Graphical models

Wei Pan (& Xiaotong Shen)

Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota, Minneapolis, MN 55455
Email: panxx014@bumn.edu

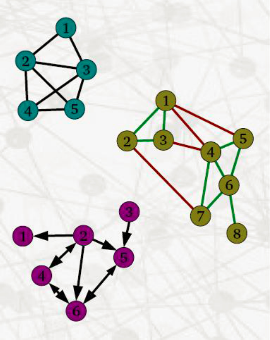
PubH 8475/Stat 8056

Outline

- Reconstruction of an undirected graph:
  Gaussian graphical model (GGM)
- Inference for an undirected graph
- Reconstruction of multiple (related) undirected graphs
- Reconstruction of a directed acyclic graph (DAG):
  - Observational data;
  - Intervention data.

Terminology

- Graph $\mathcal{G} = (V, E)$:
  - A set of nodes $V = \{v_1, \ldots, v_p\}$.
  - A set of edges or links between nodes $E = \{e_1, \ldots, e_m\}$.
  - Undirected: The edges have no direction, and the edge $\{i, j\}$ is the same as the edge $\{j, i\}$, i.e. each edge is an unordered pair of nodes.
  - Directed: The edges have direction, and the edge (i, j) is not the same as the edge (j, i), i.e. each edge is an ordered pair of nodes.



Adjacency matrix

- Graph $\mathcal{G} = (V, E) \to p \times p$ adjacency matrix
  $\boldsymbol{U} = \{U_{ij} : 1 \leq i, j \leq p\}$, where

$$U_{ij} = \begin{cases} \neq 0 & \text{if } (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases}$$

- Undirected:
  - $\boldsymbol{U}$ is symmetric, i.e., $U_{ij} = U_{ji}$.
  - $\boldsymbol{U} = \{U_{ij}\}$, U_{ij} denotes"similarity" between i & j.
- Directed:
  - $\boldsymbol{U}$: symmetric or asymmetric.
  - $\boldsymbol{U}$: directed acyclic graph (no directed cycles) $\to$ acyclicity.
  - $\boldsymbol{U}^k = 0$: maximum length of directed pathway $\leq k - 1$.
    Q: what is the meaning of $(U^k)_{ij}$?
    $(U^2)_{ij} = \sum_{m=1}^{p} U_{im} U_{mj}$

Reconstruction of an undirected graph

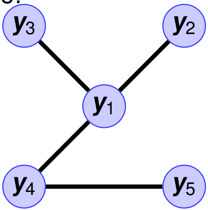
A graphical model for undirected graphs

- Pairwise relations
  - set of p variables $\Leftrightarrow$ $\boldsymbol{Y} = \left(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_p\right)$.
  - interactions $\Leftrightarrow$ conditional dependencies.
  - graph:

$$\mathcal{G} = \left(V, E\right), \ V = \{1, \cdots, p\}$$

$$(j, k) \in E \quad \text{if} \quad \boldsymbol{Y}_j \not\perp\!\!\!\perp \boldsymbol{Y}_k \mid \boldsymbol{Y}_{\setminus\{j,k\}}$$

  - example:



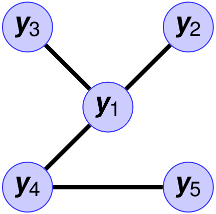
  - $\boldsymbol{y}_1 \not\perp\!\!\!\perp \boldsymbol{y}_3 \mid \boldsymbol{y}_2, \boldsymbol{y}_4, \boldsymbol{y}_5$
  - $\boldsymbol{y}_1 \perp\!\!\!\perp \boldsymbol{y}_5 \mid \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4$

- Goal: reconstruct $\mathcal{G}$ based on n i.i.d.
- Remark: in some applications, $\perp\!\!\!\perp$ may mean (conditional or marginal) uncorrelatedness.
  An example: co-expression networks.

Gaussian graphical model for undirected graphs

- Model: $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.
- Precision matrix: $\boldsymbol{\Omega} = (\omega_{jk})_{p \times p} = \boldsymbol{\Sigma}^{-1}$
- Conditional independence:

$$\boldsymbol{Y}_j \perp\!\!\!\perp \boldsymbol{Y}_k \mid \boldsymbol{Y}_{\setminus\{j,k\}} \;\Leftrightarrow\; \omega_{jk} = 0$$



$$\Longleftrightarrow \boldsymbol{\Omega} = \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} & 0 \\ \omega_{21} & \omega_{22} & 0 & 0 & 0 \\ \omega_{31} & 0 & \omega_{33} & 0 & 0 \\ \omega_{41} & 0 & 0 & \omega_{44} & \omega_{45} \\ 0 & 0 & 0 & \omega_{54} & \omega_{55} \end{bmatrix}$$

- Graph connectivity $\Longleftrightarrow$ zero offdiagonals of $\boldsymbol{\Omega}$. Estimation of zeros of $\boldsymbol{\Omega}$: covariance selection (Dempster, 1972).

Conditional independence

- $\boldsymbol{Y} = (Y_1, \cdots, Y_p)^T \sim N(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$.
- Density of $\boldsymbol{Y}$: $f(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^p det(\boldsymbol{\Sigma})}} \exp(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Omega}\boldsymbol{y})$.
- Let $Z = (Y_3, \cdots, Y_p)$ and $X = (Y_1, Y_2)$.

$$X|Z \sim N(\underbrace{\mu_X + (Z - \mu_Z)^T\Sigma_{ZZ}^{-1}\Sigma_{ZX}}_{\mu_{X|Z}=\boldsymbol{0}}, \underbrace{\boldsymbol{\Sigma}_{XX} - \Sigma_{ZX}^T\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\Sigma}_{ZX}}_{\boldsymbol{\Omega}_{XX}}).$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XZ} \\ \boldsymbol{\Sigma}_{ZX} & \Sigma_{ZZ} \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{XX} & \boldsymbol{\Omega}_{XZ} \\ \boldsymbol{\Omega}_{ZX} & \Omega_{ZZ} \end{pmatrix}.$$

