A Convex Formulation of Structural Equation Modeling (SEM) in fMRI Study

2102531 Term Project Report

Anupon Pruttiakaravanich
Auangkun Rangsikunpum
Pusit Suriyavejwongs
Tawan Luprasong
Jitkomut Songsiri

Department of Electrical Engineering
Chulalongkorn University

December 9, 2015

Abstract

Since functional magnetic resonance imaging (fMRI) technique has been discovered for a few decades, this allows human to capture the activity in brain via a signal from blood flow, known as BOLD signal. The exploring of relationship among regions inside the human brain using fMRI data is challenging and still opening for several topics. Structural equation modeling (SEM) is one of the techniques that can be used to explore relationship structure among variables. Since SEM has been developed long time ago, there are several commercial software such as `LISREL`, `S-PLUS` or `M-PLUS` that can be used to explore causal structures from data we have. In this work, we focus on exploring the brain connectivity using SEM based on path analysis. This problem can be expressed as nonlinear optimization problem which may provide us the local solution. Therefore We propose to formulate this problem to convex optimization problem which has only one optimal point and this point can be regarded as global solution. We compare the result solved by our convex problem and `LISREL` commercial software. From our experiment, we have a hypothesis that, under some conditions, the solution solved by our convex problem and and `LISREL` are same. We do our experiment on various sets of fMRI data and find the best brain connectivity structure using BIC criterion. This work is a part of 2102531 System Identification term project, semester 1 of academic year 2015.

1 Introduction

Functional magnetic resonance imaging or fMRI is a technique for measuring brain function activities using a neuroimaging technology, and is nowadays dominating the brain mapping research field since it does not have requirement of surgery, substances ingesting, ionising radiation exposing, etc. Structural equation modeling (SEM) is then one of the most popular tools for brain network modeling for fMRI data. SEM focuses only on instantaneous effects, neglecting all the time-delay lagging effects. Path analysis is one of common methods in SEM used to explore the effective connectivity which can be analyzed by many well-known commercial softwares such as `LISREL` [1], `S-PLUS` [2], `M-PLUS` [3], and `EQS` [4]. Most commercial softwares use maximum likelihood estimation which results in a nonlinear formulation having not less than one minimum point, therefore, solving with numerical methods may not obtain the global optimal solution. Moreover, the free version of commercial softwares mostly have restriction on the number of model's parameter and may not be fully functional.

In this work, we explore the mathematical description of the problem in section 2 and then the formulation in section 3 which we proposed a convex formulation. Once our problem turns into a convex problem, the optimal solution which we obtain from solving this problem can be regarded as a global optimal solution. In section 4, we look into examples of fMRI data set. In section 5, we explain how to

get brain connectivities in terms of path coefficient matrices from commercial software and our convex problem. We compare the result from both methods for verification and then we provides the conclusion in section 6.

2 Problem description

By the use of SEM only based on path analyses to explore the effective connectivity, it can be generally expressed in the form of n-dimensional linear equation as

$$Y = c + AY + \epsilon, \tag{1}$$

where $Y \in \mathbb{R}^n$ denotes observed values from n regions of interest (ROIs), or voxels, in the brain, $c \in \mathbb{R}^n$ denotes the baseline of Y, $A \in \mathbb{R}^{n\times n}$ denotes the path matrix which is to be estimated and a_{ij}, the (i,j) entry of A, is known as a path coefficient representing a causal relationship from region j to i. $\epsilon \in \mathbb{R}^n$ denotes the error from modeling and noise on each voxel which is assumed to be zero mean white noise. As seen from (1), it is a linear static model, therefore, this fMRI study neglects the effect of time and treat one time point as one sample.

Let S denote a sample or empirical covariance matrix of Y and let Σ denote covariance matrix of Y as derived from (1) as

$$\Sigma = (I - A)^{-1}\Psi(I - A)^{-T}. \tag{2}$$

And let us assume that Y is normally distributed. If we need to estimate the model parameters in sense of minimizing a distance between S and Σ, we can choose the Kullback-Leibler divergence function (see more detail in Appendix) as the fit function given by

$$d(S, \Sigma) = \log\det\Sigma + \mathbf{tr}(S\Sigma^{-1}) - \log\det S - n, \tag{3}$$

where n denotes the dimension of data. This leads to an optimization problem and it can be expressed as follows.

$$\begin{array}{ll} \underset{\Sigma,A,\Psi}{\text{minimize}} & \log\det\Sigma + \mathbf{tr}(S\Sigma^{-1}) - \log\det S - n, \\ \text{subject to} & \Sigma = (I - A)^{-1}\Psi(I - A)^{-T}, \\ & \mathbf{diag}(A) = 0., \end{array} \tag{4}$$

The constraint $\mathbf{diag}(A) = 0$, i.e., $A_{ii} = 0$, for $i = 1, 2, 3, ..., n$ means that the link from region i to i must be zero or we can say that each voxel has no effect to itself. From (4), S and n are given from data so that the last two terms in the cost function can be neglected. We can make a change of variables by letting $X = \Sigma^{-1}$. Therefore the problem (4) becomes

$$\begin{array}{ll} \underset{X,A,\Psi}{\text{minimize}} & -\log\det X + \mathbf{tr}(SX) \\ \text{subject to} & X = (I - A)^T\Psi^{-1}(I - A), \\ & \mathbf{diag}(A) = 0, \end{array} \tag{5}$$

with variables $X \in \mathbb{S}^n, A \in \mathbb{R}^{n\times n}$ and $\Psi \in \mathbb{S}^n$.

3 Problem formulation

Nonlinear optimization problem has one or more than one minimum point, therefore, solving nonlinear optimization problem may provide the local minimum point. However, if the problem is convex, solving convex optimization provides only a minimum point and this point can be regarded as a global minimum point. The problem (5) is a nonlinear optimization problem due to the quadratic equation constraint:

$$X = (I - A)^T\Psi^{-1}(I - A)$$

We propose to relax the above constraint to

$$X \succeq (I - A)^T\Psi^{-1}(I - A)$$

which is equivalent to

$$\begin{bmatrix} X & (I-A)^T \\ I-A & \Psi \end{bmatrix} \succeq 0 \tag{6}$$

by the property of Schur complement (see more detail in Appendix). The constraint (6) is now linear in X, A and Ψ and it can be considered as a linear matrix inequality (LMI). Moreover, to prevent trivial solutions and to control the noise covariance in estimation, we can add a constraint on Ψ by $\Psi \preceq \alpha I$, meaning that Ψ is bounded by αI. This leads to the final problem formulation expressed by

$$\begin{aligned} \underset{X,A,\Psi}{\text{minimize}} \quad & -\log\det X + \mathbf{tr}(SX) \\ \text{subject to} \quad & \begin{bmatrix} X & (I-A)^T \\ I-A & \Psi \end{bmatrix} \succeq 0 \\ & \Psi \preceq \alpha I, \\ & \mathbf{diag}(A) = 0, \end{aligned} \tag{7}$$

with variables $X \in \mathbb{S}^n, A \in \mathbb{R}^{n\times n}$ and $\Psi \in \mathbb{S}^n$. The problem (7) has completely turned into a convex problem in the class of semidefinite programing (SDP) (the proof is provided in Appendix). Therefore there are several efficient algorithms to solve it.

