Model selection for colored graphical models for paired data

Dung Ngoc NGUYEN & Alberto ROVERATO

Statistical methods and models for complex data — Poster Session: 21 September 2022

Paired data

Paired data problem: every variable is uniquely associated with a homologous, or twin, variable.

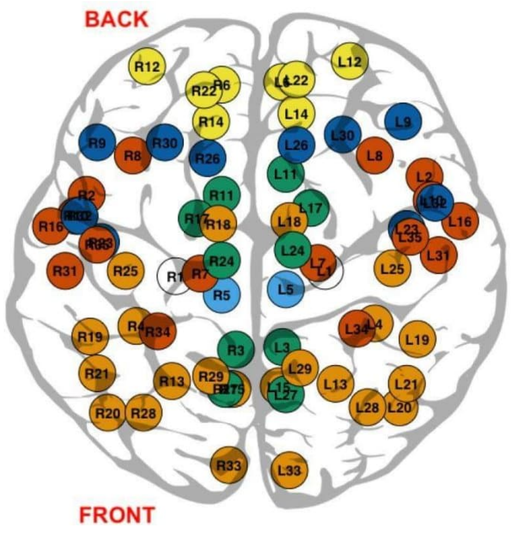


Figure 1. Example of ROI locations on the brain. Every ROI on the left hemisphere is associated with an ROI on the right hemisphere, which gives the pairs $(L_i, R_i)_{i=1,\dots,35}$. Different colors correspond to distinct brain regions.

Hence, for paired data, $\mathbf{Y}_V$ can be partitioned as $(\mathbf{Y}_L, \mathbf{Y}_R)^T$, and we consider and assume that $L = \{1, \dots, q\}$ and $R = \{1', \dots, q'\}$ where $i' = q + i$ and $q = p/2$ so that Y_i is homologous to $Y_{i'}$ with $1 \leq i \leq q$.

Gaussian graphical models (GGMs)

Let $G = (V, E)$ be an undirected graph with the vertex set V and the edge set E. Then, $\mathbf{Y}_V$ is said to satisfy the Gaussian graphical model if $\mathbf{Y}_V \sim \mathcal{N}(\mu, \Sigma)$ and $\mathbf{Y}_V$ is Markov w.r.t G, that is $(i, j) \notin E$ implies $\theta_{ij} = 0$ where $\Theta = (\theta_{ij})_{i,j \in V} = \Sigma^{-1}$.

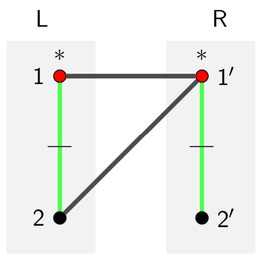
Colored GGMs for paired data

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a colored version of G where $\mathcal{V}$ is a partition of V into vertex color classes; similarly, $\mathcal{E}$ is a partition of E into edge color classes.

Colored graphs for paired data (PD-CGs)

The PD-CG $\mathcal{G}$ contains two types of color classes:

- **atomic class** that is a color class of cardinality one;
- **twin-pairing class** that is a color class containing a pair of twin vertices or a pair of twin edges.

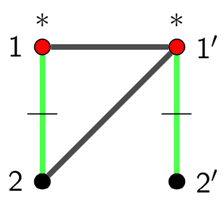


Example. Consider the PD-CG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with

$$\mathcal{V} = \{ \underbrace{\{1, 1'\}}_{twin-pairing}, \underbrace{\{2\}, \{2'\}}_{atomic} \}, \quad \mathcal{E} = \{ \underbrace{\{(1, 2), (1', 2')\}}_{twin-pairing}, \underbrace{\{(1, 1')\}, \{(2', 1')\}}_{atomic} \}.$$

RCON models for paired data (PD-RCONs)

PD-RCON models are Gaussian graphical models with additional equality constraints on the concentration matrix implied by a PD-CG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.



$$\Theta_{\mathcal{G}} = \begin{pmatrix} \theta & \theta & \theta_{11'} & 0 \\ & \theta_{22} & \theta_{21'} & 0 \\ & & \theta & \theta \\ & & & \theta_{2'2'} \end{pmatrix}$$

$$\mathcal{G} = (\mathcal{V}, \mathcal{E})$$

Challenges

Learn the graphical models for paired data:

1. learn the structure of the graph;
2. learn the colorings of the vertices;
3. learn the colorings of the edges both between and across the left and the right parts of the network.