- Inverse: $\Omega_{XX} = (\omega_{ij})_{2\times2}$, $\omega_{12} \to (1,2)$-entry of Ω_{XX}.
- Conditional density of $\boldsymbol{X}$ given $\boldsymbol{Z}$ is

$$f(\boldsymbol{x}|\boldsymbol{z}) = \frac{(\omega_{11}\omega_{22}-\omega_{12}^2)^{1/2}}{2\pi}\exp(-\frac{1}{2}(\omega_{11}y_1^2 + \omega_{22}y_2^2 + 2\omega_{12}y_1y_2))$$
$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}}\exp(-\frac{1}{2(1-\rho_{12})^2)}(y_1^2/\sigma_1^2 + y_2^2/\sigma_2^2 - 2\rho_{12}y_1y_2/\sigma_1\sigma_2)$$

- Note: $-\omega_{12}\sigma_1\sigma_2 = \frac{\rho_{12}}{(1-\rho_{12}^2)}$, $\omega_{jj}\sigma_j^2 = \frac{1}{(1-\rho_{12}^2)}$, $\rho_{12}, \sigma_j \to$ corr, var given $\boldsymbol{Z}$.
- Conditional independence of Y_1, Y_2 given rest, iff $\omega_{12} = 0$.

Conditional independence and partial correlation $\rho_{jj'}$

- Express Y_j:

$$Y_j = \sum_{j' \neq j} \beta_{jj'} Y_{j'} + \epsilon_j,$$

$$\beta_{jj'} = -\omega_{jj'} / \omega_{jj} = \rho_{jj'} \sqrt{\frac{\omega_{j'j'}}{\omega_{jj}}}.$$

Neighborhood Selection (Meinshausen & Buhlmann, 06)

- A "local" approach:
  simpler; less efficient.
- Fit p individual lasso regressions

$$\min_{\beta_{jj'}} \|\boldsymbol{Y}_j - \sum_{j' \neq j} \beta_{jj'} \boldsymbol{Y}_{j'}\|^2 + \lambda \sum_{j' \neq j} |\beta_{jj'}|, \; j = 1, \ldots, p$$

- Calculate $\hat{\rho}_{jj'} = \boldsymbol{sign}(\hat{\beta}_{jj'}) \sqrt{\hat{\beta}_{jj'} \hat{\beta}_{j'j}}$.

Maximum likelihood

- A "global" approach.
- Regularization is necessary when $p > n$, Yuan & Lin (07).
  - Single Gaussian graphical model: ($\boldsymbol{S}$: Sample covariance)

$$\left(Tr(\boldsymbol{\Omega}\boldsymbol{S}) - \log\det(\boldsymbol{\Omega})\right) + \lambda \sum_{1 \le j < k \le p} |\omega_{jk}|$$

  - Regularization for off-diagonals. Why?
  - Estimation of $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ differ dramatically in a high-d situation.

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4/3 & 2/3 & 1/3 & 1/6 \\ 2/3 & 4/3 & 2/3 & 2/3 \\ 1/3 & 2/3 & 4/3 & 2/3 \\ 1/6 & 1/3 & 2/3 & 4/3 \end{pmatrix}, \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 1 & -1/2 & 0 & 0 \\ -1/2 & 5/4 & -1/2 & 0 \\ 0 & -1/2 & 5/4 & -1/2 \\ 0 & 0 & -1/2 & 1 \end{pmatrix}.$$

  - Yuan & Lin (07) uses an interior point method.
  - Fast algorithms are developed by Friedman et al. (GLasso, 08), and Hsieh et al.(QUIC, 2013), ...

Graphical Lasso (GLasso)

- Gaussian graphical model:
- Regularized negative log-likelihood function for $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ is proportional to

$$\Big(Tr(\mathbf{\Omega}\boldsymbol{S}) - \log\det(\mathbf{\Omega})\Big) + \lambda \sum_{1\le j\le k\le p} |\omega_{jk}|. \tag{1}$$

- Note: $\sum_{1\le j\le k\le p} |\omega_{jk}| = \|\mathbf{\Omega}\|_1$ – the L_1-norm.
- When p is large or close to sample size, the sample covariance $\boldsymbol{S}$ is not a stable estimate:
- Ref: Friedman, Hastie and Tibshirani (07).

Numerical examples, GLasso

```
install.packages("glasso")
library(glasso)
set.seed(100)
    s=c(10,1,5,4,10,2,6,10,3,10)
    S=matrix(0,nrow=4,ncol=4)
    S[row(S)>=col(S)]=s
    S=(S+t(S))
    diag(S)<-10
%   zero<-matrix(c(1,3,2,4),ncol=2,byrow=TRUE)
%   a<-glasso(S,rho=0.01,zero=zero)
    a<-glasso(S,rho=1)
    a
```

L_0-regularization (Shen, Pan & Zhu, 12)

- Likelihood:

$$\Big(Tr(\boldsymbol{\Omega S}) - \log\det(\boldsymbol{\Omega})\Big) + \lambda \sum_{1\le j<k\le p} I(\omega_{jk} \neq 0).$$

- Idea: Same as before. Replace $I(\omega_{jk} \neq 0)$ by truncated L_1-function (TLP) $J_\tau(x) = \min(\frac{|x|}{\tau}, 1)$.
- Computation: DC programming+any convex method.
- R package `MGGM`: Structural Pursuit Over Multiple Undirected Graphs
  `https://rdrr.io/cran/MGGM/`

Inference for undirected graphs

Inference for Graphical Models

- Hypothesis test: $H_0 : \boldsymbol{\Omega}_B = \mathbf{0}$ vs $H_a : \boldsymbol{\Omega}_B \neq \mathbf{0}$, $B = \{(i,j)\}$ is an index set to be specified.
- Example:
  - If $B = \{(1,2)\}$, then $\Omega_B = \omega_{12}$, or

$$H_0 : \omega_{12} = 0, \quad \text{vs} \quad H_a : \omega_{12} \neq 0.$$

  - If $B = \{(1,2),(1,3),..(1,p)\}$, then $\Omega_B = (\omega_{12}, \omega_{13}, \cdots, \omega_{1p})^T$, or

$$H_0 : \omega_{12} = \cdots = \omega_{1p} = 0, \quad \text{vs} \quad H_a : \text{not}.$$

- Issues:
  - How to make a high-dimensional inference, when $p, |B| \to \infty$?
  - How to treat overparametrized models, when # par $> n$?
  - Can we use tests in a low-dimensional situation? Any modifications are needed?

