Convex Formulation for Regularized Estimation of Structural Equation Models

Anupon Pruttiakaravanich and Jitkomut Songsiri*
Department of Electrical Engineering, Faculty of Engineering
Chulalongkorn University
anupon106@gmail.com, jitkomut.s@chula.ac.th
September 18, 2018

Abstract
Path analysis is a special class of models in structural equation modeling (SEM) where it
describes causal relations among measured variables in a form of multiple linear regression. This
paper presents two alternative estimation formulations for confirmatory and exploratory SEM in
path analysis problems where a zero pattern of the estimated path coefficient matrix can explain
a causality structure of the variables. In confirmatory SEM, the original nonlinear equality
constraints of the model parameters are relaxed to an inequality, allowing us to transform the
original problem into a convex problem that can be solved by many existing efficient algorithms.
A regularized estimation formulation is proposed for exploratory SEM, where the objective
function is added with an \(\ell_1\)-type penalty of the path coefficient matrix. Under a condition on
problem parameters, we show that our optimal solution is low rank and provides an estimate
of the path matrix of the original problem. To solve our estimation problems in a convex
framework, we apply alternating direction method of multiplier (ADMM) which is shown to be
suitable for a large-scale implementation. In combination with applying model selection criteria,
the penalty parameter in the regularized estimation, controlling the density of nonzero entries
in the path matrix, can be chosen to provide a reasonable trade-off between the model fitting
and the complexity of causality structure. The performance of our approach is demonstrated in
both simulated and real data sets, and with a comparison of existing methods. Real application
results include learning causality among climate variables in Thailand where our findings can
explain known relations among air pollutants and weather variables. The other experiment is to
explore connectivities among brain regions using fMRI time series from ABIDE data sets where
our results are interpreted to explain brain network differences in autism patients.

*Corresponding author
1 Introduction

Structural equation modeling (SEM) is a class of multivariate models used for learning a causal relationship among variables called exploratory modeling, or for testing whether the model is best fit by the given data (called confirmatory modeling). A general SEM includes observed and latent variables while their relationships are explained by a linear model whose parameters explain a cause or influence from one variable to another. SEM has been used widely in behavioral researches such as in psychology [MA00], sociology [HAA11, Suk13], business [RM11] and medical research as well [MGL94, PLFI09]; see details of model and history background in SEM [Bol89, §1]. Path analysis is a special problem in SEM where it provides a model for explaining relationships among measured variables only (no latent variables). This can be more associable with scientific research where observed variables are often of primary interest. For example, one aims to explore causal relationship among brain regions from brain signals (such as fMRI data) [MGL94, BF97, BHH+00, KZC+07, CGS+11] where the entries of the path coefficient matrix in the model explain how much change in the activity of one region influences another region.

In path analysis, one applies a prior knowledge about relationship structure of variables of interest to construct a model and encode such structure as the zero structure of the path matrix in the model. The first problem type in path analysis, called confirmatory SEM is to estimate the value of nonzero entries in the path matrix and the covariance matrix of model residual errors so that the model-reproduced covariance matrix fits well with the sample covariance matrix in an optimal sense, evaluated by various types of criterion functions such as maximum likelihood, ordinary or weighted least-squares [Bol89, §4]. The second type of problem called exploratory SEM is to learn a causal structure of variables from a zero structure of the estimated path matrix. An existing approach for the exploratory SEM is to begin with a base model where a certain set of paths are affirmative but the existence of some other paths is in question. This results in a set of a few candidate models associated with different zero structures of the path matrix and the significance of the difference between these models can be determined from the $\chi^2$ statistic [Bol89, §7]. Examples of this approach can be seen in brain network study [MGL94, BF97] where only a few variables (in the order up to 10 brain regions) are selected. One can locally search for a path structure by starting from a null model (all path coefficients are zero) and sequentially allowing the coefficient corresponding to the largest Lagrangian multiplier to be nonzero [BHH+00]. The most optimal but far from feasible approaches is to perform an exhaustive search that enumerates all possible path pattern with a fixed number of paths and chooses the model corresponding to the lowest minimized maximum likelihood function [CGS+11]. It is known that the number of all possible models grows exponentially to the number of variables, so it is not feasible as the number of variables increases.

Both confirmatory and exploratory SEM problems are nonlinear optimization problems in matrix variables with quadratic equality and positive definite cone constraints. Common techniques based on Newton-Raphson or gradient descent are implemented to estimate the model parameters [Mul09, §7], [Bol89, §4] and there are many existing SEM commercial softwares such as LISREL, EQS, Mplus [RM06, JSTT00, Bol89], so in an estimation process, a starting value for the update iteration is required. Though these numerical methods work well under normal conditions, it is also known that some initial values may not lead to the convergence of the optimal solution or may stick into local minima, hence several strategies for selecting initial values have been proposed [Bol89, §4]. These include choosing an instrumental variable estimate or selecting the strength of the path coefficient magnitude. When the iterative method in these softwares does not converge, the user is suggested not to interpret the result.

In this work, we present two alternative estimation formulations for both confirmatory and exploratory SEM problems. In addition, the original nonlinear equality constraints of the model parameters are relaxed to an inequality, leading the problems transformed into convex formulations that can be solved efficiently by many existing convex program solvers where the solution is guaranteed to be the global minima. For exploratory SEM, we propose a formulation whose objective function is added with an $\ell_1$-type regularization of the path coefficient matrix, called sparse SEM. This is a known result that such formulation is regarded as a lasso formulation [HTF09] and doing so encourages many zeros in the path matrix solution, allowing us to read off the zero pattern and
interpret it as a causal structure of the variables. Solving exploratory SEM using a regularization approach was previously discussed in [JGM16] where Recticular Action Model (RAM) is the focus of model class which is more general than the model used in the path analysis problem. Our approach then can be served as a convex framework targeted to solve a special class of RAM, which will be shown later to perform better in particular cases. We solve our two formulations by the alternating direction method of multipliers or ADMM, which requires a feasible amount of memory storage suitable for large-scale implementation. Suggestion on the choice of algorithm parameter is also provided which is suggested from the properties of the proposed formulation. More importantly, we will show that, under a condition on problem parameters of both confirmatory and sparse SEM, our optimal covariance error is diagonal, meaning that errors are uncorrelated, and the optimal solution has low rank, providing an estimate of the path matrix for the original problem.

Despite a difference in our estimation formulation and the original one, we believe that our proposed formulations serve two folds. Firstly, unlike previous SEM applications that only a few variables are of interest, many applications tend to consider a much larger number of variables such as fMRI studies where the variables are neuronal activities and its number is up to thousand [BS09]. Existing approaches of learning causal structures in the exploratory SEM may experience a computational difficulty in terms of memory storage or convergence. Secondly, our solution for confirmatory SEM is obtained under an assumption of homoskedasticity of residual errors, so if this assumption holds, ours and the original solution coincide. Even if it does not hold, and our solution is then not optimal for the original problem but ours can be served as a starting value of the iterative algorithm used in the original one in case that the convergence is not obtained.

The two proposed formulations were initially developed in our prior work [PS16, Pru17] and we have adapted some details on the algorithms and included more experiments in this paper. Our paper is organized as follows. Section 2 summarizes the mathematical formulation of path analysis problem where the formulation is the maximum likelihood estimation with a quadratic equality constraint. Section 3 describes our convex formulation for confirmatory SEM and shows that the solution can be further used under the condition of having a low rank solution at optimum. Another convex formulation for exploratory SEM is proposed in Section 3.2 where an $\ell_1$ regularization is introduced in the cost objective. We show that sparse solutions are obtained and the sparsity can be controlled by a regularization parameter. To select the best optimal relation structure, the model selection procedure is explained in Section 3.4. ADMM algorithms for solving our proposed problems are explained in Section 4 where we provide some example of implementation in large-scale. Numerical experiments in Section 5 demonstrate important factors to the performance of our method from simulated data. The results of applying our estimation formulation to learn causal relation of air pollution data and brain regions are explained in Section 6. The derivation of dual problems of our formulations, some mathematical proofs and pseudo codes of algorithms are shown in Appendix, which could be omitted if the reader is familiar with the contents.

Keywords: structural equation model, convex optimization, regularization, brain connectivity

Notation. $S^n$ denotes the set of symmetric matrices of size $n \times n$ and $S_+^n$ denotes the set of positive semidefinite matrices of size $n \times n$. For a square matrix $X$, $\text{tr}(X)$ is the trace of $X$ and $\text{diag}(X)$ is a diagonal matrix containing diagonal entries of $X$. A block symmetric matrix is of the form: $X = \begin{bmatrix} X_1 & X_2^T \\ X_2 & X_4 \end{bmatrix}$. The notations $X \succ 0$ and $X \succeq 0$ refer to $X$ being positive definite, and positive semidefinite, respectively.

2 Path analysis in SEM

Structural equation modeling (SEM) starts with a set of variables involved in a study, measured variables and latent variables. Measured variables are simply the ones that can be directly measured (physical quantities), while latents are variables that cannot be directly (or exactly) measured such as intelligence, attitude, etc. Each of these variables can be regarded as either endogenous or exogenous.
An endogenous variable gets an influence from others while an exogenous variable affects the other variables. A general mathematical model in SEM explains a linear relationship from latent variables to measured variables and also includes error terms of each variable \([82\textbf{Bol89} \textbf{Mul09} \textbf{Hoy95}]\).

In practice, we are commonly interested in the application of SEM that involves only with observable variables. For this reason, we focus on a special class of model in SEM that is described by a multiple linear regression:

\[
Y = c + AY + \epsilon
\]

where \(Y \in \mathbb{R}^n\) is the measured (or observed variables), \(c \in \mathbb{R}^n\) is a constant vector representing a baseline, and \(\epsilon \in \mathbb{R}^n\) is the model error, assumed to be Gaussian distributed. The path matrix, \(A = [A_{ij}] \in \mathbb{R}^{n \times n}\) represents a dependence structure among variables in the model, i.e., if \(A_{ij} = 0\) then there is no path from \(Y_j\) to \(Y_i\). In other words, a pattern of nonzero entries in \(A\) reveals a causal structure of variables in the model. If this structure is assumed from a prior knowledge, then the problem of estimating \(A\) is called confirmatory SEM.

Let \(S\) be a sample covariance matrix of \(Y\) and \(\Sigma\) be the model-reproduced covariance matrix of \(Y\), derived from \(\Psi = \text{cov}(\epsilon)\). An estimation problem in SEM is to find \(A\) and \(\Psi\) that minimize the Kullback-Leibler divergence function

\[
d(S, \Sigma) = \log \det \Sigma + \text{tr}(SS^{-1}) - \log \det S - n
\]

meaning that \(\Sigma\) is close to \(S\), while maintaining that \(\Sigma, A\) and \(\Psi\) are constrained by \(\Psi = \text{cov}(\epsilon)\). Moreover, the structure of the path matrix is presumably encoded by a model hypothesis: i) \(A_{ij} = 0\) if there is no link from \(Y_j\) to \(Y_i\) and ii) we always have \(\text{diag}(A) = 0\), meaning that there is no path from \(Y_i\) to itself. To specify the zero structure of \(A\), we then define the associated index set \(I_A \subseteq \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}\) with properties that i) \((i, j) \in I_A\) if \(A_{ij} = 0\) and ii) \((1, 1), (2, 2), \ldots, (n, n)\} \subseteq I_A\) since \(\text{diag}(A) = 0\). In short, \(I_A\) denotes the index set of hypothetical zero entries in \(A\) and it must include the diagonal entry indices.

Given an index set \(I_A\), we define a projection operator \(P : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}\)

\[
P(X) = \begin{cases}
X_{ij}, & (i, j) \in I_A, \\
0, & \text{otherwise},
\end{cases}
\]

and denote

\[
P^c = I - P.
\]

The operators \(P^c\) and \(P\) are both self-adjoint, i.e., \(\text{tr}(Y^TP(X)) = \text{tr}(P(Y)^TX)\) and that \(P^c(P(X)) = 0\). These two projection operators will be used in the duality of our estimation formulations. With the definition of \(P\) and a change of variable \(X = \Sigma^{-1}\), the estimation problem corresponding to the confirmatory SEM is

\[
\text{minimize} \quad -\log \det X + \text{tr}(SX) - \log \det S - n
\]

subject to

\[
\begin{align*}
X &= (I - A)^T \Psi^{-1}(I - A), \\
P(A) &= 0,
\end{align*}
\]

with variables \(A \in \mathbb{R}^{n \times n}, \Psi \in \mathbb{S}^+_n\) and \(X \in \mathbb{S}^+_n\). The condition \(P(A) = 0\) basically explains the zero constraint on the entries of \(A\), and when there is no information on the path matrix, this condition reduces to \(\text{diag}(A) = 0\). The problem \(5\) is one of estimation formulations considered in SEM context \([82\textbf{Bol89}] \textbf{§4}].\) Other cost objectives are also used such as ordinary or weighted least-squares.

**Special case.** If the constraint \(P(A) = 0\) reduces to \(\text{diag}(A) = 0\) (we allow \(A\) to have as many free parameters as possible), then we can make the cost objective zero by solving \(S^{-1} = (I - A)^T \Psi^{-1}(I - A)\) where \(S\) is given while \(A\) and \(\Psi\) are free variables. In this case, we can arbitrarily make \(\Psi\) diagonal. In other words, one can always find a factor \(B\) with \(\text{diag}(B) = 1\) and a diagonal
such that \( S^{-1} = B^TDB \). Such factors can be obtained by performing an eigenvalue or \( LDL^T \) decompositions of \( S^{-1} \). In this case, the optimal path matrix to (5) is not unique; one can obtain \( A \) as dense (non-recursive model) or lower triangular matrix (recursive model). This could be problematic if one would like to read a causality structure from the zero pattern in the estimated \( A \). For this reason, it is common to assume some structure in \( A \) and diagonal structure in \( \Psi \) (meaning the error terms are uncorrelated). Specifically, the degree of freedom (df) is defined as

\[
df = \text{the number of known parameters} - \text{the number of estimated parameters}.
\]

Referring to (5), the number of known parameters is the number of entries in the sample covariance matrix and is equal to \( n(n - 1)/2 \) where \( n \) is the number of observed variables. The number of free parameters in (6) is the total number of entries in \( A \) plus the total number of entries in \( \Psi \). One can use df as a guideline for identifying the uniqueness of solution. When df is negative, the estimator may not be unique. We say that the model is identifiable if the df is nonnegative \( [RM06, \text{p. 35}] \).

Therefore, to find a unique solution of a path analysis, we must have some assumptions on the path matrix \( A \) and noise covariance \( \Psi \) to attain the nonnegative values in df as a necessary condition.

3 Convex formulations for path analysis in SEM

The non-convex property of (5) from the quadratic equality constraint is apparent and leads to a chance of obtaining a local minima when solving the problem numerically. This section describes the first contribution of our work; we propose alternative convex formulations and their dual problems for both confirmatory and exploratory SEM. We consider a special case of path analysis problem where the covariance error is allowed to be diagonal, suggesting that the residual errors of the model are assumed to be uncorrelated. The solution to our formulations is useful when it is low rank at optimum which will be shown to occur under some mild conditions on a problem parameter. The solutions to our formulation and the original problem agree when the covariance error of residuals is specified to be a multiple of the identity matrix, which is often the case in SEM applications in behavioral research.

