SUPPLEMENTARY INFORMATION for article

Mixed Graphical Models for Integrative Causal Analysis with Application to Chronic Lung Disease Diagnosis and Prognosis

Andrew J Sedgewick[1,2], Kristina Buschur[1,2], Ivy Shi[3], Joseph D. Ramsey[4], Vineet K. Raghu[5], Dimitris V. Manatakis[1], Yingze Zhang[6], Jessica Bon[6], Divay Chandra[6], Chad Karoleski[6], Frank C. Sciurba[6], Peter Spirtes[4], Clark Glymour[4], Panayiotis V. Benos[2,3,*]

Running Title: Causal Learning over Mixed Data Types

Affiliations:

[1]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA.

[2]Joint Carnegie Mellon University-University of Pittsburgh PhD Program in Computational Biology, Pittsburgh, Pennsylvania, USA.

[3]Department of Bioengineering, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.

[4]Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

[5]University of Pittsburgh, Department of Computer Science, Pittsburgh, Pennsylvania, USA

[6]University of Pittsburgh, Department of Medicine, Division of Pulmonary, Allergy and Critical Care Medicine, Pittsburgh, Pennsylvania, USA

SUPPLEMENTARY METHODS

Simulated data

We simulated data from low- and high-dimensional networks with 50 different directed graph topologies each, randomly selected using TETRAD (version 5.3.0, https://github.com/cmu-phil/tetrad), a Java package for causal modeling that uses linear or non-linear structural equation models (SEMs) to generate data from network distributions. Each of the 50 *low dimensional datasets* consisted of 500 samples, drawn from network structures of 50 variables: 25 Gaussian and 25 3-level categorical. Each of the 20 *high dimensional datasets* consisted of 100 samples drawn from network structures of 200 variables: 100 Gaussian and 100 3-level categorical. The structures are sampled uniformly from the space of all directed acyclic graphs (DAGs) with maximum node degree of 10 and a maximum of average node degree of 2. The low dimensional dataset tests the efficiency of the algorithms when the study is well powered; while the high dimensional dataset tests the efficiency of the algorithms when the number of samples is small compared to the number of variables, a condition that is frequently present in many biomedical applications.

The relationships between variables in all datasets are set up in a similar fashion to Lee and Hastie [1]. Here, for an edge $X \rightarrow Y$ we refer to X as the parent and Y as the child. Parents of the Gaussian variables contribute linearly to the mean of each child; the value of continuous parents is multiplied by an edge parameter and the value of discrete parents is associated with an edge parameter where a separate edge parameter is specified for each category of the discrete variable. Parents of discrete variables contribute log-linearly to the probabilities of each category, with separate parameters for each category of the child variable. With this set up, each edge connecting two continuous variables (*cc*) depends on 1 edge parameter, each edge connecting a continuous and a discrete variable (*cd*) depends on a vector of 3 parameters and edges connecting two discrete variables (*dd*) depend on a 3 by 3 matrix of 9 edge parameters. In order to ensure identifiability, the *cd* parameter vector, and the rows of the *dd* parameter matrix are constrained to sum to 0 leaving these edges with 2 and 6 degrees of freedom, respectively. Edge weights were drawn uniformly from the union of the regions [-1.5, -1] and [1, 1.5]. For *cc* edges the parameter is equal to the weight; for *cd* edge parameters we draw a vector three values uniformly from [0,1] and shift and scale the values so they sum to zero and the largest parameter is equal to the edge weight; for *dd* edge parameters we draw one vector of three values as with *cd* edges and set the rows of the matrix as the three permutations of this vector.

In the continuous case, zero-mean, Gaussian error terms with standard deviation uniformly drawn from the interval [1, 2], are drawn for every variable and then the variable means are resolved. In DAGs this resolution is trivial as we can start from root nodes with no parents and propagate downwards. To make this process accommodate categorical distributions, we use a uniform draw over [0, 1] as an error term for each discrete variable and this term is used to determine the value of the variable given the probabilities of each category. In generating simulated models, these probabilities that are then updated in the same way as are the means of the continuous variables. This approach ensures convergence of each discrete variable for each sample.

Biological and clinical datasets

In this paper, we used two publicly available datasets (TCGA, LGRC) that combine omics and clinical variables to show how CausalMGM can recapitulate existing knowledge. We also used a third dataset (SCCOR) to identify clinical variables that are directly linked to longitudinal lung function decline of COPD patients. Below, we describe these datasets in detail.

The **breast cancer** dataset was obtained from The Cancer Genome Atlas (TCGA)[19]. These data (BRCA, n=448 samples) included RNA-seq data normalized with RSEM for 20530 transcripts, and clinical variables (PAM50 subtypes, Progesterone, Estrogen and HER2 receptor status, and tumor and node stage codes). We used only the 500 genes with the most variant expression across all samples.

