Computational Statistics and Data Analysis 139 (2019) 164–177

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Regularized joint estimation of related vector autoregressive models

A. Skripnikov *, G. Michailidis

Department of Statistics, University of Florida, 102 Griffin-Floyd Hall, P.O. Box 118545, Gainesville, FL 32611, USA

ARTICLE INFO

Article history:
Received 24 October 2018
Received in revised form 27 March 2019
Accepted 14 May 2019
Available online 22 May 2019

Keywords:
Attention deficit hyperactivity disorder
Group lasso
Regularized estimation
Resting-state fMRI
Stability selection
Vector autoregression

ABSTRACT

In a number of applications, one has access to high-dimensional time series data on several related subjects. A motivating application area comes from the neuroimaging field, such as brain fMRI time series data, obtained from various groups of subjects (cases/controls) with a specific neurological disorder. The problem of regularized joint estimation of multiple related Vector Autoregressive (VAR) models is discussed, leveraging a group lasso penalty in addition to a regular lasso one, so as to increase statistical efficiency of the estimates by borrowing strength across the models. A modeling framework is developed that it allows for both group-level and subject-specific effects for related subjects, using a group lasso penalty to estimate the former. An estimation procedure is introduced, whose performance is illustrated on synthetic data and compared to other state-of-the-art methods. Moreover, the proposed approach is employed for the analysis of resting state fMRI data. In particular, a group-level descriptive analysis is conducted for brain inter-regional temporal effects of Attention Deficit Hyperactive Disorder (ADHD) patients as opposed to controls, with the data available from the ADHD-200 Global Competition repository.

Published by Elsevier B.V.

1. Introduction

With recent advances in technology and growing amounts of available data (e.g. click-generated web browsing data, social networks, image and video data), there is a strong interest in modeling and analysis of high-dimensional time series data. Application areas include gene regulatory network inference (Michailidis and d'Alché Buc, 2013), brain fMRI data (Song et al., 2011), macroeconomic time series forecasting and structural analysis (Bańbura et al., 2010; Lin and Michailidis, 2017), to name a few. Their common characteristic is the large number of variable relationships being analyzed relative to the time points available, thus leading to a high-dimensional statistical estimation problem. In many cases, the temporal dynamics of the data under consideration are well captured by autoregressive models, and hence the use of vector autoregressive models (VAR) enables the modeling of temporal dependencies between the variables. However, in the presence of a large number of parameters to estimate, and only a few time points, one needs to incorporate appropriate sparsity assumptions into the VAR modeling framework. To enforce sparsity, it is common to employ an ℓ_1 lasso penalty (Basu et al., 2015a,b), with theoretical properties established in the former paper.

As previously mentioned, in many applications, on top of the typical high-dimensional setting, one also has to perform estimation of time series across a moderate to large number of related subjects. As a motivating example, the area

* Correspondence to: University of Houston, College of Natural Sciences and Mathematics, Science & Research Building 1, 3507 Cullen Blvd, Room 214, Houston, TX, 77054, USA.
E-mail address: andreysk@math.uh.edu (A. Skripnikov).

https://doi.org/10.1016/j.csda.2019.05.007
0167-9473/Published by Elsevier B.V.

of medical research brings about a lot of experimental settings with multiple subjects being monitored over time, e.g. a collection of fMRI time series data for a group of patients. Looking at patients with a particular disease/disorder (e.g. Alzheimer's), it is expected that connectivity across brain regions exhibits common structural patterns. However, it has been well documented that, albeit sharing the same disorder, patients still exhibit individual patterns (Woolrich et al., 2004; Beckmann et al., 2003), which leads to subject heterogeneity.

Standard analysis pipelines for fMRI time series data typically involve estimating a network for each subject separately, with subsequent accumulation of the estimates for further group-level analysis (Narayan and Allen, 2016). That approach has been applied to study brain activity for Alzheimer's disease (Huang et al., 2010), autism (Narayan et al., 2015), Parkinson's disease (Liu et al., 2014), among others. While providing tools for group-level analysis, this approach does not incorporate the underlying assumption of similarity across subjects within the same group (e.g. cases or controls) into the estimation procedure.

A key contribution of this work is the development of a joint modeling framework that would enforce the similarity assumption into the estimation procedure, while also enabling estimation of subject-specific network effects. Proposed method increases the effective sample size for common structure estimation by borrowing strength across related time series. Additionally, the framework permits detection of most pronounced individual effects for each subject (if present). The problem of joint estimation has received attention in the literature recently, primarily focusing on the estimation of multiple graphical models. Those approaches leverage various penalties that encourage both sparsity and joint estimation of the parameters across the models; see the hierarchical penalty used in Guo et al. (2011) or fused lasso penalty in Danaher et al. (2014) and Ma and Michailidis (2016). In this work, we employ a group lasso penalty due to its ability to clearly identify a common structure across multiple subjects, while letting the magnitudes of effects vary. After having detected the common structure, a standard lasso procedure will be applied to obtain sparse estimates of subject-specific effects.

While the introduced joint estimation procedure can be used in other time series settings (economic data for cities or states with shared manufacturing and industrial features, gene expression data for matched patients, sales data for similar stores), the primary motivating application is resting-state fMRI time series data for studying the spontaneous brain temporal dynamics of various cognitive disorders. The literature previously cited on group-level inference for cognitive disorders focused on estimating functional brain connectivity (studying correlations between brain region signals) rather than inferring lead–lag relations between the brain regions. The resting-state data represent monotone within-brain fluctuations, and VAR models are suitable for capturing the temporal dynamics in such data. Further, we assume each patient's VAR model to be a perturbation of some common underlying VAR model for the group under consideration (be it subjects with disease or healthy controls), the structure of which will be estimated by our joint procedure based on a group lasso penalty. Note that certain objections have been raised in the literature on the use of VAR models for neuroimaging data (Cole et al., 2010; Ramsey et al., 2010). Nevertheless, we attempt to address some of those issues in our ADHD study in Section 4. For example, one of the six problems with the use of VAR models in application to fMRI data discussed in Ramsey et al. (2010) concerns the varying signal strength across the brain regions from multiple study participants, even though they all share a common abstract processing structure. In our framework, this issue is addressed directly via a group lasso penalty that, while enforcing a common structure, allows for variability in magnitudes of common effects. It is also worth mentioning that a joint estimation approach via regularizing penalties had been used before for brain fMRI time series data (Belilovsky et al., 2016; Chu et al., 2015), but only for functional connectivity estimation, while we apply it to infer lead–lag relations.

To introduce the joint modeling framework, consider a p-variate stationary time series $X_t^{(k)} = (x_{1,t}^{(k)}, \ldots, x_{p,t}^{(k)})^\intercal$, $t = 1, \ldots, T$, $k = 1, \ldots, K$, for K related subjects. VAR model with lag order D, or $VAR(D)$, is given by

$$X_t^{(k)} = A_1^{(k)} X_{t-1}^{(k)} + \cdots + A_D^{(k)} X_{t-D}^{(k)} + \epsilon_t^{(k)},\ \epsilon_t^{(k)} \sim N(\mathbf{0}, \sigma_{(k)}^2 \mathbf{I}_p), \tag{1}$$

$$t = D, \ldots, T, \quad k = 1, \ldots, K,$$

where $A_d^{(k)}$ is a $p \times p$ transition matrix that captures temporal effects of order d between the p variables for subject k, $d = 1, \ldots, D$, $k = 1, \ldots, K$. We further assume a diagonal error covariance matrix $\Sigma_k = \sigma_{(k)}^2 \mathbf{I}_p$, which allows us to break problem (1) into p simpler sub-problems that can be solved in parallel. In this work, we focus on the case of VAR model with lag order one ($D = 1$), so as to emphasize studying the properties of the joint estimation procedure rather than the aspects of lag order selection. The joint estimation approach starts with the assumption of common and individual components for each VAR model: $A_d^{(k)} = A_{d,C}^{(k)} + A_{d,I}^{(k)}$, $d = 1, \ldots, D$, $k = 1, \ldots, K$. Afterwards, an iterative two-stage estimation algorithm is proposed, consisting of a group lasso optimization procedure to jointly estimate the common components $\{A_{d,C}^{(k)}\}$ of the K subjects during the first stage, followed by a sparse lasso optimization procedure to estimate the individual components $\{A_{d,I}^{(k)}\}$ at stage two. The group lasso penalty effectively groups the respective elements of the transition matrices across all K subjects and either retains or excludes the whole group from the model, which guarantees a shared structure of the resulting common component estimates. Meanwhile, the residuals from the common component signal are used as data to estimate the individual structures, representing subject-specific effects, via standard lasso optimization.

The remainder of the paper is organized as follows: Section 2 describes the joint modeling framework and introduces the two-stage estimation procedure, Section 3 demonstrates the simulation study results of the joint estimation procedure

for various settings and compares its performance with other state-of-the-art methods, Section 4 provides substantial empirical application of the introduced method to resting-state fMRI data for ADHD study, while Section 5 contains concluding remarks and discussions of future work.

