Learning Directed Acyclic Graphs with Mixed Effects Structural Equation Models from Observational Data

Xiang Li [1,†], Shanghong Xie [2,†], Peter McColgan [3], Sarah J. Tabrizi [3,4], Rachael I. Scahill [3], Donglin Zeng [5] and Yuanjia Wang [2,6,*]

[1] *Statistics and Decision Sciences, Janssen Research & Development, LLC, Raritan, NJ, U.S.A.*
[2] *Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, U.S.A.*
[3] *Huntington's Disease Centre, Department of Neurodegenerative Disease, UCL Institute of Neurology, London, WC1N 3BG, UK.*
[4] *National Hospital for Neurology and Neurosurgery, Queen Square, London, United Kingdom*
[5] *Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, U.S.A.*
[6] *Departments of Psychiatry, Columbia University Medical Center, New York, NY, U.S.A.*
[†] *These authors have contributed equally to this work and are joint first authors.*

Correspondence*:
Corresponding Author
yw2016@cumc.columbia.edu

ABSTRACT

The identification of causal relationships between random variables from large-scale observational data using directed acyclic graphs (DAG) is highly challenging. We propose a new mixed-effects structural equation model (mSEM) framework to estimate subject-specific DAGs, where we represent joint distribution of random variables in the DAG as a set of structural causal equations with mixed effects. The directed edges between nodes depend on observed exogenous covariates on each of the individual and unobserved latent variables. The strength of the connection is decomposed into a fixed-effect term representing the average causal effect given the covariates and a random effect term representing the latent causal effect due to unobserved pathways. The advantage of such decomposition is to capture essential asymmetric structural information and heterogeneity between DAGs in order to allow for the identification of causal structure with observational data. In addition, by pooling information across subject-specific DAGs, we can identify causal structure with a high probability and estimate subject-specific networks with a high precision. We propose a penalized likelihood-based approach to handle multi-dimensionality of the DAG model and a fast computational algorithm to achieve desirable sparsity by hard-thresholding the edges. We theoretically prove the identifiability of mSEM. Using simulations and an application to protein signaling data, we show substantially improved performances when compared to existing methods and consistent

results with a network estimated from interventional data. Lastly, we identify gray matter atrophy networks in regions of brain from patients with Huntington's disease and corroborate our findings using white matter connectivity data collected from an independent study.

Keywords: Graphical models, Network analysis, Causal structure discovery, Heterogeneity, Regularization

1 INTRODUCTION

Directed acyclic graphs (DAGs) are used to represent the causal mechanisms of a complex system of interacting components, such as biological cellular pathways [1], gene regulatory networks [2], and brain connectivity networks [3]. The ability to identify causal relations between variables in observational data is highly challenging. Specifically, given a set of centered random variables $\boldsymbol{M} = (M_1, \cdots, M_p)'$, referred to as nodes, the causal relationship between these nodes in a DAG can be represented by a structural equation model (SEM) [4]:

$$M_j = f_j(\text{pa}(j), \varepsilon_j), \quad j = 1, \cdots, p,$$

where pa(j) is the set of parental nodes of M_j, and ε_j is a random variable representing unexplained variation. In many applications, $\boldsymbol{M}$ is assumed to follow a multivariate Gaussian distribution satisfying a linear SEM,

$$M_j = \sum_{k \in \text{pa}(j)} \theta_{jk} M_k + \varepsilon_j, \quad \varepsilon_j \sim N(0, \sigma_j^2); \quad j = 1, \cdots, p, \tag{1}$$

where $\boldsymbol{B} = (\theta_{jk})$ is referred to as an adjacency matrix.

Estimation of DAG structure (i.e., parental sets pa(j)) is non-deterministic polynomial-time hard (NP-hard) because the number of possible DAGs grows super-exponentially with the number of nodes [5]. Mainly two types of methods are proposed to tackle this challenge, namely, independence-based [e.g., 6] and score-based [e.g., 7] methods. The independence-based approaches calculate the partial correlation between any pair of nodes and perform statistical tests to assess the conditional dependence. A popular method is the PC algorithm [6], which has been proven to be uniformly consistent for estimating ultra high-dimensional, sparse DAGs [8]. The PC algorithm was modified as PC-stable to remove its dependence on node ordering [9]. A limitation of the PC algorithm is that it does not provide the proper level of multiple comparison correction and thus may lead to a large number of false positives in practice. To remedy this limitation, a hybrid, two-stage approach was proposed [PenPC, 10] that first estimates a sparse skeleton based on penalized regression and then performs a modified PC-stable algorithm on the skeleton.

The score-based approach searches for the DAG using a pre-specified score criterion, such as Bayesian Information Criterion (BIC) or penalized likelihood function. As it is not computationally feasible to search through the space of all DAGs, a two-phase greedy equivalence search algorithm explores an equivalence class based on BIC by adding and deleting edges. With additional information on node ordering, the estimation of DAG is equivalent to neighbourhood selection for which several penalized likelihood approaches have been developed [11, 12]. More recently, attempts have been made to estimate a DAG without knowing the node ordering [13, 14]. Other recent developments include leveraging asymmetric information between nodes [15, 16] or exploring the invariance property of causal relation using combined observational and interventional data [17]. Simulation studies suggest that independence-based methods perform adequately for identifying the skeleton of a DAG from observational data [18]. However, these

methods may perform worse for identifying the causal direction than some search-and-score methods that exploit the asymmetric distributional information [18].

All of the existing DAG estimation methods assume homogeneity of the causal effect of the underlying DAG model in (1) (i.e., θ_{jk} is common across individuals in the population). However, there is a growing body of evidence suggesting that biological networks may depend on subject-specific characteristics such as genomic markers [19, 20, 21]. For mental disorders, individual differences in edge strength in comorbidity networks have been widely observed [22]. Modeling heterogeneity of network effects may improve interpretability, biological relevance, and predictability. This area is much less explored with the exception of a few methods proposed to study subject-specific undirected graphical models. For example, a conditional Gaussian graphical model with covariate-adjusted mean but homogeneous precision matrix has been considered [23, 24].To characterize heterogeneous dependence structure between groups, [25] jointly estimated graphical models that share common structure but also allowed for differences between networks. Recently, instead of modeling groups separately, [26] directly incorporated covariates into an Ising model in order to build a covariate-dependent undirected graph. A common assumption of these approaches is that the dependence between two nodes is fully explained by the observed exogenous covariates. Such an assumption may not be satisfied in many biological and clinical applications due to the presence of unexplained latent residual heterogeneity representing hidden pathways between nodes. [27] proposed a Bayesian approach to estimate DAG by including non-Gaussian latent variables in a linear SEM, but does not estimate individual-specific graphs.

Our goal in this article is to develop a novel method and an efficient estimation procedure to study covariate-dependent DAGs with latent effect modification in multi-dimensional settings. Our method is based on mixed-effects SEM (mSEM) and penalized likelihood to obtain DAG structure and causal effects simultaneously. The covariates are treated as exogenous variables, and their joint distribution is not of interest. The key difference between mSEM and current approaches is that the causal effect, θ_{jk} in Model (1), is random and varies across individuals. To capture variation of the manifestation of causal relationship among individuals, our model allows the magnitude of the edge strength to be heterogeneous across subjects, while keeping the direction of causal relationship to be homogeneous. The heterogeneous causal magnitude is modeled by both fixed effects that depend on observed covariates and random effects that capture unexplained heterogeneity.

We propose a two-stage approach to estimate mSEM, whereby the first stage performs neighborhood selection by maximizing a penalized likelihood to identify a sparse skeleton, and the second stage searches for the DAG by solving an approximate ℓ_0-penalty problem via hard-thresholding within the identified skeleton, followed by an easily implemented DAG-checking procedure. We show theoretical proof of the identifiability (the graph is unique) of our model. Through extensive simulations and application to a well-known protein signaling study [1], we show substantially improved performance in terms of robustness and accuracy when compared to existing methods, including PC and penPC, and consistent performance when compared to analysis using interventional data. Lastly, we apply the proposed method to discover the causal dependence relationship among regions of brain atrophy from patients with Huntington's disease (HD) [28] and corroborate our findings in an independent study [29].

2 METHODOLOGY

For the ith subject, let $\boldsymbol{M}_i = (M_{i1}, M_{i2}, \cdots, M_{ip})'$ denote p random variables or nodes in a DAG. Let $\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, \cdots, X_{iq})'$ denote a $q+1$-dimensional vector including a constant and q exogenous

covariates that may modify the causal network among components in $\boldsymbol{M}_i$. We consider a mixed-effects model in which the causal effect depends on both fixed effects of observed variables $\boldsymbol{X}_i$ and unobserved random effects $\{\gamma_{ijk}\}$. For the jth node, the mSEM is given by:

$$M_{ij} = \sum_{k \in \text{pa}(j)} (\boldsymbol{\beta}_{jk}^T \boldsymbol{X}_i + \gamma_{ijk}) M_{ik} + \varepsilon_{ij}, \tag{2}$$

where $\boldsymbol{\beta}_{jk}$ is the vector of fixed effects (including an intercept and effects associated with $\boldsymbol{X}_i$), and γ_{ijk} is the unexplained heterogeneity of causal effects beyond $\boldsymbol{X}_i$. We assume that γ_{ijk} are independent and follow $N(0, \sigma_{jk}^2)$ and the independent error terms ε_{ij} follow $N(0, \sigma_{\varepsilon_j}^2)$. The SEM in (2) assumes that for each edge in the DAG, the causal effect is decomposed into a subject-specific fixed-effect term that depends on the exogenous covariates (i.e., $\boldsymbol{\beta}_{jk}^T \boldsymbol{X}_i$) and a subject-specific random-effect term that captures residual heterogeneity in causal effects due to other latent factors beyond $\boldsymbol{X}_i$ (i.e., γ_{ijk}). When $\boldsymbol{\beta}_{jk} = \boldsymbol{0}$ and $\sigma_{jk}^2 \neq 0$, the causal dependence between j and k is explained by unobserved latent factors but not $\boldsymbol{X}_i$. No causal effect between node j and k corresponds to $\boldsymbol{\beta}_{jk} = \boldsymbol{0}$ and $\sigma_{jk}^2 = 0$.

