Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Multivariate Analysis 97 (2006) 1451–1466
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

# Asymptotic Bayesian structure learning using graph supports for Gaussian graphical models

Guillaume Marrelec[a, b, *], Habib Benali[a, b]

[a] *UMR-S 678 Inserm/UPMC, Paris, France*
[b] *MIC/UNF, Université de Montréal, Montréal, Canada*

Received 25 August 2004
Available online 26 October 2005

**Abstract**

The theory of Gaussian graphical models is a powerful tool for independence analysis between continuous variables. In this framework, various methods have been conceived to infer independence relations from data samples. However, most of them result in stepwise, deterministic, descent algorithms that are inadequate for solving this issue. More recent developments have focused on stochastic procedures, yet they all base their research on strong a priori knowledge and are unable to perform model selection among the set of all possible models. Moreover, convergence of the corresponding algorithms is slow, precluding applications on a large scale. In this paper, we propose a novel Bayesian strategy to deal with structure learning. Relating graphs to their supports, we convert the problem of model selection into that of parameter estimation. Use of non-informative priors and asymptotic results yield a posterior probability for independence graph supports in closed form. Gibbs sampling is then applied to approximate the full joint posterior density. We finally give three examples of structure learning, one from synthetic data, and the two others from real data.


*AMS 1991 subject classification:* 62F15; 62H20; 62H05; 05C50; 62F40

*Keywords:* Bayesian analysis; Partial correlation coefficients; Gaussian graphical models; Conditional independence graphs; Gibbs sampler

## 1. Introduction

As pointed out by Dawid [2], conditional independence is believed to be fundamental knowledge in the process of statistical inference. In the theory of Gaussian graphical models [14,10], condi-

* Corresponding author. Département de Psychologie, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Canada, QC, H3C 3J7.
*E-mail address:* guillaume.marrelec@umontreal.ca (G. Marrelec).

0047-259X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2005.08.008

tional independences between continuous variables are embedded in a formal framework whose enormous advantage is its intuitive meaning. Conditional independence graphs are mathematical objects that very efficiently encode conditional independence relations, providing a precise yet vivid representation of the independence structure underlying the data. Variables are represented by vertices, direct dependences (resp. independences) between two variables by edges (resp. absence of edges) and independence patterns by graphs. In this framework, estimating independence relations from data samples reduces to inference about the independence graph underlying the data.

A crucial topic in independence analysis from data samples is computation. Working with realizations of $D$ variables brings up $D(D-1)/2$ edges and $2^{D(D-1)/2}$ potential independence graphs. As a consequence, blunt exploration of all the lattice states quickly becomes untractable, and trade-offs have to be found between exhaustive search of the parameter space and time cost. Therefore, the choice of the inference algorithm has to be given a central role. A second key point is how the inference about graphs is dealt with. Data give access to empirical interaction coefficients, that are continuous, whereas inference must be carried out on graphs, that are discrete models.

For Gaussian graphical models, various techniques have been developed to learn structure from data. Most of them are variants of that proposed by Whittaker [14] and all share the disadvantages of being stepwise, deterministic, descent algorithms, and hence are poorly adapted to solve this issue. Only recently have new stochastic methods been introduced, mainly through Bayesian analysis. In this framework, independence analysis of decomposable graphs has been widely explored, e.g., [2] or [5]. There are still several practical drawbacks to it, though. Most methods do not apply to non-decomposable graphs, which are hence a priori excluded from the analysis. Furthermore, a conditional correlation structure must be defined a priori, together with the relative weight given to this prior. Both choices have a rather important influence on the resulting graph estimate, whereas translating prior information into value for the hyperparameters is far from being obvious in most cases. Last but not least, the issue brought up by the exponentially increasing number of potential models remains unsolved, since the methods proposed require prior selection of a subset of graphs on which investigation will focus blindly, at the exclusion of all other graphs.

To be practically efficient, a structure learning technique should take into account that the most common case is when no (or very little) information is available and/or usable a priori. As such, it should be able to perform robust analysis even in a state of prior ignorance. A Bayesian analysis should hence integrate as little bias as possible into the analysis through prior information: this is achieved by use of uninformative priors [1,4]. Second, an ideal structure learning technique should also be able to consider and lead inference on the set of all $2^{D(D-1)/2}$ possible models instead of restraining the research on a potentially incorrect subset. Incidentally, meeting these two requirements would make the corresponding process operator-independent and, consequently, enable automated analysis.

In this paper, we propose a new approach for Bayesian structure learning. We first make a clear distinction concerning the mathematical object we wish to draw inference about: is it a set of conditional correlation coefficients, or is it an independence graph? Until now, attention was mostly directed towards partial correlations, and from there inference was drawn about independence models. The notion of graph only appeared through the definition of a model constraining the set of partial correlation coefficients. We explicitely introduce graphs as a discrete variable rather than a model when related to its support. The issue of model selection then boils down to parameter estimation, and the whole Bayesian machinery can be applied, leading to direct

and effective inference on the set of *all* potential graphs concerning the independence structure underlying the data.

A key issue in resolution of structure learning is integration of the constraint entailed by the positive-definiteness of the covariance matrix. This constraint has very complex implications on the set of partial correlations, on which inference is performed to gain insight into the independence structure underlying the data. We propose a pragmatic approach, based on non-informative priors and numerical calculation, that allows to take this relevant piece of information into account.

We finally show that, using this methodology, structure learning can efficiently be performed using Gibbs sampling on the set of all models.

The outline of this paper is the following. In Section 2, we introduce the notational and theoretical background of the paper. In Section 3, methods and tools are devised to draw inference about graphs. Computational issues are discussed as well. In Section 4, the features of the method are assessed using simulations. Finally, in Section 5, these new concepts are applied to two problems from the literature, showing the advantages of our approach compared to previous methods.

## 2. General background

We first introduce some important notations and definitions that will be very useful in the subsequent sections: partial correlation coefficients and independence graphs. We also recall two fundamental lemmas of asymptotic convergence and finally set the Bayesian framework that will justify our approximations.

### 2.1. Notations

Let $x$ denote a real number, $\boldsymbol{x}$ a vector, $\boldsymbol{X}$ a matrix, and $\mathbb{X}$ a set. $\mathcal{M}^+$ stands for the set of all symmetric matrices that are positive definite. For any matrix $\boldsymbol{X}$, $\mathrm{tr}(\boldsymbol{X})$ stands for the trace of $\boldsymbol{X}$. For any finite set $\mathbb{X}$, $|\mathbb{X}|$ is the cardinal of $\mathbb{X}$.

We use the general notation $G = (\mathbb{V}, \mathbb{E})$ for an undirected graph, where $\mathbb{V}$ is the vertex set and $\mathbb{E}$ the edge set. $\mathbb{F}$ stands for the edge set of the complete graph, i.e.,

$$\mathbb{F} = \{(i, j) : i, j \in \mathbb{V}, i < j\},$$

and $\bar{\mathbb{E}} = \mathbb{F} \setminus \mathbb{E}$ for the set of edges that do not appear in $G$. Note that, for notational convenience, we do not include elements of the form $(i, i)$ in the edge sets.