2. Problem formulation

We start by writing the posited VAR model (1) in standard regression form. First, we drop k from the notation, $k = 1, \ldots, K$, and show the sequence of required algebraic transformations for a single VAR(D) model:

$$X_t = A_1 X_{t-1} + \cdots + A_D X_{t-D} + \epsilon_t, \ \epsilon_t \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_p), \ t = D, \ldots, T. \tag{2}$$

The assumption on the error covariance being diagonal, $\text{cov}(\epsilon_t) = \sigma^2 \mathbf{I}_p$ for $\epsilon_t = (\epsilon_{1,t}, \ldots, \epsilon_{p,t})^\intercal$, $t = D, \ldots, T$, allows us to represent the temporal dynamics for each of the p variables as the following system of equations:

$$x_{j,t} = \sum_{l=1}^{p} (A_1[j, l]\, x_{l,t-1} + \cdots + A_D[j, l]\, x_{l,t-D}) + \epsilon_{j,t}, \ \epsilon_{j,t} \sim N(0, \sigma^2), \tag{3}$$

$$t = D, \ldots, T, \ j = 1, \ldots, p,$$

where $A_d[j, l]$ is order-d temporal effect of lth variable on jth, $l, j = 1, \ldots, p$. If we let $A_d[j, .] = (A_d[j, 1], \ldots, A_d[j, p])^\top$, $d = 1, \ldots, D$, then all $T - D + 1$ equations from (3) can be represented in a compact matrix form for each variable j respectively:

$$\underbrace{\begin{pmatrix} x_{j,T} \\ \cdots \\ x_{j,D} \end{pmatrix}}_{\tilde{X}_j} = \underbrace{\left(\begin{array}{ccc|c|ccc} x_{1,T-1} & \cdots & x_{p,T-1} & \cdots & x_{1,T-D} & \cdots & x_{p,T-D} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{1,D-1} & \cdots & x_{p,D-1} & \cdots & x_{1,0} & \cdots & x_{p,0} \end{array}\right)}_{\mathbf{Z}} \underbrace{\begin{pmatrix} A_1[j, .] \\ \cdots \\ A_D[j, .] \end{pmatrix}}_{\mathbf{A}[j,.]} + \underbrace{\begin{pmatrix} \epsilon_{j,T} \\ \cdots \\ \epsilon_{j,D} \end{pmatrix}}_{\tilde{\epsilon}_j},$$

$$\underbrace{\tilde{X}_j}_{(T-D+1)\times 1} = \underbrace{\mathbf{Z}}_{(T-D+1)\times(Dp)} \underbrace{\mathbf{A}[j, .]}_{(Dp)\times 1} + \underbrace{\tilde{\epsilon}_j}_{(T-D+1)\times 1}, \quad \tilde{\epsilon}_j \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{T-D+1}), \ j = 1, \ldots, p. \tag{4}$$

Next, reintroducing k, $k = 1, \ldots, K$, back into the notation and using the standard regression representation (4) for all K VAR models under consideration, we obtain (for $j = 1, \ldots, p$):

$$\begin{cases} \tilde{X}_j^{(1)} = \mathbf{Z}^{(1)}\, \mathbf{A}^{(1)}[j, .] + \tilde{\epsilon}_j^{(1)}, \quad \tilde{\epsilon}_j^{(1)} \sim N(\mathbf{0}, \sigma_{(1)}^2 \mathbf{I}_{T-D+1}), \\ \ldots \\ \tilde{X}_j^{(K)} = \mathbf{Z}^{(K)}\, \mathbf{A}^{(K)}[j, .] + \tilde{\epsilon}_j^{(K)}, \quad \tilde{\epsilon}_j^{(K)} \sim N(\mathbf{0}, \sigma_{(K)}^2 \mathbf{I}_{T-D+1}), \end{cases} \tag{5}$$

where $\mathbf{A}^{(k)}[j, .]$ corresponds to the temporal effects (of all D orders) that each of p variables has on the jth variable for the kth subject, $k = 1, \ldots, K$.

2.1. Decomposition into common and idiosyncratic components

A key modeling assumption is that of shared structure across all K subjects. In the model, this is manifested through similar sparsity patterns across the K transition matrices. In addition, due to natural variability among subjects, one has to account for the presence of heterogeneity in the form of certain subject-specific effects. Both of these aspects are captured by decomposing each subject's transition matrix into two parts:

$$A_d^{(k)} = A_{d,C}^{(k)} + A_{d,I}^{(k)}, \ d = 1, \ldots, D, \ k = 1, \ldots, K, \tag{6}$$

where $A_{d,C}^{(k)}$ is the common component of order-d temporal effects for kth subject, while $A_{d,I}^{(k)}$ is the idiosyncratic component. Applying this representation to equations in (5), we get (for $j = 1, \ldots, p$):

$$\begin{cases} \tilde{X}_j^{(1)} = \mathbf{Z}^{(1)} \ (\mathbf{A}_C^{(1)}[j, .] + \mathbf{A}_I^{(1)}[j, .]) + \tilde{\epsilon}_j^{(1)}, \quad \tilde{\epsilon}_j^{(1)} \sim N(\mathbf{0}, \sigma_{(1)}^2 \mathbf{I}_{T-D+1}), \\ \ldots \\ \tilde{X}_j^{(K)} = \mathbf{Z}^{(K)}\, (\mathbf{A}_C^{(K)}[j, .] + \mathbf{A}_I^{(K)}[j, .]) + \tilde{\epsilon}_j^{(K)}, \quad \tilde{\epsilon}_j^{(K)} \sim N(\mathbf{0}, \sigma_{(K)}^2 \mathbf{I}_{T-D+1}). \end{cases} \tag{7}$$

Assuming a VAR model for each subject to be a perturbation of a common underlying VAR model, we proceed to enforce the common support constraint on $\mathbf{A}_C^{(1)}[j, .], \ldots, \mathbf{A}_C^{(K)}[j, .]$, denoted by $\mathbf{A}_C^{(1)}[j, .] \approx \mathbf{A}_C^{(2)}[j, .] \approx \ldots \approx \mathbf{A}_C^{(K)}[j, .]$, $j = 1, \ldots, p$.

More formally, defining support of a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\intercal \in \mathbb{R}^m$ as $support(\boldsymbol{\beta}) = \{i : \beta_i \neq 0\}$, this assumption is equivalent to $support(\mathbf{A}_C^{(1)}[j, .]) \equiv support(\mathbf{A}_C^{(2)}[j, .]) \equiv \ldots \equiv support(\mathbf{A}_C^{(K)}[j, .])$. Moreover, the sparsity assumption is imposed on the individual components $\mathbf{A}_I^{(1)}[j, .], \ldots, \mathbf{A}_I^{(K)}[j, .]$ to recover the most important subject-specific effects. Lastly, for parameter identifiability we assume $\mathbf{A}_C^{(k)}[j, .] \perp \mathbf{A}_I^{(k)}[j, .]$ ($\mathbf{A}_C^{(k)}[j, .]$ is "perpendicular" to $\mathbf{A}_I^{(k)}[j, .]$), implying that the intersection of the supports for $\mathbf{A}_C^{(k)}[j, .]$ and $\mathbf{A}_I^{(k)}[j, .]$ is empty, or, more formally, $support(\mathbf{A}_C^{(k)}[j, .]) \cap support(\mathbf{A}_I^{(k)}[j, .]) \equiv \emptyset$, $k = 1, \ldots, K,\ j = 1, \ldots, p$. The full set of the described constraints for system (7) is outlined below:

$$\begin{cases} \mathbf{A}_C^{(1)}[j, .] \approx \mathbf{A}_C^{(2)}[j, .] \approx \ldots \approx \mathbf{A}_C^{(K)}[j, .], & j = 1, \ldots, p, \\ \mathbf{A}_I^{(k)}[j, .] \text{ - sparse}, & k = 1, \ldots, K, \quad j = 1, \ldots, p, \\ \mathbf{A}_C^{(k)}[j, .] \perp \mathbf{A}_I^{(k)}[j, .], & k = 1, \ldots, K, \quad j = 1, \ldots, p. \end{cases} \tag{8}$$

2.2. A two-stage estimation procedure

To estimate parameters $\mathbf{A}_C^{(k)}[j, .],\ \mathbf{A}_I^{(k)}[j, .],\ k = 1, \ldots, K,\ j = 1, \ldots, p$, presented in (7), while enforcing the set of aforementioned constraints (8), we formulate it as a bi-convex optimization task, which can be solved via an iterative two-stage approach that is always guaranteed to converge to a local minimum (Gorski et al., 2007; Goh et al., 1994). First, we take care of nuisance parameters $\{\sigma_{(k)},\ k = 1, \ldots, K\}$ by plugging in maximum likelihood estimators $\{\hat{\sigma}_{(k)},\ k = 1, \ldots, K\}$ that are calculated as in Lütkepohl (2005) for the system of Eqs. (1). Next, we set $\boldsymbol{\beta}_{j,C} = (\mathbf{A}_C^{(1)}[j, .], \ldots, \mathbf{A}_C^{(K)}[j, .])^\top$, $\boldsymbol{\beta}_{j,I} = (\mathbf{A}_I^{(1)}[j, .], \ldots, \mathbf{A}_I^{(K)}[j, .])^\top$, $\tilde{\mathbf{X}}_j = (\tilde{X}_j^{(1)}, \ldots, \tilde{X}_j^{(K)})^\top$, $j = 1, \ldots, p$, $\hat{D}_{\hat{\sigma}^2} = \text{Diag}(\underbrace{\hat{\sigma}_{(1)}^2, \ldots, \hat{\sigma}_{(1)}^2}_{T-D+1}, \underbrace{\hat{\sigma}_{(2)}^2, \ldots, \hat{\sigma}_{(2)}^2}_{T-D+1}, \ldots, \underbrace{\hat{\sigma}_{(K)}^2, \ldots, \hat{\sigma}_{(K)}^2}_{T-D+1})$.

