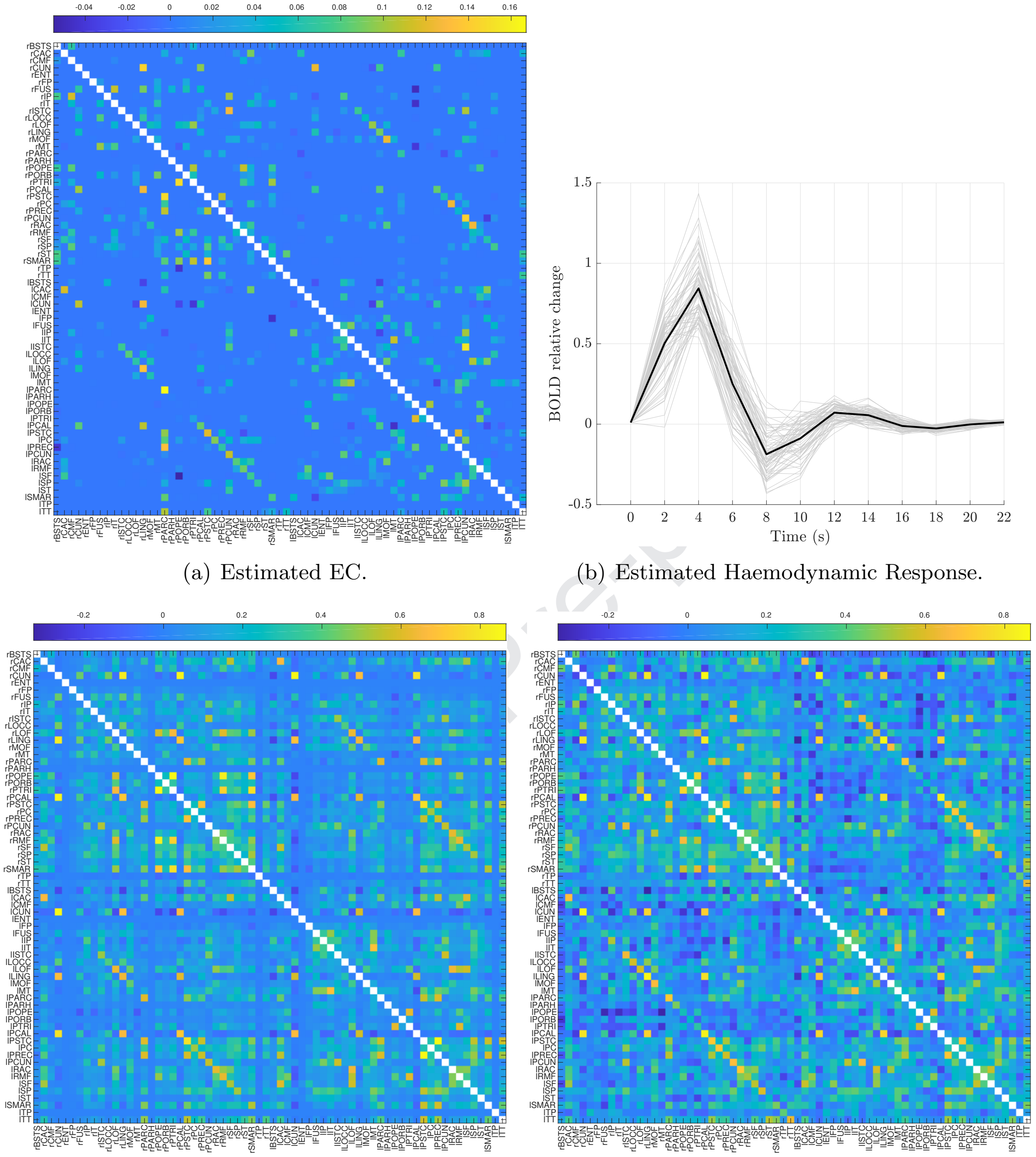
(b) Accuracy.

(c) Precision.

(d) Sensitivity.

(e) Specificity.

Figure 4: Synthetic data with 66 brain regions (nodes) and randomly drawn connectivity matrix (SNR=3, T_R=2s). Performance metrics over 20 MC runs (mean ± standard deviation) are shown for our sparse DCM as well as for the compared methods.



(a) Estimated EC.

(b) Estimated Haemodynamic Response.

(c) Estimated FC.

(d) Empirical FC.

Figure 5: Single subject analysis (Subject 17, data from Run 2). (a) Estimated Effective Connectivity (EC). (b) Estimated haemodynamic responses for each of the 66 BOLD time-series (i.e., brain regions) and the corresponding mean *(black solid line)*. (c) Functional Connectivity reconstructed from the subject's estimated EC shown in panel (a). (d) Empirical Functional Connectivity.

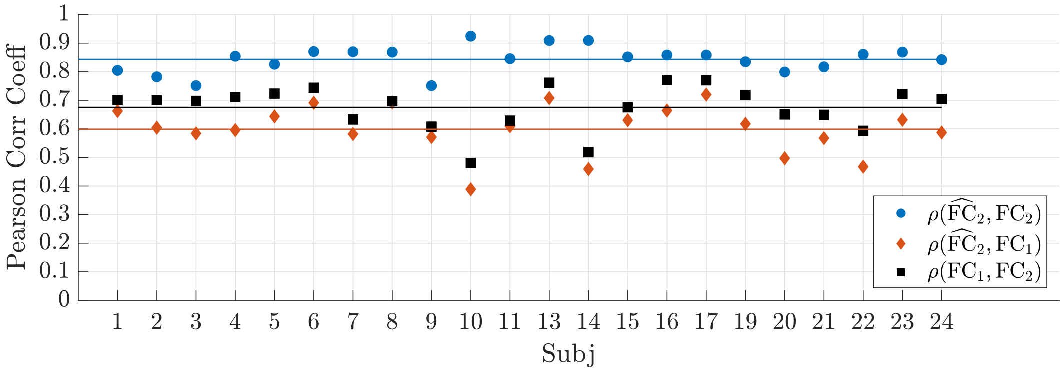


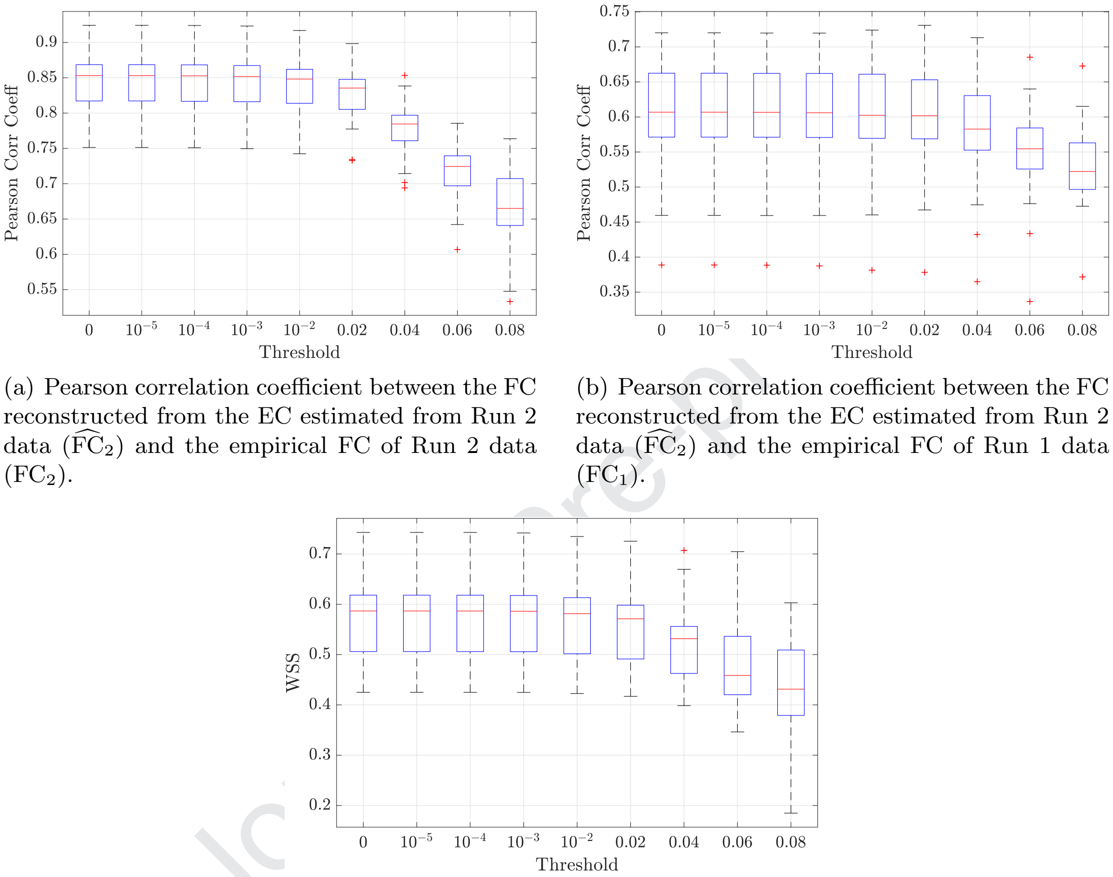
Figure 6: *Blue dots:* Pearson correlation coefficient (PearsonCC) between empirical FCs of Run 2 data (FC_2) and the functional connectivity $\widehat{\mathrm{FC}}_2$ reconstructed from the ECs estimated using the same data (see Eq. 17). *Red diamonds:* PearsonCC between empirical FCs of Run 1 data (FC_1) and the functional connectivity $\widehat{\mathrm{FC}}_2$ reconstructed from the ECs estimated using Run 2 data. *Black squares:* PearsonCC between empirical FCs of Run 1 data (FC_1) and empirical FCs of Run 2 data (FC_2). *Horizontal lines:* Corresponding average values across the entire sample of subjects.

