

Order of separability in graphs and the uPC-algorithm

Dhafer Malouche

LEGI-EPT-ESSAI,

Uty, November 7th Carthage, Tunisia,

`dhafer.malouche@essai.rnu.tn`

Sylvie Sevestre-Ghalila

U2S-ENIT, Uty El Manar, Tunisia,

MAP5, Uty Paris 5, France,

`sylvie.sevestre@math-info.univ-paris5.fr`

Bayesian Focus Week, SAMSI, Oct 30th-Nov 3rd, 2006

Outline

I Problem

II Separability in Graphs

III *faithful* Gaussian Graphical Models

IV Main result

V The uPC-algorithm

VI Simulation study

I The Problem and Data

Let $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$, where $\mathbf{X}^{(i)} = (X_1(i), \dots, X_p(i))'$, be an i.i.d sample of size n from $\mathbf{X} = (X_1, \dots, X_p) \sim \mathcal{N}_p(\mu, K^{-1})$, where K is the precision matrix.

We may have $n \leq p$.

Example Data of expression levels of genes :

- the observations are n experiments and for each experiment the expression of p genes is observed : $X_j(i)$ is the expression level of the j^{th} gene for the i^{th} expression.
- (*Wu et al., 2003*) this kind of data could be considered as sampled from $\mathcal{N}_p(\mu, \Sigma = K^{-1})$

Problem : Gene interaction detection ?

Quantification of interaction
between two genes

\longleftrightarrow **partial correlation coefficient**

$$\rho_{ij \cdot V \setminus \{i,j\}} = \frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}}$$

Conditional independence : if $\mathbf{X}_{-ij} = (X_l, l \in V \setminus \{i, j\})$

$$\rho_{ij \cdot V \setminus \{i,j\}} = 0 \iff X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_{-ij}$$

Gene interactions \iff genetic network \iff **Graphical Gaussian Model**

II Separability on graphs

An undirected graph $G = (V, E)$ is a pair of sets where V is the set of vertices and E is the set of edges. Let $i, j \in V$, we write $i \sim_G j$ when i and j are **adjacent** in G .

Let $i \not\sim_G j$ in a connected graph G and $S \subseteq V \setminus \{i, j\}$

- S is a **separator** of i and j if any path between i and j intersect S . Any $S' \supseteq S$ is also a separator of i and j .
- S is called **minimal separator** of i and j , if for any $S' \subset S$ and $S' \neq S$, S' is not a separator of i and j .
- We denote by $\text{ms}_G(i, j)$ the set of minimal separators of i, j in G .
 - If i, j are in two different connected component of G , then $\text{ms}_G(i, j) = \emptyset$.
 - We denote by $q_G(i, j)$ the common cardinality of the elements of $\text{ms}_G(i, j)$

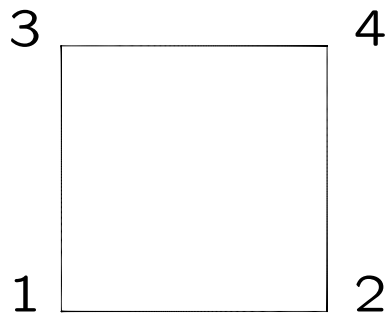
Order of separability of a graph

Definition : The order of separability of a given undirected connected and non-complete graph $G = (V, E)$ is

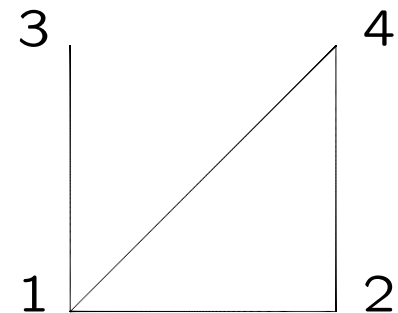
$$\text{os}(G) = \max_{i \not\sim_G j} q_G(i, j)$$

Example :

$$\text{os}(G) = 2$$



$$\text{os}(G) = 1$$



III faithful Graphical Gaussian Modeling

Whittaker (1990), Lauritzen (1996), Edwards (2000).

