An empirical Bayes procedure for the selection of Gaussian graphical models

Estimation bayésienne pour les modèles graphiques gaussiens décomposables

Jean-Michel Marin
I3M, Université Montpellier 2

joint with Sophie Donnet, Université Paris Dauphine

Introduction

The last decade has witnessed the apparition of applied problems typified by very high-dimensional variables, in marketing database or gene expression studies for instance.

Graphical modelling is a form of multivariate analysis that uses graphs to represent models.

They enable concise representations of associational and causal relations between variables under study.

There is two main types of graphical models:

- undirected graphical models;
- directed acyclic graphical models.

Lauritzen (1996)

We shall concentrate on undirected graphs.

Example of an undirected graph

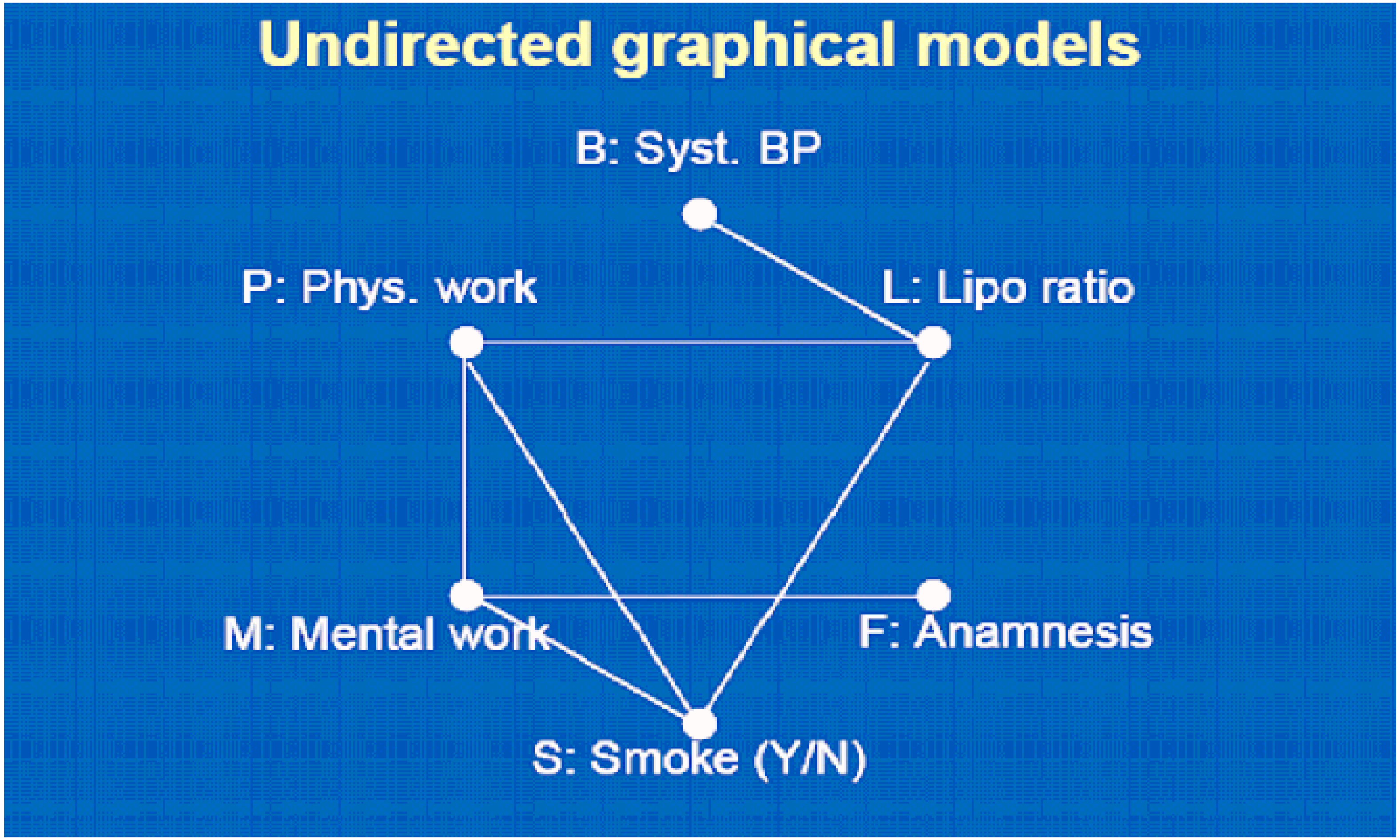
1841 employees of a car factory

6 binary variables

S: smoking (yes or not)
M: strenuous mental work (yes or not)
P: strenuous physical work (yes or not)
B: blood pressure (<140 or ≥ 140)
L: ratio of lipoproteins (<3 or ≥ 3)
F: family history of coronary heart disease (yes or not)

Madigan and Raftery (1994)

Undirected graphical models



If the graph is known, the parameters of the model are easily estimated.