3.3.3. Population Study: Effective Connectivity

Using the optimal threshold value identified in the previous section, we now proceed to (i) assess the stability of the connections across the entire sample of subjects, and (ii) briefly discuss the structure of the resulting population-level EC matrix.

We performed a one-sample t-test with FDR ($= 0.05$) correction on the ECs of the 22 subjects. The results for the two data runs are reported respectively in Figs. 8(a) and 8(b). Black and red entries denote those links for which the null hypothesis (corresponding to the absence of the link) was rejected at a significance level $\alpha = 0.05$. Brain regions are ordered according to a left-right subdivision: black and red squares respectively denote intra- and inter-hemispherical connections. The same results are reported in Fig. S7 where brain regions are grouped according to a functional atlas.

Inspection of the EC matrices suggests good agreement between runs, which is confirmed by a correlation of 0.82. Note that, intra-hemispherical connections (black squares in Fig. 8(a)) are much more frequent than inter-hemispherical ones (red squares in Fig. 8(a)). Most of the latter are actually connecting the same region in the two hemispheres (notice the red diagonals appearing in the upper right and lower left sides of Figs. 8(a)-8(b)). Most notably, the EC results reveal some directed connections that are stable either in the two hemispheres and in the two data runs. Among them, there are the links from the paracentral lobule (PARC) to the postcentral gyrus (PSTC) and to the precentral gyrus (PREC), i.e. between regions of the somatosensory-motor network, as well as to the posterior cingulate cortex (PC), suggesting a link between somatosensory-motor and default-mode networks. Moreover, we should also note the links between the auditory and the integration regions [44], i.e. those from the superior temporal cortex (ST) to the supramarginal gyrus (SMAR) and from the superior parietal cortex (SP) to SMAR. There seems also to be a strong relationship among the



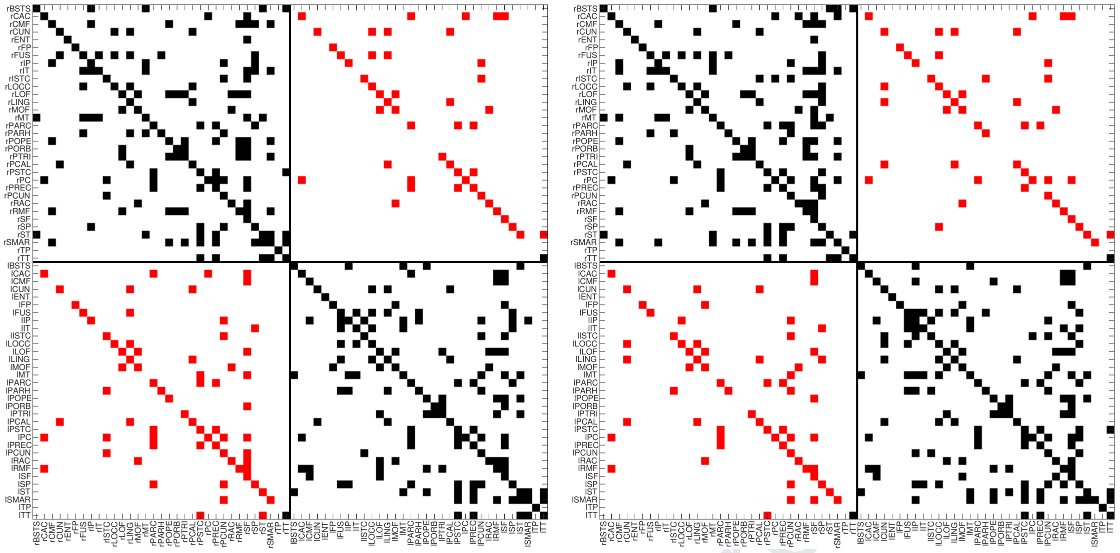
(a) Pearson correlation coefficient between the FC reconstructed from the EC estimated from Run 2 data ($\widehat{\mathrm{FC}}_2$) and the empirical FC of Run 2 data (FC_2).

(b) Pearson correlation coefficient between the FC reconstructed from the EC estimated from Run 2 data ($\widehat{\mathrm{FC}}_2$) and the empirical FC of Run 1 data (FC_1).

(c) Within Subject Similarity (WSS), measured in terms of Pearson correlation coefficient between ECs estimated from Run 1 data and Run 2 data.

Figure 7: Impact of EC thresholding on the quality of FC reconstruction and on the stability of EC estimates across fMRI runs for the entire sample of subjects (N=22). All metrics are shown as a function of EC threshold value.

.



(a) t-test on ECs - Run 1.

(b) t-test on ECs - Run 2.

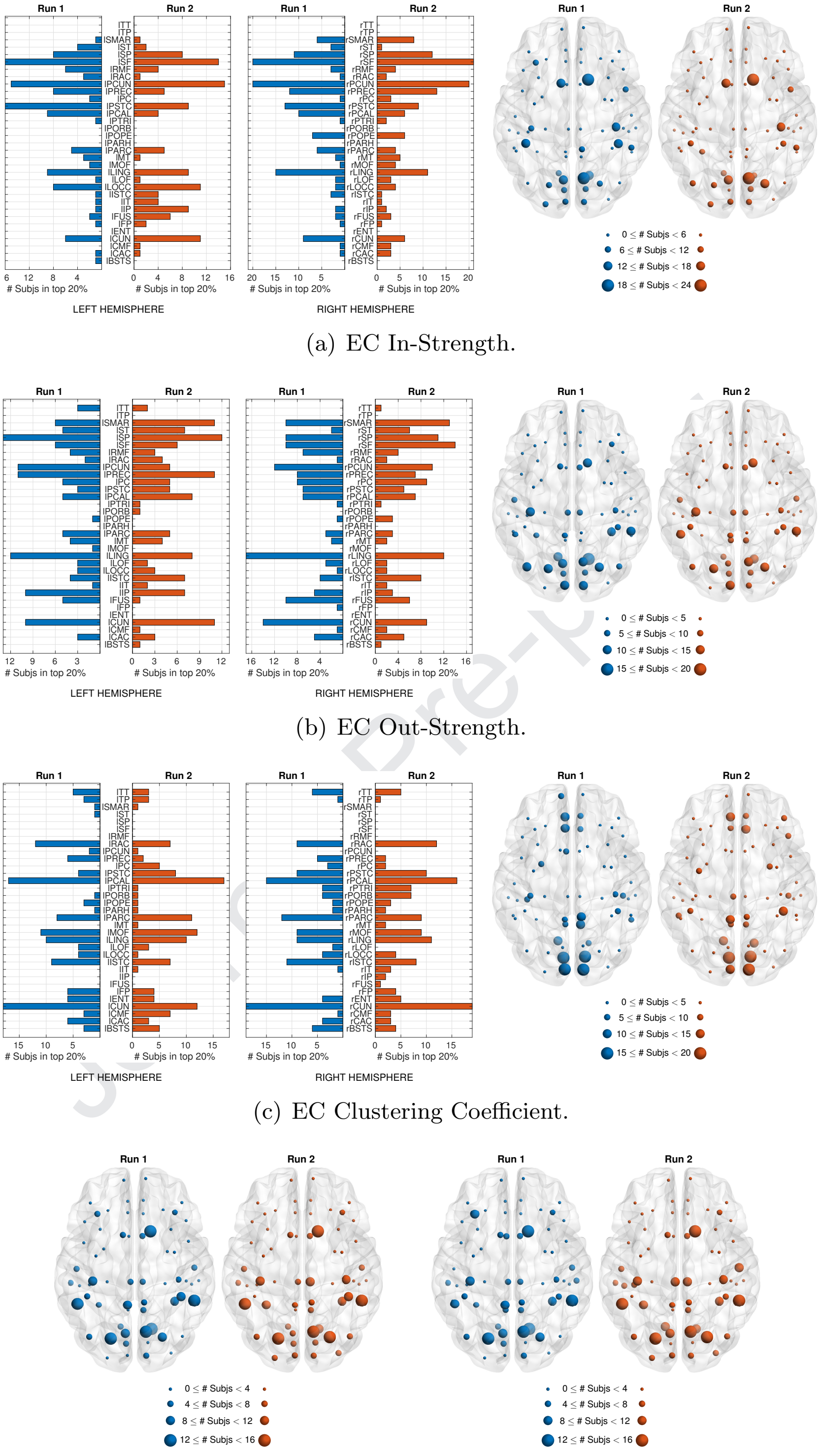
Figure 8: Population Analysis. One-sample t-test (p-value$<$ 0.05 FDR corrected) over ECs estimated using data from Run 1 (panel (a)) or Run 2 (panel (b)). The null hypothesis is rejected for colored entries (black and red squares respectively denote intra-emispheric and inter-hemispheric connections).

pars orbitalis (PORB), the pars opercularis (POPE), the pars triangularis (PTRI) and the rostral anterior cingulate cortex (RMF): PTRI is influenced by POPE, PORB and RMF. Analogously, RMF is conditioned by POPE, PTRI and PORB. Finally, PTRI affects POPE and PORB. A more in-depth network analysis is provided in the next Section.

