Journal of Neural Engineering

PAPER

Large-scale dynamic modeling of task-fMRI signals via subspace system identification

To cite this article: Cassiano O Becker *et al* 2018 *J. Neural Eng.* **15** 066016

View the article online for updates and enhancements.

IOP ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

This content was downloaded from IP address 158.130.195.27 on 27/02/2019 at 00:48

IOP Publishing Journal of Neural Engineering
J. Neural Eng. **15** (2018) 066016 (17pp) https://doi.org/10.1088/1741-2552/aad8c7

Large-scale dynamic modeling of task-fMRI signals via subspace system identification

Cassiano O Becker[1,5], **Danielle S Bassett**[1,2,3,4] **and Victor M Preciado**[1]

[1] Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, United States of America
[2] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, United States of America
[3] Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States of America
[4] Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, United States of America

E-mail: cassiano@seas.upenn.edu

Received 9 June 2018
Accepted for publication 8 August 2018
Published 28 September 2018

Abstract

Objective. We analyze task-based fMRI time series to produce large-scale dynamical models that are capable of approximating the observed signal with good accuracy. *Approach.* We extend subspace system identification methods for deterministic and stochastic state-space models with external inputs. The dynamic behavior of the generated models is characterized using control-theoretic analysis tools. To validate their effectiveness, we perform a probabilistic inversion of the identified input–output relationships via joint state-input maximum likelihood estimation. Our experimental setup explores a large dataset generated using state-of-the-art acquisition and pre-processing methods from the Human Connectome Project. *Main results.* We analyze both anatomically parcellated and spatially dense time series, and propose an efficient algorithm to address the high-dimensional optimization problem resulting from the latter. Our results enable the quantification of input–output transfer functions between each task condition and each region of the cortex, as exemplified by a motor task. Further, the identified models produce impulse response functions between task conditions and cortical regions that are compatible with typical hemodynamic response functions. We then extend subspace methods to account for multi-subject experimental configurations, identifying models that capture common dynamical characteristics across subjects. Finally, we show that system inversion via maximum-likelihood allows the time-of-occurrence of the task stimuli to be estimated from the observed outputs. *Significance.* The ability to produce dynamical input–output models might have an impact in the expanding field of neurofeedback. In particular, the models we produce allow the partial quantification of the effect of external task-related inputs on the metabolic response of the brain, conditioned on its current state. Such a notion provides a basis for leveraging control-theoretic approaches to neuromodulation and self-regulation in therapeutic applications.

Keywords: dynamical systems, task fmri, control theory, subspace system identification

Supplementary material for this article is available online

(Some figures may appear in colour only in the online journal)

[5] Author to whom any correspondence should be addressed.

 © 2018 IOP Publishing Ltd Printed in the UK

1. Introduction

Dynamical models provide a principled approach to describing the flow of energy and information in a system of interest. Generally, such models account for the effect that past history (summarized in the system's internal state) and external sources of energy and information (represented by its inputs) produce on the system's observed variables (outputs) [1, 2]. In the context of a person participating in a functional magnetic resonance imaging experiment, the externally observed variables are related to changes in blood flow resulting from neural activity, as measured by variations in the local concentration levels of oxyhemoglobin and deoxyhemoglobin on the cortex [3]. Correspondingly, the observed inputs to these models can be associated with task-related stimuli presented to the individual throughout the execution of the experiment, as well as to physiological measurements such as heart rate and respiration [4].

From a biological point of view, the relationship between such inputs and outputs is generally described as a combination of neural processes occurring across two levels. The inner (microscopic) level consists of fine-grained interactions between neurons involved in the exchange of information encoded through action potentials. The outer (mesoscopic) level, measured by the fMRI signal, consists of the aggregate effect of metabolic requirements from the inner level in terms of glucose metabolism and oxygen content in cerebral blood flow [5]. This relationship has been quantified as the localized time response of the *blood-oxygen-level-dependent* signal (BOLD) with respect to a specific stimulus onset [6], and is referred to as the BOLD-contrast *hemodynamic response function* (HRF) [7]. Notably, in comparison with the shorter time scale associated with inner-level processes, the time scale over which most of the variation of the HRF takes place is of the order of seconds. The onset of the HRF typically takes place approximately 2 s after the presentation of the stimulus, with the remaining part of the response lasting approximately 10–12 s [7]. In addition, because of an empirically observed effect of additivity in BOLD response with respect to the stimuli, this behavior can be broadly approximated by a linear input–output dynamic relation [8–11] (for a discussion of possible limitations see [12–14]). Based on these characteristics, this study seeks to estimate data-driven models capable of approximating the measured fMRI BOLD signal as a function of task-related and physiological inputs. Motivated by their analytical tractability, we consider the class of discrete-time, time-invariant linear systems, here expressed in a relatively high-dimensional output space defined by the spatial resolution of the fMRI signal.

It is known that any given class of dynamical models under consideration will impose constraints on the methods that can be applied for their estimation. In the context of fMRI time series, two commonly studied classes of models are those based on dynamic causal modeling (DCM), and those based on linear autoregressive processes (AR). The DCM approach for fMRI [15] considers a bilinear neural activity model, coupled with a nonlinear BOLD observation function. Because of the nonlinearities involved in its formulation, the associated model estimation procedure often involves an expectation maximization algorithm, which tends to limit its applicability to moderately-sized models. On the other hand, studies applying AR models to fMRI assume that the signal is generated by a recursive linear transformation of endogenous noise sources, and therefore do not account for the external input-to-output relationships that we address in this paper [16–18].

For the class of models that we consider, i.e. discrete-time time-invariant linear state-space models with external inputs, effective and reliable methods for parameter estimation have been developed [19, 20]. In particular, we take inspiration from subspace methods, which have the desirable characteristics of assured consistency and convergence to globally optimal estimates with respect to a quadratic error criteria [21, 22]. Although widely employed in many fields of engineering, to the authors' knowledge, subspace methods have received restricted attention in fMRI modeling. The few existing studies in this respect are lacking in important aspects, such as depth of analysis and experimental validation on real datasets. Earlier work [23] (followed by [24]) proposed the use of subspace methods for fMRI data, but applied them to simulated-only time series, and over a very low-dimensional space of observations (less than four regions-of-interest). The work in [25] considered a restricted configuration of subspace methods, i.e. for models with no external inputs. The goal of that study was to use resting-state (non-task) data to estimate connectivity (as opposed to dynamical models), and its scope was also restricted to a dataset with small numbers of brain regions and participants.

In this work, we analyze task-based functional magnetic resonance imaging time series using a dynamical systems approach. Based on information about the time-of-occurrence of task stimuli presentation and using the observed signal intensity on the cortical surface, we apply subspace system identification methods to produce input–output dynamical models that are capable of approximating the BOLD time series with good accuracy, as measured by the correlation between the original fMRI signal and the one generated by the model. As our experimental setup, we explore a dataset extracted from the Human Connectome Project [26], generated using state-of-the-art fMRI acquisition and pre-processing methods. We consider both anatomically parcellated and high spatial resolution versions of the fMRI time series, and propose an efficient algorithm to address the large-scale optimization problem resulting from the latter. Our analysis enables the quantification of input–output transfer function norms between each task input channel and each region of the cortex, as we exemplify by a motor task experiment. In addition, the identified models produce impulse response functions between task conditions (i.e. specific task commands issued to the subject) and cortical regions that are compatible with typical hemodynamic response functions [27]. Furthermore, we extend existing subspace system identification methods to account for multi-subject experimental configurations, identifying models that capture common dynamical characteristics across subjects, and apply them to a cohort of 100 subjects. Finally, to verify that the method consistently captures the underlying input–output relationships, we perform

a probabilistic inversion of the identified models based on a first-principles maximum likelihood derivation, thereby allowing the time-of-occurrence of the task-related inputs to be estimated from the observed outputs.

The ability to produce dynamical input–output models could have an impact in the expanding field of neurofeedback [28]. In particular, the models we produce allow the partial quantification of the effect of external task-related inputs on the metabolic response of the brain, conditioned on its current state. Such a notion provides a basis for leveraging control-theoretic approaches to neuromodulation and self-regulation in therapeutic applications, as those described in [29, 30].

The remainder of the paper is structured as follows. In section 2, we describe the model and methods proposed for system identification. In section 3, we report details of the dataset and parameters considered in the experiments that we performed. In section 4, we illustrate and discuss the main results obtained by our methods, and in section 5, we provide conclusions and suggestions for future research.

1.1. Notation

We denote by $x \in \mathbb{R}^n$ a column vector, and by $[x]_i$ its ith entry. For a matrix $X \in \mathbb{R}^{m\times n}$, $[X]_{ij}$ denotes the entry in its ith row and jth column. Also, $[X]_{i:j,:}$ indicates the sub-matrix obtained from keeping the entries from the ith to jth rows and all of its columns. The transpose of a matrix is written as X^T, i.e. $[X^\mathsf{T}]_{i,j} := [X]_{j,i}$. The Moore–Penrose pseudo-inverse of a matrix A is denoted by $A^\dagger$.

The $n \times n$ identity matrix is denoted by I_n. The vectorization operator $\mathrm{vec}(X) : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ vertically concatenates the columns of X onto a vector $x \in \mathbb{R}^{mn}$, with $\mathrm{vec}^{-1}_{m,n}(X) : \mathbb{R}^{mn} \to \mathbb{R}^{m\times n}$ denoting its inverse operator. We denote by $\mathbb{S}^n_{++}$ (resp. $\mathbb{S}^n_{+}$) the set of $n \times n$ symmetric positive definite (resp. semi-definite) matrices. The Frobenius norm of a matrix is denoted by $\|X\|_\mathcal{F} := (\sum_{i=1}^m \sum_{j=1}^n [X]_{ij}^2)^{\frac{1}{2}} = (\mathrm{Tr} X^\mathsf{T} X)^{\frac{1}{2}}$. The matrix Mahalanobis distance of X with respect to a matrix $\Psi \in \mathbb{S}_{++}$ is defined as $\|X\|_\Psi := (\mathrm{Tr} X^\mathsf{T} \Psi X)^{\frac{1}{2}}$. Further, we denote by $\langle x, y\rangle$ the (Pearson) sample correlation between two n-dimensional vectors, defined as $\langle x, y\rangle := \frac{1}{n}\sum_{k=0}^{n-1}[x(k) - \frac{1}{n}\sum_{k=0}^{n-1} x(k)][y(k) - \frac{1}{n}\sum_{k=0}^{n-1} y(k)]$.

The density function (pdf) of a random variable (r.v.) X with support on $\mathcal{X}$ is denoted by $p(x) \equiv f_X(x) := \frac{d}{dx}F_X(x)$, where $F_X(x) := P\{X \leqslant x\}$. The conditional pdf of a r.v. X given Y is denoted $p(x|y)$. A r.v. $x \in \mathbb{R}^n$ following a multivariate normal distribution with mean μ and covariance Σ is denoted $x \sim \mathcal{N}(\mu, \Sigma)$, having pdf $p(x|\mu, \Sigma) = (2\pi)^{-\frac{n}{2}}(\det \Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|x - \mu\|^2_{\Sigma^{-1}}\right)$.