3.1 Confirmatory SEM

Our prior work \( [PS16] \) applied a convex relaxation to the quadratic equality constraint of (5) and proposed the convex confirmatory SEM formulation:

\[
\begin{align*}
\text{minimize} & \quad -\log \det X + \text{tr}(SX) \\
\text{subject to} & \quad \begin{bmatrix} X & (I-A)^T \\ I-A & \Psi \end{bmatrix} \succeq 0, \\
& \quad 0 \preceq \Psi \preceq \alpha I, \quad P(A) = 0,
\end{align*}
\]

with variables \( X \in S^n, A \in \mathbb{R}^{n \times n} \) and \( \Psi \in S^n \), where \( \alpha > 0 \) is a given parameter. After relaxing the non-convex constraint, we have introduced the inequality constraint \( \Psi \preceq \alpha I \) to avoid a trivial solution in (7), e.g., \( \Psi \) can be arbitrarily large and \( A = 0 \). We justify that \( \alpha \) can serve as an upper bound on the covariance error of residual in SEM. The formulation (7) is a semidefinite programming which can be solved by many existing convex program solvers. It is noted in \( [PS16] \) that the problems (5) and (7) are no longer equivalent but the convex formulation, that is solved more efficiently, can provide useful an initial solution when solving (5) numerically.

The dual problem of the convex confirmatory SEM (7) is

\[
\begin{align*}
\text{minimize} & \quad -\log \det(S - Z_1) - 2\text{tr}(Z_2) - \alpha \text{tr}(Z_4) + n \\
\text{subject to} & \quad Z = \begin{bmatrix} Z_1 & Z_3 \\ Z_2 & Z_4 \end{bmatrix} \succeq 0, \quad P^c(Z_2) = 0,
\end{align*}
\]

with variable \( Z \in S^{2n} \) where each block \( Z_k \) has size \( n \times n \). The constraint \( P^c(Z_2) = 0 \) explains that the corresponding entries of block \( Z_2 \) to the zero entries in \( A \) are free, and the other entries of \( Z_2 \)
Proposition 1. Let $\alpha$ a criterion of a choice of $S$ that solves the feasibility problem (9). We estimate $\Sigma$ to be diagonal where $-Y$ can be chosen to be arbitrarily large. Hence, the term $\text{tr}(DY) = 0$ but $-\log \det Y \to -\infty$, leading the cost function to be unbounded below.

Problem assumption
The only assumption required in (7) on a problem parameter is that $S$ must be positive definite. Otherwise, the problem could be unbounded below. To show this, assume $S$ has the eigenvalue decomposition $S = UDU^T$. Then it follows that $\text{tr}(SX) = \text{tr}(UDU^TX) = \text{tr}(DU^TXU)$. Let $f$ be the cost function of (7) and let $Y = U^TXU$. Since $\det X = \det Y$, we can write $f(X) = f(Y) = -\log \det Y + \text{tr}(DY)$. If $S$ is positive semidefinite, then $d_{ii} = 0$ for some $i$, and we can choose $Y$ to be diagonal where $y_{ii}$ is chosen to be arbitrarily large.

Trivial dual solutions
The proposed framework (7) has two problem parameters: $S \succ 0$ and $\alpha > 0$. We have presented an important theoretical results in [PS16] that $\alpha$ becomes an impracticable approximate of $\Sigma^{-1}$ and unfavorable solution of our formulation (7). For a completeness of this paper, we state the proposition in [PS16] and provide the proof here. Let us start with the KKT conditions when zero is an optimal dual solution ($Z = 0$). Let $A = A - P(A)$ for any $A$, optimal primal solutions must satisfy

$$X = S^{-1}, \quad S^{-1} \succeq (I - \hat{A})^T \Psi^{-1}(I - \hat{A}), \quad 0 \prec \Psi \preceq \alpha I. \quad (9)$$

It suggests that under such case (trivial dual solution), and if $\alpha$ is too large, then $\Psi$ can be chosen sufficiently large, the choice of $X = S^{-1}$ is feasible and optimal, which becomes merely a simple estimate of $\Sigma^{-1}$ and unfavorable solution of our formulation (7). Therefore, the proposition addresses a criterion of a choice of $\alpha$ (that should not be too large) to avoid such solution.

Proposition 1. Let $\alpha_c = n/\text{tr}(S^{-1})$ (the harmonic mean of the eigenvalues of $S \succ 0$). If $\alpha \leq \alpha_c$ the feasibility problem (9) has no solution.

See a proof in Appendix C.1 that applies Farka’s lemma to a semidefinite programming.

To apply the formulation (7), when samples of data, $Y_1, Y_2, \ldots, Y_N$ are available, one computes $S$ (the sample covariance of $\{Y_k\}_{k=1}^N$) and choose $\alpha$, which is as of now, suggested to be less than $\alpha_c$. We can show easily that if one simple choice, the minimum eigenvalue of $S$ denoted as $\lambda_{\min}(S)$ is always less than $\alpha_c$. Suppose $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are eigenvalues of $S$. It follows that

$$\frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \cdots + \frac{1}{\lambda_n} \leq \frac{n}{\lambda_1},$$

Since the trace of a matrix is the sum of its eigenvalues, we have $\text{tr}(S^{-1}) = \sum_{k=1}^n 1/\lambda_k$ and this further implies that

$$\alpha_c = \frac{n}{\text{tr}(S^{-1})} = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \cdots + \frac{1}{\lambda_n} \geq \lambda_1 = \lambda_{\min}(S).$$

3.2 Sparse SEM with $\ell_1$-norm regularization
In exploratory SEM analysis, one aims to discover a zero structure of $A$ from the estimation process which reveals a causal structure of how one variable affects to another. An existing approach performs a local search starting from a null model (all path coefficients are zero) and sequentially allows the coefficient corresponding to the largest Lagrangian multiplier to be nonzero [BHH+00]. Another method is to start also from a null model and then add an extra path corresponding to the lowest minimized ML (maximum likelihood) discrepancy function selected among all possible paths. This scheme is referred to as tree growth as the model grows by a single entry in $A$ at a time [CGS+11].
The most optimal but far from a feasible approach is to perform a simple brute-force method (or known as forest growth) that searches through all possible patterns of zero structures in A with a fixed number of paths and chooses the model corresponding to the lowest minimized ML [CGS11]. It is known that the number of all possible models grows exponentially to the number of variables (n), so it is not feasible as the problem dimension increases.

In this section, we propose a convex formulation for exploratory SEM problem by applying a widely-used sparse optimization with $\ell_1$-norm. The effectiveness of this approach has been well-understood and found applications in many fields including system identification [Van12], statistical learning [BPC10 §6] or control [LFJ13] since the $\ell_1$-norm penalty or lasso has been introduced [HTF09 §3]. The convex formulation we propose is

$$\begin{align*}
\text{minimize} \quad & -\log \det X + \text{tr}(SX) + 2\gamma \sum_{(i,j) \notin I_A} |A_{ij}| \\
\text{subject to} \quad & \begin{bmatrix} X & (I - A)^T \\ I - A & \Psi \end{bmatrix} \succeq 0, \\
& 0 \preceq \Psi \preceq \alpha I, \quad P(A) = 0,
\end{align*} \tag{10}$$

with variables $X \in \mathbb{S}^n, A \in \mathbb{R}^{n \times n}$ and $\Psi \in \mathbb{S}^n$, and the dual of (10) is

$$\begin{align*}
\text{maximize} \quad & \log \det (S - Z_1) - 2\text{tr}(Z_2) - \alpha \text{tr}(Z_4) + n, \\
\text{subject to} \quad & \begin{bmatrix} Z_1 & Z_2^T \\ Z_2 & Z_4 \end{bmatrix} \succeq 0, \\
& |(Z_2)_{ij}| \leq \gamma, \quad \forall (i,j) \notin I_A,
\end{align*} \tag{11}$$

with variable $Z = \begin{bmatrix} Z_1 & Z_2^T \\ Z_2 & Z_4 \end{bmatrix} \in \mathbb{S}^{2n}$. Note that the derivation of dual is given in Appendix A.2. Let $h(A) = \sum_{(i,j) \notin I_A} |A_{ij}|$ in (10). We see that $h$ is a form of $\ell_1$-like penalty function (or regularization) as it resembles the 1-norm of a matrix except that only those entries not belonging to $I_A$ are penalized. Users get to hypothesize about the known location of zeros in $A$ which is encoded as the index set $I_A$. If the user has no prior knowledge about the zero locations in $A$ at all then at least we apply the constraint $P(A) = \text{diag}(A) = 0$ since there must be no path from one variable to itself. For $(i,j) \notin I_A$, then we are not uncertain if $A_{ij}$ would be zero or not, so we enforce the $\ell_1$ norm on these entries and let the regularization promote the sparsity of $A$ which is controlled by the regularization parameter $\gamma > 0$. We refer the problem (10) as sparse SEM. We also note that $h(A)$ was considered in estimation of Recticular Action Model (RAM) [JGM16] which is a more general model class than path analysis. However, we will show later in Section 5.2 that since our method was designed for solving path analysis problems, our method yields a better performance when comparing in this setting. Moreover, by exploiting theoretical results in convex analysis, many conclusions from our convex formulation can be explained analytically.

**Choice of penalty parameter**

We see that the sparsity of the optimal path coefficient $A$ can be controlled via $\gamma$, e.g., the larger $\gamma$, the sparser the matrix $A$ is. In the Appendix C.2 we prove that there exists a critical value of $\gamma$, denoted by $\gamma_{\text{max}}$ in the sense that if

$$\gamma \geq \gamma_{\text{max}} := (1/\alpha)\|P^c(\alpha I - S)\|_{\infty} \tag{12}$$

then the optimal solution of $A$ in (10) is the zero matrix. Moreover, the value of $\gamma_{\text{max}}$ can be calculated in advance and depends on $\alpha$ and $S$ (problem parameters). This means it is unnecessary to vary $\gamma$ arbitrarily in the problem, and we can use $\gamma_{\text{max}}$ as an upper bound of the range of $\gamma$ used for varying the sparsity patterns of $A$, or for controlling the sparseness of $A$.

Another property of (10) is that its cost objective is not differentiable at $A = 0$ due to the term $|A_{ij}|$. As a result, KKT conditions for non-smooth optimization problems are stated from the concept of subgradients and subgradient calculus. The non-smooth property also makes solving a high-dimensional problem nontrivial since standard gradient-based methods cannot be applied.
Problem assumptions

We will assume that $S > 0$. Otherwise, the sparse SEM problem (10) is unbounded below. To show this, let $f(X,A)$ be the cost function of (10), and assume that $S$ has the eigenvalue decomposition $S = U D U^T$. Then it follows that $\text{tr}(S X) = \text{tr}(U D U^T X) = \text{tr}(D U^T X U)$. Let $Y = U^T X U$ and since $\det X = \det Y$, we can write $f(X,A) = f(Y,A) = -\log \det Y + \text{tr}(D Y) + 2\gamma \sum_{(i,j) \notin I_A} |A_{ij}|$. To minimize $f(Y,A)$ with constraint in (10), it has at least a feasible point that yields an unbounded value in $f$ if $S \succeq 0$. For example, set $A = 0$ and then $\sum_{(i,j) \notin I_A} |A_{ij}| = 0$, but if $S \succeq 0$, then $a_{ii} = 0$ for some $i$, and one of the feasibility condition requires only $Y \succeq (1/\alpha) I$, so that we can choose $Y$ to be diagonal where $y_{ii}$ is chosen to be arbitrarily large. Hence, the term $\text{tr}(D Y) = 0$ but $-\log \det Y \to -\infty$. We comment that the assumption of the positiveness of $S$ might not be held in practice when the number of variables are much higher than the number of samples. In such case, if we replace $S$ by $\hat{S} = S + \epsilon I$ where $\epsilon > 0$, then the cost objective of (10) is

$$-\log \det X + \text{tr}(S X) + \epsilon \|X\|_* + 2\gamma h(A)$$

where $\|X\|_* = \sum_{k=1}^n \sigma_k(X)$ or the nuclear norm of $X$ (the sum of singular values of $X$), where we have used the fact that for $X > 0$, $\sigma(X) = \lambda(X)$. The nuclear norm of a matrix is well-known to be a convex approximation for the matrix rank. The new problem with the replacement of $S$ can then be interpreted as an SEM problem with a regularization term on $X$ that promotes $\text{rank}(X)$ to be small. However, $X$ cannot be low rank due to the implicit constraint in the log det function.

Solution Pathway

Since we can derive the critical regularization parameter, $\gamma_{\text{max}}$, such that for any $\gamma \geq \gamma_{\text{max}}$, the solution of $A_{ij}$ for $(i,j) \notin I_A$ reaches to zero, we plot the solution pathway by plotting the entries of $A_{ij}$ for $(i,j) \notin I_A$ against varied $\gamma$ as shown in Figure 1. This plot illustrates that as $\gamma$ increases some of $A_{ij}$ become zero and once an entry of $A$ becomes zero for a value of $\gamma$ then it stays zero as $\gamma$ increases. When $\gamma \geq \gamma_{\text{max}}$, all entries of $A_{ij}$ subject to zero constraints become zero. In short, we obtain a sparser path matrix as we increase the penalty parameter.

Figure 1: Entries of the path matrix solved for each $\gamma$. When $\gamma = \gamma_{\text{max}}$, our sparse SEM provides the sparsest solution, i.e., all entries of $A_{ij}$ for $(i,j) \notin I_A$ become zero. ©2016 IEEE. Reprinted, with permission, from the Proceedings of 2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE).

3.3 Low rank solution

Solutions of the convex confirmatory SEM and the sparse SEM are useful for the original SEM problem if $X = (I - A)^T \Psi^{-1} (I - A)$ at optimum as we can use $X$ as an estimate of $\Sigma^{-1}$. This occurs if and only if the rank of

$$W = \begin{bmatrix} X & (I - A)^T \\ (I - A) & \Psi \end{bmatrix}$$
is \( n \) at optimum. Therefore, we aim to find a relation between the parameter \( \alpha \) and the low rank optimal solutions of (7) and (10) from their the complementary slackness conditions. The result in section 3.1 gave us a hint that if \( \alpha \) is too large, then it is possible that \( \text{rank}(W) > n \) which is to be avoided. The following analysis was presented in our prior work [PS16] and we prefer to state again here for completeness. We show that if \( \alpha < \lambda_{\min}(S) \), it is often the case that \( W \) is low rank.

According to the complementary slackness conditions in sections A.1 and A.2 and from a property of trace: \( \text{tr}(AB) = 0 \iff AB = 0 \) for \( A, B \succeq 0 \), we have

\[
\begin{bmatrix}
Z_1 & Z_2^T \\
Z_2 & Z_4
\end{bmatrix}
\begin{bmatrix}
X \\
I - A
\end{bmatrix}
\begin{bmatrix}
(I - A)^T \\
\Psi
\end{bmatrix}
= 0. \tag{13}
\]

Since the columns of \( W \) are in the nullspace of \( Z \), we must have \( \text{rank}(W) = \text{nullity}(Z) \) and that \( \text{rank}(Z) = 2n - \text{rank}(W) \). In addition, the (1, 1) block of \( W \) must have rank greater than \( n \) because \( X \succ 0 \). The rank of \( W \) must satisfy \( n \leq \text{rank}(W) \leq 2n \) and therefore \( 0 \leq \text{rank}(Z) \leq n \). We obtain a low rank solution when the optimal primal and dual solutions of (7) and (10) satisfy

\[
X = (I - A)^T \Psi^{-1} (I - A) \quad \text{or equivalently} \quad \text{rank}(Z) = n.
\]

Furthermore, when this holds, \( \text{rank}(Z_4) = n \) and from the complementary slackness conditions in Appendix A it gives \( \Psi = \alpha I \), i.e., the estimated covariance error becomes a diagonal matrix. From Section 3.1, we have shown that if \( \alpha \) is smaller than \( \alpha_c = n / \text{tr}(S^{-1}) \), then the optimal dual solution is not zero. This suggests us to consider three ranges of \( \alpha \) where the rank of \( Z \) varies as shown in Figure 2. The value of \( \alpha_c \) lies somewhere in the interval that results in \( 0 < \text{rank}(Z) < n \).

