# The ratio of normalizing constants for Bayesian graphical Gaussian model selection

Mohammadi, A.; Massam, Helene; Letac, Gerard

**Publication date**
2017

**Document Version**
Final published version

# The ratio of normalizing constants for Bayesian graphical Gaussian model selection

A. Mohammadi
Tilburg University
a.mohammadi@uvt.nl

H. Massam
York University
massamh@yorku.ca

G. Letac
Université de Toulouse
gerard.letac@math.univ-toulouse.fr

June 15, 2017

**Abstract**

The ratio of normalizing constants for the $G$-Wishart distribution, for two graphs differing by an edge $e$, has long been a bottleneck in the search for efficient model selection in the class of graphical Gaussian models. We give an accurate approximation to this ratio under two assumptions: first we assume that the scale of the prior is the identity, second we assume that the set of paths between the two ends of $e$ are disjoint. The first approximation does not represent a restriction since this is what statisticians use. The second assumption is a real restriction but we conjecture that similar results are also true without this second assumption. We shall prove it in subsequent work.

This approximation is simply a ratio of Gamma functions and thus need no simulation. We illustrate the efficiency and practical impact of our result by comparing model selection in the class of graphical Gaussian models using this approximation and using current Metropolis-Hastings methods. We work both with simulated data and a complex high-dimensional real data set. In the numerical examples, we do not assume that the paths between the two end points of edge $e$ are disjoint.

**Keywords**: Bayesian Inference; Gaussian Graphical Models; Model Selection; G-Wishart distribution; Normalizing Constant; Birth-death MCMC.

# 1 Introduction

Given an undirected graph $G = (V, E)$ where $V$ is a finite set $V = \{1, \ldots, p\}$ and $E$ is the set of undirected edges, we say that the Gaussian $N(0, \Sigma)$ variable $X = (X_i, i \in V)$ is Markov with respect to $G$ if $X_i$ is independent of $X_j$ given all the other variables whenever there no edge $(i, j)$ in $E$. It is well-known that in that case the precision matrix $K = \Sigma^{-1}$ belongs to the cone $P_G$ of positive definite matrices with $K_{ij} = 0$ whenever $(i, j) \notin E$. One can then define the graphical Gaussian model Markov with respect to a given graph $G$ as the family of distributions

$$\mathcal{N}_G = \{N(0, \Sigma) \mid K = \Sigma^{-1} \in P_G\}.$$

Graphical Gaussian models form nowadays one of the basic tools used to analyze high-dimensional complex continuous data. Model selection in this class of models has thus been the topic of much research both from the frequentist and Bayesian point of view. The model is defined by both the graph $G$ and the precision matrix $K$. Model search in the frequentist framework is done by maximizing a penalized likelihood (see Friedman et al. (2008)): this yields simultaneously the best (in that sense) $G$ and $K$ by determining which entries of the covariance matrix are zero and estimating the others. In the Bayesian framework, model search has traditionally been based on the comparison of the posterior distribution of each model, each model being represented by a graph $G$. The selected models are the models with the highest posterior probabilities and the corresponding $K$ is then estimated through sampling of the posterior distribution of $K$. The conjugate prior for $K \in P_G$ is the $G$-Wishart as defined by Roverato (2002). The density of the $G$-Wishart can be written as

$$f(K \mid G) = \frac{1}{I_G(\delta, D)} |K|^{\frac{\delta-2}{2}} \exp\{-\frac{1}{2}\langle K, D\rangle\} \mathbf{1}_{P_G}(K) \tag{1}$$

where $|K|$ denotes the determinant of $K$, $\langle K, D\rangle = \mathrm{tr}(KD)$ is the inner product of $K$ and $D$ in the space of symmetric matrices and $I_G(\delta, D)$ is the normalizing constant of the $G$-Wishart which is finite for $\delta > 2$ and $D > 0$. Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ from the Gaussian distribution in $\mathcal{N}_G$, let $S = \sum_{i=1}^n x_i x_i^t$. The marginal density of $\mathbf{x}$, which is also the unnormalized posterior density of $G$ given $\mathbf{x}$, is then

$$f(\mathbf{x} \mid G) = (2\pi)^{-np/2} \frac{I_G(\delta + n, D + S)}{I_G(\delta, D)} \int_{P_G} |K|^{\frac{\delta+n-2}{2}} \exp\{-\frac{1}{2}\langle K, (D + S)\rangle\}\, dK\ .$$

To do model selection, one must run a Markov chain or some stochastic search to move through the space of graphs, see (Jones et al., 2005, Scott and Carvalho, 2008, Wong et al., 2003). This approach is not feasible for high-dimensional data for two reasons: the Markov chains are slow and the prior and posterior normalizing constants for each graph $G$ visited are hard to evaluate numerically.

Another approach to Bayesian model selection is to run a Markov chain on the joint space of $(\Sigma, G)$, see (Giudici and Green, 1999) or $(K, G)$, see (Dobra and Lenkoski, 2011, Dobra et al., 2011, Wang and Li, 2012, Mohammadi and Wit, 2015). The joint posterior distribution of $(K, G)$ is

$$f(K, G \mid \mathbf{x}) \quad = \quad \frac{|K|^{\frac{\delta+n-2}{2}}}{(2\pi)^{-np/2} I_G(\delta, D)} \exp\{-\frac{1}{2}\langle K, (D+S)\rangle\}$$

The advantage of this approach is that the posterior normalizing constant does not come into play. To compute the acceptance probabilities in the chain, the only quantity which is computationally expensive is the ratio of prior normalizing constants

$$\frac{I_{G^{-e}}(\delta, \mathbb{I}_p)}{I_G(\delta, \mathbb{I}_p)} \tag{2}$$

where $G^{-e} = (V, E^{-e})$ is the graph obtained from $G$ by removing one edge $e$ and $D = \mathbb{I}_p$. A recent paper Uhler et al. (2014) gives the exact analytic expression of these integrals. However, this expression is complicated and impossible to implement at the present time. The purpose of this paper is to give an accurate approximation to (2) which, itself, leads to accurate model selection with minimal scale-free computational burden.

In practice, the normalizing constants in (2) are computed in two ways, either using the Laplace approximation to an integral or the method given by Atay-Kayis and Massam (2005). The Laplace approximation is accurate only if the density of the prior distribution is similar in shape to a multivariate normal distribution and, like for the Wishart distribution, this requires that the shape parameter $\delta$ be high. Since, in order to minimize the impact of the prior distribution on inference, $\delta$ is traditionally chosen to be 3, which is not high, this approximation is not accurate. Atay-Kayis and Massam (2005) expressed $I_G(\delta, D)$ as the product of a constant and an expected value. For $D = \mathbb{I}_p$, this expression is

$$I_G(\delta, \mathbb{I}_p) = \prod_{i=1}^{p} 2^{\frac{\delta+\nu_i}{2}} (2\pi)^{\nu_i/2} \Gamma(\frac{\delta+\nu_i}{2}) \mathbb{E}(f_E(\psi_E)) \tag{3}$$

where, for a given order of the vertices, $\nu_i$ is the number of neighbours of vertex $i$ which have a numbering larger than or equal to $i+1$, $\psi_E$ is the incomplete upper triangular matrix with entries the free entries of the Cholesky decomposition $K = \psi^t \psi$, $f_E$ is a function depending on $G$ and more particularly on $E$. The expected value is taken with respect to a product of independent standard normal and chi-square distributions. Evaluating (2) is therefore equivalent to evaluating

$$\frac{\mathbb{E}(f_{E^{-e}}(\psi_{E^{-e}}))}{\mathbb{E}(f_E(\psi_E))}\ , \tag{4}$$

where $E^{-e}$ is the edge set of the graph $G^{-e}$ obtained from $G = (V, E)$ by removing the edge $e$. We will show that, under reasonable conditions of sparsity, we have the following approximation of (2)

$$\frac{I_{G^{-e}}(\delta, \mathbb{I}_p)}{I_G(\delta, \mathbb{I}_p)} \approx \frac{1}{2\sqrt{\pi}} \frac{\Gamma(\frac{\delta+d}{2})}{\Gamma(\frac{\delta+d+1}{2})}$$

where $d$ is the number of minimal paths of length 2 linking the two end points of $e$. Under the assumption that the minimal paths (that is paths without a chord) linking the two vertices of $e$ in $G^{-e}$ are disjoint, we show in equation (22) below that the accuracy of this approximation depends on the number and on the length of the paths linking the two end vertices of the edge $e$. We illustrate the accuracy of our approximation through simulations, using numerous configurations, and we show that if the number of paths of length 3 or more is not too large, say 5, the approximation is of the order of $10^{-2}$. A relative error of this order does not truly affect the model search. We will follow the method of Mohammadi and Wit (2015) to do a model search using both real and simulated data. We will see that using this approximation actually yields better precision for the identification of the true model than if we actually evaluate (4). This means that graphical Gaussian model selection can be done for high-dimensional data in a fast, scale-free and accurate manner with minimal computational burden.

Though, in the numerical examples considered, the assumption that the minimal paths linking the two end vertices of $e$ are disjoints is not satisfied, the approximation seems to work and the model search gives good results. We conjecture that an approximation similar to (22) holds in general. This is the topic of future work.

# 2 The ratio of prior normalizing constants

Let $G = (V, E)$ be an undirected graph and $G^{-e}$ the graph obtained by removing a given edge $e$ from $E$. Without loss of generality, in this section, we will assume that the numbering of the $p$ vertices is such that $e = (p-1, p)$. The aim of this section is to express ratio (4) in such a way that, in the next section, we can deduce an approximation to it. This is done in Proposition 2.1. As we shall see below, the accuracy of this approximation depends on the number of minimal paths between the two vertices $q = p - 1$ and $p$ of $e$. A path between $q$ and $p$ is said to be minimal if it has no chord.

In a first step, we recall how (3) was obtained in Atay-Kayis and Massam (2005). Let $\psi$ be an upper triangular matrix with positive diagonal elements such $K = \psi^t \psi$ is the Cholesky decomposition of $K$. Let $\overline{E}$ be the complement of $E$ in the set of all possible edges in a graph with vertex set $V$, i.e. $\bar{E}$ indexes the missing edges of $G$.

Similarly $\overline{E^{-e}}$ denotes the missing edges of $G^{-e}$. It was shown that

$$\psi_E = \{\psi_{ij}, (i,j) \in E\}$$

are free variables in the sense that there is a 1-1 correspondence between $\psi_E$ and $K_E = (K_{ij}, (i,j) \in E)$ and that $\psi_{\overline{E}} = \{\psi_{ij} < (i,j) \in \overline{E}\}$ can be expressed in terms of $\psi_E$ and are thus non-free variables. Then, with a change of variables from $K_E$ to $\psi_E$, we have

$$\begin{aligned} |K|^{\frac{\delta-2}{2}} \exp\{-\frac{1}{2}\langle K, D\rangle\}\mathbf{1}_{P_G}(dK) = \qquad (5) \\ 2^{\frac{\delta+\nu_i}{2}}(2\pi)^{\nu_i/2}\Gamma(\frac{\delta+\nu_i}{2})\ e^{-\frac{1}{2}\sum_{(i,j)\in\overline{E}}\psi_{ij}^2}\Big(\prod_{(i,j)\in E}\frac{1}{\sqrt{2\pi}}e^{-\frac{\psi_{ij}^2}{2}}d\psi_{ij}\Big) \\ \prod_{i=1}^{p}\frac{1}{\Gamma((\delta+\nu_i)/2)}\Big(\frac{\psi_{ii}^2}{2}\Big)^{(\delta+\nu_i)/2)}e^{-\frac{\psi_{ii}^2}{2}}d(\psi_{ii}^2)\ . \end{aligned}$$

This expression immediately yields (3) with

$$f_E(\psi_E) = e^{-\frac{1}{2}\sum_{(i,j)\in\overline{E}}\psi_{ij}^2} \qquad (6)$$

where the expected value is taken with respect to a product of independent $N(0,1)$ for $\psi_{ij}, (i,j) \in E$ and $\chi^2_{(\delta+\nu_i)/2}$ distribution for $\psi_{ii}^2, i = 1, \ldots, p$. It remains to give the expression of the non-free variables in terms of the free ones. Equation (31) in Atay-Kayis and Massam (2005), in the particular case where $D = \mathbb{I}_p$, yields

$$\psi_{ij} = -\frac{\sum_{r=1}^{i-1}\psi_{ri}\psi_{rj}}{2}, \quad i < j,\ (i,j) \in \overline{E}\ . \qquad (7)$$

This equation shows that each non-free variables $\psi_{ij}$ is equal to $-1/2$ multiplied by the sum, over the rows which are numbered 1 to $i-1$ of the products of the entries in the column $i$ and the column $j$.

