Directed Cyclic Mixed Graph Modeling for High-Dimensional Genomic Data Integration

Carel F.W. Peeters[1⋆], **Wessel N. van Wieringen**[1], **and Mark A. van de Wiel**[1]

[1]Dept. of Epidemiology & Biostatistics, VU University medical center, Amsterdam, the Netherlands

⋆cf.peeters@vumc.nl

1. Background

Graphical Modeling

A class of probabilistic models utilizing graphs to express conditional (in)dependence relations between random variables. We consider graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of a finite set $\mathcal{V}$ of vertices and set of edges $\mathcal{E}$. The vertices of the graph correspond to a collection of random variables with probability distribution P. Graphical Modeling considers pairs $(\mathcal{G}, P)$.

Dominant Approaches in Networks for Genomic Data

- Undirected (Gaussian) graphical modeling
- Modeling precision matrices structured according to a DAG
- Consider 1 omics platform at a time

Desire

- Graphical modeling of a model-structured precision matrix
- Allow for reciprocal effect and feedback cycles
- Incorporate multiple genomic platforms: miRNA and mRNA

2. Model

Model

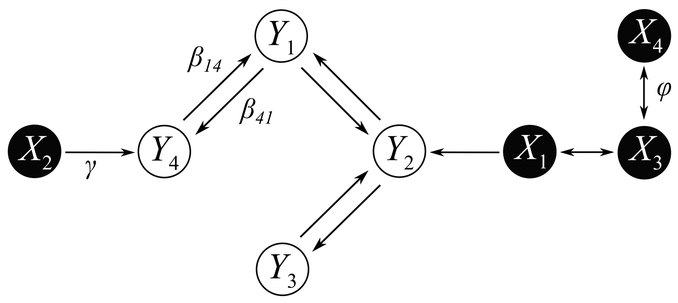
The SEM model we consider can be expressed as:

$$\mathbf{y}_i := \mathbf{B}\mathbf{y}_i + \mathbf{\Gamma}\mathbf{x}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n. \tag{1}$$

Assumptions

1. Properly preprocessed data
2. $\mathbf{y}_i \perp\!\!\!\perp \mathbf{y}_{i'}, \ \ \forall i \neq i'$
3. $\boldsymbol{\epsilon}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Psi})$, with $\mathbf{\Psi} \equiv \text{diag}[\psi_{11}, \ldots, \psi_{pp}]$, and $\psi_{jj} > 0, \forall j$
4. $\mathbf{x}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{\Phi})$, with $\mathbf{\Phi} \succ 0$
5. $\mathbf{x}_i \perp\!\!\!\perp \boldsymbol{\epsilon}_{i'}, \forall i, i'$
6. $(\mathbf{I}_p - \mathbf{B})$ is nonsingular and $\beta_{jj} = 0, \forall j$
7. $\beta_{jj'} = \beta_{j'j}, \ \ \forall\, j \neq j'$ (expression reciprocation/feedback)

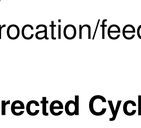
Natural Graphical Representation: Directed Cyclic Mixed Graph (DCMG)



$$\beta_{14} = \beta_{41}$$

3. The Model as a Graphical Object

Stretching the Idea of the Collider



Definition (m-separation)

Let v_l be an intermediate vertex on path $\rho_{ab} = (e_1, \ldots, e_r)$ from v_a to v_b (from a to b for short). A path ρ_{ab} from a to b in $\mathcal{G}$ is *pathwise* m-*separated* by a set of vertices $C \subseteq \mathcal{V} \setminus \{a, b\}$ iff

1. $\{v_l | v_l$ is a non-collider on $\rho_{ab}\} \cap C \neq \emptyset$; or

2. $\exists \{v_l | v_l$ is a collider on $\rho_{ab}\} \equiv S$ s.t. $S \cap C = \emptyset \ \wedge \ \textit{de}(S) \cap C = \emptyset$.