4 fMRI data

4.1 Data acquisition

The primary form of fMRI uses the blood-oxygen-level dependent (BOLD) contrast which is the change in magnetization between oxygen-rich and oxygen-poor blood. The data is acquired as image intensity versus time at each voxel which can be referred as 3D-coordinate in the brain. For an example, Figure 1 shows a time plot of a data set at the voxel with coordinate (46,64,37) in the brain.

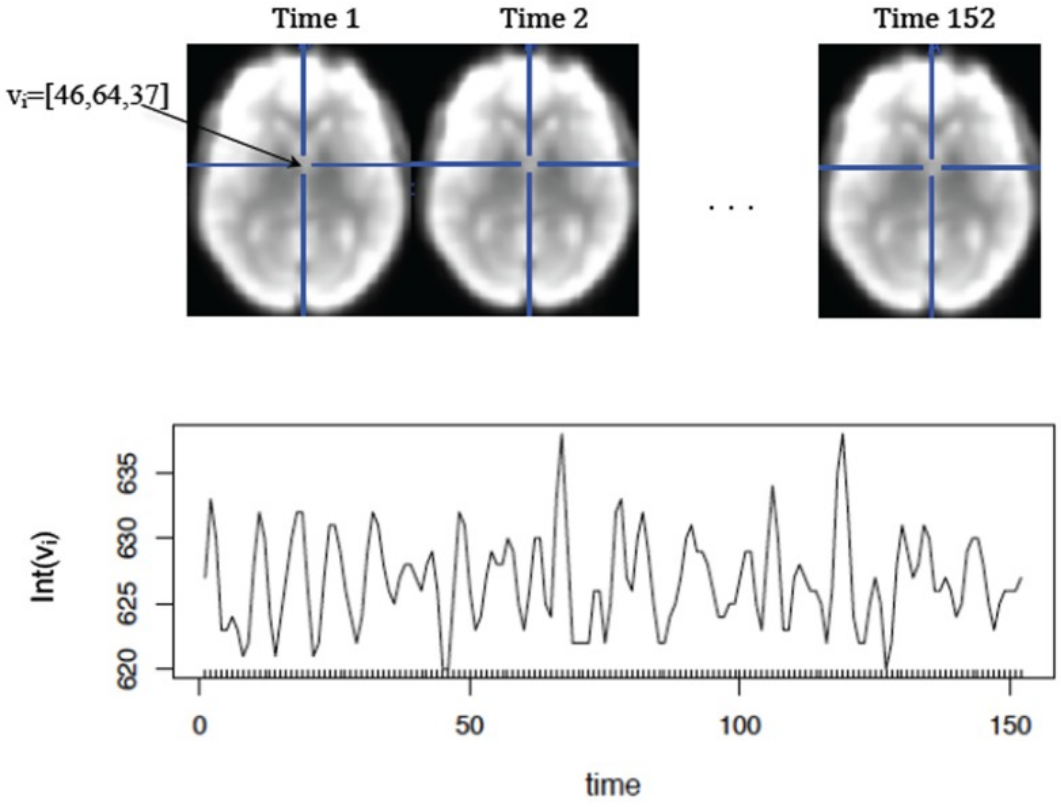


Figure 1: A plot between the intensity of the image at voxel v_i and time points [5].

4.2 Data analysis

Using brain network modeling tools such as Vector autoregression (VAR) and SEM [6] with acquisition of fMRI data set can provide us a brain mapping. In contrast to SEM, VAR focuses on time-delay lagging effects and can be solved using least-square formulation. Groups of voxels can also be defined as regions of interest (ROIs). An example of brain mapping that show the relations between the ROIs is shown in Figure 2.

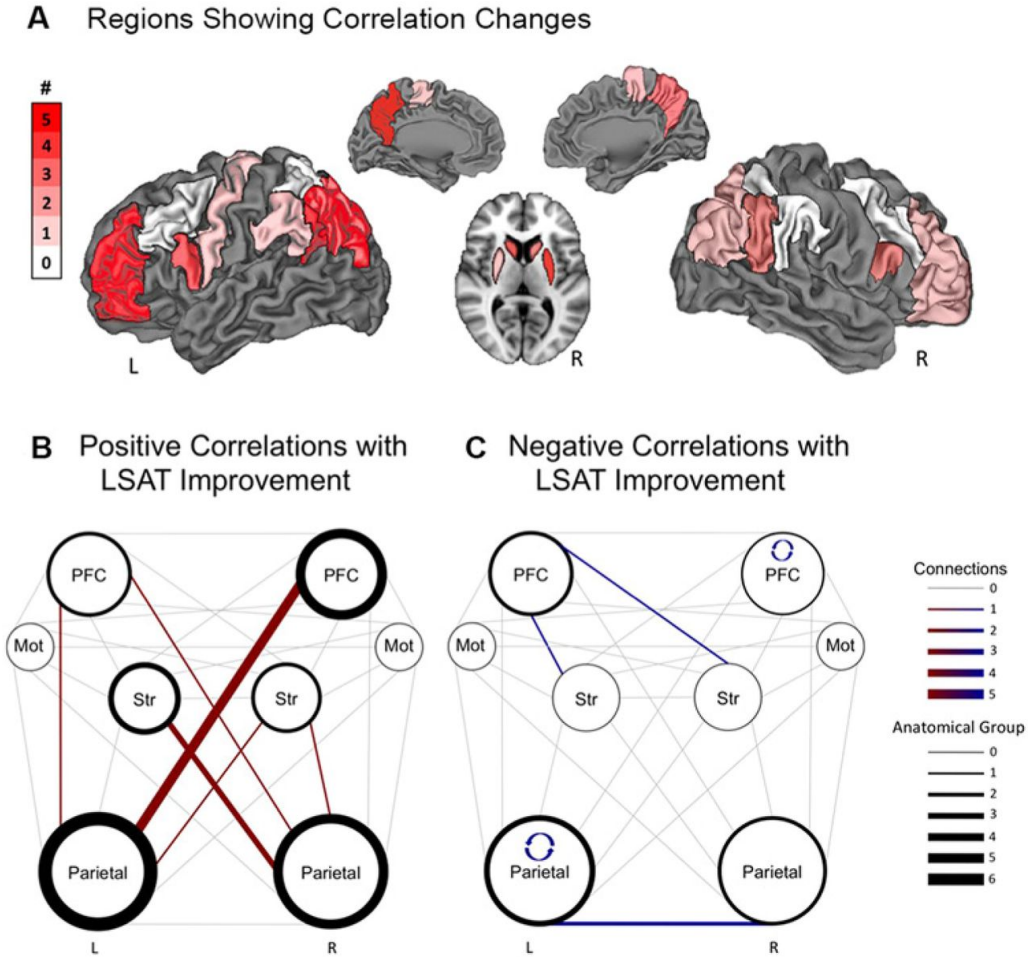


Figure 2: A brain mapping shows the relations between ROIs [7].

4.3 `StarPlus` fMRI data

`StarPlus` fMRI data is one of the free accessible raw fMRI data have collected by Marcel Just and his colleagues in Carnegie Mellon University's CCBI [8]. The experiment consists of a set of trials, and the data is partitioned into trials. For some of these intervals, subject simply rested or gazed at a fixation point on the screen. Images were collected every 500 ms. Only a fraction of the brain of each subject was imaged. The data is marked up with 25-30 ROIs. Figure 3 shows examples of resting-state `StarPlus` fMRI time series from 5 voxels. The image intensity is normalized to percentage above average signal value during control baseline (fixation trials). We will use these resting-state data in the following section.