We denote by $\mathcal{M}_0^+(G)$ the set of all matrices of $\mathcal{M}^+$, indexed by $\mathbb{V} \times \mathbb{V}$, with element $(i, j)$ equal to zero whenever $(i, j) \notin \mathbb{E}$ and $i \neq j$.

$\delta_{\mathbf{0}}(\boldsymbol{x})$ is the multidimensional delta function, such that $\delta_{\mathbf{0}}(\mathbf{0}) = 1$ and $\delta_{\mathbf{0}}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \neq \mathbf{0}$. For a set $\mathbb{A}$, $1_{\mathbb{A}}(\boldsymbol{x})$ is the characteristic function of $A$, mapping every $\boldsymbol{x} \in \mathbb{A}$ to 1, and every other $\boldsymbol{x}$ to 0.

$\mathrm{p}(x|y)$ is the probability of $x$ given $y$, $\mathrm{E}[x]$ the expectation of $x$, $\mathrm{Var}[x]$ its variance and $\mathrm{Cov}[x_1, x_2]$ the covariance between $x_1$ and $x_2$. For three variables $x_1$, $x_2$ and $x_3$, $x_1 \perp\!\!\!\perp x_2 | x_3$ means that $x_1$ and $x_2$ are independent given $x_3$. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}_0)$ represents the value of the Gaussian density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ calculated for vector $\boldsymbol{x}_0$.

### 2.2. Partial correlation coefficients

Let $\mathbb{V}$ be a finite set with $|\mathbb{V}| = D$ and $\boldsymbol{x}$ a $D$-dimensional vector indexed by $\mathbb{V}$. For $i, j \in \mathbb{V}$, the conditional correlation coefficient between $x_i$ and $x_j$ given all other variables is

given by

$$\mathrm{Corr}[x_i, x_j|\boldsymbol{x}_{\mathbb{V}\setminus\{i,j\}}] = \frac{\mathrm{Cov}[x_i, x_j|\boldsymbol{x}_{\mathbb{V}\setminus\{i,j\}}]}{\sqrt{\mathrm{Var}[x_i|\boldsymbol{x}_{\mathbb{V}\setminus\{i,j\}}]\mathrm{Var}[x_j|\boldsymbol{x}_{\mathbb{V}\setminus\{i,j\}}]}}.$$

In the special case where $\boldsymbol{x}$ is multivariate Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$—or, equivalently, concentration matrix $\boldsymbol{\Upsilon} = (\upsilon_{ij}) = \boldsymbol{\Sigma}^{-1}$—the conditional correlation coefficients are independent of the values of the conditioning variables. They are identical to the partial correlation coefficients of the partial correlation matrix $\boldsymbol{\Pi} = (\pi_{ij})$ that can be obtained from $\boldsymbol{\Upsilon}$ as [14]

$$\pi_{ij} = f_{ij}(\boldsymbol{\Upsilon}) = -\frac{\upsilon_{ij}}{\sqrt{\upsilon_{ii}\upsilon_{jj}}} \quad \text{for } i, j \in \mathbb{V}, \quad i \neq j, \tag{1}$$

and $\pi_{ii} = 1$. Since $\boldsymbol{\Pi}$ is symmetric and has unit diagonal, it can conveniently be represented by a vector $\boldsymbol{\pi}$ indexed by $\mathbb{F}$.

### 2.3. Conditional independence graphs

Consider a graph $G = (\mathbb{E}, \mathbb{V})$. A graphical model on $G$ is defined as the set of all probability distributions that satisfy the Markov conditions specified by $G$. When the distributions are multivariate normal, we speak of a Gaussian graphical model. In this case, every missing edge $(i, j) \notin \mathbb{E}$ (i.e., every relation of conditional independence $x_i \perp\!\!\!\perp x_j|\boldsymbol{x}_{\mathbb{V}\setminus\{i,j\}}$) is equivalent to setting the corresponding partial correlation $\pi_{ij}$ or, equivalently, the concentration coefficient $\upsilon_{ij}$ to zero. In other words, the concentration matrix is constrained to lie in $\mathcal{M}_0^+(G)$. The reader could refer to [14] or [10] for more details.

### 2.4. Asymptotic result

We need the following result, that is a translation in terms of density function of the asymptotic property of the partial correlation coefficients, as shown in [12,11] in a much more general setting.

Define first the parameter change $\boldsymbol{\Upsilon} \overset{h}{\mapsto} (\boldsymbol{\pi}, \boldsymbol{\omega})$ with $\boldsymbol{\pi} = f(\boldsymbol{\Upsilon})$, $\boldsymbol{\omega} = g(\boldsymbol{\Upsilon})$, and

$$\begin{cases} \pi_{ij} = f_{ij}(\boldsymbol{\Upsilon}) = -\dfrac{\upsilon_{ij}}{\sqrt{\upsilon_{ii}\upsilon_{jj}}}, \\ \omega_i = g_i(\boldsymbol{\Upsilon}) = \upsilon_{ii}, \end{cases}$$

where we set $\boldsymbol{\omega} = (\omega_i)$. This parameter change is a one-to-one mapping. Also define

$$\phi(M, \boldsymbol{A}; \boldsymbol{\omega}, \boldsymbol{\pi})\,\mathrm{d}\boldsymbol{\omega}\,\mathrm{d}\boldsymbol{\pi} = l(M, D) \cdot |\boldsymbol{\Upsilon}|^{(M-D-1)/2} \exp\left[-\frac{M}{2}\mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{\Upsilon})\right] \mathrm{d}\boldsymbol{\Upsilon} \tag{2}$$

the distribution that corresponds to an (unconstrained) inverse Wishart distribution with $M$ degrees of freedom and scale matrix $\boldsymbol{A} = (a_{ij})$ [4, Appendix A] after reparameterization $\boldsymbol{\Upsilon} \mapsto (\boldsymbol{\omega}, \boldsymbol{\pi})$. Then, Roverato and Whittaker [12] and Roverato [11] showed that

$$\psi(\boldsymbol{\pi}) = \int 1_{h(\mathcal{M}^+)}(\boldsymbol{\pi}, \boldsymbol{\omega}) \cdot \phi(M, \boldsymbol{A}; \boldsymbol{\omega}, \boldsymbol{\pi})\,\mathrm{d}\boldsymbol{\omega} \tag{3}$$

asymptotically converges towards a Gaussian distribution with mean $\boldsymbol{p}$, where $\boldsymbol{p}$ is the sample vector of partial correlations, indexed by $\mathbb{F}$, and covariance matrix $c\boldsymbol{W}$, where we set $c = 1/M$ and $\boldsymbol{W} = (w_{ij,kl})$ indexed by $\mathbb{F} \times \mathbb{F}$, so that

$$w_{ij,kl} = \begin{pmatrix} -\frac{1}{2}a_{ii}^{-3/2}a_{ij}a_{jj}^{-1/2} \\ a_{ii}^{-1/2}a_{jj}^{-1/2} \\ -\frac{1}{2}a_{ii}^{-1/2}a_{ij}a_{jj}^{-3/2} \end{pmatrix}^{\mathrm{t}} \begin{pmatrix} 2a_{ik}^2 & 2a_{ik}a_{il} & 2a_{il}^2 \\ 2a_{ik}a_{jk} & a_{ik}a_{jl} + a_{il}a_{jk} & 2a_{il}a_{jl} \\ 2a_{jk}^2 & 2a_{jk}a_{jl} & 2a_{jl}^2 \end{pmatrix} \times \begin{pmatrix} -\frac{1}{2}a_{kk}^{-3/2}a_{kl}a_{ll}^{-1/2} \\ a_{kk}^{-1/2}a_{ll}^{-1/2} \\ -\frac{1}{2}a_{kk}^{-1/2}a_{kl}a_{ll}^{-3/2} \end{pmatrix}. \tag{4}$$

