UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of
STATISTICS AND QUANTITATIVE METHODS

PhD program in
STATISTICS AND MATHEMATICAL FINANCE

XXXII CYCLE

Objective Bayes Structure Learning in Gaussian Graphical Models

Author
NIKOLAOS PETRAKIS
Supervisor
Prof. GUIDO CONSONNI
Tutor
Prof. STEFANO PELUSO

January 9, 2020

Acknowledgements

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

I would like to express my special appreciation and thanks to my advisor Professor Dr. Consonni Guido for all the support and encouragement he gave me. Without his guidance and constant feedback this PhD would not have been achievable.

I would also like to thank my tutor Professor Dr. Peluso Stefano for all his support and help for the construction of our paper and the development of the code for this PhD project.

I would like to express my gratitude to Professor Dr. Fouskakis Dimitrios for his help, support and guidance. Without him I would have never gone through this path.

I would like to thank my colleagues Bolzoni Mattia, Cappozzo Andrea and Denti Fransesco for their invaluable support through the tough times of this PhD program.

Finally, I would like to thank my partner in life, Ivon Maria Spyridi for her constant support over these three past years of the program.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Graphical models are used to represent conditional independence relationships among variables by the means of a graph, with variables corresponding to a graph's nodes. They are widely used in genomic studies (Dobra et al. 2004 and Bhadra and Mallick 2013), finance (Sohn and Kim 2012 and Carvalho and Scott 2009), energy forecasting (Wytock and Kolter 2013), among other fields.

Our interest lies in a collection of q real valued random variables $Y = \{Y_1, \cdots, Y_q\}$ from which we observe n i.i.d. q-dimensional observations which can be arranged in an $n \times q$ matrix

$$\mathbf{Y}_{n\times q} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_q) = \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix} \tag{1.1}$$

where $\mathbf{y}_i = (y_{i1}, \cdots, y_{iq})$ denotes the i-th observation and $\mathbf{Y}_j = (y_{1j}, \cdots, y_{nj})$ denotes the observations on the j-th variable.

We model the observations as $\mathbf{y}_i|\mathbf{\Sigma} \sim N_q(\mathbf{0}, \mathbf{\Sigma})$ independently over $i = 1, \cdots, n$, where $\mathbf{\Sigma}_{q\times q}$ is an unconstrained semi-positive definite matrix, and $N_q(\mathbf{0}, \mathbf{\Sigma})$ denotes the q-variate normal distribution with mean vector $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. The distribution of $\mathbf{Y}$ given the covariance matrix $\mathbf{\Sigma}$ is a special case of the Matrix Normal distribution, which will be presented in section 4.2.

Our goal is to depict the conditional independence structure of variables $\{Y_1, \cdots, Y_q\}$ by the means of an undirected graph, whose structure we assume to be unknown and to be inferred by the data at hand. This procedure in the bibliography is referred as ***Structure Learning***, which in reality is a model selection procedure that involves graphical models. There are two approaches to this problem:

1. Constraint-based approaches, where we search for a graph structure which satisfies the independence assumptions observed through the empirical distribution.

2. Score based approaches, where we define a scoring function which enable us to rank the models at hand and then identify the highest-scoring model.

This thesis will explore score-based approaches, based on posterior probabilities. We start from defining a model space which is consisted by a set of candidate graphical models; then we define a scoring function which enables us to score the different models of the model space and finally, we construct a search algorithm that will navigate through the model space to identify the optimal model that explains the problem at hand. The choise of a scoring function is crucial for optimizing the search procedure through the model space. Our approach to this problem is purely Bayesian for handling uncertainty in a more elaborate fashion. We will use estimates of posterior model probabilities for ranking the models at hand.

The underlying graph's structure is identified by observing the non-zero entires of the inverse of the covariance matrix $\mathbf{\Sigma}^{-1} = \mathbf{K}$, called, the concentration matrix. Thus, for estimating a candidate model's posterior probabilty, we first need to assign a prior on the graph itself and a prior distribution over the column-covariance matrix $\mathbf{\Sigma}$

The specification of a conditional prior on $\mathbf{\Sigma}$ is not trivial, because each graph under consideration induces a different independence structure and it affects the parameter space. In cases where it is infeasible to sucessfully elicit a subjective prior, especially in high dimensions, we resort to Objective Bayes, an approach which exploits non-informative priors. By non-informative priors, we refer to prior distributions with minimal impact on the corresponding posterior analysis.

Furthermore, the non-informative priors considered throughout this thesis are improper, i.e. do not have a finite mass and we assume that they depend on an arbitrary normallizing constant which is responsible for converting their total mass to finite (see section 3.3). These priors cannot be used for computing Bayes factors, since a pairwise model comparison will be multiplied by the ratio of their respective normallizing contants and subsequently deem the Bayes factor indeterminate (see section section 3.3).

For creating an automated Bayesian scoring technique, we resort to Objective Bayes approaches, which are initiated by an improper prior distribution and their output is a fully usable prior distributions.

Objective Bayes contributions to the field of strucutre learning can be found in Carvalho and Scott (2009), Consonni et al. (2017), Castelletti et al. (2018), where they all follow the same scoring approach based on the Fractional Bayes Factor of O'Hagan (1995). Despite its efficiency and computational conveniency, the Fractional Bayes Factor has the disadvantage of using the entire of the data twice for both prior specification and model selection.

In this thesis, we propose the use of two alternative Objective Bayes approaches for estimating posterior probabilities of models, namely the Expected Posterior prior approach of Pérez and Berger (2002) and the Power-Expected Posterior Prior approach of Fouskakis et al. (2015). Both approaches utilize the device of imaginary observations for providing usable prior distributions and are theoretically sounder than the Fractional Bayes Factor of O'Hagan (1995). They both facilitate pairwise comparison of models under consideration, by comparing them with a reference model, and then comparing the respective Bayes factors to decide which model is prefferable. Our goal is to introduce both the Expected and Power-Expected Posterior prior approaches to the field of structure learning of undirected graphical models and evaluate their performance using certain stochastic search techniques.

The remainder of the thesis is organized as follows. In chapter 2, we provide some basic notions and notations of Graph theory, Markov and Hyper-Markov laws.

In chapter 3 we describe the Model selection problem under a Bayesian point of view, focused on Objective Bayes approaches. This chapter specifically describes the importance of the use of both Expected and Power-Expected posterior prior approaches and how they can be applied under certain cases.

In chapter 4 we face the structure learning problem under an Objective Bayes point of view. We first describe how the Fractional Bayes Factor is being used for evaluating the performance of different models versus the simplest model available. Then we apply, step by step, both Power-Expected and Expected posterior prior approaches to the graphical model selection framework. We then provide the course of action required for estimating Bayes factors of candidate models versus the simplest model available.

Then, we explore standard choices for assigning a prior on any given graph under consideration. After exploring both prior specification steps we describe a stochastic search algorithm that will be used for exploring the given graphical model space. Finally, we provide applications of Expected and Power-Expected posterior prior approaches to artificially simulated datasets as well as and in real-life application. Both approaches are compared with the Fractional Bayes Factor approach of Carvalho and Scott (2009) and the Birth-Death MCMC approach of Mohammadi and Wit (2015).

In chapter 6 we describe the conclusions from the applications of Expected and Power-Expected Posterior prior approach to the structure learning problem of undirected gaussian graphical modeles as well as future directions for reasearch. Computational Appendix is provided. Finally, chapter 4 is an extended version of a paper submitted to a Special Issue Article to of Statistica Neerlandica for 2020 (Manuscript ID: 2019-050).

Chapter 2

Notions of Graph Theory and Graphical Models

By the term ***Graphical Model***, we refer to a statistical model that embodies a collection of marginal and conditional independencies which can may be summarized by means of a graph (Dawid and Lauritzen (1993)). Graphical models can incorporate richness in modeling, clarity of interpretation and expedite analysis of complex problems.

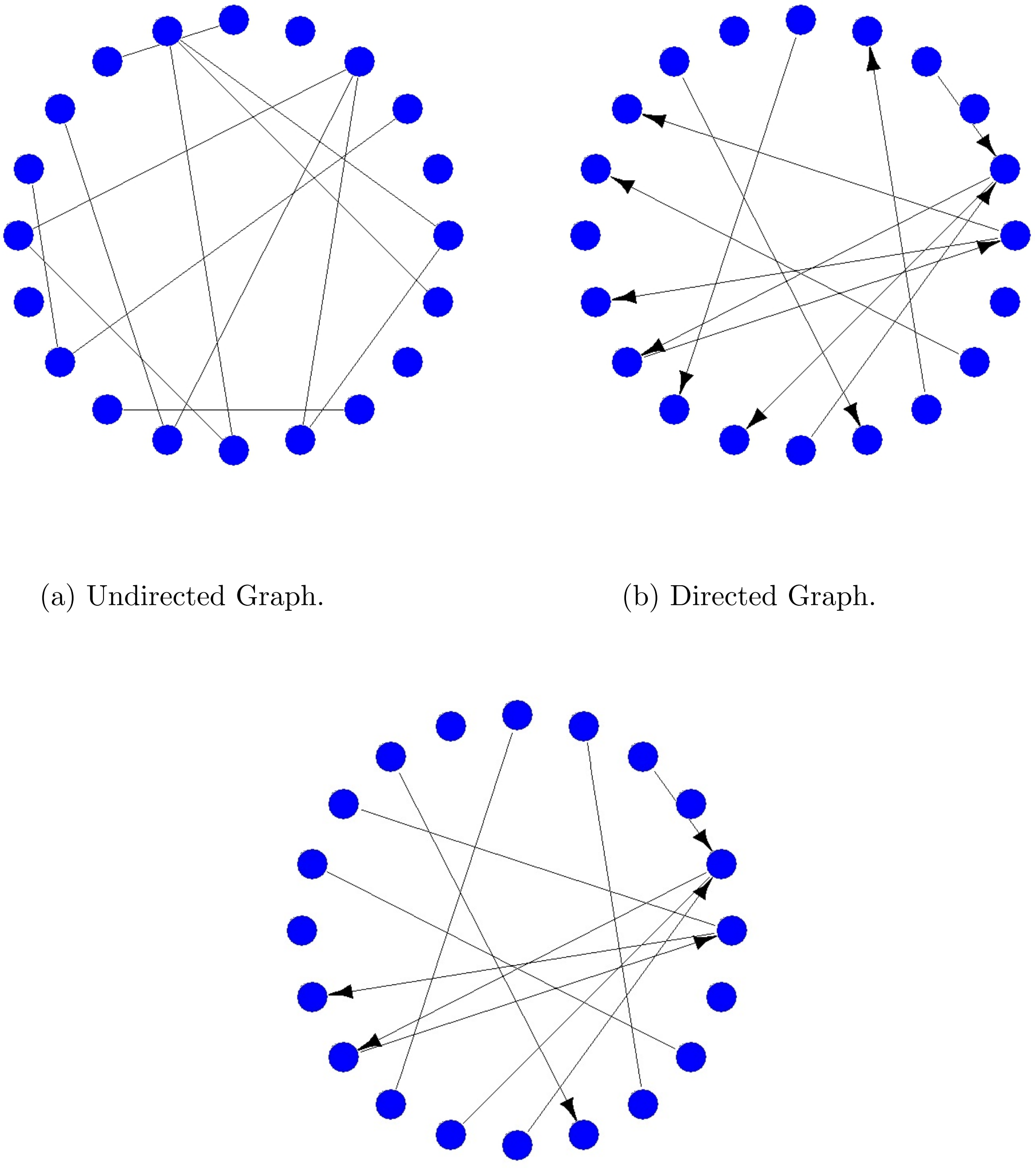
A graph, which may be either undirected, directed or a combination of them, contains nodes associated with variables on one to one correspondence under a given problem. If a directed graph is considered, then edges from a '*parent*' node lead to a '*child*' node, representing unique direct influences on the respective child node, which are independent of any other possible direct influences conditional on the other parents. If a graph is undirected, then a node is independent of any other node given its immediate neighbors. This thesis will be devoted to undirected graphs.

The goal of this chapter is to set the groundwork for chapter 4, where we introduce to the reader the concept of Gaussian graphical models. We first present the basic notions of graph theory that will be used throughout this thesis. Then we state the notion of conditional independence, which is a key inferential feature of graphical models. We then move to the description of Markov and Hyper Markov laws and how they are connected to Bayesian inference. All notions descibed, are based on Dawid and Lauritzen (1993) and Lauritzen (1996).

2.1 Notions of Graph Theory

Prior to the establishment of Markov and Hyper Markov laws, it is necessary to define basic terms and notions of Graph Theory that will be used throughout this thesis. A ***Graph*** G is characterized by a pair $G = (V, E)$, where V denotes a finite set of ***Vertices*** and E denotes a set of ***Edges*** that connect nodes of set V. Thus, a respective graph G cannot contain multiple edges or loops.

Let (a, b) denote an edge connecting nodes A and B with $A, B \in V$. If both (a, b) and (b, a) exist in E, then we will refer only to node (a, b) which will be called an ***undirected*** edge. If one of the edges does not exist in E, then the respective edge that exists in E will be called a ***directed*** edge. If a graph G is composed by undirected edges, it will be called an ***Undirected Graph*** and if the edges are directed, then it will be called a ***Directed Graph***. If a graph G contains both directed and undirected edges, it will be called ***Mixed Graph***; all three types of graphs are illustrated in Figure 2.1. This thesis will consider graphs of Figure 2.1a.



(a) Undirected Graph.

(b) Directed Graph.

(c) Mixed Graph.

Figure 2.1: Graph Examples.

An graph will be stated as ***Complete***, if all its respective vertices are connected. A graph's subset will be stated as complete, if it induces complete subgraph. A complete subgraph that is maximal with respect to $\subset$ will be called a ***Clique***. A subset $S \subset V$ will be called a ***Separator***, if all paths from a to b intersect S. The subset S is said to separate set A from B if it is an (a, b) separator for every $a \in A$ and $b \in B$.

Definition 1. *A pair (A, B) of subsets of the vertex set V of an undirected graph G is said to form a decomposition of G, if $V = A \cup B$, $A \cap B$ is complete and $A \cap B$ separates A from B.*

Thus, a graph that is decomposable, it can be successively decomposed into its respective cliques. This can be formally stated in the following definition provided by Lauritzen (1996):

Definition 2. *An undirected graph is said to be decomposable if it is complete, or if there exists a proper decomposition (A, B) into decomposable subgraphs G_A and G_B.*

The decomposition of a graph G is assumed to be proper, such that subgraphs G_A and G_B contain less vertices than the original graph G. An undirected graph G will be called ***Triangulated***, if every cycle of length $n \geq 4$ possesses a chord, i.e. two non-consecutive neighboring vertices. This leads to the following proposition:

Proposition 1. *An undirected graph is decomposable if and only if it is triangulated.*

A sequence $(C_1, \cdots, C_k)$ of complete sets in G such that $\forall j > 1$, R_j is simplicial in G_{H_j}, where

$$H_j = (C_1 \cup \cdots \cup C_j), \quad R_j = C_j \setminus H_{j-1}, \tag{2.1}$$

is said to be ***Perfect***. Elements H_j are denoted as histories and R_j are the residuals of the respective sequence. A ***Perfect Numbering*** of the vertices V of graph G is a numbering $(\alpha_1, \cdots, \alpha_k$ such that the sets

$$bd(\alpha_j) \cap \{\alpha_1, \cdots, \alpha_{j-1}\}, \quad j > 1 \tag{2.2}$$

are complete sets, i.e. the sequence $(\{\alpha_1\}, \cdots, \{\alpha_k\})$ is a perfect sequence of sets.

Thus, we obtain the following proposition from the Appendix of Dawid and Lauritzen (1993), which enables us with the ability to identify a decomposable graph:

Proposition 2. *Given an undirected graph G, the following are equivalent:*

1. *The graph G admits a perfect directed version D.*
2. *The cliques of G admit a perfect numbering.*
3. *The graph G is decomposable.*

2.2 Conditional Independence

Key feature of graphical models is the conditional independence of random variables, described by the underlying structure of a graph. The notion of conditional independence was extensively studied by Dawid (1979) and was formally defined in Dawid (1980). We provide the definition of conditional independence that was used in Dawid and Lauritzen (1993):

Definition 3. *If X, Y, Z are random variables on a probability space $(\Omega, \mathscr{A}, P)$, we say that X is* ***Conditionally Independent*** *of Y given Z under P, and write $X \perp\!\!\!\perp Y \,|\, Z\,[P]$, if for any measurable set N in the sample space of X, there exists a version of the conditional probability $P(X \in N \,|\, Y, Z)$ which is a function of Z alone.*

The relation $X \perp\!\!\!\perp Y \,|\, Z\,[P]$ may not have a probabilistic interpretation, yet as in Lauritzen (1996) p.30, it can be expressed as:

Having knowledge about Z, reading Y is irrelevant for reading X .

It further has the following properties:

1. If $X \perp\!\!\!\perp Y \,|\, Z$ then $Y \perp\!\!\!\perp X \,|\, Z$.
2. If $X \perp\!\!\!\perp Y \,|\, Z$ and U is a measurable function of X, then $U \perp\!\!\!\perp Y \,|\, Z$.
3. If $X \perp\!\!\!\perp Y \,|\, Z$ and U is a measurable function of X, then $X \perp\!\!\!\perp Y \,|\, (Z, U)$.
4. If $X \perp\!\!\!\perp Y \,|\, Z$ and $X \perp\!\!\!\perp W \,|\, (Y, Z)$ then $X \perp\!\!\!\perp (W, Y) \,|\, Z$.
5. If $X \perp\!\!\!\perp Y \,|\, Z$ and $X \perp\!\!\!\perp Z \,|\, Y$ then $X \perp\!\!\!\perp (Y, Z)$.

Lauritzen (1996) p.29 provide the notion of conditional independence, under a generic probability density f of random variables (X, Y, Z). The following statements will hold:

1. $X \perp\!\!\!\perp Y \,|\, Z \Leftrightarrow f_{XYZ}(x, y, z) = f_{XZ}(x, z) f_{ZY}(y, z)/f_Z(z)$.
2. $X \perp\!\!\!\perp Y \,|\, Z \Leftrightarrow f_{X|YZ}(x|y, z) = f_{X|Z}(x|z)$.
3. $X \perp\!\!\!\perp Y \,|\, Z \Leftrightarrow f_{XZ|Y}(x, z|y) = f_{X|Z}(x|z) f_{Z|Y}(z|y)$.
4. $X \perp\!\!\!\perp Y \,|\, Z \Leftrightarrow f_{XYZ}(x, y, z) = h(x, z)k(y, z)$ for some h, k.
5. $X \perp\!\!\!\perp Y \,|\, Z \Leftrightarrow f_{XYZ}(x, y, z) = f_{X|Z}(x|z) f_{YZ}(y, z)$.

If the density f is continuous, the above equations hold if the quantities x, y, z are well-defined i.e. the respective densities of all conditioning variables are positive.

2.3 Markov Laws for Undirected Graphs

For the remainder of this thesis, $G = (V, E)$ will always denote an undirected graph with vertex set V and edge set E, which will be assumed to be decomposable. We consider a collection of random variables $(Y_a)_{a \in V}$ having values on probability spaces $(\mathscr{Y}_a)_{a \in V}$. For a subset $A \subseteq V$, we denote $\mathscr{Y}_A = \times_{a \in A} \mathscr{Y}_a$ and $\mathscr{Y} = \mathscr{Y}_V$. Similarly $Y_A = (Y_a)_{a \in A}$. The notion of conditional independence will imply

$$A \perp\!\!\!\perp B \,|\, C \rightarrow Y_A \perp\!\!\!\perp Y_B \,|\, Y_C. \tag{2.3}$$

Lauritzen (1996) p.32 provided three distinct Markov properties under a graph $G = (V, E)$ for a collection of random variables $(y_a)_{a \in V}$. These are the following:

1. A probability measure P on $\mathscr{Y}$ is said to obey the ***Pairwise*** Markov property relative to G, if for any pair (α, β) of non-adjacent vertices

$$\alpha \perp\!\!\!\perp \beta \,|\, V \setminus \{\alpha, \beta\}. \tag{2.4}$$

2. A probability measure P on $\mathscr{Y}$ is said to obey the ***Local*** Markov property relative to G, if for any vertex $\alpha \in V$

$$\alpha \perp\!\!\!\perp V \setminus cl(\alpha) \,|\, bd(\alpha). \tag{2.5}$$

3. A probability measure P on $\mathscr{Y}$ is said to obey the ***Global*** Markov property relative to G, if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in G

$$A \perp\!\!\!\perp B \,|\, S. \tag{2.6}$$

The Markov properties are related, as described by proposition 3.4 Lauritzen (1996) p.33, in the following way:

Proposition 3. *For any undirected graph G and any probability distribution on $\mathscr{X}$, it holds that*

$$\text{Global Markov} \rightarrow \text{Local Markov} \rightarrow \text{Pairwise Markov}.$$

The global Markov property is of great importance, because it enables us to decide when two sets of variables A and B are conditionally independent given another set of variables S. Dawid and Lauritzen (1993) define the Markov property for a distribution P as follows:

Definition 4. *A distribution P on a V is called **Markov** over G if for any decomposition (A, B) of G*

$$A \perp\!\!\!\perp B \,|\, A \cap B \,[P]. \tag{2.7}$$

Let Q and R be the underlying distributions for Y_A and Y_B respectively. In order for a common underlying joint distribution to exist, having Q and R as its marginals, both Q and R must be consistent.Dawid and Lauritzen (1993) denote consistency in the following sense:

Definition 5. *We say that distributions Q over A and R over B are consistent if they both yield the same distribution over $A \cap B$.*

Lemma 1. *Suppose that the distributions Q over A and R over B are consistent. Then, there exists a unique distribution P over $A \cup B$ such that:*

1. $P_A = Q$.

2. $P_B = R$.

3. $A \perp\!\!\!\perp B \,|\, A \cap B \,[P]$.

The set of distributions that satisfy condition 3 of Lemma 1 will be denoted as $M(A,B)$. We will denote with P, the Markov combination of Q and R that satisfy all conditions of Lemma 1, and write $P = Q \star R$. If P, Q and R have density functions p, q and r respectively, then we will obtain

$$p(x) = \frac{q(Y_A)\, r(Y_B)}{q_{A\cap B}(Y_{A\cap B})}, \tag{2.8}$$

where the denominator could have been written as $r_{A\cap B}(Y_{A\cap B})$. In particular, $P \in M(A,B)$ if and only if

$$p(Y) = \frac{p_A(Y_A) p_B(Y_B)}{p_{A\cap B}(Y_{A\cap B})}. \tag{2.9}$$

The construction mechanism described above, can be extended to the case of a general decomposable graph $G = (V,E)$. Under the clique set $\mathscr{C} = \{C_1, \cdots, C_k\}$ of a decomposable graph $\mathscr{G}$ that admits a perfect ordering, we consider a pairwise consistent collection of distributions $\{Q_C : C \in \mathscr{C}\}$, where Q_C is a distribution over the respective clique C. Then, a Markov distribution P is defined having Q_C as its margins on cliques by

$$P_{C_1} = Q_{C_1}, \tag{2.10}$$

$$P_{H_{i+1}} = P_{H_i} \star Q_{C_{i+1}}. \tag{2.11}$$

Using Equation 2.10 and Equation 2.11 we can rewrite Equation 2.9 as follows

$$p(Y) = \frac{\prod_{i=1}^{k} p_{C_i}(Y_{C_i})}{\prod_{i=2}^{k} p_{S_i}(Y_{S_i})}. \tag{2.12}$$

Note that separators S_i are the same for any given perfect ordering of cliques. Each respective separator S is repeated $v(S)$ times in any sequence (S_i), where with $v(S)$ Dawid and Lauritzen (1993) denote a combinatorial index which is associated with the number of disconnected components of $G_{V\setminus S}$.

By denoting $\mathscr{S}$ the set of separators that include all $v(S)$ repetitions of each separator S, then Equation 2.12 can be alternatively provided by

$$p(x) = \frac{\prod_{C\in\mathscr{C}} p_C(Y_C)}{\prod_{S\in\mathscr{S}} p_S(Y_S)^{v(S)}}. \tag{2.13}$$

Usually, the term $v(S)$ will be omitted and we will obtain:

$$p(Y) = \frac{\prod_{C\in\mathscr{C}} p_C(Y_C)}{\prod_{S\in\mathscr{S}} p_S(Y_S)}. \tag{2.14}$$

2.4 Hyper Markov Laws for Undirected Graphs

Dawid and Lauritzen (1993) in Section 3 introduced distributional laws for a quantity θ with values in the set $M(G)$ of Markov probabilities over the undirected decomposable graph G. We will exploit these laws to describe prior and posterior distributions for graphical models. We will refer to a distribution of θ over the set $M(G)$ as a law for θ and will be denoted as $\pounds(\theta)$.

For the remainder of this chapter, the notation

$$A \perp\!\!\!\perp B \,|\, C\,[\pounds] \tag{2.15}$$

will imply

$$\theta_A \perp\!\!\!\perp \theta_B \,|\, \theta_C\,[\pounds] \tag{2.16}$$

with respect to the law $\pounds$ for θ. Thus, we state the following lemma provided by Dawid and Lauritzen (1993):

Lemma 2. *It holds that:*

1. *If $A \subseteq V$ then*

$$\theta \simeq (\theta_A,\, \theta_{V\setminus A\,|\,A}). \tag{2.17}$$

2. *If $\mathscr{C}$ is the set of cliques of G then*

$$\theta \simeq \{\theta_C : C \in \mathscr{C}\}. \tag{2.18}$$

3. *If G is collapsible onto $U \subseteq V$ and (A, B) is a decomposition of G_U, then*

$$\theta_U \simeq (\theta_A,\, \theta_B). \tag{2.19}$$

The notation $\simeq$ denotes that each member of a given relation is a function of the other.

Next, we need to define the notion of hyperconsistent laws, to lay the groundwork for the construction mechanism of Hyper Markov laws.

Definition 6. *We say that laws $\mathscr{M}$ over $A \in V$ and $\mathscr{N}$ over $B \in V$ are* ***Hyperconsistent*** *if they both induce the same law over $A \cap B$.*

If $\mathscr{M}$ and $\mathscr{N}$ rise from an appropriate marginalization of a common underlying law, then subsequently both are hyperconsistent. If $A \cap B = \emptyset$, then any pair of laws will be hyperconsistent.

Lemma 3. *Given a hyperconsistent law $\mathscr{M}$ over $A \in V$ and $\mathscr{N}$ over $B \in V$, there exists a unique law $\pounds$ over $A \cap B$ such that:*

1. *$\pounds$ is concentrated on $M(A, B)$.*
2. *$\pounds_A = \mathscr{M}$.*
3. *$\pounds_B = \mathscr{N}$.*
4. *$\theta_A \perp\!\!\!\perp \theta_B \,|\, \theta_{A\cap B}$.*

The law $\mathscr{L}$ which satisfies the conditions of Lemma 3, is called the ***Hyper Markov combination*** of $\mathscr{M}$ and $\mathscr{N}$ and will be denoted by $\mathscr{L} = \mathscr{M} \bigodot \mathscr{N}$.

The primary goal of Dawid and Lauritzen (1993) was to use this scheme to construct a law $\mathscr{L}$ over a graph G, determined by its respective clique marginal laws $\{\mathscr{L}_C : C \in \mathscr{C}\}$. To expedite this construction mechanism, Dawid and Lauritzen (1993) had to establish a restriction on law $\mathscr{L}$ comparable to the Markov requirement on the distribution θ of Y as in Definition 4.

Definition 7. *A law $\mathscr{L}(\theta)$ on $M(G)$ is called (weak)* ***Hyper Markov*** *over G, if for any decomposition (A, B) of G*

$$\theta_A \perp\!\!\!\perp \theta_B \,|\, \theta_{A\cap B}. \tag{2.20}$$

As Dawid and Lauritzen (1993) indicate, this definition is equivalent with the following:

$$\theta_{A\,|\,B} \perp\!\!\!\perp \theta_B \,|\, \theta_{A\cap B} \tag{2.21}$$
$$\theta_{A\,|\,B} \perp\!\!\!\perp \theta_{B\,|\,A} \,|\, \theta_{A\cap B}, \tag{2.22}$$

yet it is different from the corresponding pointwise property

$$\theta_A(Y_A) \perp\!\!\!\perp \theta_B(Y_B) \,|\, \theta_{A\cap B}(Y_{A\cap B}), \quad \forall Y. \tag{2.23}$$

The construction of hyper Markov laws can be achieved using the same scheme of Markov distributions. Given a set of hyperconsistent laws $\{M_C : C \in \mathscr{C}\}$ for a clique set $\mathscr{C}$ of a graph G that admits a perfect ordering, the hyper Markov distribution $\mathscr{L}$ over G having $\{M_C\}$ as its margins on cliques, must satisfy

$$\mathscr{L}_{C_1} = M_{C_1} \tag{2.24}$$

$$\mathscr{L}_{H_{i+1}} = \mathscr{L}_{H_i} \bigodot M_{C_{i+1}}. \tag{2.25}$$

Theorem 1. *The distribution defined by Equation 2.24 and Equation 2.24 is the unique hyper Markov law over G with the given hyperconsistent laws $\{M_C\}$ over clique marginals.*

An important consequence of the hyper Markov property is the following:

Theorem 2. *If the law $\mathscr{L}$ is hyper Markov over G then*

$$\theta_A \perp\!\!\!\perp \theta_B \,|\, \theta_S \,[\mathscr{L}] \tag{2.26}$$

whenever S separates A from B in G.

The hyper Markov property can be maintained under a collapsible marginalization:

Proposition 4. *If a graph G is collapsible onto A and $\mathscr{L}$ is hyper Markov over G, then $\mathscr{L}_A$ is hyper Markov over G_A.*

In order to use laws for a graphical model selection procedure, as we will describe in later parts of this thesis, we need to consider laws with stronger independence properties than the ones stated in Definition 7. Dawid and Lauritzen (1993) provide the following:

Definition 8. *A law $\pounds(\theta)$ on $M(G)$ is a* ***strong hyper Markov*** *over G, if for any decomposition (A, B)* of *G implies that*

$$\theta_{B\,|\,A} \perp\!\!\!\perp \theta_A. \tag{2.27}$$

As Dawid and Lauritzen (1993) indicate, the condition provided by 8 implies that the law $\pounds(\theta)$ is weak hyper Markov. It is stronger than the corresponding pointwise property

$$\theta_{B\,|\,A}(Y_B\,|\,Y_A) \perp\!\!\!\perp \theta_A(Y_A), \quad \forall Y. \tag{2.28}$$

Proposition 5. *A law $\pounds(\theta)$ on $M(G)$ is strong hyper Markov over G if and only if, under $\pounds$*

$$\perp\!\!\!\perp_{\theta_{A\,|\,B}, \theta_{B\,|\,A}, \theta_{A\cap B}} \tag{2.29}$$

whenever $A \cap B$ is complete and separated A from B.

We can induce a unique hyper Markov law over a graph G dependent on its clique-marginal laws, which can be formally expressed in the following proposition:

Proposition 6. *Let $\pounds$ be a hyper Markov law over G. Then $\pounds$ is strong hyper Markov if and only if, for all cliques of $\mathscr{C}$* of *G and all subsets A of $\mathscr{C}$ we have*

$$\theta_{C\setminus A\,|\,A} \perp\!\!\!\perp \theta_A\,[\pounds]. \tag{2.30}$$

2.5 Hyper Markov Laws for Bayesian inference

Hyper Markov laws can be used as prior distributions for a parameter θ and the hyper Markov property can simplify prior to posterior analysis or graphical model selection procedures. For this section, Y will denote an observation from a distribution θ with its respective law $\pounds$.

Proposition 7. *If the prior law $\pounds(\theta)$ is hyper Markov over G, then the joint distribution (Y, θ) satisfies, for any decomposition (A, B) of G,*

$$(Y_A, \theta_A) \perp\!\!\!\perp (Y_B, \theta_B)\,|\,(Y_{A\cap B}, \theta_{A\cap B}). \tag{2.31}$$

If the law $\pounds(\theta)$ is strong hyper Markov, then it also satisfies:

$$(Y_A, \theta_A) \perp\!\!\!\perp (Y_B, \theta_{B\,|\,A})\,|\,Y_{A\cap B}. \tag{2.32}$$

Dawid & Larutizen (1993) further conditioned with Y_A and Y_B the following relations Equation 2.31 and Equation 7 to obtain the following corollary

Corollary 2.1. *If the prior law of θ is hyper Markov, so is the posterior law obtained on the complete data $Y = y$. If the prior law is strong hyper Markov, so is the posterior.*

Dawid and Lauritzen (1993) note that Corollary 2.1 can be extended to the case where the data are provided as a random sample of size n from the distribution *theta*, since observations are usually assumed as i.i.d. realizations. Thus, Corollary 2.1 indicates that hyper Markov laws and strong hyper Markov laws can form conjugate families of the family $M(G)$ of Markov models over graph G.

Corollary 2.2. *If the prior law $\pounds(\theta)$ is strong hyper Markov, the posterior law of θ is the unique (strong) hyper Markov law $\pounds^*$ specified by the clique-marginal laws $\{\pounds^*_C : C \in \mathscr{C}$, where $\pounds^*_C$ is the posterior distribution of θ_C based on its prior law $\pounds_C$ and the clique-specific data $Y_C = y_C$. When densities exist, $\pi(\theta_C|y_C) \propto f(y_C|\theta_C)\pi(\theta_C)$.*

Thus, with a strong hyper Markov prior distribution, the posterior distribution can be provided piece-wise over the clique set $\mathscr{C}$ and update the marginal distribution in the same manner, a feature not available when one considers weak hyper Markov laws.

Marginal data distributions are extensively used in model selection procedures, as we will describe later on this thesis. The Markov property can be extended to a marginal distribution, i.e. distribution of the data not conditioned on parameters, using the following proposition by Dawid and Lauritzen (1993):

Proposition 8. *If the prior law of θ is strong hyper Markov, then the marginal distribution of Y is Markov.*

This property allows marginal data distributions to be decomposed accordingly to Equation 2.14 and expedite graphical model selection procedures, as we will explore in later parts of this thesis.

Chapter 3

Objective Bayes Model Selection

As mentioned in chapter 1, this thesis deals with the graphical model selection problem under an objective Bayes viewpoint. The goal of this chapter is to communicate to the reader how Bayesian statistics deal with the model selection problem, what kind of challenges one faces with in absense of information and how objective Bayes can fill this gap.

We first provide the general scope of Bayesian model selection as well as all the key quantities required. We then present a computational approach for estimating Bayes factors. Note that we describe a specific technique, namely the Importance Sampling approach, because it will be intrumental for approximating Bayes factors in chapter 4; further references provided for alternative methods.

After setting the groundwork for comparing models under a Bayesian point of view, we then present the challenges of Bayesian model selection and how objective Bayes operates. Using a non-informative setup (see section 3.3) we proceed with defining alternative Bayes factors that can facilitate the comparison of models that utilize improper priors.

Finally, we shift from the pairwise comparison of models to an automated construction of minimally informative prior distributions that allow model comparison using a common reference model (see subsection 3.5.3). Notions and strategies communicated to this chapter, will be combined with the notions of graph theory presented in chapter 2, to facilitate an objective Bayes approach to the graphical model selection problem.

3.1 Bayesian Model Selection

One of the core objectives of Statistical Science is the development of a ***Statistical Model*** for either interpreting casual relationships between characteristics of a population, like a relationship between lung cancer and dyspnoea, or predicting a future outcome, like the future price of a stock. In practice, we are called to explore a set of candidate statistical models to identify the most promising ones, in terms of either prediction or interpretation. This procedure is described as the ***Model Selection Problem*** and it has been explored extensively from both Frequentist and Bayesian point of view.

For the remainder of this chapter, we will use the following terms and notation. With the term statistical model, we refer to a family of distributions for the observable random variables (Consonni et al. (2018) Section 3.1). Let $\mathscr{M} = \{M_1, \cdots, M_k\}$, where $k \geq 2$, denote a countable set of statistical models, $Y = \{Y_1, \cdots, Y_q\}$ be a collection of q variables from which we observe n i.i.d. q-dimensional observations $\mathbf{y}_i$ for $i = 1, \cdots, n$ arranged in a data matrix $\mathbf{Y}$ as in Equation 1.1. With $f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)$, we denote the density of $\mathbf{Y}$ under model $M_j \in \mathscr{M}$. To compare two models of the set $\mathscr{M}$, we consider

$$\begin{aligned} Model\ M_i &: f(\mathbf{Y}|\boldsymbol{\theta}_i, M_i), \boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i \subset \mathbb{R}^{d_i}, \\ Model\ M_j &: f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j), \boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j \subset \mathbb{R}^{d_j}, \end{aligned}$$

where $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, with respective dimensions d_i and d_j, represent the unknown parameters of models M_i and M_j. When M_i is nested in model M_j we assume that $\boldsymbol{\theta}_j = (\boldsymbol{\theta}_i, \boldsymbol{\theta}_{j\setminus i})^T$, where $\boldsymbol{\theta}_i$ represents the parameters shared by both models M_i and M_j and $\boldsymbol{\theta}_j$ is contained only in model M_j. This specification will prove useful for specifying compatible prior distributions across models of set $\mathscr{M}$, as we will present in later sections of this chapter. Under a nested-model setup, we single out a specific model $M_0 \in \mathscr{M}$, which is nested in every other model $M_j \in \mathscr{M}$ with $j \neq 0$ and it is called the ***null model***.