2. Model and methods

2.1. System model

As our approximating model (see figure 1), we consider a discrete-time time-invariant linear system described by a state-space representation, which evolves according to

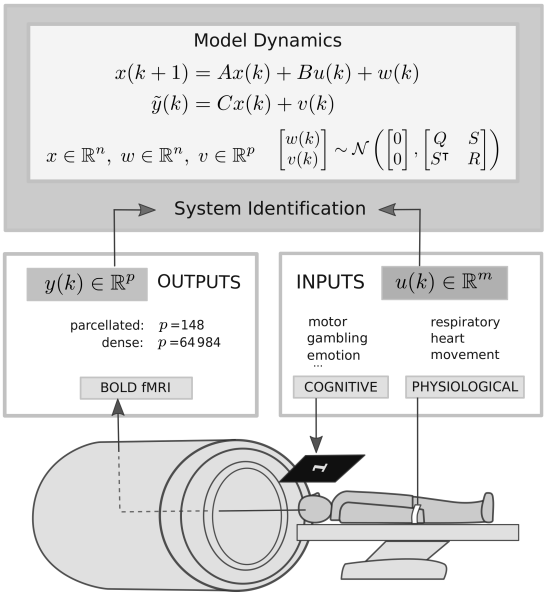

Figure 1. The system identification approach. The *cognitive* (i.e. task-related) and *physiological* inputs $u(k) \in \mathbb{R}^m$, along with the outputs $y(k) \in \mathbb{R}^p$, are presented to the subspace system identification algorithm, which produces estimates for the system matrices A, B and C (and noise matrices Q, R, and S in the stochastic case). In the experiments conducted, we have $p = 148$ for the parcellated time series, and $p = 64984$ for the spatially dense time series.

$$\begin{cases} x(k+1) &= Ax(k) + Bu(k) + w(k), \\ y(k) &= Cx(k) + v(k), \text{ for } k = 0, 1, \dots . \end{cases} \quad (1)$$

Here, the vector of outputs $y(k) \in \mathbb{R}^p$ corresponds to the observed BOLD signal intensities associated with the fMRI measurements at different regions of the brain. The dimensionality parameter p corresponds to the number of regions in the partition (parcellation) of the cortical surface, or to the surface mesh resolution, in the case of spatially dense time series (see description in section 3). The input $u(k) \in \mathbb{R}^m$ corresponds to the physiological and task-related signals observed during the experiment, whose encoding will be specified in section 2.2. In addition, the internal state variable $x(k) \in \mathbb{R}^n$ summarizes the system's past history with respect to its effect on future outputs. The matrices associated with the deterministic state-space representation (A, B, C) are referred to, respectively, as the *state transition* matrix $A \in \mathbb{R}^{n\times n}$, the *input* matrix $B \in \mathbb{R}^{n\times m}$, and the *output* matrix $C \in \mathbb{R}^{p\times n}$. They implement, in that order, the mean linear relationships between the recurrent effect of the state on itself, the effect of external inputs on the state, and the effect of the state on the observed outputs. Furthermore, such matrices correspond to the linearization [2] of a possibly nonlinear system in the close vicinity of a fixed operating point $x(k) = \bar{x}$. The variable $v(k) \in \mathbb{R}^p$ is referred to as additive *observation noise*, and is commonly associated with uncertainties in the measurement of the outputs. Correspondingly, $w(k) \in \mathbb{R}^n$ is referred to as

process noise and might, for example, account for the effect of unobserved inputs to the system. These are assumed to be zero-mean random variables jointly distributed according to

$$\begin{bmatrix} w(k) \\ v(k) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q & S \\ S^{\mathsf{T}} & R \end{bmatrix} \right), \tag{2}$$

where $Q \in \mathbb{S}^n_{++}$, $R \in \mathbb{S}^p_{++}$ and $S \in \mathbb{R}^{n\times p}$ are the matrices associated with the stochastic component of the linear system representation. The parameters of this approximating model will be estimated by subspace identification methods, subject to a model complexity constraint given by the state dimension n (a parameter of choice). In this respect, we seek estimates that, provided with the original inputs $u(k)$ and initial state estimate $\hat{x}(0)$, are able to produce output estimates $\hat{y}(k)$ that approximate the originally measured outputs $y(k)$ with good accuracy, as evaluated by their mutual correlation $\langle \hat{y}, y \rangle$.

2.2. Input encoding

In the context of a task-based fMRI experiment, we consider two classes of exogenous signals to the model, which we refer to as *cognitive* and *physiological*. Cognitive (i.e. task-related) inputs are designed to trigger different aspects of the brain's neural response, and are therefore specific to the type of experiment under investigation. Each task (e.g. MOTOR task) defines a repertoire of stimuli issued to the subject participating in the experiment, with each stimulus type being referred to as a *task condition*. Concretely, task conditions are implemented as specific visual or auditory commands issued to the subject at different times during the experiment. For example, one task condition in the MOTOR experiment is 'tap right finger' (RH), which prompts the subject to act accordingly, after the presentation of a visual cue on a screen. To encode these inputs, considering task conditions $j = 1, \ldots, m_c$ (identified with corresponding input channels $[u_c(k)]_j$), we assign $[u_c(k)]_j = 1$ if task condition j occurs on time k, and set $[u_c(k)]_j = 0$ otherwise (see figure 2). In contrast, physiological signals (e.g. heart cycles, respiration, and head movement) are usually treated as non-neural disturbances affecting the measured output signal. In this case, it is common practice to numerically remove the effect of the physiological signals from the output measurement by means of a regression procedure [4] (as we specify in section 3.2).

2.3. Identification via subspace methods

To formalize our system identification problem, we consider a batch of L input and output observations $\{u(k), y(k)\}_{k=0}^{L-1}$ generated by an *unknown* discrete-time, time-invariant linear system described by (1), i.e. our approximating model. We wish to recover estimates $(\hat{A}_T, \hat{B}_T, \hat{C}_T)$ of the deterministic system description, estimates $(\hat{Q}_T, \hat{R}_T, \hat{S}_T)$ of the noise matrices, and an estimate $\hat{x}_T(0)$ of the initial condition. Here,

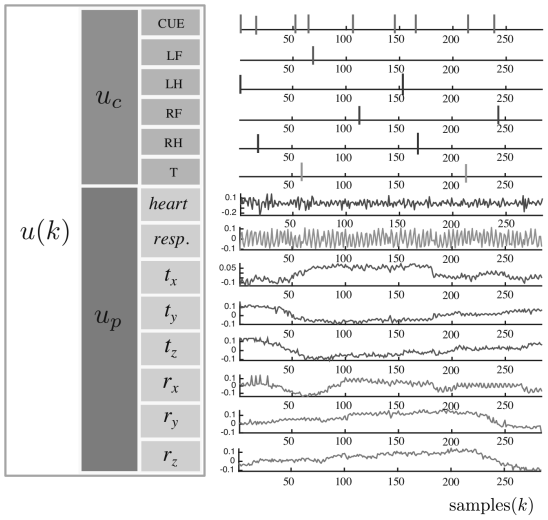

Figure 2. Input encoding of task and physiological signals for the MOTOR task. The input vector is partitioned into two blocks: the block of cognitive (i.e. task-related) input channels (u_c) and the block of physiological input channels (u_p). For u_c, a set of six task conditions is considered: CUE (a visual cue preceding the occurrence of other task conditions), LF (squeeze left toe), LH (tap left fingers), RF (squeeze right toe), RH (tap right finger), and T (move tongue). As the physiological inputs u_p, a set of eight input channels is considered: *heart* (heart signal), *resp* (breathing signal), head translation t_x, t_y, t_z (in the three spatial axes), and head rotation r_x, r_y, and r_z (in the three spatial axes).

the subscript T denotes the fact that estimates are obtained up to an invertible linear transformation of the state representation, represented by a matrix $T \in \mathbb{R}^{n\times n}$, i.e. $\hat{x}_T(k) = Tx(k)$. This transformation accounts for an invariance effect inherent to the problem, since any such T can be applied to a given state representation and still preserve the same input–output relationship [31].

To estimate system parameters, we apply and extend algorithms derived from the class of subspace identification methods [21]. Such methods rely on the construction, using the observed input and output data, of Hankel- and Toeplitz-structured matrices to establish a linear matrix equation between the data and the system parameters. The key characteristic enabling a solution to this problem lies in the notion that such matrices exhibit a specific low-rank structure that can be leveraged via a singular value decomposition, as will be seen shortly. For the sake of simplicity in our exposition, we will assume $v(k) \equiv 0$ and $w(k) \equiv 0$ in (1), and that no feedback effect is present. Details of the treatment for cases when these noise components follow normal distributions with arbitrary covariance matrices is given in the *supplementary methods* (stacks.iop.org/JNE/15/066016/mmedia) document accompanying this article.

We will begin our technical description by addressing a simple configuration, which considers the model for one subject in isolation, and whose output measurements are defined

in the lower-dimensional output spatial resolution (parcellated case). The treatment of the extended configurations (i.e. multi-subject models and large dimensionality of output measurements) will be given in sections 2.4 and 2.5, respectively.

First, we note that using (1), the output at time k with an initial condition $x(0)$ and past inputs $u(r)$, for $r = 0, \ldots, k-1$, satisfies

$$y(k) = CA^k x(0) + \sum_{r=0}^{k-1} CA^{k-r-1} Bu(r) \tag{3}$$

when $k \geqslant 0$. Using (3), a sequence of s output samples $y(k)$ at times $k = 0, \ldots, s-1$ can be written in matrix form as

$$\underbrace{\begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ \vdots \\ y(s-1) \end{bmatrix}}_{=:Y_{0,s}} = \underbrace{\begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{s-1} \end{bmatrix}}_{=:\mathcal{O}_s} x(0) + \underbrace{\begin{bmatrix} 0 & 0 & \cdots & 0 \\ CB & 0 & & 0 \\ CAB & CB & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{s-2}B & CA^{s-3} & \cdots & 0 \end{bmatrix}}_{=:\mathcal{T}_s} \underbrace{\begin{bmatrix} u(0) \\ u(1) \\ u(2) \\ \vdots \\ u(s-1) \end{bmatrix}}_{=:U_{0,s}}. \tag{4}$$

Here, $\mathcal{O}_s \in \mathbb{R}^{sp \times n}$ is a block-row matrix with each block of size $p \times n$, and $\mathcal{T}_s \in \mathbb{R}^{sp \times sm}$ is a structured block-Toeplitz matrix having each block having size $p \times m$. The block-row matrices $Y_{0,s} \in \mathbb{R}^{sp \times 1}$ and $U_{0,s} \in \mathbb{R}^{sm \times 1}$ have each block of size $p \times 1$ and $m \times 1$, respectively, with the first subscript denoting the sample time index of the block element in their first row. By horizontally concatenating N block-row matrices $Y_{i,s}$ (for $i = 0, \ldots, N-1$), we obtain the structured block-Hankel matrix $\mathcal{Y}_{0,s,N} \in \mathbb{R}^{sp \times N}$ having each block of size $p \times 1$, i.e.

$$\mathcal{Y}_{0,s,N} := \begin{bmatrix} y(0) & y(1) & \cdots & y(N-1) \\ y(1) & y(2) & \cdots & y(N) \\ \vdots & \vdots & \ddots & \vdots \\ y(s-1) & y(s) & \cdots & y(N+s-2) \end{bmatrix}.$$

Likewise, $\mathcal{U}_{0,s,N} \in \mathbb{R}^{sm \times N}$ is a block-Hankel matrix with each block of size $m \times 1$, following the same structure as in $\mathcal{Y}_{0,s,N}$. Hence, by defining the block-row matrix $X_{0,N} \in \mathbb{R}^{n \times N}$, where $X_{0,N} := [x(0)\ x(1)\ \ldots\ x(N-1)]$, we can write, from (4), the *data equation*

$$\mathcal{Y}_{0,s,N} = \mathcal{O}_s X_{0,N} + \mathcal{T}_s \mathcal{U}_{0,s,N}. \tag{5}$$

This equation will be subsequently analyzed to produce the estimates for the state-space representation matrices A, B and C, as well as for the initial state $x(0)$.

2.3.1. Estimates of A and C. We consider the *projection* matrix $\Pi^{\perp}_{\mathcal{U}_{0,s,N}} \in \mathbb{R}^{N \times N}$, defined by

$$\Pi^{\perp}_{\mathcal{U}_{0,s,N}} := I_N - \mathcal{U}^{\mathsf{T}}_{0,s,N} (\mathcal{U}_{0,s,N} \mathcal{U}^{\mathsf{T}}_{0,s,N})^{\dagger} \mathcal{U}_{0,s,N},$$

which can be explicitly computed from the set of observed inputs $\{u(k)\}_{k=0}^{N+s-2}$. By right-multiplying both sides of the data equation (5) by $\Pi^{\perp}_{\mathcal{U}_{0,s,N}}$, this projection matrix cancels the term $\mathcal{T}_s \mathcal{U}_{0,s,N}$, yielding

$$\mathcal{Y}_{0,s,N} \Pi^{\perp}_{\mathcal{U}_{0,s,N}} = \mathcal{O}_s X_N \Pi^{\perp}_{\mathcal{U}_{0,s,N}}. \tag{6}$$

We now observe that the l.h.s. of (6) can be computed based on the observed output data $\{y(k)\}_{k=0}^{N+s-2}$ and on the projection matrix $\Pi^{\perp}_{\mathcal{U}_{0,s,N}}$, which is also constructed from the known inputs. As we will see next, this term provides a basis for estimating a basis for the linear space spanned by the system's observability matrix, a key object in this method. To do so, we perform the singular value decomposition

$$U \Sigma V^{\mathsf{T}} := \mathcal{Y}_{0,s,N} \Pi^{\perp}_{\mathcal{U}_{0,s,N}}, \tag{7}$$

and produce associated matrices U_n, Σ_n, V_n where the subscript n indicates that only the left and right singular vectors associated with the n largest singular values are retained. Here, n is the dimension of the state of the system being estimated. In the deterministic case, this singular value decomposition produces exactly n nonzero singular values, so that $U \Sigma V = U_n \Sigma_n V_n$. In the presence of noise, there will be typically more than n nonzero singular values, and n becomes a parameter of choice in the method.