3.3.4. Population Study: Network Analysis

We conducted a graph theoretical analysis of the effective connectivity networks estimated using the dataset described in Sec. 2.5. Some of the most popular network measures [58, 5] were computed for each subject. These are summarized in Appendix D. For each network measure, nodes were sorted in decreasing order with respect to the chosen metric. We then determined the number of subjects in which a certain node was in the top 20% of the corresponding ranking. The same analysis was performed on the functional connectivity data. The aim of this set of analyses was threefold: first, to assess which brain regions play a relevant role within the effective connectivity graph; second, to evaluate the consistency between the measures computed for the graphs estimated from the two data runs; and third, to compare effective and functional connectivity networks.

Considering EC networks, we notice that the superior frontal cortex (SF) shows a high in-strength value (see Fig. 9(a)), suggesting that executive functions play a key role in the resting-state network. In particular, this suggests that SF is prone to be



(a) EC In-Strength.

(b) EC Out-Strength.

(c) EC Clustering Coefficient.

(d) FC Strength.

(e) FC Clustering Coefficient.

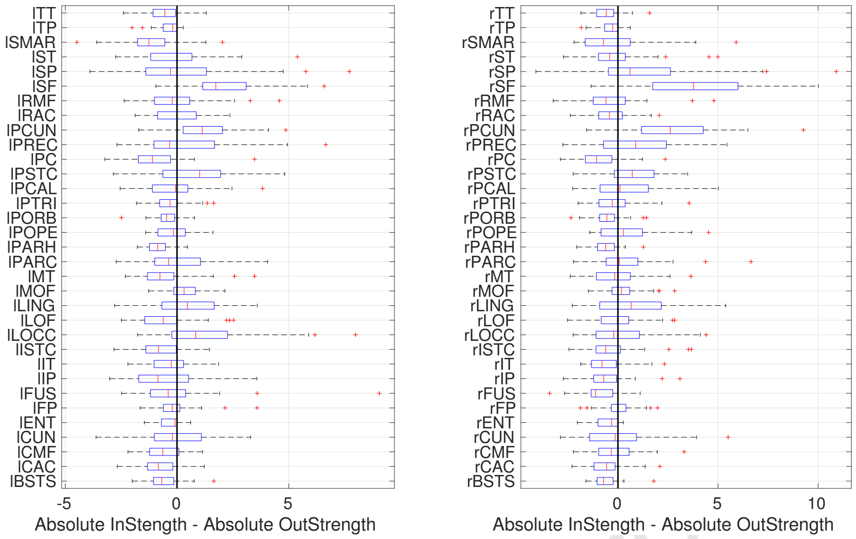
Figure 9: Number of subjects in which the weighted *strength* (or *clustering coefficient*) of a certain node is in the top 20%. (a)-(c) refer to the effective connectivity graph; (d)-(e) refer to the functional connectivity graph.

controllable by other regions, but has a weaker ability to control other areas. This is confirmed by Fig. S9(a), which refers to centrality and, to a minor extent, by Fig. 9(b), which instead considers out-strength. Moreover, the in- and out-strength of the SF node seem to be asymmetric in the two hemispheres, with the right one typically having a larger strength. This asymmetry is also evident when looking at the EC betweenness centrality. Finally, the asymmetry of the SF node can be detected in the FC strength (Fig. S8(a) and 9(d)). Other cases of asymmetry between the two hemispheres involve the precuneus (PCUN) and the pars opercularis (POPE), which appear to be more influential in the right hemisphere, according to Fig. 9(a), and the lateral occipital cortex (LOCC). The latter behaves differently in terms of both in-strength and betweenness centrality. Other significant nodes in terms of in-strength (Fig. 9(a)) and out-strength (Fig. 9(b)) are the superior parietal cortex (SP) and the lingual gyrus (LING). Not surprisingly, Figs. 9(a)-9(b) and Fig. 9(d) show a significant agreement between the nodes strengths of the effective and functional graphs.

Next, segregation is analyzed using the weighted clustering coefficient, displayed in Fig. 9(c). The cuneus (CUN) and the pericalcarine cortex (PCAL) seem relevant in segregation. A low out-degree participation coefficient and a high within-module z-score in a consistent portion of the population confirm this role. A similar conclusion can be drawn for the rostral anterior cingulate cortex (RAC). At a glance, the 3D brain plots in Figs. 9(c) and 9(e) show a clear difference between the effective and functional networks in terms of clustering coefficient: while in the functional graph the posterior brain regions are typically part of small clusters, in the effective graph this property seems to characterize only the pericalcarine cortex (PCAL) and the cuneus (CUN). Moreover, comparing Figs. 9(c) and S8(b), we can notice that the supramarginal gyrus (SMAR), the superior temporal, parietal and frontal cortex (ST,SP,SF) often tend to have small clustering coefficient. Accordingly, they seem to be associated with low local efficiency of information transfer for specialized processing (functional segregation). An opposite situation is observed for the pericalcarine cortex (PCAL) and the cuneus (CUN), which belong to the visual network.

The weighted participation coefficient and the weighted within-module z-score (Figs. S10(a), S11(b)) show that the posterior cingulate cortex (PC), the superior parietal cortex (SP) and the left middle temporal cortex (MT) are characterized by a high participation coefficient and a low within-module z-score in a significant fraction of the population, thereby facilitating integration in the effective connectivity network. Analogous properties are observed for the parahippocampal cortex (PARH) in the functional graph (compare Figs. S10(c), S11(c)).

We conclude our network analysis studying the directionality of EC graphs. In particular, sources and sinks are revealed computing the difference between the absolute in- and out-strength. Boxplots of the latter quantities are reported in Fig. 10. Several regions can be classified as sources in both hemispheres, such as the posterior cingulate cortex (PC), the pars orbitalis (PORB), the parahippocampal cortex (PARH), the caudal anterior cingulate cortex (CAC) and the bank of the superior temporal sulcus (BSTS). On the other hand, only the superior frontal cortex (SF) seems to play a relevant sink



(a) Left Hemisphere

(b) Right Hemisphere

Figure 10: Asymmetry in the effective connectivity network. Each boxplot shows the difference between the absolute in- and out-strength for a specific brain region.

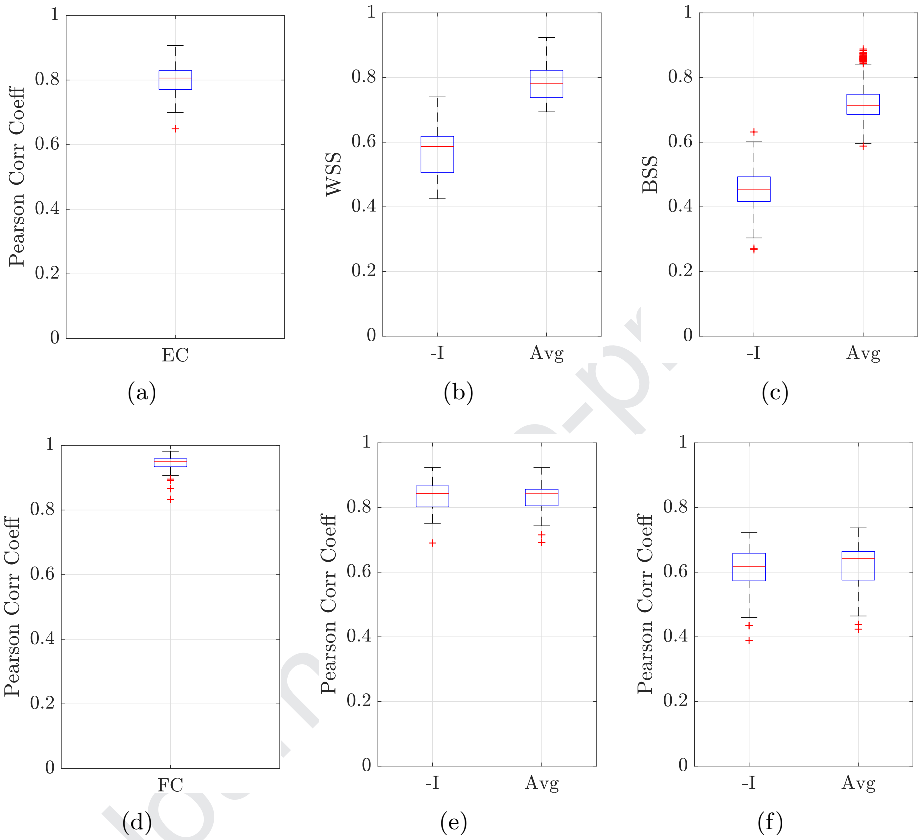
role. Some regions also show a different behavior in the two hemispheres: for instance, the left supramarginalgyrus (SMAR) mainly shows to be a source, while in the right hemisphere its function appears more variable.

