Journal of Biostatistics and Biometric Applications
Volume 3 | Issue 1
ISSN: 2348-9820

Research Article Open Access

Regularized Joint Estimation of Related VAR Models via Group Lasso

Skripnikov A[*1], **and Michailidis G**[2]

[1]Department of Mathematics, University of Houston, Houston, Texas, USA
[2]Department of Statistics, University of Florida, Gainesville, Florida, USA

[*]**Corresponding author:** Skripnikov A, Department of Mathematics, University of Houston, 7901 Cambridge St. Apt. 22, Houston, Texas, USA, Tel: 3522832960, E-mail: usdandres1@gmail.com

Citation: Skripnikov A, Michailidis G (2018) Regularized Joint Estimation of Related VAR Models via Group Lasso. J Biostat Biometric App 3(1): 104

Received Date: November 24, 2017 **Accepted Date**: February 20, 2018 **Published Date:** February 23, 2018

Abstract

In a number of applications, one has access to high-dimensional time series data on several related subjects. Natural example comes from medical experiments: brain fMRI time series data for various groups of patients, be it controls or individuals with a specific mental disorder. In this work, we discuss the problem of their regularized joint estimation, introduced via Vector Autoregressive modeling (VAR) and leveraging a group lasso penalty on top of regular lasso, so as to increase statistical efficiency of estimates by borrowing strength across the subjects. We develop a modeling framework that allows for both group level and subject-specific effects for related subjects, using group lasso to estimate the former. Besides a simulation study, we also use our approach to tackle some of the known issues of effective connectivity estimation for resting-state fMRI data. In particular, a group-level descriptive analysis is conducted for brain inter-regional temporal effects of ADHD and control patients from ADHD-200 Global Competition, and the findings are compared to those in neuroscience literature.

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

Introduction

With recent advances in technology and growing amounts of available data (click-generated web browsing data, social networks, image and video data), interest in modeling and analysis of high-dimensional time series data is at its peak. Application areas include web recommender systems for log data [1], brain fMRI data [2], gene regulatory network inference [3], macroeconomic time series forecasting and structural analysis [4], 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 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 few time points, one needs to incorporate appropriate sparsity assumptions into the VAR modeling framework. To enforce sparsity, the most classic approach over the years has implemented lasso regularizing penalty [5], with [6] studying theoretical properties for regularized estimation of VAR models in particular.

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 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 (Alzheimer's, for example), we expect their brain connectivity (both functional and effective) to have a certain common structural pattern. But at the same time, due to natural variability and subject-specific features, we usually witness certain individual patterns of temporal relationships for each patient as well. Classic approach to this problem of fMRI time series analysis is performing estimation for each patient separately, and subsequently accumulating the estimates for further group-level analysis.

This methodology has been applied to study brain activity for Alzheimer's disease [7], Autism [8], Parkinson's disease [9], among others. While providing tools for group-level analysis, this approach does not incorporate the prior knowledge of similarity among patients within the same experimental group, attributable to their shared mental condition, into the estimation procedure. The main novelty of this work is the development of joint modeling framework that would enforce the similarity assumption into the estimation procedure, while also leaving room for subject-specific effects estimation. It will allow us to increase effective sample size for common structure estimation by borrowing strength across the related time series. Additionally, the framework permits detection of most considerable individual effects of 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. Proposed approaches leveraged various penalties that encouraged both sparsity and joint estimation of the parameters across the models, see the hierarchical penalty used in [10] or fused lasso penalty in [11]. In this work we will implement 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, classic sparse lasso procedure will be applied to obtain subject-specific effect estimates.

While the introduced joint estimation procedure can be used for numerous time series settings (econometric data for cities with shared manufacturing and economic features, gene expression data for matched subjects, sales data for similar stores), the primary intended application is resting-state fMRI time series data for studying the spontaneous brain temporal dynamics of various mental diseases. All the papers on group-level inference for mental disorders mentioned in the previous paragraph dealt with estimating the functional brain connectivity (inferring the correlations between brain region signals) rather than attempting to infer effective connectivity (the temporal influence across the brain regions). The restingstate data represents monotone within-brain fluctuations and classical VAR serves well for capturing the dynamics of such data. Also, we assume each patient's VAR model to be a perturbation of some common underlying VAR model for this experimental group (be it subjects with disease or healthy controls), the structure of which will be estimated by our joint procedure via group lasso. While there has been plenty of deserved critique against using VAR modeling for causal inference in brain neuroimaging studies [12,13], we attempt to alleviate some of those issues in our ADHD study. e.g., one of the six problems for causal inference from fMRI discussed in [13] concerned varying signal strength across the brain regions for multiple study participants, even though they all share a common abstract processing structure. In our framework, this issue is addressed directly via group lasso, which enforces this common structure while leaving room for variability of effect magnitudes. It is also worth mentioning that joint estimation approach via regularizing penalties had been used before for brain fMRI time series data [14,15], but only for functional connectivity estimation, while we apply it for inferring effective connectivity.

The remainder of the paper is organized as follows: materials and methods section describes the joint modeling framework and introduces the two-stage estimation procedure, results and discussion section demonstrates the performance of the joint estimation procedure for both simulated data and the restingstate fMRI case study, while the conclusion section contains concluding remarks and potential outline of future work.

Materials and Methods

Below is the detailed outline of the methodology used to carry out the joint estimation of VAR models sharing common structure.

VAR Model Formulation

To introduce the joint modeling framework, consider p-variable stationary time series $X^t_k=(X^t_{1k}, \ldots, X^t_{pk})'$, $t=1, \ldots, T$ for $k = 1, \ldots, K$ related subjects. Then the VAR model with lag order D, or $VAR(D)$, is given by

$$X_k^t = A_k^1 X_k^{t-1} + \cdots + A_k^D X_k^{t-D} + \varepsilon_k^t, \varepsilon_k^t \sim N\left(0, \sigma_k^2 I_p\right),\ t{=}D, \ldots, T, k = 1, \ldots, K, \qquad (1)$$

where A^d_k is p x p transition matrix that captures temporal effects of order d between the p variables for subject k, d=1, . . ., D , k =1, . . ., K. Simplifying assumption of diagonal error covariance matrix $\sum_k \sigma_k^2 I_p$ will allow us to break problem (1) into p parallel problems of lower dimensions. 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 component for each VAR model: $A^d_k = A^{d,C}_k + A^{d,I}_k$, $d = 1, \ldots, D$, $k = 1, \ldots, K$. Afterwards, a two-stage estimation algorithm is proposed,consisting of group lasso optimization procedure to jointly estimate the common components $\{A_k^{d,C}\}$ of K subjects during the first stage, followed by *sparse lasso optimization procedure* to estimate the individual components $\{A_k^{d,I}\}$ on the second stage. Group lasso penalty 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 shared structure of resulting common component estimates. Meanwhile, the residuals from the common component signal are used as data to estimate the individual structures, respresenting subject-specific effects, via classic lasso procedure.

