Graphical Models For Complex Health Data (P8124)

Daniel Malinsky

Columbia University
dsm2128@cumc.columbia.edu

Some Applications in Genetics

Background for "Causal stability ranking"

The setting is gene expression data, and the overall goal is causal inference (intervention effects).

We have p genes (their expression levels/concentrations) $X_1, ..., X_p$ and a phenotype of interest Y.

The idea is that genes causally affect (regulate) each other as well as the phenotype Y. We can do experiments where we supress expression of a gene ("knockout") and see what happens to the rest.

Background for "Causal stability ranking"

- In their main example, $p = 21,326$ measurements from the organism *Arabidopsis thaliana* (a small flowering plant commonly called Thale cress) and Y is the flowering time.
- In the second example $p = 5,361$ genes from *Saccharomyces cerevisiae* for which gene knock-out data is available (the targets Y are all other genes).
- In both cases sample size $n \ll p$
- Want to know which genes are "strong" causes of Y, in the sense that there is a big difference in Y if we manipulate (knock out) gene X_i.
- It is hard and expensive to do 21K experiments, so want to find a ranking of genes by "causal strength" to prioritize a small number of important genes, which we can validate by experiment.

Background for "Causal stability ranking"

We know that genes regulate each other, and one can model that regulatory network as a kind of directed graph.

Probably in reality there is feedback (cyclic relationships) and also unmeasured confounding (latent variables), but the authors explore what happens when you ignore this and assume more simply the truth is a DAG over X.

Background for "Causal stability ranking"

Consider what you would do if you knew the true DAG. You would like to estimate, for each gene X_i:

$$\theta_i \equiv \frac{\partial}{\partial x'} \mathbb{E}[Y | \, \text{do}(X_i = x')]|_{x'=x}$$

This is the *total causal effect* of X_i on Y. It summarizes the expected change in Y under an intervention that changes the expression of gene X_i by one unit. How do you estimate this effect?

Background for "Causal stability ranking"

We previously discussed that in a given DAG $\mathcal{G}$, one set which always satisfies the backdoor criterion is $\text{Pa}(X_i, \mathcal{G})$

$$\mathbb{E}[Y|\,\text{do}(X_i = x)] = \begin{cases} E[Y] & \text{if } Y \in \text{Pa}(X_i, \mathcal{G}) \\ \int E[Y|x_i, z_i]p(z_i)dz_i & \text{if } Y \notin \text{Pa}(X_i, \mathcal{G}) \end{cases}$$

where $Z_i \equiv \text{Pa}(X_i, \mathcal{G})$. That is, you estimate the intervention effect by adjustment for (conditioning on) the parents of X_i. In the special case where the structural equations are *linear* (strong assumption!) estimating the total causal effect θ_i reduces to estimating the linear coefficient on X_i in a regression $Y \sim X_i + \text{Pa}(X_i, \mathcal{G})$.

Background for "Causal stability ranking"

Ok. So if you knew the DAG, you could estimate this regression coefficient for each X_i and rank them. A simple parameter learning problem.

However, in this application the DAG is unknown, so they also perform structure learning with the PC algorithm. That produces a CPDAG, where some edges are unoriented. So, how do you decide on the backdoor set to do adjustment?

The authors use a procedure (called IDA) that looks at all the **possible parents** of X_i and produces a **set** of estimates $\hat{\theta}_i = (\hat{\theta}_{i1}, \hat{\theta}_{i2}, \hat{\theta}_{i3}, ...)$ which represent the range of causal effects from different elements in the equivalence class $\forall i \in \{1, ..., p\}$.

They take min $|\hat{\theta}_i|$ as a *lower bound* for the true causal effect.