Difficulties

1. **Dimension of the search space,** for example:

$$\underset{\substack{\text{complete graph} \\ \text{on } p \text{ vertices}}}{1} \ll \underset{\substack{\text{complete graphs} \\ \text{for paired data}}}{2^{(p/2)^2}}$$

2. **The exploration of the space:**
   - the structure of the search space forms a lattice but it is not distributive,
   - the neighbors of a model cannot be efficiently specified.

Structure of models space of PD-CGMs

It is useful to embed search spaces with a partial order. Naturally, the order is the **model inclusion order**: a model is "larger" than any of its submodels.

Consider two PD-CGMs characterized by $\mathcal{G} = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ and $\mathcal{H} = (\mathcal{V}_\mathcal{H}, \mathcal{E}_\mathcal{H})$. Then, following [2], $\mathcal{G} \preceq_s \mathcal{H}$ if and only if

• $E_\mathcal{H} \supseteq E_\mathcal{G}$, • $\mathcal{V}_\mathcal{H} \preceq_f \mathcal{V}_\mathcal{G}$, • $\mathcal{E}_\mathcal{H} \preceq_f \mathcal{E}_\mathcal{G} \cup \{\{E_\mathcal{H} \setminus E_\mathcal{G}\}\}$,

where $\preceq_f$ is the *refinement* order and $E_\mathcal{G}, E_\mathcal{H}$ are the sets of uncolored edges of $\mathcal{G}, \mathcal{H}$, respectively.

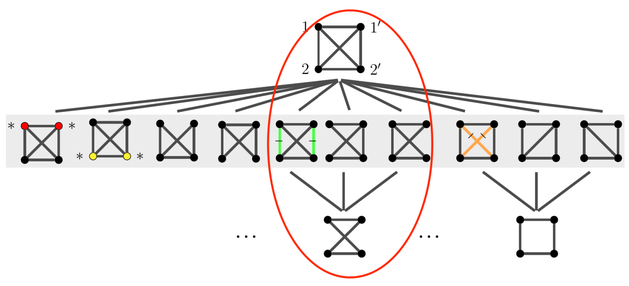


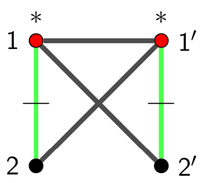
Figure 2. A part of Hasse diagram of the lattice structure of PD-CGs with 4 vertices based on the model inclusion order. The highlighted graphs are the neighbors of the model on the top. The circled graphs form the so-called diamond structure.

Therefore, the family of PD-CGMs, under the model inclusion order, forms a complete, non-distributive lattice.

Novel partial order for PD-CGs

The **twin correspondence** $\tau(\cdot)$ is a function of $i \in V$ that is $i + q$ if $i \in L$, and $i - q$ if $i \in R$. Moreover, for i, j ∈V, $\tau((i, j)) = \big(\tau(i), \tau(j)\big)$.

We say i, j are **twin vertices** i, j if $\tau(i) = j$ or $i = \tau(j)$, and $(i, j), (k, l)$ are **twin edges** if $\tau(i, j) = (k, l)$ or $(i, j) = \tau(k, l)$.



- $\mathbb{L} = \{i \in V \text{ s.t. } \{i\}, \{\tau(i)\} \in \mathcal{V}\}$, e.g. $\mathbb{L} = \{2\}$.
- $\mathbb{E} = \{(i, j) \in E \text{ s.t. } \{(i, j)\}, \{\tau(i, j)\} \in \mathcal{E}\}$, e.g. $\mathbb{E} = \{(1, 2')\}$.

An alternative and equivalent representation of PD-CGs.

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \quad \Longleftrightarrow \quad \mathcal{G} = (V, E, \mathbb{L}, \mathbb{E})$$

$\mathcal{V}$: a partition of V; $\mathcal{E}$: a partition of E

Twin order

For two PD-CGs $\mathcal{G}$ and $\mathcal{H}$, we say $\mathcal{G} \preceq_\tau \mathcal{H}$ if and only if

• $E_\mathcal{G} \subseteq E_\mathcal{H}$, • $\mathbb{L}_\mathcal{G} \subseteq \mathbb{L}_\mathcal{H}$, • $\mathbb{E}_\mathcal{G} \subseteq \mathbb{E}_\mathcal{H}$.