Standard Regression Formulation for VAR Models

Recalling the system of VAR model equations from (1), we will proceed to introduce an equivalent system of standard regression equations. 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:

$$X^t = A^1 X^{t-1} + \cdots + A^D X^{t-D} + \varepsilon^t, \varepsilon^t \sim N\left(0, \sigma^2 I_p\right),\ t{=}D, \ldots, T \qquad (2)$$

Our independence assumption for errors fet $\{\varepsilon^t_k, j=1, \ldots, p\}$, implied by diagonal error covariance structure cov(ε^t) = $\sigma^2 I_p$, $t = D, \ldots, T$, allows us to represent temporal dynamics for each of the p variables in the form of the following system of equations:

$$X_j^t = \sum_{l=1}^{p} \left(A^1[j,1]\right) X_t^{t-1} + \cdots + A^D[j,1] X_l^{t-D} + \varepsilon_j^t, \varepsilon_j^t \sim N\left(0, \sigma^2\right), t = D, \ldots, T, j = 1, \ldots, p, \qquad (3)$$

where $A^d[j,l]$ is order-d temporal effect of l^{th} variable on j^{th}, $l, j = 1; \dots, p$. If we let $A^d[j,.] = (A^d[j,1], \dots, A^d[j,p])^T$; $d = 1, \dots, D$, then all T -1 equations from (3) can be represented in a compact matrix form for each variable j respectively:

$$\underbrace{\begin{pmatrix} X_j^{\,T} \\ \dots \\ X_j^{\,D} \end{pmatrix}}_{X_j} = \underbrace{\left(\begin{array}{ccc|c|} X_1^{T-1} & \dots & X_p^{\,T-1} & \dots \\ \dots & \dots & \dots & \dots \\ X_1^{D-1} & \dots & X_p^{\,D\text{-}1} & \dots \end{array}\right.}_{B^1} \underbrace{\left.\begin{array}{ccc} X_1^{T-D} & \dots & X_p^{\,T\text{-}D} \\ \dots & \dots & \dots \\ X_1^{0} & \dots & X_p^{\,0} \end{array}\right)}_{B^D} \underbrace{\begin{pmatrix} A^1[j,.] \\ \dots \\ A^D[j,.] \end{pmatrix}}_{A[j,.]} + \underbrace{\begin{pmatrix} \varepsilon_j^{\,T} \\ \dots \\ \varepsilon_j^{\,D} \end{pmatrix}}_{\varepsilon_j}$$

$$\underbrace{\widetilde{X}_j}_{(T-D+1)\times 1} = \underbrace{\mathrm{B}}_{(T-D+1)\times D_p} \underbrace{A[j,.]}_{D_p\times 1} + \underbrace{\tilde{\varepsilon}_j,}_{(T-D+1)\times 1} \quad \tilde{\varepsilon}_j \sim N\left(0, \sigma^2 I_T\right), j = 1, \dots \mathrm{p} \tag{4}$$

Now, reinstating k, $k = 1, \dots, K$ in the notation and using the standard regression representation (4) for all K VAR models under consideration, we get:

$$\begin{cases} \tilde{X}_{1,j} = \mathrm{B}_K \mathrm{A}_1[j,.] + \tilde{\varepsilon}_{1,j}, \tilde{\varepsilon}_{1,j} \sim N(0, \sigma_1^2 I_T), & j = 1, \dots p \\ \dots \\ \tilde{X}_{K,j} = \mathrm{B}_K \mathrm{A}_K[j,.] + \tilde{\varepsilon}_{K,j}, \tilde{\varepsilon}_{K,j} \sim N(0, \sigma_K^2 I_T), & j = 1, \dots p \end{cases} \tag{5}$$

where $A_k[j,.]$ corresponds to temporal effects (of all D orders) that each of p variables has on j^{th} variable for k^{th} subject, $k = 1, \dots, K$.

Common and Individual Component

The key assumption we make is that of shared structure among the related subjects. In our model it is manifested in similar sparsity pattern across K transition matrices within the same group. Additionally, one has to account for the inevitable heterogeneity introduced by natural variability among subjects under consideration, leading to certain subject-specific effects. Both of these aspects are captured by breaking down each subject's transition matrices into two parts:

$$A_k^d = A_k^{d,C} + A_k^{d,I}, d = 1, \dots, D, k = 1, \dots K, \tag{6}$$

where $A_k^{d,C}$ is the common component of order-d temporal effects for subject k, $A_k^{d,I}$ - individual component. Applying this representation to equations in (5) we get:

$$\begin{cases} \tilde{X}_{1,j} = \mathrm{B}_K \left(\mathrm{A}_1^C[j,.] + A_1^I[j,.]\right) + \tilde{\varepsilon}_{1,j}, \tilde{\varepsilon}_{1,j} \sim N(0, \sigma_1^2 I_T), & j = 1, \dots p \\ \dots \\ \tilde{X}_{K,j} = \mathrm{B}_K \left(\mathrm{A}_K^C[j,.] + A_K^I[j,.]\right) + \tilde{\varepsilon}_{K,j}, \tilde{\varepsilon}_{K,j} \sim N(0, \sigma_K^2 I_T), & j = 1, \dots p \end{cases} \tag{7}$$

Assuming each of K related VAR model to be a perturbation of a common underlying VAR model, we enforce the common support constraint on $\{A_1^C[j,.], \dots, A_K^C[j,.]\}$. Moreover, sparsity assumption is imposed on the individual components $\{A_1^I[j,.], \dots, A_K^I[j,.]\}$ to account for most important subject-specific effects. Lastly, orthogonality constraint $A_k^{d,C} \perp A_k^{d,\mathrm{I}}$ is introduced, implying that support intersection for $A_k^{d,C}$ and $A_k^{d,\mathrm{I}}$ is an empty set, $d = 1, \dots, D$, $k = 1; \dots, K$. In combination with shared support of common components, orthogonality guarantees identifiability of the individual component. The full set of constraints for system (7) is described below:

$$\begin{cases} \mathrm{A}_1^C[j,.] \sim \mathrm{A}_2^C[j,.] \sim \cdots \sim \mathrm{A}_K^C[j,.], & j = 1, \dots, p \\ \mathrm{A}_k^I[j,.] - \text{sparse}, \; k = 1, \dots, K, & j = 1, \dots, p \\ \mathrm{A}_1^C[j,.] \perp \mathrm{A}_k^I[j,.], \; k = 1, \dots, K, & j = 1, \dots, p \end{cases} \tag{8}$$