In this work, we assume that the ordering of causal dependence or the parental sets are unknown, and propose methods to simultaneously learn the ordering and structure of DAG and the parameters in the SEM. Previous literature has pointed out that qualitative capacity claims about causal effects are invariant across different populations of subjects, whereas the quantitative claims in SEM often are population-specific [e.g., 30, Chapter 7]. Thus, we assume that the qualitative causal dependence (set of nodes and directed edges) is homogeneous among subjects while the magnitude of the edge strength varies across subjects. Presence of an edge from M_{ik} to M_{ij} is defined as $\boldsymbol{\beta}_{jk} \neq \boldsymbol{0}$ or $\sigma_{jk}^2 \neq 0$; otherwise, there is no causal effect from M_{ik} to M_{ij}. Note that when the components of $\boldsymbol{\beta}_{jk}$ associated with covariates X_{il} are zero and σ_{jk}^2 are zero, the subject-specific DAG model in (2) reduces to a homogeneous DAG model in (1). We express the model for $\boldsymbol{M}_i$ given γ_{ijk} in matrix form as

$$\boldsymbol{M}_i = \big(\boldsymbol{B}(\boldsymbol{X}_i) + \Gamma_i\big)\boldsymbol{M}_i + \boldsymbol{\varepsilon}_i \tag{3}$$

where $\boldsymbol{B}(\boldsymbol{X}_i)$ is a matrix of fixed effects with entry (j,k) as $\boldsymbol{\beta}_{jk}^T \boldsymbol{X}_i$ and the diagonal elements as zeros, Γ_i is a matrix of random effects with entry (j,k) as γ_{ijk} and the diagonal elements as zeros, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{ip})'$ is a vector of error terms. Note that the joint distribution of $\boldsymbol{M}$ in Model (3) is non-Gaussian due to random effects in Γ_i, where the asymmetric information on the distribution between nodes can facilitate inference on the causal network from the observational data.

To estimate a DAG, we use a likelihood-based approach. Given the random effects Γ_i, the conditional likelihood function of $\boldsymbol{M}_i$ is given by

$$\begin{aligned} p(\boldsymbol{M}_i; \boldsymbol{X}_i | \Gamma_i) &\propto |\boldsymbol{E}|^{-1/2} |\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{X}_i) - \Gamma_i| \times \\ &\exp\left(-\frac{1}{2}\boldsymbol{M}_i^T (\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{X}_i) - \Gamma_i)^T \boldsymbol{E}^{-1} (\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{X}_i) - \Gamma_i) \boldsymbol{M}_i\right), \end{aligned} \tag{4}$$

where $\text{Cov}[\boldsymbol{\varepsilon}_i] = \boldsymbol{E}$ is a diagonal matrix of $\sigma_{\varepsilon_j}^2$.

To simplify presentation, we introduce the notation for the vectorized Γ_i and define non-zero components of vectorized Γ_i as $\boldsymbol{\gamma}_i = \{\gamma_{ijk} : \sigma_{jk}^2 > 0\}$. Then, Γ_i can be expressed as a linear combination of components in $\boldsymbol{\gamma}_i$ as $\Gamma_i = \sum_{\sigma_{jk}^2 > 0} \gamma_{ijk} \boldsymbol{H}_{jk}$, where $\boldsymbol{H}_{jk}$ is a single-entry matrix with one entry (j,k).

Denote by $\mathrm{Cov}[\boldsymbol{\gamma}_i] = \boldsymbol{G}$ the covariance matrix of $\boldsymbol{\gamma}_i$. The observed likelihood function is given by

$$\prod_{i=1}^{n} \int_{\boldsymbol{\gamma}_i} p(\boldsymbol{M}_i; \boldsymbol{X}_i | \boldsymbol{\gamma}_i) p(\boldsymbol{\gamma}_i) \mathrm{d}\boldsymbol{\gamma}_i, \tag{5}$$

where $p(\boldsymbol{\gamma}_i) \propto |\boldsymbol{G}|^{-1/2} \exp\left(-\boldsymbol{\gamma}_i^T \boldsymbol{G}^{-1} \boldsymbol{\gamma}_i / 2\right)$.

Under the DAG assumption of no directed cycle, $\boldsymbol{B}(\boldsymbol{X}_i) + \Gamma_i$ can always be transformed into an upper diagonal matrix after some unknown permutation of the rows and columns. Therefore, the determinant $|\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{X}_i) - \Gamma_i|$ in the likelihood function (4) is one. The integral in the likelihood (5) can be explicitly calculated and the negative log-likelihood function is given by

$$l_n = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{\left(M_{ij} - \sum_{k \neq j} \left(\boldsymbol{\beta}_{jk}^T \boldsymbol{X}_i \right) M_{ik} \right)^2}{\sum_{k \neq j} \sigma_{jk}^2 M_{ik}^2 + \sigma_{\varepsilon_j}^2} + \log \left(\sum_{k \neq j} \sigma_{jk}^2 M_{ik}^2 + \sigma_{\varepsilon_j}^2 \right) \right) \tag{6}$$

up to a constant. Thus, parameter estimation in the likelihood is separable, leading to a great computational advantage. With a small number of nodes, in order to minimize the negative log-likelihood function (6), one can alternatively solve the weighted least squares to update $\{\boldsymbol{\beta}_{jk} : j = 1, \cdots, p;\ k = 1, \cdots, p\}$ and use the Newton-Raphson algorithm to update $\{\sigma_{jk}^2 : j = 1, \cdots, p;\ k = 1, \cdots, p\}$ and $\{\sigma_{\varepsilon_j}^2 : j = 1, \cdots, p\}$ until convergence. The identifiability of parameters in the model is shown in Theorem 1 in Section 2.3.

2.1 Initial Sparse Graph

With a large number of nodes, minimizing (6) would result in a full graph with all non-null estimates of $\{\boldsymbol{\beta}_{jk}\}$ and σ_{jk}^2. Without any constraint on the estimates, the graph may potentially involve many false positive edges. To accommodate the large number of nodes, we propose to use a penalized likelihood to choose an initial sparse graph skeleton and search for the optimum of (6) within this reduced graph space. Based on model (2), the marginal expectation and variance of M_{ij} are $\sum_{k \neq j} \left(\boldsymbol{\beta}_{jk}^T \boldsymbol{X}_i \right) M_{ik}$ and $\sigma_{\varepsilon_j}^2 + \sum_{k \neq j} \sigma_{jk}^2 M_{ik}^2$, respectively. Define $R_{ij} = M_{ij} - \sum_{k \neq j} \left(\boldsymbol{\beta}_{jk}^T \boldsymbol{X}_i \right) M_{ik}$. By the method of moments, we obtain initial estimates of the graph by minimizing the following objective functions $\sum_{i=1}^{n} \Big(M_{ij} - \sum_{k \neq j} \left(\boldsymbol{\beta}_{jk}^T \boldsymbol{X}_i \right) M_{ik} \Big)^2$ and $\sum_{i=1}^{n} \left(R_{ij}^2 - \sigma_{\varepsilon_j}^2 - \sum_{k \neq j} \sigma_{jk}^2 M_{ik}^2 \right)^2$ for each j with $j = 1, \cdots, p$. In order to obtain an initial sparse graph, ℓ_1-norm penalty can be included to minimize the objective function and obtain initial estimates $\{\widetilde{\boldsymbol{\beta}}_{jk}\}$, $\{\widetilde{\sigma}_{jk}^2\}$, and $\{\widetilde{\sigma}_{\varepsilon_j}^2\}$:

$$\begin{aligned} &\sum_{j=1}^{p} \left(\sum_{i=1}^{n} \Big(M_{ij} - \sum_{k \neq j} \left(\boldsymbol{\beta}_{jk}^T \boldsymbol{X}_i \right) M_{ik} \Big)^2 + \lambda_1 \sum_{k \neq j} \|\boldsymbol{\beta}_{jk}\|_1 \right), \\ &\sum_{j=1}^{p} \left(\sum_{i=1}^{n} \left(\widetilde{R}_{ij}^2 - \sigma_{\varepsilon_j}^2 - \sum_{k \neq j} \sigma_{jk}^2 M_{ik}^2 \right)^2 + \lambda_2 \sum_{k \neq j} \sigma_{jk}^2 \right), \\ &\quad \text{subject to } \ \sigma_{\varepsilon_j}^2 > 0,\ \sigma_{jk}^2 \geq 0, \end{aligned} \tag{7}$$

where $\widetilde{R}_{ij}$ is R_{ij} with β_{jk} replaced by $\widetilde{\boldsymbol{\beta}}_{jk}$, the parameter estimated from minimizing the first objective function of $\boldsymbol{\beta}$ at the current iteration. Here we use the same tuning parameter across nodes $j = 1, \cdots, p$ for illustration, although in practice node-specific tuning parameter can be used at the price of increasing computational burden. In cases where the topology of the graph varies greatly across nodes, different tuning parameters can be used. Given a regularization path with varying λ_1 and λ_2, we select the optimal λ_1^* and λ_2^* using the BIC criteria and apply the corresponding estimates as the initial skeleton. We set the edge (j, k) of the initial graph as null if $\widetilde{\boldsymbol{\beta}}_{jk} = 0$ and $\widetilde{\sigma}^2_{jk} = 0$.