The *Lung Genomics Research Consortium* (**LGRC**) contains multiple genomic datasets and clinical variables for two chronic lung diseases: chronic obstructive pulmonary disease (COPD) and interstitial lung disease (ILD). We used two data types from LGRC: gene expression profiles (15,261 probes) and clinical data for 457 patients (COPD N=215; ILD N=242). To expedite the execution time and avoid sample size problems, we only used the 530 most variant expression probes and 8 clinical variables: age, height, weight, forced expiratory volume in one second (FEV_1), forced vital capacity (FVC), gender, cigarette history, and diagnosis (COPD or ILD). Age, height, weight and the spirometry variables (FEV_1 and FVC) were divided into tertiles.

The *Pittsburgh Specialized Center of Clinically Oriented Research* (**SCCOR**) cohort consists of 747 subjects that were recruited from a larger less characterized community-based tobacco-exposed cohort, enriched for subjects with visual emphysema; 385 subjects returned for similar 2-year follow-up evaluation. Data acquisition included: 1) semi-quantitative visual and quantitative MDCT chest radiograph analyses; 2) pre- and post-bronchodilator lung function testing including spirometry, body plethysmography, impulse oscillometry and diffusing capacity; 3) extensive symptom, demographic, environmental exposure and health outcome data and incremental shuttle exercise testing; 4) blood circulating proteins. The total number of variables measured was 281.

Undirected graph learning

Lee and Hastie parameterize an MGM over p Gaussian variables (denoted by x), and q categorical variables (denoted by y), as a pairwise Markov Random Field [1]. Here we present the formulation of the log-likelihood of this model.

$$log\ p(x, y, \Theta) = \exp\left(\sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \sum_{s=1}^{p}\alpha_s x_s + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j)x_s + \sum_{j=1}^{q}\sum_{r=1}^{q}\phi_{rj}(y_r, y_j)\right) - \log(Z)$$

In this model β_{st} represents the interaction between two continuous variables, x_s and x_t; α_s represents the potential of a continuous node x_s; $\rho_{sj}(y_j)$ is a vector of parameters that correspond to the interaction between the continuous variable x_s and the categorical variable y_j indexed by the levels (i.e. categories) of the variable y_j; and $\phi_{rj}(y_r, y_j)$, is a matrix of parameters indexed by the levels of the categorical variables y_j, and y_r. In the continuous only case, this model reduces to a multivariate Gaussian model where the β_{st} parameters are entries in the precision matrix. In the categorical only case, this model is

the popular pairwise Markov random field with potentials given $\phi_{rj}(y_r, y_j)$; and it could parameterize an Ising model as in the binary-only case. Thus, the MGM model serves as a generalization of two popular uni-modal models to the multi-modal regime.

In order to avoid the computational expense of calculating the partition function Z of this model, one can optimize the negative log-pseudolikelihood ($\tilde{l}$), which is:

$$\tilde{l}(\Theta|x,y) = -\sum_{s=1}^{p} \log p\left(x_s|x_{\setminus s}, y; \Theta\right) - \sum_{r=1}^{q} \log p(y_r|x, y_{\setminus r};\ \Theta)$$

where Θ is a shorthand for all of the model parameters. To ensure a sparse model, we minimize $\tilde{l}$ respect to three sparsity penalties, λ_{cc}, λ_{cd}, λ_{dd}, depending on the edge type (continuous-continuous, continuous-discrete and discrete-discrete, respectively)

$$\text{minimize}_{\Theta}\ \tilde{l}(\Theta) + \lambda_{cc}\sum_{t<s}|\beta_{st}| + \lambda_{cd}\sum_{s,j}\left\|\rho_{sj}\right\|_2 + \lambda_{dd}\sum_{r<j}\left\|\phi_{rj}\right\|_F$$

The parameter matrices β and ϕ are symmetric, so only half of each matrix is penalized. The choice of a different sparsity penalty for each edge type is one of the improvement we did in the Lee and Hastie algorithm [9]. Lee and Hastie use an accelerated proximal gradient method [24] to solve this optimization problem.

A standard way of handling a categorical variable with L levels is to convert the variable to L-1 indicator variables where the last level is encoded by setting all indicators to zero, this is necessary to ensure the linear independence of variables in the regression problem. Lee and Hastie's MGM learning approach, uses L indicator variables (i.e. the elements of $\rho_{sj}(y_j)$ and $\phi_{rj}(y_r, y_j)$ to improve interpretability of the model, and enforces a group penalty to ensure the indicator coefficients sum to zero.