For any $P = \mathcal{N}_p(\mu, \Sigma = K^{-1})$ we associate $G = (V, E)$ where V is the set of variables, $|V| = p$ and E is constructed using the *Pairwise Markov propriety* (PMP) w.r.t G

$$\forall i, j \in V, i \not\sim_G j \iff X_i \perp\!\!\!\perp_P X_j \mid \mathbf{X}_{-ij}$$

Also, P satisfies *Global Markov propriety* (GMP) w.r.t G .

$$\text{If } S \text{ is a separator of } i, j \text{ in } G \implies X_i \perp\!\!\!\perp_P X_j \mid \mathbf{X}_S$$

Lauritzen (1996)

Theorem If P has a positive density, $(\text{GMP}) \iff (\text{PMP})$

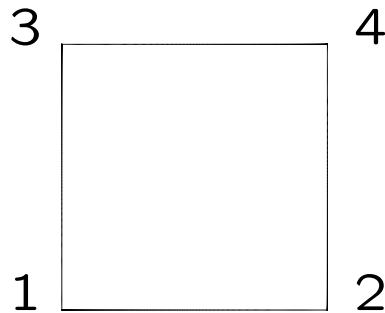
faithfulness

A probability distribution P is faithful to graph G if GMP becomes an equivalence :

$$\text{If } S \text{ is a separator of } i, j \text{ in } G \iff X_i \perp\!\!\!\perp_P X_j \mid \mathbf{X}_S$$

Example :

$$K = \begin{pmatrix} 5 & 2 & 1 & 0 \\ 2 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}$$



$$X_1 \perp\!\!\!\perp X_4 \mid \mathbf{X}_{\{2,3\}}$$

$$X_2 \perp\!\!\!\perp X_3 \mid \mathbf{X}_{\{1,4\}}$$

Independence or Covariance (bidirected) graphical models

Cox and Wermuth (1996), Richardson and Spirtes (2002)

For any $P = \mathcal{N}_p(\mu, \Sigma)$ we associate $G_0 = (V, E_0)$ where V is to the set of variables, $|V| = p$ and E_0 is constructed using the bi-*Pairwise* Markov propriety (biPMP) w.r.t G

$$\forall i, j \in V, i \not\sim_{G_0} j \iff X_i \perp\!\!\!\perp_P X_j$$

Definition : P satisfies the bi-*Global* Markov propriety (biGMP) w.r.t G_0 : $\forall i, j \in V, i \not\sim_{G_0} j$ and $S \subset V \setminus \{i, j\}$

If $V \setminus (S \cup \{i, j\})$ separates i from j in $G_0 \implies X_i \perp\!\!\!\perp_P X_j \mid \mathbf{X}_S$

Lemma : The multivariate normal distribution satisfies the biGMP

Low order partial correlation coefficient

- Let $0 \leq m \leq p - 2$. An m -partial correlation coefficient between two variables X_i and X_j is defined by

$$\rho_{ij \cdot m} = \begin{cases} \min \{ |\rho_{ij \cdot S}|, S \subseteq V \setminus \{i, j\}, |S| = m \} & \text{if } m \geq 1 \\ |\rho_{ij}| & \text{if } m = 0 \end{cases}$$

- An m -graph is an undirected graph $G_m = (V, E_m)$ associated to the distribution of X

$$i \sim_{G_m} j \iff \rho_{ij \cdot m} \neq 0$$

→ G_0 is the covariance (or independence) graph.

→ the 0-1 graph (*Wille and Bühlman (2006)*) is the graph with the set of vertices equal to $E_0 \cap E_1$

IV Main result

Let P a multivariate normal distribution faithful to a connected non-complete graph G .