Note that $\tilde{\mathbf{Z}} \in \mathbb{R}^{K(T-D+1)\times K(Dp)}$ is a block-diagonal matrix, with the kth block equal to $\mathbf{Z}^{(k)} \in \mathbb{R}^{(T-D+1)\times(Dp)}$ from Eqs. (5), $k = 1, \ldots, K$. Moreover, for an arbitrary vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m) \in \mathbb{R}^m$, let us denote $\|\boldsymbol{\beta}\|_2^2 = \sum_{l=1}^m \beta_l^2$, $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{l=1}^m \beta_l^2}$, $\|\boldsymbol{\beta}\|_1 = \sum_{l=1}^m |\beta_l|$. A large optimization problem corresponding to solving Eqs. (7) with constraints (8) is as follows

$$\begin{aligned} &\min_{\boldsymbol{\beta}_{j,C}, \boldsymbol{\beta}_{j,I}} f(\boldsymbol{\beta}_{j,C}, \boldsymbol{\beta}_{j,I}) = \\ &\min_{\boldsymbol{\beta}_{j,C}, \boldsymbol{\beta}_{j,I}} \|\hat{D}_{\hat{\sigma}^2}^{-\frac{1}{2}}(\tilde{\mathbf{X}}_j - \tilde{\mathbf{Z}}[\boldsymbol{\beta}_{j,C} + \boldsymbol{\beta}_{j,I}])\|_2^2 \ + \ \lambda_j^G \sum_{i=1}^p \|(\boldsymbol{\beta}_{j,C}^{(1)}[i], \ldots, \boldsymbol{\beta}_{j,C}^{(K)}[i])^\intercal\|_2 \\ &+ \ \lambda_j^S \|\boldsymbol{\beta}_{j,I}\|_1 \ + \ \lambda^\infty \sum_{i=1}^p \sum_{k=1}^K |\boldsymbol{\beta}_{j,C}^{(k)}[i] \cdot \boldsymbol{\beta}_{j,I}^{(k)}[i]|, \end{aligned} \tag{9}$$

where $\boldsymbol{\beta}_{j,C}^{(k)}[i] = \mathbf{A}_C^{(k)}[j, i],\ \boldsymbol{\beta}_{j,I}^{(k)}[i] = \mathbf{A}_I^{(k)}[j, i],\ k = 1, \ldots, K,\ i = 1, \ldots, p$. Here, each of the constraints from (8) is addressed via a respective penalty term. Group lasso penalty $\lambda_j^G \sum_{i=1}^p \|(\boldsymbol{\beta}_{j,C}^{(1)}[i], \ldots, \boldsymbol{\beta}_{j,C}^{(K)}[i])^T\|_2$, introduced in Yuan and Lin (2006), either shrinks ith element to zero for all K vectors $\boldsymbol{\beta}_{j,C}^{(1)}, \ldots, \boldsymbol{\beta}_{j,C}^{(K)}$, where $\boldsymbol{\beta}_{j,C}^{(k)} = \mathbf{A}_C^{(k)}[j, .]$, or estimates it to be non-zero for all K vectors. This guarantees the identical support across all K common component estimates, $\mathbf{A}_C^{(1)}[j, .] \approx \mathbf{A}_C^{(2)}[j, .] \approx \ldots \approx \mathbf{A}_C^{(K)}[j, .]$. Sparse lasso penalty $\lambda_j^S \|\boldsymbol{\beta}_{j,I}\|_1$, first introduced in Tibshirani (1996), leads to each $\mathbf{A}_I^{(k)}[j, .],\ k = 1, \ldots, K,\ j = 1, \ldots, p$, having only a few non-zero elements, hence being sparse. Penalty term $\lambda^\infty \sum_{i=1}^p \sum_{k=1}^K |\boldsymbol{\beta}_{j,C}^{(k)}[i] \cdot \boldsymbol{\beta}_{j,I}^{(k)}[i]|$ has a tuning parameter value λ^∞ set high enough, so that the Hadamard product of $\boldsymbol{\beta}_{j,C}$ and $\boldsymbol{\beta}_{j,I}$ is equal to 0. Hence, it leads to $\boldsymbol{\beta}_{j,C}^{(k)}[i] \neq 0$ implying $\boldsymbol{\beta}_{j,I}^{(k)}[i] = 0$, and, vice versa, $\boldsymbol{\beta}_{j,I}^{(k)}[i] \neq 0$ implying $\boldsymbol{\beta}_{j,C}^{(k)}[i] = 0$. This guarantees $support(\boldsymbol{\beta}_{j,C}^{(k)}) \cap support(\boldsymbol{\beta}_{j,I}^{(k)}) \equiv \emptyset$, leading to $\mathbf{A}_C^{(k)}[j, .] \perp \mathbf{A}_I^{(k)}[j, .]\ \forall\, j, k,\ j = 1, \ldots, p,\ k = 1, \ldots, K$.

For the joint function $f(\boldsymbol{\beta}_{j,C}, \boldsymbol{\beta}_{j,I})$ being optimized in (9), using definition of convexity and knowledge of operations preserving convexity (Boyd and Vandenberghe, 2004), we may show that: fixing $\boldsymbol{\beta}_{j,I}$ at some value $\hat{\boldsymbol{\beta}}_{j,I}$, $f(\boldsymbol{\beta}_{j,C}, \hat{\boldsymbol{\beta}}_{j,I})$ is a convex function of $\boldsymbol{\beta}_{j,C}$; fixing $\boldsymbol{\beta}_{j,C}$ at some value $\hat{\boldsymbol{\beta}}_{j,C}$, $f(\hat{\boldsymbol{\beta}}_{j,C}, \boldsymbol{\beta}_{j,I})$ is a convex function of $\boldsymbol{\beta}_{j,I}$. By definition (Gorski et al., 2007), it leads to $f(\boldsymbol{\beta}_{j,C}, \boldsymbol{\beta}_{j,I})$ constituting a biconvex function of its arguments $\boldsymbol{\beta}_{j,C}$ and $\boldsymbol{\beta}_{j,I}$. Below we present a two-stage algorithm performing an alternate convex search method (Gorski et al., 2007; Wendell and Hurter Jr, 1976), where, for a general biconvex function $f(\mathbf{x}, \mathbf{y})$, one alternatively updates $\mathbf{x}$ and $\mathbf{y}$ in the following manner: fix $\mathbf{y}$ at initializing value $\hat{\mathbf{y}}$, $\mathbf{y} \equiv \hat{\mathbf{y}}$; solve the convex optimization problem for $\mathbf{x}$, $\hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}} f(\mathbf{x}, \hat{\mathbf{y}})$; fix $\mathbf{x}$ at optimizer value $\hat{\mathbf{x}}$, $\mathbf{x} \equiv \hat{\mathbf{x}}$; solve the convex optimization problem for $\mathbf{y}$, $\hat{\mathbf{y}} = \text{argmin}_{\mathbf{y}} f(\hat{\mathbf{x}}, \mathbf{y})$; and so on.

Two-Stage Estimation Algorithm (**for arbitrary** j, $j = 1, \ldots, p$)

0. Initialize $\hat{\boldsymbol{\beta}}_{j,I}$ with a zero-vector, $\hat{\boldsymbol{\beta}}_{j,I} \equiv \mathbf{0}$.
1. **First stage:** Proceed to solve $min_{\boldsymbol{\beta}_{j,C}} f(\boldsymbol{\beta}_{j,C}, \hat{\boldsymbol{\beta}}_{j,I})$, which is equivalent to solving the following convex group lasso optimization criterion

$$\min_{\boldsymbol{\beta}_{j,C}} \|\hat{D}_{\hat{\sigma}^2}^{-\frac{1}{2}}([\tilde{\mathbf{X}}_j - \tilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}_{j,I}] - \tilde{\mathbf{Z}}^{(C)}\boldsymbol{\beta}_{j,C})\|_2^2 + \lambda_j^G \sum_{i=1}^{p} \|(\boldsymbol{\beta}_{j,C}^{(1)}[i], \ldots, \boldsymbol{\beta}_{j,C}^{(K)}[i])^\intercal\|_2, \tag{10}$$

where matrix $\tilde{\mathbf{Z}}^{(C)}$ is such that $\tilde{\mathbf{Z}}^{(C)}[., i] \equiv \tilde{\mathbf{Z}}[, i]$ if $\hat{\boldsymbol{\beta}}_{j,I}[i] = 0$, and $\tilde{\mathbf{Z}}^{(C)}[., i] \equiv \mathbf{0}$ if $\hat{\boldsymbol{\beta}}_{j,I}[i] \neq 0$. Using such matrix $\tilde{\mathbf{Z}}^{(C)}$ guarantees that optimizer $\hat{\boldsymbol{\beta}}_{j,C} \equiv \text{argmin}_{\boldsymbol{\beta}_{j,C}} f(\boldsymbol{\beta}_{j,C}, \hat{\boldsymbol{\beta}}_{j,I})$ will have non-overlapping support with $\hat{\boldsymbol{\beta}}_{j,I}$, $\hat{\boldsymbol{\beta}}_{j,C} \perp \hat{\boldsymbol{\beta}}_{j,I}$. We do this instead of explicitly including the penalty term $\lambda^\infty \sum_{i=1}^{p} \sum_{k=1}^{K} |\boldsymbol{\beta}_{j,C}^{(k)}[i] \cdot \boldsymbol{\beta}_{j,I}^{(k)}[i]|$ from (9), simplifying the process of optimization.
2. **Second stage:** Let $\hat{\boldsymbol{\beta}}_{j,C}$ denote the estimate of $\boldsymbol{\beta}_{j,C}$ from the first stage. Proceed to solve $min_{\boldsymbol{\beta}_{j,I}} f(\hat{\boldsymbol{\beta}}_{j,C}, \boldsymbol{\beta}_{j,I})$, which is equivalent to the following convex lasso problem