## 3. Structure learning

Let $(\boldsymbol{y}_n)_{n=1,\ldots,N}$ be $N$ independent realizations of a $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-distributed $D$-dimensional variable. Assume further that there exists a graph $G$ underlying the data, i.e., $\boldsymbol{\Upsilon} \in \mathcal{M}_0^+(G)$. In this section, the graph structure is unknown. Our objective is to calculate the posterior probability density function (pdf) of a potential graph structure and provide a numerical sampling scheme to approximate it.

Instead of making inference about a graph from estimated values of the partial correlations, we directly treat the graph as a mathematical object on which Bayesian analysis can be conducted. A convenient way to deal with graphs is to relate them to their supports. For a given conditional independence graph $G = (\mathbb{V}, \mathbb{E})$, the support $\boldsymbol{\gamma}$, indexed by $\mathbb{F}$, of the graph is defined as $\gamma_{ij} = 1$ if $(i, j) \in \mathbb{E}$, $\gamma_{ij} = 0$ otherwise. The graph support associated to a Gaussian graphical model can therefore be considered as a discrete latent variable or a state variable that is characteristic of the model.

### 3.1. Estimation of the graph support

Application of Bayes' theorem yields:

$$\mathrm{p}(\boldsymbol{\gamma}|\boldsymbol{y}) \propto \mathrm{p}(\boldsymbol{\gamma}) \cdot \mathrm{p}(\boldsymbol{y}|\boldsymbol{\gamma}). \tag{5}$$

The prior $\mathrm{p}(\boldsymbol{\gamma})$ can be chosen as desired to match the a priori information at hand. As to the likelihood $\mathrm{p}(\boldsymbol{y}|\boldsymbol{\gamma})$, it can be expanded through the marginalization and the chain rules:

$$\mathrm{p}(\boldsymbol{y}|\boldsymbol{\gamma}) = \int \mathrm{p}(\boldsymbol{\pi}, \boldsymbol{\omega}|\boldsymbol{\gamma}) \cdot \mathrm{p}(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega}\, \mathrm{d}\boldsymbol{\pi}. \tag{6}$$

### 3.2. Likelihood

According to Roverato [11], the likelihood $\mathrm{p}(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\Upsilon})$ reads

$$\mathrm{p}(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\Upsilon}) \propto |\boldsymbol{\Upsilon}|^{N/2} \exp\left[-\frac{N}{2}\mathrm{tr}(\boldsymbol{\Upsilon}\boldsymbol{S})\right],$$

where $\boldsymbol{S}$ is proportional to the sample covariance matrix,

$$\boldsymbol{S} = \sum_{n=1}^{N} (\boldsymbol{y}_n - \bar{\boldsymbol{y}})(\boldsymbol{y}_n - \bar{\boldsymbol{y}})^{\mathrm{t}},$$

and $\bar{\boldsymbol{y}}$ is the sampling mean:

$$\bar{\boldsymbol{y}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{y}_n.$$

Equivalently, consideration of a normal distribution for $\mathrm{p}(\boldsymbol{y}|\gamma, \boldsymbol{\mu}, \boldsymbol{\Upsilon})$ leads to the same result [4]. The likelihood $\mathrm{p}(\boldsymbol{y}|\gamma, \boldsymbol{\pi}, \boldsymbol{\omega})$ is exactly the same function, written as a function of $(\boldsymbol{\pi}, \boldsymbol{\omega})$, i.e.,

$$\mathrm{p}(\boldsymbol{y}|\gamma, \boldsymbol{\pi}, \boldsymbol{\omega}) \propto |h^{-1}(\boldsymbol{\pi}, \boldsymbol{\omega})|^{N/2} \exp\left[-\frac{N}{2} \operatorname{tr}(h^{-1}(\boldsymbol{\pi}, \boldsymbol{\omega})\boldsymbol{S})\right].$$

But, from Eq. (2), the right-hand side can be expressed as a $\phi$ function with parameters $M = N + D + 1$ and $\boldsymbol{A} = \boldsymbol{S}^{-1}$, leading to

$$\mathrm{p}(\boldsymbol{y}|\gamma, \boldsymbol{\pi}, \boldsymbol{\omega}) \propto \phi(N + D + 1, \boldsymbol{U}; \boldsymbol{\pi}, \boldsymbol{\omega}), \tag{7}$$

where we set $\boldsymbol{U} = \boldsymbol{S}^{-1}$.

### *3.3. Prior pdf*

Working with a parametrization $(\boldsymbol{\pi}, \boldsymbol{\omega})$, instead of $\boldsymbol{\Upsilon}$, naturally separates the component whose dimension is influenced by the graph, $\boldsymbol{\pi}$, from the one that is not, $\boldsymbol{\omega}$. Indeed, while changing from a graph to another one does not change the dimensionality of $\boldsymbol{\omega}$, yet it may drastically modify that of $\boldsymbol{\pi}$. Using the conditioning chain, we can write:

$$\mathrm{p}(\boldsymbol{\pi}, \boldsymbol{\omega}|\gamma) = \mathrm{p}(\boldsymbol{\pi}|\gamma) \cdot \mathrm{p}(\boldsymbol{\omega}|\gamma, \boldsymbol{\pi}). \tag{8}$$

#### *3.3.1. Prior for $\boldsymbol{\pi}$*

As we put it earlier, the most relevant prior information is the dimensionality of $\boldsymbol{\pi}$. For this reason, we again apply the chain rule to $\mathrm{p}(\boldsymbol{\pi}|\gamma)$, leading to

$$\mathrm{p}(\boldsymbol{\pi}|\gamma) = \mathrm{p}(\boldsymbol{\pi}_{\bar{\mathbb{E}}}|\gamma) \cdot \mathrm{p}(\boldsymbol{\pi}_{\mathbb{E}}|\gamma, \boldsymbol{\pi}_{\bar{\mathbb{E}}}),$$

where $\mathbb{E}$ is the edge set and $\boldsymbol{\pi}_{\mathbb{E}}$ stands for $(\pi_{ij})_{(i,j)\in\mathbb{E}}$. Knowing $G$ imposes a strong constraint on $\boldsymbol{\pi}_{\bar{\mathbb{E}}}$, since this vector must be zero, i.e.,

$$\mathrm{p}(\boldsymbol{\pi}_{\bar{\mathbb{E}}}|\gamma) = \delta(\boldsymbol{\pi}_{\bar{\mathbb{E}}}).$$