Description of graph search methods used in this paper

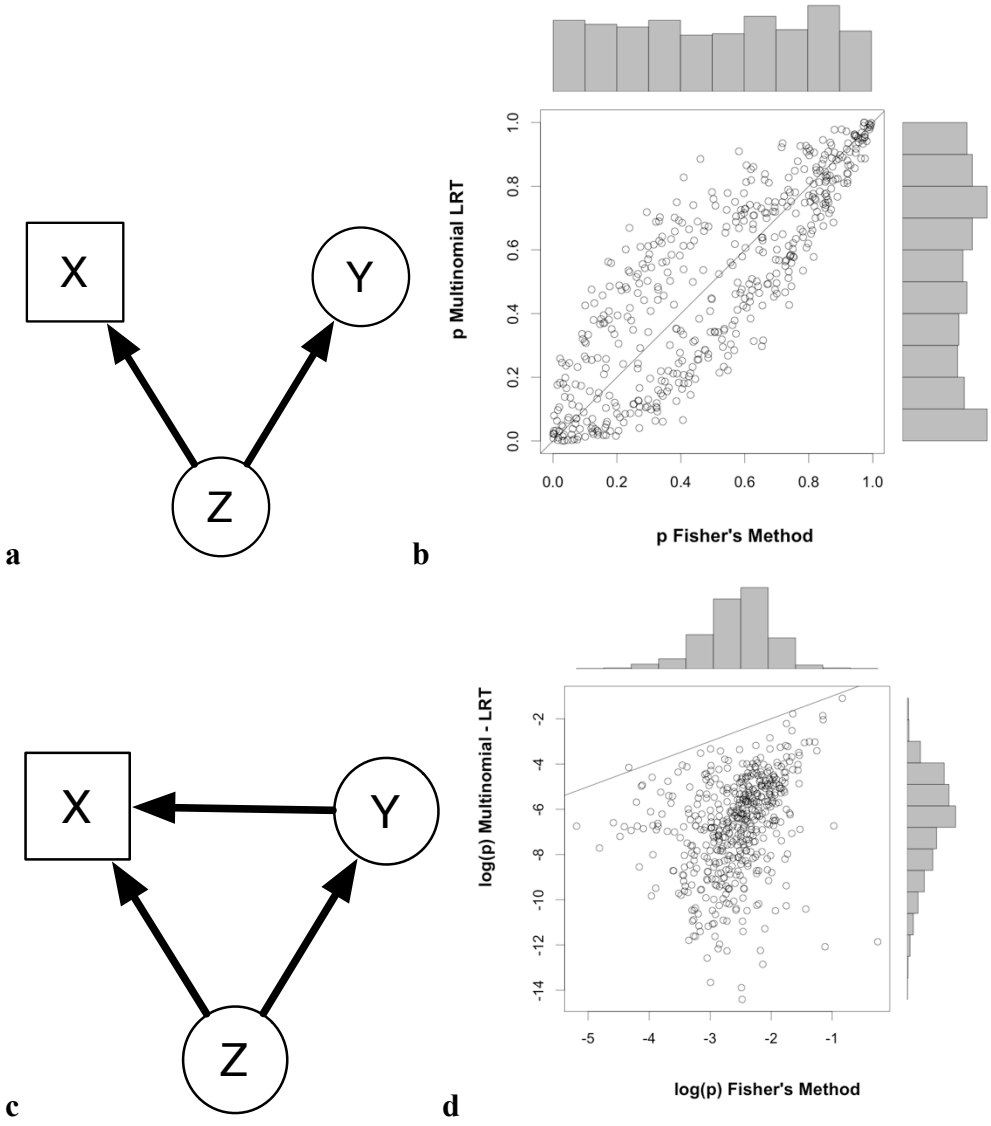
To perform directionality searches on the undirected graph from step-1 of CausalMFM, we used the PC-based methodology, supplemented by our conditional independence tests on mixed data. There are various flavors of PC algorithms. Here we make a brief presentation of the ones we used in this paper.

The **PC** algorithm and its descendants depend on conditional independence decisions that are made by a user-specified test and the α threshold (described below). PC starts with a complete graph and in <u>step 1</u> it sequentially tests all edges for independence given conditioning sets of increasing size. Starting with the empty set, these conditioning sets are subsequently made up of every set (of the given size) of common neighbors of the two nodes incident to the edge being tested. Edges that are found to be conditionally independent are immediately removed and not considered in future tests. When an edge is removed, the conditioning set that lead to the independence decision is saved. <u>Step 2</u> directs edges based on the fact that common neighbors of nodes incident to a removed edge that are not in the conditioning set must be in a v-structure ($X \rightarrow Z \leftarrow Y$). It is possible that two