Proceeding with the deterministic case, right-multiplying (7) by $V_n \Sigma_n^{-1}$ and comparing with (6) gives

$$U_n = \mathcal{Y}_{0,s,N} \Pi^{\perp}_{\mathcal{U}_{0,s,N}} V_n \Sigma_n^{-1} = \mathcal{O}_s X_N \Pi^{\perp}_{\mathcal{U}_{0,s,N}} V_n \Sigma_n^{-1} = \mathcal{O}_s T,$$

where we have defined $T := X_N \Pi^{\perp}_{\mathcal{U}_{0,s,N}} V_n \Sigma^{-1}$ as a similarity transformation matrix. Provided that T preserves the column space of $\mathcal{O}_s$ (which can be assured by the conditions presented in [20, lemma 9.1]), we may write

$$U_n = \mathcal{O}_s T = \begin{bmatrix} CT \\ CT(T^{-1}AT) \\ \vdots \\ CT(T^{-1}AT)^{s-1} \end{bmatrix} =: \begin{bmatrix} C_T \\ C_T A_T \\ \vdots \\ C_T A_T^{s-1} \end{bmatrix}, \tag{8}$$

and verify that the column space of U_n is equivalent to the column space of the extended observability matrix of the system being estimated. This fact will allow us to use the computed matrix U_n to determine the matrices $\hat{A}_T$ and $\hat{C}_T$, which are similarity-transformed estimates of the matrices A and C, as follows. First we note that (8) allows us to write

$$\begin{bmatrix} C_T \\ \vdots \\ C_T A_T^{s-2} \end{bmatrix} A_T = \begin{bmatrix} C_T A_T \\ \vdots \\ C_T A_T^{s-1} \end{bmatrix},$$

which can be equivalently written in terms of sub-matrices derived from U_n as

$$[U_n]_{1:n(s-1),1:n} A_T = [U_n]_{n+1:ns,1:n}. \tag{9}$$

This is an overdetermined linear equation in the matrix variable A_T, whose least-squares solution

$$\hat{A}_T = \arg\min_{A_T} \left\| [U_n]_{1:n(s-1),1:n} A_T - [U_n]_{n+1:ns,1:n} \right\|_F^2 \tag{10}$$

can be obtained in closed-form as

$$\hat{A}_T := \left([U_n]_{1:n(s-1),1:n}\right)^{\dagger} [U_n]_{n+1:ns,1:n},$$

where the dagger symbol denotes the Moore–Penrose pseudo-inverse matrix. Similarly, the estimate for C can be obtained by letting

$$\hat{C}_T := [U_n]_{1:p,1:n},$$

i.e, by retrieving the top $n \times n$ submatrix of U_n.

A comment is in order with respect to the set of eigenvalues $\{\lambda_i(\hat{A}_T) : i = 1, \ldots, n\}$ of the estimated state transition matrix $\hat{A}_T$. It is a known fact that an unstable system is implied if $|\lambda_i(\hat{A}_T)| > 1$ for any $i = 1, \ldots, n$ [31]. In that respect, we mention two possible approaches to enforce stability. The first, simpler, is to perform an eigen-decomposition of $\hat{A}_T$ and reconstruct it by imposing that any eigenvalues with absolute value greater than one be restricted to a stable value (i.e. to a value equal or slightly smaller than one). The second approach consists of introducing an explicit stability constraint on the minimization problem (10). This alternative, described in detail in [32], is more computationally-intensive, and involves the application of an iterative convex optimization algorithm to account for the required constraint in the resulting semi-definite program (SDP).

2.3.2. Estimates for B and x(0). Given the estimates $\hat{A}_T$ and $\hat{C}_T$, as well as the input–output data $\{y(k), u(k)\}_{k=0}^{L-1}$, one can find similarity-transformed estimates $\hat{x}_T(0)$ for $x(0)$, and $\hat{B}_T$ for B, as follows. We apply the vectorization operator in (3) and use the identity $\mathrm{vec}(LXR) = (R^{\mathsf{T}} \otimes L)\mathrm{vec}(X)$, to get

$$\begin{aligned} y(k) &= \mathrm{vec}(y(k)) \\ &= \mathrm{vec}\left(C_T A_T^k x_T(0) + \sum_{r=0}^{k-1} C_T A_T^{k-r-1} B u(r)\right) \\ &= C_T A_T^k x_T(0) + \left(\sum_{r=0}^{k-1} u(r)^{\mathsf{T}} \otimes C_T A_T^{k-r-1}\right)\mathrm{vec}(B_T), \end{aligned}$$

which is a linear equation in the variables $x_T(0)$ and $\mathrm{vec}(B_T)$. We now define a set of coefficient matrices $\{\phi(k) \in \mathbb{R}^{n(m+1)\times p}\}_{k=0}^{L-1}$, whose computation is based on the existing estimates $\hat{A}_T$ and $\hat{C}_T$, such that

$$\phi(k) := \left[\begin{array}{c|c} \hat{C}_T \hat{A}_T^k & \left(\sum_{r=0}^{k-1} u(r)^{\mathsf{T}} \otimes \hat{C}_T \hat{A}_T^{k-r-1}\right) \end{array}\right]^{\mathsf{T}}.$$

Further, we define the auxiliary parameter variable $\theta \in \mathbb{R}^{n(m+1)}$, with $\theta := [\,x_T(0)^{\mathsf{T}}\ \mathrm{vec}(B_T)^{\mathsf{T}}\,]^{\mathsf{T}}$, and note that we can find an estimate $\hat{\theta}$ by solving the minimum least squares problem

$$\hat{\theta} := \arg\min_{\theta} \quad \sum_{k=0}^{L-1} \frac{1}{2} \left\| y(k) - \phi(k)^{\mathsf{T}}\theta \right\|_2^2. \qquad (11)$$

A solution to the optimization problem in (11) can be found in closed-form by defining the vector $y_L \in \mathbb{R}^{Lp}$ with $y_L := [y(0)^{\mathsf{T}}, \cdots, y(L-1)^{\mathsf{T}}]^{\mathsf{T}}$ and the coefficient matrix $\Phi \in \mathbb{R}^{pL\times n(m+1)}$ with $\Phi := [\phi(0)|\cdots|\phi(L-1)]^{\mathsf{T}}$, such that

$$\hat{\theta} = (\Phi^{\mathsf{T}}\Phi)^{-1}\Phi y_L = \Phi^{\dagger} y_L, \qquad (12)$$

where $\Phi^{\dagger}$ is the Moore–Penrose left pseudo-inverse of Φ. Finally, the desired estimates are directly obtained from $\hat{\theta}$ by letting $\hat{x}_T(0) := [\hat{\theta}]_{1:n,1}$ and $\hat{B}_T := \mathrm{vec}_{m,n}^{-1}\left([\hat{\theta}]_{n+1:n(m+1),1}\right)$.

2.4. Dense time series

We now propose a computationally efficient method for identifying our approximating model for spatially dense time series data, i.e. the case when the output variables $y \in \mathbb{R}^p$ are defined in a space where $p \gg n, L$. In the case of the Human Connectome Project dataset explored in this paper (see description in section 3), the dense data consists of $p = 64\,984$ brain surface coordinates, with the first 32 492 being related to the left hemisphere, and the remaining 32 492 to the right hemisphere. These values become relevant from a computational aspect when we consider the steps in fitting a model using the above described subspace method. In particular, we highlight two specific steps: (i) the singular value decomposition in (7), required for the estimation of the parameters $\hat{A}_T$ and $\hat{C}_T$, and (ii) the least-squares solution in (12), required for the estimation of the parameters $\hat{B}_T$ and initial state $\hat{x}_T(0)$. In (i), the SVD is applied on the matrix $\mathcal{Y}_{0,s,N}\Pi^{\perp}_{\mathcal{U}_{0,s,N}} \in \mathbb{R}^{sp\times N}$, with e.g. $sp = 194\,952$ (for $s = 3$) and $N = 284$; whereas in (ii) the solution involves storing and calculating a pseudo-inverse of the matrix Φ of size e.g. $pN \times n(m+1) = 18.45 \cdot 10^6 \times 280$, accounting for potentially up to $5.17 \cdot 10^9$ entries. The computational cost for performing such operations is further compounded when one considers a collection of subjects, and when the behavior of the model at different parameters is analyzed, requiring many model-fitting iterations. By applying recent developments in numerical linear algebra and optimization, we propose computationally efficient solutions for both of these problems, enabling such computations to be performed within a reasonable time.

First, we address the SVD in (7) using a randomized algorithm with approximation guarantees, introduced in [33]. The principle enabling the gain in computational efficiency in the proposed approach lies in the fact that the SVD computation will be performed on a representative matrix of smaller dimensions (also called a *sketch*) $S \in \mathbb{R}^{n\times\zeta}$, which is appropriately sampled from the original matrix $\mathcal{Y}_{0,s,N}\Pi^{\perp}_{\mathcal{U}_{0,s,N}}$. The result of this SVD is then related to the original matrix decomposition $U\Sigma V^{\mathsf{T}} := \mathcal{Y}_{0,s,N}\Pi^{\perp}_{\mathcal{U}_{0,s,N}}$, yielding the desired matrices $\tilde{U}_n, \tilde{S}_n, \tilde{V}_n$ such that $\|U_n S_n V_n - \tilde{U}_n \tilde{S}_n \tilde{V}_n\|_{\mathcal{F}} \leqslant \epsilon$. It is worth noticing that the quality of the approximation can be guaranteed and controlled by the sketch size parameter ζ in the order of n/ϵ. For example, for rank $n = 30$ and tolerance $\epsilon = 3 \cdot 10^{-3}$ we select $\zeta \sim 1 \cdot 10^4$. The computational steps are presented in algorithm 1, following [34, section 5.2].

Algorithm 1. Large-scale approximate SVD.

Require: $\mathcal{Y}_{0,s,N}$, # of singular values n, sketch size ζ
1: Generate a sketch matrix $S := \text{sketch}(\mathcal{Y}_{0,s,N}, \zeta)$
2: Let $C := \mathcal{Y}_{0,s,N} S$
3: Perform QR decomposition $[Q, R] := \text{qr}(C)$
4: Perform n-SVD $[\tilde{U}_0, \tilde{S}_n, \tilde{V}_n] := \text{svd}(Q^{\mathsf{T}} \mathcal{Y}_{0,s,N}, n)$
5: Let $\tilde{U}_n := Q\tilde{U}_0$
6: **return** $\tilde{U}_n \tilde{S}_n \tilde{V}_n \approx \mathcal{Y}_{0,s,N}$

Next, we propose a computationally efficient solution for the least-squares problem (12) in the dense time series case. To do so, we address the problem in its optimization form as in (11), and describe an iterative solution using *proximal methods* [35]. The method is based on the proximal operator $\text{prox}_{f,\lambda,\theta} : \mathbb{R}^{n(m+1)} \to \mathbb{R}^{n(m+1)}$, which is defined according to the objective function in (11) as follows:

$$\begin{aligned}\text{prox}_{f,\lambda,\theta}(w) &:= \arg\min_{\theta} f(\theta) + \frac{1}{2\lambda}\|w-\theta\|_2^2 \\ &= \arg\min_{\theta} \sum_{k=0}^{L-1} \frac{1}{2}\left\|y(k) - \phi(k)^{\mathsf{T}}\theta\right\|_2^2 + \frac{1}{2\lambda}\|w-\theta\|_2^2,\end{aligned}$$

where λ is a scaling parameter. This implicit definition of the proximal operator, in terms of its quadratic minimization in θ, can be solved and alternatively represented in closed-form as

$$\begin{aligned}\text{prox}_{f,\lambda,\theta}(w) = &\left(I_{n(m+1)} + \lambda \sum_{k=0}^{L-1} \phi(k)\phi(k)^{\mathsf{T}}\right)^{-1} \\ &\times \left(\sum_{k=0}^{L-1} \phi(k)y(k) + \lambda w\right). \end{aligned} \tag{13}$$

Seeking gains in computational efficiency, we further express factors in (13) by defining a parameter matrix $W \in \mathbb{R}^{n(m+1)\times n(m+1)}$ with

$$W := \left(I_{n(m+1)} + \lambda \sum_{k=0}^{L-1} \phi(k)\phi(k)^{\mathsf{T}}\right)^{-1}, \tag{14}$$

and a parameter vector $h \in \mathbb{R}^{n(m+1)}$ where $h := \sum_{k=0}^{L-1} y(k)\phi(k)$. Importantly, we note that, for a given value of the parameter λ, both W and h can be *precomputed*, since they do not depend on the argument w of the proximal operator. In addition, since the dimensions of these parameters do not depend on the large dimension p of the output measurements $y(k)$ or the number of time samples L, these parameters can be conveniently stored and used across iterations. The proximal operator thus becomes a simple affine transformation

$$\text{prox}_{f,\lambda,\theta}(w) = W(h + \lambda w), \tag{15}$$

which can be evaluated at a much lower computational cost.

The optimization algorithm (summarized in algorithm 2) consists of iteratively applying the proximal operator on iterates $\theta^{[l]}$ of the optimization variable θ, indexed by iteration l, until a termination criterion is met (e.g. relative tolerance ξ over the norm of the difference in θ over successive iterations). With respect to the scaling parameter λ, convergence to the set of minimizers of the objective function is guaranteed provided $\lambda^{[l]} > 0$ and $\sum_{l=1}^{\infty} \lambda^{[l]} = \infty$ [35, p 143], which is satisfied by letting $\lambda^{[l]} = 1$ for all l, as we adopt in this paper.