Lemma 1 : If $\text{os}(G) = m$ then $G = G_m$

Lemma 2 : For all q, q' such that $1 \leq q \leq q' \leq p - 2$, we have $G_{q'} \subseteq G_q$

Lemma 3 : Suppose that $\text{os}(G_0) = m_0 < p - 2$ and that G_0 is a connected graph then $G_1 \subseteq G_0$

Theorem : With all these hypothesis, we have $\forall k, m \leq k \leq p - 2$,
$$G = G_k = G_m \subseteq G_{m-1} \subseteq \dots \subseteq G_1 \subseteq G_0$$

Proof of lemma 1 :

Let i and j be two vertices in G .

If $i \not\sim_G j \Rightarrow \exists S \subseteq V \setminus \{i, j\}, |S| = m, S$ separates i from j in G

$$\Rightarrow X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_S \Rightarrow \rho_{ij \cdot S} = 0. \Rightarrow \rho_{ij \cdot m} = 0$$

$$\Rightarrow i \not\sim_{G_m} j \Rightarrow G_m \subseteq G.$$

If $i \not\sim_{G_m} j \Rightarrow \rho_{ij \cdot m} = 0 \Rightarrow \exists S \subseteq V \setminus \{i, j\}, |S| = m$ such that

$$\Rightarrow X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_S, P \text{ is faithful to } G$$

$$\Rightarrow S \text{ separates } i \text{ from } j \text{ in } G$$

$$\Rightarrow i \not\sim_G j \implies G = G_m.$$

Proof of lemma 2 :

Let i and j be two vertices such that $i \not\sim_{G_q} j$ then there exists $S \subseteq V \setminus \{i, j\}$, $|S| = q$ such that

$$\rho_{ij \cdot S} = 0 \Rightarrow X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_S$$

P is faithful to G , then S separates i from j in G . Let $S' \subseteq V \setminus \{i, j\}$, $|S'| = q'$ and $S' \supseteq S$ then S' separates i from j in G .

$$\Rightarrow X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_{S'} \Rightarrow \rho_{ij \cdot S'} = 0 \Rightarrow \rho_{ij \cdot q'} = 0 \Rightarrow i \not\sim_{G_{q'}} j$$

Proof of lemma 3 :

Let $i, j \in V$ such that $i \not\sim_{G_0} j$

As $os(G_0) = m_0 \Rightarrow \exists S$ a separator of i and j such that $|S| = m_0$.

Let $k \in V \setminus (S \cup \{i, j\})$.

As $V \setminus \{i, j, k\} \supseteq S$, then $V \setminus \{i, j, k\}$ is a separator of i and j in G_0 . Using (biGMP), we conclude that

$$X_i \perp\!\!\!\perp X_j \mid X_k, \Rightarrow \rho_{ij \cdot 1} = 0 \Rightarrow i \not\sim_{G_1} j$$

Proposition $G = G_1 \cup G_2$, where $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are two connected components of G . Suppose that P is faithful to G . Let P_k be the marginalization of P on V_k . Then P_k is faithful to G_k for $k = 1, 2$.

Proof :

S is a separator in G_k of two vertices $i \not\sim_{G_k} j$

$\implies S$ is a separator in G of i and j

$\implies X_i \perp\!\!\!\perp_P X_j \mid \mathbf{X}_S$, or $S \cup \{i, j\} \subset V_k$

$\implies X_i \perp\!\!\!\perp_{P_k} X_j \mid \mathbf{X}_S$

Edge exclusion test

Let $i, j \in V$, $S \subset V \setminus \{i, j\}$, $|S| = m$ and $\hat{\rho}_{ij \cdot S}$ the sample correlation coefficient of (X_i, X_j) given \mathbf{X}_S .

$$\mathbf{H}_0^{ij \cdot S} : \rho_{ij \cdot S} = 0 \quad \text{vs} \quad \mathbf{H}_1^{ij \cdot S} : \rho_{ij \cdot S} \neq 0$$

Muirhead (1982), under $\mathbf{H}_0^{ij \cdot S}$,

$$\hat{\rho}_{ij \cdot S}^2 \sim \text{Beta} \left(\frac{1}{2}, \frac{n - m - 2}{2} \right)$$

Let $t(\alpha, n, m)$ be the $(1 - \alpha)$ -quantile of a Beta $\left(\frac{1}{2}, \frac{n - m - 2}{2} \right)$.