It is well-known (see formula (22) in Atay-Kayis and Massam (2005)) that, if $P_j, j = 1, \ldots, k$ is a perfect sequence of prime components of $G$ with corresponding separators $S_j, j = 2, \ldots, k$ and induced graphs $G_{P_j}$, then $I_G(\delta, D)$ can be decomposed as

$$I_G(\delta, D) = \frac{\prod_{j=1}^{k} I_{G_{P_j}}(\delta, D_{P_j})}{\prod_{j=2}^{k} I_{G_{S_j}}(\delta, D_{S_j})}, \qquad (8)$$

where $D_{P_j}$ and $D_{S_j}$ are the submatrices of $D$ indexed by the vertices of $P_j, j = 1, \ldots, k$ and $S_j, j = 2, \ldots, k$ respectively. In a second step, we show that the ratio in (2) can be reduced to the ratio

$$\frac{I_{G_{pq}^{-e}}(\delta, D_{G_{pq}^{-e}})}{I_{G_{pq}}(\delta, D_{G_{pq}})}.$$

where $G_{pq}^{-e}$ is the graph induced by $q, p$ and all the vertices contained in the minimal path between $q$ and $p$ and not containing the edge $(q, p)$ while $G_{pq}$ is the graph obtained from $G_{pq}^{-e}$ by adding the edge $e = (q, p)$. We say that a path between $q$ and $p$ is minimal if this path has no chord. Indeed, if a vertex $v$ is not on a path linking $q$ and $p$, it is either not linked to either $q$ or $p$ or it is linked to one of them but not both. In the first case, the graph $G$ is disconnected and we clearly only have to consider the connected part of $G$ containing $q$ and $p$. In the second case, assuming, without loss of generality, that $v$ is not linked to $p$, means that there is a separator between the prime components $P_{v,i}, i = 1, \dots k_v$ containing $v$ and the prime components $P_{p,i}, i = 1, \dots, k_p$ containing $p$. Since in both $G$ and $G^{-e}$, the prime component $P_{v,i}$ are the same, their corresponding terms will cancel out in (8) above and so will the terms for the corresponding separators. Finally if $v$ belongs to a path $\pi$ between $q$ and $p$ but not on a minimal path, then $\pi = \{q, i_1, \dots, i_v, \dots, j_v, \dots, p\}$ such that $(i_v, j_v) \neq (q, p)$ is a chord. Then the graph induced by the path $\{i_v, \dots, j_v\}$ and the chord $(i_v, j_v)$ is a prime component which will appear in both the factorization of $I_G(\delta, D)$ and $I_{G^{-e}}(\delta, D)$. This proves that in our study of (2), it is sufficient to consider a graph $G_{pq}^{-e}$ which contains only the minimal paths between $q$ and $p$. To keep notation simple, from now in this paper, we will write $G$ and $G^{-e}$ for $G_{pq}$ and $G_{pq}^{-e}$ respectively.

In a third step, we will set up the numbering of the vertices in a convenient way. From now on, we assume that $G^{-e}$ is the graph induced by the minimal paths between $q$ and $p$. Let $\Lambda$ be the set of these minimal paths. A path $\lambda \in \Lambda$ between $q = p - 1$ and $p$ with $\ell_\lambda$ vertices between $q$ and $p$ will be written

$$\lambda = \{q, 1_\lambda, 2_\lambda, \dots, \ell_\lambda, p\}.$$

We let $E_\lambda$, $V_\lambda$ and $V_\lambda^{(-1)}$be, respectively, the set of edges, the set of interior vertices of $\lambda$ and the set of interior points deprived of $1_\lambda$, i.e.

$$E_\lambda = \{(1_\lambda, q), (1_\lambda, 2_\lambda), \dots, ((l-1)_\lambda, \ell_\lambda), (\ell_\lambda, p)\}, \;\; V_\lambda = \{1_\lambda, 2_\lambda, \dots, \ell_\lambda\}, \;\; V_\lambda^{(-1)} = V_\lambda \backslash \{1_\lambda\} \; .$$

If $L = |\Lambda|$ is the total number of minimum paths, we set an arbitrary order $\lambda_1, \dots, \lambda_L$ of the path where, for convenience, we list the paths of length 2, i.e. $\ell_\lambda = 1$ last. Within each path $\lambda$, we order the vertices in $V_\lambda$ following the path. The vertices $q$ and $p$ are ranked last so that the order of the vertices is

$$1_{\lambda_1}, \dots, \ell_{\lambda_1}, 1_{\lambda_2}, \dots, \ell_{\lambda_2}, \dots, \dots, 1_{\lambda_L}, \dots, \ell_{\lambda_L}, q, p \; . \tag{9}$$

We are now ready to state the following.

**Proposition 2.1.** *let $G$, $G^{-e}$, $\Lambda, E_\lambda, V_\lambda$ and $V_\lambda^{(-1)}$ be as defined above. Assume that the paths between $q$ and $p$ are disjoint, that is, have no edge in common. Let the order*

*of the vertices in $V$ be as given in* (9). *The ratio of prior normalizing constants in* (2) *is equal to*

$$\frac{I_{G^{-e}}(\delta, \mathbb{I}_p)}{I_G(\delta, \mathbb{I}_p)} = \frac{1}{2\sqrt{\pi}} \frac{\Gamma(\delta/2)}{\Gamma((\delta+1)/2)} \frac{E\Big(\exp -\frac{1}{2}(\sum_{(i,j)\in \bar{E}} \psi_{ij}^2 + \psi_e^2)\Big)}{E(\exp -\frac{1}{2}\sum_{(i,j)\in \bar{E}} \psi_{ij}^2)}, \tag{10}$$

*where*

$$\psi_e = \sum_{\lambda \in \Lambda} (-1)^{\ell_\lambda} \frac{\prod_{a \in E_\lambda} \psi_a}{\psi_{qq} \prod_{v \in V_\lambda^{(-1)}} \psi_{vv}} \ . \tag{11}$$

*Proof.* Since $\nu_i$ is equal to the cardinality of $nb(i) \cap \{i+1, \ldots, p\}, i = 1, \ldots, p$, since the only change between $E$ and $E^{-e}$ is the edge $e$ and since the neighbours of $i_\lambda, i = 1, \ldots, \ell_\lambda, \lambda \in \Lambda$ are the same in $G$ and $G^{-e}$, given the order (9), only $\nu_q$ is different in $G$ and $G^{-e}$. Indeed $\nu_q^G = 1$ while $\nu_q^{G^{-e}} = 0$. This yields immediately equation (10). Let us now prove (11). From (7) and given (9), we have

$$\psi_e = \psi_{qp} = -\frac{\sum_{\lambda \in \Lambda} \sum_{i_\lambda \in V_\Lambda} \psi_{i_\lambda q} \psi_{i_\lambda p}}{\psi_{qq}}. \tag{12}$$

We want to compute the entries $\psi_{i_\lambda q},\ \psi_{i_\lambda p}$ in terms of

$$\psi_E = \Big(\psi_{i_\lambda i_\lambda},\ i_\lambda \in V_\lambda, \quad \psi_{i_\lambda (i+1)_\lambda},\ i_\lambda \in V_\lambda \setminus \{\ell_\lambda\}, \quad \psi_{1_\lambda, q}, \psi_{\ell_\lambda, p}, \quad \lambda \in \Lambda\Big).$$

We now note three important facts. First, the elements of the first row of the matrix $\psi$ are all zero except for those corresponding to the edges of the path $\lambda_1$, i.e.

$$\psi_{1_{\lambda_1}, v} = 0,\ v \in \cup_{\lambda \in \Lambda} V_\lambda, \ \ v \neq 1_{\lambda_1}, 2_{\lambda_1}, q. \tag{13}$$

Second, as a consequence of (13) and (7), the remaining non-free entries in all the columns of $\psi$ except for the columns $q$ and $p$, are equal to zero, i.e., for $\lambda, \lambda' \in \Lambda$ and for $\lambda$ ranked before $\lambda'$,

$$\psi_{i_\lambda, j_{\lambda'}} = 0,\ 1_\lambda < i_\lambda < j_{\lambda'},\ , j_{\lambda'} \neq q, p \ . \tag{14}$$

Third, due to the first entry $\psi_{1_{\lambda_1}, q}$ of column $q$ being free, none of the entries of column $q$ are necessarily zero. However, for each $\lambda \in \Lambda$, using iteratively (7), we see that the entries of column $p$ are zero except for the last one $\psi_{l_\lambda, p}$ which is a free variable,i.e.

$$\psi_{j_\lambda, p} = 0, \ \ j_\lambda < l_\lambda \ , \lambda \in \Lambda \ . \tag{15}$$

Applying (15) and (14), equation (12) yields

$$\psi_e = \psi_{qp}^* = -\frac{\sum_{\lambda \in \Lambda} \psi_{l_\lambda q} \psi_{l_\lambda p}}{\psi_{qq}}. \tag{16}$$

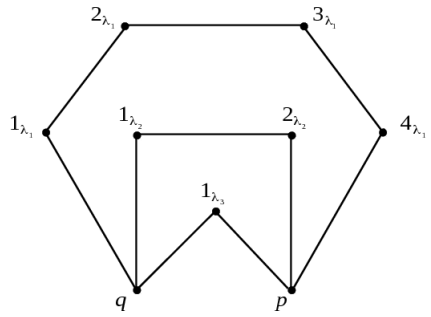


Figure 1: Graph $G_{qp}^{-e}$ for Example 1

The entries $\psi_{l_\lambda p},\ \lambda \in \Lambda$ are free. The entries $\psi_{l_\lambda q}$ are obtained by successively applying (7), (14) and the fact that $\psi_{(j-1)_\lambda, j_\lambda},\ j = 1, \ldots, (l-1)$ are free. That is

$$\begin{aligned}
\psi_{l_\lambda q} &= -\frac{\psi_{(l-1)_\lambda, l_\lambda}\psi_{(l-1)_\lambda, q}}{\psi_{(l-1)_\lambda,(l-1)_\lambda}} = +\frac{\psi_{(l-1)_\lambda, l_\lambda}\psi_{(l-2)_\lambda, l_\lambda}\psi_{(l-2)_\lambda, q}}{\psi_{(l-1)_\lambda,(l-1)_\lambda}\psi_{(l-2)_\lambda,(l-2)_\lambda}} \\
&= \ldots = (-1)^{l_\lambda - 1}\frac{\psi_{1_\lambda, q}\prod_{j=1}^{l-1}\psi_{j_\lambda,(j+1)_\lambda}}{\prod_{j=2}^{l}\psi_{j_\lambda, j_\lambda}} = (-1)^{l_\lambda - 1}\frac{\psi_{1_\lambda, q}\prod_{j=1}^{l-1}\psi_{j_\lambda,(j+1)_\lambda}}{\prod_{j=1}^{l-1}\psi_{(j+1)_\lambda,(j+1)_\lambda}}\ , \quad (17) \\
&= (-1)^{l_\lambda - 1}\psi_{1_\lambda, q}\prod_{j=1}^{l-1}\frac{\psi_{j_\lambda,(j+1)_\lambda}}{\psi_{(j+1)_\lambda,(j+1)_\lambda}}\ . \quad (18)
\end{aligned}$$

Equalities (16) and (17) together yield

$$\psi_{qp} = \sum_{\lambda\in\Lambda}(-1)^{l_\lambda}\frac{\psi_{1_\lambda, q}\psi_{l_\lambda, p}\prod_{j=1}^{l-1}\psi_{j_\lambda,(j+1)_\lambda}}{\psi_{qq}\prod_{j=2}^{l}\psi_{j_\lambda, j_\lambda}} = \sum_{\lambda\in\Lambda}(-1)^{l_\lambda}\psi_{l_\lambda, p}\frac{\psi_{1_\lambda, q}}{\psi_{qq}}\prod_{j=1}^{l-1}\frac{\psi_{j_\lambda,(j+1)_\lambda}}{\psi_{(j+1)_\lambda,(j+1)_\lambda}}\ ,$$

which is identical to (11). □

We immediately illustrate the ranking of the vertices, our notation and the expressions in (17) and (11) with an example.

**Example 2.1.** *Consider the graph of Figure 1 where for simplicity, we write $i_r$ for $i_{\lambda_r}, r = 1, \ldots, L,\ i = 1, \ldots, \ell_{\lambda_r}$ The matrix $\psi$ is as follows.*

| $1_1$ | $2_1$ | $3_1$ | $4_1$ | $1_2$ | $2_2$ | $1_3$ | $q$ | $p$ |
|---|---|---|---|---|---|---|---|---|
| $\psi_{1_11_1}$ | $\psi_{1_12_1}$ | 0 | 0 | 0 | 0 | 0 | $\psi_{1_1q}$ | 0 |
| | $\psi_{2_12_1}$ | $\psi_{2_13_1}$ | 0 | 0 | 0 | 0 | $*$ | 0 |
| | | $\psi_{3_13_1}$ | $\psi_{3_14_1}$ | 0 | 0 | 0 | $*$ | 0 |
| | | | $\psi_{4_14_1}$ | 0 | 0 | 0 | $*$ | $\psi_{4_1p}$ |
| | | | | $\psi_{1_21_2}$ | $\psi_{1_22_2}$ | 0 | $\psi_{1_2q}$ | 0 |
| | | | | | $\psi_{2_22_2}$ | 0 | $*$ | $\psi_{2_2p}$ |
| | | | | | | $\psi_{1_31_3}$ | $\psi_{1_3q}$ | $\psi_{1_3p}$ |
| | | | | | | | $\psi_{qq}$ | $*$ |
| | | | | | | | | $\psi_{pp}$ |

*where the entries marked with a $*$ are the non-free entries and are given as follows:*

$$\psi_{2_1q} = -\frac{\psi_{1_12_1}\psi_{1_1q}}{\psi_{2_12_1}}, \quad \psi_{3_1q} = \frac{\psi_{2_13_1}\psi_{1_12_1}\psi_{1_1q}}{\psi_{2_12_1}\psi_{3_13_1}}, \quad \psi_{4_1q} = -\frac{\psi_{3_14_1}\psi_{2_13_1}\psi_{1_12_1}\psi_{1_1q}}{\psi_{2_12_1}\psi_{3_13_1}\psi_{4_14_1}}$$

$$\psi_{2_2q} = -\frac{\psi_{1_22_2}\psi_{1_2q}}{\psi_{2_22_2}}$$

$$\begin{aligned}\psi_{qp} &= -\frac{1}{\psi_{qq}}\Big(\psi_{1_3q}\psi_{1_3q} + \psi_{2_2q}\psi_{2_2p} + \psi_{4_1q}\psi_{4_1p}\Big) \\ &= -\frac{\psi_{1_3q}\psi_{1_3q}}{\psi_{qq}} + \frac{\psi_{1_22_2}\psi_{1_2q}\psi_{2_2p}}{\psi_{qq}\psi_{2_22_2}} - \frac{\psi_{4_1p}\psi_{3_14_1}\psi_{2_13_1}\psi_{1_12_1}\psi_{1_1q}}{\psi_{qq}\psi_{2_12_1}\psi_{3_13_1}\psi_{4_14_1}}.\end{aligned}$$