Having no further information relative to $\boldsymbol{\pi}_{\mathbb{E}}$, except that $\boldsymbol{\pi} \in f(\mathcal{M}^+)$, one must set $\mathrm{p}(\boldsymbol{\pi}_{\mathbb{E}}|\gamma, \boldsymbol{\pi}_{\bar{\mathbb{E}}})$ to a uniform density function:

$$\mathrm{p}(\boldsymbol{\pi}_{\mathbb{E}}|\gamma, \boldsymbol{\pi}_{\bar{\mathbb{E}}}) = \frac{1}{V(G)}$$

for $\boldsymbol{\pi}_{\mathbb{E}}$ so that $\boldsymbol{\pi} \in f(\mathcal{M}^+)$, and zero otherwise. In this equation, $V(G)$ stands for the volume of $f(\mathcal{M}^+)$. Note that, having $\boldsymbol{\pi}_{\bar{\mathbb{E}}} = \mathbf{0}$, any vector $\boldsymbol{\pi} = (\boldsymbol{\pi}_{\bar{\mathbb{E}}}, \boldsymbol{\pi}_{\mathbb{E}})$ in $f(\mathcal{M}^+)$ will also belong to $f(\mathcal{M}_0^+(G))$. Finally

$$\mathrm{p}(\boldsymbol{\pi}|\gamma) = \frac{1}{V(G)} \cdot \delta(\boldsymbol{\pi}_{\bar{\mathbb{E}}}). \tag{9}$$

*3.3.2. Prior for $\boldsymbol{\omega}$*

As to $\boldsymbol{\omega}$, we only know that, given $G$ and $\boldsymbol{\pi}$, $(\boldsymbol{\pi}, \boldsymbol{\omega})$ must lie in $h(\mathcal{M}^+)$. Having no further information relative to $\boldsymbol{\omega}$, we set it to a uniform prior, i.e.,

$$\mathrm{p}(\boldsymbol{\omega}|\gamma, \boldsymbol{\pi}) \propto 1_{h(\mathcal{M}^+)}(\boldsymbol{\pi}, \boldsymbol{\omega}). \tag{10}$$

*3.3.3. Prior for $(\boldsymbol{\pi}, \boldsymbol{\omega})$*

Plugging Eqs. (9) and (10) into Eq. (8), the prior probability reads

$$\mathrm{p}(\boldsymbol{\pi}, \boldsymbol{\omega}|\gamma) \propto \frac{1}{V(G)} \cdot \delta(\boldsymbol{\pi}_{\bar{\mathbb{E}}}) \cdot 1_{h(\mathcal{M}^+)}(\boldsymbol{\pi}, \boldsymbol{\omega}). \tag{11}$$

*3.4. Posterior pdf*

Bringing together Eqs. (7) and (11) into Eq. (6) leads to

$$\begin{aligned}
\mathrm{p}(\boldsymbol{y}|\gamma) &\propto \int \left[ \frac{1}{V(G)} \cdot \delta(\boldsymbol{\pi}_{\bar{\mathbb{E}}}) \cdot 1_{h(\mathcal{M}^+)}(\boldsymbol{\pi}, \boldsymbol{\omega}) \right] \cdot \left[ \phi\,(N, \boldsymbol{S}; \boldsymbol{\pi}, \boldsymbol{\omega}) \right] \mathrm{d}\boldsymbol{\omega}\, \mathrm{d}\boldsymbol{\pi} \\
&= \frac{1}{V(G)} \int \delta(\boldsymbol{\pi}_{\bar{\mathbb{E}}}) \cdot 1_{h(\mathcal{M}^+)}(\boldsymbol{\pi}, \boldsymbol{\omega}) \cdot \phi(N, \boldsymbol{S}; \boldsymbol{\pi}, \boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega}\, \mathrm{d}\boldsymbol{\pi} \\
&= \frac{1}{V(G)} \int \delta(\boldsymbol{\pi}_{\bar{\mathbb{E}}}) \left[ \int 1_{h(\mathcal{M}^+)}(\boldsymbol{\pi}, \boldsymbol{\omega}) \cdot \phi(N, \boldsymbol{S}; \boldsymbol{\pi}, \boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\pi} \\
&= \frac{1}{V(G)} \int \delta(\boldsymbol{\pi}_{\bar{\mathbb{E}}}) \cdot \psi(\boldsymbol{\pi})\, \mathrm{d}\boldsymbol{\pi}.
\end{aligned}$$

Applying the result of Section 2.4, $\psi(\boldsymbol{\pi})$ asymptotically converges toward $\mathcal{N}(\boldsymbol{p}, c\boldsymbol{W})$, with $c = M^{-1}$ and where $\boldsymbol{W}$ is calculated as a function of $\boldsymbol{U}$. The integral then rereads:

$$\begin{aligned}
\mathrm{p}(\boldsymbol{y}|\gamma) &\overset{a}{\propto} \frac{1}{V(G)} \int \delta(\boldsymbol{\pi}_{\bar{\mathbb{E}}}) \cdot \mathcal{N}(\boldsymbol{p}, c\boldsymbol{W})\, \mathrm{d}\boldsymbol{\pi} \\
&= \frac{\mathcal{N}(\boldsymbol{p}_{\bar{\mathbb{E}}}, c\boldsymbol{W}_{\bar{\mathbb{E}}\bar{\mathbb{E}}}; \mathbf{0})}{V(G)},
\end{aligned}$$

where $\overset{a}{\propto}$ stands for "asymptotically proportional to". The asymptotic posterior pdf for $G$ is finally given by putting this last result into Eq. (5):

$$\mathrm{p}(\gamma|\boldsymbol{y}) \overset{a}{\propto} \mathrm{p}(\gamma) \cdot \frac{\mathcal{N}(\boldsymbol{p}_{\bar{\mathbb{E}}}, c\boldsymbol{W}_{\bar{\mathbb{E}}\bar{\mathbb{E}}}; \mathbf{0})}{V(G)}, \tag{12}$$

with $c = N^{-1}$.

### 3.5. Computational issues

Once that the posterior probability $\mathrm{p}(\boldsymbol{\gamma}|\boldsymbol{y})$ has been calculated in closed form, some points remain to be tackled for an efficient use. $V(G)$ must be calculated (Section 3.5.1); the posterior pdf must be approximated (Section 3.5.2); and last, estimates must be defined (Section 3.5.3).