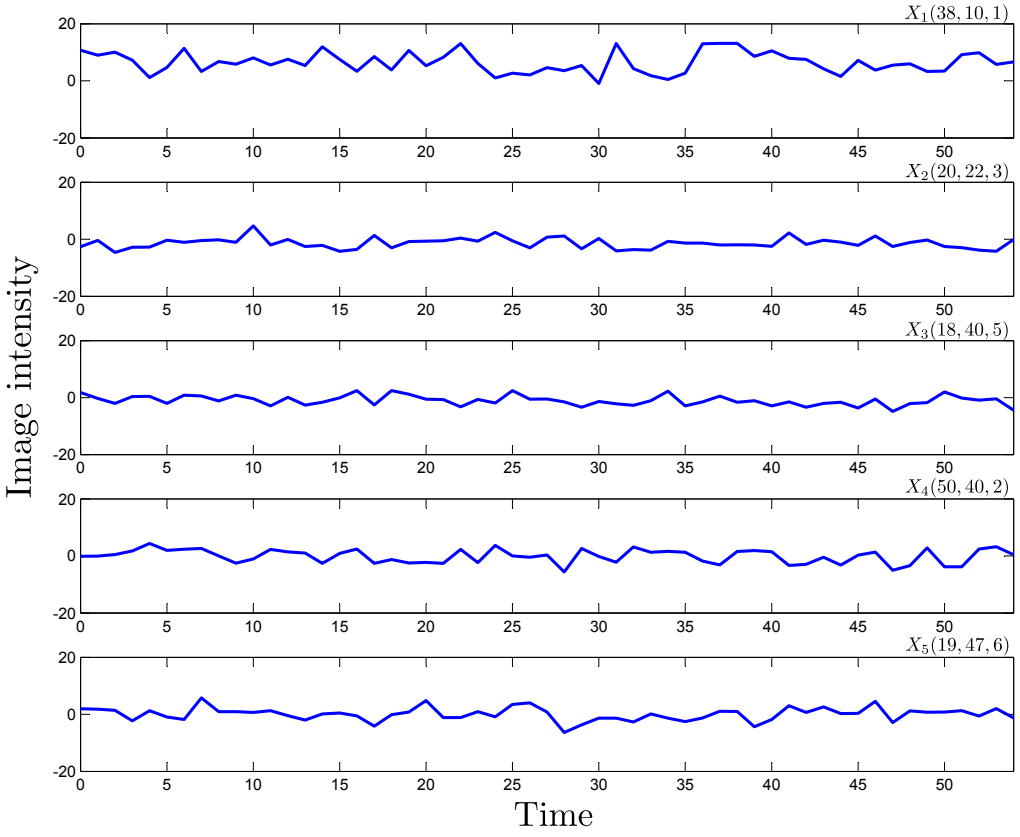


Figure 3: Examples of `StarPlus` fMRI time series from 5 voxels.

5 Solving estimating problem in SEM

5.1 Commercial softwares

There are several available software packages for solving SEM problem, for instance, `LISREL` [1], `EQS` [4], `M-PLUS` [3], and `S-PLUS` [2]. Here we would like to introduce `LISREL` for solving SEM problem. `LISREL` that we use in this work is a student version, therefore, it has some restrictions such as the restriction on the number of variables. From the beginning, when we need to perform SEM analysis using this software, it has many options to be chosen. For example, we can solve SEM problem with programming or drawing the path diagram as shown in Figure 4 and 5.

`LISREL` also allows user to choose an estimation method, for example, maximum likelihood estimation and weight least-squares estimation. Figure 6 shows estimation method selection window in `LISREL`. Three estimation methods which commonly used are unweighted least square (ULS), generalized least squares (GLS), and maximum likelihood (ML). Cost functions of these estimation methods are defined [9] as

$$\begin{aligned}
\text{ULS} &= \mathbf{tr}(S-\Sigma)^T(S-\Sigma) \\
\text{GLS} &= \mathbf{tr}(S-\Sigma)^T S^{-T} S^{-1}(S-\Sigma) \\
\text{ML} &= \log\det\Sigma - \log\det S + \mathbf{tr}(S\Sigma^{-1}).
\end{aligned}$$

The proposed formulation (7) is an ML estimation with constraints, thus, we used ML estimator.

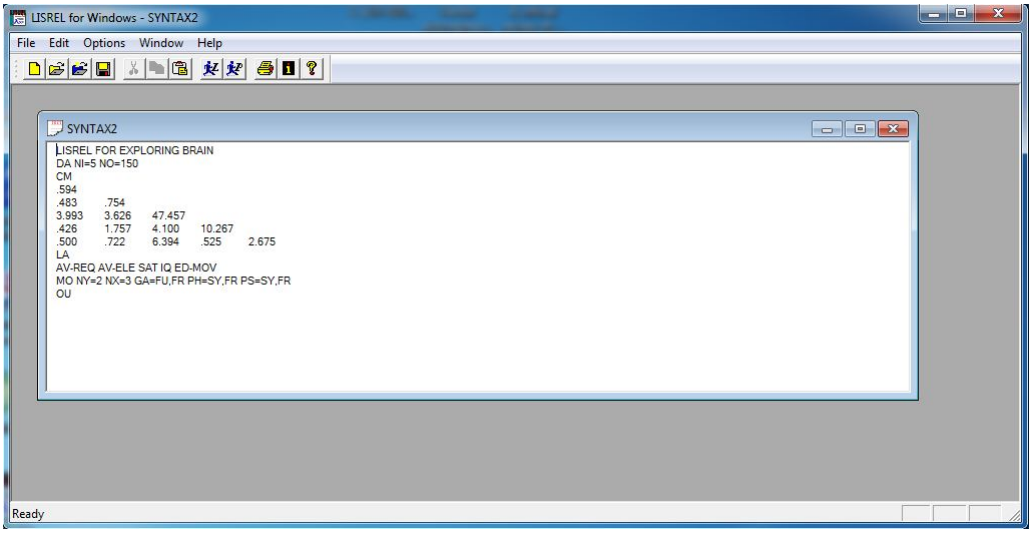


Figure 4: The `LISREL` syntax window which can be generated manually by using text editor or automatically generated from a path diagram.

