On model selection of coloured Gaussian graphical models for paired data

Dung Ngoc NGUYEN & Alberto ROVERATO

IMS International Conference on Statistics and Data Science (ICSDS) — Poster Session: December 13-16, 2022

Paired data

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

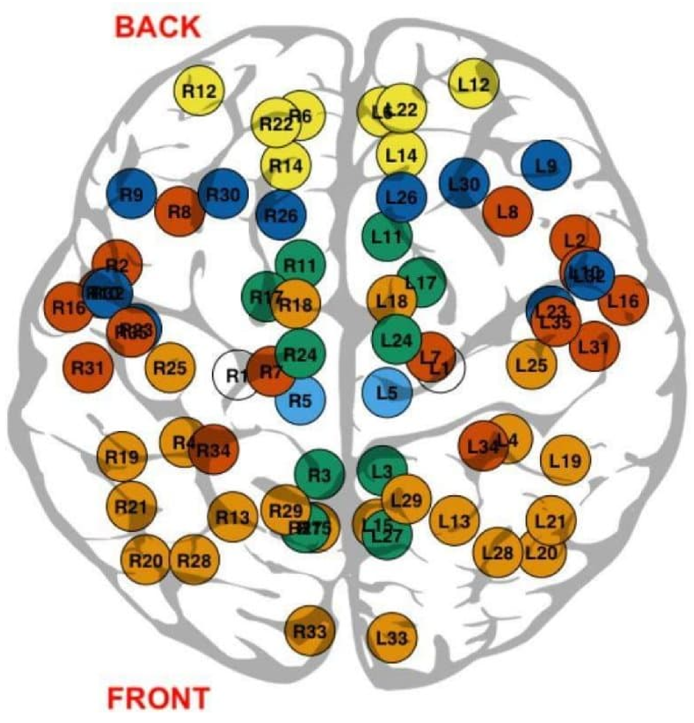


Figure 1. Example of ROI locations on the brain. Every ROI on the left hemisphere is associated with an ROI on the right hemisphere, which gives the pairs $(L_i, R_i)_{i=1,\dots,35}$. Different colors correspond to distinct brain regions.

Hence, for paired data, $\mathbf{Y}_V$ can be partitioned as $(\mathbf{Y}_L, \mathbf{Y}_R)^T$, and we consider and assume, w.l.g., 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 $1 \le i \le q$.

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 $\Theta = (\theta_{ij})_{i,j \in V} = \Sigma^{-1}$.

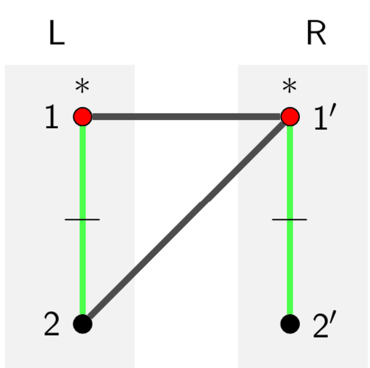
Coloured GGMs for paired data

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

Coloured 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.

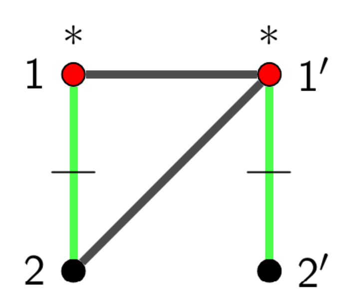


Example. Consider the PD-CG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with

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

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

Learning the graphical models for paired data requires:

1. learning the structure of the network;
2. learning the symmetries of the vertices;
3. learning the symmetries of the edges both between and across parts of the network.

Difficulties

1. **Dimension of the search space:** highly increases, e.g.,

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

2. **The exploration of the space:** considerably complex
   - the structure of the search space behaves like a partition lattice $\longrightarrow$ non-distributive,
   - the neighbors of a model cannot be efficiently specified.

Structure of model spaces of PD-CGMs

Gehrmann (2011) investigated and showed that the search space of coloured GGMs is naturally embedded with the model inclusion: a model is "larger" than any of its submodels.