Estimation Procedure: Two-Stage Approach

To enforce the set (8) of aforementioned constraints we will implement an estimation algorithm with each iteration consisting of two stages. During first stage, the common component is estimated via group lasso, while second stage uses the residuals of that common component estimate as data for sparse estimation of individual components. In order to guarantee orthogonality of resulting common and individual component estimates, we automatically eliminate the effects selected during first stage from consideration for second stage procedure. Consistent maximum likelihood estimators $\{\hat{\sigma}_k, k = 1, \dots, K\}$ will be calculated as in [16] for the system of equations in (1), and subsequently plugged in for $\{\sigma_k, k = 1, \dots, K\}$ in (7). Let us set $x_j^C = (A_1^C[j,.], \dots, A_K^C[j,.])^T$,

$x_j^I = (A_1^I[j,.], \dots, A_K^I[j,.])^T$, $\widetilde{X}_j = \left(\widetilde{X}_{1,j}, \dots, \widetilde{X}_{K,j}\right)$, $j = 1, \dots, p$,, $\hat{D}_{\hat{\sigma}} = \mathrm{Diag}\left(\underbrace{\hat{\sigma}_1^{\,2}, \dots, \hat{\sigma}_1^{\,2}}_{T-D+1}, \underbrace{\hat{\sigma}_2^{\,2}, \dots, \hat{\sigma}_2^{\,2}}_{T-D+1}, \dots, \underbrace{\hat{\sigma}_K^{\,2}, \dots \hat{\sigma}_K^{\,2}}_{T-D+1}\right)$, and $\hat{\mathrm{B}}_{K(T-D+1)\times K(D_p)}$, - block-diagonal matrix,with k^{th} block equal to $B_k \in \mathbb{R}^{(T-D+1)\times(D_p)}$ from Section 5.2, $k = 1, \dots, K$.

Two-Stage Estimation Algorithm (for arbitrary *j, j = 1, . . . , p*)

1. Initialize $\hat{\mathrm{x}}_j^I$ with a okzero-vector, use maximum likelihood estimates $\hat{\sigma}_k$ for $\sigma_k, k = 1,..., K$.

2. **First stage**: To enforce the similarity constraint from (8) for common components within vector $\mathbf{x}_j^{\mathrm{C}}$, use the following convex group lasso optimization criterion

$$\min_{\mathrm{x}_j^C} \left\| \hat{D}_{\hat{\sigma}^2}^{-\frac{1}{2}} \left(\left[\tilde{X}_j - \tilde{\mathrm{B}}\hat{\mathrm{x}}_j^I \right] - \tilde{\mathrm{B}}\mathrm{x}_j^C \right) \right\|_2^2 - \lambda_j^G \sum_{i=1}^{p} \left\| \left(\mathrm{x}_j^C\right)_{1i}, ..., \left(\mathrm{x}_j^C\right)_{Ki} \right\|_2, \tag{9}$$

3. First stage (continued): After calculating the solution path for (9) [17], pick the final common component estimate via Bayesian Information Criterion (BIC)

$$BIC(\lambda_j) = \mathrm{nlog}\left(\left\| \hat{D}_{\hat{\sigma}^2}^{-\frac{1}{2}} \left(\left[\tilde{X}_j - \tilde{\mathrm{B}}\hat{\mathrm{x}}_{\mathrm{j}}^{I} \right] - \tilde{\mathrm{B}}\hat{\mathrm{x}}_{\mathrm{j}}^{C}(\lambda_j) \right) \right\|_2^2 \right) + \log(\mathrm{n})\,\mathrm{df}_{\lambda_j}\,, \tag{10}$$

where *n* = *K(T -D*+1), $\hat{\mathrm{x}}_j^C(\lambda_j)$ - estimate corresponding to value λ_j of the solution path, df_{λ_j} - degrees of freedom for estimate $\hat{\mathrm{x}}_j^C(\lambda_j)$, calculated as in [18].

4. Second stage: Let $\hat{\mathrm{x}}_j^C$ denote the estimate of $\mathrm{x}_j^{\mathrm{C}}$ from the first stage. To enforce the $\hat{\mathrm{x}}_j^C \perp \hat{\mathrm{x}}_j^I$ constraint, let ˜B - matrix containing columns of B corresponding to zero elements of $\hat{\mathrm{x}}_j^C$. With residuals of $\hat{\mathrm{x}}_j^C$ as response, use the following convex lasso problem to estimate $\mathrm{x}_{\mathrm{J}}^{\mathrm{I}}$

$$\min_{\mathrm{x}_j^C} \left\| \hat{D}_{\hat{\sigma}^2}^{-\frac{1}{2}} \left(\left[\tilde{X}_j - \mathrm{B}\hat{\mathrm{x}}_j^C \right] - \tilde{\mathrm{B}}\mathrm{x}_j^C \right) \right\|_2^2 + \lambda_j^{SPARS} \left| \mathrm{x}_j^I \right|_1, \tag{11}$$

5. Second stage (continued): Similarly to Step 3, use BIC to pick the final estimate $\hat{\mathrm{x}}_j^I(\hat{\lambda}_j^{\mathrm{SPARS}})$ .After selecting the tuning parameter value $\hat{\lambda}_j^{\mathrm{SPARS}}$, we complement the resulting vector $\hat{\mathrm{x}}_j^I(\hat{\lambda}_j^{\mathrm{SPARS}})$ with zeros to a full individual component estimate $\hat{\mathrm{x}}_j^I$.

6. If it is the second iteration (or higher): denote $\hat{\mathrm{x}}_{\mathrm{j}} = \hat{\mathrm{x}}_j^C + \hat{\mathrm{x}}_j^I$ as the full estimate for current iteration, and $\hat{\mathrm{x}}_j^{pr}$ - full estimate from previous iteration; stop the algorithm if $\left\| \hat{\mathrm{x}}_{\mathrm{j}} - \hat{\mathrm{x}}_j^{pr} \right\|_2^2 < 10^{-2}$. Otherwise, go back to Step 2 (First stage), using $\hat{\mathrm{x}}_j^I$ calculated during Step 5.