The paths of length 2, that is, with $l_\lambda = 1$ will play a special role in the approximation of the ratio (4) (and thus (2)) that we seek to approximate. We thus write (11) as

$$\psi_e = \psi_{qp} = -A + b \tag{19}$$

where

$$A = \sum_{\lambda\in\Lambda, l_\lambda=1} \frac{\psi_{l_\lambda,q}, \psi_{\ell_\lambda,p}}{\psi_{qq}}, \quad b = \frac{1}{\psi_{qq}} \sum_{\lambda\in\Lambda,\ell_\lambda\geq 2} (-1)^{\ell_\lambda} \frac{\psi_{1_\lambda,q}\psi_{\ell_\lambda,p}\prod_{j=1}^{\ell_\lambda-1}\psi_{j_\lambda,(j+1)_\lambda}}{\prod_{j=1}^{\ell_\lambda-1}\psi_{(j+1)_\lambda,(j+1)_\lambda}}. \tag{20}$$

# 3 An approximation to the ratio of normalizing constants

## 3.1 The results

Following (5) and (6), in order to approximate the ratio (10), we need to approximate

$$\frac{\mathbb{E}\Big(\exp -\frac{1}{2}(\sum_{(i,j)\in\bar{E}}\psi_{ij}^2 + \psi_e^2)\Big)}{\mathbb{E}(\exp -\frac{1}{2}\sum_{(i,j)\in\bar{E}}\psi_{ij}^2)} = \frac{\mathbb{E}\Big(e^{-\frac{1}{2}\sum_{(i,j)\in\bar{E}}\psi_{ij}^2} e^{-\frac{A^2}{2}} e^{-Ab-\frac{b^2}{2}}\Big)}{\mathbb{E}(e^{-\frac{1}{2}\sum_{(i,j)\in\bar{E}}\psi_{ij}^2})}. \tag{21}$$

From (5), we see that $A$ is the sum of ratios of product of independent normal $N(0,1)$ by the square root of a chi-square distribution with $\delta$ degrees of freedom and similarly $b$ is the sum of ratios of products of standard normals by a product of square roots of chi-square distributions. To obtain an approximation to (21), we proceed in two steps. We will first show that $\mathbb{E}\Big(e^{-\frac{1}{2}\sum_{(i,j)\in\bar{E}}\psi_{ij}^2}e^{-\frac{A^2}{2}}\Big)$ is a good approximation to $\mathbb{E}\Big(e^{-\frac{1}{2}\sum_{(i,j)\in\bar{E}}\psi_{ij}^2}e^{-\frac{A^2}{2}}e^{-Ab-\frac{b^2}{2}}\Big)$. We will then see that $A$ is independent of $\sum_{(i,j)\in\bar{E}}\psi_{ij}^2$and thus

$$\frac{\mathbb{E}\Big(e^{-\frac{1}{2}\sum_{(i,j)\in\bar{E}}\psi_{ij}^2}e^{-\frac{A^2}{2}}e^{-Ab-\frac{b^2}{2}}\Big)}{\mathbb{E}\big(e^{-\frac{1}{2}\sum_{(i,j)\in\bar{E}}\psi_{ij}^2}\big)}\approx\frac{\mathbb{E}\Big(e^{-\frac{1}{2}\sum_{(i,j)\in\bar{E}}\psi_{ij}^2}e^{-\frac{A^2}{2}}\Big)}{\mathbb{E}\big(e^{-\frac{1}{2}\sum_{(i,j)\in\bar{E}}\psi_{ij}^2}\big)}=\mathbb{E}(e^{-\frac{A^2}{2}}),$$

which can easily be obtained explicitly. The following theorem is the main result of this paper and yields an accurate approximation to ratio (2).

**Theorem 3.1.** *For $G, G^{-e}$ as defined above and for $A$ and $b$ as defined in* (19) *and* (20)*, we have*

$$0\le 1-\frac{\mathbb{E}\Big(\exp-\frac{1}{2}(\sum_{(i,j)\in\bar{E}}\psi_{ij}^2+\psi_e^2)\Big)}{\mathbb{E}(\exp-\frac{1}{2}(\sum_{(i,j)\in\bar{E}}\psi_{ij}^2+A^2))}\le B, \tag{22}$$

*where*

$$B=\frac{2}{\pi r(\delta)}\frac{\delta}{\delta+2}\Big(\sum_{\lambda\in\Lambda_{2+}}R_\delta^{\ell_\lambda}\Big)R_{\delta+d-1},$$

*where $d$ is the number of paths $\lambda\in\Lambda$ of length 2, i.e. with $\ell_\lambda=1$, and*

$$R_\delta=\frac{\Gamma(\frac{\delta}{2})}{\sqrt{\pi}\Gamma(\frac{\delta+1}{2})},\quad r(\delta)=\frac{\Gamma(\frac{1}{2}(\delta+1))^2}{\Gamma(\frac{1}{2}\delta)\Gamma(\frac{1}{2}(\delta+2))}.$$

*Moreover, $A$ is independent of $\sum_{(i,j)\in\bar{E}}\psi_{ij}^2$ with*

$$\mathbb{E}(e^{-\frac{A^2}{2}})=\frac{\Gamma((\delta+d)/2)\Gamma((\delta+1)/2)}{\Gamma(\delta/2)\Gamma((\delta+d+1)/2)} \tag{23}$$

*and, with an accuracy given by* (22)*,*

$$\frac{I_{G^{-e}}(\delta,D)}{I_G(\delta,D)}\approx\frac{1}{2\sqrt{\pi}}\frac{\Gamma(\frac{\delta+d}{2})}{\Gamma(\frac{\delta+d+1}{2})}\ . \tag{24}$$

The approximation (24) gives an analytically explicit expression for ratio (2) and thus removes the need to do the simulations that were previously necessary to the evaluation of (4) in model search. This makes the search algorithm scale free. We

will illustrate this fact in Section 4. Before doing so and before giving the proof of Theorem 3.1 above, we give, in Table 1 the value of $B$ under different scenarios for graphs having five different paths between $q$ and $p$.

For $\delta = 3$,

$$B = \frac{8}{3\pi} \frac{\sum_{\lambda \in \Lambda_{2+}} R_\delta^{\ell_\lambda}}{R_{\delta+d}}.$$

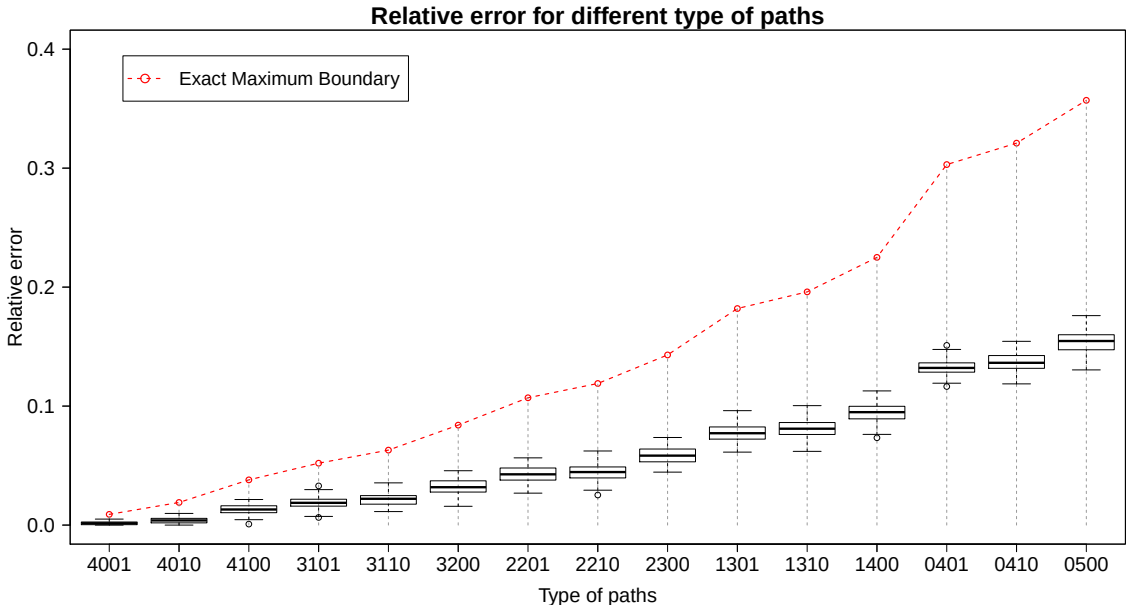


Figure 2: Boxplot and exact maximum boundary for the relative errors of different type of paths between node $q$ and $p$. Boxplot is based on the Monte Carlo estimation with 1000 iterations for the ratio of the prior normalizing constants, based on 100 replications. Here we have 15 different type of paths ways (in $x$ axis), in which e.g. "4001" means 4 paths of size 2, 0 path of size 3, 0 path of size 4, 1 path of size 5.

The computations in Table 1 and Figure 2 have been done with 100 replications and 1000 iterations for the Monte-Carlo algorithm.

## 3.2 The proof

We express here $A$ and $b$ in terms of $A_1, b_1, b_{1\lambda}, \lambda \in \Lambda$

$$A = -\sum_{\lambda \in \Lambda, l_\lambda = 1} \frac{\psi_{l_\lambda, q}, \psi_{l_\lambda, p}}{\psi_{qq}} = -\frac{1}{\psi_{qq}} \sum_{\lambda \in \Lambda, l_\lambda = 1} \psi_{l_\lambda, q}, \psi_{l_\lambda, p} = \frac{A_1}{\sqrt{Q_\delta}},$$

$$b = \frac{1}{\psi_{qq}} \sum_{\lambda \in \Lambda, l_\lambda \geq 2} (-1)^{l_\lambda} \psi_{1_\lambda, q} \psi_{l_\lambda, p} \prod_{j_\lambda = 1}^{l_\lambda - 1} \frac{\psi_{j_\lambda, (j+1)_\lambda}}{\psi_{(j+1)_\lambda, (j+1)_\lambda}} = \frac{b_1}{\sqrt{Q_\delta}} = \frac{\sum_{\lambda \in \Lambda, l_\lambda \geq 2} b_{1\lambda}}{\sqrt{Q_\delta}},$$

Table 1: Values of $B$ under various configurations for $L = 5$, $\delta = 3$. All the figures have been rounded to the 4-th decimal.

| $d = 4, \ell_5$ | $\ell_5 = 2$ | $\ell_5 = 3$ | $\ell_5 = 4$ |
|---|---|---|---|
| $R_{\delta+d-1}$ | 0.339 | 0.339 | 0.339 |
| $\sum_{\lambda\in\Lambda_{2+}} R_\delta^{\ell_\lambda}$ | 0.25 | 0.125 | 0.0625 |
| B | 0.038 | 0.019 | 0.009 |
| $1 - \frac{I_1}{I_2}$ | 0.0136 (0) | 0.0036 (0) | 0.0014 (0) |
| | | | |
| $d = 3, \ell_4, \ell_5$ | 11122 | 11123 | 11124 |
| $R_{\delta+d-1}$ | 0.375 | 0.375 | 0.375 |
| $\sum_{\lambda\in\Lambda_{2+}} R_\delta^{\ell_\lambda}$ | 0.5 | 0.375 | 0.312 |
| B | 0.084 | 0.063 | 0.052 |
| $1 - \frac{I_1}{I_2}$ | 0.0333 (0) | 0.0213 (0) | 0.0186 (0) |
| | | | |
| $d = 2, \ell_3, \ell_4, \ell_5$ | 11222 | 11223 | 11224 |
| $R_{\delta+d-1}$ | 0.424 | 0.424 | 0.424 |
| $\sum_{\lambda\in\Lambda_{2+}} R_\delta^{\ell_\lambda}$ | 0.75 | 0.625 | 0.562 |
| B | 0.143 | 0.119 | 0.107 |
| $1 - \frac{I_1}{I_2}$ | 0.0591 ($10^{-4}$) | 0.0452 (0) | 0.0411 (0) |
| | | | |
| $d = 1, \ell_2, \ell_3, \ell_4, \ell_5$ | 12222 | 12223 | 12224 |
| $R_{\delta+d-1}$ | 0.5 | 0.5 | 0.5 |
| $\sum_{\lambda\in\Lambda_{2+}} R_\delta^{\ell_\lambda}$ | 1 | 0.875 | 0.8125 |
| B | 0.225 | 0.196 | 0.182 |
| $1 - \frac{I_1}{I_2}$ | 0.0944 ($10^{-4}$) | 0.0797 (0) | 0.076 ($10^{-4}$) |
| | | | |
| $d = 0, \ell_1, \ell_2, \ell_3, \ell_4, \ell_5$ | 22222 | 22223 | 22224 |
| $R_{\delta+d-1}$ | 0.636 | 0.636 | 0.636 |
| $\sum_{\lambda\in\Lambda_{2+}} R_\delta^{\ell_\lambda}$ | 1.25 | 1.125 | 1.062 |
| B | 0.357 | 0.321 | 0.303 |
| $1 - \frac{I_1}{I_2}$ | 0.1541 ($10^{-4}$) | 0.1369 (0) | 0.1323 ($10^{-4}$) |

where

$$A_1 = \sum_{\lambda\in\Lambda, l_\lambda=1} \psi_{l_\lambda,q}, \psi_{l_\lambda,p}, \quad b_{1\lambda} = \sum_{\lambda\in\Lambda, l_\lambda\geq 2} (-1)^{l_\lambda} \psi_{1_\lambda,q}\psi_{l_\lambda,p} \prod_{j_\lambda=1}^{l_\lambda-1} \frac{\psi_{j_\lambda,(j+1)_\lambda}}{\psi_{(j+1)_\lambda,(j+1)_\lambda}}, \quad Q_\delta = \psi_{qq}^2.$$