Algorithm 2. Proximal method for dense time series.

Require: Scaling parameter λ, tolerance ξ
1: Let $\theta^{[1]} = 0$
2: **while** $\|\theta^{[l+1]} - \theta^{[l]}\|_2 / \|\theta^{[l]}\|_2 \geqslant \xi$ **do**
3: $\quad \theta^{[l+1]} = \text{prox}_{f,\lambda,\theta}(\theta^{[l]})$
4: **end while**
5: **return** $\hat{\theta} = \theta^{[l]} = [\hat{x}_T(0)^{\mathsf{T}}, \text{vec}(\hat{B}_T)^{\mathsf{T}}]^{\mathsf{T}}$

2.5. Multi-subject identification

We consider a set of subjects $\{h : h \in 1, \ldots, q\}$ associated with a given task, and we seek a *common* state-space representation across all subjects. We denote the input block-Hankel matrices associated with subject h as $\mathcal{Y}_{s,N_h}^{(h)} \equiv \mathcal{Y}_{0,s,N_h}^{(h)}$, dropping the first subscript for brevity, and allowing for different block column dimensions N_h. Correspondingly, we denote the output block-Hankel matrix $\mathcal{U}_{s,N_h}^{(h)} \equiv \mathcal{U}_{0,s,N_h}^{(h)}$ and block-row state matrices $X_{N_h}^{(h)} \equiv X_{0,N_h}^{(h)}$. We can write a *multi-subject* data equation, analogous to (5), as

$$\begin{aligned}\left[\mathcal{Y}_{s,N_1}^{(1)} | \ldots | \mathcal{Y}_{s,N_q}^{(q)}\right] = &\mathcal{O}_s \left[X_{N_1}^{(1)} | \ldots | X_{N_q}^{(q)}\right] \\ &+ \mathcal{T}_s \left[\mathcal{U}_{s,N_1}^{(1)} | \ldots | \mathcal{U}_{s,N_q}^{(q)}\right].\end{aligned}$$

In this case, since the functional dependency between the system dependent matrices $\mathcal{O}_s$ and $\mathcal{T}_s$ is the same as the one in (5), the steps for obtaining the *common* matrix estimates $\hat{A}_T$ and $\hat{C}_T$ are exactly the same as those presented in section 2.3, in this case considering $\mathcal{Y}_{0,s,N} := [\mathcal{Y}_{s,N_1}^{(1)} | \ldots | \mathcal{Y}_{s,N_q}^{(q)}]$, $X_{0,N} = [X_{N_1}^{(1)} | \ldots | X_{N_q}^{(q)}]$ and $\mathcal{U}_{0,s,N} := [\mathcal{U}_{s,N_1}^{(1)} | \ldots | \mathcal{U}_{s,N_q}^{(q)}]$. In contrast, because the internal state $x_T^{(h)}(k)$, and in particular the initial state $x_T^{(h)}(0)$ is specific for each subject, the estimation step involving the *common* parameter B_T and the *individual* condition $x_T^{(h)}(0)$ requires us to examine the individual equations

$$y^{(h)}(k) = \hat{C}_T \hat{A}_T^k x_T^{(h)}(0) + \sum_{r=0}^{k-1} \hat{C}_T \hat{A}_T^{k-r-1} B_T u^{(h)}(r) \tag{16}$$

for each $h = 1, \ldots, q$. We note that these equations are coupled through the common parameter B_T. To address this, we define $b := \text{vec}(B_T)$, $x_h \equiv x_T^{(h)}(0)$, and

$$Y_h := \begin{bmatrix} y^{(h)}(0) \\ y^{(h)}(1) \\ \vdots \\ y^{(h)}(L-1) \end{bmatrix}, \; \mathcal{O}_L := \begin{bmatrix} \hat{C}_T \\ \hat{C}_T\hat{A}_T \\ \vdots \\ \hat{C}_T\hat{A}_T^{L-1} \end{bmatrix}.$$

We also define $\mathcal{K}_h = [K^{(h)}(0)^{\mathsf{T}}, \cdots, K^{(h)}(L-1)^{\mathsf{T}}]^{\mathsf{T}}$, where

$$K^{(h)}(k) := \sum_{r=0}^{k-1} u^{(h)}(r)^{\mathsf{T}} \otimes \hat{C}_T \hat{A}_T^{k-r-1}.$$

Then, we propose to solve the quadratic problem

$$\min_{\{x_h\}_{h=1}^q, b} \quad \sum_{h=1}^q \tfrac{1}{2} \|Y_h - \mathcal{O}_L x_h - \mathcal{K}_h b\|_2^2 \tag{17}$$

by proximal algorithms. Similarly to (13), the corresponding proximal operators can be derived in closed-form, respectively for b and $x_{0,h}$, as:

$$\begin{aligned}\text{prox}_{f,\lambda,b}(w) = &\left(I_{mn} + \lambda \sum_{t=1}^q \mathcal{K}_h^{\mathsf{T}} \mathcal{K}_h\right)^{-1} \\ &\times \left(w - \lambda \sum_{h=1}^q \mathcal{K}_h^{\mathsf{T}} \mathcal{K}_h b + \lambda \sum_{h=1}^q \mathcal{K}_h^{\mathsf{T}} Y_i\right), \\ \text{prox}_{f,\lambda,x_h}(w) = &\left(I_n + \lambda \mathcal{O}_L^{\mathsf{T}} \mathcal{O}_L\right)^{-1} \\ &\times \left(w - \lambda \mathcal{O}_L^{\mathsf{T}} \mathcal{K}_h b + \lambda \mathcal{O}_L^{\mathsf{T}} Y_h\right).\end{aligned}$$

We can now solve (17) by recursively evaluating

$$\begin{cases} x_h^{(l+1)} & = \text{prox}_{f,\lambda,x_h}(x_h^{(l)}),\ h = 1, \ldots, q \\ b^{(l+1)} & = \text{prox}_{f,\lambda,b}(b^{(l)}), \end{cases}$$

until the stopping criterion is met. Finally, we can retrieve the estimates $\hat{B}_T := \text{vec}_{m,n}^{-1}(b)$ and $\hat{x}_T^{(h)}(0) = x_h$, as desired.

3. Experimental description

3.1. The Human Connectome Project dataset

The dataset that we explore in this paper is derived from the Human Connectome Project (HCP) [36], as part of the HCP 900 subjects release. The acquisition and pre-processing pipelines are discussed in [37], and described in detail in [38]. Here, we provide a brief summary of the aspects that are most relevant to our problem.

3.1.1. Acquisition and pre-processing. Functional MRI data were acquired using a Gradient-echo EPI sequence, at a TR (sampling rate) of 720 ms, and additional parameters TE = 33.1 ms, flip angle = 52°, FOV = 208 × 180 mm (RO × PE), 72 slices, 2.0 mm isotropic views, multi-band acceleration factor of 8, echo spacing of 0.58 ms and a BW of 2290 Hz Px^{-1}. The pre-processing of the data comprised the following steps: gradient unwarping, motion correction, field-map-based EPI distortion correction, brain-boundary-based registration of EPI to structural T1-weighted scan, non-linear (FNIRT) registration into MNI152 space, and grand-mean intensity normalization. These procedures were executed according to the HCP data analysis pipelines using the FSL and FreeSurfer software packages, and following the steps described in [37].

3.1.2. Surface registration and spatial encoding. The HCP project provides cortical neuroimaging data in a surface-constrained format, as an alternative to the standard volumetric format. A motivation for the surface format is that distances defined in the geodesic surface of the brain are more neurobiologically relevant than the distances evaluated in volumetric space. In addition, in the surface format, the voxels of the cortical gray matter ribbon are projected onto a registered surface mesh with a standard number of vertices that is more efficiently encoded and stored. The coordinates defined by this mesh are referred to as *grayordinates*. In this system, each brain hemisphere is represented by 32 492 grayordinates, which are appended by a volumetric representation of subcortical structures, in sum providing 91 282 spatial points per fMRI frame. Excluding the subcortical grayordinates (not analyzed in this paper), we refer to this spatially dense representation of 64 984 grayordinates per time frame as DENSE. Alternatively, the spatial representation of the surface can be summarized into regions-of-interest (ROI) by the use of a standard cortical atlas. More specifically, we adopted the Destrieux 2009 atlas, which is composed of 74 regions per hemisphere, following internationally accepted nomenclature and criteria [39]. In this case, the signal intensities in all grayordinates associated with a region are averaged together to account for the representative signal intensity at that region. Following the HCP nomenclature, we refer to this type of spatial encoding as the APARC parcellation.

3.1.3. Physiological data. Cardiac signals were measured by a pulse oximeter, sampled at 400 Hz (288 samples per frame of functional image), and synchronized with the scanner. Respiratory signals, also sampled at 400 Hz, were measured by an elastic respiratory belt transducer. These signals account for oscillations occurring typically at 60 to 100 cycles a minute for the cardiac signals, and at 6 to 20 cycles per minute for the respiratory signals. Head motion was acquired by an optical motion tracking camera system in real-time using an infrared camera mounted in the scanner bore. The estimates of motion parameters were derived from a rigid-body transformation to the SBRef image acquired at the start of each fMRI scan, and comprise six parameters: trans_x, trans_y, trans_z, rot_x(deg), rot_y(deg), and rot_z(deg).

3.1.4. The motor task. In the MOTOR task [40], a set of six task conditions is considered. The participants are presented with (i) visual cues that ask them to tap their (ii) left or (iii) right fingers, squeeze their (iv) left or (v) right toes, or (vi) move their tongue. The task is intended to map motor areas in the cortex, and its design is inspired by [41], where its motivation for an examination of cognitive function is discussed. Each task run has a duration of 3 min 34 s, corresponding to 284 samples in time, i.e. 284 full spatial acquisition frames. The time-of-occurrence of each task condition is recorded with a precision of milliseconds, enabling its consideration as input data in our model.

3.2. Time series pre-processing

3.2.1. Centering and normalizing. For each region $i = 1, \ldots, p$, the output samples $\{[y(k)]_i\}_{k=0}^{L-1}$ were subtracted

by their mean values $\bar{y}_i := 1/L \sum_{k=0}^{L-1} [y(k)]_i$ and scaled by $(1/\sigma_i)$, with $\sigma_i^2 := 1/L \sum_{k=0}^{L-1} ([y(k)]_i - \bar{y}_i)^2$ being the squared sample standard deviation.

3.2.2. Filtering. Within the spectrum of the fMRI BOLD signal, a specific frequency sub-band is predominantly associated with neural activity [42]. We therefore consider two alternatives of the frequency content of the signal, which we refer to as ALL-PASS and BAND-PASS. In the ALL-PASS case, the whole frequency band of the signal is preserved; in the BAND-PASS case, we attenuate the frequencies outside the 0.06–0.12 Hz band by applying a band-pass filter, defined as follows. We design an order 50 FIR-type filter using MATLAB's 'equiripple' method, so as to achieve a 20 dB attenuation outside the pass band. The initial stop and pass frequencies considered were $f_{s1} = 0.04$ Hz, $f_{p1} = 0.06$ Hz, and the final stop and pass frequencies were $f_{p2} = 0.12$ Hz and $f_{s2} = 0.15$ Hz, respectively. Both filter alternatives are explored in the experiments analyzed in this paper.

3.2.3. Physiological signals. We perform a linear regression of the physiological inputs on the outputs, using cognitive inputs $u_c(k) \in \mathbb{R}^{m_c}$, and physiological inputs $u_p(k) \in \mathbb{R}^{m_p}$, defined as

$$u_c(k) := \begin{bmatrix} \text{cue}(k) \\ \text{lf}(k) \\ \text{lh}(k) \\ \text{rh}(k) \\ \text{rh}(k) \\ \text{t}(k) \end{bmatrix} \quad u_p(k) := \begin{bmatrix} \text{heart}(k) \\ \text{resp}(k) \\ \text{trans}_x(k) \\ \text{trans}_y(k) \\ \text{trans}_z(k) \\ \text{rot}_x(k) \\ \text{rot}_y(k) \\ \text{rot}_z(k) \end{bmatrix},$$

with $m_c = 6$ and $m_p = 8$. To do so, we build a data matrix $U_p \in \mathbb{R}^{m_p \times L}$ containing the physiological inputs, $U_p := [u_p(0) \dots u_p(L-1)]$, and a data matrix $Y \in \mathbb{R}^{p \times L}$ containing the measured outputs, $Y := [y(0) \dots y(L-1)]$. We then find a matrix of linear regression coefficients $\hat{H} \in \mathbb{R}^{p \times m_p}$ corresponding to the least-squares solution to

$$\hat{H} := \arg\min_H \|Y - HU_p\|_{\mathcal{F}}^2 = YU_p^{\dagger}.$$

Finally, we define the set of *physiologically regressed* outputs $\{y_r(k)\}_{k=0}^{L-1}$ by

$$[\, y_r(0) \dots y_r(L-1)] := Y - \hat{H}U_p = Y(I_L - U_p^{\dagger}U_p),$$

i.e. by removing the direct linear prediction of the physiological inputs on the outputs.