Start with an estimated CPDAG

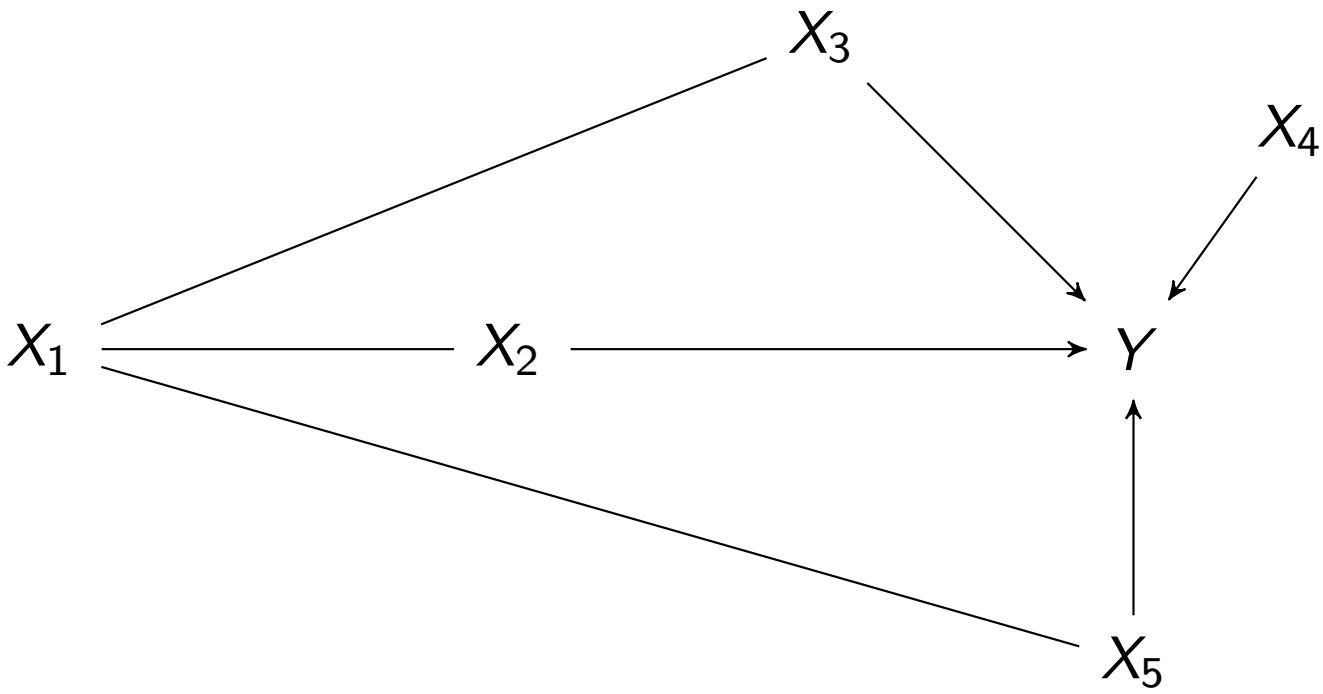


Figure: A CPDAG representing a Markov equivalence class of DAGs

Enumerate the DAGs and estimate causal effects

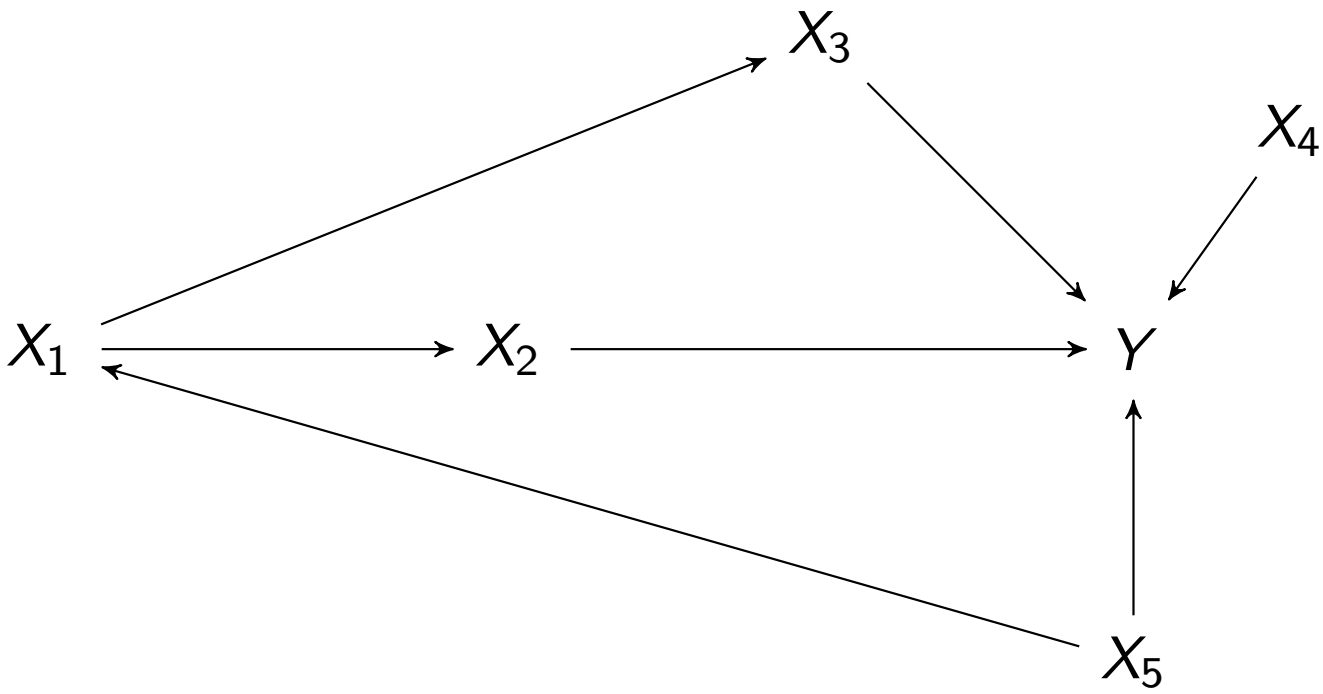


Figure: A DAG in the equivalence class