Results and Discussion

Simulation Study

As mentioned in the introduction, we will emphasize studying VAR models of order D=1 in order to focus on the properties of joint estimation procedure rather than on model order selection:

$$X_k^t = A_k X_k^{t-1} + \varepsilon_k^t, \varepsilon_k^t \sim N\left(0, \sigma_k^2 I_p\right),\ t = 1,....,T,\ k = 1,...,K, \tag{12}$$

Evaluation of the Estimation Approach: We will evaluate our two-stage estimation approach by generating a group of $p \times p$ matrices {A_k=A^C_K+A^I_k,*k=1, . . . ,K*} such that $A_1^C = ... = A_K^C \equiv A^C$ and $A_k^I \in \mathbb{R}^{p\times p}$ is an arbitrary sparse matrix, *k* = 1, . . . ,*K*. Denoting estimated matrix as $\hat{\mathrm{A}} = \left(\hat{a}_{i,j}\right)_{p\times p}$ and the true matrix as $\mathrm{A} = \left(\hat{a}_{i,j}\right)_{p\times p}$, we use the following metrics to evaluate performance of our estimation procedure:

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

$$FP = \frac{\sum_{1\le i<j\le p} I\left(a_{i,j} = 0, \hat{a}_{i,j} \ne 0\right)}{\sum_{1\le i<j'\le p} I\left(a_{i,j} = 0\right)}, \quad TN\text{=1-}FP,$$

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

$$FN = \frac{\sum_{1\le i<j\le p} I\left(a_{i,j} \ne 0, \hat{a}_{i,j} = 0\right)}{\sum_{1\le i<j\le p} I\left(a_{i,j} \ne 0\right)}, \quad TP\text{=1-}FN,$$

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

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

We calculate all three of these measures for common component estimates $\hat{A}_C^k$ and individual component estimates $\hat{A}_I^k$, $k = 1, ..., K$. Low values of **FP** and **FN** (near 0) and high values of **MC** (near 1) would indicate good performance. Additionally, an average number of iterations needed for convergence is provided for the two-stage estimation algorithm described above.

Simulation Results and Discussion: Matrices ($A_k = A^C_k + A^I_k; k = 1, . . . , K$) are generated with spectral radius of 0.4, with non-zero effects having magnitude of at least 0.2. This would guarantee stationarity of generated time series, separate noise from the signal, and likely replicate the features of **fMRI** data from **ADHD** study (no estimated effects had magnitude larger than 0.4). Data is generated with independent $N(0,1)$ errors and signal-to-noise ratio of one (as in ADHD study). Multiple settings are considered by varying numbers p of variables per subject, T of observed time points, and K of subjects 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 individual component is set to 2-3% (of total number of matrix elements), implying a moderate level of heterogeneity. An example of generated transition matrices for the case of $p = 20$ can be found in the supplement. Cases of higher and lower spectral radius (meaning magnitude of non-zero effects) are considered in Appendix 9.1 and the supplement. Hard threshold of 0.02 is applied to final estimates to alleviate noise, all results are averaged over 50 replicates. Table 1 below describes how increase in number T of observed time points per subject affects the performance.

T	FP (comm)	FN (comm)	MC (comm)	FP (ind)	FN (ind)	MC (ind)	N Iter
p=20, *K*=10							
50	0.05(0.02)	0.15(0.22)	0.82(0.17)	0(0)	0.94(0.03)	0.16(0.05)	2.16(1.31)
80	0.06(0.02)	0.05(0.08)	0.89(0.06)	0(0)	0.88(0.06)	0.25(0.07)	2.17(1.05)
120	0.09(0.02)	0.01(0.02)	0.9(0.01)	0(0)	0.87(0.08)	0.24(0.1)	2.12(0.85)
p=30, *K*=10							
80	0.04(0.01)	0.01(0.03)	0.95(0.03)	0(0)	0.89(0.04)	0.23(0.05)	2.15(1.08)
120	0.06(0.01)	0(0.01)	0.94(0.01)	0(0)	0.89(0.04)	0.24(0.05)	2.13(0.97)
150	0.07(0.01)	0(0)	0.93(0)	0(0)	0.93(0.03)	0.19(0.04)	2.06(0.51)
p=40, *K*=10							
120	0.05(0.01)	0.01(0.01)	0.95(0.01)	0(0)	0.88(0.06)	0.24(0.06)	2.22(1.45)
150	0.06(0.01)	0(0)	0.94(0.01)	0(0)	0.92(0.04)	0.2(0.05)	2.13(0.95)
200	0.07(0.01)	0(0)	0.93(0.01)	0(0)	0.95(0.04)	0.14(0.06)	2.09(0.93)

Table 1: Two-stage procedure performance for increasing number T of time points, spectral radius of 0.4

We witness the tendency of denser common component estimates, which is manifested in decreasing false negative rate (e.g. from 15 to 1% for p=20) and increasing false positive rate (from 5 to 9% for p=20) within each setting. Meanwhile, the quality of individual component estimation is inconsistent, highly depending on the false positives of common component absorbing some of the subject-specific effects due to orthogonality assumption (see $p = 40$ setting). It shows that for a moderate signal strength (spectral radius of 0.4) the proposed method is more applicable for high-dimensional settings, with number T of time points not being much higher than the number p of variables per subject (making it a great tool for neuroimaging group studies). Another aspect is reflected in the Table 2 below - the effect of increasing the number K of subjects per group, which would allow us to borrow more strength across the group for estimation.