Literature

- Inference for GGM. Jankova & van de Geer (2016)
- Debiased Lasso approach (Zhang & Zhang, 14): Bias correction for low-d parameters.
- Debiased GLasso
  - GLasso: $\hat{\mathbf{\Omega}} = \arg\min_{\mathbf{\Omega}} \Big(Tr(\mathbf{\Omega}\boldsymbol{S}) - \log\det(\mathbf{\Omega}) \Big) + \lambda \sum_{1 \le j \le k \le p} |\omega_{jk}|$
  - $\hat{T} = \hat{\Omega} + \underbrace{(\hat{\mathbf{\Omega}} - \hat{\mathbf{\Omega}}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Omega}})}_{\text{bias corr}}$, $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Omega}}^{-1}$.
  - Asym: $\sqrt{n}(\hat{T}_{ij} - \Omega_{ij})/\sigma_{ij} \to N(0,1)$ when $\lambda \approx \sqrt{\log p/n}$, where $\sigma_{ij}^2 = Var(\hat{T}_{ij})$.
- Issues: How to utilize dependence of multi-components?

Constrained likelihood ratio (Zhu, Shen, & Pan, 20)

- Regularizing only nuisance parameters.
- Higher test efficiency for testing multiple parameters.
- Reducing potential bias due to regularization.
- Test:

$$H_0 : \omega_{ij} = 0, (i,j) \in B \quad (\Omega_B = \mathbf{0}) \quad \text{vs} \quad H_a : \exists (i,j) \in B, \omega_{ij} \neq 0.$$

Constrained MLEs $\widehat{\boldsymbol{\Omega}}^{(0)}$ (H_0) & $\widehat{\boldsymbol{\Omega}}^{(1)}(H_a)$:

$$\begin{aligned} \widehat{\boldsymbol{\Omega}}^{(0)} &= \operatorname{argmin}_{\sum_{(i,j)\notin B} J_T(|\omega_{ij}|)\le K, \boldsymbol{\Omega}_B = 0} Tr(\boldsymbol{S}\boldsymbol{\Omega}) - \log det(\boldsymbol{\Omega}), \\ \widehat{\boldsymbol{\Omega}}^{(1)} &= \operatorname{argmin}_{\sum_{(i,j)\notin B} J_T(|\omega_{ij}|)\le K} Tr(\boldsymbol{S}\boldsymbol{\Omega}) - \log det(\boldsymbol{\Omega}), \end{aligned}$$

- $J_T(z) = \min\left(\frac{|z|}{\tau}, 1\right)$, TLP (Truncated L_1-penalty).
- Estimate (K, τ) by a cross-validation (CV) criterion based on the full model.

Null distributions

- Under regularity conditions on $p, n, |B|$, and Ω^0,
  - Asymptotic normality: If $|B|$ is fixed,

$$\sqrt{n}(\widehat{\Omega}_B^{(1)} - \Omega_B^0) \xrightarrow{d} N(0, \underbrace{\Gamma_B}_{\text{Fisher info}}),$$

  - Wilk's Theorem: If $\omega_{ij} = 0$ for $(i,j) \in B$ & $|B|$ is fixed, then

$$2\Big[L_n(\widehat{\Omega}^{(1)}) - L_n(\widehat{\Omega}^{(0)})\Big] \xrightarrow{d} \chi^2_{|B|}.$$

  - Generalized Wilk's Theorem: If $\omega_{ij} = 0$ for $(i,j) \in B$ & $|B| \to \infty$ as $n \to \infty$, then

$$(2|B|)^{-1/2}\Big[2\Big[L_n(\widehat{\Omega}^{(1)}) - L_n(\widehat{\Omega}^{(0)})\Big] - |B|\Big] \xrightarrow{d} N(0,1).$$

Comments

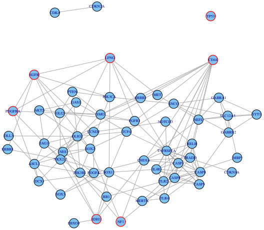
- LR tests (can handle varying dimensions) are more preferable in terms of the power compared to the debias-test. The asymptotic distribution can be the χ^2 or normal depending on the degrees of freedom.
- (Generalized) Wilk's Theorem is generalized to a high-d situation provided that nuisance parameters have sparse structures.

Reconstruction of multiple undirected graphs

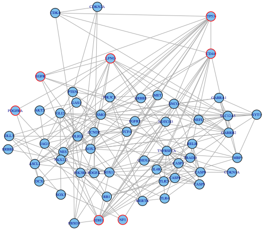
Example: Multiple networks of 4 subtypes of cancers

- 11,861 genes
- 200 patients

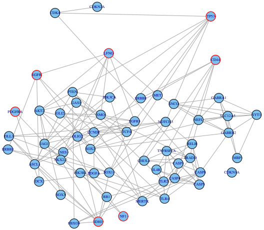
- 4 subtypes
- multiple networks
- similar overall structure



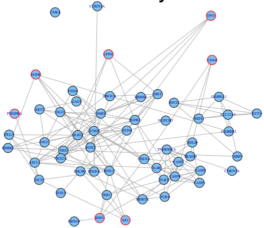
Classical



Mesenchymal



Proneural



Neural

Multiple Gaussian graphical models

- Motivation: Data contains sub-populations
- Model: independent

$$\boldsymbol{Y}_{n_1}^{(l)}, \cdots, \boldsymbol{Y}_{n_l}^{(l)} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_l), l = 1, \cdots, L$$

- Graphs:

$$\mathcal{G}_1, \cdots, \mathcal{G}_L$$

- Parameters of interest:

$$\boldsymbol{\Omega}_l = \boldsymbol{\Sigma}_l^{-1}$$

- Assumptions: $\Omega_1, \cdots, \Omega_L$ are similar.
- Goal: Encourage similarity among $\boldsymbol{\Omega}_l$'s.

