Bayesian Methods for Gaussian Graphical Models

by

Nikolaos Mitsakakis

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Dalla Lana School of Public Health
University of Toronto

Copyright © 2010 by Nikolaos Mitsakakis
Abstract

Bayesian Methods for Gaussian Graphical Models

Nikolaos Mitsakakis
Doctor of Philosophy
Graduate Department of Dalla Lana School of Public Health
University of Toronto
2010

This thesis contributes to the field of Gaussian Graphical Models by exploring either numerically or theoretically various topics of Bayesian Methods in Gaussian Graphical Models and by providing a number of interesting results, the further exploration of which would be promising, pointing to numerous future research directions.

Gaussian Graphical Models are statistical methods for the investigation and representation of interdependencies between components of continuous random vectors. This thesis aims to investigate some issues related to the application of Bayesian methods for Gaussian Graphical Models. We adopt the popular $G$-Wishart conjugate prior $W_G(\delta, D)$ for the precision matrix. We propose an efficient sampling method for the $G$-Wishart distribution based on the Metropolis Hastings algorithm and show its validity through a number of numerical experiments. We show that this method can be easily used to estimate the Deviance Information Criterion, providing a computationally inexpensive approach for model selection.

In addition, we look at the marginal likelihood of a graphical model given a set of data. This is proportional to the ratio of the posterior over the prior normalizing constant. We explore methods for the estimation of this ratio, focusing primarily on applying the Monte Carlo simulation method of path sampling. We also explore numerically the effect of the completion of the incomplete matrix $D^V$, hyperparameter of the $G$-Wishart distribution, for the estimation of the normalizing constant.
We also derive a series of exact and approximate expressions for the Bayes Factor between two graphs that differ by one edge. A new theoretical result regarding the limit of the normalizing constant multiplied by the hyperparameter $\delta$ is given and its implications to the validity of an improper prior and of the subsequent Bayes Factor are discussed.
Dedication

To my family
Acknowledgements

I would like to thank my supervisors Michael Escobar and Hélène Massam for their invaluable guidance and continuous help and support. I would also like to thank my committee member Rafal Kustra for originally motivating me to join the PhD program and for first inspiring my interest on Gaussian Graphical Models. Finally, I would like to thank my wife Serena for her support and patience during the course of my studies.
# Contents

1 Introduction .................................................. 1

2 Background .................................................. 3
   2.1 Graph Theory ........................................... 3
      2.1.1 General ............................................... 3
      2.1.2 Graphs and Linear Algebra .......................... 5
   2.2 Gaussian Graphical Models ............................... 6
      2.2.1 Introduction ........................................ 6
      2.2.2 Bayesian approach and the $G$-Wishart prior .......... 9
   2.3 Model Search and Selection and Covariance Estimation .... 13

3 Estimation of the Normalizing Constant of $W_G(\delta, D)$ .... 15
   3.1 Introduction - Motivation ................................ 15
   3.2 Previous approaches ...................................... 16
      3.2.1 Monte Carlo method .................................. 16
      3.2.2 Laplace approximation ................................ 19
   3.3 New proposed approaches .................................. 20
      3.3.1 Using estimation of the marginal likelihood ............ 20
      3.3.2 Using Path Sampling .................................. 21
      3.3.3 Numerical results .................................... 32
   3.4 Monte Carlo estimation method and matrix $D$ ............... 35
3.5 Conclusions .......................................................... 44

4 Sampling from G-Wishart distribution 45
  4.1 Introduction - Motivation ........................................ 45
  4.2 Existing methods .................................................. 46
  4.3 New proposed method .............................................. 48
    4.3.1 Background on the general Metropolis-Hastings algorithm .... 48
    4.3.2 Description of the proposed MH-based sampling method ....... 50
    4.3.3 Experiments ................................................... 51
    4.3.4 Application for model selection using DIC ................. 61

5 Bayes Factors 65
  5.1 Definition and importance in Model Selection ................. 65
  5.2 Expression for decomposable graphs ............................ 66
  5.3 Bayes Factors for non-decomposable graphs .................... 69
  5.4 Approximation of Bayes Factors ............................... 70
  5.5 Limit of $\delta \cdot I_G(\delta, D)$ and applications ............... 73

6 Conclusions and future directions 79

Bibliography 81
List of Tables

3.1 Results from numerical experiment comparing path-sampling and Monte Carlo based estimates of the log ratio of normalizing constants. Standard errors are reported for the Monte Carlo based estimates, with NE denoting “non-estimable” errors. True values are also shown for decomposable graphs. 36

4.1 Results from numerical experiment comparing DIC and other related parameters between two graphical models using iris virginica data . . . . . 64
List of Figures

3.1 Scatterplot between true values of log ratio of normalizing constants and their Monte Carlo estimate .......................... 34
3.2 Plot comparing the ranking of graphs (in decreasing order) based on the true value of the log ratio of normalizing constants and on the path-sampling estimate ........................................ 37
3.3 Convergence of the Monte Carlo estimate of the normalizing constant, when $D$ and $D^c$ are used ........................................... 39
3.4 Contour plot of the logarithm of the variance estimate against the values of $D_{14}$ and $D_{23}$. The black dot indicates the variance corresponding to $D^c$ 41
3.5 Contour plot of $|D|$ against the values of $D_{14}$ and $D_{23}$. The black dot indicates the variance corresponding to $D^c$ .......................... 42
3.6 Scatter plot of $|D|$ and $\log(Var(f_D(\psi^v)))$, when $D_{14}$ and $D_{23}$ vary. The black dot corresponds to $D^c$ ........................................... 43
4.1 Non-decomposable graph used in Experiment 1 ......................... 54
4.2 Non-decomposable graph used in Experiments 2 and 3 ................. 54
4.3 Non-decomposable graph used in Experiments 4 and 5 .................. 55
4.4 QQ-plot, acceptance rate plot and trace plot for Experiment 1 ......... 55
4.5 Plots for Experiment 2 using random $D$ ..................................... 56
4.6 Plots for Experiment 2 using $D^c$ .............................................. 57
4.7 Plots for Experiment 3 using random $D = I + X^T X$ .................... 58
4.8  Plots for Experiment 4 using $D = I + X^T X$  

4.9  Plots for Experiment 4 using $D = (I + X^T X)^c$  

4.10 Scatterplot between the log ratio of the normalizing constants, $\lambda$, and the value of -DIC
Chapter 1

Introduction

Gaussian Graphical Models (GGMs) are statistical methods for the investigation of continuous random vectors and the interdependencies between their components. Recent development of various experimental techniques (such as microarrays) have given rise to large amount of continuous multivariate datasets, in disciplines like biology or finance. Investigation of conditional interdependencies underlying the mechanisms generating those datasets is often desirable. Gaussian Graphical Models can be used for such endeavour. Particularly in genomics, they have been used for the investigation of how the cell works, by using gene expression or other high-throughput data. Given a set of such data, one problem of interest is the identification of a graphical model that fits the data well. In that model, the existence of an edge connecting the nodes corresponding to the expressions of two genes would mean that those two genes are associated directly, while the lack of an edge will indicate the conditional independence of these two genes and the lack of direct association. A model for a gene network would then be constructed from the dataset. After one or more models are selected for the given dataset, a subsequent problem of interest is the estimation of the covariance matrix or a function of it, such as the values of the partial correlations. Those estimates could be used as a measure of how strong the association between two genes is, with larger partial correlation indicating
stronger conditional dependence. Similar problems can be tackled when applying GGMs to data from other fields, such as portfolio funds performances in finance.

In this thesis we approach these problems adopting a Bayesian framework. Assuming that the data come from a multivariate Gaussian distribution, the conjugate prior for the precision matrix can be used, corresponding to a distribution named $G$-Wishart. Various problems related to this distribution remain still unsolved, and some of them I am exploring in this thesis. The main topics of my thesis are:

- Application of path sampling for the estimation of the ratio of normalizing constants of the $G$-Wishart

- Investigation of the role of the hyperparameter matrix $D$ for the estimation of the normalizing constant

- Introduction of a novel method for sampling from the $G$-Wishart distribution

- Study of the asymptotic behavior of the normalizing constant, with application to the Bayes factor
Chapter 2

Background

2.1 Graph Theory

2.1.1 General

A graph is a pair $G = (V, E)$ where $V$ is the set of vertices and $E \subseteq V \times V$ is the set of edges. A graph is called undirected if its edges have no “direction”, i.e. if there is no distinction between $(i, j)$ and $(j, i)$, when $(i, j) \in E$. The notation $i \sim_G j$ is then used to indicate that vertex $v_i$ is linked to vertex $v_j$. In this thesis every graph is assumed to be undirected, unless otherwise indicated. In addition to the notation $E$, the edges of a graph can be represented by the set $\hat{V} = \{(i, j), i < j, \text{ such that } i = j \text{ or } (i, j) \in E\}$. It is clear that $\hat{V}$ contains all the vertex-pairs that correspond to missing edges of $G = (V, E)$. A sequence of vertices $v_{i_1}, v_{i_2}, v_{i_3}, \ldots, v_{i_{k-1}}, v_{i_k}$ is called a path that connects vertices $v_{i_1}$ and $v_{i_k}$ if $i_1 \sim i_2, i_2 \sim i_3, \ldots, i_{k-1} \sim i_k$. A path is called proper if all vertices $v_{i_1}, v_{i_2}, \ldots, v_{i_k}$ are distinct. A proper path that connects one vertex with itself is called cycle. The number of vertices in a cycle is called length of the cycle. Let $v_{i_1}, v_{i_2}, \ldots, v_{i_{k-1}} = v_{i_1}$ a cycle of length $k$. Any edge connecting two distinct vertices $v_{i_l}, v_{i_m}$ from $\{v_{i_1}, v_{i_2}, \ldots, v_{i_k}\}$, where $l \neq m + 1, l \neq m - 1$, is called a chord of the cycle. For simplicity, for the rest of the thesis we will refer to a vertex with its index, e.g. $i$ or $j$, instead of $v_i$ or $v_j$. 
Given the graph \( G = (V, E) \) and \( A \subset V \), the induced subgraph \( G_A \) is the graph \( G(A, E_A) \), where \( E_A = E \cap \{(i, j) : i, j \in A, i \neq j\} \). \( G_A \) is proper subgraph of \( G \) if \( A \) is a proper subset of \( V \), i.e. there are vertices in \( V \) that do not belong to \( A \). A graph \( G = (V, E) \) is called complete if every vertex in \( V \) is connected by some edge with every other vertex in \( V \). A complete subgraph of \( G \) that is not proper subgraph of any other complete subgraph is called a clique. If \( A, B, C \) are subsets of \( V \), \( B \) is called separator of \( A \) and \( C \) if every path that connects any vertex of \( A \) with any vertex of \( C \) passes through \( B \). A triplet \( (A, B, C) \) of disjoint subsets of \( V \) is called decomposition of \( G \), if (a) \( C \) is separator of \( A, B \) and (b) \( C \) is complete. A decomposition is called proper if both \( A \) and \( B \) are not empty. A (sub)graph that does not accept a decomposition is called prime.

A graph \( G \) is called decomposable if it is complete or it assumes a proper decomposition \((A, B, C)\) such that the induced subgraphs \( G_{AC} \) and \( G_{BC} \) are both decomposable. An equivalent but much easier to use characterization of a decomposable graph is related to the existence of chordless cycles. It can be proved ([Lauritzen, 1996]) that a graph is non-decomposable if and only if it contains chordless cycles of length larger than 3.

**Definition** Let \( V_1, V_2, \ldots, V_k \) a sequence of subsets of \( V \) and let \( H_i = V_1 \cup \ldots \cup V_i \), \( S_i = H_{i-1} \cap V_i \) and \( R_i = V_i \setminus H_{i-1} \). The sequence is called perfect if (a) for all \( j > 1 \) there is \( i < j \) such that \( S_j \subseteq V_i \) and (b) the sets \( S_i \) are complete for all \( i > 1 \). Condition (a) is called running intersection property. The sets \( H_i, S_i \) and \( R_i \) are called histories, separators and residuals, respectively.

It can be proven (Proposition 2.17 in [Lauritzen, 1996], Theorem 3.3 in [Leimer, 1993]) that if \( G \) is decomposable, its cliques \( C(G) \) can be ordered in a perfect sequence, \( C_1, C_2, \ldots, C_k \). It can also be proven ([Leimer, 1993]) that all the prime subgraphs of any arbitrary graph (possibly not decomposable) can be ordered into a perfect sequence, \( P_1, P_2, \ldots, P_k \) (see also [Dobra and Fienberg, 2000] for a discussion). The prime subgraphs (some of which can be cliques) are often called prime components of \( G \).
2.1.2 Graphs and Linear Algebra

In this section we present some definitions, notations and results from Linear Algebra and Matrix Theory that are important in Graph Theory. In this thesis we denote with \( M^+ \) the set of all \( p \times p \) positive definite matrices. Given a graph \( G = (V, E) \) we denote with \( M^+(G) \) the set of all \( p \times p \) positive definite matrices \( X \), for which \( X_{ij} = 0 \) whenever \( i \not\sim_G j \); in other words

\[
M^+(G) = \{ X \in M^+ | X_{ij} = 0, \forall (i, j) \in \bar{V} \}
\]

(2.1)

Let \( G = (V, E) \) a graph with \( |V| = p \). If \( C \) is a subset of \( W = \{ (i, j) | i, j \in V, i \leq j \} \), the \( C \)-incomplete symmetric matrix \( \Gamma^C \) is the symmetrized matrix with all the elements \( \Gamma_{ij} \) specified for \( (i, j) \in C \) and all the remaining elements unspecified. If it is possible to fill a \( C \)-incomplete matrix \( \Gamma^C \) in a way that the completed matrix \( \Gamma \) is positive definite, we say that \( \Gamma^C \) admits a positive completion. The set \( M^+_V(G) \) denotes all the \( V \)-incomplete symmetric matrices that admit a positive completion. If the completed matrix \( \Gamma \) is such that \( \Gamma^{-1} \in M^+(G) \), it is called a PD-completion of \( \Gamma^V \). The following fundamental Theorem indicates that the spaces \( M^+_V(G) \) and \( M^+(G) \) are isomorphic.

**Theorem 2.1.1** ([Grone et al., 1984]) Let \( G = (V, E) \) be a graph and \( \Gamma^V \in M^+_V(G) \) a \( V \)-incomplete matrix that admits a positive completion. Then there exists a unique PD-completion \( \Gamma \), such that \( \Gamma^{-1} \in M^+(G) \).

The isomorphism between the sets \( M^+_V(G) \) and \( M^+(G) \) is due to the mapping that corresponds \( \Gamma^V \) to its PD-completion \( \Gamma \). This mapping is well defined and bijective according to Grone’s theorem. It can be seen (e.g. in Proposition 2 of [Atay-Kayis and Massam, 2005]) that the elements \( \Gamma_{ij}, (i, j) \in \bar{V} \) are functions of the elements \( \Gamma_{ij}, (i, j) \in V \), which are called the *free* elements of \( \Gamma \).

We also need the following definitions and notations. With \( M^\triangleleft \) we denote the set of all upper triangular matrices with positive diagonal elements. With \( M^\triangleleft_V(G) \) we denote
the set of $\mathcal{V}$-incomplete upper triangular matrices with positive diagonal elements. The following important proposition defines an isomorphism between $M^+(G)$ and $M_*^<(G)$.

**Proposition 2.1.2** ([Roverato, 2002], [Atay-Kayis and Massam, 2005]) Let $x \in M^+(G)$ and $x = \phi^T \phi$ its Choleski decomposition. For $(i, j) \in \mathcal{V}$,

$$
\phi_{ij} = \begin{cases} 
0 & \text{if } i = 1 \\
-\sum_{k=1}^{i-1} \frac{\phi_{ik} \phi_{kj}}{\phi_{ii}} & \text{otherwise}.
\end{cases} \quad (2.2)
$$

The proof can be found in [Roverato, 2002] and [Atay-Kayis and Massam, 2005]. This proposition indicates that the elements of $\phi$ that correspond to the missing edges of $G$ can be uniquely determined by the elements of $\phi$ that correspond to the edges of $G$. For that reason, the latter are called “free” elements and the others are called “non-free”. The Equation 2.2 can be used for the completion of the $\mathcal{V}$-incomplete upper triangular matrix $\phi^\mathcal{V} \in M_*^<(G)$ to the upper triangular matrix $\phi \in M^<$. This completion mechanism is called $T$-completion. One can now define the mapping

$$
M_*^<(G) \ni \phi^\mathcal{V} \mapsto \pi(\phi^\mathcal{V}) = K \in M^+(G),
$$

where $K = \phi^T \phi$ and $\phi = T(\phi^\mathcal{V})$ is the $T$-completion of $\phi^\mathcal{V}$. Due to Equation 2.2 this mapping is well defined and 1-1. It is also “onto” as, if $K \in M^+(G)$ and $K = \phi^T \phi$, $K = \pi(\phi^\mathcal{V})$ where $\phi^\mathcal{V}$ is the projection of $\phi$ to $M_*^<(G)$. Therefore the mapping $\pi$ defines an isomorphism between $M^+(G)$ and $M_*^<(G)$.

## 2.2 Gaussian Graphical Models

### 2.2.1 Introduction

In this section we introduce the concept of a Gaussian Graphical Model and present some of its important properties. We first describe the concept of conditional independence.
Let $X, Y$ and $Z$ be random variables with some joint distribution $P$. We say that $X$ and $Y$ are conditionally independent given $Z$ if

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

(2.3)

If $X, Y, Z$ admit a joint density $f$, the above relation is equivalent to

$$f_{XY|Z}(X=x, Y=y|Z=z) = f_X(Z=x|Z=z) f_Y(Z=y|Z=z)$$

(2.4)

We then use the notation

$$X \perp \!
\!
\!
\perp Y | Z$$

(2.5)

This concept can be extended to random vectors or collections of variables. If $X_A, X_B, X_C$ are collections of random variables indexed with the sets $A, B, C$, we say that $X_A, X_B$ are conditionally independent given $X_C$ if

$$P(X_A, X_B|X_C) = P(X_A|X_C)P(X_B|X_C),$$

(2.6)

in which case we use the notation

$$X_A \perp \!
\!
\!
\perp X_B | X_C$$

(2.7)

We can now present two properties that characterize how a random vector is associated with a graph. Let $G = (V, E)$ a graph with $|V| = p$ and $X$ a $p$-dimensional random vector, the components of which are indexed according to the vertices of $G$, $V$. We say that the distribution of $X$, $P$ is pairwise Markov with respect to $G$ if

$$\forall i,j \in V, i \not\sim_G j \Rightarrow i \perp \!
\!
\!
\perp j|V \setminus \{i,j\}$$

(2.8)

We say that the distribution of $X$, $P$ is global Markov with respect to $G$ if

$$\forall A, B, C \subset V, C \text{ separates } A, B \Rightarrow A \perp \!
\!
\!
\perp B | C$$

(2.9)

We also have the following proposition:
Proposition 2.2.1 If the distribution $P$ is global Markov with respect to graph $G$ with density $p$, then for any decomposition $(A, B, C)$ of $G$,

$$p(X) = \frac{p_{A\cup C}(X_{A\cup C})p_{B\cup C}(X_{B\cup C})}{p_C(X_C)}$$

(2.10)

Proof It is a simplified version of Proposition 3.17 of [Lauritzen, 1996].

It can be easily seen that the global Markov property implies the pairwise Markov property. It can also be proven (see e.g. [Lauritzen, 1996]) that if the distribution $P$ has a positive and continuous density, e.g. $P$ is Gaussian, the pairwise Markov property implies the global Markov property. Therefore for Gaussian distributions, we do not distinguish between pairwise and global, and we use the term Markov property.

Given a graph $G = (V, E)$, and a random vector $X$ with components indexed from $V$ as above, the Gaussian Graphical Model of $X$ with graph $G$ is the class $\mathcal{M}_G$ of all multivariate Gaussian distributions that are Markov with respect to $G$.

