Senior Project Report 2102499 Year 2018

Learning Group Differences in Brain Networks using Statistical Methods

Parinthorn Manomaisaowapak ID 5830312921

Advisor: Assist. Prof. Jitkomut Songsiri

Department of Electrical Engineering

Faculty of Engineering

Chulalongkorn University

Abstract

Effective brain connectivity is a measure of causal interaction between brain regions. We applied concept of Granger causality (GC) on vector autoregressive (VAR) model to explain effective brain connectivity which can be represented as a matrix called GC matrix. We aim to learn zero pattern of GC matrix to extract significant effective brain connections. We considered the sparse estimation approach to learn sparsity patterns of GC matrix. However, if we have data from many groups, individual estimation of GC matrix does not use any prior knowledge on group similarity. For this reason, we studied three joint estimation formulation to estimate multiple sparse GC matrices from multiple groups simultaneously. The first formulation returns models that share the same pattern of GC matrix, while the second formulation provide models that have both common and differential patterns. The problem is a least-squares estimation with a sum of ℓ_2 norm regularization and known as group lasso problem. The third formulation promotes the estimated models that partially share the same parameters value. The problem is a least-squares estimation with two regularization terms; sum of ℓ_2 norm of parameters and ℓ_2 norm of parameter differences. These three formulations are tested on simulated data to find performance and consistency of each formulation in different scenarios. The results confirm that each formulation performed at its best when the assumed prior on data is correct.

Keywords: Fused lasso, Group lasso, Sparse vector autoregressive model, Granger causality

Contents

1 Introduction

Effective brain connectivity is one of brain topological structure definitions that reflects how each region inside human brain interacts to each other. By learning brain connectivity pattern of patients who have neurological disorders such as Alzheimer's disease, autism, long term traumatic brain injury (TBI), we may detect those disorders before it occurs to raise medical attention beforehand. This is what make us interested in learning brain connectivity. There are two popular measure on effective brain connectivity which are Granger causality (GC) based and Gaussian graphical model based. GC based connectivity is a way to measure directed relation between regions in the brain that the forward connection and reverse connection between two regions may not be the same. GC measures can be characterized on zero patterns of vector autoregressive (VAR) and can be represented as a matrix that entry i, j is GC measure from region j to region i. Gaussian graphical model treated data as static model and used zero pattern of inverse covariance matrix to represent conditional independence structure. This definition of effective brain connectivity is undirected. After we are able to learn brain connectivity of those who have neurological disorders, we can find the differences of those who have the diseases and those who are healthy.

In this project, we considered effective brain connectivity using GC measure on sparse VAR model. After brain connectivity of each subject is learned, the differences of the healthy subjects from who are not can be detected. We consider two frameworks, statistical framework and sparse estimation frame work. In statistical framework, the connectivity matrices are computed from each patient in each group individually and find average value of brain connectivity measures as the representation of the whole group. After that we will perform hypothesis test using Hotelling's T^2 statistics between group whether there are statistical significant different. In sparse estimation framework, It brought us to interest because of three reasons. The first reason is based on problem dimensionality such as in fMRI data that every voxel (3D pixel) contain signal that is time series. Even if the problem dimensionality were reduced by clustering all those signal as the regions of interest (ROIs), the dimension of the model still be high. The estimation of Gaussian graphical models usually results in a dense effective brain connectivity [THW$^+$16] which is hard to interpret. One way to reduce complexity of the problem is to extract significant connectivity in the model. In graphical model, extracting significant connectivity can done by penalizing the off-diagonal parameters of brain connectivity matrix in the log-likelihood optimization problem by ℓ_1 norm penalty. The result is well-known that the final solution tends to have more zero entries which makes model more parsimonious. This formulation is called graphical lasso [FHT07]. In our statistical framework, we excluded the insignificant connections by using hypothesis test. Our second reason is that sparse estimation framework is a regularized optimization problem. The regularization can be interpreted as adding prior knowledge to the estimation as in Bayesian statistics. If the prior knowledge is true, the quality of estimation can be maintained with lower number of samples. This leads to our final reason. The sparse estimation is able to estimate multiple models simultaneously where we can put regularization terms such that each model have relation to another model. This is called joint estimation of brain connectivity.

There are several ways to learn multiple effective brain connectivities such as learning brain connectivity on graphical model or on VAR model. There are many literature that proposed joint estimation of multiple brain connectivities based on Gaussian graphical model. In [THW$^+$16], they proposed method to estimate multiple graphical models simultaneously to exploit the similarity and sparsity of the models in the way that the connection strength of brain connectivity of similar model need not to be equal. In [DWW14], they proposed two penalty functions that can encourage likeness of brain connectivity and also encourage each model to have different sparsity pattern. Those two formulations they proposed are called *fused graphical lasso* and *group graphical lasso*. Both penalty functions are able to encourage each model to have different sparsity pattern and able to encourage likeness of effective brain connectivity. *Fused graphical lasso* formulation defines likeness of effective brain connectivity by having similar brain connectivity strength. The likeness in *group graphical lasso* formulation is defined by common sparsity pattern that all brain connectivities in the group have in common. In [GLMZ11], they proposed method to re-parametrized parameters of multiple inverse covariance matrix such that decomposed parameters into product of non-negative term and some value. The non-negative term is shared to all models. Sum of the non-negative term penalization can be interpreted as ℓ_1 regularization and be understood as common sparsity pattern encouragement. Another decomposed parameters were penalized to encourage difference in sparsity pattern in each model. They also proved that their cost function has same minimizer as regularization with single non-convex function that can encourage both common sparsity pattern and different sparsity pattern. These literature have brain connectivity based on Gaussian graphical model. Gaussian graphical models used conditional independence to learn sparsity pattern of brain connectivity. However, bio-signals such as fMRI, EEG are generated from unknown dynamical models but graphical model treats data as random vectors or treats brain signals as generated from static model which may not be realistic assumption. We consider sparse VAR estimation as one of dynamical models to be more realistic. There are multiple literature sparse VAR joint estimation. In [Son15], the author exploited the sparsity pattern in brain connectivity matrix and estimates all group's brain network simultaneously using regular ℓ_1 regularization method to increase sparsity of VAR parameters that directly relate to structure of effective brain connectivity, while controlling the brain network to be similar in multiple groups. However, if true model parameters are not close to each other, this formulation can produce parameters bias. In [SM18], they proposed two-step estimation to estimate two part of VAR parameters which are parameters that are common to all models and parameters that are specific to each model. They formulated the first stage problem as group lasso on multiple VAR models estimation to force common sparsity. In second step, they fitted VAR independently to the residuals that came from model bias of estimated common VAR parameters. However, all model parameters were not estimated simultaneously. The estimated parameters in each step are optimal in their estimation step but all parameters need not to be optimal. We proposed a method to estimate all model parameters of GC matrix simultaneously which all estimated parameters are optimal.

We consider two frameworks that are based on VAR model: statistical and sparse estimation frameworks. In statistical framework, the overall process will be described in *Figure* 1. In statistical framework The data will be arranged into two groups, patient group and control group. In each group, we will estimate GC matrix of each trial individually, the group GC matrix will be obtained by averaging all GC matrices through all trials in the group and the difference will be detected by statistical test. We will use Multivariate Granger Causality toolbox (MVGC) [BL14] to compute Granger causality matrix as a measure of effective brain connectivity denoted as GC matrix. The GC matrix is based on vector autoregressive model because of model simplicity and EEG data are time series which are dynamic model. The statistical approach is already finished [MS18]. In sparse brain estimation framework, the overall process will be described in *Figure* 2. In this framework, we study three regularization techniques in [Son17] and we extended group fused lasso formulation in this work in the sense that our formulation is more general. We consider only estimation on simulated data in sparse estimation framework because the accuracy and performance of the formulation can be directly measured. However, the group differences in this framework can be obtained through comparison of the pattern in the GC matrix using knowledge in neuroscience. In this report, we presented sparse estimation framework only and the detail of this framework is in methodology section.

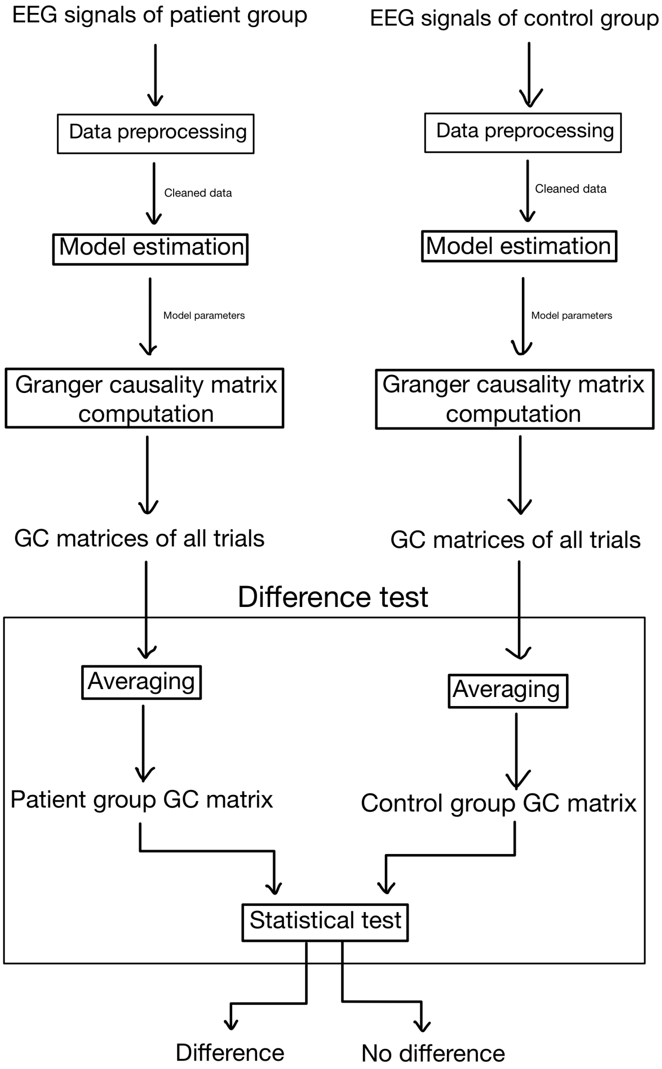


Figure 1: Statistical framework for learning brain network differences. The results of this framework is in [MS18]

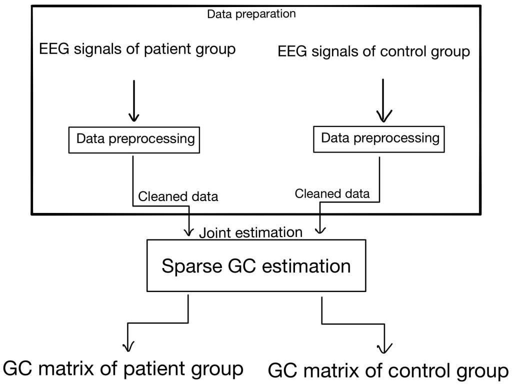


Figure 2: Sparse estimation framework for learning brain network differences.

2 Objectives

The objectives of this project are

1. To estimate brain network using Granger causality concept from EEG or fMRI data.
2. To compare brain network difference between two groups, control and patient group.

3 Background

This section contains two topics. First topic describes common knowledge of Granger causality measure in its general form and the second topic describe how sparsity pattern in vector autoregressive model's coefficients can infer to the sparsity pattern of Granger causality matrix.

3.1 Granger causality estimation

Granger causality is a concept that test if the past of one time series can help to predict another time series in sense of reducing the residual variance of the predicted time series. In this project, we will compute Granger causality based on vector autoregressive model. Vector autoregressive model order p is defined as

$$y(t) = \sum_{k=1}^{p} A_k y(t-k) + e(t) \tag{1}$$

where $y(t), e(t) \in \mathbb{R}^n, A_k \in \mathbb{R}^{n \times n}$ AR model can be expressed in state-space representation as

$$x(t+1) = \begin{bmatrix} A_1 & A_2 & \dots & A_{p-1} & A_p \\ I & 0 & \dots & 0 & 0 \\ \vdots & \ddots & & \vdots & \vdots \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \dots & I & 0 \end{bmatrix} x(t) + \begin{bmatrix} e(t+1) \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{2}$$

with $x(t) = \begin{bmatrix} y(t-1)^T & y(t-2)^T & \dots & y(t-p)^T \end{bmatrix}^T$, and the output equation is

$$y(t) = \begin{bmatrix} I & 0 & \dots & 0 & 0 \end{bmatrix} x(t), \tag{3}$$

VAR model parameter can be estimated by ordinary least square methods or solve via Yule-Walker equation [BJ76].