#### 3.5.1. Estimation of $V(G)$

The parameter $V(G)$ in Eq. (12) does not have any closed form. We here propose a numerical scheme to approximate it. Since all off-diagonal elements of a partial correlation matrix lie in $[-1, 1]$, an upper bound for $V(G)$ is $2^{D(D-1)}$. Hence, if we define $V_D = 2^{D(D-1)/2}$, we can estimate the fraction $k(G) = V(G)/V_D$ by use of a rejection sampling scheme. Drawing $L$ (typically 10,000) samples of $\boldsymbol{\pi}_{\mathbb{E}}$'s uniformly in the hypercube $[-1, 1]^{|\mathbb{E}|}$, $k(G)$ can be approximated by

$$k(G) \approx \frac{1}{L} \sum_{l=1}^{L} 1_{f(\mathcal{M}^+)}(\boldsymbol{\pi}^{[l]}).$$

Furthermore, it can be shown (cf. Appendix A) that $\boldsymbol{\pi}$ belongs to $f(\mathcal{M}^+)$ if and only if $2\boldsymbol{I} - \boldsymbol{\Pi}$ is in $\mathcal{M}^+$, which relates this set to the so-called elliptope [6,9,7] and provides a straightforward test for $\boldsymbol{\pi} \in f(\mathcal{M}^+)$.

Note that $V(G)$ is an a priori parameter and, as such, does not depend on the data. These values could therefore be calculated independently, once and for all. However, we would have to calculate and store $2^{D(D-1)/2}$ such volumes, which would quickly prove to be infeasible for large $D$. We rather take advantage of the sampling scheme developed (see next section), that essentially concentrates on the most probable graphs: Each time the probability of a graph is required for the first time in Gibbs sampling, we calculate the corresponding volume and then store it in case the probability of the same graph is required again.

#### 3.5.2. Sampling issues

Even though the joint posterior pdf given by Eq. (12) does not belong to any known pdf family, stochastic simulation can be utilized to approximate it by implementing a Gibbs sampler [13]. More precisely, we follow the scheme proposed in [4]:

- obtain 100 graphs by the following sampling scheme: every edge of each graph is sampled independently, with a probability $1/2$ to be present.
- from these 100 graphs, sample 10 graphs using importance resampling; these graphs are taken as seeds for the Gibbs sampler;
- starting with each seed, for each iteration $n$, alternately sample each edge support $\gamma_{ij}^{[n]}$ according to the conditional posterior pdf $\mathrm{p}(\gamma_{ij}|\boldsymbol{y}, \boldsymbol{\gamma}_{\mathbb{F}\setminus(i,j)})$ obtained from Bayes' theorem:

$$\begin{aligned}\mathrm{p}(\gamma_{ij}|\boldsymbol{y}, \boldsymbol{\gamma}_{\mathbb{F}\setminus(i,j)}) &= \frac{\mathrm{p}(\gamma_{ij}, \boldsymbol{\gamma}_{\mathbb{F}\setminus(i,j)}|\boldsymbol{y})}{\mathrm{p}(\boldsymbol{\gamma}_{\mathbb{F}\setminus(i,j)}|\boldsymbol{y})} \\ &\propto \mathrm{p}(\gamma_{ij}, \boldsymbol{\gamma}_{\mathbb{F}\setminus(i,j)}|\boldsymbol{y}),\end{aligned}$$

  which is given by Eq. (12).

Gelman et al. [4] showed that the convergence can be tracked by comparing the within- and between-variances of the samples and that the joint posterior pdf can then be approximated by the frequency histogram resulting from the second half of the samples.

*3.5.3. Estimates*

Since a joint posterior pdf of many variables is hard to interpret, two estimators are proposed here as "summaries".

The Maximum a posteriori graph (MAP graph) is defined as

$$\gamma^{\mathrm{MAP}} = \mathrm{argmax}_{\gamma}\,[\mathrm{p}(\gamma|\boldsymbol{y})].$$

It is obtained by a global and objective criterion, and is independent of any threshold.

Another estimate is the mean graph, $\mathrm{E}[\gamma]$. Though not a graph, this estimate allows more flexibility in the interpretation, since its values can vary between 0 and 1. Besides, its components are equal to the marginal posterior pdfs and have an important value as such.

## 4. Simulations

This section assesses the features of the structure learning relative to the asymptotic results assumed, and the ability to work on the whole set of possible independence graphs. The programs were developed with Matlab 6.0® (The MathWorks, Inc.) and run on a Sun SPARC Ultra 10 workstation. In the following, a uniform prior will be assumed for graph supports.

### *4.1. Data*

To assess structure learning, we simulated 100 samples from a 4-dimensional variable with the following concentration matrix and corresponding partial correlation matrix:

$$\boldsymbol{\Upsilon} = \begin{pmatrix} 1 & -0.4 & 0 & -0.4 \\ -0.4 & 1 & -0.4 & 0 \\ 0 & -0.4 & 1 & -0.4 \\ -0.4 & 0 & -0.4 & 1 \end{pmatrix}, \quad \text{i.e., } \boldsymbol{\Pi} = \begin{pmatrix} 1 & 0.4 & 0 & 0.4 \\ 0.4 & 1 & 0.4 & 0 \\ 0 & 0.4 & 1 & 0.4 \\ 0.4 & 0 & 0.4 & 1 \end{pmatrix}.$$

The corresponding graph, shown in Fig. 1(a), is the simplest example of a non-decomposable graph and, as such, very few methods would be able to estimate it correctly from the data.

### *4.2. Structure learning*

Two points of interest were assessed here: the pertinence of the asymptotic posterior and the accuracy of the sampling scheme detailed in Section 3.5.2. First, the probability of each of the 64 possible graphs was calculated using Eq. (12). The results are given in Fig. 1(b). The most probable graph was the true underlying structure, with a probability well above any other graph. The two second-more probable graphs were very similar to the MAP, still reinforcing the validity of the estimate. When the amount of data increased, further simulations (not shown here) showed that the probability of the true graph was observed to increase and tend to 1. Concerning the numerical approximation, the Gibbs sampler converged after only 200 steps on the proposed dataset. The sample MAP estimate recovered the true independence graph, and the mean estimate given by the sampler also approximated the mean of the posterior pdf very accurately, as shown in Fig. 1(c). Finally, comparison of sample and exact joint pdfs exhibited no visible difference in Fig. 1(b).