3.3.5. Population Study: Algorithm Initialization

The impact of different initializations of our algorithm (see Sec. 2.3 and Appendix B) is now empirically evaluated. In particular two possible initializations of effective connectivity matrix A are considered. When no a priori information is available, a simple choice would be $A^{(0)} = -I_n$[6]. On the other hand, if some prior knowledge on the effective connectivity is available, this could be exploited, e.g. setting $A^{(0)}$ to some "average" network.

These two initialization strategies for Algorithm 1 were compared using the empirical dataset illustrated in Sec. 2.5. The results of this study are reported in Fig. 11, where the two strategies are respectively denoted with "-I" and "Avg". To implement the latter we followed a "leave-one-out strategy": for each subject we set $A^{(0)}$ equal to the average of the ECs estimated from the remaining subjects in the dataset. The top row in Fig. 11 evaluates the impact of the two initializations directly on EC, while the bottom row focuses on the resulting model FC (see Eq. (17)). Specifically, Fig. 11(a) shows the PearsonCCs between the ECs returned by the two strategies over the population, while Fig. 11(d) contains the same comparison performed on the model FCs. Despite a moderate agreement between the estimated ECs, there is a very high consistency

[6] The results illustrated in the previous sections were achieved by means of this initialization.



(a) (b) (c)

(d) (e) (f)

Figure 11: Impact of Effective Connectivity initialization on the estimation algorithm: comparison between initializing with the identity matrix (-I) and with the average of the ECs estimated from other subjects (Avg). (a) Pearson correlation coefficient (PearsonCC) between ECs estimated from the two initializations. (b) Within Subject Similarity. (c) Between Subject Similarity. (d) PearsonCC between the FCs reconstructed from the ECs estimated from the two initializations. (e) PearsonCC between empirical FCs and the FCs reconstructed from the ECs estimated using the same data run. (f) PearsonCC between empirical FCs and the FCs reconstructed from the ECs estimated using the other data run.

between the FCs. This finding is confirmed by the PearsonCC between the model FCs and the empirical ones, respectively computed on the estimation BOLD time-series and on a new data run (Figs. 11(e)-11(f)). It is therefore apparent that the initialization of A does not affect the performance in terms of functional connectivity. Nonetheless, the initialization strategy does affect the estimated effective connectivity itself: depending on the starting point, the EM algorithm could converge to different local minima, giving rise to the discrepancies that can be observed in Fig. 11(a). Not surprisingly, exploiting the a-priori information to initialize the algorithm ("Avg") favours similarity among the inferred ECs. This behaviour is shown in Figs. 11(b)-11(c) which respectively report the within-subject and between-subject similarity, measured in terms of Pearson correlation coefficient between the inferred ECs.

We conducted the same analysis reported in Sec. 3.3.4 on the effective connectivity networks returned using the initialization denoted with "Avg". The results are reported in Fig. S12. There is a significant agreement on the in-strength, out-strength and clustering coefficient. The increased between-subject similarity observed in Fig. 11(c) achieved using "Avg" is also revealed in the bar plots in Fig. S12(a): compared to Fig. 9(a), fewer nodes have a large strength value across subjects, thus reflecting a reduced variability within the population. This trend is less apparent in Figs. S12(b) and S12(c). Finally, the results returned by the two initialization strategies significantly agree in terms of clustering coefficient (compare Figs. 9(c) and S12(c)), while some discrepancies can be observed when evaluating the nodes out-strength (compare Figs. 9(b) and S12(b)).

Concerning the remaining parameters in θ (see Eq. (16)), as well as the hyperparameters γ (see the discussion in Sec. 2.3), it is fair to say that the choice of initialization has only a marginal (if any) effect. The only exception regards the initialization of the standard deviation σ of endogenous fluctuation process $v(t)$. To this purpose we have developed a tailored procedure, which we experimentally verified to be rather robust (see the description in Appendix B).

4. Discussion

In this work we introduced *sparse DCM*, a novel method to estimate effective connectivity from resting-state fMRI data. Our method stands out against state-of-the-art contributions thanks to its ability to infer whole-brain graphs comprising tens of regions (66 in the experiments reported here). This was made possible by the use of a linearized model for haemodynamics and of a sparsity inducing mechanism, which automatically prunes irrelevant connections. In this way, contrary to most existing techniques, there is no need to perform a selection of candidate network structures that typically also relies on the information about structural connectivity.

Key steps underlying our approach are the discretization and statistical linearisation of the haemodynamic model [6, 28, 76]. The latter have allowed to transform the non-linear continuous time DCM into a discrete-time linear state space model. The linearized haemodynamic response accounts for empirical priors available in the literature for physiological parameters which define the nonlinear model in Eqs. (4)-(8). We then developed an EM-like algorithm to estimate this linear generative model and in particular the effective connectivity matrix.

We demonstrated the face validity of the novel method by means of numerical experiments performed on two synthetic scenarios, consisting of 7 and 66 regions (nodes), respectively. Sparse DCM was compared to several state-of-the-art methods, including spectral DCM [26, 53, 54], Multivariate Granger Causality (MVGC) [2], Fast Greedy Equivalence Search (FGES) [49, 47], Lingam [64], Fast Adjacency Skewness algorithm (FASK) [61], an optimized version of the CCD routine [56] and the “Peter and Clark” method [69, 9]. When considering the simpler scenario comprising 7 regions, our method proved to be superior to the competitors in terms of accuracy and sensitivity, while it was among the two best-performing approaches when looking at specificity, precision and RMSE. Among the other methods, MVGC yielded a competitive performance when considering specificity and precision, but this result turned out to be due to its tendency to overestimate the sparsity degree of the networks. In extreme cases, which were often detected in the larger-scale scenario consisting of 66 regions, MVGC returned a completely disconnected network. This feature makes MVCG unreliable to our purpose. In the “large” (66 regions) synthetic scenario our method outperformed all the compared approaches especially in terms of RMSE, accuracy and sensitivity.

Using the 7 regions synthetic scenario we also evaluated the sensitivity of our approach with respect to the data SNR and to the sampling frequency. As expected, sensitivity and, to a minor extent, accuracy and precision improve as SNR increases. On the other hand, the data sampling time seemed to have minor impact on the performance of sparse DCM. We envisage that future advances in the technology of fMRI scanners would, on the one hand, increase the SNR levels in the measured BOLD time-series and, on the other hand, increase the image acquisition frequency. Some developments in this sense have already been achieved by exploiting high magnetic field strengths [16], ultrafast imaging [77, 86] or by considering the confounds due to magnetic field fluctuations [4]. According to our study in the synthetic 7 regions setup, we believe that our effective connectivity estimation method will strongly benefit from these technological advances.

The proposed algorithm was also applied on empirical BOLD time-series measured in 22 healthy adults [44] to infer whole-brain effective connections. Two resting-state fMRI runs were available for each subject, thereby allowing us to estimate distinct effective connectivity matrices for each run. These two matrices showed good agreement across individuals, as measured by the Pearson correlation coefficient. The consistency of the effective connectivity estimated across runs was also confirmed when looking at the connections that are found to be stable (i.e., statistically reliable) across individuals.

We also performed a graph-theoretical analysis on the whole-brain effective connectivity graphs estimated from the empirical data. The same study was also conducted on the undirected FC graphs thus highlighting analogies and discrepancies between effective and functional networks. We believe that this preliminary investigation might serve as a possible pipeline for future studies focusing on the brain’s functional organization and on the pattern of directed interactions. However, we warn the reader that the graph measures computed in this paper should be taken with great caution.

We anticipate that the availability of an inference algorithm for whole-brain effective connectivity will serve as seed for stimulating further applications of graph theory to directed brain networks, which represents a largely unexplored area of computational neuroscience.

We emphasize that effective connectivity models, differently from functional and structural networks, also encode directionality in the connection. Our results suggest that directed connections play a key role in the evaluation of the small-word properties of the brain. Different values of the clustering coefficient can change the importance of the regions when functional segregation processes are considered, while different path length measures could indicate a different vision of the functional integration properties of the regions in rapidly combining specialized information. Thus, knowledge about the directionality of the links between brain regions could give an additional value to the interpretation of how brain networks are organized and how they generate complex dynamics. Moreover, the recent introduction of control theory methods into neuroscience is perfectly suited to the case of directed networks [40] but it remains highly controversial for undirected networks such as connectomes generated from diffusion imaging data (see [79] for a lively debate).

The empirical dataset was also exploited to investigate the impact of different initialization strategies for the sparse DCM algorithm. In particular, we discussed how previously estimated effective connectivity profiles could be used to initialize the estimation procedure. The exploitation of this prior information can, on the one hand, significantly reduce the computational effort required by the estimation routine and, on the other hand, increase both the within- and between-subject similarity of the inferred effective connectivity matrices. These findings appear particularly valuable in prospective clinical applications when, for instance, a new patient has to be screened and effective connectivity patterns from other patients (with a similar clinical profile) are already available.