Next, we will introduce the concept of Granger causality test. Let's consider a multivariate AR(1) process $y_i(t), y_j(t)$, note that both are vector. We want to investigate if $y_i(t)$ is depended only in its own past value, not from past of $y_j(t)$. The full fitted VAR model is

$$\hat{y}_i(t) = A_{ii} y_i(t-1) + A_{ij} y_j(t-1)$$

$$\hat{y}_j(t) = A_{ji} y_i(t-1) + A_{jj} y_j(t-1)$$

which can be expressed as

$$\hat{y}(t) = \begin{bmatrix} A_{ii} & A_{ij} \\ A_{ji} & A_{jj} \end{bmatrix} y(t-1). \tag{4}$$

Note that $y(t) = \begin{bmatrix} y_i(t)^T & y_j(t)^T \end{bmatrix}^T$. The model that remove the testing parameter that represent the connection between y_i and y_j is called the reduced model. In this scenario, the reduced model is

$$\hat{y}^R(t) = \begin{bmatrix} A_{ii}^R & 0 \\ A_{ji}^R & A_{jj}^R \end{bmatrix} y^R(t-1). \tag{5}$$

One can understand that more complex model has more flexibility, *i.e.* more parameters to be determined, than a simpler model. In this case, the full model is more complex than the reduced model therefore, the residual's variance of full model should be smaller than the reduced model. In other words, the full model is expected to have more model quality than the reduced model. However, if residual's variance of reduced model is close to the full model, this can be inferred that the removed parameter is not significant. As in (5), the parameter A_{ij} implies direct impact from the past of y_j to current y_i. If A_{ij} is removed and model quality does not reduce, then the past of y_j does not have direct effect on y_i. On the contrary, if the model quality is reduced on reduced model. Then y_j should have direct effect on y_i but to answer how the strong the effect is, at this point, the concept of Granger causality arises as the measure of the strength. In multivariate Granger causality, the residual variance will be changed into the measure that reflects how large the residual covariance matrix is. In [BBS10], they used the concept of generalized variance which is determinant of covariance matrix and denoted the uses of total variance, *i.e.* trace of covariance matrix. The Granger causality measure is defined as

$$\mathcal{F}_{ij} = \log \frac{\det \Sigma_{ii}^R}{\det \Sigma_{ii}}. \tag{6}$$

Granger causality can be tested by a log ratio of the generalized variance which represents model quality of the reduced model compared to the full model. this measure is, in general, defined by (6) which is multivariate version of Granger causality with residual covariance matrix of reduced model Σ_{ii}^R and the residual covariance matrix of full model Σ_{ii}. Since condition $\mathcal{F}_{ij} = 0$ is denoted as j th time series does not have Granger causal relation to i th time series. This is where we can exploit from. If the model is based on VAR model in equation (1), the following statement will hold [Lüt05].

$$\mathcal{F}_{ij} = 0 \leftrightarrow (A_k)_{ij} = 0, k = 1, 2, ..., p. \tag{7}$$

From this point of view, the sparsity pattern in GC matrix can be used in VAR estimation. Only ordinary least square estimation of VAR model cannot inferred to zero patterns in GC matrix because it will be hardly exact zero in least square solution. By using knowledge on statement (7), we can put a regularization on the VAR model's cost function to force $(A_k)_{ij} = 0$ for all lags to be zero together. The selection of regularization function is crucial and will be discussed in next section.

3.2 Joint sparse VAR estimation

In sparse estimation framework, it is known that the effective brain connectivity is expected to have sparse structure [Spo07] and human brain is expected to have similar effective brain connectivity in healthy brain area therefore, in model estimation, the sparse structure can be achieved by adding ℓ_1 norm [HTW15] of parameters as penalty term with penalty parameter to control sparsity to optimization formulation in parameter estimation. The similar structure can be achieved by adding ℓ_1 norm of parameters differences between groups so that the parameters are more likely not to change a lot causing the structure of both groups is similar to each other and the difference of brain connectivity can be observed. For simplification in explanation, the example of estimation method for only two models has a general structure as

$$\underset{\theta_1,\theta_2}{\text{minimize}} \quad f_1(\theta_1) + f_2(\theta_2) + \lambda g_1(\theta_1, \theta_2) + \gamma g_2(\theta_1, \theta_2) \tag{8}$$

where θ_1, θ_2 are the parameters that inferred to brain connectivity matrices such as VAR parameters or inverse covariance matrix in graphical model. If the parameters that coupled between multiple brain signals are sparse, the causal inference between them is expected to be sparse. $f_i(\cdot)$ is defined as cost function that represent the goodness of fit of model. In this case λ and γ are regularization parameters. $g_1(\cdot)$ is function that penalized the nonzero model parameters in the way that the solution of the optimization problem is sparse. One of function that has this property is ℓ_1 norm which is referred as lasso or group lasso penalty [HTW15]. The explicit form is

$$g_1(\theta_1, \theta_2) = \|\theta_1\|_1 + \|\theta_2\|_1$$

. $g_2(\theta_1, \theta_2)$ is function that penalized the difference in model parameters θ_1 and θ_2 which ℓ_1 norm of difference in model parameters can be used. The explicit form is

$$g_2(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_1$$

. However, the penalization of model parameters difference can be bias if the value of parameters in each model is not close to each other. The term $g_2(\cdot)$ can be changed to the form that encourage common sparsity pattern in GC matrix. The explicit form is

$$g_2(\theta_1, \theta_2) = \sum_i \left\| \begin{bmatrix} (\theta_1)_i \\ (\theta_2)_i \end{bmatrix} \right\|_2$$

when multiple models are simultaneously estimated where i is index of the parameters that infer to index in GC matrix. This also known as the regularization term in group lasso problem. The likeness among the models cannot be achieved if the models were individually estimated. This extension will be discussed in the following section.

4 Methodology

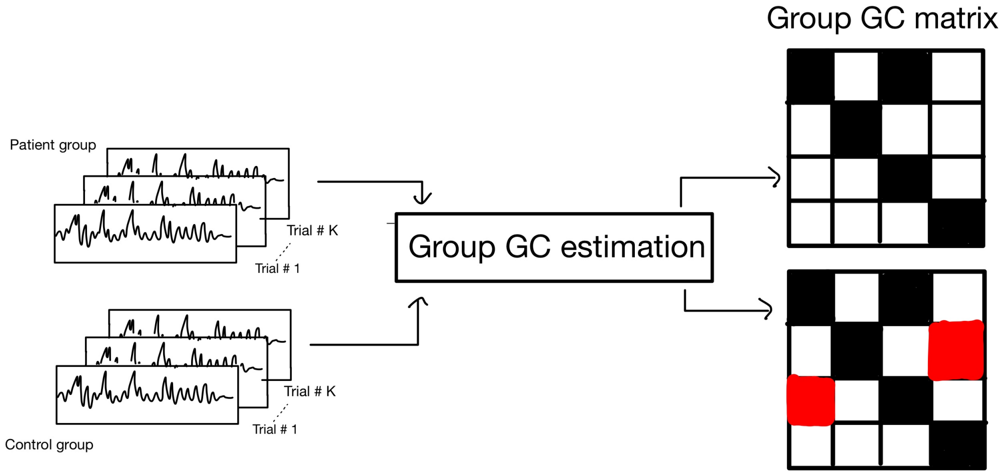


Figure 3: The detail each steps of methodology in sparse estimation framework.

The detail of sparse estimation framework will be described in section 4.1. In this framework, we will compute group GC matrix directly from two groups simultaneously. We considered 3 steps as follow

1. Jointly sparse VAR estimation of brain networks

   This step will describe how to achieve sparse VAR estimation for multiple groups and how regularization should be.

2. Model selection for learning brain networks

   This step will select the tuning parameters which play crucial role in consistency of zero patterns.

3. Simulated data generation

   We will describe how we generated data to test each formulation in this step.

The brain network differences may be determined by comparing the common sparsity pattern of GC matrices of multiple models between group. However, The sparse estimation will only be tested on simulated data to verify performance index of each formulation. Methodology of this framework will follow as *Figure* 3. The estimation method is different from statistical framework because the sparse estimation framework estimates GC matrices of all models simultaneously which allows the regularization on coupling terms between models and this will introduce likeness of model such as similarity in parameters value or shared common sparsity index in GC matrices. The difference of brain network can be determined whether in qualitative, the result interpretation by background knowledge in neuroscience or in quantitative such as the statistical test or the distance measure between brain network of two groups [VBK$^+$10], note that the distance does not necessary to be Euclidean distance. However, we only focus on the estimation process in this framework. The steps of this framework are described in the following section.

4.1 Jointly sparse VAR estimation of brain networks

In VAR estimation, we considered the Ordinary least square (OLS) solution. Ordinary least square is a solution of the over determined system. In this case, the linear system is

$$\begin{bmatrix} y(p+1) & y(p+2) & \dots & y(N) \end{bmatrix} = \begin{bmatrix} A_1 & \dots & A_p \end{bmatrix} \begin{bmatrix} y(p) & y(p+1) & \dots & y(N-1) \\ \vdots & \vdots & \dots & \vdots \\ y(2) & y(3) & \dots & y(N-p+1) \\ y(1) & y(2) & \dots & y(N-p) \end{bmatrix} \tag{9}$$

This is in the form $Y = AH$, $Y \in \mathbb{R}^{n \times (N-p)}$, $H \in \mathbb{R}^{pn \times (N-p)}$, $A_i \in \mathbb{R}^{n \times n}$ where n is the dimension of time series and N is number of data points, p is model order.

Then the least square optimization formulation is

$$\underset{A}{\text{minimize}} \quad (1/2)\|Y - AH\|_F^2 \tag{10}$$

The least square solution $\hat{A}$ can be solved analytically by solving the normal equation.

$$\hat{A}(HH^T) = YH^T \tag{11}$$

For sparse VAR estimation, it is known that the zero pattern at index (i,j) in GC matrices can be inferred that the value of $(A_q)_{ij} = 0 \ \forall q = 1,...,p$ which has been proved in [Lüt05]. To learn sparsity pattern of GC matrix, we can regularized the sum of euclidean norm of vector $\begin{bmatrix} (A_1)_{ij} & \cdots & (A_p)_{ij} \end{bmatrix}$ for each $i \neq j$. The regularized cost function will be

$$\underset{A}{\text{minimize}} \quad (1/2)\|Y - AH\|_F^2 + \lambda \sum_{i \neq j} \|(A_1)_{ij} \quad \cdots \quad (A_p)_{ij}\|_2 \tag{12}$$

which is called Group lasso[YL06] that is the extension of original lasso. However, for multiple VAR models estimation, the individual estimation of VAR model alone does not use prior knowledge on how each model is relate to each other such as having common sparsity structure or having similar parameters value, so the regularization terms that encourage those relation arise as the estimation of multiple VAR models with joint penalty. The regularized least square formulation of multiple VAR model that are estimated independently [Son17] is

$$\underset{A}{\text{minimize}} \quad \sum_{k=1}^{K} (1/2)\|Y^{(k)} - A^{(k)}H^{(k)}\|_F^2 + \lambda \sum_{i \neq j} \sum_{k=1}^{K} \|(A_1^{(k)})_{ij} \quad \cdots \quad (A_p^{(k)})_{ij}\|_2. \tag{13}$$

To have more compact representation and further formulation development, consider the following. Define $B_{ij}^{(k)} = \begin{bmatrix} (A_1^{(k)})_{ij} & (A_2^{(k)})_{ij} & \cdots & (A_p^{(k)})_{ij} \end{bmatrix}^T$. By using this notation, the cost function in equation (13) can be expressed as

$$\underset{A}{\text{minimize}} \quad \sum_{k=1}^{K} (1/2)\|Y^{(k)} - A^{(k)}H^{(k)}\|_F^2 + \lambda \sum_{i \neq j} \sum_{k=1}^{K} \|B_{ij}^{(k)}\|_2. \tag{14}$$

By seeing that the term $\|B_{ij}^{(k)}\|_2$ is the individually regularization of each model, we define this as differential regularization because it encourages single model to have different sparsity from another. At this point, we proposed another two regularization terms that encourage likeness between models, the common sparsity structure and similar parameters value structure. The common sparsity structure can be achieved by grouping variables that will cause entry (i,j) of GC matrix to be zero to all models resulting in the terms $B_{ij}^{(k)} = 0, \forall k$ which can be formulated as group lasso problem as follow.