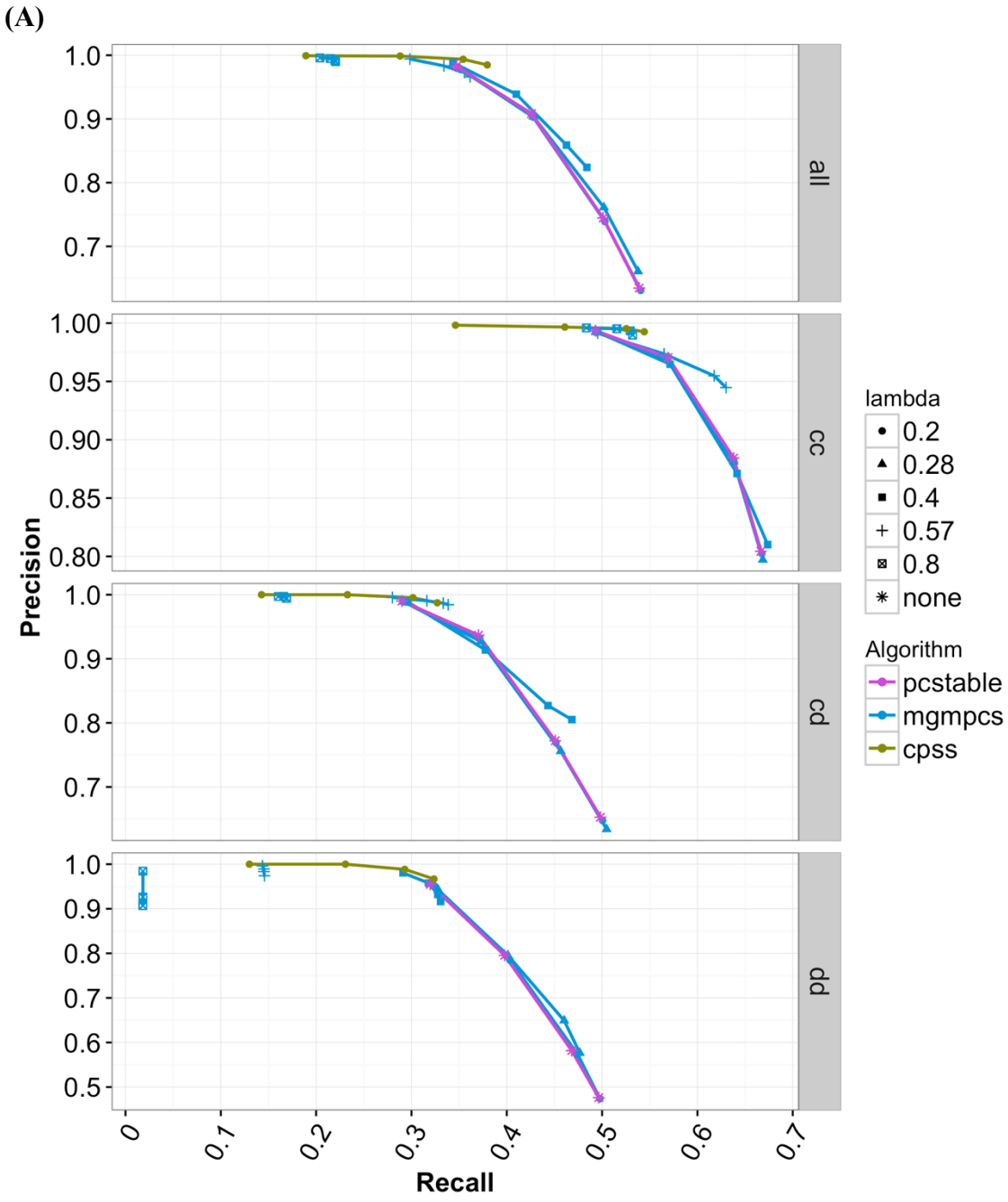
implied v-structures will induce conflicting edge directions. Step 3 further directs edges based on a set of rules that ensure the directions will not induce any cycles or new v-structures [22]. **PC-stable** modifies PC by waiting to update the edge removals in phase 1 until all tests for a given conditioning set size are completed. This leads to an output that is independent of variable ordering and allows for parallelization of the independence tests. In the TETRAD implementation of PC-stable, direction conflicts result in a bi-directed edge: $X \leftrightarrow Y$.

CPC-stable [20] is the variable order independent variant of **Conservative PC** [23] which revises step 2 of PC, described above to perform conditional independence tests with all possible conditioning sets between two nodes, X and Y, that have had an edge between them removed. The conditioning sets are determined by taking subsets of neighbors of the two nodes found in the skeleton graph returned by step 1 of PC. For any node, Z that is incident to both X and Y, the v-structure is only predicted if Z is not in any separating set S such that $X \perp Y \mid S$. Otherwise no direction is predicted from this triplet of nodes. If Z participates in some sets that result in the conditional independence of X and Y and some that result in a conditional dependence, the ambiguity is recorded. Since the change to the PC algorithm takes place after adjacency has been determined, PC and CPC algorithms will produce the same adjacency predictions.