![Figure 2: The effect of α on the solutions of the convex confirmatory SEM. Left. rank(Z) as α varies. Right. Error between X and the low rank approximation as α varies. Lines with the same color correspond to the result from using the same n and shown as results of many trials. ©2016 IEEE. Reprinted, with permission, from the Proceedings of 2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE).](image-url)

We have shown in Section 3.1 that the minimum eigenvalue of \( S \) always lies on the left to the harmonic mean of the eigenvalues of \( S \). If we choose \( \alpha = \lambda_{\min}(S) \), then it is often the case that \( \text{rank}(Z) = n \). This advises us to include a constraint \( \Psi \succeq \lambda_{\min}(S)I \succeq S \) into the estimation problem where we justify that the covariance error is controlled to be less than the covariance of the variables. We can consider \( n = 5, 10, 20 \) and vary \( \alpha \in [0.5 \lambda_{\min}(S), 5 \lambda_{\min}(S)] \) and solve (7) with each of \( \alpha \). Figure 2 illustrates that the error between \( X \) and \( (I - A)^T \Psi^{-1} (I - A) \) increases as \( \alpha \) increases and is zero when \( \alpha \) is sufficiently small relatively to the minimum eigenvalue of \( S \). It is also noted that when \( \alpha \) is too large, then \( \text{rank}(W) > n \) and there could be many optimal solutions \( \Psi \) that is strictly less than \( \alpha I \). In this case, the solution \( X \) is not unique.
3.4 Exploratory SEM

As we see in Section 3.2 that controlling the regularization parameter in the sparse SEM problem can provide path matrix solutions with various sparsity patterns. If $\gamma$ is large then the path matrix $A$ contains many zeros, resulting in a simple interpretation of the estimated causal structure but the goodness of fit becomes smaller. Therefore, choosing an appropriate value of $\gamma$ is a trade-off between choosing the solution to explain a causal structure in a simple way and to best describe data in a certain level. We adopt a criterion for selecting a suitable choice of $\gamma$ from a statistical approach. For instance, Akaike Information Criterion (AIC) [Aka87, Aka11], Akaike’s Final Prediction-Error Criterion (FPE) [Aka87] or Bayesian Information Criterion (BIC) [HTF09]. BIC is known to prefer a simpler model since the penalty on the model complexity is higher relatively to other criterions. Moreover, it can be shown that BIC chooses the correct model with probability reaching to one when the number of sample sizes grows to infinity. AIC is the first model selection criterion which has been widely accepted and is known to perform poorly when the number of sample sizes is small compared with the number of effective parameters. To solve this problem, the corrected AIC or AICc was developed to improve the performance of AIC [HT93]. Kullback Information Criterion (KIC) is a recent model selection criterion based on Kullback’s symmetric divergence [Cav99]. Similarly to AICc, the corrected KIC (KICc) [Seg06] was developed to reduce bias and improve model selection for a small-sample setting. For these reasons, we compare the performance of each model selection criterion, i.e., BIC, AIC, AICc, KIC and KICc for SEM which are given by

- $\text{BIC} = -2\mathcal{L} + d \log N$, 
- $\text{AIC} = -2\mathcal{L} + 2d$, 
- $\text{AICc} = -2\mathcal{L} + \frac{2dN}{N - d - 1}$, 
- $\text{KIC} = -2\mathcal{L} + 3d$, 
- $\text{KICc} = -2\mathcal{L} + \frac{(d + 1)(3N - d - 2)}{N - d - 2} + \frac{d}{N - d}$,

where $d$ is the number of effective parameters of the model, $N$ is the number of samples and

$$\mathcal{L} = \frac{N}{2} \left( -\log \det \hat{\Sigma} - \text{tr}(S\hat{\Sigma}^{-1}) \right)$$

is the log-likelihood function of samples $Y_1, Y_2, \ldots, Y_N$. We note that this is a fair comparison between each criterion since each of them consists of two terms: the goodness of fit and the complexity.

To learn the best causal structure of path matrices, we can choose a range of $\gamma$ and then solve sparse SEM (10) for each of those values, resulting in the estimated path matrices having different sparsity patterns ranging from densest to sparsest. Each of the estimated sparsity patterns is then used as a sparsity constraint on $A$ in the convex confirmatory SEM (7) and we solve for the optimal path matrix $A$, yielding a candidate model. We repeat this process using all the values of $\gamma$ and obtain a set of candidate models; each of which is labelled by a model selection criterion score and the best model is the one with the minimum score. In short, we use the sparse SEM to select a finite number of sparsity patterns in $A$ and use the convex confirmatory SEM to provide the best estimate of the path matrix corresponding to the sparsity pattern selected from the model selection criterion score. This procedure is illustrated in Figure 3.

4 ADMM algorithms

In this section, we present efficient numerical methods for solving our two estimation formulations, the convex confirmatory SEM and the sparse SEM. The ADMM method is a type of proximal gradient methods [PB14] that applies a splitting technique on the cost objective of convex problems and introduces some auxiliary variables. Mostly, a key success of ADMM is to add indicator functions corresponding to nonlinear constraints where the update steps, requiring computing proximal operator, can often turn into an almost analytical form. In the following, most proximal operators involved in our problem is the projection operator of $v$ on a set $C$, defined by $\Pi_C(v) = \arg\min_{x \in C} \|x - v\|_2$. We follow the ADMM standard format and implementation guidelines in [BPC+10].
4.1 ADMM for the convex confirmatory SEM

By the changes of variables: $X_1 = X, X_2 = I - A, X_4 = \Psi$, the problem (7) is equivalent to

$$
\begin{align*}
\text{minimize} & \quad -\log \det X_1 + \text{tr}(SX_1) \\
\text{subject to} & \quad X = \begin{bmatrix} X_1 & X_2^T \\ X_2 & X_4 \end{bmatrix} \succeq 0, \\
& \quad 0 \preceq X_4 \preceq \alpha I, \quad P(X_2) = I,
\end{align*}
$$

with variable $X \in S^{2n}$ where $X_1, X_4 \in S^n$ and $X_2 \in \mathbb{R}^{n \times n}$. To rearrange (15) into ADMM format, we define functions $f : S^{2n} \to \mathbb{R}$, $g_1 : S^{2n} \to \mathbb{R}$ and $g_2 : \mathbb{R}^{2n \times 2n} \to \mathbb{R}$ where

$$
f(X) = -\log \det(X_1) + \text{tr}(SX_1), \quad g_1(U) = \begin{cases} 0, & U \succeq 0 \\
\infty, & \text{otherwise} \end{cases}, \quad g_2(V) = \begin{cases} 0, & P(V_2) = I, 0 \preceq V_4 \preceq \alpha I, \\
\infty, & \text{otherwise} \end{cases}.
$$

The ADMM format of the problem (15) is

$$
\text{minimize} \quad f(X) + g_1(U) + g_2(V) \quad \text{subject to} \quad X - U = 0, \quad X - V = 0,
$$

with variables $X, U$ and $V \in S^{2n}$. The ADMM algorithm starts with forming the augmented Lagrangian defined by

$$
L_\rho(X, U, V, Y_1, Y_2) = f(X) + g_1(U) + g_2(V) \\
+ \text{tr}(Y_1^T(X - U)) + \text{tr}(Y_2^T(X - V)) + \frac{\rho}{2} \|X - U\|_F^2 + \frac{\rho}{2} \|X - V\|_F^2,
$$
where $\rho$ is the ADMM penalty parameter whose value relates to the speed of convergence and enforces the equality constraint. Let us denote $X$ and $X^+$ the variables in current and next iteration. The update rule of ADMM is to minimize $L_\rho(X, U, V, Y, Z)$ with respect to $X, U, V$ alternately.

$$X^+ = \arg\min_X f(X) + \text{tr}(Y^T(X - U)) + \text{tr}(Y^T(X - V)) + \frac{\rho}{2} \|X - U\|^2_F + \frac{\rho}{2} \|X - V\|^2_F,$$

$$U^+ = \arg\min_U g_1(U) + \text{tr}(Y^T(X - U)) + \frac{\rho}{2} \|X - U\|^2_F,$$

$$V^+ = \arg\min_V g_2(V) + \text{tr}(Y^T(X - V)) + \frac{\rho}{2} \|X - V\|^2_F,$$

$$Y_i^{+} = Y_1 + \rho(X^+ - U^+), \quad Y_2^+ = Y_2 + \rho(X^+ - V^+)$$

where we will show that the update rules for $X, U$ and $V$ can be expressed into a closed form, so that we can compute these updates efficiently.

**X-update.** A general problem form of solving $X$ update is

$$\min_X -\log\det(X_1) + \text{tr}(S X_1) + \rho \|X - M\|^2_F,$$

where $M = \frac{1}{2}(U + V) - \frac{1}{2\rho}(Y_1 + Y_2) \in S^{2n}$. This is an unconstrained problem (with an implicit constraint), so the zero gradient condition is

$$\nabla_X L_\rho(X, U, V, Y_1, Y_2) = \begin{bmatrix} -X_1^{-1} + S & 0 \\ 0 & 0 \end{bmatrix} + 2\rho (X - M) = 0,$$

(16)

with an implicit constraint from the domain of $f$ that $X_1 > 0$. Suppose $M = \begin{bmatrix} M_1 & M_2^T \\ M_2 & M_4 \end{bmatrix}$. We can apply the method based on an eigenvalue decomposition from [BPC+10][6.5] to show that the zero gradient condition on the $(1, 1)$ block:

$$2\rho X_1 - X_1^{-1} = 2\rho M_1 - S,$$

(17)

can be achieved with a positive definite $X_1$. The detail starts with taking an eigenvalue decomposition on the RHS of (17), providing

$$2\rho X_1 - X_1^{-1} = Q \Lambda Q^T,$$

(18)

where $Q$ is an eigenvector matrix of $2\rho M_1 - S$ and the eigenvalues, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. We multiply $Q^T$ on the left and $Q$ on the right of (18) and use the orthogonality of $Q$, and (18) becomes

$$2\rho \tilde{X}_1 - \tilde{X}_1^{-1} = \Lambda,$$

(19)

where $\tilde{X}_1 = Q^T X_1 Q$. The above equation leads us to find the positive number of $(\tilde{X}_1)_{ii}$ satisfying $2\rho (\tilde{X}_1)_{ii} - (\tilde{X}_1^{-1})_{ii} = \lambda_i$. The positive root of quadratic polynomial of diagonal entries of (19) is

$$(\tilde{X}_1)_{ii} = \frac{\lambda_i + \sqrt{\lambda_i^2 + 8\rho}}{4\rho}$$

and therefore $X_1 = Q \tilde{X}_1 Q^T$ is optimal. Other blocks of $X$ are simply $X_2 = M_2$ and $X_4 = M_4$. In conclusion, the solution of the $X$-update is given by

$$X^+ = \begin{bmatrix} Q \tilde{X}_1 Q^T & M_2^T \\ M_2 & M_4 \end{bmatrix}.$$
**U-update.** In this step, we solve the problem
\[
\underset{U \geq 0}{\text{minimize}} \quad \|U - M\|_F^2, \quad M = X + \frac{1}{p}Y_1.
\]
This is a projection problem onto the positive definite cone which is known to be given by \(\Pi_{C}(A) = \sum_{i}(\lambda_i)^{+}u_iu_i^T\) where \(\sum_{i}\lambda_iu_iu_i^T\) is the eigenvalue decomposition of \(A\), i.e., we discard all the modes corresponding to the negative eigenvalues of \(A\). Hence
\[
U^+ = \Pi_{C}(M) \text{ where } C = S_{+}^{2n}.
\]

**V-update.** A general form of the problem in this step is
\[
\underset{V}{\text{minimize}} \quad \|V - M\|_F^2, \quad \text{subject to } P(V_2) = I, \quad 0 \preceq V_4 \preceq \alpha I,
\]
where \(M = (X + \frac{1}{p}Y_2)\). From the two constraints in the problem, we can write a feasible \(V\) as
\[
V = \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} P(V_2) + P^c(V_2) \\ V_4 \end{bmatrix} = \begin{bmatrix} V_1 & I + P^c(V_2) \\ V_4 & \alpha I \end{bmatrix}.
\]
Then, the cost function
\[
\|V - M\|_F^2 = \|V_1 - M_1\|_F^2 + 2\|I + P^c(V_2) - M_2\|_F^2 + \|V_4 - M_4\|_F^2,
\]
is minimized by the optimal \(V\) given by
\[
V = \begin{bmatrix} M_1 \\ P^c(M_2) + I \end{bmatrix} \Pi_{C}(M_4).
\]
where \(\Pi_{C}(A) = \sum_{i}\min(\max(\lambda_i, 0), \alpha)u_iu_i^T\) is the projection of \(A\) onto \(C = \{X \in S^n \mid 0 \preceq X \preceq \alpha I\}\), i.e., we project each eigenvalue of \(A\) onto the interval \(0 \leq \lambda_i \leq \alpha\) and recompose \(A\) with the corresponding eigenvectors.

We summarize the pseudo codes of this algorithm in Appendix B.1.

### 4.2 ADMM for sparse SEM

As we mentioned that the cost objective of the sparse SEM is not differentiable at \(A = 0\) due to the term \(|A_{ij}|\). Due to this non-smooth property, standard gradient-based methods cannot be applied, so we apply ADMM to solve the problem. By the change of variables:
\[
X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}, \quad X_4 = \Psi, \quad X_2 = I - Z, \quad Z = A,
\]
the problem becomes
\[
\begin{aligned}
&\underset{X,Z}{\text{minimize}} & & -\log \det X_1 + \text{tr}(SX_1) + 2\gamma \sum_{(i,j) \notin I_A} |Z_{ij}| \\
&\text{subject to} & & 0 \preceq X_4 \preceq \alpha I, \quad X_2 = I - Z, \quad P(Z) = 0,
\end{aligned}
\] (20)
with variables \(X \in S^{2n}\) and \(Z \in R^{n \times n}\). Let us define functions \(f : S^{2n} \to R, g_1 : R^{n \times n} \to R, g_2 : S^{2n} \to R\) and \(g_3 : R^{2n \times 2n} \to R\) given by
\[
\begin{align*}
f(X) &= -\log \det(X_1) + \text{tr}(SX_1), \quad g_1(Z) = \sum_{(i,j) \notin I_A} |Z_{ij}| \\
g_2(U) &= \begin{cases} 0, & U \geq 0 \\
\infty, & \text{otherwise} \end{cases}, \quad g_3(V) = \begin{cases} 0, & 0 \preceq V_4 \preceq \alpha I \\
\infty, & \text{otherwise} \end{cases}.
\end{align*}
\]
Then the problem (20) can be arranged into ADMM format as

\[
\begin{align*}
\text{minimize} \quad & f(X) + 2\gamma g_1(Z) + g_2(U) + g_3(V) \\
\text{subject to} \quad & X_2 = I - Z, \quad F(Z) = 0, \\
& X - U = 0, \quad X - V = 0,
\end{align*}
\]

with variables \(X, U, V \in \mathbb{S}^{2n}\) and \(Z \in \mathbb{R}^{n \times n}\). The augmented Lagrangian of this problem is

\[
L_\rho(X, Z, U, V, Y_1, Y_2, Y_3) = f(X) + 2\gamma g_1(Z) + g_2(U) + g_3(V) + \text{tr}(Y_1^T(X_2 + Z - I)) + \text{tr}(Y_2^T(X - U)) + \text{tr}(Y_3^T(X - V)) + \frac{\rho}{2}\|X_2 + Z - I\|^2_F + \frac{\rho}{2}\|X - U\|^2_F + \frac{\rho}{2}\|X - V\|^2_F.
\]

Similarly, we denote \(X^+\) and \(X\) the variables in current and next iteration, respectively. The update rule of ADMM is given by

\[
X^+ = \arg\min_X \left( f(X) + \text{tr}(Y_1^T(X_2 + Z - I)) + \text{tr}(Y_2^T(X - U)) + \text{tr}(Y_3^T(X - V)) + \frac{\rho}{2}\|X_2 + Z - I\|^2_F + \frac{\rho}{2}\|X - U\|^2_F + \frac{\rho}{2}\|X - V\|^2_F, \right)
\]

\[
Z^+ = \arg\min_Z \left( 2\gamma g_1(Z) + \text{tr}(Y_1^T(X_2 + Z - I)) + \frac{\rho}{2}\|X_2 + Z - I\|^2_F, \right)
\]

\[
U^+ = \arg\min_U \left( g_2(U) + \text{tr}(Y_2^T(X - U)) + \frac{\rho}{2}\|X - U\|^2_F, \right)
\]

\[
V^+ = \arg\min_V \left( g_3(V) + \text{tr}(Y_3^T(X - V)) + \frac{\rho}{2}\|X - V\|^2_F, \right)
\]

\[
Y_1^+ = Y_1 + \rho(X^+ + Z^+ - I), \quad Y_2^+ = Y_2 + \rho(X^+ - U^+), \quad Y_3^+ = Y_3 + \rho(X^+ - V^+),
\]

where we can show that the \(X, Z, U\) and \(V\) updates can be derived into a closed form as follows.