All the entries of $\psi_E$ appearing above are independent and those appearing in $b_{1\lambda}, \lambda \in \Lambda, \ell_\lambda \geq 2$ are different from those appearing in $A_1$. Thus $A_1$, $\sum_{\lambda\in\Lambda, l_\lambda\geq 2} b_{1\lambda}$ and $Q_\delta$ are stochastically independent. For ease of notation, from now on, we will write $\Lambda_{2+}$ for the set of paths in $\Lambda$ that have $\ell_\lambda \geq 2$. We define $D$ and $D_\lambda$ as follows:

$$D = \sum_{(i,j)\in\bar{E}} \psi_{ij}^2 = \sum_{\lambda\in\Lambda_{2+}} \sum_{k=2}^{l_\lambda} \Big\{ (-1)^{k-1} \psi_{1_\lambda,q} \prod_{j_\lambda=1}^{k-1} \frac{\psi_{j_\lambda,(j+1)_\lambda}}{\psi_{(j+1)_\lambda,(j+1)_\lambda}} \Big\}^2 = \sum_{\lambda\in\Lambda_{2+}} D_\lambda$$

To prove (22), we thus have to find an upper bound for

$$|1 - \frac{I_1}{I_2}| = |1 - \frac{E(e^{-\frac{D}{2}} e^{-\frac{(A_1+b_1)^2}{2Q_\delta}})}{E(e^{-\frac{D}{2}} e^{-\frac{A_1^2}{2Q_\delta}})}| \, .$$

In the sequel, we will often use the Gaussian equality which expresses that if $Z \sim N(0,\sigma^2)$, then

$$\mathbb{E}(e^{itZ}) = \int_{-\infty}^{+\infty} e^{itz} e^{-\frac{z^2}{2\sigma^2}} \frac{dz}{\sigma\sqrt{2\pi}} = e^{-\frac{\sigma^2 t^2}{2}} \, . \tag{25}$$

Applying (25) with $t = A_1 + b_1$ and $\sigma^2 = \frac{1}{Q_\delta}$, we have

$$\begin{aligned}
\mathbb{E}(e^{-\frac{D}{2}} e^{-\frac{(A_1+b_1)^2}{2Q_\delta}}) &= \mathbb{E}(e^{-\frac{\sum_{\lambda\in\Lambda_{2+}} D_\lambda}{2}} e^{-\frac{(A_1+\sum_{\lambda\in\Lambda_{2+}} b_{1\lambda})^2}{2Q_\delta}}) \\
&= \mathbb{E}\Big(e^{-\frac{\sum_{\lambda\in\Lambda} D_\lambda}{2}} \int_{-\infty}^{+\infty} e^{iA_1x + i\sum_{\lambda\in\Lambda_{2+}} b_{1\lambda}} e^{-\frac{Q_\delta x^2}{2}} \sqrt{Q_\delta}) \frac{dx}{\sqrt{2\pi}} \\
&= \int_{-\infty}^{+\infty} \mathbb{E}(e^{iA_1x}) \prod_{\lambda\in\Lambda_{2+}} \mathbb{E}(e^{-\frac{D_\lambda}{2} + ib_{1\lambda}x}) f(x) \frac{dx}{\sqrt{2\pi}}
\end{aligned}$$

where the last equality is due to the fact that $A_1$ is independent of $b_{1\lambda}, Q_\delta$ and $D$, and $f(x) = \mathbb{E}(e^{-\frac{Q_\delta x^2}{2}} \sqrt{Q_\delta})$. We note that $\frac{Q_\delta}{2} \sim \gamma_{\frac{\delta}{2}}$ where $\gamma_{\frac{\delta}{2}}$ denotes the Gamma distribution with parameters $(\frac{\delta}{2}, 1)$ and with density given by (36). Thus

$$\begin{aligned}
f(x) &= E(e^{-\frac{Q_\delta x^2}{2}} \sqrt{Q_\delta}) = \frac{1}{\Gamma(\delta/2)} \int_0^{+\infty} e^{-yx^2 - y} \sqrt{2y} y^{\delta/2-1} dy \\
&= \frac{\sqrt{2}\Gamma((\delta+1)/2)}{\Gamma(\delta/2)} (1+x^2)^{-\frac{\delta+1}{2}}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned} E(e^{-\frac{D}{2}}e^{-\frac{A_1^2}{2Q_\delta}}) &= \int_{-\infty}^{+\infty} E(e^{iA_1x}) \prod_{\lambda\in\Lambda_{2+}} E(e^{-\frac{D_\lambda}{2}}) f(x)dx \\ &= \prod_{\lambda\in\Lambda_{2+}} E(e^{-\frac{D_\lambda}{2}}) \int_{-\infty}^{+\infty} E(e^{iA_1x}) f(x)dx \ . \end{aligned}$$

Thus

$$|1-\frac{I_1}{I_2}| = |\frac{I_2-I_1}{I_2}| = \frac{\int_{-\infty}^{+\infty} E(e^{iA_1x})\Big\{E\Big(e^{-\sum_{\lambda\in\Lambda}\frac{D_\lambda}{2}}(1-\prod_{\lambda\in\Lambda_{2+}} e^{+ib_{1\lambda}x})\Big)\Big\} f(x)dx}{\prod_{\lambda\in\Lambda_{2+}} E(e^{-\frac{D_\lambda}{2}}) \int_{-\infty}^{+\infty} E(e^{iA_1x}) f(x)dx}$$

Consider independent identically distributed random variables $X_1, \ldots, X_n, \ldots$ such that $X_1 \sim Z/\sqrt{Q}$ with $Z \sim N(0,1)$ independent of $Q$ which is chi-square distributed with $\delta+1 \geq 4$ degrees of freedom. For $\ell = \ell_\lambda, \lambda \in \Lambda_{2+}$, we define

$$S_\ell = X_1^2 + X_1^2X_2^2 + \cdots + (X_1 \ldots X_{\ell-1})^2, \quad B_\ell = X_1X_2 \ldots X_{\ell-1}.$$

We see that for each $\lambda \in \Lambda_{2+}$,

$$\begin{aligned} D_\lambda &\sim N_{1_\lambda q} S_{\ell_\lambda} \\ b_{1\lambda} &\sim N_{1_\lambda q} N_{\ell_\lambda p} B_{\ell_\lambda}. \end{aligned}$$

where $N_{1_\lambda q}, N_{\ell_\lambda p}$ are independent $N(0,1)$ random variables, independent of $X_1, \ldots, X_\ell, \ldots$. We note that, from the independence of the entries of $\psi_E$, we have that

$$(b_{1\lambda}, D_\lambda, N_{1_\lambda q}, N_{\ell_\lambda p}), \quad \lambda \in \Lambda_{2+}$$

are mutually independent.

Omitting the index $\lambda$ on $\ell_\lambda$, and simplifying $N_{1_\lambda q}$ to $N_q$ and $N_{\ell_\lambda p}$ to $N_p$, we define

$$g_\ell(x) = \mathbb{E}\left(e^{-\frac{N_q^2 S_\ell}{2} + iN_pN_qB_\ell x}\right). \tag{26}$$

Then

$$\begin{aligned} I_1 &= \int_{-\infty}^{\infty} E(e^{iA_1x}) \prod_{\lambda\in\Lambda_{2+}} g_{\ell_\lambda}(x) f(x)dx \\ I_2 &= \int_{-\infty}^{\infty} E(e^{iA_1x}) \prod_{\lambda\in\Lambda_{2+}} g_{\ell_\lambda}(0) f(x)dx \end{aligned}$$

and

$$|\frac{I_2-I_1}{I_2}| \;=\; \frac{\int_{-\infty}^{\infty}\mathbb{E}(e^{iA_1x})\Big(\prod_{\lambda\in\Lambda_{2+}} g_{\ell_\lambda}(0)-\prod_{\lambda\in\Lambda_{2+}} g_{\ell_\lambda}(x)\Big)f(x)dx}{\Big(\prod_{\lambda\in\Lambda_{2+}} g_{\ell_\lambda}(0)\Big)\int_{-\infty}^{\infty}\mathbb{E}(e^{iA_1x})f(x)dx}\,. \qquad (27)$$

We note that the quantity $\mathbb{E}(e^{iA_1x})$ has been computed in (43) and is positive. We also note that by (25),

$$\begin{aligned}\mathbb{E}\left(e^{-\frac{N_q^2S_\ell}{2}+iN_pN_qB_\ell x}\right) &= \mathbb{E}\Big\{\mathbb{E}\left(e^{-\frac{N_q^2S_\ell}{2}+iN_pN_qB_\ell x}\mid N_q,S_\ell,B_\ell\right)\Big\}\\ &= \mathbb{E}\Big\{e^{-\frac{N_q^2S_\ell}{2}}\mathbb{E}\left(e^{iN_pN_qB_\ell x}\mid N_q,S_\ell,B_\ell\right)\Big\}=\mathbb{E}\Big(e^{-\frac{N_q^2S_\ell}{2}-\frac{(N_qB_\ell x)^2}{2}}\Big),\end{aligned}$$

which shows that $0\le g_\ell(x)\le g_\ell(0)$ and that $0\le\frac{I_2-I_1}{I_2}$. By (39) applied to $a_\lambda = g_{\ell_\lambda}(0)$, $b_\lambda = g_{\ell_\lambda}(x)$ and $R = E(e^{-\frac{N_q^2X_1^2}{2}})$, we have

$$\prod_{\lambda\in\Lambda_{2+}} g_{\ell_\lambda}(0)-\prod_{\lambda\in\Lambda_{2+}} g_{\ell_\lambda}(x)\le\Big(\prod_{\lambda\in\Lambda_{2+}} g_{\ell_\lambda}(0)\Big)\sum_{\lambda\in\Lambda_{2+}}\frac{(g_{\ell_\lambda}(0)-g_{\ell_\lambda}(x))}{g_{\ell_\lambda}(0)}.$$

Writing $\ell$ for $\ell_\lambda$, we have that

$$\begin{aligned} g_\ell(0)-g_\ell(x) &= E\Big(e^{-\frac{N_q^2S_\ell}{2}}(1-e^{iN_pN_qB_\ell x})\Big)\\ &\le E\Big(e^{-\frac{N_q^2S_\ell}{2}}|N_pN_qX_1\dots X_{\ell-1}|\Big)|x|\\ &= E\Big(e^{-\frac{N_q^2S_\ell}{2}}|N_qX_1|\Big)E\Big(|N_pX_2\dots X_{\ell-1}|\Big)|x|\\ &= \frac{2}{\pi}\frac{\delta}{\delta+2}R_\delta^\ell\;|x|,\end{aligned}$$

where the inequality is due to the fact that $|1-e^{iN_pN_qB_\ell x}|\le|N_pN_qB_\ell|\;|x|$ and the last equality is obtained using (40) and (41). Moreover, by (42), we have that $r(\delta)\le g_\ell(0)$. Thus

$$\frac{g_\ell(0)-g_\ell(x)}{g_\ell(0)} \;\le\; \frac{2}{\pi r(\delta)}\frac{\delta}{\delta+2}R_\delta^\ell$$

Thus equation (27) yields

$$\begin{aligned} 0\le\frac{I_2-I_1}{I_2} &\le \frac{\frac{2}{\pi r(\delta)}\frac{\delta}{\delta+2}\Big(\sum_{\lambda\in\Lambda_{2+}}R_\delta^{\ell_\lambda}\Big)\int_{-\infty}^{\infty}E(e^{iA_1x})|x|f(x)dx}{|\int_{-\infty}^{\infty}E(e^{iA_1x})f(x)dx|}\\ &\le \frac{2}{\pi r(\delta)}\frac{\delta}{\delta+2}\Big(\sum_{\lambda\in\Lambda_{2+}}R_\delta^{\ell_\lambda}\Big)\frac{\int_{-\infty}^{\infty}E(e^{iA_1x})|x|f(x)dx}{|\int_{-\infty}^{\infty}E(e^{iA_1x})f(x)dx|}\\ &= \frac{2}{\pi r(\delta)}\frac{\delta}{\delta+2}\Big(\sum_{\lambda\in\Lambda_{2+}}R_\delta^{\ell_\lambda}\Big)R_{\delta+d-1}\end{aligned}$$

where the last equality is due to (44).

# 4 Model selection using the Birth and Death algorithm of Mohammadi and Wit (2015) and approximation (24)

We will now recall the Birth and Death Markov chain Monte Carlo algorithm developed by Mohammadi and Wit (2015) and show at which point of this algorithm we use approximation (24). This approximation, as we shall see in the next two sections, makes model selection faster and allows us to do model selection for high-dimensional data. Let $X^{(1:n)} = (X^{(1)}, X^{(2)}, ..., X^{(n)})$ denote the matrix of observed data where $X^{(i)}, i = 1, \ldots, n$ are independent random vector from the $N_p(0, \Sigma)$ distribution with $K = \Sigma^{-1} \in P_G$. The likelihood function is then

$$Pr(K|X^{(1:n)}, G) \propto |K|^{n/2} \exp\left\{-\frac{1}{2}\text{tr}(KS)\right\},$$

where $S = X^{(1:n)}X^{(1:n)^\top}$. As outlined in the introduction, we choose, as the conjugate prior for $K \in P_G$, the $G$-Wishart distribution with density (1). If we put the uniform prior on the space of graphs with $p$ vertices, the joint posterior distribution of $(K, G)$ is then

$$f(K, G \mid X^{(1:n)}) \quad = \quad \frac{|K|^{\frac{\delta+n-2}{2}}}{(2\pi)^{-np/2} I_G(\delta, D)} \exp\{-\frac{1}{2}\langle K, (D + S)\rangle\}. \tag{28}$$

Various search algorithms using the joint distribution of $(K, G)$ have been proposed in the Bayesian literature: see for example the reversible jump MCMC (henceforth abbreviated RJMCMC) algorithm by Lenkoski and Dobra (2011), Dobra et al. (2011), the double reversible jump MCMC algorithm of Wang and Li (2012) or the double birth-death MCMC (henceforth abbreviated BDMCMC) algorithm by Mohammadi et al. (2017). The RJMCMC is based on an ergodic discrete-time Markov chain. It can efficiently explore the graph space only if the acceptance rate is high which is not always the case. in such cases that acceptance rate in the RJMCMC algorithm is small the algorithm poorly mixing in the graph space and as a result multi-model distributions are defectively approximated. As an alternative, Mohammadi and Wit (2015) proposed the BDMCMC which is based on a continuous-time Markov process. Starting from a graph $G = (V, E)$ and a precision matrix $K$, this sampling scheme explores the graph space through two types of moves: jumping to a larger dimension by adding an edge (birth) or to a lower dimension by deleting an edge (death).