We now recall an important property of the Gaussian distribution that connects conditional independence with the precision matrix (i.e. the inverse covariance matrix): if $K$, written as

$$K = \begin{bmatrix} K_{11} & \cdots & K_{1p} \\ \vdots & \ddots & \vdots \\ K_{p1} & \cdots & K_{pp} \end{bmatrix},$$

is the precision matrix of a multivariate gaussian distribution, the components $X_i, X_j$ are conditionally independent given the rest of the components $(X_k)_{k \in V \setminus \{i,j\}}$ if and only if the $(i,j)$ entry of $K$ is equal to 0. This can easily be seen if we consider the conditional joint distribution of $(X_i, X_j)$, given $(X_k)_{k \in V \setminus \{i,j\}}$, known to be a bivariate normal distribution with covariance matrix

$$\Sigma_{i,j,V \setminus \{i,j\}} = \begin{bmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{bmatrix}^{-1} = \frac{1}{K_{ii}K_{jj} - K_{ij}^2} \begin{bmatrix} K_{jj} & -K_{ji} \\ -K_{ij} & K_{ii} \end{bmatrix}.$$
Therefore $X_i, X_j$ are conditionally independent given $(X_k)_{k \in V \setminus \{i,j\}}$ if and only if

$$[\Sigma_{ij, V \setminus \{i,j\}}]_{12} = 0,$$

i.e. if and only if $K_{ij} = 0$. In summarizing

$$\forall i, j \in V, X_i \perp \!\!\!\! \perp X_j | X_{V \setminus \{i,j\}} \iff K_{ij} = 0 \quad (2.11)$$

Therefore, a graph $G$ defines a Gaussian Graphical Model as the class of all multivariate normal distributions with precision matrix $K$ with zero entries for all the missing edges of $G$, or otherwise $K \in M^+(G)$. For simplicity and without loss of generality, we assume that the distributions have zero mean, reaching therefore to the formal definition

$$\mathcal{M}_G = \{N_p(0, \Sigma) | \Sigma^{-1} \in M^+(G)\} \quad (2.12)$$

### 2.2.2 Bayesian approach and the $G$-Wishart prior

Given a multivariate dataset, such as gene expression data, one is often interested in the identification of a graphical gaussian model (or a set of models) describing the structure of the data. This often includes the identification of the conditional independencies between the random variables that generated the dataset, as well as the estimation of the covariance (or precision) matrix. In the Bayesian framework, dealing with those problems requires the choice of a prior for the covariance (or precision) matrix.

Given a graph $G$ a sample $x = (x^{(1)}, \ldots, x^{(n)})$ from the graphical model $\mathcal{M}_G$ has density

$$p(x|K) = \frac{|K|^{n/2}}{(2\pi)^{np/2}} \exp\{-\frac{1}{2} \langle K, u \rangle\}, \quad (2.13)$$

where $K \in M^+(G)$, $u = x^T x = \sum_{i=1}^n x^{(i)} x^{(i)T}$ and $\langle a, b \rangle$ denotes the inner product of matrices $a, b$, equal to the trace of $a^T b$.

If $f(K|G)$ is the prior of $K$, the marginal likelihood of $G$ given a data set $x$ is given by

$$p(x|G) = \int_{M^+(G)} p(x|K, G) f(K|G) dK \quad (2.14)$$
It is known that the random quantity $u$ in Equation 2.13 follows the Wishart distribution. Since it belongs to an exponential family, with parameter $K$, one choice of prior for $K$ is the $DY$-conjugate prior, which is described in [Diaconis and Ylvisaker, 1979]. The conjugate family for $K$ has density

$$f(K|G, \delta, D) \propto |K|^{(\delta-2)/2} \exp\{-\frac{1}{2} \langle K, D \rangle\},$$

(2.15)

where $\delta \in \mathbb{R}, D \in M^+$. In order for this density to be proper, the values of $\delta$ and $D$ need to be such that the normalizing constant

$$I_G(\delta, D) = \int_{K \in M^+(G)} |K|^{(\delta-2)/2} \exp\{-\frac{1}{2} \langle K, D \rangle\} dK$$

(2.16)

is finite.

The distribution of the $DY$-conjugate prior presented in Equation 2.15 has been named $G$-Wishart ([Roverato, 2002], [Atay-Kayis and Massam, 2005], [Lenkoski and Dobra, 2008]), due to its similarities with the Wishart distribution, while being restricted by the structure of the graph $G$. The notation used for this distribution is $W_G(\delta, D)$. Under the selection of $G$-Wishart prior, the marginal likelihood $p(x|G)$ also depends on the values of $\delta$ and $D$. Whenever we need to emphasize this dependency we use the notation $p(x|G, \delta, D)$ instead.

When $G$ is complete, there are no restrictions in $K$ and $W_G(\delta, D)$ becomes the Wishart distribution, while its normalizing constant is equal to

$$I_G(\delta, D) = 2^{(\delta+p-1)p/2} \frac{\Gamma_p[(\delta + p - 1)/2]}{|D|^{(\delta+p-1)/2}},$$

(2.17)

where $p$ is the number of nodes of $G$.

When a matrix $K \in M^+(G)$ follows the $G$-Wishart distribution, its inverse, $\Sigma = K^{-1}$, follows a distribution that has been named $Hyper$ $Inverse$ $Wishart$ ([Roverato, 2002]). The notation $HIW_G(\delta, D)$ is used for this distribution. When $K \in M^+(G)$, the matrix $\Sigma$ is determined uniquely by its “free” elements, $\Sigma_{i,j}, (i, j) \in \mathcal{V}$. Therefore, we can say
that the actual parameter of $HIW_G(\delta, D)$ is not the complete matrix $\Sigma$ but rather the incomplete matrix $\Sigma^V$. The density of $HIW_G(\delta, D)$ is

$$HIW_G(\Sigma^V|\delta, D) = \frac{1}{I_G(\delta, D)}|\Sigma|^{-(\delta-2)/2}|J(K \mapsto \Sigma^V)| \exp \left( \frac{1}{2} (\Sigma^{-1}, D) \right),$$ (2.18)

where $J(K \mapsto \Sigma^V)$ is the Jacobian for the change of variable $K \mapsto \Sigma^V$ and $\Sigma$ denotes the PD-completion of $\Sigma^V$.

The value of $J(K \mapsto \Sigma^V)$ has been previously studied in [Roverato and Whittaker, 1998]. It has been found (see also Proposition 1 of [Roverato, 2002]) that $|J(K \mapsto \Sigma^V)| = 2^p |\text{Iss}(\Sigma)_V|^{-1}$, where $\text{Iss}(\Sigma)_V$ is the submatrix, identified by the edge set $V$ of $G$, of the *Isserlis matrix* $\text{Iss}(\Sigma)$ ([Roverato and Whittaker, 1998], [Isserlis, 1918]). If $\Gamma$ is a symmetric invertible matrix indexed by $V \times V$, the Isserlis matrix $\text{Iss}(\Gamma)$ is the symmetric matrix indexed by $W = \{(i,j) : i, j \in V, i \leq j\}$, such that

$$\{\text{Iss}(\Gamma)\}_{(i,j),(r,s)} = \gamma_{ir}\gamma_{js} + \gamma_{is}\gamma_{jr}$$ (2.19)

Hyper Inverse Wishart and G-Wishart distributions and their normalizing constant has been previously studied (see e.g. [Roverato, 2002], [Atay-Kayis and Massam, 2005]). Here we now provide a number of important properties that we need for this thesis:

1. If $G$ is decomposable graph and $C_1, C_2, \ldots, C_k, S_2, \ldots, S_k$, a perfect ordering of its cliques and corresponding separators, the density of $\Sigma^V$ can be decomposed to

$$HIW_G(\Sigma^V|\delta, D) = \frac{\prod_{i=1}^k IW(\Sigma_{C_i}|\delta, D_{C_i})}{\prod_{i=2}^k IW(\Sigma_{S_i}|\delta, D_{S_i})},$$ (2.20)

where $\Sigma_{C_i}, D_{C_i}, \Sigma_{S_i}, D_{S_i}$ are the submatrices of $\Sigma$ and $D$ with respect to the indices corresponding to $C_i$ and $S_i$ respectively. Since $C_i$ are complete subgraphs of $G$, the matrices $\Sigma_{C_i}$ are complete and their distribution is the regular Inverse Wishart distribution, $IW(\delta, D)$. Similar situation occurs for the separators $S_i$.  

2. The above proposition is generalized to any not necessarily decomposable graph $G$, when prime components are used instead of cliques. Given a perfect ordering of its
prime components $P_1, P_2, \ldots, P_k$ and corresponding separators $S_2, \ldots, S_k$, we have

$$HIW_G(\Sigma^V|\delta, D) = \prod_{i=1}^k HIW_{G_{P_i}}(\Sigma_{P_i}|\delta, D_{P_i}) \prod_{i=2}^k HIW_{G_{S_i}}(\Sigma_{S_i}|\delta, D_{S_i}), \quad (2.21)$$

where $HIW_{G_{S_i}}(\Sigma_{S_i}|\delta, D_{S_i}) = IW(\Sigma_{S_i}|\delta, D_{S_i})$ since $S_i$ are complete. Additionally, whenever $P_i$ is complete, $HIW_{G_{P_i}}(\Sigma_{P_i}|\delta, D_{P_i})$ becomes $IW(\Sigma_{P_i}|\delta, D_{P_i})$.

3. Due to the above, the normalizing constant of a graph $G$ with prime components $P_1, P_2, \ldots, P_k$ and corresponding separators $S_2, \ldots, S_k$ can be decomposed in the same fashion:

$$I_G(\delta, D) = \prod_{i=1}^k I_{G_{P_i}}(\delta, D_{P_i}) \prod_{i=2}^k I_{G_{S_i}}(\delta, D_{S_i}), \quad (2.22)$$

where the notation $I_{G_{S_i}}$ is replaced by $I_{|S_i|}$ since the normalizing constant depends only on the size $|S_i|$ of the complete subgraph $S_i$. When $P_i$ is complete, $I_{G_{P_i}}$ can also be replaced by $I_{|P_i|}$.

4. If $K$ follows the prior distribution $W_G(\delta, D)$ and given a set of data $x$ of size $n$, the posterior distribution of $K$ is $W_G(\delta^*, D^*)$, where

$$\delta^* = \delta + n$$
$$D^* = D + x^T x \quad (2.23)$$

One can see that since

$$p(K|x, G) = \frac{p(x|K, G)p(K|G)}{p(x|G)} = \frac{|K|^{n/2} \exp \left\{-\frac{1}{2}(K, x^T x)\right\} |K|^{(\delta-2)/2} \exp \left\{-\frac{1}{2}(K, D)\right\}}{(2\pi)^{np/2} p(x|G) I_G(\delta, D)} \quad (2.24)$$

$$= \frac{|K|^{(\delta+n-2)/2} \exp \left\{-\frac{1}{2}(K, D + x^T x)\right\}}{(2\pi)^{np/2} p(x|G) I_G(\delta, D)} \quad (2.25)$$

from which we can observe that the posterior density of $K$ given $x$ has the form of the $G$-Wishart with parameters $\delta^* = \delta + n$ and $D^* = D + x^T x$. Therefore the normalizing constant of the posterior distribution is $I_G(\delta^*, D^*)$ and it is often referred to as posterior normalizing constant (while the constant of the prior is referred to
as prior normalizing constant). Equation 2.23 can be used for an interpretation of the parameters of \( W_G(\delta, D) \). Scalar parameter \( \delta \) plays the role of the sample size of a hypothetical prior dataset, while matrix \( D \) plays the role of its (un-normalized) sample covariance matrix.

5. From Equation (2.26) we can see that given a dataset \( x \), the marginal likelihood of \( G \) is equal to

\[
p(x|G) = \frac{1}{(2\pi)^{np/2}} \frac{I_G(\delta^*, D^*)}{I_G(\delta, D)}
\]  

(2.27)

6. If \( \Sigma^v \sim H I W_G(\delta, D) \) for any complete subgraph \( C \), the marginal distribution of \( \Sigma_C \) is \( I W(\delta, D_C) \) ([Roverato, 2002])

### 2.3 Model Search and Selection and Covariance Estimation

Given a dataset \( x \) one is often interested in identifying one (or a number of) sufficiently plausible graphical model(s) for the data. Various methods have been previously proposed for this objective. Most of these methods aim to select graphs with higher posterior probability \( p(G|x) \). Some of those methods such as Metropolis-Hastings, Shotgun Stochastic Search (used in [Jones et al., 2005]) and Mode Oriented stochastic search (MOSS, described in [Lenkoski and Dobra, 2008]), perform a model search where at each step the posterior probability of a candidate model, \( G_{new} \), is compared with the posterior probability of the “current state” model \( G_{cur} \). This comparison is performed by the value of the ratio

\[
r_{G_{new}:G_{cur}} = \frac{p(G_{new}|x)}{p(G_{cur}|x)} = \frac{p(x|G_{new})\pi(G_{new})}{p(x|G_{cur})\pi(G_{cur})},
\]  

(2.28)

where \( \pi(G) \) denotes the “prior” probability for the graphical model \( G \). This ratio depends on the ratio of the marginal likelihoods \( p(x|G_{new})/p(x|G_{cur}) \), which as we will discuss further in Chapter 5, is equal to the Bayes Factor. Since the marginal likelihood for
a particular graphical model is the ratio of the posterior over the prior normalizing constant (multiplied by a constant, see Equation 2.27), one can conclude that the value of the normalizing constant determines in a large degree the outcome of the model search. Therefore, its calculation or estimation is essential for performing the search. Chapter 3 of this thesis is dedicated to methods of estimation of the normalizing constant (or of the ratio of the posterior over the prior).

Once one model (or a set of models) has been selected as fitting the data well, an additional question of interest is often the estimation of the covariance matrix (or its inverse, the precision matrix) or functions of it. Working in the Bayesian framework, this problem is related to the estimation of the posterior mean (or other moments) of the covariance matrix. If methods for generating exact solutions for those problems do not exist, as it is the case for non-decomposable models, estimation can be performed if a method for sampling from the posterior distribution is available. Here we focus on the problem of estimating the posterior mean of the precision matrix, \( E(K|\mathbf{x}, G) \), or of functions of it. A simple Monte Carlo estimate can be given with the use of an efficient sampling method from the \( \mathcal{G} \)-Wishart distribution. Existing and new proposed methods for sampling from \( \mathcal{W}_\mathcal{G}(\delta, D) \) are discussed in Chapter 4.
Chapter 3

Estimation of the Normalizing Constant of $W_G(\delta, D)$

3.1 Introduction - Motivation

As we already discussed in Chapter 2, many methods used for the selection between two or more graphical models for a given $p$-dimensional dataset $x$ are based on the comparison of the posterior probability $p(G|x)$. Using the Bayes Theorem we can write

$$p(G|x) = \frac{p(x|G)\pi(G)}{\sum_{g \in \mathcal{G}_p} p(x|g)\pi(g)},$$

(3.1)

where $\pi(G)$ denotes the prior probability of $G$ and $\mathcal{G}_p$ the space of all possible graphs with $p$ vertices. The quantity $p(G|x)$ can be calculated if the marginal likelihood $p(x|g)$ and prior $\pi(g)$ is known for all $g$ in $\mathcal{G}_p$. When model selection is performed between graphical models $G_1$ and $G_2$, the ratio of the posterior probabilities results to

$$\frac{p(G_1|x)}{p(G_2|x)} = \frac{p(x|G_1)\pi(G_1)}{p(x|G_2)\pi(G_2)},$$

(3.2)

and therefore the values of marginal likelihood and prior probability of only the two graphs under comparison are needed.
From this discussion it becomes apparent that the value of the marginal likelihood is necessary for calculating the posterior probability of a graph and for performing model selection. We already know from Equation 2.27 that the marginal likelihood is equal to the ratio of the posterior over the prior normalizing constant, multiplied by a constant. Therefore the value of the normalizing constant (or of the ratio of the posterior over the prior) is essential for Gaussian graphical model analysis. We have already discussed in Chapter 2 that the normalizing constant for non-decomposable graphs cannot be computed exactly. Instead, it needs to be estimated in an efficient manner. In this Chapter we first present existing approaches for the estimation of the normalizing constant. Then we describe various attempts for application of a Bayesian computation method, path sampling, for the estimation of the ratio of the normalizing constants. We close this Chapter with some numerical experiments investigating how the completion of the incomplete matrix $D^\nu$ hyperparameter of the $G$-Wishart distribution, influences the Monte Carlo estimation method.

### 3.2 Previous approaches

#### 3.2.1 Monte Carlo method

Perhaps the most popular method for the estimation of the normalizing constant of the $G$-Wishart distribution is the Monte Carlo method described in [Atay-Kayis and Massam, 2005]. This method uses a sequence of changes of variables in order to express the integral $I_G(\delta, D)$ as the expected value of a random quantity (of which we can obtain random samples), multiplied by a constant. The first bijective mapping is between the sets $M^+(G)$ and $M^\alpha_\nu(G)$, where $K \in M^+(G)$ is mapped to the $\nu$-incomplete upper-triangular matrix $\phi^\nu$, such that $K = \phi^T \phi$, where $\phi$ is the unique T-completion of $\phi^\nu$. The second bijective mapping is between the set $M^\alpha_\nu(G)$ and itself, mapping $\phi^\nu$ to $\psi^\nu$, where $\psi^\nu$ is the projection of the upper triangular matrix with positive diagonal elements $\psi$, such that
\(\psi = \phi T^{-1}\) and \(T^T T\) is the Choleski decomposition of the hyperparameter positive definite matrix \(D\). It can be proven that the non-free elements of \(\psi\), i.e. those that correspond to the missing edges of \(G\) can be determined uniquely by the free elements, i.e. those that correspond to the edges of \(G\) and those in the diagonal of \(\psi\), which proves that the mapping is bijection. The recursive mechanism that gives the values of \(\psi_{ij}, (i, j) \in \hat{V}\) from the values of \(\psi^V\) is given in equation (31) in [Atay-Kayis and Massam, 2005], which we include here for completeness reasons: if \((1, s) \in \hat{V}\),

\[
\psi_{1s} = \sum_{j=1}^{s-1} (-\psi_{1j} h_{js}), \tag{3.3}
\]

while for \((r, s) \in \hat{V}, r > 1,\)

\[
\psi_{rs} = \sum_{j=r}^{s-1} (-\psi_{rj} h_{js}) - \sum_{i=1}^{r-1} \left( \psi_{ir} + \sum_{j=i}^{r-1} \psi_{ij} h_{jr} \right) \left( \psi_{is} + \sum_{j=i}^{s-1} \psi_{ij} h_{js} \right), \tag{3.4}
\]

where \(h_{ij} = t_{ij}/t_{jj}\), \(t_{ij}\) being the \((i, j)\) entry of matrix \(T\).

After applying this change of variable to the integral

\[
I_G(\delta, D) = \int_{M^+(G)} |K|^{(\delta - 2)/2} \exp\left\{-\frac{1}{2}(K, D)\right\} dK \tag{3.5}
\]

it becomes

\[
I_G(\delta, D) = C_{\delta,D} \cdot E[f_D(\psi^V)], \tag{3.6}
\]

where

\[
f_D(\psi^V) = \exp \left( -\frac{1}{2} \sum_{(i, j) \in \hat{V}} \psi_{ij}^2 \right), \tag{3.7}
\]

and \(C_{\delta,D}\) a constant that depends on \(\delta\) and \(D\). The expected value in Equation 3.6 is calculated based on the random variable \(\psi^V\), the distribution of which is such that

\[
\psi_{ii} \sim \chi_{\delta + \nu_i}, \text{ for } i = 1, \ldots, p, \tag{3.8}
\]

and

\[
\psi_{ij} \sim N(0, 1), \text{ for } (i, j) \in V, i \neq j \tag{3.9}
\]
and the components of $\psi^V$ are pairwise independent. With $\chi_{\delta+\nu_i}$ we denote the $\chi$ distribution (square root of $\chi^2$ distribution) with $\delta + \nu_i$ degrees of freedom. $\nu_i$ is equal to the number of nodes $j$ that are connected with $i$ with an edge, and that $j > i$. The normalizing constant can then be estimated by

$$ I_G(\delta, D) = C_{\delta, D} \cdot \frac{1}{N} \sum_{i=1}^{N} f_D(\psi^V_i), $$

(3.10)

where $\psi^V_l$, $l = 1, \ldots, N$ are independent samples from $\prod_{i=1}^{p} \chi_{\delta+\nu_i} \times N_{|E|}(0_{|E|}, I_{|E|})$.