- **Formulation C**

  Define $\mathrm{C}_{ij} = \begin{bmatrix}(B_{ij}^{(1)})^T & (B_{ij}^{(2)})^T & \cdots & (B_{ij}^{(K)})^T\end{bmatrix}^T$. The formulation for common sparsity structure will be

$$\underset{A}{\text{minimize}} \quad \sum_{k=1}^{K} (1/2)\|Y^{(k)} - A^{(k)}H^{(k)}\|_F^2 + \lambda \sum_{i\neq j} \|C_{ij}\|_2. \tag{15}$$

  By adding differential term, the formulation (15) became

- **Formulation D**

$$\sum_{k=1}^{K} (1/2)\|Y^{(k)} - A^{(k)}H^{(k)}\|_F^2 + \lambda_1 \sum_{i\neq j}\sum_{k=1}^{K} \|B_{ij}^{(k)}\|_2 + \lambda_2 \sum_{i\neq j} \|C_{ij}\|_2. \tag{16}$$

- **Formulation S**

  Furthermore, the likeness of multiple models can be described by the similarity in VAR parameters value. In [Son17], the author proposed group fused lasso regularization. However, the regularization technique used in [Son17] is limited only if the similar models are consecutive to each other. We extended by changing group fused lasso order to match all models together for forcing similarity across all models. We propose

$$\sum_{k=1}^{K} (1/2)\|Y^{(k)} - A^{(k)}H^{(k)}\|_F^2 + \lambda_1 \sum_{i\neq j}\sum_{k=1}^{K} \|B_{ij}^{(k)}\|_2 + \lambda_2 \sum_{k<k'}\sum_{i\neq j} \|B_{ij}^{(k)} - B_{ij}^{(k')}\|_2. \tag{17}$$

This formulation encourages both differential sparsity and similarity of all models. In brain data such as fMRI of multiple group of patients where each group represents a different condition. The patients in the same group should have likeness in their brain connectivity but each subject is also a different person. So, it should be better if both likeness and differential regularization terms. We will also compare a formulation that assumes only common sparsity pattern.

4.2 Algorithm

To make the optimization problem (15),(16), (17) to be compatible with existing algorithm used in [Son17], the matrix form of optimization problem can be vectorized. We vectorized in the way that is easy to implement as a group lasso and group fused lasso problem. The vectorized version of (15),(16), (17) cost function will be

- **Vectorized formulation C** 15

$$(1/2)\|b - Gx\|_2^2 + \lambda_1\|Px\|_A + \lambda\|Px\|_B \tag{18}$$

- **Vectorized formulation D** 16

$$(1/2)\|b - Gx\|_2^2 + \lambda_1\|Px\|_A + \lambda_2\|Px\|_B \tag{19}$$

- **Vectorized formulation S** 17

$$(1/2)\|b - Gx\|_2^2 + \lambda_1\|Px\|_A + \lambda_2\|Dx\|_A \tag{20}$$

where $x = \begin{bmatrix}C_{11}^T & C_{12}^T & \cdots & C_{n,n-1}^T & C_{n,n}^T\end{bmatrix}^T$, Px is off-diagonal terms of x and D is difference matrix that is equivalent to fused term in (17). Given $Z = (z_1, \cdots, z_L), z_k \in \mathcal{R}^w$. $\|Z\|_A = \sum_{k=1}^L \|z_k\|_2$ for $w = p$ and $\|Z\|_B = \sum_{k=1}^L \|z_k\|_2$ for $w = pK$. b is vectorized version of Y.

Problems 18, 19, 20 are known optimization problems as lasso type optimization problem. Both problems are convex optimization problems but non-differentiable. The discontinuity of gradient cause gradient method to fail. By convexity of the problem, we may use **cvx** software [Inc12] to solve the problem. However, cvx solver suffers from memory storage problem for large scale problem. We used ADMM algorithm derived in [Son17] to solve 18, 19, 20. One thing to decide is how to select ADMM penalty parameter to achieve good performance on convergence rate. By using rule of thumb, we fixed the penalty to be 10 times more than the highest value of regularization parameters.

4.3 Model selection for learning brain networks

The crucial step to estimate models correctly is the regularization penalty selection step. There are literature on two penalty parameters selection for various models such as elastic-net, structured lasso. In [LLSL18] they tuned elastic-net penalty by 5-fold cross-validation and find maximum average-over-5-fold area under curve (AUC) of ROC plot among given range of parameters and then selected the parameters that is within 1 standard deviation from the parameters that yielded maximum average-over-5-fold AUC for more parsimonous model. In [GMCW13], they selected two parameters from problem group-subgroup lasso that is an extension of original group lasso problem which has three parameters $(\lambda, \alpha_1, \alpha_2)$ but can be reduced to two parameters as their special case$(\alpha_1 + \alpha_2 \leq 1 \rightarrow \alpha_1 + \alpha_2 = 1)$. They have parameter that control sparsity level(λ) and parameters that are trade-off sparsity level of each other (α_1, α_2). They derived a bound that gives sparsest solution, or all zero solution then selects λ by fixing α_1, α_2 and select λ^* that minimize Mallow's C_p criterion

$$M_n(p^*) = \frac{SSE_{p^*}}{\hat{\sigma}^2} - n + 2p^* \tag{21}$$

where p^* is degree of freedom of estimation, which is nonzero entries in the estimated parameters, SSE is sum squared error. And α_1, α_2 were selected by 10-fold cross validation. However, These methods contained cross validation which is computationally infeasible in large scale problem [CGC12] so the information theoretic criterion such as AIC, BIC are preferred. In [WC18], they proposed jointly estimated VAR models that have term fused lasso between models controlled by λ_1 and λ_2 for lasso. In each model, they used BIC criterion described as

$$\mathrm{BIC}(\lambda_1, \lambda_2) = -2\mathcal{L}(\lambda_1, \lambda_2) + \mathrm{df}(\lambda_1, \lambda_2) \log (N) \tag{22}$$

$$\mathcal{L} = -\frac{N}{2} \log \det \Sigma - \frac{1}{2} \mathrm{tr}(e^T \Sigma^{-1} e); e = Y - AH \tag{23}$$

where $\mathcal{L}$ is non-constant term of log-likelihood function of the VAR model and df is the effective degree of freedom of VAR parameters which, in this literature, is number of non-zero VAR parameters. However, In [ZHT07], The degree of freedom is a measure of model complexity and the consistency of information theoretic criterion is depended on choice of effective degree of freedom. In their work, the original lasso problem, the df is given by the number of nonzero of estimators which was derived from assumption that required uncorrelated Gaussian measurement $(y \sim \mathcal{N}(\mu, \sigma^2 I)$ in $y = Ax + \epsilon)$. The measurement from autoregressive process are expected to be correlated to each other in the sense of dynamic model that depended on their past value. In spite of this assumption, there are other literature that still used effective degree of freedom of VAR model or other variation of lasso problem as number of nonzero of estimators such as in [GMCW13], [RXZ13], [WC18]. There are literature that estimated effective degree of freedom of the sparse estimation. As a group lasso problem in [YL06], they proposed method to estimate degree of freedom as

$$\tilde{df} = \sum_j I(\|\beta_j\| > 0) + \sum \frac{(\|\beta_j\|}{(\|\beta_j^{LS}\|}(p_j - 1) \tag{24}$$

where I is indicator function, β_j is group of predictors size p_j as in group lasso problem and β_j^{LS} is least square solution of those group of predictors. However, in [SM18], they proposed regularized joint estimation on multiple VAR model with BIC for penalty selection but they calculated the degree of freedom as in [BH09]. In fused-lasso problem, the degree of freedom can be estimated by number of the nonzero identical parameters or zero value difference but the value is nonzero [JTT10].

Each model selection method can used to select regularization variables that is the best in the sense of optimal trade-off between model fitting and model complexity and some in sense of consistency under some conditions. In sparse estimation on real data set, we cannot ensure that the selected regularization parameters are good enough to reject all insignificant predictors or accept all significant predictors. At this point, the concept of stability selection is introduced. The stability selection [MB10] is the method that gives bound of number of false discoveries or false nonzero predictors, by subsampling the data set and estimate the model with subsampled data at single regularization level multiple time to estimate probability of accepting the nonzero value. To be more clear, this method finds probability that the estimated predictors are nonzero (significant). Those predictors that have high probability will be called stable predictors. This method can be used as model selection criteria or used with another method to control amount of false nonzero estimation. In our formulation, we preferred to used BIC over K fold cross-validation because computational complexity and BIC selection is consistent [CG11]. In formulation C, D, we will use degree of freedom as the number of effective parameters. In formulation S, we will compare two choices on degree of freedom, the number of effective parameters and the number of nonzero parameters that are equivalent or are fused to each other. To be clear, the choices of degree of freedom are

- df = number of nonzero parameters in the model (Lasso).
- df = number of fused nonzero parameters in the model (fused Lasso).

In the experiment, we will use grid search in each formulation to find optimal regularization parameters by BIC score. In formulation C, we perform 1D-search on λ_2 in formulation C. We used range of the search as searching until we obtained the sparsest solution. In [Son17], they have derived a necessary and sufficient bound of the value λ_2 in formulation C. We will denote the bound as λ_c. The 1D searching will vary λ_2 as $\alpha\lambda_c$ where $\alpha \in [0, 1]$. In formulation D, we will use the this bound to vary λ as $\lambda_1 = \alpha\lambda_c$ and $\lambda_2 = \beta\lambda_c$ where α, β are in range [0,1]. In formulation S, we empirically selected λ_c to be the same as in formulation C and D because the bound we derived for this formulation is only necessary for the sparsest solution in value and in difference between model or the most parsimonious solution so that solution will necessary to have λ that satisfy the bound (e.g. more than the bound) but the bound does not sufficient to conclude that the λ above the bound will be the most parsimonious solution.

In problem 18, 19, 20, we rearranged those parameters as

$$x = \begin{bmatrix} x_p \\ xq \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_L \end{bmatrix}, \quad G = \begin{bmatrix} G_p & G_q \end{bmatrix} = \begin{bmatrix} G_1 & \cdots & G_L . \end{bmatrix} \tag{25}$$

where x_p are VAR coefficients that are penalized in penalty term and x_q are diagonal terms of VAR coefficients.

The critical value (λ_c) in (15) that is derived in [Son17] has formula as

$$\lambda_c = \max_{k \in \{1,2,\ldots,L\}} \|G_k^T(b - G_q(G_q^T G_q)^{-1} G_q^T b)\|_2 \tag{26}$$

which can be evaluated directly from data and $x_p = 0$ is minimizer of problem (15) if and only if $\lambda \geq \lambda_c$.

We also proved critical λ_1, λ_2 in (16) as

$$(\lambda_1\sqrt{K} + \lambda_2)_c = \max_{k \in \{1,2,\ldots,L\}} \|G_k^T(b - G_q(G_q^T G_q)^{-1} G_q^T b)\|_2 \tag{27}$$

which is necessary bound for λ_1, λ_2 in sense that if $x_p = 0$ is minimizer of problem (16) then $(\lambda_1\sqrt{K} + \lambda_2) \geq (\lambda_1\sqrt{K} + \lambda_2)_c$.