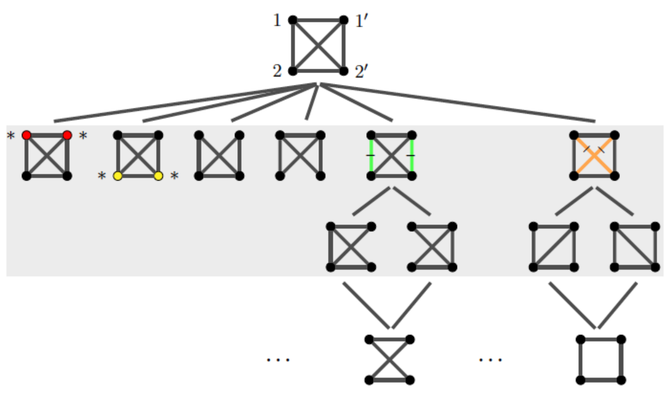


Figure 3. A part of Hasse diagram of the lattice structure of PD-CGs with 4 vertices based on the twin order. The highlighted graphs are the neighbors of the model on the top.

Theorem. The family of PD-CGs under the twin order forms a complete and distributive lattice.

Proposition. For two PD-CGs $\mathcal{G}, \mathcal{H}$, if $\mathcal{G} \preceq_s \mathcal{H}$ then $\mathcal{G} \preceq_\tau \mathcal{H}$.

Backward elimination stepwise procedure with coherent steps

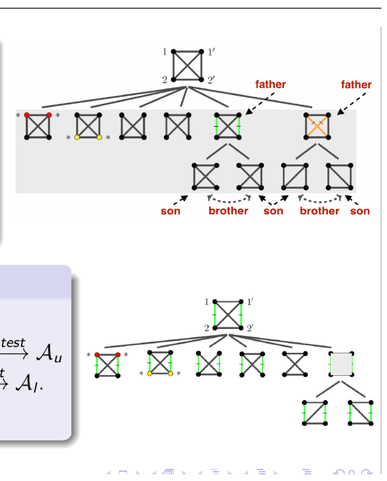
Initial stage

1. $\mathcal{M}^\star \longleftarrow$ the saturated model
2. Neighboring submodels:
   U) $\mathcal{N}_u \xrightarrow{test} \mathcal{A}_u$
   L) $\mathcal{N}_l \xrightarrow{coherence} \mathcal{N}_l' \xrightarrow{test} \mathcal{A}_l$
3. Update $\mathcal{M}^\star$ from $\mathcal{A}_u \cup \mathcal{A}_l$

Iterative stages

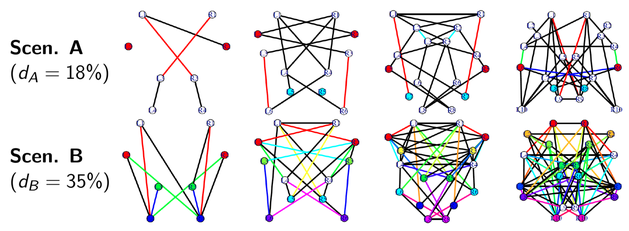
1. Neighboring submodels:
   U) $\mathcal{N}_u = \mathcal{M}^\star \wedge \mathcal{A}_u \cup \{\text{son}_1 \wedge \text{son}_2\} \xrightarrow{test} \mathcal{A}_u$
   L) $\mathcal{N}_l' = \mathcal{M}^\star \wedge \mathcal{A}_l \setminus \{\text{its brother}\} \xrightarrow{test} \mathcal{A}_l$.
2. Update $\mathcal{M}^\star$ from $\mathcal{A}_u \cup \mathcal{A}_l$.

Note: $m \wedge A = \{m \wedge a, \ \forall a \in A\}$.



Numerical experiment

- We generate 100 independent samples with different numbers of variables p varying in $\{8, 12, 16, 20\}$. The recorded results are taken on average over 20 simulated data sets.



Recorded results

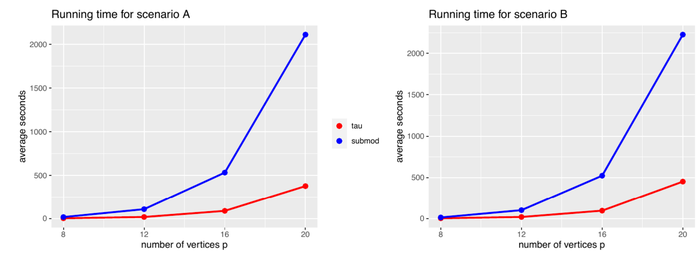


Figure 4. Elapsed time from the backward elimination procedures based on the twin order $\preceq_\tau$ (illustrated in red) and the model inclusion $\preceq_s$ (illustrated in blue) of two scenarios A (on the left) and B (on the right).