• A new edge $e \in \bar{E}$ is added as an independent Poisson process with rate $B_e(G, K)$

and with overall birth rate $B(G,K) = \sum_{e \in \notin E} B_e(G,K)$. Therefore, the probability of a birth event is

$$Pr(\text{birth of edge } e) = \frac{B_e(G,K)}{B(G,K) + D(G,K)}, \qquad \text{for each } \ e \notin E, \tag{29}$$

where $D(G,K)$ is defined below. If the birth of an edge $e$ occurs the process moves to a new graph $G^{+e} = (V, E \bigcup \backslash e)$ with a new $K^{+e} \in \mathbb{P}_{G^{+e}}$.

• An edge $e \in E$ is deleted as an independent Poisson process with rate $D_e(G,K)$ and with overall death rate $D(G,K) = \sum_{e \in \notin E} D_e(G,K)$. Thus, the probability of a death event is

$$Pr(\text{death of edge } e) = \frac{D_e(G,K)}{B(G,K) + D(G,K)}, \qquad \text{for each } \ e \in E. \tag{30}$$

If the death of an edge $e$ occurs the process moves to a new graph $G^{-e} = (V, E \backslash e)$ with new $K^{-e} \in \mathbb{P}_{G^{-e}}$.

Mohammadi and Wit (2015, Theorem 3.1) show that the BDMCMC sampling scheme has the joint posterior distribution $Pr(G,K|X^{(1:n)})$ as stationary distribution if for all edges

$$D_e(G^{+e}, K^{+e}) Pr(G^{+e}, K^{+e}|\mathbf{X}) = B_e(G,K) Pr(G,K|\mathbf{X}), \tag{31}$$

(see Preston (1976)).

From the detailed balance condition (31), the birth and death rates can be chosen as follows

$$B_e(G,K) = \min\left\{ \frac{Pr(G^{+e}, K^{+e}|\mathbf{X})}{Pr(G,K|\mathbf{X})}, 1 \right\}, \quad \text{for each } \ e \notin E, \tag{32}$$

$$D_e(G,K) = \min\left\{ \frac{Pr(G^{-e}, K^{-e}|\mathbf{X})}{Pr(G,K|\mathbf{X})}, 1 \right\}, \quad \text{for each } \ e \in E. \tag{33}$$

Based on the above rates we determine our BDMCMC sampling algorithm as follows:

Computing the birth and death rates (32) and (33) in Step 1 constitutes the main computational difficulty of the BDMCMC algorithm given above. Mohammadi and Wit (2015), shows that (33) can be written as

$$D_e(G,K) = \min\left\{ \frac{I_G(\delta, D)}{I_{G^{-e}}(\delta, D)} H(K, D^*, e), 1 \right\}, \tag{34}$$

where $e = (i,j)$,

$$H(K, D^*, e) = (\frac{D^*_{jj}}{2\pi a_{11}})^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}\left[ \langle D^*(K^0 - K^1) \rangle - (D^*_{ii} - \frac{D^{*2}_{ij}}{D^*_{jj}}) a_{11} \right] \right\},$$

**Algorithm 1** BDMCMC algorithm for GGMs

Input: a graph $G = (V, E)$ and a precision matrix $K \in P_G$

Output: Samples from the joint posterior distribution (28) and the weight of graphs.

For $N$ iterations, do

**1. Sample from the graph space.**

For all possible edges

1.1. Calculate (32) and (33)

End for

1.2. Calculate the waiting time as the weight of graph $G$ that is

$$W(G, K) = 1/(B(G, K) + D(G, K))$$

1.4. Simulate the type of jump (birth or death) using (29) and (30)

**2. Sample from the precision matrix space.** Use the exact sampling algorithm proposed by Lenkoski (2013).

End for

---

$a_{11} = k^0_{11} - k^1_{11}$ and

$$K^0 = \begin{bmatrix} k_{ii} & 0 \\ 0 & K_{j,V\setminus j}(K_{V\setminus j,V\setminus j})^{-1}K_{V\setminus j,j} \end{bmatrix},$$

and $K^1 = K_{e,V\setminus e}(K_{V\setminus e,V\setminus e})^{-1}K_{V\setminus e,e}$.

The computation of $H(K, D^*, e)$ in (34) is easy. However computing the ratio of normalizing constant $\frac{I_G(\delta,D)}{I_{G^{-e}}(\delta,D)}$ has been the subject of much research. In particular, Wang and Li (2012) and Cheng and Lenkoski (2012) developed an approach which borrows ideas from the exchange algorithm (Murray et al., 2012) and the double Metropolis-Hastings (MH) algorithm (Liang, 2010) to compute this ratio. It has been observed that in a high-dimensional setting, for large $p$, the double MH sampler becomes computationally inefficient due to the curse of dimensionality.

In the numerical experiments described in the following two sections, for $D = \mathbb{I}_p$, we use Theorem 3.1 to obtain the following simplified expression of (34):

$$D_e(G, K) = \min\left\{2\sqrt{\pi}\frac{\Gamma((\delta + d + 1)/2)}{\Gamma((\delta + d)/2)}H(K, D^*, e), 1\right\}.$$

We will see below that using this approximation makes the BDMCMC computations much faster and allows for model selection in high-dimension. To illustrate the efficacy of this approximation, we will perform model selection using the BDMCMC both ways, that is, to compute the ratio of prior normalizing constant, we use the double MH algorithm and we use the approximation of Theorem 3.1.

# 5 Simulation study

In this section we compare the performance of the BDMCMC model selection algorithm when using double Metropolis-Hastings to compute $\frac{I_G(\delta,D)}{I_{G^{-e}}(\delta,D)}$ and when using the approximation of Theorem 3.1 for simulated data. We consider four different graph structures (see Figure 3):

1. **Scale-free:** A graph which has a power-law degree distribution generated by the Barabási-Albert algorithm (Albert and Barabási, 2002).

2. **Random_p:** A graph in which edges are randomly generated from independent Bernoulli distributions with mean equal to $p$.

3. **Random_2p:** The same as the Random_p graph for Bernoulli distributions with mean equal to $2p$.

4. **Cluster:** A graph in which the number of clusters is $max\{2, [p/20]\}$. Each cluster has the same structure as the **Random_p** graph.

For each graph, the corresponding precision matrix $K$ is generated from the $G$-Wishart $W_G(3, \mathbb{I}_p)$. For each graph, we consider various scenarios based on the number of nodes $p \in \{50, 100, 150\}$ and the sample size $n \in \{p, 2p\}$. We draw $n$ independent samples from the normal $N_p(0, K)$ distribution. All simulations are performed using **BDgraph** *R* package (Mohammadi and Wit, 2017b).

For each scenario, we evaluate the performance of the BDMCMC algorithm with both approximation approaches, Theorem 3.1 or double Metropolis-Hastings. For each scenario, we run the BDMCMC algorithm with both approximations $100,000$ times and $60,000$ iterations as burn-in.

Briefly, graph selection is as follows. Following Mohammadi and Wit (2015), based on Bayesian model averaging, we first estimate the posterior probabilities for all possible edges $e = (i, j)$ in the graph by using a Rao-Blackwellized estimate (Cappé et al., 2003, Section 2.5) as follows

$$Pr(e \in E|data) = \frac{\sum_{t=1}^{N} 1(e \in E^{(t)}) W(K^{(t)})}{\sum_{t=1}^{N} W(K^{(t)})}, \tag{35}$$

where $N$ is the number of BDMCMC iteration and $W(K^{(t)})$ are the weights of the graph $G^{(t)}$ with precision matrix $K^{(t)}$. Then, by using the median probabilities model of Barbieri and Berger (2004), the selected graph is a graph with edges for which the estimated posterior probabilities are greater than 0.5.

To evaluate the performance of graph structure learning, we report ROC curves, which depict the true positive rate (TPR) as a function of the false postive rate

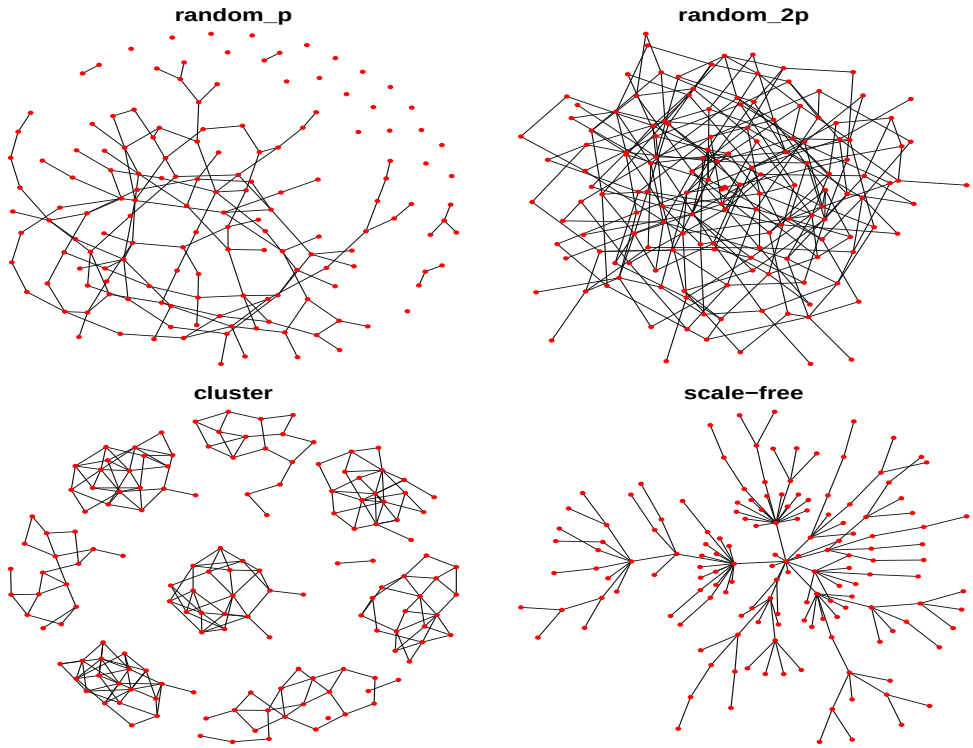


Figure 3: An illustration of the 6 different undirected graph structures for the cases with $p = 150$, as a number of nodes.

(FPR), the sensitivity, specificity and Matthews correlation coefficients (MCC) measurements, which are defined as follows

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives, respectively. Sensitivity and Specificity return a value between 0 and 1 and MCC returns a value between $-1$ and $+1$. For all three measures the larger the values are, the better the result is. For more details see Baldi et al. (2000).

The simulation results are summarized in Tables 2 and Figures 4, 5, 6, 7. As we can see, in almost all the cases the performance of the BDMCMC algorithm based on both approximations is the same. Only in some few cases, the BDMCMC algorithm with our approximation performs slightly better than the double Metropolis-Hastings

approximation: this happens especially when $p$ is increased, e.g. **Random_p**: $p = 150$ and $n = 150$ in Figure 6. It is mainly due to convergence problems of the double Metropolis-Hastings algorithm in high dimensions (see (Liang, 2010)).

In summary, our simulation study shows that the approximation of Theorem 3.1 performs well especially for high-dimensional sparse graphs, which is the case for many real world application such as in genetics.