Multiple Gaussian graphical models

- Model: $\mathbf{Y}_1^{(l)}, \cdots, \mathbf{Y}_{n_l}^{(l)} \sim N(\mathbf{0}, \mathbf{\Omega}_l^{-1}), l = 1, \cdots, L$
- Joint log-likelihood:

$$\sum_{l=1}^{L} \frac{n_l}{2}\Big(- Tr(\mathbf{\Omega}_l \mathbf{S}_l) + \log\det(\mathbf{\Omega}_l)\Big)$$

- Penalty for Sparsity:

$$\lambda_1 \sum_{1 \le j < k \le p} \sum_{l=1}^{L} J_\tau(|\omega_{jkl}|)$$

- Penalty for grouping:

$$\lambda_2 \sum_{1 \le j < k \le p} \sum_{l \sim l'} J_\tau\Big(|\omega_{jkl} - \omega_{jkl'}|\Big)$$

- Grouping over graph $\mathcal{G}^* = (V^*, E^*)$:
  - $V^* = \{1, \cdots, L\}, l \sim l' \Leftrightarrow (l, l') \in E^*$

    $E^* = \Big\{(l, l') \mid |l - l'| \le 1\Big\}$ — serial (fused) graph

    $E^* = \Big\{(l, l') \mid 1 \le l < l' \le L\Big\}$ — complete graph

L_0-regularization—Truncated ℓ_1 penalty[1]

- Non-convex penalty: truncated ℓ_1 penalty (*TLP*)

$$J_\tau(x) = \min\left(\frac{|x|}{\tau}, 1\right), \ \tau > 0$$

- Relation to ℓ_0:

$$\lim_{\tau \to 0} J_\tau(x) = \mathbb{I}(x \neq 0)$$

- Advantages over ℓ_1:
  - better model selection
  - nearly unbiased

[1] **Shen, Pan \& Zhu, 2012**.

Multiple Gaussian graphical models

- Penalized maximum likelihood:

$$\min . \left(\sum_{l=1}^{L} \frac{n_l}{2} Tr(\boldsymbol{\Omega}_l \boldsymbol{S}_l) - \log \det(\boldsymbol{\Omega}_l) \right) + \sum_{1 \le j < k \le p} p_{jk}\left(\omega_{jk1}, \cdots, \omega_{jkL}\right) \right)$$

- Zhu, Shen & Pan (2014):
  - *TLP* + *nonconvex grouping*:

$$p_{jk}\left(\omega_{jk1}, \cdots, \omega_{jkL}\right) = \lambda_1 \sum_{l=1}^{L} J_\tau\left(|\omega_{jkl}|\right) + \lambda_2 \sum_{l \sim l'}^{L} J_\tau\left(|\omega_{jkl} - \omega_{jkl'}|\right)$$

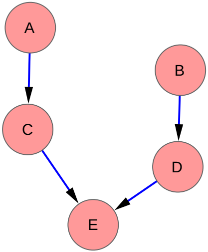
  - (Convex) Lasso version:

$$p_{jk}\left(\omega_{jk1}, \cdots, \omega_{jkL}\right) = \lambda_1 \sum_{l=1}^{L} |\omega_{jkl}| + \lambda_2 \sum_{l \sim l'}^{L} |\omega_{jkl} - \omega_{jkl'}|$$

Causal discovery: DAG reconstruction

Directed acyclic graphical (DAG) model

- A DAG is a directed graph without directed cycles.
  - Nodes correspond to primary variables $(Y_1, \cdots, Y_p)$.
  - Directed edges represent causal (parent-child) relations, $Y_i \to Y_j$.



Adjacency matrix:

$$\begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 0 & 0 & 0 & 0 & 0 \\ B & 0 & 0 & 0 & 0 & 0 \\ C & * & 0 & 0 & 0 & 0 \\ D & 0 & * & 0 & 0 & 0 \\ E & 0 & 0 & * & * & 0 \end{array}$$

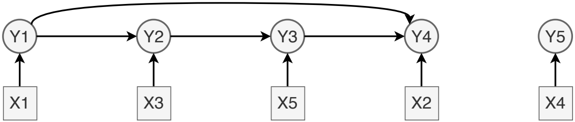
- Local Markov Property specifies a DAG: given its parents, a node is conditionally independent of its non-descendants.

$$Y_j = f_j(Y_{\mathrm{pa}(j)}, \varepsilon_j), \quad j = 1, \ldots, p,$$

$\mathrm{pa}(j)$: parent variables of Y_j; ε_j: error.

Terminology

- Parent-child relation: Y_i is a parent of Y_j: $Y_i \to Y_j$.
- Leaf: no children (terminal node). Root: No parent.
- Ancestral relation: Y_i is an ancestor of Y_j if a ν-directed pathway $Y_i = Y_{k_0} \to Y_{k_1} \to \ldots \to Y_{k_\nu} = Y_j$; $\nu \geq 1$: $Y_i \rightsquigarrow Y_j$.
- Immediate parent-child relation: $\nu < 2$, $Y_i \Rightarrow Y_j$. Special case of ancestral relation.
- Ex: $Y_1 \rightsquigarrow Y_j$, $Y_1 \Rightarrow Y_2 \Rightarrow Y_3 \Rightarrow Y_4$.

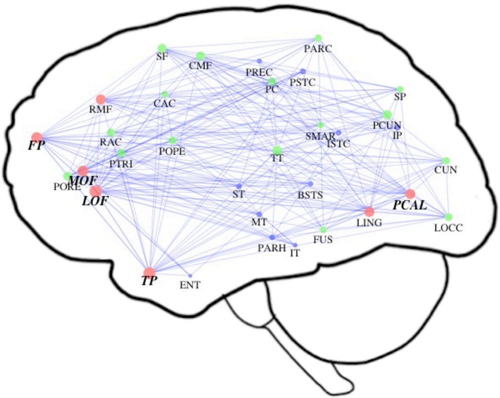


Why DAG models?