For each model $M_j \in \mathscr{M}$, we assign a prior probability $\pi(M_j)$ which represents our prior belief that M_j is the true model, and a prior distribution $\pi(\boldsymbol{\theta}_j|M_j)$ on the parameter vector $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j \subset \mathbb{R}^{d_j}$. By using Bayes theorem, we obtain the ***Posterior Model Probability*** of model M_j by

$$\pi(M_j|\mathbf{Y}) = \frac{\pi(M_j) m_j(\mathbf{Y})}{\sum_{M_l \in \mathscr{M}} \pi(M_l) m_l(\mathbf{Y})}. \tag{3.1}$$

The quantity $m_j(\mathbf{Y})$ denotes the marginal likelihood of the data vector $\mathbf{Y}$ under model $M_j \in \mathscr{M}$ and it is obtained by integrating over the parameter vector $\boldsymbol{\theta}_j$, i.e.

$$m_j(\mathbf{Y}) = \int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)\pi(\boldsymbol{\theta}_j|M_j) d\boldsymbol{\theta}_j. \tag{3.2}$$

The most common approach for comparing two models $M_j, M_i \in \mathscr{M}$ is the production of the ***Posterior Odds*** (Jeffreys (1961)) of model M_j versus model M_i , defined by

$$\begin{aligned} PO_{M_j:M_i} &= \frac{\pi(M_j|\mathbf{Y})}{\pi(M_i|\mathbf{Y})} = \frac{m_j(\mathbf{Y})}{m_i(\mathbf{Y})} \times \frac{\pi(M_j)}{\pi(M_i)} \\ &= BF_{M_j:M_i}(\mathbf{Y}) \times O_{M_j:M_i}, \end{aligned} \tag{3.3}$$

which represent the odds that model M_j will perform better than model M_i, based on their respective posterior model probabilities. The quantity

$$BF_{M_j:M_i}(\mathbf{Y}) = \frac{m_j(\mathbf{Y})}{m_i(\mathbf{Y})} \tag{3.4}$$

denotes the ***Bayes Factor*** (***BF***) of model M_j versus model M_i and represents the evidence in favor of model M_j based on information provided by the data $\mathbf{Y}$. The quantity

$$O_{M_j:M_i} = \frac{\pi(M_j)}{\pi(M_i)} \tag{3.5}$$

denotes the ***Prior Odds*** of model M_j versus model M_i which represents our prior belief that model M_j will perform better than model M_i. Using the Bayes factor of Equation 3.4 we can define in a straightforward manner the Bayes factor of model M_i versus model M_j by

$$BF_{M_i:M_j}(\mathbf{Y}) = \frac{1}{BF_{M_j:M_i}(\mathbf{Y})}. \tag{3.6}$$

The Bayes factor is not influenced by prior model probabilities, thus by assigning a uniform probability across all models of $\mathscr{M}$, i.e. we assume that every model $M_j \in \mathscr{M}$ has the exact same prior probability, the Posterior odds of Equation 3.3 are reduced to the Bayes factor of Equation 3.4. Using Equation 3.3 we can provide another interpretation for the Bayes factor, that is, the ratio of Posterior to Prior Odds.

Larger values of Posterior odds (or Bayes factors) correspond to stronger evidence in favor of model M_j against model M_i and smaller values indicate the opposite behavior. Jeffreys (1961) and Kass and Raftery (1995) provide specific thresholds of Bayes and log-Bayes factors that measure the amount of evidence in favor of each model.

After establishing the pairwise comparison of candidate models of set $\mathscr{M}$, we can define the Posterior model probabilities of a candidate model $M_j \in \mathscr{M}$ using Posterior Odds of every model of $\mathscr{M}$ versus null model, i.e.

$$\begin{aligned}\pi(M_j|\mathbf{Y}) &= \frac{\pi(M_j)m_j(\mathbf{Y})}{\sum_{M_l \in \mathscr{M}} \pi(M_l)m_l(\mathbf{Y})} \\ &= \frac{\pi(M_j)m_j(\mathbf{Y})}{\sum_{M_l \in \mathscr{M}} \pi(M_l)m_l(\mathbf{Y})} \frac{\pi(M_0)m_0(\mathbf{Y})}{\pi(M_0)m_0(\mathbf{Y})} \\ &= \frac{PO_{M_j:M_0}}{\sum_{M_l \in \mathscr{M}} PO_{M_l:M_0}}. \end{aligned} \tag{3.7}$$

The model that achieves the highest posterior probability compared to every other model of set $\mathscr{M}$, is defined as the ***Maximum a-Posteriori Model*** and we will refer to it as the ***MAP*** model.

The main disadvantage of the posterior model probabilities, compared to posterior odds and Bayes factors, is their decline over similar models. If one considers even larger model spaces, then the posterior model probabilities decrease, even for the MAP model. A useful suggestion by Consonni et al. (2018) Section 3.2 is, besides the posterior model probabilities, to investigate the posterior odds or the Bayes factors of each model considered against the MAP model.

3.2 Monte Carlo Estimation of Bayes Factors

The derivation of posterior model probabilities requires the calculations of BFs using marginal likelihoods of relation Equation 3.2, which can be analytically tractable only in low dimension settings or under a conjugate setup; for larger dimension settings, numerical approximations or Monte Carlo schemes are deployed. The most well known numerical approximations are the Laplace approximation as was used by Tierney and Kadane (1989) and the Schwarz criterion (Schwarz (1978)), but both these approximations are plausible only for certain simple problems as indicated by Berger and Pericchi (1998a).

An alternative approach is provided using MCMC methods in conjunction with Importance Sampling, which will only be presented in this subsection since it will be used for the utilization of Expected and Power-Expected Posterior prior approaches; see subsection 3.5.4

Under a given model $M_j \in \mathscr{M}$, a simple Monte Carlo estimator of the marginal likelihood of Equation 3.2 can be provided by

$$\widehat{m_j}(\mathbf{Y}) = \frac{1}{R}\sum_{r=1}^{R} f(\mathbf{Y}|\boldsymbol{\theta}^{(r)}, M_j) \tag{3.8}$$

where $\{\boldsymbol{\theta}^{(r)}, r = 1, \cdots, R\}$ is a sample obtained by the prior distribution $\pi(\boldsymbol{\theta}_j|M_j)$. When one considers diffused priors or highly concentrated likelihoods, this rough average of likelihoods will return unstable estimations, because it will be guided by a few large values of the likelihood. Moreover, this estimator will have large variances and its convergence to its actual value will be very slow. To obtain more accurate Monte Carlo estimations of the marginal likelihood Equation 3.2, we deploy ***Importance Sampling***.

Let $g^*(\boldsymbol{\theta}_j|M_j)$ be an importance density, which will be used to generate a sample of $\{\boldsymbol{\theta}^{(r)}, r = 1, \cdots, R\}$. Then, the marginal likelihood Equation 3.2 can be estimated by

$$\widehat{m}_j(\mathbf{Y}) = \frac{\sum_{r=1}^{R} w_r^* f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)}{\sum_{r=1}^{R} w_r^*} \tag{3.9}$$

where $w_r^* = \pi(\boldsymbol{\theta}_j|M_j)/g^*(\boldsymbol{\theta}_j|M_j)$. One intriguing choice of the importance density $g^*(\boldsymbol{\theta}_j|M_j)$ is the posterior distribution of $\pi(\boldsymbol{\theta}_j|\mathbf{Y}, M_j)$, which results to the following estimator

$$\widehat{m}_j(\mathbf{Y}) = \left\{\frac{1}{R}\sum_{r=1}^{R}[f(\mathbf{Y}|\boldsymbol{\theta}_j^{(r)}, M_j)]^{-1}\right\}^{-1}. \tag{3.10}$$

More information on this topic is provided by Kass and Raftery (1995) Section 4.3.

3.3 The Objective Bayes Approach

The derivation of Bayes factors requires the specification of a prior distribution $\pi(\boldsymbol{\theta}_j|M_j)$ for the parameters $\boldsymbol{\theta}_j$ under each model $M_j \in \mathscr{M}$. There are two stances to be adopted to assign prior distributions; a subjective stance, where one can successfully elicit a prior distribution to quantify his knowledge and beliefs for $\boldsymbol{\theta}_j$ and an objective stance where one depicts his prior ignorance or lack of knowledge for $\boldsymbol{\theta}_j$ and the prior distributions are used up to suitably defining statistical models. These stances are formally called ***Subjective Bayes*** and ***Objective Bayes*** approaches.

The Objective Bayes approach has received a great deal of criticism by researchers in the Bayesian community regarding its philosophical viewpoints (Berger (2004) Section 1.2). In real life applications, it may be difficult for a practitioner to encapsulate his/her prior opinion into a suitable prior distribution. This complication rises when we consider high-dimensional parameter spaces, where it is infeasible to illustrate the dependence structure among parameters through a prior distribution. It is also possible that for several applications there is no prior knowledge and yet we must depict our prior ignorance to a prior distribution. Thus, in this thesis, we will focus only on Objective Bayes approaches.

The distributions that are used under an Objective Bayes approach are called ***non-informative prior distributions***. For the remainder of this thesis, we will refer to a non-informative prior distribution for a parameter vector $\boldsymbol{\theta}_j$ under model $M_j \in \mathscr{M}$ as $\pi^N(\boldsymbol{\theta}_j|M_j)$. When we consider ***improper*** prior distributions, i.e. distributions with total mass not finite, we assume that they depend from an arbitrary normalizing constant c_j which is responsible for converting their total mass to finite. Non-informative distributions are characterized by the fact that their effect on the posterior analysis is minimal and our results are guided by the data at hand.

The main difficulties one meets in search of automatic default Objective Bayes model selection procedures according to Berger and Pericchi (2001) Section 1.5, are the following:

1. ***Computational cost can be enormous***.
   When the parameter or model space grows rapidly, the computational burden for the derivation of Bayes factors and Posterior odds can become enormous, especially in variable selection and graphical model selection problems.

2. ***Use of improper priors on different parameter spaces, yields indeterminate answers***.
   If we consider models $M_i, M_j \in \mathscr{M}$ and their respective improper non-informative priors $\pi^N(\boldsymbol{\theta}_i|M_i), \pi^N(\boldsymbol{\theta}_j|M_j)$, the Bayes factor of model M_j versus model M_i will be provided by:

$$\begin{aligned} BF_{M_j:M_i}(\mathbf{Y}) &= \frac{m_j^N(\mathbf{Y})}{m_j^N(\mathbf{Y})} \\ &= \frac{\int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)\pi^N(\boldsymbol{\theta}_j|M_j)}{\int f(\mathbf{Y}|\boldsymbol{\theta}_i, M_i)\pi^N(\boldsymbol{\theta}_i|M_i)} \\ &= \frac{\int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)c_j\pi^N(\boldsymbol{\theta}_j|M_j)}{\int f(\mathbf{Y}|\boldsymbol{\theta}_i, M_i)c_i\pi^N(\boldsymbol{\theta}_i|M_i)} \\ &= \frac{c_j}{c_i}\frac{\int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)\pi^N(\boldsymbol{\theta}_j|M_j)}{\int f(\mathbf{Y}|\boldsymbol{\theta}_i, M_i)\pi^N(\boldsymbol{\theta}_i|M_i)} \\ &= \frac{c_j}{c_i} BF^N_{M_j:M_i}(\mathbf{Y}). \end{aligned} \tag{3.11}$$

We observe that with different parameter spaces between models, the Bayes factor of M_j versus model M_i depends from the ratio of arbitrary constants c_j/c_i, thus we are not able to extract any information regarding which model performs better.

As Berger and Pericchi (2001) indicate, even when we compare models with identical parameter spaces it is not rational to assume $c_j = c_i$ since we refer to a completely arbitrary constant. If we consider similar model spaces, it may not be prohibitive to set $c_j = c_i$, yet it is not suggested. Solution to this issue will be investigated extensively in the following sections.

3. ***Model parameters do not maintain their meaning across models and prior distributions should adapt***.
   This complication usually rises in the variable selection and Berger and Pericchi (2001) provided a simple example of a variable selection problem, where they consider two candidate models having one covariate in common. If one decides to assign the same prior distribution regarding the coefficient of the common covariate, it will result in irrational results since under the second model there exists positive correlation among the two given covariates.

Under an objective Bayes appraoch, we start by defining a default non-informative prior distributions and we present the most important of them as were described by Consonni et al. (2018) Section 2. These are the following:

1. ***Uniform prior distribution***.
   With a uniform prior distribution, we assign a prior to a continuous parameter which provides equal probability over equal sized intervals. Yet as Consonni et al. (2018) indicate, the uniform prior is not invariant under reparametrization and in real life applications is impossible to find a natural parametrization for a given model. If the model space is not bounded then the uniform prior will be improper, thus we cannot be certain that the posterior will be proper.

2. ***Invariant prior distribution***.
   Since the uniform prior is not invariant under re-parametrization, the development of a suitable objective prior that would be invariant under a certain

class of transformation was necessary. Following the notation provided by Consonni et al. (2018), let $(\mathbf{P}, \boldsymbol{\Theta})$ denote a statistical model for the observation X, where $\mathbf{P}$ denotes a family of distributions and $\boldsymbol{\theta}$ is the set of parameters. Consider the transformation $Y = s(X)$ with distribution model $\mathbf{P}$ and set of parameters $\boldsymbol{\Lambda}$. Since the family of distributions $\boldsymbol{P}$ is common for both X and Y, the model is invariant to a transformation $s(\cdot)$. If we define our prior through family $\mathbf{P}$, then the priors π_{θ} of $\boldsymbol{\theta}$ and π_{λ} of $\boldsymbol{\lambda}$ must be defined such that $\mathbf{P}^{\pi_\theta}\{\boldsymbol{\theta} \in A\} = \mathbf{P}^{\pi_\lambda}\{\boldsymbol{\lambda} \in A\}, \forall A$. This attribute is named ***Context Invariance*** in Dawid (2006) and represents a very strong requirement, since the structure of $\mathbf{P}$ remains the same regardless of the framework applied.

3. ***Matching prior distribution***.
   The matching prior is based on the principle that a non-informative prior should have similar inferential capabilities with some standard frequentist approaches. If we consider an example where we compare credible and confidence intervals, the probability matching prior should lead to posterior probabilities of certain regions which are the identical or approximately equal with their respective frequentist coverage probabilities; more details on Datta and Mukerjee (2004).

4. ***Maximum entropy prior***.
   The maximum entropy prior is a distribution which maximizes the entropy over a class of priors under some certain constraints. Let $\pi(\boldsymbol{\theta})$ denote a distribution, then its entropy will be provided by

$$Ent(\pi) = -\int_{\Theta} \pi(\boldsymbol{\theta}) \log \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{3.12}$$

   which is able to quantify the absence of information from a non-informative prior distribution.

   This approach is based on two steps. First, one defined a class Γ of candidate prior distributions characterized on k constraints, which can be in the form of quantiles or moments and can be expressed as

$$\mathbb{E}(g_j(\boldsymbol{\theta})) = \mu_j, \;\; j = 1, \cdots, k, \tag{3.13}$$

   where $g_j(\cdot)$ is a set of suitable functions. To finalize the selection of the maximum entropy prior, one should select the element of the set Γ that maximizes the entropy $Ent(\pi)$.

5. ***Jeffreys and reference prior***.
   Before the emergence of MCMC methods, the most common choice of a prior distribution under an objective Bayes framework was the Jeffreys prior, which is provided by

$$\pi^J(\boldsymbol{\theta}) \propto det(I(\boldsymbol{\theta}))^{1/2}. \tag{3.14}$$

   The term $I(\boldsymbol{\theta})$ denotes the Fisher information matrix, where under the assumption of a continuous parameter space $\boldsymbol{\Theta}$, its elements are provided by

$$I_{ij}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\Big(\frac{\partial^2}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \log f(\mathbf{Y}|\boldsymbol{\theta})\Big), \tag{3.15}$$

where the expectation $\mathbb{E}_{\boldsymbol{\theta}}$ denotes the expected value over the sampling space for a given value of the parameter $\boldsymbol{\theta}$, and $\mathbf{Y}$ is the observable random variable.

Jeffreys prior is remains invariant under re-parametrization and enjoys many optimality properties in the absence of nuisance parameters. It achieves the maximum asymptotic divergence between the prior and the posterior for $\boldsymbol{\theta}$ under several different metrics, and for scalar $\boldsymbol{\theta}$ Jeffreys prior is a second order matching prior.

Though it is the most widely used objective prior distribution, it has some certain flaws that need to be taken into account. First, it does not always lead to proper posterior distributions for all possible datasets, as was proved with counterexamples in Ye and Berger (1991) and Berger et al. (2001). Jeffreys prior was originally developed for scalar parameters, yet for larger dimensions, it may lead to incoherence and paradoxes, as was pointed out by Dawid et al. (1973).

Second, it was suggested by Jeffreys to separately deal with location parameters. If $\boldsymbol{\theta} = \{\phi, \lambda\}$ where ϕ denotes a vector of location parameters, then it is advised to keep ϕ fixed and use $\pi^J(\boldsymbol{\theta}) \propto (det(\lambda))^{1/2}$. This prior is called by Kass and Wasserman (1996) non-location Jeffreys prior.

Third, when we consider a low dimensional function $\psi(\boldsymbol{\theta})$ of the entire parameter vector $\boldsymbol{\theta}$, there is no guarantee of a "satisfactory" behavior of Jeffreys approach. The word ‘satisfactory’ is used from Consonni et al. (2018) to describe the ability of Jeffreys approach to producing statistical procedures which correspond to frequentist procedures for the repeated sampling framework.

3.4 Model Selection Under Objective Bayes Approach

Following the description of default prior distributions, we proceed with establishing the groundwork that will lead to a well-defined construction of standard Objective Bayes model selection procedures.

3.4.1 Principles for Objective Bayes Model Comparison

Bayarri et al. (2012) proposed a general framework in terms of criteria, under which an objective prior distribution should operate. Note, that several of these criteria are applicable only in nested model comparison. These where described extensively in Consonni et al. (2018) Section 3.3 as follows:

1. ***Propriety***: The prior distribution of each specific model parameter conditionally on the common ones, $\pi(\boldsymbol{\theta}_{j\setminus 0}|\boldsymbol{\theta}_0, M_j)$, must be proper in order to deliver identifiable Bayes factors without arbitrary normalizing constants.

2. ***Model Selection Consistency***: If M_j is the model that generated the data, then the posterior probability of M_j should converge to 1 as the sample size grows to infinity. Though it is one of the strongest requirements for developing

objective priors, yet it cannot provide evidence in favor of a candidate objective prior among several that satisfy model selection consistency.

3. ***Information Consistency***: If there exists a sequence of identical-sized datasets such that the likelihood ratio between model M_j and model M_0 goes to infinity, then the respective sequence of Bayes factors will also go to infinity. This form of consistency was originally described in Berger and Pericchi (2001) under a conjugate prior setup for location with unknown scale.

4. ***Intrinsic Consistency Criterion***: Effects that rise from model structure, such as sample size, should diminish as n grows to infinity leading to a limiting proper prior

5. ***Predictive Matching***: Under a minimal sample size an Objective Bayes model selection procedure should not be able to discriminate among two competing models, with the resulting Bayes factor converging to one for all minimally sized samples.

6. ***Measurement Invariance***: Answers provided by an Objective Bayes model selection procedure should not be affected by changes in measurement units.

7. ***Group Invariance Criterion***: If models M_j and M_0 remain invariant with respect to a group of transformation Γ_0, then the respective conditional priors $\pi(\boldsymbol{\theta}_{j\setminus 0}|\boldsymbol{\theta}_0, M_j)$ must be defined accordingly to provide an invariant marginal distribution $f(\mathbf{Y}|\boldsymbol{\theta}_0, M_j)$ under Γ_0.

Prior Compatibility

The assignment of prior distributions is focused on encapsulating the level of uncertainty regarding a model parameter, yet it should also contain features that are relevant across models of set $\mathscr{M}$. This feature can be formally expressed as ***Prior Compatibility***; see Consonni and Veronese (2008). In essence, the prior compatibility ensures that prior distributions are related across models, which is not a requirement, each being conditional on a given model. Usually, it is applied in comparison of nested models with different parameter spaces, where all models are compared to a benchmark model (e.g. the null model). In this way, prior compatibility can be achieved between each model and the null model and by implication between any pair of models.

Compatibility was originally proposed to alleviate the sensitivity of model comparison to prior specifications and allow the extraction of multiple prior distributions when a large number of models is considered. Yet, it can influence the construction of Objective Prior distributions, e.g. the Expected Posterior Priors that we will describe in latter parts of this chapter, where all prior distributions are defined up to a common predictive measure.

3.4.2 Partial Bayes Factors

Before the development of automated Objective Bayes model selection procedures, the use of improper default priors was allowable by deploying a specific type of Bayes factors, namely ***Partial Bayes Factors***. Key feature of these variations was the use of a part of the data as a training sample to eliminate the indeterminacy arising from ratios of arbitrary constants. The remainder of the data would be deployed to facilitate a standard model selection procedure by constructing BFs and Posterior odds. Most well known partial Bayes factors are the Fractional Bayes Factor of O'Hagan (1995) and the Intrinsic Bayes Factor of Berger and Pericchi (1996).

Fractional Bayes Factor

O'Hagan (1995) introduced a partial Bayes factor whose primary aim was to operate in absence of concrete prior information and deal with the main issue of the dependence of BFs from ratios of arbitrary constants. This approach was name the ***Fractional Bayes Factor*** (***FBF***) approach and it is based on a simple and intuitive idea. One *fraction* of the full sample likelihood will be used for model specification, i.e. to specify suitable prior distributions, and the remaining part of the full sample likelihood will be used for facilitating the model selection procedure. Let $b = b(n)$, where $0 < b < 1$, be the fraction parameter which is a fraction of the data. Given two models M_j and M_i of set $\mathscr{M}$, we define the Fractional Bayes factor of model M_j versus model M_i as

$$FBF_{M_j:M_i}(b, \mathbf{Y}) = \frac{Q_j(b, \mathbf{Y})}{Q_i(b, \mathbf{Y})}, \tag{3.16}$$

where

$$Q_j(b, \mathbf{Y}) = \frac{\int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)\pi^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)^b\pi^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}, \tag{3.17}$$

$\forall j \in \{1, \cdots, |\mathscr{M}|\}$. The quantity $f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)^b$ represents the likelihood function under model M_j raised to the power b. We can rewrite $Q_j(b, \mathbf{Y})$ as

$$\begin{aligned} Q_j(b, \mathbf{Y}) &= \frac{\int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)\pi^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)^b\pi^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j} \\ &= \frac{\int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)^{1-b}f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)^b\pi^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{m_j(\mathbf{Y}; b)} \\ &= \int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)^{1-b}\frac{f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)^b\pi^N(\boldsymbol{\theta}_j)}{m_j(\mathbf{Y}; b)}d\boldsymbol{\theta}_j \\ &= \int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)^{1-b}\pi^F(\boldsymbol{\theta}_j|\mathbf{Y}, b)d\boldsymbol{\theta}_j. \end{aligned} \tag{3.18}$$

where $m_j(\mathbf{Y}; b)$ represents the marginal likelihood of model M_j using the fraction b of the likelihood, $f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)^b$, and $\pi^F(\boldsymbol{\theta}_j|\mathbf{Y}, b)$ is the implied the ***Fractional Prior*** of model M_j. The fractional prior is actually the posterior distribution of $\boldsymbol{\theta}_j$ given the fraction b of the likelihood $f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)^b$ and the default prior under model M_j. Using this form we can easier distinguish the two fractions of the likelihood function and their respective roles on the model selection procedure. Thus, we are able to produce automatically prior distributions for any given model $M_j \in \mathscr{M}$ without any restrictions on model comparison, i.e. comparing only nested models.

It is evident that the Fractional Bayes factor is heavily dependent on the choice of the fraction parameter b. Under an objective Bayes approach, b must be relatively small, in order to minimize the effect of the fractional prior. O'Hagan (1995) provided three basic choices of the fractional parameter b based on the robustness of the Fractional Bayes factor, which are

1. $b = m_0/n$ when there is not any concern regarding robustness,
2. $b = n^{-1}\max\{m_0, \sqrt{n}\}$ when there is high concern regarding robustness,
3. $b = n^{-1}\max\{m_0, \log n\}$ as a more common choice,

where m_0 represents a minimal training sample with $m_0 < n$ such that the fractional likelihood will be well defined.

One disadvantage of the FBF approach, besides the double usage of data, is the absence of a formal definition for the selection of the fraction parameter b. As Berger and Pericchi (2001) Section 5.1 indicate, one may argue that the FBF approach does not correspond to an actual Bayes factor asymptotically. The choice of b is determined in practice, where the most usual choice is the one provided by O'Hagan (1995) $b = m_0/n$, when m_0 can be identified. Berger and Pericchi (2001) also provided applications in which the FBF couldn't cope with certain cases (see Sections 4.1 and 4.4 of the respective paper), which eventually led to more evolved definitional issues of the approach.

Intrinsic Bayes Factor

One of the goals of an Objective Bayes approach is to utilize fully automatic model selection procedures, that are able to compare models of different dimensions and non-nested models. That was achieved up to a point, with the use of conventional prior distributions or crude approximations of BFs, without proper Bayesian basis. Berger and Pericchi (1996) successfully tackled these issues with the development of the ***Intrinsic Bayes Factor*** (***IBF***) and variations of it. It was the first automated model selection procedure that actually led to the development of objective prior distributions, namely the Intrinsic Priors, which will be extensively explored in the following section.

The IBF uses a part of the data as a training set for eliminating the indeterminacy that arises from the use of improper priors and the rest of the data for employing model selection procedures. Let $\mathbf{Y}$ be a vector or observations and $\boldsymbol{\theta}_j$ be the parameter vector under model $M_j \in \mathscr{M}$. We define $\mathbf{Y}(l)$ as a training sample, which is actually a sub-sample of the original data $\mathbf{Y}$. Then, the posterior distribution of $\boldsymbol{\theta}_j$ given the training sample $\mathbf{Y}(l)$ will be provided by

$$\pi^N(\boldsymbol{\theta}_j|\mathbf{Y}(l), M_j) = \frac{f(\mathbf{Y}(l)|\boldsymbol{\theta}_j, M_j)\pi^N(\boldsymbol{\theta}_j|M_j)}{m^N(\mathbf{Y}(l))}. \tag{3.19}$$

The idea is to use this posterior distribution, under the assumption that it is proper, as a prior distribution over $\boldsymbol{\theta}_j$, and then calculate Bayes factors with the remaining data $\mathbf{Y}(-l)$. Thus the IBF of model M_j versus model M_i conditional on $\mathbf{Y}(l)$ will be provided by

$$IBF_{M_j:M_i}(\mathbf{Y}(l)) = BF^N_{M_j:M_i}(\mathbf{Y})BF^N_{M_i:M_j}(\mathbf{Y}(l)), \tag{3.20}$$

where

$$BF^N_{M_i:M_j}(\mathbf{Y}(l)) = \frac{m_i^N(\mathbf{Y}(l))}{m_j^N(\mathbf{Y}(l))}. \tag{3.21}$$

It is evident that the Intrinsic Bayes factor of Equation 3.20 is not influenced by arbitrary normalizing constants, since the two components of Equation 3.20 provide inverse ratios that cancel out. The use of the training sample $\mathbf{Y}(l)$ is allowed only if the marginal likelihood in the denominator of Equation 3.19 is proper.

This leads to the definition of a proper training sample, that is the training sample $\{\mathbf{Y}(l) : 0 < m_j^N(\mathbf{Y}(l)) < \infty, \forall M_j \in \mathscr{M}\}$. A training sample $\mathbf{Y}(l)$ is called minimal training sample if there is no subset of it that is proper and they are necessary for keeping the impact of the prior distribution at a minimum. Usually, the size of a minimal training sample is defined by the number of the identifiable parameters of all models in $\mathscr{M}$.

Since the training sample $\mathbf{Y}(l)$ is consisted by a part of the original data $\mathbf{Y}$, we can define many different minimal training samples, thus we define as

$$\mathscr{Y}_T = \{\mathbf{Y}(1), \mathbf{Y}(2), \cdots, \mathbf{Y}(L)\} \tag{3.22}$$

the set of all available minimal training samples. The IBF, as was defined in Equation 3.20, heavily depends on the choice of a minimal training sample $\mathbf{Y}(l) \in \mathscr{Y}_T$ through the Bayes factor $BF^N_{M_i:M_j}(\mathbf{Y}(l))$. In order to alleviate the IBF from its dependence on the choice of a minimal training sample and increase its stability, one can provide averages (arithmetic and geometric) of the IBF with respect to $\mathscr{Y}_T$. Using Equation 3.20, we define the ***Arithmetic*** (***AIBF***) and ***Geometric*** (***GIBF***) Intrinsic Bayes Factor of model M_j versus model M_i as

$$IBF^A_{M_j:M_i}(\mathbf{Y}) = BF^N_{M_j:M_i}(\mathbf{Y})\frac{1}{L}\sum_{l=1}^{N} BF^N_{M_i:M_j}(\mathbf{Y}(l)) \tag{3.23}$$

and

$$IBF^G_{M_j:M_i}(\mathbf{Y}) = BF^N_{M_j:M_i}(\mathbf{Y})\Big(\prod_{l=1}^{N} BF^N_{M_i:M_j}(\mathbf{Y}(l))\Big)^{1/L} \tag{3.24}$$

respectively. Generally, the geometric mean is always less or equal with the arithmetic mean, which also applies in this context, i.e. $IBF^G_{M_j:M_i}(\mathbf{Y}) \leq IBF^A_{M_j:M_i}(\mathbf{Y})$, therefore the geometric IBF will support the simpler model to a greater extent than the arithmetic IBF.

The Arithmetic and Geometric IBFs are applied when a considerable amount of data is available. In cases where we operate with few data, the averages of Arithmetic and Geometric IBF can have a large variances which leads to unstable IBFs. Berger and Pericchi (1996) introduced the ***Expected Arithmetic (EAIBF)*** and

Geometric Intrinsic Bayes factors (EGIBF) to deal with this complication, which are provided by replacing the averages of Equation 3.23 and Equation 3.24 by their expectations, evaluated at the respective MLE. Thus, the EAIBF and EGIBF of model M_j versus model M_i will be provided by

$$IBF^{EA}_{M_j:M_i}(\mathbf{Y}) = BF^{N}_{M_j:M_i}(\mathbf{Y})\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}^{M_j}_{\hat{\boldsymbol{\theta}}_j}[BF^{N}_{M_i:M_j}(\mathbf{Y}(l))] \tag{3.25}$$

and

$$IBF^{EG}_{M_j:M_i}(\mathbf{Y}) = BF^{N}_{M_j:M_i}(\mathbf{Y})\exp\left\{\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}^{M_j}_{\hat{\boldsymbol{\theta}}_j}[logBF^{N}_{M_i:M_j}(\mathbf{Y}(l))]\right\}, \tag{3.26}$$

where both o f the above expectations are provided with respect to model M_j and $\boldsymbol{\theta}_j$ is replaced by its respective MLE $\hat{\boldsymbol{\theta}}_j$. These versions of the IBF are useful when one operates with larger data-sets, since the computational cost that arises with the excessive summation terms of Arithmetic and Geometric IBF can be significantly reduced.

The presented versions of the IBF crucial for the development of the Intrinsic prior approach and they can be used for nested and non-nested model comparison. Nested model comparison is favored though because the resulting IBFs correspond to reasonable intrinsic priors. For a default and more robust IBF approach, Berger and Pericchi (1998b) developed the Median IBF, which is defined by

$$IBF^{MED}_{M_j:M_i}(\mathbf{Y}) = BF^{N}_{M_j:M_i}(\mathbf{Y})Med[BF^{N}_{M_i:M_j}(\mathbf{Y}(l))] \tag{3.27}$$

where Med denotes the median.

3.5 Construction of Objective Prior Distributions

The development of Partial BFs expedited the use of improper prior distributions in model selection procedures, yet these approaches lacked a fundamental Bayesian basis. This led researchers to invent methodologies for constructing Objective prior distributions that correspond to actual BFs, starting from a default non-informative prior distribution. In the following sub-section, we describe extensively three of the most important approaches for constructing Objective prior distributions.

3.5.1 Intrinsic Priors

One of the most important features of the IBF is its correspondence to actual BFs under an appropriate proper prior. If this prior exists, it will be called the ***Intrinsic Prior*** (***IP***). The IP approach provides practical benefits, like alleviation of the computational cost and increased stability, since one can directly use an Intrinsic Prior in place of a default non-informative prior and provide the corresponding BF without depending on training samples. The development of the IPs is progressed through two conditions.

Condition 1. As the sample size goes to infinity and for prior distributions in an appropriate class Γ, the BF of model M_j versus model M_i can be approximated by

$$BF_{M_j:M_i}(\mathbf{Y}) = BF^N_{M_j:M_i}(\mathbf{Y}) \frac{\pi(\hat{\boldsymbol{\theta}}_j|M_j)\pi^N(\hat{\boldsymbol{\theta}}_i|M_i)}{\pi^N(\hat{\boldsymbol{\theta}}_j|M_j)\pi(\hat{\boldsymbol{\theta}}_i|M_i)}(1+o(1)) \tag{3.28}$$

where $\hat{\boldsymbol{\theta}}_j$ and $\hat{\boldsymbol{\theta}}_i$ are the MLE's under M_j and M_i respectively. For defining the IPs, Berger & Perrichi equate Equation 3.20 with either Equation 3.23 or Equation 3.24, resulting to

$$BF^N_{M_j:M_i}(\mathbf{Y}) \frac{\pi(\hat{\boldsymbol{\theta}}_j|M_j)\pi^N(\hat{\boldsymbol{\theta}}_i|M_i)}{\pi^N(\hat{\boldsymbol{\theta}}_j|M_j)\pi(\hat{\boldsymbol{\theta}}_i|M_i)}(1+o(1)) = \tilde{BF}_{M_i:M_j}(\mathbf{Y}(l)) \tag{3.29}$$

where $\tilde{BF}_{M_i:M_j}(\mathbf{Y})$ denotes of either arithmetic or geometric average of $BF_{M_i:M_j}(\mathbf{Y}(l))$. After establishing the first condition , Berger and Pericchi (1996) had to provide necessary assumptions for clarifying the limiting behavior of the quantities included in Equation 3.29

Condition 2. As the sample size n goes to infinity, the following hold

- Under model $M_j, \hat{\boldsymbol{\theta}}_j \to \boldsymbol{\theta}_j, \hat{\boldsymbol{\theta}}_i \to \psi_i(\boldsymbol{\theta}_j)$, and $\tilde{BF}_{M_i:M_j}(\mathbf{Y}(l)) \to BF^*_{M_j}(\boldsymbol{\theta}_j)$
- Under model $M_i, \hat{\boldsymbol{\theta}}_i \to \boldsymbol{\theta}_i, \hat{\boldsymbol{\theta}}_j \to \psi_j(\boldsymbol{\theta}_i)$, and $\tilde{BF}_{M_i:M_j}(\mathbf{Y}(l)) \to BF^*_{M_i}(\boldsymbol{\theta}_i)$

where for $k = i, j$

$$BF^*_{M_k}(\boldsymbol{\theta}_k) = \lim_{L\to\infty} \mathbb{E}^{M_k}_{\boldsymbol{\theta}_k}\Big\{\frac{1}{L}\sum_{l=1}^{L} BF^N_{M_i:M_j}(\mathbf{Y}(l))\Big\} \tag{3.30}$$

for the arithmetic case, and

$$BF^*_{M_k}(\boldsymbol{\theta}_k) = \lim_{L\to\infty} \exp \mathbb{E}^{M_k}_{\boldsymbol{\theta}_k}\Big\{\frac{1}{L}\sum_{l=1}^{L} \log BF^N_{M_i:M_j}(\mathbf{Y}(l))\Big\} \tag{3.31}$$

for the geometric case. If $\mathbf{Y}(l)$ are exchangeable, then limit terms and averages over L can be removed. The term $\psi_j(\boldsymbol{\theta}_i)$ denotes the limit of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_j(\mathbf{Y})$ under model M_i at the point $\boldsymbol{\theta}_i$ and similarly $\psi_i(\boldsymbol{\theta}_j)$ denotes the limit of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_i(\mathbf{Y})$ under model M_j at the point $\boldsymbol{\theta}_j$ (see Dmochowski (1994)).

Then, by using Condition 2 and passing the limit in Equation 3.29 first under M_j and then under M_i, we obtain the ***Intrinsic Equations***

$$\frac{\pi^I(\boldsymbol{\theta}_j|M_j)\pi^N(\psi_i(\boldsymbol{\theta}_j)|M_i)}{\pi^N(\boldsymbol{\theta}_j|M_j)\pi^I(\psi_i(\boldsymbol{\theta}_i)|M_i)} = B^*_{M_j}(\boldsymbol{\theta}_j) \tag{3.32}$$

and

$$\frac{\pi^I(\psi_j(\boldsymbol{\theta}_i)|M_j)\pi^N(\boldsymbol{\theta}_i|M_i)}{\pi^N(\psi_j(\boldsymbol{\theta}_i)|M_j)\pi^I(\boldsymbol{\theta}_i|M_i)} = B^*_{M_i}(\boldsymbol{\theta}_i), \tag{3.33}$$

whose solutions lead to the derivation of the Intrinsic priors $(\pi^I(\boldsymbol{\theta}_j|M_j), \pi^I(\boldsymbol{\theta}_i|M_i))$. So, if we use these priors directly to compute the respective Bayes factor, i.e. $BF^I_{M_j:M_i}(\mathbf{Y})$, we will obtain asymptotically equivalent results with any variation of a respective IBF. Berger and Pericchi (1996) add that these solutions are not necessary unique nor proper.

If one considers a nested model comparison case where model M_i is nested in model M_j, then the solution of the Intrinsic Equations will be provided by

$$\pi^I(\boldsymbol{\theta}_i|M_i) = \pi^N(\boldsymbol{\theta}_i|M_i) \tag{3.34}$$

$$\pi^I(\boldsymbol{\theta}_j|M_j) = \pi^N_{M_j}(\boldsymbol{\theta}_j|M_j)B^*_{M_j}(\boldsymbol{\theta}_j). \tag{3.35}$$

Thus, if we can order all available models of set $\mathscr{M}$ in terms of complexity, we can obtain for each $M_j \in \mathscr{M}$ its respective intrinsic prior when compared with the null M_0. Dmochowski (1994) provides more elaborate characterizations in the nested model case, whilst for non-nested model comparison Moreno et al. (2014) provide a general framework for certain cases under the assumption of exchangeable random variables.

3.5.2 Expected Posterior Priors

Pérez and Berger (2002) introduced a new model selection procedure, with subjective and objective implementation, namely the ***Expected-Posterior Prior*** (***EPP***) which is based on a basic principle. Any prior distribution under each respective model, will be defined through a common underlying predictive distribution using MCMC methods. The main objective of the EPP approach is the construction of appropriately compatible prior distributions across the set $\mathscr{M}$, a trait not shared by several established model selection schemes. The development of the EPP approach is based on imaginary training samples, where their basic principle is to assume an imaginary experiment with a well-defined dataset that will alleviate any arbitrary constants involved in the construction of BFs, when improper default prior are considered.

By imaginary observations we refer to observations generated by an artificial experimenet based on an appropriate dataset, which will used to eliminate the depedence of the Bayes factor of Equation 3.11 to the ratio of arbitrary normallizing constants. Origins of the notion of imaginary observations can be found in Good (1950), Spiegelhalter and Smith (1982). Imaginary data can be either fixed or randomly generated. In this thesis we will focus on more recent approaches which treat imaginary observations as stochastic components, i.e. randomly generated observations based on an appropriate dataset which will have minimal impact to the posterior analysis.