Enumerate the DAGs and estimate causal effects

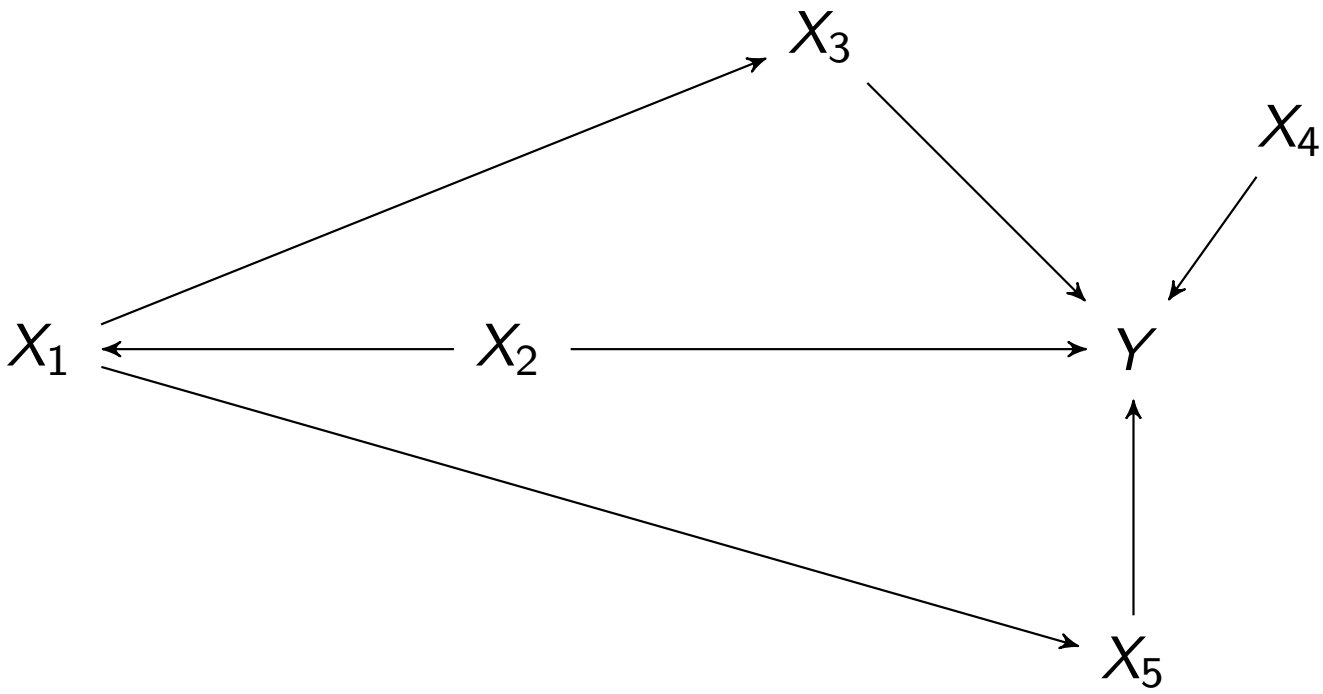


Figure: A DAG in the equivalence class

Enumerate the DAGs and estimate causal effects

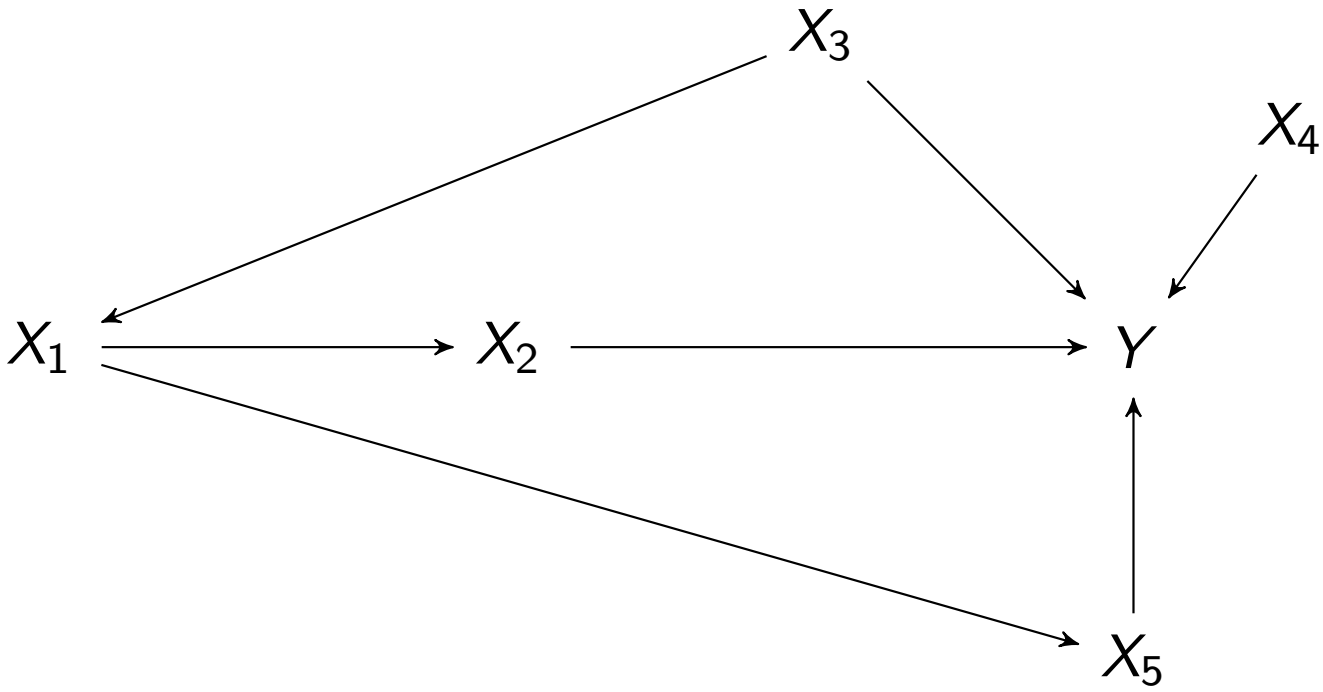


Figure: A DAG in the equivalence class

Let's look at the paper...