- Causal relations modeling:
  - Tools for mediation analysis: Exposure→Mediators→Outcome.
- Applications:
  - Brain network analysis: Effective connectivity of ROI's–casual influences between neurons to explain regional effects in terms of interregional connectivity.
  - Gene regulatory networks: Regulatory relations between genes.
  - Insurance, Marketing, Decision support systems, ...
- Bayesian or causal networks.

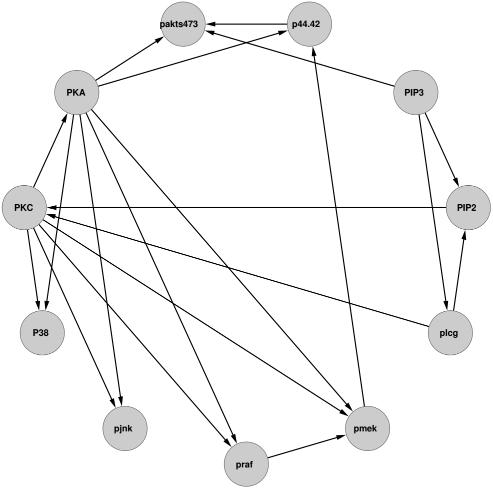
Brain network analysis example

- Functional connectivity
- 30 regions of interest



Cell signaling example

- 11 proteins
- 20 edges



Gaussian models

- Structural equations:

$$Y_j = \sum_{k \neq j} U_{jk} Y_k + \varepsilon_j, \quad \varepsilon_j \overset{ind}{\sim} N(0, \sigma_j^2); \quad j = 1, \ldots, p, \qquad (2)$$

- Parameter: $\boldsymbol{U} = (U_{ij})$ is a real-valued adjacency matrix.
- Causal discovery (Structure learning): Reconstruction of a DAG from data
  - Estimation of $\boldsymbol{U}$ & casual order of $Y_1, \cdots, Y_p$ simultaneously–challenging, could be high-dimensional ($p > n$).
  - Can this be done? To what extent? Identifiability.

Identifiability

- Equal variances: If $\sigma_1 = \cdots = \sigma_p = \sigma$, $\boldsymbol{U}$ is identifiable (Peters & Bühlmann, 13).
- Example: given $Y_1 \sim Y_2$, what is the causal direction?
  - I. Hidden confounding: $Y_1 \Longleftarrow Z \Longrightarrow Y_2$.
  - II. No hidden confounding (in the **current** context):
    i) If $Y_1 \Longleftarrow Y_2$, then
    $Y_1 = Y_2\beta_{21} + \epsilon_1$ and $Y2 = \epsilon_2$,
    $var(Y_1) = var(Y_2\beta_{21}) + var(\epsilon_1) > var(\epsilon_1) = var(\epsilon_2) = var(Y_2)$.
    ii) If $Y_1 \Longrightarrow Y_2$, then ...
- Remarks: it will be easier if i) ϵ's are not normal, or ii) relationships are non-linear.
  ii): Additive noise model (ANM),
  If in truth $Y_1 = f(Y_2) + \epsilon_1$ with ϵ_1 indep of Y_2, then cannot write $Y_2 = g(Y_1) + \epsilon_2$ with ϵ_2 indep of Y_1.
  Example: If $Y_1 = Y_2^2 + \epsilon_1$, then $Y_2 = \sqrt{Y_1 - \epsilon_1} = ...$
  In practice, fit a nonparametric reg model, then test the independence b/w the residuals and the predictor (Jiao et al 18).

Existing methods for observational models

- Search-and-score: Use a model selection criterion to enumerate directions stepwisely.
  Hill Climbing (HC, Korb & Nicholson, 03), Entropy (De Campos,07).
  Comments: Super-exponential candidate DAGs: $O(p^p)$, lack of theory.
- Test-based: Sequential independence tests through edge deletion.
  PC (Spirtes & Glymour, 00).
  Comments: Super-exponential tests in the worst case:$O(p^p)$, Strong faithfulness assumption: restrictive (Uhler et al., 13).
- L_1-regularization: Identify links and choose possible directions.
  Fu & Zhou (JASA, 13), Huang, et. al (IEEE, 13).
- Challenges:
  Computation: Infeasible. Super-exponential DAGs (roughly $p!2^{p^2}$, p is # node). Statistical accuracy: Low due to a huge number of enumerations.

PC algorithm for DAG skeleton

- Principle:
  - If no edge exist between X_1 & X_2 (no local Markov property), in either direction, then X_1 is neither X_2's parent nor its child. But any variable is independent of its non-descendants given its parents. Thus $X_1 \perp X_2 | S$ for some set of variables S.
  - Suppose the converse is true: if $X_1 \perp X_2 | S$, then there cannot be an edge between X_1 and X_2. So there is an edge between X_1 and X_2 iff we cannot make dependence between them to go away, no mater what we condition on.

PC algorithm for DAG skeleton

- Start with a complete undirected graph (with an edge b/w any two nodes).
- For each pair X_1 and X_2, see if $X_1 \perp X_2$. If so, remove the edge between X_1 and X_2.
- For each X_1 and X_2 that are still connected, and each third variable Z; see if $X_1 \perp X_2 | Z$. If so, remove the edge between X_1 and X_2.
- For each X_1 and X_2 that are still connected, and each third or fourth variables Z_1 and Z_2, see if $X_1 \perp X_2 | Z_1, Z_2$. If so, remove their edge.
- ...
- For each X_1 and X_2 that are still connected, see if $X_1 \perp X_2$ given the $p - 2$ other variables. If so, remove their edge.