The detail of derivation on our critical value in problem (16) is in the appendix.

4.4 Simulated data generation

We assume three different scenarios that should be a representation of real data such as fMRI data. In groups of healthy subjects and groups that have brain condition, brain connectivity of subject in the same group should have common patterns that contributed by their conditions and the individual condition will contribute their difference in brain connectivity. From these reasons, we generated data in 3 types to test each formulation, note that we select VAR lag at $p = 2$ because we aim for sparsity level selection. We generated zero patterns to reflect the zero index in GC matrix as in statement (7). Because of this, entries of VAR coefficients in (1) will be zero along all lags. Each formulation should perform best if their data type is the same as the assumptions on the formulation. In simulated data generation, we randomized the coefficients in the way that eigenvalue of dynamic matrix in (2) is inside unit circle to assure stability of VAR models. The algorithm is simple. By exploiting that the stability of single sparse VAR should be affected by diagonal term of $A_1, \cdots, A_p$ most and the nonzero in diagonal term is negligible then the characteristic equation of dynamic matrix in (2) will be $\prod_{i=1}^{n}(z^p(A_1)_{(i,i)} + z^{p-1}(A_2)_{(i,i)} + \cdots + (A_p)_{(i,i)}) = 0$. We generated this characteristic equation from roots generated in unit circle and then randomly add value to the guaranteed stability diagonal VAR coefficients to generalize. Note that randomly adding value to the VAR coefficients may put the process unstable, so we check stability after randomly adding. If the process is unstable, we will repeat the process. This algorithm may not work in high-dimensional setting such as too many lags, high time-series dimension or in dense model setting. We generated three different ground truth models with lag $p = 2$ and $K = 4$. The ground truth types are as follows.

1. Common-sparsity type ground truth

   VAR coefficient matrices of all models have same sparsity patterns but the nonzero value need not to be equal. To be more compact, the zero pattern in this type are generated as

$$(A_q)_{ij}^{(k)} = 0, \ \ q = 1, ..., p, \ \ k = 1, ..., K. \tag{28}$$

2. Differential-sparsity type ground truth

   The parameters of all models are generated as in common type ground truth but we will add entries of VAR coefficient randomly with given proportion to matrix dimension. We will call this proportion as differential density.

3. Similar-coefficients type ground truth

   VAR coefficient matrices of all models have similar value with randomly added the entries with given differential density.

Each of the ground truth varies nonzero VAR parameters density to be $0.1, 0.3$ for common type. Other types have nonzero VAR parameters density of only 0.1 and have differential density of 0.1 or 0.3 instead.

5 Experimental results

We setup three experiments based on three different assumptions on data. The data are generated from common type ground truth model, differential type ground truth model and similar type ground truth model. Each data that were generated in different type of ground truth were estimated by three formulation which are formulation **C, D, S** to test the performance of each formulation with different assumption on data. In real data such as fMRI, or bio-signal, the data are limited and to take this in to account, we selected the data points of time series to be three times more than number of model parameter as the rule of thumb suggested to use more than ten times of model parameters. We also varied two different density on all ground truth types. In ground truth type common, we varied nonzero density with $d = 0.1, 0.3$. In ground truth type differential and similar, we fixed nonzero density with $d = 0.1$ and varied differential density which is proportion to different value added into the coefficients to be $d = 0.1, 0.3$. The setting on data can be described as in *Figure* 4.

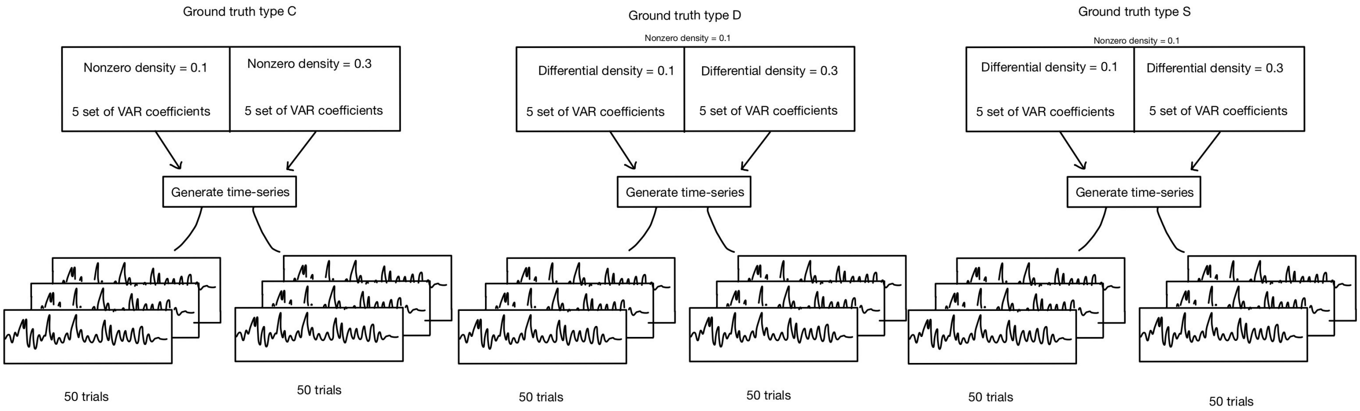
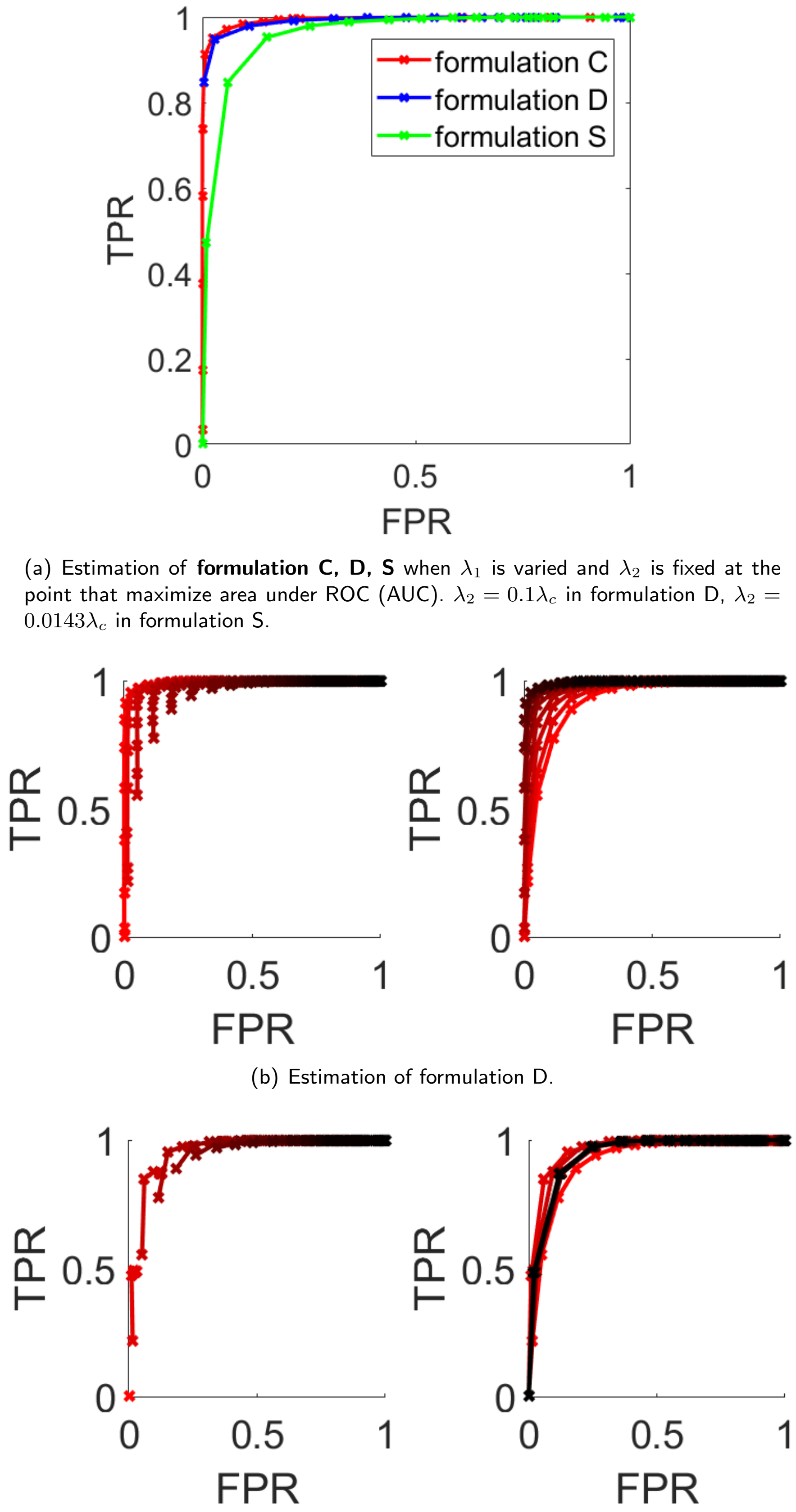


Figure 4: Data on each types of group truth.

To compare each formulation performance, we used ROC curve and area under curve (AUC), parameter bias as performance indices. Parameters that were estimated as zero are considered as positive result and vice versa. Parameter bias will be calculated as $\frac{\|A_{true}-A_{est}\|_F}{\|A_{true}\|_F}$. To find average performance, the data of this experiment were generated by 5 set of VAR coefficients and in each set, we generated 10 new realization of multivariate time series, so the data in each type were generated 50 trials. This should be enough to see overall performance of each formulation.

Experiment 1: Common type ground truth

This experiment, we will explore how good each formulation perform on data set #1 that **all models have identical sparsity pattern**. The results of this experiment are in *Figure* 5 for nonzero density of 0.1 and *Figure* 6 for nonzero density of 0.3. All formulation estimates common type ground truth very well but with formulation C and D, the ROC is better than formulation S. In *Figure* 5 (b), 6 (b), It can be clearly seen that if we increases λ_2 in formulation D, the ROC curve will move upward. In 5 (c), 6 (c), It can be seen that if we increase λ_2 which controls similarity of all models, the improvement on increasing λ_2 on the left plot is at peak on second value of λ_2 and then starts to decay. Since formulation C is special case of formulation D when $\lambda_1 = 0$, formulation D performed as good as formulation C as expected. Since the data that have only common structure in GC pattern, the way ROC curve of formulation D is improved by increment of λ_2 is as expected because λ_2 in formulation D is regularization parameter that controls common sparsity of all models. The non-monotonicity improvement in ROC of formulation S may come from the zero parameters that came from λ_1 are fused with another nonzero parameters. Both parameters can be zero or nonzero. If the parameters become zero and the true structure at those index are sparse then they improve ROC. In some value of λ_2, the decay of ROC may came from the same scenario but after fusion, the zero parameters became nonzero. The coefficients when regularization parameter λ_2 is increased to some point cannot improve any similarity because all models are identical which is expected because we penalized them. All results tend to confirm our hypothesis because formulation C has best performance in the sense of area under ROC curve and formulation C is special case of formulation D where $\lambda_1 = 0$. In formulation S, the ROCs are the worst as hypothesized.