If C pathwise *m*-separates every path from a to b, then a to b are said to be m-*separated* given C. If C does not *m*-separate a from b, then a and b are said to be m-*connected* given C.

Some Results

- The model (1) is identified under our assumptions
- Denote the set of normal probability distributions that satisfy system (1) by $\mathcal{P}$. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the associated DCMG. Then all $P \in \mathcal{P}$ are global $\mathcal{G}$-markov and the class $\mathcal{P}$ is (given a Faithfulness assumption) Markov perfect w.r.t. $\mathcal{G}$

Implications

The definition allows one to read all the conditional (in)dependencies off the DCMG. The DCMG tied to model (1) is a true graphical object. Thus, *we can use the machinery of graphical modeling to solve the reverse engineering problem: for given data, can we find the DCMG?*

4. Approach

Step 1: Regularization

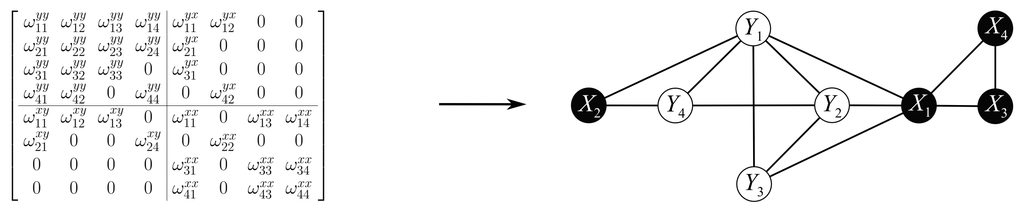
- Let $\hat{\mathbf{\Sigma}}$ denote the sample covariance matrix on $\mathbf{y}_i$ and $\mathbf{x}_i$
- When $(p+q) \approx n$ or $(p+q) > n$, $\hat{\mathbf{\Sigma}}$ is ill-behaved or singular and $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Sigma}}^{-1}$ is undefined
- The following (proper ℓ_2) penalized ML estimator is always well-behaved and p.d.:

$$\hat{\mathbf{\Omega}}(\lambda) = \left\{ \left[\lambda \mathbf{I}_{p+q} + \frac{1}{4}(\hat{\mathbf{\Sigma}} - \lambda \mathbf{T})^2 \right]^{1/2} + \frac{1}{2}(\hat{\mathbf{\Sigma}} - \lambda \mathbf{T}) \right\}^{-1},$$

where $\mathbf{T}$ denotes a p.d. symmetric target matrix and where the penalty $\lambda \in (0, \infty)$.

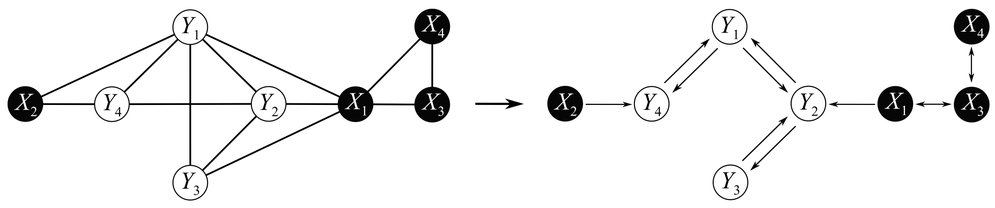
Step 2: Determine Support Precision Matrix

- Test for vanishing partial correlations to obtain $\hat{\mathbf{\Omega}}(\lambda)^0$: A sparse representation of $\hat{\mathbf{\Omega}}(\lambda)$
- Use local false discovery rate procedure