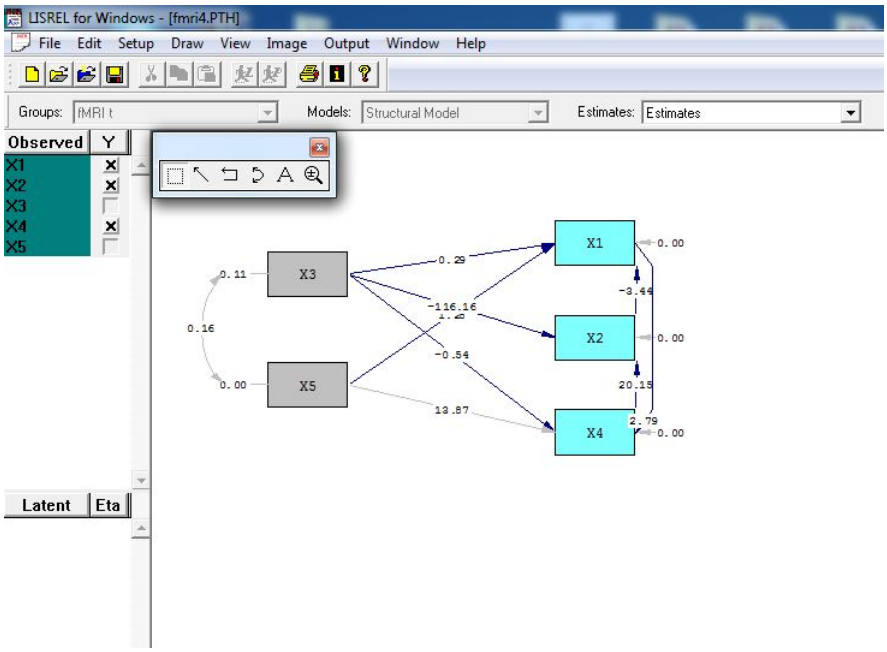


Figure 5: The path diagram window in `LISREL`.

In this work, we try to solve SEM using fMRI data. This problem is based on path analysis, therefore, we prefer to use LISREL with a path diagram to find the parameters. Firstly, we have to define our assumption about the causal relationship among all the ROIs as shown in the Figure 7. X_i denotes the region i in the brain.

From `StarPlus` fMRI data, we randomly pick 5 voxels (defined as X_1 to X_5). One time point is treated as one sample, and we have 50 time points of X_i meanning that we have 50 samples. We compute the sample covariance matrix of X by `MATLAB` (`cov`(X) command) and then import to `LISREL`. In this step we can solve this problem and the result is shown in Figure 8.

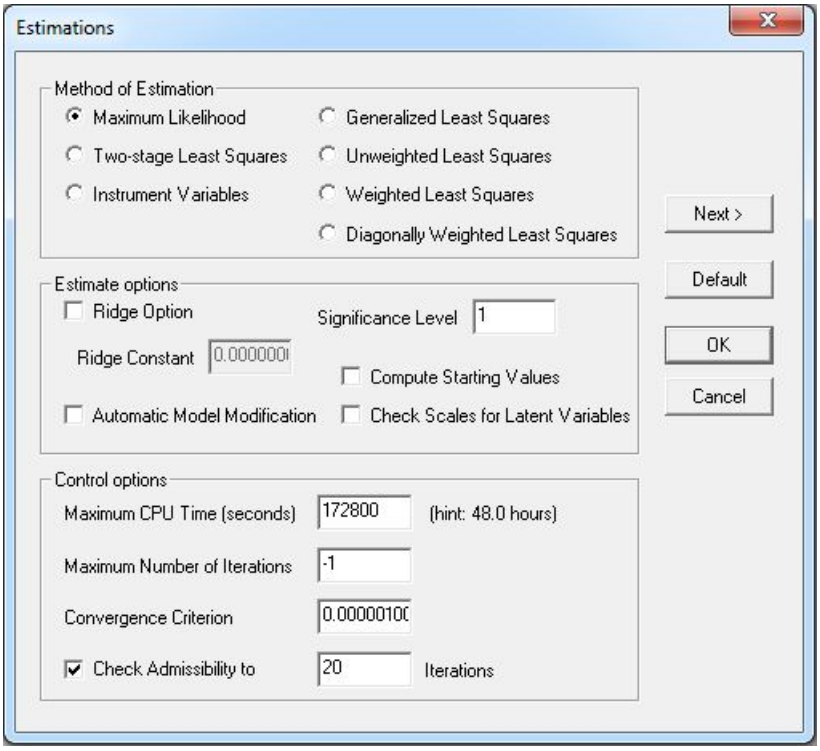


Figure 6: The estimation method selection window in LISREL.

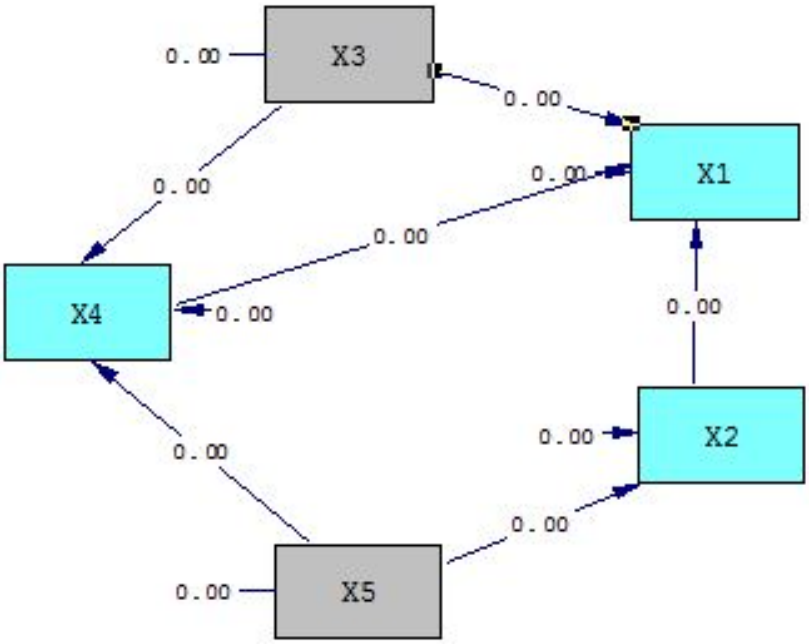


Figure 7: A presumed path diagram for 5 observed variables.

In Figure 8, df stand for degrees of freedom, RMSEA stands for root mean square error of approximation, and Chi-Square is the likelihood ratio chi-square (χ^2). Since the test statistic of maximum likelihood estimation is χ^2 distributed, in hypothesis testing, a non-significant χ^2 implies that there is no significant discrepancy between the estimated covariance matrix and the empirical covariance matrix. Therefore, a non-significant χ^2 indicates that the model fits with the data [9].

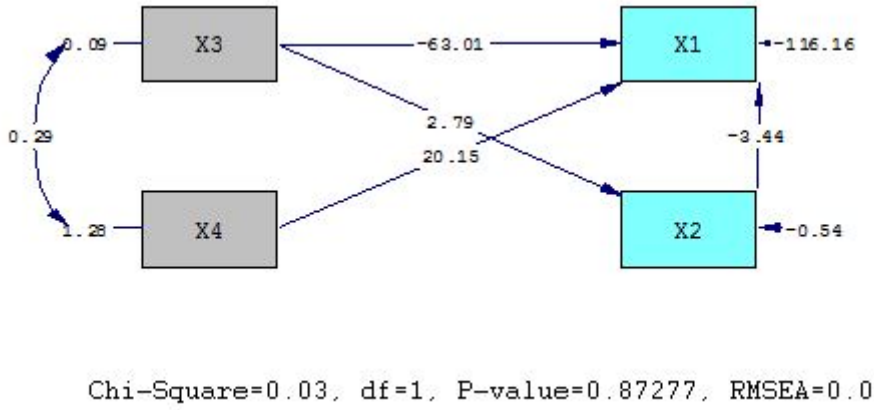


Figure 8: The solved path diagram from the sampled fMRI data.

Once we obtain the path coefficient that minimize the distance between $\mathbf{cov}(X)$ and its sample covariance, we try to use `MATLAB` to express the above picture in 3 dimensions with real coordinates of the chosen voxels, and the result is shown in Figure 9.