Table 1. Performance measures of the model selection procedure for the lattice structure equipped by the partial orders $\preceq_\tau$ and $\preceq_s$.

Scenario	p	Order	Graph structure #edges	ePPV%	eTPR%	eTNR%	Symmetries #sym	sPPV%	sTPR%	sTNR%	Time(s)
A	8	$\preceq_\tau$	7(2)	76.68	100.00	91.52	2(1)	41.67	95.00	89.44	4.02
		$\preceq_s$	7(2)	75.41	100.00	91.30	2(1)	46.67	95.00	85.56	17.20
	12	$\preceq_\tau$	17(3)	71.22	97.92	90.37	6(1)	15.99	90.00	87.61	18.77
		$\preceq_s$	17(3)	70.23	98.75	90.00	5(1)	17.34	90.00	83.91	108.55
	16	$\preceq_\tau$	27(4)	74.83	88.64	92.70	9(1)	18.53	85.00	89.43	89.10
		$\preceq_s$	28(4)	70.98	87.05	91.48	8(1)	19.32	77.50	84.77	532.02
	20	$\preceq_\tau$	44(8)	64.24	82.21	89.49	16(3)	13.47	70.00	86.18	378.74
		$\preceq_s$	46(7)	60.11	78.97	88.04	13(3)	11.97	51.67	80.00	2102.20
B	8	$\preceq_\tau$	11(2)	84.54	89.50	89.72	5(1)	64.08	93.33	92.50	3.78
		$\preceq_s$	11(2)	83.59	89.00	89.44	4(1)	64.83	85.00	85.83	14.56
	12	$\preceq_\tau$	23(4)	81.78	80.00	89.65	9(2)	56.28	79.17	87.35	19.43
		$\preceq_s$	23(4)	81.25	78.48	89.53	7(2)	63.26	73.33	83.53	101.94
	16	$\preceq_\tau$	34(5)	72.49	57.86	87.63	12(2)	52.38	64.00	86.09	96.02
		$\preceq_s$	31(4)	74.50	55.24	89.49	9(2)	63.36	54.00	82.97	522.97
	20	$\preceq_\tau$	51(9)	69.74	53.41	87.02	18(2)	48.17	54.38	84.07	451.71
		$\preceq_s$	48(7)	67.81	48.64	87.22	12(2)	52.97	39.38	78.98	2226.35

Concluding remarks:

- The model selection procedure on the twin lattice $\preceq_\tau$ is considerably faster than the similar approach on $\preceq_s$, as shown in Figure 4 and Table 1.
- With $p = 36$, the procedure with the twin order ≈ 7 hours whereas the existing procedure is infeasible.
- The procedure with the twin order tends to perform better when many symmetries are present.

Application on fMRI data

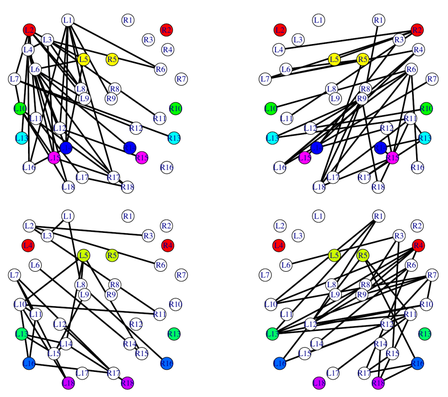


Figure 5. Colored graphical representations for 36 brain regions in anterior temporal and frontal lobes between two hemispheres.

References

[1] Davey, B. A. and Priestley, H. A. (2002) *Introduction to lattices and order*. Cambridge University Press.

[2] Gehrmann, H. (2011) Lattices of graphical Gaussian models with symmetries. *Symmetry*, **3(3)**, 653 – 679.

[3] Hojsgaard, S. and Lauritzen, S. L. (2008) Graphical Gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70(5)**, 1005 – 1027.

[4] Ranciati, S., Roverato, A. and Luati, A. (2021) Fused graphical lasso for brain networks with symmetries. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **70(5)**, 1299 – 1322.

[5] Roverato, A. and Nguyen, D. N. (2022) Model inclusion lattice of colored Gaussian graphical models for paired data. *Proceedings of Machine Learning Research*.

[6] Nguyen, D. N. and Roverato, A. Stepwise model search for multiple Gaussian graphical models for paired data *(working paper)*.

Contact information

Dung Ngoc NGUYEN, Postdoctoral Research Fellow.
Department of Statistical Sciences, University of Padova.
ngocdung.nguyen@unipd.it
https://ngocdung-nguyen.github.io/
https://github.com/NgocDung-NGUYEN