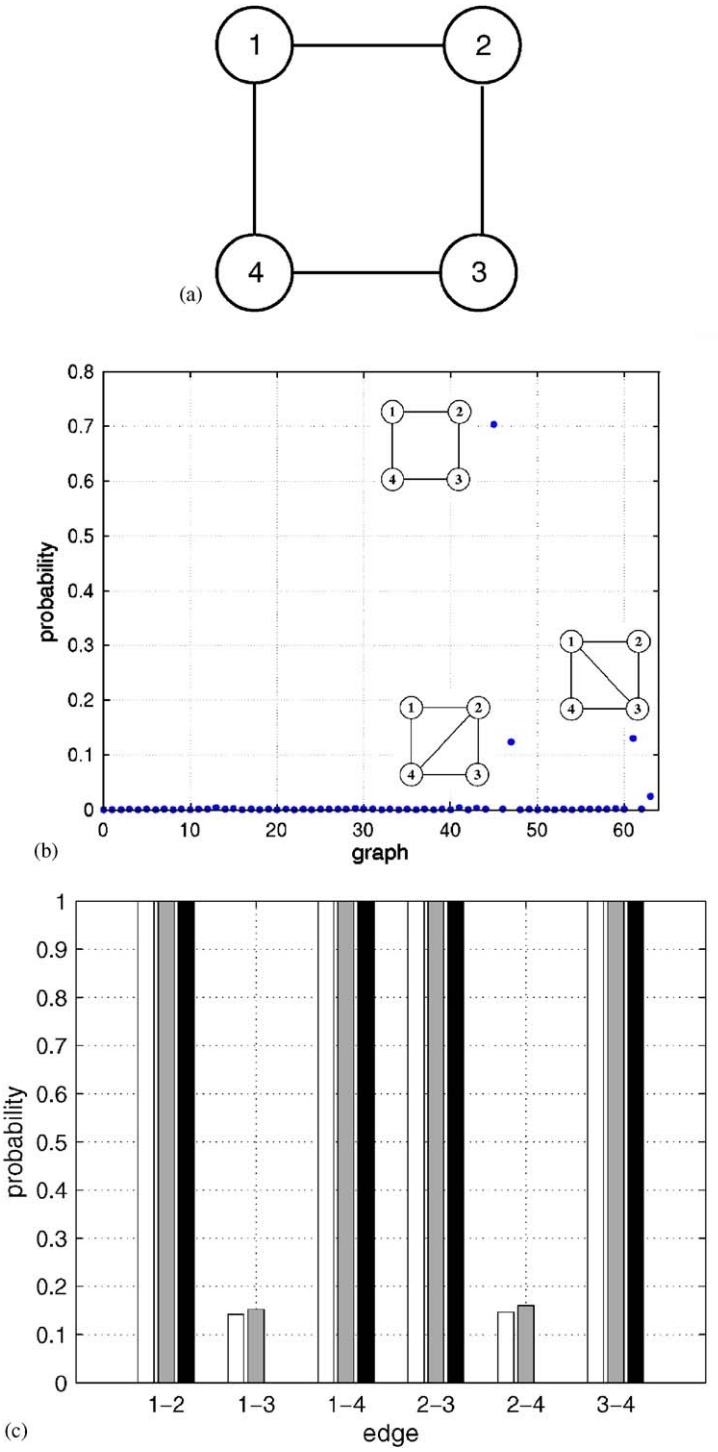



Fig. 1. Structure learning: (a) Independence graph. (b) Asymptotic posterior pdf for all possible structures calculated for each graph. $x$-coordinate is defined by binary-to-decimal conversion of $(\gamma_{12}, \gamma_{13}, \gamma_{14}, \gamma_{23}, \gamma_{24}, \gamma_{34})$. For instance, the true graph $(\gamma_{12} = 1, \gamma_{13} = 0, \gamma_{14} = 1, \gamma_{23} = 1, \gamma_{24} = 0, \gamma_{34} = 1)$ has an $x$-coordinate of bin2dec$(101101) = 45$. (c) For each edge $i - j$: proportion of graphs with $\gamma_{ij} = 1$ after convergence of Gibbs sampling (white bar); $(\mathrm{E}[\boldsymbol{\gamma}|\boldsymbol{y}])_{ij}$ from direct calculation (gray bar); sample MAP graph (black bar).

## 5. Real data

In this section, we developed two examples from the literature: "Frets' Heads" and the HIV Study example.

### 5.1. "Frets' Heads"

"Frets' Heads" [14] is an easy to understand, yet computationally hard problem. Frets reported measurements of the head length and head breadth of the first and second adult son in a sample of $N = 25$ families. The corresponding partial correlation matrix is reported in Table 1. This example is of great interest, in the sense that it allows to stress the various problems arising in model selection, and especially the importance of the optimization method used. Whittaker [14] showed that the minimal graph differed depending on the method used: selection based on the edge exclusion deviances, backward elimination with deviance difference stopping rule, backward elimination with overall deviance stopping rule, or two-step procedure. Giudici and Green [5] proposed a treatment for the set of decomposable graphs.

Structure learning was performed. The algorithm converged in less than 1500 steps, and the corresponding results are summarized in Figs. 2 and 3. There was no difference between approximated (from Gibbs sampling) and exact probabilities. The MAP estimate found here is in accordance with Whittaker [14], with a probability of 0.2176, but other graphs were also found to have a non-negligible probability.

### 5.2. HIV study data

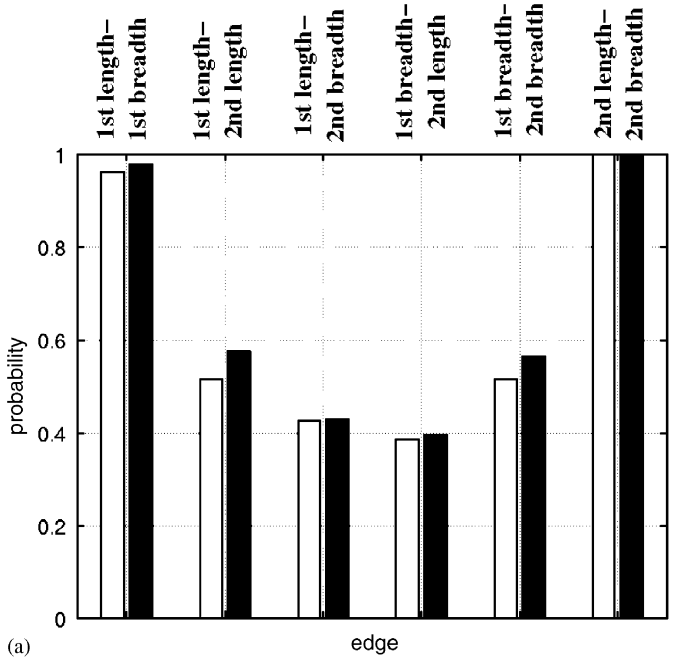
This data set was used in [11] to exemplify the use of the Gaussian approximation for the partial correlation matrix. It originates from a study investigating early diagnosis of HIV infection in children from HIV positive mothers. The variables are related to various measures on blood and its components: $x_1$ and $x_2$ immunoglobin G and A, respectively; $x_4$ the platelet count; $x_3$, $x_5$ lymphocyte B and T4, respectively; and $x_6$ the T4/T8 lymphocyte ratio. The observed partial correlation matrix is given in Table 2. The model assumed in [11] is given in Fig. 4(a). The results of the structure learning are summarized in Figs. 4(b), 4(c), and 5.

From both the posterior marginal probabilities and the most probables graphs, there is overwhelming support in favor the presence of links 1–2, 1–5, 1–6, 2–5, 3–5, 3–6, and 5–6. In contrast, links 2–4 and 3–4 have low probability. Interestingly, while using the model pictured in Fig. 4(a), Roverato [11] acknowledged that the correlations corresponding to $x_4$ had strong probability around zero and hypothesized that the model was overparametrized. Our study nicely supports this conjecture.