$$\alpha = \Pr(\text{reject } \mathbf{H}_0^{ij \cdot S} \mid \mathbf{H}_0^{ij \cdot S} \text{ true}) = \Pr(\hat{\rho}_{ij \cdot S}^2 \geq t(\alpha, n, m) \mid \mathbf{H}_0^{ij \cdot S} \text{ true})$$

V uPC algorithm

Some related references

- *Meek (1996)*, Chapter 6, PhD-thesis : based on the maximum degree of a graph
- *Spirtes, Glymour and Scheines (2000)* : gives the names of PC-algorithm
- *Kalish and Bühlman (2005)* : estimating skeleton of DAGs, based on the search of neighbors.

A package on **R** : `pcalg`

uPC-algorithm

Step 0 Let $m = 0$

Step 1 Estimating G_m

1. If $m = 0$, for all $i \neq j$, reject the edge (i, j) in \hat{G}_0 if $\hat{\rho}_{ij}^2 \leq t(\alpha, n, 0)$

2. If $m > 0$

2.a. use conComp to find the connected components of

\hat{G}_{m-1} :

$$\hat{G}_{m-1}^{(1)} = (V_{m-1}^{(1)}, E_{m-1}^{(1)}), \dots, \hat{G}_{m-1}^{(s)} = (V_{m-1}^{(s)}, E_{m-1}^{(s)})$$

2.b. Let $l = 1, \dots, s$

2.a. α If $|V_{m-1}^{(s)}| < m - 2$, then $\hat{G}_{m-1}^{(s)} = \hat{G}_m^{(s)}$

2.a. β Else, let $(i, j) \in E_{m-1}^{(l)}$, reject (i, j) in \hat{G}_m when $\exists S, |S| = m$ and $S \subseteq V_{m-1}^{(l)} \setminus \{i, j\}$ such that $\hat{\rho}_{ij.S}^2 \leq t(\alpha, n, m)$

Step 3 Stop if $\hat{G}_{m-1} = \hat{G}_m$ else $m = m + 1$ and return to Step 1,2.

VI Simulation Study

- We write a uPC package on **R** (still in progress)
- We use GeneTS package (on **R**, *Opgen-Rhein, Schaefer and Strimmer (2005)*) for sampling data.
- Number of vertices, $p = 10, 20$. Number of observations from $n = 20, 40, \dots, 100$. Number of replications $s = 50$. Percentage of present edges $\eta = 30\%$. Tuning parameter $\alpha = 5\%$.

False positive rate (fpr), true discovery rate (tdr) and true positive rate (tpr)

	Rejected Edges	Accepted Edges	
Edges Absents	$h_0 - V = h - R - U$	V	h_0
Edges Presents	U	$h_1 - U = R - V$	h_1
	$h - R$	R	h

$$fpr = E \left(\frac{V}{h_0} \right), \quad tdr = E \left(\frac{h_1 - U}{R}, R > 0 \right) \quad \text{and} \quad tpr = E \left(\frac{h_1 - U}{h_1} \right)$$

Results

		fpr		tdr	
$p = 10$	n	uPC	0-1	uPC	0-1
	20	0.0052	0.0155	0.9673	0.9328
	40	0.0013	0.0413	0.9927	0.8927
	60	0.0026	0.0561	0.9893	0.8769
	80	0.0013	0.0632	0.9956	0.8685
	100	0.0026	0.0800	0.9921	0.8432

		fpr		tdr	
$p = 20$	n	uPC	0-1	uPC	0-1
	20	0.0176	0.0180	0.6220	0.6201
	40	0.0101	0.0141	0.8484	0.8253
	60	0.0074	0.0146	0.9157	0.8772
	80	0.0039	0.0167	0.9605	0.8941
	100	0.0042	0.0188	0.9640	0.8980

- Changing Tuning parameter α , $p = 20$, $n = 1000$, $\eta = 15\%$ and $s = 50$.