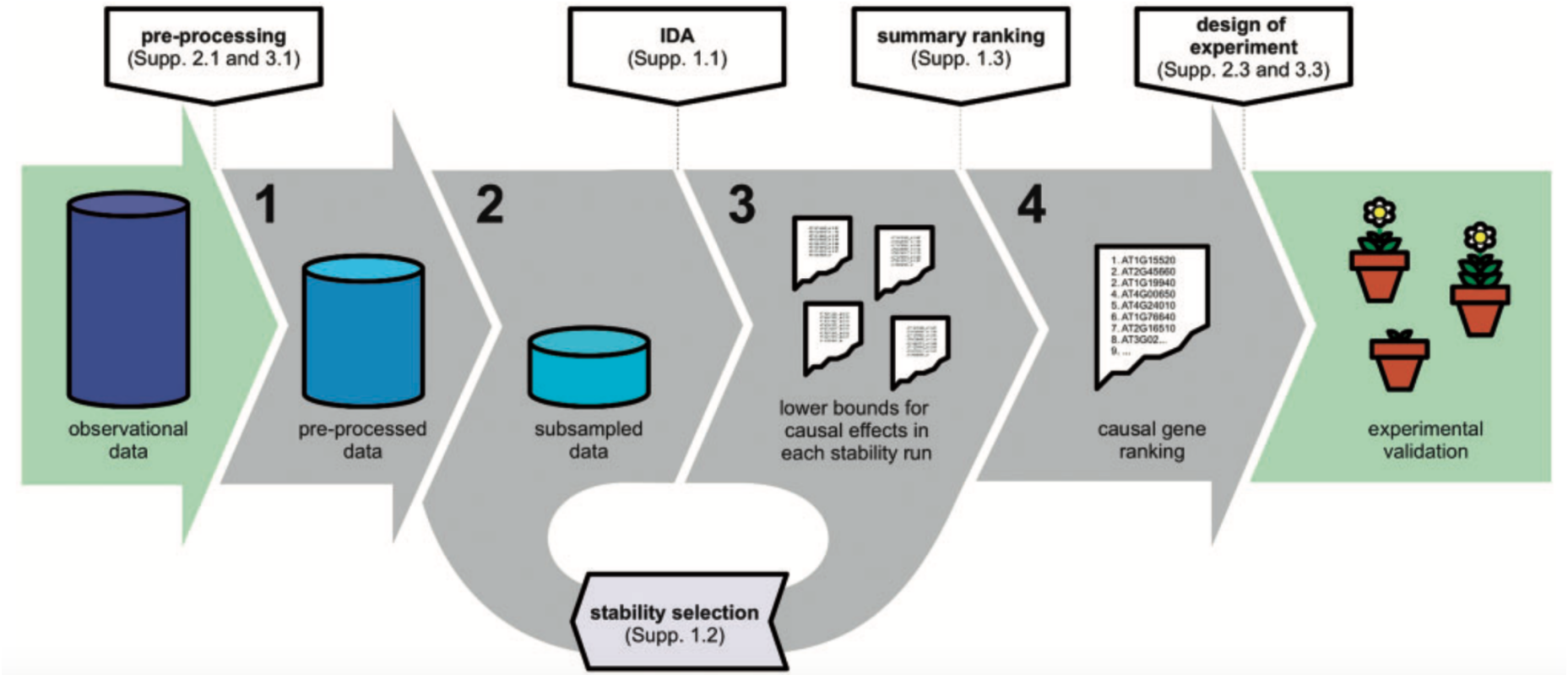
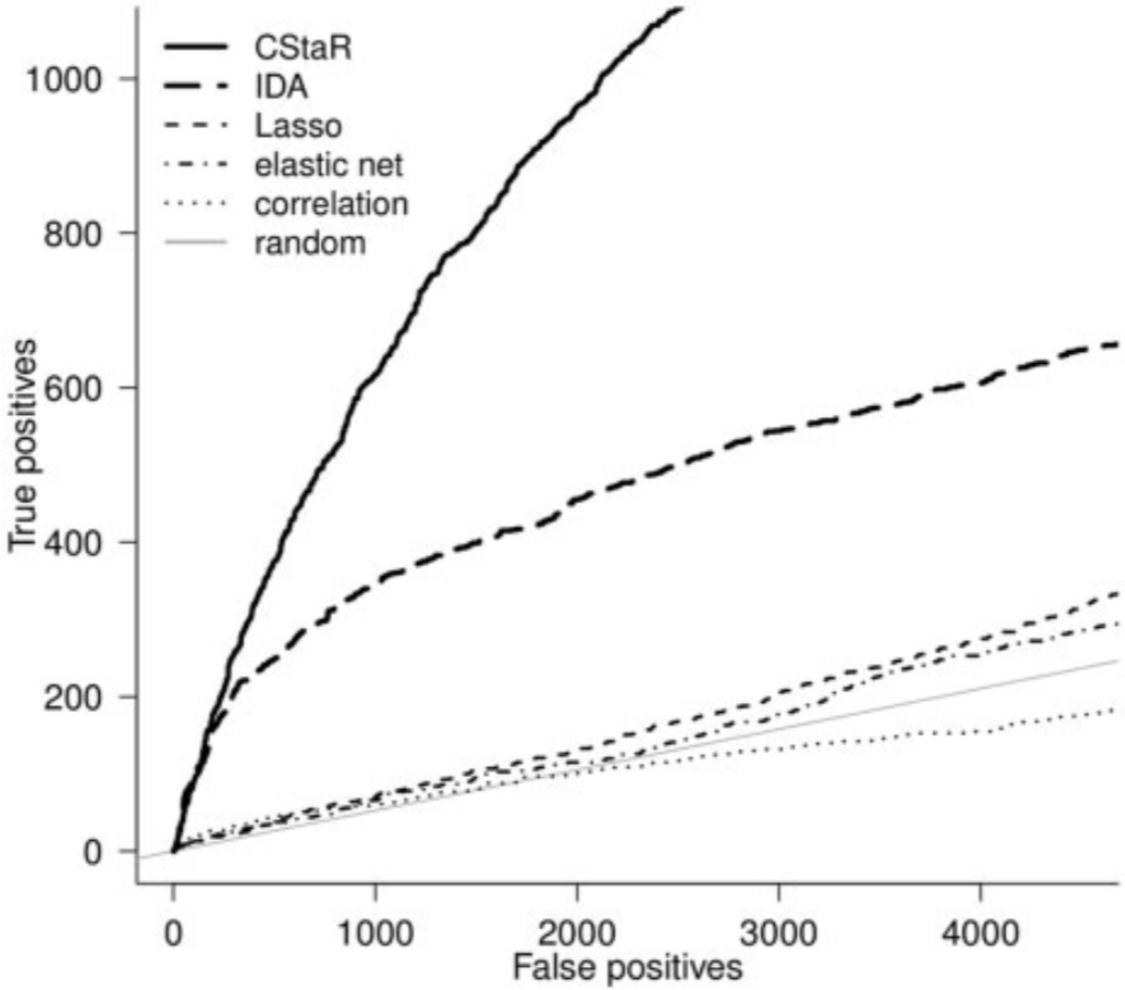