**X-update.** This step involves solving the optimization problem

\[
\begin{align*}
\text{minimize} \quad & -\log(\det(X_1)) + \text{tr}(SX_1) + \frac{\rho}{2}\|X_2 - H\|^2_F + \rho\|X - M\|^2_F,
\end{align*}
\]

where \(H = I - Z - \frac{1}{\rho}Y_1 \in \mathbb{R}^{n \times n}\) and \(M = \frac{1}{2}(U + V) - \frac{1}{2\rho}(Y_2 + Y_3) \in \mathbb{S}^{2n}\). Suppose \(M = \begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix}\).

The zero gradient condition of the above unconstrained problem is

\[
\nabla_X L_\rho = \begin{bmatrix} X_1^{-1} + S & 0 \\ 0 & 0 \end{bmatrix} + \rho \begin{bmatrix} 0 & (X_2 - H)^T \\ X_2 - H & 0 \end{bmatrix} + 2\rho \begin{bmatrix} X_1 - M_1 & 2(X_2 - M_2) \\ 2(X_2 - M_2)^T & 2(X_2 - M_2) \\ X_4 - M_4 \end{bmatrix} = 0, \quad (21)
\]

with an implicit constraint from domain of \(f\) that \(X_1 \succ 0\). It is clear that we can choose \(X_1, X_2\) and \(X_4\) independently to satisfy (21). To find \(X_1\) that satisfies (21) and \(X_1 \succ 0\), we can apply the same technique [BPC+10, §6.5] as presented in the X-update of ADMM for confirmatory SEM. To this end, we compute the eigenvalue decomposition: \(2\rho M_1 - S = QAQ^T\) where \(Q\) is the eigenvector matrix, and \(A\) is diagonal and contains the eigenvalues. Then define \(\bar{X}_1\) a diagonal matrix whose entries are \((\lambda_i + \sqrt{\lambda_i^2 + 8\rho})/4\rho\) and compute \(X_1 = Q\bar{X}_1Q^T\) which is guaranteed to be positive definite. The X-update is therefore given by

\[
X^+ = \begin{bmatrix} \frac{1}{2}(Q\bar{X}_1Q^T) & \frac{1}{2}(H + 4M_2) \\ \frac{1}{2}(H + 4M_2)^T & M_4 \end{bmatrix}.
\]

**Z-update.** In this step, we are required to solve the problem of the form:

\[
\begin{align*}
\text{minimize} \quad & 2\gamma \sum_{i,j \notin I_A} |Z_{ij}| + \frac{\rho}{2}\|Z - M\|^2_F, \quad \text{subject to} \quad P(Z) = 0
\end{align*}
\]
where $M = (I - X_2 - \frac{1}{\rho}Y_1)$. The problem is separable in $Z_{ij}$ where the $(i, j)$ term that is not constrained by $P(Z) = 0$ can be solved from a simple problem: minimize $|z| + (\rho/2)|z - m|^2$, typically known as finding a proximal operator of $f(x) = |x|$. The solution of minimizing the above problem can be performed by element-wise soft thresholding, defined by

$$Z_{ij} = S_{2\gamma/\rho}(M_{ij})$$

where $S_k(a)$ is called soft thresholding operator, defined by

$$S_k(a) = \begin{cases} 
  a - k, & a > k, \\
  0, & |a| \leq k, \\
  a + k, & a < -k,
\end{cases}$$

or equivalently $S_k(a) = \max(|a| - k)\text{sign}(a)$. In conclusion, the $Z$-update is given by

$$Z_{ij}^+ = S_{2\gamma/\rho}(M_{ij}) \text{ for } (i, j) \notin I_A, \quad \text{and } Z_{ij}^+ = 0 \text{ for } (i, j) \in I_A.$$

**U-update.** The problem in this step is of the form:

$$\min_{U \succeq 0} \frac{\rho}{2} \|U - M\|_F^2, \quad M = X + \frac{1}{\rho}Y_2.$$  

The solution is therefore the projection onto the positive definite cone: $U^+ = \Pi_C(M)$ where $C = S_{2n}^+. \hfill \square$

**V-update.** The optimization in this step is of the form:

$$\min_V \frac{\rho}{2} \|V - M\|_F^2, \text{ subject to } 0 \preceq V_4 \preceq \alpha I, \text{ where } M = X + \frac{1}{\rho}Y_3$$

We can choose $V_1, V_2, V_4$ in $V = \begin{bmatrix} V_1 & V_2^T \\ V_2 & V_4 \end{bmatrix}$ independently where obviously, $V_1 = M_1, V_2 = M_2$ are optimal solutions. For $V_4$, it is to find a projection onto a positive definite cone with an upper bound (similarly to the $V$-update in ADMM for confirmatory SEM). In conclusion,

$$V^+ = \begin{bmatrix} M_1 & M_2^T \\ M_2 & \Pi_C(M_4) \end{bmatrix}$$

where $C = \{X \in S^n \mid 0 \preceq X \preceq \alpha I\}$.

We summarize the pseudo codes of this algorithm in Appendix B.2.

### 4.3 Choice of ADMM parameter

The choice of ADMM parameter, $\rho$, affects the convergence speed and an optimal selection is still an open question. Recent progresses also include an adaptive formula of $\rho$ varying upon the primal and dual residuals [BPC+10] or the rule motivated by the Barzilai-Borwein spectral method [XFG17]. However, this paper do not follow this direction but we have experimented and found a good practical choice of $\rho$ that leads to a convergence by a reasonable number of iterations. When considering [10], we found that good choices of $\rho$ vary by the problem scale, determined by $(S, \alpha)$. When the minimum eigenvalue of $S$ is very small ($\leq 10^{-3}$), ADMM seems to converge slowly for $\rho = 10, 100$. We show that if we scale the problem [10] properly by $\beta$, the choice of $\rho = 10\beta$ works well in practice. Hence, we need a result that solutions from scaled and original problems can be interchangeably obtained.

**Proposition 2.** Let $(X, Z)$ be primal and dual optimal solutions of [10] and [11], respectively using problem parameter $(S, \alpha)$ where

$$X = \begin{bmatrix} X_1 & X_2^T \\ X_2 & X_4 \end{bmatrix}, \quad X_2 = I - A, \quad X_4 = \Psi.$$
Moreover, let \((\tilde{X}, \tilde{Z})\) the primal and dual optimal solutions of \((10)\) and \((11)\), respectively using problem parameter \((\tilde{S}, \tilde{\alpha})\) where \(\tilde{S} = \beta S\) and \(\tilde{\alpha} = \beta \alpha\). Then, \((X, Z)\) and \((\tilde{X}, \tilde{Z})\) are related by

\[
\tilde{X} = \begin{bmatrix} X_1/\beta & X_2^T \\ X_2 & \beta X_4 \end{bmatrix}, \quad \tilde{Z} = \begin{bmatrix} \beta Z_1 & Z_2^T \\ Z_2 & Z_4/\beta \end{bmatrix}.
\]

(22)

In other words, we can obtain the optimal solution of the scaled sparse SEM instantly from the solution of unscaled problem, and vice versa.

The proof of this result is described in Appendix C.3. To apply Proposition 2, we scale the sparse SEM by \(\beta = 1/\lambda_{\min}(S)\), so that the minimum eigenvalue of \(\tilde{S} = \beta S\) is one and \(\tilde{\alpha} = 1\). As a result, we solve the scaled sparse SEM using \((\tilde{S}, \tilde{\alpha})\) with a heuristic choice of \(\rho = 10\beta\) by ADMM algorithm and retrieve the solution to the original problem using (22).

4.4 Algorithm performance

To see the performance of each algorithm in solving our both convex confirmatory SEM and sparse SEM, we generate data with \(n = 50, 100, \ldots, 1000\), using 50 samples of \(S\) for each \(n\) and solve the problems (7) and (10) using ADMM algorithms. The computer’s specification used in this experiment is: CPU : Intel Core i5-6400 (2.7 GHz), RAM : 16GB DDR4 BUS2133, HDD : SATA III 7200 RPM (1TBs), OS : WINDOWS10-64bit Education. Solving either convex confirmatory SEM or sparse SEM with dimension \(n\) involves total number of variables in \(X\), \((n(n+1)/2\) plus number variable in \(\Psi\) and the number of of paths in \(A\).

5 Results on generated data

This section we describe the performance of exploratory SEM presented in Section 3.4 with the scheme in Figure 3. Our goal is to examine how well we can estimate nonzero and zero entries in the path matrix from the data generated from a true model encoded with a pattern of the true path matrix. Throughout this section, we choose \(A_{true}\) having random sparsity patterns and generate measurements from \(Y = (I - A_{true})^{-1}e\) where \(e\) is normally distributed with variance of 0.1. The matrix \(S\) is then computed as the sample covariance of \(Y\). Options of data generating process are
number of samples and density of non-zero elements in $A_{true}$. In the estimation process, as required by the constraint $P(A) = 0$, we make assumptions about zero locations in $A$, varied by patterns and as percentage of all zeros. To examine the performance, we apply typical measures: True positives (TP), True negative (TN), False positives (FP), False negatives (FN), TP rate (TPR), FP rate (FPR), where positives are non-zero entries in $A$ and negatives are zeros in $A$, through a receiver operating characteristic or ROC curve [Alp11][§19.7].

5.1 Performance of sparse SEM

In the first experiment, we observe the effect of i) percentage of known number of zeros in the path matrix and ii) sample size of measurements. The assumption of known zeros in $A$ relates to the degree of freedom (df). The number of known parameters is $n(n-1)/2$ and the number of estimated parameters is the sum of free entries in $A$ and the number of entries in $\Psi$. In the first case, we assume the location of zeros in $A_{true}$ is known in the amount of 0%, 20%, 50%, 65% and 80% of all zeros. In the second case, we set the number of known zeros of $A$ to 50% and vary the sample sizes as $N = 100, 1000, 5000$.

ROC curves are shown in Figure 5 where one line of this plot is obtained by varying $\gamma$ from 0 to $\gamma_{max}$ and the values of TPR and FPR are averaged from 25 runs. This result illustrates that if we do not have any assumptions on zero structure in $A_{true}$, our accuracy is about 50% (TP rate is approximately equal to FP rate), but if we have more assumptions on zero structure in $A_{true}$, the accuracy is improved. We note that if no known location of zeros is assumed (df is negative) then the problem [10] may not have a unique solution for a given $\gamma$, then the estimated zero structure of $A$ may not be the same as the true matrix. However, if we assume more location of zeros (df is zero or positive), then the problem [10] could have a unique solution, and there exists a value of $\gamma$ that yields a satisfactorily accurate zero structure in $\hat{A}$. Moreover, when the number of observations increases, the performance of our sparse SEM is also significantly improved.

Figure 5: ROC curves of sparse SEM as we vary the regularization parameter, $\gamma$. Left. Knowing more correct zero structure in $A_{true}$ provides the better accuracy of our learning causal structure method. Right. When the number of observation increases, the performance of our exploratory SEM also significantly increases.

In the second experiment, we explore four main aspects that could influence the estimation results. These factors are sparsity density of the true model ($A_{true}$), the number of sample sizes ($N$), the number of known zero locations used in the estimation, and the choice of model selection scores. The experiments are then designed to investigate the effects of these factors as explained below.
(a) **False Positive (FP) error.** When $A_{\text{true}}$ is dense, FP from all model selection criterions tends to decrease when we use more knowledge about zero location in $A_{\text{true}}$ into the estimation process, but it increases when $N$ grows high as all model selection criterions tend to select the denser $\hat{A}$. In the case of small $N$, AICc provides the minimum FP error.

(b) **False Negative (FN) error.** Using more knowledge about zero location in $A_{\text{true}}$ into the estimation process barely affects the change of FN, but it can be improved when $N$ grows.

(c) **Total error.** Main portion of the total error comes from FP so it tends to decrease when we have more assumption about true zero location in $A_{\text{true}}$.

Figure 6: Averaged FP, FN and total error from 50 runs of sample covariance matrix $S$, when $A_{\text{true}}$ is dense. AIC provides the minimum error when $N$ is small.
(a) **False Positive (FP) error.** FP from all model selection criterions tends to decrease when we use the knowledge about zero location in $A_{true}$ into the estimation process, but it increases when $N$ grows as all model selection criterions tend to select a denser $\hat{A}$. In the case of small $N$, BIC, AICc and KICc provide the better accuracy as the true model is sparse.

(b) **False Negative (FN) error.** Using more knowledge about zero location in $A_{true}$ into the estimation process barely affects the change of FN, but it can be improved when $N$ grows.

(c) **Total error.** Main portion of the total error comes from FP so it tends to decrease when we have more assumption about true zero location in $A_{true}$.

Figure 7: Averaged FP, FN and total error from 50 runs of sample covariance matrix $S$, when $A_{true}$ is sparse. BIC, AICc and KICc provide the lower total error when $N$ is small.
1. The sparsity density of $A_{\text{true}}$. In this experiment, we generate $A_{\text{true}}$ with two sparsity levels, 50% and 80% and observe a relation between the sparsity pattern of $A$ that minimizes BIC score and the error rate. If we compare FN from Figures 6 and 7 for moderate sample size, $N = 100$, we see that our method gives less FN when $A_{\text{true}}$ is sparse and less FP when $A_{\text{true}}$ is dense. Unavoidable errors as FP (when $A_{\text{true}}$ is sparse) and FN (when $A_{\text{true}}$ is dense) are commonly seen since these type of errors occur against the hypothesis of the true model. Moreover, when $A_{\text{true}}$ is dense, Figure 6 shows that using AIC leads to the minimum total error since this score is prone to use a dense model (which agrees with the assumption of the true model). Similarly, when $A_{\text{true}}$ is sparse, Figure 7 confirms that using the scores penalizing more on the model complexity such as BIC, AICc, KICc yields a lower total error.

2. The number of samples. In the experiments, we use $N = 100$ (moderate size) and $N = 100,000$ (large sample size) to examine the asymptotic properties of the estimates. When $N$ is large, Table 1 confirms that the selected $\gamma$ is closer to zero since the sparse SEM (as a regularized problem) should yield the solution closer to that of non-regularized problem. The selected $\gamma$ is also larger when the true model is sparse. Moreover, Figures 6 and 7 report that FP is not improved (since the solutions are denser), but FN obviously decreases when $N$ increases, showing that our regularized formulation is robust to false negative errors.

3. The percentage of known zero locations in the estimation. To examine this factor, the experiments are performed with the percentage of known zeros of 0%, 20% and 50%. The first two values correspond to the problem with negative df where the sparse SEM solution could be not unique implying that the estimated zero pattern may not be as accurate as when knowing more zero locations. This, in principle, should affects FP, so Figures 6 and 7 also supports this hypothesis, that is, when we know more about the true zero locations in $A_{\text{true}}$, FP and overall total error decrease, but FN seems to be indifferent.