Concerning the related literature, our work can be considered part of a restricted number of contributions dealing with the inference of whole-brain effective connectivity from resting-state fMRI data. These include the spectral DCM approach proposed by [54], where principal components of the functional connectivity were exploited to define the prior covariance assigned to the effective connectivity matrix, and in turn to constrain the number of DCM parameters to be estimated. This allowed to invert DCMs consisting of 36 regions. However, the required computational effort remained significant, thus making the application to larger DCMs still questionable. The DCM framework was also recently developed in order to deal with whole-brain effective connectivity estimation from task-based fMRI data [18]. The procedure, called regression-DCM, was applied to a network comprising up to 104 nodes. It also exploited a sparsity inducing prior on effective connections tuned by free energy minimization. However, differently from our approach, the haemodynamic response was held fixed, thus not accounting for region- and subject-specific variations of the mapping between neural activity and BOLD signal [35, 1]. We regard this as a major limitation of regression-DCM, making it possibly very sensitive in applications on atypical brains (e.g. post stroke) where the haemodynamic response may be severely altered in lesioned areas. Finally, though regression-DCM could potentially deal with resting-state data, it does not explicitly account for the endogenous fluctuations that are assumed to drive neural activity at rest, thus making it not directly applicable to the estimation of resting-state effective connectivity. Nonetheless, a comparison between regression-DCM and our inference procedure based on a linear DCM would help in assessing the impact of both haemodynamic variability and of endogenous fluctuations modelling. Outside the DCM framework, models that attempt to establish Granger-type causality directly on observed BOLD signals have been developed. For instance, [32] encoded effective connectivity in an Ornstein-Uhlenbeck model and estimated effective connections by fitting the model stationary covariances to the empirical ones computed from resting-state fMRI observations. Despite being very computationally efficient, this procedure does not include an haemodynamic model and has to be provided with a prior structure for the effective connectivity network (e.g., using a structural connectivity matrix). Nevertheless, this approach has been successful in retrieving signatures for subject identification [42] and in task recognition [31]. We believe that a comparison between this technique and our inversion scheme would reveal whether haemodynamic modelling is relevant in the context of effective connectivity or might be neglected. Whole-brain estimates can also be obtained by resorting to a class of model-free methods, typically referred to as "Bayesian Nets" [3]. These approaches include the "Peter and Clark" algorithm, the Cyclic Causal Discovery procedure (CCD), Greedy Equivalence Search (GES) and fast GES (FGES). Even if these methods are very fast, they typically return only an equivalence class of graphs, whose members can be distinguished only using further assumptions. Nonetheless, FGES was recently applied to all the cortical voxels in a resting-state fMRI scan (around 51000 voxels) [47]. [15] used FGES as a preliminary step to infer the set of candidate structures and exploited it to subsequently derive subject-specific networks. In this case, a comprehensive comparison between these model-free techniques and our approach would provide insights on the importance of modeling when dealing with brain directed interactions. We plan to conduct this and the aforementioned comparisons in a future contribution.

In addition to the detailed comparisons with other state-of-the-art techniques, future developments of sparse DCM include a study about the modelling of brain endogenous fluctuations that are supposed to drive the neural activity at rest. Despite these are typically assumed to be scale-free processes [20, 72, 65], our model considered them as Gaussian white noise. A preliminary investigation on the plausibility of this assumption was already conducted in a recent work [46], where also first-order autoregressive models for endogenous fluctuations were considered. However, we believe that a more in-depth analysis should be conducted, considering larger datasets and more complex autoregressive models.

The ultimate goal of this work would regard clinical applications and, in particular, the possibility to detect individual differences in the effective connectivity profiles of patients that are predictive of the clinical outcomes. In future contributions we plan to further analyze the plausibility of the linear model and, subsequently, to apply our estimation procedure to fMRI data measured in neurological subjects in order to characterize the discrepancies between healthy and damaged brains.

Acknowledgments

Part of this work has been supported by the University of Padova through the Project BIRD162411/16 “Statistical learning methods for estimating the effective connectivity of human brain networks”.

We would like to acknowledge the work [18] which was for us a source of inspiration. The authors would also like to thank the anonymous reviewers for many insightful comments and suggestions which have helped us to improve the paper.

Appendix A: Statistical Linearisation of the haemodynamic response

To derive the linear model of the haemodynamic response, we first compute a population of typical responses generated by the non-linear model (4)-(8). Then, we define g_i as a linear combination of their empirical mean and of the first p principal components of their sample covariance matrix, that is $h_i = H\alpha_i$, $i = 1, ..., n$. While the coefficients $\alpha_i \in \mathbb{R}^{p+1}$ have to be estimated from the given fMRI data, the matrix H is constructed through the following steps:

1. Sample $\theta_h^{(j)}$, $j = 1, ..., N_s$ from the empirical Gaussian distributions given in Table 1 of the seminal work [25].
2. For each $\theta_h^{(j)}$ compute, with some abuse of terminology, the impulse response of the non-linear model (4)-(8), i.e. the output (say $b(k)$) when the input $x(k) = \delta(k)$ and $\delta(k)$ is the Kronecker delta function. Let $b^{(j)}$ be the corresponding output sampled at rate $1/T_R$ and truncated at length s.
3. Compute the empirical mean $\bar{b} = \frac{1}{N_s}\sum_{j=1}^{N_s} b^{(j)}$.
4. Compute the empirical covariance matrix $\bar{\Sigma}_b \in \mathbb{R}^{s\times s}$:

$$\bar{\Sigma}_b = \frac{1}{N_s}\sum_{j=1}^{N_s}(b^{(j)} - \bar{b})(b^{(j)} - \bar{b})^\top. \tag{A.30}$$

5. Compute the eigenvalue decomposition of $\bar{\Sigma}_b$, $\bar{\Sigma}_b = USU^\top$, where $S := \text{diag}(s_1, ..., s_s)$ and $U := [u_1 \ \cdots \ u_s]$.
6. Define H as $H := [\bar{b}\ u_1\ u_2\ \cdots\ u_p]$ where $p << s$.
7. Model b as $h_i = H\alpha_i$.

Exploiting the fact that the empirical covariance $\bar{\Sigma}_b$ is close to be low rank, with only a small number (p) of significant singular values $\{s_j\}_{j=1}^p$, the coefficient vectors α_i are assigned the Gaussian prior

$$p(\alpha_i) \sim \mathcal{N}(\mu_\alpha, \Sigma_\alpha) \qquad \mu_\alpha := [1\ 0\ \cdots\ 0]^\top \qquad \Sigma_\alpha := \text{diag}(\epsilon, s_1, ..., s_p) \quad i = 1, ..., n$$

so that the h_i's match the empirical statistics $\bar{b}$ and $\bar{\Sigma}_b$; ϵ is a small positive constant to guarantee the invertibility of Σ_α. Clearly, the final prior for h_i will be $h_i \sim \mathcal{N}(\bar{b}, H\Sigma_\alpha H^\top)$.

Appendix B: EM algorithm

The model parameters in (12) are estimated using the EM algorithm detailed in Algorithm 1 below. The inputs to this algorithm are the fMRI data $\{y(k)\}_{k=1}^{N}$ and the prior for the haemodynamic impulse responses h_i, that is H, μ_α and Σ_α (see the discussion in Appendix A). An initial guess for the parameters θ has also to be provided. The latter aspect has been thoroughly discussed in Section 3.3.5.

More specifically, we initialize the connectivity matrix A as $A^{(0)} = -I_n$ and set $h_i^{(0)} = \bar{b}$, $i = 1, ..., n$ (the empirical mean from the prior). An initial value for the variance σ^2 of the endogenous fluctuations $v(t)$ is chosen as follows:

1. deconvolve the fMRI time-series $\{y(k)\}_{k=1}^{N}$ with $\bar{b}$ in order to have a first estimate of the neural time-series $\{x(k)\}_{k=1}^{N}$;
2. model each $\{x_i(k)\}_{k=1}^{N}$, $i = 1, ..., n$ as an AR(3) model;
3. set σ^2 as the sample mean of the estimated noise variance of the n AR models estimated at step 2.

Finally, the variances λ_i^2, $i = 1, .., n$, of the measurement noise $e(k)$ are initialized at one tenth of the empirical variance of the corresponding BOLD time-series $\{y_i(k)\}_{k=1}^{N}$.
The hyper-parameters $\{\gamma_i^{(0)}\}_{i=1}^{n^2}$ are also assigned a starting value, according to any a-priori knowledge available on the effective connectivity network. For instance, structural connectivity can be exploited at this stage, by setting to non-zero the γ_i's corresponding to structural links and to a small quantity (e.g. $\epsilon \approx 10^{-6}$) all the others. If no a-priori knowledge is available, the same value can be assigned to all $\{\gamma_i^{(0)}\}_{i=1}^{n^2}$.
After the initialization, each iteration of Algorithm 1 consists in the application of the RTS smoother (whose routine is reported in Algorithm 2) to compute the function $\mathcal{Q}(\theta, \theta^{(l)})$, which is then maximized to update the parameter estimate $\theta^{(l+1)}$ (Step 4 of Algorithm 1). The objective function also includes the priors for A and $\{\alpha_i\}_{i=1}^{n}$ (which shape the haemodynamic responses $\{h_i\}_{i=1}^{n}$). Note that the shorthand notation $a := \text{vec}(A^\top)$ is used. The new estimates $\sigma^{(l+1)}$ and $A^{(l+1)}$ are then used at Step 5 to update the covariance matrix $Q^{(l+1)}$ and in turn the hyper-parameters $\{\gamma_i^{(l+1)}\}_{i=1}^{n^2}$ at Step 6. The details about the derivation of the update equation are provided in Appendix C .