3.3. System identification parameters

The system identification method presented in section 2.3 depends on the parameters s and N, associated with the block-Hankel matrices, as well as on the parameter n, corresponding to the dimension of the state representation of the system. We note that, for a given value of the parameter s (the number of block-rows) and number of samples L, the structure of the Hankel matrix allows a maximum value for the parameter N to be $N_{\max} := L - s + 1$. In the following experiments, unless otherwise noted, we adopt $s = 3$. Given that $L = 284$ for the MOTOR task, we have $N = N_{\max} = 282$. Finally, we set the sketch size $\zeta = 1 \cdot 10^4$ (yielding tolerance $\epsilon \sim 1 \cdot 10^{-3}$), and proximal algorithm stopping tolerance $\xi = 1 \cdot 10^{-6}$.

4. Results

4.1. Task-fMRI time series is well approximated with a low degrees-of-freedom model

In this section, we analyze the quality of the approximated output signal when the method described in section 2.3 is applied individually to each subject $h = 1, \dots, q$ in the dataset. For each subject, we consider experimental data collected from the MOTOR task in the form of a set of inputs $\{u^{(h)}(k)\}_{k=0}^{L-1}$ and a set of outputs $\{y^{(h)}(k)\}_{k=0}^{L-1}$. First, we obtain the individual system matrix estimates $(\hat{A}_T^{(h)}, \hat{B}_T^{(h)}, \hat{C}_T^{(h)})$ and the initial state $\hat{x}_T^{(h)}(0)$ estimate, given chosen values for (n, s, N). Next, using the original inputs $\{u^{(h)}(k)\}_{k=0}^{L-1}$ and the initial condition $x_T^{(h)}(0)$, we produce the outputs $\{\hat{y}^{(h)}(k)\}_{k=0}^{L-1}$ as the response generated by the estimated system $(\hat{A}_T^{(h)}, \hat{B}_T^{(h)}, \hat{C}_T^{(h)})$. We then measure the quality of the approximation obtained for each region $i = 1, \dots, p$, by computing the Pearson correlation coefficient between the original and the approximated time series, i.e. $\gamma_i^{(h)} := \langle [y^{(h)}]_i, [\hat{y}^{(h)}]_i \rangle$.

The results from performing this procedure, when considering the parcellated times series ($p = 148$) for both the ALL-PASS and BAND-PASS filters, are presented in figure 3. The dimension of the state was defined to be $n = 25$, by inspection of the singular values in (7), fulfilling the condition that $sp > n$. We display the original and reconstructed outputs for the three regions with highest $\gamma_i^{(h)}$ (left) and for the three regions with $\gamma_i^{(h)}$ around the median across regions (right), considering the subject h whose *across-regions* average correlation coefficient (i.e. $\frac{1}{p}\sum_{i=1}^{p} \gamma_i^{(h)}$) was in the median of the *across-subjects* distribution. We observe that the approximate output captures the main features of the BOLD signal in all cases, with especially high accuracy for the BAND-PASS filter.

In figure 4, we build the sample covariance matrices $F \in \mathbb{S}^p$ where each entry $[F]_{i,j}$ is given by the sample correlation coefficient between the output signal at every pair of regions $(i, j) \in \{1, \dots, p\}^2$, i.e. $[F]_{i,j} := \langle [y]_i, [y]_j \rangle$. We also compare F obtained from the original inputs $y(k)$ with $\hat{F}$ obtained from the approximated outputs $\hat{y}(k)$ by displaying the matrix of the absolute values of their difference, which were found to be of low magnitude on average.

In figure 5, we consider the *dense* times series ($p = 64\,984$), and apply the method described in section 2.4 with both the

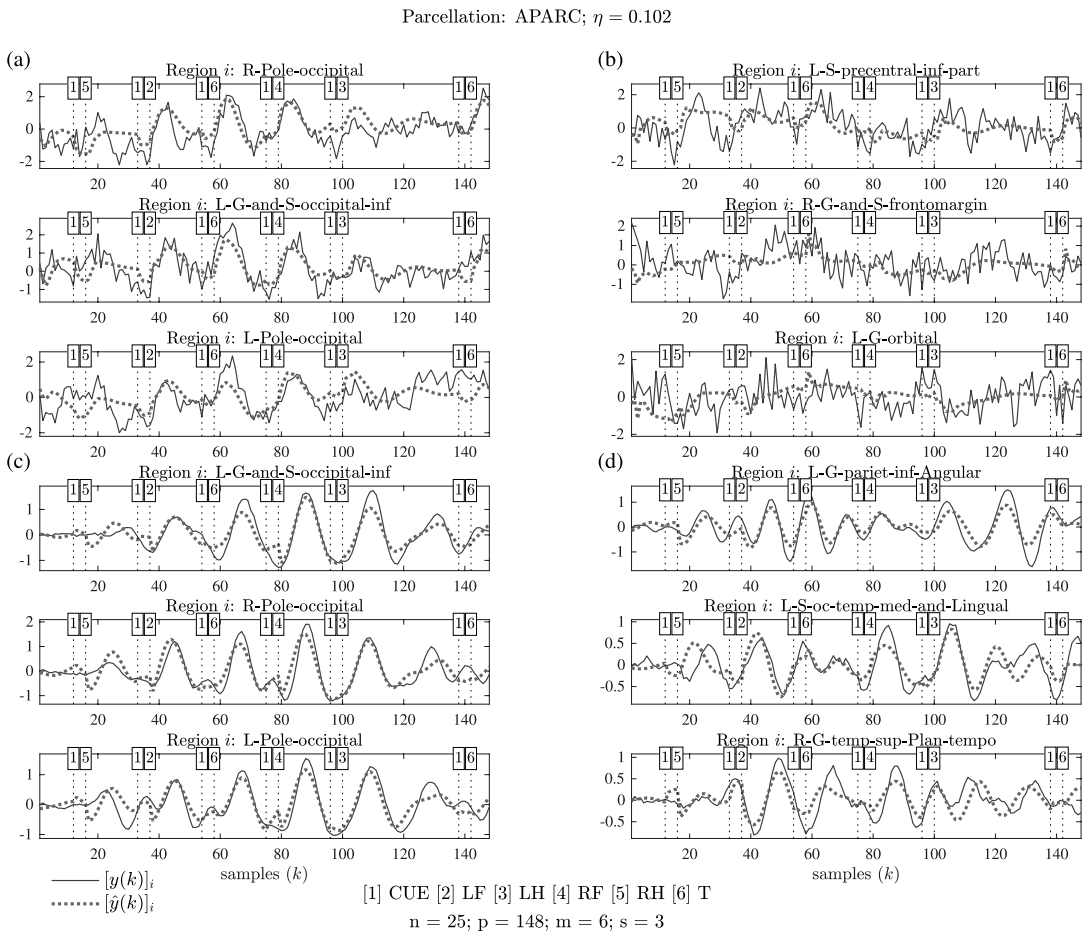

Figure 3. Original and reconstructed output times series for the MOTOR task with the APARC parcellation, for filters ALL-PASS (a) and (b), and BAND-PASS (c) and (d). Values are presented for the subject h whose *across-regions* average correlation coefficient (i.e. $\frac{1}{p}\sum_{i=1}^{p}\gamma_i^{(h)}$) was in the median of the *across-subjects* distribution. In (a) and (c), we display values for the reconstructed outputs $\{[y(k)]_i\}_{k=0}^{L-1}$ for regions $i \in \mathcal{I}_{\text{high}} \subset \{1,\ldots,p\}$ with $p = 148$, where the subset $\mathcal{I}_{\text{high}}$ consists of the three regions for which the correlation $\gamma_i^{(h)}$ was *highest*. In (b) and (d), values correspond to the reconstructed outputs $\{[y(k)]_i\}_{k=0}^{L-1}$ for regions $i \in \mathcal{I}_{\text{med}} \subset \{1,\ldots,p\}$, where the subset $\mathcal{I}_{\text{med}}$ consists of the three regions for which the correlation $\gamma_i^{(h)}$ was adjacent to the *median*. The vertical dotted lines across each plot mark the time-of-occurrence of each task condition, with the task condition number being indicated in the small box on top.

ALL-PASS and BAND-PASS filters, considering the state dimension parameter $n = 35$. We display the original and reconstructed outputs for the three regions with highest $\gamma_i^{(h)}$ and the three regions with $\gamma_i^{(h)}$ around the median across regions, for the subject h whose *across-regions* average correlation coefficient (i.e. $\frac{1}{p}\sum_{i=1}^{p}\gamma_i^{(h)}$) was the median in the *across-subjects* distribution. We observe that, in this finer parcellation, the noise content of the ALL-PASS filter is relatively high, while the approximation quality for the BAND-PASS filter achieves an accuracy that is comparable to the one observed in the APARC parcellation.

Finally, to evaluate the approximation capacity of the models, we examine the ratio between the number of degrees of freedom allowed by the parameters of the model against the number of constraints imposed by the input and output observations, taking the deterministic system model as a reference. Considering L samples, we have that the set of inputs $u(k) \in \mathbb{R}^m$ and outputs $y(k) \in \mathbb{R}^p$ generates $L(p+m)$ constraints. On the side of the model, the matrices (A, B, C) correspond, respectively, to (n^2,nm,pn) scalar parameters, i.e. a total of $n(n+m+p)$ degrees of freedom. Therefore, the deterministic degree of parametrization of the models can be quantified by the *ratio of the degrees of freedom*

$$\eta := \frac{n(n+m+p)}{L(p+m)}. \tag{18}$$

When applied to our case, for single-task models, with $p = 148$, $m = 6$, $n = 25$, and $L = 284$ we have $\eta \approx 0.123$. Furthermore, we note that the degrees of freedom ratio decreases linearly with

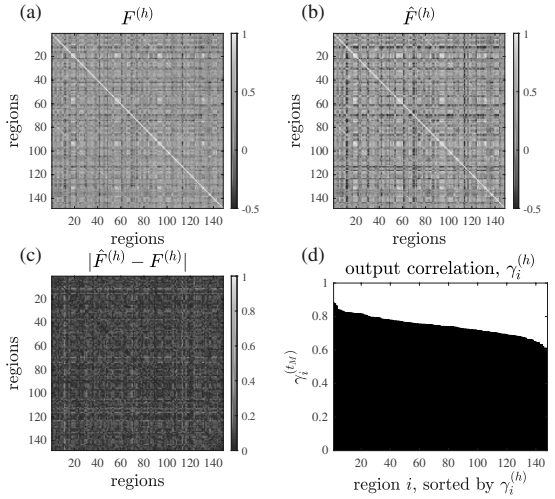

Figure 4. Covariance matrices $F^{(h)} \in \mathbb{S}^p$, where each entry $[F^{(h)}]_{i,j}$ is given by the correlation coefficient $[F^{(h)}]_{i,j} := \langle [y^{(h)}]_i, [y^{(h)}]_j \rangle$ between the output signal at every pair of regions $(i,j) \in \{1,\ldots,p\}^2$. We display, in (a), $F^{(h)}$ obtained from the original inputs $y(k)$; in (b), $\hat{F}^{(h)}$ obtained from the approximated outputs $\hat{y}(k)$; and in (c), the absolute entry-wise error from $|F^{(h)} - \hat{F}^{(h)}|$. In (d), we present the full distribution of output correlations $\gamma_i^{(h)}$ over regions $i = 1,\ldots,p$, for the same subject.

the number of samples L. For dense time series, where $p \gg m, n$ we have that $\eta \approx n/L$. In this case, for $n = 25$ and $L = 284$ we have $\eta \approx 0.102$, while for $n = 35$ we have $\eta \approx 0.123$. Based on the average correlations obtained for both the parcellated and dense time series, we can conclude that the models derived from the subspace methods are able to approximate the task-fMRI signal with a low degrees-of-freedom model.

4.2. Input–output transfer function identified is compatible with hemodynamic response function

We refer as *hemodynamic response function* (HRF), at a cortical point-of-interest and with respect to a stereotyped stimulus, to the time sequence of BOLD values occurring at that cortical point following the presentation of that stimulus. In terms of dynamical systems, this quantity can be also defined as an impulse response function [27]. In this regard, the state-space representation (A, B, C) contains the required information to generate the approximated impulse response on every brain region (or grayordinate) $i = 1,\ldots,p$ due to an impulse (i.e. task condition event occurrence) at any input channel $j = 1\ldots,m$. Subsequently, we denote the impulse response (IR) for an individual h due to an impulse at the jth input channel at time index $k = 0$ by $y^{(h)}(k)\big|_{\delta(j)}$. The IR corresponds to the system's output computed according to (16) for an input defined as $\delta(j) : [u(0)]_j = 1$ and $[u(k)]_j = 0$ for $k > 0$, applied at the jth input channel, with $x(0) = 0$. With the above definitions, in figure 6, we investigate the IR associated with two specific sets of regions of the cortex, OCCIPITAL and PARACENTRAL.