Table 1. Top 25 findings by CStaR for the *A. thaliana* data

	Gene	Summary rank	Median effect	Maximum expression	Error (PCER)	Name/annotation
1	AT2G45660	1	0.60	5.07	0.0032	*SOC1*
2	AT4G24010	2	0.61	5.69	0.0033	*ATCSLG1*
3	**AT1G15520**	**2**	**0.58**	**5.42**	**0.0033**	***PDR12***
4	AT3G02920	5	0.58	7.44	0.0041	*RPA2B*
5	**AT5G43610**	**5**	**0.41**	**4.98**	**0.0069**	***ATSUC6***
6	AT4G00650	7	0.48	5.56	0.0051	*FRI*
7	**AT1G24070**	**8**	**0.57**	**6.13**	**0.0040**	***ATCSLA10***
8	**AT1G19940**	**9**	**0.53**	**5.13**	**0.0045**	***ATGH9B5***
9	**AT3G61170**	**9**	**0.51**	**5.12**	**0.0044**	**PPR protein**
10	**AT1G32375**	**10**	**0.54**	**5.21**	**0.0045**	**F-box protein**
11	**AT2G15320**	**10**	**0.50**	**5.57**	**0.0047**	**LRR protein**
12	**AT2G28120**	**10**	**0.49**	**6.45**	**0.0054**	**Nodulin protein**
13	AT2G16510	13	0.50	10.7	0.0050	*AVAP5*
14	AT3G14630	13	0.48	4.87	0.0056	*CYP72A9*
15	**AT1G11800**	**15**	**0.51**	**6.97**	**0.0053**	**Endonuclease**
16	AT5G44800	16	0.32	6.55	0.0079	*CHR4*
17	AT3G50660	17	0.40	7.60	0.0078	*DWF4*
18	AT5G10140	19	0.30	10.3	0.0085	*FLC*
19	AT1G24110	20	0.49	4.66	0.0071	Peroxidase
20	**AT2G27350**	**20**	**0.48**	**7.06**	**0.0067**	***OTLD1***
21	AT1G27030	20	0.45	10.0	0.0075	Unknown protein
22	**AT2G28680**	**22**	**0.46**	**5.23**	**0.0072**	**Cupin protein**
23	AT3G16370	23	0.43	12.4	0.0099	Lipase/hydrolase
24	AT5G25640	23	0.33	5.59	0.0091	Serine protease
25	AT1G30120	24	0.46	9.97	0.0077	*PDH-E1 BETA*

The genes are ranked by increasing summary rank, where ties are sorted according to the estimated median total causal effect taken over 100 stability runs (third column).