Table 2: Graph structure learning performance of the BDMCMC algorithm with our approximation (BD) and double Metropolis-Hastings approximation (BD-DMH). The table presents the measures over 50 repetitions.

| | | | Specificity | | Sensitivity | | MCC | |
|---|---|---|---|---|---|---|---|---|
| p | n | graph | BD | BD-DMH | BD | BD-DMH | BD | BD-DMH |
| 50 | 50 | cluster | 0.6 | 0.61 | 0.9 | 0.9 | 0.38 | 0.39 |
| | | random_p | 0.62 | 0.61 | 0.9 | 0.91 | 0.33 | 0.34 |
| | | random_2p | 0.59 | 0.6 | 0.87 | 0.86 | 0.35 | 0.34 |
| | | scale-free | 0.66 | 0.65 | 0.9 | 0.9 | 0.34 | 0.34 |
| | 100 | cluster | 0.69 | 0.67 | 0.95 | 0.96 | 0.56 | 0.58 |
| | | random_p | 0.7 | 0.7 | 0.94 | 0.95 | 0.47 | 0.5 |
| | | random_2p | 0.65 | 0.64 | 0.93 | 0.94 | 0.51 | 0.52 |
| | | scale-free | 0.71 | 0.7 | 0.94 | 0.95 | 0.47 | 0.51 |
| 100 | 100 | cluster | 0.68 | 0.67 | 0.93 | 0.93 | 0.4 | 0.4 |
| | | random_p | 0.72 | 0.7 | 0.94 | 0.94 | 0.37 | 0.38 |
| | | random_2p | 0.63 | 0.63 | 0.9 | 0.89 | 0.33 | 0.32 |
| | | scale-free | 0.69 | 0.68 | 0.92 | 0.92 | 0.32 | 0.32 |
| | 200 | cluster | 0.75 | 0.74 | 0.96 | 0.97 | 0.57 | 0.6 |
| | | random_p | 0.78 | 0.77 | 0.96 | 0.97 | 0.49 | 0.52 |
| | | random_2p | 0.71 | 0.7 | 0.95 | 0.96 | 0.5 | 0.52 |
| | | scale-free | 0.76 | 0.75 | 0.96 | 0.97 | 0.45 | 0.49 |
| 150 | 150 | cluster | 0.72 | 0.71 | 0.93 | 0.93 | 0.38 | 0.38 |
| | | random_p | 0.75 | 0.74 | 0.94 | 0.94 | 0.32 | 0.33 |
| | | random_2p | 0.67 | 0.66 | 0.92 | 0.91 | 0.32 | 0.31 |
| | | scale-free | 0.73 | 0.72 | 0.93 | 0.93 | 0.29 | 0.29 |
| | 300 | cluster | 0.79 | 0.78 | 0.96 | 0.97 | 0.52 | 0.55 |
| | | random_p | 0.82 | 0.82 | 0.97 | 0.97 | 0.47 | 0.51 |
| | | random_2p | 0.74 | 0.73 | 0.96 | 0.96 | 0.47 | 0.48 |
| | | scale-free | 0.77 | 0.76 | 0.96 | 0.96 | 0.38 | 0.39 |

# 6 Real world example: human gene expression data

We apply the BDMCMC algorithm using the approximation of Theorem 3.1 to the human gene expression dataset which is available in the **BDgraph** *R* package. The dataset was collected from 60 individuals of Northern and Western European ancestry from Utah, whose genotypes are available at ftp://ftp.sanger.ac.uk/pub/genevar. For the individuals, the gene expression in Blymphocyte cells are measured by the Human-6 Expression BeadChips (Stranger et al., 2007). The raw data first were background corrected and quantile normalized across four replicates of a single individual, then, median normalized across all individuals. Finally, the 100 most variable probes among the $47,293$ probes have been chosen, corresponding to different transcript (Bhadra and Mallick, 2013, Mohammadi and Wit, 2015). The data has been studied in the literature (Stranger et al., 2007, Chen et al., 2008, Bhadra and Mallick, 2013, Gu et al., 2015, Mohammadi and Wit, 2015, 2017a), since it is an interesting case study for recovering underlying graph structures.

We run the BDMCMC algorithm with 100000 times and 60000 iterations as a burn-in. We present the results in Figures 8 and 10. The estimated posterior probabilities of existing links between each gene pair are presented under matrix form in Figure 8. These probabilities are calculated using (35) based on model averaging (Madigan et al., 1996). We present the result as a selected graph in Figure 10 where each edge selected has posterior probability greater than 0.8. Our estimated graph contains most of the interactions between genes found in Bhadra and Mallick (2013), Gu et al. (2015) and Mohammadi and Wit (2015, 2017a).

# Appendix 1

## General

We have the following two propositions. The standard gamma distribution with shape parameter $a > 0$ is denoted by

$$\gamma_a(dx) = \frac{1}{\Gamma(a)} e^{-x} x^{a-1} 1_{(0,\infty)}(x) dx. \tag{36}$$

**Proposition Appendix .1.** *Let $X, Y, Z$ be independent random variables with respective distributions $\gamma_a$, $\gamma_b, \gamma_c$. Then*

$$E\left(e^{-XY/Z}\right) = \frac{\Gamma(a+c)\Gamma(b+c)}{\Gamma(c)\Gamma(a+b+c)} \tag{37}$$

*Proof.*

$$\begin{aligned}
E\left(e^{-XY/Z}\right) &= E\left(E\left(e^{-XY/Z}\right)|X,Z\right) = E\left(\frac{1}{(1+\frac{X}{Z})^b}\right) \\
&= \frac{1}{\Gamma(a)}E\left(\int_0^\infty e^{-x}x^{a-1}(\frac{1}{(1+\frac{x}{Z})^b}dx\right) \\
&= \frac{1}{\Gamma(a)}E\left(Z^a\int_0^\infty e^{-Zu}\frac{u^{a-1}}{(1+u)^b}du\right) \\
&= \frac{1}{\Gamma(a)}\int_0^\infty \frac{u^{a-1}}{(1+u)^b}\left(E(Z^a e^{-Zu}\right)du \\
&= \frac{1}{\Gamma(a)\Gamma(c)}\int_0^\infty \frac{u^{a-1}}{(1+u)^b}\left(\int_0^\infty e^{-zu}z^{a+c-1}dz\right)du \\
&= \frac{\Gamma(a+c)}{\Gamma(a)\Gamma(c)}\int_0^\infty \frac{u^{a-1}}{(1+u)^{a+b+c}}du = \frac{\Gamma(a+c)\Gamma(b+c)}{\Gamma(c)\Gamma(a+b+c)}
\end{aligned}$$

□

**Proposition Appendix .2.** *Let $U_1,\ldots,U_k,V_1\ldots,V_k,Q$ be independent random variables such that $U_i$ and $V_j$ are $N(0,1)$ and such that $Q$ is chisquare distributed with degree of freedom $\delta$. We give an elementary proof of the following fact:*

$$E\left(e^{-\frac{1}{2Q}\left(\sum_{i=1}^k U_iV_i\right)^2}\right) = \frac{\Gamma((\delta+k)/2)\Gamma((\delta+1)/2)}{\Gamma(\delta/2)\Gamma((\delta+k+1)/2)}$$

*Proof.* This proposition is a direct consequence of the Proposition Appendix .1 above. We apply (37) to $Z = Q/2$ and $c = \delta/2$, to $Y = V_1^2/2$ and $b = 1/2$ and to

$$X = \frac{1}{2}(U_1^2+\cdots+U_k^2)$$

and $a = k/2$. The important remark is the fact that

$$\sum_{i=1}^k U_iV_i \sim (U_1^2+\cdots+U_k^2)^{1/2}V_1. \tag{38}$$

To see this we compute the Laplace transforms of both sides of (38). As a consequence $XY/Z \sim \frac{1}{2Q}\left(\sum_{i=1}^k U_iV_i\right)^2$. □

**Lemma Appendix .1.** *Let $a_1,\ldots,a_n$ and $b_1,\ldots,b_n$ be complex numbers such that $|b_i| \le |a_i|,\ i=1,\ldots,n$. Then*

$$|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i| \le \left(\prod_{i=1}^n |a_i|\right)\sum_{i=1}^n \frac{|a_i-b_i|}{|a_i|} \tag{39}$$

*Proof.*

$$
\begin{aligned}
|\prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i| &= |(\prod_{i=1}^{n} a_i - b_1 \prod_{i=2}^{n} a_i) + (b_1 \prod_{i=2}^{n} a_i - b_1 b_2 \prod_{i=3}^{n} a_i) + \ldots + (\prod_{i=1}^{n-1} b_i a_n - \prod_{i=1}^{n} b_i)| \\
&\leq \sum_{j=1}^{n} |\prod_{i=1}^{j-1} b_i \prod_{i=j}^{n} a_i - \prod_{i=1}^{j} b_i \prod_{i=j+1}^{n} a_i| = \sum_{j=1}^{n} |(\prod_{i=1}^{j-1} b_i \prod_{i=j+1}^{n} a_i)||a_j - b_j| \\
&\leq \sum_{j=1}^{n} |(\prod_{i=1}^{j-1} a_i \prod_{i=j+1}^{n} a_i)||a_j - b_j| = \prod_{i=1}^{n} |a_i||1 - \frac{b_j}{a_j}|
\end{aligned}
$$

□

**Lemma Appendix .2.** *Let $X_1, \ldots, X_{\ell-1}$ be independent identically distributed random variables such that $X_1 \sim Z/\sqrt{Q}$ with $Z \sim N(0,1)$ independent of $Q$ which is chi-square distributed with $\delta + 1 \geq 4$ degrees of freedom. Let $N_p$ and $N_q$ also be standard normal $N(0,1)$ random variables, mutually independent and independent of $X_1, \ldots, X_{\ell-1}$. Let $R_\delta = \frac{\Gamma(\frac{\delta}{2})}{\sqrt{\pi}\Gamma(\frac{\delta+1}{2})}$. We then have the following expected values:*

$$
E(e^{-\frac{N_q^2 X_1^2}{2}} |N_q X_1|) = \sqrt{\frac{2}{\pi}} \frac{\delta}{\delta + 2} R_\delta \tag{40}
$$

*and*

$$
E(|N_p X_1 \ldots X_{\ell-1}|) = \sqrt{\frac{2}{\pi}} R_\delta^{\ell-1} \, . \tag{41}
$$

*Proof.* Let us first prove (40). The variable $X_1^2$ follows a beta distribution of the second kind with parameter $\alpha = 1/2$ and $\beta = (\delta+1)/2$. The density of such a variable is $f(u) = \frac{1}{B(\alpha,\beta)} \frac{u^{\alpha-1}}{(1+u)^{\alpha+\beta}} \mathbf{1}_{(0,+\infty)}$. Thus

$$
\begin{aligned}
E(e^{-\frac{N_q^2 X_1^2}{2}} |N_q X_1|) &= E\{E(e^{-\frac{N_q^2 X_1^2}{2}} |N_q X_1| \;\; |X_1)\} \\
E(e^{-\frac{N_q^2 X_1^2}{2}} |N_q X_1| \;\; | \, X_1) &= |X_1| \int_{-\infty}^{+\infty} |u| e^{-\frac{u^2(1+X_1^2)}{2}} \frac{du}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}} \frac{|X_1|}{(1+X_1^2)} \\
E(e^{-\frac{N_q^2 X_1^2}{2}} |N_q X_1|) &= \sqrt{\frac{2}{\pi}} \int_0^{+\infty} \frac{1}{B(1/2,(\delta+1)/2)} \frac{u^{2/2-1}}{(1+u)^{(\delta+2+2)/2}} du \\
&= \sqrt{\frac{2}{\pi}} \frac{B(1,\delta/2+1)}{B(1/2,(\delta+1)/2)} \\
&= \sqrt{\frac{2}{\pi}} \frac{\delta}{\delta+2} R_\delta
\end{aligned}
$$

Let us now consider (41). Using the mutual independence of $N_p, X_1, \ldots, X_{\ell-1}$ and the fact that $X_i^2, i = 1, \ldots, \ell-1$ follow a beta distribution of the second kind with parameters $\frac{1}{2}, \frac{\delta+1}{2}$, we have

$$E(|N_p X_1 \ldots X_{\ell-1}|) = E(|N_p|)E(|X_1|)^{\ell-1} = \frac{\sqrt{2}}{\sqrt{\pi}}\Big(\frac{\Gamma(\frac{\delta}{2})}{\sqrt{\pi}\Gamma(\frac{\delta+1}{2})}\Big)^{\ell-1} = \sqrt{\frac{2}{\pi}}R_\delta^{\ell-1}$$

□

**Proposition Appendix .3.** *Let $X_1, \ldots, X_n, \ldots$ $S_\ell$ and $B_\ell$ as above.*

1. *For $-\frac{1}{2} < s < \frac{1}{2}(\delta+1)$ one has*

$$E(X_1^{2s}) = \frac{\Gamma(\frac{1}{2}+s)}{\Gamma(\frac{1}{2})}\frac{\Gamma(\frac{1}{2}(\delta+1)-s)}{\Gamma(\frac{1}{2}(\delta+1))}.$$

   *In particular if $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ then $E(\log X_1^2) = \psi(\frac{1}{2}) - \psi(\frac{1}{2}(\delta+1)) < 0$ and the variance of $\log X_1^2)$ is $\psi'(\frac{1}{2}) + \psi'(\frac{1}{2}(\delta+1))$.*

2. *$\lim_\ell (X_1 \ldots X_\ell)^{2/\ell} = e^{E(\log X_1^2)} < 1$, $\lim_\ell B_\ell = 0$ and $S = \lim_\ell S_\ell$ exists almost surely and is finite.*

3. *The distribution of $S$ is $\beta^{(2)}_{\frac{1}{2},\frac{1}{2}\delta}$*

4. *Denote*

$$r(\delta) = \frac{\Gamma(\frac{1}{2}(\delta+1))^2}{\Gamma(\frac{1}{2}\delta)\Gamma(\frac{1}{2}(\delta+2))}.$$

   *If $N \sim N(0,1)$ is independent of $X_1$ and $S$ then $E(e^{-\frac{1}{2}N^2 S}) = r(\delta)$, $E(e^{-\frac{1}{2}N^2 X_1^2}) = r(\delta+1)$. In particular if $g\ell(x)$ is defined by (26) then*

$$r(\delta) < g_\ell(0) < r(\delta+1). \tag{42}$$

*Proof.* 1) If $U \sim \beta^{(2)}_{a,b}$ clearly for $-a < s < b$ one has

$$E(U^s) = \frac{1}{B(a,b)}\int_0^\infty \frac{u^{a+s-1}du}{(1+u)^{a+s+b-s}} = \frac{B(a+s,b-s)}{B(a,b)} = \frac{\Gamma(a+s)}{\Gamma(a)}\frac{\Gamma(b-s)}{\Gamma(b)}.$$

As a consequence

$$E(\log U) = \frac{d}{ds}E(U^s)|_{s=0} = \frac{\Gamma'(a)}{\Gamma(a)} - \frac{\Gamma'(b)}{\Gamma(b)} = \psi(a) - \psi(b).$$

Furthermore the variance of $\log U$ is

$$\frac{d^2}{ds^2}\log E(U^s)|_{s=0} = \psi'(a) + \psi'(b).$$

Since $X_1 \sim Z/\sqrt{Q}$ then $X_1^2 \sim \frac{1}{2}\,Z^2 \times \frac{2}{Q}$ is the quotient of gamma variates of parameters $a = \frac{1}{2}$ and $b = \frac{1}{2}\,(\delta+1)$ and this proves 1).