2.2 Algorithms for Estimating DAG with Mixed-Effects Model (DAG-MM) and Justification

The initial graph, although asymptotically consistent[31], may not satisfy the DAG constraint due to that estimated $\widehat{\boldsymbol{\beta}}_{jk} \neq 0$ and $\widehat{\boldsymbol{\beta}}_{kj} \neq 0$ or $\widehat{\sigma}_{jk} \neq 0$ and $\widehat{\sigma}_{kj} \neq 0$. Define graph $\boldsymbol{A}$ (set of nodes, edges, and edge strength) as the set of non-null edges $\left\{(j,k) : \sqrt{\|\boldsymbol{\beta}_{jk}\|_2^2/q + \sigma^2_{jk}} > 0\right\}$ in the skeleton resulting from (7). Let $\boldsymbol{\theta}_{\boldsymbol{A}} = \left\{\boldsymbol{\beta}_{jk}, \sigma^2_{jk} : (j,k) \in \boldsymbol{A};\ \sigma^2_{\varepsilon_j} : j = 1, \cdots, p\right\}$ be the parameters for graph $\boldsymbol{A}$ and $n_{\boldsymbol{A}}$ be the number of non-zero edges of $\boldsymbol{A}$. To obtain a sparse DAG, a direct approach is to constrain the number of edges in the graph by optimizing a regularized likelihood:

$$\min\ l_n(\boldsymbol{\theta}_{\boldsymbol{A}}), \quad \text{subject to } \ \boldsymbol{A} \text{ is a DAG and } n_{\boldsymbol{A}} < C, \tag{8}$$

where C is a tuning parameter controlling the number of edges in $\boldsymbol{A}$. The constraints in (8) guarantee the estimated graph is a DAG and also perform edge selection. However, the optimization in (8) is NP-hard, because one needs to evaluate all possible graphs that satisfy the constraint $n_{\boldsymbol{A}} < C$. Furthermore, the computational challenge is elevated due to the acyclic constraint.

Instead, we perform hard-thresholding to approximately solve the ℓ_0-norm constrained optimization problem in (8). Specifically, after the estimates in $\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}}$ are obtained for a given graph skeleton $\boldsymbol{A}$, we perform hard-threshold on the estimated edge weights by removing the edge with the smallest $\sqrt{\|\widehat{\boldsymbol{\beta}}_{jk}\|_2^2/q + \widehat{\sigma}^2_{jk}}$ from $\boldsymbol{A}$ and then update the graph $\boldsymbol{A}$. Given an updated graph $\boldsymbol{A}$, we then start from the estimates obtained in the previous iteration and update the estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}}$. This procedure continues until some criterion of optimality is met. In our implementation, we use BIC as the criterion to select the optimal graph.

The above procedure can be summarized into a DAG-MM algorithm (described in Algorithm 1). The tasks include identifying graph structure (set of nodes and edges), direction of edges, and edge strength. DAG-MM consists of three main steps: estimation of sparse skeleton and edge strength, edge orientation, and iterative DAG building. In the first step, each node's Markov blanket is identified by penalized likelihood and edge strength is obtained. In the second step, edge orientation is performed by removing directionalities with weak dependence (computed from fixed-effects parameters and variances of random effects). In the third step, an iterative procedure performs edge pruning using the norm of the edge connection strength and searches for the DAG that satisfies the acyclic constraint using a general and fast routine described in a DAG-Checking algorithm (described in Algorithm 2 in Supplementary Material Section S1).

Algorithm 1 is computationally efficient for several reasons: the sparse skeleton reduces the search space of DAGs; ranking by the magnitude of edge effects provides search paths in the DAG space; selection criteria BIC is only calculated when the log-likelihood (6) is the correct model (i.e., the acyclic constraint is satisfied); and the optimal graph is selected from candidate DAGs. We observe empirically that the full

Algorithm 1: DAG with mixed model (DAG-MM)

1.Sparse skeleton: Estimate an initial sparse graph $\boldsymbol{A}_I$ by solving the objective function (7). Obtain the estimates $\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}_I}$ by minimizing (6) for $\boldsymbol{A}_I$.
2.Edge orientation: Initialize $\boldsymbol{A}_R = \boldsymbol{A}_I$. For (j,k) belongs to $\{(j,k) : (j,k) \in \boldsymbol{A}_R \text{ and } (k,j) \in \boldsymbol{A}_R\}$, prune the initial graph:
a.Calculate $c_{jk} = \sqrt{\|\widehat{\boldsymbol{\beta}}_{jk}\|_2^2/q + \widehat{\sigma}_{jk}^2}$ and $r_{jk} = c_{jk}/c_{kj}$ for all $(j,k) \in \{(j,k) : (j,k) \in \boldsymbol{A}_R \text{ and } (k,j) \in \boldsymbol{A}_R\}$.
b.Remove the edge (j,k), where $(j,k) = \arg\min_{j,k} r_{jk}$; update $\boldsymbol{A}_R = \boldsymbol{A}_R \setminus (j,k)$.
c.Update the estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}_R}$ by minimizing (6) for $\boldsymbol{A}_R$.
3.Iterative DAG building: Initialize $\boldsymbol{A}_1 = \boldsymbol{A}_R$. For $i = 1, \cdots, p*(p-1)/2$ or until $\boldsymbol{A}_i = \emptyset$, search DAG with hard-thresholding:
a.Update the estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}_i}$ by minimizing (6) for $\boldsymbol{A}_i$.
b.Calculate BIC if $\boldsymbol{A}_i$ is a DAG.
c.Perform edge pruning by removing the edge (j,k) with the smallest $\sqrt{\|\widehat{\boldsymbol{\beta}}_{jk}\|_2^2/q + \widehat{\sigma}_{jk}^2}$. Obtain the updated graph $\boldsymbol{A}_{i+1} = \boldsymbol{A}_i \setminus (j,k)$, and check whether $\boldsymbol{A}_{i+1}$ satisfies acyclic constraint by Algorithm 2 if $\boldsymbol{A}_i$ is not a DAG.

graph shrinks to a DAG very fast in only a few iterations of the third step. For implementation, we have developed main routines in C++ codes with an R interface (R program available upon request).

2.3 Rationale of DAG-MM Algorithm and Theoretical Result

Essentially DAG-MM uses the likelihood function as the objective function in the optimization and thus belongs to the class of score-based approaches for estimating DAG. Similar to other score-based methods in this class [7], the search is performed locally at each iteration. The first step provides a sparse skeleton and consistent initial estimators of DAG edge strength through moment estimation, with the magnitude of estimated effects close to the truth parameter values. In the second step, the direction that maximizes the network edge strength is selected. The rationale is that the overall edge strength under the correct direction is greater than the strength under the incorrect one (which is close to null effect). In the third step, the DAG with the lowest BIC objective function is selected. Under the identifiability result in Theorem 1 shown below, the optima is uniquely identified, and the DAG-MM algorithm may converge in a local neighborhood of true parameters.

Next, we prove the identifiability of the DAG-MM procedure. Here we omit the subscript i corresponding to subjects. For any matrix $B = \{\boldsymbol{\beta}_{jk}\}_{j,k}$ and $\Sigma = \{\sigma_{jk}^2\}_{j,k}$, we call (B, Σ) to be compatible with DAG, denoted by $(B, \Sigma) \sim DAG$, if the edge pair (j,k) such that $\boldsymbol{\beta}_{jk} \neq 0$ or $\sigma_{jk} \neq 0$ forms a DAG network. Furthermore, we use $L(B, \Sigma, \theta)$ to denote the likelihood function associated with (B, Σ) using the SEM, where $\theta = (\sigma_{\varepsilon_1}^2, ..., \sigma_{\varepsilon_p}^2)^T$. Note that if $(B, \Sigma) \sim DAG$, then $|I - B(X) - \Gamma| = 1$, so

$$L(B, \Sigma, \theta) = \exp\left\{ -\sum_{j=1}^{p} \left[\frac{(M_j - \sum_{k \neq j} (\boldsymbol{\beta}_{jk}^T X) M_k)^2}{\sum_{k \neq j} \sigma_{jk}^2 M_k^2 + \sigma_{\varepsilon_j}^2} + \log(\sum_{k \neq j} \sigma_{jk}^2 M_k^2 + \sigma_{\varepsilon_j}^2) \right] \right\}. \tag{9}$$

In Theorem 1, we prove the identifiability by assuming $\sigma_{\varepsilon_j}^2 > 0$ for any $j = 1, ..., p$.