$$\min_{\boldsymbol{\beta}_{j,I}} \|\hat{D}_{\hat{\sigma}^2}^{-\frac{1}{2}}([\tilde{\mathbf{X}}_j - \tilde{\mathbf{Z}}^{(I)}\hat{\boldsymbol{\beta}}_{j,C}] - \tilde{\mathbf{Z}}\boldsymbol{\beta}_{j,I})\|_2^2 + \lambda_j^S \|\boldsymbol{\beta}_{j,I}\|_1, \tag{11}$$

where matrix $\tilde{\mathbf{Z}}^{(I)}$ is such that $\tilde{\mathbf{Z}}^{(I)}[., i] \equiv \tilde{\mathbf{Z}}[, i]$ if $\hat{\boldsymbol{\beta}}_{j,C}[i] = 0$, and $\tilde{\mathbf{Z}}^{(I)}[., i] \equiv \mathbf{0}$ if $\hat{\boldsymbol{\beta}}_{j,C}[i] \neq 0$. Analogously to $\tilde{\mathbf{Z}}^{(C)}$ in the first stage, using such matrix $\tilde{\mathbf{Z}}^{(I)}$ guarantees that $\hat{\boldsymbol{\beta}}_{j,C} \perp \hat{\boldsymbol{\beta}}_{j,I}$.
3. If it is the second iteration (or higher): denote $\hat{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\beta}}_{j,C} + \hat{\boldsymbol{\beta}}_{j,I}$ as the full estimate for current iteration, and $\hat{\boldsymbol{\beta}}_j^{pr}$ – full estimate from previous iteration; stop the algorithm if $\|\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j^{pr}\|_2^2 < 10^{-4}$. Otherwise, go back to the first stage, using $\hat{\boldsymbol{\beta}}_{j,I}$ calculated during second stage.

The described two-stage iterative approach performs the alternate convex search method, which would lead to a local optimizer of a biconvex function $f(\boldsymbol{\beta}_{j,C}, \boldsymbol{\beta}_{j,I})$ for fixed values of tuning parameters λ_j^G and λ_j^S. The specific values for λ_j^G and λ_j^S, in their turn, can be selected via heuristic approach of running the aforementioned two-stage algorithm for a few iterations, but adding a tuning parameter selection step to each of the two stages as follows:

1. **First stage (continued):** Calculate the solution path of λ_j^G values for optimization task (10) as in Yuan and Lin (2006). Pick the tuning parameter value $\hat{\lambda}_j^G$ that minimizes the Bayesian Information Criterion (BIC)

$$BIC(\lambda_j) = n \log(\|\hat{D}_{\hat{\sigma}^2}^{-\frac{1}{2}}([\tilde{\mathbf{X}}_j - \tilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}_{j,I}] - \tilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}_{j,C}(\lambda_j^G))\|_2^2) + \log(n)\, \text{df}_{\lambda_j^G}, \tag{12}$$

where $n = K(T - D + 1)$, $\hat{\boldsymbol{\beta}}_{j,C}(\lambda_j^G)$ – estimate of $\boldsymbol{\beta}_{j,C}$ corresponding to value λ_j^G of the solution path, $\text{df}_{\lambda_j^G}$ – degrees of freedom for estimate $\hat{\boldsymbol{\beta}}_{j,C}(\lambda_j^G)$, calculated as in Breheny and Huang (2009).
2. **Second stage (continued):** Obtain a solution path of λ_j^S values for optimization task (11) as in Tibshirani et al. (2011), and, similarly to the first stage, pick the tuning parameter value $\hat{\lambda}_j^G$ that minimizes the BIC. Degrees of freedom $\text{df}_{\lambda_j^S}$ in this case are calculated as the number of non-zero elements in the estimate $\hat{\boldsymbol{\beta}}_{j,I}(\lambda_j^S)$.

We implement this extended version of a two-stage approach for the first few iterations (in the numerical work presented, we used it for the first five iterations) in order to obtain a good data-driven choice of tuning parameter values $\hat{\lambda}_j^G$ and $\hat{\lambda}_j^S$. Then, setting $\lambda_j^G \equiv \hat{\lambda}_j^G$ and $\lambda_j^S \equiv \hat{\lambda}_j^S$ for optimization tasks (9)–(11), we simply execute the original version of the two-stage algorithm, which is guaranteed to converge to a local minimum of $f(\boldsymbol{\beta}_{j,C}, \boldsymbol{\beta}_{j,I})$. To improve our chances of landing in a global minimum, we could potentially look at multiple different initialization values for estimate $\hat{\boldsymbol{\beta}}_{j,I}$ at step 0 of our two-stage approach, which could serve as an interesting topic for future work (see related work in Lin et al. (2016)).

3. Performance evaluation

We focus our evaluation studies on *VAR* models of order $D = 1$ in order to assess the performance of the proposed joint estimation procedure, rather than focusing on model order selection:

$$X_t^{(k)} = A^{(k)} X_{t-1}^{(k)} + \epsilon_t^{(k)}, \ \epsilon_t^{(k)} \sim N(\mathbf{0}, \sigma_{(k)}^2 \mathbf{I}_p), \quad t = 1, \ldots, T, \quad k = 1, \ldots, K. \tag{13}$$

The employed performance metrics are presented next, where $\hat{A} = (\hat{a}_{i,j})_{p \times p}$ denotes the transition matrix estimate, and $A = (a_{i,j})_{p \times p}$ – the matrix of true values:

Table 1
Two-stage procedure performance for increasing subject group size K.

K	FP (Com)	FN (Com)	MC (Com)	NFD (Com)	FP (Ind)	FN (Ind)	MC (Ind)	NFD (Ind)
				$p = 20$	$T = 80$			
10	0.06(0.02)	0.05(0.08)	0.89(0.06)	0.5(0.18)	0(0)	0.92(0.07)	0.19(0.08)	0.98(0.02)
20	0.03(0.02)	0.06(0.08)	0.92(0.06)	0.51(0.19)	0(0)	0.65(0.06)	0.46(0.05)	0.89(0.02)
50	0.01(0.01)	0.06(0.1)	0.94(0.08)	0.51(0.21)	0(0)	0.43(0.11)	0.63(0.08)	0.8(0.05)
				$p = 30$	$T = 120$			
10	0.06(0.01)	0(0)	0.94(0)	0.35(0.04)	0(0)	0.95(0.03)	0.15(0.04)	0.99(0.01)
20	0.02(0)	0(0)	0.98(0)	0.3(0.01)	0(0)	0.51(0.04)	0.56(0.03)	0.83(0.02)
50	0.01(0)	0(0)	0.99(0)	0.3(0.07)	0(0)	0.18(0.02)	0.83(0.02)	0.66(0.01)
				$p = 40$	$T = 150$			
10	0.06(0.01)	0(0)	0.94(0.01)	0.39(0.1)	0(0)	0.96(0.04)	0.12(0.07)	0.99(0.01)
20	0.02(0.01)	0(0)	0.97(0)	0.34(0.11)	0(0)	0.5(0.09)	0.57(0.07)	0.83(0.04)
50	0.01(0)	0(0)	0.99(0)	0.29(0.1)	0(0)	0.19(0.04)	0.82(0.04)	0.64(0.01)

Note: Triplets of rows correspond to the same setting in terms of p and T. E.g. rows 3–5 correspond to the case of $p = 20$, $T = 80$, rows 7–9 correspond to setting $p = 30$, $T = 120$, etc. Means and standard deviations are shown for performance metrics discussed at the top of Section 3, with 50 replicates per each setting.

- False Positive (FP) and True Negative (TN) rates:

$$FP = \frac{\sum_{1\le i,j\le p} I(a_{i,j} = 0, \hat{a}_{i,j} \neq 0)}{\sum_{1\le i,j\le p} I(a_{i,j} = 0)}, \quad TN = 1 - FP.$$

- False Negative (FN) and True Positive (TP) rates:

$$FN = \frac{\sum_{1\le i,j\le p} I(a_{i,j} \neq 0, \hat{a}_{i,j} = 0)}{\sum_{1\le i,j\le p} I(a_{i,j} \neq 0)}, \quad TP = 1 - FN.$$

- Matthews Correlation Coefficient (MC, geometric mean of FP and FN):

$$MC = \frac{TP \times TN \; - \; FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

- Normalized Frobenius Difference (NFD):

$$NFD = \frac{\|\hat{A} - A\|_F}{\|A\|_F} = \frac{\sqrt{\sum_{1\le i,j\le p}(\hat{a}_{i,j} - a_{i,j})^2}}{\sqrt{\sum_{1\le i,j\le p} a_{i,j}^2}}.$$

The first three measures – FP, FN, MC – describe how well the matrix structure (the positioning of non-zero elements) is estimated, which is our priority in this work. Low values of FP, FN (near 0) and high values of MC (near 1) are indicative of a good performance. In the meantime, NFD evaluates how well the matrix element magnitudes are estimated. Lower values of NFD (closer to 0) correspond to better performance. When only evaluating the performance of the two-stage estimation algorithm introduced in Section 2, all four measures are calculated for both the common component estimates $\hat{A}_C^{(k)}$ and individual component estimates $\hat{A}_I^{(k)}$, $k = 1, \ldots, K$. Meanwhile, when comparing this two-stage approach to the other state-of-the-art methods, due to the fact that those other methods do not implement a breakdown into common and individual components, we calculate the four measures just for the full matrix estimates $\hat{A}^{(k)} = \hat{A}_C^{(k)} + \hat{A}_I^{(k)}$, $k = 1, \ldots, K$.