SUPPLEMENTARY FIGS

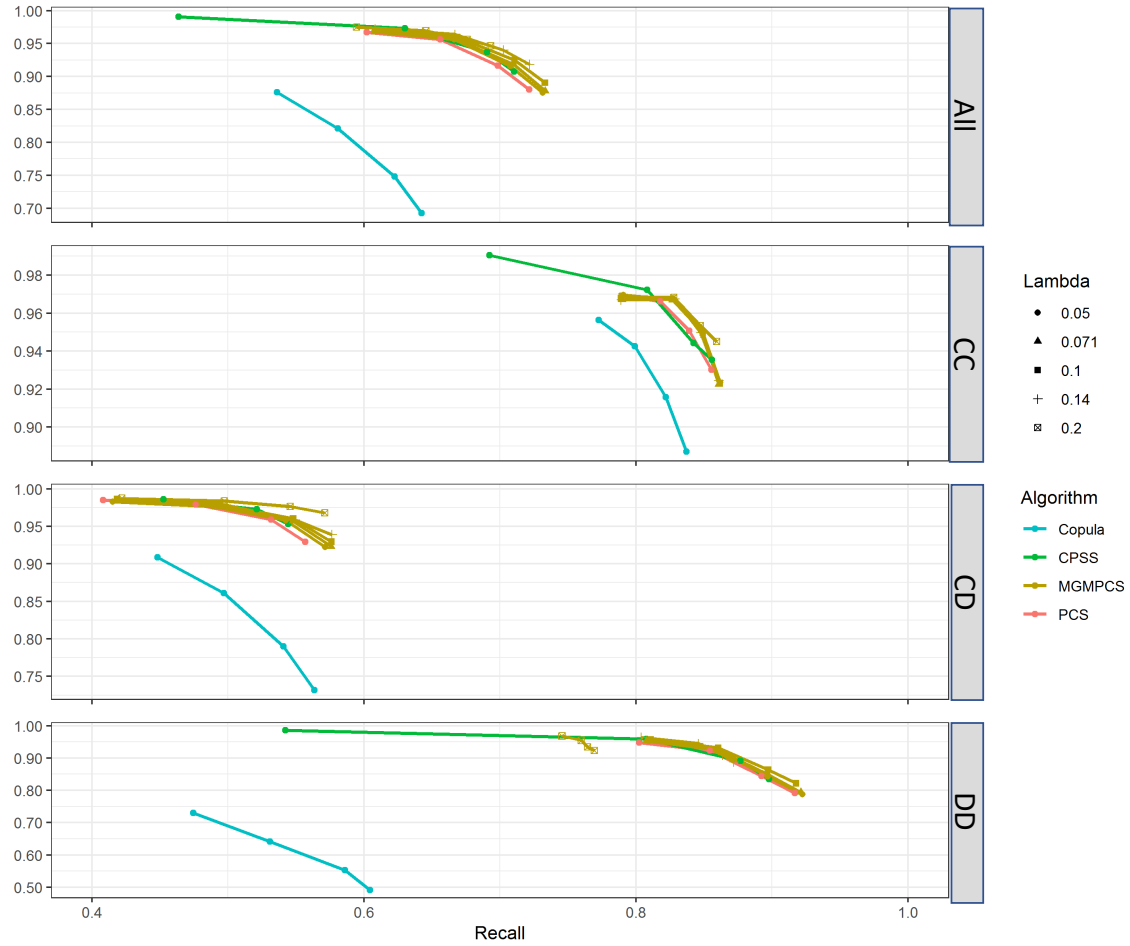


Supplementary Fig S1 Comparison of independence tests p-values to reject the null hypothesis that $X \perp Y \mid Z$ in data generated from a network structure where the null hypothesis is true (**a**,**b**) and a structure where the null hypothesis is false (**c**,**d**). Independence tests were performed on 500 datasets of 100 samples from a discrete variable X and continuous variables Y and Z, with edge weight for $Z \rightarrow Y$ set to 1 and weights for $Y \rightarrow X$ and $Z \rightarrow X$ set to [.75, -1.5, .75].