THEOREM 1. *Assume that $P(\boldsymbol{\beta}^T X = 0) < 1$ for any $\boldsymbol{\beta} \neq 0$, i.e., X is full rank with positive probability. Let $(B_0, \Sigma_0, \theta_0)$ be the true values in the underlying true DAG, and let $ch_0(k)$ denote the set of child nodes of the node k. Assume that for all nodes k, $\sum_{j \in ch_0(k)} (\beta_{0jk}^T X)^2$ is not a constant (heterogeneity assumption) across nodes. Suppose $(B, \Sigma, \theta) \sim DAG$ and $L(B, \Sigma, \theta) = L(B_0, \Sigma_0, \theta_0)$. Then, $B_0 = B$, $\Sigma_0 = \Sigma$, and $\sigma^2_{0\varepsilon_j} = \sigma^2_{\varepsilon_j}$ for $j = 1, ..., p$.*

The proof of the theorem is in the Supplementary Material Section S2. The heterogeneity assumption implies that when there are multiple child nodes, their squared edge strengths from fixed effects are different across parental nodes. When there is a single child node, the edge strengths are different across subpopulations defined by covariates $\boldsymbol{X}$.

3 SIMULATION STUDIES

We performed comprehensive simulations to evaluate DAG-MM with varying sample sizes, $n = 200, 500, 1000$, and varying number of nodes, $p = 20, 50, 100$. We let $\sigma^2_{\varepsilon_{2*j-1}} = 1.0$ and $\sigma^2_{\varepsilon_{2*j}} = 0.5$, and the dimension of exogenous covariates $\boldsymbol{X}$ is 3: two of them are continuous variables that follow the standard normal distribution $N(0, 1)$, and the other is a binary variable that follows the Bernoulli distribution, $Bernoulli(0.5)$. Note that there are at most $p * (p-1) * (q+1) + p$ parameters to be estimated. For example, the total number of parameters is 1540 when $p = 100$ and $q = 3$. We fixed 12 non-zero edges as shown in Figure 1 (black edges), and the other features were independent noise variables. For the non-null edges, we let $\boldsymbol{\beta}_{jk} = (-0.5, 1.0, -1.5)$ and $\sigma^2_{jk} = 0.5$. Several settings were considered in our simulations:

1.Fixed effects only: $\boldsymbol{\beta}_{jk} = (-0.5, 1.0, -1.5)$ and $\sigma^2_{jk} = 0$ for $(j, k) \in \boldsymbol{A}^0$.
2.Random effects only: $\boldsymbol{\beta}_{jk} = \mathbf{0}$ and $\sigma^2_{jk} = 0.5$ for $(j, k) \in \boldsymbol{A}^0$.
3.Mixed effects 1: $\boldsymbol{\beta}_{jk} = (-0.5, 1.0, -1.5)$ and $\sigma^2_{jk} = 0.5$ for $(j, k) \in \boldsymbol{A}^0$.
4.Mixed effects 2: $\boldsymbol{\beta}_{jk} = (-0.5, 1.0, -1.5)$ for $(j, k) \in \big\{(1, 2), (1, 4), (4, 5), (7, 8), (8, 10), (11, 12), (12, 13), (14, 15)\big\}$ and $\sigma^2_{jk} = 0.5$ for $(j, k) \in \big\{(1, 2), (1, 3), (1, 4), (2, 3), (6, 7), (8, 9), (8, 10), (12, 13)\big\}$.
5.Homogeneous, constant effects without covariates or random effects: we include a column of ones into $\boldsymbol{X}_i$. $(\beta_{jk,2}, ..., \beta_{jk,q+1})' = \mathbf{0}$, $\sigma^2_{jk} = 0$, $\beta_{jk,1} = 1$ for $(j, k) \in \big\{(1, 2), (1, 4), (4, 5), (7, 8), (8, 10), (12, 13\big\}$, and $\beta_{jk,1} = -1$ for $(j, k) \in \big\{(1, 3), (2, 3), (6, 7), (8, 9), (11, 12), (14, 15)\big\}$.

In each simulation, we compared DAG-MM with the commonly used PC algorithm [32] and a two-step penalized version of the PC algorithm, penPC [10]. We used the default settings in R-packages "pcalg" and "penPC" for these alternative methods (e.g., with $\alpha = 0.1$). The edge selection performance was assessed by the number of true positive (TP) edges and false positive (FP) edges, taking into consideration the direction (i.e., an edge with a wrong direction will be counted as false). To evaluate the estimation of causal effects, we calculated the root sum squared (RSS) error of $\{\widehat{\boldsymbol{\beta}}_{jk}\}$, $\{\widehat{\sigma}^2_{jk}\}$, and $\{\widehat{\sigma}^2_{\varepsilon_j}\}$, which is defined as $RSS(\widehat{\boldsymbol{\beta}}) = \sqrt{\sum_{j \neq k} \|\widehat{\boldsymbol{\beta}}_{jk} - \boldsymbol{\beta}_{jk}\|_2^2}$, $RSS(\widehat{\sigma}^2) = \sqrt{\sum_{j \neq k} \left(\widehat{\sigma}^2_{jk} - \sigma^2_{jk}\right)^2}$, and $RSS(\widehat{\sigma}^2_{\varepsilon}) = \sqrt{\sum_{j=1}^{p} \left(\widehat{\sigma}^2_{\varepsilon_j} - \sigma^2_{\varepsilon_j}\right)^2}$, respectively.

The simulations were repeated 100 times for each setting.

Table 1 summarizes the number of TP and FP edge selections. The initial graph selection (i.e., performing steps 1 and 2 in Algorithm 1) correctly identified the true edges for all settings with TP edges very close

to 12, but also selected many FP edges. Starting from the initial graph, the DAG-MM procedure can retain almost all the TP edges and also remove most FP edges, with a FP rate close to 0. Note that there are 9900 edges in total when $p = 100$, and DAG-MM can still select the 12 true edges from a total of 9900 edges (0.05%). With a small sample size of $n = 200$, the performance of DAG-MM remains to be satisfactory, except in Setting 2. Setting 2 is more difficult because all edges involve latent effects. DAG-MM selects about 40% of TP edges when $n = 200$ and selects almost all true edges when the sample size increases to $n = 1,000$, without including FP edges. PC and penPC algorithms are designed for Setting 5 - constant effect without any covariates. As expected, they perform the best for Setting 5 but not other settings, and penPC selects fewer FP edges than PC algorithm due to an initial penalized regression step. However, for Setting 5, DAG-MM significantly outperforms the two PC algorithms in terms of fewer FP. Figure 1 visualizes the number of times (greater than one) that an edge is selected in the simulations. The visualization shows that DAG-MM performs satisfactorily and correctly identifies the true network structure in all settings. In contrast, penPC identifies many edges with incorrect direction and includes many more FP edges.

Next, we examined the estimation performance of the strength of the connection. Table 2 shows the RSS for parameters $\boldsymbol{\beta}$, σ^2, and σ_ε^2. Overall, RSS decreases to small values as sample size n increases. The increase in the number of features p affects the estimation of variance components σ^2 and σ_ε^2 more than $\boldsymbol{\beta}$. The results may suggest that for large p, including more samples improves the estimation performance of the individual-level heterogeneity associated with γ_{ijk}.

The computing time for DAG-MM is highly manageable. For example, in simulation Setting 5, the running time (averaged over 100 replicates) for simulated data with $n = 1000$ is 0.4 seconds for $p = 20$, 1.2 seconds for $p = 50$, and 4.4 seconds for $p = 100$, compared to 3.2, 16.8, and 66.5 seconds, respectively, for the penPC algorithm.

4 APPLICATIONS TO PROTEIN SIGNALING NETWORK AND BRAIN DEPENDENCE NETWORK

4.1 Protein Signaling Network

Our first application involved a study that examined the interaction between major mitogen-activated protein kinase (MAPK) pathways in human CD4+ T cells. Using intracellular multicolor flow cytometry, single-cell protein expression levels were measured for 11 proteins in the MAPK pathways in [1]. Six experiments were performed using different stimuli, each targeting a different protein in the selected pathway [1], and thus both interventional and unperturbed observational data were available for our application. Various data-driven methods were proposed to estimate the protein signaling networks, including Bayesian network analyses [1, 33] and ICP using combined interventional and observational data [17], and results were compared with a consensus network in the literature [33, 17].

In our analyses, we applied DAG-MM to learn the causal signaling network using unperturbed, observational data only. The observational data consisted of 2594 observations and were pre-processed using a standard arcsinh transformation for biological interpretability. DAG-MM with fixed effects only (DAG-MM1) and with mixed effects (DAG-MM2) were applied. Our results were compared with those obtained using the PC algorithm as reported in Kalisch et al. [32] and with ICP as reported in Meinshausen et al. [17] for both interventional and observational data. Table 3 summarizes the number of selected edges by each method and whether these edges were also previously reported in the literature. Treating the edges previously identified as “gold standard”, DAG-MM2 reduces the number of FP edges to a greater extent

than DAG-MM1. PC and ICP identified a similar number of true positive edges as DAG-MM2, but with a higher number of FP edges. In Figure 2, we compare DAG-MM2 with ICP. The skeleton of DAG-MM2 and ICP is almost identical, with DAG-MM2 identifying one more edge, Plcg $\rightarrow$ PIP3. Two edges were in the reverse direction of those reported in literature, which might due to feedback loops that are expected to be present in this system [17]. The striking similarity of DAG-MM2 identified from observational data alone and ICP using interventional data suggests robustness and the ability of the former to infer causal relationships from observational data by including random effects.