The design of the simulation studies is as follows: matrices $\{A^{(k)} = A_C^{(k)} + A_I^{(k)}, \; k = 1, \ldots, K\}$ are generated with spectral radius of 0.4, with non-zero effects having magnitude of at least 0.2. This would guarantee stationarity of the generated time series, separate noise from the signal, and replicate to a large extent features of the fMRI data presented in Section 4 (no estimated effects had magnitude larger than 0.4). The resulting time series data are generated by adding independent $N(0, 1)$ errors and setting the signal-to-noise ratio at one (as in the ADHD study of Section 4). Multiple settings are considered by varying the number of time series p per subject, and number of subjects K in a group. All the diagonals are generated to be non-zero, while the edge density of elements in the common component is set to 5% (of off-diagonal elements) for $p = 20$, and 2% for $p = 30, \; 40$. The edge density for the individual component is set to 2%–3% (of the total number of matrix elements), implying a moderate level of heterogeneity. A threshold of 0.02 is applied to the elements of the resulting estimates to eliminate noisy entries. Finally, the metrics are obtained by averaging over 50 replicates.

Given that our joint estimation procedure was developed with the goal of borrowing strength across multiple subjects, its efficiency is best observed when gradually increasing the number K of related subjects. On simulation settings for the case of increasing time series length T, see results in the supplement. In Table 1, we demonstrate the impact on performance when the number K of jointly estimated subjects grows from 10 up to 50. With an increase in the

Table 2
Comparative simulation study across four methods: regular lasso (LASSO), sparse group lasso (SGL), group bridge (GB) and our two-stage approach (2ST).

Setting	Method	FP	FN	MC	NFD
$K = 20$ $p = 20,\ T = 80.$	LASSO	0.02(0.02)	0.42(0.42)	0.61(0.33)	0.71(0.21)
	SGL	0.3(0.05)	0.05(0.08)	0.68(0.04)	0.55(0.11)
	GB	0.03(0.01)	0.13(0.09)	0.85(0.07)	0.5(0.16)
	2ST	0.02(0.01)	0.14(0.08)	0.85(0.06)	0.62(0.16)
$K = 20,$ $p = 40, T = 150.$	LASSO	0.01(0)	0.11(0.18)	0.9(0.15)	0.49(0.14)
	SGL	0.2(0.03)	0.02(0.04)	0.79(0.02)	0.45(0.08)
	GB	0.02(0)	0.03(0.02)	0.95(0.02)	0.32(0.08)
	2ST	0.02(0.01)	0.03(0.01)	0.95(0.01)	0.49(0.08)
$K = 50,$ $p = 20, T = 80.$	LASSO	0.02(0.02)	0.43(0.41)	0.61(0.32)	0.71(0.21)
	SGL	0.36(0.05)	0.04(0.05)	0.64(0.04)	0.55(0.11)
	GB	0.03(0.01)	0.17(0.13)	0.81(0.1)	0.54(0.18)
	2ST	0.01(0.01)	0.13(0.1)	0.87(0.08)	0.61(0.17)
$K = 50,$ $p = 40, T = 150.$	LASSO	0.01(0)	0.08(0.17)	0.92(0.14)	0.47(0.13)
	SGL	0.23(0.03)	0.02(0.02)	0.76(0.01)	0.45(0.05)
	GB	0.03(0)	0.05(0.02)	0.92(0.02)	0.35(0.08)
	2ST	0.01(0)	0.02(0)	0.97(0)	0.42(0.07)

Note: Each of four rows corresponds to four methods applied to the same setting. E.g. rows 2–5 correspond to setting $p = 20, T = 80, K = 20$, rows 6–9 correspond to setting $p = 40, T = 150, K = 20$, etc. Means and standard deviations are shown for performance metrics discussed at the top of Section 3, with 50 replicates per each setting.

number of subjects per group, we see a clear improvement in the estimation of the common component structure (FP decreasing and MC increasing in each setting), which also leads to better estimates for the individual effects (FN increasing and MC decreasing, as well). This aspect stems from the identifiability constraint imposed on the common and individual components: each false positive in the common component may correspond to a misplaced true positive from individual component, directly causing a false negative therein. Therefore, less false positives when estimating the common component lead to less false negatives for the individual component estimates. Nonetheless, the FN numbers for the individual component do not get very close to zero. This is a direct result of enforcing the assumption from (8) on individual components being very sparse. Relaxing that assumption, albeit decreasing FN, would lead to an increase in FP for individual components, which, by the identifiability assumption employed, leads to missing true group-level effects (see supplement for extra simulation results confirming this claim). Given that we prioritize correctly detecting the group-level effects and would only like to account for the most critical subject-specific effects, it is preferable to have FN in individual components, as opposed to FN in the common component. Additionally, it may potentially stem from initializing the individual components with a zero-vector in the two-stage approach, which leads to a local minimum of the joint function (9). On the other hand, when looking at how well the overall estimate $\hat{A} = \hat{A}_C + \hat{A}_I$ does with respect to the true matrix A, one may notice that both FP and FN get close to 0 for certain settings (see Table 2 where we compare our two-stage approach with three competing methods). As it pertains to the algorithm convergence properties, after completing tuning parameter selection iterations, the algorithm converges for each setting with an average of 2–3 iterations needed across 50 replicates.

To evaluate the performance of the proposed joint modeling approach compared to other established regularized estimation methods, we proceed with a comparative simulation study that includes:

- A regular Lasso approach (Tibshirani, 1996), which uses only a sparse penalty norm $\|\boldsymbol{\beta}\|_1 = \sum_l |\beta_l|$ to obtain sparse estimates for each subject separately, without accounting for grouping aspects.
- A Sparse Group Lasso (Simon et al., 2013), which adds a single-variable sparsity penalty term on top of a regular group lasso from Yuan and Lin (2006), thereby allowing for sparsity on both the group and individual levels. Albeit sounding similar to our proposed approach, the sparse group lasso does not employ a decomposition into common and individual components. It instead applies both group- and individual-level sparsity penalties directly to the original VAR transition matrix.
- A Group Bridge (Huang et al., 2009), which implements a bridge-type group-level penalty norm $\|\boldsymbol{\beta}\|_1^\gamma = (\sum_l |\beta_l|)^\gamma,\ 0 < \gamma < 1$, instead of the classical Euclidean norm $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_l \beta_l^2}$ used in the original group lasso (Yuan and Lin, 2006). Similarly to the sparse group lasso (Simon et al., 2013), it does not lead to a decomposition of the VAR transition matrix into common and individual components.

Tuning parameter values for regular lasso and group bridge were selected via BIC, as we did for our two-stage approach. Meanwhile, for sparse group lasso we used cross-validation due to the formula for degrees of freedom not being available. Table 2 demonstrates the performance of all four methods – regular lasso, sparse group lasso, group bridge, two-stage approach – when estimating the full VAR transition matrix A.

In terms of structure estimation (FP, FN, MC), for the settings with $K = 20$ subjects, the group bridge method achieves equivalent performance to our two-stage approach, while the other two methods lag considerably behind. Meanwhile, when increasing the number of subjects to $K = 50$, we see that the proposed two-stage approach clearly outperforming all competitors, since it borrows strength across many similar subjects. It reaches the highest MC values of 0.87 and 0.97 for the $K = 50, p = 20, T = 80$ and $K = 50, p = 40, T = 150$ settings, respectively. On the other hand, in terms of element magnitude estimation (measured via NFD), the group bridge algorithm exhibits an advantage over all remaining methods. Nevertheless, in the setting of $K = 50, p = 40, T = 150$ our two-stage approach outperforms both the regular lasso and sparse group lasso in NFD, on top of already providing the best quality of structure estimation among all methods.

4. Application to brain fMRI data for ADHD study

We analyze resting-state fMRI data from two groups: one of healthy controls, and one of patients exhibiting a certain cognitive disorder. It is expected for subjects within both groups to have similar patterns in the lead–lag relationships amongst their brain regions, while also displaying certain subject-specific effects. The idea of shared structure had been used for fMRI data before (see Chu et al. (2015), Belilovsky et al. (2016)), but the focus was on contemporaneous dependence between brain regions (functional connectivity). Our estimation procedure attempts to describe temporal dynamics across those regions (effective connectivity).

Note that VAR modeling of the cross-region relationships within the brain has received attention in the literature recently. Even though there has been work done on applying VAR models to estimate effective connectivity of the brain (Friston et al., 2003; Goebel et al., 2003), various drawbacks have been pointed out for such an approach. For example, a review paper on advances and pitfalls in resting-state fMRI analysis by Cole et al. (2010) emphasizes the variations in hemodynamic delay across the regions, which may introduce bias into any attempt of estimating temporal effects. Meanwhile, Ramsey et al. (2010) proceed to lay out six detailed reasons for caution when inferring temporal causal effects from fMRI data, one of them being potential heterogeneity across different experimental sites and inability to capture causal relations due to neural activity processes occurring more quickly than the sampling rate of fMRI measurements. Nevertheless, our modeling approach alleviates some of the issues discussed above. First of all, the group lasso directly addresses one of the drawbacks in Ramsey et al. (2010) by capturing the common abstract processing structure, such as which regions of the brain influence which other regions, while also allowing for varying strengths of those influences across the patients. As a reminder, it selects a particular relationship that is common for all the patients, but the exact effect magnitude can differ across the board. We also proceed to select the studies with same repetition times (TR) of fMRI measurements, otherwise risking to jointly estimate temporal effects of different lags. Additionally, the individual component accounts for subject-specific influences, e.g. age or handedness. Lastly, in order to avoid the experiments yielding different regions of interest (ROI) for different subjects, we make sure to use a standard unified brain atlas for region assignment across all the patients.