Table 1
Frets' Heads example

| | | | | |
|---|---|---|---|---|
| $x_1$ | 1 | | | |
| $x_2$ | 0.4252 | 1 | | |
| $x_3$ | 0.2225 | 0.1319 | 1 | |
| $x_4$ | 0.1523 | 0.2247 | 0.6256 | 1 |
| | $x_1$ | $x_2$ | $x_3$ | $x_4$ |

Partial correlation matrix.



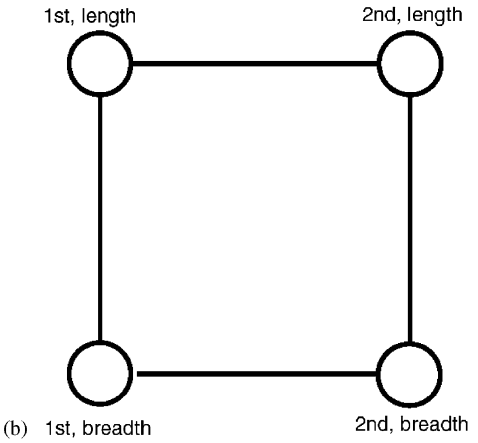




Fig. 2. "Frets' Heads" example: (a) Comparison of $(\mathrm{E}[\gamma|\boldsymbol{y}])_{ij}$ from the real posterior pdf (white bar) and estimated from Gibbs sampling (black bar). (b) MAP estimators for the independence graph.

## 6. Discussion

The new developments proposed in this paper enabled us to consider independence graphs as mathematical objects of their own and draw inference about them in a very robust, yet flexible way. The Bayesian approach has the dramatic advantage that uncertainty can very easily be embedded in the analysis *and* in the programming. No prior information is needed about potential graphs, since structure learning is performed on the whole set of models. On the other

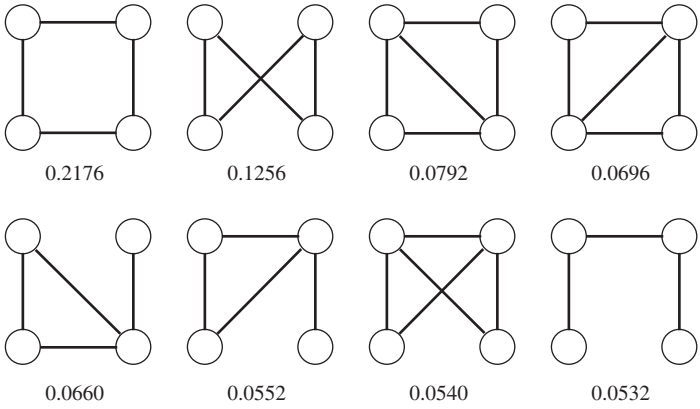



Fig. 3. "Frets' Heads" example. Eight most probable graphs, with corresponding probabilities.

Table 2
HIV Study example

| | | | | | | |
|---|---|---|---|---|---|---|
| $x_1$ | 1 | | | | | |
| $x_2$ | 0.483 | 1 | | | | |
| $x_3$ | 0.220 | 0.057 | 1 | | | |
| $x_4$ | −0.040 | −0.133 | 0.149 | 1 | | |
| $x_5$ | 0.253 | −0.124 | 0.523 | 0.179 | 1 | |
| $x_6$ | −0.276 | −0.314 | −0.183 | 0.064 | 0.213 | 1 |
| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |

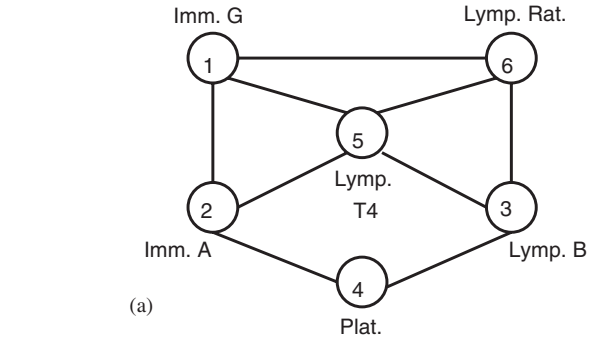
Partial correlation matrix.

hand, prior information concerning structural relationships, such as independence, can easily be incorporated.

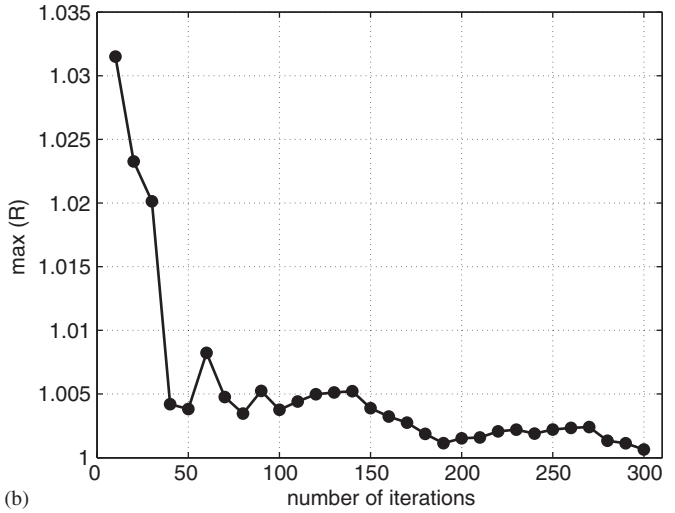
The algorithm introduced has the advantage of performing a stochastic and global search. Its speed could be improved if, instead of calculating $V(G)$ through a rejection sampling scheme, a closed form were available for the constraint volume. Different graphs can have the same constraint set, and hence the same associated volume, depending on their structural properties. A first step in this direction could be done by application of results developed in [8]. Other more effective sampling schemes could also be developed, that concentrate on a set smaller than the hypercube. For such bounding sets, see for instance [7]. Another possibility would be to replace $V(G)$ by a crude approximation, e.g., $2^{|\mathbb{E}|}$, into Eq. (12) to perform the Gibbs sampling, and then recalculate the posterior probability of all graphs that have been obtained with the sampling according to the true asymptotic form. Using $2^{|\mathbb{E}|}$ instead of $V(G)$ and throwing away some constraint also allows to obtain a crude approximation of $\mathrm{p}(\gamma_{ij}|\boldsymbol{y})$:

$$\mathrm{p}(\gamma_{ij} = 1|\boldsymbol{y}) \approx \frac{\frac{1}{2}}{\frac{1}{2} + \mathcal{N}(p_{ij}, cw_{ij,ij}; 0)}.$$

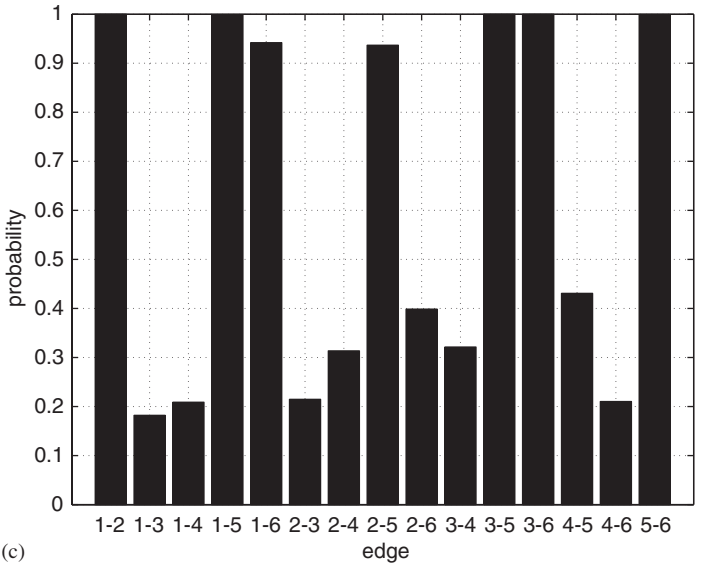
This approximation could, in turn, be used to sample the first 100 graphs, instead of a—certainly too diffuse—probability of 1/2. This would speed up the number of iterations required for the algorithm to converge.