4.2 Brain Gray Matter Atrophy Dependence Network

Our second application involved a study on atrophy networks in the brains of patients with HD. HD is a monogenic neurodegenerative disorder caused by an expansion of the CAG trinucleotide (≥ 36) in the *huntingtin* gene [34]. The hallmark of HD neuropathology is brain atrophy, in terms of gray matter loss within the striatum and white matter loss around the striatum [35]. While evidence shows that selective brain regions undergo atrophy at different rates [28], it is unknown how these regional atrophies depend on one another and act together as disease progresses. In this application, we aimed to construct brain atrophy dependence networks using data collected from a large natural-history study of HD progression, PREDICT-HD [28], and we aimed to corroborate findings in an independent study, TRACK-ON [36]. Subcortical gray matter loss of volume and gray matter cortical thinning were considered as measures of brain atrophy and hallmarks of HD. Thus, we examined dependencies between rates of volume loss and cortical thinning in different brain regions.

For the PREDICT-HD study, we included individuals who carried an expansion of the CAG trinucleotide in the hungtington gene and thus were at risk of HD but had not been diagnosed at baseline. Data consisted of 824 subjects with 68 cortical regions of interest (ROI) and 22 subcortical ROIs measured by structural magnetic resonance imaging (MRI). Longitudinal assessments were obtained from these subjects with a median follow-up period of 3.9 years. The details of MRI data segementation, preprocessing, and study design are in [28]. A linear mixed-effects model with subject-specific random intercepts and random slopes was used to estimate the rate of volumetric change and the rate of cortical thickness change at each ROI for each subject. Rates of change at ROIs form the nodes in the brain atrophy dependence network. Because CAG repeats and age are two variables with substantial contribution to HD, a covariate based on the CAG-age product [CAPs score in 37] was created to indicate a subject's risk of receiving a diagnosis of HD (low, medium, and high risk). Baseline age was dichotomized into two groups (young versus old) based on the median split. A total of seven covariates was included (high risk, medium risk, baseline age group, sex, and baseline clinical measures: total functioning capacity [TFC], total motor score [TMS], symbol digit modalities test [SDMT]).

Potentially there are 462 edges (involving 4,180 parameters) for the subcortical gray matter volumetric atrophy network and 4556 potential edges (involving 41,072 parameters) for the cortical gray matter thickness network. The proposed DAG-MM identified 5 connections (Supplementary Material Section 3 Table S1) from the subcortical network (e.g., left thalamus to right accumbens, and right pallidum to left putamen), which suggests that most subcortical ROI atrophy rates do not depend on other ROIs. In contrast, a denser network was identified for the cortical thickness network, with 58 connections identified (Supplementary Material Section 3 Table S2), suggesting that cortical thinning acts in a more concerted fashion, consistent with the neuroimaging literature on cortical networks in HD [38]. PenPC identified a very dense network for both subcortical volumes (92 edges) and cortical thickness networks (480 edges).

Due to its non-sparseness and difficulty in interpretation, we omit results from PenPC and report DAG-MM in the subsequent presentation.

ROIs were further organized into modules related to HD pathology as in [29] for better interpretation. We present these results in Figure 3, where the modular-wise strength of the connection was computed as the total strength of connections within a module (summation of β_{jk} between all pairs of connected nodes (j, k) in the same module) or between two modules (summation of β_{jk} between all pairs of connected nodes (j, k) for j in one module and k in the other). Figure 3 shows that the two strongest connections in the average modular graph (with covariates fixed at the sample averages) are the inter-hemispheric links between the left and right temporal regions and between the left and right motor-occipital-parietal regions. For within-modular connection, the right side motor-occipital-parietal module has the strongest strength.

We also examined differences between the networks for high-risk group versus low-risk group, and medium-risk versus low-risk (other covariates fixed at the sample average). For the high- versus low-risk group comparison (Figure 3), the largest difference is in the inter-hemispheric temporal regions. Most within-module and between-module connections show a loss of strength in the high-risk group. For example, a large loss of intra-modular connections within the right motor-occipital-parietal, right temporal, left fronto-cingulate is seen. A loss of between-module connections is observed between the left and right motor-occipital-parietal regions and between the left fronto-cingulate and left and right temporal regions. A minor gain of connection is seen within and between a few modules. A similar trend with a milder effect is present for most connections when contrasting medium-risk and low-risk groups. When comparing older adults with younger adults, most connections show a loss of strength in the older group (Figure 3). The largest loss in the intra-modular connections is in the right temporal region. A loss of between-module connections occurs between the left and right fronto-cingulate regions, between the left fronto-cingulate and left and right temporal regions, between the left fronto-cingulate and left motor-occipital-parietal regions, and between the right fronto-cingulate and right temporal regions.

In Supplementary Material Section 3 Figure S1, we show the node-wise DAGs and the difference of the estimated network between groups with different baseline risk of HD diagnosis. At the nodal level, we see a loss of connection in the high-risk group and older group in a large number of links. The connection with the largest difference is L.caudalmiddlefrontal $\Rightarrow$ L.rostralmiddlefrontal (based on L2-norm). When effects are aggregated from nodes within modules, group differences are more apparent (Figure 3). The strength of connections between nodes is summarized in Supplementary Material Section 3 Table S1 and Table S2. Among all covariates, the three covariates with the largest effects aggregated across all connections (based on L_2-norm) are high-low risk group contrast, medium-low risk contrast, and older-younger adult contrast. Substantial heterogeneity of connections due to latent factors not captured by covariates is observed for almost all links (represented by σ^2 in Supplementary Material Section 3 Table S1 and Supplementary Material Section 3 Table S2). We show the variation of the heterogeneous effects (standard deviation: σ_{jk}) of connections in Supplementary Material Section 3 Figure S2. The connection with the highest variation is L.caudalmiddlefrontal $\Rightarrow$ L.posteriorcingulate. This analysis demonstrates substantial heterogeneity of the brain dependence networks among individuals.

4.2.1 A Validation Study Using Independent Samples

We sought to corroborate our estimated cortical gray matter network using white matter cortical connectivity network data obtained from an independent study, TRACK-ON [36, 29]. TRACK-ON is a longitudinal study of premanifest HD, with 84 subjects and a median follow-up length of 1.89 years. White matter structural connection network was constructed from diffusion tensor imaging (DTI) technology,

and connection strengths between pairs of nodes were computed by probabilistic tractography. A similar algorithm as PREDICT-HD was used to define regions of interest, and the same method that was used to partition nodes into HD pathology also informed modules [29]. Detailed information on the study design and data pre-processing can be found in [29]. With longitudinal DTI measurements available, a linear mixed-effects model was used to compute the rate of change in connections between nodes and their p-values. Baseline connection, CAG, age, gender, motor score, SDMT, and TFC were included as covariates. Nodes were classified into modules by the same method as the structural MRI network. Inter-modular connection was defined as present if at least c pairs ($c = 1, 2$) of nodes (each node resides in the module being considered) were connected after the false discovery rate (FDR) correction ($q < 0.1$). Presence of intra-modular connection was defined similarly based on the number of pairs of nodes connected (with $q < 0.1$) within a module. In total, 30 white matter atrophy connections were identified after the FDR correction.

Supplementary Material Section 3 Table S3 summarizes the module-wise white matter structural connectivity network estimated from the DTI technology. The average modular gray matter atrophy network and the white matter connection network both indicate a strong intra-modular connection in the right-side motor-occipital-parietal region and a strong interhemispheric connection in the left and right motor-occipital-parietal regions, whereas a weak connection (or no connection for the white matter network) was present in the left side of the same module. For some of the other four modules, the intra-modular connection strength for gray matter and white matter appears to be complementary: a stronger link in the former corresponds to a weaker link in the latter. For example, connections between the right temporal and right motor-occipital-parietal regions and between the left and right temporal regions show a moderate to strong dependence in the gray matter network, but are absent in the white matter network. The link between the right-fronto-cingulate region is strong in the white matter network, but weak in the gray matter network. These observations might suggest a mechanism that constrains the total modular connections in the gray matter and white matter networks; thus, a strong connection in one correlates with a weak connection in the other.

We evaluated the consistency of the gray matter cortical network (obtained by DAG-MM2 statistical modeling) with the white matter cortical structural connectivity network (directly measured by DTI technology). The overall operational characteristics of the gray matter network are reported in Table 4, treating the white matter network as the reference since white matter connections were directly measured by DTI. Due to a potential complementary effect on the total number of connections between and within modules, the number of connections in the gray matter and white matter networks is negatively correlated. Thus, we computed the sensitivity as $P(L \leq l | C \geq c)$, where L denotes the number of links in the gray matter network, and C denotes the number of links in the white matter network. We fixed the connectivity threshold of the white matter network at $c = 1$ or $c = 2$, and we evaluated the overall consistency of the gray matter network across all levels of threshold l by computing the AUC across l. The AUC is 0.80 (95%CI: 0.61, 0.99) at $c = 1$ and 0.75 (95%CI: 0.48, 1.00) at $c = 2$. Using a higher threshold c increases sensitivity, but with a slightly decreased specificity and a slightly lower AUC. These results show that at the modular level, the gray matter cortical atrophy network estimated by DAG-MM has adequate consistency with the white matter structural connectivity network.