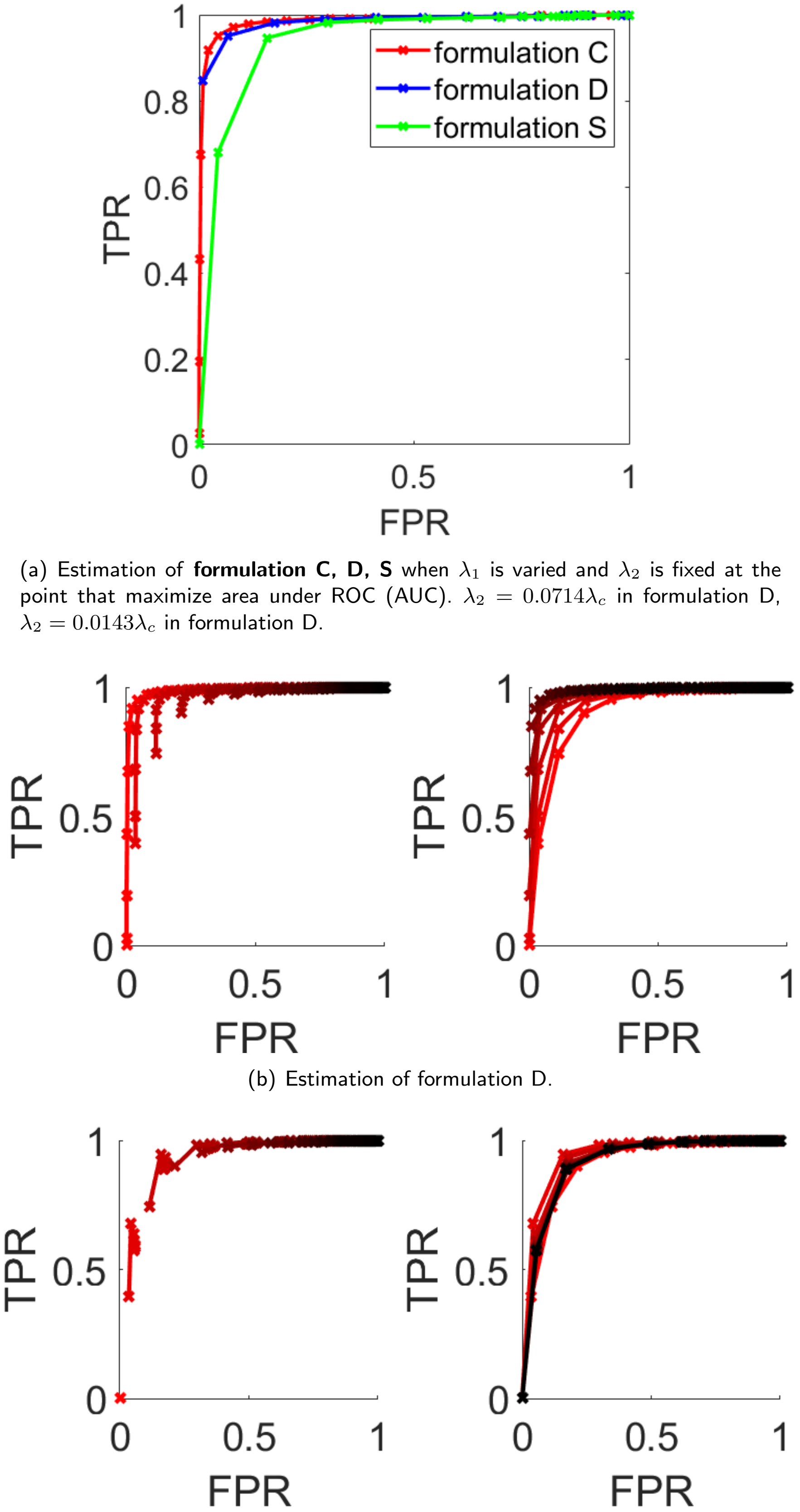


(a) Estimation of **formulation C, D, S** when λ_1 is varied and λ_2 is fixed at the point that maximize area under ROC (AUC). $\lambda_2 = 0.1\lambda_c$ in formulation D, $\lambda_2 = 0.0143\lambda_c$ in formulation S.

(b) Estimation of formulation D.

(c) Estimation of formulation S.

Figure 5: ROC plot of formulation **C, D, S** on common type ground truth with **nonzero density of 0.1**. λ_1 is fixed and varies λ_2 on left side and λ_2 is fixed and varies λ_1 on the right right side in both (b), (c). Darker color in (b), (c) denotes increments in varying λ.



(a) Estimation of **formulation C, D, S** when λ_1 is varied and λ_2 is fixed at the point that maximize area under ROC (AUC). $\lambda_2 = 0.0714\lambda_c$ in formulation D, $\lambda_2 = 0.0143\lambda_c$ in formulation D.

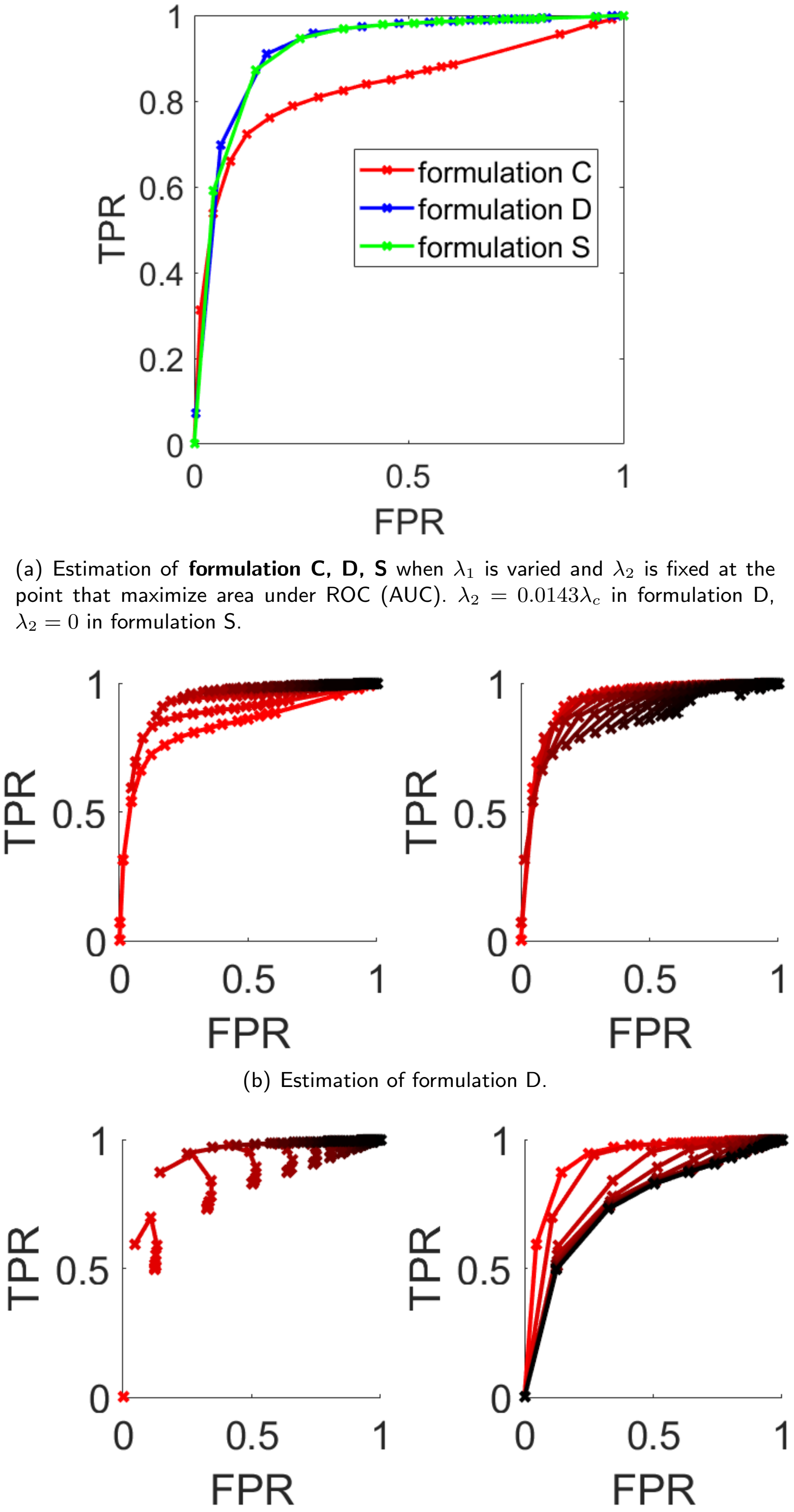
(b) Estimation of formulation D.

(c) Estimation of formulation S.

Figure 6: ROC plot of formulation **C, D, S** on common type ground truth with **nonzero density of 0.3**. (Left) fixed λ_1, varied λ_2. (Right) fixed λ_2, varied λ_1 in both (b), (c). Darker color in (b), (c) denotes increments in varying λ

Experiment 2: Differential type ground truth

We explored how each formulation performed on data set $\#2$ that **every models shared GC sparsity pattern but each model also has its own different GC sparsity pattern**. The results of this experiment are in *Figure* 7 for differential density of 0.1 and 0.3 for *Figure* 8. Formulation C is the worst of all as expected because there are different GC patterns of each model. In *Figure* 7 (a), formulation S performs better in average in the sense of having more area under ROC but it performs best when λ_2 is zero. In other words, It encouraged only different sparsity pattern or individually estimated sparse model. If we compared formulation D and S in the sense of moderate λ_2, the formulation D performed better than formulation S because in *Figure* 7, the ROC curve of formulation D at low λ_1 , moderate λ_2 is above ROC curve of formulation S. In our hypothesis, the formulation S should have more parameter bias when estimating differential type ground truth than formulation D. The result of parameter bias is shown in *Figure* 9. It can be clearly seen that the parameter bias increased as λ_2 in formulation S increased significantly more than when λ_2 in formulation D increased. In *Figure* 9 (c), we show that the averaged parameter bias obtained from various λ_1 of formulation D is less than formulation S in most of λ_2. The jumping point of parameter bias in formulation D is because the regularization level (λ_2) is high enough to remove significant parameters in VAR model. This does not happen in formulation S because if models are mostly identical, the significant parameters depended on λ_1 while in formulation D, it depends on both λ_1, λ_2.

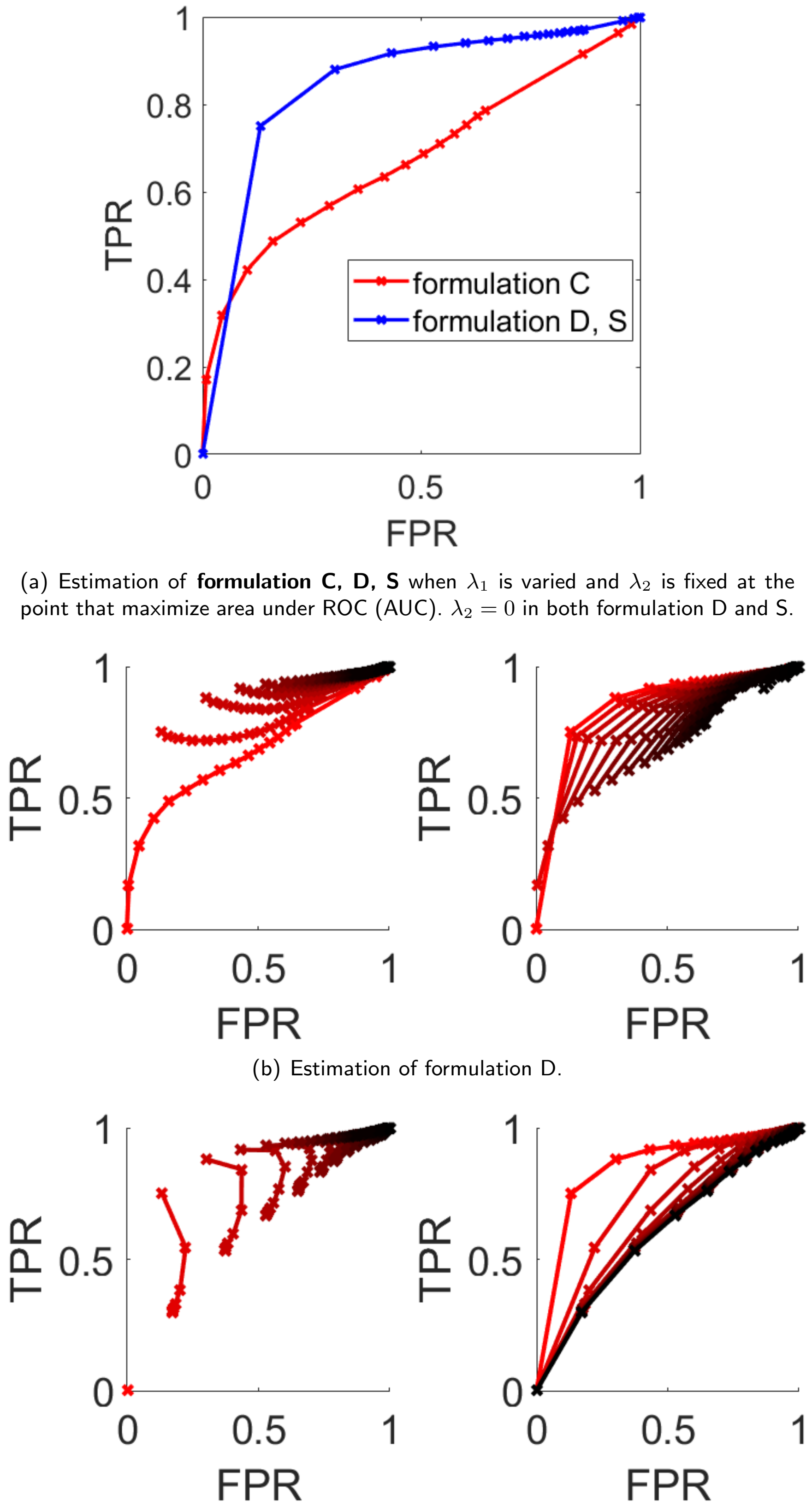


(a) Estimation of **formulation C, D, S** when λ_1 is varied and λ_2 is fixed at the point that maximize area under ROC (AUC). $\lambda_2 = 0.0143\lambda_c$ in formulation D, $\lambda_2 = 0$ in formulation S.