We are closing this section with a result that is directly related to definition of the normalizing constant, instead of its estimation. This result regards the constraint on the parameter $\delta$ under which $I_G(\delta, D)$ is well defined. It is given in this section because its proof uses some of the derivations from [Atay-Kayis and Massam, 2005], which were originally used for the Monte Carlo estimation method.

There has been a tradition in the literature (e.g. [Atay-Kayis and Massam, 2005], [Roverato, 2002], [Lenkoski and Dobra, 2008]) for the value of $\delta$ to be chosen to be larger than 2, following Theorem 1 of [Diaconis and Ylvisaker, 1979]. The lemma below shows that values larger than 0 are sufficient for $I_G(\delta, D) < \infty$.

**Lemma 3.2.1** The normalizing constant $I_G(\delta, D)$ is a positive real number if $D$ is positive definite and $\delta > 0$.

**Proof** The proof uses some of the derivations of [Atay-Kayis and Massam, 2005] and the changes of variables, originally used for the estimation of the normalizing constant. As we discuss in the beginning of this section, the variable $K$ is changed to the $V$-incomplete triangular matrix $\psi^V$, where $V$ denotes the set of edges of $G$. The integral then can be written

$$ I_G(\delta, D) = C_{\delta, D} \cdot \int \exp \left( -\frac{1}{2} \sum_{(i,j) \in V} \psi^2_{ij} \right) \pi(\psi^V) d\psi^V, $$

(3.11)

where $\{\psi_{ij}\}_{(i,j) \in V}$ are functions of $\psi^V$ that are given in Equations (3.3) and (3.4). Those functions depend on $G$ and $D$, but not on $\delta$. Also, the constant $C_{\delta, D}$ is always finite and
if \( \delta > 0 \) the density \( \pi(\psi^V) \), as given from Equations 3.8, 3.9, is well defined. Additionally, we have

\[
0 < \exp \left( -\frac{1}{2} \sum_{(i,j) \in \overline{V}} \psi_{ij}^2 \right) < 1
\]

and therefore from Equation 3.11

\[
0 < I_G(\delta, D) < C_{\delta,D} \cdot \int \pi(\psi^V) d\psi^V = C_{\delta,D} < \infty \tag{3.12}
\]

### 3.2.2 Laplace approximation

Assuming that \( f \) is a smooth function of a \( p \)-dimensional variable \( \omega \), maximized at \( \hat{\omega} \), with negative definite Hessian at \( \hat{\omega} \), the Laplace approximation of the integral

\[
I = \int \exp\{f(\omega)\} d\omega
\]

is given by

\[
\bar{I} = \exp\{f(\hat{\omega})\}(2\pi)^{p/2}|\Sigma(\hat{\omega})|^{1/2}, \tag{3.13}
\]

where \( \Sigma(\hat{\omega}) = -(D^2f(\hat{\omega}))^{-1} \), is minus the inverse of the Hessian of \( f \) evaluated at \( \hat{\omega} \) [Tierney and Kadane, 1986]. This can be easily seen if we apply the Taylor expansion of \( f(\omega) \) around \( \hat{\omega} \). Keeping only the first three terms the Taylor expansion gives the approximation:

\[
f(\omega) \approx f(\hat{\omega}) + (\omega - \hat{\omega})^T Df(\hat{\omega}) + \frac{1}{2}(\omega - \hat{\omega})^T D^2f(\hat{\omega})(\omega - \hat{\omega}), \tag{3.14}
\]

where \( Df \) is the vector of first partial derivatives of \( f \) (the gradient) and \( D^2f \) is the matrix of second partial derivatives of \( f \) (the Hessian). Since \( \hat{\omega} \) maximizes \( f \), \( Df(\hat{\omega}) = 0 \), so Equation 3.14 becomes

\[
f(\omega) \approx f(\hat{\omega}) + \frac{1}{2}(\omega - \hat{\omega})^T D^2f(\hat{\omega})(\omega - \hat{\omega}), \tag{3.15}
\]
Using Equation 3.15 an approximation of the integral $I$ is

$$
\tilde{I} = \exp\{f(\hat{\omega})\} \int \exp \left\{ -\frac{1}{2}(\omega - \hat{\omega})^T (-D^2 f(\hat{\omega})) (\omega - \hat{\omega}) \right\} d\omega
$$

(3.16)

If we call $\Sigma(\hat{\omega}) = -(D^2 f(\hat{\omega}))^{-1}$, from the density of the $p$-variate normal distribution, we have that

$$
\int \exp \left\{ -\frac{1}{2}(\omega - \hat{\omega})^T (-D^2 f(\hat{\omega})) (\omega - \hat{\omega}) \right\} d\omega = (2\pi)^{p/2} |\Sigma(\hat{\omega})|^{1/2}
$$

(3.17)

This method can be used to approximate the normalizing constant $I_G(\delta, D)$, when $\delta \neq 2$, by first writing it as

$$
\int_{M^+(G)} \exp\{f_{\delta,D}(K)\} dK,
$$

where $f_{\delta,D}(K) = -\frac{1}{2}\{(2-\delta) \log |K| + \langle K, D \rangle \}$ ([Lenkoski and Dobra, 2008]). The Laplace approximation then becomes

$$
\hat{I}_G(\delta, D) = \exp\{f_{\delta,D}(\hat{K})\} (2\pi)^{|V|/2} |F_{\delta,D}(\hat{K})|^{1/2},
$$

(3.18)

where $\hat{K}$ is the mode of $W_G(\delta, D)$, $V$ denotes the set of ordered pairs-edges of $G$, and $F_{\delta,D}(\hat{K})$ is minus the inverse of the Hessian matrix associated with $f_{\delta,D}$, evaluated at $\hat{K}$. To be noted here that this method needs the value of the mode $\hat{K}$ of $W_G(\delta, D)$. When $G$ is decomposable, $\hat{K}$ can be calculately exactly ([Lauritzen, 1996], [Lenkoski and Dobra, 2010]). When $G$ is non-decomposable an exact solution does not exist but it can be estimated using the Iterative Proportional Scaling algorithm (IPS) described in [Speed and Kiiveri, 1986]. See also [Lenkoski and Dobra, 2010].

### 3.3 New proposed approaches

#### 3.3.1 Using estimation of the marginal likelihood

In this section we present a method for the estimation of the marginal likelihood $p(x|G)$. If needed, the ratio of the “posterior” over the “prior” normalizing constant can then be
Chapter 3. Estimation of the Normalizing Constant of $W_G(\delta, D)$

estimated, using the equation (2.27). This method requires the ability to sample from the $G$-Wishart distribution. We already know that the marginal likelihood of graph $G$, given a $p$-dimensional dataset $X$ of size $n$, is equal to

$$p(X|\delta, D, G) = \int_{K \in M^+(G)} p(X|K, G, \delta, D)p(K|G, \delta, D)dK$$

$$= E_{G,\delta,D}(L(K; X, G, \delta, D)),$$  \hspace{1cm} (3.19)  \hspace{1cm} (3.20)

where the last expected value is with respect to $W_G(\delta, D)$, and $L(K; X, G, \delta, D)$ is the likelihood of $K$ given the data $X$ and $G, \delta$. Since we are assuming a Gaussian Graphical Model, we have that

$$L(K; X, G, \delta, D) = \frac{|K|^{n/2}}{(2\pi)^{np/2}} \exp\{-\frac{1}{2} \langle K, X^TX \rangle\}.$$  \hspace{1cm} (3.21)

Using (3.20) we can estimate the marginal likelihood by the Monte Carlo estimate

$$\hat{p}(X|\delta, D, G) = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{|K^{(i)}|^{n/2}}{(2\pi)^{np/2}} \exp\{-\frac{1}{2} \langle K^{(i)}, X^TX \rangle\}\right]$$

where $K^{(i)}$ are random draws from $W_G(\delta, D)$ (methods of sampling from $W_G(\delta, D)$ without the exact value of the normalizing constant are presented in Chapter 4).

Finally, if $r$ denotes the ratio $I_G(\delta + n, D + X^TX)/I_G(\delta, D)$, from (2.27) we have

$$\frac{I_G(\delta + n, D + X^TX)}{I_G(\delta, D)} = (2\pi)^{np/2}p(X|\delta, D, G),$$  \hspace{1cm} (3.22)

and from Equation (3.21), we have the estimate

$$\hat{r} = \frac{1}{m} \sum_{i=1}^{m} \left[|K^{(i)}|^{n/2} \exp\{-\frac{1}{2} \langle K^{(i)}, X^TX \rangle\}\right]$$  \hspace{1cm} (3.23)

3.3.2 Using Path Sampling

General

*Path sampling* has been proposed ([Gelman and Meng, 1998], [Chen et al., 2000]) as a method of estimation of the ratio of normalizing constants. Given a random quantity
ω the distribution of which depends on a parameter θ, the density of ω can be written as $p(\omega|\theta) = q(\omega|\theta)/z(\theta)$, where $z(\theta)$ denotes the normalizing constant (depending on θ) and $q(\omega|\theta)$ the unnormalized density of ω, given θ. Without loss of generality one can assume that $\theta \in [0, 1]$ and that we are interested in $r = z(1)/z(0)$. We have that

$$\lambda = \log(r) = \int_0^1 \frac{d}{d\theta} \log z(\theta) d\theta \quad (3.24)$$

But we also have

$$\frac{d}{d\theta} \log z(\theta) = \frac{1}{z(\theta)} \frac{dz(\theta)}{d\theta} = \frac{1}{z(\theta)} \frac{d}{d\theta} \int q(\omega|\theta) \mu(d\omega) \quad (3.25)$$

and assuming that we can interchange integration with differentiation the above becomes

$$\int \frac{1}{z(\theta)} \frac{d}{d\theta} q(\omega|\theta) \mu(d\omega) = \int \frac{1}{q(\omega|\theta)} \left( \frac{d}{d\theta} q(\omega|\theta) \right) p(\omega|\theta) \mu(d\omega) = E_\theta \left[ \frac{d}{d\theta} \log q(\omega|\theta) \right] \quad (3.26)$$

By setting

$$U(\omega, \theta) = \frac{d}{d\theta} \log q(\omega|\theta) \quad (3.27)$$

equation (3.26) leads to

$$\lambda = \int_0^1 E_\theta[U(\omega, \theta)] d\theta \quad (3.28)$$

Let now the general case where θ in $[\theta_0, \theta_1]$ and we are interested in $r = z(\theta_1)/z(\theta_0)$. Given a continuous path $\theta(t) : [0, 1] \mapsto [\theta_0, \theta_1]$, such that $\theta(0) = \theta_0$ and $\theta(1) = \theta_1$, we have that
\[ \lambda = \log(r) = \int_{\theta_0}^{\theta_1} E_\theta \left[ \frac{d}{d\theta} \log q(\omega|\theta) \right] d\theta \]  
(3.29)

\[ = \int_0^1 E_{\theta(t)} \left[ \theta'(t) \frac{d}{d\theta} \log q(\omega|\theta(t)) \right] dt \]  
(3.30)

\[ = \int_0^1 E_{\theta(t)}[\theta'(t)U(\omega, \theta(t))] dt \]  
(3.31)

\[ = \int_0^1 \left( \int \theta'(t)U(\omega, \theta(t))p(\omega|t)d\omega(t) \right) dt, \]  
(3.32)

where \( \theta'(t) \) the derivative of \( \theta(t) \) with respect to \( t \). If we put a uniform prior \( \pi \) on \( t \in [0,1] \), we have \( p(\omega, t) = p(\omega|t)\pi(t) = p(\omega|t) \), so the equation 3.32 becomes

\[ \lambda = \int_0^1 \int \theta'(t)U(\omega, \theta(t))p(\omega, t)d\omega dt, \]  
(3.33)

so \( \lambda \) is equal to the expected value of \( \frac{d\theta}{dt}U(\omega, \theta(t)) \), with respect to the joint distribution of \( (\omega, t) \). One can then use a Monte Carlo estimate for \( \lambda \),

\[ \hat{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \theta'(t_i)U(\omega_i, \theta(t_i)) \]  
(3.34)

where \( t_i \) are sampled randomly from the \( Uniform([0,1]) \) and \( \omega_i \) are sampled by \( p(\omega|\theta(t_i)) \). The estimate and its variance depends on the choice of the path \( \theta(t) \).

We can extend this method to the multivariate case, were

\[ \theta(t) = (\theta_1(t), \theta_2(t), \ldots, \theta_d(t)). \]

By denoting

\[ \theta'_i(t) = \frac{d\theta_i(t)}{dt}, \]

and

\[ U_k(\omega, \theta) = \frac{\partial}{\partial \theta_k} \log q(\omega|\theta) \]

equation (3.32) becomes

\[ \lambda = \int_0^1 E_{\theta(t)} \left[ \sum_{k=1}^{d} \theta'_k(t)U_k(\omega, \theta(t)) \right] dt \]  
(3.35)
Chapter 3. Estimation of the Normalizing Constant of $W_G(\delta, D)$

Respectively, the estimate from equation (3.34) becomes

$$\hat{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \left[ \sum_{k=1}^{d} \theta'_k(t_i) U_k(\omega_i, \theta(t_i)) \right], \quad (3.36)$$

where, as in the univariate case, the $t_s$’s are sampled randomly from $Uniform([0, 1])$ and $\omega_i$ are sampled by $p(\omega | \theta(t_i))$.

The implementation and approach presented above can be used mainly for the estimation of the ratio of normalizing constants corresponding to two different parameter values ($\theta_1$ and $\theta_2$ denoted above) when sampling from the density $p(\omega | \theta)$ is feasible. If this is not the case, path sampling can still be used but with different implementation. See [Gelman and Meng, 1998] for details.

Application to GGM - Traditional path sampling approach

As mentioned above, graph model selection methods discussed here are based on the maximization of the posterior density of the graph, $p(G|X = x)$, which is proportional to $p(X|G) \pi(G)$. It is therefore of high interest to estimate $p(X|G)$ (sometimes called marginal likelihood). In this section we present an application of path sampling for the estimation of the ratio of normalizing constants $I_G(\delta^*, D^*)/I_G(\delta, D)$ where $\delta^* = \delta + n$, $D^* = D + U$. We consider the path $\theta(t) = (\delta(t), D(t)) = (\delta + tn, D + tU)$, where $t \in [0, 1]$. It is obvious that the normalizing constant is well defined for all $t \in [0, 1]$. This is because for every $t \in [0, 1]$ it is $\delta + tn > 0$ and $D + tU$ is a symmetric positive definite matrix. We have that

$$\frac{d\delta(t)}{dt} = n \quad (3.37)$$

$$\frac{dD_{kl}(t)}{dt} = U_{kl} \quad (3.38)$$

We also have that

$$U_\delta(K, \delta, D) = \frac{\partial \log(|K|^{(\delta-2)/2} \exp\left\{-\frac{1}{2} \langle K, D \rangle\right\})}{\partial \delta}$$

$$= \frac{\partial \frac{1}{2} \log |K| - \frac{1}{2} \langle K, D \rangle}{\partial \delta} = \frac{\log |K|}{2}, \quad (3.40)$$
and

\[ U_{Dkl}(K, \delta, D) = \frac{\partial \log(|K|^{(\delta-2)/2} \exp\{-\frac{1}{2}(K, D)\})}{\partial D_{kl}} = -K_{kl}, \tag{3.41} \]

where \( k < l, (k, l) \in V \) (the set of edges of \( G \)) and \( U_{Dkk}(K, \delta, D) = -\frac{1}{2}K_{kk}. \) Applying these formulas to (3.36) we get

\[ \hat{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \left[ n \frac{\log |K(i)|}{2} - \sum_{k<l} U_{kl} K_{kl}^{(i)} - \frac{1}{2} \sum_{k=1}^{p} U_{kk} K_{kk}^{(i)} \right], \tag{3.42} \]

where \( K^{(i)} \) denotes a draw from \( W_G(\delta + t_i n, D + t_i U) \), given that \( t_i \) is a random draw from \( Uniform([0, 1]) \). As it was mentioned in the previous subsection, this method requires sampling from the G-Wishart distribution with respect to graph \( G \).

**Application to GGM - Path sampling and Monte Carlo hybrid approach**

Here we present a different approach for the estimation of the ratio of normalizing constants. This approach imitates up to some degree the development followed by the path sampling method, and it is motivated by the expression of \( I_G(\delta, D) \) presented in [Atay-Kayis and Massam, 2005]. According to that paper, the normalizing constant is equal to a constant \( C_{\delta, D} \) multiplied by the expected value of \( f_D(y) \) where \( y \) is a random vector of known density depending on \( \delta \), and \( f_D \) is a real function depending only on \( D \). It is interesting to notice here that, while \( y \) depends only on \( \delta \), the function \( f_D \) depends only on \( D \) (see section 3.2.1 and Equation 3.7). We can then write

\[ I_G(\delta, D) = C_{\delta, D} \cdot \int f_D(y)p(y|\delta)dy \tag{3.43} \]

Our goal is the estimation of the ratio \( I_G(\delta^*, D^*)/I_G(\delta, D) \). Since the constant \( C_{\delta, D} \) is computable for every \((\delta, D)\), we just need to estimate the ratio \( \int f_{D^*}(y)p(y|\delta^*)dy/ \int f_D(y)p(y|\delta)dy \).

Before we tackle this specific problem, we proceed with a development for a more general context: given a density \( p(x|\theta) \), where \( \theta \in \mathbb{R}^d \), and a continuous function in \( \mathbb{R}^d, f(x, \theta) \), we are interested in estimating the ratio

\[ r = \frac{\int f_{\theta_1}(x)p(x|\theta_1)dx}{\int f_{\theta_0}(x)p(x|\theta_0)dx}. \]
Chapter 3. Estimation of the Normalizing Constant of \( W_G(\delta, D) \)

If we call \( z(\theta) = \int f(x, \theta)p(x|\theta)dx \), the logarithm of \( r, \lambda \), is equal to

\[
\lambda = \int_0^1 \frac{d}{dt} \log z(\theta(t)) dt, \tag{3.44}
\]

where \( \theta(t) \) a continuous path from \([0, 1]\) to \([\theta_0, \theta_1]\). Expanding (3.44) we get

\[
\lambda = \int_0^1 \frac{1}{z(\theta(t))} \frac{dz(\theta(t))}{dt} dt \tag{3.45}
\]

\[
= \int_0^1 \frac{1}{z(\theta(t))} \left( \sum_{i=1}^d \frac{\partial z(\theta(t))}{\partial \theta_i} \frac{d\theta_i(t)}{dt} \right) dt \tag{3.46}
\]

For the partial derivatives of \( z(\theta) \) we have:

\[
\frac{\partial z(\theta)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \left[ \int f(x, \theta)p(x|\theta)dx \right] \tag{3.47}
\]

\[
= \int \frac{\partial f(x, \theta)}{\partial \theta_i} p(x|\theta) + f(x, \theta) \frac{\partial p(x|\theta)}{\partial \theta_i} dx \tag{3.48}
\]

\[
= \int \left[ \frac{\partial \log f(x, \theta)}{\partial \theta_i} + \frac{\partial \log p(x|\theta)}{\partial \theta_i} \right] f(x, \theta)p(x|\theta)dx \tag{3.49}
\]

By combining this result with (3.46) and assuming a \([0, 1]\)-uniform prior for \( t \), we end up with

\[
\lambda = \int_0^1 \frac{1}{z(\theta(t))} \left\{ \sum_{i=1}^d \left( \int \left[ \frac{\partial \log f(x, \theta)}{\partial \theta_i} + \frac{\partial \log p(x|\theta)}{\partial \theta_i} \right] f(x, \theta)p(x|\theta)dx \right) \frac{d\theta_i(t)}{dt} \right\} dt \tag{3.50}
\]

\[
= \int_0^1 \int \frac{1}{z(\theta(t))} \sum_{i=1}^d \frac{d\theta_i(t)}{dt} \left( \frac{\partial \log f(x, \theta)}{\partial \theta_i} + \frac{\partial \log p(x|\theta)}{\partial \theta_i} \right) f(x, \theta)p(x|\theta)dx dt
\]

At this point we can choose different directions depending on the implementation. We also notice that, since \( t \sim Uniform([0, 1]) \) and \( \theta(\cdot) \) is a bijection, \( p(x|\theta) = p(x|t) = p(x,t) \).
**Method 1**

One way, and using the notation

\[ G(x, \theta) = \sum_{i=1}^{d} \frac{d\theta_i(t)}{dt} \left( \frac{\partial \log f(x, \theta)}{\partial \theta_i} + \frac{\partial \log p(x|\theta)}{\partial \theta_i} \right), \]  

(3.51)

is to write

\[ \lambda = E_{x,t} \left[ \frac{1}{z(\theta(t))} f(x, \theta) G(x, \theta) \right], \]  

(3.52)

where the last expected value is with respect to the joint distribution of \( t \) and \( x \).