However, a quite challenging issue is the determination of the set of most appropriate graphs for a given dataset.

We consider this problem and the case of decomposable Gaussian graphical models

Dawid and Lauritzen (1993)

Plan

- Background on Bayesian model selection
- Background on decomposable Gaussian graphical models
- Bayesian tools for Gaussian graphical models
- An empirical Bayes procedure via the SAEM-MCMC algorithm
- A new Metropolis-Hastings sampler to explore the space of graphs
- Numerical experiments

Background on Bayesian model selection

Several models available for the same observation

$$\mathfrak{M}_i : \mathbf{y} \sim f_i(\mathbf{y}|\theta_i), \qquad i \in \mathfrak{I}$$

where $\mathfrak{I}$ can be finite or infinite

Probabilise the entire model/parameter space

- allocate probabilities p_i to all models $\mathfrak{M}_i$
- define priors $\pi_i(\theta_i)$ for each parameter space Θ_i
- compute

$$\mathbb{P}(\mathfrak{M}_i|\mathbf{y}) = \frac{p_i \displaystyle\int_{\Theta_i} f_i(\mathbf{y}|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\displaystyle\sum_j p_j \int_{\Theta_j} f_j(\mathbf{y}|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

- take largest $\mathbb{P}(\mathfrak{M}_i|\mathbf{y})$ to determine "best" model, or use averaged predictive

$$\sum_j \mathbb{P}(\mathfrak{M}_j|\mathbf{y}) \int_{\Theta_j} f_j(\mathbf{y}'|\theta_j,\mathbf{y})\pi_j(\theta_j|\mathbf{y})\mathrm{d}\theta_j$$

Background on decomposable Gaussian graphical models

Let $\mathcal{G} = (V, E)$ be an undirected graph:

- $V = \{1, \ldots, p\}$ is the vertex set;
- $E \subseteq \{(i, j) : 1 \leq i < j \leq p\}$ is the edge set: if $(a, b) \in E$ then vertices a and b are adjacent in $\mathcal{G}$.

A graph or subgraph is complete if all its vertices are joined by an edge.

A complete subgraph that is not contained within another complete subgraph is called a clique.

Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be the set of cliques of $\mathcal{G}$.

An ordering of all the cliques $(C_1, \ldots, C_k)$ is said to be perfect if the vertices of each clique C_i also contained in any previous clique $C_1, \ldots, C_{i-1}$ are all members of one previous clique; that is $\forall i = 2, 3, \ldots, k$,

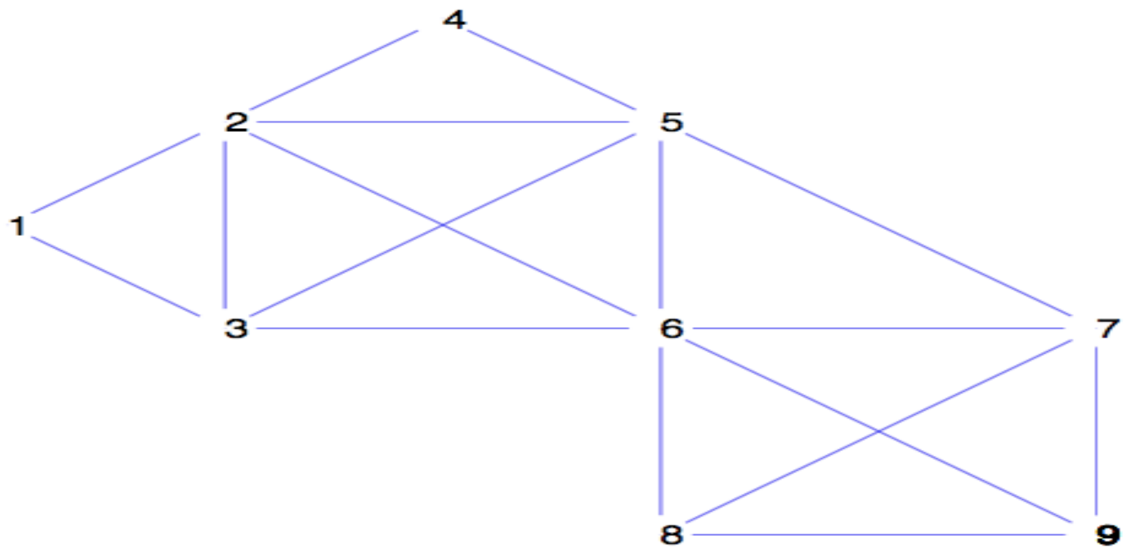
$$S_i = C_i \cap \cup_{j=1}^{i-1} C_i \subseteq C_h$$

for some $h = h(i) \in \{1, 2, \ldots, i-1\}$.