Appendix C: Iterative Reweighted Procedure for Hyperparameters Update

Step 6 of Algorithm 1 updates the hyper-parameters $\{\gamma_i\}_{i=1}^{n^2}$ of the prior for the connectivity matrix A, i.e. $a \sim \mathcal{N}(\mathbf{0}, \text{diag}(\gamma_1, ..., \gamma_{n^2}))$, adapting the reweighted procedure proposed by [85] for linear regression models of the form

$$\mathrm{x} = \Phi a + \mathrm{w}$$

where a is to be estimated from the noisy observations x, while Φ is the regressors matrix and w is the noise vector.

Algorithm 1 Estimation of parameters θ through EM

```
Inputs: $\{y(k)\}_{k=1}^{N}$, $H$, $\mu_\alpha$, $\Sigma_\alpha$
Initialization: Initialize $\theta^{(0)}$ and $\{\gamma_i^{(0)}\}_{i=1}^{n^2}$, set $l = 0$
repeat
    Apply Algorithm 2 to get $\hat{\mathbf{x}}^s(k)$, $\mathbf{P}^s(k)$, $\mathbf{G}(k)$, $k = 1, ..., N$
    Compute $\mathcal{Q}(\theta, \theta^{(l)})$ using Eq. (27)
    $\theta^{(l+1)} = \arg\max_{\theta\in\Omega}\ \mathcal{Q}(\theta, \theta^{(l)})$
    $Q^{(l+1)} = \sigma^{(l+1)2} \int_0^{T_R} e^{A^{(l+1)}\tau} e^{A^{(l+1)^\top}\tau} d\tau$
    $\gamma_i^{(l+1)} = -(\gamma_i^{(l)})^2\ \phi_i^\top (\Phi\Gamma^{(l)}\Phi^\top + Q^{(l+1)} \otimes I_N)^{-1}\phi_i + (a_i^{(l+1)})^2 + \gamma_i^{(l)}, \quad i = 1, ..., n^2$
    $l = l + 1$
until $\|A^{(l)} - A^{(l-1)}\|_F / \|A^{(l)}\|_F$ is sufficiently small
Outputs: $\theta^{(l)}$
```

Algorithm 2 RTS Smoother

```
Inputs: $\{y(k)\}_{k=1}^{N}$; $\mathbf{A}, \mathbf{H}$ in Eq. (13), $\mathbf{Q}$ and $R$ in Eqs. (14), (15).
Forward Recursion
Initialize: $\hat{\mathbf{x}}(0) = 0$, $\mathbf{P}(0) = I_{ns}$
for $k = 1, ..., N$ do
    $\hat{\mathbf{x}}^-(k) = \mathbf{A}\hat{\mathbf{x}}(k-1)$
    $\mathbf{P}^-(k) = \mathbf{A}\mathbf{P}(k-1)\mathbf{A}^\top + \mathbf{Q}$
    $S(k) = \mathbf{H}\mathbf{P}^-(k)\mathbf{H}^\top + R$
    $K(k) = \mathbf{P}^-(k)\mathbf{H}^\top S^{-1}(k)$
    $\hat{\mathbf{x}}(k) = \hat{\mathbf{x}}^-(k) + K(k)[y(k) - \mathbf{H}\hat{\mathbf{x}}^-(k)]$
    $\mathbf{P}(k) = \mathbf{P}^-(k) - K(k)S(k)K^\top(k)$
Backward recursion
Initialize: $\hat{\mathbf{x}}^s(N) = \hat{\mathbf{x}}(N)$, $\mathbf{P}^s(N) = \mathbf{P}(N)$
for $k = N-1, ..., 0$ do
    $\hat{\mathbf{x}}^-(k+1) = \mathbf{A}\hat{\mathbf{x}}(k)$
    $\mathbf{P}^-(k+1) = \mathbf{A}\mathbf{P}(k)\mathbf{A}^\top + \mathbf{Q}$
    $\mathbf{G}(k) = \mathbf{P}(k)\mathbf{A}^\top[\mathbf{P}^-(k+1)]^{-1}$
    $\hat{\mathbf{x}}^s(k) = \hat{\mathbf{x}}(k) + \mathbf{G}(k)[\hat{\mathbf{x}}^s(k+1) - \hat{\mathbf{x}}^-(k+1)]$
    $\mathbf{P}^s(k) = \mathbf{P}(k) + \mathbf{G}(k)[\mathbf{P}^s(k+1) - \mathbf{P}^-(k+1)]\mathbf{G}^\top(k)$
Outputs: $\hat{\mathbf{x}}^s(k)$, $\mathbf{P}^s(k)$, $\mathbf{G}(k)$, $k = 1, ..., N$
```

In our setup this model is obtained after linearizing the original state update Eq. (9) which is non-linear as a function of $a = \mathrm{vec}(A^\top)$. Namely, define the matrices

$$X_+ := \begin{bmatrix} x^\top(2) \\ \vdots \\ x^\top(N) \end{bmatrix}, \qquad X := \begin{bmatrix} x^\top(1) \\ \vdots \\ x^\top(N-1) \end{bmatrix}, \qquad W := \begin{bmatrix} w^\top(1) \\ \vdots \\ w^\top(N-1) \end{bmatrix}. \tag{C.31}$$

Eq. (9) can be rewritten in the non-linear regression form

$$X_+ = X e^{A^\top T_R} + W. \tag{C.32}$$

Then, using the approximation $e^{A^\top T_R} \simeq I + A^\top T_R$, we obtain:

$$\Delta X = X A^\top T_R + W \tag{C.33}$$

where $\Delta X = X_+ - X$. Using the vectorization operator, we can rewrite (C.33) in linear regression form,

$$\mathrm{x} = \Phi a + \mathrm{w} \tag{C.34}$$

where $\mathrm{x} := \mathrm{vec}(\Delta X)$, $\Phi = [\phi_1 \ \cdots \ \phi_{n^2}] := (I \otimes X) T_R$, $a := \mathrm{vec}(A^\top)$, $\mathrm{w} := \mathrm{vec}(W)$. Therefore, we can update the hyper-parameters $\{\gamma_i\}_{i=1}^{n^2}$ as suggested in [85], see Step 6 of Algorithm 1.

Appendix D: Network Measures of Brain Connectivity

The estimated effective connectivity A can be interpreted as the weighted adjacency matrix of a directed graph between different brain regions, where each link corresponds to a directed influence of one area on another one. Specifically, the set of vertexes (or nodes) $\mathcal{V}$ coincides with the set of monitored brain areas ($|\mathcal{V}| = n$), while we say that region i is influenced by region j if the (i, j)-th entry of matrix A, say A_{ij}, is non-zero. To the purpose of computing network indexes we define the matrix $E := A - \mathrm{diag}(A)$, i.e. E coincides with A on the off-diagonal entries and has zeros on the main diagonal. We also define the binary adjacency matrix $\bar{E}$, obtained from E by setting to 1 its non-zero entries. We evaluate the estimated graph in terms of three types of metrics, which quantify the degree of centrality of each node within the network (measures of *centrality*), as well as the presence of clusters (measures of *segregation*) and the ease with which brain regions communicate (measures of *integration*).

The most common centrality measure is the so-called node weighted *degree*: since we deal with a directed graph, we can distinguish between the weighted *in-degree* d_i^{in} and the weighted *out-degree* d_i^{out} which sum the weights of the links coming in and out from a certain node, respectively:

$$d_i^{in} = \sum_{j \in \mathcal{V}} |E_{ij}|, \qquad d_i^{out} = \sum_{j \in \mathcal{V}} |E_{ji}|. \tag{D.35}$$

These quantities are also known as *in-strength* (d_i^{in}) and *out-strength* (d_i^{out}). As a second centrality measure we consider the weighted *betweenness centrality*, i.e. the fraction of shortest paths in the network which pass through a given node. Namely, for vertex i it is defined as

$$b_i = \frac{1}{(n-1)(n-2)} \sum_{\substack{h,j \in \mathcal{V} \\ h \neq j, h \neq i, j \neq i}} \frac{\rho_{hj}(i)}{\rho_{hj}} \tag{D.36}$$

where ρ_{hj} is the number of shortest paths between nodes h and j, while $\rho_{hj}(i)$ is the number of shortest paths between h and j which pass through i. The shortest path length between vertexes i and j is defined as

$$l_{ij} = \sum_{E_{uv} \in \mathcal{P}_{i \to j}} \tilde{E}_{uv}, \qquad \tilde{E}_{uv} = \frac{1}{E_{uv}} \tag{D.37}$$

where $\mathcal{P}_{i \to j}$ denotes the directed shortest path from i to j. According to Eq. (D.37), stronger connections are interpreted as shorter distances.