We can now use a Monte Carlo estimate of this expected value,

\[ \hat{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \theta'_j(t_i) f_j(x_i, \theta(t_i)) \left[ f_j(x_i, \theta(t_i)) + \ell_j(x_i, \theta(t_i)) \right] \frac{1}{m_i} \sum_{k=1}^{m_i} f_j(x_i, \theta(t_i)), \]  

(3.53)

where

\[ \theta'_j(t) = \frac{d\theta_j(t)}{dt}, \]

\[ f_j(x, \theta) = \frac{\partial \log f(x, \theta)}{\partial \theta_j} \]

and

\[ \ell_j(x, \theta) = \frac{\partial \log p(x|\theta)}{\partial \theta_j}, \]

and where \( t_i, i = 1, \ldots, N \) are independent samples from \( Uniform([0, 1]) \) and for a given \( i, x_i, x_i^k, k = 1, \ldots, m_i \) are independent samples from \( p(x|\theta(t_i)) \).

**Method 2**

We can also use a different implementation route, not using the joint density of \( x \) and \( t \).

We can rewrite equation (3.50) as

\[ \lambda = \int_0^1 \frac{1}{z(\theta(t))} \left[ \int G(x, \theta) f(x, \theta)p(x|\theta)dx \right] dt, \]  

(3.54)

which can be written as

\[ \lambda = E_t \left[ \frac{E[f(x, \theta(t))G(x, \theta(t))|t]}{E[f(x, \theta(t))|t]} \right] \]  

(3.55)
In order to make sure that this quantity is a real number we have to make sure that both
\( E[f(x, \theta(t))G(x, \theta(t))|t] \) and \( E[f(x, \theta(t))|t] \) are real numbers for every \( 0 \leq t \leq 1 \). The
quantity \( E[f(x, \theta(t))|t] \) is proportional to the normalizing constant of \( W_G(\delta(t), D(t)) \), de-
noted by \( I_G(\delta(t), D(t)) \), and therefore it always exists. The quantity \( E[f(x, \theta(t))G(x, \theta(t))|t] \)
is equal to the derivative of \( I_G(\delta(t), D(t)) \) with respect to \( t \).

We now apply this method for the estimation of the ratio
\[
 r = \frac{\int f_D^*(y)p(y|\delta^*)dy}{\int f_D(y)p(y|\delta)dy}
\]
for a given graph \( G \). For reasons that will become clearer below, we substitute the
hyperparameter \( D \) by the upper triangular matrix \( H = (h_{ij})_{1 \leq i < j \leq p} \), with every diagonal
element equal to 1. This matrix is determined by the upper triangular matrix \( T =
(t_{ij})_{1 \leq i \leq j \leq p} \), for which \( D^{-1} = T^T T \), by the equation
\[
 h_{ij} = t_{ij}/t_{jj}.
\]

Additionally, for consistency with the development in the general context, we substitute
the notation \( f_H(y) \) by \( f(y, H) \). We recall here that \( y = (\psi_{(i,j)})_{(i,j) \in \mathcal{V}} = \psi^\mathcal{V} \), where \( \mathcal{V} \) are the
lexicographically ordered vertex-pairs corresponding to the edges of \( G \) and same-vertex
pairs. We also recall that
\[
f_D(\psi^\mathcal{V}) = \exp\{-\frac{1}{2} \sum_{(i,j) \in \mathcal{V}} \psi_{ij}^2\},
\]
where each one of \( \psi_{ij}, (i, j) \in \mathcal{V} \) is a function of \( \psi^\mathcal{V} \) that depends on \( H \). We now
denote with \( \theta \) the vector \((\delta, H)\), and with \( H^* = (h^*_{ij})_{1 \leq i \leq j \leq p} \) the upper triangular matrix
concerning to \( D+U \). We define the path \( \theta(t) = (\theta_\delta(t), \theta_H(t)) = (\delta+tn, H+t(H^*-H)) \),
with \( 0 \leq t \leq 1 \). It is clear that \( \theta(0) = (\delta, H), \theta(1) = (\delta^*, H^*) \). We now have the following
identities:
Chapter 3. Estimation of the Normalizing Constant of $W_G(\delta, D)$

\[ \theta'_\delta = \frac{d\theta_i}{dt} = n \]  
\[ \theta'_{h_{ij}} = \frac{dh_{ij}(t)}{dt} = h_{ij}^* - h_{ij} \]  
\[ f_\delta(\psi^V, H) = \frac{\partial f(\psi^V, H)}{\partial \delta} = 0 \]  
\[ (3.56) \]  
\[ (3.57) \]  
\[ (3.58) \]  
\[ (3.59) \]

If we use the notation

\[ f_{ij}(\psi^V, H) = \frac{\partial \log f(\psi^V, H)}{\partial h_{ij}} \]  
\[ (3.60) \]

and following equation (3.50), the logarithm of $r$ is given by

\[
\lambda = E_{\psi^V, t} \left[ \frac{f(\psi^V, H)}{z(\theta(t))} \left\{ \theta'_\delta(t) \left( \frac{\partial \log f(\psi^V, H)}{\partial \delta} + \frac{\partial \log p(\psi^V|G, \delta, H)}{\partial \delta} \right) + \sum_{1 \leq i < j \leq p} \theta'_{h_{ij}} \left( f_{ij}(\psi^V, H) + \frac{\partial \log p(\psi^V|D, \delta, H)}{\partial h_{ij}} \right) \right\} \right]
\]
\[ = E_{\psi^V, t} \left[ \frac{f(\psi^V, H)}{z(\theta(t))} \left\{ \theta'_\delta(t) \frac{\partial \log p(\psi^V|G, \delta, H)}{\partial \delta} + \sum_{1 \leq i < j \leq p} \theta'_{h_{ij}} f_{ij}(\psi^V, H) \right\} \right] \]
\[ (3.61) \]

since the density of $\psi^V$ depends only on $\delta$ and not on $H$, and therefore

\[ \frac{\partial \log p(\psi^V|G, \delta, H)}{\partial h_{ij}} = 0. \]

For each $i, j$ where $1 \leq i < j \leq p$, we have that

\[ f_{ij}(\psi^V, H) = - \sum_{(k,l) \in \bar{V}} \psi_{kl} \frac{\partial \psi_{kl}}{\partial h_{ij}}. \]  
\[ (3.62) \]

We recall here that $\psi_{kl}$ are functions of $\psi^V$ that depend on $H$. We also need to note here that the partial derivatives $\frac{\partial \psi_{kl}}{\partial h_{ij}}$ can be computed recursively using the equations (3.3) and (3.4), as follows: for $(1, t) \in \bar{V}$,

\[ \frac{\partial \psi_{1s}}{\partial h_{ij}} = - \sum_{k=1}^{s-1} \left( \frac{\partial \psi_{1k}}{\partial h_{ij}} h_{ks} + \psi_{1k} I(k = i, s = j) \right), \]  
\[ (3.63) \]
while for the more general case with \((r, s) \in \bar{V}, r > 1,\)

\[
\frac{\partial \psi_{rs}}{\partial h_{ij}} = - \sum_{k=r}^{s-1} \left( \frac{\partial \psi_{rk}}{\partial h_{ij}} h_{ks} + \psi_{rk} I(k = i, s = j) \right)
\]

\[-\frac{1}{\psi_{rr}} \sum_{i=1}^{r-1} \left\{ \left( \frac{\partial \psi_{ir}}{\partial h_{ij}} + \sum_{k=l}^{r-1} \left( \frac{\partial \psi_{lk}}{\partial h_{ij}} h_{kr} + \psi_{lk} I(k = i, r = j) \right) \right) \left[ \psi_{ts} + \sum_{k=l}^{s-1} \psi_{lk} h_{ks} \right] \]

\[+ \left( \frac{\partial \psi_{ts}}{\partial h_{ij}} + \sum_{k=l}^{s-1} \left( \frac{\partial \psi_{lk}}{\partial h_{ij}} h_{ks} + \psi_{lk} I(k = i, s = j) \right) \right) \left[ \psi_{tr} + \sum_{k=l}^{r-1} \psi_{lk} h_{kr} \right] \} (3.64)
\]

Since \(\psi^V = (\psi_{ii})_{1 \leq i \leq p}, (\psi_{ij})_{(i,j) \in \bar{V}, i \neq j},\)

where \(\psi_{ij} \sim N(0, 1),\) for \((i, j) \in \bar{V}, i \neq j,\) and

\(\psi_{ii} \sim \sqrt{\frac{2}{\chi^2_{p+1}}}, i = 1, \ldots, p,\) and \((\psi_{ii})_{1 \leq i \leq p}, (\psi_{ij})_{(i,j) \in \bar{V}, i \neq j}\)

are pairwise independent, the density of \(\psi^V\) is given by

\[
p(\psi^V|G, \delta) = p((\psi_{ii})_{1 \leq i \leq p}, (\psi_{ij})_{(i,j) \in \bar{V}, i \neq j})|G, \delta) (3.65)
\]

\[= \prod_{i=1}^{p} \frac{1}{2^{(\delta+\nu_i)/2-1}} \frac{1}{\Gamma[(\delta + \nu_i)/2]} \psi_{ii}^{\delta+\nu_i-1} \exp(-\frac{1}{2} \psi_{ii}^2) \cdot \prod_{(i,j) \in \bar{V}, i \neq j} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} \psi_{ij}^2) (3.66)
\]

Therefore we have that

\[
\frac{\partial \log p(\psi^V|G, \delta)}{\partial \delta} = \frac{1}{2} \sum_{i=1}^{p} \left[ \log \left( \frac{\psi_{ii}}{2} \right) - \Psi \left( \frac{\delta + \nu_i}{2} \right) \right],
\]

where \(\Psi\) denotes the digamma function, the derivative of \(\log \Gamma\) function.

We can then use (3.53) and (3.64) for the estimator of \(\lambda\) as follows:

\[
\hat{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\psi^V(i), T(t_i))}{z(\bar{\theta}(t_i))} \left\{ \frac{n}{2} \sum_{k=1}^{p} \left[ \log \left( \frac{\psi_{kk}}{2} \right) - \Psi \left( \frac{\delta + t_i n + \nu_k}{2} \right) \right] + \sum_{1 \leq k < l \leq p} (h_{kl}^* - h_{kl}) f_{kl}(\psi^V(i), T(t_i)) \right\},
\]

where \(t_i, i = 1, \ldots, N\) are independent samples from \(Uniform([0, 1]), \psi^V(i)\) are samples from \(p(\psi^V|G, \delta(t_i), T(t_i)),\) the values of \(f_{kl}(\psi^V(i), T(t_i))\) can be computed by equation (3.62) and, for each \(t_i\) the estimate \(z(\bar{\theta}(t_i))\) can be given by

\[
\frac{1}{m_i} \sum_{k=1}^{m_i} f \left( [\psi^V]^k(i), T(t_i) \right),
\]
where, for a given $i$, $[\psi^{\mathcal{V}}]^{k}(i), k = 1, \ldots, m_{i}$ are independent samples from $p(\psi^{\mathcal{V}}|\delta(t_{i}), T(t_{i}))$.

A computationally more expensive but potentially more stable implementation is the one based on expression (3.55). If we use the notation

$$G(\psi^{\mathcal{V}}(i), T(t_{i})) = \frac{n}{2} \sum_{k=1}^{p} \log \left( \frac{\psi_{kk}}{2} \right) - \Psi \left( \frac{\delta + t_{i} \eta + \nu_{k}}{2} \right) + \sum_{1 \leq k < l \leq p} (h_{kl}^{*} - h_{kl}) f_{kl}(\psi^{\mathcal{V}}(i), T(t_{i})), $$

we can draw for each $t_{i}$ a set of $m_{i}$ random samples from $p(\psi^{\mathcal{V}}|t_{i})$ used for the estimation of $z(\theta(t_{i}))$ and another set of $m'_{i}$ independent samples for the estimation of $E_{t_{i}}[G(\psi^{\mathcal{V}}, T(t_{i})) f(\psi^{\mathcal{V}}(i), T(t_{i}))].$

Method 3

An alternative way of estimating the quantity in equation (3.54) is by treating the integral

$$z(\theta) = \int f(x, \theta)p(x|\theta)dx$$

as the normalizing constant of the density function

$$p^{*}(x|\theta) = \frac{f(x, \theta)p(x|\theta)}{z(\theta)}.$$ 

In that case we have

$$\lambda = \int_{0}^{1} \int G(x, t)p^{*}(x|t)dxdt \quad (3.69)$$

We can then follow a direct sampling approach (see [Gelman and Meng, 1998]), where the first step is to sample $t$ from $\text{Uniform}([0, 1])$. Given $t$, sampling from $p^{*}(x|t)$ is not straightforward since we do not have an exact expression of the density or other way to sample. On the other hand, since we know the unnormalized density $q^{*}(x|t) = f(x|t)p(x|t)$ we can employ an iterative simulation approach, applying for instance the Metropolis-Hastings algorithm for the generation of random samples from $p^{*}(x|t)$ given $t$, in a series of nested loops, like for instance:
Chapter 3. Estimation of the Normalizing Constant of $W_G(\delta, D)$

- For $i = 1, \ldots, n_1$
  - Sample $t_i \sim Uniform([0, 1])$
  - For $j = 1, \ldots, n_2$
    - Sample $x_{ij} \sim p^*(x|t_i)$
  
- Output sample $\{x_{ij}\}_{ij}$

Then we can use the estimate

$$
\hat{\lambda} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} G(x_{ij}, t_i).
$$

(3.70)

3.3.3 Numerical results

In this subsection the results from an experimental application of the path sampling-based method are presented and compared with the results from Monte Carlo method.

The objective of the experiment was the estimation of the log ratio of the normalizing constants (posterior over prior), denoted with $\lambda$, for all possible graphical models for a given data set. For the experiment, Iris flower data set ([Fisher, 1936]) was used. This 4-dimensional data set contains the measurements of the length and width of sepals and petals from 150 samples of Iris flowers. The samples are taken equally from three species of the flower, Iris setosa, Iris virginica and Iris versicolor. For our experiment only the 50 samples of Iris virginica were used. For simplicity the values of $D = I_4, \delta = 3$ were used for the estimation. The estimates from both methods were also compared with the true values, which are known for a decomposable graph. Since $p = 4$, there are $2^{p(p-1)/2} = 2^6 = 64$ possible four-node graphical models. From those, only 3 models were non-decomposable, corresponding to 4-cycle graphs. The no-edge model was excluded from the experiment. For the path sampling-based method, the implementation based on Equation 3.68 was used, with values $N = 100, m_i = 100, i = 1, \ldots, 100$. Also, for the Monte Carlo method 5000 iterations were used for each estimation of the prior and
posterior normalizing constants. In addition to the estimate of the log ratio of the
normalizing constants (posterior over prior), the standard error was also calculated. For
this, the multivariate Delta method was used ([Knight, 1999]): let $X, Y$ random variables
such that $\tilde{I}_G(\delta, D) = \tilde{Y}, \tilde{I}_G(\delta^*, D^*) = \tilde{X}$. Then by the bivariate Central Limit Theorem
we have
\[
\sqrt{n} \left( \begin{pmatrix} \tilde{X}_n \\ \tilde{Y}_n \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \right) \rightarrow Z \sim N_2(\mathbf{0}, C),
\]
where $\mu_X = E(X), \mu_Y = E(Y)$ and
\[
C = \begin{pmatrix} \sigma_X^2 & \sigma_{YX} \\ \sigma_{YX} & \sigma_Y^2 \end{pmatrix}
\]
is the covariance matrix of $(X, Y)$. The application of the Delta method gives
\[
\sqrt{n} \left( \log \frac{\tilde{X}_n}{\tilde{Y}_n} - \log \frac{\mu_X}{\mu_Y} \right) \rightarrow N(\mathbf{0}, D(\mu_X, \mu_Y)^T C D(\mu_X, \mu_Y)),
\]
where $D(\mu_X, \mu_Y)$ is the matrix of the partial derivatives of $g(x, y) = \log x - \log y$, evaluated at $(x, y) = (\mu_X, \mu_Y)$, and which is equal to
\[
D(\mu_X, \mu_Y) = \begin{pmatrix} 1/\mu_X \\ -1/\mu_Y \end{pmatrix}
\]
Since
\[
D(\mu_X, \mu_Y)^T C D(\mu_X, \mu_Y) = \frac{\sigma_X^2}{\mu_X^2} - 2 \frac{\sigma_{XY}}{\mu_X \mu_Y} + \frac{\sigma_Y^2}{\mu_Y^2}
\]
the standard error of $\hat{\lambda} = \log r = \log \left( \frac{\tilde{I}_G(\delta^*, D^*)}{\tilde{I}_G(\delta, D)} \right)$ is equal to
\[
se(\hat{\lambda}) = \sqrt{\frac{1}{n} \left( \frac{\sigma_X^2}{\mu_X^2} - 2 \frac{\sigma_{XY}}{\mu_X \mu_Y} + \frac{\sigma_Y^2}{\mu_Y^2} \right)},
\]
and it can be estimated by replacing $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY}$ by their sample estimates.

Table 3.1 presents the results from the experiment. For some graphs the standard
error of the Monte Carlo based estimate of the log ratio was not estimable due to the
very small values of the estimates of $\mu_X, \mu_Y$ and the very high values of $\sigma_X^2, \sigma_Y^2, \sigma_{XY}$. 

From the table one can immediately see that the path sampling-based method in this implementation does not perform well in comparison to the Monte Carlo method. In order to assess the performance of the Monte Carlo method, we focus on graphs with a reasonable fit to the data (i.e. of relatively large marginal likelihood values) and we arbitrary decided to select graphs with true $\lambda$ value larger than -100. As we can see in Table 3.1, Monte Carlo method gives a very poor estimate for four of those selected graphs (28, 29, 32, 51). For the rest, the Monte Carlo estimates are of the same order of magnitude with the true values, as the scatterplot in Figure 3.1 shows.

![Figure 3.1: Scatterplot between true values of log ratio of normalizing constants and their Monte Carlo estimate](image)

In order to investigate further the performance of the path sampling-based method we compare the rankings of the selected graphs based on the true value and the path
sampling estimate. Similar rankings would indicate that model selection based on the estimates may be accurate, even though the actual estimate are not close to the true values. Unfortunately this is not the case from the results of our experiments and there does not seem to be a trend or relationship between the two rankings, as the plot in Figure 3.2 show.

We close this section with the conclusion that our experiments show that the proposed path-sampling based method did not offer any improvement for the estimation of the ratio of normalizing constants, over previously proposed Monte Carlo method. Further investigation and potentially different implementation is needed in order to fully explore the usefulness of path sampling for the estimation of the ratio of normalizing constants.

### 3.4 Monte Carlo estimation method and matrix $D$

There has been some controversy on the exact form of the matrix $D$, the “scale” hyperparameter of the $G$-Wishart distribution. Although using the same source ([Diaconis and Ylvisaker, 1979]), some authors ([Atay-Kayis and Massam, 2005], [Lenkoski and Dobra, 2008]) indicate that $D^{-1}$ is sufficient to belong to $M^+(G)$ in order for $G$-Wishart to be well defined, while others ([Roverato, 2002]) indicate that the true hyperparameter of the $G$-Wishart distribution is not the matrix $D$, but instead the incomplete matrix $D^\mathcal{V}$ (as described in section 2.1.2), where $\mathcal{V}$ are the edges of $G$. The latter approach seems to also be consistent with the following proposition:

**Proposition 3.4.1** Let $G$ graph. Let $D$ and $D'$ positive definite matrices such that $D^\mathcal{V} = (D')^\mathcal{V}$ and $\delta > 0$. Then for any matrix $K \in M^+(G)$ we have that $f(K|G, \delta, D) = f(K|G, \delta, D')$, where $f$ is the density function of the $G$-Wishart distribution with parameters $\delta$ and $D$.