$\mathcal{S} = \{S_2, \ldots, S_k\}$ is the set of separators associated to the perfect ordering $\{C_1, \ldots, C_k\}$.

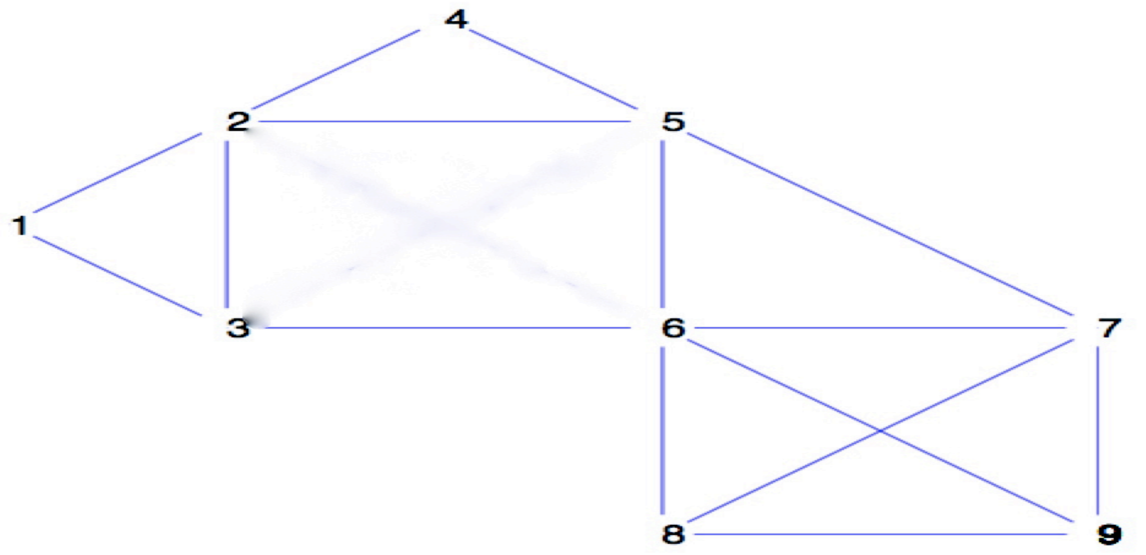
If an undirected graph admits a perfect ordering it is said to be decomposable.

The following graph (used as benchmark in the following) is decomposable.



$k = 5$, $C_1 = \{1, 2, 3\}$, $C_2 = \{2, 3, 5, 6\}$, $C_3 = \{2, 4, 5\}$, $C_4 = \{5, 6, 7\}$ and $C_5 = \{6, 7, 8, 9\}$, $S_2 = \{2, 3\}$, $S_3 = \{2, 5\}$, $S_4 = \{5, 6\}$ and $S_5 = \{6, 7\}$.

If $(2,6) \notin E$ and $(3,5) \notin E$, the graph is not decomposable any more.



$k = 5$, $C_1 = \{1,2,3\}$, $C_2 = \{2,4,5\}$, $C_3 = \{3,6\}$, $C_4 = \{5,6,7\}$ and $C_5 = \{6,7,8,9\}$

With p vertices, the number of possible edges is $T = \frac{p(p-1)}{2}$ and the total number of graphs is 2^T.

The total number of decomposable graphs with p vertices can be calculated for moderate values of p, for instance:

if $p = 6$ there is $32,768$ graphs and $18,154$ are decomposable (around 55%);

if $p = 8$, there is $268,435,456$ graphs and $30,888,596$ are decomposable (around 12%).

A pair (A, B) of subsets of the vertex set V of an undirected graph $\mathcal{G}$ is said to form a decomposition of $\mathcal{G}$ if

- $V = A \cup B$;
- $A \cap B$ is complete;
- $A \cap B$ separates A from B (any path from a vertex in A to a vertex in B goes through $A \cap B$).

To each vertex $v \in V$, we associate a random variable y_v.

For $A \subseteq V$, $\mathbf{y}_A = (y_v)_{v \in A}$ indicates the collection of random variables $\{y_v : v \in A\}$. To ease the notation, let $\mathbf{y} = \mathbf{y}_V$.