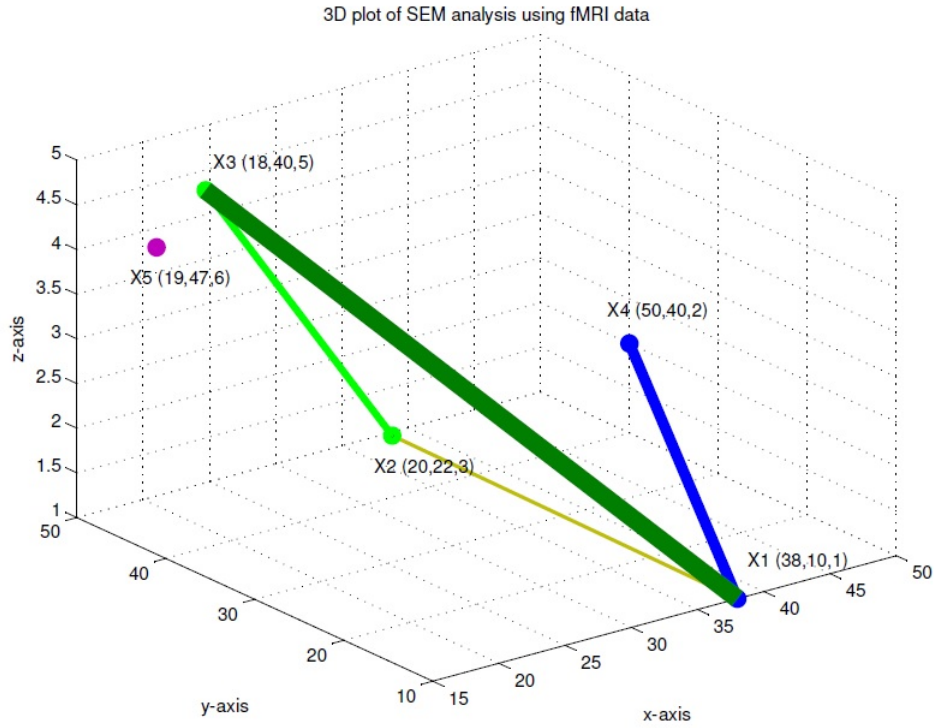


Figure 9: The 3D-diagram model of SEM analysis using `StarPlus` fMRI data. The line weight represents the connections strength between voxels.

5.2 Convex program

To solve the convex problem from (7), we use `CVX` [10, 11]. `CVX` is a `MATLAB`-based modeling system for convex optimization. `CVX` turns `MATLAB` into a modeling language, allowing constraints and objectives to be specified using standard `MATLAB` expression syntax. From `StarPlus` fMRI data, we randomly pick 5 voxels (defined as X_1 to X_5). One time point is treated as one sample, and we have 55 time points of X_i meaning that we have 55 samples. We try to solve the problem with $n = 5$, $\alpha = 1$, and S is sample covarince matrix of 5 voxels. S and the result of A are shown below.

$$S = \begin{bmatrix} 13.1776 & -0.8818 & -0.8368 & -1.6614 & -0.7553 \\ -0.8818 & 3.5246 & -0.1580 & -0.7471 & 0.2329 \\ -0.8368 & -0.1580 & 3.0973 & 0.1068 & 1.4363 \\ -1.6614 & -0.7471 & -0.1068 & 5.9328 & 0.1014 \\ -0.7553 & 0.2329 & 1.4363 & 0.1014 & 5.3736 \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & -0.3046 & -0.2384 & -0.2938 & -0.0507 \\ -0.0636 & 0 & -0.0740 & -0.1107 & 0.0447 \\ -0.0365 & -0.0543 & 0 & -0.0070 & 0.1835 \\ -0.1194 & -0.2155 & -0.0187 & 0 & 0.0138 \\ -0.0168 & 0.0709 & 0.3974 & 0.0113 & 0 \end{bmatrix}$$

The `CVX` code for solving convex problem is as follows.

```
cvx_begin sdp
    variables X(n,n) A1(n,n) Phi(n,n)
    minimize -log_det(X)+trace(S*X)
    subject to
        [X , (eye(n)-A1)' ; (eye(n)-A1) , Phi ] >= 0;
        %constraint on Phi
        Phi <= alpha*eye(n);
        %constraint on A
        diag(A1) == 0;
cvx_end
```

5.3 Comparison of solving methods

As `LISREL` needs at least one independent variable and can not explore all possibility of paths between observed variables, we start from using X_5 as an independent variable and X_1, X_2, X_3 and X_4 as dependent variables. Because we try to confirm the result from `CVX`, then, we need to fix diagonal covariance of the error of prediction matrix, denotes as $\texttt{diag}(\Psi_l)$, to 1 as same as the constraint in convex problem from previous experiment. As seen in Figure 10, `LISREL` shows us that this structure has 15 degrees of freedom, that means we can construct at most 15 paths. After that, we construct 13 paths between variables as shown in Figure 11 and then we get the result as shown in Figure 12.

The estimated path coefficient can be seen from the output file. It can be shown in matrix form as A_l.

$$A_l = \begin{bmatrix} 0 & -0.305 & -0.250 & -0.294 & -0.055 \\ -0.064 & 0 & -0.092 & -0.111 & 0.061 \\ 0 & 0 & 0 & 0.008 & 0.267 \\ -0.120 & -0.214 & -0.011 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{8}$$

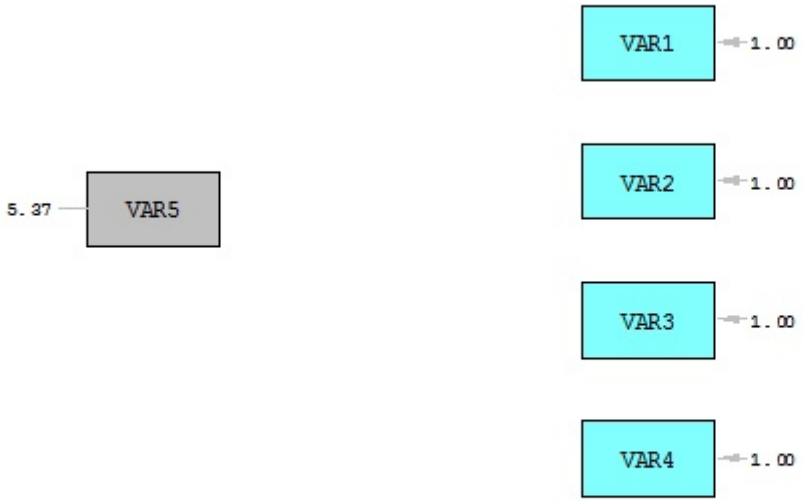


Figure 10: The degree of freedom of this structure is 15 which imply that we can construct at most 15 paths.