(a)




(b)




(c)

Fig. 4. HIV Study example: (a) Model assumed. (b) Convergence monitoring. (c) Marginal probabilities.

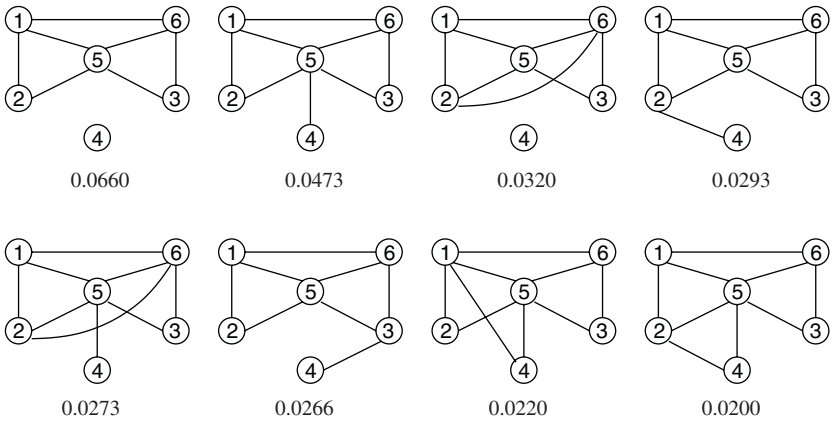



Fig. 5. HIV Study example. Eight most probable graphs, with corresponding probabilities.

## 7. Conclusion

In this paper, we proposed that independence graphs be treated as mathematical objects by considering their supports. From there, the issue of model selection became one of parameter estimation that could easily be handled in a Bayesian framework. Applying asymptotic results, we were able to derive closed forms for the joint posterior probability of a graph to be the independence structure underlying a data set. We also showed how this approach could be implemented for a full statistical analysis on the space of all possible models.

Further research includes optimization and development of this technique for mass analysis, in order to retrieve independence structure from systems having a huge amount of variables and data.

## Acknowledgements

The authors are grateful to Mélanie Pélégrini-Issac and Pierre Bellec for helpful comments and to Monique Laurent for guiding them in their research on elliptopes.

## Appendix A. Proof of the results for $f(\mathcal{M}^+)$

If $(2\boldsymbol{I} - \boldsymbol{\Pi}) \in \mathcal{M}^+$, then $\boldsymbol{\Upsilon} = 2\boldsymbol{I} - \boldsymbol{\Pi}$ is a positive definite concentration matrix whose associated partial correlation matrix is $\boldsymbol{\Pi}$, and thus $\boldsymbol{\pi} \in f(\mathcal{M}^+)$. Conversely, if $\boldsymbol{\pi} \in f(\mathcal{M}^+)$, then there exists a concentration matrix $\boldsymbol{\Upsilon}$ so, that $\boldsymbol{\pi} = f(\boldsymbol{\Upsilon})$. Let $\boldsymbol{\Pi}$ be the corresponding matrix and $\boldsymbol{\Omega} = (\omega_{ij}) = 2\boldsymbol{I} - \boldsymbol{\Pi}$. For any $D$-dimensional vector $\boldsymbol{x} = (x_i) \neq \boldsymbol{0}$, the following holds:

$$\sum_{i,j} \omega_{ij} x_i x_j = \sum_{i,j} \frac{\upsilon_{ij}}{\sqrt{\upsilon_{ii}\upsilon_{jj}}} x_i x_j = \sum_{i,j} \upsilon_{ij} \left(\frac{x_i}{\sqrt{\upsilon_{ii}}}\right)\left(\frac{x_j}{\sqrt{\upsilon_{jj}}}\right) > 0,$$

since $\boldsymbol{\Upsilon}$ is positive definite. $\boldsymbol{\Omega}$ is hence also positive definite, and $(2\boldsymbol{I} - \boldsymbol{\Pi}) \in \mathcal{M}^+$.

## References

[1] G.L. Bretthorst, An introduction of parameter estimation using Bayesian probability, in: P. Fougere (Ed.), Maximum Entropy and Bayesian Methods, Kluwer Academic Publishers, Dordrecht, 1990, pp. 53–79, URL http://bayes.wustl.edu/glb/bib.html.
[2] A.P. Dawid, Conditional independence, In: S. Kotz, C.B. Read, D.L. Banks, (Eds.), Encyclopedia of Statistical Sciences, vol. 2. Wiley, New York, 1998, pp. 146–155.
[3] A.P. Dawid, S.L. Lauritzen, Hyper Markov laws in the statistical analysis of decomposable graphical models, Ann. Statist. 21 (1993) 1272–1317.
[4] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, Chapman & Hall, London, 1998.
[5] P. Giudici, P.J. Green, Decomposable graphical Gaussian model determination, Biometrika 86 (1999) 785–801.
[6] M. Laurent, On a positive semidefinite relaxation of the cut polytope, Linear Algebra Appl. 223/224 (1995) 439–461.
[7] M. Laurent, A tour d'horizon on positive demidefinite and Euclidean distance matrix completion problems, Fields Inst. Commun. 18 (1998) 51–76.
[8] M. Laurent, On the sparcity order of a graph and its deficiency in chordality, Combinatorica 21 (2001) 543–570.
[9] M. Laurent, S. Poljak, F. Rendl, Connections between semidefinite relaxations of the max-cut and stable set problems, Math. Progr. 77 (1997) 225–246.
[10] S.L. Lauritzen, Graphical Models, Oxford University Press, Oxford, 1996.
[11] A. Roverato, Asymptotic prior to posterior analysis for graphical Gaussian models, in: M. Vichi, O. Opitz (Eds.), Classification and Data Analysis, Springer, Berlin, 1999, pp. 335–342.
[12] A. Roverato, J. Whittaker, The Isserlis matrix and its application to non-decomposable graphical Gaussian models, Biometrika 85 (1998) 711–725.
[13] J.J.K.O. Ruanaidh, W.J. Fitzgerald, Numerical Bayesian Methods Applied to Signal Processing, Statistics and Computing, Springer, New York, 1996.
[14] J. Whittaker, Graphical Models in Applied Multivariate Statistics, Wiley, Chichester, 1990.