First, we examine the OCCIPITAL pole regions, which are areas associated with the visual cortex, whose HRF is commonly studied [6]. In figure 6(a), the *across-subjects* average impulse response function, i.e. $\bar{y}(k)\big|_{\delta(j)} := \frac{1}{q}\sum_{t=1}^{q} y^{(h)}(k)\big|_{\delta(j)}$, is presented for all $j = 1,\ldots,6$ input channels (corresponding to the six different task conditions). If compared in terms of their overall form (time-to-peak and total duration), it can be said that the HRF's obtained are in good agreement with the HRF's typically reported for that cortical area [6].

Next, we examine the IR for an area associated with motor control [43, 44], the PARACENTRAL lobule and sulcus in the Destrieux atlas [39]. In line with the previous experiment, we computed and presented in figure 6(b) the across-subjects average IR for that region, for all of the six task conditions. One notable aspect is the crossed association between a high amplitude of the response in the *left-hemisphere* region ('L-G-and-S-paracentral') for the RF (squeeze *right* toe) task condition. The corresponding crossed association in the symmetric *right-hemisphere* region ('R-G-and-S-paracentral' versus LF, squeeze *left* toe) is also present, possibly indicating some form of lateralization effect associated with motor function [45, 46]. Overall, we note that both the shape and time scales of the IR's obtained are also compatible with HRF's reported in the literature (see [27]).

4.3. Pairwise input–output $\mathcal{H}_2$ norm reveals region activation

The state-space formulation allows the representation of the multi-input multi-output transfer function of the system, which can be analyzed in terms of the typical and maximum rates of signal energy transfer between inputs and outputs, i.e. between task-related inputs and ensuing BOLD signal response. Formally, the transfer function $H(z)$ associated with the state-space parameters (A, B, C) is given by

$$H(z) = C(zI - A)^{-1}B,$$

where $H(z)$ is a $p \times m$ matrix whose entries are scalar transfer functions, i.e. rational polynomials, of the complex variable z. Given the transfer functions, one can associate an $\mathcal{H}_2$ norm $\|[H(z)]_{i,j}\|_2$ to each input–output pair $(i,j) \in \{1,\ldots,p\} \times \{1,\ldots,m\}$, which can be computed, for stable systems and under a stationarity assumption, as

$$\|[H(z)]_{i,j}\|_2^2 = \sum_{k=0}^{\infty} \left[y(k)\big|_{\delta(j)}\right]_i \left[y(k)\big|_{\delta(j)}\right]_i^*$$

where the symbol $*$ denotes complex conjugation. Such expression accounts for the sum of the squared absolute values of the ith output response $[y(k)\big|_{\delta(j)}]_i$ due to an impulse applied at the jth input channel. Intuitively, this $\mathcal{H}_2$ norm accounts for the total output energy observed at output region i as a result of a unit energy impulse input at channel j. In terms of the input–output brain models, this corresponds

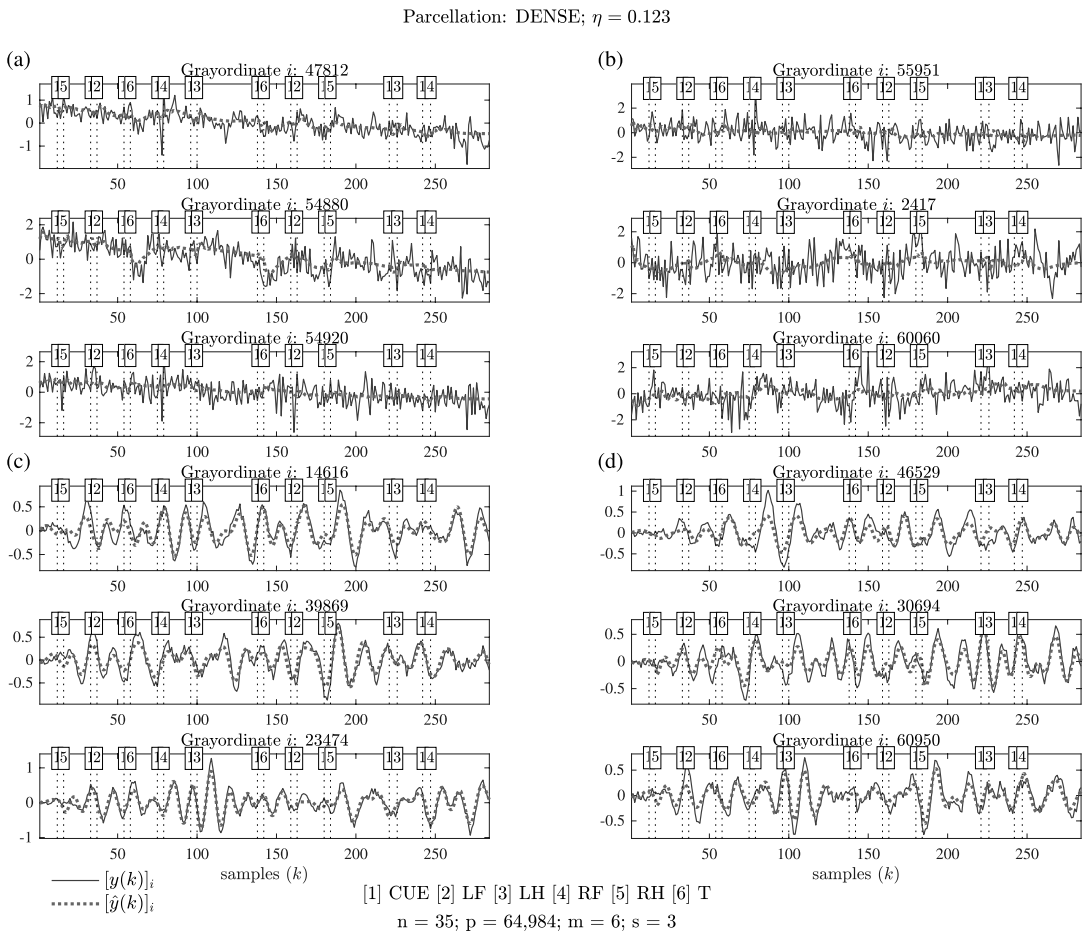

Figure 5. Original and reconstructed output times series for the MOTOR task with the DENSE parcellation, for filters ALL-PASS ((a) and (b)), and BAND-PASS ((c) and (d)). Values are presented for the subject h whose *across-regions* average correlation coefficient (i.e. $\frac{1}{p}\sum_{i=1}^{p}\gamma_i^{(h)}$) was in the median of the *across-subjects* distribution. In (a) and (c), we display values for the reconstructed outputs $\{[y(k)]_i\}_{k=0}^{L-1}$ for regions $i \in \mathcal{I}_{\text{high}} \subset \{1,\ldots,p\}$ with $p = 64\,984$, where the subset $\mathcal{I}_{\text{high}}$ consists of the three regions for which the correlation $\gamma_i^{(h)}$ was *highest*. In (b) and (d), values correspond to the reconstructed outputs $\{[y(k)]_i\}_{k=0}^{L-1}$ for regions $i \in \mathcal{I}_{\text{med}} \subset \{1,\ldots,p\}$, where the subset $\mathcal{I}_{\text{med}}$ consists of the three regions for which the correlation $\gamma_i^{(h)}$ was adjacent to the *median*. The vertical dotted lines across each plot mark the time-of-occurrence of each task condition, with the task condition number being indicated in the small box on top.

to the BOLD signal energy observed at brain region i as a result of the occurrence of a task condition event at input channel j. Alternatively, starting from the system parameters (A, B, C), the $\mathcal{H}_2$ norm associated with the (i,j) output-input pair $\|[H(z)]_{i,j}\|_2$ can be calculated analytically, for stable systems, as follows. Denoting by $B_j := [B]_{:,j}$ (the jth column of the input matrix B), we first compute a solution $W_j \in \mathbb{S}_+^n$ to the discrete-time Lyapunov equation [31, section 12.3]:

$$AW_jA^{\mathsf{T}} - W_j + B_jB_j^{\mathsf{T}} = 0. \tag{19}$$

Given W_j as the solution to (19), and denoting $C_i := [C]_{i,:}$ (the ith row of the output matrix C), we thus have

$$\|[H(z)]_{i,j}\|_2^2 = C_iW_jC_i^{\mathsf{T}}.$$

In figure 7, we display a surface plot projecting the pairwise input–output $\mathcal{H}_2$ norm values for the *dense* time series for all grayordinates $i = 1,\ldots,p$ $(p = 64\,984)$, considering the input j corresponding to the RH condition (i.e. 'tap right finger'). We note that the regions in the neighborhood of the primary motor cortex present a predominantly high $\mathcal{H}_2$ norm value, suggesting a higher engagement of these regions with the RH cognitive input as a task condition [47].

4.4. Multi-subject model identifies common dynamic characteristics across subjects

To evaluate the quality of the approximated output for the multi-subject model identification presented in section 2.5,

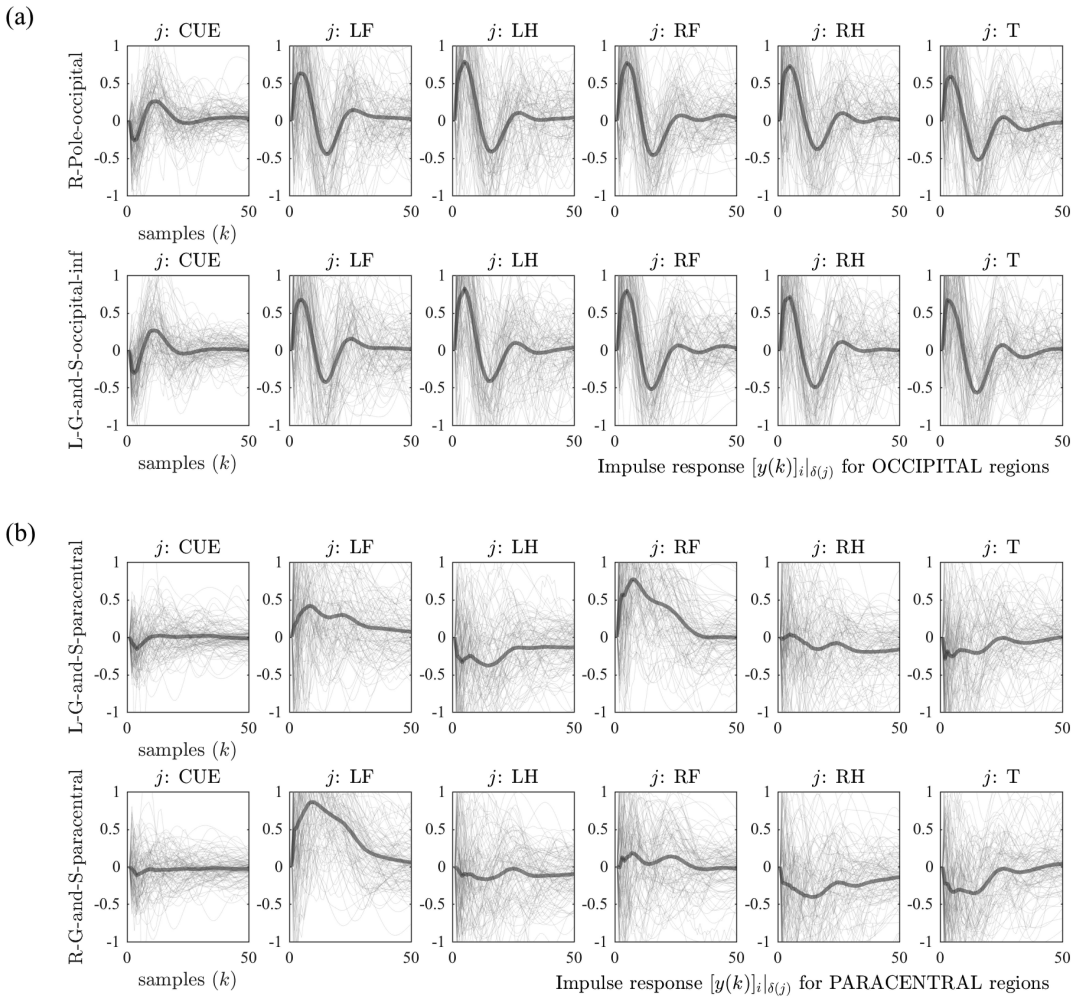

Figure 6. In (a), the *across-subjects* average impulse response function, i.e. $\frac{1}{q}\sum_{h=1}^{q} y^{(h)}(k)\big|_{\delta(j)}$, is presented for all six input channels (in each column), corresponding to the six different task conditions $\{\text{CUE}, \text{LF}, \text{LH}, \text{RF}, \text{RH}, \text{T}\}$, for the ALL-PASS filter. The regions considered (index i, as different rows) are those achieving the highest average correlation $\frac{1}{q}\sum_{h=1}^{q}\gamma_i^{(h)}$ across subjects $h = 1, \ldots, q$. In (b), we display the corresponding average impulse response functions for the left and right paracentral lobule and sulcus regions. One notable aspect is the crossed association between a high intensity response in the *left-hemisphere* region ('L-G-and-S-paracentral') for the 'squeeze right toe' task condition (RF). The corresponding crossed association in the symmetric right-hemisphere region ('R-G-and-S-paracentral' versus LF), is also present.