Let $\mathbf{Y}^* = (\mathbf{y}^*_1, \cdots, \mathbf{y}^*_n)^T$ denote the ***imaginary data*** observations which arise independently from a random variable $\mathbf{Y}^*$ on sample space $\mathscr{Y}^*$. We will assume that both random variables $\mathbf{Y}^*$ and $\mathbf{Y}$ are i.i.d. random variables on a common sample space $\tilde{\mathscr{Y}}$. For a given model M_j we will denote $f(\mathbf{Y}^*|\boldsymbol{\theta}_j, M_j)$ the likelihood of the imaginary observations $\mathbf{Y}^*$. Starting with the default improper prior $\pi^N(\boldsymbol{\theta}_j|M_j)$, the posterior distribution of $\boldsymbol{\theta}_j$ under model M_j given $\mathbf{Y}^*$ will be provided by

$$\pi^N(\boldsymbol{\theta}_j|\mathbf{Y}^*, M_j) = \frac{f(\mathbf{Y}^*, |\boldsymbol{\theta}_j, M_j)\pi^N(\boldsymbol{\theta}_j|M_j)}{\int f(\mathbf{Y}^*|\boldsymbol{\theta}_j, M_j)\pi^N(\boldsymbol{\theta}_j|M_j)d\boldsymbol{\theta}_j} \tag{3.36}$$

Since the imaginary data are not included in the data at hand, the posterior distribution of Equation 3.36 can be solely used for model specification purposes.

Given a suitable predictive measure $m^*(\mathbf{Y}^*)$, we define the ***Expected Posterior Prior*** density of $\boldsymbol{\theta}_j$ under M_j with respect to $m^*(\mathbf{Y}^*)$ as

$$\pi^{EPP}(\boldsymbol{\theta}_j|M_j) = \int \pi^N(\boldsymbol{\theta}_j|\mathbf{Y}^*, M_j)m^*(\mathbf{Y}^*)d\mathbf{Y}^*. \tag{3.37}$$

Using the EPP density of Equation 3.37 we can define the BF of model M_j versus model M_i as

$$BF^{EPP}_{M_j:M_i}(\mathbf{Y}) = \frac{m_j^{EPP}(\mathbf{Y})}{m_i^{EPP}(\mathbf{Y})}, \tag{3.38}$$

where

$$m_j^{EPP}(\mathbf{Y}) = \int f(\mathbf{Y}|\boldsymbol{\theta}_j, M_j)\pi^{EPP}(\boldsymbol{\theta}_j|M_j)d\boldsymbol{\theta}_j. \tag{3.39}$$

The imaginary data $\mathbf{Y}^*$ are actually used to convert a default improper prior distribution to usable posterior distribution, yet certain restrictions must be introduced to ensure that the dimension of $\mathbf{Y}^*$ and the predictive distribution $m^*(\cdot)$ are well-defined. Pérez and Berger (2002) in Section 2 introduced the following

$$0 < m_j^N(\mathbf{Y}^*) < \infty \qquad 0 < \mathbb{E}_{M_j}\left\{\frac{m^*(\mathbf{Y}^*)}{m_j^N(\mathbf{Y}^*)}\Big|\boldsymbol{\theta}_j\right\} < \infty, \tag{3.40}$$

which must hold for all $\boldsymbol{\theta}_j$; the expectation $\mathbb{E}_{M_j}$ is provided with respect to the density $f(\mathbf{Y}^*|\boldsymbol{\theta}_j, M_j)$. Thus, if Equation 3.40 hold, then the EPP of Equation 3.37 for every model $M_j \in \mathscr{M}$ exists and is well defined.

The dimension of $\mathbf{Y}^*$ is considered to be the minimal dimension such that Equation 3.40 hold for each model under consideration. If the predictive density $m^*(\cdot)$ is not proper then the EPP will also be improper, yet it is not prohibitive to consider this scenario as we will be seeing later in this section.

If one considers the combined data $\tilde{\mathbf{Y}} = (\mathbf{Y}, \mathbf{Y}^*)$ and define

$$m_j^N(\mathbf{Y}|\mathbf{Y}^*) = \frac{m_j^N(\mathbf{Y}, \mathbf{Y}^*)}{m_j^N(\mathbf{Y}^*)}, \tag{3.41}$$

then we can re-write the marginal likelihood of Equation 3.39 as

$$m_j^{EPP}(\mathbf{Y}) = \int m_j^N(\mathbf{Y}|\mathbf{Y}^*)m^*(\mathbf{Y}^*)d\mathbf{Y}^*, \tag{3.42}$$

which leads to a useful alternative form of the BF of Equation 3.38, i.e.

$$BF^{EPP}_{M_j:M_i}(\mathbf{Y}) = \frac{m_j^{EPP}(\mathbf{Y})}{m_i^{EPP}(\mathbf{Y})} = \frac{\int m_j^N(\mathbf{Y}|\mathbf{Y}^*)m^*(\mathbf{Y}^*)d\mathbf{Y}^*}{\int m_i^N(\mathbf{Y}|\mathbf{Y}^*)m^*(\mathbf{Y}^*)d\mathbf{Y}^*} \tag{3.43}$$

3.5.3 Choice of $m^*(\cdot)$

It is evident that the choice of an appropriate predictive density $m^*(\cdot)$ is key to the EPP approach. Pérez and Berger (2002) viewed $m^*(\cdot)$ as 'arising from beliefs as to how a training sample would behave', in order to also be able to elicit $m^*(\cdot)$ based on subjective knowledge about the problem at hand. We will describe the most important choice as provided by Pérez and Berger (2002); for further choices we refer the reader to Pérez and Berger (2002) Section 3.

Base-Model Approach

An intuitive and widely applicable choice for the predictive density $m^*(\cdot)$, is the ***Base-Model*** approach which uses a base-model as a reference for defining every EPP distribution. The base-model should be the simplest one available and nested in every other model, thus we use the null model M_0 as defined earlier sections of this chapter. In reality, we will be defining EPP distributions for every model of set $\mathscr{M}$ and we will proceed with pairwise model comparisons through the base model, in a sense that

$$BF_{M_j:M_i}(\mathbf{Y}) = BF_{M_j:M_0}(\mathbf{Y})/BF_{M_i:M_0}(\mathbf{Y}). \tag{3.44}$$

Thus, as indicated by Perez and Berger, the predictive density for the imaginary training samples $\mathbf{Y}^*$ will be provided by

$$m^*(\mathbf{Y}^*) = m_0^N(\mathbf{Y}^*) = \int f(\mathbf{Y}^*|\boldsymbol{\theta}_0, M_0)\pi^N(\boldsymbol{\theta}_0|M_0)d\boldsymbol{\theta}_0. \tag{3.45}$$

Thus, the EPP for model M_0 with respect to the predictive density of Equation 3.45 under the base model approach, will be provided by

$$\pi^{EPP}(\boldsymbol{\theta}_0|M_0) = \int \pi^N(\boldsymbol{\theta}_0|\mathbf{Y}^*, M_0)m^*(\mathbf{Y}^*)d\mathbf{Y}^* = \pi^N(\boldsymbol{\theta}_0|M_0) \tag{3.46}$$

which is identical to the default improper prior under the null model M_0. For any other model $M_j \in \mathscr{M}$ its respective EPP will be given by

$$\pi^{EPP}(\boldsymbol{\theta}_j|M_j) = \int \pi^N(\boldsymbol{\theta}_j|\mathbf{Y}^*, M_j)m_0^N(\mathbf{Y}^*)d\mathbf{Y}^* \tag{3.47}$$

This approach was highly motivated by the observation that this version of the EPP coincides with the respective version of the IPs under the AIBF as presented earlier, providing to the EPP procedure a Bayesian justification needed to be considered as a pure Objective Bayesian model selection technique. Given a model $M_j \in \mathscr{M}$ such that M_0 is nested in M_j and $\mathbf{Y}^*$ be an exchangeable imaginary training sample, the respective IP as in Equation 3.34 will be provided by

$$\begin{aligned}
\pi^I(\boldsymbol{\theta}_j|M_j) &= \pi^N(\boldsymbol{\theta}_j|M_j)B^*_{M_j}(\boldsymbol{\theta}_j) \\
&= \pi^N(\boldsymbol{\theta}_j|M_j)\mathbb{E}^{M_j}_{\boldsymbol{\theta}_j}\left[\frac{m_0^N(\mathbf{Y}^*)}{m_j^N(\mathbf{Y}^*)}\right] \\
&= \pi^N(\boldsymbol{\theta}_j|M_j)\int \frac{m_0^N(\mathbf{Y}^*)}{m_j^N(\mathbf{Y}^*)} f(\mathbf{Y}^*|\boldsymbol{\theta}_j, M_j)d\mathbf{Y}^* \\
&= \int \pi^N(\boldsymbol{\theta}_j|M_j)\frac{m_0^N(\mathbf{Y}^*)}{m_j^N(\mathbf{Y}^*)} f(\mathbf{Y}^*|\boldsymbol{\theta}_j, M_j)d\mathbf{Y}^* \\
&= \int \frac{f(\mathbf{Y}^*|\boldsymbol{\theta}_j, M_j)\pi^N(\boldsymbol{\theta}_j|M_j)}{m_j^N(\mathbf{Y}^*)} m_0^N(\mathbf{Y}^*)d\mathbf{Y}^* \\
&= \int \pi^N(\boldsymbol{\theta}_j|\mathbf{Y}^*, M_j)m_0^N(\mathbf{Y}^*)d\mathbf{Y}^* \qquad (3.48)
\end{aligned}$$

which actually is the EPP for $\boldsymbol{\theta}_j$ under model M_j. Pérez and Berger (2002) indicate, that it is not necessary for M_0 to be nested in M_j, as long as M_0 remains the simplest available in terms of complexity e.g. comparison of Exponential with Weibull distribution; for more information we refer to Pérez and Berger (2002) Section 3.2 Example 2.

3.5.4 Computational Aspects

By observing the form of an EPP distribution of Equation 3.37, Pérez and Berger (2002) in Section 4 claimed that they can be viewed as a two-stage hierarchical prior distributions, whereas the first-stage prior will be the posterior distribution $\pi^N(\boldsymbol{\theta}_j|\mathbf{Y}^*, M_j)$, and respectively the second-stage prior will be the predictive density $m^*(\mathbf{Y}^*)$. Using this observation, Pérez and Berger (2002) deduced, that the EPP approach could be integrated by MCMC schemes.

If the posterior density $\pi^N(\boldsymbol{\theta}_j|\mathbf{Y}^*, M_j)$ is available in a closed-form expression then the potential computational cost can be significantly reduced. We will present only the most important computational strategy provided by Pérez and Berger (2002) which will be used in a later section of this thesis.

Direct Simulation and Importance Sampling

If Equation 3.41 is available in a closed-form expression, then Equation 3.43 can be used directly for deriving BFs with EPPs. Let $m^*(\cdot)$ be a proper density (e.g. under the E-EPP approach), then consider R i.i.d. samples $\mathbf{Y}^{*(r)}, r = 1, \cdots, R$ generated from $m^*(\cdot)$ and then approximate the BF of a model M_j versus M_i by

$$\widehat{BF}_{M_j:M_i}(\mathbf{Y})\frac{\sum_{r=1}^R m_j^N(\mathbf{Y}|\mathbf{Y}^{*(r)})}{\sum_{r=1}^R m_i^N(\mathbf{Y}|\mathbf{Y}^{*(r)})} \qquad (3.49)$$

In order to provide all the BFs or Posterior odds we seek, we need to generate a suitable size R of imaginary observations.

If the predictive density $m^*(\cdot)$ is not a proper density, then we resort to an Importance sampling scheme to approximate BFs. Let $\mathbf{Y}^{*(r)}, r = 1, \cdots, R$ be R i.i.d. samples from an importance density $g(\mathbf{Y}^*)$, then the respective BF of M_j versus M_i will be provided by

$$\widehat{BF}_{M_j:M_i}(\mathbf{Y}) = \frac{\sum_{r=1}^{R} m_j^N(\mathbf{Y}|\mathbf{Y}^{*(r)})m^*(\mathbf{Y}^{*(r)})/g(\mathbf{Y}^{*(r)})}{\sum_{r=1}^{R} m_i^N(\mathbf{Y}|\mathbf{Y}^{*(r)})m^*(\mathbf{Y}^{*(r)})/g(\mathbf{Y}^{*(r)})}. \tag{3.50}$$

Under the base-model approach, we can use as importance density $g(\mathbf{Y}^*) = m_0^N(\mathbf{Y}^*|\mathbf{Y})$. Thus, above approximation will be given by

$$\widehat{BF}_{M_j:M_0}(\mathbf{Y}) = \frac{1}{R}\sum_{r=1}^{R} BF^N_{M_j:M_0}(\mathbf{Y}|\mathbf{Y}^{*(r)}), \tag{3.51}$$

where

$$BF^N_{M_j:M_0}(\mathbf{Y}|\mathbf{Y}^{*(r)}) = \frac{1}{R}\sum_{r=1}^{R} \frac{m_j^N(\mathbf{Y}|\mathbf{Y}^{*(r)})}{m_0^N(\mathbf{Y}|\mathbf{Y}^{*(r)})}. \tag{3.52}$$

Under this importance density, we will construct all respective BFs using the same sample of $\mathbf{Y}^{*(r)}$. In most cases the generation procedure of imaginary training samples from the predictive density $m_0^N(\mathbf{Y}^*|\mathbf{Y})$ will be feasible, yet the derivation of Equation 3.52 may be difficult if the respective predictive densities $m_j^N(\mathbf{Y}^*|\mathbf{Y})$ are not readily computable. In cases such as these, we choose the importance density of the numerator of Equation 3.52 to be $g(\mathbf{Y}^*) = m_j^N(\mathbf{Y}^*|\mathbf{Y})$ and for the denominator of Equation 3.52 $g(\mathbf{Y}^*) = m_0^N(\mathbf{Y}^*|\mathbf{Y})$ since we operate under the base-model approach. Thus, the resulting BF approximation will be provided by,

$$\widehat{BF}_{M_j:M_i}(\mathbf{Y}) = BF^N_{M_j:M_0}(\mathbf{Y})\frac{1}{R}\sum_{r=1}^{R} BF^N_{M_0:M_j}(\mathbf{Y}^{*(r)}), \tag{3.53}$$

where $\mathbf{Y}^{*(r)}, r = 1, \cdots, R$ are R i.i.d. samples from $m_j^N(\mathbf{Y}^*|\mathbf{Y})$ and the BFs are provided with the respective marginal likelihoods using the standard non-informative distribution of $\pi^N(\boldsymbol{\theta}_j|M_j)$. The approximation part of Equation 3.53 can be provided quite easily due to the low dimension of the training samples $\mathbf{Y}_r^*$ and in certain cases in can be derived in a closed form expression even for nonstandard distributions. It may be difficult to provide $BF^N_{M_j:M_0}(\mathbf{Y})$, yet it has to be calculated only once.

A disadvantage of this approach is the accumulated computational cost which rises with the continuous generation process of imaginary observations $\mathbf{Y}^*$ under each model $M_j \in \mathscr{M}$, since a new sample $\mathbf{Y}^{*(r)}$ from $m_j^N(\mathbf{Y}^*|\mathbf{Y})$ has to be generated for each model under consideration. If the predictive densities $m_j^N(\mathbf{Y}^*|\mathbf{Y})$ are not available in a closed-form expression, we can deploy a Gibbs sampling scheme for generating importance samples which can be performed as follows:

Finally, the most attractive features of the EPP approach can be summarized as follows:

- All parameter prior distributions are specified through a common predictive distribution $m^*(\mathbf{Y}^*)$.

```
For $r = 1, \cdots, R$:
    • Generate $\boldsymbol{\theta}^{(r)}$ from $\pi^N(\boldsymbol{\theta}_j|\mathbf{Y}, M_j)$.
    • Generate a sample $\mathbf{Y}^{*(r)}$ from $f(\mathbf{Y}^*|\boldsymbol{\theta}^{(r)}, M_j)$.
```

- Computations can be performed through MCMC schemes due to the probabilistic construction of the EPP approach.
- Direct use of improper priors is allowable due to nullification of arbitrary normalizing constants.
- As Consonni and Veronese (2008) indicate "*the EPP approach is a method to make priors compatible across models, through their dependence on a common marginal data distribution; thus this methodology can be applied also with subjectively specified (proper) prior distributions*".
- Multiple model comparison becomes feasible, because the EPP methodology produces fixed prior distributions for each model, feature not shared by all default model selection methods.
- Production of equivalent results with other established model selection approaches such as the IP approach, providing a necessary Bayesian justification.

3.5.5 Power- Expected Posterior Priors

The EPP approach of Pérez and Berger (2002) provides tempting results based on imaginary training samples which arise from a common predictive distribution, starting from a default improper prior. In the variable selection context of Gaussian regression models, where one needs to consider imaginary observations for the design matrix $\mathbf{X}^*$, the EPP approach requires one or more training samples to be chosen from the data-provided design matrix $\mathbf{X}$. Thus, one must investigate the size, the selection procedure and the influence of the imaginary observations.

These three issues led to the development of the ***Power-Expected Posterior Prior*** approach (***PEPP***) by Fouskakis et al. (2015), where they constructed minimally-informative prior distributions without being affected by the imaginary training samples to the same extent as the EPP approach. The PEPP approach is actually an evolution of the EPP approach, using ideas from the Power-Priors of Ibrahim and Chen (2000) and the unit-information prior approach of Kass and Wasserman (1995), tailored to the variable selection context.

For this section we will redefine some basic notions adapted to the variable selection problem of Gaussian regression models. We define $\mathscr{M} = \{M_1, \cdots, M_k\}$, where $k \geq 2$, denote a countable set of regression models. Under any model $M_j \in \mathscr{M}$ we consider the parameters $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \sigma_j^2)$ and likelihood

$$\mathbf{Y}|\mathbf{X}_j, \boldsymbol{\beta}_j, \sigma_j^2, M_j \sim N_n(\mathbf{X}_j\boldsymbol{\beta}_j, \sigma_j^2\mathbf{I}_n), \tag{3.54}$$

where $\mathbf{Y} = (y_1, \cdots, y_n)$ is a vector of real-valued responses, $\mathbf{X}_j$ is the $n \times d_j$ design matrix that contains the values of the explanatory variables, $\mathbf{I}_n$ denotes the $n \times n$

identity matrix, $\boldsymbol{\beta}_j$ is a vector of length d_j containing the effects of the respective covariates in model M_j on the response $\mathbf{Y}$ and σ_j^2 is the error variance of model M_j. We will present the Power-Expected Posterior prior approach based on the variable selection problem, since it was developed for dealing with the issues described and in section 4.6 we will adapt it to the graphical model selection problem of undirected decomposable graphical models.

We denote with $\mathbf{Y}^*$ the imaginary data vector of size m, and the $\mathbf{X}^*_{m\times(p+1)}$ the respective design matrix of size $m \times (p+1)$ where p denotes the number of the available covariates. Note that the design matrix $\mathbf{X}^*_{m\times(p+1)}$ is not consisted of imaginary observations and we will study how it will be structured. Under the EPP approach, the respective EPP will depend on $\mathbf{X}^*$ and not on $\mathbf{Y}^*$ since the imaginary observations will be integrated out.

The selection of a minimal training sample is crucial for constructing minimally informative prior distributions, yet the selection of a suitable one remains an open question, since it depends on the number of models under considerations and on the number of covariates. Usually, it will be specified either from the dimension of the full model i.e. the model where all covariates are present, or the larger model in every pairwise model comparison.

If one decides to use a minimal training sample, then all possible subsets of minimal training samples must be incorporated to the model selection procedure through averages of BFs, resulting in even larger computational cost. As an alternative option, one could randomly select subsets of minimal training samples, but this will result to extra Monte Carlo noise to the model selection procedure. The PEPP approach efficiently tackles the side-effects of selecting a suitable training sample using the following setup.

Following the rationale of the EPP approach, the likelihoods contained in the EPP prior will be raised to the power $1/\delta$ and then will be density-normalized. Fouskakis et al. (2015) default choice is $\delta = m$ and will represent information equal up to one data point. Thus, the resulting prior will provide the same sufficient statistics of the fully observed data but compressed into one data point. The second and most important suggestion by Fouskakis et al. (2015), is to consider the size of the imaginary data vector $m = n$ and subsequently set the design matrix $\mathbf{X}^*_{m\times(p+1)} = \mathbf{X}_{n\times(p+1)}$, which significantly reduces the computational cost that rises from the dependence on the imaginary training samples. Note, that Fouskakis et al. (2015) indicate that in reality δ can have any value in $[p+2, n]$, which also subsequently restricts PEPP approach to cases only where $p << n$.

For a given model $M_j \in \mathscr{M}$ let $\pi^N(\boldsymbol{\beta}_j, \sigma_j^2 | \mathbf{X}_j^*, M_j)$ denote the respective default improper prior for the parameters $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \sigma_j^2)$. Then the ***Power-Expected Posterior Prior*** is defined by

$$\pi^{PEPP}(\boldsymbol{\beta}_j, \sigma_j^2 | \mathbf{X}_j^*, \delta, M_j) = \int \pi^N(\boldsymbol{\beta}_j, \sigma_j^2 | \mathbf{Y}^*, \delta, M_j) m^N(\mathbf{Y}^* | \mathbf{X}_0^*, \delta, M_0) d\mathbf{Y}^*, \tag{3.55}$$

where

$$\pi^N(\boldsymbol{\beta}_j, \sigma_j^2 | \mathbf{Y}^*, \delta, M_j) = \frac{f(\mathbf{Y}^* | \boldsymbol{\beta}_j, \sigma_j^2, M_j; \mathbf{X}_j^*, \delta) \pi^N(\boldsymbol{\beta}_j, \sigma_j^2 | \mathbf{X}_j^*, M_j)}{m_j^N(\mathbf{Y}^* | \mathbf{X}_j^*, \delta)}. \tag{3.56}$$

The likelihood $f(\mathbf{Y}^*|\boldsymbol{\beta}_j, \sigma_j^2, \mathbf{X}_j^*, \delta, M_j)$ is the EPP likelihood raised to the power $1/\delta$ and density-normalized i.e.

$$\begin{aligned} f(\mathbf{Y}^*|\boldsymbol{\beta}_j, \sigma_j^2, M_j; \mathbf{X}_j^*, \delta) &= \frac{f(\mathbf{Y}^*|\boldsymbol{\beta}_j, \sigma_j^2, M_j; \mathbf{X}_j^*)^{\frac{1}{\delta}}}{\int f(\mathbf{Y}^*|\boldsymbol{\beta}_j, \sigma_j^2, M_j; \mathbf{X}_j^*)^{\frac{1}{\delta}} d\mathbf{Y}^*} \\ &= \frac{f_{N_m}(\mathbf{Y}^*; \mathbf{X}_j^*\boldsymbol{\beta}_j, \sigma_j^2\mathbf{I}_m)^{\frac{1}{\delta}}}{\int f_{N_m}(\mathbf{Y}^*; \mathbf{X}_j^*\boldsymbol{\beta}_j, \sigma_j^2\mathbf{I}_m)^{\frac{1}{\delta}} d\mathbf{Y}^*} \\ &= f_{N_m}(\mathbf{Y}^*; \mathbf{X}_j^*\boldsymbol{\beta}_j, \delta\sigma_j^2\mathbf{I}_m), \end{aligned} \tag{3.57}$$

where $f_{N_d}(\mathbf{Y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of the d-dimensional Normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ evaluated at $\mathbf{Y}$.

The marginal distribution $m^N(\mathbf{Y}^*|\mathbf{X}_j^*, \delta, M_j)$ appearing in Equation 3.56 for $j = 0$ is the prior predictive distribution evaluated at $\mathbf{Y}^*$ of model M_j with the power-likelihood as defined in Equation 3.57 under its respective default baseline prior, i.e.

$$m_j^N(\mathbf{Y}^*|\mathbf{X}_j^*, \delta) = \int f_{N_m}(\mathbf{Y}^*; \mathbf{X}_j^*\boldsymbol{\beta}_j, \delta\sigma_j^2\mathbf{I}_m)\pi^N(\boldsymbol{\beta}_j, \sigma_j^2|\mathbf{X}_j^*, M_j)d\boldsymbol{\beta}_j d\sigma_j^2. \tag{3.58}$$

Fouskakis et al. (2015) adopted the base-model approach of Pérez and Berger (2002), where they consider $m^*(\mathbf{Y}^*) = m_0^N(\mathbf{Y}^*|\mathbf{X}_j^*, \delta)$, which is a natural choice under the variable-selection context. Using Equation 3.55 and Equation 3.56 we can re-write the PEPP prior of Equation 3.55 under model M_j as

$$\pi^{PEPP}(\boldsymbol{\beta}_j, \sigma_j^2|\mathbf{X}_j^*, \delta, M_j) = \pi^N(\boldsymbol{\beta}_j, \sigma_j^2|\mathbf{X}_j^*, M_j) \int \frac{m_0^N(\mathbf{Y}^*|\mathbf{X}_0^*, \delta)}{m_j^N(\mathbf{Y}^*|\mathbf{X}_j^*, \delta)} f(\mathbf{Y}^*|\boldsymbol{\beta}_j, \sigma_j^2, M_j; \mathbf{X}_j^*, \delta)d\mathbf{Y}^*. \tag{3.59}$$

Therefore, using Equation 3.59 the posterior distribution of the model parameters $(\boldsymbol{\beta}_j, \sigma_j^2)$ given $\mathbf{Y}$ under model M_j will be provided by

$$\pi^{PEPP}(\boldsymbol{\beta}_j, \sigma_j^2|\mathbf{Y}, M_j; \mathbf{X}_j, \mathbf{X}_j^*, \delta) \propto \int \pi^N(\boldsymbol{\beta}_j, \sigma_j^2|\mathbf{Y}, \mathbf{Y}^*, M_j; \mathbf{X}_j, \mathbf{X}^*, \delta)\times \tag{3.60}$$

$$\times\, m_j^N(\mathbf{Y}|\mathbf{Y}^*; \mathbf{X}_j, \mathbf{X}_j^*, \delta)m_0^N(\mathbf{Y}^*|\mathbf{X}_0^*, \delta)d\mathbf{Y}^*, \tag{3.61}$$

where $\pi^N(\boldsymbol{\beta}_j, \sigma_j^2|\mathbf{Y}, \mathbf{Y}^*, M_j; \mathbf{X}_j, \mathbf{X}^*, \delta)$ and $m_j^N(\mathbf{Y}|\mathbf{Y}^*; \mathbf{X}_j, \mathbf{X}_j^*, \delta)$ denote the posterior distribution of $(\boldsymbol{\beta}_j, \sigma_j^2)$ and marginal likelihood of model M_j respectively, under the data $\mathbf{Y}^*$ using design matrix $\mathbf{X}_j$ with respect to the prior $\pi^N(\boldsymbol{\beta}_j, \sigma_j^2|\mathbf{Y}^*, M_j; \mathbf{X}_j^*, \delta)$ - i.e. the posterior distribution of the parameters $(\boldsymbol{\beta}_j, \sigma_j^2)$ given the imaginary data $\mathbf{Y}^*$ under the power Normal likelihood as defined in Equation 3.57 and with respect to the default baseline prior $\pi^N(\boldsymbol{\beta}_j, \sigma_j^2|\mathbf{X}_j^*, M_j)$.

Using the latest version of the PEPP prior of Equation 3.60, we provide the marginal likelihood of the data $\mathbf{Y}$ under model $M_j \in \mathscr{M}$ by

$$m_j^{PEPP}(\mathbf{Y}|\mathbf{X}_j, \mathbf{X}_j^*, \delta) = m_j^N(\mathbf{Y}|\mathbf{X}_j, \mathbf{X}_j^*) \int \frac{m_j^N(\mathbf{Y}^*|\mathbf{Y}, \mathbf{X}_j, \mathbf{X}_j^*, \delta)}{m_j^N(\mathbf{Y}^*|\mathbf{X}_j^*, \delta)} m_0^N(\mathbf{Y}^*|\mathbf{X}_0^*, \delta)d\mathbf{Y}^*, \tag{3.62}$$

where $m_j^N(\mathbf{Y}|\mathbf{X}_j, \mathbf{X}_j^*)$ is the marginal likelihood of the actual data under model M_j with respect to the default baseline prior $\pi^N(\boldsymbol{\beta}_j, \sigma_j^2|\mathbf{X}_j^*, M_j)$. Therefore, we are able to provide Posterior Odds and BFs using this expression of the marginal likelihood, yet only when it can be analytically tractable.

Fouskakis et al. (2015) in Section 3 provide two MCMC approaches for estimating the marginal likelihood of Equation 3.62, plus two more less successful approaches in the Appendix of the respective paper. These approaches are the following:

1. Generate imaginary data $\mathbf{Y}^{*(r)}, r = (1, \cdots, R)$ from $m_j^N(\mathbf{Y}^*|\mathbf{Y}, \mathbf{X}_j, \mathbf{X}_j^*, \delta)$ and estimate the marginal likelihood of Equation 3.62 by

$$\widehat{m}_j^{PEPP}(\mathbf{Y}|\mathbf{X}_j, \mathbf{X}_j^*, \delta) = m_j^N(\mathbf{Y}|\mathbf{X}_j, \mathbf{X}_j^*)\left[\frac{1}{R}\sum_{r=1}^{R}\frac{m_0^N(\mathbf{Y}^{*(r)}|\mathbf{X}_0^*, \delta)}{m_j^N(\mathbf{Y}^{*(r)}|\mathbf{X}_j^*, \delta)}\right]. \tag{3.63}$$

2. Generate imaginary data $\mathbf{Y}^{*(r)}, r = (1, \cdots, R)$ from $m_j^N(\mathbf{Y}^*|\mathbf{Y}, \mathbf{X}_j, \mathbf{X}_j^*, \delta)$ and estimate the marginal likelihood of Equation 3.62 by

$$\begin{aligned}\widehat{m}_j^{PEPP}(\mathbf{Y}|\mathbf{X}_j, \mathbf{X}_j^*, \delta) = m_0^N(\mathbf{Y}|\mathbf{X}_0, \mathbf{X}_0^*)\times \\ \times\frac{1}{R}\sum_{r=1}^{R}\frac{m_j^N(\mathbf{Y}|\mathbf{Y}^{*(r)}; \mathbf{X}_j, \mathbf{X}_j^*, \delta)}{m_0^N(\mathbf{Y}|\mathbf{Y}^{*(r)}; \mathbf{X}_0, \mathbf{X}_0^*, \delta)}\frac{m_0^N(\mathbf{Y}^{*(r)}|\mathbf{Y}; \mathbf{X}_0, \mathbf{X}_0^*, \delta)}{m_j^N(\mathbf{Y}^{*(t)}|\mathbf{Y}; \mathbf{X}_j, \mathbf{X}_j^*, \delta)}.\end{aligned} \tag{3.64}$$

3. Generate imaginary data $\mathbf{Y}^{*(r)}, r = (1, \cdots, R)$ from $m_0^N(\mathbf{Y}^*|\mathbf{X}_0^*, \delta)$ and estimate the marginal likelihood of Equation 3.62 by

$$\widehat{m}_j^{PEPP}(\mathbf{Y}|\mathbf{X}_j, \mathbf{X}_j^*, \delta) = m_j^N(\mathbf{Y}|\mathbf{X}_j, \mathbf{X}_j^*)\frac{1}{R}\sum_{r=1}^{R}\frac{m_j^N(\mathbf{Y}^{*(r)}|\mathbf{Y}, \mathbf{X}_j, \mathbf{X}_j^*, \delta)}{m_j^N(\mathbf{Y}^{*(r)}|\mathbf{X}_j^*, \delta)}. \tag{3.65}$$

4. Generate imaginary data $\mathbf{Y}^{*(r)}, r = (1, \cdots, R)$ from $m_0^N(\mathbf{Y}^*|\mathbf{Y}, \mathbf{X}_0, \mathbf{X}_0^*, \delta)$ and estimate the marginal likelihood of Equation 3.62 by

$$\widehat{m}_j^{PEPP}(\mathbf{Y}|\mathbf{X}_j, \mathbf{X}_j^*, \delta) = m_j^N(\mathbf{Y}|\mathbf{X}_0, \mathbf{X}_0^*)\frac{1}{R}\sum_{r=1}^{R}\frac{m_j^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, \mathbf{X}_j, \mathbf{X}_j^*, \delta)}{m_0^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, \mathbf{X}_0, \mathbf{X}_0^*, \delta)} \tag{3.66}$$

Under the first two approaches, the imaginary observations are generated from the posterior predictive distribution under M_j, thus one must repeat the generation process according to the model under consideration. Approaches (3) and (4) are simpler, since one imaginary sample must be generated using the prior and posterior predictive distribution under the null model M_0. Yet, as Fouskakis et al. (2015) indicate, the provided estimates will have large Monte-Carlo error, since the imaginary data are generated from importance functions that do not exploit completely the information provided by the data (Approach (3)) or the stochastic structure of model M_j (Approaches (3) and (4)). Therefore, Approaches (1) and (2) are preferred over (3) and (4) since they produce more stable estimates of the marginal likelihood.

Fouskakis et al. (2015) provided two basic setups based on different baseline priors, the Jeffreys prior and Zellners g-prior. Under these approaches as presented in the Appendix Section 2 of the respective paper, the marginal likelihood $\widehat{m}_j^{PEPP}(\mathbf{Y}|\mathbf{X}_j, \mathbf{X}_j^*, \delta)$ is available in a closed form expression as a multivariate Student distribution. If we cannot identify a closed form expression, we can alternatively use the Gibbs Sampling scheme as in the EPP approach, where the imaginary observations will be generated as follows:

For $r = 1, \cdots, R$:

- Generate $\boldsymbol{\theta}^{(r)}$ from $\pi^N(\boldsymbol{\theta}_j|\mathbf{Y}, M_j)$.
- Generate a sample $\mathbf{Y}^{*(r)}$ from $f(\mathbf{Y}^*|\boldsymbol{\theta}^{(r)}, \delta, M_j)$.

3.6 Priors on Model Space $\mathscr{M}$

In the previous sections of this chapter, we studied extensively ways on constructing Objective prior distributions regarding the parameter vector $\boldsymbol{\theta}_j$ under any model $M_j \in \mathscr{M}$. In order to provide Posterior odds, one must also define the model prior probabilities $\pi(M_j)$, $\forall M_j \in \mathscr{M}$. If we consider the $\mathscr{M}$-closed view, then a standard choice which depicts prior ignorance is to assume a uniform distribution across the model space $\mathscr{M}$, i.e.

$$\pi(M_j) = \frac{1}{|\mathscr{M}|}, \tag{3.67}$$

where $|\mathscr{M}|$ denotes the cardinality of set $\mathscr{M}$. Though it is a simple approach, it has received a great deal of criticism since it disregards the structural features of a given model, such as sparsity, dimensionality or collinearity of predictors under the variable selection context.

As Consonni et al. (2018) indicate, under the variable selection context, once can assign a random prior probability of inclusion ω for each predictor such that $\pi(M_j|\omega) = \omega^{p_j}(1-\omega)^{n-p_j}$, where p_j denotes the number of covariates included under model M_j. Then, using as hyper-prior for ω a Beta distribution such that $\omega \sim Beta(\alpha_\omega, b_\omega)$, the resulting model prior probability will be provided by

$$\pi(M_j) = \frac{B(\alpha_\omega + pi_j, b_\omega + p - \pi_j)}{B(\alpha_\omega, b_\omega)}, \tag{3.68}$$

which is called the ***Beta*** - ***Binomial*** prior on model space $\mathscr{M}$. If one sets $\alpha_\omega = b_\omega = 1$ then the prior distribution for ω, would be a Uniform distribution, resulting to model prior probabilities

$$\pi(M_j) = \frac{1}{p+1}\binom{p}{p_j}^{-1}, \tag{3.69}$$

which leads to a uniform prior probability on model size

$$\pi(\{M_j \in \mathscr{M} : p_j = k\}) = \frac{1}{k+1}, k = 0, 1, \cdots, p. \tag{3.70}$$

Choosing a uniform prior on ω, will not penalize model complexity. One other consideration made from Wilson et al. (2010), is setting $\alpha_\omega = 1$ and $b_\omega = \lambda_p$, where $\lambda > 0$. Under this choice, the resulting model prior will have expectation $1/\lambda$ and similar behavior to a geometric distribution when one considers low dimensional models. In this way, this prior provides an approximate penalization $log(1 + \lambda)$ for each covariate added to the model.

Chapter 4

Objective Bayes in Structure Learning

4.1 Introduction

In the chapter 2, we established the notion of graphical models and how they can be used to describe conditional independence relationships among variables through the means of a graph. In reality, the underlying graph's structure is unknown and has to be inferred by the data at hand. This procedure is known in the bibliography as ***Structural Learning*** and it can be approached by both Frequentist and Bayesian point of view, where the latter will be explored, since it allows for more delicate handling of uncertainty.

For expediting a Bayesian approch to the problem of structural learning, we first set a prior on a given graph G and then a prior distribution to the covariance matrix $\mathbf{\Sigma}_{q\times q}$ given G on q nodes. The choice of a prior distribution over $\mathbf{\Sigma}_{q\times q}$ can become an arduous task for two main reasons:

- Different graphs imply different independence structures that influence the parameters space and improper priors cannot be directly used.
- With an increasing number of variables the parameter space to grow super-exponentially and eliciting a prior distribution is impossible.

To alleviate these dependencies, we set our focus on Obejctive Bayes procedures which provide prior distributions starting from improper priors. Notable contributions of OB community to structure learning can be found in Carvalho and Scott (2009), Consonni et al. (2017) and Castelletti et al. (2018), which are based on the Fractional Bayes Factor of O'Hagan (1995). The Fractional Bayes factor, despite its computationally convenient setup, is uses twice the available data at hand.

In this chapter, we apply the Expected and Power-Expected Posterior Prior approaches of Pérez and Berger (2002) and Fouskakis et al. (2015) respectively, to the structural learning problem of undirected decomposable graphical models. The remainder of the chapter will be structured as follows. First, we describe the application of Fractional Bayes factor to the structural learning problem, as applied by Carvalho and Scott (2009). Then we introduce the Expected and Power-Expected

Posterior Prior approach to the structural learning problem, followed by some standard choices for priors on graphs. Next, we discuss about computational strategies for implementing structural learning processes, where we focus on FINCS algorithm as developed by Carvalho and Scott (2008). Finally, we test the performance of the proposed approaches on diverse simulation scenarios, as well as in a protein-signaling data application.