K	FP (comm)	FN (comm)	MC (comm)	FP (ind)	FN (ind)	MC (ind)	N Iter
p=20, *T*=80							
10	0.06(0.02)	0.05(0.08)	0.89(0.06)	0(0)	0.88(0.06)	0.25(0.07)	2.17(1.05)
20	0.03(0.02)	0.06(0.08)	0.92(0.06)	0(0)	0.6(0.07)	0.5(0.05)	2.81(1.94)
40	0.01(0.01)	0.05(0.09)	0.95(0.07)	0(0)	0.43(0.11)	0.63(0.08)	3.32(2.19)
p=30, *T*=120							
10	0.06(0.01)	0(0.01)	0.94(0.01)	0(0)	0.89(0.04)	0.24(0.05)	2.13(0.97)
20	0.03(0)	0(0)	0.97(0)	0(0)	0.43(0.04)	0.63(0.03)	3.1(1.59)
40	0.01(0)	0(0)	0.99(0)	0(0)	0.19(0.03)	0.82(0.02)	3.75(2.45)
p=40, *T*=150							
10	0.06(0.01)	0(0)	0.94(0.01)	0(0)	0.92(0.04)	0.2(0.05)	2.13(0.95)
20	0.03(0.01)	0(0)	0.97(0)	0(0)	0.43(0.07)	0.63(0.05)	3.21(1.18)
40	0.01(0.01)	0(0)	0.99(0)	0(0)	0.16(0.06)	0.85(0.05)	3.95(2.35)

Table 2: Two-stage procedure performance for increasing subject group size K, spectral radius of 0.4

With increase in the number of subjects per group we see a clear improvement in the common component estimation, which by the structure of our two-stage estimation approach leads to better estimates for individual effects as well, both of those aspects being reflected in respective **MC** values approaching

1. Results for higher spectral radius of 0.6 from Appendix and supplementary materials reaffirm the preference to increasing the number K of subjects over obtaining more time points T per subject. Meanwhile, in the the case of very small spectral radius (0.25, with minimum allowed effect magnitude of 0.1), we learn to emphasize collecting higher number T of time points per subject. Due to smaller signal from true underlying temporal effects being mixed up with the noise, longer time series is necessary for better temporal signal detection (see Appendix). On the other hand, increasing the number of observed subjects prior to obtaining long enough time series may lead to bigger confusion between signal and noise (see supplementary materials).

To sum up, the simulation studies demonstrate the preference for increasing the number of subjects in a study for moderate to high temporal signal strength, as opposed to increasing the number of observed time points, which is preferred for smaller temporal effect magnitudes. As a side note, the estimation algorithm always converges in under 10 iterations, averaging about three iterations per run.

Application to Resting-State Brain fMRI Data for ADHD Study

Primary intended application for the introduced framework of common structure is resting-state brain f**MRI** time series data in mental disorder studies. We look at experiments dealing with evaluation of inter-regional temporal effects within the brain for patients diagnosed with a certain mental disease (e.g., ADHD, Autism, etc), comparing those to a control group. We expect the patients from the diseased group (likewise controls) to have similar tendencies in brain temporal dynamics, while also displaying certain patient-specific features. The idea of shared structure had been used for f**MRI** data before [14,15], but it estimated contemporaneous dependence between brain regions (functional connectivity). Our estimation procedure attempts to describe temporal dynamics across those regions (effective connectivity).

Issues with Estimating Effective Connectivity of the Brain: Attempts at causal modeling of the cross-region relationships within the brain have been heavily scrutinized in the literature over the recent years. Even though there has been work done on dynamic causal modeling for estimating effective connectivity of the brain ([19, 20]), the more recent literature is quick to pinpoint various drawbacks for such approach. For example, a review paper on advances and pitfalls in resting-state f**MRI** analysis by [12] emphasizes the variations in haemodynamic delay across the regions which may introduce bias into any attempt of estimating causality. [13] proceeds to lay out six detailed reasons for caution when inferring causality for f**MRI** analysis, some of them being potential heterogeneity across different experimental sites and inability to capture causal relations due to neural activity processes occuring quicker than the sampling rate of f**MRI** measurements. Nevertheless, we will attempt to alleviate some of the issues discussed above with out joint estimation approach. First of all, the group lasso directly addresses one of the issues in [13] 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 f**MRI** measurements, otherwise risking to jointly estimate temporal effects of different lags. Additionally, the individual component accounts for subject-specific influences, e.g. sex 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.

Details of ADHD Study Setup: We will consider the resting-state brain fMRI time series data for 20 ADHD patients and 20 controls from the ADHD-200 Global Competition provided by Python module nilearn. The sample was collected across five experimental sites, three of which (NYU Child Study Center, Peking University and Radboud University for NeuroIMAGE study) shared a TR of 2.0s, with the other two (Oregon Health and Science University, Kennedy Krieger Institute) having a longer TR of 2.5s. 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 the subjects. That leaves us with 12 ADHD subjects (all males; age mean ± SD = 13.85 ± 3.83) and 12 controls (all males; age mean ± SD = 13.72 ± 3.72), respectively. All sites reported the 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 [21]. As mentioned in the discussion above, we proceed to parcellate the brain into 39 regions according to Multi-Subject Dictionary Learning atlas (MSDL), and following further the steps of [21] we summarize the signal over those regions via the mean of voxels weighted by gray matter probabilistic segmentation. Both the anatomical locations of the regions and resulting extracted time series can be found on Figures 1 and 2, with details on scientific names for each of those brain regions contained in Supplementary Table 3 of Appendix. Moreover, Figure 3 demonstrates autocorrelation plots corresponding to time series extracted for one of the regions. Obvious spike at the first time lag of PACF (bottom right plot) serves as a strong AR(1) signature, which is present for vast majority of brain regions under consideration. This leads us to believe in VAR(1) model, emphasized in our simulation studies, being the appropriate tool for this data.

Considering the lack of literature on distributional characterization and asymptotical properties of estimates resulting from group lasso procedure, we defer to conducting descriptive analysis via *stability selection*. The concept of stability selection was introduced in [22], with its main idea being the use of *bootstrap*, and subsequent accumulation of the results over all bootstrapped samples for further 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 obser

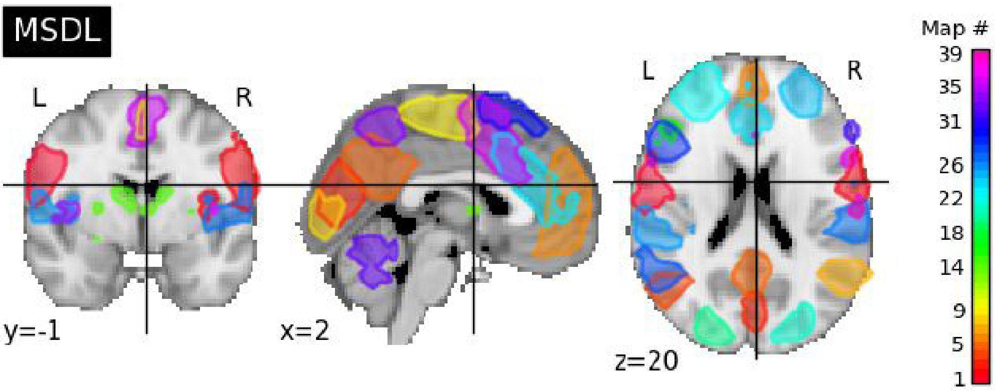


Figure 1: Brain regions according to MSDL atlas (Supplementary Table 3 for region names)