2) The series $S = \sum_{n=1}^{\infty} X_1^2 \dots X_n^2$ converges almost surely, by the $u_n^{1/n}$ Cauchy criteria, since

$$(X_1 \dots X_\ell)^{2/\ell} = \exp \frac{1}{\ell}\sum_{i=1}^{\ell} \log X_i^2$$

converges almost surely to $e^{E(\log X_1^2)} < 1$ from the law of large numbers.

3) Suppose that $U \sim \beta_{a,b}^{(2)}$ and $V$ are independent and suppose that $b > a$. Then $U(1+V) \sim V$ if and only if $V \sim \beta_{a,b-a}^{(2)}$. See Chamayou and Letac (1991), Example 9. One applies this to $U = X_1^2$, $V = S$, $a = \frac{1}{2}$ and $b = \frac{1}{2}\,(\delta+1)$ since $S' = \sum_{i=2}^{\infty} X_2^2 \dots X_i^2 \sim S$ and $X_1^2(1+S') = S$.

4) The first part is an immediate consequence of Proposition 1. The second part is clear since $X_1^2 \le S_\ell < S$. □

It is useful to mention that

$$r(3) = \frac{8}{3\pi} = 0.84..,\ r(4) = \frac{9\pi}{32} = 0.88..,\ r(5) = \frac{128}{45\pi} = 0.90..,\ r(6) = \frac{75\pi}{256} = 0.92..,$$

From Abramovitz and Stegun page 258 we have the following informations. If $n \ge 1$ is an integer then, using the Euler constant $\gamma = 0.57721...$

$$\psi(\frac{1}{2}) = -\gamma - 2\log 2,\ \psi(\frac{1}{2}+n) = \psi(\frac{1}{2}) + 2(1+\frac{1}{3}+\dots+\frac{1}{2n-1}),\ \psi(n) = -\gamma + (1+\dots+\frac{1}{n}).$$

This leads to the following values of $E(\log X_1^2)$ according to $\delta$ :

$$-2.3863\ (\delta=3);\ -2.766\ (\delta=4);\ -2.8863\ (\delta=5).$$

Similarly, for computing the variance of $\log X_1^2$ one uses Abramovitz and Stegun page 260

$$\psi'(\frac{1}{2}) = \frac{\pi^2}{2},\ \psi'(\frac{1}{2}+n) = \frac{\pi^2}{2} - 4(1+\frac{1}{3}+\dots+\frac{1}{(2n-1)^2}),\ \psi(n) = \frac{\pi^2}{6} - (1+\dots+\frac{1}{n^2}).$$

For instance for $\delta = 3$ the variance of $\log X_1^2$ is $\sigma^2 = \frac{2\pi^2}{3} - 1 = 5.79..$ and $\sigma = 2.36...$

Let us now compute an integral we need.

**Lemma Appendix .3.**

$$E(e^{iA_1x}) = E(e^{i(U_1V_1+\ldots+U_dV_d)}) = \prod_{j=1}^{d} E(e^{iU_jV_j}) = E(e^{-\frac{U_1^2x^2}{2}})^d = \frac{1}{(1+x^2)^{d/2}} \qquad (43)$$

*and*

$$\frac{\int_{-\infty}^{+\infty} E(e^{iA_1x})|x|f(x)dx}{\int_{-\infty}^{+\infty} E(e^{iA_1x})f(x)dx} = R_{\delta+d-1}\ . \qquad (44)$$

*Proof.*

$$\rho_{\delta+d} \;=\; \frac{\int_{-\infty}^{+\infty} E(e^{iA_1x})|x|f(x)dx}{\int_{-\infty}^{+\infty} E(e^{iA_1x})f(x)dx} = \frac{\int_{-\infty}^{+\infty} \frac{1}{(1+x^2)^{d/2}}|x|\frac{1}{(1+x^2)^{(\delta+1)/2}}dx}{\int_{-\infty}^{+\infty} \frac{1}{(1+x^2)^{d/2}}\frac{1}{(1+x^2)^{(\delta+1)/2}}dx}$$

Letting $u = x^2,\ dx = \frac{1}{2\sqrt{u}}du$, we have

$$\rho_{\delta+d} \;=\; \frac{\int_0^{+\infty} \frac{u^{1/2-1/2}}{(1+u)^{(\delta+d+1)/2}}}{\int_0^{+\infty} \frac{u^{1/2-1}}{(1+u)^{(\delta+d+1)/2}}} = \frac{B(1,(\delta+d+1)/2)}{B(1/2,(\delta+d)/2)} = R_{\delta+d-1}$$

□

# Appendix A Estimating the distribution of $B_\ell$

Since $X_1^2$ has a particular $\beta^{(2)}$ distribution it is reasonable to consider more generally iid random variables $U_1,\ldots,U_n$ with distribution $\beta^{(2)}_{a,b}$ such that $b > a$, with mean $m = \psi(a) - \psi(b)$ and variance $\sigma^2 = \psi' a + \psi'(b)$. Therefore we are interested in saying something of

$$\Pr(U_1 \ldots U_n < x) = \Pr\left(\frac{1}{\sigma\sqrt{n}}\left(\sum_{i=1}^{n}\log U_i) - nm\right) < \frac{\log x - nm}{\sigma\sqrt{n}}\right) \sim \Phi(\frac{\log x - nm}{\sigma\sqrt{n}}),$$

where $\Phi$ is the distribution function of $N(0,1)$. And we apply this to $a = \frac{1}{2}$, $b = \frac{1}{2}(\delta+1)$ and $n = \ell + 1$.

If $n$ is rather small, like $n = 4$ it may seem questionable to use the central limit theorem. However, one can point out that the density of $\log U_1$ is

$$\frac{1}{B(a,b)}\frac{1}{(1+e^{-x})^a}\frac{1}{(1+e^x)^b}$$

has a bell curve for $a = \frac{1}{2}$ and $b = 2$ and the approximation by the central limit theorem should be good for $n > 3$.

# References


Albert, R. and A.-L. Barabási (2002). Statistical mechanics of complex networks. *Reviews of modern physics 74*(1), 47.

Atay-Kayis, A. and H. Massam (2005). A monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models. *Biometrika 92*(2), 317–335.

Baldi, P., S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics 16*(5), 412–424.

Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. *Annals of Statistics*, 870–897.

Bhadra, A. and B. K. Mallick (2013). Joint high-dimensional bayesian variable and covariance selection with an application to eqtl analysis. *Biometrics 69*(2), 447–457.

Cappé, O., C. Robert, and T. Rydén (2003). Reversible jump, birth-and-death and more general continuous time markov chain monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(3), 679–700.

Chamayou, J.-F. and G. Letac (1991). Explicit stationary distributions for compositions of random functions and products of random matrices. *Journal of Theoretical Probability 4*(1), 3–36.

Chen, L., T. Tong, and H. Zhao (2008). Considering dependence among genes and markers for false discovery control in eqtl mapping. *Bioinformatics 24*(18), 2015–2022.

Cheng, Y. and A. Lenkoski (2012). Hierarchical Gaussian graphical models: Beyond reversible jump. *Electronic Journal of Statistics 6*, 2309–2331.

Dobra, A. and A. Lenkoski (2011). Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics 5*(2A), 969–993.

Dobra, A., A. Lenkoski, and A. Rodriguez (2011). Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association 106*(496), 1418–1433.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Giudici, P. and P. Green (1999). Decomposable graphical gaussian model determination. *Biometrika 86*(4), 785–801.

Gu, Q., Y. Cao, Y. Ning, and H. Liu (2015). Local and global inference for high dimensional nonparanormal graphical models. *arXiv preprint arXiv:1502.02347*.

Jones, B., C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science 20*(4), 388–400.

Lenkoski, A. (2013). A direct sampler for g-wishart variates. *Stat 2*(1), 119–128.

Lenkoski, A. and A. Dobra (2011). Computational aspects related to inference in gaussian graphical models with the g-wishart prior. *Journal of Computational and Graphical Statistics 20*(1), 140–157.

Liang, F. (2010). A double metropolis–hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation 80*(9), 1007–1022.

Madigan, D., A. E. Raftery, C. Volinsky, and J. Hoeting (1996). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*, pp. 77–83.

Mohammadi, A., F. Abegaz Yazew, E. van den Heuvel, and E. C. Wit (2017). Bayesian modelling of dupuytren disease using gaussian copula graphical models. *Journal of Royal Statistical Society-Series C 66*(3), 629–645.

Mohammadi, A. and E. Wit (2015). Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis 10*(1), 109–138.

Mohammadi, A. and E. Wit (2017a). BDgraph: An R package for Bayesian structure learning in graphical models. *Journal of Statistical Software*. accepted.

Mohammadi, A. and E. Wit (2017b). *BDgraph: Bayesian Structure Learning in Graphical Models using Birth-Death MCMC*. R package version 2.38.

Murray, I., Z. Ghahramani, and D. MacKay (2012). Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*.

Preston, C. J. (1976). Special birth-and-death processes. *Bulletin of the International Statistical Institute 46*, 371–391.

Roverato, A. (2002). Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics 29*(3), 391–411.

Scott, J. G. and C. M. Carvalho (2008). Feature-inclusion stochastic search for gaussian graphical models. *Journal of Computational and Graphical Statistics 17*(4), 790–808.

Stranger, B. E., A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, et al. (2007). Population genomics of human gene expression. *Nature genetics 39*(10), 1217–1224.

Uhler, C., A. Lenkoski, and D. Richards (2014). Exact formulas for the normalizing constants of wishart distributions for graphical models. *arXiv preprint arXiv:1406.4901*.

Wang, H. and S. Li (2012). Efficient gaussian graphical model determination under g-wishart prior distributions. *Electronic Journal of Statistics 6*, 168–198.

Wong, F., C. K. Carter, and R. Kohn (2003). Efficient estimation of covariance selection models. *Biometrika*, 809–830.

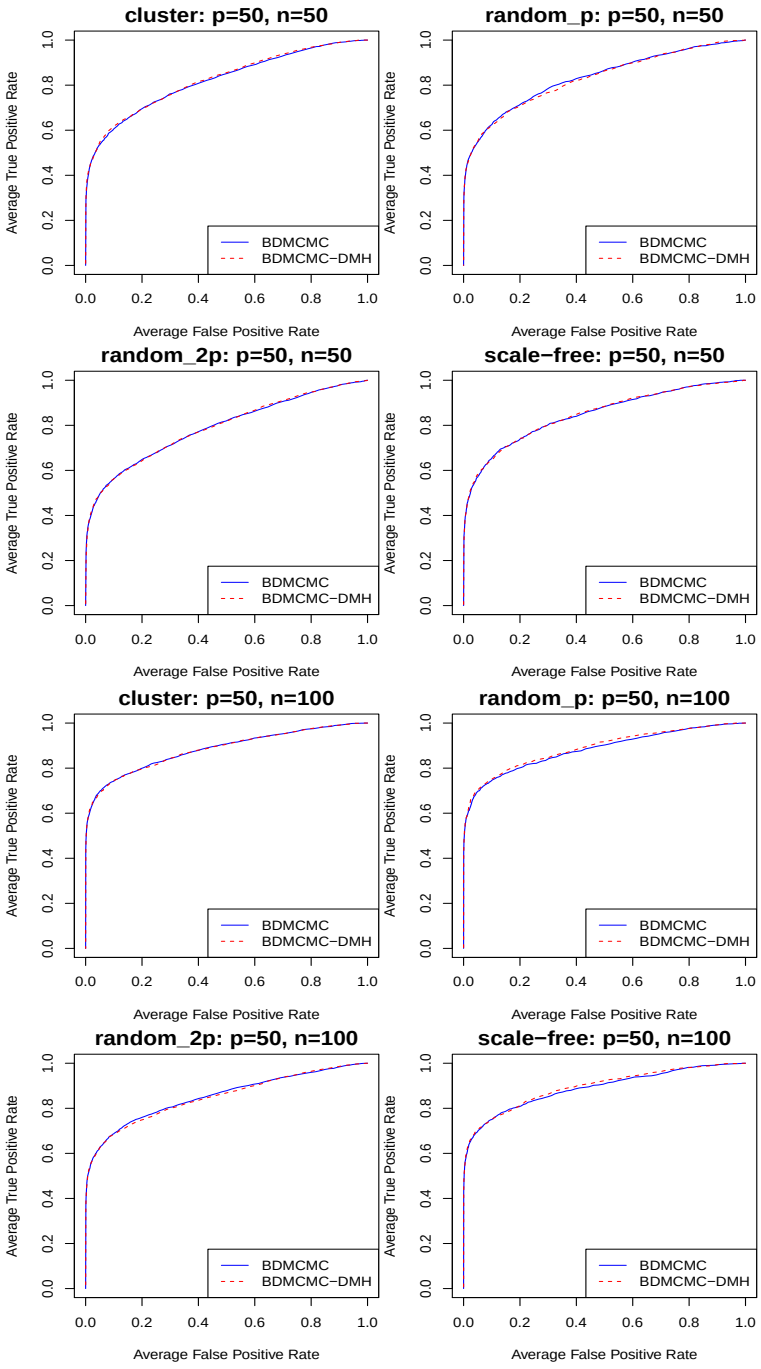


Figure 4: ROC curves for BDMCMC algorithm with the approximation of Theorem 3.1 (BDMCMC) and BDMCMC algorithm with double Metropolis-Hastings (BDMCMC-DMH), with 50 replications. Here, $p = 50$, $n \in \{50, 100\}$, and 6 different graph structures.