4. The choice of model selection scores. We consider AIC (tend to choose dense models), AICc (adjusted for finite sample size), BIC, KIC and KICc scores (tend to choose simpler models). Figures 6 and 7 have verified us that AIC tends to yield the minimum total error when $A_{\text{true}}$ is dense, and conversely, the choices of BIC, AICc and KICc tends to provide the total error lower than other criterions when $A_{\text{true}}$ is sparse.

5.2 Comparison with existing method

The package in R was developed to solve structural equation modeling with regularization term including both ridge and the least absolute shrinkage and selection operator (lasso), proposed by [16]. Regsem uses Recticular Action Model (RAM) notation to derive an implied covariance matrix. The parameters of general SEM will be translated into three matrices: the filter matrix ($F$), the direct path matrix ($A$), the covariance matrix of variables ($\Psi$). In details, regsem can be used to solved an optimization problem:

\[
\minimize \quad - \log \det X + \text{tr}(SX) - \log \det S - n + \gamma \sum_{(i,j) \notin I_A} |A_{ij}|
\]

subject to

\[
X = F^{-T}(I - A)^T\Psi^{-1}(I - A)F^{-1},
\]

\[
P(A) = 0,
\]

with variables $\Sigma$, $A$ and $\Psi$. It generalizes (i) to include i) $\ell_1$-regularization penalty on $A$ and ii) the filter matrix, $F$, since RAM model also considers latent variables. However, our problem aims
to find the relationship among observed variable only. Therefore, to be able to make a comparison between regsem and our method, we explore a structure of a recursive path model which is the simplest model in RAM, described by a set of linear equations:

\[
\begin{align*}
w &= Hx + \epsilon_1, \\
x &= Jv + \epsilon_2, \\
v &= Kz + \epsilon_3,
\end{align*}
\]  

(24)

where \(w, x, v, z\) are observed vectors and \(H, J, K\) are coefficient matrices. These equations can be written in \(y = Ay + u\) as displayed in Figure 8 where \(A\) has a certain sparse structure.

In our estimation process, the constraint \(P(A) = 0\) is encoded according to the structure of \(A_{true}\) derived in Figure 8 that corresponds to the degree of freedom of 42. For Regsem, it is based on the lavaan package, which is a general SEM software program. This command requires at least four arguments, i.e., our model generated by lavaan library, type of estimation formulation (lasso or ridge regression), entries to be penalized and a penalty parameter. The argument pars_pen = \(c(1 : 24)\) in regsem function is to set the number of penalized entries in the path matrix \(A\); here, 24 entries, according to its structure in Figure 8. We also use this function to find a lower bound of regularization parameter by selecting the minimum value of \(\gamma\) that forces all entries in \(A\) to be zeros. For each data trial, we solve both sparse SEM and regularized SEM by varying 50 values of regularization parameter in the range of \([0, \gamma_{\text{max}}]\).
ROC curves averaged from 100 trials shown in Figure 9 illustrate that our sparse SEM can achieve more accuracy than Regsem in both moderate and high sample size settings. Though the Regsem problem (23) estimates the matrix variables in a more general model than our methods, when it comes to a special case that reduces to a path analysis problem that contains observed variables only, our method, customized to path analysis problem, should perform better. Regsem can be more advantageous when solving a more general SEM problem.

6 Real-world application

6.1 Air pollution and weather data

In this experiment, we explore a relation structure among eleven climate variables including greenhouse gas (SO$_2$, NO$_2$, O$_3$, CO), solar radiation, relative humidity, temperature, particulate matter (PM$_{10}$), pressure and wind speed, in Bangkok, the capital of Thailand, during Summer (15 February to 15 May, 2007-2009). The hourly data were measured from eight stations located in residential and financial sites in the central Bangkok. The data are standardized to have the zero mean and unit variance and are split into training and validation sets with the ratio of 2 : 1.

Figure 10 illustrates an example of climate time series in summer. The plots show that radiation, temperature and O$_3$ are significant during 7 a.m. to 3 p.m. of the day, so we select this interval of 8 hours to be the measurement, resulting in 8 time points of each variable in a day.

Figure 10: A sample plot of climate time series, measured at Nonsi station, Bangkok (the center of Thailand), from 1 a.m. to 12 p.m.

In the estimation process, if identifiability of the model is encouraged, a zero degree of freedom (6) should be obtained. Hence, in this application, it requires setting 55 entries in the estimated path matrix to be zero encoded in the constraint $P(A) = 0$. In order to obtain a reasonable constraints, we perform a partial correlation analysis on $y$ using partialcorr in MATLAB to find insignificant relations among variables via the zero pattern in the sample partial correlation matrix (using a statistical test with the significance level of 0.01) and impose the corresponding entries in $A$ to zero.

We perform our estimation process described in Section 3.4 and the path matrices from model selection scheme of each station are selected. The nonzero entries in $A$ and their magnitude can define a graphical model explaining relationships of variables and their strength of connections. These relation patterns are different from stations to stations; see Figure 11a, so we compare a common network from all stations using similarity scores ranging from 50% to 100% shown in Figure 11b. The similarity score is the percentage of the number of common links from all stations in relative to the number of all links. Dominant connections with similarity score of 100% are
(a) The structure of the optimal path matrix from each station when zero constraints of $A$ from a partial correlation analysis is applied.

(b) A common relationship pattern of variables from all stations with various similarity scores.

Figure 11: The structures of optimal path matrices ($A$) from each station and graphical model explaining relationships among the eleven climate variables, encoded from the optimal path matrix.

CO - NO$_2$, Radiation-Temperature, RH-Temperature

(RH denotes relative humidity.) Firstly, a strong relation between CO and NO$_2$ is supported by a known fact of a combustion reaction from car congestion in city areas. This appears in our finding as the red color between the two variables from all stations in Figure 11a. Secondly, the relation between RH and temperature is known to be inverse to each other, given that the moisture content in the air is constant. Relative humidity is defined as the ratio of the water vapor pressure to the equilibrium vapor pressure at a given temperature. As the temperature rises, the capacity of air to hold water increases, so if the actual amount of water in the air does not change, then the relative humidity falls down. Mathematical expressions explaining relationship between these two variables can be found in [Law05]. Our result finds an inverse relationship between RH and temperature as noted in the blue value from all stations in Figure 11a and are also consistent to the trend of time series plots in Figure 10. Thirdly, a positive dependency between temperature and radiation indicated by the red color in Figure 11a agree with natural characteristics of solar radiation; as the sun rises up (and hence more radiation), the air temperature is increased. Nevertheless, in the area of meteorology, the daily solar radiation is typically explained by an increasing function of the diurnal range of air temperature, $\Delta T$ (the difference between maximum and minimum temperature) [LML+09].

In addition to the above three relationships, the estimated structures with similarity score of 75% in Figure 11b include relations between

CO-PM$_{10}$, PM$_{10}$-O$_3$, Temperature-O$_3$, RH-O$_3$, PM$_{10}$-Temperature, NO$_2$-SO$_2$,

RH-Radiation, RH $\rightarrow$ Pressure, Radiation-Pressure, Temperature-Pressure.
The findings from [JL99] using a regression analysis explain that the wind speed and temperature have inverse dependence on radiation attenuation; RH and pollution have a direct influence on radiation attenuation; and that the reduction of solar radiation in the rainy season due to pollution wash-out effect is not significantly different from the dry season. The relations among temperature, relative humidity, and radiation are consistent with several researches; [JL99] where RH was found to have a direct influence on radiation. The connection between PM$_{10}$ and O$_3$ corresponds to the finding in [JCS+15] (but they considered PM$_{2.5}$ instead of PM$_{10}$.) It is known that vehicle emissions are the main source of CO and NO$_2$ and especially diesel vehicles can emit particular matter. Moreover, areas containing burning process such as fossil fuel combustion can be major sources of NO$_2$ and SO$_2$. PM in Bangkok can be produced in the area of high temperature such as from automobiles and biomass burning in residential areas [CNLK08]. These facts relate with our findings of connection between CO-PM$_{10}$, PM$_{10}$-Temperature and NO$_2$-SO$_2$ where the last two relationships are present altogether in many stations; see Figure 11a. As temperature increases, the air is expanding and its density in that area becomes less, resulting in a decreasing variation of air pressure. This agrees with the minus sign of the coefficient in the path matrix from temperature to pressure in Figure 11a. We then conclude that the dependence structure learned from our model mostly agree with known characteristics of air pollutants and meteorological variables and share the same findings from previous studies.

6.2 fMRI data

In this experiment, we aim to explore a common brain network between control and autism groups under resting-state condition learned from Autism Brain Imaging Data Exchange (ABIDE) data set [MYL+14, AS17]. This data set contains 1112 fMRI images, collected across more than 24 international brain imaging laboratories. We select 46 images from autism group and 40 images from control group, in total of 86 images provided by University of Michigan. Demographic information about characteristics, including age, gender, and handedness, of 86 subjects are summarized in Table 2. According to qualitative data, the value of 0 and 1 are assigned to male (M) and female (F), respectively, for gender; moreover, are assigned to left-handedness (L) and right-handedness (R), respectively, for handedness. Then p-values from two-sample t-test for equal mean with significance level of 0.05 show that both groups have significantly the same characteristics.

|                        | Control group | Autism group | p-value |
|------------------------|---------------|--------------|---------|
| Age (mean ± SD)        | 13.3826 ± 2.5979 | 13.1275 ± 1.9765 | 0.6139  |
| Handedness (R/L)       | 38/8          | 35/5         | 0.5332  |
| Gender (M/F)           | 33/13         | 34/6         | 0.1426  |

The functional preprocessing of our data has been done via Preprocessed Connectomes Project (PCP) [CPT11], using Configurable Pipeline for the Analysis of Connectomes (CPAC). It performs a structural preprocessing of skull-stripping using AFNIs 3dSkullStrip, three-issue type brain segmentation using FSLs FAST, skull-stripped brain normalization to MNI152 with linear and non-linear registrations using ANTs. For functional preprocessing of fMRI images, some of the first volumes were not dropped. Then it performs a slice timing correction and motion realignment, respectively. The image intensity was normalized by 4D global mean and a band-pass filtering in the range of 0.01-0.1 Hz is applied. All images from every subject are transformed from the original to MNI152 template. To reduce the dimension of our data, we extracted mean time-series for a set of regions of interest (ROIs) using Automated Anatomical Labeling (AAL) template [TMPLP+02] which was fractionated to functional resolution (3 x 3 x 3mm$^3$) via nearest-neighbor interpolation.

After preprocessing data, we have 90 time series from 90 ROIs with 249 time points, corresponding to $Y \in \mathbb{R}^{90 \times 249}$. The sample covariance matrix is computed from $Y$ and 25 log scale values of $\gamma$ are selected in $\{0, 10^{-4}, \ldots, \gamma_{\text{max}}\}$ where $\gamma_{\text{max}}$ is the $\gamma$ that penalizes all entries in $A$ to be zero. A prior assumption on the zero locations of $A$ is set as $P(A) = \text{diag}(A) = 0$. Path matrices,
\( A \in \mathbb{R}^{90 \times 90} \), estimated from our approach can be represented as a graphical model containing 90 nodes whose labels are shown in Table 3 in Appendix D. For each value of \( \gamma \), we can specify a common optimal path matrix by searching for some positions of nonzero entry that appear more than 90% from all subjects in each group. Most frequently appeared links are denoted as significant connections between brain regions. We compute a common optimal path matrix (among all subjects in one group) by discarding insignificant entries and averaging only significant entries over all subjects. We select three values of \( \gamma \) as \( \gamma = 0.0025 \gamma_{\text{max}}, \gamma = 0.0182 \gamma_{\text{max}} \) and \( \gamma = 0.135 \gamma_{\text{max}} \) to produce three common path matrices with different density levels. Figure 12 shows brain networks from the autism groups, constructed by BrainNet Viewer [XWH]. It illustrates that our method generally provides nested graph structures as \( \gamma \) varies, i.e., noticeable connections in the sparse structure also exist in the moderately sparse and dense structures.

![Brain networks](images/brain_networks.png)

Figure 12: A common brain network from the autism group with various graph density levels. The link widths vary with the magnitudes of entries in the estimated path matrix. The red and blue link represent the positive and negative magnitude, respectively.

![Brain maps](images/brain_maps.png)

Figure 13: Comparing the brain maps between autism and control groups. Black edges are common links that appear in both groups; red edges are ones found only in the autism group, and blue edges are links detected only in the control group.

To draw some conclusions on the differences of brain networks between two groups, we select the sparsest structures and compare the brain graphs in Figure 13. An interesting observation is the autism brain graph has less number of connections as we see many ROI relations disappear from the control group. The connections that appear only in the autism group are

\[
\text{SFGdor}(L) \leftrightarrow \text{SFGmed}(L), \quad \text{PoCG}(R) \leftrightarrow \text{PreCG}(R), \quad \text{PUT}(R) \leftrightarrow \text{INS}(R), \quad \text{IOG}(R) \leftrightarrow \text{IOG}(L), \\
\text{SPG}(L) \leftrightarrow \text{SPG}(R), \quad \text{SPG}(R) \leftrightarrow \text{PCUN}(L), \quad \text{PUT}(R) \leftrightarrow \text{PAL}(R), \quad \text{SOG}(R) \leftrightarrow \text{CUN}(R), \quad \text{MOG}(L) \leftrightarrow \\
\]
Moreover, connections only found in the control group are

\[
\text{MFG}(R) \leftarrow \text{MFG}(L), \text{ORBsupmed}(R) \leftarrow \text{ORBsupmed}(L), \text{DCG}(R) \leftarrow \text{ACG}(R), \text{PHG}(L) \leftarrow \text{PHG}(R), \text{HIP}(R) \leftarrow \text{PHG}(R), \text{PCUN}(R) \leftarrow \text{CUN}(R), \text{PCUN}(L) \leftarrow \text{SPG}(L), \text{IPL}(R) \leftarrow \text{SPG}(R), \text{CUN}(R) \leftarrow \text{PCUN}(R), \text{ROL}(R) \leftarrow \text{HES}(R), \text{PUT}(R) \leftarrow \text{PAL}(L), \text{STG}(R) \leftarrow \text{HES}(R), \text{TPOmid}(R) \leftarrow \text{TPOmid}(L), \text{ITG}(R) \leftarrow \text{ITG}(L), \text{ORBmid}(R) \leftarrow \text{ORBsup}(R), \text{LING}(L) \leftrightarrow \text{CAL}(L), \text{SOG}(R) \leftrightarrow \text{CUN}(L), \text{MOG}(R) \leftrightarrow \text{SOG}(L), \text{IOG}(R) \leftrightarrow \text{MOG}(R), \text{PoCG}(L) \leftrightarrow \text{PoCG}(R), \text{IPL}(L) \leftrightarrow \text{SPG}(L), \text{SMG}(L) \leftrightarrow \text{SMG}(R), \text{CAU}(L) \leftrightarrow \text{CAU}(R), \text{STG}(L) \leftrightarrow \text{HES}(L), \text{STG}(L) \leftrightarrow \text{STG}(R) \text{ and } \text{MTG}(L) \leftrightarrow \text{MTG}(R). \]

From the lists above, it is worth comparing our results with previous studies on some of the missing connections that are no longer present in the autism group, which are among temporal gyrus group (ITG(R) \leftarrow ITG(L), \text{STG}(L) \leftrightarrow \text{STG}(R) \text{ and } \text{MTG}(L) \leftrightarrow \text{MTG}(R)). The inferior temporal gyrus (ITG) contributions are a processing of visual stimuli in object recognition [KW12]. The superior temporal gyrus (STG) has been known to involve in the auditory process and developing language but recently, has been implicated as a key factor in social cognition [BMN+07, JMK+10] which is an important consideration of autism patients. The main function of middle temporal gyrus (MTG) is exactly unknown but it helps on some processes about recognition of known faces and accessing word meaning while reading [AH13]. These results are supported by the experiment from [CRZ+17] in which they showed the decreasing in functional connectivity among these areas in autism group. The experiments in [CRZ+17] was designed to study a disrupt change of state and strength of connectivities in autism group, referred to as abnormal connectivities, which are reported to be found in ITG and STG area as well. This is consistent with [CRZ+15]. They concluded that the middle temporal gyrus (MTG) is a region that implicates in face expression, gesture representation impairments and theory of mind impairments in autism. We also found that relations from the precuneus, the basal ganglia, the anterior cingulate cortex (PCUN(R) \leftrightarrow \text{CUN}(R), \text{CAU}(L) \leftrightarrow \text{CAU}(R) \text{ and } \text{DCG}(R) \leftarrow \text{ACG}(R)) are still missing from the brain network in the autism group. Caudate nucleus (CAU), one of the components in the basal ganglia, affects to many nonmotor functions such as procedural learning or inhibitory control of action [NHM09], while precuneus (PCUN) is involved in self-consciousness, such as reflective self-awareness [KNL02, LLC+04], and memory task, such as responding what people have seen based on what they have remembered [WRGB03]. These results agree with [SWQ17] who expressed brain networks as conditional independence graphs and showed that, in autism group, the edges linking to the precuneus, the basal ganglia, the anterior cingulate cortex and the medial frontal cortex, are mostly affected. Moreover [CKKJ06] found underconnectivity among ROIs linking to precuneus in autism group.