In this chapter we introduce the notion of Gaussian graphical models. We first provide the general setup of a Gaussian graphical model, using the notions established in chapter 2, and we then describe the basic

4.2 Distributions

Let $Y = \{Y_1, \cdots, Y_q\}$ from which we observe n i.i.d. q-dimensional observations which can be arranged in an $n \times q$ matrix

$$\mathbf{Y}_{n\times q} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_q) = \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix} \tag{4.1}$$

where $\mathbf{y}_i = (y_{i1}, \cdots, y_{iq})$ denotes the i-th observation and $\mathbf{Y}_j = (y_{1j}, \cdots, y_{nj})$ denotes the observations on the j-th variable. We model the observations as $\mathbf{y}_i|\mathbf{\Sigma} \sim N_q(\mathbf{0}, \mathbf{\Sigma})$ independently over $i = 1, \cdots, n$, where $\mathbf{\Sigma}_{q\times q}$ is an unconstrained semi-positive definite matrix, and $N_q(\mathbf{0}, \mathbf{\Sigma})$ denotes the q-variate normal distribution with mean vector $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$.

We say that a random matrix $\mathbf{Y}_{n\times q}$ follows the matrix normal distribution with mean matrix $\mathbf{M}_{n\times q}$, row covariance matrix $\mathbf{R}_{n\times n}$ and column covariance matrix $\mathbf{\Sigma}_{q\times q}$, when $vec(\mathbf{Y})$ follows the multivariate normal distributions with mean vector $vec(\mathbf{M})$ and covariance matrix $\mathbf{\Sigma} \otimes \mathbf{R}$; $\otimes$ denotes the kronecker product. Note that $vec(\mathbf{Y})$ denotes the vectorization of matrix $\mathbf{Y}$ i.e. its conversion to a column vector, which is attained by stacking the columns of matrix $\mathbf{Y}_{n\times q}$ on top of one another and obtaining a $mn \times 1$ column vector. To denote that the random matrix $\mathbf{Y}$ follows the matrix normal distribution, we will write

$$\mathbf{Y}_{n\times q} \sim MN_{n,q}(\mathbf{M}, \mathbf{R}, \mathbf{\Sigma}), \tag{4.2}$$

and for the scope of this thesis, we will consider $\mathbf{M} = \mathbf{0}_{n\times q}$ and $\mathbf{R} = \mathbf{I}_n$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix and $\mathbf{0}_{n\times q}$ denotes an $n \times q$ matrix with zero entries only. Therefore, given that $\mathbf{Y} \sim MN_{n,q}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma})$, the density of matrix $\mathbf{Y}$ given matrix $\mathbf{\Sigma}$, will be provided by

$$f(\mathbf{Y}|\mathbf{\Sigma}) = \frac{det(\mathbf{\Sigma})^{-n/2}}{(2\pi)^{nq/2}} \exp\Big\{ -\frac{1}{2} tr(\mathbf{\Sigma}^{-1}\mathbf{S})\Big\}, \tag{4.3}$$

where $det(\cdot)$ denotes the determinant of a matrix, $tr(\cdot)$ denotes the trace of a matrix and $\mathbf{S} = \mathbf{Y}^T\mathbf{Y}$. For more details on the matrix normal distribution see Gupta and Nagar (2000).

Now, let $\mathbf{\Sigma}$ be a $q \times q$ unconstrained s.p.d. random matrix. We will write $\mathbf{\Sigma} \sim IW_q(b, \mathbf{D})$ to denote that $\mathbf{\Sigma}$ follows an Inverse-Wishart distribution with density

$$\pi(\mathbf{\Sigma}|b, \mathbf{D}) = \boldsymbol{K} det(\mathbf{\Sigma})^{-(b/2+q)} \exp\Big\{ -\frac{1}{2} tr(\mathbf{\Sigma}^{-1}\mathbf{D})\Big\}, \tag{4.4}$$

where

$$\mathbf{K} = \frac{det(\mathbf{D})}{2^{bq/2}\Gamma_q\left(\frac{b}{2}\right)}, \tag{4.5}$$

having $\mathbf{\Sigma}$ s.p.d, and $\pi(\mathbf{\Sigma}) = 0$ otherwise, $\mathbf{D}_{q\times q}$ is a s.p.d matrix, $b > q - 1$ is a scalar and

$$\Gamma_q\Big(\frac{b}{2}\Big) = \pi^{\frac{q(q-1)}{4}} \prod_{j=1}^{q} \Gamma\Big(\frac{b}{2} + \frac{1+j}{2}\Big) \tag{4.6}$$

denotes the multivariate gamma function evaluated at $b/2$.

Consider the Gaussian multivariate setup provided by Equation 4.2. Following Consonni et al. (2017), we start from the improper prior

$$\pi^N(\mathbf{\Sigma}) \propto det(\mathbf{\Sigma})^{-\frac{a+q-1}{2}}, \tag{4.7}$$

which can provide several default distributions based on different values of parameter a. For the remaining part of this thesis, consider $a = q + 1$, which corresponds to the default prior also used by Carvalho and Scott (2009). Starting from the prior of Equation 4.7, the posterior distribution of $\mathbf{\Sigma}$ given the data matrix $\mathbf{Y}$ will be provided by

$$\pi^N(\mathbf{\Sigma}|\mathbf{Y}) \propto det(\mathbf{\Sigma})^{-(n/2+q)} exp\Big\{ -\frac{1}{2} tr(\mathbf{\Sigma}^{-1}\mathbf{S})\Big\}, \tag{4.8}$$

which corresponds to the kernel of an Inverse Wishart distribution with degrees of freedom parameter $b = n$ and scale matrix $\mathbf{D} = \mathbf{S}$. For the posterior distribution of $\mathbf{\Sigma}$ given the data matrix $\mathbf{Y}$ to be proper, we require for $b = n > q - 1$ and for the scale matrix $\mathbf{\Sigma}$ to be s.p.d. This requirement provides us with a restriction which indicates that we must consider application with at least $n = q$ observations. With the notation $\pi^N(\mathbf{\Sigma}|\mathbf{Y})$, we will be reffering to the posterior distribution of $\mathbf{\Sigma}$ given the data matrix $\mathbf{Y}$ which is provided using an improper prior.

Using Equation 4.3 and Equation 4.7, the marginal density of the data matrix $\mathbf{Y}$ is provided by

$$m^N(\mathbf{Y}) = (2\pi)^{nq/2}\Gamma_q\Big(\frac{n+q-1}{2}\Big) det\Big(\frac{1}{2}\mathbf{S}\Big)^{\frac{n+q-1}{2}}. \tag{4.9}$$

Again, we will use the notation $m^N(\mathbf{Y})$ to denote that the marginal likelihood is calculated with respect an improper prior, in our case the prior defined in Equation 4.7, $\pi^N(\mathbf{\Sigma})$; See appendix (??).

4.3 Gaussian Graphical Models and Prior Laws

In this section we introduce to the reader the notion of gaussian graphical models which are decomposable. For the remainder of this thesis, we consider a collection of decomposable graphical models $\mathscr{G} = \{G_1, \cdots, G_k\}, k \geq 2$ with clique set $\mathscr{C}$ and separator set $\mathscr{S}$ on q nodes.

Dempster (1972) introduced the notion of covariance selection models, which assumes a zero-mean multivariate normal population specified by zero entries on the inverse of the covariance matrix $\mathbf{K} = \mathbf{\Sigma}^{-1}$. Using the pattern of the zero entries in $\mathbf{K}$, the pairwise conditional independence structure of the entertained variables can be associated with an undirected decomposable graph $G = (V, E)$. A Gaussian graphical model is defined by assuming a matrix Normal distribution on $\mathbf{Y}$, adhered to the Markov property with respect to a graph G, where the underlying graph's structure will be represented through the non-zero elements of the precision matrix $\mathbf{K}$. A Gaussian graphical model will be formally stated as

$$\mathbf{Y}|\mathbf{\Sigma}, G \sim MN_{n,q}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma}). \tag{4.10}$$

For the covariance matrix to be well-defined, we will write $\mathbf{\Sigma} \in M^+(G)$ to denote that $\mathbf{\Sigma}$ takes values in the space $M(G)$, whose elements are positive definite. We will say that $\mathbf{\Sigma}$ is Markov with respect to the graph G to indicate that conditional independence structure of the graph G is represented through the off-diagonal zero entries of the precision matrix $\mathbf{K}$. In Figure 4.1 we provide a graphical representation of the model, for $q = 10$ nodes, which assumes no conditional independence structure restrictions and will be described as the ***null graphical model*** (or as the ***saturated model*** (Lauritzen (1996) p. 124)) and will be denoted as $G_0 = (V, E_0)$. Note, that the null graphical model will be used later on this chapter for establishing the base-model approach of subsection 3.5.3 to the graphical model selection problem of undirected decomposable graphical models.

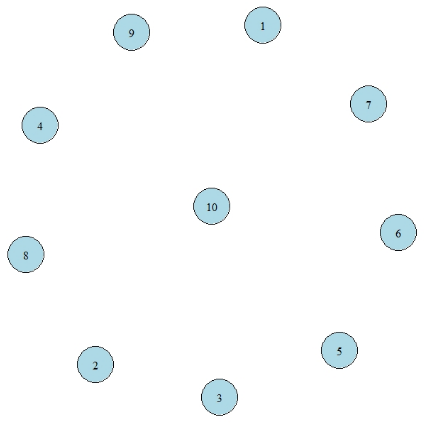


Figure 4.1: Graphical representation of the null graphical model for $q = 10$ nodes.

The density of a multivariate Gaussian graphical model with respect to the undircted decomposable graph G, using the clique factorization of Equation 2.14, will be provided by

$$f(Y|\boldsymbol{\Sigma}, G) = \frac{\prod_{C \in \mathscr{C}} f(\mathbf{Y}_C|\boldsymbol{\Sigma}_C, G)}{\prod_{S \in \mathscr{S}} f(\mathbf{Y}_S|\boldsymbol{\Sigma}_S, G)}, \tag{4.11}$$

where under each clique $C \in \mathscr{C}$ we assume $\mathbf{Y}_C \sim MN(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma}_C)$ and under each separator $S \in \mathscr{S}$ $\mathbf{Y}_C \sim MN(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma}_C)$. Each matrix $\mathbf{Y}_C$ under each clique $C \in \mathscr{C}$ is composted by selecting the columns of matrix $\mathbf{Y}$ that represents the variables contained in the respective clique C; similarly for any given separator $S \in \mathscr{S}$. Each matrix $\boldsymbol{\Sigma}_C$ under each clique $C \in \mathscr{C}$ is composted by partitioning matrix $\boldsymbol{\Sigma}$ into blocks corresponding to the variables contained in clique $C \in \mathscr{C}$; similary for any given separator $S \in \mathscr{S}$. The precision matrix $\mathbf{K}$ obeys the same factorization per cliques and separators as the covariance matrix $\boldsymbol{\Sigma}$.

Prior Distributions for Gaussian Graphical Models

Following section 4.2, a common choice for prior distribution over $\boldsymbol{\Sigma}$ that admits a conjugate setup, is the Inverse Wishart distribution. Dawid and Lauritzen (1993) constructed hyper Markov laws based on Wishart and Inverse Wishart distributions, to expedite a conjugate setup on the Gaussian graphical model framework. We say that $\boldsymbol{\Sigma}$ follows a Hyper-Inverse Wishart distribution and write $\boldsymbol{\Sigma}|b, \mathbf{D} \sim HIW_G(b, \mathbf{D})$, when the density of $\boldsymbol{\Sigma}$ under any given graphical model $G \in \mathscr{G}$, using the clique-separator factorization of Equation 2.14, is provided by

$$\pi(\boldsymbol{\Sigma}|b, D, G) = \frac{\prod_{C \in \mathscr{C}} \pi(\boldsymbol{\Sigma}_C|b, \mathbf{D}_C, G)}{\prod_{S \in \mathscr{S}} \pi(\boldsymbol{\Sigma}_S|b, \mathbf{D}_S, G)}, \tag{4.12}$$

implying that under each clique $C \in \mathscr{C}$, $\boldsymbol{\Sigma}_C$ follows an Inverse Wishart distribution, such that $\boldsymbol{\Sigma}_C|b, \mathbf{D}_C \sim IW_{|C|}(b, \mathbf{D}_C)$, with density

$$\pi(\boldsymbol{\Sigma}_C|b, \mathbf{D}_C, G) = K_C\, det(\boldsymbol{\Sigma}_C)^{-(b/2+|C|)} exp\Big\{ -\frac{1}{2} tr(\boldsymbol{\Sigma}_C^{-1}\mathbf{D}_C)\Big\}, \tag{4.13}$$

where

$$K_C = \frac{det(\mathbf{D}_C)}{2^{b|C|/2}\Gamma_{|C|}(\frac{b}{2})};$$

$\mathbf{D}_C, \boldsymbol{\Sigma}_C \in M^+(G)$, $tr(\cdot)$ denotes the trace of a matrix and $|\cdot|$ denotes cardinality of a set.

Roverato (2000) studied the properties of the Cholesky decomposition of a matrix observation provided by the hyper Inverse Wishart distribution. More precisely, he provided an alternative parametrization of a decomposable graphical model based on the pattern of zero entries on the upper triangle matrix of a cholesky decomposition. A significant contribution of this work, was the definition of the inverse of a hyper Inverse Wishart distribution, namely the G-conditional Wishart distribution.

Under the null graphical model G_0, the covariance matrix $\boldsymbol{\Sigma}$ will follow an Inverse Wishart distribution $IW(b, \mathbf{D})$ as in Equation 4.4. If we consider the change of variable $K = \mathbf{\Sigma}^{-1}$ with $\mathbf{\Sigma} \sim IW(\delta, \mathbf{B})$, then as provided by Roverato (2000) p.102, the distribution of $\mathbf{K}$ would be a Wishart distribution such that $\mathbf{K} \sim W(\delta + q - 1, \mathbf{A})$ with density

$$\pi(\mathbf{K}|b, \mathbf{A}) = h(b, q)\frac{det(\mathbf{K})^{(\delta-2)/2}}{det(\mathbf{A})^{(\delta+q-1)/2}} \exp\Big\{ - \frac{1}{2}tr(\mathbf{K}\mathbf{A}^{-1})\Big\}, \tag{4.14}$$

where $\mathbf{A} = \mathbf{B}^{-1}$ and $h(\delta, q)$ as provided in **??**.

Under a decomposable graph $G = (V, E)$, with $\mathbf{B}^{-1} = \mathbf{A} \in M^{+}(G)$, by Lauritzen (1996) we get

$$\mathbf{B}^{-1} = \sum_{i=1}^{k}[\mathbf{B}_{C_i}^{-1}]^0 - \sum_{i=2}^{k}[\mathbf{B}_{S_i}^{-1}]^0, \tag{4.15}$$

$$det(\mathbf{B}) = \frac{\prod_{i=1}^{k} det(\mathbf{B}_{C_i})}{\prod_{i=2}^{k} det(\mathbf{B}_{S_i})}, \tag{4.16}$$

where $[\mathbf{B}_{C_i}^{-1}]^0$ denotes the $q \times q$ matrix structured by filling zero entries around B_C, in order to obtain a full dimension matrix. Using Equation 4.15 and Equation 4.16, the density of $\mathbf{\Sigma}$ in **??** can be written as

$$\pi(\mathbf{\Sigma}|\delta, \mathbf{D}, G) = h_G(\delta, q)\left(\frac{\prod_{i=1}^{k} \det(\mathbf{\Sigma}_{C_j})^{|C_j|+1}}{\prod_{i=2}^{k} \det(\mathbf{\Sigma}_{S_j})^{|S_j|+1}}\right)^{-1}\left(\frac{\prod_{i=1}^{k} \det(B_{C_j})^{|C_j|+1}}{\prod_{i=2}^{k} \det(B_{S_j})^{|S_j|+1}}\right)^{1/2} \times \tag{4.17}$$

$$\times \frac{det(\mathbf{\Sigma})^{-(\delta-2)/2}}{det(B)^{-(\delta-2)/2}} \exp\{-\frac{1}{2}tr(\mathbf{\Sigma}^{-1}\mathbf{B})\} \tag{4.18}$$

Using **??** and Equation 2.14, the normalizing constant of **??**) can be re-written as:

$$h_G(\delta, q) = \frac{\prod_{i=1}^{k} h(\delta, C_i)}{\prod_{i=2}^{k} h(\delta, S_i)} \tag{4.19}$$

$$= \frac{\prod_{i=2}^{k} \Gamma_{|S_j|}\left(\frac{\delta+|S_i|-1}{2}\right)}{\prod_{i=1}^{k} \Gamma_{|C_i|}\left(\frac{\delta+|C_j|-1}{2}\right)} \times \frac{\prod_{i=1}^{k} 2^{-|C_i|(\delta+|C_i|-1)/2}}{\prod_{i=2}^{k} 2^{-|S_i|(\delta+|S_i|-1)/2}}, \tag{4.20}$$

This computational cost required for providing the normalizing constant $h(\delta, q)$ using the expression in Equation 4.19 can be intense, thus Roverato (2000) p.106 provided an alternative formulation, given by

$$h_G(\delta, q) = (2\pi)^{-v/2} \prod_{i=1}^{p} \frac{2^{-(\delta+v_i)/2}}{\Gamma\left(\frac{\delta+v_i}{2}\right)}, \tag{4.21}$$

where

$$v = \sum_{j=1}^{k} c_j(c_j - 1)/2 - \sum_{j=2}^{k} s_j(s_j - 1)/2. \tag{4.22}$$

In order to define the inverse of the hyper Inverse Wishart distribution, it is necessary to consider the change of variables $\mathbf{K} = \mathbf{\Sigma}^{-1}$, which will actually transform the elements σ_{ij} of $\mathbf{\Sigma}$ matrix, such that $(i,j) \in E$, to non-zero entries of matrix $\mathbf{K}$. The Jacobian matrix J for this transformation is provided by Roverato and Whittaker (1998), that is

$$|J| = \frac{\prod_{i=1}^{k} det(\mathbf{\Sigma}_{C_i})^{|C_i|+1}}{\prod_{i=2}^{k} det(\mathbf{\Sigma}_{S_i})^{|S|_i|+1}}. \tag{4.23}$$

Thus, the density of $\mathbf{K}$ will be provided by substituting $\mathbf{\Sigma} = K^{-1}$ in Equation 4.17, multiplied by the Jacobian matrix of Equation 4.23, resulting to

$$\pi(\mathbf{K}|\delta, \mathbf{A}, G) \propto det(K)^{(\delta-2)/2} \exp\Big\{ -\frac{1}{2}tr(KA^{-1})\Big\}, \tag{4.24}$$

where $A = B^{-1}$ and $K, A \in M^{+}(G)$. If G is the complete graph, then the density of K will be a Wishart distribution. If G is not the complete graph, then the density of K will be proportionate to a Wishart distribution, conditioned to the event $K \in M^{+}(G)$. This distribution is stated by Roverato (2000) as the G-'textbf*conditional Wishart distribution* (or G-Wishart distribution) and is denoted by $K \sim W_G(\delta + |V| - 1, A)$. This distribution will be useful when we will be called to generate hyper-Inverse Wishart observations in later parts of this thesis.

4.4 Objective Bayes in Undirected Decomposable Graphical Models

Given the data setup of Equation 4.1 and a set of undirected decomposable graphical models $\mathscr{G}$, we assume the Gaussian graphical model setup of Equation 4.10. Under any graph $G \in \mathscr{G}$ we consider the improper prior for $\mathbf{\Sigma}$

$$\pi^{N}(\mathbf{\Sigma}|G) \propto \frac{\prod_{C \in \mathscr{C}} det(\mathbf{\Sigma}_C)^{-|C|}}{\prod_{S \in \mathscr{S}} det(\mathbf{\Sigma}_S)^{-|S|}}, \tag{4.25}$$

where the covariance matrix $\mathbf{\Sigma}$ will live in $M^{+}(G)$, exploits the same factorization over cliques and separators as in Equation 4.11 and will expedite a computationally convenient and conjugate setup moving forward. If we wish to proceed with the comparison of two competing models $G_i, G_j \in \mathscr{G}$, it will be required to provide the Bayes factor of G_j versus G_i, but it would result to indeterminate Bayes factors as per Equation 3.11. To this end, there are several applications of the Fractional Bayes Factor of O'Hagan for alleviating the indeterminancy cause by the ratio of arbitrary constants, with applications to DAG models as per Consonni et al. (2017), Castelletti et al. (2018), and to undirected decomposable graphical models as per Carvalho and Scott (2009). We will present the latter, that is the application of the Fractional Bayes Factor to the graphical model selection of undirected Gaussian graphical models since it falls under the context of this thesis.

4.4.1 Fractional Bayes Factor approach

Following the setup provided in Equation 4.10 and the data structure of Equation 4.1, we consider two competing models $G_i, G_j \in \mathscr{G}$. Starting from the improper prior of Equation 4.25 and letting g denoting the fraction parameter as in Equation 3.4.2, the Fractional Bayes factor of G_j versus G_i will be provided by

$$FBF_{G_j:G_i}(g,\mathbf{Y}) = \frac{Q_j(g,\mathbf{Y})}{Q_i(g,\mathbf{Y})} \tag{4.26}$$

where $Q_j(b,\mathbf{Y})$ represents the fractional marginal likelihood of the data matrix $\mathbf{Y}$ with respect to G_j and is provided by

$$Q_j(g,\mathbf{Y}) = \frac{\int f(\mathbf{Y}|\mathbf{\Sigma},G_j)\pi^N(\mathbf{\Sigma}|G_j)d\mathbf{\Sigma}}{\int f(\mathbf{Y}|\mathbf{\Sigma},G_j)^g\pi^N(\mathbf{\Sigma}|G_j)d\mathbf{\Sigma}}. \tag{4.27}$$

Following Carvalho and Scott (2009), the fractional marginal likelihood of Equation 4.27 will be provided by

$$Q_j(g,\mathbf{Y}) = (2\pi)^{-nq/2}\frac{h(G_j,gn,g\mathbf{Y}^T\mathbf{Y})}{h(G_j,b,\mathbf{Y}^T\mathbf{Y})} \tag{4.28}$$

where for any given graph $G \in \mathscr{G}$, $g > 0$ and $D \in M^+(G)$,

$$h(G,g,D) = \frac{\prod_{C\in\mathscr{C}} det(\frac{1}{2}D_C)^{\frac{g+|C|-1}{2}}\Gamma_{|C|}(\frac{g+|C|-1}{2})^{-1}}{\prod_{S\in\mathscr{S}} det(\frac{1}{2}D_S)^{\frac{g+|S|-1}{2}}\Gamma_{|S|}(\frac{g+|S|-1}{2})^{-1}}. \tag{4.29}$$

Fractional Bayes factor allows us to facilitate direct pairwise model comparison as well as comparing models through a common baseline model. Carvalho and Scott (2009) in section 4.2 of the respective paper, present the Fractional Bayes factor of any given model $G \in \mathscr{G}$ versus the null graphical model G_0 which can also provide with pairwise model comparison through the null gaphical model, as we will later perform under Expected and Power Expected Posterior prior approach.

4.5 Expective Posterior Prior Approach

The Expected Posterior prior approach is an automated prior generation procedure, by utilizing imaginary data (see subsection 3.5.3). Let $\mathbf{Y}^*$ denote be the $m \times q$ matrix consisted by these observations, in a similar fashion as in Equation 4.1. We let $\mathbf{Y}$ and $\mathbf{Y}^*$ to be considered independent on a common sample space $\mathscr{Y}$. Thus, starting with the default prior of Equation 4.25 under any given graph $G \in \mathscr{G}$, the posterior distribution of $\mathbf{\Sigma}$ given $\mathbf{Y}^*$ is provided by

$$\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*,G) = \frac{f(\mathbf{Y}^*|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)}{\int f(\mathbf{Y}^*|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)d\mathbf{\Sigma}}, \tag{4.30}$$

where where $f(\mathbf{Y}^*|\mathbf{\Sigma},G)$ represents the density of $\mathbf{Y}^*$ under model G as in Equation 4.11. Thus, given a predictive density $m^*(\cdot)$, the Expected Posterior prior of $\mathbf{\Sigma}$ under $G \in \mathscr{G}$ will be provided by

$$\pi^{EPP}(\mathbf{\Sigma}|G) = \int \pi^N(\mathbf{\Sigma}|\mathbf{Y}^*,G)m^*(\mathbf{Y}^*)d\mathbf{Y}^*. \tag{4.31}$$

For the $\pi^{EPP}(\mathbf{\Sigma}|G)$ to be well defined given a density $m^*(\cdot)$ on the sample space $\mathscr{Y}^*$, following the conditions of Pérez and Berger (2002) that were presented in Equation 3.40, we assume for all $\mathbf{\Sigma} \in M^+(G)$ that

$$0 < m^N(\mathbf{Y}^*|G) < \infty, \qquad 0 < E_G\left[\frac{m^*(\mathbf{Y}^*)}{m^N(\mathbf{Y}^*|G)}\middle|\mathbf{\Sigma}\right] \tag{4.32}$$

hold; the expectation is derived with respect to $f(\mathbf{Y}^*|\mathbf{\Sigma}, G)$.

As we previously described in subsection 3.5.2, the size of the imaginary data is to be maintained at a minimum to ensure that their effect to the respective posterior analysis is to be kept at a minimum. Therefore, we need to define a suitable size for the number of rows m of the imaginary data matrix $\mathbf{Y}^*$. Given a decomposable graph $G \in \mathscr{G}$, the imaginary data matrix $\mathbf{Y}^*$ will comply with the same structure and laws of the observation matrix $\mathbf{Y}$ as in Equation 4.1. Under each clique $C \in \mathscr{C}$, the matrix $\mathbf{Y}^*_C$ follows a Matrix Normal distribution such that $(\mathbf{Y}^*_C|\mathbf{\Sigma}_C, G) \sim MN_{m\times q}(\mathbf{0}, \mathbf{I}_m, \mathbf{\Sigma}_C)$. By utilizing the improper prior Equation 4.25, the posterior distribution of $\mathbf{\Sigma}_C$ given $\mathbf{Y}^*_C$ under each $C \in \mathscr{C}$, will be an Inverse Wishart distribution, with degrees of freedom parameter $b = m$ and scale matrix $D = \mathbf{S}^*_C$, where $\mathbf{S}^*_C = \mathbf{Y}^{*^T}_C \mathbf{Y}^*_C$.

For the posterior density of $\mathbf{\Sigma}_C$ given $\mathbf{Y}^*_C$ to be proper, we need to ensure the following conditions. First, the degrees of freedom parameter b, which is represented by the number of rows m of $\mathbf{Y}^*$ must always be $m \geq |C|$, $\forall C \in \mathscr{C}$, $\forall G \in \mathscr{G}$ and respectively, the matrix $\mathbf{S}_C$ must be s.p.d under each $C \in \mathscr{C}$, which is true for every $C \in \mathscr{C}$. Thus, by scaling up to the maximal clique that one can meet in $\mathscr{G}$, that is under the full graph (figure ?? for $q = 10$ node example) is $|C_{max}| = q$. Therefore, for obtaining a proper posterior density of $\mathbf{\Sigma}_C$ given $\mathbf{Y}^*_C$ for each $C \in \mathscr{C}$ under each $G \in \mathscr{G}$, we require to have degrees of freedom parameter $m \geq q$.

Using the EPP of Equation 4.31 we can directly compute marginal likelihoods by

$$m^{EPP}(\mathbf{Y}|G) = \int f(\mathbf{Y}|\mathbf{\Sigma}, G)\pi^{EPP}(\mathbf{\Sigma}|G)d\mathbf{\Sigma} \tag{4.33}$$

and calculate Bayes Factors and Posterior model odds.

4.5.1 Base Model Approach

As we described earlier in subsection 3.5.3, EPP approach facilitates and automates OB model selection using a reference model. Graphical model selection as well, is well suited for this approach, since it is easy to specify a reference model, that is the null graphical model G_0. Thus, the predictive density $m^*(\cdot)$ will be defined as the marginal density of the observations under model G_0, i.e.

$$m^*(\mathbf{Y}^*|G_0) = m^N(\mathbf{Y}^*|G_0) \propto \int f(\mathbf{Y}^*|\mathbf{\Sigma}, G_0)\pi^N(\mathbf{\Sigma}|G_0)d\mathbf{\Sigma}; \tag{4.34}$$

see further Appendix A.0.12.Thus, the EPP under the base-model will reduce to its respective default prior i.e.

$$\pi^{EPP}(\mathbf{\Sigma}|G_0) = \pi^N(\mathbf{\Sigma}|G_0). \tag{4.35}$$

Under a general Graph $G \in \mathscr{G}$ the respective EPP will be provided by

$$\pi^{EPP}(\mathbf{\Sigma}|G) = \int \pi^N(\mathbf{\Sigma}|\mathbf{Y}^*, G) m^N(\mathbf{Y}^*|G_0) d\mathbf{Y}^*. \tag{4.36}$$

Though each respective EPP contains an arbitrary constant due to the use of the baseline prior $\pi^N(\mathbf{\Sigma}|G_0)$, we can calculate Bayes factors and Posterior odds, since this arbitrary constant is common for every model of set $\mathscr{G}$.

The predictive density $m^*(\mathbf{Y}^*|G_0)$ is improper, thus it cannot be directly used to generate the imaginary data matrix $\mathbf{Y}^*$. Thus, we resort to the Importance Sampling scheme to approximate Bayes factors of any given model $G \in \mathscr{G}$ versus G_0, as previously described in subsection 3.5.4.

Following Equation 3.41, the marginal density of sample data matrix $\mathbf{Y}$ given the imaginary data matrix $\mathbf{Y}^*$ under a graph $G \in \mathscr{G}$, will be provided by

$$m^N(\mathbf{Y}|\mathbf{Y}^*, G) = \frac{m^N(\mathbf{Y}, \mathbf{Y}^*|G)}{m^N(\mathbf{Y}^*|G)}, \tag{4.37}$$

where the marginal densities of Equation 4.37 are derived with respect to the baseline prior of Equation 4.25 i.e.

$$m^N(\mathbf{Y}, \mathbf{Y}^*|G) \propto \int f(\mathbf{Y}, \mathbf{Y}^*|\mathbf{\Sigma}, G)\pi^N(\mathbf{\Sigma}|G) d\mathbf{\Sigma} \tag{4.38}$$

and

$$m^N(\mathbf{Y}^*|G) \propto \int f(\mathbf{Y}^*|\mathbf{\Sigma}, G)\pi^N(\mathbf{\Sigma}|G) d\mathbf{\Sigma}. \tag{4.39}$$

Note that both marginal likelihoods of Equation 4.38 and Equation 4.39 are provided as a proportion of the right part of the equations described, yet Equation 4.37 can be expressed with the equal sign since both numerator and denominator contain the same arbitrary normallizing constant. Thus, the marginal likelihood of Equation 4.33 can be re-written as

$$m^{EPP}(\mathbf{Y}|G) = \int m^N(\mathbf{Y}|\mathbf{Y}^*, G) m^*(\mathbf{Y}^*|G_0) d\mathbf{Y}^*; \tag{4.40}$$

see further Appendix A.0.2. Now consider $\mathbf{Y}^{*^{(1)}}, \cdots, \mathbf{Y}^{*^{(R)}}$ be a sample of independent and identically distributed observations of $m \times q$ imaginary data matrices from an importance density $g(\mathbf{Y}^*)$ and let G and $G^{'}$ be two competing graphical models of $\mathscr{G}$. We approximate the Bayes factor of G versus $G^{'}$ using the generated imaginary observations, by

$$BF^{EPP}_{G:G^{'}}(\mathbf{Y}) \approx \widehat{BF}^{EPP}_{G:G^{'}}(\mathbf{Y}) = \frac{\sum_{l=1}^{r} m^N(\mathbf{Y}|\mathbf{Y}^*_l, G) m^*(\mathbf{Y}^*_l)/g(\mathbf{Y}^*_l)}{\sum_{l=1}^{r} m^N(\mathbf{Y}|\mathbf{Y}^*_l, G^{'}) m^*(\mathbf{Y}^*_l)/g(\mathbf{Y}^*_l)}, \tag{4.41}$$

where the marginal densities are provided using Equation 4.37.

Following the base-model approach, we can use as importance density $g(\mathbf{Y}^*|G_0) = m^N(\mathbf{Y}^*|\mathbf{Y}, G_0)$, which is provided by reverting $\mathbf{Y}$ with $\mathbf{Y}^*$ in Equation 4.37. Having $G^{'} = G_0$, Equation 4.41 is reduced to

$$BF^{EPP}_{G:G_0}(\mathbf{Y}) \approx \widehat{BF}^{EPP}_{G:G_0}(\mathbf{Y}) = \frac{1}{R}\sum_{r=1}^{R} BF^N_{G:G_0}(\mathbf{Y}|\mathbf{Y}^{*(r)}), \tag{4.42}$$

where the Bayes factor inside the sum expression will be provided for every $l = 1, \cdots, R$

$$BF^N_{G:G_0}(\mathbf{Y}|\mathbf{Y}^{*(r)}) = \frac{m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G)}{m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G_0)} \tag{4.43}$$

and the marginal likelihoods are given by Equation 4.37; see further Appendix A.0.4. If the computational cost of Equation 4.43 is heavy due to the computation of $m^N(\mathbf{Y}|\mathbf{Y}^*, G)$, we can use as importance density in the numerator of Equation 4.42 $g(\mathbf{Y}^*|G) = m^N(\mathbf{Y}^*|\mathbf{Y}, G)$ and replace the importance density of the denominator with $g(\mathbf{Y}^*|G_0) = m^N(\mathbf{Y}^*|\mathbf{Y}, G_0)$. So Equation 4.42 will be reduced to

$$BF^{EPP}_{G:G_0}(\mathbf{Y}) \approx \widehat{BF}^{EPP}_{G:G_0}(\mathbf{Y}) = BF^N_{G:G_0}(\mathbf{Y})\frac{1}{R}\sum_{r=1}^{R} BF^N_{G_0:G}(\mathbf{Y}^{*(r)}), \tag{4.44}$$

where

$$BF^N_{G:G_0}(\mathbf{Y}) = \frac{m^N(\mathbf{Y}|G)}{m^N(\mathbf{Y}|G_0)} \tag{4.45}$$

and

$$BF^N_{G_0:G}(\mathbf{Y}^{*(r)}) = \frac{m^N(\mathbf{Y}^{*(r)}|G_0)}{m^N(\mathbf{Y}^{*(r)}|G)}; \tag{4.46}$$

see further Appendix A.0.3. The Bayes factors included in Equation 4.44 are calculated with respect to the marginal densities that make use of the baseline improper prior of Equation 4.25.

Thus, the Posterior Model Odds of any given model $G \in \mathscr{G}$ versus G_0, are provided by

$$PO^{EPP}(G : G_0) = BF^{EPP}_{G:G_0}(\mathbf{Y})O(G : G_0) \tag{4.47}$$

where $O(G : G_0)$ indicate the prior Model Odds, which will be provided in latter parts of this Section. Therefore, using Equation 4.47 we are able to provide evidence of every model $G \in \mathscr{G}$ versus the null graphical model G_0 and furthermore, we perform pairwise model comparisons by comparing pairwise comparisons of evidences versus G_0.

Finally, we need to define a suitable generation procedure of imaginary matrices for facilitating the base model approach, where we resort to the Gibbs sampling scheme provided in subsection 3.5.4. Thus, we first need to re-write the importance density $g(\mathbf{Y}^*|G) = m^N(\mathbf{Y}^*|\mathbf{Y}, G)$ as

$$m^N(\mathbf{Y}^*|\mathbf{Y}, G) = \int f(\mathbf{Y}^*|\mathbf{\Sigma}, G)\pi^N(\mathbf{\Sigma}|\mathbf{Y}, G)d\mathbf{\Sigma}, \tag{4.48}$$

where $f(\mathbf{Y}^*|\mathbf{\Sigma}, G)$ represents the density of $\mathbf{Y}^*$ under model G and $\pi^N(\mathbf{\Sigma}|\mathbf{Y}, G)$ is the posterior distribution of $\mathbf{\Sigma}$ given data $\mathbf{Y}$ under model $G \in \mathscr{G}$; see further Appendix A.0.5. Using this expression facilitates, under the graph $G \in \mathscr{G}$, the imaginary data generation procedure, will be structured as:

```
For $r = 1, \cdots, R$
    1. Generate $\mathbf{\Sigma}^{(r)}$ from posterior $\pi^N(\mathbf{\Sigma}|\mathbf{Y}, G)$.
    2. Generate $\mathbf{Y}^{*(r)}$ from $f(\mathbf{Y}^*|\mathbf{\Sigma}^{(r)}, G)$.
```

Note that if we decide to use $g(\mathbf{Y}^*) = m^N(\mathbf{Y}^*|\mathbf{Y}, G_0)$ as importance density, then every Bayes factor will be calculated with respect to an one-time generated sample of imaginary data. Though, if we decide to use $g(\mathbf{Y}^*) = m^N(\mathbf{Y}^*|\mathbf{Y}, G)$ as importance density for each respective model G, then we must generate a different sample under each model G.

4.6 Power Expected Posterior Prior Approach

As we previously described in subsection 3.5.5, the PEPP approach was developed for alleviating the dependency of EPP to computationally costly averages over submatrices of imaginary data. Under the graphical model selection problem we dont face the same constraints and we can directly use an imaginary data matrix. We are intrested to the feature of PEPP, that is reducing the effect of the imaginary data to the posterior analysis.

Let us consider a sample data matrix $\mathbf{Y}$ as provided in Equation 4.1 and let $\mathscr{G}$ denote the entire collection of all undirected decomposable Gaussian graphical models on q nodes and G_0 be the null graphical model. Given a graph $G \in \mathscr{G}$ we consider the improper default prior of Equation 4.11.

Let $(\mathbf{y}_1^*, \cdots, \mathbf{y}_m^*)$ be m independent imaginary observations and $\mathbf{Y}^*$ be the $m \times q$ matrix consisted by these observations, in a similar fashion as Equation 4.1. We let $\mathbf{Y}$ and $\mathbf{Y}^*$ to be considered independent on a common sample space $\mathscr{Y}$, as in EPP approach of section 4.5.