PC algorithm

- Skeleton of a DAG: an undirected graph ignoring directions of arrows.
- Identifying the skeleton:
  - From complete graph G, $l = -1$,
  - $l = l + 1$,
  - repeat
    - select (new) ordered pair of adjacent nodes $X_1, X_2 \in G$.
    - select (new) neighborhood N of X_1 with size l (if possible)
    - if X_1, X_2 are conditional independence given N, save $N \in M$; delete edge $X_1, X_2 \in G$.
  - until all ordered pairs have been tested; until all neighborhoods are of size smaller than l.
- Finding the DAG: The skeleton can be directed using some rules.
- Test H_0: $\rho_{X_1,X_2|N} = 0$ vs H_a. Test stat: $Z = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{X_1,X_2|N}}{1-\hat{\rho}_{X_1,X_2|N}}\right)$, reject if $\sqrt{n-|N|-3}|Z| > \Phi^{-1}(1-\alpha/2)$ for significance α, $\hat{\rho}$: Sample partial correlation.
- Fisher's transformation: $Z \sim N(0, 1/\sqrt{n-|N|-3})$ under H_0 assuming normality between X_1, X_2 given N.

PC algorithm, Consistency

- (n, p): Sample size, # nodes,
- Distribution: $(X_1, \cdots, X_p) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.
- Nodes: $p = O(n^a)$ with $0 \leq a < \infty$,
- Max # neighbors: $O(n^{1-b})$ with $0 < b < 1$ (sparse),
- Strong faithfulness: $S \subset V \setminus \{i, j\}$,

$$\min_{i,j}\{|Corr(X_i, X_j|X_S)| : Corr(X_i, X_j|X_S) \neq 0\} \geq \kappa;$$

  where $\kappa = O(n^{-d})$ (larger than $n^{-1/2}$), $0 < d < \frac{b}{2}$.
- Thm (Kalisch & Bühlmann, 07, Uhler, Raskutti, Bühlmann, & Yu, 13): Under these assumptions, if $n \to \infty$, then

$$P(\widehat{CPDAG} \neq \text{true CPDAG}) \to 0.$$

- CPDAG (Completed partial DAG): an equivalent class of DAG.

PC algorithm, continued

- R-implementation
- Function pc() in R-package: pcalg:
  https://cran.r-project.org/web/packages/pcalg/index.html
  https://cran.r-project.org/web/packages/pcalg/pcalg.pdf
- R-function pdag2dag: Extend a Partially Directed Acyclic Graph (PDAG) to a DAG:
  https://www.rdocumentation.org/packages/pcalg/versions/2.7-4/topics/pdag2dag
- Reference: Dor and Tarsi (1992). (May not be always possible. Check to see if extendable)

Maximum likelihood

- Global approach: constrained maximum likelihood to estimate all directions simultaneously.
  - Complexity: super-exponentially many candidate DAGs (NP) $(\exp(cp \log p))$.
  - Acyclicity: DAG requirement: Need constraints to solve. Without constraints: Not causal relations.

- Large problem: Achieved reconstruction consistency for DAG's structure as $n, p \to +\infty$, when identifiable.

Constrained maximum likelihood

- Linear causal relations: Parameter: $(\boldsymbol{U} = (U_{ij}), \sigma^2)$

$$Y_j = \sum_{k \neq j} U_{jk} Y_k + \varepsilon_j, \quad \varepsilon_j \overset{ind}{\sim} N(0, \sigma^2); \quad j = 1, \ldots, p,$$

- Constrained maximum likelihood (Yuan, Shen, Pan & Wang, 19): $l(\boldsymbol{U}, \sigma) \to l(\boldsymbol{U})$ by separating $\boldsymbol{U}$ from σ^2. Given a $n \times p$ data matrix $\boldsymbol{Y}$,

$$\min_{\boldsymbol{U}} l(\boldsymbol{U}) = \frac{1}{2} \sum_{j=1}^{p} \sum_{i=1}^{n} \Big(y_{ij} - \sum_{k \neq j} y_{ik} U_{jk} \Big)^2$$

$$\text{subj to } \sum_{j \neq k} I(U_{jk} \neq 0) \leq \kappa, (\textit{sparsity}),$$

$$\boldsymbol{U} \text{ Acyclicity (5) },$$

  $\kappa > 0$: an integer-valued tuning parameter.

- Alternative: Zheng, Dan, Aragam, Ravikumar and Xing (2020).

Acyclicity

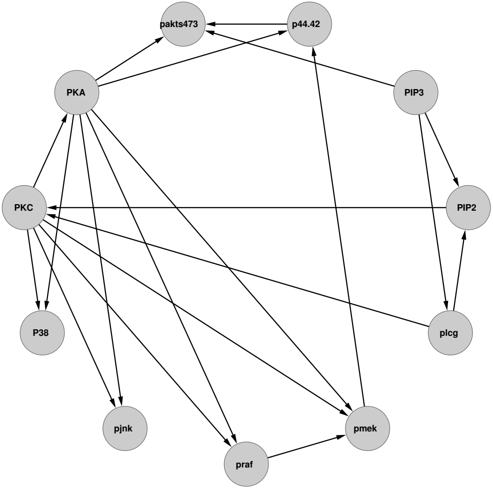
- Yuan, Shen, Pan, & Wang (19): Difference convex programming +constraint reduction (primal/dual)–global method.
  - Acyclicity:

$$\sum_{j_1=j_{L+1}:1\leq k\leq L} I(U_{j_k j_{k+1}} \neq 0) \leq L-1; L=2,\cdots,p. \tag{3}$$

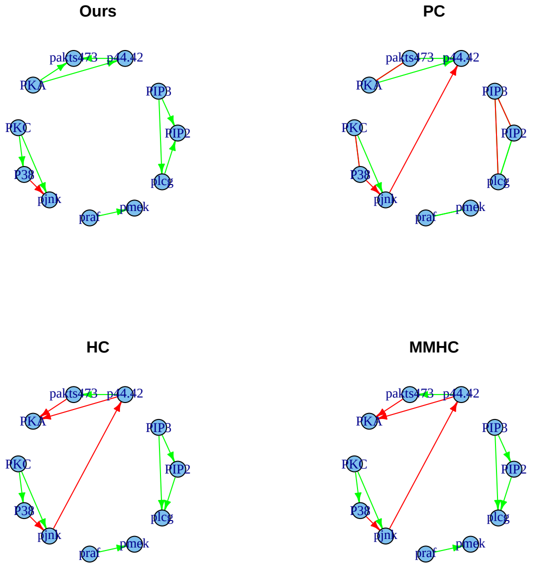
  - Guarantee DAG. Conjecture: DC $\rightarrow$ global minimizer with prob $\rightarrow 1$ as $n, p \rightarrow \infty$.
- R-implementation of constrained MLE: R-package: clrdag https://cran.r-project.org/web/packages/clrdag/index.html