5 DISCUSSION

In this article, we propose a statistical framwork for estimating DAGs with mixed effects in multi-dimensional settings, referred to as DAG-MM. The framework captures covariate-dependent causal effects,

along with residual effect modification, by building a series of mSEMs. Our framework is a two-stage approach, which first obtains a sparse initial skeleton (undirected graph) and then searches for DAG through a solution path within the selected skeleton and an easily implemented DAG checking procedure. The DAG-MM method is computationally efficient and has shown satisfactory performance, especially for edge selection and orientation, in both simulation studies and real-data applications. The advantage arises when taking into account the covariate-dependent structure and residual heterogeneous effects through the use of random variables. Specifically, the joint distribution of the nodes in model (2) are non-Gaussian due to these random effects and their multiplicative form with the other nodes. This asymmetry in the joint distribution permits the identification of causal relationships from observational data, which we formally prove in Theorem 1. We note that the edge orientation is more accurate than PC and its derivatives, which assume a symmetric joint distribution. For computation, the regularized likelihood-based approach identifies a sparse skeleton in an efficient fashion.

In the analyses of brain atrophy dependence network in patients with HD, some modules of the gray matter network estimated from the PREDICT-HD study share similarity with the white matter connectivity network estimated from an independent study. For some other modules, the results suggest a complementary mechanism that constrains the total modular-wise connections in gray and white matter networks. In the second application, the protein signaling network constructed from DAG-MM with observational data and invariance causal prediction (ICP) with interventional data [17] is highly similar. The latter approach assumes causal relationships remain invariant under interventions that do not directly target a cause. This similarity suggests that the random effects in mSEM may serve as a random perturbation of the node distribution. Under the invariance assumption, the true causal effects are stable under such perturbation, and thus, DAG-MM generates similar results as ICP, but with only observational data.

The network structure among nodes can be further parametrized to incorporate prior information about the causal effects. For example, the knowledge on pathways in the gene regulatory network available in public databases or discoverable in published literature can be included by removing or adding the edge between nodes j and k or by restricting the edge direction from j to k. Model (3) can handle this structure by specifying some values of $\boldsymbol{\beta}_{jk}$ or/and σ^2_{jk} as zero. Another extension is to analyze temporal data $\boldsymbol{M}_i$ with two time points t_0 and t_1, where the desirable temporal ordering corresponds to removing all edges from $\boldsymbol{M}_i(t_1)$ to $\boldsymbol{M}_i(t_0)$ and modeling the effect from $\boldsymbol{M}_i(t_0)$ to $\boldsymbol{M}_i(t_1)$.

DAG-MM can be extended to handle multiple types of data, including neuroimaging, protein, and other biomarker measures of different scales, in a regression framework by choosing the appropriate regression for each data type. When the dimension of covariates $\boldsymbol{X}$ is high (e.g., large number of genomic measures), feature selection can be imposed on $\boldsymbol{\beta}$ in order to choose important covariates. Here, we use mSEMs to estimate network connectivity, but we did not differentiate the fixed effects from the random effects. Our main algorithm is a backward selection method and does not allow edge addition. To overcome this issue, one may start DAG-MM from multiple skeletons, which is an approach that provides a more stable edge selection. Lastly, other interesting extensions include direct modeling of a dynamic network among $\boldsymbol{M}(t)$ to allow for time-varying network structure and associate network connections with clinical outcomes.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

Wang and Zeng designed and oversaw the study. Li and Xie developed algorithm, implemented the study, and carried out the statistical analysis. McColgan, Tabrizi, and Scahill provided DTI data, discussed results, and gave the biological insights. All authors participated in writing the manuscript.

ACKNOWLEDGEMENTS

This research is support by U.S. NIH grants NS082062 and NS073671 and Wellcome Trust Grant 103437/Z/13/Z. The authors wish to thank Raymund A C Roos, Alexandra Durr, Blair R Leavitt and the Track, PREDICT Investigators for their contribution to collect TRACK-HD and PREDICT-HD data, and NINDS dbGap data repository (accession number phs000222.v3).

SUPPLEMENTARY MATERIALS

Algorithm 2, Proof of Theorem 1, Supplementary Material Tables, and Figures referenced in Sections (2.2, 2.3, 4) are available in the Supplement Materials.

REFERENCES

[1] Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308** (2005) 523–529.

[2] Ud-Dean SM, Heise S, Klamt S, Gunawan R. Trace+: Ensemble inference of gene regulatory networks from transcriptional expression profiles of gene knock-out experiments. *BMC Bioinformatics* **17** (2016) 252.

[3] Friston KJ. Functional and effective connectivity: a review. *Brain Connectivity* **1** (2011) 13–36.

[4] Pearl J. Causality: Models, Reasoning, and Inference. *New York, NY* (2009).

[5] Robinson RW. Counting labeled acyclic digraphs. Harary F, editor, *New Directions in the Theory of Graphs: Proc. Third Ann Arbor Conference on Graph Theory* (New York: Academic Press) (1971), 239–273.

[6] Spirtes P, Glymour CN, Scheines R. *Causation, Prediction, and Search* (MIT press) (2000).

[7] Heckerman D, Geiger D, Chickering DM. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20** (1995) 197–243.

[8] Kalisch M, Bühlmann P. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research* **8** (2007) 613–636.

[9] Colombo D, Maathuis MH. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research* **15** (2014) 3741–3782.

[10] Ha MJ, Sun W, Xie J. Penpc: A two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs. *Biometrics* **72** (2016) 146–155. doi:10.1111/biom.12415.

[11] Shojaie A, Michailidis G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97** (2010) 519–538. doi:10.1093/biomet/asq038.

[12] Yuan Y, Shen X, Pan W. Maximum likelihood estimation over directed acyclic Gaussian graphs. *Statistical Analysis and Data Mining* **5** (2012) 523–530.

[13] Aragam B, Zhou Q. Concave penalized estimation of sparse Gaussian Bayesian networks. *Journal of Machine Learning Research* **16** (2015) 2273–2328.

[14] Han SW, Chen G, Cheon MS, Zhong H. Estimation of directed acyclic graphs through two-stage adaptive Lasso for gene network inference. *Journal of the American Statistical Association* **111** (2016) 1004–1019.

[15] Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* **7** (2006) 2003–2030.

[16] Luo R, Zhao H. Bayesian hierarchical modeling for signaling pathway inference from single cell interventional data. *The Annals of Applied Statistics* **5** (2011) 725.

[17] Meinshausen N, Hauser A, Mooij JM, Peters J, Versteeg P, Bühlmann P. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences* **113** (2016) 7361–7368.

[18] Smith SM, Miller KL, Salimi-Khorshidi G, Webster M, Beckmann CF, Nichols TE, et al. Network modelling methods for fMRI. *Neuroimage* **54** (2011) 875–891.

[19] Brown JA, Terashima KH, Burggren AC, Ercoli LM, Miller KJ, Small GW, et al. Brain network local interconnectivity loss in aging APOE-4 allele carriers. *Proceedings of the National Academy of Sciences* **108** (2011) 20760–20765.

[20] Langfelder P, Cantle JP, Chatzopoulou D, Wang N, Gao F, Al-Ramahi I, et al. Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. *Nature Neuroscience* **19** (2016) 623–633.

[21] Bohlken MM, Brouwer RM, Mandl RC, Van den Heuvel MP, Hedman AM, De Hert M, et al. Structural brain connectivity as a genetic marker for Schizophrenia. *JAMA Psychiatry* **73** (2016) 11–19.

[22] Fleeson W, Furr RM, Arnold EM. An agenda for symptom-based research. *Behavioral and Brain Sciences* **33** (2010) 157–157.

[23] Yin J, Li H. A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics* **5** (2011) 2630.

[24] Cai TT, Li H, Liu W, Xie J, et al. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100** (2013) 139–156.

[25] Guo J, Cheng J, Levina E, Michailidis G, Zhu J. Estimating heterogeneous graphical models for discrete data with an application to roll call voting. *The Annals of Applied Statistics* **9** (2015) 821.

[26] Cheng J, Levina E, Wang P, Zhu J. A sparse Ising model with covariates. *Biometrics* **70** (2014) 943–953.

[27] Shimizu S, Bollen K. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions. *The Journal of Machine Learning Research* **15** (2014) 2629–2652.

[28] Paulsen JS, Long JD, Johnson HJ, Aylward EH, Ross CA, Williams JK, et al. Clinical and biomarker changes in premanifest Huntington disease show trial feasibility: a decade of the PREDICT-HD study. *Frontiers in Aging Neuroscience* **6** (2014) 78.

[29] McColgan P, Seunarine KK, Gregory S, Razi A, Papoutsi M, Long JD, et al. Topological length of white matter connections predicts their rate of atrophy in premanifest Huntington's disease. *JCI Insight* **2** (2017).

[30] Woodward J. *Making things happen: A theory of causal explanation* (Oxford university press) (2005).