α	fpr	tdr	tpr
0.001	0.0000	1.0000	0.8525
0.01	0.0002	0.9984	0.8600
0.05	0.0021	0.9871	0.8621
0.10	0.0051	0.9695	0.8669
0.30	0.0234	0.8745	0.8786

- Changing Tuning parameter α , $p = 10$, $n = 1000$, $\eta = 40\%$ and $s = 50$.

α	fpr	tdr	tpr
0.001	0.0007	0.9986	0.7289
0.010	0.0015	0.9973	0.7756
0.050	0.0044	0.9922	0.8356
0.100	0.0156	0.9749	0.8589
0.300	0.0504	0.9240	0.8867

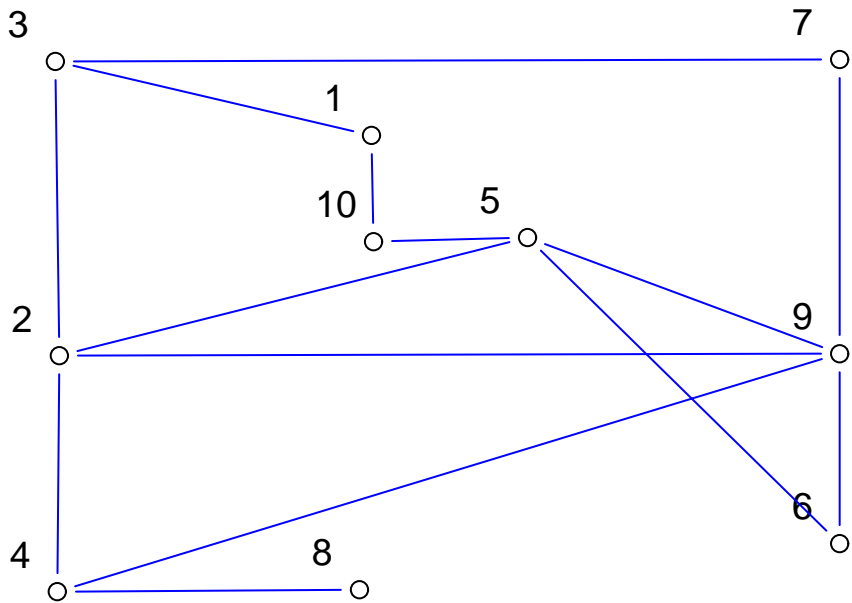
Estimating $os(G)$

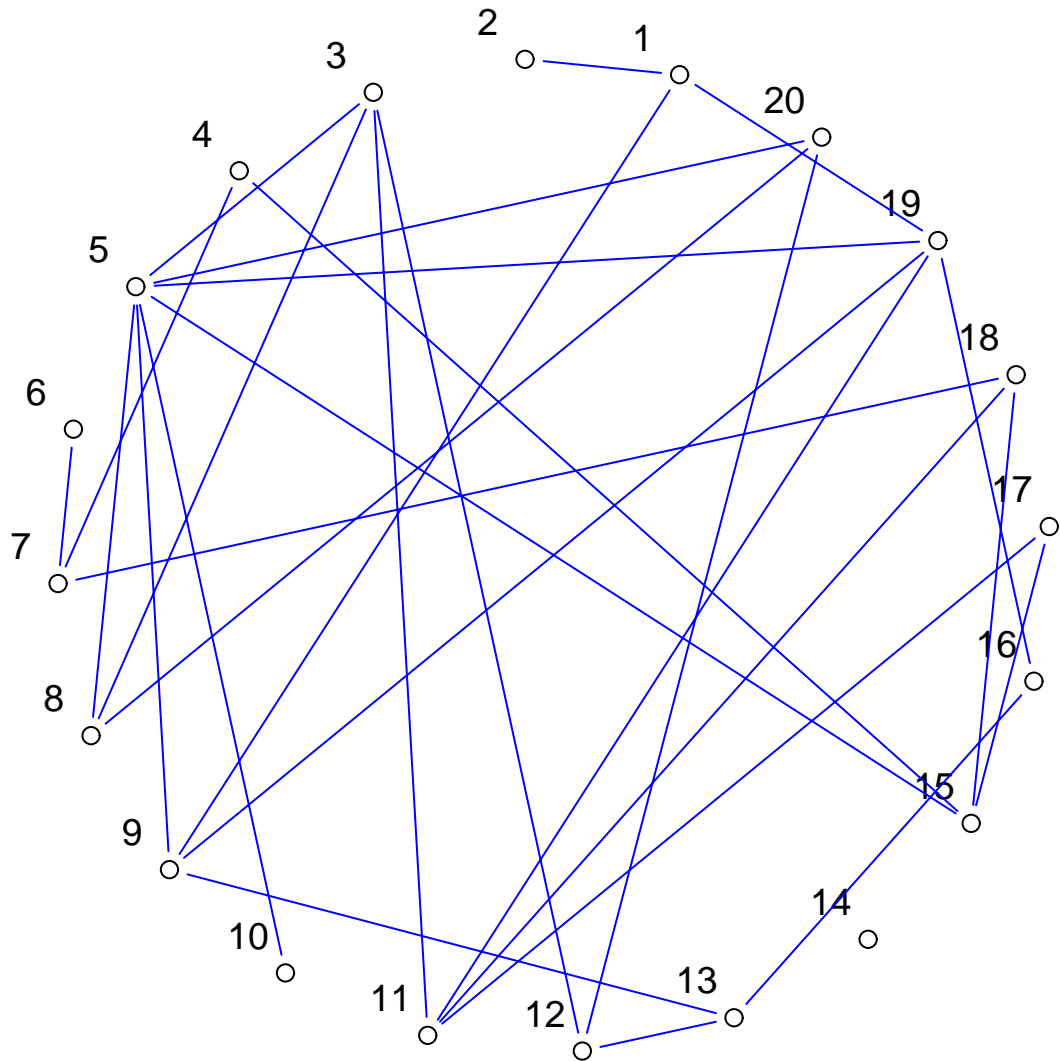
- $p = 10$, $\eta = 30\%$ and the $os(G)=3$ (G the generated graph)

	\widehat{m}	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$
frequency	2	0.56	0.48	0.22	0.36	0.42
	3	0.44	0.38	0.56	0.46	0.48

- $p = 20$, $\eta = 15\%$ and the $os(G)=3$ (one isolated vertice)

	\widehat{m}	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$
frequency	2	0.64	0.08	0.04	0.02	
	3	0.32	0.66	0.42	0.46	0.46
	4	0.04	0.26	0.48	0.36	0.38





References

- Wu, X. and Ye, Y. and Subramanian, K.R. Interactive Analysis of Gene Interactions Using Graphical Gaussian Model, *3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*, 2003