Other centrality metrics are the *within-module degree z-score* and the *participation coefficient*, which are based on a preceding partition of the network into a set of non-overlapping modules (or clusters) M. The weighted within-module in-degree z-score of node i is defined as

$$z_i^{in} = \frac{d_i^{in}(m_i) - \mu_{d^{in}}(m_i)}{\sigma_{d^{in}}(m_i)} \tag{D.38}$$

where m_i is the module containing node i and $d_i^{in}(m_i)$ is the weighted within-module in-degree of node i, i.e. the weighted sum of links entering i from vertexes in module m_i. $\mu_{d^{in}}(m_i)$ and $\sigma_{d^{in}}(m_i)$ are respectively the mean and the standard deviation of the within-module m_i weighted in-degree distribution. The weighted within-module out-degree z-score is analogously defined, replacing $d_i^{in}(m_i), \mu_{d^{in}}(m_i), \sigma_{d^{in}}(m_i)$ with $d_i^{out}(m_i), \mu_{d^{out}}(m_i), \sigma_{d^{out}}(m_i)$.

The weighted in-degree participation coefficient is given by

$$pc_i^{in} = 1 - \sum_{m \in M} \left(\frac{d_i^{in}(m)}{d_i^{in}} \right)^2 \tag{D.39}$$

where $d_i^{in}(m)$ is the weighted sum of the links entering node i from all vertexes in module m. The definition of the out-degree participation coefficient follows the same principle. Combined together, the within-module degree z-score and the participation coefficient provide information about the role of a certain node in facilitating network segregation or integration. Specifically, a node with high within-module degree z-score and low participation coefficient is a so-called provincial hub, that is, it favors segregation. On the other hand, a vertex with low within-module degree z-score and high participation coefficient is a connector hub, meaning that it encourages integration. In addition to the combined evaluation of these two metrics, we consider a further measure of segregation, known as weighted *clustering coefficient*, i.e. the weighted fraction of triangles around a node. Specifically, it is defined as

$$cc_i = \frac{\left[\left(E^{[\frac{1}{3}]} + (E^{\top})^{[\frac{1}{3}]}\right)^3\right]_{ii}}{2\left[\bar{d}_i^{tot}(\bar{d}_i^{tot} - 1) - 2\left(\bar{E}_{ii}\right)^2\right]} \tag{D.40}$$

where the notation $E^{[\alpha]}$ denotes the element wise exponentiation of matrix E, i.e. $\left[E^{[\alpha]}\right]_{ij} = E_{ij}^{\alpha}$, and $\bar{d}_i^{tot} = \bar{d}_i^{in} + \bar{d}_i^{out}$ and $\bar{d}_i^{in} = \sum_{j\in\mathcal{V}} \bar{E}_{ij}$, $\bar{d}_i^{out} = \sum_{j\in\mathcal{V}} \bar{E}_{ji}$.

[1] S. Badillo, T. Vincent, and P. Ciuciu. Group-level impacts of within-and between-subject hemodynamic variability in fMRI. *Neuroimage*, 82:433–448, 2013.

[2] L. Barnett and A. K. Seth. The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Neuroscience Methods*, 223:50 – 68, 2014.

[3] N. Z. Bielczyk, S. Uithol, T. v. Mourik, P. Anderson, J. C. Glennon, and J. K. Buitelaar. Disentangling causal webs in the brain using functional magnetic resonance imaging: A review of current approaches. *Network Neuroscience*, pages 1–63, 2018.

[4] S. Bollmann, L. Kasper, S. J. Vannesjo, A. O. Diaconescu, B. E. Dietrich, S. Gross, K. E. Stephan, and K. P. Pruessmann. Analysis and correction of field fluctuations in fMRI data using field monitoring. *Neuroimage*, 154:92–105, 2017.

[5] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186, 2009.

[6] R. B. Buxton, E. C. Wong, and L. R. Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic resonance in medicine*, 39(6):855–864, 1998.

[7] A. Carter, S. Astafiev, L. C.E., L. Connor, J. Rengachary, M. Strube, D. Pope, G. Shulman, and M. Corbetta. Resting interhemispheric functional magnetic resonance imaging connectivity predicts performance after stroke. *Ann. Neurol.*, 67(3):365–375, 2010.

[8] A. Chiuso and G. Pillonetto. A bayesian approach to sparse dynamic network identification. *Automatica*, 48(8):1553 – 1565, 2012.

[9] D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[10] R. Cubo, A. Medvedev, and M. Astrom. Model-based optimization of individualized deep brain stimulation therapy. *IEEE Design Test*, 33(4):74–81, Aug 2016.

[11] J. Daunizeau, O. David, and K. E. Stephan. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *Neuroimage*, 58(2):312–322, 2011.

[12] O. David, S. J. Kiebel, L. M. Harrison, J. Mattout, J. M. Kilner, and K. J. Friston. Dynamic causal modeling of evoked responses in eeg and meg. *Neuroimage*, 30(4):1255–1272, 2006.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[14] X. Di and B. B. Biswal. Identifying the default mode network structure using dynamic causal modeling on resting-state functional magnetic resonance imaging. *Neuroimage*, 86:53–59, 2014.

[15] J. Dubois, H. Oya, J. M. Tyszka, M. Howard, F. Eberhardt, and R. Adolphs. Causal mapping of emotion networks in the human brain: Framework and preliminary findings. *bioRxiv*, page 214486, 2017.

[16] J. H. Duyn. The future of ultra-high field mri and fMRI for study of the human brain. *Neuroimage*, 62(2):1241–1248, 2012.

[17] A. Fornito, A. Zalesky, and M. Breakspear. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159, 2015.

[18] S. Frässle, E. I. Lomakina, L. Kasper, Z. M. Manjaly, A. Leff, K. P. Pruessmann, J. M. Buhmann, and K. E. Stephan. A generative model of whole-brain effective connectivity. *Neuroimage*, 2018.

[19] S. Frässle, E. I. Lomakina, A. Razi, K. J. Friston, J. M. Buhmann, and K. E. Stephan. Regression DCM for fMRI. *Neuroimage*, 2017.

[20] F. Freyer, K. Aquino, P. A. Robinson, P. Ritter, and M. Breakspear. Bistability and non-gaussian fluctuations in spontaneous cortical activity. *Journal of Neuroscience*, 29(26):8512–8524, 2009.

[21] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny. Variational free energy and the laplace approximation. *Neuroimage*, 34(1):220–234, 2007.

[22] K. Friston and W. Penny. Post hoc bayesian model selection. *Neuroimage*, 56(4):2089–2099, 2011.

[23] K. Friston, K. Stephan, B. Li, and J. Daunizeau. Generalised filtering. *Mathematical Problems in Engineering*, 2010, 2010.

[24] K. J. Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

[25] K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[26] K. J. Friston, J. Kahan, B. Biswal, and A. Razi. A DCM for resting state fMRI. *Neuroimage*, 94:396–407, 2014.

[27] K. J. Friston, V. Litvak, A. Oswal, A. Razi, K. E. Stephan, B. C. van Wijk, G. Ziegler, and P. Zeidman. Bayesian model reduction and empirical bayes for group (dcm) studies. *Neuroimage*, 128:413–431, 2016.

[28] K. J. Friston, A. Mechelli, R. Turner, and C. J. Price. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *Neuroimage*, 12(4):466–477, 2000.

[29] K. J. Friston, N. Trujillo-Barreto, and J. Daunizeau. DEM: a variational treatment of dynamic systems. *Neuroimage*, 41(3):849–885, 2008.

[30] H. Garnier and L. Wang. *Identification of Continuous-time Models from Sampled Data*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[31] M. Gilson, G. Deco, K. Friston, P. Hagmann, D. Mantini, V. Betti, G. L. Romani, and M. Corbetta. Effective connectivity inferred from fMRI transition dynamics during movie viewing points to a balanced reconfiguration of cortical interactions. *Neuroimage*, 2017.

[32] M. Gilson, R. Moreno-Bote, A. Ponce-Alvarez, P. Ritter, and G. Deco. Estimation of directed effective connectivity from fMRI functional connectivity hints at asymmetries of cortical connectome. *PLOS computational biology*, 12(3):e1004762, 2016.

[33] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258, 2003.

[34] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns. Mapping the structural core of human cerebral cortex. *PLOS biology*, 6(7):e159, 2008.

[35] D. A. Handwerker, J. M. Ollinger, and M. D'Esposito. Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21(4):1639–1651, 2004.

[36] J. Kahan, M. Urner, R. Moran, G. Flandin, A. Marreiros, L. Mancini, M. White, J. Thornton, T. Yousry, L. Zrinzo, M. Hariz, P. Limousin, K. Friston, and T. Foltynie. Resting state functional MRI in Parkinson's disease: the impact of deep brain stimulation on "effective" connectivity. *Brain*, 137(4):1130–1144, 2014.

[37] S. J. Kiebel, M. I. Garrido, R. Moran, C.-C. Chen, and K. J. Friston. Dynamic causal modeling for eeg and meg. *Human brain mapping*, 30(6):1866–1876, 2009.