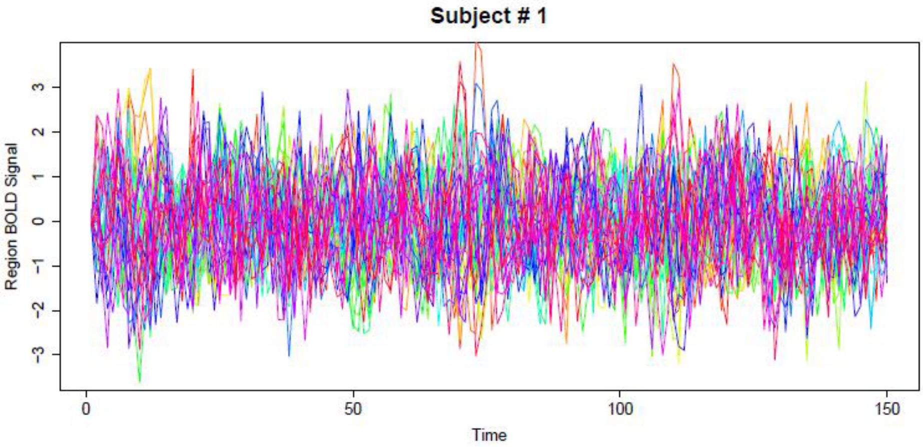


Figure 2: Extracted time series for 39 brain regions for a single subject

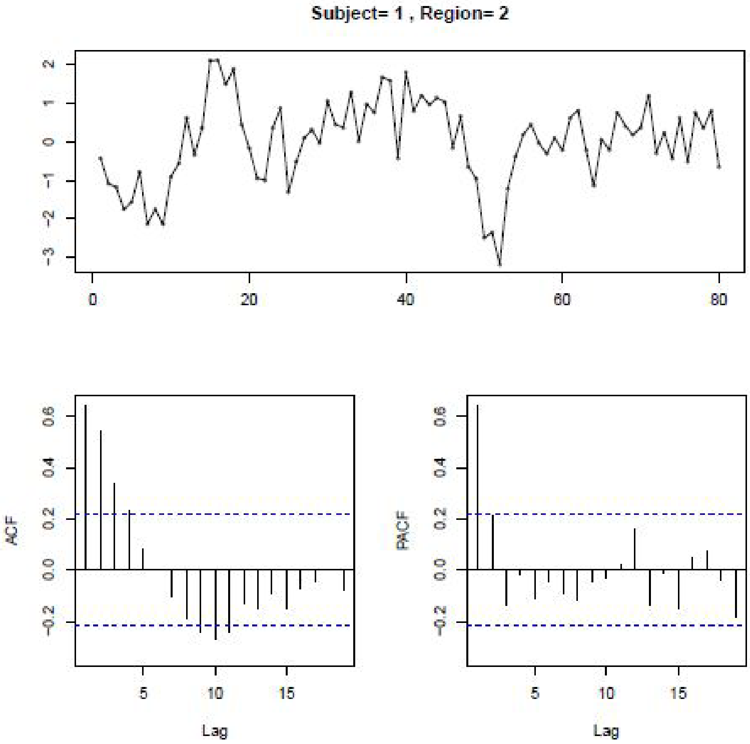
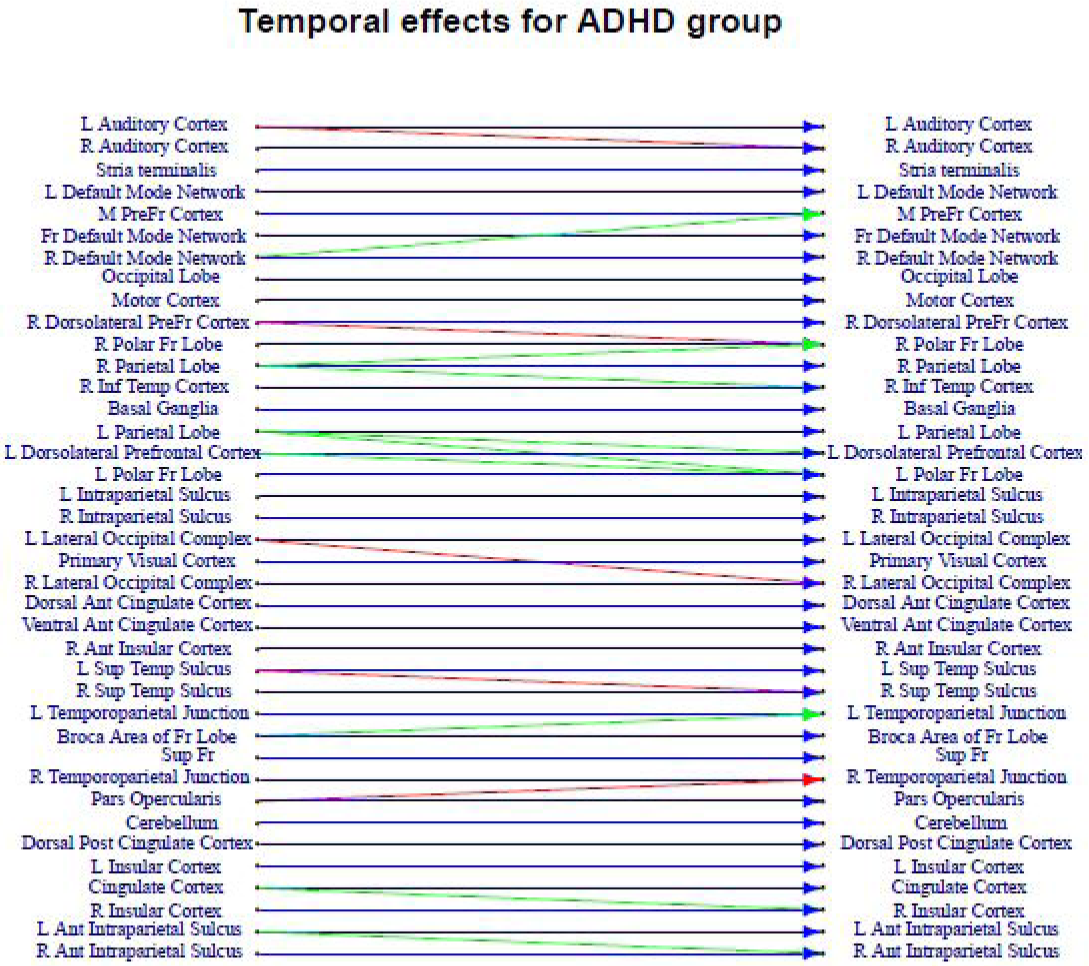