In conclusion, our result represents that, in the brain network, some circuits relating to cognitive process, such as social interaction, face and image recognition, learning process or working memory, are missing from the autism group but exist in the control group. Our findings are satisfied with many previous studies.

7 Conclusions

This paper have proposed two convex formulations for confirmatory and exploratory SEM which find applications of learning causality among variables based on path analysis models. Our formulations relax a quadratic matrix equality into inequality and show that the solution of our formulation can be a solution to the original problem under homoskedastic assumptions of noise in the model. The proposed scheme of exploratory SEM exploits the feature of sparse estimation introduced by adding an $\ell_1$ regularization to the estimation function. Causality structures encoded from sparsity patterns of the path matrix can be obtained by sweeping values of regularization parameters in a specific range, derived analytically. Applying efficient ADMM algorithms allows us to solve the problems in large-scale including hundreds of variables, while it is not common to see existing SEM softwares solve such cases. Numerical results have shown that i) the percentage of known zeros of $A$ is a
key factor for decreasing FP, regardless of the sparsity density in the true model, ii) FN can be improved mostly by increasing sample size, iii) the total error mainly comes from FP and iv) the choice of model selection criterions could depend on the sparsity density in the true model. That is, BIC, AICc, and KICc provide better accuracies if the true path matrix is sparse while AIC performs best when the true model is dense. A comparison between our approach with Regsem method has been set in a fair setting, while considering a special case of SEM model. Our method yields higher accuracies where this desirable results could benefit from either our formulation or the algorithms. Results on real data sets show that the findings of causality structures coincides with the previous studies of applications. Most relations between climate variables learned from our model can be supported from environmental facts and research findings from literature. While in the brain studies, a ground-truth causal network of brain activities are not completely known, our findings are comparable with the literature with some extents. The brain networks from the control and autism groups are different in some brain regions including temporal gyrus group, precuneus, and caudate nucleus areas.

8 Acknowledgement

We would like to thank Department of City Planning, Bangkok Metropolitan Administration, and Pollution Control Department, Bangkok, for providing relevant information in the air pollution experiment.

References

[AH13] D.J. Acheson and P. Hagoort. Stimulating the brain’s language network: syntactic ambiguity resolution after ms to the inferior frontal gyrus and middle temporal gyrus. *J Cogn Neurosci*, 25(10):1664–1677, 2013.

[Aka87] H. Akaike. Factor analysis and AIC. *Psychometrika*, 52(3):317–332, 1987.

[Aka11] H. Akaike. Akaike’s Information Criterion. In *International Encyclopedia of Statistical Science*, pages 25–25. Springer, 2011.

[Alp14] E. Alpaydin. *Introduction to Machine Learning*. MIT press, 2014.

[AS17] D.M. Adriana and M. Stewart. Autism Brain Imaging Data Exchange (ABIDE). [http://fcon_1000.projects.nitrc.org/indi/abide/](http://fcon_1000.projects.nitrc.org/indi/abide/) 2017.

[BF97] C. Büchel and K.J. Friston. Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral cortex*, 7(8):768–778, 1997.

[BHH*00] E. Bullmore, B. Horwitz, G. Honey, M. Brammer, S. Williams, and T. Sharma. How good is good enough in path analysis of fMRI data? *NeuroImage*, 11(4):289–301, 2000.

[BMN+07] E.D. Bigler, S. Mortensen, E.S. Neeley, S. Ozonoff, L. Krasny, M. Johnson, J. Lu, S.L. Provencal, W. McMahon, and J.E. Lainhart. Superior temporal gyrus, language function, and autism, developmental neuropsychology. *Developmental Neuropsychology*, 31(2):217–238, 2007.

[Bol89] K.A. Bollen. *Structural equations with latent variables*. John Wiley & Sons, 1989.

[BPC+10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

[BS09] E. Bullmore and O. Sporns. Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.
[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. Available: [www.stanford.edu/~boyd/cvxbook](http://www.stanford.edu/~boyd/cvxbook).

[Cav99] J.E. Cavanaugh. A large-sample model selection criterion based on Kullback’s symmetric divergence. *Statistics & Probability Letters*, 42(4):333–343, 1999.

[CGS+11] G. Chen, D.R. Glen, Z.S. Saad, J.P. Hamilton, M.E. Thomason, I.H. Gotlib, and R.W. Cox. Vector autoregression, structural equation modeling, and their synthesis in neuroimaging data analysis. *Computers in biology and medicine*, 41:1142–1155, 2011.

[CKKJ06] V.L. Cherkassky, K.R. Kana, T.A. Keller, and M.A. Just. Functional connectivity in a baseline resting-state network in autism. *Neuroreport*, 17(16):1687–1690, 2006.

[CNLK08] N. Chuersuwan, S. Nimrat, S. Lekphet, and T. Kerdkumrai. Levels and major sources of PM$_{2.5}$ and PM$_{10}$ in Bangkok Metropolitan Region. *Environment international*, 34(5):671–677, 2008.

[CP11] C.R. Cameron and B. Pierre. Preprocessed Connectomes Project (PCP). [http://preprocessed-connectomes-project.org/abide/](http://preprocessed-connectomes-project.org/abide/), 2011.

[CRZ+15] W. Cheng, E.T. Rolls, J. Zhang, J. Feng, and H. Gu. Autism: reduced connectivity between cortical areas involved in face expression, theory of mind, and the sense of self. *Brain*, 139:3269–3309, 2015.

[CRZ+17] W. Cheng, E.T. Rolls, J. Zhang, W. Sheng, L. Ma, L. Wan, Q. Luo Luo, and J. Feng. Functional connectivity decreases in autism in emotion, self, and face circuits identified by knowledge-based enrichment analysis. *neuroimage*, 148:169–178, 2017.

[HAA11] H.M. Hassan and M.A. Abdel-Aty. Analysis of drivers’ behavior under reduced visibility conditions using a structural equation modeling approach. *Transportation research part F*, 14:614–625, 2011.

[Hoy95] R.H. Hoyle. *Structural equation modeling: Concepts, issues, and applications*. Sage Publications, 1995.

[HT93] C. Hurvich and C. Tsai. A corrected Akaike information criterion for vector autoregressive model selection. *Journal of time series analysis*, 14(3):271–279, 1993.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition, 2009.

[JCS+15] I. Jhun, B.A. Coull, J. Schwartz, B. Hubbell, and P. Koutrakis. The impact of weather changes on air quality and health in the United States in 1994–2012. *Environmental Research Letters*, 10(8):084009, 2015.

[JGM16] R. Jacobucci, K.J. Grimm, and J.J. McArdle. Regularized structural equation modeling. *Structural equation modeling: a multidisciplinary journal*, 23(4):555–566, 2016.

[JL99] E. Jauregui and E. Luyando. Global radiation attenuation by air pollution and its effects on the thermal climate in mexico city. *International Journal of Climatology*, 19(6):683–694, 1999.

[JMK+10] R.J. Jou, N.J. Minshew, M.S. Keshavan, M.P. Vitale, and A.Y. Hardan. Enlarged right superior temporal gyrus in children and adolescents with autism. *Brain Res*, 1360:205–212., 2010.

[JSTT00] K. Jöreskog, D. Sörbom, S. Du Toit, and M. Du Toit. *LISREL 8: New statistical features*. Scientific Software International, 2000.

28
[KNL02] T.W. Kjaer, M. Nowak, and H.C. Lou. Reflective self-awareness and conscious states: PET evidence for a common midline parietofrontal core. *NeuroImage*, 17(2):1080–1086, 2002.

[KW12] B. Kolb and I.Q. Whishaw. *An Introduction to Brain and Behavior*. Worth Publishers, 2012.

[KZC07] J. Kim, W. Zhu, L. Chang, P.M. Bentler, and T. Ernst. Unified structural equation modeling approach for the analysis of multisubject, multivariate functional MRI data. *Human Brain Mapping*, 28(2):85–93, 2007.

[Law05] M. G. Lawrence. The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications. *Bulletin of the American Meteorological Society*, 86(2):225–233, 2005.

[LFJ13] F. Lin, M. Fardad, and M.R. Jovanovic. Design of optimal sparse feedback gains via the alternating direction method of multipliers. *IEEE Transactions on Automatic Control*, 58(9):2426–2431, 2013.

[LLC04] H.C. Lou, B. Luber, M. Crupain, J.P. Keenan, M. Nowak, T.W. Kjaer, H.A. Sackeim, and S.H. Lisanby. Parietal cortex and representation of the mental Self. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6827–6832, 2004.

[LML09] X. Liu, X. Mei, Y. Li, Q. Wang, J.R. Jensen, Y. Zhang, and J.R. Porter. Evaluation of temperature-based global solar radiation models in China. *Agricultural and Forest Meteorology*, 149(9):1433–1446, 2009.

[MA00] R.C. MacCallum and J.T. Austin. Applications of structural equation modeling in psychological research. *Annual review of psychology*, 51(1):201–226, 2000.

[MGL94] A.R. Mcintosh and F. Gonzalez-Lima. Structural equation modeling and its application to network analysis in functional brain imaging. *Human Brain Mapping*, 2(1-2):2–22, 1994.

[Mul09] S.A. Mulaik. *Linear causal modeling with structural equations*. CRC Press, 2009.

[MYL14] A. Di Martino, C.G. Yan, Q. Li, E. Denio, F.X. Castellanos, K. Alaerts, J.S. Anderson, M. Assaf, S.Y. Bookheimer, M. Dapretto, et al. The Autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659, 2014.

[NHM09] E. Nestler, S. Hyman, and R. Malenka. *Molecular Neuropharmacology: A Foundation for Clinical Neuroscience*. New York: McGraw-Hill Medical, second edition, 2009.

[PB14] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[PLFI09] L.R. Price, A.R. Laird, P.T. Fox, and R.J. Ingham. Modeling dynamic functional neuroimaging data using structural equation modeling. *Structural Equation Modeling*, 16(1):147–162, 2009.

[Pru17] A. Pruttiakaravanich. Convex formulations for path analysis problems in structural equation modeling. Master’s thesis, Chulalongkorn University, 2017.

[PS16] A. Pruttiakaravanich and J. Songsiri. A convex formulation for path analysis in structural equation modeling. In *Proceedings of the 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pages 282–287. IEEE, 2016.
[RM06] T. Raykov and G.A. Marcoulides. *A First Course in Structural Equation Modeling*. Lawrence Erlbaum Associates, Inc., second edition, 2006.

[RM11] U. Ramanathan and L. Muyldermans. Identifying the underlying structure of demand during promotions: a structural equation modeling approach. *Expert systems with applications*, 38:5544–5552, 2011.

[Seg06] A. Seghouane. Vector autoregressive model-order selection from finite samples using Kullback’s symmetric divergence. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 53(10):2327–2335, 2006.

[Suk14] N.M. Suki. Passenger satisfaction with airline service quality in Malaysia: A structural equation modeling approach. *Research in transportation business & management*, 10:26–32, 2014.

[SWQ17] C. Singh, B. Wang, and Y. Qi. A Constrained, Weighted-L1 Minimization Approach for Joint Discovery of Heterogeneous Neural Connectivity Graphs. *CoRR*, abs/1709.04090, 2017.

[TMLP+02] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15:273–289, 2002.

[Van12] L. Vandenberghe. Convex optimization techniques in system identification. In *Proceedings of the IFAC Symposium on System Identification*, pages 71–76, 2012.

[WRGB03] M. Wallentin, A. Roepstorff, R. Glover, and N. Burgess. Parallel memory systems for talking about location and age in precuneus, caudate and Broca’s region. *NeuroImage*, 32(4):1850–1864, 2003.

[XFG17] Z. Xu, M.A.T. Figueiredo, and T. Goldstein. Adaptive ADMM with spectral penalty parameter selection. *arXiv:1605.07216v5 [cs.LG]*, 2017.

[XWH] M. Xia, J. Wang, and Y. He. BrainNet Viewer: A Network visualization tool for human brain connectomics. [http://www.nitrc.org/projects/bnv/](http://www.nitrc.org/projects/bnv/)
A  Dual problems

A.1  Dual problem of confirmatory SEM

This section describes the derivation of the dual and its KKT conditions. Let \( Z = \begin{bmatrix} Z_1 & Z_2^T \\ Z_2 & Z_4 \end{bmatrix} \in S^{2n}, \Omega \in S^n \) and \( U \in \mathbb{R}^{n \times n} \) be the Lagrange multipliers of the constraints

\[
\begin{bmatrix} X & (I-A)^T \\ I-A & \Psi \end{bmatrix} \succeq 0, \quad \Psi \preceq \alpha I, \quad P(A) = 0
\]

respectively. The Lagrangian of the problem (10) is

\[
L(X, A, \Psi, Z, \Omega, U) = -\log \det X + \text{tr}(SX) - \text{tr}(Z_1 X) - 2 \text{tr}(Z_2) + 2 \text{tr}(Z_2^T A) - \text{tr}(Z_4 \Psi) + \text{tr}(\Omega \Psi) - \alpha \text{tr}(\Omega) - 2 \text{tr}(U^T P(A)). \quad (25)
\]

The infimum of \( L \) with respect to the primal variables can be determined as follows. The term in \( L \) that is a function of \( \Psi \) is \( \text{tr}(\Omega \Psi) - \text{tr}(Z_4 \Psi) \). This function is linear in \( \Psi \), so the infimum is zero provided that \( \Omega = Z_4 \). The term in \( L \) that is a function of \( X \) is given by \( -\log \det X + \text{tr}(SX) - \text{tr}(Z_1 X) \) which can be minimized when its gradient with respect to \( X \) is zero. This gives \( -X^{-1} + S - Z_1 = 0 \) or that \( X = (S - Z_1)^{-1} \) and therefore,

\[
\inf_X \{-\log \det X + \text{tr}(SX) - \text{tr}(Z_1 X)\} = \log \det(S - Z_1) + n.
\]

Lastly, the infimum of the term in \( L \) that is a function of \( A \) (up to the scaling factor 2) is given by

\[
\inf_A \{ \text{tr}(Z_2^T A) - \text{tr}(U^T P(A)) \} = \inf_A \{ \text{tr}(Z_2^T A) - \text{tr}(P(U)^T A) \}
\]

where we have used the fact that the operator \( P \) defined is self-adjoint, i.e., \( \text{tr}(U^T P(A)) = \text{tr}(P(U)^T A) \). Hence, the expression is linear in \( A \), so the infimum is zero provided that \( Z_2 = P(U) \). This means the \((i,j)\) entries of \( Z_2 \) for \((i,j) \in I_A \) are free variables, and the other entries of \( Z_2 \) must be zero. This can be written in the matrix format as \( P^*(Z_2) = 0 \). The minimized Lagrangian with respect to the primal variables provides us the dual function

\[
g(Z) = \log \det(S - Z_1) + n - 2 \text{tr}(Z_2) - \alpha \text{tr}(Z_4)
\]

with the domain constraints: \( Z \succeq 0, \quad P^*(Z_2) = 0 \). The dual is the problem of maximizing the dual function which is obtained directly.