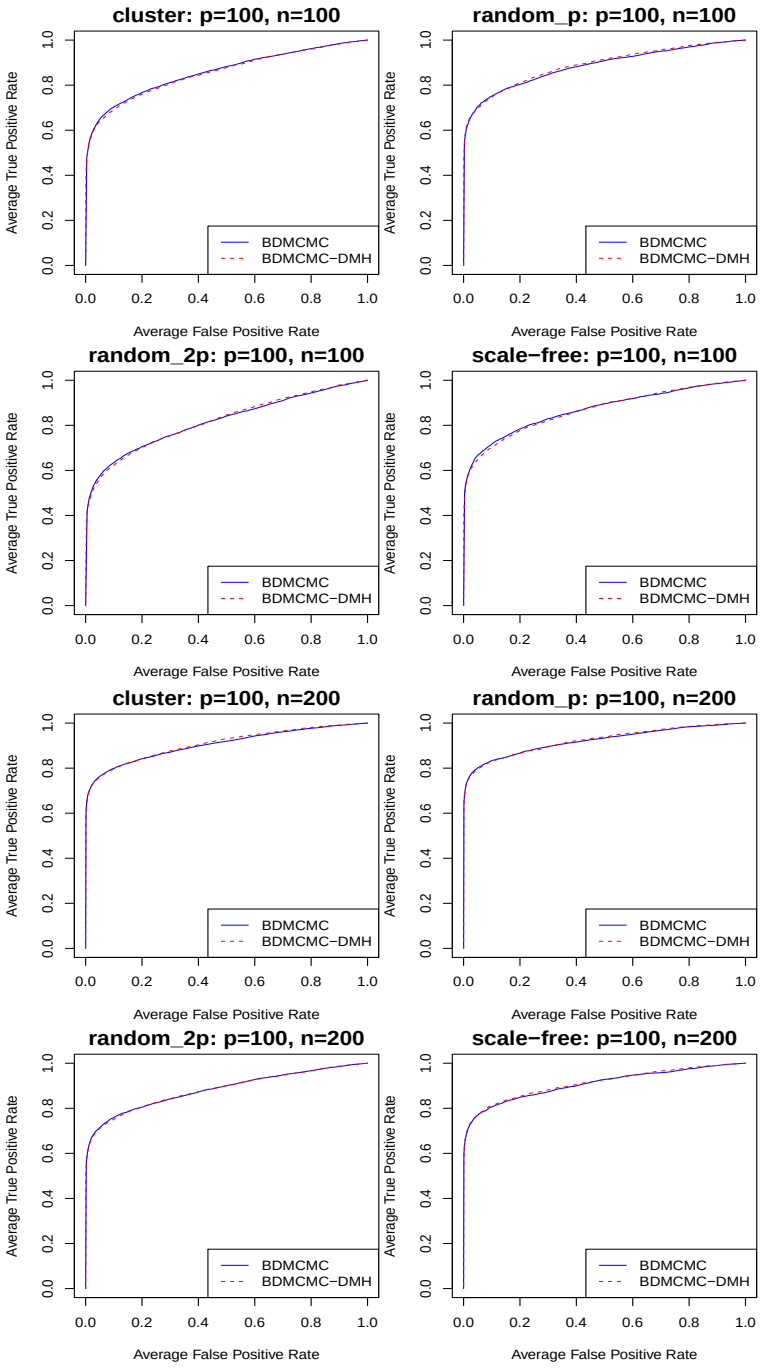


Figure 5: ROC curves for BDMCMC algorithm with the approximation of Theorem 3.1 (BDMCMC) and BDMCMC algorithm with double Metropolis-Hastings (BDMCMC-DMH), with 50 replications. Here, $p = 100$, $n \in \{100, 200\}$, and 6 different graph structures.

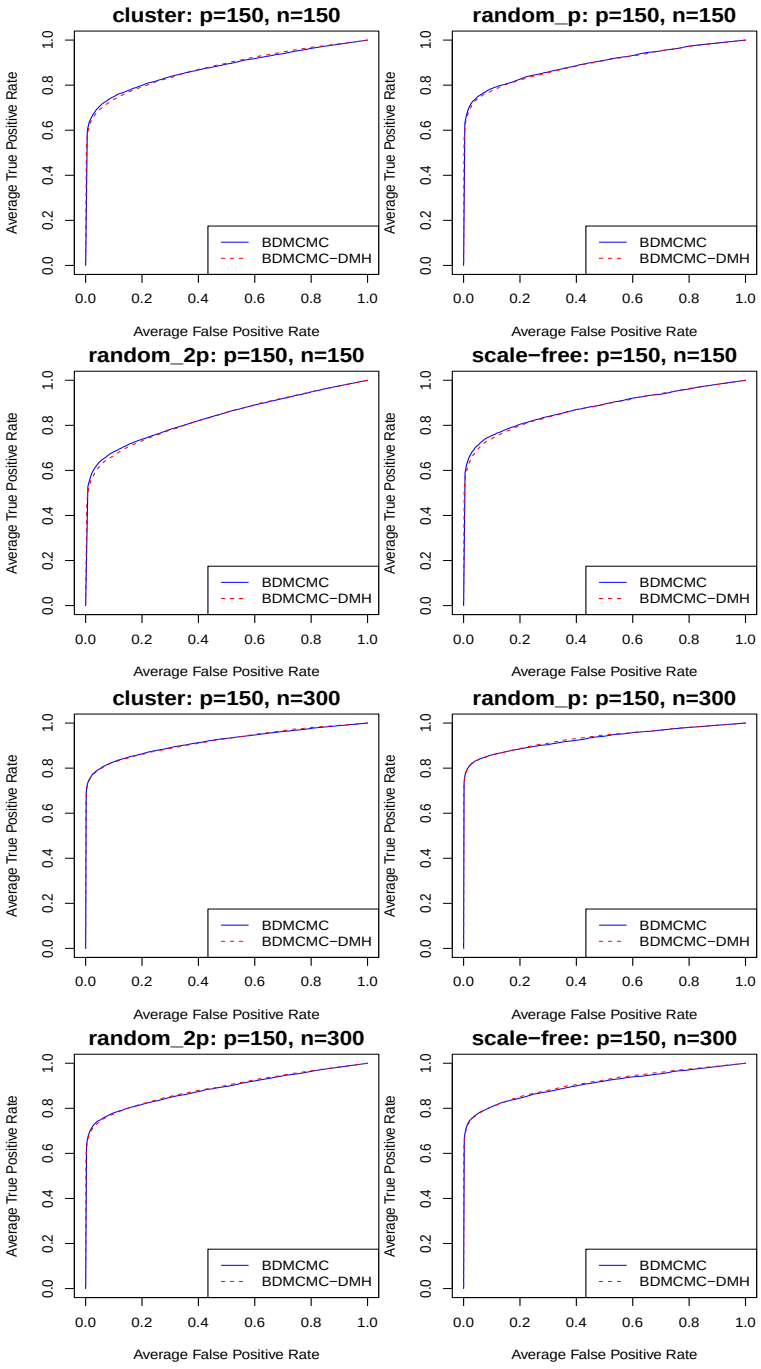


Figure 6: ROC curves for BDMCMC algorithm with the approximation of Theorem 3.1 (BDMCMC) and BDMCMC algorithm with double Metropolis-Hastings (BDMCMC-DMH), with 50 replications. Here, $p = 150$, $n \in \{150, 300\}$, and 6 different graph structures.

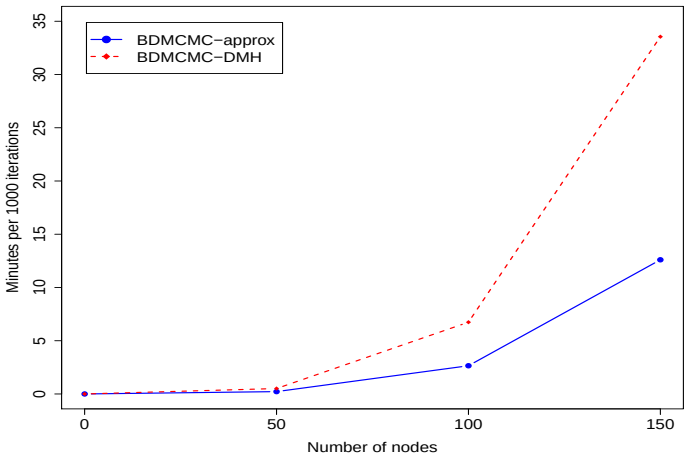


Figure 7: Plot which present time execution of BDMCMC algorithm with the approximation of Theorem 3.1 (BDMCMC) and with double Metropolis-Hastings (BDMCMC-DMH). Time is per minutes for 1000 iterations for different number of variables ($p = 50, 100, 150$).

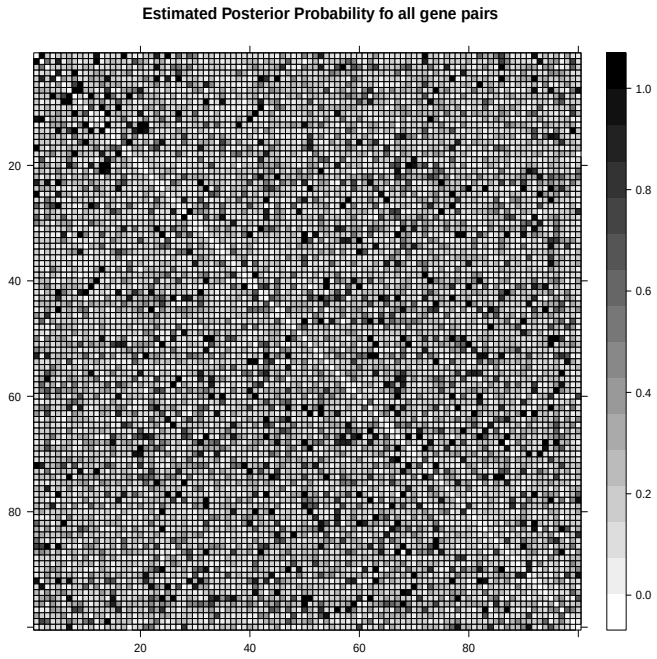


Figure 8: Image of estimated posterior probability of all gene pairs, according to the result of BDMCMC algorithm with our approximation based on model averaging using formula 35.

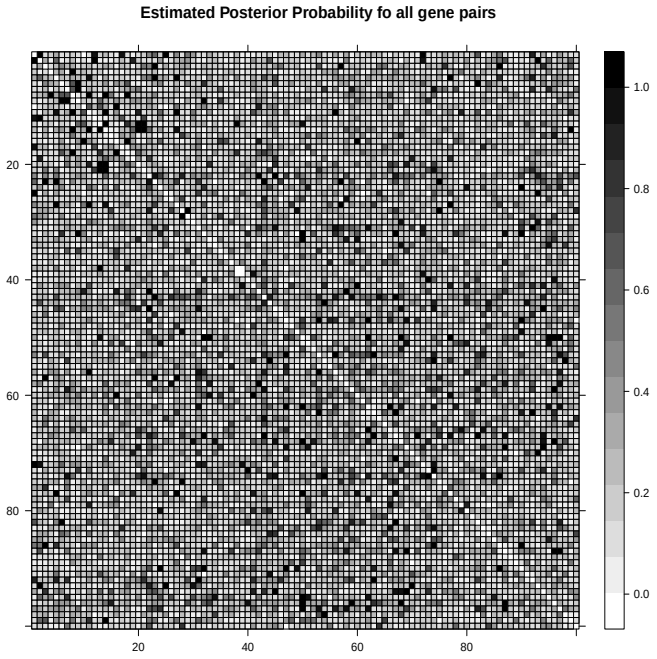


Figure 9: Image of the estimated posterior probability for all gene pairs, according to the result of the BDMCMC algorithm with the double Metropolis-Hastings approximation based on model averaging using formula 35.

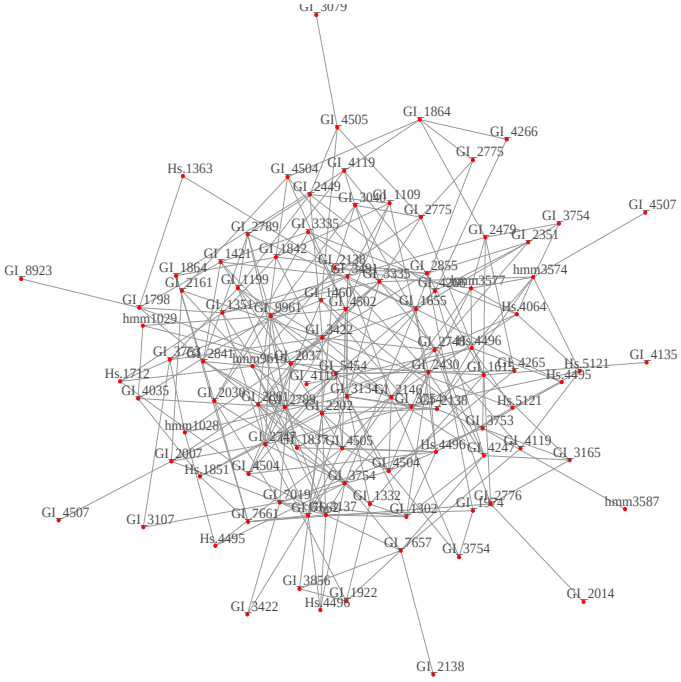


Figure 10: Visualization of the selected graph, according to the result of the BDM-CMC algorithm based on model averaging using (35). The selected graph includes the edges with the estimated posterior probability greater than 0.8.