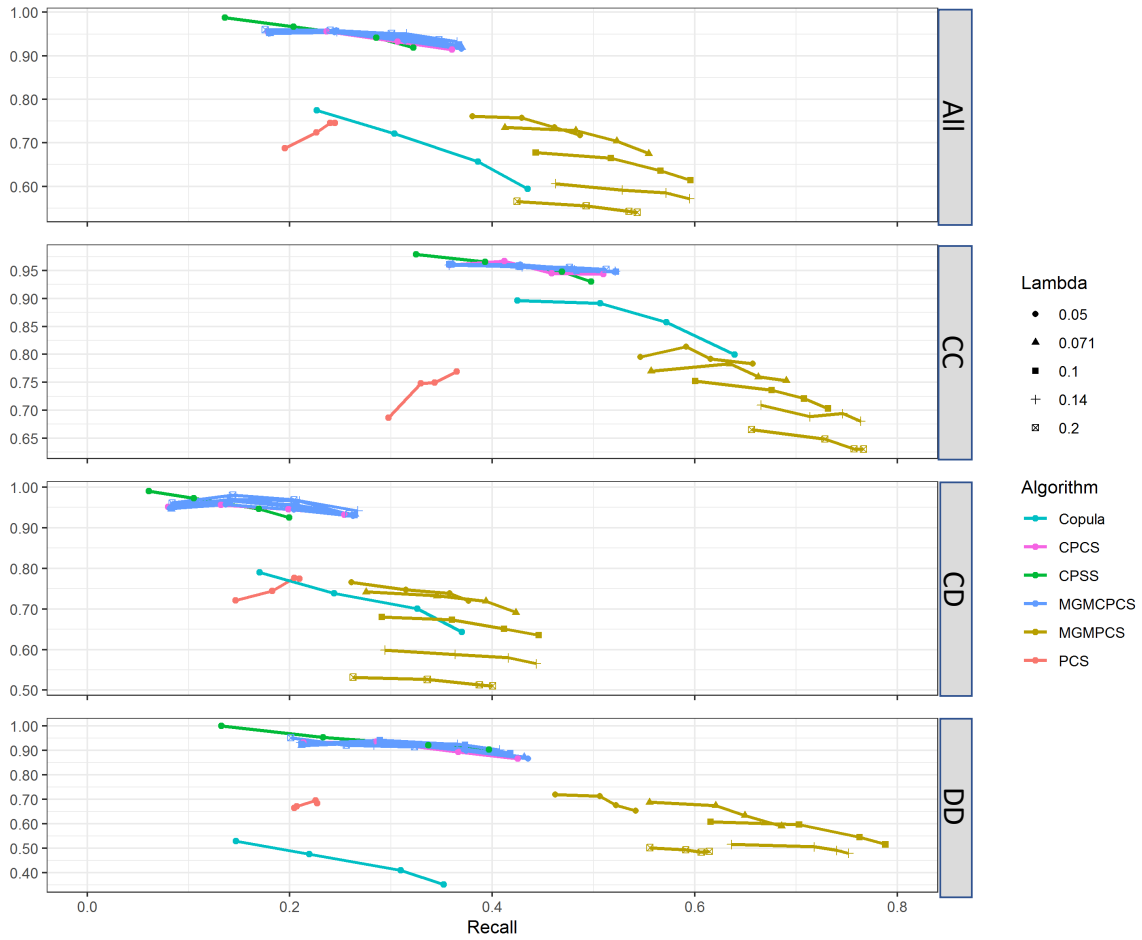


Supplementary Fig S2. (A) Precision-Recall curves of edge adjacency recovery on high-dimensional dataset for $0.2 \leq \lambda \leq 0.8$ (represented by different shaped points) and $0.001 \leq \alpha \leq 0.1$. For a given setting of λ, the different settings of α are connected by lines with colors corresponding to the algorithm. The CPSS line shows the settings of error rate $q \in \{.001, .01, .05, .1\}$.