[38] M. M. Klein, R. Treister, T. Raij, A. Pascual-Leone, L. Park, T. Nurmikko, F. Lenz, J.-P. Lefaucheur, M. Lang, M. Hallett, M. Fox, M. Cudkowicz, A. Costello, D. B. Carr, S. S. Ayache, and A. L. Oaklander. Transcranial magnetic stimulation of the brain: guidelines for pain treatment research. *Pain*, 156(9):1601–1614, 09 2015.

[39] B. Li, J. Daunizeau, K. E. Stephan, W. Penny, D. Hu, and K. Friston. Generalised filtering and stochastic DCM for fMRI. *Neuroimage*, 58(2):442–457, 2011.

[40] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabasi. Controllability of complex networks. *Nature*, 473:167–73, 05 2011.

[41] A. M. Lozano and N. Lipsman. Probing and regulating dysfunctional circuits using Deep Brain Stimulation. *Neuron*, 77(3):406–424, 2013.

[42] V. Pallares, A. Insabato, A. Sanjuan, S. Kuehn, D. Mantini, G. Deco, and M. Gilson. Extracting orthogonal subject-and condition-specific signatures from fMRI data using whole-brain effective connectivity. *Neuroimage*, 178:238–254, 2018.

[43] H.-J. Park and K. Friston. Structural and functional brain networks: from connections to cognition. *Science*, 342(6158):1238411, 2013.

[44] A. Ponce-Alvarez, G. Deco, P. Hagmann, G. L. Romani, D. Mantini, and M. Corbetta. Resting-state temporal synchronization networks emerge from connectivity topology and heterogeneity. *PLOS computational biology*, 11(2):e1004100, 2015.

[45] G. Prando, M. Zorzi, A. Bertoldo, and A. Chiuso. Estimating effective connectivity in linear brain network models. In *56th IEEE Conference on Decision and Control*, 2017.

[46] G. Prando, M. Zorzi, A. Bertoldo, and A. Chiuso. The role of noise modeling in the estimation of resting-state brain effective connectivity. In *Proceedings of SYSID 2018*, July 2018.

[47] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3(2):121–129, 2017.

[48] J. Ramsey, J. Zhang, and P. Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 06 2006.

[49] J. D. Ramsey. Scaling up greedy causal search for continuous variables. *arXiv preprint arXiv:1507.07749*, 2015.

[50] J. D. Ramsey, S. J. Hanson, C. Hanson, Y. O. Halchenko, R. A. Poldrack, and C. Glymour. Six problems for causal inference from fMRI. *Neuroimage*, 49(2):1545–1558, 2010.

[51] H. E. Rauch, C. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

[52] A. Razi and K. J. Friston. The connected brain: causality, models, and intrinsic dynamics. *IEEE Signal Processing Magazine*, 33(3):14–35, 2016.

[53] A. Razi, J. Kahan, G. Rees, and K. J. Friston. Construct validation of a DCM for resting state fMRI. *Neuroimage*, 106:1–14, 2015.

[54] A. Razi, M. L. Seghier, Y. Zhou, P. McColgan, P. Zeidman, H.-J. Park, O. Sporns, G. Rees, and K. J. Friston. Large-scale DCMs for resting-state fMRI. *Network Neuroscience*, 1(3):222–241, 2017.

[55] Z. Y. M. P. Z. P. P. H. S. O. R. G. F. K. Razi A, Seghier ML. Large-scale DCMs for resting-state fMRI. *Network Neuroscience*, 1(3):222–241, 2017.

[56] T. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 454–461. Morgan Kaufmann Publishers Inc., 1996.

[57] T. S. Richardson, P. Spirtes, et al. *Automated discovery of linear feedback models.* Carnegie Mellon, 1996.

[58] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[59] S. Ryali, T. Chen, K. Supekar, T. Tu, J. Kochalka, W. Cai, and V. Menon. Multivariate dynamical systems-based estimation of causal brain interactions in fMRI: Group-level validation using benchmark data, neurophysiological models and human connectome project data. *Journal of neuroscience methods*, 268:142–153, 2016.

[60] S. Ryali, K. Supekar, T. Chen, and V. Menon. Multivariate dynamical systems models for estimating causal interactions in fMRI. *Neuroimage*, 54(2):807–823, 2011.

[61] R. Sanchez-Romero, J. D. Ramsey, K. Zhang, M. R. Glymour, B. Huang, and C. Glymour. Causal discovery of feedback networks with functional magnetic resonance imaging. *bioRxiv*, page 245936, 2018.

[62] S. Särkkä. *Bayesian filtering and smoothing*, volume 3. Cambridge University Press, 2013.

[63] M. L. Seghier and K. J. Friston. Network discovery with large DCMs. *Neuroimage*, 68:181–191, 2013.

[64] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

[65] C.-W. Shin and S. Kim. Self-organized criticality and scale-free properties in emergent functional neural networks. *Physical Review E*, 74(4):045101, 2006.

[66] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis*, 3(4):253–264, 1982.

[67] J. S. Siegel, L. E. Ramsey, A. Z. Snyder, N. V. Metcalf, R. V. Chacko, K. Weinberger, A. Baldassarre, C. D. Hacker, G. L. Shulman, and M. Corbetta. Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke. *Proceedings of the National Academy of Sciences*, 113(30):E4367–E4376, 2016.

[68] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. Network modelling methods for fMRI. *Neuroimage*, 54(2):875–891, 2011.

[69] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[70] O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Hilgetag. Organization, development and function of complex brain networks. *Trends in cognitive sciences*, 8(9):418–425, 2004.

[71] O. Sporns, G. Tononi, and R. Kötter. The human connectome: a structural description of the human brain. *PLOS computational biology*, 1(4):e42, 2005.

[72] C. J. Stam and E. A. De Bruin. Scale-free dynamics of global functional connectivity in the human brain. *Human brain mapping*, 22(2):97–109, 2004.

[73] C. J. Stam and J. C. Reijneveld. Graph theoretical analysis of complex networks in the brain. *Nonlinear biomedical physics*, 1(1):3, 2007.

[74] K. E. Stephan, W. D. Penny, J. Daunizeau, R. J. Moran, and K. J. Friston. Bayesian model selection for group studies. *Neuroimage*, 46(4):1004–1017, 2009.

[75] K. E. Stephan, W. D. Penny, R. J. Moran, H. E. den Ouden, J. Daunizeau, and K. J. Friston. Ten simple rules for dynamic causal modeling. *Neuroimage*, 49(4):3099–3109, 2010.

[76] K. E. Stephan, N. Weiskopf, P. M. Drysdale, P. A. Robinson, and K. J. Friston. Comparing hemodynamic models with DCM. *Neuroimage*, 38(3):387–401, 2007.

[77] R. Stirnberg, W. Huijbers, D. Brenner, B. A. Poser, M. Breteler, and T. Stöcker. Rapid whole-brain resting-state fMRI at 3 t: Efficiency-optimized three-dimensional epi versus repetition time-matched simultaneous-multi-slice epi. *Neuroimage*, 163:81–92, 2017.

[78] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[79] C. Tu, R. P. Rocha, M. Corbetta, S. Zampieri, M. Zorzi, and S. Suweis. Warnings and caveats in brain controllability. *Neuroimage*, 176:83 – 91, 2018.

[80] V. Ushakov, M. G. Sharaev, S. I. Kartashov, V. V. Zavyalova, V. M. Verkhlyutov, and B. M. Velichkovsky. Dynamic causal modeling of hippocampal links within the human default mode network: Lateralization and computational stability of effective connections. *Frontiers in human neuroscience*, 10:528, 2016.

[81] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston. Effective connectivity: influence, causality and biophysical modeling. *Neuroimage*, 58(2):339–361, 2011.

[82] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston. Effective connectivity: Influence, causality and biophysical modeling. *NeuroImage*, 58(2):339 – 361, 2011.

[83] M. P. Van Den Heuvel and H. E. H. Pol. Exploring the brain network: a review on resting-state fMRI functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

[84] H. H. Weerts, P. M. V. den Hof, and A. G. Dankers. Identifiability of linear dynamic networks. *Automatica*, 89:247 – 258, 2018.

[85] D. Wipf and S. Nagarajan. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):317–329, 2010.

[86] J. Xu, S. Moeller, E. J. Auerbach, J. Strupp, S. M. Smith, D. A. Feinberg, E. Yacoub, and K. Uğurbil. Evaluation of slice accelerations using multiband echo planar imaging at 3 t. *Neuroimage*, 83:991–1001, 2013.

[87] Y. Yuan, G.-B. Stan, S. Warnick, and J. Goncalves. Robust dynamical network structure reconstruction. *Automatica*, 47(6):1230 – 1235, 2011. Special Issue on Systems Biology.

[88] Z. Yue, J. Thunberg, L. Ljung, and J. M. Gonçalves. Identification of sparse continuous-time linear systems with low sampling rate: Exploring matrix logarithms. *CoRR*, abs/1605.08590, 2016.

[89] C. Zhou, L. Zemanová, G. Zamora, C. C. Hilgetag, and J. Kurths. Hierarchical organization unveiled by functional connectivity in complex brain networks. *Physical review letters*, 97(23):238103, 2006.

[90] M. Zorzi and R. Sepulchre. AR identification of latent-variable graphical models. *IEEE Transactions on Automatic Control*, 61(9):2327–2340, 2016.