The probability distribution of $\mathbf{y}$ is said to be Markov with respect to $\mathcal{G}$, if for any decomposition (A, B) of $\mathcal{G}$, $\mathbf{y}_A$ is independent of $\mathbf{y}_B$ given $\mathbf{y}_{A \cap B}$ (global Markov property).

A graphical model is a family of distributions on $\mathbf{y}$ which are Markov with respect to a graph.

A Gaussian graphical model is such that

$$\mathbf{y}|\mathcal{G},\Sigma_{\mathcal{G}} \sim \mathcal{N}_p\left(0_p,\Sigma_{\mathcal{G}}\right)\ , \tag{1}$$

where Σ_G is a positive definite matrix which ensures that the distribution of $\mathbf{y}$ is Markov with respect to $\mathcal{G}$.

$\Sigma_{\mathcal{G}}$ ensures that the distribution of $\mathbf{y}$ is Markov if and only if

$$(i,j)\notin E \Longleftrightarrow \left(\Sigma_{\mathcal{G}}^{-1}\right)_{(i,j)} = 0\,.$$

Dempster (1972) (covariance selection models)

In a Gaussian graphical model, the global, local and pairwise Markov properties are equivalent.

Local Markov property: every variable is conditionally independent of the remaining, given its neighbours.

Pairwise Markov property: any non-adjacent pair of random variables are conditionally independent given the remaning.

The mean parameter is typically set to zero: the data we analyze will be expressed as deviation from the sample mean.

We observe a sample $\mathbf{y}^1, \ldots, \mathbf{y}^n$ from (1) (the data are centered).

We would like to identify the set of most relevant graphs.

For the considered multivariate random phenomenon, we are interested in the set of most relevant conditional independence structures.

$\Longrightarrow$ explore huge graph space.

Bayesian tools for Gaussian graphical models

We consider the Bayesian paradigm.

Conditionally on $\mathcal{G}$, we use a Hyper-Inverse Wishart (HIW) distribution associated to the graph $\mathcal{G}$ as prior distribution on $\Sigma_{\mathcal{G}}$:

$$\Sigma_{\mathcal{G}} | \mathcal{G}, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}} \sim \text{HIW}_{\mathcal{G}} \left(\delta_{\mathcal{G}}, \Phi_{\mathcal{G}} \right)$$

where $\delta_{\mathcal{G}} > 0$ and $\Phi_{\mathcal{G}}$ is a $p \times p$ symmetric positive definite matrix.

Dawid and Lauritzen (1993), Giudici and Green (1999), Armstrong et al. (2006)

Conditionally on $\mathcal{G}$, the HIW distribution is conjugate

$$\Sigma_{\mathcal{G}} | \mathbf{y}^1, \ldots, \mathbf{y}^n, \mathcal{G}, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}} \sim \mathrm{HIW}_{\mathcal{G}} \left(\delta_{\mathcal{G}} + n, \Phi_{\mathcal{G}} + \sum_{i=1}^{n} \mathbf{y}^{\mathbf{i}} \left(\mathbf{y}^{\mathbf{i}} \right)^{\mathrm{T}} \right) . \qquad (2)$$

Moreover, for such a prior,

$$f(\mathbf{y}^1, \ldots, \mathbf{y}^n | \mathcal{G}, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}}) = \frac{h_{\mathcal{G}}(\delta_{\mathcal{G}}, \Phi_{\mathcal{G}})}{(2\pi)^{-np/2} h_{\mathcal{G}} \left(\delta_{\mathcal{G}} + n, \Phi_{\mathcal{G}} + \sum_{i=1}^{n} \mathbf{y}^{i} \left(\mathbf{y}^{i} \right)^{\mathrm{T}} \right)}$$

where $h_{\mathcal{G}}$ is the normalizing constant of the HIW distribution associated to the graph $\mathcal{G}$.

Roverato (2002) extends Hyper-Inverse Wishart distribution to non-decomposable case.

Let $\mathbf{Y} = (\mathbf{y}^1, \ldots, \mathbf{y}^n)$ and $S_{\mathbf{Y}} = \sum_{i=1}^n \mathbf{y}^i \left(\mathbf{y}^i\right)^{\mathrm{T}}$.

If we assume a uniform prior distribution in the space of graphs, $\pi(\mathcal{G}) \propto 1$:

$$\pi\left(\mathcal{G}|\mathbf{Y}, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}}\right) \propto f\left(\mathbf{Y}|\mathcal{G}, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}}\right).$$

Uniform distribution on the space of graphs typically not satisfactory: with p vertices, the number of possible edges is equal to $\frac{p(p-1)}{2}$ and, for an uniform prior over all graphs, the prior number of edges has mode around $\frac{p(p-1)}{4}$.