KKT conditions. If strong duality holds, then \( X, A, \Psi \) and \( Z \) are optimal if and only if the following conditions hold.

- **Zero gradient of the Lagrangian:** \( X = (S - Z_1)^{-1} \).
- **Primal feasibility:** \( (I-A)^T \Psi^{-1} (I-A) \preceq X, \quad 0 \prec \Psi \preceq \alpha I, \quad P(A) = 0 \).
- **Dual feasibility:** \( Z \succeq 0, Q(Z) = 0 \).
- **Complementary Slackness condition:**

\[
\text{tr} \left( \begin{bmatrix} Z_1 & Z_2^T \\ Z_2 & Z_4 \end{bmatrix} \begin{bmatrix} X & (I-A)^T \\ I-A & \Psi \end{bmatrix} \right) = 0, \quad \text{tr}(Z_4 (\Psi - \alpha I)) = 0.
\]
\[ \begin{bmatrix} X \\ I - A \end{bmatrix} \begin{bmatrix} (I - A)^T \\ \Psi \end{bmatrix} \geq 0, \quad \Psi \preceq \alpha I, \quad P(A) = 0 \]

respectively. With the notation \( h(A) = \sum_{(i,j) \notin I_A} |A_{ij}| \), the Lagrangian of (10) is

\[
L(X, A, \Psi, Z, \Omega, U) = -\log \det X + \text{tr}(SX) + 2\gamma h(A) - \text{tr}(Z_1X) - 2\text{tr}(Z_2) + 2\text{tr}(Z_2^T A) - \text{tr}(Z_4 \Psi) + \text{tr}(\Omega \Psi) - \alpha \text{tr}(\Omega) - 2\text{tr}(U^T P(A)).
\]

The infimum of \( L \) with respect to the primal variables can be determined as follows. The term in \( L \) that is a function of \( \Psi \) is \( \text{tr}(\Omega \Psi) - \alpha \text{tr}(\Omega) \) which can be minimized when its gradient with respect to \( \Psi \) is zero. This gives \( -X^{-1} + S - Z_1 = 0 \) or that \( X = (S - Z_1)^{-1} \) and

\[
\inf_X \{ -\log \det X + \text{tr}(SX) - \text{tr}(Z_1X) \} = \log \det(S - Z_1) + n.
\]

Lastly, the infimum of the term in \( L \) that is a function of \( A \) is given by

\[
\inf_A \left\{ 2\gamma h(A) + 2\text{tr}(Z_2^T A) - 2\text{tr}(U^T P(A)) \right\} = \inf_A \left\{ 2\gamma h(A) + 2\text{tr}(Z_2^T A) - 2\text{tr}(P(U)^T A) \right\} = 2\gamma h^*(\frac{P(U) - Z_2}{\gamma})
\]

where we have used the conjugate function of \( h \), denoted by \( h^* \) and given by

\[
h^*(Y) = \begin{cases} 
0, & Y_{ij} = 0 \text{ for } (i, j) \in I_A \text{ and } \max_{(i,j) \notin I_A} |Y_{ij}| \leq 1 \\
\infty, & \text{otherwise}.
\end{cases}
\]

(See a derivation the conjugate function in [Pru17].) Therefore, the infimum of \( L \) that involves \( A \) is zero provided that

\[
P(P(U) - Z_2) = 0, \quad \|P^c(P(U) - Z_2)\|_{\infty} \leq \gamma.
\]

Since the entries in \( U \) can be chosen arbitrarily, and \( P(P(U) - Z_2) = P(U) - P(Z_2) = 0 \), we can say that \( P(Z_2) \) contains free entries. Moreover, \( P^c(P(U) - Z_2) = P^c(P(U)) - P^c(Z_2) = -P^c(Z_2) \). We then conclude that the infimum of \( L \) with respect to \( A \) is zero when \( \|P^c(Z_2)\|_{\infty} \leq \gamma \) or equivalently that \( |Z_2|_{ij} \leq \gamma \) for \( (i, j) \notin I_A \). The minimized Lagrangian with respect to the primal variables provides us the dual function

\[
g(Z) = \log \det(S - Z_1) + n - 2\text{tr}(Z_2) - \alpha \text{tr}(Z_4)
\]

with the domain constraints: \( Z \preceq 0, \quad \|P^c(Z_2)\|_{\infty} \leq \gamma \).

**KKT conditions.** If strong duality holds, then \( X, A, \Psi \) and \( Z \) are optimal if and only if the following conditions hold.

- **Zero gradient of the Lagrangian:** \( X = (S - Z_1)^{-1} \).
- **Primal feasibility:** \((I - A)^T \Psi = 0, \quad 0 \preceq \Psi \preceq \alpha I, \quad P(A) = 0 \).
- **Dual feasibility:** \( Z = \begin{bmatrix} Z_1 \\ Z_2 \\ Z_4^T \end{bmatrix} \succeq 0, \quad \|P^c(Z_2)\|_{\infty} \leq \gamma \).
- **Complementary slackness condition:**

\[
\text{tr} \left( \begin{bmatrix} Z_1 \\ Z_2 \\ Z_4^T \end{bmatrix} \begin{bmatrix} X \\ I - A \end{bmatrix} \begin{bmatrix} (I - A)^T \\ \Psi \end{bmatrix} \right) = 0, \quad \text{tr} \left( Z_4 \Psi - \alpha I \right) = 0.
\]
B Pseudo codes of algorithms

We provide the summary of ADMM algorithms used to solve the problems (15) and (20).

B.1 ADMM for solving the convex confirmatory SEM (15).

All variables in the algorithms consist of $X, U, V, Y_1, Y_2 \in \mathbb{S}^{2n}$. Problem parameters are $\alpha > 0$ and the covariance matrix $S$. The algorithm parameter is $\rho > 0$.

Initialization. Choose $\alpha = \lambda_{\min}(S)$ and $Y_1, Y_2, U$ as identity matrix and $V$ such that $V_1 = 0, V_2 = I, V_4 = \alpha I$.

Repeat
1. Set $M = \frac{1}{2}(U + V) - \frac{1}{2\rho}(Y_1 + Y_2)$
2. Perform eigenvalue decomposition: $2\rho M_1 - S = QAQ^T, \Lambda = \text{diag}(\lambda)$.
3. Set a temporary diagonal matrix $\tilde{X}$ where $(\tilde{X})_{ii} = (\lambda_i + \sqrt{\lambda_i^2 + 8\rho})/4\rho$, for $i = 1, \ldots, n$
4. $X^+ = \begin{bmatrix} Q\tilde{X}Q^T & M_2^T \\ M_2 & M_4 \end{bmatrix}$
5. $U^+ = \Pi_C(X + Y_1/\rho)$ where $C = \mathbb{S}^{2n}_+$
6. Set $M = X + Y_2/\rho$
7. $V^+ = \begin{bmatrix} M_1 & P^c(M_2^T) + I \\ P^c(M_2) + I & \Pi_C(M_4) \end{bmatrix}$ where $C = \{X \in \mathbb{S}^n \mid 0 \preceq X \preceq \alpha I\}$
8. $Y_1^+ = Y_1 + \rho(X^+ - U^+)$ and $Y_2^+ = Y_2 + \rho(X^+ - V^+)$

Until the primal residual ($r$) and dual residual ($s$) are less than some tolerance values:

$$\|r\|_F = \left\| \begin{bmatrix} X - U \\ X - V \end{bmatrix} \right\|_F \leq \epsilon_{\text{pri}}, \quad \|s\|_F = \rho \left\| \begin{bmatrix} X^+ - X \\ U^+ - U \\ V^+ - V \end{bmatrix} \right\|_F \leq \epsilon_{\text{dual}}.$$

The tolerance values $\epsilon_{\text{pri}}$ and $\epsilon_{\text{dual}}$ can be computed by

$$\epsilon_{\text{pri}} = 2n\epsilon_{\text{abs}} + \epsilon_{\text{rel}} \max\{|\|X\|_F, |\|U\|_F, |\|V\|_F\|\}, \quad \epsilon_{\text{dual}} = 2n\epsilon_{\text{abs}} + \epsilon_{\text{rel}} \max\{|\|Y_1\|_F, |\|Y_2\|_F\|\},$$

where $\epsilon_{\text{abs}} = 10^{-8}$ and $\epsilon_{\text{rel}} = 10^{-8}$ are chosen.

B.2 ADMM for solving the sparse SEM (20)

All variables in the algorithms consist of $X, U, V, Y_1, Y_2, Y_3 \in \mathbb{S}^{2n}$ and $Z, Y_1 \in \mathbb{R}^{n \times n}$. Problem parameters are $\gamma > 0, \alpha > 0$ and the covariance matrix $S$. The algorithm parameter is $\rho > 0$.

Initialization. Choose $\alpha = \lambda_{\min}(S)$ and $Y_1, Y_2, Y_3, U$ as identity matrix, $Z = 0$, and $V$ such that $V_1 = 0, V_2 = I, V_4 = \alpha I$. 

33
We will prove by contradiction: if \( \alpha C \) Mathematical proofs

Repeat

1. Set \( M = \frac{1}{2}(U + V) - \frac{1}{2\rho}(Y_2 + Y_3) \) and \( H = I - Z - Y_1/\rho \)
2. Perform eigenvalue decomposition: \( 2\rho M_1 - S = QAQ^T, \Lambda = \text{diag}(\Lambda) \).
3. Set a temporary diagonal matrix \( X_1 \) where \( (X_1)_{ii} = (\lambda_i + \sqrt{\lambda_i^2 + 8\rho})/4\rho \), for \( i = 1, \ldots, n \)
4. \( X^+ = \begin{bmatrix} IQ_1Q^T \\ \frac{1}{2}(H + 4M_2) \\ M_4 \end{bmatrix} \)
5. Set \( M = I - X_2 - Y_1/\rho \)
6. \( Z_{ij} = -S_{2i}/\rho(M_{ij}) \) for \( (i, j) \notin I_A \) and \( Z_{ij} = 0 \) for \( (i, j) \in I_A \)
7. \( U^+ = \Pi_C(X + Y_2/\rho) \) where \( C = S^n_+ \)
8. Set \( M = X + Y_3/\rho \)
9. \( V^+ = \begin{bmatrix} M_1 \\ M_2 \\ \Pi_C(M_4) \end{bmatrix} \) where \( C = \{ X \in S^n | 0 \preceq X \preceq \alpha I \} \)
10. \( Y_1^+ = Y_1 + \rho(X_2^+Z - I), Y_2^+ = Y_2 + \rho(X^+ - U^+), \) and \( Y_3^+ = Y_3 + \rho(X^+ - V^+) \)

Until the primal residual \( (r) \) and dual residual \( (s) \) are less than some tolerance values:

\[
\|r\|_F = \left\| \begin{bmatrix} X - U \\ X - V \end{bmatrix} \right\|_F \leq \epsilon^{\text{pri}}, \quad \|s\|_F = \rho \left\| \begin{bmatrix} X^+ - X \\ U^+ - U \\ V^+ - V \end{bmatrix} \right\|_F \leq \epsilon^{\text{dual}}.
\]

The tolerance values \( \epsilon^{\text{pri}} \) and \( \epsilon^{\text{dual}} \) can be computed by

\[
\epsilon^{\text{pri}} = 2n \epsilon^{\text{abs}} + \epsilon^{\text{rel}} \max \{ \|X\|_F, \|Z\|_F, \|U\|_F, \|V\|_F \}, \quad \epsilon^{\text{dual}} = 2n \epsilon^{\text{abs}} + \epsilon^{\text{rel}} \max \{ \|Y_1\|_F, \|Y_2\|_F, \|Y_3\|_F \},
\]

where \( \epsilon^{\text{abs}} = 10^{-6} \) and \( \epsilon^{\text{rel}} = 10^{-6} \) are chosen.

## C Mathematical proofs

### C.1 Proof of Proposition 1

We will prove by contradiction: if \( \alpha \leq \alpha_c \) and then \([\mathbb{P}]\) has a solution by applying a generalization of Farka’s lemma to semidefinite programming \([\mathbb{BV04}]\).

**Lemma 3.** \([\mathbb{BV04}]\) The system

\[
Z \succeq 0, \quad \text{tr}(GZ) > 0, \quad \text{tr}(F_i Z) = 0, i = 1, 2, \ldots, n
\]

is a strong alternative for the nonstrict LMI: \( \sum_{i=1}^n x_i F_i + G \preceq 0 \), if the matrices \( F_i \) satisfy \( \sum_{i=1}^n v_i F_i \succeq 0 \) implies that \( \sum_{i=1}^n v_i F_i = 0 \).

The feasibility problem \([\mathbb{P}]\) can be expressed as an LMI

\[
\begin{bmatrix}
S^{-1} & (I - \tilde{A})^T & 0 \\
I - \tilde{A} & \Psi & 0 \\
0 & 0 & \alpha I - \Psi
\end{bmatrix} \succeq 0
\]

(27)
or equivalently, $G + \sum_{ij} A_{ij} F_{ij} + \sum_{ij} \Psi_{ij} H_{ij} \preceq 0$ where

$$G = \begin{bmatrix} -S^{-1} & -I & 0 \\ -I & 0 & 0 \\ 0 & 0 & -\alpha I \end{bmatrix}, \quad \sum_{ij} A_{ij} F_{ij} = \begin{bmatrix} 0 & \tilde{A}^T \\ \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}, \quad \sum_{ij} \Psi_{ij} H_{ij} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\Psi & 0 \\ 0 & 0 & \Psi \end{bmatrix}. $$

We note that $A_{ij}$ and $\Psi_{ij}$ are the $(i,j)$ entries of $\tilde{A}$ and $\Psi$, respectively. The matrices $F_{ij}$ and $H_{ij}$ are the common choice of standard basis matrices that make up to the above summations. To describe more details, let $E_{ij}$ be a standard basis matrix for set of $n \times n$ matrices with zero diagonals and $S_{ij}$ be a standard basis matrix for $S^n$. In other words, the entries of $E_{ij}$ are all zero except that the $(i,j)$ entry is 1. Similarly, the entries of $S_{ij}$ are all zero except that the $(i,j)$ and $(j,i)$ entries are 1. The expressions of $F_{ij}$ and $H_{ij}$ are

$$F_{ij} = \begin{bmatrix} 0 & E_{ij}^T \\ E_{ij} & 0 \end{bmatrix}, \quad \text{for } (i,j) \notin I_A, \quad H_{ij} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -S_{ij} & 0 \\ 0 & 0 & S_{ij} \end{bmatrix}, \quad \text{for } i \geq j = 1, 2, \ldots, n.$$

From Lemma 3, the LMI (27) has no solution if and only if $\exists U \succeq 0, U \neq 0$ such that

$$\text{tr}(GU) \geq 0, \quad \text{tr}(F_{ij} U) = 0, \quad \text{for } (i,j) \notin I_A, \quad \text{tr}(H_{ij} U) = 0, \quad \text{for } i \geq j.$$

In what follows, we will show that there always exists such matrix $U$ under the condition $\alpha \leq \alpha_c$. For scalars $\gamma$ and $\beta$ with $\beta \geq 0$ and $\gamma \neq 0$, we construct a positive semidefinite matrix $U$ of the form

$$U = \begin{bmatrix} (\gamma^2/\beta)I & \gamma I & 0 \\ \gamma I & \beta I & 0 \\ 0 & 0 & \beta I \end{bmatrix}.$$

With this choice, we can easily check that $\text{tr}(F_{ij} U) = 0$ regardless of $I_A$ (as long as $I_A$ contains the indices of diagonal entries of $A$), and that $\text{tr}(H_{ij} U) = 0$. We also see that

$$\sum_{ij} A_{ij} F_{ij} + \sum_{ij} \Psi_{ij} H_{ij} = \begin{bmatrix} 0 & \tilde{A}^T \\ \tilde{A} & -\Psi \\ 0 & 0 \end{bmatrix} \succeq 0$$

implies that $\Psi = 0$ and consequently conclude that $\tilde{A} = 0$ because 0 is in the leading $(1,1)$ block. Lastly, the condition $\text{tr}(GU) \geq 0$ is expressed as

$$\frac{\text{tr}(S^{-1})}{\beta} \left( \gamma^2 + \frac{2n\beta}{\text{tr}(S^{-1})} \gamma + \frac{n}{\text{tr}(S^{-1})}\alpha \beta^2 \right) \leq 0. \quad (28)$$

The above quadratic polynomial in $\gamma$ can be expressed in terms of $\alpha$ and $\alpha_c$ as

$$\gamma^2 + 2\alpha_c \beta \gamma + \alpha \alpha_c \beta^2 \leq 0. $$

Therefore, if $\alpha \leq \alpha_c$ then we can always choose any negative real value of $\gamma$ in the interval

$$(-\alpha_c \beta (1 + \sqrt{1 - \alpha/\alpha_c}), -\alpha_c \beta (1 - \sqrt{1 - \alpha/\alpha_c})), $$

so that (28) is satisfied. This concludes that if $\alpha \leq \alpha_c$ the alternative of (27) always has a solution. This completes the proof.