[31] Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *The annals of statistics* (2006) 1436–1462.

[32] Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P, et al. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* **47** (2012) 1–26.

[33] Mooij J, Heskes T. Cyclic causal discovery from continuous equilibrium data. *arXiv preprint arXiv:1309.6849* (2013).

[34] O'Donovan MC. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72** (1993) 971–983.

[35] Ross CA, Aylward EH, Wild EJ, Langbehn DR, Long JD, Warner JH, et al. Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nature Reviews Neurology* **10** (2014) 204–216.

[36] Klöppel S, Gregory S, Scheller E, Minkova L, Razi A, Durr A, et al. Compensation in preclinical Huntington's disease: Evidence from the TRACK-ON HD study. *EBioMedicine* **2** (2015) 1420–1429.

[37] Zhang Y, Long JD, Mills JA, Warner JH, Lu W, Paulsen JS. Indexing disease progression at study entry with individuals at-risk for Huntington disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **156** (2011) 751–763.

[38] He Y, Chen Z, Evans A. Structural insights into aberrant topological patterns of large-scale cortical networks in Alzheimer's disease. *The Journal of Neuroscience* **28** (2008) 4756–4766.

Table 1. Simulation results of graph edge selection performance (TP: average number of true positive edges; FP: average number of false positive edges; FN: average number of false negative edges) using the initial DAG selection, DAG-MM procedure, PC algorithm, and penPC algorithm for various sample sizes n and numbers of features p.

		Initial graph			DAG-MM			PC			penPC		
		$p = 20$	$p = 50$	$p = 100$	$p = 20$	$p = 50$	$p = 100$	$p = 20$	$p = 50$	$p = 100$	$p = 20$	$p = 50$	$p = 100$
		Setting 1 - fixed effects only											
TP	$n = 200$	12.0	12.0	12.0	12.0	12.0	12.0	1.8	1.4	1.2	2.7	2.5	2.4
	$n = 500$	12.0	12.0	12.0	12.0	12.0	12.0	2.0	1.7	1.4	3.0	3.0	3.0
	$n = 1000$	12.0	12.0	12.0	12.0	12.0	12.0	2.1	1.6	1.4	3.2	2.9	3.0
FP	$n = 200$	33.1	77.9	162.8	0.2	0.0	0.1	9.3	18.7	48.6	18.1	26.5	46.0
	$n = 500$	32.9	69.7	69.2	0.0	0.0	0.0	9.8	19.9	51.2	19.6	28.1	42.8
	$n = 1000$	25.2	44.1	74.1	0.0	0.0	0.0	9.9	20.7	53.2	19.7	28.8	40.7
FN	$n = 200$	0.0	0.0	0.0	0.0	0.0	0.0	10.2	10.6	10.8	9.3	9.5	9.6
	$n = 500$	0.0	0.0	0.0	0.0	0.0	0.0	10.1	10.4	10.6	9.0	9.0	9.0
	$n = 1000$	0.0	0.0	0.0	0.0	0.0	0.0	9.9	10.4	10.6	8.8	9.1	9.0
		Setting 2 - random effects only											
TP	$n = 200$	11.5	11.3	10.7	6.9	5.3	3.7	0.6	0.4	0.2	0.6	0.4	0.3
	$n = 500$	12.0	11.9	11.9	10.4	10.3	10.0	0.5	0.3	0.2	0.6	0.3	0.2
	$n = 1000$	12.0	12.0	12.0	11.3	11.3	11.3	0.3	0.2	0.1	0.4	0.3	0.2
FP	$n = 200$	57.2	130.5	215.7	1.1	2.2	3.1	3.4	15.4	46.1	4.0	14.5	31.5
	$n = 500$	56.3	96.4	167.0	0.3	0.8	1.4	3.5	15.7	49.7	3.9	13.3	25.9
	$n = 1000$	49.8	115.6	109.3	0.0	0.2	0.2	3.6	16.9	51.7	4.3	14.7	25.0
FN	$n = 200$	0.5	0.8	1.3	5.1	6.8	8.3	11.5	11.7	11.8	11.4	11.6	11.7
	$n = 500$	0.0	0.1	0.1	1.6	1.7	2.0	11.5	11.8	11.8	11.4	11.7	11.8
	$n = 1000$	0.0	0.0	0.0	0.7	0.7	0.7	11.7	11.8	11.9	11.6	11.7	11.8
		Setting 3 - mixed effects 1											
TP	$n = 200$	12.0	12.0	12.0	12.0	12.0	11.9	1.7	1.4	1.1	2.6	2.5	2.4
	$n = 500$	12.0	12.0	12.0	12.0	12.0	12.0	1.9	1.6	1.4	3.0	2.9	2.8
	$n = 1000$	12.0	12.0	12.0	12.0	12.0	12.0	2.1	2.0	1.4	3.1	3.2	3.0
FP	$n = 200$	114.8	228.7	362.5	0.0	0.2	0.7	8.9	18.4	47.5	17.1	26.8	44.1
	$n = 500$	109.6	266.8	431.3	0.0	0.0	0.0	9.2	19.7	51.2	17.9	27.9	41.6
	$n = 1000$	138.9	185.8	326.4	0.0	0.0	0.0	9.4	20.2	53.8	18.5	29.1	40.5
FN	$n = 200$	0.0	0.0	0.0	0.0	0.0	0.1	10.3	10.6	10.9	9.4	9.5	9.6
	$n = 500$	0.0	0.0	0.0	0.0	0.0	0.0	10.1	10.5	10.7	9.0	9.2	9.2
	$n = 1000$	0.0	0.0	0.0	0.0	0.0	0.0	9.9	10.0	10.6	8.9	8.8	9.0
		Setting 4 - mixed effects 2											
TP	$n = 200$	11.7	11.4	11.2	10.7	10.3	9.8	0.6	0.6	0.4	1.1	1.1	0.9
	$n = 500$	11.9	11.8	11.7	11.6	11.3	11.1	0.5	0.4	0.4	1.0	1.0	0.9
	$n = 1000$	12.0	12.0	11.9	11.6	11.6	11.5	0.6	0.5	0.5	1.1	1.2	1.1
FP	$n = 200$	81.7	121.4	237.9	0.6	2.3	5.5	8.0	18.3	47.8	12.2	22.2	41.3
	$n = 500$	56.1	155.5	258.0	0.1	0.4	0.8	8.3	18.9	50.8	12.7	21.6	36.4
	$n = 1000$	92.1	96.0	161.6	0.0	0.1	0.1	8.2	19.7	52.4	13.5	22.2	33.8
FN	$n = 200$	0.3	0.6	0.8	1.3	1.7	2.2	11.4	11.4	11.7	10.9	10.9	11.1
	$n = 500$	0.1	0.2	0.3	0.4	0.7	0.9	11.5	11.6	11.6	11.0	11.0	11.1
	$n = 1000$	0.0	0.0	0.1	0.4	0.4	0.5	11.4	11.5	11.6	10.9	10.8	10.9
		Setting 5 - homogeneous (constant effect)											
TP	$n = 200$	12.0	12.0	12.0	11.9	12.0	11.9	11.2	10.8	10.3	11.9	11.9	11.8
	$n = 500$	12.0	12.0	12.0	12.0	12.0	12.0	11.8	11.5	11.0	12.0	12.0	12.0
	$n = 1000$	12.0	12.0	12.0	12.0	12.0	12.0	11.9	11.6	11.3	12.0	12.0	12.0
FP	$n = 200$	37.3	24.6	68.9	0.3	1.3	5.0	6.8	14.7	42.6	12.4	19.3	36.9
	$n = 500$	23.2	49.6	22.2	0.2	0.3	0.8	5.7	14.5	43.7	12.6	19.3	33.9
	$n = 1000$	36.3	22.6	31.8	0.0	0.7	3.6	5.3	14.5	45.3	12.5	19.2	30.4
FN	$n = 200$	0.0	0.0	0.0	0.1	0.0	0.1	0.8	1.3	1.7	0.1	0.2	0.2
	$n = 500$	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.5	1.0	0.0	0.0	0.0
	$n = 1000$	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.8	0.0	0.0	0.0

[29] reported a modular white matter network obtained by comparing connectivity in patients with HD and healthy controls and applying FDR adjustment (Figure 2 in [29]). When we compare our results to theirs, we see similarity, in terms of connections between the left and right temporal regions and between the left and right motor-occipital-parietal regions.

Table 2. Simulation results of root sum-squared (RSS) error of parameters for the connection strength estimated by DAG-MM under various sample sizes n and numbers of features p.