Wong, Carter and Kohn (2003), Jones et al. (2005), Armstrong et al. (2009), Carvalho and Scott (2009)

An alternative to the naive uniform prior is to set a Bernouilli distribution of parameter r on the inclusion or not of each edge:

$$\pi(\mathcal{G}|r) \propto r^{k_{\mathcal{G}}}(1-r)^{\frac{p(p-1)}{2}-k_{\mathcal{G}}} ,$$

where $k_{\mathcal{G}}$ is the number of edges of $\mathcal{G}$. The parameter r has to be calibrate. If $r = 1/2$, this prior resumes to the uniform one.

We deduce easily that

$$\pi\left(\mathcal{G}|\mathbf{Y}, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}}, r\right) \propto \frac{h_{\mathcal{G}}(\delta_{\mathcal{G}}, \Phi_{\mathcal{G}})}{h_{\mathcal{G}}(\delta_{\mathcal{G}}+n, \Phi_{\mathcal{G}}+S_{\mathbf{Y}})}\pi(\mathcal{G}|r)\,. \tag{3}$$

An empirical Bayes procedure via the SAEM-MCMC algorithm

(3) is sensible to the specification of the hyper-parameters $\delta_{\mathcal{G}}$, $\Phi_{\mathcal{G}}$ and r!

Typically, $\delta_{\mathcal{G}} = \delta$ and $\Phi_{\mathcal{G}} = \Phi$.

Different strategies:

Giudici and Green (1999) and others propose to fix $r = 1/2$ and use a hierarchical prior modeling: δ and Φ are considered as random quantities and a prior distribution is assigned on δ and Φ.

The difficulty with this approach is that the prior distributions on δ and Φ also depend on hyper-parameters...

Jones et al. (2005) and others fix the values of δ, Φ and r using some heuristics more or less justified and never completely satisfactory.

r is set to $\frac{1}{p-1}$ encouraging sparse graphs and $\delta = 3$ which is the minimal integer such that the first moment of the prior distribution on $\Sigma_{\mathcal{G}}$ exists.

They set $\Phi = \tau I_p$ and using the fact that the mode of the marginal prior for each variance terms σ_{ii} is equal to $\tau/(\delta+2)$, τ is fixed to $\delta + 2$ if the data set is standardized.

Armstrong et al. (2009) fix $\delta = 4$ assessing that such a value gives a suitably non-informative prior for $\Sigma_{\mathcal{G}}$ and use a hierarchical prior modeling on Φ and r.

They consider different possibilities for Φ, all of the form $\Phi = \tau A$ where the matrix A is fixed. In all cases, for the hyper-parameter τ, they use a uniform prior distribution on the interval $[0, \Gamma]$ where Γ is very large.

Finally, they also use a hierarchical prior on r : $r \sim \beta(1,1)$, which leads to

$$\pi(\mathcal{G}) \propto \binom{\frac{p(p-1)}{2}}{k_{\mathcal{G}}}^{-1}$$

by integration.

Carvalho and Scott (2009) also use a hierarchical prior on r such that

$$\pi(\mathcal{G}) \propto \begin{pmatrix} \frac{p(p-1)}{2} \\ k_{\mathcal{G}} \end{pmatrix}^{-1} .$$

They suggest a HIW g-prior approach with $g = 1/n$. This approach consists of fixing $\delta = 1$ and $\Phi = S_{\mathbf{Y}}/n$.

δ measures the amount of information in the prior relative to the sample (see (2)), we propose to fix $\delta = 1$.

The prior weight is the same as the weight of one observation.

Moreover, we propose to standardize the data and to use $\Phi = \tau I_p$.

This choice encourages sparse graph (on average each variable has major interactions with a relative small number of other variables).

τ and r play the role of shrinkage factors: important to choose τ and r to be on the appropriate scale!

Empirical Bayes strategy: we propose to fix τ and r to their maximum likelihood estimations.

How to calculate the maximum likelihood estimates of τ and r?

We use a Markov Chain Monte Carlo (MCMC) version of the Stochastic Approximation EM (SAEM) algorithm.

Delyon, Lavielle and Moulines (1999), Kuhn and Lavielle (2004)

The maximization of $f\left(\mathbf{Y}|\tau, r\right)$ can not be done in closed form.

$$f(\mathbf{Y}|\tau, r) \propto \sum_{\mathcal{G}\in\mathcal{D}_p} \left\{ \frac{h_{\mathcal{G}}(\delta, \tau I_p)}{h_{\mathcal{G}}(n+\delta, \tau I_p + S_{\mathbf{Y}})} \right\} \pi(\mathcal{G}|r)\,.$$

The observed data $\mathbf{Y} = \left(\mathbf{y}^1, \ldots, \mathbf{y}^n\right)$ are issued from the partial observations of the complete data $(\mathbf{Y}, \mathcal{G}, \Sigma_{\mathcal{G}})$.