CStaR
IDA
Lasso
elastic net
correlation
random
True positives
0
200
400
600
800
1000
0
1000
2000
3000
4000
False positives

Some nice things about this paper

- They don't simply apply a structure learning procedure once and estimate some quantity once. Rather, they subsample repeatedly and report estimates that are stable under the subsampling procedure (genes ranked highly with high frequency). They also vary q, the ranking quantile. This is a way of ensuring (or trying to) that their results are not merely artifacts.
- They can validate their procedure with follow-up experiments (not always possible).
- They compare with alternative procedures, even ones that seem naive (like using pairwise correlations to find important genes).
- They are relatively clear about what assumptions they're making (DAG, linearity, etc.) though most of this is in the supplement.

Drawbacks

- They don't discuss much the limitations of their method/assumptions.
- They don't explain why they use $\alpha = 0.1$ with PC.
- They might also have tried alternative structure learning procedures (for DAGs, PAGs, cyclic graphs?)
- They might have tried nonparam/nonlinear regression methods in the estimation step (though for small n, perhaps not feasible).

Wang et al. (2016) FastGGM paper

- Their FastGGM method is a scalable alg for estimating Σ^{-1} (they call this Ω, we called it K) in settings with $n \ll p$
- The algorithm involves solving a sequence of scaled lasso regressions, this requires an initial estimate of σ, and an interative process estimating β and σ in turn
- Final estimate of each ω_{ij} comes from $(\frac{1}{n}\hat{\epsilon}_{ij}^T\hat{\epsilon}_{ij})^{-1}$ regression residuals
- Can get p-value and CI for each $\hat{\omega}_{ij}$ based on $\sqrt{n(\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2)}(\hat{\omega}_{ij} - \omega_{ij}) \xrightarrow{D} N(0,1)$

Wang et al. gene expression app

- Gene expression microarray data from lymphoblastoid cells
- Two groups: 258 asthmatic children and 134 healthy children
- After removing various genes (why?) had $p = 9938$ genes for each group
- Applied FastGGM alg to data from each group a constructed networks w/ edge if partial corr met FDR < 0.01 threshold