**Proof** The proof is a result of the fact that $\langle D, K \rangle = \langle D', K \rangle$, since in the calculation
Chapter 3. Estimation of the Normalizing Constant of $W_G(\delta, D)$

### Table 3.1: Results from numerical experiment comparing path-sampling and Monte Carlo based estimates of the log ratio of normalizing constants. Standard errors are reported for the Monte Carlo based estimates, with NE denoting “non-estimable” errors. True values are also shown for decomposable graphs.

<table>
<thead>
<tr>
<th>Graph #</th>
<th>True</th>
<th>Monte Carlo (NE)</th>
<th>Path Sampling</th>
<th>Graph #</th>
<th>True</th>
<th>Monte Carlo (NE)</th>
<th>Path Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-368.8185</td>
<td>-671.6712 (NE)</td>
<td>-Inf</td>
<td>33</td>
<td>-16.0158</td>
<td>-40.8261 (0.9995)</td>
<td>-5.082E+65</td>
</tr>
<tr>
<td>2</td>
<td>-258.2368</td>
<td>-427.9107 (NE)</td>
<td>-Inf</td>
<td>34</td>
<td>-10.5874</td>
<td>-10.341 (0.5195)</td>
<td>-9.138E+29</td>
</tr>
<tr>
<td>3</td>
<td>-221.8661</td>
<td>-1914.405 (NE)</td>
<td>-1.248E+92</td>
<td>35</td>
<td>-14.5547</td>
<td>-18.9258 (1)</td>
<td>-7.364E+24</td>
</tr>
<tr>
<td>4</td>
<td>-260.2136</td>
<td>-4601.1951 (NE)</td>
<td>-Inf</td>
<td>36</td>
<td>-22.1015</td>
<td>-23.8348 (0.2346)</td>
<td>-3.008E+04</td>
</tr>
<tr>
<td>5</td>
<td>-271.7277</td>
<td>-1938.5995 (NE)</td>
<td>-8.205E+168</td>
<td>37</td>
<td>-120.6281</td>
<td>-1026.4691 (NE)</td>
<td>-1.527E+87</td>
</tr>
<tr>
<td>6</td>
<td>-266.2893</td>
<td>-454.3852 (NE)</td>
<td>-Inf</td>
<td>38</td>
<td>-60.4490</td>
<td>-108.9103 (0.9922)</td>
<td>-1.092E+22</td>
</tr>
<tr>
<td>7</td>
<td>-270.1122</td>
<td>-4901.16 (NE)</td>
<td>-Inf</td>
<td>39</td>
<td>-20.6404</td>
<td>-23.8348 (0.2346)</td>
<td>-3.008E+04</td>
</tr>
<tr>
<td>8</td>
<td>-111.1299</td>
<td>-607.5899 (NE)</td>
<td>8.087E+65</td>
<td>40</td>
<td>-64.4163</td>
<td>-147.7101 (1)</td>
<td>-8.451E+36</td>
</tr>
<tr>
<td>9</td>
<td>-149.6318</td>
<td>-2981.7025 (NE)</td>
<td>-Inf</td>
<td>41</td>
<td>-131.5242</td>
<td>-3186.9505 (NE)</td>
<td>-1.527E+87</td>
</tr>
<tr>
<td>10</td>
<td>-160.9915</td>
<td>-739.5805 (NE)</td>
<td>-1.210E+73</td>
<td>42</td>
<td>-70.5020</td>
<td>-111.7621 (0.9979)</td>
<td>-8.184E+08</td>
</tr>
<tr>
<td>11</td>
<td>-155.7175</td>
<td>-2715.4453 (NE)</td>
<td>-5.747E+302</td>
<td>43</td>
<td>-12.5113</td>
<td>-39.8031 (1)</td>
<td>-1.349E+80</td>
</tr>
<tr>
<td>12</td>
<td>-159.5304</td>
<td>-3824.1024 (NE)</td>
<td>-1.534E+285</td>
<td>44</td>
<td>-7.974</td>
<td>-7.3032 (0.3231)</td>
<td>-2.696E+14</td>
</tr>
<tr>
<td>13</td>
<td>-113.2611</td>
<td>-656.7167 (NE)</td>
<td>-1.230E+76</td>
<td>45</td>
<td>-11.1648</td>
<td>-8.6491 (0.9999)</td>
<td>-1.203E+35</td>
</tr>
<tr>
<td>14</td>
<td>-124.6207</td>
<td>-1101.7358 (NE)</td>
<td>-1.026E+138</td>
<td>46</td>
<td>-9.4978</td>
<td>-4.8182 (0.1333)</td>
<td>-2.603E+10</td>
</tr>
<tr>
<td>15</td>
<td>-119.3468</td>
<td>-521.1263 (NE)</td>
<td>-2.263E+37</td>
<td>47</td>
<td>-9.7774</td>
<td>-9.8154 (0.3029)</td>
<td>1.017E+09</td>
</tr>
<tr>
<td>16</td>
<td>-123.1597</td>
<td>-950.72 (NE)</td>
<td>-1.730E+44</td>
<td>48</td>
<td>NA</td>
<td>-2.8395 (0.0434)</td>
<td>-1.580E+02</td>
</tr>
<tr>
<td>17</td>
<td>-163.1227</td>
<td>-809.5669 (NE)</td>
<td>-5.447E+59</td>
<td>49</td>
<td>-43.4098</td>
<td>-94.1779 (0.9895)</td>
<td>-2.432E+27</td>
</tr>
<tr>
<td>18</td>
<td>-157.5398</td>
<td>-4048.8264 (NE)</td>
<td>-Inf</td>
<td>50</td>
<td>NA</td>
<td>-73.5793 (0.993)</td>
<td>-1.575E+10</td>
</tr>
<tr>
<td>19</td>
<td>-161.5072</td>
<td>-3800.9336 (NE)</td>
<td>-Inf</td>
<td>51</td>
<td>-41.7608</td>
<td>-1066.722 (NE)</td>
<td>-Inf</td>
</tr>
<tr>
<td>20</td>
<td>-169.2084</td>
<td>-650.4983 (NE)</td>
<td>-4.066E+34</td>
<td>52</td>
<td>-53.4629</td>
<td>-75.0598 (1)</td>
<td>-9.206E+08</td>
</tr>
<tr>
<td>21</td>
<td>-173.0213</td>
<td>-1176.4562 (NE)</td>
<td>-7.746E+157</td>
<td>53</td>
<td>NA</td>
<td>-5.5257 (0.0591)</td>
<td>-1.893E+02</td>
</tr>
<tr>
<td>22</td>
<td>-167.5929</td>
<td>-3850.1228 (NE)</td>
<td>-Inf</td>
<td>54</td>
<td>-12.0231</td>
<td>-10.8815 (0.8826)</td>
<td>-9.002E+18</td>
</tr>
<tr>
<td>23</td>
<td>-109.8712</td>
<td>-674.5624 (NE)</td>
<td>-2.887E+58</td>
<td>55</td>
<td>-5.3901</td>
<td>-5.4837 (0.0492)</td>
<td>-2.053E+02</td>
</tr>
<tr>
<td>24</td>
<td>-13.7700</td>
<td>-43.5883 (1)</td>
<td>-1.384E+170</td>
<td>56</td>
<td>-18.1089</td>
<td>-18.4214 (0.0957)</td>
<td>-1.755E+02</td>
</tr>
<tr>
<td>25</td>
<td>-8.6106</td>
<td>-9.9009 (0.5161)</td>
<td>-1.373E+23</td>
<td>57</td>
<td>-55.2517</td>
<td>-90.6209 (0.6329)</td>
<td>-3.949E+11</td>
</tr>
<tr>
<td>26</td>
<td>-12.4235</td>
<td>-12.6921 (0.344)</td>
<td>3.421E+09</td>
<td>58</td>
<td>-5.5347</td>
<td>-3.422 (0.0752)</td>
<td>-4.456E+02</td>
</tr>
<tr>
<td>27</td>
<td>-52.3865</td>
<td>-86.0816 (1)</td>
<td>-1.612E+34</td>
<td>59</td>
<td>-8.5187</td>
<td>-11.0925 (0.9007)</td>
<td>-1.275E+26</td>
</tr>
<tr>
<td>28</td>
<td>-46.9581</td>
<td>-1987.7043 (NE)</td>
<td>-Inf</td>
<td>60</td>
<td>-2.0002</td>
<td>-2.0347 (0.0295)</td>
<td>-1.625E+02</td>
</tr>
<tr>
<td>29</td>
<td>-50.9254</td>
<td>-2043.856 (NE)</td>
<td>-Inf</td>
<td>61</td>
<td>-0.9552</td>
<td>-0.9311 (0.0152)</td>
<td>-1.698E+02</td>
</tr>
<tr>
<td>30</td>
<td>-152.1693</td>
<td>-598.8999 (NE)</td>
<td>-5.967E+72</td>
<td>62</td>
<td>-38.2126</td>
<td>-51.6562 (1)</td>
<td>-2.438E+08</td>
</tr>
<tr>
<td>31</td>
<td>-62.8251</td>
<td>-109.9118 (0.9924)</td>
<td>-5.002E+20</td>
<td>63</td>
<td>-2.8565</td>
<td>-2.8581 (0.0077)</td>
<td>-1.576E+02</td>
</tr>
<tr>
<td>32</td>
<td>-57.0111</td>
<td>-2012.2958 (NE)</td>
<td>-Inf</td>
<td>64</td>
<td>-5.5257</td>
<td>-0.9311 (0.0152)</td>
<td>-1.698E+02</td>
</tr>
</tbody>
</table>

Table 3.1: Results from numerical experiment comparing path-sampling and Monte Carlo based estimates of the log ratio of normalizing constants. Standard errors are reported for the Monte Carlo based estimates, with NE denoting “non-estimable” errors. True values are also shown for decomposable graphs.
Figure 3.2: Plot comparing the ranking of graphs (in decreasing order) based on the true value of the log ratio of normalizing constants and on the path-sampling estimate.
of \langle D, K \rangle all the entries of \( D \) that correspond to missing edges in \( G \) are annihilated by the zero entries of \( K \).

Despite this theoretical evidence there is a tendency in the literature that \( D^{-1} \) should belong to \( M^+(G) \), i.e. \( [D^{-1}]_{ij} = 0 \), if \((i, j)\) is a missing edge of \( G \). More specifically, in [Lenkoski and Dobra, 2008] it is clearly stated that for the estimation of the “posterior” normalizing constant given data \( X \) and prior hyperparameter matrix \( D_0 \), the matrix \( D = D_0 + X^T X \), needs to be replaced by \( D^c \) where \( D^c \) is a positive definite matrix that satisfies the conditions:

\[
D^c_{ij} = D_{ij}, \text{ for } (i, j) \in \mathcal{V} \tag{3.77}
\]

\[
(D^c)^{-1}_{ij} = 0, \text{ for } (i, j) \in \bar{\mathcal{V}} \tag{3.78}
\]

This system can be solved with the use of the Iterative Proportional Scaling algorithm ([Speed and Kiiveri, 1986]). One can see that \( D^c \) is the PD-completion of the \( \mathcal{V} \)-incomplete matrix \( D^\mathcal{V} \). Although in theory, based on Proposition 3.4.1, \( I_G(\delta, D) = I_G(\delta, D^c) \), the use of matrix \( D^c \) has been supported by some numerical results (Dobra, A., personal communication) showing that the estimation of the normalizing constant \( I_G(\delta, D) \) is improving when the matrix \( D \) is replaced by \( D^c \). In a first simple example we investigated the stability of the Monte Carlo estimate for a random matrix \( D \) and the corresponding \( D^c \). We used the non-decomposable 5-node graph showing in Figure 4.2, \( \delta = 3 \) and a random positive definite symmetric 5 × 5 matrix \( D \). We also calculated the matrix \( D^c \) (using the IPS algorithm) and for each matrix we estimated the normalizing constant using an increasing number of iterations. Figure 3.3 shows the values of the two estimates as we increase the number of samples used for the estimation. It is clear that the stability of the estimate is increasing when \( D^c \) is used.

This section describes some experiments for the investigation of this problem. First, recall that

\[
I_G(\delta, D) = C_D \cdot E(f_D(\psi^\mathcal{V}))
\]
Figure 3.3: Convergence of the Monte Carlo estimate of the normalizing constant, when $D$ and $D^c$ are used
where the distribution of $\psi^V$ is a product of independent Normal and $\chi$ distributions, and

$$f_D(\psi^V) = \exp\left\{-\frac{1}{2} \sum_{(i,j) \in \bar{V}} \psi_{ij}^2(\psi^V, D)\right\},$$

where the notation $\psi_{ij}(\psi^V, D)$ indicates that for $(i, j) \in \bar{V}$ the entries $\psi_{ij}$ are functions of $\psi^V$, depending on $D$. The estimation, as presented in [Atay-Kayis and Massam, 2005] relies on the Law of Large Numbers, which ensures that the sample average converges to the true mean. Nevertheless, the rate of convergence depends on the size of the variance of the distribution. We hypothesize that the quantity $C_D \cdot f_D(\psi^V)$ has larger variance than $C_{D^c} \cdot f_{D^c}(\psi^V)$, although the expected value is the same.

In order to investigate this problem we consider the simple graph of four-cycle. We then compute the sample variance of the function $f_D(\psi^V)$ for a given value of $D$, when $\psi^V$ follows the distribution described in Equations 3.8, 3.9 and in [Atay-Kayis and Massam, 2005]. We start with a random positive definite symmetric matrix $D_0$ and vary the entries $(1, 4)$ and $(2, 3)$ covering a large range of values, containing the values of $D^c$. We then plot the contour of the variance value against the pair values of $(D_{14}, D_{23})$. We observe that variance of $f_D(\psi^V)$ takes its lowest value when $D = D^c$. The same result is observed for different values of $D_0$. One example of the contour plot is shown in Figure 3.4.

One possible explanation for this phenomenon could be the relationship between $\text{Var}(f_D(\psi^V))$ and the determinant of $D$. It is known that given a graph $G$ and a matrix $D_0$, among all matrices $D'$ such that $D_{ij} = D'_{ij}$ when $(i, j)$ is an edge of $G$, matrix $D^c$ has the largest determinant ([Grone et al., 1984]). This was also explored numerically and confirmed, as the contour plot of $\text{det}(D)$ against the values of $(D_{14}, D_{23})$ in Figure 3.5 illustrates.

Finally, the scatterplot in Figure 3.6 suggests a close to monotone relationship between $\text{Var}(f_D(\psi^V))$ and $|D|$.

Based on the results of these experiments, we can conclude that there is evidence that matrix $D^c$ maximizes the stability of the Monte Carlo estimate of $I_G(\delta, D)$ over all the
Figure 3.4: Contour plot of the logarithm of the variance estimate against the values of $D_{14}$ and $D_{23}$. The black dot indicates the variance corresponding to $D^c$. 
Figure 3.5: Contour plot of $|D|$ against the values of $D_{14}$ and $D_{23}$. The black dot indicates the variance corresponding to $D^c$. 
Figure 3.6: Scatter plot of $|D|$ and $\log(Var(f_D(\psi^V)))$, when $D_{14}$ and $D_{23}$ vary. The black dot corresponds to $D^c$. 

Chapter 3. Estimation of the Normalizing Constant of $W_G(\delta, D)$
completions of $D^V$. As a next step, this hypothesis needs to be further investigated and justified theoretically.

3.5 Conclusions

In this Chapter we described various existing approaches for the estimation of the Normalizing Constant of the $G$-Wishart distribution. We also presented some new approaches mainly based on path sampling, a Bayesian computation method. Our application of path sampling and our numerical results did not show any improvement over the existing Monte Carlo method. Nevertheless, we believe that a more careful and customized application may be promising, and therefore further research work on this topic may be worth pursuing.

Furthermore, in this Chapter we presented evidence from numerical experiments that the stability of the Monte Carlo estimate of the normalizing constant is maximized when $D \in M^*_+(G)$. This claim requires further investigation and theoretical justification.
Chapter 4

Sampling from G-Wishart distribution

4.1 Introduction - Motivation

The ability to sample from the G-Wishart (or the Hyper Inverse Wishart) distribution is important and has many applications, including the estimation of the normalizing constant (following for instance the path sampling method and equation 3.42). Another application is the estimation of the covariance matrix and functions of it. In the Bayesian context, one is often interested in the posterior mean of \( \Sigma \) (or \( K \)) given a set of data \( x \) of size \( n \) and a selected model \( G \), denoted by \( E(\Sigma|x,G) \). Since the posterior distribution of \( K \) is given by \( W_G(\delta + n, D + x^T x) \), we can estimate \( E(\Sigma|x,G) \) by the quantity \( \hat{J} \), where

\[
\hat{J} = \frac{1}{N} \sum_{i=1}^{N} (K_i)^{-1},
\]

(4.1)

where \( K_i \) are random samples from \( W_G(\delta + n, D + x^T x) \). In a similar manner, for any function of \( K \), \( f \), we estimate the posterior mean of \( f(K) \) by

\[
\hat{J}_f = \frac{1}{N} \sum_{i=1}^{N} f(K_i),
\]

(4.2)
where $K_i$ are random samples from $W_G(\delta + n, D + x^T x)$. For those estimators to be implemented, the use of a sampling method from $G$-Wishart distribution is necessary. In this Chapter we present existing methods for sampling from the $G$-Wishart distribution and propose a new method based on the Metropolis-Hastings algorithm, accompanied by the results of a number of experiments investigating its performance. At the end of the Chapter we also give an application for model selection where sampling from $G$-Wishart is used.

4.2 Existing methods

Sampling from the Hyper Inverse Wishart distribution for decomposable graphs has been previously proposed in [Carvalho et al., 2007]. If $C_1, S_2, C_2, \ldots, S_k, C_k$ is a perfect ordering of cliques and separators of $G$ and $R_i = C_i \setminus S_i, i = 2, \ldots, k$ we can write

$$\Sigma_{C_i} = \begin{pmatrix} \Sigma_{S_i} & \Sigma_{S_i, R_i} \\ \Sigma_{R_i, S_i} & \Sigma_{R_i} \end{pmatrix}$$

(4.3)

$$D_{C_i} = \begin{pmatrix} D_{S_i} & D_{S_i, R_i} \\ D_{R_i, S_i} & D_{R_i} \end{pmatrix}$$

(4.4)

Also let $\Sigma_{R_i|S_i} = \Sigma_{R_i} - \Sigma_{R_i, S_i} \Sigma_{S_i}^{-1} \Sigma_{S_i, R_i}$, $D_{R_i|S_i} = D_{R_i} - D_{R_i, S_i} D_{S_i}^{-1} D_{S_i, R_i}$. Then the sampling scheme is as follows:

1. sample $\Sigma_{C_1} \sim IW(\delta, D_{C_1})$ and get the values for $\Sigma_{S_2}$

2. for $i = 2, \ldots, k$,

   (a) sample $\Sigma_{R_i|S_i} \sim IW(\delta + |S_i|, D_{R_i|S_i}), U_i \sim N(D_{R_i, S_i} D_{S_i}^{-1} \Sigma_{R_i|S_i} \otimes D_{S_i}^{-1})$, with $\otimes$ denoting the Kronecker product

   (b) compute $\Sigma_{R_i, S_i} = U_i \Sigma_{S_i}$ and $\Sigma_{R_i} = \Sigma_{R_i|S_i} + \Sigma_{R_i, S_i} \Sigma_{S_i}^{-1} \Sigma_{S_i, R_i}$, therefore get $\Sigma_{C_i}$
This method cannot be used when the graph is not decomposable, since in that case the marginals for the submatrices of Σ corresponding to prime components are not following the Inverse Wishart distribution. Recently a Gibbs Sampling based method has been proposed for the generation of random samples for the $G$-Wishart distribution [Asci and Piccioni, 2007]. This method is an application of a more general algorithm titled Block Gibbs Sampler that was first developed based on the theory of cuts in exponential families and natural conjugate densities, in [Piccioni, 2000]. When applied to Gaussian Graphical Models, cuts are the cliques of the graph and the Block Gibbs Sampler becomes the Bayesian Iterative Proportional Scaling (BIPS) ([Gelman et al., 1995]). For this cyclical method, a number of iterations equal to the number of cliques of the graph are needed for the generation of one sample point. If $C_1, C_2, \ldots, C_k$ is the set of cliques of $G$, The algorithm is as follows:

- Set $K^0 = I_p$
- For $r = 0, 1, \ldots$ do
  1. Set $K^{r,0} = K^r$
  2. For each $j = 1, \ldots, k$:
     a. Sample $A$ from $W_{[C_j]}(\delta, D_{C_j})$
     b. Set $K^{r,j}$ so that $K^{r,j}_{C_j,C_j} = A + K^{r,j-1}_{C_j,C_j} \left[ K^{r,j-1}_{\bar{C}_j,C_j} \right]^{-1} K^{r,j-1}_{\bar{C}_j,C_j}$, while $K^{r,j}_{\kappa,\lambda} = K^{r,j-1}_{\kappa,\lambda}$, for $(\kappa, \lambda) \notin C_j \times C_j$
  3. Set $K^{r+1} = K^{r,k}$

After some burn-in time $r_0$, the sequence $(K^r)_{r \geq r_0}$ is a set of random samples from $W_G(\delta, D)$.