Cell signaling example

- 11 proteins
- 20 edges
- Data: 679 measurements



Analysis of cell signaling data

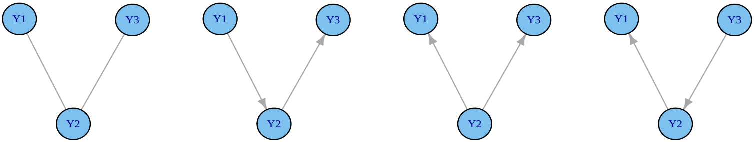


Interventional models

- Add q intervention variables $\{X_1, X_2, \ldots, X_q\}$ into (2).

$$Y_j = \sum_{k \neq j} U_{jk} Y_k + \sum_{l=1}^{q} W_{jl} X_l + \varepsilon_j, \quad \varepsilon_j \sim N(0, \sigma_j^2); \quad j = 1, \ldots, p. \tag{4}$$

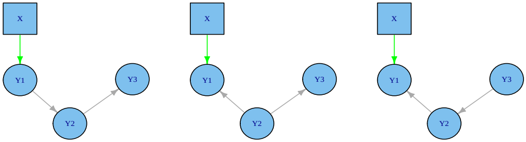
- Unknown interventions: Unknown location and strength W_{lj}.
- Before intervention:



CPDAG (a) (b) (c)

Effect of intervention

- After intervention:



(a) (b) (c)

- can identify (a) from other two if ...
- Cause → outcome: $Y_3 = \alpha Y_2 + \beta Y_1 + \gamma X + Z$.
- What kind of interventions should work?

Instrument and non-instrument interventions

- Intervention: $X_l \to Y_j$ if $W_{lj} \neq 0$ in (7). ($Y_j \nrightarrow X_l$ by prior knowledge but not from model).
  - Instrument: if it satisfies that
    - **(A) Relevance:** intervenes on at least one primary variable.
    - **(B) Exclusion:** does not intervene with more than one primary variables.
  - Non-instrument: not (A) (invalid intervention) or not (B): (multiple: $X_l \to Y_j$, $X_l \to Y_k$, ...

Assumptions for model identifiability

- Thm (Li, Shen, Pan, 20) Model (7) is identifiable if
  - **(1A) (Non-degeneracy)** $E\boldsymbol{X}\boldsymbol{X}^\top$ is positive definite, $\boldsymbol{X} = (X_1, \cdots, X_q)^\top$.
  - **(1B) (Intervention effectiveness)** $Cov(Y_j, X_l \mid \boldsymbol{X}_{\{1,\ldots,q\}\setminus\{l\}}) \neq 0$ when $X_l \to Y_i$ ($Y_i \Rightarrow Y_j$) or, X_l intervenes on an immediate parent of Y_j.
  - **(1C) (Instrument adequacy)** Each primary variable is intervened by at least one instrument.
- No distributional assumption on intervention $\boldsymbol{X}$ (discrete or continuous).
- If either of (1A)-(1C) breaks down, the model is not identifiable.
- Key idea: a peeling algorithm.
  - Identifying all ancestors including parents.
  - Given identifying ancestors, determine parents.
- Can draw inference.
- An application: Zilinskas R, Li C, Shen X, Pan W, Yang T. (2024). Inferring a directed acyclic graph of phenotypes from GWAS summary statistics. *Biometrics*.

Peeling algorithm

- In (7), rewrite $\boldsymbol{V} = \boldsymbol{W}(\boldsymbol{I} - \boldsymbol{U})^{-1}$ as $\boldsymbol{V}^\top$:

$$\boldsymbol{Y} = \boldsymbol{V}^\top \boldsymbol{X} + \varepsilon_V, \quad \varepsilon_V = (\boldsymbol{I} - \boldsymbol{U}^\top)^{-1}\varepsilon \sim N(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}), \qquad (5)$$

- $\boldsymbol{V}_{l\bullet} \in \mathbb{R}^p$ & $\boldsymbol{V}_{\bullet j} \in \mathbb{R}^q$: l-th row & j-th column vectors of $\boldsymbol{V}$.
- Prop: (Causal discovery via $\boldsymbol{V}$) Under Assumptions 1(A)-1(C),
  - (A) $V_{lj} \neq 0$ means X_l intervenes on Y_j or an ancestor of Y_j;
  - (B) $Y_j \to$ leaf node (no children) iff there exists an instrument X_l such that $V_{lj} \neq 0$ & $\|V_{l\bullet}\|_0 = 1$; $l = 1, \ldots, q$.
  - (C) If $V_{lj} \neq 0$ & X_l is an instrument of Y_k, then Y_k is ancestor of Y_l.
- Insight:
- $V_{lj} = \sum_{k=1}^{p} W_{lk}(\underbrace{(\boldsymbol{I})_{kj}}_{par} + \underbrace{(\boldsymbol{U})_{kj}}_{gra-par} + \cdots + \underbrace{(\boldsymbol{U}^{p-1})_{kj}}_{anc})$.
- $V_{lj} \neq 0$ if there exist k, r such that $W_{lk} \neq 0$ and $(\boldsymbol{U}^r)_{kj} \neq 0$.