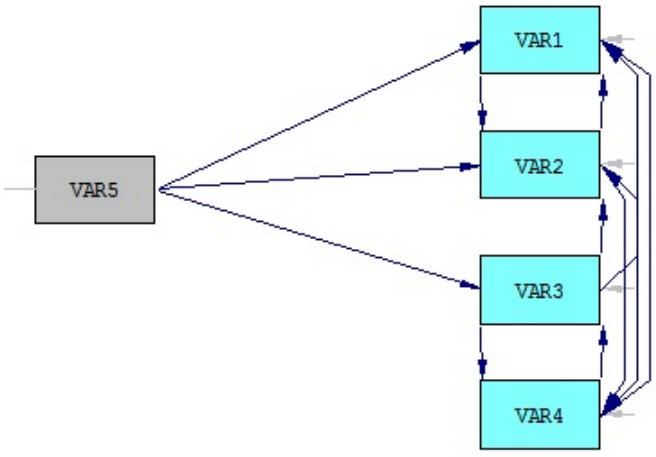


Figure 11: A presumed path diagram that contains 13 paths.

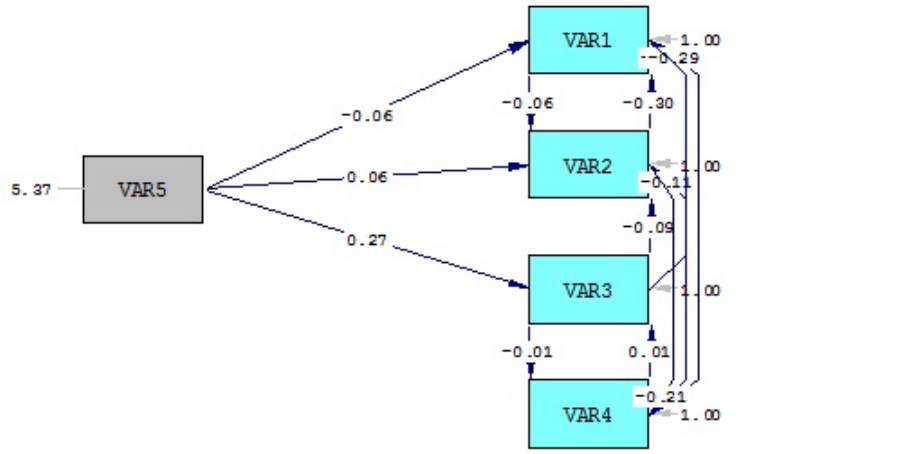


Figure 12: A result of path diagram solved by `LISREL`.
L

After that, we estimate the path coefficient matrix by using `CVX`, A_m. In order to make A_m comparable to A_l, we add new constraints on A_m by setting some of its entries to be zero as same as A_l. In this case, the new constraints are $a_{31}, a_{32}, a_{45}, a_{51}, a_{52}, a_{53}, a_{54} = 0$. The result is shown below.

$$A_m = \begin{bmatrix} 0 & -0.3047 & -0.2498 & -0.2937 & -0.0550 \\ -0.0635 & 0 & -0.0923 & -0.1109 & 0.0612 \\ 0 & 0 & 0 & 0.0077 & 0.2671 \\ -0.1198 & -0.2141 & -0.0114 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{9}$$

It can be seen that when $\alpha = \texttt{diag}(\Psi_l) = 1$, A_m and A_l are very close. Subsequently, we vary the value of α and Ψ at the same time to show that A_m and A_l are still close. The result is shown as a norm of the difference of A_m and A_l.

α, $\texttt{diag}(\Psi_l)$	$\|A_m - A_l\|$
$0.25\lambda_{\min}(S)$	8.2559×10^{-4}
$0.5\lambda_{\min}(S)$	8.2734×10^{-4}
$0.75\lambda_{\min}(S)$	7.1767×10^{-4}
$\lambda_{\min}(S)$	8.6710×10^{-4}

Table 1: Norm of the difference of A_m and A_l using different value of α and $\texttt{diag}(\Psi_l)$

From the result of the simulation, we have a hypothesis that for $\alpha \leq \lambda_{\min}(S)$, the path coefficient matrix obtained from `LISREL` and `CVX` are quite equal.

5.4 Model selection with BIC score

From the last section, we have seen from our several simulations that if we choose $\alpha \leq \lambda_{\min}(S)$, then the path coefficient matrices, A, from `LISREL` and `CVX` can have similar structures. Therefore, we can use `CVX` to solve the exploring connectivity problem if we choose the suitable value of α. This leads to the reason that we will choose $\alpha = \lambda_{\min}(S)$ for other experiments.

In this section we explore the brain connectivity from 3 fMRI data sets by using `CVX` solving to find the path coefficient matrix A. In our experiment, we set $\alpha = \lambda_{\min}(S)$ and $n = 5$. Then we obtain 3 dependence structures from 3 data sets as shown below.

$$A_1 = \begin{bmatrix} 0 & -0.2985 & -0.2420 & -0.2776 & -0.0487 \\ -0.0327 & 0 & -0.0561 & -0.0656 & 0.0293 \\ -0.0136 & -0.0288 & 0 & -0.0016 & 0.0910 \\ -0.0918 & -0.1976 & -0.0097 & 0 & 0.0125 \\ -0.0119 & 0.0651 & 0.3930 & 0.0092 & 0 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 0 & 0.2443 & 0.2640 & 0.0078 & -0.1234 \\ 0.0080 & 0 & -0.0491 & -0.0358 & -0.0038 \\ 0.0403 & -0.2278 & 0 & -0.016 & 0.0305 \\ 0.0164 & -0.2286 & -0.0221 & 0 & 0.0270 \\ -0.0381 & -0.0355 & 0.0618 & 0.0397 & 0 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 0 & -0.2133 & -0.0827 & 0.1729 & 0.2284 \\ -0.0174 & 0 & -0.0276 & 0.0610 & -0.0023 \\ -0.0168 & -0.0686 & 0 & -0.0038 & 0.0991 \\ 0.0334 & 0.1442 & -0.0036 & 0 & -0.0535 \\ 0.0506 & -0.0063 & 0.1080 & -0.0614 & 0 \end{bmatrix}$$

where A_i denotes the path coefficient from model i. Once we obtain A from each model, all of the matrices A are dense and they do not imply the significant dependence structure. Then we would like to make it sparser by setting some entries of A to be zero. For each A, we have normalized all the entries in A to be in range $[-1, 1]$ and choose the value called *threshold*. If absolute value of some entries in A is lower than threshold, then those entries are eliminated by setting them to be zero. For example, we choose the threshold to be 0.1 and it can be expressed as mathematical equation below

$$a_{ij} = \begin{cases} 0 & \text{if } |a_{ij}| < 0.1 \\ a_{ij} & \text{otherwise} \end{cases}$$

From this step, we have 3 sparser dependence structures from 3 data sets as shown below

$$A_1 = \begin{bmatrix} 0 & -0.2985 & -0.2420 & -0.2776 & -0.0487 \\ 0 & 0 & -0.0561 & -0.0656 & 0.0293 \\ 0 & 0 & 0 & 0 & 0.0910 \\ -.0918 & -0.1976 & 0 & 0 & 0 \\ -0.0119 & 0.0651 & 0.3930 & 0.0092 & 0 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 0 & 0.2443 & 0.2640 & 0.0078 & -0.1234 \\ 0 & 0 & -0.0491 & -0.0358 & 0 \\ 0.0403 & -0.2278 & 0 & 0 & 0.0305 \\ 0 & -0.2286 & 0 & 0 & 0.0270 \\ -0.0381 & -0.0355 & 0.0618 & 0.0397 & 0 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 0 & -0.2133 & -0.0827 & 0.1729 & 0.2284 \\ 0 & 0 & -0.0276 & 0.0610 & 0 \\ 0 & -0.0686 & 0 & 0 & 0.0991 \\ 0.0334 & 0.1442 & 0 & 0 & -0.0535 \\ 0.0506 & 0 & 0.1080 & -0.0614 & 0 \end{bmatrix}$$