4.1. The ADHD study setup

We consider the resting-state brain fMRI time series data from 20 ADHD patients and 20 controls obtained from the ADHD-200 Global Competition that can be retrieved using the Python module *nilearn*. The subjects came from five experimental sites, three of which (NYU Child Study Center, Peking University and Radboud University for NeuroIMAGE study) shared a TR of 2.0 s, with the other two (Oregon Health and Science University, Kennedy Krieger Institute) having a longer TR of 2.5 s. As our model aims at inferring temporal effects within the brain, we proceed to exclude the last two studies from consideration in order to have consistent measurement repetition times across all subjects. That leaves us with 12 ADHD subjects (all males; age mean $\pm$ SD $= 13.85 \pm 3.83$) and 12 controls (all males; age mean $\pm$ SD $= 13.72 \pm 3.72$), respectively. All sites reported a signal to noise ratio of one. The data pre-processing steps include corrections for delay in slice acquisition and motion, filtering to remove high frequencies, data standardization and detrending (for more details see Varoquaux and Craddock (2013)). As mentioned in the discussion above, we proceed to parcellate the brain into 39 regions according to the Multi-Subject Dictionary Learning atlas (MSDL), and subsequently follow the processing steps outlined in Varoquaux and Craddock (2013). We summarize the signal over those regions via the mean of voxel time series, weighted by gray matter probabilistic segmentation. Both the anatomical locations of the regions and resulting extracted time series can be found in Figs. 1 and 2, with details on scientific names for each of those brain regions contained in Table 3 of Appendix A.1. Moreover, Fig. 3 depicts autocorrelation plots corresponding to time series extracted for one of the regions. The obvious spike at the first time lag of PACF (bottom right plot) serves as a strong AR(1) signature, which is present for the vast majority of brain regions under consideration. This leads us to believe that a VAR(1) model would be sufficient, which also motivated the design of the simulation studies presented in Section 3.

Considering the lack of literature on distributional characterization and asymptotical properties of estimates resulting from the group lasso procedure, we employed stability selection to evaluate consistency of estimated temporal effects between brain regions. The concept of stability selection was introduced in Meinshausen and Bühlmann (2010), with its main idea being the use of bootstrap, and subsequent accumulation of the results over all bootstrapped samples for further

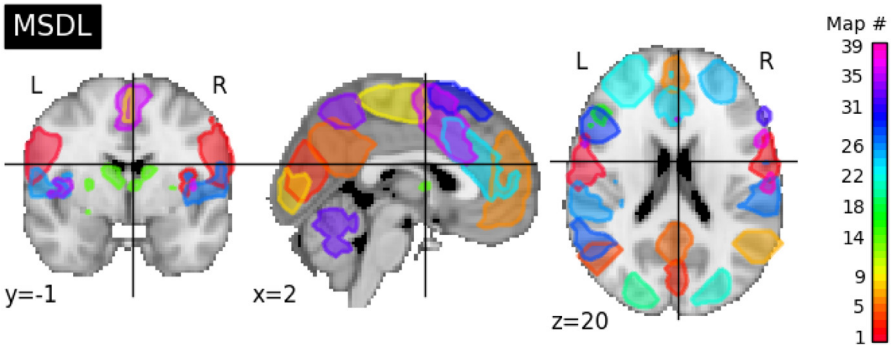

Fig. 1. Brain regions according to MSDL atlas (see Table 3 for region names).

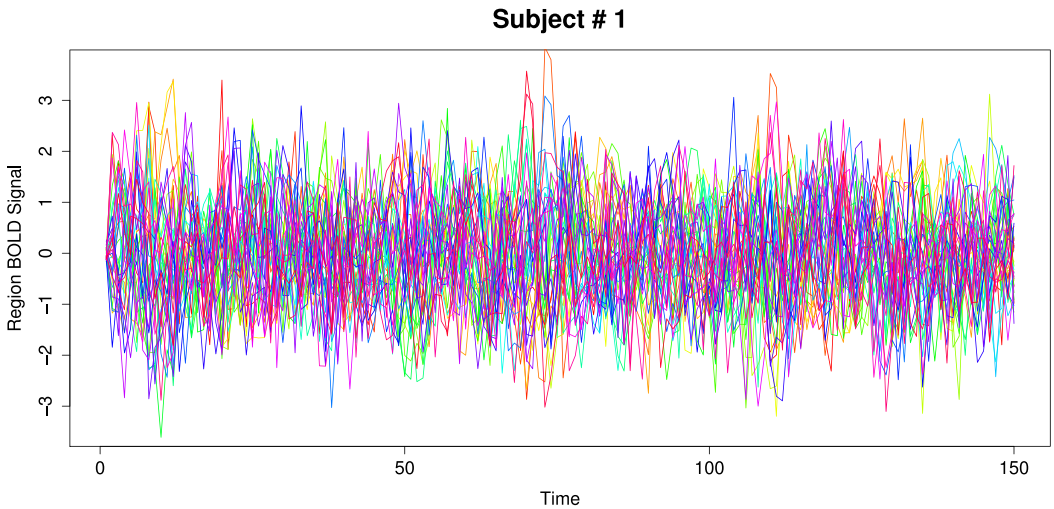

Fig. 2. Extracted time series for 39 brain regions for a single subject.

analysis. Subsampling leads to controlling the family-wise type I error rate in multiple testing for finite sample sizes, which is considerably more important for high-dimensional problems than an asymptotic statement with the number of observations tending to infinity. When dealing with time series data, one has to generate subsamples while retaining the dependence structure among the observations. Time series block bootstrapping techniques (Härdle et al., 2003; Bühlmann and Künsch, 1999) take care of this task, providing us with $B = 100$ estimates per each subject along the way. It allows for evaluation of their stability and leads to a more reliable descriptive analysis of temporal relationships between the brain regions.

4.2. ADHD study results

We ran the two-stage algorithm from Section 2.2 to jointly estimate temporal dependencies across brain regions for patients of the same group, and focused our attention on common component estimates. Fig. 4 depicts resulting temporal relationships between the brain regions in a form of a directed graph: 39 nodes on the left correspond to each brain region's blood-oxygen-level dependent signal (BOLD) at time t, 39 nodes on the right — at time $t + 1$. Each directed edge from left to right node represents temporal effect consistently appearing in the common component estimate across all bootstrapped samples for a patient group (ADHD or control). In particular, we calculated proportions of times each temporal effect appeared in the common component estimate (out of total number of bootstrapped estimates) and thresholded the results at 0.75, only showing the most consistent effects.

Fig. 4 depicts both the ADHD and control groups having strong autocorrelation effects for each brain region (blue edges), which is to be expected. As it pertains to inter-regional temporal effects, we have a fair amount of those that are present in both groups (red edges) and ones that are specific to a certain group (green edges). The magnitudes of detected temporal effects were (mean $\pm$ SD) 0.23 ± 0.1 for autocorrelation effects, 0.07 ± 0.05 for inter-regional effects. Going

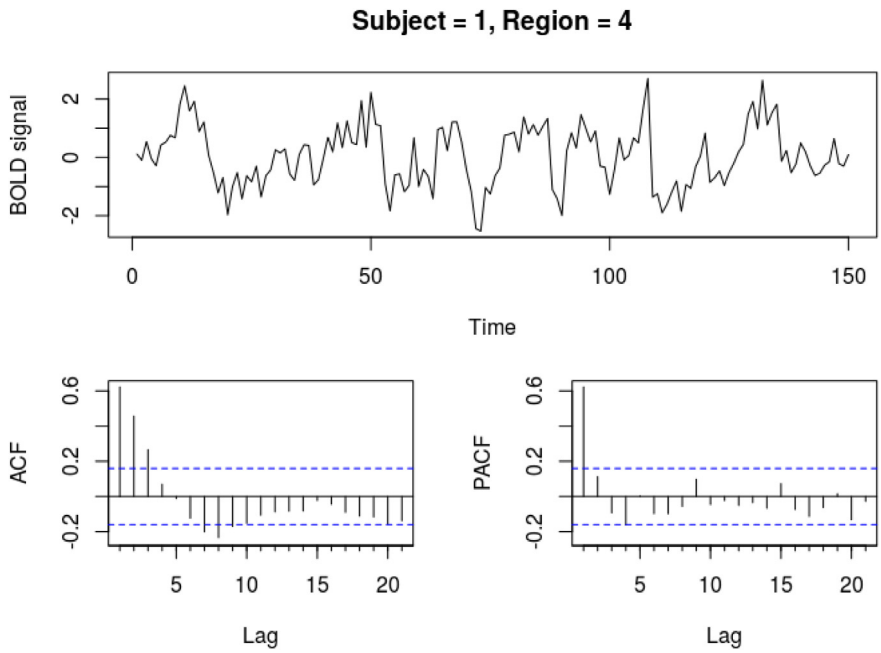

Fig. 3. Autocorrelation plots for a single region time series: top — time series plot, bottom left — autocorrelation plot (ACF), bottom right — partial autocorrelation plot (PACF).

from top to bottom, we start with auditory cortex network (top two regions) and witness left and right auditory cortex influencing each other for controls, while only working in one direction for ADHD patients. As described in Serrallach et al. (2016), a lot of studies have shown auditory problems for ADHD-diagnosed patients, which may be reflected in the lack of communication within the auditory cortex network for their group in Fig. 4. The default mode network (DMN) has been shown to experience irregularities for ADHD patients, which leads to disruptions in cognitive performance and resulting lapses of attention (one of the main symptoms of the disease). In particular, Cortese et al. (2012) discovered the patterns of hyperactivation, while Weissman et al. (2006) and Sato et al. (2012) pointed to deficits in its deactivation, which reduces sensitivity to stimulus. Our results partly reflect those ideas by showing more activity in DMN for ADHD group: on top of autocorrelation effects, it also contains a cross-region temporal effect of right DMN to medial prefrontal cortex.