Prior to the implementetion of the PEPP approch to the graphical model selection context of undirected decomposable graphical models, we first need to define the notion of PEPP likelihood , that is the likelihood of the imaginary data matrix $\mathbf{Y}^*$. As seen in previous section, the likelihood of the imaginary data matrix $\mathbf{Y}^*$ under the EPP approach is Matrix Normal distribution such that $\mathbf{Y}^* \sim MN_{m\times q}(\mathbf{0}, \mathbf{I}_m, \mathbf{\Sigma})$ where $\mathbf{\Sigma} \in M^+(G)$. So, following section subsection 3.5.5, the PEPP likelihood will be given by the EPP likelihood of $\mathbf{Y}^*$ raised in the power of $\frac{1}{\delta}$ and density-normalized, i.e.

$$f(\mathbf{Y}^*|\mathbf{\Sigma}, \delta, G) = \frac{f(\mathbf{Y}^*|\mathbf{\Sigma}, G)^{1/\delta}}{\int f(\mathbf{Y}^*|\mathbf{\Sigma}, G)^{1/\delta} d\mathbf{Y}^*} = MN_{m\times q}(\mathbf{0}, I_m, \delta\mathbf{\Sigma}) \tag{4.49}$$

which exploits the factorization under cliques and separators as in Equation 4.11; see further Appendix A.0.1. Thus, under a given graph $G \in \mathscr{G}$, for each clique $C \in \mathscr{C}$ we obtain that $\mathbf{Y}^*_C \sim MN_{m\times|C|}(\mathbf{0}, \mathbf{I}_m, \delta\mathbf{\Sigma}_C)$ and similarly for each separator $S \in \mathscr{S}$, $\mathbf{Y}^*_S \sim MN_{m\times|S|}(\mathbf{0}, \mathbf{I}_m, \delta\mathbf{\Sigma}_S)$. As previously stated before in subsection 3.5.5, Fouskakis et al. (2015) default choice for the size of the imaginary data is $m = n$, they explicitely state that, under the variable selection context, m can have any value between $p + 2$ and n, and subsequently $\delta \in [p + 2, n]$. As we will describe in subsection 5.2.2, in our context we consider the minimal value available based on the computational cost of the respective approach.

Therefore, under any graph $G \in \mathscr{G}$ the posterior distribution of $\mathbf{\Sigma}$ given the imaginary data matrix $\mathbf{Y}^*$ with respect to the default improper prior of Equation 4.25, will be provided by

$$\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*, \delta, G) = \frac{f(\mathbf{Y}^*|\mathbf{\Sigma}, \delta, G)\pi^N(\mathbf{\Sigma}|G)}{m^N(\mathbf{Y}^*|\delta, G)}; \tag{4.50}$$

see further Appendix A.0.11. The quantity $m^N(\mathbf{Y}^*|\delta, G)$ is the marginal likelihood of the imaginary data matrix $\mathbf{Y}^*$ under model $G \in \mathscr{G}$, which is provided by

$$m^N(\mathbf{Y}^*|\delta, G) = \int f(\mathbf{Y}^*|\mathbf{\Sigma}, \delta, G)\pi^N(\mathbf{\Sigma}|G)d\mathbf{\Sigma}. \tag{4.51}$$

Thus, the Power Expected Posterior Prior of $\mathbf{\Sigma}$ under any model $G \in \mathscr{G}$ is defined by

$$\pi^{PEPP}(\mathbf{\Sigma}|\mathbf{Y}^*, \delta, G) = \int \pi^N(\mathbf{\Sigma}|\mathbf{Y}^*, \delta, G)m^N(\mathbf{Y}^*|\delta, G_0)d\mathbf{Y}^*, \tag{4.52}$$

where $\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*, \delta, G)$ is defined by Equation 4.50 and $m^N(\mathbf{Y}^*|\delta, G_0)$ by Equation 4.51 for the case of $G = G_0$; see further Appendix A.0.13. The marginal likelihood that is defined under model G_0, acts as the predictive density $m^*(\cdot)$, as similarly defined under EPP procedure. Therefore, for the implementation of the PEPP approach to our contest, we wil be using the base model approach as in subsection 4.5.1.

By using Equation 4.51, Equation 4.52 could be written as,

$$\pi^{PEPP}(\mathbf{\Sigma}|\mathbf{Y}^*, \delta, G) = \pi^N(\mathbf{\Sigma}|G) \int \frac{m^N(\mathbf{Y}^*|\delta, G_0)}{m^N(\mathbf{Y}^*|\delta, G)} f(\mathbf{Y}^*|\mathbf{\Sigma}, \delta, G)d\mathbf{Y}^*, \tag{4.53}$$

If we consider a data matrix $\mathbf{Y}$, then the posterior distribution of $\mathbf{\Sigma}$ under model the G with respect to the PEPP prior $\pi^{PEPP}(\mathbf{\Sigma}|\mathbf{Y}^*, \delta, G)$, will be provided by

$$\pi^{PEPP}(\mathbf{\Sigma}|\mathbf{Y}, \delta, G) \propto \int \pi^N(\mathbf{\Sigma}|\mathbf{Y}, \mathbf{Y}^*, \delta, G)m^N(\mathbf{Y}|\mathbf{Y}^*, \delta, G)m^N(\mathbf{Y}^*|\delta, G_0)d\mathbf{Y}^*, \tag{4.54}$$

where $\pi^N(\mathbf{\Sigma}|\mathbf{Y},\mathbf{Y}^*,\delta,G)$ is the posterior distribution of $\mathbf{\Sigma}$ under model G and $m^N(\mathbf{Y}|\mathbf{Y}^*,\delta,G)$ is the marginal likelihood of model G, respectively, using data $\mathbf{Y}$ with respect to the prior distribution $\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*,\delta,G)$ as defined in Equation 4.50; see further Appendix A.0.11.

So now we are able to proceed with the model selection procedure by deriving Posterior model odds based on the PEPP procedure. Given a data matrix $\mathbf{Y}$, the marginal likelihood of $\mathbf{Y}$ under a model G with respect to the PEPP prior of Equation 4.53, will be provided by

$$\begin{aligned} m^{PEPP}(\mathbf{Y}|\delta,G) &= \int f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^{PEPP}(\mathbf{\Sigma}|\mathbf{Y}^*,\delta,G)d\mathbf{\Sigma} \\ &= m^N(\mathbf{Y}|G)\int \frac{m^N(\mathbf{Y}^*|\delta,G_0)}{m^N(\mathbf{Y}^*|\delta,G)} m^N(\mathbf{Y}^*|\mathbf{Y},\delta,G)d\mathbf{Y}^* \end{aligned} \tag{4.55}$$

where these marginals are defined as in Equation 4.37; see further Appendix A.0.7. Thus, the Bayes factor a random graph $G \in \mathscr{G}$ against the null graphical model G_0 will be provided by

$$BF^{PEPP}_{G:G_0}(\mathbf{Y},\delta) = \frac{m^{PEPP}(\mathbf{Y}|\delta,G)}{m^{PEPP}(\mathbf{Y}|\delta,G_0)} \tag{4.56}$$

Fouskakis et al. (2015) in Section 3 propose two similar simulation schemes for approximating the marginal likelihood of Equation 4.55. We will apply the first proposed scheme which is compatible with the Importance Sampling procedure that Pérez and Berger (2002) Section 4.2. Thus, the approximation of the marginal likelihood of Equation 4.55 using as importance density $g(\mathbf{Y}^*) = m^N(\mathbf{Y}^*|\mathbf{Y},\delta,G)$ will be structured as:

- Generate $\mathbf{Y}^{*(1)},\cdots,\mathbf{Y}^{*(R)}$ from $m^N(\mathbf{Y}^*|\mathbf{Y},\delta,G)$.
- Estimate the marginal likelihood of Equation 4.55 by:

$$m^{PEPP}(\mathbf{Y}|\delta,G) \approx \widehat{m}^{PEPP}(\mathbf{Y}|\delta,G) = m^N(\mathbf{Y}|G)\frac{1}{R}\sum_{r=1}^{R}\frac{m^N(\mathbf{Y}^{*(r)}|\delta,G_0)}{m^N(\mathbf{Y}^{*(r)}|\delta,G)}. \tag{4.57}$$

Therefore, the Bayes factor of a graphical model $G \in \mathscr{G}$ versus the null graph

G_0 can be approximated by:

$$\begin{aligned}
BF^{PEPP}_{G:G_0}(\mathbf{Y}|\delta) &\approx \widehat{BF}^{EPP}_{G:G_0}(\mathbf{Y}) \\
&= \frac{\widehat{m}^{PEPP}(\mathbf{Y}|\delta, G)}{\widehat{m}^{PEPP}(\mathbf{Y}|\delta, G_0)} \\
&= \frac{m^N(\mathbf{Y}|G)\frac{1}{R}\sum_{r=1}^{R}\frac{m^N(\mathbf{Y}^{*(r)}|\delta,G_0)}{m^N(\mathbf{Y}^{*(r)}|\delta,G)}}{m^N(\mathbf{Y}|G_0)\frac{1}{R}\sum_{r=1}^{R}\frac{m^N(\mathbf{Y}^{*(r)}|\delta,G_0)}{m^N(\mathbf{Y}^{*(r)}|\delta,G_0)}} \\
&= \frac{m^N(\mathbf{Y}|G)}{m^N(\mathbf{Y}|G_0)}\frac{1}{R}\sum_{r=1}^{R}\frac{m^N(\mathbf{Y}^{*(r)}|\delta, G_0)}{m^N(\mathbf{Y}^{*(r)}|\delta, G)} \\
&= BF^N_{G:G_0}(\mathbf{Y})\frac{1}{R}\sum_{r=1}^{R} BF^N_{G_0:G}(\mathbf{Y}^{*(r)}, \delta), \qquad (4.58)
\end{aligned}$$

which is identical with the approximation provided under the EPP approach in Equation 4.44. By using Equation 4.58 we are able to provide directly Posterior Model Odds of any given model $G \in \mathscr{G}$ against G_0, by

$$PO^{PEPP}(G : G_0) = BF^{PEPP}_{G:G_0}(\mathbf{Y})O(G : G_0) \qquad (4.59)$$

where $O(G : G_0)$ represent the prior Model Odds, which will be explored in a latter section of this chapter. The generation procedure of the imaginary data matrices $\mathbf{Y}^{*(1)}, \cdots, \mathbf{Y}^{*(R)}$ is identical to the Importance sampling scheme that Pérez and Berger (2002) Section 4.2 used and we presented in subsection 4.5.1. by re-writing the importance density $g(\mathbf{Y}^*)$ as

$$g(\mathbf{Y}^*) = m^N(\mathbf{Y}^*|\mathbf{Y}, \delta, G) = \int f(\mathbf{Y}^*|\mathbf{\Sigma}, \delta, G)\pi^N(\mathbf{\Sigma}|\mathbf{Y}, \delta, G)d\mathbf{\Sigma}, \qquad (4.60)$$

the Gibbs-Sampling scheme for generating the imaginary observations $\mathbf{Y}^{*(1)}, \cdots, \mathbf{Y}^{*(R)}$ from $m^N(\mathbf{Y}^*|\mathbf{Y}, \delta, G)$ will be structured as:

```
For $r = 1, \cdots, R$.
    1. Generate $\mathbf{\Sigma}^{(r)}$ from $\pi^N(\mathbf{\Sigma}|\mathbf{Y}, \delta, G)$.
    2. Generate $\mathbf{Y}^{*(r)}$ from $f(\mathbf{Y}^*|\mathbf{\Sigma}^{(r)}, \delta, G)$.
```

4.7 Feature-Inclusion Stochastic Search Algorithm

There are several Bayesian and non-Bayesian computational strategies for performing structure learning in real-life applications, where the choice of a suitable one is highly dependent to the size of the problem at hand. Bayesian approaches, such as the reversible jump MCMC of Giudici and Green (1999) or the approach of Jones et al. (2005), provide adequate results in small to moderate sized problems. As the number of nodes grows larger, useful alternatives can be found in Dobra et al. (2004) or Mohammadi and Wit (2015) where they developed the Birth Death MCMC. Regarding non-Bayesian approaches, L_1-regularization methods can be applied as developed by Meinshausen and Buhlmann (2006) or the graphical lasso of Friedman et al. (2008).

In the bayesian field, researchers faced two conflicting issues regarding the exploration of the graphical model space $\mathscr{G}$. Results from Giudici and Green (1999) and Wong et al (2003) provide that navigating throughout $\mathscr{G}$ using an edge-at-a-time moves is swifter than performing multiple-edge moves. This is due to the local structure implied by the hyper-inverse Wishart distribution, which allows to evaluate a candidate graph fast since it will differ from the previous one visited only by two cliques and one separator at most.

On the other hand, performing one-edge-at-a-time moves that maintain decomposability is not sufficient. The exploration of the graphical model space $\mathscr{G}$ is not adequate, since one requires a huge amount of stepwise decomposable moves to visit an acceptable number of candidate models that will be enough to yield satisfactory convergence on edge inclusion posterior probability estimates due to the multimodality problem of graphical model space $\mathscr{G}$. In reality means, that different graphs imply different topologies of space $\mathscr{G}$ that are not well connected, in a sense a decomposable graph can reach specific areas of $\mathscr{G}$, a constraint which tends to get worse with an increasing number of nodes.

The contradiction above, implies that one must cleverly combine the computational benefit of a stepwise move with a solid strategy to navigate around the graphical model space $\mathscr{G}$, using an efficient mixture of stepwise and global moves that will explore good portions of $\mathscr{G}$ which will also contain promising models. This led Carvalho and Scott (2008) to the development of a serial algorithm for applying a stochastic search procedure over $\mathscr{G}$, by combing local, resampling and global moves based on estimates of posterior edge inclusion probabilities.

The algorithm developed was named FINCS (Feature Inclusion Stochastic Search), with origins in the work of Berger and Molina (2004), where they applied a similar principal for the variable selection probelm. FINCS bases its reasoning on a simple and intuitive observation: higher posterior edge-inclusion probability estimates correspond to stronger conditional independence relationships, thus they are well suited for guiding the stochastic search algorithm, instead of randomly moving in the graphical model space. Therefore, if there are certain edges that increase the estimated posterior probability of already visited models, it is highly probable that they will guide towards other good performing models

Before establishing the stochastic search procedure operated by FINCS, we need to define the estimated posterior probability of a visited model $G^{(t)}$ and the estimated posterior edge inclusion probabilities of G_t. Let $\mathscr{G}^{(T)} = \{G^{(1)}, \cdots, G^{(t)}\} \subseteq \mathscr{G}$, with $t \geq 1$, denote a collection of visited decomposable graphical models by a stochastic search algorithm. The estimated posterior probability of the visited model G_t is provided by

$$\hat{\pi}(G^{(t)}|\mathbf{Y}) = \frac{BF_{G^{(t)}:G_0}(\mathbf{Y})\pi(G^{(t)})}{\sum_{j=1}^{t} BF_{G^{(j)}:G_0}(\mathbf{Y})\pi(G^{(j)})} \tag{4.61}$$

which is provided in comparison to the null model G_0 and $\pi(G^{(t)})$ represents the prior probability of $G^{(t)}$. We will report the estimated posterior probability of a visited model based on Bayes factors of each visited model versus the null graphical model G_0, since both EPP and PEPP are facilitated using the base model approach. Let e_{ij} denote the ij_{th}-edge of graph $G^{(t)}$. The estimated posterior edge inclusion probability of e_{ij} is provided by

$$\hat{q}_{ij} = \sum_{G \in \mathscr{G}^{(T)}} I_{e_{ij} \in G}\, \hat{\pi}(G|\mathbf{Y}). \tag{4.62}$$

Using these estimates, FINCS algorithm incorporates three different kinds of moves across the model space $\mathscr{G}$:

- Local Move: Starting from $G^{(t)} \in \mathscr{G}$, move to graph $G^{(t+1)}$ by randomly adding or deleting an edge, given that $G^{(t+1)}$ will be decomposable. The addition of edges will be applied in proportion to estimates of posterior edge inclusion probabilities at state (t) and the deletionof edges is applied in inverse proportion of the respective edge-inclusion probabilities.

- Resampling Move: Starting from a graph $G^{(t)} \in \mathscr{G}$, revisit one of $\{G^{(1)}, \cdots, G^{(t-1)}\}$ in proportion to their estimated posterior model probabilities, and then start performing local moves from that point.

- Global Move: Move to a new area of the graphical model space using a randomized median triangulation pair, using the following strategy:

  1. Start from the null graph and consider the addition of every possible edge in proportion to their posterior edge inclusion probabilities. The resulted graph G_N will usually be a non-decomposable graph. The median graph can be alternatively used i.e. the graph including edges with posterior edge inclusion probabilities greater than 0.5.
  2. Enclose G_N in $G^- \subset G_N \subset G^+$, where no edge can be added to G^- or deleted from G^+ in a way such that both graphs can maintain their decomposability.
  3. Move to one of the above in proportion to their estimated posterior probabilities and start performing local moves from that point.

Regardless of the move performed, $G^{(t+1)}$ is dealt like it has not been visited before and is being used for updating the estimated posterior edge inclusion probabilities.

FINCS algorithm has two important advantages compared to standard MCMC approaches. First, it incorporates a mixture of moves across the model space and not explicitly using local moves, leading to more efficient exploration of the graphical model space $\mathscr{G}$. The resampling move allows us to consider alternative routes of exploration and the global moves can bridge different areas of $\mathscr{G}$, which would be difficult to link using local moves exclusively. Second, it operates as a purely heuristic search algorithm, without requiring the convergence of the algorithm to a stationary distribution or always maintain a certain level of acceptance probabilities. Following the experimental studies performed by Carvalho and Scott (2008), in every simulation setup considered FINCS algorithm always explored a greater area of $\mathscr{G}$ compared to Metropolis approaches.

4.8 Priors on Graphs

For the calculation of Posterior model odds of any given $G \in \mathscr{G}$ versus the independence graph G_0, besides the respective Bayes factor, we need to calculate the prior model odds. One standard choice, is applying a uniform probability over the graphical model space $\mathscr{G}$, i.e. each graph will have the same prior probability

$$\pi(G) = \frac{1}{|G|}, \quad \forall G \in \mathscr{G}, \tag{4.63}$$

which is sensible in absence of prior information. Yet, following Giudici and Green (1999) section 1.3, this prior probability will favor middle sized graphs due to their vast number compared to others of $\mathscr{G}$.

A useful alternative, is to focus on the prior probability of an edge being included to the graph instead of the graph itself, which is similar to the feature inclusion probability of the variable selection problem. Let r denote the probability of an edge being included in the graph. Following Dobra et al. (2004), if we consider a binomial prior over the probability of inclusion r, the prior o a given graph $G \in \mathscr{G}$ is provided by

$$\pi(G) \propto r^k (1-r)^{m-k}, \tag{4.64}$$

where k denotes the number of edges being included in $G \in \mathscr{G}$ and $m = q(q-1)/2$ indicate the maximum number of possible edges. The prior is provided in proportion to the kernel of the distributions, since the constant of proportionality is identical for every $G \in \mathscr{G}$, thus it can be omitted and alleviate the computational cost required for its calculation. This approach is amenable, if we acquire prior information regarding the number of edges included in the true graph, which is highly unlikely in real-life application.

To avoid the incremental computational cost required for the estimation of r, Carvalho and Scott (2009) consider applying a Beta prior over r, such that

$$\pi(G) \propto \frac{B(\alpha + k, \beta + m - k)}{B(\alpha, \beta)}, \tag{4.65}$$

where $B(\cdot, \cdot)$ denotes the Beta function. By considering a $\alpha = \beta = 1$, we assign a uniform prior over the edge inclusion probability r, such that

$$\pi(G) \propto \frac{1}{m+1} \binom{m}{k}^{-1}. \tag{4.66}$$

As Carvalho and Scott (2009) indicate, this prior automatically penalizes the inclusion of false positive edges, as the dimension grows larger. In this thesis, this prior will be used for obtaining posterior model estimates, using EPP and PEPP.

Chapter 5

Experiments

In this chapter we evaluate the ability of EPP and PEPP to infer an undirected graph from simulated data and protein-signaling data. Furthermore, we compare EPP and PEPP with the Fractional Bayes Factor approach (FBF) of Carvalho and Scott (2009) and the Birth Death MCMC (BDMCMC), a methodology introduced by Mohammadi and Wit (2015) and implemented by `BDgraph` package.

5.1 Preliminaries

Our goal is, given a graphical model space $\mathscr{G}$ and a data matrix $\mathbf{Y}$, to obtain insights regarding the posterior model probability $\pi(G|\mathbf{Y})$ under any $G \in \mathscr{G}$ (see also chapter 4). We either perform a full enumeration of $\mathscr{G}$, that is the calculation of each posterior model probability of every distinct $G \in \mathscr{G}$, or with a stochastic search of $\mathscr{G}$, that is the exploration of the graphical model space for the most promising models based on estimations of posterior probabilities. The decision on which of the two will be applied, is purely based on the size of $\mathscr{G}$. With fewer number of nodes under consideration, the size of $\mathscr{G}$ is smaller therefore a full enumeration framework is feasible, whereas with greater number of nodes, the size of $\mathscr{G}$ grows superexponentially and a stochastic search framework is considered. We provide examples for both frameworks.

Uncer a full enumberation framework, we consider $\mathscr{G} = \{G_0, \cdots, G_L\}$ to denote the graphical model space of L distinct undirected decomposable graphical models, with G_0 being the null model and nested to all other graphical models of $\mathscr{G}$. The posterior probability of a given $G \in \mathscr{G}$, following Equation 3.7 is calculated by

$$\pi(G|\mathbf{Y}) = \frac{PO_{G:G_0}}{\sum_{G_l \in \mathscr{G}} PO_{G_l:G_0}}, \tag{5.1}$$

where for every $G \in \mathscr{G}$

$$PO_{G:G_0} = BF_{G:G_0}(\mathbf{Y}) O_{G:G_0}; \tag{5.2}$$

see further Appendix A.0.9. Under a stochastic search framework, after visiting a set of K distinct models $\mathscr{G}^* = \{G^{(0)}, \cdots, G^{(K)}\} \subseteq \mathscr{G}$ based on the output of a stochastic search algorithm of $T >= K$ iterations, we estimate the posterior model probabilities $\hat{\pi}(G|\mathbf{Y})$ of any $G \in \mathscr{G}^*$, using Equation 5.1, by replacing $\mathscr{G}$ with $\mathscr{G}^*$.

Under the graphical model selection framework, we will evaluate the output of the stochastic search algorithm based on the ***median probability (graphical) model***, which can be constructed by including the all edges with posterior edge inclusion probability greater that 0.5; the notion of the median probability model was originally defined in Barbieri and Berger (2004) for the variable selection problem, providing to better predictive performance. An edge posterior inclusion probability of an edge e_{ij}, given the data matrix $\mathbf{Y}$, is provided by

$$\hat{q}_{ij} = \sum_{G \in \mathscr{G}^*} I_{e_{ij} \in G} \hat{\pi}(G|\mathbf{Y}). \tag{5.3}$$

Thus, the median probability graphical model is defined as the graph containing all underected edges such that $\hat{q}_{ij} \geq 0.5$. Note, that it is not guaranteed for the median probability graphical model to be decomposable, but it will be used throughout this chapter for: 1) provide performance metrics of the proposed methodologies under simulated data (see section 5.2; (2) for depicting conditional independence relationships based on protein-signalling data (see section 5.3).

5.1.1 Posterior Odds formulations

In this subsection we provide analytical the analytical expressions required for approximation Bayes factors of any given graph $G \in \mathscr{G}$ versus the null graphical model G_0, under EPP and PEPP approach; for the FBF approach we provide the formulation provided by Carvalho and Scott (2009).

Let us consider a set of decomposable graphical model $\mathscr{G}$ on q nodes and data matrix as in Equation 4.1. As we previously stated in subsection 4.5.1 and section 4.6, we adopt the base-model approach where we obtain posterior model estimated through the comparison of candidate models versus the null graphical model G_0. Under the PEPP approach, the approximation of the Bayes factor of a given model $G \in \mathscr{G}$ versus the null graphical model G_0, will be provided by

$$\widehat{BF}^{PEPP}_{G:G_0}(\mathbf{Y}, \delta) = K(\mathbf{Y}, G)\, H(\mathbf{Y}, G) \sum_{r=1}^{R} K(\mathbf{Y}^*, G)\, H(\mathbf{Y}^*, G), \tag{5.4}$$

where for a data matrix $\mathbf{X}_{n \times q}$ and an undirected decomposable graph G we define

$$K(\mathbf{X}, G) = \frac{\prod_{C \in \mathscr{C}} \Gamma_{|C|}(\frac{n+|C|-1}{2})}{\prod_{S \in \mathscr{S}} \Gamma_{|S|}(\frac{n+|S|-1}{2})} \Gamma^{-q}\left(\frac{n}{2}\right) \tag{5.5}$$

and

$$H(\mathbf{X}, G, \delta) = \prod_{j=1}^{q} det(\frac{1}{2\delta}\mathbf{S}_j)^{\frac{n}{2}} \frac{\prod_{C \in \mathscr{C}} det(\frac{1}{2\delta}\mathbf{S}_C)^{-\frac{n+|C|-1}{2}}}{\prod_{S \in \mathscr{S}} det(\frac{1}{2\delta}\mathbf{S}_S)^{-\frac{n+|S|-1}{2}}}; \tag{5.6}$$

see further Appendix A.0.10. For obtaining the respective Bayes factor approximation under the EPP approach, we set $\delta = 1$ in Equations Equation 5.4, Equation 5.5

and Equation 5.6. Last, following Carvalho and Scott (2009), the Fractional Bayes factor of any given model $G \in \mathscr{G}$ versus the null graphical model G_0, will be provide by

$$
\begin{aligned}
FBF_{G:G_0}(\mathbf{Y}) = K \frac{\prod_{j=1}^{q} det(\frac{1}{2}\mathbf{S}_j)^{n/2}}{\prod_{j=1}^{q} det(\frac{g}{2}\mathbf{S}_j)^{gn/2}} & \frac{\prod_{C\in\mathscr{C}} det(\frac{g}{2}\mathbf{S}_C)^{\frac{gn+|C|-1}{2}}}{\prod_{S\in\mathscr{S}} det(\frac{g}{2}\mathbf{S}_S)^{\frac{gn+|S|-1}{2}}} \\
\times & \frac{\prod_{S\in\mathscr{S}} det(\frac{1}{2}\mathbf{S}_S)^{\frac{n+|S|-1}{2}}}{\prod_{C\in\mathscr{C}} det(\frac{1}{2}\mathbf{S}_C)^{\frac{n+|C|-1}{2}}} \qquad (5.7)
\end{aligned}
$$

where

$$
K = \left(\frac{\Gamma(\frac{gn}{2})}{\Gamma(\frac{n}{2})}\right)^q \frac{\prod_{C\in\mathscr{C}} \Gamma_{|C|}(\frac{n+|C|-1}{2})}{\prod_{S\in\mathscr{S}} \Gamma_{|S|}(\frac{n+|S|-1}{2})} \frac{\prod_{S\in\mathscr{S}} \Gamma_{|S|}(\frac{gn+|S|-1}{2})}{\prod_{C\in\mathscr{C}} \Gamma_{|C|}(\frac{gn+|C|-1}{2})} \qquad (5.8)
$$

We will use the quantities in Equation 5.4 and Equation 5.7 for estimating posterior edge inclusion probabilities and posterior model probabilities of visited models by a stochastic search, further explained in the following section. All scores above will be calculated using logarithms. For the generation process of imaginary observations under both PEPP and EPP approach we use thefollowing:

For $r = 1, \cdots, R$.

1. Generate $K^{(r)}$ from $W_G(n + q - 1, (\mathbf{S}/\delta)^{-1})$.
2. Generate $\mathbf{Y}^{*(r)}$ from $MN_{m\times q}(\mathbf{0}, I_m, \delta {K^{(r)}}^{-1})$.

Note again, in order to generate imaginary observations under EPP approach, we use the above described scheme using $\delta = 1$

5.1.2 Our approach and FINCS setup

For applying EPP and PEPP approaches, we utilize the FINCS algorithm of Carvalho and Scott (2008) with two modifications. First, both approaches add a layer of computational cost due to the importance sampling scheme required for the approximation of the respective Bayes factor. Thus, following Altomare et al. (2013), we modify the global move by deterministically selecting the median graph and then perform a local move from that point. If the median graph is not a decomposable model, we use a minimal graph triangulation scheme to get a decomposable subgraph of hypergraph of it and then perform a local move.

Second, our experience showed that we can reach to an optimal result quite fast, thus we keep the number of iterations as small as possible, instead of performing superfluous simulations runs. So our version of FINCS algorithm is structured as follows:

For $t = 1, \cdots, T$ (iterations):

1. For a given model G_t generate importance samples following Gibbs scheme.
2. Estimate Bayes factor under EPP or PEPP using Equation 5.4
3. Update posterior edge inclusion probabilities.
4. Propose a new model following FINCS logic.

For mixing local, resampling and global moves, we follow the guidelines provided by Carvalho and Scott (2008), where every 10 iteration a resampling move is applied and every 50 iterations a global move is applied.

Finally, for the PEPP approach described in section 4.6 we will use $m = q$ and subsequently $\delta = q$. We are allowed to consider any $m \in [q, n]$ as per Fouskakis et al. (2015) and we further discuss our reasoning on section 5.2.3.

5.2 Simulations

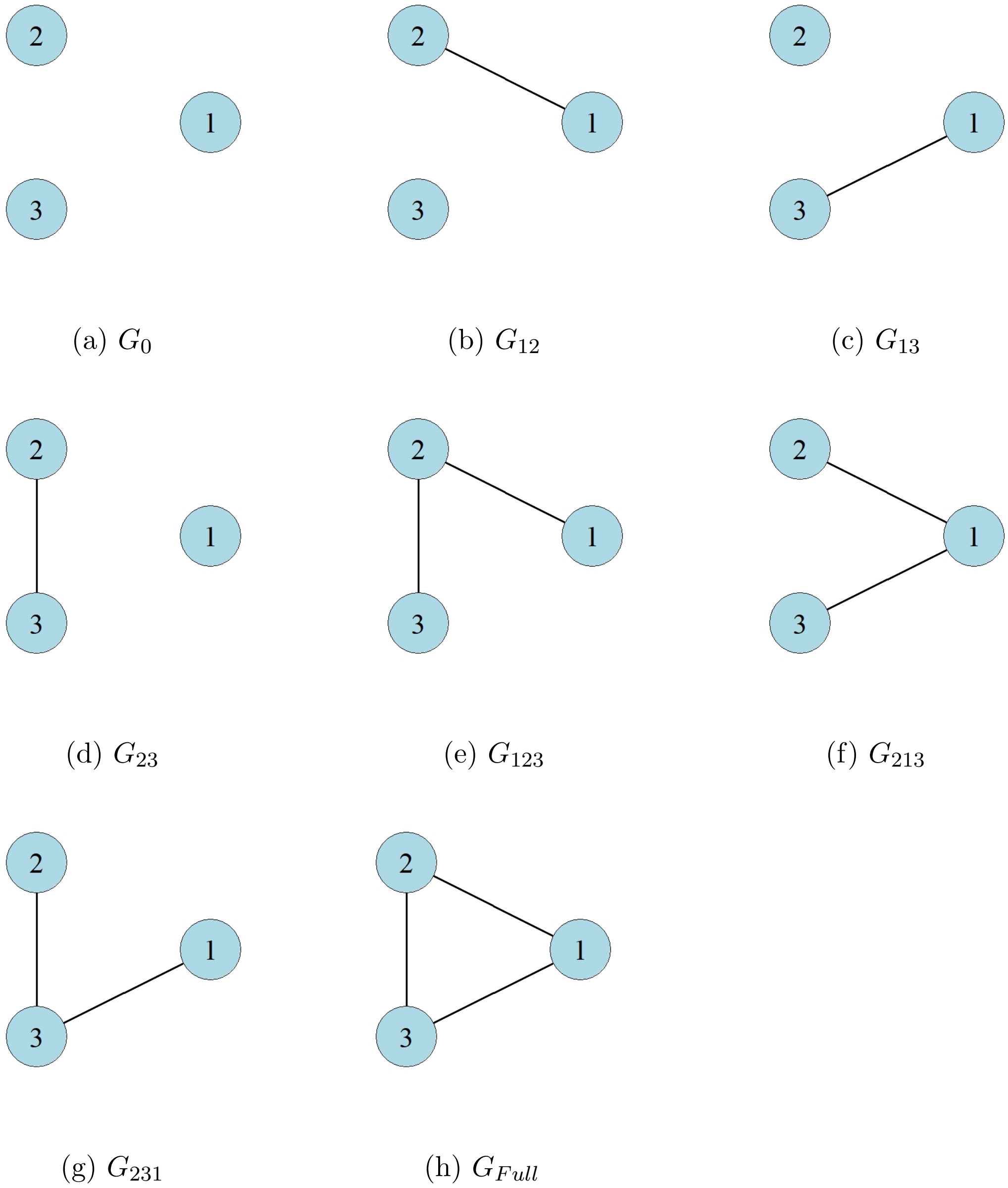
In this section we apply EPP and PEPP to simulated datasets and compare their performance with FBF and BDMCMC. We first provide information about the simulation framework considered and the data generation process. Then, we provide details about the full enumeration and stochastic search simulation framework and data generation process, and finally, we present our results based on diverse simulation scenarios.

5.2.1 Simulation Framework and Data Generation Process

A simulation framework will be characterized by a pair (q, n), where u $q \in \{3, 10, 20, 30\}$ is the number of nodes and $n \in \{100, 300, 500\}$ is the number of observations. For the case of $q = 3$ we will perform a full enumeration approach, since the graphical model space $\mathscr{G}$ contains only 8 models, and for the remaining number of nodes we resort to a stochastic search approach using FINCS algorithm (see subsection 5.1.2).

Full Enumeration Approach

Under the full enumeration approach the graphical model space will be consisted by 8 graphs i.e. $\mathscr{G} = \{G_0, G_{12}, G_{13}, G_{23}, G_{123}, G_{213}, G_{231}, G_{full}\}$. In Figure 5.1 we describe all graphs under consideration. For each simulation pair $\{q, n\}$, we generate 40 datasets having as true model G_{213} (see (f) Figure 5.1).



(a) G_0 (b) G_{12} (c) G_{13}

(d) G_{23} (e) G_{123} (f) G_{213}

(g) G_{231} (h) G_{Full}

Figure 5.1: Model Space of Undirected Graphical Model Space $\mathscr{G}$ under 3 nodes.

The data generation process will be the following:

1. Generate a precision matrix K from the G-Wishart distribution having $b = 10$ degrees of freedom and scale matrix $D = \mathbf{I}_q + G_{213} * 0.6$, using `bdgraph.sim` function of `BDgraph` package.
2. Generate a data matrix $\mathbf{Y}_{n\times q} \sim MN_{n\times q}(\mathbf{0}, \mathbf{I}_n, \mathbf{K}^{-1})$.

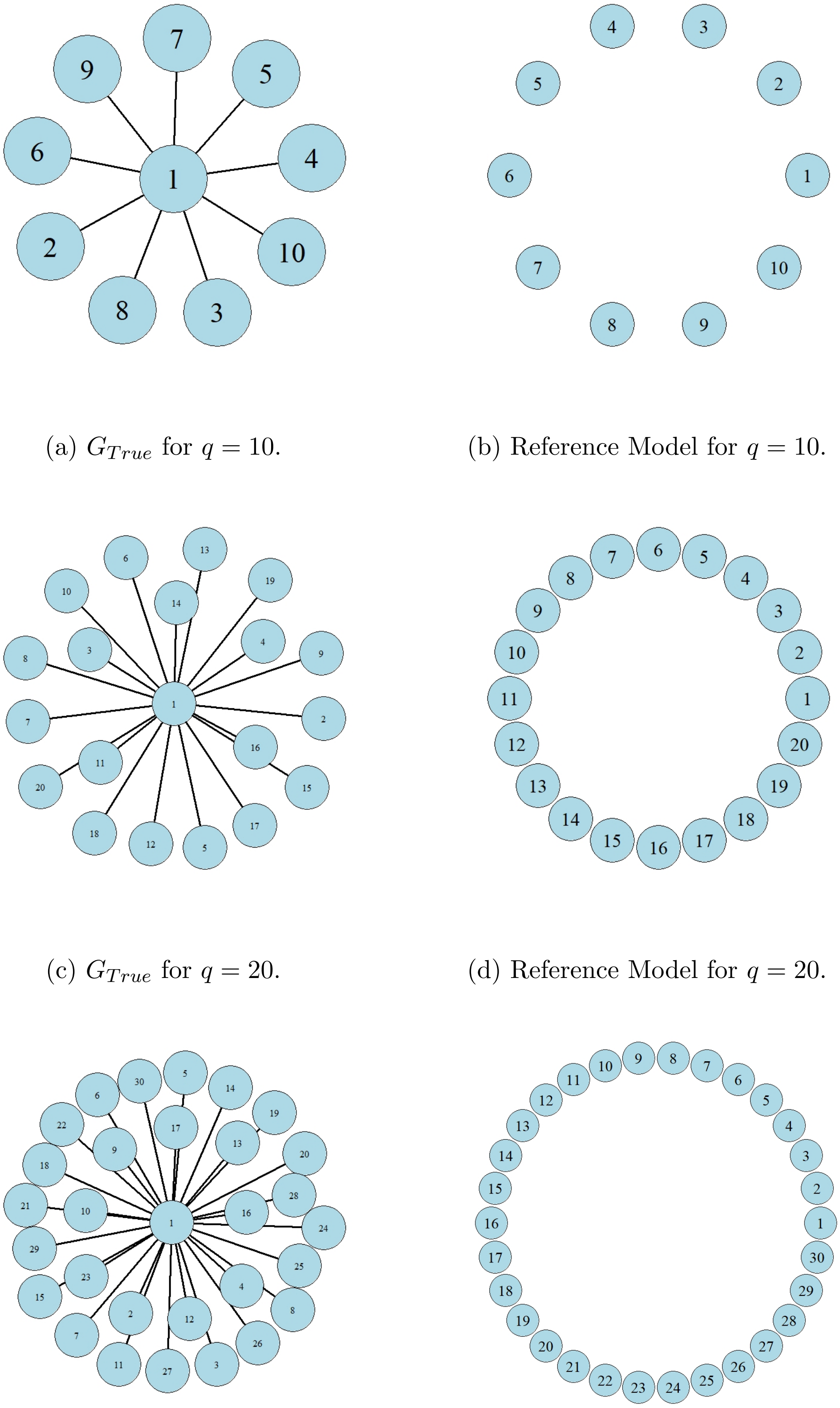
Stochastic Search Approach

For each simulation pair under the stochastic search framework, we consider two different simulation scenarios, namely the *Random Scenario* and the *Star Scenario*. Under the Random Scenario, we generate a total of 40 datasets corresponding to 40 true undirected graphical models, not guaranteed to be decomposable. Following Peters and Buhlmann (2014), for a given q we generate a random undirected graph G_{True}, with probability of edge inclusion $p_{edge} = 3/(2q-2)$. Throughout this section, G_{True} will denote the graphical model that was used to generate the data. The data generation process under the Random Scenario will be the following:

1. Generate an undirected graph G_{True} using `bdgraph.sim` function of `BDgraph` package
2. Generate a precision matrix K from the G-Wishart distribution having $b = 10$ degrees of freedom and scale matrix $D = \mathbf{I}_q$, using `bdgraph.sim` function of `BDgraph` package.
3. Generate a data matrix $\mathbf{Y}_{n\times q} \sim MN_{n\times q}(\mathbf{0}, \mathbf{I}_n, \mathbf{K}^{-1})$.