In our experiment, to see the effect of threshold parameter and the sparsity pattern of the path matrices, A, from 3 data sets, we vary this threshold parameter to be 0.1, 0.2, 0.3, 0.4 and 0.5 sequentially. From this step, we will have many different structures. In this case, BIC score plays important role for selecting the most suitable model from all models of each data set. It is suggested to use BIC score because it performs well for selecting the correct number of factor in exploratory factor analysis when the number of factor is small [12]. The BIC score for SEM is given by

$$\mathrm{BIC}_i = \hat{G}_i + d_i \log N \tag{10}$$

where
$\hat{G}_i = (N-1)(\log\det\hat{\Sigma} + \mathbf{tr}(S\hat{\Sigma}^{-1}) - \log\det S - n)$,
d_i : the number of effective parameters of model i,
N : the number of sample data.

The result of our simulation of each data set with varying threshold parameters as mentioned above is shown in the Table 2.

threshold	$\mathbf{BIC}_i = \hat{G}_i + d_i \log N$		
	data set 1	data set 2	data set 3
0.1	254.8017	315.7840	143.2459
0.2	240.4732	285.8395	135.5335
0.3	236.1721	279.2249	121.0622
0.4	236.1721	279.2249	114.2358
0.5	236.1721	277.7280	109.1742

Table 2: BIC score at different selection threshold for each data set

From Table 2, we would like to reasonably choose the minimal BIC score from each data set to select the best model to use. The data set 1 we obtain minimal BIC score is 236.1721 at threshold 0.3, data set 2 BIC score is 277.7280 at threshold 0.5 and data set 3 BIC score is 109.1742 at threshold 0.5. The best models with sparse structure in the sense of minimum BIC score for each data set are shown in Figure 13.

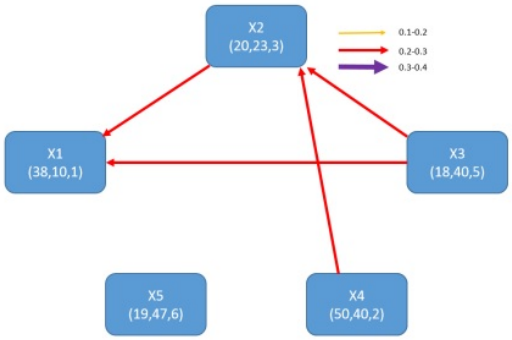
6 Conclusion

A convex formulation of SEM in fMRI study has been proposed and the global optimal solution of path coefficient matrix, thus, can certainly be obtained using numerical methods. Experiments have been introduced, using both proposed formulation and well known `LISREL` software, then, compared together for verification. From our simulation, the requirement for validity of proposed formulation is that noise covariance must be bounded by the lowest eigenvalue of sample covariance matrix of observed fMRI data set *i.e.*, $\alpha \leq \lambda_{\min}(S)$. Obtained path coefficient matrices are then made sparser by eliminating all entries in A which have the absolute value lower than threshold value after they are normalized to be in range $[-1, 1]$. To select the suitable model, we use the BIC score as our criterion and choose the one that has the minimum BIC score as our best model for each data set.

$$A_1 = \begin{bmatrix} 0 & -0.2985 & -0.2420 & -0.2776 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -0.1976 & 0 & 0 & 0 \\ 0 & 0 & 0.3930 & 0 & 0 \end{bmatrix}$$

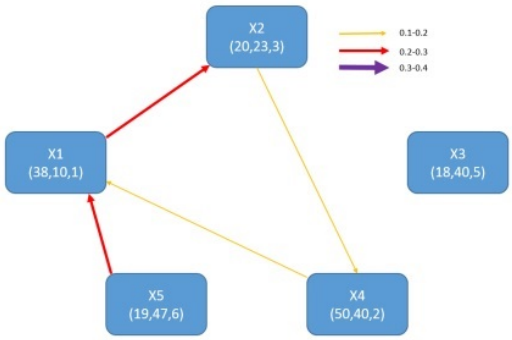
(a)

$$A_2 = \begin{bmatrix} 0 & 0.2443 & 0.2640 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -0.2278 & 0 & 0 & 0 \\ 0 & -0.2286 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$



(b)

$$A_3 = \begin{bmatrix} 0 & -0.2133 & 0 & 0.1729 & 0.2284 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1442 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$



(c)

Figure 13: The dependence structures of the best model in the sense of minimum BIC score from a) data set 1, b) data set 2 and c) data set 3

7 Appendix

7.1 Convex problem

An optimization problem is said to be convex obtimization if

1. Cost function is a convex function
2. Constraint set is convex

Let us recall the problem (7) again as follows.

$$\begin{aligned} \underset{X,A,\Psi}{\text{minimize}} \quad & -\log\det X + \mathbf{tr}(SX) \\ \text{subject to} \quad & \begin{bmatrix} X & (I-A)^T \\ I-A & \Psi \end{bmatrix} \succeq 0, \\ & \Psi \preceq \alpha I, \\ & \mathbf{diag}(A) = 0, \end{aligned}$$

with variables $X \in \mathbb{S}^n, A \in \mathbb{R}^{n\times n}$ and $\Psi \in \mathbb{S}^n$. In this section, we would like to show that the above problem is a convex optimization problem by showing the cost function is a convex function and all of three constraint are convex. Let us define cost function, $g(x) = f(X) + h(X)$, where $f(X) = -\log\det X$ and $h(X) = \mathbf{tr}(SX)$.

Proof: $f(x) = -\log\det X$ is a convex function.
Definition : $f : \mathbf{R^n} \to \mathbf{R}$ is convex if and only if the function $g : \mathbf{R} \to \mathbf{R}$,

$$g(t) = f(x+tv),\ \mathbf{dom}g = \{t | x+tv \in \mathbf{dom}f\}$$

is convex in t for any $x \in \mathbf{dom}f, v \in \mathbf{R^n}$.

For $f(X) = -\log\det X$, $\mathbf{dom}X = \mathbf{S^n_{++}}$.

$$\begin{aligned} g(t) = -\log\det(X+tV) &= -\log(\det(X)\det(X^{-1/2}(X+tV)X^{-1/2})) \\ &= -\log\det(X) - \log\det(I + tX^{-1/2}VX^{-1/2}) \\ &= -\log\det(X) - \sum_{i=1}^{n}\log(1+t\lambda_i) \end{aligned}$$

where λ_i are the eigenvalues of $X^{-1/2}VX^{-1/2}$.

This is a sum of constant and convex function (negative logarithm function); hence $g(t)$ is convex. So $f(X)$ is convex.

Proof: $h(X) = \mathbf{tr}(SX)$ is a convex function.
Definition : $f : \mathbf{R^n} \to \mathbf{R}$ is convex if $\mathbf{dom}f$ is convex set and

$$f(\theta x_1 + (1-\theta)x_2) \le \theta f(x_1) + (1-\theta)f(x_2)$$

for all $x_1, x_2 \in \mathbf{dom}f, 0 \le \theta \le 1$.