Estimation of $\boldsymbol{V}$

- For $j = 1, \cdots, p$,

$$\widehat{\boldsymbol{V}}_{\bullet j} = \arg\min_{\boldsymbol{V}_{\bullet j}} (2n)^{-1} \sum_{i=1}^{n} (Y_{ij} - \boldsymbol{V}_{\bullet j}^{\top} \boldsymbol{X}_i)^2 \quad \text{s.t.} \quad \sum_{l=1}^{q} I(V_{lj} \neq 0) \leq K_j; \tag{6}$$

- $1 \leq K_j \leq q \rightarrow$ tuning parameter controlling sparsity & chosen by CV.
- Variable selection (TLP, DC programming)

Peeling algorithm for identifying all ancestral relationships

(1) **(Initialization)** $\hat{\boldsymbol{V}}^{[1]} = \hat{\boldsymbol{V}}$. Begin iteration $h = 0, \cdots$:

(2) **(Leaf-IV pairs)**

(a) Identify rows of $\hat{\boldsymbol{V}}$ with smallest ℓ_0-norm. Restore the indices in $A^{[h]} = \left\{l^* : l^* = \arg\min \|\widehat{V}_{l\bullet}^{[h]}\|_0\right\}$ for all IVs associated with leaf variables.

(b) Identify largest absolute element index of the rows for each $l^* \in A^{[h]}$: $B_{l^*}^{[h]} = \left\{j^* : j^* = \arg\max \left|\widehat{V}_{l^* j}^{[h]}\right|\right\}$ for any $l^* \in A^{[h]}$ to identify all leaf-IV $X_{l^*} \to Y_{j^*}$ pairs.

(3) **(Ancestral relationships)** Identify ancestral relationships $Y_{j^*} \rightsquigarrow Y_k$ if $\widehat{V}_{l^* k} \neq 0$ for all $l^* \in A^{[h]}$ such that $X_{l^*} \to Y_{j^*}$ & Y_k has been already removed for $k \in B^{[h-1]}$.

(4) **(Peeling-off)** Remove leaf-IV pairs. Let $\hat{\boldsymbol{V}}^{[h+1]} = \hat{\boldsymbol{V}}^{[h]}_{\setminus\left(A^{[h]}, B^{[h]}\right)}$, where $\hat{\boldsymbol{V}}^{[h]}_{\setminus\left(A^{[h]}, B^{[h]}\right)}$ is a submatrix by removing rows & columns indexed by $A^{[h]}$ and $B^{[h]}$ from $\hat{\boldsymbol{V}}^{[h]}$.

- [5] **(Termination)** $h \to h + 1$ & repeat Steps 2-4 until removing all Y_j's.

Identifying Pa(j) from An(j)

- Structure eq: $Y_j = \sum_{k \in \mathrm{Pa}(j)} Y_k + \sum_{l \in \mathrm{Int}(j)} W_{jl} X_l + \varepsilon_j$;
- Constrained regression:
  $Y_j = \sum_{k \in \mathrm{An}(j)} U_{jk} Y_k + \sum_{l \in \mathrm{Int}(\mathrm{An}(j))} W_{jl} X_l + \varepsilon_j$.
- For $j = 1, \cdots, p$,

$$(\hat{U}_{jk}, \hat{W}_{jl}) = \arg\min_{U_{jk}, W_{jl}} (2n)^{-1} \sum_{i=1}^{n} (Y_{ij} - \sum_{k \in \mathrm{An}(j)} U_{jk} Y_{ik} - \sum_{l \in \mathrm{Int}(\mathrm{An}(j))} W_{jl}$$

$$\text{s.t.} \quad \sum I(|U_{jk}| \neq 0) + I(|W_{jl}| \neq 0) \leq K_j^{'};$$

- $\mathrm{An}(j) = \{k : \hat{U}_{jk} \neq 0\}$.

Extension: Interventional models with confounders

- Chen L, Li C, Shen X, Pan W (2023). Discovery and Inference of a Causal Network with Hidden Confounding. JASA.
- Add q intervention variables $\{X_1, X_2, \ldots, X_q\}$ into (2).

$$Y_j = \sum_{k \neq j} U_{jk} Y_k + \sum_{l=1}^{q} W_{jl} X_l + h_j + \varepsilon_j, \quad \varepsilon_j \sim N(0, \sigma_j^2); \quad j = 1, \ldots, p. \tag{7}$$

- $h_1, \cdots, h_p \sim N(0, \Sigma)$: unmeasured confounders.
- Unmeasured confounders: h_j; $j = 1, \ldots, p$.
- Unknown interventions: Unknown location and strength W_{lj}.
- Model is not identifiable without IVs. Use IV to treat confounding effects.
- Alternative: Li C, Shen X, Pan W. (2023). Nonlinear causal discovery with confounders. JASA. As in (7),

$$Y_j = f_j(Y_{pa(j)}) + h_j + \varepsilon_j,$$

References

- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *Elements of Statistical Learning*. Springer. Chapter 17
- Jordan, M.I. (2004). Graphical models. *Statistical Sciences*, **19**, 140-155
- Janková, Jana and van de Geer, Sara (2018). Inference in high-dimensional graphical models. arXiv preprint arXiv:1801.08512.
- Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.
- d'Aspremont, A., Banerjee, O., and Ghaoui. EL. (2008) First-order methods for sparse covariance selection. *SIAM. J. Matrix Anal. & Appl.*, **30**, 56.

- Li, C., Shen, X., and Pan, W. (2020). Likelihood ratio tests for a large directed acyclic graph. *Journal of American Statistical Association*. **115**, 1304-1319.
- Li, C., Shen, X., and Pan, W. (2023). Inference for a large directed acyclic graph with unspecified interventions. arXiv:2110.03805. *Journal of Machine Learning Research*, 24, 73.
- Guo, J., Levina, E., Michailidis, G. and Zhu (2010). Joint estimation of multiple graphical models. *Biometrika*, **98**, 1-15.
- Mazumder, R., and Hastie, T. (2012) The graphical lasso: New insights and alternatives. *Electrical Journal of Statistics*, **6**, 2125-2149.
- Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances, *Biometrika*, 101, 219-228.
- Zhu, Y., Shen, X. and Pan, W. (2014). Structural pursuit over multiple undirected graphs. *Journal of American Statistical Association*. **109**, 1683-1696.