- Whittaker, J., *Graphical models in Applied Multivariate Statistics.*, Wiley, 1990
- Lauritzen, S. L., *Graphical Models*, New York : Oxford University Press., 1996
- Edwards, D., *Introduction to graphical modelling*, Springer texts in statistics, 2000

- Cox, D. R. and Wermuth, N. , *Multivariate Dependencies : Models, Analysis and Interpretations*, Chapman and Hall, 1996
- Richardson, T. S. and Spirtes, P., Ancestral graph Markov models, *Ann. Statist.* , **30**, 962-1030, 2002
- Wille, A. and Bühlmann, P., Low-Order Conditional Independence Graphs for Inferring Genetic Networks, journal=*Statistical Applications in Genetics and Molecular Biology*, Volume 5, Issue 1, 1-32, 2006.
- Muirhead, R.J. *Aspects of multivariate statistical theory*. Wiley:New-York, 1982.
- Meek, C. Relating Graphical Frameworks, Ph-D, Chapter 6, 1996.

- P. Spirtes, C. Glymour and R. Scheines, *Causation, Prediction and Search*, The MIT Press, 2nd edition, 2000.
- M. Kalish and P. Bühlman, Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *preprint* 2005.
- Opgen-Rhein, R., Schaefer, J. and Strimmer K. GeneTS, a package on **R**, 2005