Sampling using BIPS poses various limitations. The algorithm needs the set of cliques of $G$ to be enumerated. This problem is known to be NP-hard ([McHugh, 1990]). Also, significant computational time is needed since for the generation of one sample a series
of matrix inversions is needed (see step 2b above). In the following section we propose a new sampling method that does not suffer from those limitations. This method is based on the Metropolis Hastings (MH) algorithm.

4.3 New proposed method

4.3.1 Background on the general Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a special type of Markov chain Monte Carlo method and it can be used for the simulation and sampling from complex distributions. Over the years many variations of the algorithm have been developed and studied in theory and various applications ([Chib and Greenberg, 1995], [Chen et al., 2000]). The algorithm is performing an iterative generation of proposed samples, some of which are accepted as part of the simulated sample of the distribution. In general, for the application of the algorithm, we need

- the density $\pi(x)$ of the target distribution

- a proposal density $Q(x'; x^t)$, where $x^t$ denotes the current state of the sampler and $x'$ the proposed state

It is assumed that we can sample from $Q(x'; x^t)$, generating proposed samples $x'$ given the current sample $x^t$. Then, the proposed sample $x'$ is accepted with probability $\alpha(x'; x^t)$, which is equal to

$$\min \left[ \frac{\pi(x')Q(x^t; x')}{\pi(x^t)Q(x'; x^t)}, 1 \right],$$

if $\pi(x^t)Q(x'; x^t) > 0$, otherwise $\alpha(x'; x^t) = 1$. It is important to notice here that the normalizing constant of the target density $\pi(x)$ is not needed to be known, as it cancels out in the ratio expression of $\alpha(x^t; x')$. This is a property of the algorithm that makes it particularly attractive.
Another important issue of the algorithm is the choice of the proposal density $Q(x'; x^t)$. In the original implementation of the algorithm the proposal density was selected such that $Q(x'; x^t) = Q(x^t; x')$, in which case the algorithm is performing a random walk. A subsequent implementation raised this constraint suggesting proposal-generating densities that do not depend on the current state, i.e. $Q(x'; x^t) = Q(x')$. The term independence chain was adopted by [Tierney, 1994] to characterize this type of implementation. Another useful choice is offered when for the target density

$$\pi(x) \propto w(x)h(x),$$

where $h(x)$ a distribution which we can sample from, and $w(x)$ a real function. In that case choosing $h(x)$ as the proposal density is convenient, simplifying the computation of the acceptance probability $\alpha(x'; x^t)$ to

$$\min \left[ \frac{w(x')}{w(x^t)}, 1 \right].$$

[Mengersen and Tweedie, 1996] has shown that if $w(x)$ is uniformly bounded, the chain is uniformly ergodic (the strongest convergence rate condition in use). See [Chib and Greenberg, 1995] for a discussion on this topic.

The performance of the algorithm is evaluated based on how well it is sampling the target distribution and, more specifically, how well the generated samples are covering the probability mass of the target distribution. Authors have often drawn an association between the behaviour of the chain and the acceptance rate, the percentage of times the proposed sample is accepted in the chain. Under specific assumptions for target and proposal densities, optimal values for acceptance rate have been previously proposed ([Chib and Greenberg, 1995]). Nevertheless, for complex multivariate distributions no such recipes can be offered without a risk.
4.3.2 Description of the proposed MH-based sampling method

Here we present our proposed method and describe how the MH algorithm can be used for sampling from G-Wishart distribution. We make use of the fact that, given a positive definite matrix \( D \), there is bijection mapping between \( M^+(G) \) and \( M^\triangleleft(G) \), the space of all the upper triangle matrices incomplete with respect to the graph \( G \). The mapping is described in detail in [Atay-Kayis and Massam, 2005] and also in section 3.2.1 in the thesis. In summary, each \( K \) in \( P^+(G) \) is mapped to \( \psi^V \), the projection of the upper triangle matrix \( \psi \) on to the space of \( V \)-incomplete matrices, where \( \psi = \phi T^{-1} \), and \( \psi, T \) upper triangle matrices such that \( K = \phi^T \phi, D^{-1} = T^T T \). Conversely, equations [3.3, 3.4] can be used for the completion of \( \psi^V, \psi \), and given \( \psi \), \( K \) is given by \( (\psi^T)^T \psi T \). The density of \( \psi^V \) is such that

\[
p(\psi^V | G, D) \propto \exp\left\{-\frac{1}{2} \sum_{(i,j) \in \bar{V}} \psi^2_{ij}\right\} \times \prod_{i=1}^{p} \chi_{\delta+\nu_i} \times N_{|E|}(0_{|E|}, I_{|E|})
\]

(4.5)

The proposed MH algorithm described here can be used to sample from the distribution of \( \psi^V \). Then, samples from G-Wishart can be obtained using the mapping described above. The proposed algorithm is an independence chain, with proposal density

\[
Q(\psi^V) = \prod_{i=1}^{p} \chi_{\delta+\nu_i} \times N_{|E|}(0_{|E|}, I_{|E|})
\]

(4.6)

The acceptance probability is equal to

\[
\min\left\{ \frac{f((\psi^V)')} {f(\psi^V)}, 1 \right\},
\]

where

\[
f(\psi^V) = \exp\left\{-\frac{1}{2} \sum_{(i,j) \in \bar{V}} \psi^2_{ij}\right\}.
\]

Since \( f(\psi^V) \) is uniformly bounded (by 1) the chain is uniformly ergodic.

We now present the pseudo code of the method:

- Initialize the chain by sampling \( \psi^V_0 \) from \( Q \), as in Equation 4.6; set \( \psi^V_{\text{cur}} = \psi^V_0 \)
• For $i = 1, 2, \ldots, \text{MAX}$ do:

1. Sample $\psi^\mathcal{V}_{\text{prop}}$ from $Q$

2. Set

\[
\log \alpha = \frac{1}{2} \sum_{(i,j) \in \mathcal{V}} \{ (\psi^\mathcal{V}_{\text{cur}})^2_{ij} - (\psi^\mathcal{V}_{\text{prop}})^2_{ij} \}
\]

3. If $\log \alpha > 0$ then do:

- $\psi^\mathcal{V}_{\text{cur}} = \psi^\mathcal{V}_{\text{prop}}$
- $\text{accept} = 1$

4. Else do:

- Sample $b$ from $\text{Bernoulli}(\alpha)$
- If $b = 1$ then do:
  * $\psi^\mathcal{V}_{\text{cur}} = \psi^\mathcal{V}_{\text{prop}}$
  * $\text{accept} = 1$
- Else $\text{accept} = 0$

For each $\psi^\mathcal{V}_{\text{cur}}$ there is a unique corresponding matrix $K_{\text{cur}}$ that can be constructed following the reverse procedure of what it is described in the beginning of this section. The sequence of $K_{\text{cur}}$ comprises a sample from the $G$-Wishart distribution.

### 4.3.3 Experiments

**Design**

In order to investigate the performance of the MH sampling method under different scenarios we conducted a number of numerical experiments. We made use of the fact that, for a graph $G$ that contains complete prime components, exact distributions are available for comparison. If $K \sim W_G(\delta, D)$ and $C$ is a complete prime component of $G$, we know that the submatrix $\Sigma_C$, where $\Sigma = K^{-1}$, follows the Inverse Wishart distribution,
with parameters $\delta$ and $D_C$ ([Roverato, 2002]). If \{\(K^{(i)}\)\}_i is the sample generated by the MH method, we compute the determinants

\[
d^{(i)} = \det((\Sigma^{(i)}_C)^{-1}),
\]

(4.7)

where $\Sigma^{(i)} = [K^{(i)}]^{-1}$. The empirical distribution of the sample \{\(d^{(i)}\)\}_i is then compared with the empirical distribution of $\det(X)$, where $X \sim W(\delta, D_C)$. Q-Q plots can be used for the comparison.

The experiments that we have conducted are the following:

**Experiment 1**

Graph: A seven-nodes non-decomposable graph (see Figure 4.1)

Parameters: $D = I$ (Identity matrix); $\delta = 3$

Objective: Investigation of the performance of the MH sampling method in a non-decomposable graph

Evaluation tools: QQ-plot, Acceptance rates, trace plot

**Experiment 2**

Graph: A five-nodes non-decomposable graph (see Figure 4.2)

Parameters: (i) $D = D_{rand}$, a random symmetric positive definite matrix and (ii) $D = D_{c rand}$; $\delta = 3$

Objective: To measure the performance of the MH sampler when $D$ is random ($D^{-1} \notin M^+(G)$) and how this performance changes when $D$ is replaced by $D^c$

Evaluation tools: QQ-plot, Acceptance rates, trace plot, autocorrelation

**Experiment 3**

Description: A five-node non-decomposable graph $G$ is assumed and a simulated dataset $X$ is generated from a multivariate Normal distribution with zero mean and covariance matrix $\Sigma$ such that $\Sigma^{-1} \in M^+(G)$. Assuming an un-informative prior ($D_0 = I$), sampling from the posterior is performed using the MH algorithm
Graph: A five-nodes non-decomposable graph, same as in Experiment 2
Parameters: (i) $D = D_0 + X^T X$ and (ii) $D = (D_0 + X^T X)^c; \delta = 3 + n$
Objective: To measure the performance of the MH sampler in the simple case where the data are compatible with the graph and the prior is un-informative
Evaluation tools: QQ-plot, Acceptance rates, trace plot, autocorrelation

Experiment 4
Graph: The four-cycle non-decomposable graph (see Figure 4.3)
Parameters: (i) $D = D_{\text{rand}}$, a random symmetric positive definite matrix and (ii) $D = D_{\text{rand}}^c; \delta = 3$
Objective: To investigate the performance of the MH sampler for random $D$ and $D^c$ for a simple non-complete primary component
Evaluation tools: Acceptance rates, trace plot, autocorrelation, histogram (no QQ-plot is used since true distribution is not known)

Experiment 5
Graph: The four-cycle non-decomposable graph (as in Experiment 4)
Parameters: $D = I_4 + X^T X; \delta = 3 + n$, where $X$ is the *Iris virginica* dataset, described in Section 3.3.3
Objective: to compare the performance of MH method with the Block Gibbs Sampling method, by comparing the posterior expected value estimates of $K$ and the posterior density estimates of $\det(K)$.

Results
From the results of Experiment 1 shown in Figure 4.4 we see that under the simple case of $D = I$ and despite the fact that $G$ is non-decomposable, the sampling distribution is very similar to the true distribution. Also the acceptance rate is quite high, suggesting
Figure 4.1: Non-decomposable graph used in Experiment 1

Figure 4.2: Non-decomposable graph used in Experiments 2 and 3
Chapter 4. Sampling from G-Wishart distribution

55

Figure 4.3: Non-decomposable graph used in Experiments 4 and 5

that the samples are approximately independent. The trace plot indicates that the chain seems to be mixing well.

Figure 4.4: QQ-plot, acceptance rate plot and trace plot for Experiment 1

Experiment 2 results (in Figures 4.5 and 4.6) show that with a random $D$, where $D^{-1} \notin M^+(G)$, the algorithm is not performing well. Using $D^c$ instead of $D$ seems to solve the problem and the sampling works fine.

Figure 4.7 shows the results from Experiment 3, using $D = I + X^T X$. It is interesting
Figure 4.5: Plots for Experiment 2 using random $D$
Figure 4.6: Plots for Experiment 2 using $D^c$
to observe that the algorithm is working well without using the PD-completion of $D^V$, although $D = (I + X^T X)^c$ improves the sampling (data not shown). This is because the prior is un-informative ($D_0 = I$) and the simulated data used are generated from the assumed graph. Therefore the sample covariance matrix, $X^T X$, is close to belong to $M^+(G)$, and the same holds for $I + X^T X$.

![QQ-plot for log(det)](image)
![traceplot of log(det)](image)
![acceptance rate](image)
![Autocorrelation of log(det)](image)

Figure 4.7: Plots for Experiment 3 using random $D = I + X^T X$

Experiment 4 has a similar design with Experiment 3, with the exception that there is no true distribution available for comparison. One can see that when $(I + X^T X)^c$ is used instead of $I + X^T X$ acceptance rate is increasing and autocorrelation is decreasing, suggesting that sampling is overall improving. Figures 4.8 and 4.9 show the results.

Experiment 5 showed that the mean estimate generated from the MH method is very close to the estimate from the Block Gibbs sampler. The posterior density estimates of
Figure 4.8: Plots for Experiment 4 using $D = I + X^T X$
Figure 4.9: Plots for Experiment 4 using $D = (I + X^T X)^c$
det(K) from the two methods are very similar, nevertheless the estimate from MH shows signs of multiples modes. This could be because the chain is occasionally “trapped” in probability areas and it does not always move around well. The application of “thinning” of the chain could remediate this problem.

Following the results of those experiments we can conclude that the MH algorithm seems to provide a reliable method for sampling from the $G$-Wishart distribution, with relatively small computational expense.

### 4.3.4 Application for model selection using DIC

In this section we present an application of the proposed MH-based sampling method to model selection. The selection is performed with the use of the Deviance Information Criterion (DIC) ([Spiegelhalter et al., 2002]), a criterion similar in spirit to the AIC and BIC. It provides with a measure of how well a particular model fits some existing data, while taking into account how easy it is for the model to fit the data (how many parameters it effectively has). It can therefore be used for Bayesian model selection. Given a set of data $x$ and unknown parameter $\theta$, deviance is defined as $D(\theta) = -2\log(p(x|\theta)) + C$, where $p(x|\theta)$ is the likelihood function and $C$ a constant that depends only on the data.

Under a specific model and posterior distribution of $\theta$ given $x$ one can define the expectations $\bar{\theta} = E_{\theta|x}[\theta]$ and $\bar{D} = E_{\theta|x}[D(\theta)]$. The Deviance Information Criterion can be calculated as

$$DIC = p_D + \bar{D},$$

where $p_D = \bar{D} - D(\bar{\theta})$. The quantity $p_D$ is a measure of how easily the model fits the data, playing the role of the effective number of parameters in the model ([Spiegelhalter et al., 2002]). The larger $p_D$ is the easier it is for the model to fit the data. On the other hand the quantity $\bar{D}$ measures how well the model fits the data, with larger $\bar{D}$ indicating worse fit. Overall, models with smaller DIC are preferred to those with larger DIC values. One of the characteristics of DIC that makes it attractive over other model selection criteria
is the fact that it can be calculated when an MCMC or other sampling method for the posterior distribution of $\theta$ is available. In that case $\bar{\theta}$ and $\bar{D}$ can be estimated by

$$\frac{1}{N} \sum_{i=1}^{N} \theta_i$$

and

$$\frac{1}{N} \sum_{i=1}^{N} D(\theta_i)$$

drespectively, where $\theta_i, i = 1, \ldots, N$ are samples from the posterior distribution of $\theta$ given $x$. The standard error of $\bar{D}$ can be estimated by

$$\sqrt{\frac{\sum_{i=1}^{N} (D(\theta_i) - \bar{D})^2}{N(N-1)}}.$$

In Gaussian Graphical Models, the precision matrix $K$ plays the role of the parameter $\theta$, with known posterior distribution $W_G(\delta + n, D + x^T x)$. The DIC of a graph $G$ can be defined as in Equation 4.8, and $\bar{K}$ and $\bar{D}$ can be estimated after sampling efficiently from $W_G(\delta+n, D+x^T x)$ with the use of the proposed MH-method. We now present a numerical example of the calculation of DIC for various graphical models. We use the *iris virginica* data and calculate the DIC values for all 63 four-node graphs, similar to the experiment in section 3.3.3. We use the popular choices of $\delta = 3$ and $D = I_4$. Under each model we generate 10,000 samples from the posterior distribution of $K$, using the MH-based sampling method. The initial 2,000 iterations are discarded as “burn-in”. Using those samples we estimate $\bar{K}$ and $\bar{D}$, and subsequently calculate the DIC values. We examine the consistency of the DIC values with the values of the log of the ratio of normalizing constants (equal to the marginal likelihood $p(x|G)$ up to a constant multiplying factor), denoted with $\lambda$. Figure 4.10 shows a scatterplot of the values of $\text{-DIC}$ against $\lambda$. The values of the two measures seem to be well correlated, which gives the indication that model selection based on DIC value is similar to the one based on the marginal likelihood. Similar indication is offered by the observation that 9 out of the 10 best models based on DIC belong to the set of the 10 best models according to $\lambda$ values.
Figure 4.10: Scatterplot between the log ratio of the normalizing constants, $\lambda$, and the value of -DIC
Table 4.1: Results from numerical experiment comparing DIC and other related parameters between two graphical models using *iris virginica* data

<table>
<thead>
<tr>
<th></th>
<th>$p_D$</th>
<th>$\bar{D}$</th>
<th>se($\bar{D}$)</th>
<th>DIC</th>
<th>user time (sec)</th>
<th>system time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_1$:</td>
<td>chain 1</td>
<td>8.073</td>
<td>379.410</td>
<td>0.065</td>
<td>387.483</td>
<td>8.912</td>
</tr>
<tr>
<td>$G_1$:</td>
<td>chain 2</td>
<td>7.182</td>
<td>378.643</td>
<td>0.057</td>
<td>385.825</td>
<td>9.098</td>
</tr>
<tr>
<td>$G_1$:</td>
<td>chain 3</td>
<td>7.346</td>
<td>378.539</td>
<td>0.060</td>
<td>385.885</td>
<td>9.025</td>
</tr>
<tr>
<td>$G_1$:</td>
<td>average</td>
<td>7.534</td>
<td>378.864</td>
<td>0.061</td>
<td>386.397</td>
<td>9.012</td>
</tr>
<tr>
<td>$G_2$:</td>
<td>chain 1</td>
<td>9.185</td>
<td>379.858</td>
<td>0.061</td>
<td>389.043</td>
<td>8.178</td>
</tr>
<tr>
<td>$G_2$:</td>
<td>chain 2</td>
<td>9.173</td>
<td>379.811</td>
<td>0.061</td>
<td>388.984</td>
<td>7.981</td>
</tr>
<tr>
<td>$G_2$:</td>
<td>chain 3</td>
<td>9.071</td>
<td>379.694</td>
<td>0.060</td>
<td>388.765</td>
<td>8.101</td>
</tr>
<tr>
<td>$G_2$:</td>
<td>average</td>
<td>9.143</td>
<td>379.788</td>
<td>0.061</td>
<td>388.931</td>
<td>8.087</td>
</tr>
</tbody>
</table>

We also investigate the repeatability and stability of DIC by focusing on the comparison of two graphs, the four-cycle shown in Figure 4.3, denoted as $G_1$ and the complete four nodes graph, denoted as $G_2$. In each case, we use three randomly initialized chains. In each chain the initial 2,000 iterations are discarded as “burn-in”. The results from this experiment are presented in Table 4.1. From the results we can see that model $G_1$ has smaller values for both $p_D$ and $\bar{D}$ than model $G_2$ and it is therefore preferred for the used data.
Chapter 5

Bayes Factors

5.1 Definition and importance in Model Selection

Bayes Factor between two models $M_1$ and $M_2$ given a set of data $Y$ is defined as the ratio of the marginal likelihood for $M_1$ over the marginal likelihood for $M_2$. In the case of Gaussian Graphical Model selection, models are corresponding to graphs. Given two graphs $G_1$ and $G_2$ and data set $Y$ the Bayes factor is given by $p(Y|G_1)/p(Y|G_2)$, i.e. the ratio of the marginal likelihoods.