Both ventral attention networks, right (RVAN, from right dorsolateral prefrontal down to right inferior temporal cortex) and left (LVAN, from left parietal down to left polar frontal lobe), demonstrate plenty of distinct activity patterns for two groups. As mentioned in Cortese et al. (2012), in ADHD studies for children they found regions of hypoactivation as well as hyperactivation in VAN. Hypoactive regions manifest ADHD-related deficits in detecting and adjusting to environmental irregularities, while hyperactive ones underline distractability — one of the most crucial ADHD-symptoms. Meanwhile, in dorsal attention network (DAN, left and right intraparietal sulcus) we see better communication for controls, which reinforces results of Castellanos et al. (2008) and Tomasi and Volkow (2012), both pointing to abnormalities and lack of interactions in DAN as one of characteristics for ADHD patients. Additionally, Sigi Hale et al. (2007) emphasize enhanced intraparietal sulcus activation for controls compared to ADHD patients. As for the visual network (VN, primary visual cortex, right and left occipital complex), it was shown to be a discriminative area when comparing ADHD and control groups in Zhu et al. (2008), with Castellanos and Proal (2012) unveiling lower connectivity patterns in visual and occipital cortexes of ADHD subjects. In our case, we witness considerably more activity for the controls with a temporal cross-effect for right/left occipital complexes, and visual cortex influencing both of those regions as well.

Salience network (SN, dorsal/ventral anterior cingulate cortex and anterior insular cortex) for controls demonstrates higher endogenous activity (ventral anterior cingulate cortex affecting the other two regions), while also being heavily influenced by the cingulate insular network (CIN, cingulate, right and left insular cortex). According to Cortese et al. (2012), this network plays crucial role in such executive processes as decision-making and processing information from the external factors, deficiencies in which are very characteristic for ADHD. Additionally, Weissman et al. (2006) point to association between attention lapses with reduced pre-stimulus activity in anterior cingulate regions, giving credence to the lack of communication between those regions of the SN for ADHD group in our study (for controls, in the meantime, also showing more communication and being influenced by members of CIN network). Left and right superior temporal sulcus regions display temporal cross-effect between them for the control group, while only showing a left to right effect for ADHD, reaffirming the results of Cortese et al. (2012) who claim hypoactivity in temporal regions for ADHD patients.

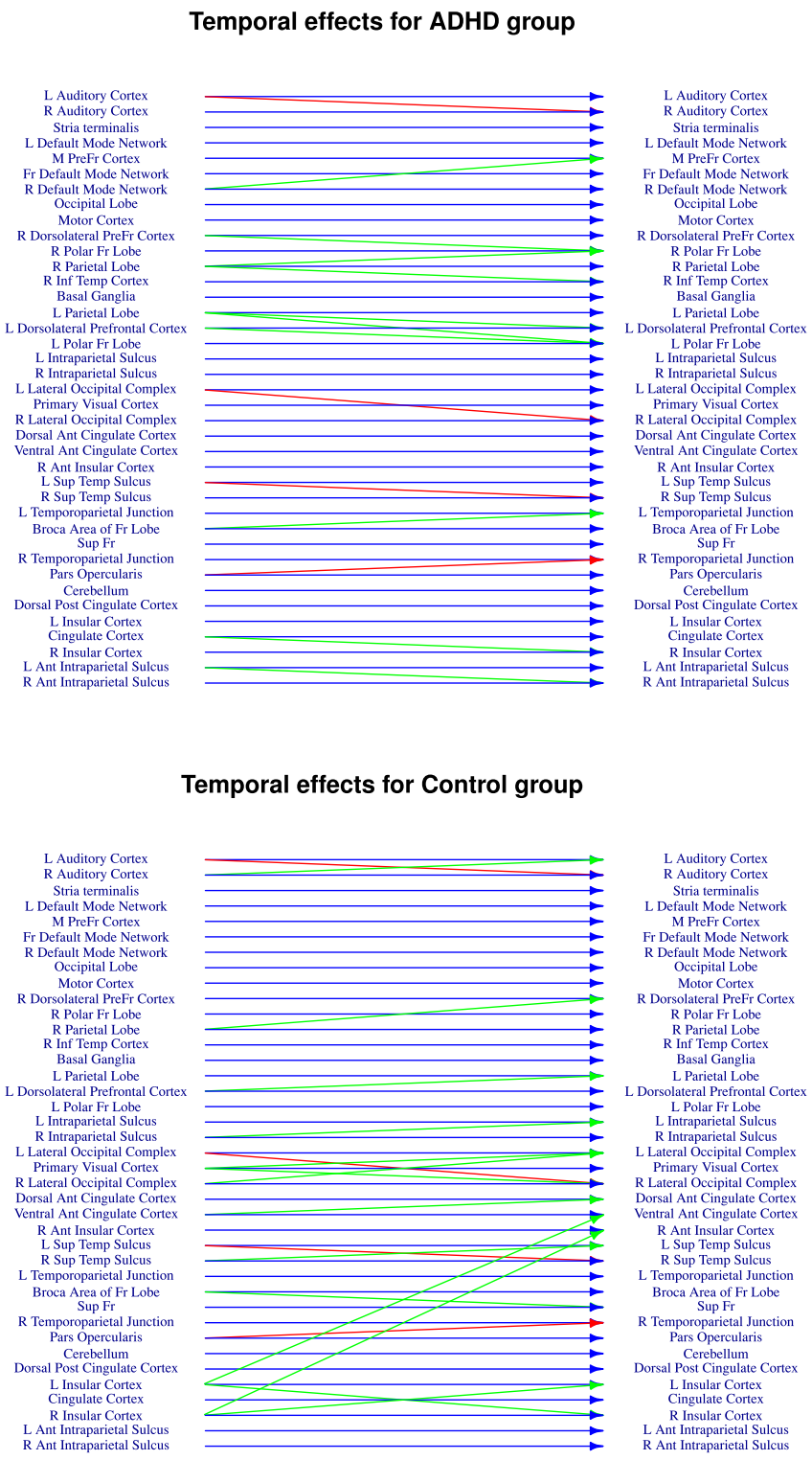

Fig. 4. Directed graph of temporal effects for ADHD patients (top) and controls (bottom). Nodes on the left — brain regions at time t, right — at time $t+1$. Blue edge — region's autocorrelation effect; red — inter-regional effect present for both patient groups; green — group-specific inter-regional effect . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The last network demonstrating considerable activity is the aforementioned cingular insular network (CIN). Here, ADHD patients show distinct influence of cingulate cortex region on the right insular cortex, which agrees with hypothesis of cingulate cortex hyperactivation for ADHD subjects shown in Cortese et al. (2012). In the meantime, controls have a temporal cross-effect between right and left insular cortexes, with both of the regions influencing the members of Salience Network.

Regarding the idiosyncratic component estimates, that resulted from our two-stage procedure described in Section 2.2, there happened to be no consistent subject-specific effects for either of the study participants (none were picked in more than 20% of bootstrapped samples). This may be attributable in part to the lack of heterogeneity in our data set (all the subjects being teenage boys), and the simulation study from Section 3 for a similar setting ($p = 40,\ T = 150,\ K = 10$) showing propensity for false negatives in individual component estimates.

5. Conclusion and future work

In this work, we present a joint estimation procedure for a setting with multiple related VAR models being perturbations of a single underlying common VAR model. A strong motivating application is the analysis of brain fMRI time series data for mental disorder studies with patients sharing the same mental status (disease or healthy control). The final estimates are provided by a two-stage algorithm that breaks down the temporal signal into common and individual components, uses all subjects to jointly estimate a common component via group lasso, and subtracts the common VAR signal to estimate individual components via sparse lasso. The performance on simulated data is shown to be consistent for common component across most of the settings, while individual component estimation gradually gets better with increase in the number of subjects being jointly estimated. Proposed approach also showed superior performance in terms of matrix structure estimation when compared to other state-of-the-art methods. Moreover, having explicitly defined the concept of a common component, our two-stage estimation approach makes it much easier to capture and interpret the group-level effects in the brain fMRI data study for ADHD and control patients, respectively.

The most significant extension would be developing a hypothesis testing framework for the joint estimation procedure presented in Section 2. There is not enough literature on distributional properties of estimates resulting from group lasso procedure, which is necessary for both testing the significance of temporal relationships in a single group, and comparing strength of temporal relationships across different groups. One of the papers addressing the issues of group-level inference for regularized estimation is Narayan et al. (2015), where it is being referred to as Population Post Selection Inference. It outlines a testing procedure for group-level effects that accounts for uncertainties introduced by both the regularization and inter-subject variability. Unfortunately, they do not use joint estimation approach and group lasso penalty, estimating brain networks separately with regular lasso, and therefore not directly applying to our procedure. Other aspects in need of further study include: the exploration of different initializations for individual component estimate to increase chances of the presented two-stage iterative procedure converging to a global minimum; the evaluation of the joint estimation procedure's ability to deal with VAR models of various lag orders, along with developing a lag order selection method.

Acknowledgments

The work of the first author was supported in part by Graduate Fellowship, USA and by National Science Foundation, USA [grant DMS-1632730]. The work of the second author was supported in part by National Science Foundation, USA [grants CCF 1540093 and IIS 1632730] and National Institutes of Health, USA [grant 5 R01 GM114029-03].