(b) Estimation of formulation D.

(c) Estimation of formulation S.

Figure 7: ROC plot of formulation **C, D, S** on differential type ground truth with **differential density of 0.1** . (Left) fixed λ_1, varied λ_2. (Right) fixed λ_2, varied λ_1 in both (b), (c). Darker color in (b), (c) denotes increments in varying λ

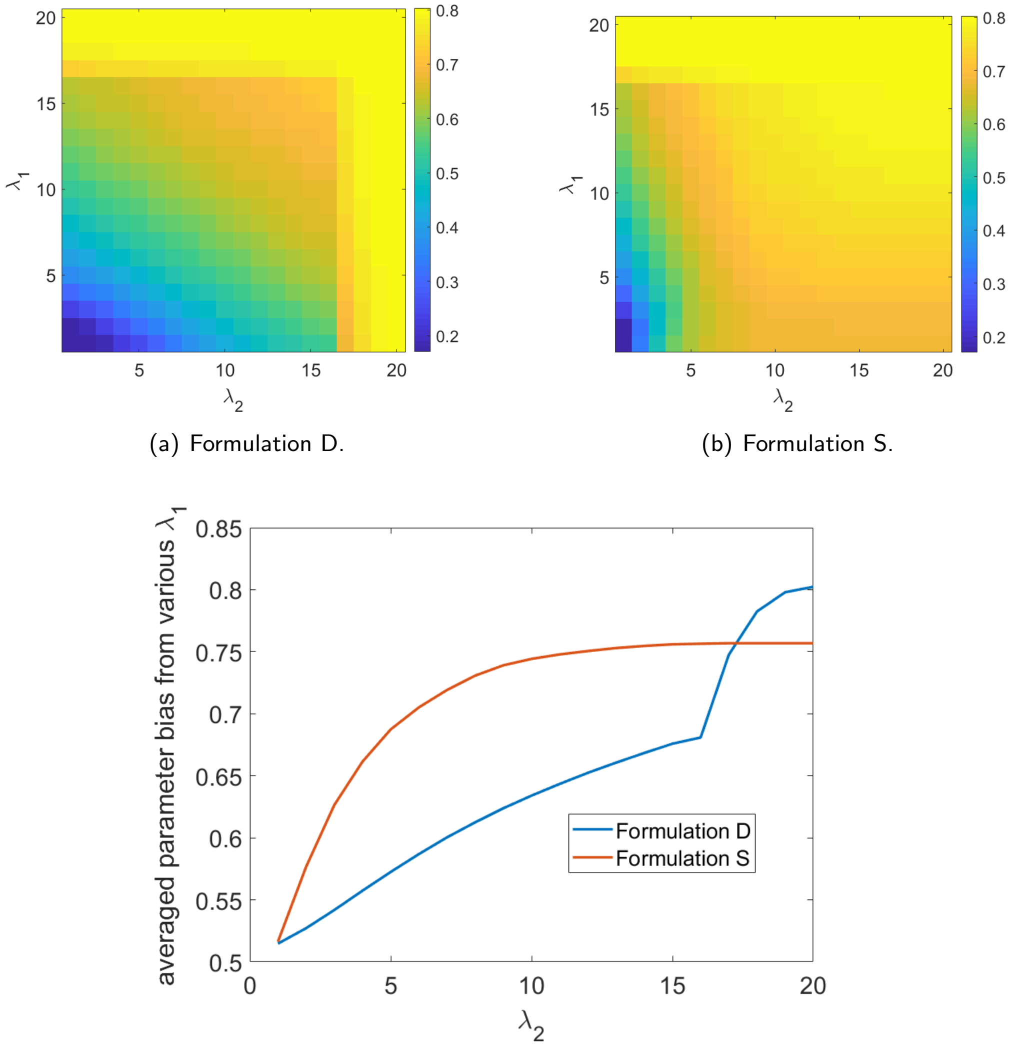


(a) Estimation of **formulation C, D, S** when λ_1 is varied and λ_2 is fixed at the point that maximize area under ROC (AUC). $\lambda_2 = 0$ in both formulation D and S.

(b) Estimation of formulation D.

(c) Estimation of formulation S.

Figure 8: ROC plot of formulation **C, D, S** on differential type ground truth with **differential density of 0.3**. (Left) fixed λ_1, varied λ_2. (Right) fixed λ_2, varied λ_1 in both (b), (c). Darker color in (b), (c) denotes increments in varying λ



(a) Formulation D.

(b) Formulation S.

(c) Parameter bias averaged over λ_1 comparison.

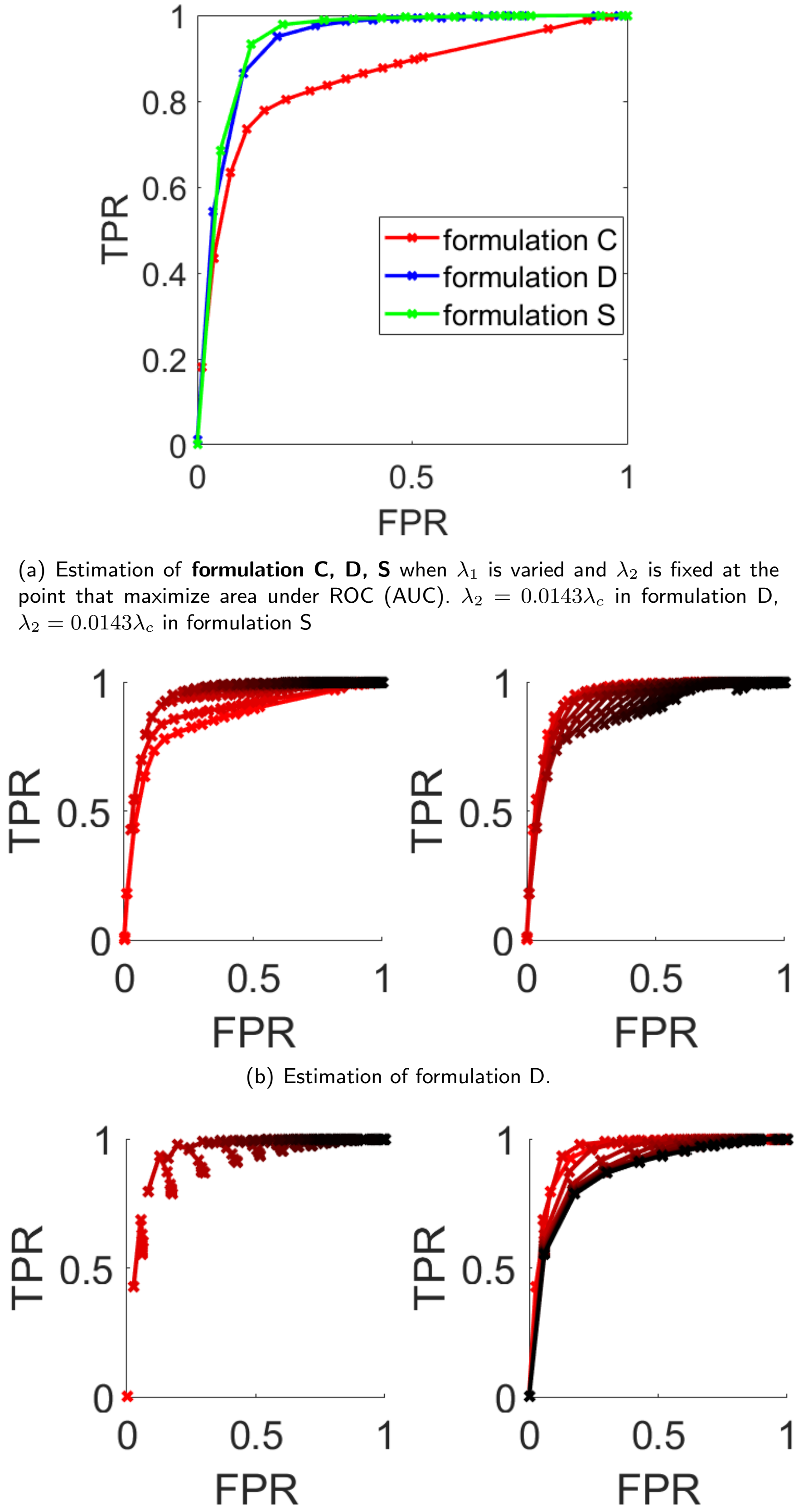
Figure 9: Parameter bias of differential type ground truth estimated by **formulation D** and **S**.

Experiment 3: Similar type ground truth

The goal of this experiment is to test all formulation base on data set #3 that **have similar parameters and some different sparsity patterns**. The results of this experiment are in *Figure* 10 for differential density 0.1 and *Figure* 11 for differential density 0.3. In *Figure* 10 (a), formulation C have lowest area under ROC of all formulation. In *Figure* 10 (b), (c), formulation S and D give ROC that is similar with ROC form their formulation in experiment 2. This can be inferred that the formulation S and D can perform equally which is plausible because the data were generated differently only in the value of parameters. The value of parameters in ground truth type S in each models is close to each other whereas the ground truth type D do not necessary to have this characteristics. Although The estimation is without similarity encouragement, the estimated value of nonzero parameters, however, should be close to each other because the similarity came from the least square solution. In other words, if true models have similarity in nonzero value entries, the least square estimation should explain these similarity because they are significant predictors that reduces model bias.

After each experiment, we used BIC as model selection criteria. The regularization parameters that minimized BIC score over grid of those regularization parameters will be selected. the result is in *Figure* 12. The results suggested that the number of fused nonzero entries as degree of freedom tends to give higher TPR and higher FPR than the number of nonzero as degree of freedom in experiment 1, 2. Experiment 3 have same accuracy in low differential density but in high differential density, the TPR drops and FPR is increased significantly. From the result of all experiment, same ground truth type as formulation type has a better accuracy which is as we hypothesized. From degree of freedom comparison between number of nonzero entries and number of fused nonzero entries in formulation S, the results yielded that by using degree of freedom as nonzero entries, the selection by BIC score tends to have better accuracy and parameter bias.

In each experiment, the example of the GC patterns in formulation C, D, S is in the *Figure* 13. In formulation S, we show first lag VAR coefficients to illustrate similarity of estimated parameters between models. The pattern of GC matrix in formulation C is the same for all models. The patterns of GC matrix in formulation D has both common structure which all models shared together and differential structure in each model. Formulation S have close value of all models. Then we can concluded that results of each formulation is as hypothesized.