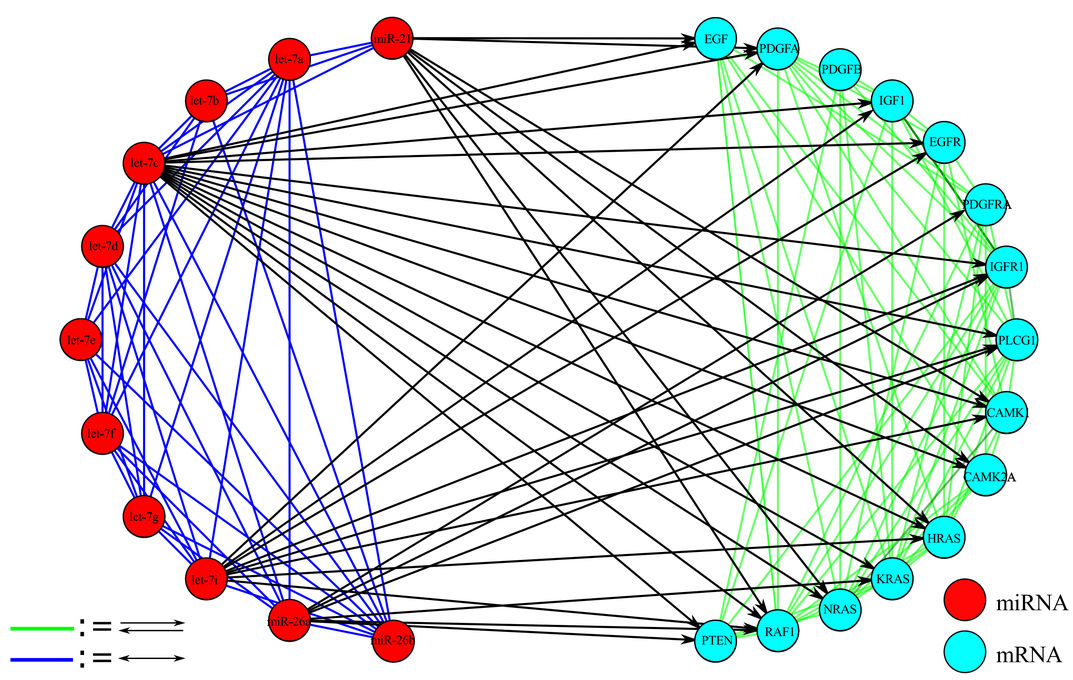
Step 3: Find Cyclic Directed Mixed Graph

- From $\hat{\mathbf{\Omega}}(\lambda)^0$ we find $\hat{\mathbf{\Theta}} = \{\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}, \hat{\mathbf{\Psi}}, \hat{\mathbf{\Phi}}\}$ such that $\mathbf{\Omega}(\hat{\mathbf{\Theta}})$ is as close as possible to $\hat{\mathbf{\Omega}}(\lambda)^0$
- Inverse variance lemma and identification proposition imply simple iterative algorithm



5. Application: Glioblastoma Multiforme

- Aggressive malignant primary human brain tumor
- miRNA and mRNA data from The Cancer Genome Atlas
- Retained features implied in progression of glial cell to GBM (as defined by KEGG)
- 350 samples, sample covariance poorly conditioned



References

[1] Koster, J.T.A. (1996). Markov properties of nonrecursive causal models. Annals of Statistics, **24**: 2148–2177

[2] Peeters, C.F.W., Bilgrau, A.E., & van Wieringen, W.N. (2015). `rags2ridges`: Ridge estimation of precision matrices from high-dimensional data. `R` package 2.0. Available from: http://cran.r-project.org/web/packages=rags2ridges

[3] Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. Scandinavian Journal of Statistics, **30**: 145–157.

[4] van Wieringen, W.N., & Peeters, C.F.W. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. Computational Statistics & Data Analysis, **103**: 284–303.

Abbreviations: DAG = directed acyclic graph; DCMG = directed cyclic mixed graph; GBM = Glioblastoma Multiforme; KEGG = Kyoto Encyclopedia of Genes and Genomes; miRNA = micro RNA; mRNA = messenger RNA; RNA = ribonucleic acid; SEM = simultaneous equation model