### C.2 Derivation of $\gamma_{\text{max}}$

We will show that there exists a critical value of $\gamma$, denoted by $\gamma_{\text{max}}$ such that if $\gamma \geq \gamma_{\text{max}}$, then the optimal solution of $A$ in (10) is the zero matrix. The derivation of $\gamma_{\text{max}}$ is, in fact, derived from
the KKT conditions given in \( \mathbf{A.2} \) and under an assumption that the optimal primal solution is low rank. If the optimal primal solution has low rank, i.e.,

\[
\text{rank} \left( \begin{bmatrix} \mathbf{X} & (\mathbf{I} - \mathbf{A})^T \\ \mathbf{I} - \mathbf{A} & \Psi \end{bmatrix} \right) = n
\]

then it follows from the complementary slackness conditions that \( \text{rank} \mathbf{Z} = n \) and \( \text{rank} \mathbf{Z}_4 = n \), so \( \mathbf{Z}_4 \) is invertible. This further implies that \( \Psi = \alpha \mathbf{I} \). Since we aim to characterize the dual feasibility condition when we obtain the sparsest solution of \( \mathbf{A} \), we set \( \mathbf{A} = 0 \) in the optimal condition, then the matrix

\[
\begin{bmatrix} \mathbf{X} & (\mathbf{I} - \mathbf{A})^T \\ \mathbf{I} - \mathbf{A} & \Psi \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \mathbf{I} \\ \mathbf{I} & \alpha \mathbf{I} \end{bmatrix}
\]

has rank \( n \) if and only if \( \mathbf{X} = (1/\alpha) \mathbf{I} \). From zero gradient of the Lagrangian condition, \( \mathbf{Z}_1 = \mathbf{S} - \alpha \mathbf{I} \). Substitute this in the slackness condition,

\[
\begin{bmatrix} \mathbf{X} & \mathbf{I} \\ \mathbf{I} & \alpha \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{S} - \alpha \mathbf{I} & \mathbf{Z}_2^T \\ \mathbf{Z}_2 & \mathbf{Z}_4 \end{bmatrix} = \mathbf{0}
\]

from which we can solve for \( \mathbf{Z}_2 \) as \( \mathbf{Z}_2 = (1/\alpha) (\alpha \mathbf{I} - \mathbf{S}) \) and dual feasibility condition becomes

\[
\gamma \geq (1/\alpha) \| P^c(\alpha \mathbf{I} - \mathbf{S}) \|_\infty.
\]

(29)

In conclusion, we have shown that if \( \mathbf{A} = 0 \) is the optimal solution to (10) and the optimal primal solution has rank \( n \), then (29) must be fulfilled. Since KKT conditions are sufficient and necessary conditions for the optimality of a convex problem. As a result, we can set

\[
\gamma_{\text{max}} = (1/\alpha) \| P^c(\alpha \mathbf{I} - \mathbf{S}) \|_\infty
\]

(30)

as the critical value of \( \gamma \), and conclude that for any \( \gamma \geq \gamma_{\text{max}} \), the optimal solution \( \mathbf{A} \) must be zero.

### C.3 Solution of scaled sparse SEM

We provide a proof of Proposition 2. It is straightforward to show that if \((\mathbf{X},\mathbf{Z})\) satisfies the KKT conditions of the unscaled problem (10) using parameter \((\mathbf{S},\alpha)\) then \((\tilde{\mathbf{X}},\tilde{\mathbf{Z}})\) provided in the statement also satisfies the KKT conditions of the scaled problem (10) using parameter \((\tilde{\mathbf{S}},\alpha)\). The KKT conditions of the unscaled problem we have to check are:

- **Primal feasibility**: \( \tilde{\mathbf{X}} \succeq 0 \) and \( \tilde{\mathbf{X}}_4 \preceq \tilde{\alpha} \mathbf{I} \).
- **Dual feasibility**: \( \tilde{\mathbf{Z}} \succeq 0 \) and \( \| P^c(\tilde{\mathbf{Z}}_2) \|_\infty \leq \gamma \).
- **Zero gradient of the Lagrangian**: \( \tilde{\mathbf{X}}_1 = (\mathbf{S} - \mathbf{Z}_1)^{-1} \).
- **Complementary slackness condition**: \( \tilde{\mathbf{X}} \tilde{\mathbf{Z}} = \mathbf{0} \) and \( \tilde{\mathbf{Z}}_4(\tilde{\mathbf{X}}_4 - \tilde{\alpha} \mathbf{I}) = \mathbf{0} \).

Each of the above conditions can be examined as follows provided that the KKT conditions of the unscaled problem are satisfied.

- **Primal feasibility**: Given that \( \mathbf{X} \succeq 0 \) which is equivalent to

\[
\mathbf{X}_1 - \mathbf{X}_2^T \mathbf{X}_4^{-1} \mathbf{X}_2 \succeq 0, \quad \mathbf{X}_4 \succ 0,
\]

by the Schur complement. For any \( \beta > 0 \), it follows that the above inequalities are preserved under positive scaling:

\[
\beta \mathbf{X}_4 \succ 0, \quad \mathbf{X}_1/\beta - \mathbf{X}_2^T (\beta \mathbf{X}_4)^{-1} \mathbf{X}_2 \succeq 0
\]

which are equivalent to \( \tilde{\mathbf{X}} \succ 0 \) by the Schur complement when \( \tilde{\mathbf{X}} \) is given in (22). Moreover, it is obvious that if \( \mathbf{X}_4 \preceq \alpha \mathbf{I} \) then \( \tilde{\mathbf{X}}_4 = \beta \mathbf{X}_4 \preceq \beta \alpha \mathbf{I} = \tilde{\alpha} \mathbf{I} \).
• Dual feasibility: Given that $Z \succeq 0$, by the Schur complement, we have
$$Z_1 - Z_4^T Z_4^{-1} Z_2 \succeq 0, \quad Z_4 \succ 0.$$ 
It follows the same way that the above inequalities can be scaled by a positive scalar, and therefore are equivalent to
$$\beta Z_4 / \beta \succ 0, \quad \beta Z_1 - Z_2^T (Z_4 / \beta)^{-1} Z_2 \succeq 0.$$ 
By the Schur complement, this means we also have $\bar{Z} \succ 0$ when $\bar{Z}$ is given in (22). Next, since $\bar{Z}_2 = Z_2$, we immediately have $\|P_c(\bar{Z}_2)\|_\infty \leq \gamma$.

• Zero gradient of the Lagrangian: If $X_1 = (S - Z_1)^{-1}$, then we scale both sides by $1 / \beta$ and obtain $X_1 / \beta = (\beta (S - Z_1))^{-1}$, which is the same as $\bar{X}_1 = (\bar{S} - \bar{Z}_1)^{-1}$.

• Complementary slackness: By a simple algebraic operation, we check that if
$$\begin{bmatrix} Z_1 & Z_2^T \\ Z_2 & Z_4 \end{bmatrix} \begin{bmatrix} X_1 & X_2^T \\ X_2 & X_4 \end{bmatrix} = 0,$$ 
then for any $\beta \neq 0$, we also have
$$\begin{bmatrix} \beta Z_1 & Z_2^T \\ Z_2 & Z_4 / \beta \end{bmatrix} \begin{bmatrix} X_1 / \beta & X_2^T \\ X_2 & \beta X_4 \end{bmatrix} = 0.$$ 
Moreover, if $Z_4 (X_4 - \alpha I) = 0$ then $(1/\beta Z_4)(\beta X_4 - \beta \alpha I) = \bar{Z}_4 (\bar{X}_4 - \bar{\alpha} I) = 0$.

In addition to above results, we can check from (22) that if $(X, Z)$ have the low rank properties: $X_4 = X_1 - X_2^T X_4^{-1} Z_2$ and $Z_4 = Z_1 - Z_2^T Z_4^{-1} Z_2$, then $(\bar{X}, \bar{Z})$ also have low rank.

**D Brain Region of Interest (ROI)**
Table 3: Names of region of interests (ROIs) in fMRI connectivity modeling according to Automated Anatomical Labeling (AAL) template.

| No. | Name                                         | No.  | name                                         |
|-----|----------------------------------------------|------|----------------------------------------------|
| 1   | Left precentral gyrus (PreCG.L)              | 2    | Right precentral gyrus (PreCG.R)             |
| 3   | Left superior frontal gyrus (SFGdor.L)       | 4    | Right superior frontal gyrus (SFGdor.R)      |
| 5   | Left superior frontal gyrus, orbital part (ORB-sup.L) | 6    | Right superior frontal gyrus, orbital part (ORB-sup.R) |
| 7   | Left middle frontal gyrus (MFG.L)            | 8    | Right middle frontal gyrus (MFG.R)           |
| 9   | Left middle frontal gyrus, orbital part (ORB-mid.L) | 10   | Right middle frontal gyrus, orbital part (ORB-mid.R) |
| 11  | Left inferior frontal gyrus, pars opercularis (IFGoperc.L) | 12   | Right inferior frontal gyrus, pars opercularis (IFGoperc.R) |
| 13  | Left inferior frontal gyrus, pars triangularis (IFGtriang.L) | 14   | Right inferior frontal gyrus, pars triangularis (IFGtriang.R) |
| 15  | Left inferior frontal gyrus, pars orbitalis (ORBInf.L) | 16   | Right inferior frontal gyrus, pars orbitalis (ORBInf.R) |
| 17  | Left Rolandic operculum (ROL.L)              | 18   | Right Rolandic operculum (ROL.R)             |
| 19  | Left supplementary motor area (SMA.L)        | 20   | Right supplementary motor area (SMA.R)       |
| 21  | Left olfactory cortex (OLF.L)                | 22   | Right olfactory cortex (OLF.R)               |
| 23  | Left medial frontal gyrus (SFGmed.L)         | 24   | Right medial frontal gyrus (SFGmed.R)        |
| 25  | Left medial orbitofrontal cortex (ORB-supmed.L) | 26   | Right medial orbitofrontal cortex (ORB-supmed.R) |
| 27  | Left gyrus rectus (REC.L)                    | 28   | Right gyrus rectus (REC.R)                   |
| 29  | Left insula (INS.L)                          | 30   | Right insula (INS.R)                         |
| 31  | Left anterior cingulate gyrus (ACG.L)        | 32   | Right anterior cingulate gyrus (ACG.R)       |
| 33  | Left midcingulate area (DCG.L)               | 34   | Right midcingulate area (DCG.R)              |
| 35  | Left posterior cingulate gyrus (PCG.L)       | 36   | Right posterior cingulate gyrus (PCG.R)      |
| 37  | Left hippocampus (HIP.L)                     | 38   | Right hippocampus (HIP.R)                    |
| 39  | Left parahippocampal gyrus (PHG.L)           | 40   | Right parahippocampal gyrus (PHG.R)          |
| 41  | Left amygdala (AMYG.L)                       | 42   | Right amygdala (AMYG.R)                      |
| 43  | Left calcarine sulcus (CAL.L)                | 44   | Right calcarine sulcus (CAL.R)               |
| 45  | Left cuneus (CUN.L)                          | 46   | Right cuneus (CUN.R)                         |
| 47  | Left lingual gyrus (LING.L)                  | 48   | Right lingual gyrus (LING.R)                 |
| 49  | Left superior occipital (SOG.L)              | 50   | Right superior occipital (SOG.R)             |
| 51  | Left middle occipital gyrus (MOG.L)          | 52   | Right middle occipital gyrus (MOG.R)         |
| 53  | Left inferior occipital cortex (IOG.L)       | 54   | Right inferior occipital cortex (IOG.R)      |
| 55  | Left fusiform gyrus (FFG.L)                  | 56   | Right fusiform gyrus (FFG.R)                 |
| 57  | Left postcentral gyrus (PoCG.L)              | 58   | Right postcentral gyrus (PoCG.R)             |
| 59  | Left superior parietal lobule (SPG.L)        | 60   | Right superior parietal lobule (SPG.R)       |
| 61  | Left inferior parietal lobule (IPL.R)        | 62   | Right inferior parietal lobule (IPL.R)       |
| 63  | Left supramarginal gyrus (SMG.L)             | 64   | Right supramarginal gyrus (SMG.R)            |
| 65  | Left angular gyrus (ANG.L)                   | 66   | Right angular gyrus (ANG.R)                  |
| 67  | Left precuneus (PCUN.L)                      | 68   | Right precuneus (PCUN.R)                     |
| 69  | Left paracentral lobule (PCL.L)              | 70   | Right paracentral lobule (PCL.R)             |
| 71  | Left caudate nucleus (CAU.L)                 | 72   | Right caudate nucleus (CAU.R)                |
| 73  | Left putamen (PUT.L)                         | 74   | Right putamen (PUT.R)                        |
| 75  | Left globus pallidus (PAL.L)                 | 76   | Right globus pallidus (PAL.R)                |
| 77  | Left thalamus (THAL)                         | 78   | Right thalamus (THAL.R)                      |
| 79  | Left transverse temporal gyrus (HES.L)       | 80   | Right transverse temporal gyrus (HES.R)      |
| 81  | Left superior temporal gyrus (STG.L)         | 82   | Right superior temporal gyrus (STG.R)        |
| 83  | Left superior temporal pole (TPOsup.L)       | 84   | Right superior temporal pole (TPOsup.R)      |
| 85  | Left middle temporal gyrus (MTG.L)           | 86   | Right right middle temporal gyrus (MTG.R)    |
| 87  | Left middle temporal pole (TPOmid.L)         | 88   | Right middle temporal pole (TPOmid.R)        |
| 89  | Left inferior temporal gyrus (ITG.L)         | 90   | Right inferior temporal gyrus (ITG.R)        |