Under the Star scenario, following Mohammadi and Wit (2015), we generate a total of 40 datasets, where the true graph is constructed by setting all nodes adjacent to the firest node. In Figure 5.2 we provide a graphical representation of G_{True} and the reference model, under the Star graph scenario for the cases of $q \in \{10, 20, 30\}$. The data generation process under the Star Scenario will be the following:

1. Generate a precision matrix $\mathbf{K}$ from the G-Wishart distribution having $b = 10$ degrees of freedom and scale matrix $D = \mathbf{I}_q$, based on the graphical structure which complies to the Star Graph scenario. Matrix $\mathbf{K}$ is generated using `bdgraph.sim` function of `BDgraph` package.
2. Generate a data matrix $\mathbf{Y}_{n\times q} \sim MN_{n\times q}(\mathbf{0}, \mathbf{I}_n, \mathbf{K}^{-1})$.



(a) G_{True} for $q = 10$.

(b) Reference Model for $q = 10$.

(c) G_{True} for $q = 20$.

(d) Reference Model for $q = 20$.

(e) G_{True} for $q = 30$.

(f) Reference Model for $q = 30$.

Figure 5.2: Stochastic Search Approach. G_{True} and Reference model under Star Graph Scenario for $q \in \{10, 20, 30\}$.

For both Random Scenario and Star Scenario, we use as reference model the null graphical model G_0, under the respective number of nodes under considerations, as presented for Star Scenario in Figure 5.2.

5.2.2 FINCS and Computational Setup

To apply EPP and PEPP approaches for $q \in \{10, 20, 30\}$, we will utilize the FINCS algorithm (see subsection 5.1.2) with two minor modifications. First, due to the accumulating processing cost arising from the Bayes factor approximations, we perform the global move of FINCS, by deterministically selecting the median graph (see section 5.1), similarly as Altomare et al. (2013).

Second, we control the number of iterations through pilot runs, in order to keep them as small as possible, since our experimental studies show that our algorithm can reach the optimal model choice quite fast, rendering superfluous runs. To this end, we choose $T = \{3000, 6000, 9000\}$ for $q \in \{10, 20, 30\}$ respectively. For every simulation pair (q, n) of the stochastic search approach, we keep fixed the amount of iterations where a respampling and global moves are performed, following the guidelines provided by Carvalho and Scott (2008). More specificaly, we perform a resampling move every 10 iterations and a global move every 50. Based on several pilot approaches of FINCS algorithm, we chose to generate $R = 20$ importance samples for approximating Bayes factors of any provided $G \in \mathscr{G}$ versus G_0. This choice was based on the computational cost arising from the importance sampling estimation of Bayes factors (see Equation 5.4) and the ability of FINCS to return an optimal solution.

For the BDMCMC approach we follow the choices of Mohammadi and Wit (2015) ,where the total number of iterations $T = 60000$ with a burn-in period of 30000 iterations. The remaining parameters were selected by the baseline choises developed in the `BDgraph` package.The BDMCMC approach was applied using the `bdgraph.sim` function of `BDgraph` package. Note that BDMCMC is a fully Bayesian transdimensional method which performs structural learning in an explicit Bayesian context, rather than using Bayes factors as per our approach and it is applicable to all types of graphical model, compared to our approach which only focuses on undirected decomposable graphical models.

Under every scenario and approach under consideration, we evaluate the performance of EPP, PEPP and benchmark approaches in identifying the graphical structure of G_{True} in terms of ***Structural Hamming Distance*** (SHD), which described the number of insertions and deletions for converting the estimated undirected graph to G_{True}, and F_1-score, provided by

$$F_1 = \frac{2TP}{2TP + FP + FN}, \tag{5.9}$$

where TP, FP and FN denote the number of true positive, false positive and false negative edges identified respectively. The F_1-score lives in $[0, 1]$ and is used as a classification measure of edges identified, with values closer to 1 indicate better classification of edges, whereas values closer to 0 indicate worse classifications. For SHD, values closer to 0 indicate better performances.

5.2.3 Full Enumeration Results

In this subsection we present the results from the application of EPP and PEPP to the full enumeration simulation framework presented earlier in this section, and we compare their performance with the benchmark methods FBF. First, in Figure 5.3 we present the posterior probability of each model of $\mathscr{G}$ (see Figure 5.1).

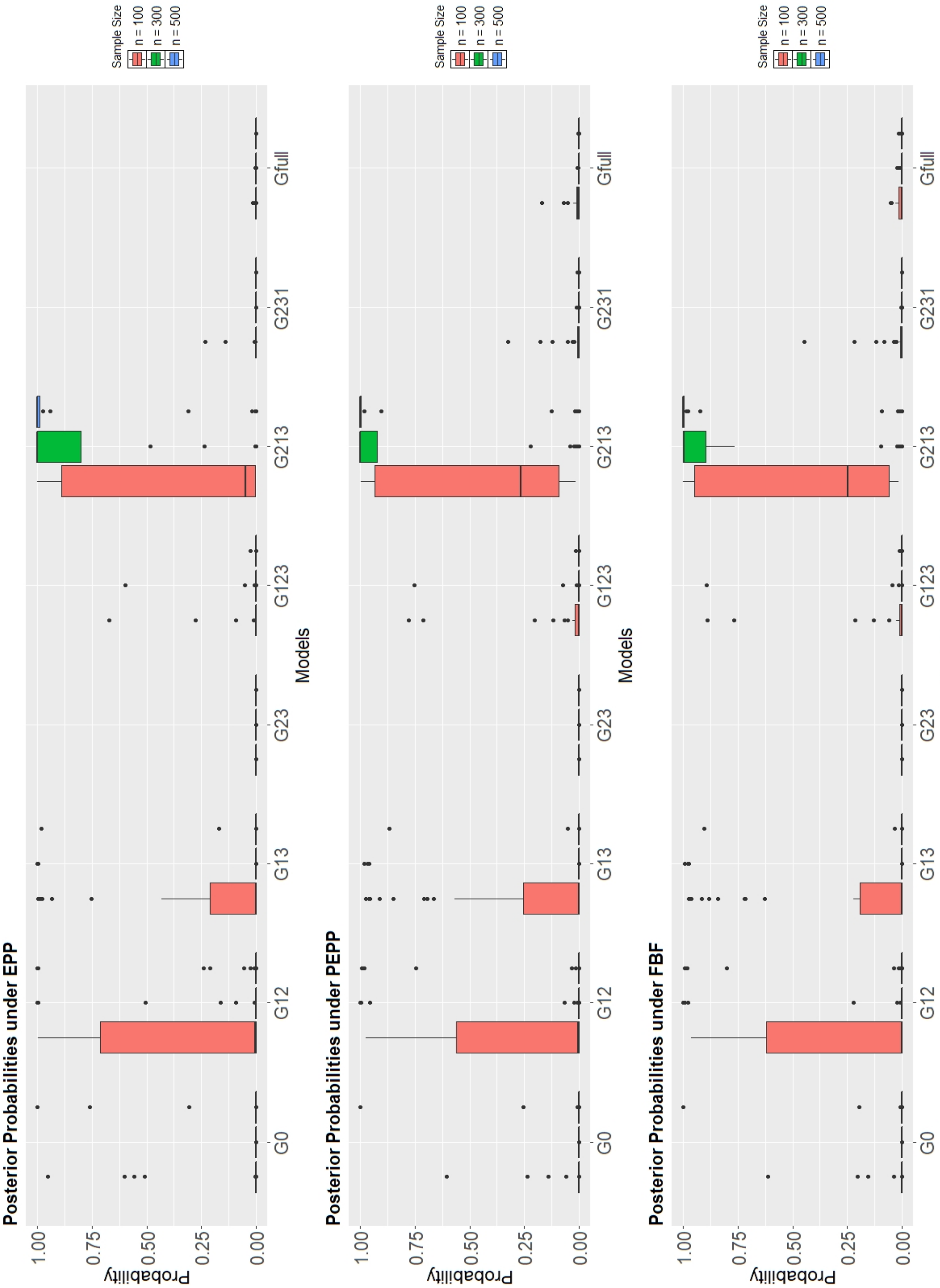


Figure 5.3: Full Enumeration Approach. Model Posterior Probability under EPP, PEPP and FBF approach for $n \in \{100, 300, 500\}$.

We observe that for small sample sizes, i.e. $n = 100$ all three approaches return model posterior probabilities with high variability. More precisely, all three approaches provide posterior probabilities for modes G_{213}, which represents G_{True} and G_{12}, G_{13}, which are nested in G_{213}. As the sample size increases, i.e. $n \in \{300, 500\}$ all three methods under consideration, provide stronger evidence for G_{213} with posterior probabilities over 0.9. One of the remarks in the end of section 3.1, was that model posterior probabilities, decline over similar models and in this simulated example, G_{12} and G_{13} are nested to G_{213}, therefore similar, thus in some cases they obtain higher posterior model probabilities. We also observe that for the case of $n = 300$, EPP provides greater variability compared to PEPP and FBF due to the effect of the imaginary data, compared to PEPP which contains their effect.

Next, in Figure 5.4 we provide boxplots of the posterior edge inclusion probabilities under EPP, PEPP and FBF for each sample size under consideration. Results are in line with the ones provided Figure 5.3.For almost half of the simulated datasets, for small sample sizes, i.e. $n = 100$, we obtain posterior edge inclusion probabilities near 1. More precisely, we obtain these posterior edge inclusion probabilites, yet with high variability, for edges $1 - 2$ and $1 - 3$ which are present in G_{True} and for edge $2 - 3$ which is absent, we obtain posterior edge inclusion probabilities near 0. As the sample size increases, for $n \in \{300, 500\}$, posterior edge inclusion probabilities for edges present in G_{True} are concentrated in 1 and in 0 otherwise.

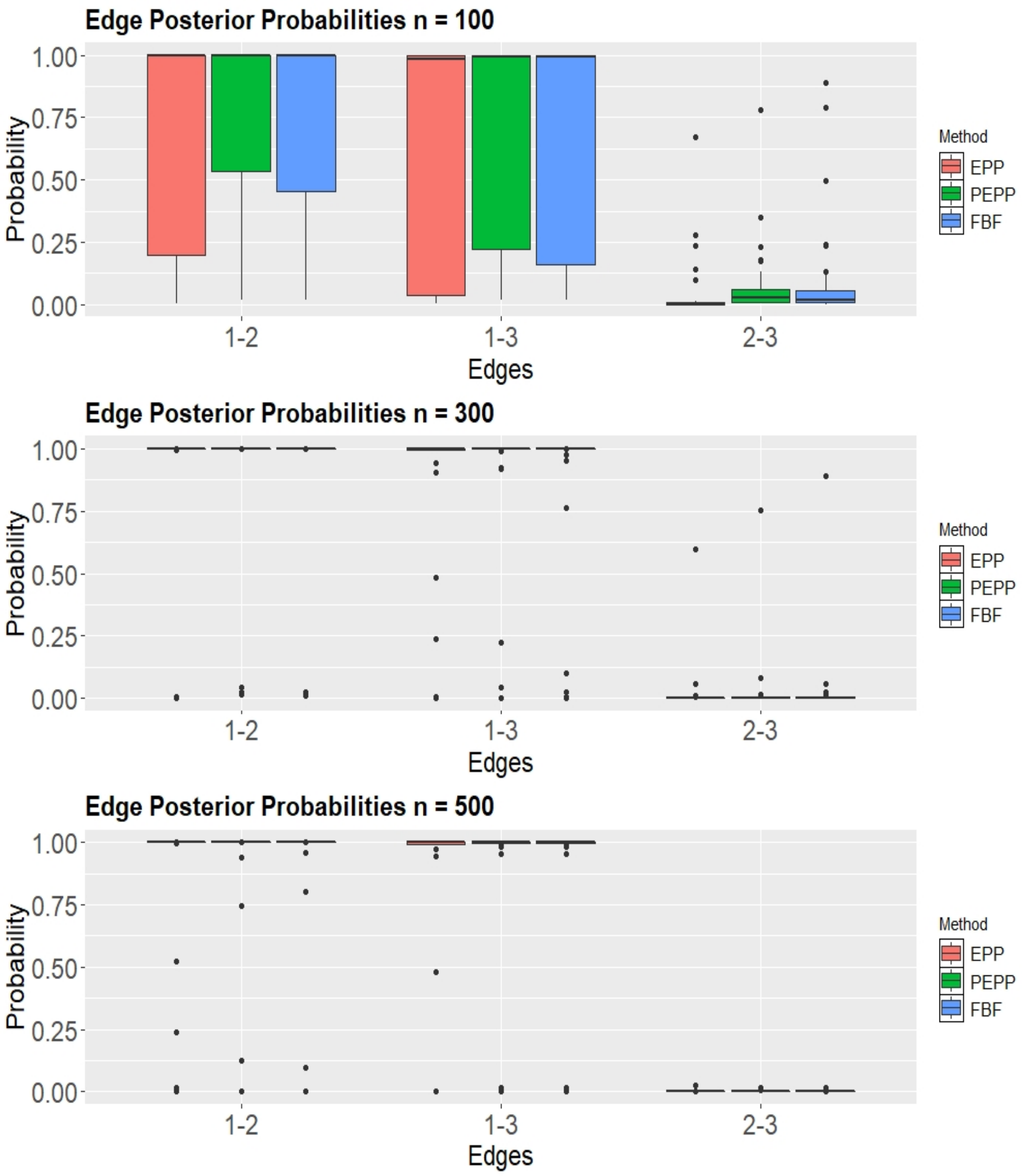


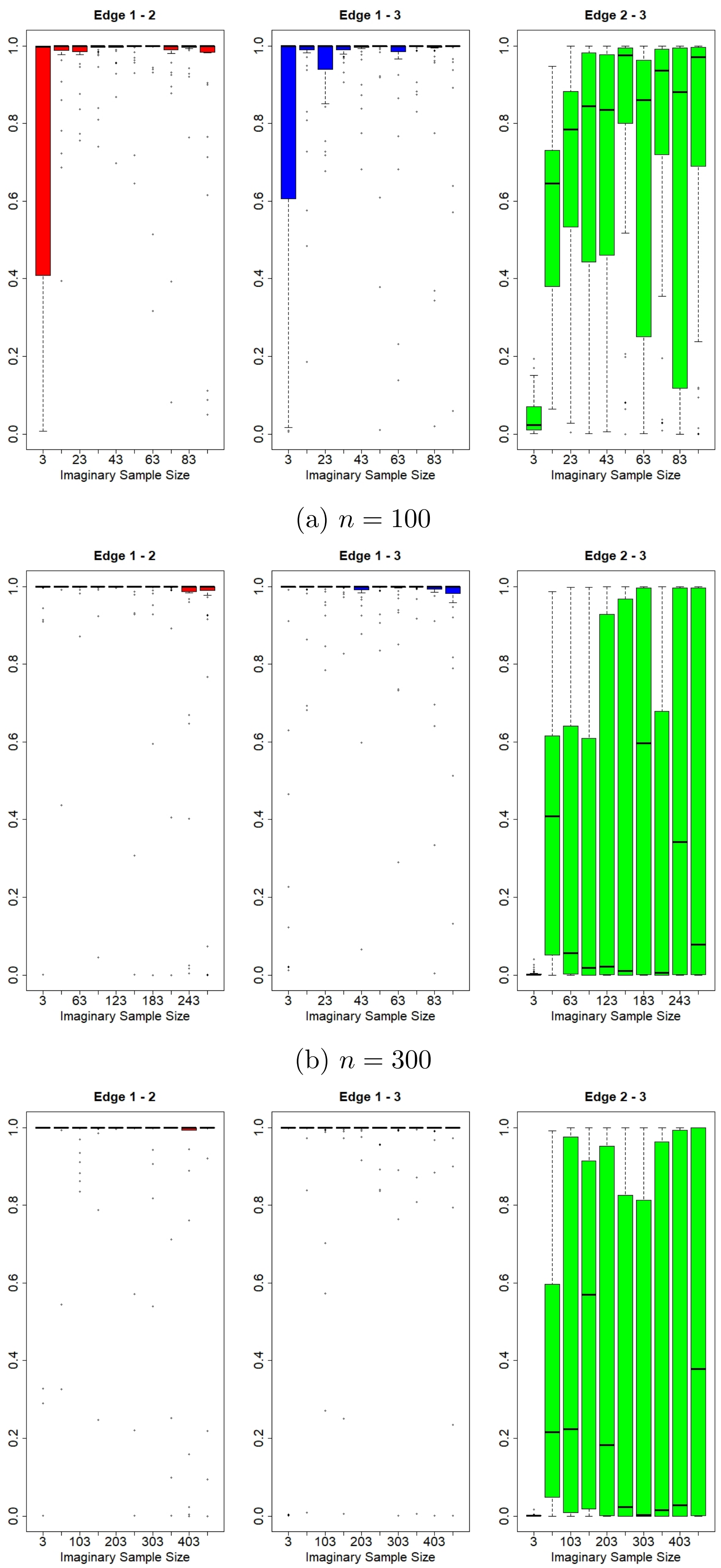
Figure 5.4: Full Enumeration Approach. Edge Inclusionl Posterior Probabilities under EPP, PEPP and FBF approach for $n \in \{100, 300, 500\}$.

Sensitivity Analysis on m under PEPP

Next, in Figure 5.5 we provide a sensitivity analysis for the size of imaginary observations m under the PEPP approach. Since we set $\delta = m$, we are interested in observing the behavior of edge inclusion posterior probabilities as m, and subsequently δ grows. We generate an extra of 40 simulated datasets using the exact same scheme as before, using $R = 20$ importance samples for approximating Bayes factors. We first observe that when we consider small sample sizes, i.e. $n = 100$ and for a small amount of imaginary observations, the posterior edge inclusion probabilities behave as we previously seen in Figure 5.4. As the sample size increases and the number of imaginary observations increases as well, edges that are included in G_{True} behave as expected, i.e. their posterior edge inclusion probabilities are concentrated in 1.

For edges that are not included in G_{True}, specificaly for edge $2 - 3$, we observe an abnormal behavior as the size of imaginary observarions m increases. When m is kept at a minimum, i.e. $m = 3$, the behavior of the posterior edge inclusion probabilities is as expected. As the imaginary sample size increases, we observe a huge amount of variability without a clear indication that the edge inclusion posterior probabilities are trending towards 0 or 1. Furthermore, we performed the exact same experiment using more exhaustive approximations, where we considered $R = 1000$ importance samples for approximating the Bayes factors of models of $\mathscr{G}$ versus G_0 and in Figure 5.6 we provide boxplots of edge inclusion posterior probabilities and we observe the same behavior as in Figure 5.5.

Following the findings of this sensitivity analysis, we opt to keep the size of imaginary observations m at a minimum for PEPP as well, for ensuring the stability of our approach and controlling the computational cost as well. Further more, the amount of importance samples considered, i.e. $R = 20$ does not seem to affect the analysis applied by PEPP, therefore we will keep it at a minimum for controlling the computational cost of the stochastic search approach as well.

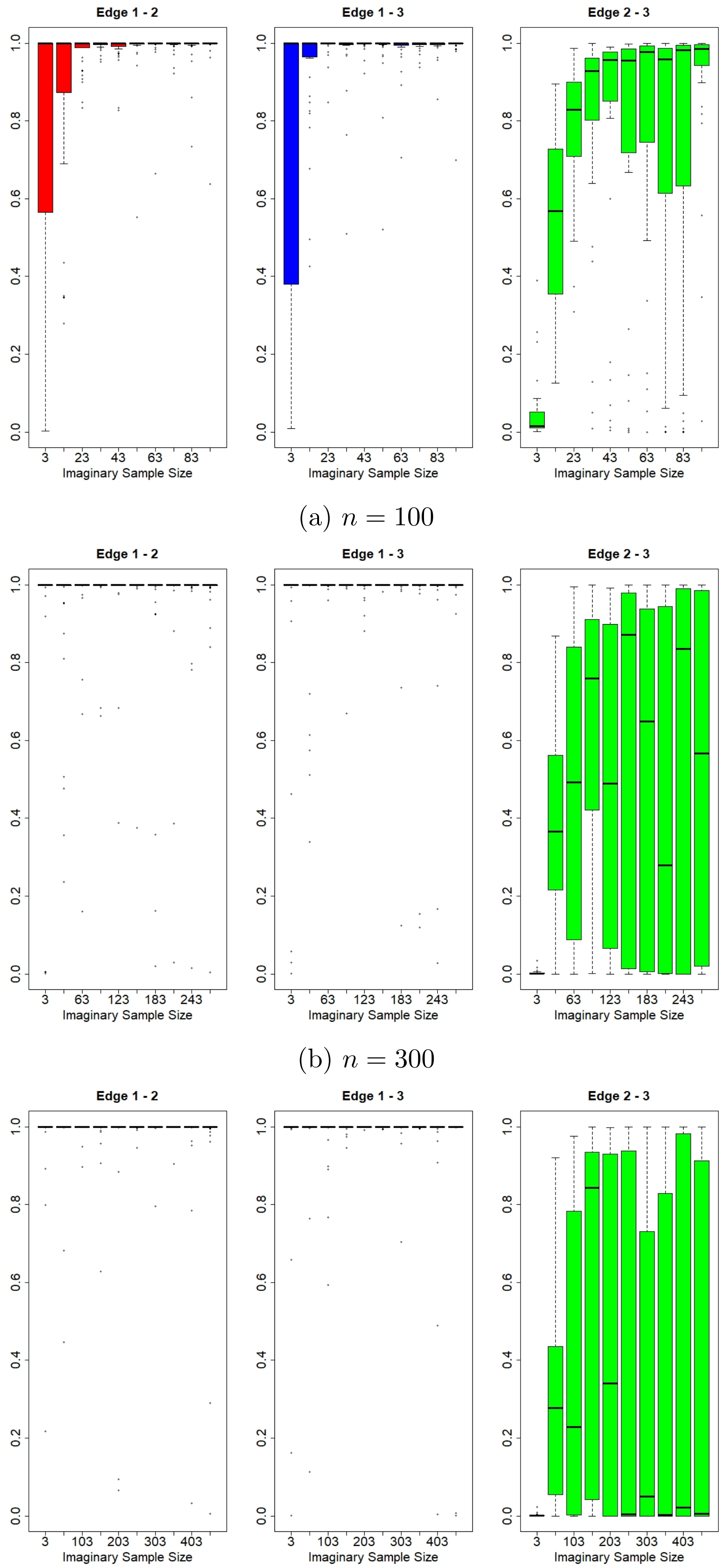


(a) $n = 100$

(b) $n = 300$

(c) $n = 500$

Figure 5.5: Full Enumeration Approach. Sensitivity Analysis for PEPP for varying imaginary sample size m and for $n \in \{100, 300, 500\}$ and $R = 20$. Edge Inclusion Posterior Probabilities over 40 simulated datasets with real graph G_{213}.



(a) $n = 100$

(b) $n = 300$

(c) $n = 500$

Figure 5.6: Full Enumeration Approach. Sensitivity Analysis for PEPP for varying imaginary sample size m and for $n \in \{100, 300, 500\}$ and $R = 1000$. Edge Inclusion Posterior Probabilities over 40 simulated datasets with real graph G_{213}.

5.2.4 Stochastic Search Approach

In this subsection we present the results from the application of EPP and PEPP to the simulation scenarios presented earlier in this section, and we compare their performance with the benchmark methods FBF and BDMCMC. To evaluate the performance of each method under consideration, we measure the SHD between the median probability graphical model and the respective G_{True}. In Figure 5.7 we report the boxplots of the SHDs over the 40 simulated datasets for every simulation pair (q, n) considered by the Random Scenario.

We observe that the performances of every method under consideration improves as the sample size n increases, and deteriorate as the number of nodes q increases. An interesting finding is that PEPP distances are in all cases smaller than those provided by EPP, which signals that the compression feature of imaginary observations that PEPP provides, returns better estimations in terms of SHD.

Compared to the other benchmark approaches, PEPP reveals a performance comparable to FBF, especially for sample sizes without suffering from double usage of data as in FBF. Therefore, we believe that PEPP is a valid Bayesian alternative to FBF approach, since it can provide similar performance using a more theoretically sound procedure.

Finally, we observe that BDMCMC method performs worse, but it tends to improve as the number of nodes increases. We need to note that BDMCMC performs structure learning in an explicit Bayesian context, rather that using Bayes factors as per our approach and it is applicable to all kinds of graphical modes, whereas we are restricted to decomposable graphical models. Furthermore, the output of BDMCMC is far richer MCMC output, and therefore it requires higher computational time and higher sample sizes. In Figure 5.8 we report the boxplots of the SHDs over the 40 simulated datasets for every simulation pair (q, n) considered by the Star scenario, with findings similar to the Random scenario provided before.

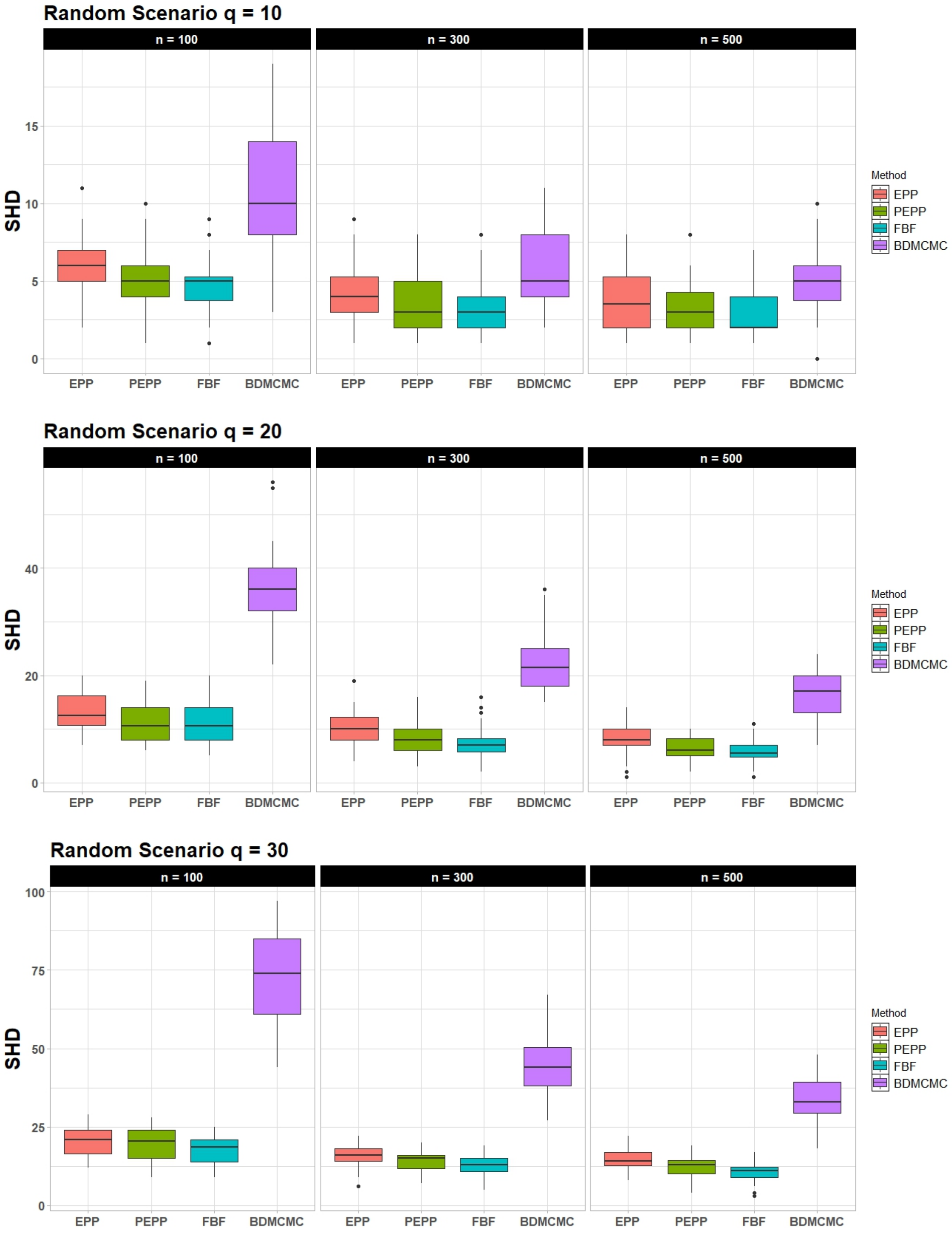


Figure 5.7: Simulation study under Random Setup. Structural Hamming distances between the estimated Undirected graphs and true Undirected graphs, over 50 datasets for number of nodes $q = \{10, 20, 30\}$ and sample size $n = \{100, 300, 500\}$. The performances are measured for our intermediate output, the median probability model under EPP and PEPP, the median probability model under FBF and BDMCMC and the final output structure under MB approach.

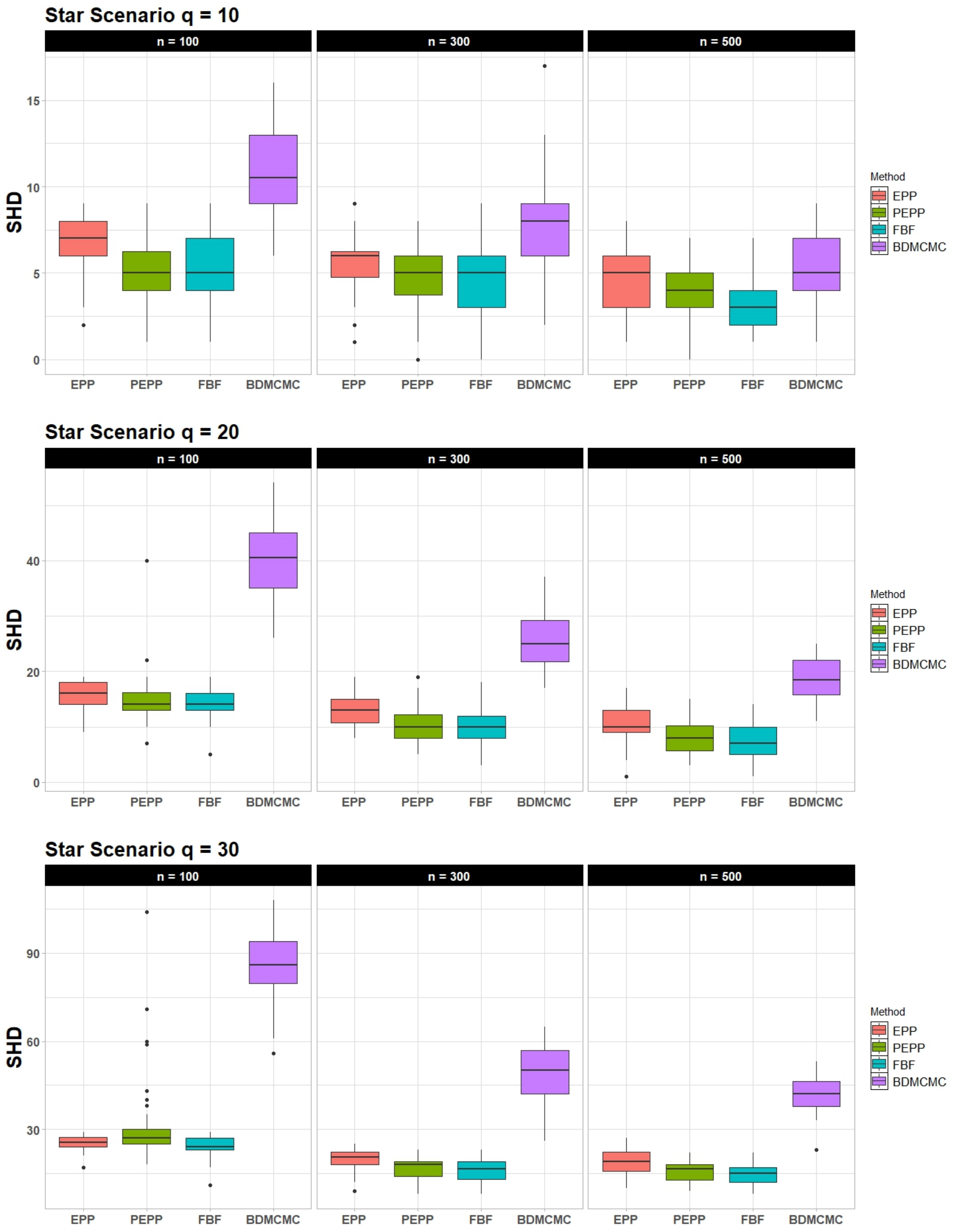


Figure 5.8: Simulation study under AR-1 Setup. Structural Hamming distances between the estimated Undirected graphs and true Undirected graphs, over 50 datasets for number of nodes $q = \{10, 20, 30\}$ and sample size $n = \{100, 300, 500\}$. The performances are measured for our intermediate output, the median probability model under EPP and PEPP, the median probability model under FBF and BDMCMC and the final output structure under MB approach.

In Table 5.1 and Table 5.2 we present the mean and variance of F1 scores under Random and Star graph simulaton scenario respectivelly, for each method under consideration. We observe that for small sample sizes, i.e. $n = 100$, PEPP approach provides better F_1-score than all other approaches under consideration. As the sample size increases, i.e. $n = \{300, 500\}$, FBF approach is the best performer and the PEPP approach closely follows.

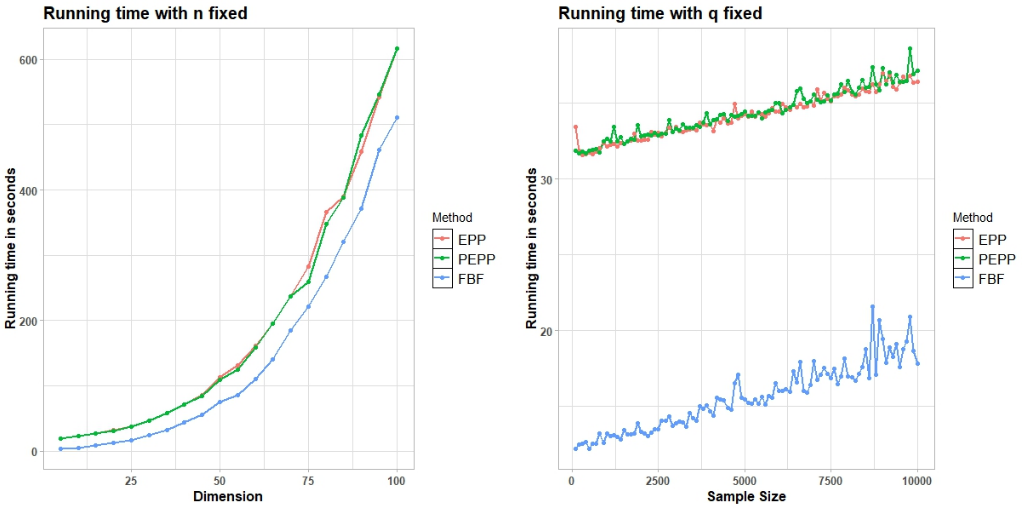
Case Studies			Approaches			
Scenario	q	n	EPP	PEPP	FBF	BDMCMC
Random	10	100	0.31 (0.05)	0.47 (0.06)	**0.51 (0.06)**	0.50 (0.02)
		300	0.54 (0.05)	0.65 (0.03)	**0.71 (0.03)**	0.66 (0.02)
		500	0.59 (0.07)	0.68 (0.05)	**0.74 (0.04)**	0.70 (0.02)
	20	100	0.25 (0.02)	**0.48 (0.02)**	0.41 (0.03)	0.36 (0.01)
		300	0.47 (0.02)	0.62 (0.01)	**0.66 (0.02)**	0.50 (0.01)
		500	0.59 (0.02)	0.71 (0.01)	**0.76 (0.01)**	0.59 (0.01)
	30	100	0.23 (0.01)	**0.42 (0.01)**	0.37 (0.02)	0.31 (0.01)
		300	0.44 (0.02)	0.55 (0.01)	**0.59 (0.01)**	0.44 (0.01)
		500	0.51 (0.01)	0.60 (0.01)	**0.68 (0.01)**	0.50 (0.01)

Table 5.1: Simulated data. Means of F_1-score (variances in parentheses) under the Random Scenario.

Case Studies			Approaches			
Scenario	q	n	EPP	PEPP	FBF	BDMCMC
Star	10	100	0.33 (0.06)	**0.57 (0.05)**	**0.57 (0.05)**	0.54 (0.01)
		300	0.55 (0.05)	0.64 (0.03)	**0.66 (0.03)**	0.63 (0.01)
		500	0.62 (0.04)	0.71 (0.01)	**0.78 (0.01)**	0.73 (0.01)
	20	100	0.25 (0.03)	**0.47 (0.02)**	0.41 (0.03)	0.37 (0.00)
		300	0.45 (0.04)	0.60 (0.04)	**0.63 (0.04)**	0.51 (0.00)
		500	0.61 (0.03)	0.72 (0.02)	**0.75 (0.02)**	0.61 (0.01)
	30	100	0.22 (0.03)	**0.38 (0.02)**	0.29 (0.04)	0.29 (0.00)
		300	0.48 (0.02)	0.62 (0.01)	0.64 (0.01)	0.45 (0.00)
		500	0.52 (0.01)	0.64 (0.01)	**0.68 (0.01)**	0.50 (0.00)

Table 5.2: Simulated data. Means of F_1-score (variances in parentheses) under the Star Scenario.

We next investigate the computational time required for performing FINCS algorithm under EPP, PEPP and FBF, as function of sample size n and number of nodes q. In the left panel of Figure 5.9 we report the time in seconds needed for FINCS algorithm to perform 500 iterations for $n = 500$, $R = 20$ generated importance samples for estimating Bayes factors under EPP and PEPP and number of nodes q varying betwen 5 and 100 nodes, whilst in the right panel of Figure 5.9 we report the time in seconds needed for FINCS to perform 500 iterations for $q = 20$, $R = 20$ generated importance samples and sample size n varying between 500 and 10000. Algorithms were run on a `Toshiba Satellite 12GB RAM 2.6GHZ + 2.6GHS 8-processor unit (4 processors + 4 mirrors)`. We observe that with an increasing number of nodes q, the computational time increases exponentially, whilst with an increasing number of sample size n the computational time grows in a much lower rate.



(a) Run-time as a function of q.

(b) Run-time as a function of n.