Let $\theta = (\tau, r)$, EM algorithm:

$$Q(\theta|\theta') = \mathbb{E}_{\Sigma_{\mathcal{G}},\mathcal{G}} \left\{ \ln f(\mathbf{Y}, \mathcal{G}, \Sigma_{\mathcal{G}}|\theta) | \mathbf{Y}, \theta' \right\}\,.$$

At iteration k, the E-step is the evaluation of $Q_k(\theta) = Q(\theta \,|\, \widehat{\theta}_{k-1})$ while the M-step updates $\widehat{\theta}_{k-1}$ by maximizing $Q_k(\theta)$.

For cases where the E-step is intractable, Delyon, Lavielle and Moulines (1999) propose the SAEM algorithm.

The E-step is replaced by a stochastic approximation of $Q_k(\theta)$.

At iteration k, the E-step is divided into a simulation step $\left(\mathcal{G}^{(k)}, \Sigma_{\mathcal{G}}^{(k)}\right)$ and a Stochastic Approximation step:

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left[\ln f(\mathbf{Y}, \mathcal{G}^{(k)}, \Sigma_{\mathcal{G}}^{(k)} | \widehat{\theta}_{k-1}) - Q_{k-1}(\theta)\right],$$

where $(\gamma_k)_{k\in\mathbb{N}}$ is a sequence of positive numbers decreasing to zero.

In gaussian graphical models, we can not generate directly a realization from the conditional distribution of $(\mathcal{G}, \Sigma_{\mathcal{G}})$ given $\mathbf{Y}$ and $\widehat{\theta}_{k-1}$.

For such cases, Kuhn and Lavielle (2004) suggest to replace the simulation step by a MCMC scheme:

generate M realizations from an ergodic Markov chain with stationary distribution $\mathcal{G}, \Sigma_{\mathcal{G}} | \mathbf{Y}, \widehat{\theta}_{k-1}$ and use the last simulation in the SAEM algorithm.

It is very easy to generate a realization from $\Sigma_{\mathcal{G}}|\mathcal{G}, \mathbf{Y}, \widehat{\theta}_{k-1}$.

Moreover, the pdf of $\mathcal{G}|\mathbf{Y}, \widehat{\theta}_{k-1}$ is available up to a normalizing constant.

In the MCMC step of the SAEM-MCMC algorithm, we will generate M realizations from an ergodic Markov chain with stationary distribution $\mathcal{G}|\mathbf{Y}, \widehat{\theta}_{k-1}$ and use the last simulation to generate $\Sigma_{\mathcal{G}}$.

Once the sequence of $\widehat{\theta}_k$ converges, we use only the MCMC algorithm to explore the space of graphs.

A new Metropolis-Hastings sampler to explore the space of graphs

We propose a new Metropolis-Hastings algorithm.

Let K denote the empirical correlation matrix.

At iteration t of the algorithm,

1) Choose at random to delete or add an edge to $\mathcal{G}^{(t-1)}$;

a) If delete, enumerate $G^{-}_{\mathcal{G}^{(t-1)}}$ and generate $\mathcal{G}^p$ using the following distribution

$$\mathbb{P}(\mathcal{G}^p = \{\mathcal{G}^{(t-1)} \setminus (i,j)\} | \mathcal{G}^{(t-1)}) \propto \frac{1}{|K(i,j)|} \, ;$$

b) If add, enumerate $G^{+}_{\mathcal{G}^{(t-1)}}$ and generate $\mathcal{G}^p$ using the following distribution

$$\mathbb{P}(\mathcal{G}^p = \{\mathcal{G}^{(t-1)} \cup (i,j)\} | \mathcal{G}^{(t-1)}) \propto |K(i,j)| \, ;$$

2) Calculate the acceptance probability $\rho(\mathcal{G}^{(t-1)}, \mathcal{G}^p)$;

3) With probability $\rho(\mathcal{G}^{(t-1)}, \mathcal{G}^p)$, accept $\mathcal{G}^p$ and set $\mathcal{G}^{(t)} = \mathcal{G}^p$, otherwise reject $\mathcal{G}^p$ and set $\mathcal{G}^{(t)} = \mathcal{G}^{(t-1)}$;