we consider multiple subject *splits*, i.e. random partitions of the set of subjects. More formally, we partition the set of subjects $\mathcal{S} = \{1, \ldots, q\}$ into training and testing subsets, as follows. We define $\ell = 1, \ldots, \ell_{\text{MAX}}$ *training subject splits* $\hat{\mathcal{S}}^{(\ell)} \subset \mathcal{S}$, where each subject split $\hat{\mathcal{S}}^{(\ell)}$ is obtained by uniformly sampling a subset of $\hat{q} < q$ subjects from the total set of individuals $\mathcal{S} = \{1, \ldots, q\}$, without replacement. For each training subject split, we define a corresponding *training data split* $\hat{\mathcal{D}}^{(\ell)} = \{y^{(\ell)}(k), u^{(\ell)}(k)\}_{\ell \in \hat{\mathcal{S}}^{(\ell)}}$ containing all time samples $k = 1, \ldots, L-1$ of the inputs and outputs, for each subject in that training subject split. Similarly, we define the $\ell = 1 \ldots, \ell_{\text{MAX}}$ *testing subject splits* $\tilde{\mathcal{S}}^{(\ell)} = \mathcal{S} \setminus \hat{\mathcal{S}}^{(\ell)}$, as well as their associated data-splits $\tilde{\mathcal{D}}^{(\ell)}$ containing $\tilde{q} = q - \hat{q}$ testing subjects per split, with corresponding input and output samples.

For each training subject split $\hat{\mathcal{S}}^{(\ell)}$, we find one *common* model $\{\hat{A}_T^{(\ell)}, \hat{B}_T^{(\ell)}, \hat{C}_T^{(\ell)}\}$ by applying the procedure described in section 2.5 to the training data split $\hat{\mathcal{D}}^{(\ell)}$. Then, for each subject $h \in \hat{\mathcal{S}}^{(\ell)}$, we produce the estimated training outputs $\hat{y}_\ell^{(h)} \equiv \{\hat{y}^{(h)}(k)\}_{k=0}^{L-1}$, using the inputs $u^{(h)} \in \hat{\mathcal{D}}^{(\ell)}$ and the initial state $\hat{x}_T^{(h)}(0)$. We then compare $\hat{y}_\ell^{(h)}$ with the original outputs $y_\ell^{(h)} \equiv \{y^{(h)}(k)\}_{k=0}^{L-1}$ to compute the per subject, per-region *training* approximation correlation $\hat{\gamma}_i^{(h)}(\ell) := \langle [y_\ell^{(h)}]_i, [\hat{y}_\ell^{(h)}]_i \rangle$ in that split. Likewise, we use the same common model $(\hat{A}_T^{(\ell)}, \hat{B}_T^{(\ell)}, \hat{C}_T^{(\ell)})$ to compute the estimated testing outputs $\tilde{y}_\ell^{(h)} \equiv \{\tilde{y}^{(h)}(k)\}_{k=0}^{L-1}$ for $h \in \tilde{\mathcal{S}}^{(\ell)}$, and compare with the original outputs $y_\ell^{(h)}$, to compute the *testing* approximation correlation $\tilde{\gamma}_i^{(h)}(\ell) := \langle [y_\ell^{(h)}]_i, [\tilde{y}_\ell^{(h)}]_i \rangle$ in that split.

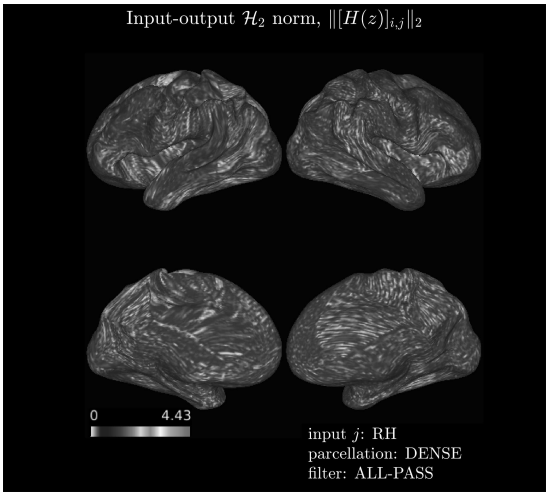

Figure 7. We display a projection of the input–output $\mathcal{H}_2$ system norm $\|[H(z)]_{i,j}\|_2$ onto the cortical surface for the DENSE parcellation and ALL-PASS filter. The jth input corresponds to the RH task condition, i.e. 'tap right finger', where each output i corresponds to a surface grayordinate, i.e. $i = 1, \ldots, p$, with $p = 64\,984$.

The results obtained by applying these definitions are illustrated in figure 8. We consider the state dimension parameter $n = 35$, and a number of total splits $\ell_{\text{MAX}} = 3$, defined over a base set of $q = 100$ total subjects and having $\hat{q} = 30$ and $\tilde{q} = 70$ training and testing subjects, respectively. More specifically, we summarize the distribution of *per-region* training correlation values across all splits $\bigcup_\ell \hat{\gamma}_i^{(h)}(\ell)$, for the regions whose *across-subjects* average correlation $\frac{1}{\ell_{\text{MAX}}\tilde{q}} \sum_{\ell=1}^{\ell_{\text{MAX}}} \sum_{h \in \hat{\mathcal{S}}^{(\ell)}} \hat{\gamma}_i^{(h)}(\ell)$ was highest. We can argue that the multi-subject models capture common dynamic characteristics across individuals, since the in-sample and out-of-sample correlations achieved comparable values. Furthermore, the regions with the highest correlations were the ones associated with the occipital areas (implicated in visual processing), as was the case for the previous experiments. Finally, to compare its approximating capacity with the single-subject model, we can derive the multi-subject degrees of freedom ratio

$$\eta_{\text{multi}} := \frac{n(n+m+p)}{\hat{q}L(p+m)},$$

for which, with $n = 35$, we have $\eta_{\text{multi}} \approx 0.003$, a significantly lower value.

4.5. Probabilistic inversion of the identified stochastic models enables input estimation

If the stochastic component of the identified models is considered (see description in *supplementary methods*), we may perform a probabilistic inversion of the input–output relationship that they capture. In particular, we may derive inverted models that are able to express the probability of occurrence

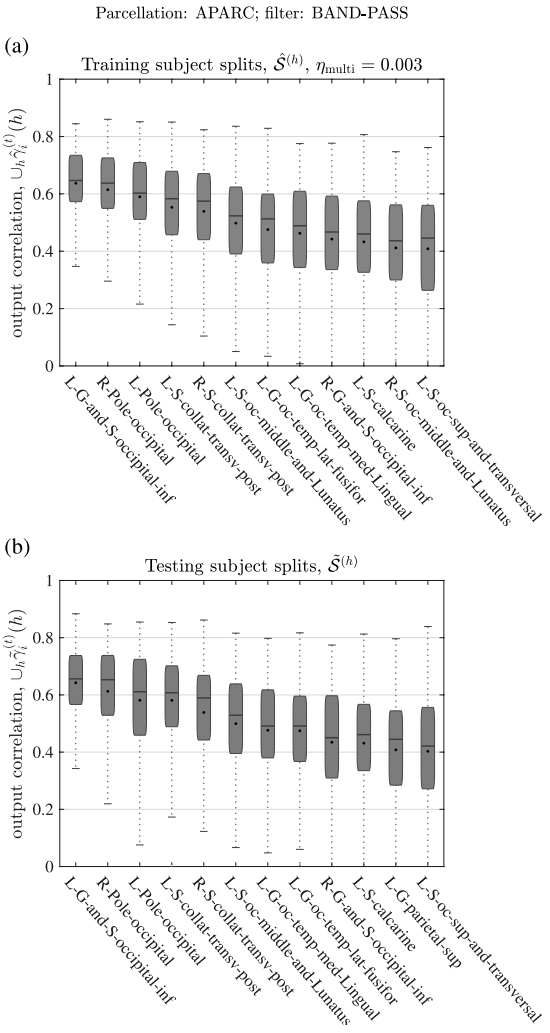

Figure 8. In (a), we present the *across-subjects* distribution of *training* correlations $\cup_h \hat{\gamma}_i^{(h)}$ for the multi-subject model, obtained for $\ell_{\text{MAX}} = 3$ data splits. The distribution of the correlation is presented for the ten regions whose *across-subjects* average correlation was highest. In (b), the corresponding *testing* correlations $\cup_h \tilde{\gamma}_i^{(h)}$ are presented.

of specific input values given output observations. Taking inspiration in [48] and seeking greater clarity of exposition, we derive a first-principles joint state-input estimation formulation based on maximum likelihood estimation criteria [49]. For that purpose, we resort to the full stochastic subspace identification algorithm (an extension to the algorithm presented in section 2.3), whereby estimates for the covariance matrices Q, R, and S (and initial state mean m_0 and covariance Q_0) are produced (see also algorithm 4.3 in [21] or section 9.6 in [20]). Given estimates for the full parametrization $\Theta = (A, B, C, Q, R, S, m_0, Q_0)$, the system model in (1) provides us with the elements to express the joint density $p(y, x|u, \Theta)$ of the states $x := \{x(k)\}_{k=0}^{L-1}$ and outputs

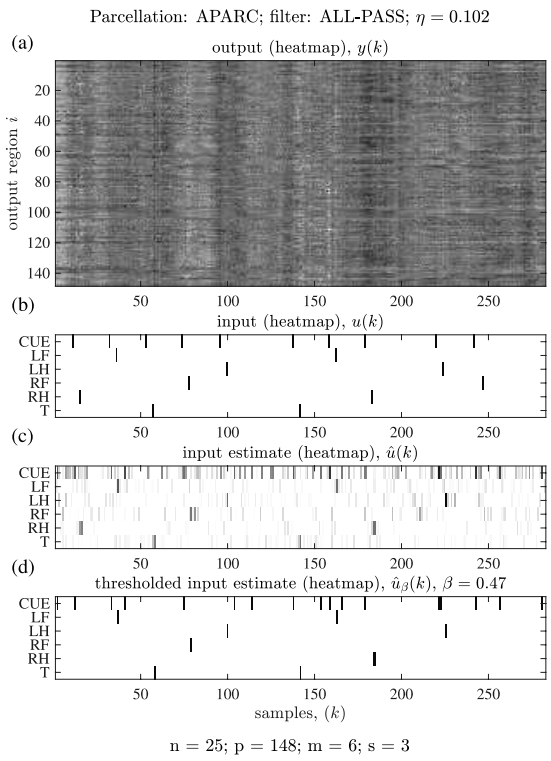

Figure 9. Joint input and state estimation through probabilistic inversion. In (a), we present the original outputs $y(k)$; in (b), the original inputs $u(k)$; and in (c), the recovered input estimates $\hat{u}$. In (d), we display the *thresholded* input estimates, obtained by setting $[\hat{u}_\beta(k)]_j = 1$ if $[\hat{u}(k)]_j \geqslant \beta$ and $[\hat{u}_\beta(k)]_j = 0$ otherwise.