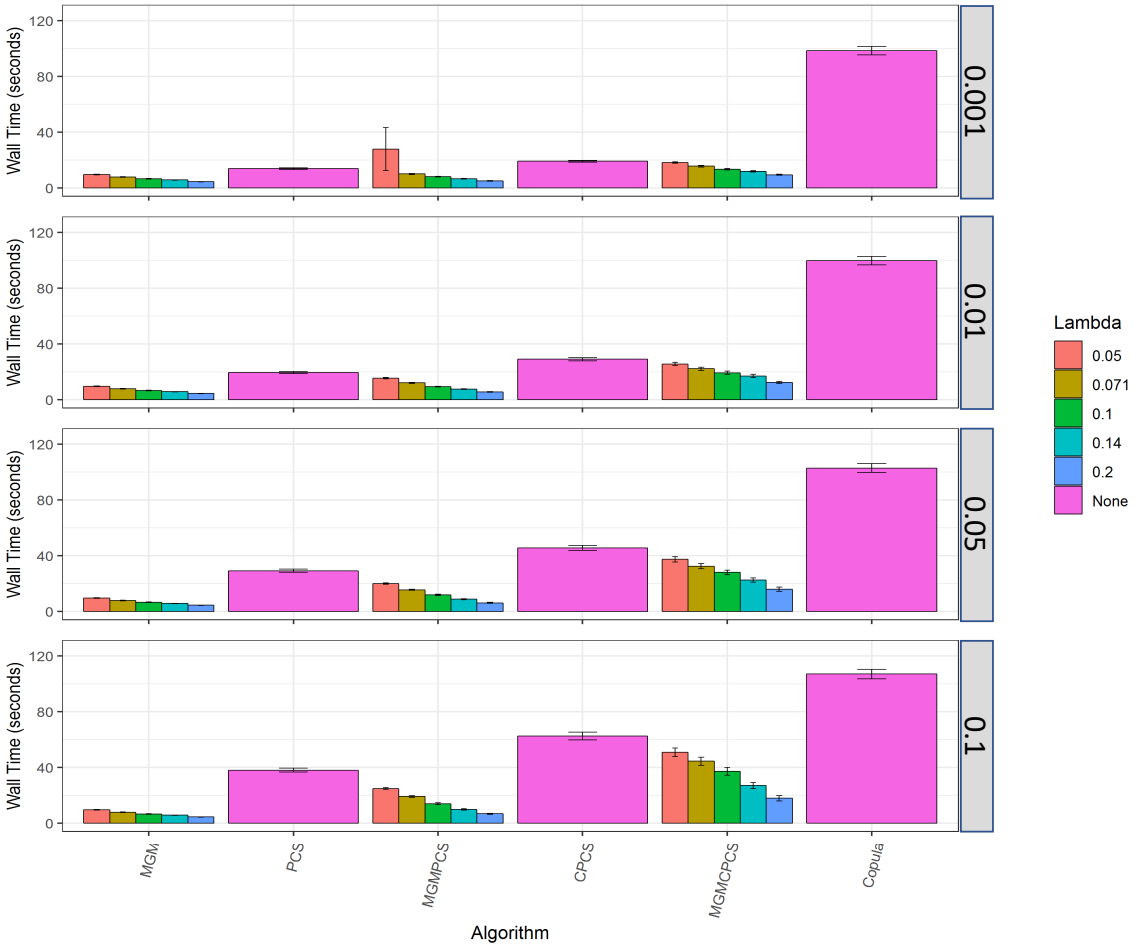
(B)



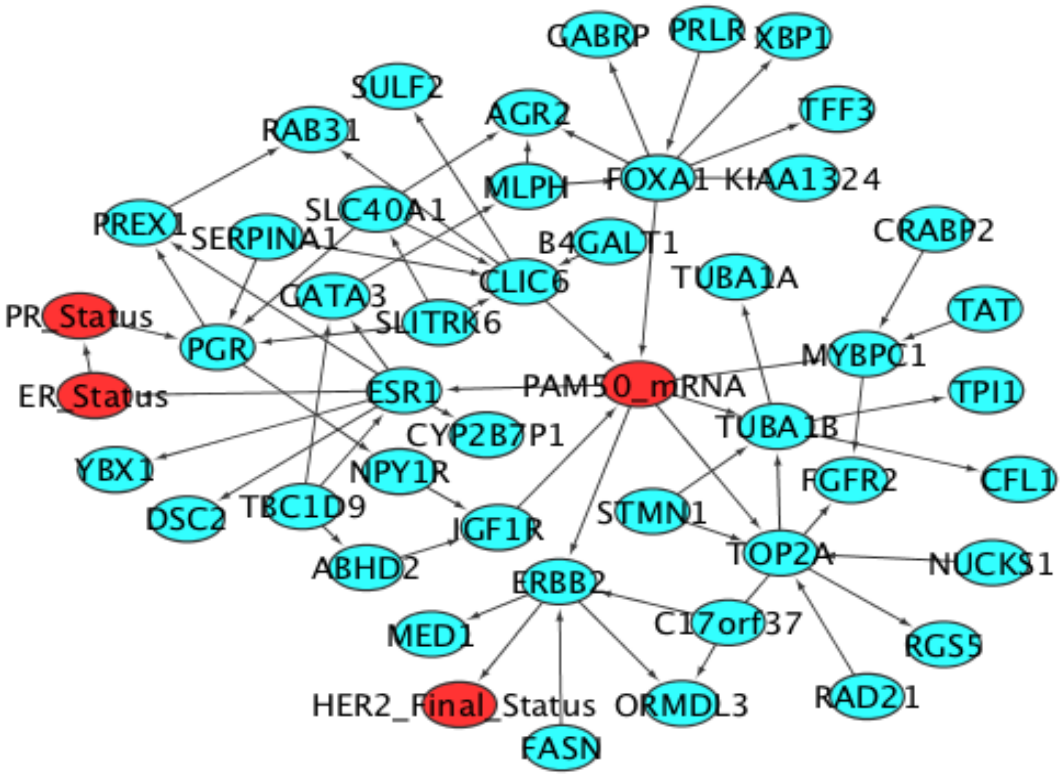
Supplementary Fig S2. (B) Precision-Recall curves of edge adjacency recovery on low dimensional dataset for $.05 \leq \lambda \leq .2$ and $.001 \leq \alpha \leq .1$. and the full range of algorithms and edge types. *Copula*: Copula PC method; *MGMCPCS*: MGM-CPC_Stable; *PCS*: PC_Stable.