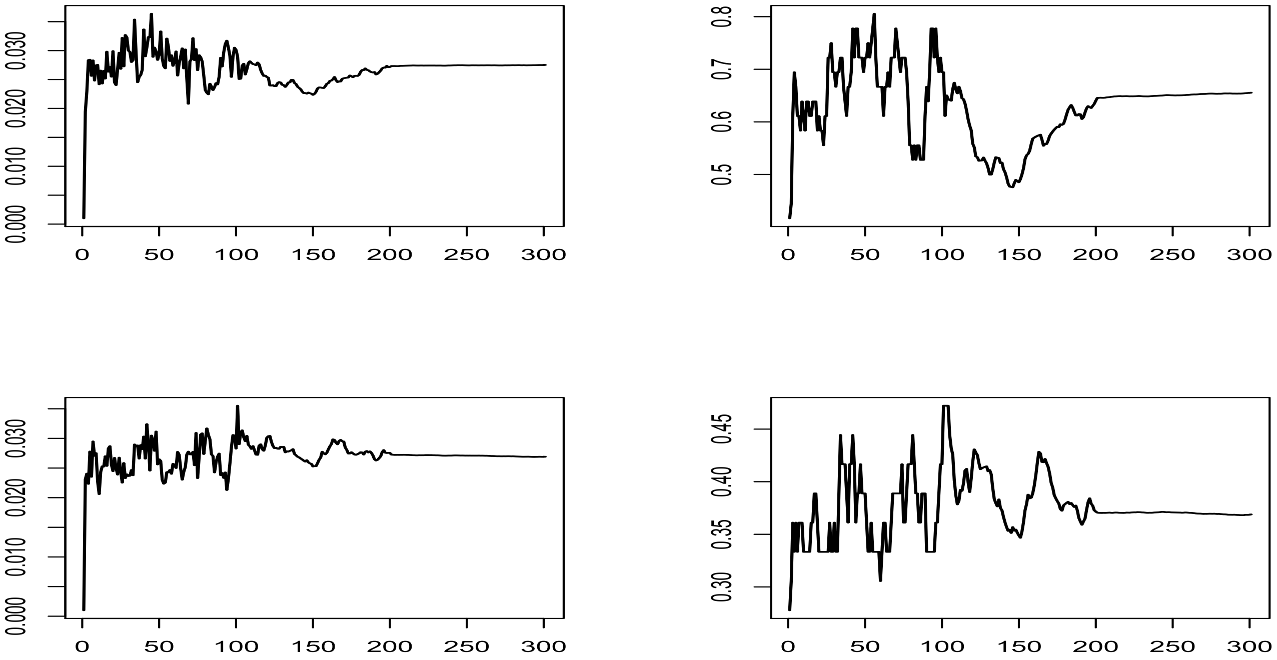
Numerical experiments

Simulated datasets

$p = 9$, $n = 100$, $\delta = 1$ and $\tau = 0.03$.

$\gamma_k = 1$ during the first iterations $1 \leq k \leq 100$ and $\gamma_k = (k - 100)^{-1}$ during the subsequent iterations.

$M = 500$ during the 5 first iterations and then $M = 10$.



Simulated datasets: evolution of the SAEM-MCMC $\hat{\tau}^{(k)}$ estimations (left) and $\hat{r}^{(k)}$ estimations (right) on 2 datasets

Real datasets

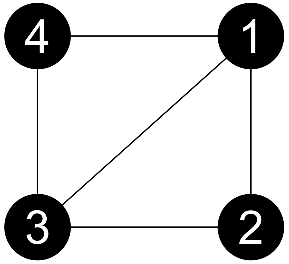
Fret's heads dataset contains head measurements on the first and the second adult son in a sample of 25 families.

The 4 variables are the head length of the first son, the head breadth of the first son, the head length of the second son and the head breadth of the second son.

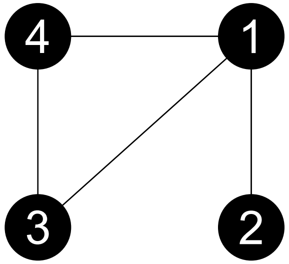
In this case $p = 4$ and 61 graphes are decomposable among the 64 possibles graphes.

On this dataset we aim at proving that the hyper-parameters τ and r has to be carefully chosen.