Figure 5.9: Simulated data. Computational time (in seconds) of 500 iterations of EPP, PEPP and FBF, as a function of q for $n = 500$ (left panel) and as a function of the sample size n for a fixed number of nodes $q = 20$ (right panel).

5.3 Real Data Application

In this section, we apply both EPP and PEPP approaches to the protein signalling data set provided by Sachs et al. (2005). In their original work, Sachs et al. (2005) aimed to infer a signle DAG using these data, where Friedman et al. (2008) used them to infer an undirected graph and Castelletti et al. (2018) used the same data for performing structure learning of interventional essential graphs. Recently, Peterson et al. (2015) analyzed the data for inferring multiple graphs under each experimental condition, allowing for the possibility of shared structural features among the estimated graphs. We share the same view with Peterson et al. (2015) and our goal will be to infer an undirected graph for each experimental condition and identify the most common edges amongst all resulted graphs.

The data provided by Sachs et al. (2005), are based on simultaneous measurements of multiple phosphorylated proteins and phospholipid componenets contained in individual primary human immune system cells. Our interest lies in $q = 11$ phosphorylated proteins and phospholipids, which are observed after applying nine different experimental conditions, resulting to nine different datasets containing observations that share the same experimental characteristics.

We assume that the joint distribution of the data is a multivariate normal distribution as in previous works which analyzed the same data. In Figure 5.10 we provide the estimated graphs under each experimental condition from Peterson et al, where indices $(i) - (ix)$

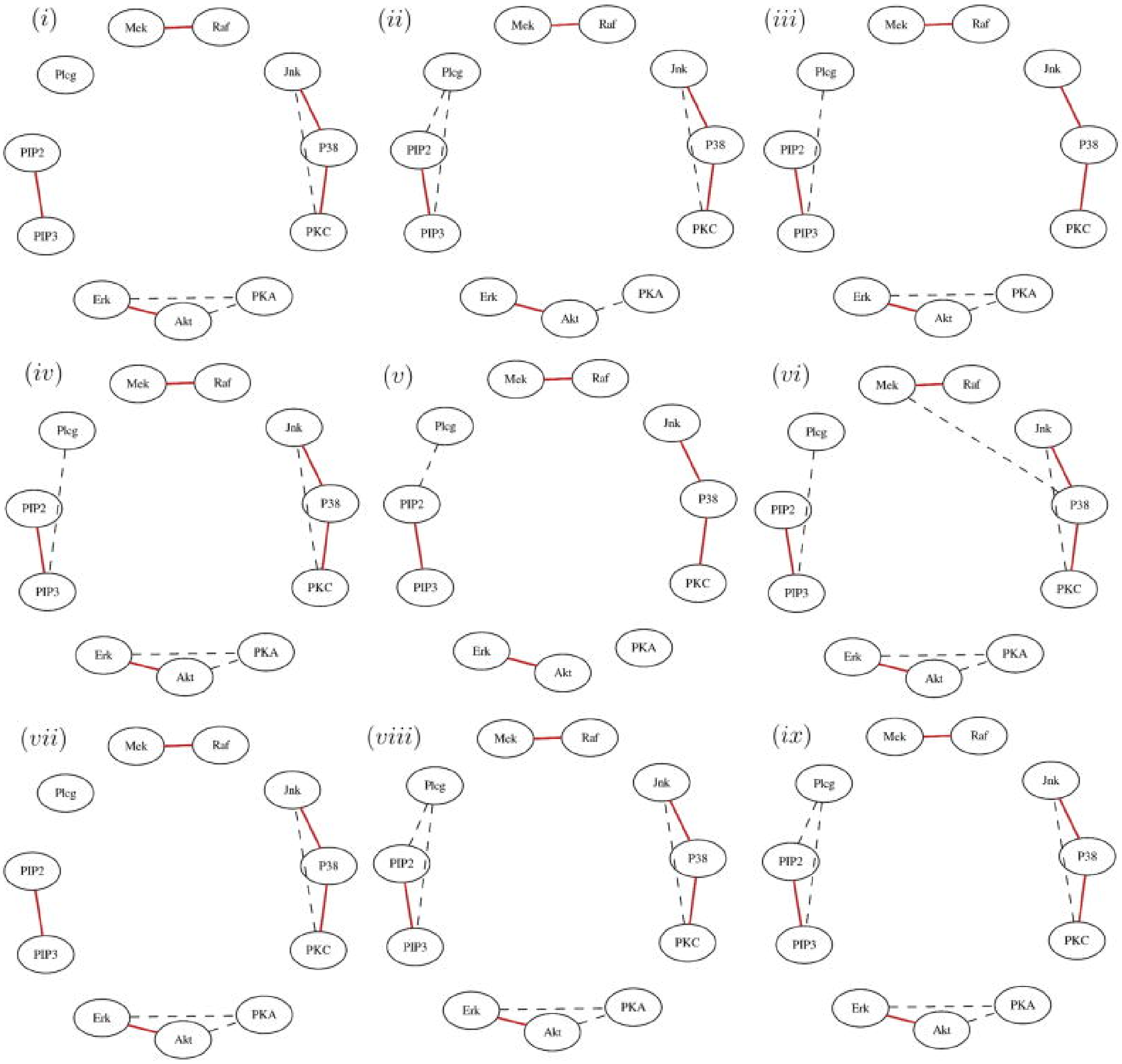
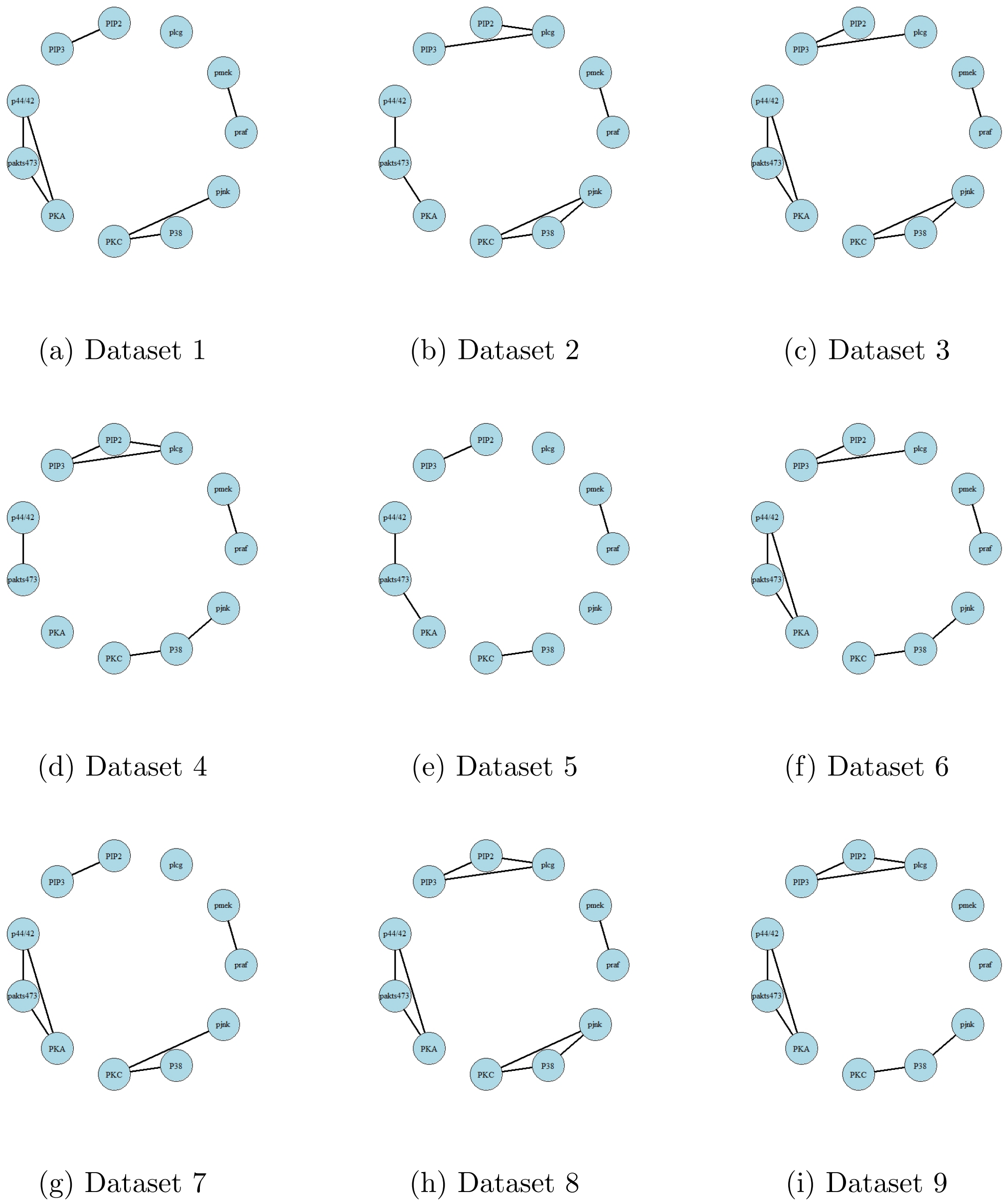


Figure 5.10: Estimated undirected graphs per dataset under Peterson et al. (2015).

We apply EPP and PEPP approaches on each dataset provided using the setup established in the simulation studies for $q = 10$, where we consider $T = 3000$ iterations of FINCS algorithm, applying resampling moves every 10 iterations and global moves every 50 iterations, and returning the median probability graph (see subsection 5.2.2). In Figure 5.11 and Figure 5.12 we report the inferred undirected graphs under each experimental conditions under EPP and PEPP respectivelly. We perform sixteen rounds of FINCS algorithm and we then consider the average of the posterior edge inclusion probabilities among these 16 runs.



(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

(d) Dataset 4

(e) Dataset 5

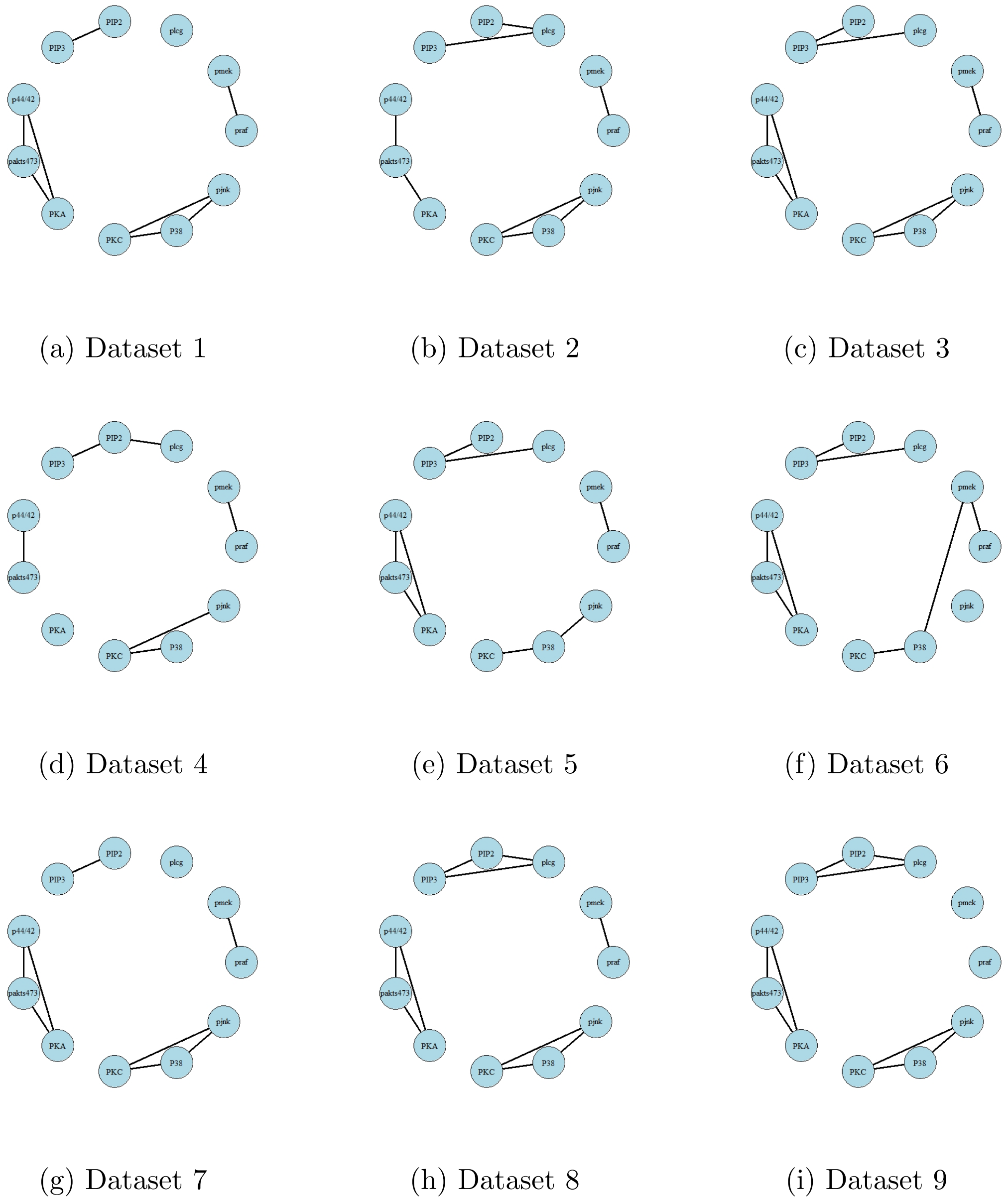
(f) Dataset 6

(g) Dataset 7

(h) Dataset 8

(i) Dataset 9

Figure 5.11: Protein Signalling data. Estimated median probability graphs using EPP approach.



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

(d) Dataset 4 (e) Dataset 5 (f) Dataset 6

(g) Dataset 7 (h) Dataset 8 (i) Dataset 9

Figure 5.12: Protein Signalling data. Estimated median probability graphs using PEPP approach.

Furthermore, in Table 5.3 and Table 5.4, we report the amount of edge appearances in the nine datasets for EPP and PEPP respectivelly. The maximum amount an edge can appear is nine times. We consider an edge as significant if it appears in 8 or 9 datasets. We observe that PEPP approach returns the same significant edges as per Peterson et al. (2015), that is the edges *praf - pmek*, *PIP2 - PIP3*, *pakts473 - p44/42*, *pakts473 - PKA* and *P38 - PKC*. EPP approach returns one edge less than PEPP approach, namely the connection between *pakts473 -PKA*, yet it occurs more frequently than the other non-significant edges.

Variables	praf	pmek	plcg	PIP2	PIP3	p44/42	pakts473	PKA	PKC	P38	pjnk
praf	0	**8**	0	0	0	0	0	0	0	0	0
pmek	**8**	0	0	0	0	0	0	0	0	0	0
plcg	0	0	0	4	6	0	0	0	0	0	0
PIP2	0	0	4	0	**8**	0	0	0	0	0	0
PIP3	0	0	6	**8**	0	0	0	0	0	0	0
p44/42	0	0	0	0	0	0	**9**	6	0	0	0
pakts473	0	0	0	0	0	**9**	0	**8**	0	0	0
PKA	0	0	0	0	0	6	**8**	0	0	0	0
PKC	0	0	0	0	0	0	0	0	0	**9**	7
P38	0	0	0	0	0	0	0	0	**9**	0	7
pjnk	0	0	0	0	0	0	0	0	7	7	0

Table 5.3: Edge apperances per dataset under PEPP approach.

Variables	praf	pmek	plcg	PIP2	PIP3	p44/42	pakts473	PKA	PKC	P38	pjnk
praf	0	**8**	0	0	0	0	0	0	0	0	0
pmek	**8**	0	0	0	0	0	0	0	0	0	0
plcg	0	0	0	4	4	0	0	0	0	0	0
PIP2	0	0	4	0	**8**	0	0	0	0	0	0
PIP3	0	0	4	**8**	0	0	0	0	0	0	0
p44/42	0	0	0	0	0	0	**9**	6	0	0	0
pakts473	0	0	0	0	0	**9**	0	7	0	0	0
PKA	0	0	0	0	0	6	7	0	0	0	0
PKC	0	0	0	0	0	0	0	0	0	**9**	3
P38	0	0	0	0	0	0	0	0	**9**	0	6
pjnk	0	0	0	0	0	0	0	0	3	6	0

Table 5.4: Edge apperances per dataset under EPP approach.

In Table 5.5, we report the SHD of the infered graphs between PEPP and EPP, FBF and BDMCMC, using PEPP as the benchmark method. We observe that besides dataset 5, all approaches return similar inferred graphs and more specifically they return identical graphs for datasets 3 and 8.

Dataset	EPP	FBF	BDMCMC	# PEPP	# EPP	# FBF	# BDMCMC
1	2	1	1	8	6	9	9
2	0	2	3	8	8	10	11
3	0	0	0	9	9	9	9
4	2	1	1	6	6	7	7
5	1	5	5	6	5	11	11
6	1	1	1	8	7	9	9
7	2	1	1	8	6	9	9
8	0	0	0	10	10	10	10
9	2	1	1	9	7	10	10

Table 5.5: Protein Signaling data. Structural Hamming distances between the estimated undirected graph under Power-Expected Posterior prior versus every alternative method and total number of edges under each approach.

Following the findings of Table 5.5, in Table 5.6 we provide the estimated posterior probability of MAP model under PEPP, EPP and FBF for datasets 3 and 8, where all three methods resulted to the same graph. We observe that PEPP and EPP associate a higher estimated posterior probability to the estimated MAP model, i.e. the model with the highest posterior probability, compared to the FBF approach. Thus, we deduce that PEPP and EPP can more easily distinct the optimal model relative to FBF. Results of BDMCMC are ommited since the output of the `bdgraphsim` function returns only estimations of posterior edge inclusion probabilities, rather than model estimated posterior probabilities.

Dataset	EPP	PEPP	FBF
3	0.96	0.97	0.3
8	0.96	0.96	0.3

Table 5.6: Protein Signalling data. Estimated Posterior Probability of the MAP model for EPP, PEPP and FBF, for datasets 3 and 8.

Chapter 6

Conclusions and further research

Graphical models are used for depicting conditional independence relationships among a given set of variables. In real-life situations, we do not know the structure of the underlying graph and we use the data at hand to infer the graph's structure. Objective Bayes approaches are well-suited for the structure learning problem of undirected decomposable gaussian graphical models, since it is difficult to subjectively specify a prior over vast parametric spaces, constrained by graphical structure.

In this thesis we introduced the Expected and Power Expected Posterior prior approaches to the structure learning problem of undirected Gaussian graphical models, using a specific class of models, that is the decomposable models. Furthermore, we applied a modified version of FINCS algorithm of Carvalho and Scott (2008) for exploring a given graphical model space $\mathscr{G}$ of decomposable graphical models on q nodes. We applied the proposed methodologies, namely EPP and PEPP, to artificially simulated datasets and in protein-signaling data (Sachs et al. (2005)) and compared their performance with the FBF approach (Carvalho and Scott (2009)) and BDMCMC approach (Mohammadi and Wit (2015)).

The output of both EPP and PEPP, returns posterior estimates of edge inclusion probabilities and in small settings in can return graphical model posterior probabilities. In the simulated scenarios illustrated in chapter 5 we observed that PEPP approach is highly competitive with the FBF approach and outperforms both EPP and BDMCMC approaches. More specifically, for the 3-node example, EPP and PEPP perform the same as FBF approach and PEPP returns slightly higher posterior edge inclusion probabilities.

When we scale-up to higher dimensions, i.e. for $q \in \{10, 20, 30\}$, in both simulation scenarios considered, PEPP approach performs similarly to the FBF approach and again outperforms EPP and BDMCMC in terms of SHD. When it comes to classification ability, PEPP performs better than the other three approaches when considering small sample sizes. As the sample size increases, PEPP follows-up closely FBF approach and outperforms EPP and BDMCMC.

The protein-signaling data provided by Sachs et al. (2005) were observed under nine different experimand conditions. They have been analyzed as one common dataset (Friedman et al. (2008)) or seperately as per Peterson et al. (2015) for identifying shared features among graphs. We shared the view of Peterson et al. (2015) and inferred a singe graph per dataset, for identifying common edges among the nine datasets. Results showed that PEPP returns the exact same common edges as per Peterson et al. (2015) and EPP returns one edge less. Furthermore, we observed that under a stochastic search approach, both EPP and PEPP support the optimal identified model with greater evidence than FBF, resulting to shorter algorithmic runs and subsequently lower computational cost.

Key difference between EPP-PEPP and the Fractional Bayes Factor approach proposed in the literature is that both are based on the use of imaginary observations, thus avoiding double use of available data. In their core, they utilize improper parameter prior distributions when it is difficult to successfully elicit a subjective one, alleviating the indeterminacy in Bayes factors arising from the existence of arbitrary normalizing constants. For both EPP and PEPP, the base-model approach is considered since it can efficiently adapt to the field of graphical model selection. The advantage of PEPP over EPP is that the former reduces the effect of imaginary data , leading to more accurate estimation, compared to EPP.

EPP and PEPP are restrictive in terms of applications since they both operate in examples where we have $n > q$. In their current state, both are not feasible for higher dimensions, since the computational cost required for their implementation is grows exponentially as the number of variables increase. We further investigate a full transition of C++ routines for obtaining cliques and separators.

In all simulation scenarios under considerations, EPP and PEPP provided high variability in the results, when we studies small datasets i.e. $n = 100$. More specifically, in the full enumeration approach we observed huge variabilities in posterior model probabilites and posterior edge inclusion probabilities under EPP, PEPP. For PEPP, and eventually EPP, increasing the number of importance samples for the estimation of Bayes factors, did not result to reducing the variability (Figure 5.5). Regarding the variability of FBF approach, one can alternatively consider more restrictive fraction parameter. In our study we consider $g = q/n$, whereas we a more conservative choice is $g = 1/n$. Robustifying both EPP and PEPP approach in the graphical model selection context is currently under investigation.

Another open issue is the instability of parameter δ under PEPP approach, as presented in section 5.2.3. Fouskakis et al. (2015) provide evidence that the choice of δ does not influence the posterior analysis in the variable selection context, where PEPP was originally created for. Yet, in the graphical model selection context thats not the case and as we described in our sensitivity analysis, with greater size of imaginary observations, we obtain greater instability on posterior edge inclusion probabilities of non-significant edges. We are currently investigating alternative expressions of parameter δ for robustifying the posterior analysis.

After resolving the open issues stated above, following Consonni et al. (2017), we are interested in futher expanding EPP and PEPP to the covariate-adjusted graphical model selection framework of undirected decomposable graphical models.

Bibliography

Altomare, D., Consonni, G., and La Rocca, L. (2013). Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. *Biometrics*, 69:478–487.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32:870–897.

Bayarri, M. J., Berger, J. O., Forte, A., and Garcia-Donato, G. (2012). Criteria for bayesian model choice with application to variable selection. *The Annals of Statistics*, 40:1550–1577.

Berger, J. (2004). The case for objective bayesian analysis. *Bayesian Analysis*, 1(1):1–17.

Berger, J. and Pericchi, L. (1998a). Accurate and stable bayesian model selection: the median intrinsic bayes factor. *Sankhya*, 60:1–18.

Berger, J. and Pericchi, L. (1998b). Accurate and stable bayesian model selection: The median intrinsic bayes factor. *Sankya*, 60(B):1–18.

Berger, J. O., De Oliveira, V., and Sanso, B. (2001). Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96:1361–1374.

Berger, J. O. and Molina, G. (2004). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica*, 59(1):3–15.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91:109–122.

Berger, J. O. and Pericchi, L. R. (2001). Objective bayesian methods for model selection: Introduction and comparison. *IMS Lecture Notes - Monograph Series*, 38.

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

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

Carvalho, C. and Scott, J. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96:497–512.

Castelletti, F., Consonni, G., Marco, D. V., and Peluso, S. (2018). Learning Markov equivalence classes of directed acyclic graphs: An objective Bayes approach. *Bayesian Analysis*, 13:1235–1260.

Consonni, G., Fouskakis, D., Ntzoufras, I., and Liseo, B. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13:627–679.

Consonni, G., La Rocca, L., and Peluso, S. (2017). Objective Bayes covariate-adjusted sparse graphical model selection. *Scandinavian Journal of Statistics*, 44:741–764.

Consonni, G. and Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, 23:332–353.

Datta, G. S. and Mukerjee, R. (2004). Probability matching priors: Higher order asymptotics. *Lecture Notes in Statistics. Sringer, New York.*

Dawid, A., M., S., and J., Z. (1973). Marginalization paradoxes in bayesian and structural inference (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 35:189–233.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15.

Dawid, A. P. (1980). Conditional independence for statistical operations. *The Annals of Statistics*, 8(3):598–617.

Dawid, A. P. (2006). Invariant prior distributions. *Encyclopedia of Statistical Sciences, Wiley, New York.*

Dawid, P. and Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 3:1272–1317.

Dempster, A. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Dmochowski, J. (1994). Intrinsic priors via kullback-leibler geometry. *Technical Report Purdue University.*

Dobra, A., Jones, B., Hans, C., Nevins, J., and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Mult. Anal*, 90:126–212.

Fouskakis, D., Ntzoufras, I., and Draper, D. (2015). Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Analysis*, 10:75–107.

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

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

Good, I. (1950). *Probability and the Weighting of Evidence.* London, UK: Charles Griffin.

Gupta, A. K. and Nagar, D. K. (2000). *Matrix variate distributions.* Chapman & Hall/CRC.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *J. Statist. Sci*, 1:46–60.

Jeffreys, H. (1961). *The Theory of Probability.* Oxford University Press.

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

Kass, R. E. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90:928–934.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91:1343–1369.

Lauritzen, S. L. (1996). *Graphical models.* Oxford University Press.

Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Mohammadi, R. and Wit, E. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10:109–138.

Moreno, E., Bertolino, F., and Racugno, W. (2014). An intrinsic limiting procedure for model selection and hypotheses testing. *Journal of the American Statistical Association*, 93:1451–1460.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:99–138.

Pérez, J. and Berger, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, 89:491–511.

Peters, J. and Buhlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101:219–228.

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110:159–174.

Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika*, 87:99–122.

Roverato, W. and Whittaker, J. (1998). The isserlis matrix and its application to non-decomposable graphical gaussian models. *Biometrika*, 85(3):711–725.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein signaling networks derived from multiparameter single-cell data. *Science*, 308:523– 529.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Sohn, K.-A. and Kim, S. K. (2012). Joint estimation of structured sparsity and output structure in multiple - output regression via inverse - covariance regularization. *In Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 22:1081 – 1089.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log- linear models with vague prior information. *Journal of the Royal Statistical Society: Series B*, 44(3):377–387.

Tierney, L. and Kadane, J. B. (1989). Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84(407):710–716.

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., and Schildkraus, J. M. (2010). Bayesian model search and multilevel inference for snp association studies. *Annals of Applied Statistics*, 4:1342–1364.

Wytock, M. and Kolter, Z. (2013). Sparse Gaussian conditional random fields. algorithms, theory, and application to energy forecasting. *In Proceedings of the 30th International Conference on Machine Learning*, 28:1265–1273.

Ye, K. and Berger, J. O. (1991). Noninformative priors for inferences in exponential regression models. *Biometrika*, 78:645–656.

Appendix A

Appendix

A.0.1 Likelihood of Equation 4.49

The PEPP Likelihood of $\mathbf{Y}^*$ under a model $G \in \mathscr{G}$ will be provided by

$$f(\mathbf{Y}^*|\mathbf{\Sigma}, \delta, G) = \frac{f(\mathbf{Y}^*|\mathbf{\Sigma}, G)^{1/\delta}}{\int f(\mathbf{Y}^*|\mathbf{\Sigma}, G)^{1/\delta} d\mathbf{Y}^*}.$$

Then

$$\begin{aligned} f(\mathbf{Y}^*|\mathbf{\Sigma}, G)^{1/\delta} &= \left(\frac{\prod_{C \in \mathscr{C}} f(\mathbf{Y}^*_C|\mathbf{\Sigma}_C, G)}{\prod_{S \in \mathscr{S}} f(\mathbf{Y}^*_S|\mathbf{\Sigma}_S, G)} \right)^{1/\delta} \\ &= \frac{\prod_{C \in \mathscr{C}} f(\mathbf{Y}^*_C|\mathbf{\Sigma}_C, G)^{1/\delta}}{\prod_{S \in \mathscr{S}} f(\mathbf{Y}^*_S|\mathbf{\Sigma}_S, G)^{1/\delta}}. \end{aligned}$$

Under each clique $C \in \mathscr{C}$ (and separator $S \in \mathscr{S}$) we have that $\mathbf{Y}^*_C \sim MN_{m \times |C|}(\mathbf{0}, I_m, \mathbf{\Sigma}_C)$, then

$$\begin{aligned} f(\mathbf{Y}^*_C|\mathbf{\Sigma}_C, G)^{1/\delta} &= \left(\frac{det(\mathbf{\Sigma}_C)^{-n/2}}{(2\pi)^{m|C|/2}} exp\Big\{ -\frac{1}{2} tr(\mathbf{\Sigma}_C^{-1}\mathbf{S}_C) \Big\} \right)^{1/\delta} \\ &= \frac{det(\mathbf{\Sigma}_C)^{-\frac{m/\delta}{2}}}{(2\pi)^{\frac{m|C|/\delta}{2}}} exp\Big\{ -\frac{1}{2} tr((\delta\mathbf{\Sigma}_C)^{-1}\mathbf{S}_C) \Big\}. \end{aligned}$$

The integral expression will be provided by

$$\int f(\mathbf{Y}^*|\mathbf{\Sigma}, G)^{1/\delta} d\mathbf{Y}^* = \frac{\prod_{C \in \mathscr{C}} \int f(\mathbf{Y}^*_C|\mathbf{\Sigma}_C, G)^{1/\delta} d\mathbf{Y}^*_C}{\prod_{S \in \mathscr{S}} \int f(\mathbf{Y}^*_S|\mathbf{\Sigma}_S, G)^{1/\delta} d\mathbf{Y}^*_S}.$$

So we have that

$$\begin{aligned} &\int f(\mathbf{Y}^*_C|\mathbf{\Sigma}_C, G)^{1/\delta} d\mathbf{Y}^*_C = \int \frac{det(\mathbf{\Sigma}_C)^{-\frac{m/\delta}{2}}}{(2\pi)^{\frac{m|C|/\delta}{2}}} exp\Big\{ -\frac{1}{2} tr((\delta\mathbf{\Sigma}_C)^{-1}\mathbf{S}_C) \Big\} d\mathbf{Y}^*_C \\ &= \frac{det(\mathbf{\Sigma}_C)^{-\frac{m/\delta}{2}}}{(2\pi)^{\frac{m|C|/\delta}{2}}} \frac{(2\pi)^{m|C|/2}}{det(\delta\mathbf{\Sigma}_C)^{-m/2}} \int \frac{det(\mathbf{\Sigma}_C)^{-m/2}}{(2\pi)^{m|C|/2}} exp\Big\{ -\frac{1}{2} tr((\delta\mathbf{\Sigma}_C)^{-1}\mathbf{S}_C) \Big\} d\mathbf{Y}^*_C \\ &= \frac{det(\mathbf{\Sigma}_C)^{-\frac{m/\delta}{2}}}{(2\pi)^{\frac{m|C|/\delta}{2}}} \frac{(2\pi)^{m|C|/2}}{det(\delta\mathbf{\Sigma}_C)^{-m/2}} \end{aligned}$$

Thus we obtain that

$$\begin{aligned}\frac{f(\mathbf{Y}_C^*|\mathbf{\Sigma}_C,G)^{1/\delta}}{\int f(\mathbf{Y}_C^*|\mathbf{\Sigma}_C,G)^{1/\delta}d\mathbf{Y}_C^*} &= \frac{det(\mathbf{\Sigma}_C)^{-\frac{m/\delta}{2}}}{(2\pi)^{\frac{m|C|/\delta}{2}}}exp\Big\{-\frac{1}{2}tr((\delta\mathbf{\Sigma}_C)^{-1}\mathbf{S}_C)\Big\}\frac{det(\mathbf{\Sigma}_C)^{\frac{m/\delta}{2}}}{(2\pi)^{-\frac{m|C|/\delta}{2}}}\frac{(2\pi)^{-m|C|/2}}{det(\delta\mathbf{\Sigma}_C)^{m/2}}\\ &= \frac{det(\delta\mathbf{\Sigma}_C)^{-m/2}}{(2\pi)^{m|C|/2}}exp\Big\{-\frac{1}{2}tr((\delta\mathbf{\Sigma}_C)^{-1}\mathbf{S}_C)\Big\}.\end{aligned}$$

Thus we conclude that under each clique $C \in \mathscr{C}$ (and separator $S \in \mathscr{S}$), the PEPP likelihood of the imaginary data matrix $\mathbf{Y}_C^*$ is a Matrix Normal Likelihood such that $\mathbf{Y}_C^* \sim MN_{m\times|C|}(\mathbf{0}, I_m, \delta\mathbf{\Sigma}_C)$. Therefore, $\mathbf{Y}|\mathbf{\Sigma} \sim MN_{n\times q}(\mathbf{0}, \mathbf{I}.\delta\mathbf{\Sigma})$.

A.0.2 Proof of Equation 4.40

$$\begin{aligned}m^{EPP}(\mathbf{Y}|G) &= \int f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^{EPP}(\mathbf{\Sigma}|G)d\mathbf{\Sigma}\\ &= \int f(\mathbf{Y}|\mathbf{\Sigma},G)\int \pi^N(\mathbf{\Sigma}|\mathbf{Y}^*,G)m^*(\mathbf{Y}^*|G_0)d\mathbf{Y}^*d\mathbf{\Sigma}\\ &= \int f(\mathbf{Y}|\mathbf{\Sigma},G)\int \frac{f(\mathbf{Y}^*|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)}{m^N(\mathbf{Y}^*|G)}m^*(\mathbf{Y}^*|G_0)d\mathbf{Y}^*d\mathbf{\Sigma}\\ &= \int\int \frac{f(\mathbf{Y}|\mathbf{\Sigma},G)f(\mathbf{Y}^*|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)}{m^N(\mathbf{Y}^*|G)}m^*(\mathbf{Y}^*|G_0)d\mathbf{Y}^*d\mathbf{\Sigma}\\ &= \int\int \frac{f(\mathbf{Y},\mathbf{Y}^*|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)}{m^N(\mathbf{Y}^*|G)}m^*(\mathbf{Y}^*|G_0)d\mathbf{Y}^*d\mathbf{\Sigma}\\ &= \int \frac{\int f(\mathbf{Y},\mathbf{Y}^*|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)d\mathbf{\Sigma}}{m^N(\mathbf{Y}^*|G)}m^*(\mathbf{Y}^*|G_0)d\mathbf{Y}^*\\ &= \int \frac{m^N(\mathbf{Y},\mathbf{Y}^*|G)}{m^N(\mathbf{Y}^*|G)}m^*(\mathbf{Y}^*|G_0)d\mathbf{Y}^*\\ &= \int m^N(\mathbf{Y}|\mathbf{Y}^*,G)m^*(\mathbf{Y}^*|G_0)d\mathbf{Y}^*\end{aligned}$$

A.0.3 Proof of passage from Equation 4.41 to Equation 4.44

Consider relation Equation 4.41 where we set the importance density of the numerator to be $g(\mathbf{Y}^*) = m^N(\mathbf{Y}^{*(r)}|\mathbf{Y},G)$ and the importance density of the denominator to be $g(\mathbf{Y}^{*(r)}) = m^N(\mathbf{Y}^{*(r)}|\mathbf{Y},G_0)$. Thus, for $m^*(\mathbf{Y}^{*(r)}) = m^N(\mathbf{Y}^{*(r)}|G_0)$ and $G^{'} = G_0$

(15) will be structured as

$$
\begin{aligned}
\hat{BF}^*_{G:G_0}(\mathbf{Y}) &= \frac{\sum_{r=1}^{R} m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G)m^N(\mathbf{Y}^{*(r)}|G_0)/m^N(\mathbf{Y}^{*(r)}|\mathbf{Y}, G)}{\sum_{r=1}^{R} m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G_0)m^N(\mathbf{Y}^{*(r)}|G_0)/m^N(\mathbf{Y}^{*(r)}|\mathbf{Y}, G_0)} \\
&= \frac{\sum_{r=1}^{R} \frac{m^N(\mathbf{Y},\mathbf{Y}^{*(r)}|G)}{m^N(\mathbf{Y}^{*(r)}|G)} m^N(\mathbf{Y}^{*(r)}|G_0)\frac{m^N(\mathbf{Y}|G)}{m^N(\mathbf{Y}^{*(r)},\mathbf{Y}|G)}}{\sum_{r=1}^{R} \frac{m^N(\mathbf{Y},\mathbf{Y}^{*(r)}G_0)}{m^N(\mathbf{Y}^{*(r)}|G_0)} m^N(\mathbf{Y}^{*(r)}|G_0)\frac{m^N(\mathbf{Y}|G_0)}{m^N(\mathbf{Y}^{*(r)},\mathbf{Y}|G_0)}} \\
&= \frac{1}{R}\frac{m^N(\mathbf{Y}|G)}{m^N(\mathbf{Y}|G_0)} \sum_{r=1}^{R} \frac{m^N(\mathbf{Y}^{*(r)}|G_0)}{m^N(\mathbf{Y}^{*(r)}} \\
&= BF^N_{G:G_0}(\mathbf{Y})\frac{1}{R}\sum_{r=1}^{r} BF^N_{G_0:G}(\mathbf{Y}^{*(r)}).
\end{aligned}
$$

A.0.4 Proof of passage from Equation 4.41 to Equation 4.42

Consider Equation 4.41 where the predictive density will be provided by $m^*(\mathbf{Y}^*_i) = m^N(\mathbf{Y}^*_i|G_0)$ and the importance density will be provided by $g(\mathbf{Y}^*_i) = m^N(\mathbf{Y}^*_i|\mathbf{Y})$. Thus, by letting $G^{'} = G_0$ Equation 4.41 will be structured as

$$
\begin{aligned}
\hat{BF}^*_{G:G_0}(\mathbf{Y}) &= \frac{\sum_{r=1}^{R} m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G)m^*(\mathbf{Y}^{*(r)})/g(\mathbf{Y}^{*(r)})}{\sum_{r=R}^{r} m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G_0)m^*(\mathbf{Y}^{*(r)})/g(\mathbf{Y}^{*(r)})} \\
&= \frac{\sum_{r=1}^{R} m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G)m^N(\mathbf{Y}^{*(r)}|G_0)/m^N(\mathbf{Y}^{*(r)}|\mathbf{Y}, G_0)}{\sum_{r=1}^{R} m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G_0)m^N(\mathbf{Y}^{*(r)}|G_0)/m^N(\mathbf{Y}^{*(r)}|\mathbf{Y}, G_0)} \\
&= \frac{\sum_{r=1}^{R} m^N(\mathbf{Y}^{*(r)}|G_0)\frac{m^N(\mathbf{Y}|\mathbf{Y}^{*(r)},G)}{m^N(\mathbf{Y}^{*(r)}|\mathbf{Y},G_0)}}{\sum_{r=1}^{R} \frac{m^N(\mathbf{Y},\mathbf{Y}^{*(r)}|G_0)}{m^N(\mathbf{Y}^{*(r)}|G_0)} m^N(\mathbf{Y}^{*(r)}|G_0)\frac{m^N(\mathbf{Y}|G_0)}{m^N(\mathbf{Y}^{*(r)},\mathbf{Y}|G_0)}} \\
&= \frac{1}{R}\frac{\sum_{r=1}^{R} m^N(\mathbf{Y}^{*(r)}|G_0)\frac{m^N(\mathbf{Y}|\mathbf{Y}^{*(r)},G)}{m^N(\mathbf{Y}^{*(r)}|\mathbf{Y},G_0)}}{m^N(\mathbf{Y}|G_0)} \\
&= \frac{1}{R}\sum_{r=1}^{R} \frac{m^N(\mathbf{Y}^{*(r)}|G_0)}{m^N(\mathbf{Y}|G_0)}\frac{m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G)}{\frac{m^N(\mathbf{Y}^{*(r)},\mathbf{Y}|G_0)}{m^N(\mathbf{Y}|G_0)}} \\
&= \frac{1}{R}\sum_{r=R}^{r} \frac{m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G)}{\frac{m^N(\mathbf{Y},\mathbf{Y}^{*(r)}|G_0)}{m^N(\mathbf{Y}^{*(r)})}} \\
&= \frac{1}{R}\sum_{r=1}^{R} \frac{m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G)}{m^N(\mathbf{Y}|\mathbf{Y}^{*(r)}, G_0)} = \frac{1}{R}\sum_{r=1}^{R} BF^N_{G:G_0}(\mathbf{Y}|\mathbf{Y}^{*(r)}).
\end{aligned}
$$

A.0.5 Proof of Equation 4.60

$$\begin{aligned}\int f(\mathbf{Y}^*|\mathbf{Y},\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|\mathbf{Y},G)d\mathbf{\Sigma} = \\ &= \int f(\mathbf{Y}^*|\mathbf{\Sigma},G)\frac{f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)}{\int f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)d\mathbf{\Sigma}}d\mathbf{\Sigma} \\ &= \frac{\int f(\mathbf{Y}^*|\mathbf{\Sigma},G)f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)d\mathbf{\Sigma}}{\int f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)d\mathbf{\Sigma}} \\ &= \frac{\int f(\mathbf{Y}^*,\mathbf{Y}|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)d\mathbf{\Sigma}}{\int f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)d\mathbf{\Sigma}} \\ &= \frac{m^N(\mathbf{Y}^*,\mathbf{Y}|G)}{m^N(\mathbf{Y}|G)} \\ &= m^N(\mathbf{Y}^*|\mathbf{Y},G).\end{aligned}$$

Note that we exploited the property that $\mathbf{Y}$ and $\mathbf{Y}^*$ arise independently on a common sample space.