	$\boldsymbol{\beta}$			σ^2			σ_ε^2		
	$p = 20$	$p = 50$	$p = 100$	$p = 20$	$p = 50$	$p = 100$	$p = 20$	$p = 50$	$p = 100$
Setting 1 - fixed effects only									
$n = 200$	0.312	0.305	0.358	0.077	0.087	0.102	0.388	0.576	0.832
$n = 500$	0.184	0.188	0.189	0.045	0.045	0.049	0.232	0.360	0.505
$n = 1000$	0.130	0.131	0.130	0.037	0.034	0.032	0.162	0.253	0.357
Setting 2 - random effects only									
$n = 200$	0.527	0.501	0.479	1.347	1.708	1.977	1.065	1.307	1.601
$n = 500$	0.353	0.369	0.386	0.739	0.815	0.943	0.523	0.631	0.714
$n = 1000$	0.254	0.270	0.270	0.461	0.496	0.485	0.294	0.400	0.458
Setting 3 - mixed effects 1									
$n = 200$	0.606	0.667	1.063	0.635	0.857	4.058	0.523	0.693	0.962
$n = 500$	0.367	0.362	0.363	0.391	0.355	0.348	0.347	0.433	0.564
$n = 1000$	0.254	0.261	0.262	0.259	0.264	0.248	0.227	0.300	0.387
Setting 4 - mixed effects 2									
$n = 200$	0.559	0.624	0.797	0.963	1.336	2.359	0.649	0.888	1.243
$n = 500$	0.333	0.345	0.365	0.478	0.593	0.783	0.375	0.462	0.627
$n = 1000$	0.234	0.229	0.233	0.351	0.389	0.427	0.252	0.333	0.436
Setting 5 - homogeneous (constant effect)									
$n = 200$	0.358	0.348	0.447	0.125	0.303	0.837	0.479	0.647	0.936
$n = 500$	0.157	0.166	0.157	0.073	0.112	0.205	0.251	0.389	0.530
$n = 1000$	0.098	0.143	0.226	0.048	0.054	0.093	0.172	0.269	0.363

Table 3. Comparison with previously identified causal relationships. Total number of edges previously identified in the literature is 34. ICP [17] used both observational and interventional data. Proposed DAG-MM1 (fixed effects only) and DAG-MM2 (mixed effects) used only observational data.

Reported†	PC	ICP	DAG-MM1	DAG-MM2
Yes	8	10	8	9
No	4	5	10	2*

†: whether an edge was previously reported in the literature. *: edges in reverse direction of those reported in the literature.

Table 4. Operating characteristics of cortical gray matter atrophy dependence network evaluated against the white matter structural connectivity network treated as the reference.

c	AUC (95%CI)	Sensitivity	Specificity	PPV
1	0.80 (0.61, 0.99)	0.57	1.00	1.00
2	0.75 (0.48, 1.00)	0.75	0.85	0.75

Figure 1. Frequency of edges selected in 100 simulations. Edge width is proportional to the number of times an edge is identified in simulations. Black: true positive edges; Blue: false positive edges; Red: false negative edges (true edges that were never selected).

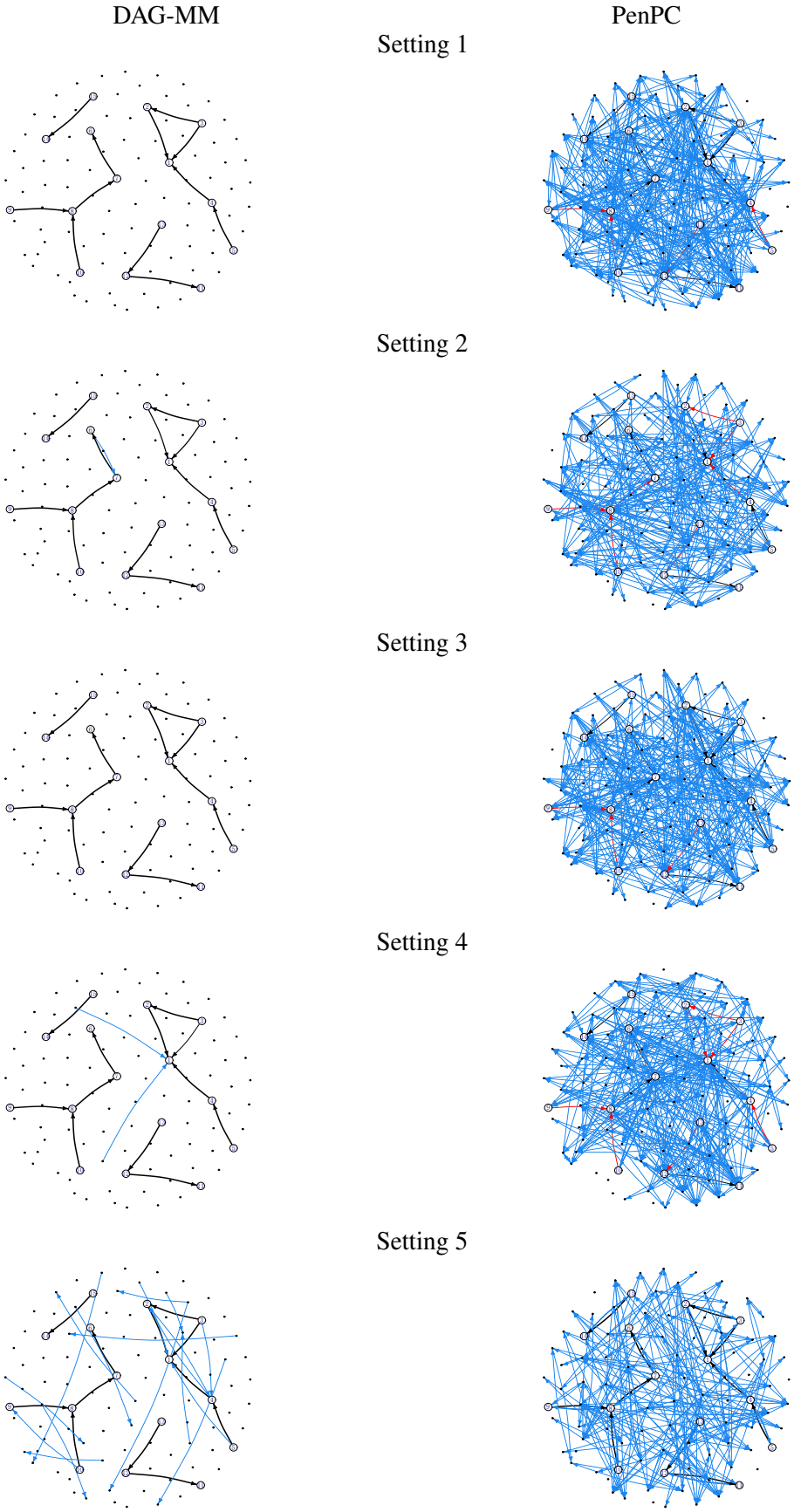


Figure 2. Estimated protein signaling network. Black: edges identified by DAG-MM2 and also reported previously; Blue: edges are identified by DAG-MM2 but not reported previously; Gray: edges previously reported edges but not identified by DAG-MM2.

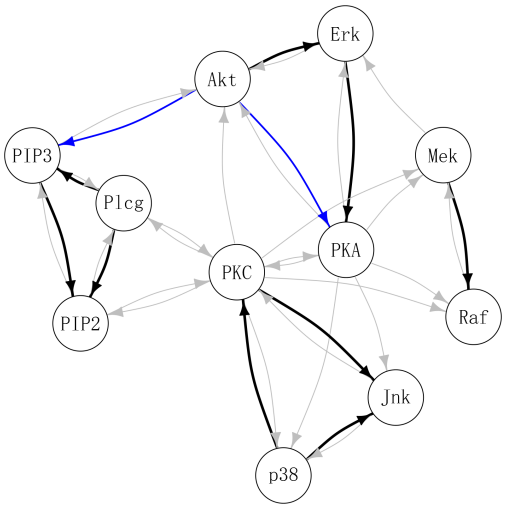
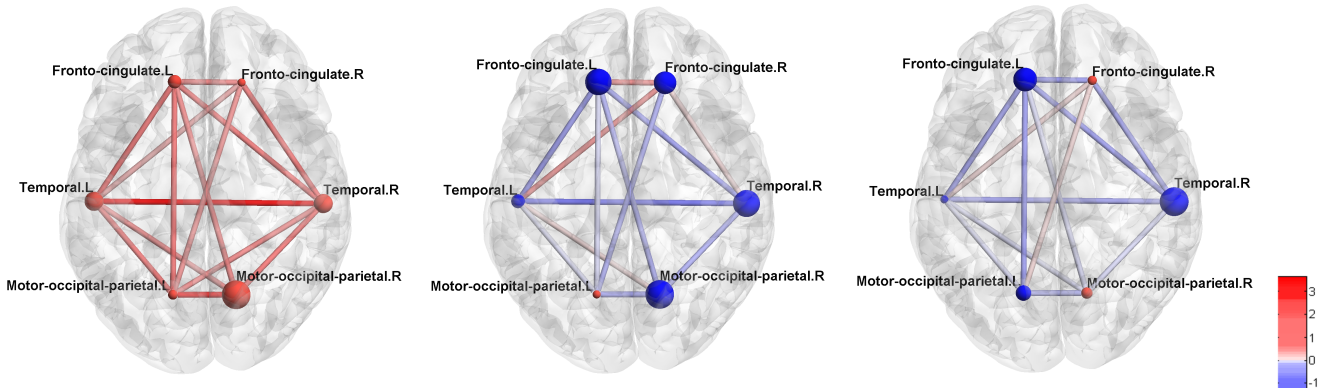


Figure 3. Estimated cortical thickness atrophy dependence network (organized into modules). The node size is proportional to the intra-modular connection strength (edge effects) and scaled within each subfigure. Red nodes: positive effects. Blue nodes: negative effects.



Population Average High versus Low Risk Older versus Younger