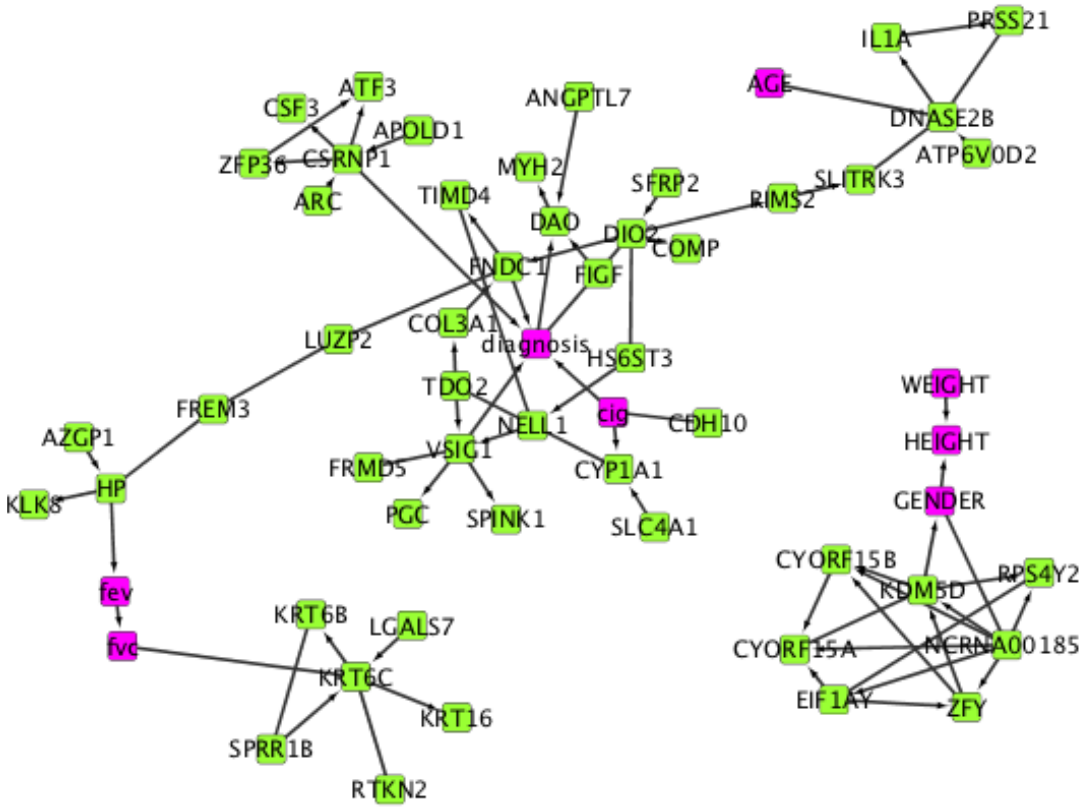
Supplementary Fig S3 Precision-Recall curves of edge direction recovery on low dimensional dataset for $.05 \leq \lambda \leq .2$ and $.001 \leq \alpha \leq .1$. and the full range of algorithms and edge types. *Copula*: Copula PC method; *CPCS*: CPC-Stable; *MGMCPCS*: MGM-CPC_Stable; *MGMPCS*: MGM-PC_Stable; *PCS*: PC_Stable.



Supplementary Fig S4 Running times of search algorithms on low dimensional datasets. Directed search steps were run in parallel on a 4-core laptop.



Supplementary Fig S5. Integrative analysis of omics and clinical variables in the TCGA breast cancer dataset. First and second neighbors of PAM50 breast cancer subtype variable. Variable types include mRNA (green) and clinical (red) variables. MGM-PCS parameters: $\lambda = .2$, $\alpha = .05$.



Supplementary Fig S6. Integrative analysis of omics and clinical variables in the LGRC dataset. First and second neighbors of clinical variables in the LGRC dataset using edges selected based on stability selection derived FDR < .1 . Variable types include mRNA (green) and clinical (red) variables. The genes connected to *Gender* in the bottom right of the graph are Y-chromosome genes.