Results

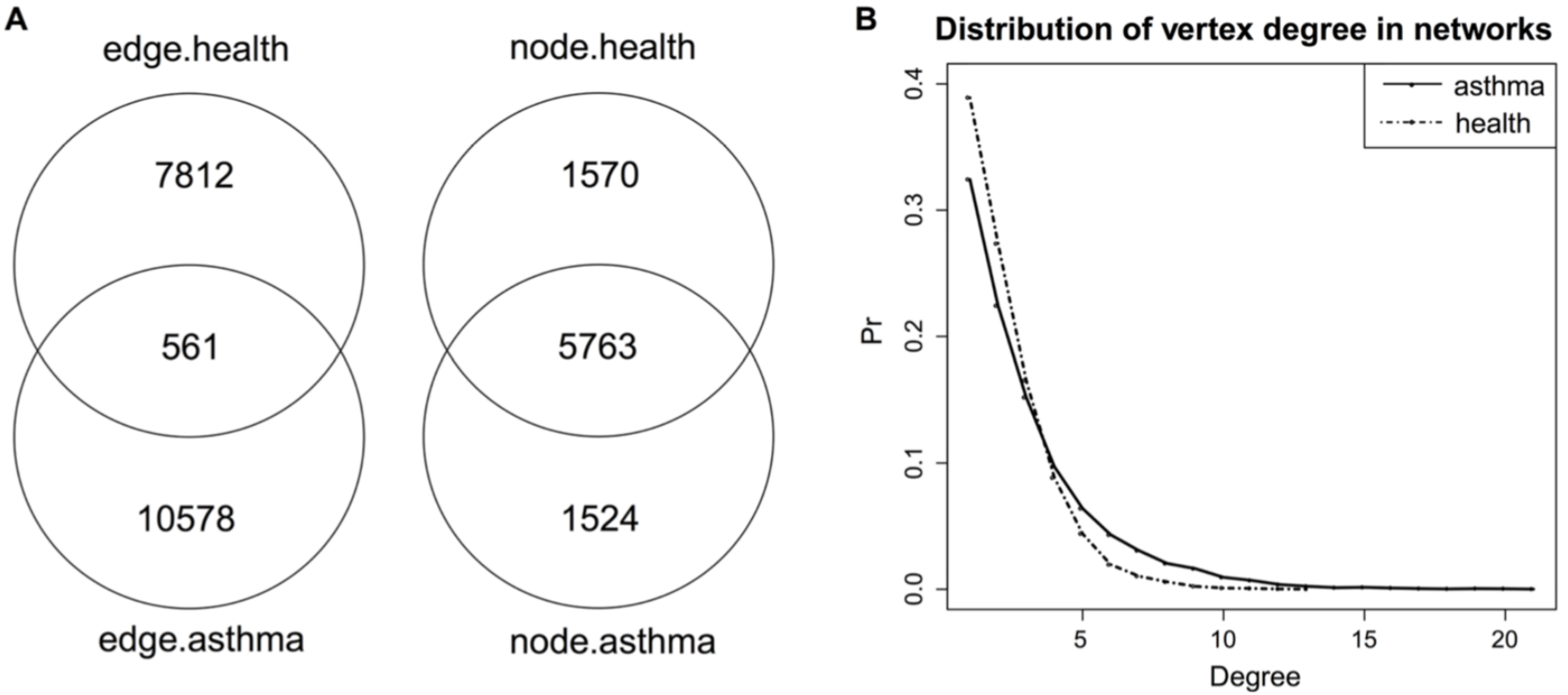
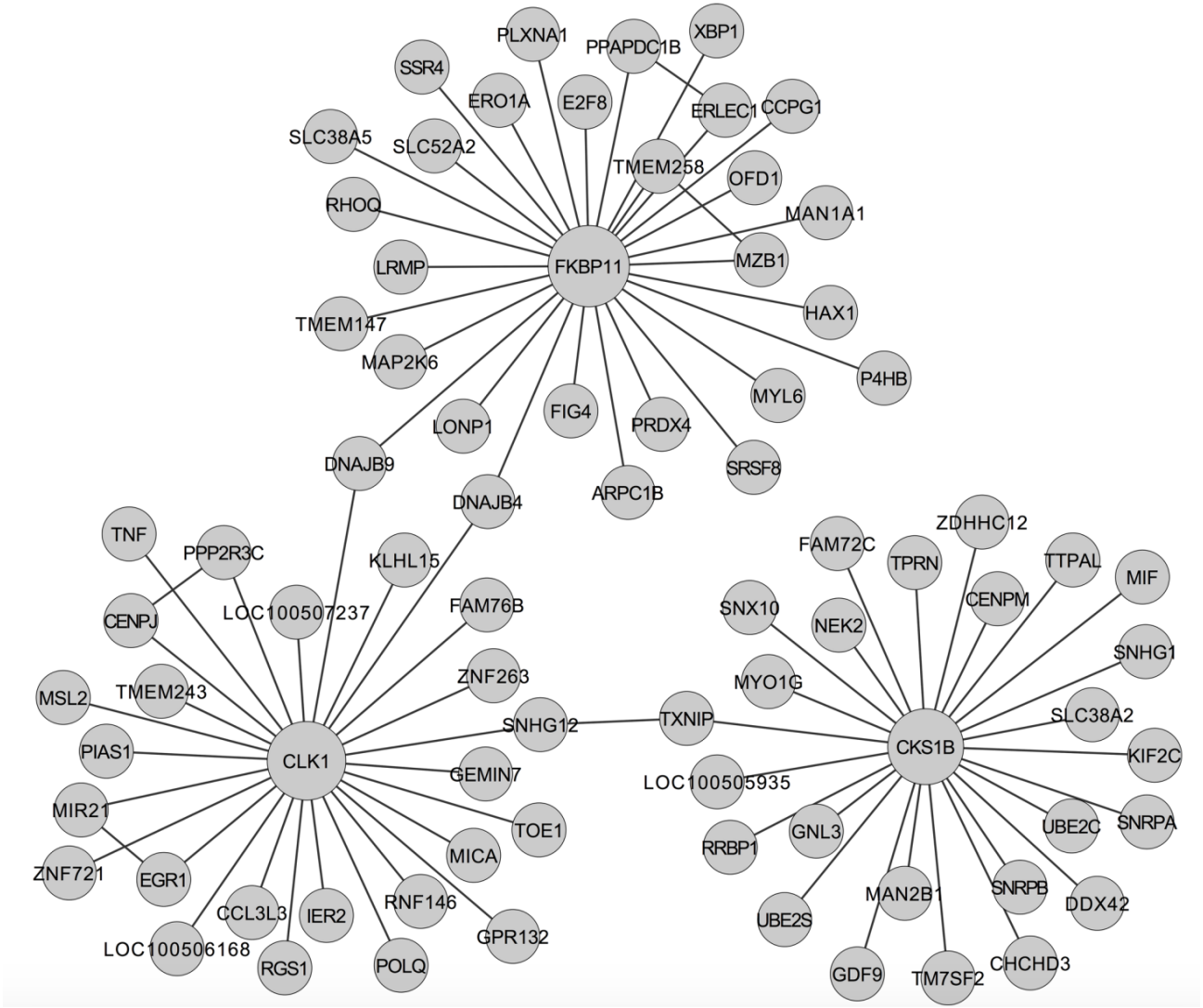


Fig 2. Comparing gene association networks under asthmatic and healthy conditions. A) Venn diagram of the edges and nodes in the asthmatic and healthy networks. B) Distributions of vertex degree in the two networks.

Results



Differential network (top 3 hubs)

Wang et al. protein network app

- $n = 59$ Alzheimer's disease (AD) subjects and $p = 192$ protein levels
- A bit unclear what these protein levels repesent (weighted average of peptide measures) and protein labels are not informative
- "of interest to examine the synaptic protein networks for the AD subjects and to detect modules of highly interconnected proteins from the network to elucidate the disease physiology"
- Bootstrapped FastGGM 200 times and averaged the partial correlations
- Performed clustering to identify clusters of co-dependent networks