From Equation (2.27) the ratio becomes

$$BF_{G_1;G_2} = \frac{I_{G_1}(\delta + n, D + U)}{I_{G_2}(\delta + n, D + U)} \cdot \frac{I_{G_2}(\delta, D)}{I_{G_1}(\delta, D)}, \quad (5.1)$$

where $U = Y^TY$. Many model selection methods compare graphical models based on the ratio of their posterior probabilities. Since

$$\frac{p(G_1|y)}{p(G_2|y)} = \frac{\pi(G_1)}{\pi(G_2)} \cdot \frac{p(y|G_1)}{p(y|G_2)}, \quad (5.2)$$

the value of the Bayes Factor between two graphs determines the selection of the model. When the $G$-Wishart prior is used for the precision matrix $K$, the value of the Bayes Factor depends on the values of the hyperparameters $\delta$ and $D$. It is therefore of high
interest to determine the effect of those parameters to the Bayes Factor between two graphs, given a set of data. In the following sections we provide expressions for the Bayes Factors between two graphs that differ by one edge. This kind of model comparison is of practical use since most of model search methods are performing comparisons of “neighbouring” graphs, i.e. graphs that differ by one edge.

5.2 Expression for decomposable graphs

In this section we develop expressions for the Bayes Factor of two “neighbouring” graphs that are both decomposable. We denote with $G_1, G_0$ the graphs with and without the edge, respectively. Previously (e.g. [Roverato and Whittaker, 1998]) it has been shown that if both $G_1$ and $G_0$ are decomposable and assuming the $G$-Wishart prior

$$BF_{G_1:G_0} = \frac{p(Y|C\ \{u\}, \delta, D)p(Y|C\ \{v\}, \delta, D)}{p(Y|C, \delta, D)},$$

(5.3)

where $C$ is the unique clique of $G_1$ containing the edge $(u, v)$. Using the equation (2.27) we can express the Bayes Factor above using the normalizing constants corresponding to the cliques $C, C\ \{u\}, C\ \{v\}, C\ \{u, v\}$. We also know (see [Atay-Kayis and Massam, 2005]) that for a complete graph $G$ and hyperparameters $\delta, D$, the normalizing constant is equal to

$$I_G(\delta, D) = 2^{\tau p/2} \Gamma_p(\tau/2) |D|^{\tau/2},$$

(5.4)

where $\tau = \delta + p - 1$ and $\Gamma_p(\alpha) = \pi^{p(p-1)/4} \prod_{i=0}^{p-1} \Gamma(\alpha - i/2)$, the multivariate $\Gamma$ function.

In the case of a complete graph, using equations (2.27) and (5.4), $p(Y|G)$ becomes

$$p(Y|G) = \frac{1}{(2\pi)^{np/2}} \frac{2^{(\tau+n)p/2} \Gamma_p((\tau+n)/2)|D|^{\tau/2}}{2^{\tau p/2} \Gamma_p(\tau/2)|D^*|^{(\tau+n)/2}}$$

$$= \frac{1}{\pi^{np/2}} \frac{\Gamma_p((\tau+n)/2)|D|^{\tau/2}}{\Gamma_p(\tau/2)|D^*|^{(\tau+n)/2}}$$

(5.6)
Equation (5.3) then becomes

$$BF_{G_1\gamma_0} = \frac{\left[ \frac{\Gamma_{p-1}}{\Gamma_p} \right]^{(\frac{(\tau-1)+n}{2})}}{\left[ \frac{\Gamma_{p-2}}{\Gamma_p} \right]^{(\frac{(\tau-1)}{2})}} \cdot \frac{|D_{-u,-u}|^{(\frac{n}{2})} |D_{v,-v}|^{(\frac{n}{2})} |D^{*}_{u,-u}|^{(\frac{n}{2})} |D^{*}_{v,-v}|^{(\frac{n}{2})}}{|D_{u,-u}|^{(\frac{n}{2})} |D_{v,-v}|^{(\frac{n}{2})} |D^{*}_{u,-u}|^{(\frac{n}{2})} |D^{*}_{v,-v}|^{(\frac{n}{2})}}$$

We also note that if $\alpha$ is a real number,

$$\frac{\left[ \frac{\Gamma_{p-1}}{\Gamma_p} \right]^{\alpha}}{\left[ \frac{\Gamma_{p-2}}{\Gamma_p} \right]^{\alpha}} = \frac{\left[ \frac{\Gamma_{p-1}^{\alpha}}{\Gamma_p^{\alpha}} \right]^{\frac{(p-1)}{4}} \prod_{i=0}^{p-2} \Gamma(\alpha - 1/2 - i/2)^{\frac{(p-2)}{4}} \prod_{i=0}^{p-1} \Gamma(\alpha - i/2)^{\frac{(p-1)}{4}}}{\prod_{i=0}^{p-2} \Gamma(\alpha - i/2)^{\frac{(p-1)}{4}} \prod_{i=0}^{p-1} \Gamma(\alpha - i/2)^{\frac{(p-1)}{4}}}$$

Therefore equation (5.8) becomes

$$BF_{G_1\gamma_0} = \frac{\Gamma(\frac{n+1}{2}) \Gamma(\frac{r}{2})}{\Gamma(\frac{n+1}{2}) \Gamma(\frac{r-1}{2})} \cdot \frac{|D_{-u,-u}|^{-1/2} |D_{v,-v}|^{-1/2} |D^{*}_{u,-u}|^{(n-2)/2} |D^{*}_{v,-v}|^{(n-2)/2}}{|D_{u,-u}|^{(n-2)/2} |D^{*}_{u,-u}|^{(n-2)/2} |D^{*}_{v,-v}|^{(n-2)/2}}$$

We can now make use of the following notation and properties: first, we denote $Q = D^{-1}$. Second, we know that, given that $Q$ exists,

$$Q \cdot |D| = \text{adj}D,$$

where $\text{adj}D = (d_{i,j}^*)$ is the classical adjoint of $D$, with element $d_{i,j}^* = (-1)^{i+j} |D_{i,-j}|$.

Third, we also know (see Theorem 2.5.2 of [Prasolov, 1996]) that

$$|(\text{adj}D)_{I,J}| = \epsilon |D_{-J,-I}||D|^{r-1},$$

where $I, J$ are two subsets of $1, \ldots, p$ of the same cardinality $r$ and $\epsilon = 1$ or $-1$.

We then have

$$|D_{uv,-uv}| |D| = |(\text{adj}D)_{uv,uv}| = d_{u,v}^* d_{v,u}^* - d_{u,v}^* d_{v,u}^* = |D_{u,-u}||D_{-v,-v}| - |D_{u,-u}||D_{-v,-u}|.$$
Therefore we have
\[
\frac{|D_{-u,u}||D_{-v,v}|}{|D_{-uv,-uv}|D} = \frac{|D_{-u,u}||D_{-v,v}|}{|D_{-u,u}||D_{-v,v}| - |D_{-u,v}||D_{-v,u}|}
\]
\[= \frac{Q_{u,u}Q_{v,v}}{Q_{u,u}Q_{v,v} - Q_{u,v}Q_{v,u}},\] (5.16)
where the last equality holds since \(Q_{i,j} = |D|^{-1}(-1)^{i+j}|D_{-i,-j}|, \forall i, j \in \{1, \ldots, p\} \).

If we interpret \(D\) and \(D^*\) as un-normalized sample covariance matrices of hypothetical prior and posterior datasets (see comment in Property 4 of \(G\)-Wishart in section 2.2.2) the above ratio becomes equal to
\[
\frac{|D_{-u,u}||D_{-v,v}|}{|D_{-uv,-uv}|D} = \frac{1}{1 - \pi_{u,v}^2}, \quad \text{(5.17)}
\]
where \(\pi_{u,v}\) is the prior partial correlation of \(u, v\) (see e.g. \([Cox \ and \ Wermuth, \ 1996]\)).

Denoting with \(\pi_{u,v}^*\) the posterior partial correlation of \(u, v\) we have
\[
r_D = \frac{|D_{-u,u}||D_{-v,v}|}{|D_{-uv,-uv}|D} = \frac{1 - (\pi_{u,v}^*)^2}{1 - \pi_{u,v}^2}. \quad \text{(5.18)}
\]
The ratio \(r_D\) is larger, equal or smaller than 1 if \(|\pi_{u,v}^*|\) is smaller, equal or larger than \(|\pi_{u,v}|\) respectively.

We can use the above expressions in order to calculate the value at which \(BF_{G_1:G_0}\) converges when \(\delta \to \infty\). For this we use the Sterling’s approximation
\[
\Gamma(x) \approx \sqrt{2\pi}x^{\frac{1}{2}}e^{-x}.
\]

Using equation (5.14) we then have
\[
\lim_{\delta \to \infty} \frac{\Gamma\left(\frac{\tau+n-1}{2}\right)\Gamma\left(\frac{\tau}{2}\right)}{\Gamma\left(\frac{\tau+n}{2}\right)\Gamma\left(\frac{\tau-1}{2}\right)} = \frac{(\tau + n - 1)^{\frac{\tau+n-2}{2}} \tau^{\frac{\tau-1}{2}}}{(\tau + n)^{\frac{\tau+n-1}{2}} (\tau - 1)^{\frac{\tau-2}{2}}}
\]
\[= \frac{1}{\tau^{\frac{\tau-1}{2}}}, \quad \text{(5.19)}
\]
\[
\frac{(\tau+n-1)^{\frac{\tau+n-2}{2}} \tau^{\frac{\tau-1}{2}}}{(\tau+n)^{\frac{\tau+n-1}{2}} (\tau-1)^{\frac{\tau-2}{2}}}
\]
\[= \frac{1}{\tau^{\frac{\tau-1}{2}}}, \quad (5.20)
\]
\[
\frac{1}{(1 + \frac{n-1}{\tau})^{\frac{\tau+n-2}{2}} (1 - \frac{1}{\tau})^{\frac{\tau-2}{2}}}
\]
\[= \frac{e^{-\frac{n-1}{\tau}}}{e^{-\frac{1}{\tau}e^{-\frac{\tau}{2}}} = 1, \quad (5.21)}
\]
Therefore the limit of $BF_{G_1:G_0}$ when $\delta \to \infty$ is $\infty$ if $|\pi_{u,v}^*| < |\pi_{u,v}|$, 0 if $|\pi_{u,v}^*| > |\pi_{u,v}|$, and if $|\pi_{u,v}^*| = |\pi_{u,v}|$ it is equal to

$$\lim_{\delta \to \infty} BF_{G_1:G_0} = \frac{|D_{u,v}^{*}|^{n+2} |D_{u,v}|^{\frac{p-2}{2}} |D_{u,v,-uv}|^{\frac{p-2}{2}} |D_{uv,-uv}|^{\frac{n-2}{2}} |D_{u,v,-uv}|^{-1}}{|D_{u,v}^{*}|^{(n-1)/2} |D_{u,v,-uv}|^{(n-2)/2} |D_{uv,-uv}|^{-1}} \quad (5.23)$$

When $\delta \to 0$, $\tau \to p - 1$, the limit of $BF_{G_1:G_0}$ is

$$\lim_{\delta \to 0} BF_{G_1:G_0} = \frac{\Gamma\left(\frac{p+n-2}{2}\right)\Gamma\left(\frac{p-1}{2}\right)}{\Gamma\left(\frac{p+n-1}{2}\right)\Gamma\left(\frac{p-2}{2}\right)} \cdot \frac{|D_{u,v}^{*}| |D_{u,v}|^{\frac{p-2}{2}} |D_{u,v,-uv}|^{\frac{p-2}{2}} |D_{uv,-uv}|^{\frac{n-2}{2}} |D_{u,v,-uv}|^{-1}}{|D_{u,v}^{*}|^{(n-1)/2} |D_{u,v,-uv}|^{(n-2)/2} |D_{uv,-uv}|^{-1}} \quad (5.24)$$

$$\frac{|D_{u,v}^{*}| |D_{u,v}|^{\frac{p-2}{2}} |D_{u,v,-uv}|^{\frac{p-2}{2}} |D_{uv,-uv}|^{\frac{n-2}{2}} |D_{u,v,-uv}|^{-1}}{|D_{u,v}^{*}|^{(n-1)/2} |D_{u,v,-uv}|^{(n-2)/2} |D_{uv,-uv}|^{-1}} = \frac{\Gamma\left(\frac{p+n-2}{2}\right)\Gamma\left(\frac{p-1}{2}\right)}{\Gamma\left(\frac{p+n-1}{2}\right)\Gamma\left(\frac{p-2}{2}\right)} \cdot (1 - (\pi_{u,v})^2)^\frac{n-2}{2} \cdot \frac{1 - (\pi_{u,v})^2}{1 - (\pi_{u,v})^2} \cdot \frac{Q_{u,v}^{*}}{Q_{u,v}^{*}} \quad (5.26)$$

where the last equality is due to the Equation 5.17.

### 5.3 Bayes Factors for non-decomposable graphs

The results of the previous section can be generalized to not necessarily decomposable graphs, given that some conditions are met. This can be done with the use of the following proposition:

**Proposition 5.3.1** Let $G, G'$ be graphs on the same set of vertices $V$ and $A, B \subset V$ such that (i) $G_B = G'_B$ and (ii) $(A \setminus B, B \setminus A, A \cap B)$ is a decomposition for both $G$ and $G'$. Then $BF_{G:G'} = BF_{G_A:G'_A}$.

**Proof** From Equation 2.10 we know that

$$p(X|G) = \frac{p(X_A|G_A)p(X_B|G_B)}{p(X_{A\cup B}|G_{A\cup B})}, \quad (5.27)$$

as well as

$$p(X|G') = \frac{p(X_A|G'_A)p(X_B|G'_B)}{p(X_{A\cup B}|G'_{A\cup B})}, \quad (5.28)$$
Using equations 5.27, 5.28, the Bayes factor becomes

\[ BF_{G,G'} = \frac{p(X|G)}{p(X|G')} = \frac{p(X_A|G_A)p(X_B|G_B)p(X_{A\cup B}|G'_{A\cup B})}{p(X_{A\cup B}|G'_{A\cup B})} \frac{p(X_A|G_A)p(X_B|G_B)}{p(X_A|G'_{A})p(X_B|G'_B)} = BF_{G_A,G'_A} \] (5.29)

\[ \frac{p(X|G_A)}{p(X|G'_A)} = BF_{G_A:G'_A} \] (5.30)

If \( G_0, G_1 \) meet the conditions of Proposition 5.3.1 and \( A_0, A_1 \) are decomposable, then, according to the proposition, \( BF_{G_1:G_0} = BF_{A_1:A_0} \), and so any result and expression of Section 5.2 is applied for \( BF_{G_1:G_0} \).

### 5.4 Approximation of Bayes Factors

In this section we employ the Laplace approximation in order to compute an approximate expression for the Bayes Factor between graphs \( G_0 \) and \( G_1 \). From expression 3.18 we have that when \( \delta \neq 2 \),

\[ \tilde{I}_G(\delta, D) = (2\pi)^{\frac{|V|}{2}} |\hat{K}|^{\frac{\delta-2}{2}} \exp \left\{ -\frac{1}{2} < \hat{K}, D > \right\} |F_{\delta,D}(\hat{K})|^{\frac{1}{2}}, \] (5.31)

where \( \hat{K} \) indicates the point that maximizes the function \( f_{\delta,D}(K) = -\frac{1}{2}(2-\delta) \log |K| + \langle K, D \rangle \) and \( F_{\delta,D}(\hat{K}) \) is minus the inverse of the Hessian of \( f_{\delta,D} \) evaluated at \( \hat{K} \). Since the density of \( K \) is proportional to \( \exp\{f_{\delta,D}(K)\} \), \( \hat{K} \) is also the mode of \( W_G(\delta, D) \). The mode \( \hat{K} \) is the solution of the following system:

\[ K_{ij} = 0, \text{ for } (i,j) \in \tilde{V} \] (5.32)

\[ \frac{\partial f_{\delta,D}(K)}{\partial K_{ij}} = 0, \text{ for } (i,j) \in V \] (5.33)

For \( (i,j) \in V \) we have

\[ \frac{\partial f_{\delta,D}(K)}{\partial K_{ij}} = -\frac{1}{2}(2-\delta) \log |K| + \langle K, D \rangle \right\} = -\frac{1}{2} tr \left\{ [(2-\delta)K^{-1} + D] L_{ij} \right\}, \] (5.34)

where \( L_{ij} \) is a symmetric \( p \) by \( p \) matrix with entries equal to 1 at the positions \( (i,j) \) and \( (j,i) \) and 0 anywhere else ([Harville, 2007],[Lenkoski and Dobra, 2010]).
System 5.33 then becomes

\[ K_{ij} = 0, \text{ for } (i, j) \in \mathcal{V} \]  
\[ K_{ij}^{-1} = \frac{D_{ij}}{\delta - 2}, \text{ for } (i, j) \in \mathcal{V} \]  

(5.35) (5.36)

By definition, the solution of the system gives \( \hat{K} = (\delta - 2)(D^c)^{-1} \), where \( D^c \) is the PD-completion of the matrix \( D^V \) with respect to the graph \( G \).