Consider two PD-RCONs characterized by $\mathcal{G} = (\mathcal{V}_{\mathcal{G}}, \mathcal{E}_{\mathcal{G}})$ and $\mathcal{H} = (\mathcal{V}_{\mathcal{H}}, \mathcal{E}_{\mathcal{H}})$. Then, $\mathcal{G} \preceq_s \mathcal{H}$ 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 uncoloured edges of $\mathcal{G}, \mathcal{H}$, respectively.

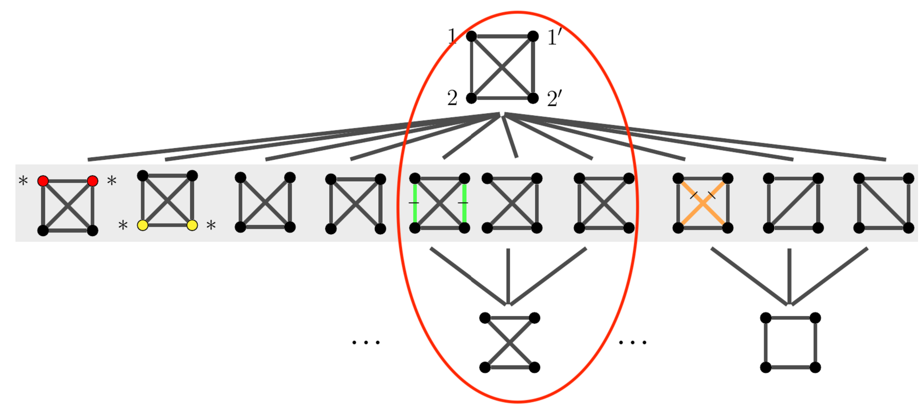


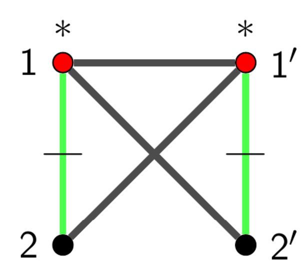
Figure 2. A part of Hasse diagram of lattice structure of PD-CGs with 4 vertices based on the model inclusion. The highlighted graphs are the neighbours of the model on the top. The circled graphs form the so-called diamond structure.

The family of PD-RCONs, under the model inclusion, forms a complete, non-distributive lattice, see *Roverato and Nguyen (2022)*.

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 $\in V$, $\tau((i,j)) = \big(\tau(i), \tau(j)\big)$.

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.

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

a partition of V — 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

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

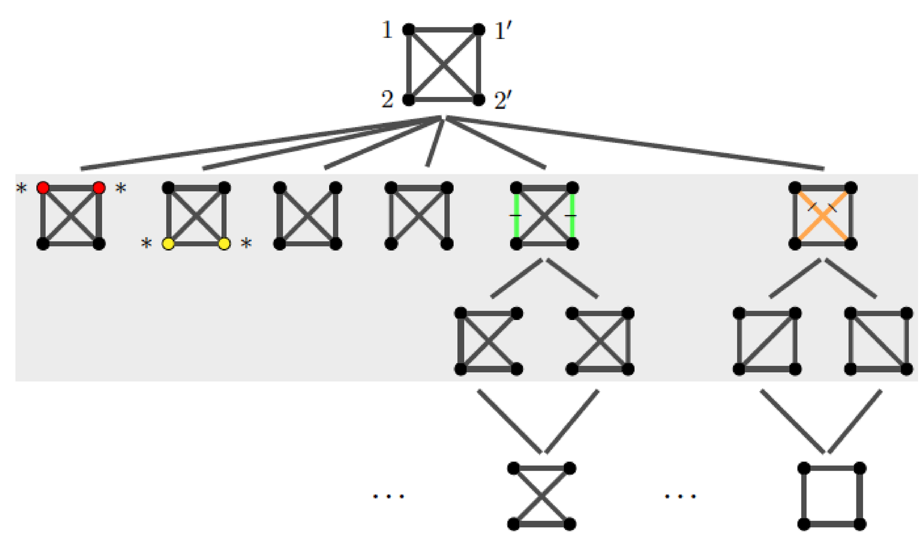


Figure 3. A part of Hasse diagram of the lattice structure of PD-CGs with 4 vertices based on the twin order. The highlighted graphs are the neighbours of the model on the top.