Figure 3: Autocorrelation plots for a single region time series: top - time series plot, bottom left - autocorrelation plot (ACF), bottom right - partial autocorrelation plot (PACF)

further 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 obser

ADHD Study Results and Discussion: Supplementary Table 2 demonstrates both 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 ± SD) 0:23 ± 0:1 for autocorrelation effects, 0:07± 0:05 for inter-regional effects. Going 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 [25], 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 on Supplementary Table 2. 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, discovered the patterns of hyperactivation [26], while [27,28] 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 [26], 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 [29,30], both pointing to abnormalities and lack of interactions in DAN as one of characteristics for ADHD patients. Additionally, emphasizes enhanced intraparietal sulcus activation for controls compared to ADHD patients [31]. 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 [32], with [33] 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.



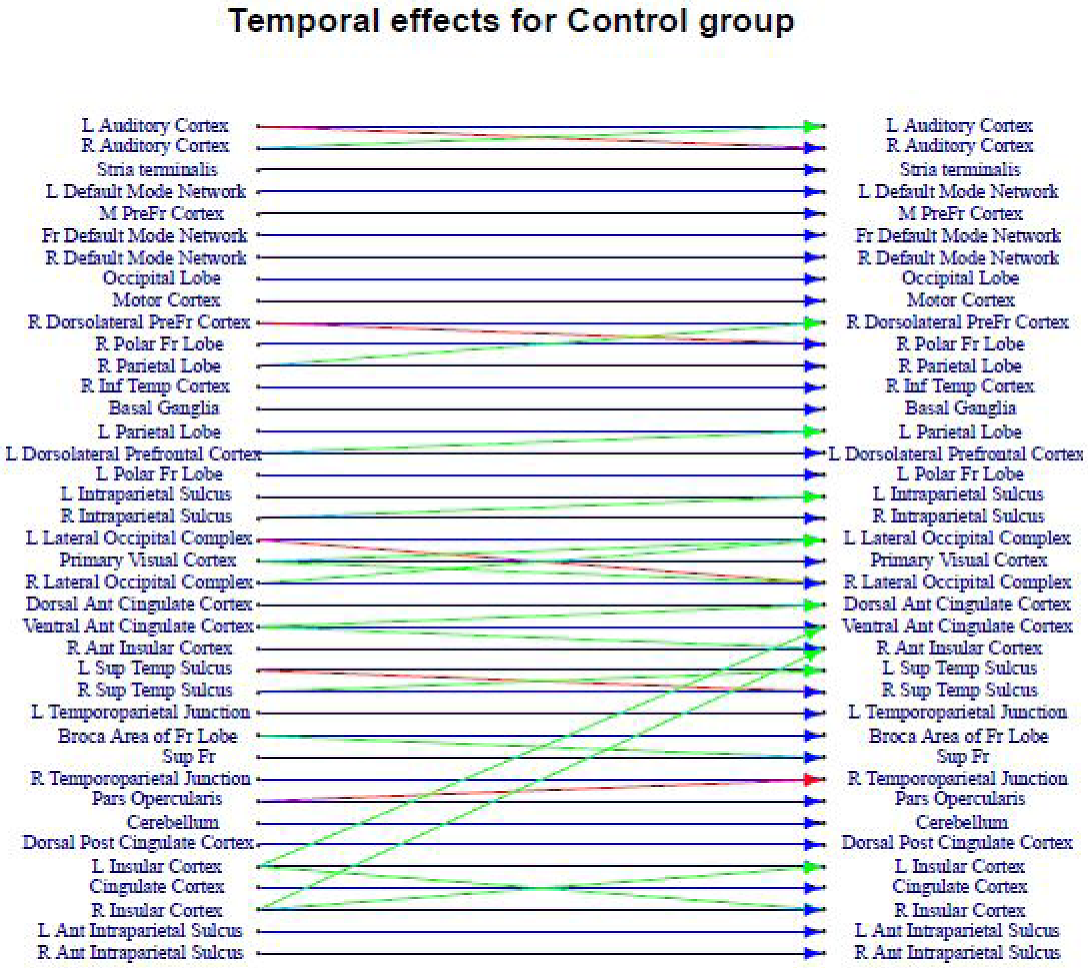


Figure 4: Directed graph of temporal effects for ADHD patients (top) and controls (bottom). Nodes on the left - brain regions at time t, right - at time t +1. Blue edge - region's autocorrelation effect; red - inter-regional effect present for both patient groups; green - group-specific inter-regional effect

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 [26], 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, point to association between attention lapses with reduced pre-stimulus activity in anterior cingulate regions [27], reinforcing the lack of inter-regional activation of those regions of the SN for ADHD group (while showing more communcation and also being influenced by members of CIN network for controls) in our study. Left and right superior temporal sulcus regionsdisplay temporal cross-effect between them for the control group, while only showing a left to right effect for ADHD, reaffirming the results of [26] who claim hypoactivity in temporal regions for ADHD patients. The last network showing 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 [26]. 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.

Referring to the individual component estimates that resulted from our two-stage procedure described, there happened to be no consistent subject-specific effects for either of the study participants (non were picked in more than 20% of bootstrapped samples). This may be attributable in part to the lack of heterogeneity in our dataset (all the subjects being teenage boys), and the simulation study for a similar setting (p=40; T=150; K=10) showing propensity for false negatives in individual component estimates.

Conclusion

In this work we present a regularized joint estimation procedure for the setting with multiple related VAR models being perturbations of a single underlying common VAR model. Even though the idea of joint estimation has been applied to precision matrices in [11], and regularized approach has been used to produce sparse estimates of VAR models in [6], they were yet to be combined into a joint procedure for estimating several homogeneous multivariate time series. Such approach allows one to borrow strength across multiple related subjects (time series) to produce more informed and, due to sparsity, more interpretable estimates. Primary intended application is brain fMRI time series 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 component, 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 componentestimation gradually gets better with increase in either the number of subjects per group, or observed time points per subjects (depending on generated strength of temporal signal).