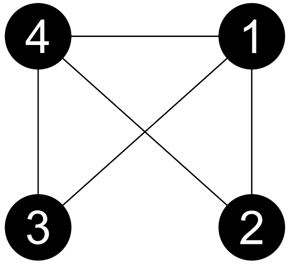
Jones et al. (2005)



0.24076

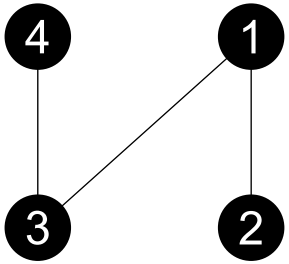


0.16924

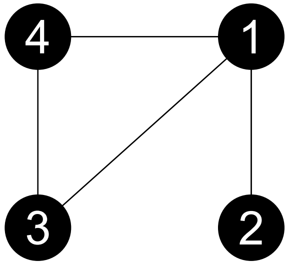


0.11761

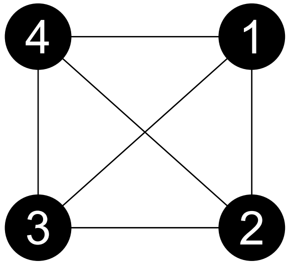
Carvalho and Scott (2009)



0.30512

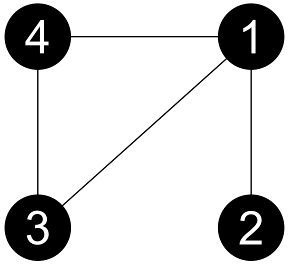


0.19979

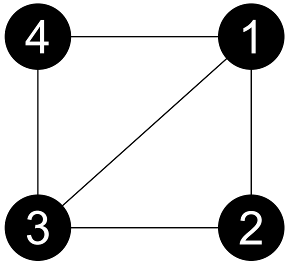


0.10813

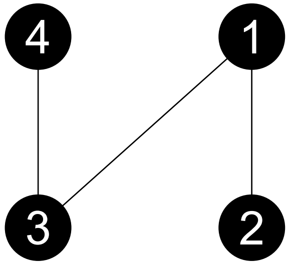
SAEM



0.28613



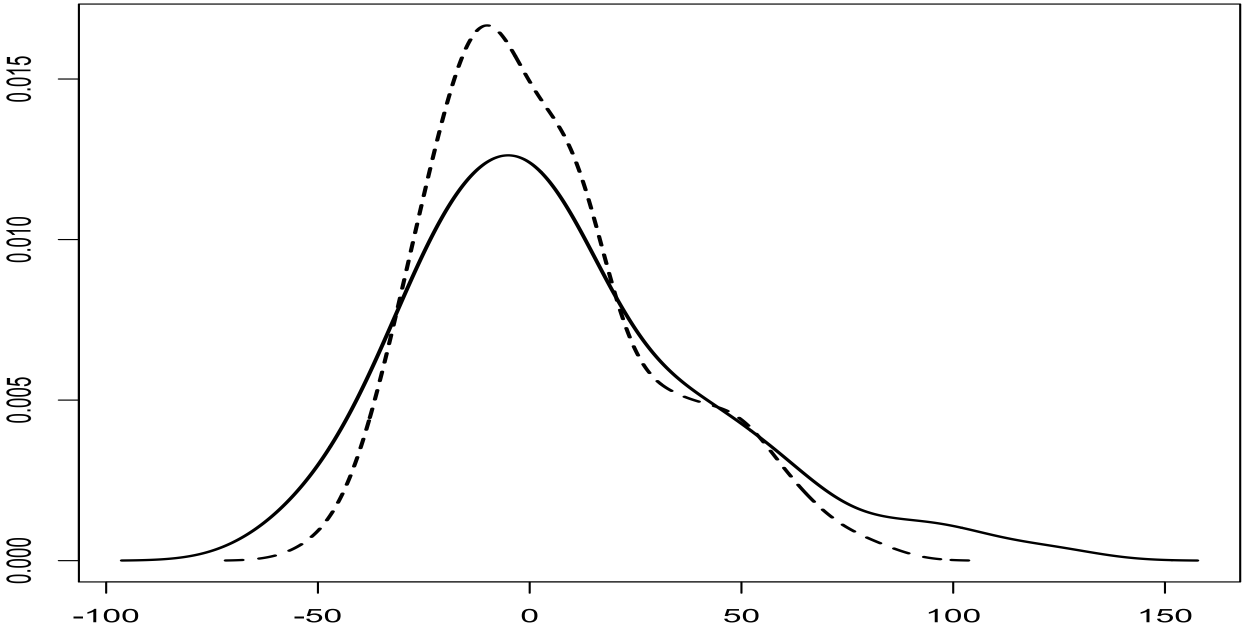
0.18219



0.1264

In the **Fowl bones dataset**, bone measurements are taken on $n = 276$ white leghorn fowl. The 6 variables are skull length, skull breadth, humerous (wings), ulna (wings), femur (legs) and tibia (legs).

We aim at illustrating the fact that a careful choise of the transition kernel in the MCMC algorithm ensures a better exploration of the support of the posterior distribution.



Fowl bones data set: densities of the relative errors on the posterior probabilities for the 107 most probable graphs. *add and delete* kernel in solid line and data-driven kernel in dashed line.