For h(X) = $\mathbf{tr}(SX)$, S is a constant positive definite, and $\mathbf{dom}X = \mathbf{S^n_{++}}$ which is convex set

$$\begin{aligned} h(X) &= \mathbf{tr}(SX) \\ h(\theta X_1 + (1-\theta)X_2) &= \mathbf{tr}(\theta SX_1 + (1-\theta)SX_2) \\ &= \mathbf{tr}(\theta SX_1) + \mathbf{tr}((1-\theta)SX_2) \\ &= \theta\mathbf{tr}(SX_1) + \mathbf{tr}(SX_2) - \theta\mathbf{tr}(SX_2) \\ h(\theta X_1 + (1-\theta)X_2) &= \theta h(X_1) + (1-\theta)h(X_2) \end{aligned}$$

Therefore $h(X)$ is a convex function.

Proof: $f(X) + h(X)$ is a convex function.

For $f(X)$ and $h(X)$ are convex functions.

$$\begin{aligned}(f+h)(\theta X_1 + (1-\theta)X_2) &= f(\theta X_1 + (1-\theta)X_2) + h(\theta X_1 + (1-\theta)X_2) \\ &\leq \theta f(X_1) + (1-\theta)f(X_2) + \theta h(X_1) + (1-\theta)h(X_2) \\ &\leq \theta(f+h)(X_1) + (1-\theta)(f+h)(X_2)\end{aligned}$$

Therefore $f(X) + h(X)$ is a convex function.

Proof: The constraints of problem (7) are convex.

Consider the constraints

$$\begin{bmatrix} X & (I-A)^T \\ I-A & \Psi \end{bmatrix} \succeq 0, \tag{11}$$

$$\begin{aligned} \Psi &\preceq \alpha I \\ \alpha I - \Psi &\succeq 0, \end{aligned} \tag{12}$$

$$\mathbf{diag}(A) = 0. \tag{13}$$

(11) and (12) are inequality constraints, and for all solutions of cost function (X, A, Ψ) these constraints are sets of positive definite matrix; hence (11) and (12) are convex. (13) is a linear equality constraint; hence (13) is convex.

Because the cost function and the constraints are a convex function and convex sets, respectively, this problem has completely turned into the convex optimization problem.

7.2 Schur complement of matrix

In the theory of matrices, there is the use of Schur complements in several contexts and they also appear in various important theorems. Let us define a matrix $M \in \mathbb{S}^n$ which is partitioned by

$$M = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

where $A \in \mathbb{S}^k$. If A is invertible, then the matrix

$$S = C - B^T A^{-1} B,$$

is called the *Schur complment* of A in M. An important property of the Shur complement is positive definiteness and positive semidefiniteness property in which they are explained by

$$M \succ 0 \iff A \succ 0 \text{ and } C - B^T A^{-1} B \succ 0,$$

and

$$\text{If } A \succ 0, \text{ then } M \succeq 0 \iff C - B^T A^{-1} B \succeq 0.$$

7.3 Kullback-Leibler divergence

In probability theory, the Kullback-Leibler divergence function is a well-known distance function for measuring the difference between two probability distribution. Let $f_1(x)$ and $f_2(x)$ be continuous distribution, then by definition, the Kullback-Leibler divergence is defined as

$$d(f_1(x), f_2(x)) = \int [\log\left(\frac{f_1(x)}{f_2(x)}\right)] f_1(x) dx. \tag{14}$$

From this definition, we can show that the Kullback-Leibler divergence between two multivariate Gaussian distributions $f_1(x)$ and $f_2(x)$, with mean μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2, is given by

$$d(f_1(x), f_2(x)) = \frac{1}{2}\{\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - n + \mathbf{tr}(\Sigma_2^1\Sigma_1) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2), \tag{15}$$

where n denotes the dimension of variable. Let us start with the definition of Kullback-Leibler divergence from (14)

$$\begin{aligned}
d(f_1(x), f_2(x)) &= \int [\log\left(\frac{f_1(x)}{f_2(x)}\right)] f_1(x)dx, \\
&= \int [\log(f_1(x)) - \log(f_2(x))] f_1(x)dx, \\
&= \int \frac{1}{2}[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)] f_1(x)dx, \\
&= \mathbf{E}\left[\frac{1}{2}[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)]\right], \\
d(f_1(x), f_2(x)) &= \frac{1}{2}\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \mathbf{E}[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)] + \mathbf{E}[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)]\right].
\end{aligned}$$

If we consider the term $\mathbf{E}[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)]$ and $\mathbf{E}[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)]$, we can write them as

$$\mathbf{E}[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)] = \mathbf{E}[\mathbf{tr}((x-\mu_1)(x-\mu_1)^T\Sigma_1^{-1})] = \mathbf{E}[\mathbf{tr}(I_n)] = n,$$

and

$$\mathbf{E}[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)] = (\mu_1-\mu_2)^T\Sigma_2(\mu_1-\mu_2) + \mathbf{tr}(\Sigma_1\Sigma_2^{-1}).$$

Therefore

$$d(f_1(x), f_2(x)) = \frac{1}{2}\left[\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - n + \mathbf{tr}(\Sigma_1\Sigma_2^{-1}) + (\mu_1-\mu_2)^T\Sigma_2(\mu_1-\mu_2)\right],$$

which shows our statement in (15). Note that, from our proof, we have used some properties from section 8 of The Matrix Cookbook [13]. For another example, it can be shown that the Kullback-Leibler divergence of two Bernoulli random variables p and q is given by

$$d(f(k;p), f(k;q)) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right),$$

where $f(k;p)$ denotes the probability mass function of Bernoulli distribution over possible outcomes k.

References

[1] K. Jöreskog and Sörbom, "LISREL 8.8 for windows." http://www.ssicentral.com/lisrel/, 2006.

[2] "S-PLUS." http://www.solutionmetrics.com.au/products/splus/default.html.

[3] "M-PLUS." http://www.statmodel.com/.

[4] "EQS." http://www.mvsoft.com/eqs60.htm.

[5] "Analytic programming with fMRI data: a quick-start guide for statisticians using r." http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3938835/.

[6] G. Chen, D. Glen, Z. Saad, J. Hamilton, M. Thomason, I. Gotlib, and R. Cox, "Vector autoregression, structural equation modeling, and their synthesis in neuroimaging data analysis," *Computers in Biology and Medicine*, vol. 41, pp. 1142–1155, 2011.

[7] "Resting-state fMRI: a window into human brain plasticity." http://nro.sagepub.com/content/20/5/522.full.pdf+html".

[8] "Starplus fMRI data." http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/.

[9] E. Kelloway, *Using LISREL for structural equation modeling.* SAGE, 1998.

[10] M. Grant and S. Boyd, "Graph implementations for non smooth convex programs," in *Recent advances in learning and control* (V. Blondel, S. Boyd, and H. Kimura, eds.), Lecture notes in control and information sciences, pp. 95–110, Springer-Verlag Limited, 2008. http://stanford.edu/~boyd/graph_dcp.html.

[11] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," mar 2014.

[12] K. Preacher and E. Merkle, "The problem of model selection uncertainty in structural equation modeling," *Psychological Methods*, vol. 17, no. 1, pp. 1–14, 2012.

[13] K. Petersen and M.S.Pedersen, "The Matrix Cookbook." hhttp://matrixcookbook.com.