While producing certain results that find confirmations in the literature, as discussed in the ADHD study results section, our methodology is not devoid of oversimplified, at times questionable, assumptions. First of all, by implementing VAR modeling framework we are not accounting for possibility of a non-linear temporal relationship between the brain regions. Further investigation would need to be done in order to establish whether the relationship is linear, and if not-how to model the case of non-linearity. Moreover, a simple averaging of a signal across the voxels within a brain region is done in order to get each separate time series in our ADHD study, which might be a severe oversimplification of true underlying signal structure of the brain. At last, while the iterative two-stage procedure introduced in Section 5.4 converged in practice (for all the replicates of both simulated and the ADHD study), we didn't explicitly state the conditions on the time series data in order to show its' convergence in theory. Same goes for the oracle properties of the estimates resulting upon that convergence, settling for the heuristic study results.

The most significant extension would be developing a hypothesis testing framework for the presented joint estimation procedure. 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 introduces a testing procedure for group-level effects that accounts for uncertainties introduced by both the regularization and inter-subject variability. Unfortunately, they don't use joint estimation approach and group lasso penalty, estimating brain networks separately with classic lasso, and therefore not applying directly to our procedure. The other aspect in need of further studying is the joint estimation procedure's ability to deal with VAR models of various lag orders, along with developing a lag order selection method.

Acknowledgement

This work was supported in part by Graduate Fellowship and by National Science Foundation [grantDMS-1632730].

References

1. Pavlov DY, Pennock DM (2003) A maximum entropy approach to collaborative filtering in dynamic, sparse, high-dimensional domains. Adv Neural Inf Process Syst, USA.

2. 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: e17191.

3. Michailidis G, d'Alche-Buc F (2013) Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. Math Biosci 246: 326-34.

4. Banbura M, Giannone D, Reichlin L (2010) Large Bayesian vector auto regressions. J Appl Econom 25: 71-92.

5. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Stats Soc Series B (Methodological) 58: 267-88.

6. Basu S, Michailidis G (2015) Regularized estimation in sparse high-dimensional time series models. Ann Statist 43: 1535-67.

7. Shuai H, Jing L, Liang S, Jieping Y, Adam F, et al. (2010) Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. NeuroImage 50: 935-49.

8. Narayan M, Allen GI, Tomson S (2015) Two sample inference for populations of graphical models with applications to functional connectivity. Stat ME 1502.03853.

9. Liu A, Chen X, Wang ZJ, McKeown MJ (2014) Time varying brain connectivity modeling using FMRI signals. In Acoustics, Speech and Signal Processing (ICASSP). IEEE Int conf pp.2089-2093.

10. Guo J, Levina E, Michailidis G, Zhu J (2011) Joint estimation of multiple graphical models. Biometrika 98: 1-15.

11. Danaher P, Wang P, Witten DM (2013) The joint graphical lasso for inverse covariance estimation across multiple classes. J Royal Stats Soc Series B (Statistical Methodology) 76: 373-97.

12. Cole DM, Smith SM, Beckmann CF (2010) Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. Front Syst Neurosci 4: 8.

13. Ramsey JD, Hanson SJ, Hanson C, Halchenko YO, Poldrack RA, et al. (2010) Six problems for causal inference from fMRI. Neuroimage 49: 1545-58.

14. Belilovsky E, Varoquaux G, Blaschko MB (2016) Testing for differences in Gaussian graphical models: applications to brain connectivity. In Advances in NIPS pp.595-603.

15. Chu SH, Lenglet C, Parhi KK (2015) Joint brain connectivity estimation from diffusion and functional MRI data. SPIE Medical Imaging 9413: 941321.

16. Lutkepohl H (2005) New introduction to multiple time series analysis, Springer Science & Business Media.

17. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J Royal Stats Soc Series B (Statistical Methodology) 68: 49-67.

18. Breheny P, Huang J (2009) Penalized methods for bi-level variable selection. Stat Interface 2: 369-80.

19. Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. NeuroImage 19: 1273-302.

20. Goebel R, Roebroeck A, Kim DS, Formisano E (2003) Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. Magn Reson Imaging 21: 1251-61.

21. Varoquaux G, Craddock RC (2013) Learning and comparing functional connectomes across subjects. NeuroImage 80: 405-15.

22. Meinshausen N, Buhlmann P (2010) Stability selection. J Royal Stats Soc Series B (Statistical Methodology) 72: 417-73.

23. Hardle W, Horowitz J, Kreiss JP (2003) Bootstrap methods for time series. ISI 71: 435-59.

24. Buhlmann P, Kunsch HR (1999) Block length selection in the bootstrap for time series. Comput Stat Data Anal 31: 295-310.

25. Serrallach B, Groß C, Bernhofs V, Engelmann D, Benner J, et al. (2016) Neural biomarkers for dyslexia, ADHD, and ADD in the auditory cortex of children. Front Neurosci 10: 324.

26. Cortese S, Kelly C, Chabernaud C, Proal E, Di Martino A, et al. (2012) Toward systems neuroscience of ADHD: a meta-analysis of 55 fMRI studies. Am J Psychiatry 169: 1038-1055.

27. Weissman DH, Roberts K, Visscher K, Woldorff M (2006) The neural bases of momentary lapses in attention. Nat Neurosci 9: 971-8.

28. Sato JR, Hoexter MQ, Castellanos XF, Rohde LA (2012) Abnormal brain connectivity patterns in adults with ADHD: a coherence study. PLoS One 7: e45671.

29. Castellanos FX, Margulies DS, Kelly C, Uddin LQ, Ghaffari, M, et al. (2008) Cingulate-precuneus interactions: a new locus of dysfunction in adult attention-deficit/hyperactivity disorder. Biol Psychiatry 63: 332-7.

30. Tomasi D, Volkow ND (2012) Abnormal functional connectivity in children with attentiondeficit/hyperactivity disorder. Biol Psychiatry 71: 443-50.

31. Hale TS, Bookheimer S, McGough JJ, Phillips JM, McCracken JT (2007) Atypical brain activation during simple & complex levels of processing in adult ADHD: an fMRI study. J Atten Disord 11: 125-39.

32. Zhu CZ, Zang YF, Cao QJ, Yan CG, He Y (2008) Fisher discriminative analysis of resting-state brain function for attentiondeficit/hyperactivity disorder. Neuroimage 40: 110-20.

33. Castellanos FX, Proal E (2012) Large-scale brain systems in ADHD: beyond the prefrontal-striatal model. Trends Cogn Sci 16: 17-26.