$y := \{y(k)\}_{k=0}^{L-1}$ as a function of the inputs $u := \{u(k)\}_{k=0}^{L-1}$. Therefore, to perform the intended probabilistic inversion, we apply Bayes' theorem to express

$$p(u, x|y, \Theta) = \frac{p(y, x|u, \Theta)p(u, \Theta)}{p(y, \Theta)}, \tag{20}$$

and, as an estimation criterion, look for estimates $\hat{x}$ and $\hat{u}$ such that

$$\{\hat{u}, \hat{x}\} := \arg\max_{u,x} p(u, x|y, \Theta).$$

We now proceed to find an explicit form for the likelihood function $p(y, x|u, \Theta)$ required on the r.h.s of (20). For compactness, we denote $x_k \equiv x(k)$, $y_k \equiv y(k)$, $u_k \equiv u(k)$, and the system and noise matrices as

$$\Upsilon := \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \text{ and } \Psi := \begin{bmatrix} Q & S \\ S^{\mathsf{T}} & R \end{bmatrix},$$

with $\Upsilon \in \mathbb{R}^{(n+p)\times(n+m)}$ and $\Psi \in \mathbb{S}_{++}^{n+p}$. Further, we assume a prior distribution over the first observation, i.e. $x_0 \sim \mathcal{N}(m_0, Q_0)$. Importantly, we note that the set of equation (1) establish a Markovian property that allows us to write

$$p(y, x|u, \Theta) = p(x_0|m_0, Q_0)\prod_{k=0}^{L-1} p\left(\begin{bmatrix} x_{k+1} \\ y_k \end{bmatrix} \middle| \begin{bmatrix} x_k \\ u_k \end{bmatrix}, \Upsilon, \Psi\right),$$

given its sequential dependence structure. By the process and noise models assumed in (1) and (2), this relationship is equivalently described, for each sample k, by a normal distribution

$$\begin{bmatrix} x_{k+1} \\ y_k \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} A & B \\ C & 0 \end{bmatrix}\begin{bmatrix} x_k \\ u_k \end{bmatrix}, \begin{bmatrix} Q & S \\ S^{\mathsf{T}} & R \end{bmatrix}\right) = \mathcal{N}\left(\Upsilon\begin{bmatrix} x_k \\ u_k \end{bmatrix}, \Psi\right)$$

having a log-likelihood given by

$$\log p\left(\begin{bmatrix} x_{k+1} \\ y_k \end{bmatrix} \middle| \begin{bmatrix} x_k \\ u_k \end{bmatrix}, \Upsilon, \Psi\right) = -\frac{n+p}{2}\log 2\pi - \frac{1}{2}\log\det\Psi - \frac{1}{2}\left\|\begin{bmatrix} x_{k+1} \\ y_k \end{bmatrix} - \Upsilon\begin{bmatrix} x_k \\ u_k \end{bmatrix}\right\|_{\Psi^{-1}}^2.$$

By expressing the log-likelihood of the initial state and combining the sum of the log-likelihood over the successive time samples, we have that the complete log-likelihood can be written as

$$\begin{aligned} \log p(u, x|y, \Theta) &\propto -\frac{1}{2}\log\det Q_0 - \frac{1}{2}\|x_0 - m_0\|_{Q_0^{-1}}^2 \\ &- \frac{1}{2}\sum_{k=1}^{N}\log\det\Psi - \frac{1}{2}\sum_{k=0}^{L-1}\left\|\begin{bmatrix} x_{k+1} \\ y_k \end{bmatrix} - \Upsilon\begin{bmatrix} x_k \\ u_k \end{bmatrix}\right\|_{\Psi^{-1}}^2 \\ &=: \mathcal{L}(u, x|y, \Theta). \end{aligned} \tag{21}$$

For simplicity, we assume a flat prior probability for u and discard the probability term on the denominator of (20) (since it is independent of u and x). The problem of finding maximum likelihood estimates for the inputs $\hat{u} = \{\hat{u}_k\}_{k=1}^N$ and state $\hat{x} = \{\hat{x}_k\}_{k=1}^N$ can thus be expressed as

$$\{\hat{u}, \hat{x}\} := \arg\max_{u,x} \mathcal{L}(u, x|y, \Theta),$$

which, by (21), is a convex quadratic optimization problem in u and x.

In figure 9, we present the estimation results for $\hat{u}$, obtained when the above estimation criterion is applied. We consider a model identified on data from a subject h whose average estimated output correlation $\langle \hat{y}^{(h)}, y^{(h)} \rangle$ was in the median of the distribution over $h = 1, \ldots, q$, for the population of $q = 100$ subjects. It can be seen that, although the input estimate presents a significant noise component, by applying a fixed threshold to the magnitude of the signal, the binary content of the original signal can be recovered with reasonable accuracy.

5. Conclusion

We have proposed the use of dynamical input–output models for discrete linear state-space systems as a framework to analyze task-fMRI times series. Such input–output relationships are based on a specific encoding of task-related inputs that uses information about the time-of-occurrence of task conditions.

In addition, we provided means to numerically characterize this relationship by a set of matrix parameters, whose estimation algorithm relies on subspace identification methods. We used a comprehensive dataset comprising time series from multiple subjects, whose spatial configuration followed both a regional parcellation and a spatially dense representation. In the latter configuration, the estimation of the parameters becomes a large-scale optimization problem, for which we proposed a numerically efficient algorithm.

One of the advantages of the state representation of a dynamical system model is that it enables rich analyses of the system's behavior. One example is the generation of impulse responses, which can be associated with the hemodynamic response functions and are readily computable from the system representation. Another example is the calculation of input–output system norms, which provide a principled manner to quantify the dynamic effect of task-related inputs on the intensity of the BOLD response at different regions of the cortex.

Given the dynamic nature of these models, we expect that our approach might have an impact in the emerging field of neurofeedback. In the scenario where clinical stimulation of the subject's brain is studied (e.g. transcranial magnetic stimulation), our method might be used to measure and compare the effects of potentially induced modifications in the BOLD dynamics. We also envision an application where our method is applied in real time, where both the system dynamics and the underlying state are measured and used to guide interventions and assess treatment efficacy.

Acknowledgments

This work was supported by the National Science Foundation, grants CRCNS BCS-1441502, CAREER PHY-1554488, IIS-1447470, and CAREER-ECCS-1651433, and by CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil. Data were provided (in part) by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

ORCID iDs

Cassiano O Becker https://orcid.org/0000-0003-4280-7411
Danielle S Bassett https://orcid.org/0000-0002-6183-4493
Victor M Preciado https://orcid.org/0000-0001-9998-8730

References

[1] Kailath T 1980 *Linear Systems* (Englewood Cliffs, NJ: Prentice-Hall)
[2] Khalil H K 2002 *Nonlinear Systems* 3rd edn (Upper Saddle River, NJ: Prentice-Hall)
[3] Ogawa S, Lee T M, Nayak A S and Glynn P 1990 Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields *Magn. Reson. Med.* **14** 68–78
[4] Huettel S A, Song A W and McCarthy G 2014 *Functional Magnetic Resonance Imaging* 3rd edn (Sunderland, MA: Sinauer Associates)
[5] Logothetis N K, Pauls J, Augath M, Trinath T and Oeltermann A 2001 Neurophysiological investigation of the basis of the fMRI signal *Nature* **412** 150–7
[6] Buxton R B, Uludağ K, Dubowitz D J and Liu T T 2004 Modeling the hemodynamic response to brain activation *NeuroImage* **23** S220–33
[7] Buckner R L 1998 Event-related fMRI and the hemodynamic response *Hum. Brain Mapp.* **6** 373–7
[8] Boynton G M, Engel S A, Glover G H and Heeger D J 1996 Linear systems analysis of functional magnetic resonance imaging in human V1 *J. Neurosci.* **16** 4207–21
[9] Dale A M and Buckner R L 1997 Selective averaging of rapidly presented individual trials using fMRI *Hum. Brain Mapp.* **5** 329–40
[10] Birn R M and Bandettini P A 2005 The effect of stimulus duty cycle and 'off' duration on BOLD response linearity *NeuroImage* **27** 70–82
[11] Boynton G M, Engel S A and Heeger D J 2012 Linear systems analysis of the fMRI signal *NeuroImage* **62** 975–84
[12] Toyoda H, Kashikura K, Okada T, Nakashita S, Honda M, Yonekura Y, Kawaguchi H, Maki A and Sadato N 2008 Source of nonlinearity of the BOLD response revealed by simultaneous fMRI and NIRS *NeuroImage* **39** 997–1013
[13] de Zwart J A, van Gelderen P, Jansma J M, Fukunaga M, Bianciardi M and Duyn J H 2009 Hemodynamic nonlinearities affect BOLD fMRI response timing and amplitude *NeuroImage* **47** 1649–58
[14] Liu Z, Rios C, Zhang N, Yang L, Chen W and He B 2010 Linear and nonlinear relationships between visual stimuli, EEG and BOLD fMRI signals *NeuroImage* **50** 1054–66
[15] Friston K J, Harrison L and Penny W 2003 Dynamic causal modelling *NeuroImage* **19** 1273–302
[16] Harrison L, Penny W D and Friston K 2003 Multivariate autoregressive modeling of fMRI time series *NeuroImage* **19** 1477–91
[17] Rogers B P, Katwal S B, Morgan V L, Asplund C L and Gore J C 2010 Functional MRI and multivariate autoregressive models *Magnetic Resonance Imaging* **28** 1058–65
[18] Samdin S B, Ting C M, Ombao H and Salleh S H 2017 A unified estimation framework for state-related changes in effective brain connectivity *IEEE Trans. Biomed. Eng.* **64** 844–58
[19] Ljung L 1999 *System Identification: Theory for the User* (New York: Prentice Hall)
[20] Verhaegen M and Verdult V 2007 *Filtering and System Identification: a Least Squares Approach* (Cambridge: Cambridge University Press)
[21] Van Overschee P and De Moor B 1996 *Subspace Identification for Linear Systems: Theory-Implementation-Applications* (Dordrecht: Kluwer)
[22] Katayama T 2006 *Subspace Methods for System Identification* (Berlin: Springer)
[23] Tauchmanova J and Hromcik M 2008 Subspace identification methods and fMRI analysis *30th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society* (IEEE) pp 4427–30
[24] Nováková J, Hromčík M and Jech R 2012 Dynamic causal modeling and subspace identification methods *Biomed. Signal Process. Control* **7** 365–70

[25] Bakhtiari S K and Hossein-Zadeh G A 2012 Subspace-based identification algorithm for characterizing causal networks in resting brain *NeuroImage* **60** 1236–49
[26] Van Essen D C *et al* 2012 The human connectome project: A data acquisition perspective *NeuroImage* **62** 2222–31
[27] Aguirre G K, Zarahn E and D'esposito M 1998 The variability of human, BOLD hemodynamic responses *NeuroImage* **8** 360–9
[28] Weiskopf N 2012 Real-time fMRI and its application to neurofeedback *NeuroImage* **62** 682–92
[29] Sitaram R *et al* 2016 Closed-loop brain training: the science of neurofeedback *Nat. Rev. Neurosci.* **18** 86–100
[30] Bassett D S and Khambhati A N 2017 A network engineering perspective on probing and perturbing cognition with neurofeedback *Ann. New York Acad. Sci.* **1396** 126–43
[31] Hespanha J P 2009 *Linear Systems Theory* (Princeton, NJ: Princeton University Press)
[32] Miller D N and De Callafon R A 2013 Subspace identification with eigenvalue constraints *Automatica* **49** 2468–73
[33] Halko N, Martinsson P G and Tropp J A 2011 Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions *SIAM Rev.* **53** 217–88
[34] Wang S 2015 A practical guide to randomized matrix computations with MATLAB implementations (arXiv:1505.07570)
[35] Parikh N *et al* 2014 Proximal algorithms *Found. Trends® Optim.* **1** 127–239
[36] Van Essen D C *et al* 2013 The WU-Minn human connectome project: an overview *NeuroImage* **80** 62–79
[37] Glasser M F *et al* 2013 The minimal preprocessing pipelines for the human connectome project *NeuroImage* **80** 105–24
[38] Washington University in St. Louis-University of Minnesota (WU-Minn) Consortium 2015 Human Connectome Project 900 Subjects Data Release: Reference Manual (www.humanconnectome.org/study/hcp-young-adult/document/900-subjects-data-release, retrieved 20 August 2018)
[39] Destrieux C, Fischl B, Dale A and Halgren E 2010 Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature *NeuroImage* **53** 1–15
[40] Barch D M *et al* 2013 Function in the human connectome: Task-fMRI and individual differences in behavior *NeuroImage* **80** 169–89
[41] Yeo B T *et al* 2011 The organization of the human cerebral cortex estimated by intrinsic functional connectivity *J. Neurophysiol.* **106** 1125–65
[42] Bassett D S, Wymbs N F, Porter M A, Mucha P J, Carlson J M and Grafton S T 2011 Dynamic reconfiguration of human brain networks during learning *Proc. Natl Acad. Sci.* **108** 7641–6
[43] Wander J D, Blakely T, Miller K J, Weaver K E, Johnson L A, Olson J D, Fetz E E, Rao R P and Ojemann J G 2013 Distributed cortical adaptation during learning of a brain–computer interface task *Proc. Natl Acad. Sci.* **110** 10818–23
[44] Buneo C A and Andersen R A 2006 The posterior parietal cortex: Sensorimotor interface for the planning and online control of visually guided movements *Neuropsychologia* **44** 2594–606
[45] Shabbott B A and Sainburg R L 2008 Differentiating between two models of motor lateralization *J. Neurophysiol.* **100** 565–75
[46] Mutha P K, Haaland K Y and Sainburg R L 2013 Rethinking motor lateralization: specialized but complementary mechanisms for motor control of each arm *PloS One* **8** e58582
[47] Kandel E R *et al* 2000 *Principles of Neural Science* 4th edn (New York: McGraw-Hill)
[48] Gillijns S and De Moor B 2007 Unbiased minimum-variance input and state estimation for linear discrete-time systems *Automatica* **43** 111–6
[49] Casella G and Berger R L 2002 *Statistical Inference* vol 2 (Pacific Grove, CA: Duxbury)