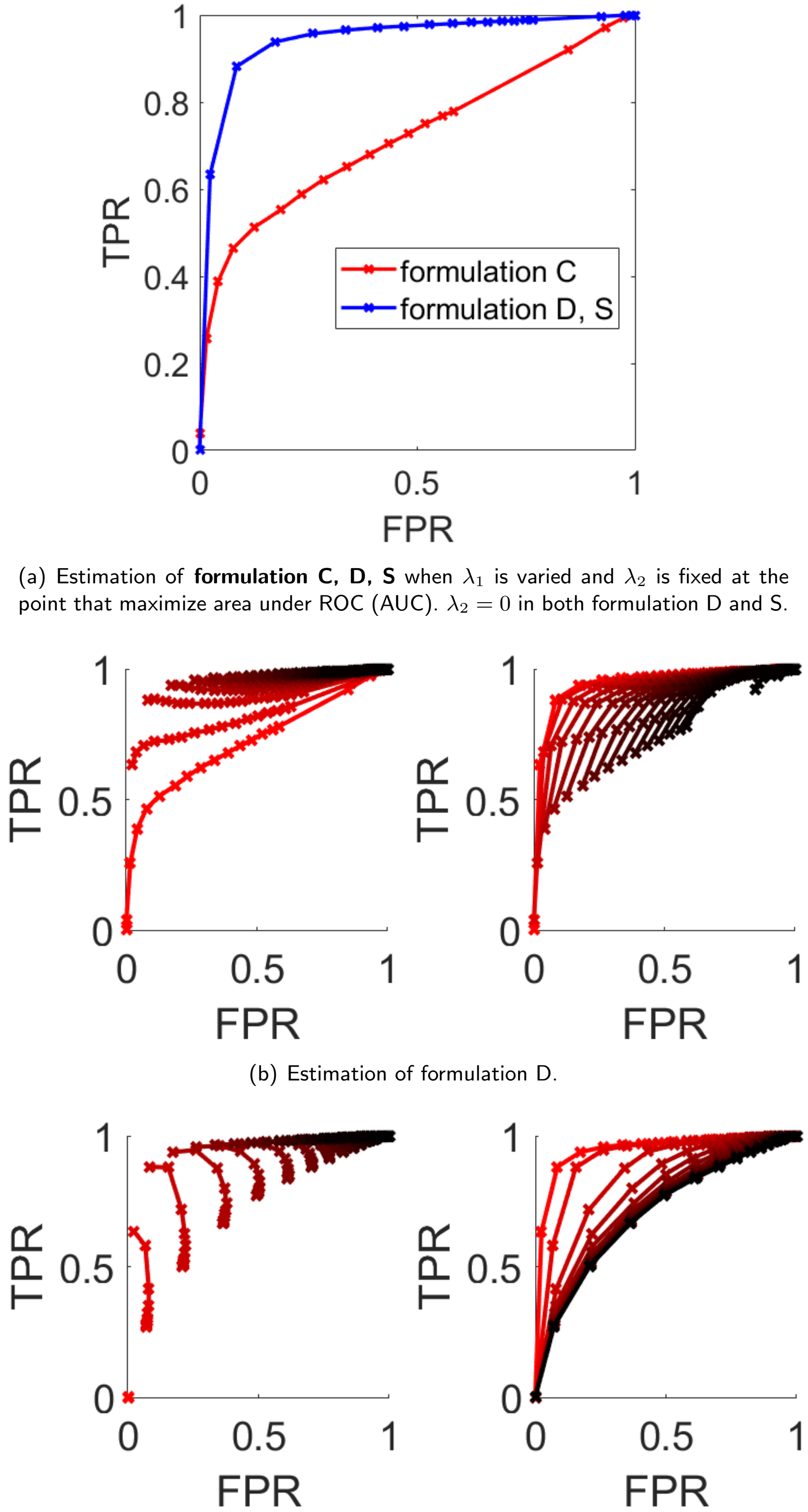


(a) Estimation of **formulation C, D, S** when λ_1 is varied and λ_2 is fixed at the point that maximize area under ROC (AUC). $\lambda_2 = 0.0143\lambda_c$ in formulation D, $\lambda_2 = 0.0143\lambda_c$ in formulation S

(b) Estimation of formulation D.

(c) Estimation of formulation S.

Figure 10: ROC plot of formulation **C, D, S** on similar type ground truth (Experiment 3) with **differential density of 0.1**. (Left) fixed λ_1, varied λ_2. (Right) fixed λ_2, varied λ_1 in both (b), (c). Darker color in (b), (c) denotes increments in varying λ

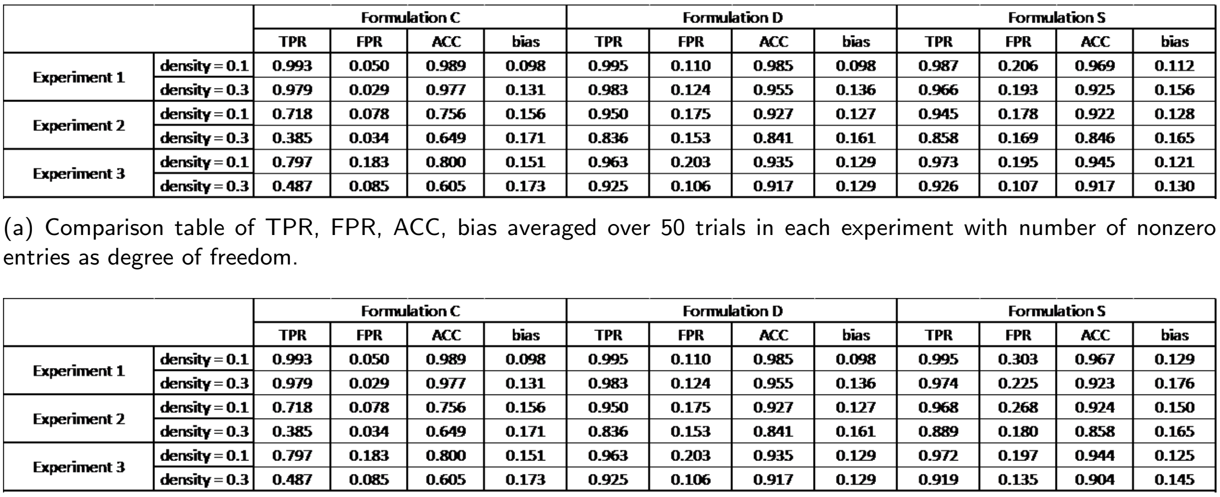


(a) Estimation of **formulation C, D, S** when λ_1 is varied and λ_2 is fixed at the point that maximize area under ROC (AUC). $\lambda_2 = 0$ in both formulation D and S.

(b) Estimation of formulation D.

(c) Estimation of formulation S.

Figure 11: ROC plot of formulation **C, D, S** on similar type ground truth (Experiment 3) with **differential density of 0.3**. (Left) fixed λ_1, varied λ_2. (Right) fixed λ_2, varied λ_1 in both (b), (c). Darker color in (b), (c) denotes increments in varying λ



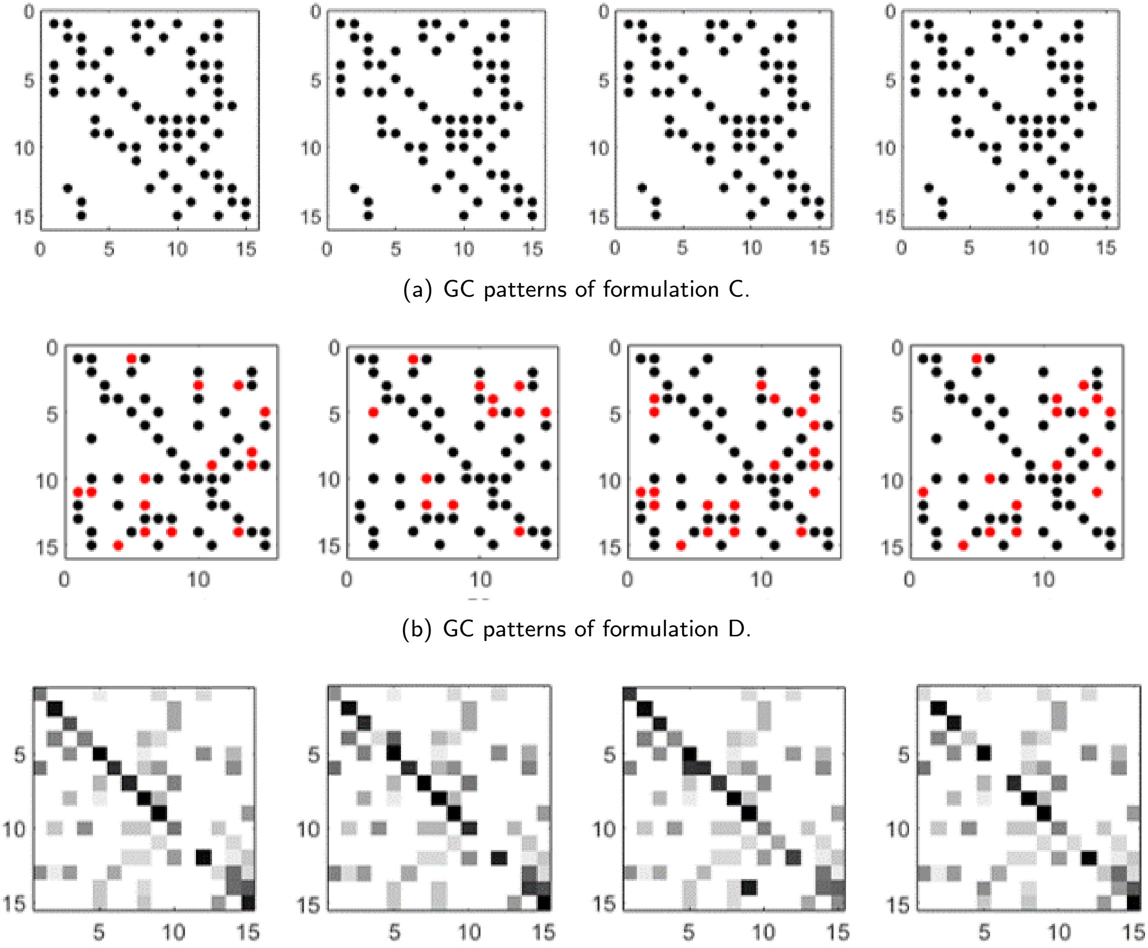
		Formulation C				Formulation D				Formulation S			
		TPR	FPR	ACC	bias	TPR	FPR	ACC	bias	TPR	FPR	ACC	bias
Experiment 1	density = 0.1	0.993	0.050	0.989	0.098	0.995	0.110	0.985	0.098	0.987	0.206	0.969	0.112
	density = 0.3	0.979	0.029	0.977	0.131	0.983	0.124	0.955	0.136	0.966	0.193	0.925	0.156
Experiment 2	density = 0.1	0.718	0.078	0.756	0.156	0.950	0.175	0.927	0.127	0.945	0.178	0.922	0.128
	density = 0.3	0.385	0.034	0.649	0.171	0.836	0.153	0.841	0.161	0.858	0.169	0.846	0.165
Experiment 3	density = 0.1	0.797	0.183	0.800	0.151	0.963	0.203	0.935	0.129	0.973	0.195	0.945	0.121
	density = 0.3	0.487	0.085	0.605	0.173	0.925	0.106	0.917	0.129	0.926	0.107	0.917	0.130

(a) Comparison table of TPR, FPR, ACC, bias averaged over 50 trials in each experiment with number of nonzero entries as degree of freedom.