A.0.6 Proof of Equation 4.54

$$\begin{aligned}&\pi^{PEPP}(\mathbf{\Sigma}|\mathbf{Y},\delta,G) \propto f(\mathbf{Y}^*|\mathbf{\Sigma},G)\pi^{PEPP}(\mathbf{\Sigma}|\mathbf{Y}^*,\delta,G) \\ &\propto f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)\int \frac{m^N(\mathbf{Y}^*|\delta,G_0)}{m^N(\mathbf{Y}^*|\delta,G)}f(\mathbf{Y}^*|\mathbf{\Sigma},\delta,G)d\mathbf{Y}^* \\ &\propto \int f(\mathbf{Y}|\mathbf{\Sigma},G)\frac{f(\mathbf{Y}^*|\mathbf{\Sigma},\delta,G)\pi^N(\mathbf{\Sigma}|G)}{m^N(\mathbf{Y}^*|\delta,G)}m^N(\mathbf{Y}^*|\delta,G_0)d\mathbf{Y}^* \\ &\propto \int f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*,\delta,G)m^N(\mathbf{Y}^*|\delta,G_0)d\mathbf{Y}^* \\ &\propto \int \pi^N(\mathbf{\Sigma}|\mathbf{Y},\mathbf{Y}^*,\delta,G)m^N(\mathbf{Y}|\mathbf{Y}^*,\delta,G)m^N(\mathbf{Y}^*|\delta,G_0)d\mathbf{Y}^*.\end{aligned}$$

A.0.7 Proof of Equation 4.55

$$\begin{aligned}
m^{PEPP}(\mathbf{Y}|\delta,G) &= \int f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^{PEPP}(\mathbf{\Sigma}|\mathbf{Y}^*,\delta,G)d\mathbf{\Sigma}\\
&= \int f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^{N}(\mathbf{\Sigma}|G)\int \frac{m^N(\mathbf{Y}^*|\delta,G_0)}{m^N(\mathbf{Y}^*|\delta,G)} f(\mathbf{Y}^*|\mathbf{\Sigma},\delta,G)d\mathbf{Y}^*d\mathbf{\Sigma}\\
&= \int \frac{m^N(\mathbf{Y}^*|\delta,G_0)}{m^N(\mathbf{Y}^*|\delta,G)} \int f(\mathbf{Y}|\mathbf{\Sigma},G)f(\mathbf{Y}^*|\mathbf{\Sigma},\delta,G)\pi^N(\mathbf{\Sigma}|G)d\mathbf{\Sigma}d\mathbf{Y}^*\\
&= \int \frac{m^N(\mathbf{Y}^*|\delta,G_0)}{m^N(\mathbf{Y}^*|\delta,G)} \int f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*,\delta,G)m^N(\mathbf{Y}^*|\delta,G)d\mathbf{\Sigma}d\mathbf{Y}^*\\
&= \int m^N(\mathbf{Y}^*|\delta,G_0)m^N(\mathbf{Y}|\mathbf{Y}^*,\delta,G)d\mathbf{Y}^*\\
&= \int \frac{m^N(\mathbf{Y},\mathbf{Y}^*|\delta,G)}{m^N(\mathbf{Y}^*|\delta,G)}m^N(\mathbf{Y}^*|\delta,G_0)d\mathbf{Y}^*\\
&= \int \frac{m^N(\mathbf{Y},\mathbf{Y}^*|\delta,G)}{m^N(\mathbf{Y}|G)}\frac{m^N(\mathbf{Y}|G)}{m^N(\mathbf{Y}^*|\delta,G)}d\mathbf{Y}^*\\
&= m^N(\mathbf{Y}|G)\int \frac{m^N(\mathbf{Y}^*|\mathbf{Y},\delta,G)}{m^N(\mathbf{Y}^*|\delta,G)}m^N(\mathbf{Y}^*|\delta,G_0)d\mathbf{Y}^*.
\end{aligned}$$

A.0.8 Posterior distribution of Equation 4.30

The posterior distribution of Equation 4.30 under a graph $G \in \mathscr{G}$ will be provided by

$$\begin{aligned}
\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*,G) &\propto f(\mathbf{Y}^*|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)\\
&\propto \frac{\prod_{C\in\mathscr{C}} f(\mathbf{Y}^*_C|\mathbf{\Sigma}_C,G)\pi^N(\mathbf{\Sigma}_C|G)}{\prod_{S\in\mathscr{S}} f(\mathbf{Y}^*_S|\mathbf{\Sigma}_S,G)\pi^N(\mathbf{\Sigma}_S|G)}\\
&\propto \frac{\prod_{C\in\mathscr{C}} det(\mathbf{\Sigma}_C)^{-m/2}\exp\Big\{-\frac{1}{2}tr\big(\mathbf{\Sigma}_C^{-1}\mathbf{S}^*_C\big)\Big\}\det(\mathbf{\Sigma}_C)^{-|C|}}{\prod_{S\in\mathscr{S}} det(\mathbf{\Sigma}_S)^{-m/2}\exp\Big\{-\frac{1}{2}tr\big(\mathbf{\Sigma}_S^{-1}\mathbf{S}^*_S\big)\Big\}\det(\mathbf{\Sigma}_S)^{-|S|}}\\
&\propto \frac{\prod_{C\in\mathscr{C}} det(\mathbf{\Sigma}_C)^{-(m/2+|C|)}\exp\Big\{-\frac{1}{2}tr\big(\mathbf{\Sigma}_C^{-1}\mathbf{S}^*_C\big)\Big\}}{\prod_{S\in\mathscr{S}} det(\mathbf{\Sigma}_S)^{-(m/2+|S|)}\exp\Big\{-\frac{1}{2}tr\big(\mathbf{\Sigma}_S^{-1}\mathbf{S}^*_S\big)\Big\}}.
\end{aligned}$$

Given the form of an Inverse-Wishart distribution, we deduce that the kernel of the distribution in the nominator represents an Inverse-Wishart distribution with parameters $b = m$, $p = |C|$ and $D_C = \mathbf{S}^*_C$. Similarly the kernel of the distribution of the denominator will be an Inverse-Wishart with parameters $b = m$, $p = |S|$ and $D_S = \mathbf{S}^*_S$. Thus, the posterior distribution of $\mathbf{\Sigma}$ given $\mathbf{Y}^*$ under an graph G will be a $HIW_G(m, \mathbf{S}^*)$ where $\mathbf{S}^* = \mathbf{Y}^{*^T}\mathbf{Y}^*$ and $b = m$.

A.0.9 Proof of Equation 5.1

Let $G \in \mathscr{G}$ and G_0 be the independence graph. Then the Posterior Odds of G versus G_0 will be given by

$$PO(G : G_0) = BF_{G:G_0}(\mathbf{Y})O(G : G_0).$$

The Posterior Model Probability of any $G \in \mathscr{G}$ will be provided by

$$\begin{aligned}\pi(G|\mathbf{Y}) &= \frac{\pi(G)m(\mathbf{Y}|G)}{\sum_{G_l \in \mathscr{G}} \pi(G_l)m(\mathbf{Y}|G_l)} \\ &= \frac{\pi(G)m(\mathbf{Y}|G)}{\sum_{G_l \in \mathscr{G}} \pi(G_l)m(\mathbf{Y}|G_l)} \frac{\pi(G_0)m(\mathbf{Y}|G_0)}{\pi(G_0)m(\mathbf{Y}|G_0)} \\ &= \frac{O_{G:G_0} BF_{G:G_0}(\mathbf{Y})}{\sum_{G_l \in \mathscr{G}} O_{G_l:G_0} BF_{G_l:G_0}(\mathbf{Y})} \\ &= \frac{PO(G : G_0)}{\sum_{G_l \in \mathscr{G}} PO(G_l : G_0)}.\end{aligned}$$

A.0.10 Bayes Factor of Equation 5.4

The Bayes factor of Equation 5.4 (fpr the case of $\delta = 1$) is provided by

$$BF_{G:G_0}(\mathbf{Y}) \approx \widehat{BF}^{EPP}_{G:G_0}(\mathbf{Y}) = BF^N_{G:G_0}(\mathbf{Y})\frac{1}{r}\sum_{l=1}^{r} BF^N_{G_0:G}(\mathbf{Y}^*_l).$$

The logarithm of the above-mentioned Bayes factor will be provided by

$$\begin{aligned}\log(\widehat{BF}^{EPP}_{G:G_0}(\mathbf{Y})) &= log(BF^N_{G:G_0}(\mathbf{Y})) + log\Big(\sum_{l=1}^{r} BF^N_{G_0:G}(\mathbf{Y}^*_l)\Big) - log(r) \\ &= log\Big(\frac{m^N(\mathbf{Y}|G)}{m^N(\mathbf{Y}|G_0)}\Big) + log\Big(\sum_{l=1}^{r} \frac{m^N(\mathbf{Y}^*_l|G_0)}{m^N(\mathbf{Y}^*_l|G)}\Big) - log(r).\end{aligned}$$

Note that $\mathbf{Y}_{n\times p} \sim MN_{n\times q}(\mathbf{0}, I_n, \mathbf{\Sigma})$, $\mathbf{Y}^*_{m\times q} \sim MN_{m\times q}(\mathbf{0}, I_n, \mathbf{\Sigma})$.The posterior distribution of $\mathbf{\Sigma}|\mathbf{Y}, G$ is a $HIW(n, \mathbf{S})$. The marginal density of the data matrix $\mathbf{Y}$ under any given undirected decomposable graph $G \in \mathscr{G}$, with respect to the setup provided in this report, will be provided by

$$\begin{aligned}m^N(\mathbf{Y}|G) &= \frac{f(\mathbf{Y}|\mathbf{\Sigma}, G)\pi^N(\mathbf{\Sigma}|G)}{\pi^N(\mathbf{\Sigma}|\mathbf{Y}, G)} \\ &= \frac{\prod_{C\in\mathscr{C}} \frac{f(\mathbf{Y}_C|\mathbf{\Sigma}_C,G)\pi^N(\mathbf{\Sigma}_C|G)}{\pi^N(\mathbf{\Sigma}_C|\mathbf{Y}_C,G)}}{\prod_{S\in\mathscr{S}} \frac{f(\mathbf{Y}_S|\mathbf{\Sigma}_S,G)\pi^N(\mathbf{\Sigma}_S|G)}{\pi^N(\mathbf{\Sigma}_S|\mathbf{Y}_S,G)}}.\end{aligned}$$

Under each clique $C \in \mathscr{C}$ we obtain that

$$
\begin{aligned}
&\frac{f(\mathbf{Y}_C|\mathbf{\Sigma}_C, G)\pi^N(\mathbf{\Sigma}_C|G)}{\pi^N(\mathbf{\Sigma}_C|\mathbf{Y}_C, G)} = \\
&= \frac{det(\mathbf{\Sigma}_C)^{-n/2}(2\pi)^{-n|C|/2}\exp\Big(-\tfrac{1}{2}tr(\mathbf{\Sigma}_C^{-1}\mathbf{S}_C)\Big)det(\mathbf{\Sigma}_C)^{-|C|}}{2^{-\frac{n+|C|-1}{2}|C|}\Gamma^{-1}_{|C|}(\frac{n+|C|-1}{2})det(\mathbf{S}_C)^{\frac{n+|C|-1}{2}}det(\mathbf{\Sigma}_C)^{\frac{n+|C|-1+|C|+1}{2}}\exp\Big(-\tfrac{1}{2}tr(\mathbf{\Sigma}_C^{-1}\mathbf{S}_C)\Big)} \\
&= \frac{\Gamma_{|C|}(\frac{n+|C|-1}{2})}{(2\pi)^{n|C|/2}(\frac{1}{2})^{\frac{n+|C|-1}{2}|C|}det(\mathbf{S}_C)^{\frac{n+|C|-1}{2}}} \\
&= \frac{\Gamma_{|C|}(\frac{n+|C|-1}{2})}{(2\pi)^{n|C|/2}det(\frac{1}{2}\mathbf{S}_C)^{\frac{n+|C|-1}{2}}} \\
&= \Gamma_{|C|}(\frac{n+|C|-1}{2})(2\pi)^{-n|C|/2}det(\frac{1}{2}\mathbf{S}_C)^{-\frac{n+|C|-1}{2}}.
\end{aligned}
$$

Similarly, we have the same expression for each separator $S \in \mathscr{S}$, thus the marginal likelihood of $\mathbf{Y}$ under the decomposable graph G will be provided by

$$
\begin{aligned}
m^N(\mathbf{Y}|G) &= \frac{\prod_{C\in\mathscr{C}}\Gamma_{|C|}(\frac{n+|C|-1}{2})(2\pi)^{-n|C|/2}det(\frac{1}{2}\mathbf{S}_C)^{-\frac{n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}\Gamma_{|S|}(\frac{n+|S|-1}{2})(2\pi)^{-n|S|/2}det(\frac{1}{2}\mathbf{S}_S)^{-\frac{n+|C|-1}{2}}} \\
&= (2\pi)^{-nq/2}\frac{\prod_{C\in\mathscr{C}}\Gamma_{|C|}(\frac{n+|C|-1}{2})det(\frac{1}{2}\mathbf{S}_C)^{-\frac{n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}\Gamma_{|S|}(\frac{n+|S|-1}{2})det(\frac{1}{2}\mathbf{S}_S)^{-\frac{n+|C|-1}{2}}}.
\end{aligned}
$$

Given the array of imaginary data, we get in a similar fashion the marginal likelihood of each matrix $\mathbf{Y}_l^*$ by

$$
\begin{aligned}
m^N(\mathbf{Y}_l^*|G) &= \frac{\prod_{C\in\mathscr{C}}\Gamma_{|C|}(\frac{m+|C|-1}{2})(2\pi)^{-m|C|/2}det(\frac{1}{2}\mathbf{S}_C^*)^{-\frac{m+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}\Gamma_{|S|}(\frac{m+|S|-1}{2})(2\pi)^{-m|S|/2}det(\frac{1}{2}\mathbf{S}_S^*)^{-\frac{m+|C|-1}{2}}} \\
&= (2\pi)^{-mq/2}\frac{\prod_{C\in\mathscr{C}}\Gamma_{|C|}(\frac{m+|C|-1}{2})det(\frac{1}{2}\mathbf{S}_C^*)^{-\frac{m+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}\Gamma_{|S|}(\frac{m+|S|-1}{2})det(\frac{1}{2}\mathbf{S}_S^*)^{-\frac{m+|C|-1}{2}}}.
\end{aligned}
$$

Under the independence graph G_0, the marginal likelihood of the data matrix $\mathbf{Y}$ will be provided by

$$
m^N(\mathbf{Y}|G_0) = \Gamma^q(\frac{n}{2})(2\pi)^{-nq/2}\prod_{j=1}^{q}det(\frac{1}{2}\mathbf{S}_j)^{-\frac{n}{2}}
$$

and the marginal likelihood of the imaginary data matrix $\mathbf{Y}_l^*$ will be provided by

$$
m^N(\mathbf{Y}_l^*|G_0) = \Gamma^q(\frac{m}{2})(2\pi)^{-mq/2}\prod_{j=1}^{q}det(\frac{1}{2}\mathbf{S}_j^*)^{-\frac{m}{2}}.
$$

Thus by, using log terms we obtain

$$
\begin{aligned}
\log(\widehat{BF}^{EPP}_{G:G_0}(\mathbf{Y})) &= log(BF^N_{G:G_0}(\mathbf{Y})) + log\Big(\sum_{l=1}^{r} BF^N_{G_0:G}(\mathbf{Y}^*_l)\Big) - log(r) \\
&= log\Big(\frac{m^N(\mathbf{Y}|G)}{m^N(\mathbf{Y}|G_0)}\Big) + log\Big(\sum_{l=1}^{r}\frac{m^N(\mathbf{Y}^*_l|G_0)}{m^N(\mathbf{Y}^*_l|G)}\Big) - log(r) \\
&= \sum_{C\in\mathscr{C}}\Big[log\Big(\Gamma_{|C|}(\frac{n+|C|-1}{2})\Big) - \frac{n+|C|-1}{2}log(det(\frac{1}{2}\mathbf{S}_C))\Big] - \\
&- \sum_{S\in\mathscr{S}}\Big[log\Big(\Gamma_{|S|}(\frac{n+|S|-1}{2})\Big) - \frac{n+|S|-1}{2}log(det(\frac{1}{2}\mathbf{S}_S))\Big] - \\
&- qlog(\Gamma(\frac{n}{2})) + \frac{n}{2}\sum_{j=1}^{q} det(\frac{1}{2}\mathbf{S}_j) + log\Big(\sum_{l=1}^{r}\frac{m^N(\mathbf{Y}^*_l|G_0)}{m^N(\mathbf{Y}^*_l|G)}\Big) - log(r).
\end{aligned}
$$

A.0.11 Posterior distribution of Equation 4.50

$$
\pi^N(\mathbf{\Sigma}|\mathbf{Y},\delta,G) = \frac{\prod_{C\in\mathscr{C}}\pi^N(\mathbf{\Sigma}_C|\mathbf{Y}_C,\delta,G)}{\prod_{S\in\mathscr{S}}\pi^N(\mathbf{\Sigma}_S|\mathbf{Y}_S,\delta,G)}.
$$

Then, under each clique $C \in \mathscr{C}$ (and separator $S \in \mathscr{S}$) we obtain that

$$
\begin{aligned}
\pi^N(\mathbf{\Sigma}_C|\mathbf{Y}_C,\delta,G) &\propto f(\mathbf{Y}_C,\delta,G)\pi^N(\mathbf{\Sigma}_C|G) \\
&\propto det(\delta\mathbf{\Sigma}_C)^{-m/2}exp\Big\{-\frac{1}{2}tr((\delta\mathbf{\Sigma}_C)^{-1}\mathbf{S}_C)\Big\}det(\mathbf{\Sigma}_C)^{-|C|} \\
&\propto det(\mathbf{\Sigma}_C)^{-(m/2+|C|)}exp\Big\{-\frac{1}{2}tr(\mathbf{\Sigma}_C^{-1}\delta^{-1}\mathbf{S}_C)\Big\}
\end{aligned}
$$

which is an Inverse Wishart Kernel with $p = |C|$, $b = m$ and $D = \delta^{-1}\mathbf{S}_C$. Thus, the posterior distribution of $\mathbf{\Sigma}$ given $\mathbf{Y}$ under a graph G will be a Hype-Inverse Wishart with degrees of freedom parameter $b = m$ and scale matrix $\delta^{-1}\mathbf{S}$.

A.0.12 Marginal likelihood of Equation 4.34

Let $\mathbf{Y}^*_{m\times q}$ be the data matrix consisted by m independent vectors of q-dimensional imaginary data as in (1) and $G_0 = (V, E_0)$ where $E_0 = \{\emptyset\}$ will represent the indepedence graph. Thus, the marginal distribution of $\mathbf{Y}^*$ under the independence graph G_0 will be derived by,

$$
\begin{aligned}
m^N(\mathbf{Y}^*|G_0) &= \int f(\mathbf{Y}^*|\mathbf{\Sigma},G_0)\pi^N(\mathbf{\Sigma}|G_0)d\mathbf{\Sigma} \\
&= \prod_{C\in\mathscr{C}_0}\frac{\int f(\mathbf{Y}^*_C|\mathbf{\Sigma}_C,G_0)\pi^N(\mathbf{\Sigma}_C|G_0)d\mathbf{\Sigma}_C}{\int f(\mathbf{Y}^*_S|\mathbf{\Sigma}_S,G_0)\pi^N(\mathbf{\Sigma}_S|G_0)d\mathbf{\Sigma}_S}.
\end{aligned}
$$

Now we change our focus on the integral expression, where the likelihood of $\mathbf{Y}_C^*$ will be provided my a $MN_{|C|\times q}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma}_C)$ and $\pi^N(\mathbf{\Sigma}_C|G_0)$ will be provided by (7). Thus, under each clique $C \in \mathscr{C}_0$ the integral will be provided by,

$$\begin{aligned}
&\int f(\mathbf{Y}_C^*|\mathbf{\Sigma}_C, G_0)\pi^N(\mathbf{\Sigma}_C|G_0)d\mathbf{\Sigma}_C = \\
&= \int \frac{det(\mathbf{\Sigma}_C)^{-m/2}}{(2\pi)^{m|C|/2}} \exp\Big\{-\frac{1}{2}tr\big(\mathbf{\Sigma}_C^{-1}\mathbf{S}_C^*\big)\Big\} det(\mathbf{\Sigma}_C)^{-|C|}d\mathbf{\Sigma}_C \\
&= \frac{1}{(2\pi)^{m|C|/2}} \int det(\mathbf{\Sigma}_C)^{-(m/2+|C|)} \exp\Big\{-\frac{1}{2}tr\big(\mathbf{\Sigma}_C^{-1}\mathbf{S}_C^*\big)\Big\} d\mathbf{\Sigma}_C.
\end{aligned}$$

We deduce that the kernel inside the integral expression represents and Inverse-Wishart distribution with parameters $b = m$, $p = |C|$ and $D_C = \mathbf{S}_C^*$ (similarly for each separator $S \in \mathscr{S}_0$). So we obtain,

$$\begin{aligned}
\int f(\mathbf{Y}_C^*|\mathbf{\Sigma}_C, G_0)\pi^N(\mathbf{\Sigma}_C|G_0)d\mathbf{\Sigma}_C &= \frac{1}{(2\pi)^{m|C|/2}} \frac{2^{(m+|C|-1)|C|/2}\Gamma_{|C|}(\frac{m+|C|-1}{2})}{det(\mathbf{S}_C^*)^{(m+|C|-1)/2}} \\
&= (2\pi)^{-m|C|/2} \frac{\Gamma_{|C|}(\frac{m+|C|-1}{2})}{det(\frac{1}{2}\mathbf{S}_C^*)^{(m+|C|-1)/2}},
\end{aligned}$$

where $\Gamma_C(\cdot)$ is the multivariate Gamma function. Since we operate under the independence graph, it is evident that each node will be a clique of its own, i.e. $\mathscr{C}_0 = \{\mathbf{Y}_1^*, \cdots, \mathbf{Y}_q^*\}$. As indicated by Lauritzen (1996) pp. 90, if the empty set is a separator, it should be included in the set $\mathscr{S}_0$, so under the independence graph G_0 we have that $\mathscr{S}_0 = \{\emptyset\}$. Also, by using Lauritzen (1996) we have that $f(\mathbf{Y}_S^*|\mathbf{\Sigma}_S, G_0) \equiv 1$ and we let $\pi^N(\mathbf{\Sigma}_S|G_0) \equiv 1$. Thus the marginal distribution of $\mathbf{Y}^*$ under the independence graph G_0 will be provided by

$$\begin{aligned}
m^N(\mathbf{Y}^*|G_0) &= \prod_{j=1}^{q} (2\pi)^{-m/2} \frac{\Gamma(\frac{m}{2})}{det(\frac{1}{2}\mathbf{S}_j^*)^{m/2}} \\
&= (2\pi)^{-mq/2}\Gamma^q(\frac{m}{2}) \prod_{j=1}^{q} det(\frac{1}{2}\mathbf{S}_j^*)^{-m/2},
\end{aligned}$$

where S_j is structured by the the j-the column of $\mathbf{Y}^*$ matrix, i.e. $\mathbf{S}_j = \mathbf{Y}_j^T\mathbf{Y}_j$. Finally, we will write

$$m^N(\mathbf{Y}^*|G_0) \propto (2\pi)^{-mq/2}\Gamma^q(\frac{m}{2}) \prod_{j=1}^{q} det(\frac{1}{2}\mathbf{S}_j^*)^{-m/2}$$

because this expression will be restricted by the arbitrary normalizing constant which arises from the use of the baseline prior $\pi^N(\mathbf{\Sigma}|G_0)$.

A.0.13 Marginal likelihood of Equation 4.6 for $G = G_0$

The marginal likelihood of $\mathbf{Y}^*$ under a graph $G \in \mathscr{G}$ will be provided by

$$m^N(\mathbf{Y}^*|\delta, G) = \frac{f(\mathbf{Y}^*|\mathbf{\Sigma}, \delta, G)\pi^N(\mathbf{\Sigma}|G)}{\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*, \delta, G)}$$

where $\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*,\delta,G)$ is the posterior distribution of $\mathbf{\Sigma}$ given the imaginary data matrix $\mathbf{Y}^*$ under model G, as provided by Section 4.4 of the present Appendix. Under each clique $C \in \mathscr{C}$ (and separator $S \in \mathscr{S}$) we have that

$$
\begin{aligned}
&\frac{f(\mathbf{Y}^*_C|\mathbf{\Sigma}_C,\delta,G)\pi^N(\mathbf{\Sigma}_C|G)}{\pi^N(\mathbf{\Sigma}_C|\mathbf{Y}^*_C,\delta,G)} = \\
&= \frac{det(\delta\mathbf{\Sigma}_C)^{-m/2}}{(2\pi)^{m|C|/2}} exp\Big\{-\frac{1}{2}tr((\delta\mathbf{\Sigma}_C)^{-1}\mathbf{S}^*_C)\Big\} det(\mathbf{\Sigma}_C)^{-|C|} \times \\
&\times \frac{2^{\frac{m+|C|-1}{2}|C|}\Gamma_{|C|}(\frac{m+|C|-1}{2})}{det(\delta^{-1}\mathbf{S}^*_C)^{\frac{m+|C|-1}{2}} det(\mathbf{\Sigma}_C)^{-(m/2+|C|)} exp\Big\{\mathbf{\Sigma}_C^{-1}\delta^{-1}\mathbf{S}^*_C\Big\}} \\
&= \frac{\delta^{-m|C|/2}2^{\frac{m+|C|-1}{2}|C|}\Gamma_{|C|}(\frac{m+|C|-1}{2})}{(2\pi)^{m|C|/2}det(\frac{1}{\delta}\mathbf{S}^*_C)^{\frac{m+|C|-1}{2}}} \\
&= (2\pi\delta)^{-m|C|/2}\frac{\Gamma_{|C|}(\frac{m+|C|-1}{2})}{det(\frac{1}{2\delta}\mathbf{S}^*_C)^{\frac{m+|C|-1}{2}}}.
\end{aligned}
$$

As indicated byLauritzen (1996), if the empty set is a separator, it should be included in the set $\mathscr{S}_0$, so under the independence graph G_0 we have that $\mathscr{S}_0 = \{\emptyset\}$. Also, by usingLauritzen (1996) we have that $f(\mathbf{Y}^*_S|\mathbf{\Sigma}_S,G_0) \equiv 1$ and we let $\pi^N(\mathbf{\Sigma}_S|G_0) \equiv 1$. Thus the marginal distribution of $\mathbf{Y}^*$ under the independence graph G_0 will be provided by

$$
\begin{aligned}
m^N(\mathbf{Y}^*|\delta,G_0) &= \prod_{j=1}^{q}(2\pi\delta)^{-m/2}\frac{\Gamma(\frac{m}{2})}{det(\frac{1}{2\delta}\mathbf{S}^*_j)^{\frac{m}{2}}} \\
&= (2\pi\delta)^{-mq/2}\Gamma^q(m/2)\prod_{j=1}^{q}det(\frac{1}{2\delta}\mathbf{S}^*_j)^{-\frac{m}{2}}
\end{aligned}
$$

and we will write

$$
m^N(\mathbf{Y}^*|\delta,G_0) \propto (2\pi\delta)^{-mq/2}\Gamma^q(m/2)\prod_{j=1}^{q}det(\frac{1}{2\delta}\mathbf{S}^*_j)^{-\frac{m}{2}}
$$

due to the arbitrary normalizing constant that arises from the use of the improper baseline prior $\pi^N(\mathbf{\Sigma}|G_0)$.

A.0.14 Marginal Likelihood of Equation 4.39 under any $G \in \mathscr{G}$

The marginal likelihood of the data matrix $\mathbf{Y}$ under a graph $G \in \mathscr{G}$, will be provided by

$$
\begin{aligned}
m^N(\mathbf{Y}|G) &= \frac{f(\mathbf{Y}|\mathbf{\Sigma},G)\pi^N(\mathbf{\Sigma}|G)}{\pi^N(\mathbf{\Sigma}|\mathbf{Y},G)} \\
&= \frac{\prod_{C\in\mathscr{C}}\frac{f(\mathbf{Y}_C|\mathbf{\Sigma}_C,G)\pi^N(\mathbf{\Sigma}_C|G)}{\pi^N(\mathbf{\Sigma}_C|\mathbf{Y}_C,G)}}{\prod_{S\in\mathscr{S}}\frac{f(\mathbf{Y}_S|\mathbf{\Sigma}_S,G)\pi^N(\mathbf{\Sigma}_S|G)}{\pi^N(\mathbf{\Sigma}_S|\mathbf{Y}_S,G)}}.
\end{aligned}
$$

Under each clique $C \in \mathscr{C}$ we obtain that

$$
\begin{aligned}
&\frac{f(\mathbf{Y}_C|\mathbf{\Sigma}_C, G)\pi^N(\mathbf{\Sigma}_C|G)}{\pi^N(\mathbf{\Sigma}_C|\mathbf{Y}_C, G)} = \\
&= \frac{det(\mathbf{\Sigma}_C)^{-n/2}(2\pi)^{-n|C|/2}\exp\Big(-\tfrac{1}{2}tr(\mathbf{\Sigma}_C^{-1}\mathbf{S}_C)\Big)det(\mathbf{\Sigma}_C)^{-|C|}}{2^{-\frac{n+|C|-1}{2}|C|}\Gamma_{|C|}^{-1}(\frac{n+|C|-1}{2})det(\mathbf{S}_C)^{\frac{n+|C|-1}{2}}det(\mathbf{\Sigma}_C)^{\frac{n+|C|-1+|C|+1}{2}}\exp\Big(-\tfrac{1}{2}tr(\mathbf{\Sigma}_C^{-1}\mathbf{S}_C)\Big)} \\
&= \frac{\Gamma_{|C|}(\frac{n+|C|-1}{2})}{(2\pi)^{n|C|/2}(\frac{1}{2})^{\frac{n+|C|-1}{2}|C|}det(\mathbf{S}_C)^{\frac{n+|C|-1}{2}}} \\
&= \frac{\Gamma_{|C|}(\frac{n+|C|-1}{2})}{(2\pi)^{n|C|/2}det(\frac{1}{2}\mathbf{S}_C)^{\frac{n+|C|-1}{2}}} \\
&= \Gamma_{|C|}(\frac{n+|C|-1}{2})(2\pi)^{-n|C|/2}det(\frac{1}{2}\mathbf{S}_C)^{-\frac{n+|C|-1}{2}}.
\end{aligned}
$$

Similarly, we have the same expression for each separator $S \in \mathscr{S}$, thus the marginal likelihood of $\mathbf{Y}$ under the decomposable graph G will be provided by

$$
\begin{aligned}
m^N(\mathbf{Y}|G) &= \frac{\prod_{C\in\mathscr{C}}\Gamma_{|C|}(\frac{n+|C|-1}{2})(2\pi)^{-n|C|/2}det(\frac{1}{2}\mathbf{S}_C)^{-\frac{n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}\Gamma_{|S|}(\frac{n+|S|-1}{2})(2\pi)^{-n|S|/2}det(\frac{1}{2}\mathbf{S}_S)^{-\frac{n+|C|-1}{2}}} \\
&= (2\pi)^{-nq/2}\frac{\prod_{C\in\mathscr{C}}\Gamma_{|C|}(\frac{n+|C|-1}{2})det(\frac{1}{2}\mathbf{S}_C)^{-\frac{n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}}\Gamma_{|S|}(\frac{n+|S|-1}{2})det(\frac{1}{2}\mathbf{S}_S)^{-\frac{n+|C|-1}{2}}}.
\end{aligned}
$$

A.0.15 Marginal Likelihood of Equation 4.6

The marginal likelihood of $\mathbf{Y}^*$ under a graph $G \in \mathscr{G}$ will be provided by

$$
m^N(\mathbf{Y}^*|\delta, G) = \frac{f(\mathbf{Y}^*|\mathbf{\Sigma}, \delta, G)\pi^N(\mathbf{\Sigma}|G)}{\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*, \delta, G)}
$$

where $\pi^N(\mathbf{\Sigma}|\mathbf{Y}^*, \delta, G)$ is the posterior distribution of $\mathbf{\Sigma}$ given the imaginary data matrix $\mathbf{Y}^*$ under model G. Under each clique $C \in \mathscr{C}$ (and separator $S \in \mathscr{S}$) we have that

$$
\begin{aligned}
&\frac{f(\mathbf{Y}_C^*|\mathbf{\Sigma}_C, \delta, G)\pi^N(\mathbf{\Sigma}_C|G)}{\pi^N(\mathbf{\Sigma}_C|\mathbf{Y}_C^*, \delta, G)} = \\
&= \frac{det(\delta\mathbf{\Sigma}_C)^{-m/2}}{(2\pi)^{m|C|/2}}exp\Big\{-\frac{1}{2}tr((\delta\mathbf{\Sigma}_C)^{-1}\mathbf{S}_C^*)\Big\}det(\mathbf{\Sigma}_C)^{-|C|}\times \\
&\times \frac{2^{\frac{m+|C|-1}{2}|C|}\Gamma_{|C|}(\frac{m+|C|-1}{2})}{det(\delta^{-1}\mathbf{S}_C^*)^{\frac{m+|C|-1}{2}}det(\mathbf{\Sigma}_C)^{-(m/2+|C|)}exp\Big\{\mathbf{\Sigma}_C^{-1}\delta^{-1}\mathbf{S}_C^*\Big\}} \\
&= \frac{\delta^{-m|C|/2}2^{\frac{m+|C|-1}{2}|C|}\Gamma_{|C|}(\frac{m+|C|-1}{2})}{(2\pi)^{m|C|/2}det(\frac{1}{\delta}\mathbf{S}_C^*)^{\frac{m+|C|-1}{2}}} \\
&= (2\pi\delta)^{-m|C|/2}\frac{\Gamma_{|C|}(\frac{m+|C|-1}{2})}{det(\frac{1}{2\delta}\mathbf{S}_C^*)^{\frac{m+|C|-1}{2}}}.
\end{aligned}
$$

Thus, the marginal of $\mathbf{Y}^*$ under any model $G \in \mathscr{G}$ will be provided by

$$m^N(\mathbf{Y}^*|G) = (2\pi\delta)^{-mq/2} \frac{\prod_{C\in\mathscr{C}} \Gamma_{|C|}(\frac{m+|C|-1}{2}) det(\frac{1}{2\delta}\mathbf{S}^*_C)^{-\frac{m+|C|-1}{2}}}{\prod_{S\in\mathscr{S}} \Gamma_{|S|}(\frac{m+|S|-1}{2}) det(\frac{1}{2\delta}\mathbf{S}^*_S)^{-\frac{m+|S|-1}{2}}}$$

and we will write

$$m^N(\mathbf{Y}^*|G) \propto (2\pi\delta)^{-mq/2} \frac{\prod_{C\in\mathscr{C}} \Gamma_{|C|}(\frac{m+|C|-1}{2}) det(\frac{1}{2\delta}\mathbf{S}^*_C)^{-\frac{m+|C|-1}{2}}}{\prod_{S\in\mathscr{S}} \Gamma_{|S|}(\frac{m+|S|-1}{2}) det(\frac{1}{2\delta}\mathbf{S}^*_S)^{-\frac{m+|S|-1}{2}}}$$

due to the arbitrary normalizing constant that arises from the use of the improper default prior $\pi^N(\mathbf{\Sigma}|G)$.