Appendix

A.1. MSDL atlas annotation

In Table 3 you can see the full scientific names for brain regions included in Multi-Subject Dictionary Learning (MSDL) atlas. The enumeration corresponds to the order in which regions show up in Fig. 1.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2019.05.007.

File Supplement.pdf Contains details on binconvexity of optimization task (9) and extra simulation results.

Table 3
MSDL atlas brain region names.

Region #	Region name	Region #	Region name
1	L Auditory Cortex	21	Primary Visual Cortex
2	R Auditory Cortex	22	R Lat Occ Complex
3	Stria terminalis	23	Dorsal Ant Cing Cortex
4	L Default Mode Network	24	Ventral Ant Cing Cortex
5	Med Prefr Cortex	25	R Ant Insular Cortex
6	Fr Default Mode Network	26	L Sup Temp Sulcus
7	R Default Mode Network	27	R Sup Temp Sulcus
8	Occ Lobe	28	L Temp-Par Junction
9	Motor Cortex	29	Broca Area of Fr Lobe
10	R Dorso-Lat Prefr Cortex	30	Sup Fr
11	R Polar Fr Lobe	31	R Temp-Par Junction
12	R Par Lobe	32	Pars Opercularis
13	R Inf Temp Cortex	33	Cerebellum
14	Basal Ganglia	34	Dorsal Post Cing Cortex
15	L Par Lobe	35	L Insular Cortex
16	L Dorso-Lat	36	Cing Cortex
17	L Polar Fr Lobe	37	R Insular Cortex
18	L Intra-Par Sulcus	38	L Ant Intra-Par Sulcus
19	R Intra-Par Sulcus	39	R Ant Intra-Par Sulcus
20	L Lat Occ Complex		

Note: The common abbreviations used are 'L' — left, 'R' — right, 'Lat' — lateral, 'Ant' — anterior, 'Fr' — frontal, 'Med' — medial, 'Post' — posterior, 'Sup' — superior, 'Inf' — inferior, 'Temp' — temporal, 'Par' — parietal, 'Occ' — occipital, 'Cing' — cingulate.

References

Bańbura, M., Giannone, D., Reichlin, L., 2010. Large bayesian vector auto regressions. J. Appl. Econometrics 25 (1), 71–92.

Basu, S., Michailidis, G., et al., 2015a. Regularized estimation in sparse high-dimensional time series models. Ann. Statist. 43 (4), 1535–1567.

Basu, S., Shojaie, A., Michailidis, G., 2015b. Network granger causality with inherent grouping structure. J. Mach. Learn. Res. 16 (1), 417–453.

Beckmann, C.F., Jenkinson, M., Smith, S.M., 2003. General multilevel linear modeling for group analysis in fmri. Neuroimage 20 (2), 1052–1063.

Belilovsky, E., Varoquaux, G., Blaschko, M.B., 2016. Testing for differences in gaussian graphical models: applications to brain connectivity. In: Advances in Neural Information Processing Systems. pp. 595–603.

Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge university press.

Breheny, P., Huang, J., 2009. Penalized methods for bi-level variable selection. Stat. Interface 2 (3), 369.

Bühlmann, P., Künsch, H.R., 1999. Block length selection in the bootstrap for time series. Comput. Statist. Data Anal. 31 (3), 295–310.

Castellanos, F.X., Margulies, D.S., Kelly, C., Uddin, L.Q., Ghaffari, M., Kirsch, A., Shaw, D., Shehzad, Z., Di Martino, A., Biswal, B., et al., 2008. Cingulate-precuneus interactions: a new locus of dysfunction in adult attention-deficit/hyperactivity disorder. Biol. Psychiatry 63 (3), 332–337.

Castellanos, F.X., Proal, E., 2012. Large-scale brain systems in adhd: beyond the prefrontal–striatal model. Trends Cogn. Sci. 16 (1), 17–26.

Chu, S.-H., Lenglet, C., Parhi, K.K., 2015. Joint brain connectivity estimation from diffusion and functional mri data. In: SPIE Medical Imaging. International Society for Optics and Photonics, p. 941321.

Cole, D.M., Smith, S.M., Beckmann, C.F., 2010. Advances and pitfalls in the analysis and interpretation of resting-state fmri data. Front. Syst. Neurosci. 4.

Cortese, S., Kelly, C., Chabernaud, C., Proal, E., Di Martino, A., Milham, M.P., Castellanos, F.X., 2012. Toward systems neuroscience of adhd: a meta-analysis of 55 fmri studies. Am. J. Psychiatry 169 (10), 1038–1055.

Danaher, P., Wang, P., Witten, D.M., 2014. The joint graphical lasso for inverse covariance estimation across multiple classes. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (2), 373–397.

Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19 (4), 1273–1302.

Goebel, R., Roebroeck, A., Kim, D.-S., Formisano, E., 2003. Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. Magn. Reson. Imaging 21 (10), 1251–1261.

Goh, K., Turan, L., Safonov, M., Papavassilopoulos, G., Ly, J., 1994. Biaffine matrix inequality properties and computational methods. In: Proceedings of 1994 American Control Conference-ACC'94, Vol. 1. IEEE, pp. 850–855.

Gorski, J., Pfeuffer, F., Klamroth, K., 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. Math. Methods Oper. Res. 66 (3), 373–407.

Guo, J., Levina, E., Michailidis, G., Zhu, J., 2011. Joint estimation of multiple graphical models. Biometrika asq060.

Härdle, W., Horowitz, J., Kreiss, J.-P., 2003. Bootstrap methods for time series. Internat. Statist. Rev. 71 (2), 435–459.

Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Wu, T., Chen, K., Reiman, E., Initiative, A.D.N., et al., 2010. Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. Neuroimage 50 (3), 935–949.

Huang, J., Ma, S., Xie, H., Zhang, C.-H., 2009. A group bridge approach for variable selection. Biometrika 96 (2), 339–355.

Lin, J., Basu, S., Banerjee, M., Michailidis, G., 2016. Penalized maximum likelihood estimation of multi-layered gaussian graphical models. J. Mach. Learn. Res. 17 (1), 5097–5147.

Lin, J., Michailidis, G., 2017. Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. J. Mach. Learn. Res. 18 (1), 4188–4236.

Liu, A., Chen, X., Wang, Z., McKeown, M.J., 2014. Time varying brain connectivity modeling using fmri signals. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, pp. 2089–2093.

Lütkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer Science & Business Media.

Ma, J., Michailidis, G., 2016. Joint structural estimation of multiple graphical models. J. Mach. Learn. Res. 17 (1), 5777–5824.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (4), 417–473.

Michailidis, G., d'Alché Buc, F., 2013. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. Math. Biosci. 246 (2), 326–334.

Narayan, M., Allen, G.I., 2016. Mixed effects models for resampled network statistics improves statistical power to find differences in multi-subject functional connectivity. Front. Neurosci. 10, 108.
Narayan, M., Allen, G.I., Tomson, S., 2015. Two sample inference for populations of graphical models with applications to functional connectivity. arXiv preprint arXiv:1502.03853.
Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., Glymour, C., 2010. Six problems for causal inference from fmri. Neuroimage 49 (2), 1545–1558.
Sato, J.R., Hoexter, M.Q., Castellanos, X.F., Rohde, L.A., 2012. Abnormal brain connectivity patterns in adults with adhd: a coherence study. PLoS One 7 (9), e45671.
Serrallach, B., Groß, C., Bernhofs, V., Engelmann, D., Benner, J., Gündert, N., Blatow, M., Wengenroth, M., Seitz, A., Brunner, M., et al., 2016. Neural biomarkers for dyslexia, adhd, and add in the auditory cortex of children. Front. Neurosci. 10.
Sigi Hale, T., Bookheimer, S., McGough, J.J., Phillips, J.M., McCracken, J.T., 2007. Atypical brain activation during simple & complex levels of processing in adult adhd: an fmri study. J. Attention Disorders 11 (2), 125–139.
Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A sparse-group lasso. J. Comput. Graph. Statist. 22 (2), 231–245.
Song, S., Zhan, Z., Long, Z., Zhang, J., Yao, L., 2011. Comparative study of svm methods combined with voxel selection for object category classification on fmri data. PLoS One 6 (2), e17191.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 267–288.
Tibshirani, R.J., Taylor, J.E., Candes, E.J., Hastie, T., 2011. The Solution Path of the Generalized Lasso. Stanford University.
Tomasi, D., Volkow, N.D., 2012. Abnormal functional connectivity in children with attention-deficit/hyperactivity disorder. Biol. Psychiatry 71 (5), 443–450.
Varoquaux, G., Craddock, R.C., 2013. Learning and comparing functional connectomes across subjects. Neuroimage 80, 405–415.
Weissman, D.H., Roberts, K., Visscher, K., Woldorff, M., 2006. The neural bases of momentary lapses in attention. Nature Neurosci. 9 (7), 971.
Wendell, R.E., Hurter Jr, A.P., 1976. Minimization of a non-separable objective function subject to disjoint constraints. Oper. Res. 24 (4), 643–657.
Woolrich, M.W., Behrens, T.E., Beckmann, C.F., Jenkinson, M., Smith, S.M., 2004. Multilevel linear modelling for fmri group analysis using bayesian inference. Neuroimage 21 (4), 1732–1747.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (1), 49–67.
Zhu, C.-Z., Zang, Y.-F., Cao, Q.-J., Yan, C.-G., He, Y., Jiang, T.-Z., Sui, M.-Q., Wang, Y.-F., 2008. Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. Neuroimage 40 (1), 110–120.