		Formulation C				Formulation D				Formulation S			
		TPR	FPR	ACC	bias	TPR	FPR	ACC	bias	TPR	FPR	ACC	bias
Experiment 1	density = 0.1	0.993	0.050	0.989	0.098	0.995	0.110	0.985	0.098	0.995	0.303	0.967	0.129
	density = 0.3	0.979	0.029	0.977	0.131	0.983	0.124	0.955	0.136	0.974	0.225	0.923	0.176
Experiment 2	density = 0.1	0.718	0.078	0.756	0.156	0.950	0.175	0.927	0.127	0.968	0.268	0.924	0.150
	density = 0.3	0.385	0.034	0.649	0.171	0.836	0.153	0.841	0.161	0.889	0.180	0.858	0.165
Experiment 3	density = 0.1	0.797	0.183	0.800	0.151	0.963	0.203	0.935	0.129	0.972	0.197	0.944	0.125
	density = 0.3	0.487	0.085	0.605	0.173	0.925	0.106	0.917	0.129	0.919	0.135	0.904	0.145

(b) Comparison table of TPR, FPR, ACC, bias averaged over 50 trials in each experiment with fused nonzero parameters as degree of freedom.

Figure 12: Averaged accuracy comparison table of BIC selected regularization parameters in each experiment.



(a) GC patterns of formulation C.

(b) GC patterns of formulation D.

(c) GC patterns and VAR coefficient value of formulation S drawn from first lag of VAR coefficients in each model.

Figure 13: Sparsity pattern on GC matrices of 4 models estimated by each formulation. The black dots in (a), (b) are common for all models and red dots in (b) are entries that are different from black dots. *Figure* (c) shows that the estimated VAR parameters are fused or close to each other and the zero entries denote sparsity pattern in its GC matrix.

6 Conclusion

In this project, we aim to learn zero pattern of brain connectivity matrix based on Granger causality in multiple VAR models simultaneously. We have shown that if we extend from learning single model multiple times to learning multiple models at once, we can include prior knowledge on those models in the estimation. The prior we included in our estimation are

1. All models have identical GC pattern (formulation C).
2. All models have two parts of GC pattern. First is the common GC pattern that all models shared together. Second is the differential GC pattern that makes GC matrices of models are different (formulation D).
3. All models have two parts of GC pattern. First is the models are similar to each other in the way that value of VAR coefficient is close. Second is the differential GC pattern that makes GC matrices of models are different (formulation S).

We tested each formulation on simulated data based on three different types of ground truth that have the same assumptions as their formulation counterpart assumed. Our results yielded that the formulation performed at best if their data have same properties as the assumptions on the formulation. However, the performance of each formulation depended on the regularization parameters selection. We selected the regularization parameters that minimized BIC score so that the solution is the optimal trade-off between complexity and fitness of model. Anyway, the choices of complexity terms are not trivial and must be derived from the formulation but for simplicity, we used number of nonzero parameters as the degree of freedom in complexity term for formulation C, D and used number of fused nonzero parameters for formulation S.

At this point, we concluded that the formulation C can performed well if all models have strong common sparsity structure that have no different sparsity at all but if the true structure of the models contains a different sparsity structure, formulation C may performed poorly depended on how different the structure is in each model. In formulation D, it can be seen that it is a generalization of formulation C which allows different sparsity structure in the models perform best when compared to formulation C and S. In formulation S, we assumed that all models have close parameters value. Formulation S regularized the cost function to forced each model to have close parameters value. This will performed best if the true parameters are having close value on each models. If this is not true, the bias of the estimation may arises. Such as in ground truth type D that does not assume all models have parameters close value when estimating using formulation S, the bias will increase.

References

[BBS10] Adam B. Barrett, Lionel Barnett, and Anil K. Seth. Multivariate Granger causality and generalized variance. *Physical Review E*, 81:041907, Apr 2010.

[BH09] Patrick Breheny and Jian Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2:369–380, 07 2009.

[BJ76] G.E.P. Box and G.M. Jenkins. *Time series analysis: forecasting and control*. Holden-Day series in time series analysis and digital processing. Holden-Day, 1976.

[BL14] Seth AK Barnett L. The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Neuroscience Methods*, 2 2014.

[CG11] Arijit Chakrabarti and Jayanta K. Ghosh. AIC, BIC and recent advances in model selection. In Prasanta S. Bandyopadhyay and Malcolm R. Forster, editors, *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*, pages 583 – 605. North-Holland, Amsterdam, 2011.

[CGC12] Julien Chiquet, Yves Grandvalet, and Camille Charbonnier. Sparsity with sign-coherent groups of variables via the cooperative-lasso. *The Annals of Applied Statistics*, 6(2):795–830, 06 2012.

[DWW14] Patrick Danaher, Pei Wang, and Daniela Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 76:373–397, 03 2014.

[FHT07] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 12 2007.

[GLMZ11] Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98 1:1–15, 2011.

[GMCW13] Tanya Garcia, Samuel Mller, Raymond Carroll, and Rosemary Walzem. Identification of important regressor groups, subgroups, and individuals via regularization methods: Application to gut microbiome data. *Bioinformatics (Oxford, England)*, 30, 10 2013.

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

[Inc12] CVX Research Inc. CVX: Matlab software for disciplined convex programming, version 2.0. `http://cvxr.com/cvx`, August 2012.

[JTT10] Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39, 05 2010.

[LLSL18] Jie Liu, Gangning Liang, Kimberly D. Siegmund, and Juan Pablo Lewinger. Data integration by multi-tuning parameter elastic net regression. *BMC Bioinformatics*, 19(1):369, Oct 2018.

[Lüt05] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer, Berlin, 2005.

[MB10] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[MS18] P. Manomaisaowapak and J. Songsiri. Learning group differences in brain networks from eeg signals using hotelling's T^2. Technical report, Control system research laboratory, Electrical engineering department, Chulalongkorn university, 2018. `https://github.com/parinthorn/Academic-project/blob/master/eeg_usm.pdf`.

[RXZ13] Yunwen Ren, Zhiguo Xiao, and Xinsheng Zhang. Two-step adaptive model selection for vector autoregressive processes. *Journal of Multivariate Analysis*, 116:349 – 364, 2013.

[SM18] Skripnikov and George Michailidis. Regularized joint estimation of related VAR models via group lasso. *Journal of Biostatistics and Biometric Applications*, 3(2), 2018.

[Son15] J. Songsiri. Learning multiple Granger graphical models via group fused lasso. In *2015 10th Asian Control Conference (ASCC)*, pages 1–6, May 2015.

[Son17] J. Songsiri. Estimations in learning Granger graphical models with application to fMRI time series. Technical report, CSRL, Electrical engineering department, Chulalongkorn university, 2017.

[Spo07] O. Sporns. Brain connectivity. *Scholarpedia*, 2(10):4695, 2007. revision #91084.

[THW+16] Qinghua Tao, Xiaolin Huang, Shuning Wang, Xiangming Xi, and Li Li. Multiple Gaussian graphical estimation with jointly sparse penalty. *Signal Processing*, 128:88 – 97, 2016.

[VBK+10] Gaël Varoquaux, Flore Baronnet, Andreas Kleinschmidt, Pierre Fillard, and Bertrand Thirion. Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In Tianzi Jiang, Nassir Navab, Josien P. W. Pluim, and Max A. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, pages 200–208, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[WC18] I. Wilms and C. Croux. An algorithm for the multivariate group lasso with covariance estimation. *Journal of Applied Statistics*, 45(4):668–681, 2018.

[YL06] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 02 2006.

[ZHT07] Hui Zou, Trevor Hastie, and Robert Tibshirani. On the degrees of freedom of the lasso. *Annals of Statistics*, 35(5):2173–2192, 10 2007.

7 Appendix

7.1 MATLAB programs

This section included our MATLAB function

7.1.1 Data generation

- gen_VAR_param.mat [Son17] : generates K VAR models with specified relation such as common, differential, similar
- GEN_MODEL_TIMESERIES.mat : generates time series from given VAR coefficient

7.1.2 Estimation

- veccoefmatgroup.m [Son17] : vectorizes problem (10) to the fitting term in (19).
- offdiag.m [Son17] : generates P matrix in (19).
- diffmat.m : generate D matrix in (20).
- efficient_diff.m : find Dx in (20) using indexing instead of sparse matrix multiplication.
- efficient_vect.m : vectorizes VAR coefficients in 4D to vectorized form using indexing instead of looping.
- H_gen.m : generates H matrix.
- lambdamax_grouplasso.m [Son17] : finds λ_c in problem (15).
- fusedlasso_ADMMoff.m [Son17]: solves problem (20).
- grouplasso_sharedsp.m [Son17] : solves problem (19).
- refit.m : solve constrained least square in (10) with given sparsity patterns by grouplasso_sharedsp.m.
- refit_fused.m : solve constrained least square in (10) with given sparsity pattern and fused pattern by fusedlasso_ADMMoff.m.
- BIC.m find BIC score of estimation.

7.2 Derivation of necessary bound of formulation D

Consider cost function

$$(1/2)\|b - Gx\|_2^2 + \lambda_1\|Px\|_A + \lambda_2\|Px\|_B \tag{29}$$

In [Son17], the author derived the sufficient and necessary bound for the special case of this cost function, $\lambda_1 = 0$. The necessary bound for this problem can be derived from subdifferential calculus as follow.

We reparametrized the problem 29 the same way as in [Son17] that separate penalized parameters out of parameters that are not penalized as

$$x = \begin{bmatrix} x_p \\ xq \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_L \end{bmatrix}, \quad G = \begin{bmatrix} G_p & G_q \end{bmatrix} = \begin{bmatrix} G_1 & \cdots & G_L \end{bmatrix} \tag{30}$$

The KKT condition of cost function 29 will be

$$0 \in \begin{bmatrix} G_1^T \\ \vdots \\ G_k^T \\ \vdots \\ G_L^T \end{bmatrix} (b - G_p x_p^* - G_q x_q^*) + \begin{bmatrix} \lambda_1 a + \lambda_2 b \\ 0 \end{bmatrix} \tag{31}$$

where a, b are subgradient of regularization terms which are sum of 2 norm of different size, first is block size p, the second is block size pK. We will use notation as, b_j to denotes j th block of b that has $n^2 - n$ blocks and a_{ij} as i th block of a that has $(n^2 - n)k$ blocks and j denotes index of b_j as a_{ij} is subset of b_j. The KKT condition, can now split into multiple matrix equation in each chunk of a_{ij}, b_j. As we know that if $x_j = 0$ then b_j can be any vector that has euclidean norm less than 1. If norm of b_j is less than 1 then norm of a_{ij} will also less than 1. At this point, we can rearrange the KKT condition to

$$0 \in -G_k^T(b - G_p x_p^* - G_q x_q^*) + \lambda_1 \begin{bmatrix} a_{1,j} \\ \vdots \\ a_K \end{bmatrix} + \lambda_2 b_j \tag{32}$$

$$\|G_k^T(b - G_p x_p^* - G_q x_q^*)\|_2 = \|\lambda_1 \begin{bmatrix} a_{1,j} \\ \vdots \\ a_K \end{bmatrix} + \lambda_2 \|b_j\|_2 \tag{33}$$

We want to find relation that gives $x_p^* = 0$ then (31) gives $x_q^* = (G_q^T G_q)^{-1} G_q^T b$ then the equation 33 for x_p reduced to

$$\|G_k^T(b - G_q(G_q^T G_q)^{-1} G_q^T b)\|_2 = \|\lambda_1 \begin{bmatrix} a_{1,j} \\ \vdots \\ a_K \end{bmatrix} + \lambda_2 b_j\|_2 \tag{34}$$

$$\|G_k^T(b - G_q(G_q^T G_q)^{-1} G_q^T b)\|_2 \leq \lambda_1 + \| \begin{bmatrix} a_{1,j} \\ \vdots \\ a_K \end{bmatrix} \|_2 + \lambda_2 \|b_j\|_2; \text{triangle inequality} \tag{35}$$

$$\|G_k^T(b - G_q(G_q^T G_q)^{-1} G_q^T b)\|_2 \leq \lambda_1 \sqrt{K} + \lambda_2; x_p^* = 0 \tag{36}$$

At this point, the necessary bound of sparsest solution for problem 29 can be expressed as

$$(\lambda_1 \sqrt{K} + \lambda_2)_{crit} = \max_{k \in \{1,2,\ldots,L\}} \|G_k^T(b - G_q(G_q^T G_q)^{-1} G_q^T b)\|_2 \tag{37}$$