Results

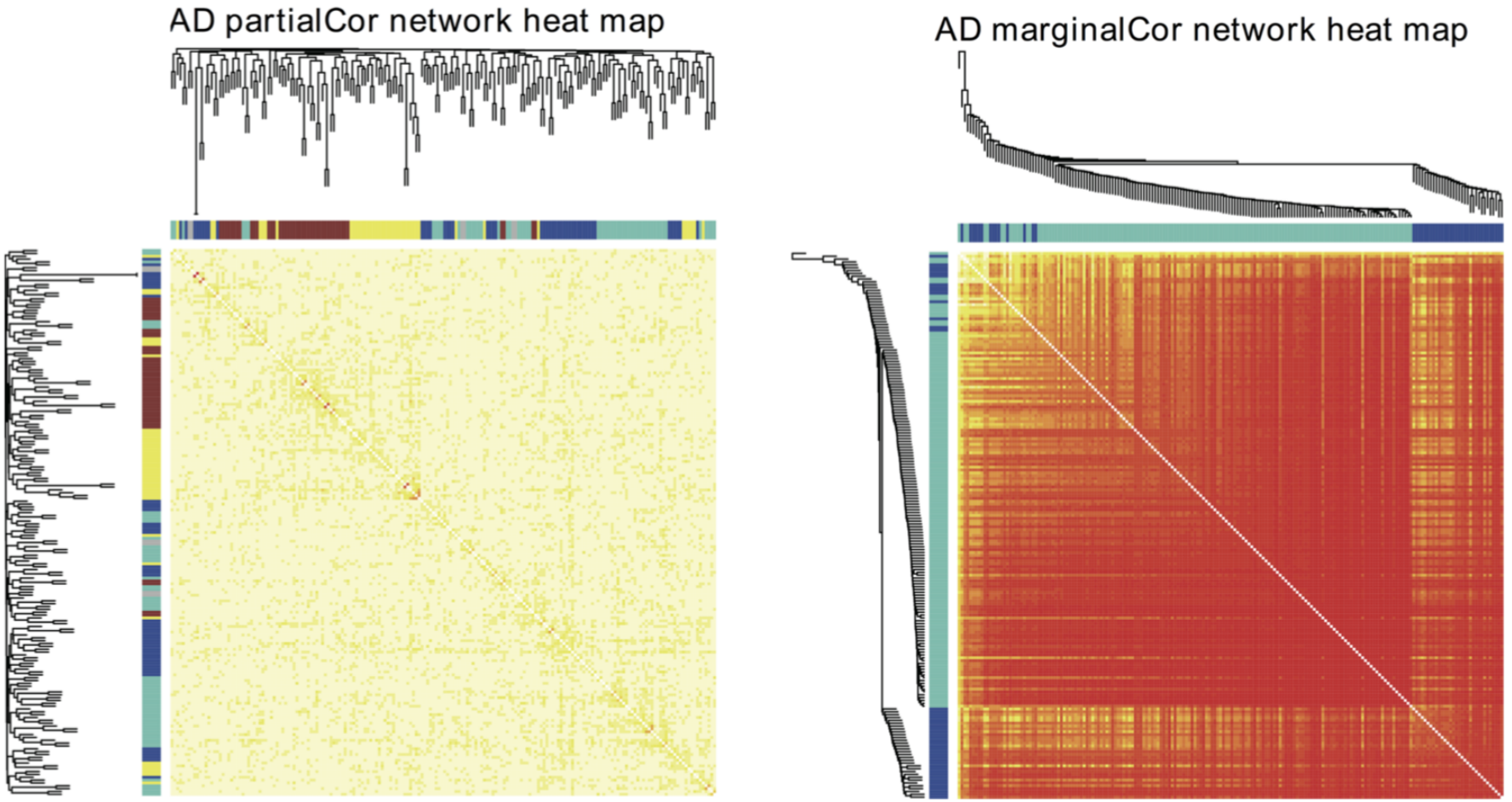


Fig 4. Heat maps of synaptic protein network in AD cohort where red indicates stronger correlation and the white indicates weaker correlation. The left and top color bars indicate the module membership of each protein (grey colored proteins do not belong to any module), with the corresponding hierarchical clustering dendrograms plotted. The left is the heat map based on partial correlations and the right is the heat map based on marginal correlations.

Interpretation issues

"Since the partial-correlation-based network quantifies the correlation between each pair of proteins with the effects of other proteins excluded, it only keeps the correlations due to direct causal relationships between the protein pairs..." (p. 12) $\Rightarrow$ **this is false!** (why?)

Interpretation issues

"Since the partial-correlation-based network quantifies the correlation between each pair of proteins with the effects of other proteins excluded, it only keeps the correlations due to direct causal relationships between the protein pairs..." (p. 12) $\Rightarrow$ **this is false!** (why?)

Edges in an MRF may correspond to "moralized" colliders in the underlying directed graph, e.g., if $X_i \rightarrow X_k \leftarrow X_j$ in DAG then $X_i - X_j$ in MRF. Also unmeasured variables may induce associations between X_i, X_j without there being any causal effect in either direction. So an edge $X_i - X_j$ in MRF can correspond to a number of different causal/non-causal arrangements!

Interpretation issues

"Analyzing differential networks between conditions can help to elucidate the molecuar mechanisms of complex biological processes... identifying novel biomarkers or biological pathways for further experimental evaluation..." (p. 14)

"compared the topological changes"

"these proteins are worthy of additional verifications for their interaction and functions..."