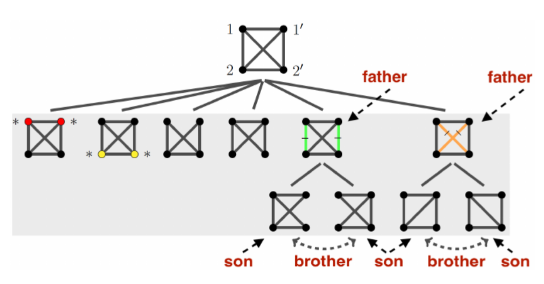
Theorem. The family of PD-CGs under the twin order forms a complete and distributive lattice.

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

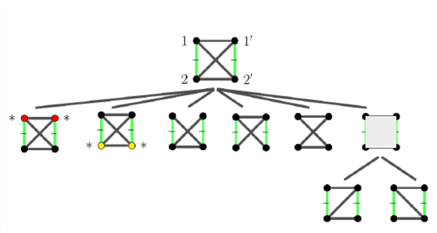
Initial stage

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 $\mathcal{A}_u \cup \mathcal{A}_l$.



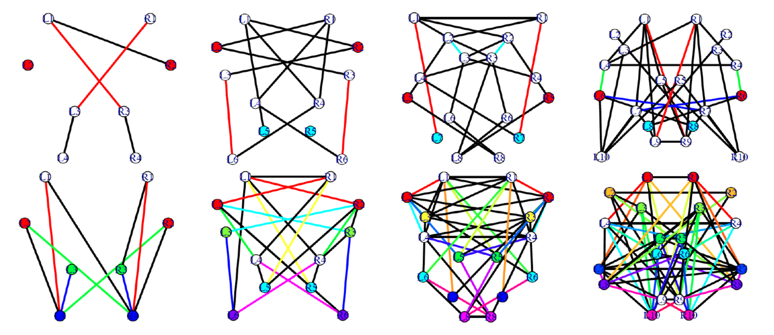
Note: $m \wedge A = \{m \wedge a, \ \forall a \in A\}$.

Numerical experiment

- We generate 100 independent samples with different numbers of variables p varying in $\{8, 12, 16, 20\}$. Figure 1 summarizes the average results over all 20 repetitions of the simulated data.

Scen. A ($d_A = 18\%$)

Scen. B ($d_B = 35\%$)



Recorded results

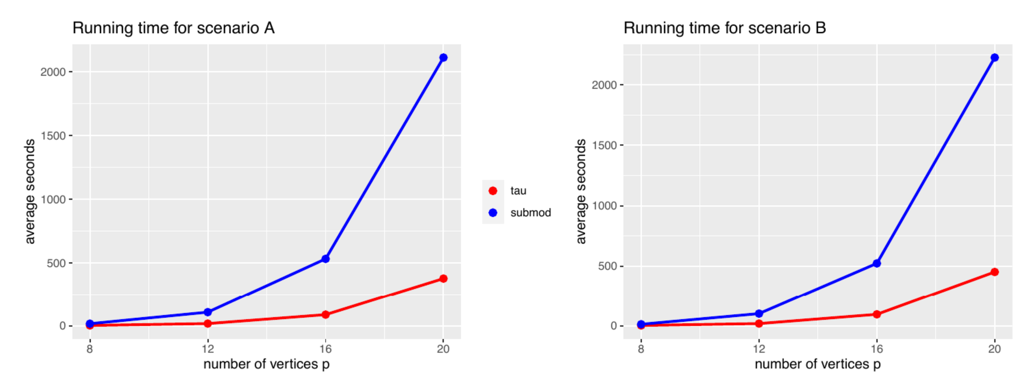


Figure 4. Computational time from the stepwise procedures on the twin lattice $\preceq_\tau$ (illustrated in red) and the model inclusion lattice $\preceq_s$ (illustrated in blue) of two scenarios A (on the left) and B (on the right).

Table 1. Performance measures of the stepwise procedures for the structures of the lattices equipped by the partial orders $\preceq_\tau$ and $\preceq_s$.