For the matrix \( F_{\delta,D}(\hat{K}) \) we first notice that

\[ \frac{\partial^2 f_{\delta,D}(K)}{\partial K_{ij} \partial K_{lm}} = -\frac{\delta - 2}{2} \text{tr} \left\{ (K^{-1} L_{ij} K^{-1} L_{lm}) \right\} = -\frac{\delta - 2}{2} \|s\|_{ss(K^{-1})}(i,j),(l,m), \]  

(5.37)

and therefore we have

\[ F_{\delta,D}(\hat{K}) = -(D^2 f_{\delta,D}(\hat{K}))^{-1} = \left[ (\delta - 2)^{-1} ||s\|_{ss(D^c)V^V}^{-1} \right]^{-1} \]  

(5.38)

From Equation 5.31 we have

\[ \tilde{I}_G(\delta, D) = (2\pi)^{|V|/2} |(\delta - 2)(D^c)^{-1}|^{\frac{\delta - 2}{2}} \exp \left\{ -\frac{1}{2} \text{tr}[(\delta - 2)I_p] \right\} \cdot |(\delta - 2)^{-1} ||s\|_{ss(D^c)V^V}^{-1}^{\frac{1}{2}} \]

\[ = (2\pi)^{|V|/2} (\delta - 2)^{|p(\delta - 2)/2|} |(D^c)|^{-\frac{\delta - 2}{2}} \exp \left\{ -\frac{1}{2} p(\delta - 2) \right\} |(D^c)|^{-\frac{\delta - 2}{2}} \exp \left\{ -\frac{1}{2} p(\delta - 2) \right\} |I_s(D)|^{-\frac{1}{2}} \]

\[ = |(D^c)|^{-\frac{\delta - 2}{2}} |I_s(D)|^{-\frac{1}{2}} \exp \left\{ \frac{1}{2} |V| \log 2\pi + p\delta' \log \delta' + |V| \log \delta' - p\delta' \right\} \]

\[ = |(D^c)|^{-\frac{\delta - 2}{2}} |I_s(D)|^{-\frac{1}{2}} 2^{p/2} \exp \left\{ \frac{1}{2} |V| \log(2\pi\delta') + p\delta' \log(\delta') + \log(2) - \delta' \right\}, \]

where \( \delta' = \delta - 2 \). Therefore we can write

\[ [\tilde{I}_G(\delta, D)]^{-1} = |I_s(D)|^{1/2} |D^c|^{(\delta - 2)/2} 2^{p/2} \tilde{h}_G(\delta, D), \]  

(5.39)

where

\[ \tilde{h}_G(\delta, D) = \exp \left\{ -\frac{1}{2} |V| \log(2\pi\delta') + p\delta' \log(\delta') + \log(2) - \delta' \right\} \]  

(5.40)
which it also used in Equation 6 in [Roverato, 2002]. Equation 5.39 can be used for the approximation of the Bayes Factor: assuming a scale matrix \( D \), we use the notation \( D^0, D^1 \) for the PD-completion of \( D^V \) with respect to the graphs \( G_0, G_1 \) respectively. We also use the notation \( D^* = D + X^T X, \delta^* = \delta + n \). The approximation then becomes:

\[
\tilde{BF}_{G_1;G_0} = \frac{\text{Iss}(D^1)_{\mathcal{V}_1 \mathcal{V}_1}^{1/2}|D^1|^{(\delta-2)/2}\tilde{h}_{G_1}(\delta,D)|\text{Iss}((D^*)^0)_{\mathcal{V}_0 \mathcal{V}_0}^{1/2}|(D^*)^0|^{(\delta^*-2)/2}\tilde{h}_{G_1}(\delta^*,D^*)}{\text{Iss}((D^*)^1)_{\mathcal{V}_1 \mathcal{V}_1}^{1/2}|(D^*)^1|^{(\delta^*-2)/2}\tilde{h}_{G_1}(\delta^*,D^*)}|\text{Iss}(D^0)_{\mathcal{V}_0 \mathcal{V}_0}^{1/2}|D^0|^{(\delta-2)/2}\tilde{h}_{G_1}(\delta,D)}
\]

(5.41)

We now have

\[
\tilde{h}_{G_1}(\delta,D)\tilde{h}_{G_1}(\delta^*,D^*) = \exp \left\{ -\frac{1}{2} \left[ |\mathcal{V}| \log(2\pi\delta') + p\delta'\log(\delta') + p\log(2) - p\delta' + (|\mathcal{V} - 1|) \log(2\pi(\delta' + n)) + p(\delta' + n)\log(\delta' + n) + p\log(2) - p(\delta' + n) \right. \\
- |\mathcal{V}| \log(2\pi(\delta' + n)) - p(\delta' + n)\log(\delta' + n) - p\log(2) + p(\delta' + n) \\
- \left. (|\mathcal{V} - 1|) \log(2\pi\delta') - p\delta'\log(\delta') - p\log(2) + p\delta' \right]\}
\]

\[
= \exp \left\{ -\frac{1}{2} \left[ \log(2\pi\delta') - \log(2\pi(\delta' + n)) \right] \right\} = \left( \frac{\delta' + n}{\delta'} \right)^{1/2},
\]

Equation 5.41 then becomes

\[
\tilde{BF}_{G_1;G_0} = \left( \frac{\delta' + n}{\delta'} \right)^{1/2} \frac{\text{Iss}(D^1)_{\mathcal{V}_1 \mathcal{V}_1}^{1/2}|D^1|^{(\delta-2)/2}|\text{Iss}((D^*)^0)_{\mathcal{V}_0 \mathcal{V}_0}^{1/2}|(D^*)^0|^{(\delta^*-2)/2}}{\text{Iss}((D^*)^1)_{\mathcal{V}_1 \mathcal{V}_1}^{1/2}|(D^*)^1|^{(\delta^*-2)/2}|\text{Iss}(D^0)_{\mathcal{V}_0 \mathcal{V}_0}^{1/2}|D^0|^{(\delta-2)/2}}
\]

(5.42)

Equation 5.42 can be greatly simplified when \( D^0 = D^1 \) and \( (D^*)^0 = (D^*)^1 \), although this condition does not necessarily hold. If it holds Equation 5.41 becomes:

\[
\tilde{BF}_{G_1;G_0} = \left( \frac{\delta' + n}{\delta'} \right)^{1/2} \frac{\text{Iss}(D^c)_{\mathcal{V}_1 \mathcal{V}_1}^{1/2}|\text{Iss}((D^*)^c)_{\mathcal{V}_0 \mathcal{V}_0}^{1/2}}{\text{Iss}((D^*)^c)_{\mathcal{V}_1 \mathcal{V}_1}^{1/2}|\text{Iss}(D^c)_{\mathcal{V}_0 \mathcal{V}_0}^{1/2}}
\]

(5.43)

If \( e \) denotes the edge differing \( G_1 \) from \( G_0 \), it holds \( \mathcal{V}_1 = \mathcal{V}_0 \cup \{ e \} \). Then, if we use the notation \( i_e \) for the index of \( e \) in matrix \( \text{Iss}(D)_{\mathcal{V}_1 \mathcal{V}_1} \), the \( M_{i_e,i_e} \) minor is equal to \( |\text{Iss}(D)_{\mathcal{V}_0 \mathcal{V}_0}| \). Therefore

\[
|\text{Iss}(D)_{\mathcal{V}_0 \mathcal{V}_0}| = |\text{Iss}(D)_{\mathcal{V}_1 \mathcal{V}_1}|^{-1} \cdot |\text{Iss}(D)_{\mathcal{V}_1 \mathcal{V}_1}|.
\]
Hence 5.43 becomes
\[
\widetilde{BF}_{G_1;G_0} = \left( \frac{\delta' + n}{\delta'} \right)^{1/2} \left\{ \frac{[\text{Iss}(D^*)_{V_1;V_1}]_{i_e,i_e}^{-1}}{[\text{Iss}(D)_{V_1;V_1}]_{i_e,i_e}^{-1}} \right\}^{1/2}.
\]
(5.44)

From this formula we can conclude that under the condition of \(D^0 = D^1\) and \((D^*)^0 = (D^*)^1\) and with sufficiently large sample size \(n\), \(\widetilde{BF}_{G_1;G_0} > 1\), which indicates that model \(G_1\) is selected over \(G_0\).

### 5.5 Limit of \(\delta \cdot I_G(\delta, D)\) and applications

In Bayesian analysis we are often interested in priors that put as little as possible weight on prior information. Those “uninformative” priors are often improper, in the sense that they do not constitute a proper distribution. In the case of \(G\)-Wishart prior we have seen in Chapter 2 that the hyperparameter \(\delta\) needs to be positive in order for the distribution to be proper. Also, we have given the interpretation that the value of \(\delta\) in a sense characterizes the amount of “prior” information or knowledge that we want to add to the analysis, in comparison to the available dataset. More specifically, \(\delta\) plays the role of the sample size of a hypothetical prior dataset, which is to be added to the sample size of the available data set in order to give the sample size of the “posterior” hypothetical dataset (as suggested by the relationship \(\delta^* = \delta + n\)). This recommends that the smaller the value of \(\delta\) the smaller the amount of “prior” knowledge or information that we are bringing into the analysis. In that sense, a value of \(\delta = 0\) will correspond to an “uninformative” prior, which will be also improper. This improper prior would be of limited usefulness if Bayes Factors were not well defined and could not be used for model selection. In this Section we prove that this is not the case and that Bayes Factors can be well defined and used for model selection, when an improper prior with \(\delta = 0\). In a more strict sense we prove that the limit of the Bayes Factor as \(\delta\) approaches 0 is finite.

This interesting result derives from the following new Theorem that characterizes the limiting behaviour of \(I_G(\delta, D)\) when \(\delta \to 0\).
Theorem 5.5.1 For a given graph \( G(E, V) \) with \( p = |V| \), and \( D \) a \( p \)-by-\( p \) positive definite matrix we have that
\[
\lim_{\delta \to 0^+} \delta \cdot I_G(\delta, D) = c,
\]
where \( c \in \mathbb{R}^+ \).

Before we prove Theorem 5.5.1 we present a series of lemmas:

Lemma 5.5.2 If \( G(E, V) \) is a complete graph with \( p = |V| \), and \( D \) a \( p \)-by-\( p \) positive definite matrix we have that
\[
\lim_{\delta \to 0^+} \delta \cdot I_G(\delta, D) = c,
\]
where \( c \in \mathbb{R}^+ \).

Proof Since \( G \) is complete we know ([Atay-Kayis and Massam, 2005]) that
\[
I_G(\delta, D) = 2^{(\delta + p - 1)/2} \frac{\Gamma_p[(\delta + p - 1)/2]}{|D|^{(\delta + p - 1)/2}},
\]
where \( \Gamma_p(\alpha) = \pi^{p(p-1)/4} \prod_{i=0}^{p-1} \Gamma(\alpha - i/2) \), the multivariate \( \Gamma \) function. We then have
\[
\lim_{\delta \to 0^+} \delta \cdot I_G(\delta, D) = \lim_{\delta \to 0^+} \delta \pi^{p(p-1)/4} \prod_{i=0}^{p-1} \Gamma((\delta + p - 1 - i)/2) \frac{2^{(\delta + p - 1)/2}}{|D|^{(\delta + p - 1)/2}} \quad (5.45)
\]
\[
= \pi^{p(p-1)/4} \prod_{i=0}^{p-2} \Gamma((p - 1 - i)/2) \frac{2^{(p-1)/2}}{|D|^{(p-1)/2}} \cdot \lim_{\delta \to 0^+} \delta \Gamma((\delta/2) \quad (5.46)
\]
We now that for every \( x > 0 \) it is \( x \Gamma(x) = \Gamma(x + 1) \), therefore \( \lim_{x \to 0^+} x \Gamma(x) = \Gamma(1) = 1 \).

Also, if we call \( p - 1 = q \), Equation 5.46 becomes
\[
\lim_{\delta \to 0^+} \delta \cdot I_G(\delta, D) = \pi^{q/2} \pi^{q(q-1)/4} \prod_{i=0}^{q-1} \Gamma((q - i)/2) \frac{2^{(p-1)/2+1}}{|D|^{(p-1)/2}} \quad (5.47)
\]
\[
= \pi^{q/2} \Gamma_q(q/2) \frac{2^{(p-1)/2+1}}{|D|^{(p-1)/2}} = \pi^{(p-1)/2} \Gamma_{p-1}((p - 1)/2) \frac{2^{(p-1)/2+1}}{|D|^{(p-1)/2}} \quad (5.48)
\]
The last quantity in Equation 5.48 is a positive real number, which completes the proof.

The following lemma extends this result to not necessarily complete prime components of \( G \):
Lemma 5.5.3 If $G(E,V)$ is a prime graph (that does not take a decomposition) with $p = |V|$, and $D$ a $p$-by-$p$ positive definite matrix we have that

$$
\lim_{\delta \to 0^+} \delta \cdot I_G(\delta, D) = c,
$$

where $c \in \mathbb{R}^+$. 

Proof From [Atay-Kayis and Massam, 2005] we know that the normalizing constant of an arbitrary graph $G$ given the hyperparameters $\delta, D$ is equal to

$$
I(G, D) = 2^p \prod_{i=1}^p t_{ii}^{\delta+b_i-1} \int \exp \left( -\frac{1}{2} \sum_{(i,j) \in V} \psi_{ij}^2 \right) \prod_{i=1}^p (\psi_{ii}^2)^{\delta+i-1/2} \exp \left( -\frac{1}{2} \sum_{i=1}^p \psi_{ii}^2 \right) \prod_{(i,j) \in V, i \neq j} d\psi_{ij}
$$

Also, by looking at the equations that give the non-free $\psi$’s from the free ones, we make the observation that no $\psi_{ij}$ depends on the value of $\psi_{pp}$. Also, $\nu_p = 0$. Therefore, we can split the quantity in 5.50 into

$$
I(G, D) = 2^p \prod_{i=1}^p t_{ii}^{\delta+b_i-1} \int \exp \left( -\frac{1}{2} \sum_{(i,j) \in V} \psi_{ij}^2 \right) \prod_{i=1}^{p-1} (\psi_{ii}^2)^{\delta+i-1/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{p-1} \psi_{ii}^2 \right) \prod_{(i,j) \in V, i \neq j} d\psi_{ij} \times I_2(\delta)
$$

where

$$
I_2(\delta) = \int_0^{+\infty} (x^2)^{(\delta-1)/2} \exp \left( -\frac{1}{2} x^2 \right) dx
$$

Changing the variable from $x$ to $y = -x^2/2$ we get $dx = x^{-1}dy$ and

$$
I_2(\delta) = \int_0^{+\infty} x^{\delta-2} \exp (-y) dy = 2^{\delta/2-1} \int_0^{+\infty} y^{\delta/2-1} \exp(-y)dy = 2^{\delta/2-1} \Gamma(\delta/2)
$$

We are now ready to calculate the $\lim_{\delta \to 0^+} \delta \cdot I_G(\delta, D)$. We use equation 5.52 and first observe that the quantity

$$
I_1(\delta) = 2^p \prod_{i=1}^p t_{ii}^{\delta+b_i-1} \int \exp \left( -\frac{1}{2} \sum_{(i,j) \in V} \psi_{ij}^2 \right) \prod_{i=1}^{p-1} (\psi_{ii}^2)^{\delta+i-1/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{p-1} \psi_{ii}^2 \right) \prod_{(i,j) \in V, i \neq j} d\psi_{ij}
$$

(5.55)
exists when $\delta = 0$, and it is $I_1(0) > 0$. This is due to the fact that after the application of Perfect Elimination Ordering we can guarantee for the vertices of $G$ that $\nu_i \geq 1$, for $i = 1, \ldots, p - 1$. Therefore we can follow an argument similar to the one used in Lemma 3.2.1 and prove that $I_1(0) > 0$. On the other hand we have

$$\lim_{\delta \to 0^+} \delta \cdot I_2(\delta) = \lim_{\delta \to 0^+} \delta \cdot 2^{\delta/2 - 1} \Gamma(\delta/2) = \lim_{\delta \to 0^+} 2^{\delta/2} \Gamma(\delta/2 + 1) = 1$$

Therefore we now have that

$$\lim_{\delta \to 0^+} \delta \cdot I(\delta, D) = I_1(0) = 2^p \prod_{i=1}^{p} b_{ii}^{b_{ii} - 1} \int \exp \left( -\frac{1}{2} \sum_{(i,j) \in V} \psi_{ij}^2 \right) \prod_{i=1}^{p-1} (\psi_{ii}^2)^{(\nu_i - 1)/2}$$

$$\times \exp \left( -\frac{1}{2} \sum_{i=1}^{p-1} \psi_{ii}^2 \right) \exp \left( -\frac{1}{2} \sum_{(i,j) \in V, i \neq j} \psi_{ij}^2 \right) \prod_{i=1}^{p-1} d\psi_{ii} \prod_{(i,j) \in V, i \neq j} d\psi_{ij},$$

and the proof is complete. \[\Box\]

We are now ready to prove Theorem 5.5.1.

**Proof** Let assume that $(P_j, j = 1, \ldots, k)$ and $(S_j, j = 2, \ldots, k)$ is a perfect ordering of the set of prime components and its corresponding set of separators ([Lauritzen, 1996]).

We then know that the normalizing constant of $G$ can be written as

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

From Lemmata 5.5.2 and 5.5.3 we have that

$$\lim_{\delta \to 0^+} \delta \cdot I_{P_j}(\delta, D_{P_j}) = c_j > 0, j = 1, \ldots, k$$

and

$$\lim_{\delta \to 0^+} \delta \cdot I_{S_j}(\delta, D_{S_j}) = d_j > 0, j = 2, \ldots, k.$$
Therefore, we have

\[
\lim_{\delta \to 0^+} \delta \cdot I_G(\delta, D) = \lim_{\delta \to 0^+} \delta \cdot \frac{\prod_{j=1}^{k} I_{G_{Pj}}(\delta, D_{Pj})}{\prod_{j=2}^{k} I_{G_{Sj}}(\delta, D_{Sj})} 
\]

(5.60)

\[
= \lim_{\delta \to 0^+} \frac{\delta^{k-1} \prod_{j=1}^{k} I_{G_{Pj}}(\delta, D_{Pj})}{\delta^{k-1} \prod_{j=2}^{k} I_{G_{Sj}}(\delta, D_{Sj})} 
\]

(5.61)

\[
= \frac{\prod_{j=1}^{k} \lim_{\delta \to 0^+} \delta \cdot I_{G_{Pj}}(\delta, D_{Pj})}{\prod_{j=2}^{k} \lim_{\delta \to 0^+} \delta \cdot I_{G_{Sj}}(\delta, D_{Sj})} 
\]

(5.62)

\[
= \frac{\prod_{j=1}^{k} c_j}{\prod_{j=2}^{k} d_j} 
\]

(5.63)

and since the last quantity is a positive real number, the proof is complete. □

This result can be used for determining the limiting value of the Bayes Factor, as \( \delta \) approaches 0.

**Proposition 5.5.4** If \( G_1, G_2 \) are graphs, the limit

\[
\lim_{\delta \to 0^+} BF_{G_1:G_2} 
\]

(5.64)

is a real positive number.

**Proof** From Equation 5.1 we have

\[
\lim_{\delta \to 0^+} BF_{G_1:G_2} = \lim_{\delta \to 0^+} \frac{I_{G_1}(\delta + n, D + U) I_{G_2}(\delta, D)}{I_{G_1}(\delta + n, D + U) I_{G_2}(\delta, D)} 
\]

(5.65)

\[
= \lim_{\delta \to 0^+} \frac{\delta \cdot I_{G_2}(\delta, D)}{\delta \cdot I_{G_1}(\delta, D)} \cdot \frac{I_{G_1}(\delta + n, D + U)}{I_{G_2}(\delta + n, D + U)} 
\]

(5.66)

\[
= \frac{\lim_{\delta \to 0^+} \delta \cdot I_{G_2}(\delta, D)}{\lim_{\delta \to 0^+} \delta \cdot I_{G_1}(\delta, D)} \cdot \frac{I_{G_1}(n, D + U)}{I_{G_2}(n, D + U)} 
\]

(5.67)

From Theorem 5.5.1 we have \( \lim_{\delta \to 0^+} \delta \cdot I_{G_2}(\delta, D) \in \mathbb{R}^+ \) and \( \lim_{\delta \to 0^+} \delta \cdot I_{G_1}(\delta, D) \in \mathbb{R}^+ \), and since \( I_{G_1}(n, D + U) \in \mathbb{R}^+ \), \( I_{G_2}(n, D + U) \in \mathbb{R}^+ \) we have that \( \lim_{\delta \to 0^+} BF_{G_1:G_2} \in \mathbb{R}^+ \). □

One of the implications of this result is related to the choice of a “flat” prior. Flat priors are often selected as less informative, especially when the desire is to maximize the influence of the data to the posterior distribution. It is often the case that flat
priors are not proper, as in the case of G-Wishart, when choosing $\delta = 0$. In that case, the prior is not a proper density function and the normalizing constant is not a finite number. Nevertheless, traditional Model Selection procedures can be applied even when an improper prior is chosen. The selection between graphs $G_1, G_2$ can be made by using the limit $\lim_{\delta \to 0^+} BF_{G_1;G_2}$, which according to Proposition 5.5.4 is finite. In addition, the posterior distribution will be $G$-Wishart with parameter $\delta^* = n$ and therefore it will be always proper.

Those results indicate that in Gaussian Graphical models the use of an improper prior (with $\delta = 0$) does not generate significant problems in model selection and it is a valid and often convenient choice.
Chapter 6

Conclusions and future directions

In this thesis we investigated various issues related to model selection and precision matrix estimation in Gaussian Graphical Models, under the Bayesian framework. We propose different approaches in applying the method of path sampling for the estimation of the ratio of posterior over prior normalizing constants of the $G$-Wishart distribution. Despite the lack of improvement over the traditional Monte Carlo estimation method, we believe that a more careful application of path sampling or other similar Bayesian computation method can be proven beneficial.

We also investigated the issue of completing the incomplete matrix $D^V$ before we apply the Monte Carlo estimation method for the normalizing constant $I_G(\delta, D^V)$. Through a number of numerical experiments we find that PD-completion provides the most stable estimate of the true value, a result that seems to be in accordance to the common practice in the literature. As a next step, a more thorough theoretical investigation is needed.

One major contribution of the thesis is the proposal of a new sampling method from the $G$-Wishart distribution, based on the Metropolis-Hastings algorithm. First series of experiments showed satisfactory results and improved efficiency over existing sampling methods, such as the Block Gibbs Sampler. In addition, efficient sampling using this method can be used for the estimation of the Deviance Information Criterion (DIC),
which, as our experiments suggest, can provide with a computationally inexpensive but sufficiently stable alternative to Bayes Factors for model selection. Further investigation is needed for fine tuning of the method in order to improve its performance.

Finally, in this thesis we investigated the Bayes Factors between two graphs that differ by one edge. Expressions for either the exact Bayes Factor value or an approximation were developed. Of particular theoretical interest and practical application is a new result regarding the limit of the normalizing constant multiplied by the parameter $\delta$, as $\delta$ approaches 0. This result suggests that the use of improper $G$-Wishart priors does not pose significant problems in model selection.
Bibliography