S	p	Order	Graph structure #edges	ePPV%	eTPR%	eTNR%	Symmetries #sym	sPPV%	sTPR%	sTNR%	Time(s)	#models
A	8	$\preceq_t$	7(2)	76.68	100.00	91.52	2(1)	41.67	95.00	89.44	4	273
		$\preceq_s$	7(2)	75.41	100.00	91.30	2(1)	46.67	95.00	85.56	17	580
	12	$\preceq_t$	17(3)	71.22	97.92	90.37	6(1)	15.99	90.00	87.61	19	1300
		$\preceq_s$	17(3)	70.23	98.75	90.00	5(1)	17.34	90.00	83.91	109	2985
	16	$\preceq_t$	27(4)	74.83	88.64	92.70	9(1)	18.53	85.00	89.43	89	4245
		$\preceq_s$	28(4)	70.98	87.05	91.48	8(1)	19.32	77.50	84.77	532	10554
	20	$\preceq_t$	44(8)	64.24	82.21	89.49	16(3)	13.47	70.00	86.18	379	10212
		$\preceq_s$	46(7)	60.11	78.97	88.04	13(3)	11.97	51.67	80.00	2102	27356
B	8	$\preceq_t$	11(2)	84.54	89.50	89.72	5(1)	64.08	93.33	92.50	3	264
		$\preceq_s$	11(2)	83.59	89.00	89.44	4(1)	64.83	85.00	85.83	15	486
	12	$\preceq_t$	23(4)	81.78	80.00	89.65	9(2)	56.28	79.17	87.35	19	1230
		$\preceq_s$	23(4)	81.25	78.48	89.53	7(2)	63.26	73.33	83.53	102	2729
	16	$\preceq_t$	34(5)	72.49	57.86	87.63	12(2)	52.38	64.00	86.09	96	4259
		$\preceq_s$	31(4)	74.50	55.24	89.49	9(2)	63.36	54.00	82.97	523	10247
	20	$\preceq_t$	51(9)	69.74	53.41	87.02	18(2)	48.17	54.38	84.07	452	10300
		$\preceq_s$	48(7)	67.81	48.64	87.22	12(2)	52.97	39.38	78.98	2226	26960

Concluding remarks:

- **Accuracy:** The two procedures have similar behaviour in terms of the identification of zeros.
- The procedure with the twin order tends to perform better when many symmetries are present.
- **Efficiency:** The computational time required by the procedure on the twin lattice is $15 - 20\%$ of the time required by the procedure on the model inclusion lattice.
- With $p = 36$, the procedure with the twin order ≈ 7 hours whereas the existing procedure is infeasible.

Application to fMRI data

Subject 14 (19 years old)

Subject 15 (57 years old)

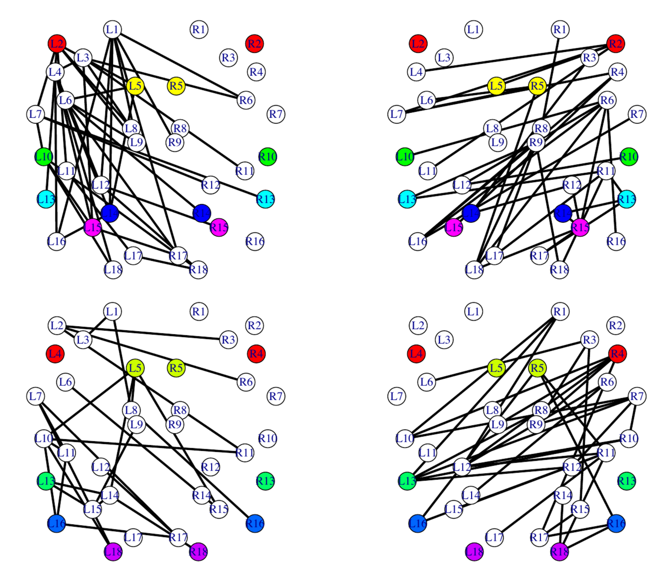


Figure 5. Coloured graphical representations for 36 brain regions in anterior temporal and frontal lobes between two hemispheres.

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 coloured Gaussian graphical models for paired data. *Proceedings of The 11th International Conference on Probabilistic Graphical Models*, **PMLR 186**, 133 – 144.

[6] Roverato, A. and Nguyen, D. N. 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`